What is Mineral Informatics?

Anirudh Prabhu[1,1], Shaunna M. Morrison[1,1], Peter Fox[2,2], Xiaogang Ma[3,3], Michael L. Wong[1,1], Jason Williams[1,1], Kenneth N. McGuinness[4,4], Sergey Krivovichev[5,5], Kerstin Annette Lehnert[6,6], Jolyon Paul Ralph[7,7], Barbara Lafuente[8,8], Robert Downs[9,9], Michael Walter[1,1], and Robert M Hazen[10,10]

[1]Carnegie Institution for Science
[2]Rensselaer Polytechnic Institute
[3]University of Idaho
[4]Rutgers University
[5]Russian Academy of Sciences
[6]Columbia University
[7]Mindat.org
[8]SETI
[9]University of Arizona
[10]Carnegie Institute for Science

November 30, 2022

Abstract

Minerals are information-rich materials that offer researchers a glimpse into the evolution of planetary bodies. Thus it is important to extract, analyze, and interpret this abundance of information in order to improve our understanding of the planetary bodies in our solar system and the role our planet's geosphere played in the origin and evolution of life. Over the past decades, data-driven efforts in mineralogy have seen a gradual increase. The development and application of data science and analytics methods to mineralogy, while extremely promising, has also been somewhat ad-hoc in nature. In order to systematize and synthesize the direction of these efforts, we introduce the concept of "Mineral Informatics". Mineral Informatics is the next frontier for researchers working with mineral data. In this paper, we present our vision for Mineral Informatics, the X-Informatics underpinnings that led to its conception, the needs, challenges, opportunities, and future directions. The intention of this paper is not to create a new specific field or a sub-field as a separate silo, but to document the needs of researchers studying minerals in various contexts and fields of study, to demonstrate how the systemization and increased access to mineralogical data will increase cross- and interdisciplinary studies, and how data science and informatics methods are a key next step in integrative mineralogical studies.

What is Mineral Informatics?

Authors: Anirudh Prabhu[1], Shaunna Morrison[1], Peter Fox[2], Xiaogang Ma[3], Michael L. Wong[1], Jason Williams[1], Kenneth N. McGuinness[4], Sergey Krivovichev[5], Kerstin Lehnert[6], Jolyon Ralph[7], Barbara Lafuente[8], Robert T. Downs[9], Michael Walter[1], Robert Hazen[1]

[1]Earth and Planets Laboratory, Carnegie Institution for Science, 5241 Broad Branch Rd NW, Washington, DC 20015.

[2]Tetherless World Constellation, Rensselaer Polytechnic Institute, 110 8th St, Troy, NY 12180.

[3]Department of Computer Science, University of Idaho, 875 Perimeter Dr, Moscow, ID 83844.

[4]Department of Biochemistry and Molecular Biology, Rutgers University, 57 US Highway 1. New Brunswick, NJ 08901-8554.

[5]Kola Science Centre of the Russian Academy of Sciences, Leninskiy Prospekt, 14, Moscow, Russia, 119991.

[6]Lamont-Doherty Earth Observatory, Columbia University, 61 Rte 9W, Palisades, NY 10964.

[7]Mindat.org, 1113 Cambridge Hill Lane, Keswick, VA 22947-2749.

[8]Carl Sagan Center, SETI Institute, 189 Bernardo Ave, Suite 200, Mountain View, CA 94043.

[9]Department of Geosciences, University of Arizona, 1040 E 4th St, Tucson, AZ 85721.

Abstract: Minerals are information-rich materials that offer researchers a glimpse into the evolution of planetary bodies. Thus, it is important to extract, analyze, and interpret this abundance of information in order to improve our understanding of the planetary bodies in our

22   solar system and the role our planet's geosphere played in the origin and evolution of life. Over

23   the past decades, data-driven efforts in mineralogy have seen a gradual increase. The

24   development and application of data science and analytics methods to mineralogy, while

25   extremely promising, has also been somewhat ad-hoc in nature. In order to systematize and

26   synthesize the direction of these efforts, we introduce the concept of "Mineral Informatics".

27   Mineral Informatics is the next frontier for researchers working with mineral data. In this paper,

28   we present our vision for Mineral Informatics, the X-Informatics underpinnings that led to its

29   conception, the needs, challenges, opportunities, and future directions. The intention of this

30   paper is not to create a new specific field or a sub-field as a separate silo, but to document the

31   needs of researchers studying minerals in various contexts and fields of study, to demonstrate

32   how the systemization and increased access to mineralogical data will increase cross- and

33   interdisciplinary studies, and how data science and informatics methods are a key next step in

34   integrative mineralogical studies.

35

36   Keywords: Minerals; X-Informatics; Data; Data Science; Information; Scientific Discovery.

37   Word Count: 15714

38

39

40   1. Introduction

41

42   The potential for data-driven methods to make novel, unintuitive, and groundbreaking

43   discoveries in Earth and planetary science research will only grow as the volume and

2

44   variety of data increases with time. Mineralogy, in particular, is ripe for the application of

45   data-driven methods. Minerals form as a result of their unique chemical and physical

46   conditions and, in the process, retain information regarding their formation information

47   that offers an opportunity to study the complex geologic and biologic past of these

48   planetary bodies (Prabhu et al. 2021a).

49

50   Mineralogy has been the subject of scientific curiosity and study for millennia (Needham

51   1986; Bandy & Bandy 1955). In addition to their roles as captivating specimens for

52   collection and study, minerals and their ores are essential in the survival and

53   industrialization of humankind (Murray 1995; Coates 1985). This interest and utility has

54   led to the characterization and systemization of mineralogy and mineral occurrence on

55   Earth and other planetary bodies (Dana & Dana 1895; Hazen & Morrison 2021; Bragg and

56   Bragg 1913; Strunz & Tennyson 1941; Lafuente et al. 2015; Lehnert et al. 2000). As a result

57   of this rich history of scientific investigation, vast amounts of information are available on

58   the occurrence and attributes of minerals. These data provide a robust platform for the

59   analysis of more complex, multidimensional, and larger mineralogical systems; the

60   integration of heterogeneous data types, linking to data from other fields of science; and

61   predictive, data-driven scientific exploration - all of which leads to the answering of

62   complex, multidisciplinary questions. The potential of data-driven mineralogical research

63   has been exemplified by important scientific advances in the last decade. Recent

64   discoveries have demonstrated periodicity of mineral formation and diversification

65   associated with supercontinent assemble (Bradley 2011; Voice et al. 2011; Nance et al.

66   2014; Hazen et al. 2014), an association of mineral redox state to the oxidation of Earth's

67   atmosphere (Liu et al. 2021; Hummer et al. 2022; Large et al. 2022), and that much of

68   Earth's mineral inventory is the direct or indirect result of interactions with water and/or

69   biology (Hazen & Morrison 2021, 2022), as well as the prediction of the number of as-yet

70   undiscovered mineral species (e.g., Hazen et al. 2015, Hystad et al. 2019; Hystad et al.

71   2015), the chemical composition of minerals on Mars (Morrison et al. 2018 a-c), and the

72   location of undiscovered mineral deposits (Prabhu et al. 2019; Morrison et al 2022 (in

73   prep)). Mineralogy is rapidly entering the data-driven era, tackling previously

74   unanswerable questions, and demonstrates the need and opportunity for a symbiotic

75   relationship between the mineralogy and the fields of data science and informatics.

76

77   Data-driven efforts in mineralogy have been gradually increasing in the past decades and

78   there are some promising studies that have helped researchers uncover the patterns

79   hidden in the data and have led to scientific discoveries (Morrison et al 2017; Hazen et al.

80   2019; Prabhu et al. 2019; Morrison et al 2020; Hazen & Morrison 2020, 2022; Zhao et al.

81   2020; Gregory et al. 2019; Boujibar et al. 2021; Hystad et al. 2021). While still nascent,

82   application of data science and data analytics methods in mineralogy shows a promising

83   trajectory. The development of these methods and advances in the past have been

84   somewhat ad-hoc in nature. However, development of mineral informatics can be guided

85   in a more deliberate and systematic way by taking into account the underpinnings from

86   information theory and data science advances, as exemplified by collaborations in other

87   fields, including biology, medicine, chemistry, and astronomy. We believe this is the start

88   of a new era in mineralogy, where utilizing data-driven methods to answer mineralogical

89   (and broader scientific) questions takes center stage.

90

91   In this paper, we take a high-level look at our vision for "Mineral Informatics", the

92   underpinnings that led to its conception, the needs, challenges and opportunities for this

93   emerging field of mineral informatics. We also discuss the implications such advances will

94   have on the field of mineralogy.

95

96

97   2. Foundation provided by X-informatics

98

99   Informatics studies the structure, algorithms, behavior, and interactions of natural and

100   artificial systems that store, process, access and communicate information (Fox 2011).

101   The term informatics has often been used in conjunction with the name of a

102   domain/discipline, for example, Bioinformatics, Geoinformatics, Astroinformatics, and

103   Cheminformatics. In the past, researchers with expertise in a specific domain worked on

104   processing and engineering information systems designed for that domain only. But in

105   the last decade, informatics has gained a much wider visibility across a range of disciplines

106   (Prabhu 2018). This wider visibility is in large part due to successful efforts at

107   systematizing the core (i.e., discipline neutral) aspects of informatics, for example, use-

108   cases, human-centered design, iterative approaches, information models etc. (Fox 2020).

109   The core methods of informatics are used as a foundation to explore raw data and extract

110   information from the data that lead to scientific discoveries. As the volume and

111   complexity of the data increase, so does the need for utilizing the solid foundations

112   provided by informatics methods and combining them with needs of the specific domain

113   to pursue data-driven scientific discoveries.

114

115   Mineral informatics is a nascent approach compared to fields like Bioinformatics, Medical

116   Informatics, and Geoinformatics that have been pursued for decades (Collen 1986; Sinha

117   2006; Fox et al. 2006; Gauthier 2019). The intention of this paper is not to create a new

118   specific field or a sub-field as a separate silo, but to think of and document the needs of

119   researchers studying minerals in various contexts and how data science and informatics

120   methods are a key next step in mineralogical studies. We also need to learn from the

121   successes and failures of more mature domains that have applied the informatics

122   approach. Lastly, a very important factor to keep in mind is the truly interdisciplinary and

123   important questions that can be explored by studying minerals. So, while the term

124   "mineral informatics" may seem like creating a new subclass of geoinformatics, we assert

125   that we are instead tying in various disciplines that use minerals as a key part of the

126   pursuit for answers to big science questions.

127

128   3. A methodology for mineral informatics explorations

129

130   In this paper, we present a general methodology for mineral informatics (see Figure 1).

131   This methodology, adapted from Fox & McGuinness (2008), includes all the steps typically

132 followed in a data-driven scientific exploration. This approach was created for mineral

133 informatics but, as is the case with many data science and informatics approaches, is

134 transferable and applicable to other domains.

135 Most informatics explorations start in one of two ways: 1) Scientists have a research

136 question they want to answer, or 2) scientists have data ready to be explored. In the

137 second case, we perform preliminary data exploration, which helps generate new

138 hypotheses and research questions based on interesting trends and anomalies in the

139 data.

140

141 Once a specific research question has been selected for scientific exploration, we start by

142 dividing the large problem into smaller more tractable parts. Next, we iteratively develop

143 use cases for every one of these parts. A "use case" is a documented collection of possible

144 sequences of actions and interactions between a system and its users in pursuit of a

145 particular goal. Identification and development of use cases helps to define the needs

146 (e.g., data, personnel, infrastructure) for this data-driven approach. The next steps in the

147 methodology includes creating (or assigning roles to an already established) an

148 interdisciplinary team to conduct the data-driven research.

149

150 Next, we inventory the preliminary dataset and/or existing mineral data resources (See

151 section 5a) to determine if they are what is necessary for the desired exploration. In some

152 cases, we need to collect, compile, and extract data from other repositories or sources

153 like scientific literature, websites, digital PDFs, and experimental results. We then create

154 an information model to better understand and mediate data from heterogeneous

155 sources and data types, which provides a holistic picture of the relationships between the

156 various data sources, types, and attributes. The information model allows us to extract

157 the datasets and data attributes most relevant to answering the desired research

158 question. Note that this step differs from the statistical and machine learning approaches

159 used for feature selection.

160

161 We then begin applying data analytics methods (i.e., data visualization as well as

162 descriptive, predictive, and prescriptive analysis) to identify and explore patterns and

163 anomalies seen in the data. A team of domain and data scientists iteratively examine the

164 results of the analytics methods and use their respective expertise to (1) provide

165 interpretations and/or insight, and/or (2) recommend changes to the analysis. The data

166 analysis and scientific interpretation are usually done over multiple iterations with small

167 modifications to the approach, algorithms, and/or code to explore different aspects of

168 the data.

169

170 If scientists come to an agreement that parts of the analysis would be widely used in the

171 larger community, they can choose to generalize and adapt their work into a system,

172 technology, or infrastructure. This development can include creation of tools, code

173 snippets, reusable workflows, R packages, Python libraries, and other resources.

174 Irrespective of whether there is a decision to create a general tool, technology, or

175   package, we recommended using rapid prototyping coding practices[1] (Gordon & Bieman

176   1995) for data science and informatics activities.

177

178   After obtaining the desired results from our data analysis, it is important to disseminate

179   and effectively communicate the research products generated by mineral informatics

180   explorations. Research products can include datasets, code, scientific literature,

181   executable workflows, etc. Establishing best practices for disseminating research

182   products is an ongoing effort, especially in the geoscience community. Datasets can be

183   published as part of a data paper, or be assigned their own DOIs by data repositories like

184   Zenodo, Dryad, Figshare, or Dataverse (Assante et al. 2016). Existing mineral data

185   repositories like the EarthChem Library (ECL), Astromat, and the Open Data Repository

186   (ODR) (See section 5a) also provide DOIs for datasets deposited by researchers.

187   Additionally, some journals host data associated with their publications. Similar to

188   releasing data used in scientific exploration, code can be maintained and released in many

189   ways, including Github (with a persistent identifier pointing to the repository), figshare,

190   or Zenodo. Saving executable code for an experiment in an interactive environment like

191   Jupyter or R notebooks adds to the reproducibility of the code and of the scientific

192   workflow in general (Prabhu & Fox 2021). Dissemination of scientific advances through

193   scientific publications has been practiced for more than 300 years (Fyfe et al. 2015). In

194   addition to journal publications, conference proceedings, preprint servers (such as arXiv,

---

[1] https://medium.com/convergeml/rapid-prototyping-in-machine-learning-ffe79f023aec

195   ESSOAr, and EarthArXiv), and even press releases associated with publications have

196   considerably improved the landscape of disseminating research products.

197

198   The final stage of our informatics methodology follows the sharing of the research

199   products. If researchers follow FAIR and Open Science practices (Wilkinson et al. 2016;

200   Stall et al. 2019; Ramachandran et al. 2020) not only for the dissemination of their

201   scientific results, but also during the use case development, information modeling, and

202   analysis stages, then it becomes easier to evolve, improve, redesign, or adapt your work.

203   Ongoing research and recommendations on designing FAIR and Open scientific workflows

204   will help improve the methodology of data-driven exploration (e.g., Prabhu & Fox 2021;

205   Kluyver et al 2016; Sandve et al. 2013).

206

207   It is important to evaluate the outcomes at almost every stage of the informatics

208   methodology. The evaluation method or metric used at each stage will be significantly

209   different, but it is important to stop at the end of every stage and assess not only the

210   progress made, but also lessons learnt for future iterations in the same exploration or the

211   beginning of a different exploration. For example, a data collection/resource may be

212   evaluated based on a set of quality criteria (e.g., Prabhu et al. 2021a), but results from the

213   data analysis may need to use quantitative metrics to evaluate results from a descriptive,

214   prescriptive, or predictive model (e.g., Statnikov et al. 2008; Tomasev & Radovanovic

215   2016; Hossin & Sulaiman 2015; Zhou et al. 2021). Established evaluation methods exist

216   for each stage of the informatics methodology, and we recommend following those

217 established best practices and standards set by the scientific community. Issues found

218 during evaluation will need to be documented in the use case and thus improve the data-

219 driven exploration during the next iteration or redesign of the approach.

220

221

222

223

224
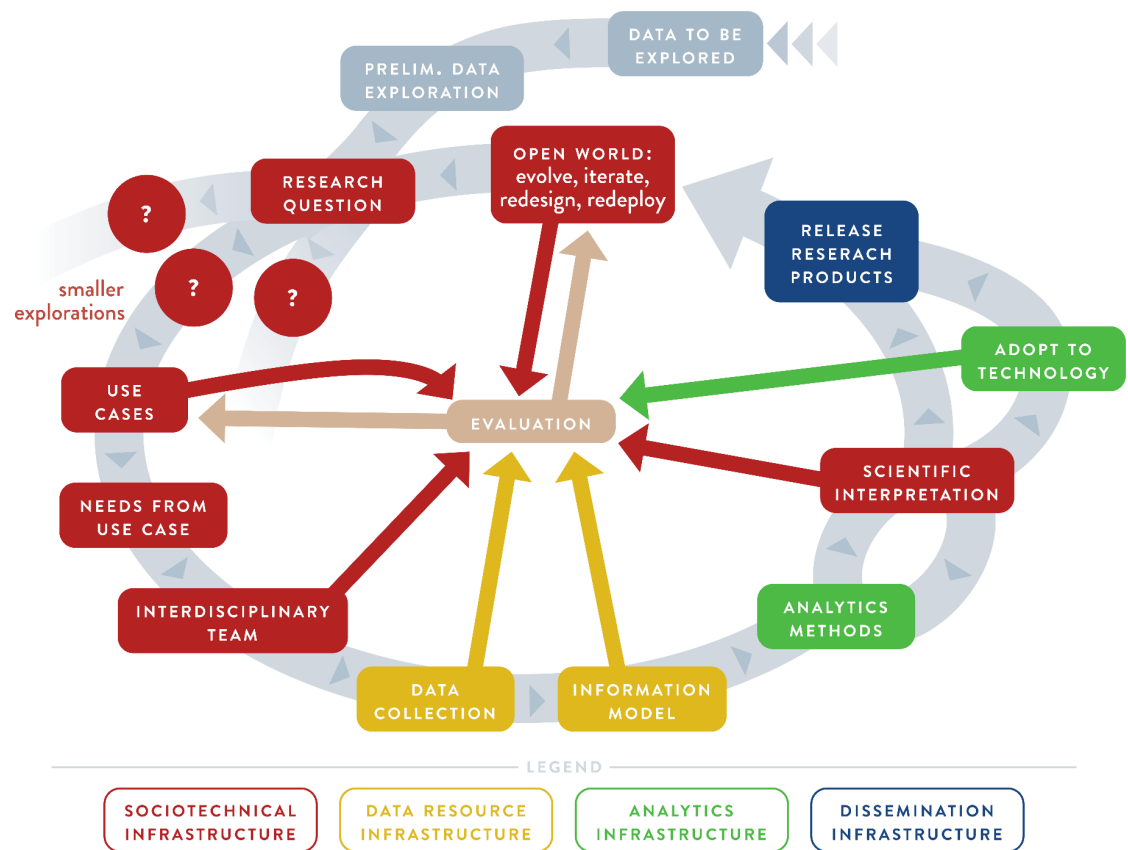
225 **[Figure 1: Mineral Informatics Methodology]**

226 **Caption: Mineral informatics methodology adapted from Semantic Web methodology**

227 **by Fox & McGuinness (2008).**

228

229   4. Challenges and opportunities in Mineral Informatics

230   Mineral informatics methods not only systematize the mineral data landscape, but also

231   provide a path to answering some of the big, interdisciplinary scientific questions. Figure

232   2 gives an example of the domains influenced by the research questions being broached
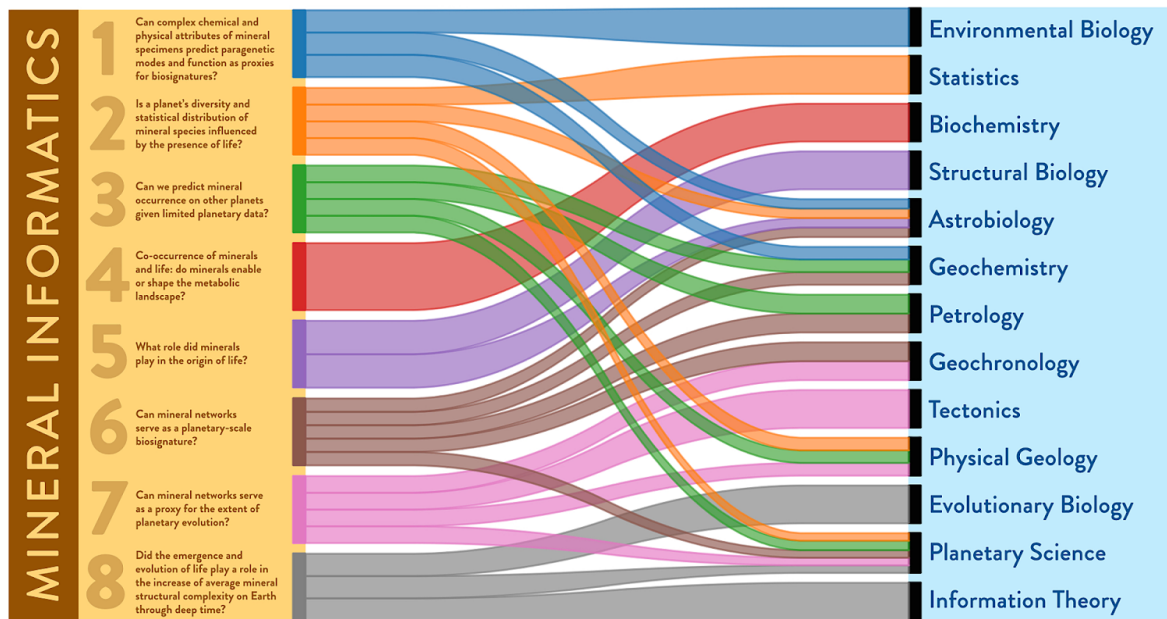
233   with mineral informatics methods.

234

235   Figure 2: Interdisciplinary research questions related where mineral informatics play a

236   key role.

237

238   a. Major scientific questions that can be addressed with mineral informatics

239   i. **Can complex chemical and physical attributes of mineral specimens**

240   **predict paragenetic modes and function as proxies for biosignatures?**

241   Minerals record the physical, chemical, and, in some cases, biological

242   conditions of their paragenetic modes (i.e., formational and alteration

243   environments). This information is stored in the myriad attributes of

244   mineral specimens, including major, minor, and trace elements, isotopic

245   ratios, texture, and grain size. Therefore, conditions of mineralization,

246   including whether or not there was biological input, can be characterized

247   with cluster analysis performed on the various properties of mineral

248   samples (Gregory et al. 2019). Furthermore, robust classification schemes

249   can be developed from the clustering models that will enable prediction

250   not only of the geologic environment of formation but also of any biogenic

251   origins (Hazen 2019). Therefore, this work will deconvolve our

252   understanding of the minerals that formed in environments influenced by

253   life from those that formed under strictly abiotic conditions.

254

255   **ii. Is a planet's diversity and statistical distribution of mineral species**

256   **influenced by the presence of life?**

257   Life creates unique niches of chemical disequilibrium for minerals to

258   exploit. These processes likely drove a significant fraction of the mineral

259   diversity we see on Earth today, influencing the spatial and temporal

260   patterns of mineral distribution (Hazen et al. 2018; Morrison et al. 2020;

261   Hazen & Morrison 2022). These trends on Earth and other planetary bodies

262   can be modeled, compared, and used to develop statistical biosignatures

263		and abiosignatures regularizing the diversity and distribution of mineral
264		species across a planetary body (Hystad et al. 2019) and provide models
265		for planetary-scale mineralogical biosignatures of inhabited worlds.
266		
267	iii.	**Can we predict mineral occurrence on other planets given limited**
268		**planetary data?**
269		From orbital infrared spectroscopy, we have obtained global or near-global
270		datasets of the mineralogy of other terrestrial worlds, including Mars,
271		Mercury, Vesta, and Ceres (Murchie et al. 2009; de Sanctis et al. 2012;
272		Ehlmann & Ewards 2014; Namur & Charlier 2017; Prettyman et al. 2019).
273		Informatics methods, such as association analysis, can be used to predict
274		the existence of minerals that cannot be detected from space. By
275		understanding mineral affinities for assemblages, localities, and
276		geochemical parameters, we may be able to use a sparse mineralogical
277		dataset to anticipate future discoveries (Prabhu et al. 2019), but first a
278		robust small/sparse-data framework must be developed (see Section 6a).
279		Enhancing predictive capabilities will help to prioritize landing sites for
280		future landers and rovers with broad science goals that relate to
281		mineralogy, like understanding planetary history or searching for signs of
282		life. Such predictions would be strategically important because
283		interplanetary missions cost hundreds of millions to billions of dollars and
284		take years to decades to develop, build, and launch.

285

286   We also have geochemical indicators of the mineralogy of the ice-covered

287   ocean world Enceladus from plume flybys and E-ring analyses performed

288   by the Cassini spacecraft (Postberg et al. 2008; Waite et al. 2017; Glein &

289   Waite 2020). Mineral informatics methods can help predict the mineral

290   composition of ice-covered ocean worlds, whose mineralogy is

291   planetologically and perhaps astrobiologically relevant but cannot be

292   accessed directly in the near future.

293

iv. Co-occurrence of minerals and life: do minerals enable or shape the

metabolic landscape?

296   Minerals play a key role in biological redox transformations.

297   Microorganisms, (e.g., of the genus *Geobacter*) are able to use metals in

298   their environment to power their metabolisms (Childers et al. 2002).

299   Several studies have suggested deep similarities between minerals and

300   metalloenzymes (Nitschke et al. 2013; Zhao et al. 2020; McGuinness et al.

301   2022 (In Review)). Thus, minerals may play an important role in shaping

302   the metabolic landscape of ecosystems by providing electron

303   donors/acceptors or raw materials (Novikov & Copley 2013) that

304   organisms assimilate to create metalloenzymes. Mineral informatics

305   methods may be able to elucidate connections between minerals and

306   biology. If minerals are found to be critical in shaping which metabolisms

307	occur/do not occur in certain environments, this may allow for the
308	prediction of metabolisms in terrestrial and extraterrestrial environments
309	for which we have mineralogical data.
310	
311	**v. What role did minerals play in the origin of life?**
312	Several studies have posited that minerals played a critical role at the
313	emergence of life on Earth, whether by influencing the homochirality of
314	organic molecules or performing redox transformations and carbon
315	fixation (Hazen & Scholl 2003; Hazen 2005; Hazen & Sverjensky 2010;
316	Nitschke et al. 2013; Russell et al. 2018). Others have suggested that clays
317	and other minerals with layered structures may have been the first self-
318	replicating entities (Cairns-Smith & Hartmann 1986; Cairns-Smith 1990;
319	Greenwell & Coveney, 2006; Brack 2013), though these hypotheses have
320	not been confirmed experimentally (Bullard et al., 2007; Krivovichev et al.,
321	2012). Mineral informatics, combined with phylogenetics, geology, and
322	laboratory experiments, could be informative for deducing the likely
323	role(s) that minerals played at the origin of life in Earth's deep past. If
324	certain minerals are found to be uniquely critical to the emergence of life
325	on Earth, this would have profound implications for the emergence of life
326	on other planetary bodies where those minerals may or may not occur.
327	The origin of life from a non-living substance involves considerable jump in
328	the informational (static) complexity of the underlying molecular

329	structures, which should be taken into account in any possible scenario of
330	molecular (r)evolution that led to the appearance of self-replicating living
331	entities. The sudden rise in structural complexity corresponds to the drop
332	in configurational entropy (Krivovichev, 2016). Can the (local) entropic
333	changes associated with the origin of life be measured quantitatively and
334	understood using mineral informatics data?
335	
336	**vi. Can mineral networks serve as a planetary-scale biosignature?**
337	Roughly half of all known minerals are mediated by biology and 34% are
338	exclusively biotic (Hazen & Morrison 2022; Hazen et al. 2021a; Morrison et
339	al. 2021). Many of these minerals are formed when life opens up a new
340	compositional space for the planet, such as the Great Oxidation Event
341	(Hazen et al. 2008; Sverjensky & Lee 2010). However, some of this biogenic
342	chemical space may be abiotically accessed on other worlds. Abundant
343	atmospheric O_2, for instance, may be abiotically generated by various star–
344	planet interactions (Meadows et al. 2018 and references therein). Earth
345	and planetary mineral network analysis may reveal whether mineral
346	networks of environmental, biological, geochemical, and mineralogical
347	attributes can distinguish living from nonliving worlds.
348	
349	**vii. Can mineral networks serve as a proxy for the extent of planetary**
350	**evolution?**

351   Mineralogical evolution occurs when processes create new pressure–

352   temperature–compositional regimes where solids can form (Hazen et al.

353   2008; Cleland et al. 2021; Hazen et al. 2021a; Hazen & Morrison 2020).

354   Each stage of mineral evolution expands the network of mineralogy

355   through the introduction of new minerals, localities, and paragenetic

356   modes. The network of martian mineralogy, therefore, is thought to be a

357   subset of the network of Earth's mineralogy, due to the halting or slowing

358   of mineral-generating geological processes on Mars. One can consider

359   Mars and Earth to be two points along a spectrum of terrestrial worlds

360   whose geological (and biological) activities have differed in temporal

361   extent. A hypothetical world where plate tectonics was sustained for ~1

362   Gyr but then ceased should have a mineral network that surpasses Mars's

363   mineral diversity, but is still a subset of Earth's. In this way, mineral

364   informatics helps us interpret the extent of a planet's mineralogical

365   network as a record of ancient and extinct processes, revealing a planet's

366   geological history.

367

368   When considering exoplanetary systems where element ratios (e.g., C:O or

369   Mg:Si) differ greatly from those of our own solar system, this linear

370   spectrum on which Mars and Earth lie becomes a multidimensional phase

371   space (Unterborn et al. 2016; Unterborn and Panero 2019; Hinkel and

372   Unterborn 2018; Putrika et al. 2021). Understanding mineral networks

373	from an informatics point of view may help to predict how planetary
374	mineralogy might evolve in vastly different geochemical contexts.
375	
376	**viii. Did the emergence and evolution of life play a role in the increase of**
377	**average mineral structural complexity on Earth through deep time?**
378	It has been shown that complexity of Earth's mineral kingdom increased
379	gradually during planetary evolution (Krivovichev et al., 2018), but it is
380	unclear whether this trend is related to the increase in complexity in the
381	course of biological evolution. The average structural complexity of
382	minerals on the abiotic moon, for example, does not follow the same trend
383	of increasing complexity through time. Minerals are relatively less complex
384	than biological organisms, both in terms of their static (Krivovichev 2013,
385	2015) and functional (Hazen et al., 2007) complexities. However, Since life
386	and the mineral kingdom co-evolved, the character of the evolution of
387	mineral complexity on Earth (Krivovichev et al., 2018) may have been
388	influenced by biological activity, and is thereby a potential bio-signature.
389	
390	b. Successful Use Cases in Mineral Informatics
391	**i. The evolution of mineralizing environments, as characterized by their**
392	**myriad, complex attributes**
393	Mineralization, and associated formational environments, vary
394	significantly across Earth and neighboring planetary bodies, as well as

395 throughout the different historical stages of planetary evolution. These

396 stages and environmental parameters dictate the types of mineralization

397 that occur and, likewise, leave their mark in the complex chemical and

398 physical attributes of the resulting mineral specimens. Understanding the

399 changing characteristics of mineralizing environments spatially and

400 temporally across our planetary systems requires the examination of huge

401 volumes of mineralogical information. The beginning steps of this work

402 included a survey of all formational environments of ~5700 known mineral

403 species, resulting in a compiled dataset ripe for exploration (Hazen and

404 Morrison 2022; Hazen et al., 2022). Initial exploration has led to the

405 discovery that (1) 80% of all mineral species formed through processes that

406 involved water; (2) 50% of minerals formed through processes directly or

407 indirectly related to biology, with 34% of minerals forming exclusively

408 through biotic processes; (3) 42% of minerals contain one or more rare

409 elements (e.g., REE, PGE, As, Mo, Sn), elements which all together

410 represent only 0.01% of crustal atoms; and (4) most minerals have only

411 one (59%) or two (24%) modes of formation, with a few notable

412 exceptions, including pyrite with the most modes of formation at 21

413 (Hazen and Morrison 2022).

414

415 An additional component of this work involves analyzing those myriad

416 attributes of mineral specimens via cluster analysis to relate their complex

417   characteristics to their modes of formation, thereby determining the

418   natural kind clustering of these mineral systems. There are many such

419   projects underway, including those examining the formation of pyrite

420   (Gregory et al. 2019; Zhang et al., 2019), garnet minerals (Chiama et al.,

421   2020; 2022a; 2022b(in prep)), spinel oxide phases (Hindrichs et al., 2022),

422   and presolar SiC (Boujibar et al., 2021; Hystad et al. 2021). Boujibar et al.

423   (2021) performed cluster analysis on a range of isotopic data from presolar

424   SiC grains in order to examine and compare the origins of these materials.

425   This study made several exciting discoveries - while the clustering model

426   agreed with previously defined grain types and origins in several aspects,

427   there were notable and important deviations, including: a division of one

428   grain type into three distinct types based on varying metallicity of the

429   parent star, the arbitrary nature of certain divisions in systems that are

430   continuous rather than discrete, the observation that asymptotic giant

431   branch (AGB) stars with narrow ranges of mass and metallicity tend to have

432   enhanced production of SiC, and enrichment in ^{15}N and ^{26}Al that is not

433   explained by existing AGB models.

434

435   Next steps: This exploration of mineralizing environments and their

436   characteristics not only provides an opportunity to integrate data from

437   heterogeneous sources and types (e.g., X-ray diffraction, electron

438   microprobe analysis, inductively coupled plasma mass spectrometry), but

439	also to link data from different fields of science to better understand
440	mineral paragenesis. Handling heterogeneous data is a challenge
441	(Reichman et al. 2011; Wang 2017) and many researchers have been
442	actively working on using heterogeneous data for their analysis by creating
443	methods, approaches, and pipelines to seamlessly clean, integrate,
444	process, and analyze data (Wang 2017; Zhang et al. 2018; Beneventano &
445	Bergamaschi 2004; Wiederhold 1999; Nazabal 2020). Additionally, the
446	exploration conducted by Boujibar et al. (2021), provided another use case
447	to test out machine learning methods on sparse data sets, thereby aiding
448	in the eventual development of a sparse data framework (see Section 5a
449	for more details).
450	
451	**ii. Mineral association analysis**
452	Prediction of the locations of as yet undiscovered mineral deposits has
453	long been a point of great scientific and economic interest. Mineralization
454	and mineral co-occurrence across the varied geologic terrains of Earth and
455	other planetary bodies has a level of complexity that makes prediction of
456	mineral locations, or even the mineral inventory at a locality of interest,
457	difficult. However, recent advances in the mineral locality data resources
458	(e.g., mindat.org and the Mineral Evolution Database) have provided an
459	opportunity to begin tackling this tough problem with machine learning.
460	Association analysis can be used to create a recommender system (Burke

461   et al. 2011; Shah et al. 2017) that generates association rules based on

462   known co-occurrences and these rules can be queried to determine the

463   likelihood of currently unknown co-occurrences. In the case of minerals,

464   we can query our mineral association rules to predict: A) previously

465   unknown locations of a mineral species, B) previously unknown locations

466   of mineral assemblages, including those that represent analog

467   environments for study, and C) the mineral inventory at a locality of

468   scientific interest. The mindat.org team have conducted preliminary

469   explorations using pairwise associations to predict the occurrence of

470   certain minerals on Earth.

471

472   Next Steps: Mineral association analysis provides new types of data

473   problems. We need to modify the association analysis algorithms to better

474   handle larger mineral occurrence datasets. For example, our models can

475   currently handle only 2,473 minerals occurring in 87,306 localities (Prabhu

476   et al. 2019; Morrison et al. 2022 (in prep)), but there are at present ~5760

477   mineral species in the International Mineralogical Association's (IMA) list

478   of approved mineral species (https://rruff.info/ima/), which occur in more

479   than 375,000 localities (https://www.mindat.org/stats.php). In addition to

480   improving the scalability of association analysis methods, we also need to

481   work on the dimensionality and reducing the minimum support of our

482   method. For example, our method currently develops rules containing 4

483	minerals at a time, but there are localities with more than 50 coexisting
484	minerals. Therefore, an important next step in our research is to increase
485	the dimensionality of the association analysis method to handle more
486	complex mineral assemblages. We also need to adapt our methods to
487	enable inclusion of rarer mineral species that are known to occur in 17 or
488	fewer localities (Prabhu et al. 2019). Lastly, we are currently developing a
489	new approach to evaluate association rule mining methods (Prabhu et al.
490	2021b).
491	
492	**iii. Martian crystal chemistry**
493	The scientific payload onboard the NASA Mars Science Laboratory (MSL)
494	rover, *Curiosity*, is the one of the most advanced instrument suites ever
495	landed on another planet. Part of this payload is the CheMin X-ray
496	diffraction (XRD) instrument, which is used to characterize the mineralogy
497	of rock and soil samples. CheMin is capable of identifying mineral phases
498	present in samples, as well as their abundances and, for phases with an
499	abundance $\geq$ 1–3 wt %, their unit-cell parameters. While there are
500	instruments that analyze the bulk composition of martian samples, there
501	is no instrument that directly measures the chemical composition of these
502	mineral phases. However, in compiling data resources on mineral unit-cell
503	parameters and compositions measured on Earth, the CheMin XRD
504	patterns and resulting mineralogical data are used to predict the

24

505	composition of the mineral phases observed on the martian surface
506	(Morrison et al. 2018a-b).
507	
508	These initial studies, as those predating it, used unit-cell parameters to
509	predict mineral composition in chemically limited systems, generally 2- or
510	3-element systems such as Fe-Mg olivine or Mg-Fe-Ca pyroxene (Morrison
511	et al. 2018a-b). This limitation was due to the complexity of the
512	compositional and structural parameter space when for or more elements
513	are considered together. One way to develop a model that accounts for
514	the complexity associated with multi-component systems and predicts the
515	chemical composition of crystalline phases based on their crystallographic
516	parameters is by using Label Distribution Learning (LDL) (Geng et al. 2013,
517	2014; Geng 2016). LDL is a machine learning algorithm originally created
518	for facial recognition applications. When the approach was adapted for
519	application to crystallographic and chemical parameters, it resulted in a
520	model that accurately predicted the multi-component chemical
521	compositions (up to 12 elements, in some mineral systems) of samples
522	based solely on their unit-cell parameters (Morrison et al. 2018c). This
523	crystal-chemical method has expanded the capability of XRD on spacecraft
524	to that of a powerful chemical analysis tool, such as an electron
525	microprobe, and has dramatically deepened our understanding of the
526	geologic history of Mars.

527

528   Next steps: This exploration was the initial inspiration that motivated us to

529   create a framework for small and sparse data (See section 6.a. for more

530   details). In addition to our work developing a framework for small and

531   sparse data, we will also need to develop methods to evaluate the

532   accuracy of predictions made by our data models. This evaluation will

533   attempt to address sources of uncertainty and how that affects our

534   predictions. The LDL evaluation method being developed will address

535   uncertainty of measurement (instrument errors), uncertainty from

536   sampling (various sampling strategies to train predictive models), and

537   most interestingly, scope compliance (Klas 2018) of the LDL method.

538

539   **iv. Machine Learning Majorite Barometer**

540   Diamond-hosted majoritic garnet inclusions provide important insights in

541   processes that occur in Earth's deep mantle. Majoritic garnets provide the

542   most accurate estimates for diamond formation pressures because

543   laboratory experiments have shown that garnet chemistry varies as a

544   function of pressure (Thomson et al. 2021; Akaogi & Akimoto 1977; Irifune

545   1987; Collerson et al., 2010; Wijbrans et al., 2016; Beyer et al., 2017).

546   Thomson et al. (2021) show that none of the available barometers in the

547   literature reliably reproduce the pressures of experimentally synthesized

548   majoritic garnet over the entire pressure-temperature-composition space

549   investigated. Hence, they developed a barometer built using machine

550   learning algorithms (specifically random forest regression) and

551   experimental training data. This machine learning approach, tested with

552   various cross-validation methods, produces a barometer with a much

553   improved fit to the experimental data, especially at the highest pressures

554   and at extremes of composition space, and thus provides more reliable

555   estimates of formation pressures of diamond-hosted majoritic inclusions.

556   Applying the machine learning barometer to the global database of

557   diamond-hosted inclusions reveals that their formation occurs over

558   specific depth intervals that can be related to melting and decarbonation

559   of subducted oceanic crust.

560

561   Next Steps: While the machine learning approach improved the fit to the

562   available experimental data, it also revealed regions in pressure,

563   temperature, and most critically, composition space where the

564   experimental data set is sparse. Because many of the mineral inclusions

565   have compositions lying near or within sparse data regions, uncertainty

566   remains as to whether the barometer is accurately capturing their pressure

567   (and depth) of origin. Experiments can now be targeted to these specific

568   P-T-X regimes for an even more improved barometer. Machine learning

569   methods also can be used to predict the compositional variables that

570   correlate most strongly with changes in pressure, leading to an improved

571   crystal chemical and thermodynamic understanding of pressure-sensitive

572   substitutions in garnet. These methods can also be applied to other

573   mineral thermometers and barometers where large experimental datasets

574   are fitted to extract thermodynamic solution parameters.

575

576   **v. Comparison of mineral and protein metal clusters**

577   Understanding the evolutionary stages of biology on a geological timescale

578   is hampered by the propensity of organic matter to degrade within

579   thousands of years without leaving physical fossil records. To understand

580   how life evolved over the course of billions of years, proxy data are

581   required.

582

583   At least five observations suggest that minerals can act as a source of proxy

584   data from which to infer how biology evolved: 1) biology and geology are

585   intimately connected, for instance, cellular organisms excrete minerals as

586   metabolic end products (hazenite; Yang et al. 2011; greigite; Gorlas et al.

587   2018) and cellular organisms transmit electrons to and from minerals (Shi

588   et. al. 2016), 2) cellular organisms and minerals use transition metals (Fe,

589   Mn, Co, Mo, Cu, V, W, Ni) to perform electron transfer reactions, 3) mineral

590   surfaces are hypothesized and shown to be capable of prebiotic reactions

591   similar to those that extant proteins perform (Wachtershauser 1988,

592   Novikov & Copley 2013), 4) minerals are similar to the rings of a tree in that

593   they provide information (e.g., temperature, humidity, etc.) about the

594   environment of formation, and 5) metal cluster structures of extant

595   proteins were observed to be so similar to the structure of bulk mineral

596   metal clusters as to be considered vestiges of minerals that were co-opted

597   and assimilated into biological systems (Russell & Hall 1997, Nitschke et al.

598   2013, Zhao et al. 2020).

599

600   Access to large mineral and protein structure databases allows the

601   potential to understand how mineral and protein metal clusters are

602   connected. Connecting the mineral world with biology will allow a deeper

603   understanding of how geology and biology co-evolved. Directly quantifying

604   metal cluster similarity between minerals and proteins is a challenge due

605   to comparing the finite protein cluster to a periodic lattice of a mineral.

606   Solutions using graph-based methods have been proposed (Zhao et al.

607   2020; McGuinness et al. 2022 (in review)). Each solution compared

608   subgraphs of mineral and protein metal clusters, however without

609   including metal coordination, and mineral dimensionality (2D-layer vs 3D

610   lattice) metal clusters were quantified as being highly similar (Zhao et al.

611   2020). Subsequent studies, building off the quantitative pioneering work

612   of Zhao et al., included these chemically important characteristics and

613   found FeS minerals and protein were significantly less similar (McGuinness

614   et al. 2022 (in review)) than previously proposed (Russell & Hall 1997,

615	Nitschke et al. 2013) Even though McGuinness et al. 2022 show that FeS
616	mineral lattices and protein metal clusters are not structurally similar, this
617	method has not been applied to other metal types such as Ni or Cu.
618	Applying the method developed by McGuinness et al. 2022 to additional
619	metal types may help understand the extent to which proteins and
620	minerals co-evolved as cellular metabolism and minerals became more
621	complex (Moore et al. 2017; Krivovichev et al. 2018).
622	
623	Next Steps: An additional step towards a potentially more clear
624	understanding of how minerals and proteins are related is to compare
625	mineral surface and protein metal cluster structures. Mineral surfaces
626	expose the chemically active components that may have catalyzed
627	biologically relevant products under hydrothermal conditions on early
628	Earth (Novikov & Copley 2013). Comparing the surface properties of
629	minerals to the chemical properties of protein metal clusters might
630	elucidate the extent to which minerals acted as primitive enzymes at the
631	dawn of life. Did biology co-opt the chemical configuration of the
632	chemically active surface of minerals to reproduce the reactions that were
633	possible abiotically? Or did biology incorporate and reconfigure metal
634	building blocks (e.g., 2Fe2S) to meet growing cellular needs? Answering
635	these questions is challenging because mineral surfaces are complex,
636	subject to relaxation, are chemically active, display complexly irregular

637 surface topologies, and are affected by many solution conditions (pH,

638 salinity, temperature, etc.) Alternatively, there also exists the possibility

639 that protein metal clusters do not bear any significant resemblance to

640 minerals (neither surface, nor lattice structure), suggesting an alternative

641 pathway and relationship between mineralogy and biology in which

642 biology acts independently, only relying on minerals for the feedstock (i.e.,

643 metals) to nucleate the information-rich systems that remain far from

644 equilibrium.

645

646 5. Mineral Information Systems

647

648   a. A non-exhaustive list of open access mineral data resources

649   The mineral data resources we have chosen to highlight below are open and

650   among the most widely used in the community. There are many other useful and

651   important mineral data resources that are not yet available as open resources.

652

653   i. **International Mineralogical Association** (IMA) list of approved minerals

654   (https://rruff.info/ima/) - a searchable database of mineral species

655   information, including chemical formula, unit-cell parameters, paragenetic

656   modes, and links to other important mineralogical data resources (e.g.,

657   American Mineralogist Crystal Structure Database, mindat.org, Mineral

658   Evolution Database, Handbook of Mineralogy)

ii. **The RRUFF Project** (https://rruff.info/) - a mineral library and database of chemical, spectral, and diffraction data for mineral species (Lafuente et al. 2015).

iii. **The Mineral Evolution Database** (MED; https://rruff.info/Evolution/) - a database of mineral locality and age information, with ~200,000 species/locality/age records extracted primarily from scientific literature and mindat.org. (Golden et al. 2016; Golden 2019)

iv. **The American Mineralogist Crystal Structure Database** (AMCSD; http://rruff.geo.arizona.edu/AMS/amcsd.php) - a crystal structure database that includes every structure published in the *American Mineralogist*, *The Canadian Mineralogist*, *European Journal of Mineralogy*, and *Physics and Chemistry of Minerals*, as well as selected datasets from other journals.

v. **The Handbook of Mineralogy** - a five volume set with each of the 4988 pages dedicated to a mineral species description, with information such as crystallographic and physical attributes, microprobe chemical analyses, paragenetic mode and locality information, and select references.

vi. **Mindat.org** - the world's largest open database of minerals, rocks, meteorites and the localities from which they were found.[2]

---

[2] https://www.mindat.org/

678   vii. **Mineral Properties Database** (MPD; https://odr.io/MPD) - a database of

679   various mineral attributes including age, color, redox state, structural

680   complexity, and method of discovery.

681   viii. **Evolutionary System of Mineralogy Database** (ESMD;

682   https://odr.io/esmd) - a database containing measured geochemical and

683   physical characteristics of mineral samples, including major, minor, trace

684   elements as well as isotopic ratios. (Chiama et al. 2022a)

685   ix. **The CheMin Database** (https://odr.io/chemin) - a database containing the

686   X-ray diffraction data from martian rock and soil samples analyzed by the

687   CheMin instrument onboard the NASA Mars Science Laboratory.

688   x. **The Astromaterials Data System** (AstroMat; https://www.astromat.org/)

689   - a data infrastructure that stores, curates, and provides access to

690   laboratory data acquired on samples curated in the NASA Johnson Space

691   Center Astromaterials Collection, including the Apollo lunar samples and

692   the Antarctic meteorite collection (Lehnert et al. 2019).

693   xi. **EarthChem** (https://earthchem.org/) - a data system providing open data

694   services to the geochemical, petrological, mineralogical, and related

695   communities, including data preservation, discovery, access, and

696   visualization.

697   xii. **GEOROC** (Geochemistry of Rocks of the Oceans and Continents;

698   http://georoc.mpch-mainz.gwdg.de/georoc/) - a global geochemical

699		database containing published chemical and isotopic data as well as
700		extensive metadata for rocks, minerals and melt/fluid inclusions.
701	xiii.	**MetPetDB** (https://tw.rpi.edu/project/MetPetDB) - a relational database
702		and repository for global geochemical data on and images collected from
703		metamorphic rocks from the earth's crust.
704	xiv.	**The Planetary Data System** (PDS; https://pds.nasa.gov/) - a long-term
705		archive of digital data products returned from NASA's planetary missions,
706		and from other kinds of flight and ground-based data acquisitions,
707		including laboratory experiments.
708	xv.	**Mineral RI** (https://odr.io/mineralRI): a database containing the refractive
709		indices minerals and synthetic compounds. (Shannon et al. 2017).
710		
711	b.	Mineral Information Models
712		The global research community of mineralogy has made impressive progress on
713		information models for database construction and data sharing in the past
714		decades. From the point of view of data management, a good information model
715		should be correct, complete, and consistent. An effective way for information
716		modeling in real-world practice is to follow or adapt existing community
717		agreements or standards on mineralogy, such as those on the physical, chemical,
718		and biological characteristics of minerals. For instance, the Database of Mineral
719		Properties (https://rruff.info/ima/) maintained by the International Mineralogical
720		Association (IMA) keeps an up-to-date list of mineral species. The main

721   components in the information model include mineral name, chemistry, mineral

722   groups, origins, paragenetic mode, IMA status, relevant references, and links to

723   external sources such as mindat.org, Google Images, and Wikipedia.

724

725   As open data and data-driven studies are increasingly accepted in the geoscience

726   community, many databases in the field of mineralogy also increase the visibility

727   of their information model and build machine interfaces for data query, access,

728   and download. For instance, the RRUFF database (https://rruff.info) has

729   integrated records of Raman spectra, X-ray diffraction, and chemistry data for

730   minerals. The user interface enables data query through mineral name and

731   chemistry includes/excludes. Interested users can also contact the database

732   manager for batch data download and sharing. Mindat.org

733   (https://www.mindat.org) is another widely used database in the field of

734   mineralogy. Its construction and maintenance follow a crowd-sourcing style.

735   Besides the physical and chemical attributes of mineral species, a unique attribute

736   on mindat.org is a comprehensive list of the localities where that mineral species

737   has been found. In the past years, many research activities benefited from the

738   open data shared by mindat.org. As each of those open databases has its own

739   focus and information model, scientists in large-scale research activities often

740   need to collect data from multiple sources. Recently, researchers in

741   geoinformatics and data science also discussed the need for a more

742   comprehensive mineral information model to document the extensive facets of

743   mineral data, such as the global earth mineral inventory (GEMI) proposed by

744   Prabhu et al. (2021a). Complementing these efforts are initiatives using semantic

745   technologies to build knowledge graphs for mineral species, as a preparation to

746   explore new ways for annotating and discovering mineral data shared on the

747   Internet (Brodaric and Richard, 2020).

748

749   The FAIR (findable, accessible, interoperable, and accessible) data principles

750   (Wilkinson et al. 2016) are now widely accepted in geoscience. Information

751   models are an important part of FAIR data. More community efforts, such as

752   through IMA, the Mineralogical Society of America (MSA), and the

753   Geoinformation Committee of the International Union of Geological Sciences

754   (IUGS-CGI), are needed to promote the quality and usefulness of the model

755   outputs.

756

757

758 6. Informatics Innovations Needed for Mineralogy.

759

760   The previous sections of this paper (and many other informatics papers focusing on

761   various domains) have clearly emphasized the value that informatics methods provide to

762   their respective domains (Lord et al. 2004; Goble & Stevens 2008; Heberling et al. 2021;

763   Collen 1986; Gauthier et al. 2019). However, a point often missed or overlooked in

764   scientific literature discussions is that innovations in data science and informatics are

765 usually driven by diverse datasets available in various domains and the needs of the use-

766 cases utilizing those datasets. In this section we discuss some of the interesting data

767 science challenges we observed while working with mineral data to try to answer some

768 of the unanswered questions in geoscience.

769

770 Mineral data provide interesting and unique problems that limit the usability of existing

771 machine learning methods meant to extract meaningful information from data.

772

773   a. Small and Sparse Data Framework:

774   It has been widely publicized that we live in the "Age of Big Data" (Lohr 2012; Wise

775   & Shaffer 2015; Yu 2016; Wachter 2019; Borgman et al. 2008), and understandably

776   there has been a lot of research done into scaling-up algorithms, methods,

777   software, and hardware needed to enable the exploration and use of very large

778   datasets to gain valuable information. This focus has led to the creation and

779   constant improvement of "big data frameworks", which provide a roadmap on

780   how to work with large datasets. However, mineralogy, along with many other

781   fields in Earth and planetary sciences, provide a plethora of small and sparse

782   datasets that do not fall into the realm of big data. These datasets therefore

783   require the application of methodologies that lie outside the focus of traditional

784   big data researchers. The next major hurdle for mineral informatics (and

785   geoinformatics in general) is to work towards creating a framework for small and

786   sparse data.

787

For example, mineral data collected by the CheMin X-ray diffractometer onboard the Mars Science Laboratory (Morrison et al. 2018a; Rampe et al. 2018) has few data points, having analyzed ~40 samples, each with around a dozen mineral species (as of January 2022). The CheMin team used small (on the order of dozens to a few hundred data points) datasets of mineral composition and associated unit-cell parameters to build models capable of predicting the basic chemical composition of major mineral phases observed on Mars, based solely on their unit-cell parameters (Morrison et al. 2018a-b). However, the team wished to push their chemical prediction further - to predict complex, multi-element mineral compositions for the martian crystallographic data. In order to do so, Morrison et al. (2018c) assembled datasets of laboratory-analyzed complex, multi-element mineral compositions and unit-cell parameters, which contained only a few hundred data points for each of the major mineral groups identified by CheMin. Morrison et al. (2018c) used the small data Label Distribution Learning approach to predict complex chemical compositions (up to 12 elements, in some mineral systems) of mineral samples collected by the CheMin instrument based on the unit-cell parameters of these samples (See section 3b for more details). Significantly more work can be done here to increase the accuracy and performance of these models and such complex datasets with small sample sizes provide an interesting and rare challenge to data scientists.

808

809   Mineral geochemistry often contains information related to the geologic,

810   chemical, and/or biological processes and materials that went into their formation

811   and any subsequent weathering and alteration. However, geochemical data are

812   inherently sparse due to chemical variability in geologic deposits and materials,

813   different elemental affinities amongst different mineral species, and analytical

814   bias introduced by research aims or instrument limitations. The resulting

815   frequency of "missing values" makes many geochemical datasets unsuitable for

816   use with existing algorithms designed for complete or near-complete datasets. A

817   prime example of the sparseness of geochemical data is the garnet dataset

818   compiled by Chiama et al. (2020, 2022b(in prep)), which contains over 95,000

819   geochemical analyses of garnet group mineral samples collected from a variety of

820   sources, ranging from large repositories (EarthChem, RRUFF, MetPetDB) to

821   individual peer-reviewed literature. Even a compiled and curated dataset such as

822   this is considered sparse, largely due to the chemical variability amongst the

823   various garnet mineral species, resulting in missing values in the chemical

824   compositions of these samples (Chiama et al. 2022a). For example, of the 95,000

825   analyses compiled, only 5 major elements (Mg, Fe, Ca, Al, and Si) are present

826   and/or reported in most samples, while other elements, including Mn, Cr, and Ti,

827   are much less common throughout the dataset. An additional contribution to this

828   sparseness is that studies may not analyze for all elements in a sample (e.g.,

829   limited to elements of interest, difficulty measuring light elements), resulting in

830   missing values for which it is not known whether that element is present. Thus,

831   while analyzing these data (using descriptive, prescriptive, or predictive methods)

832   we need to take into account these missing values and their effect on the results.

833   Sparse data is not a problem new or unique to mineral data (Greenland et al. 2000;

834   Greenland et al. 2016; Sweeting et al. 2004; Rogers et al. 2018), but, as is the

835   theme for the rest of this paper, we must learn from the successes and failures of

836   other domains in addressing sparse data (Shepperd & Cartwright 2001; Katz 1987;

837   Uzuner 2009; Derczynski et al. 2013).

838

839   Other examples of small and sparse data challenges can be encountered in efforts

840   to understand other planets and moons including Venus and Titan through their

841   mineralogy and geochemistry. Frigid Titan's exotic mineralogy, with water ice as a

842   principal rock-forming mineral, oceans of liquid hydrocarbons, and varied

843   postulated organic minerals, is mostly understood through laboratory analogs

844   (e.g., Fegley et al. 1992; Gilmore et al. 2017; Bullock & Grinspoon 1996; Hashimoto

845   & Abe 2005; Treiman & Bullock 2012; Zolotov 2018; Hazen 2018; Maynard-Casely

846   et al. 2018; Cable et al. 2021).

847

848   Small and sparse datasets are a common occurrence in Earth and planetary

849   science. Despite the limitations of the available information, the answers to key

850   scientific questions are tied to these datasets. Therefore, an effort to create a

851   framework to handle small and/or sparse data will be highly beneficial to scientific

852   research in Earth and planetary science. Many researchers are working on "High-

853   Dimensional, Small Sample Size" (HDSSS) or "High-Dimensional, Low Sample Size"

854   (HDLSS) and its use in data analytics (Liu et al. 2017; Golugula et al. 2011; Shen et

855   al. 2016; Yata and Aoshima 2012; Hall et al. 2005). However, this area of research

856   has received much less attention compared to its big data counterpart, and hence

857   has lacked the synthesis and generalization that comes with the popularity and

858   maturity of well-established fields. The aforementioned examples (including

859   section 3 and 5), clearly demonstrate how such a framework would open paths

860   for exploring very important scientific questions within and beyond mineralogy.

861

862   b. Data Discovery

863   An increasing trend of data science in recent years is doing research with open

864   data shared by others (Fox and Hendler, 2014). Several recent scientific advances

865   in mineral informatics also reflect that trend (e.g., Hazen et al., 2019). From the

866   point of view of data users, an ideal situation is that they can efficiently find data

867   portals on the Internet, datasets on the portals, or subsets of the data. In

868   comparison, from the point of view of data providers and data managers, they

869   need to organize the data with shared community standards, detailed metadata,

870   and persistent and stable facilities to increase the reusability. As illustrated in the

871   FAIR data principles for open data (Wilkinson et al., 2016), the first two key points

872   to consider are the findability and accessibility of data. Correspondingly, three key

873   technical items arise here. The first item is the metadata schema for describing

874   the datasets. While there are many common-purpose metadata schemas, such as

875	the Dublin Core, for describing datasets, for domain-specific data such as those in
876	mineralogy there can also be specific metadata elements. The second item is the
877	identifier for the datasets. Similar to the Digital Object Identifier (DOI) for
878	publications, datasets shared on the Internet should also have specific identifiers
879	to enable persistent and stable discoverability. The third item with respect to
880	findability and accessibility is the protocol for retrieving metadata through the
881	identifier of datasets. Community efforts such as DataCite (Brase, 2009) have
882	made solid progress toward that goal. Nevertheless, the wide implementation of
883	those best practices for open data in geosciences, including mineralogy, still need
884	more time. It is also important to remember that appropriate scientific credit must
885	be given at every stage of informatics methodology, from the acquisition of data,
886	to data analytics, and finally the dissemination of the research products produced
887	by the data analysis.

888

889	A very recent technical development regarding data discovery is the Dataset
890	Search Engine released by Google (Noy et al., 2019), which is able to index millions
891	of datasets on thousands of data portals, including their identifiers or Web links.
892	End users of the dataset search engine
893	(https://datasetsearch.research.google.com) have integrated access to thousands
894	of data portals. When a dataset is found on the engine, users can go to its original
895	data portal page through the identifier or Web link and then download. The
896	Google Dataset Search Engine is built on the top of Schema.org, which is designed

897   as a comprehensive metadata schema for annotating digital objects on the Web.

898   The annotated objects, such as datasets, will then be indexed by the search

899   engines. As its usage is expanding, Schema.org also provides space for extending

900   the metadata elements of certain objects. A potential here is to have specific

901   metadata elements designed for datasets of mineralogy, and this should be based

902   on community collaboration. In the past few years, the EarthCube community has

903   leveraged a list of open geoscience data portals to develop the GeoCODES search

904   engine (McHenry et al., 2021). It is also based on Schema.org but has made

905   extensions specifically for the registration and discovery of geoscience data. Any

906   future efforts on the findability and accessibility of open mineralogy data can

907   absolutely benefit from the technical structure and experience of GeoCODES.

908   Community agreements and standards, such as those developed by IMA, MSA,

909   and IUGS-CGI, as well as best practices in existing data portals, such as those in

910   RRUFF and mindat.org, will also be helpful to enrich the metadata of open

911   mineralogy data.

912

913   c. Data Processing

914   Dozens of data repositories contain a wealth of mineralogical information (see

915   section 5a) from which large data resources can be extracted. Web-scraping

916   algorithms allow for the retrieval and storage of large amounts of data from web

917   sources (Glez-Peña et al. 2014; Zhao 2017). Scraping algorithms in scripting

918   languages such as Python or R allow users to extract and compile large amounts

919   of data from web sources or journal articles in minutes or seconds, but the

920   structure (or lack thereof) of web pages can slow the production of new data

921   resources. Open-access mineral databases tend to be very contributor friendly;

922   thus, users can pick and choose which data to include for a particular entry.

923   Recognizing the inconsistencies in the storage and representation of mineral

924   attribute data within and across different mineralogical databases is essential

925   when compiling large mineral datasets from open data sources.

926   Webpages associated with Webmineral and Mindat have hierarchical structures

927   made up of Hypertext Markup Language (HTML), Extensible Markup Language

928   (XML), or Cascading Style Sheet (CSS) elements that allow for the selection of

929   nodes that can contain specific data a user is interested in (Gunawan et al., 2019).

930   The ubiquitous occurrence or rarity of mineral, relative interest among the

931   scientific community in a mineral, as well as the age of discovery cause significant

932   differences in the amount of information available for a mineral, driving the

933   differences in the structure of these webpages. Webpages and digital PDFs

934   associated with the *Handbook of Mineralogy* have very little structure, which

935   places more importance on the use of keywords (e.g., space group or crystal

936   system) or separators (e.g., each mineral attribute or property introduced may

937   have a semicolon preceding the associated description) in the compilation of data.

938   Nested conditional statements (i.e., if-else statements) are useful for compiling

939   data from web databases that have variable or no structure, but this approach can

940   be more time-consuming and prone to error. Some headers may be reused such

941   as "beta (ß)" which is used as a descriptor of the refractive indices in biaxial

942   minerals (e.g., Frazier et al., 1963; Gunter, 1992) and it can also refer to the

943   geometry of the unit-cell of dimensions (e.g., Grove and Hazen, 1974; Nesse,

944   1991).

945   d. Quantifying and Correcting Bias

946   Critical to all of these aspects of data resource development and use is an

947   understanding of and, where possible, modeling of the biases that exist in each of

948   these systems. For example, significant biases occur in mineral sampling based on

949   the physical appearance of the phase (e.g., large, brightly colored, euhedral

950   crystals), the economic value, the scientific interest, proximity to major

951   universities or research centers, and analytical technology. Such biases can be

952   corrected with models of each of these parameters (Hystad et al. 2019; Hazen et

953   al. 2016; Grew et al. 2017). Natural preservational biases is more complex, as it

954   involves geologic history and mineral properties (e.g., chemistry, solubility,

955   hardness), but work is underway to begin unraveling the history of preservational

956   biases in mineral systems on Earth and other planetary bodies (Liu et al. 2019).

957

958   7. Informatics research as a socio-technical system

959

960   Research in the field of informatics is heavily dependent on the interactions

961   between the data scientists and domain scientists (e.g., mineralogists, planetary

962   scientists) (Ma et al. 2017). Conducting and applying informatics research is very

much a socio-technical system (Herrmann 2011). It is as much about the researchers, their interactions, the hypotheses generated, and interpreting of results from visualizations or models as it is about the data, the algorithms, and the models. Collaborations in informatics include many iterations between data and domains scientists starting from data explorations and the problem formulation to interpreting the results and documenting the scientific insights learnt from the data.

We recommend starting an informatics exploration with an in-person or virtual "datathon" (Anslow et al. 2016; Fritz et al. 2020). During this datathon, which usually lasts a day or two, collaborators mainly focus on:

1. Interactions and discussions between data scientists and domain scientists to frame their goals and expectations.
2. Documenting the research questions to be explored.
3. Collating the data resources required to explore the documented research questions.
4. Exploring the methods needed to examine the data (both analytically and visually).
5. Constructing a roadmap for dividing the research question and tasks into smaller, more tractable parts.
6. Leveraging descriptive, prescriptive, and predictive methods to gain preliminary insights from the data.

985	7. Forming short term and long-term goals based on the preliminary results.
986	8. Documenting the shortcomings of the methods explored and why these
987	roadblocks hamper scientific exploration.
988	9. Documenting the innovation needs of both data science and domain
989	science methods to overcome the previously documented hurdles.
990	

991   Not all of these steps need to be done during the two-day datathon; steps 1 and

992   2 can be completed beforehand. The main goal of conducting a datathon is to

993   expedite and streamline the initial data exploration to gain preliminary results that

994   can be examined by the domain scientist, while also allowing the data scientist to

995   explore and understand the intricacies of the data at hand. Additionally, all

996   collaborators gain an understanding of the shortcomings, needs, and

997   opportunities of their data and of the current methods to address the desired

998   scientific questions. This inventory of needs and opportunities in both the data

999   science and domain science can result in a datathon output of a list of projects and

1000   publications spurred by the creative and iterative processes of this closely

1001   collaborative effort.

1002   After the initial datathon, each collaborator (or group of collaborators) has a plan

1003   of action for the projects and subtasks within the project they are leading.

1004   Subsequent communication and collaboration usually follows the preferred

1005   working model of the team. For example, weekly meetings between the group to

1006   discuss advances in the project, or email communications between the team for

1007   the same purpose. The steps taken after the datathon and methods to

1008   communicate and collaborate change depending on the work style and comfort

1009   levels of the collaborators. General recommendations for this step include

1010   "science of team science" best practices advocated by many communities (NASEM

1011   2015).

1012

1013   8. Vision for the future

1014

1015   a. Implications

1016   Durable and information-rich, minerals are the only ancient relics that offer direct,

1017   solid glimpses of eons of planetary transformation (Hazen 2021). It is important to

1018   extract the abundance of information contained in these mineral samples to

1019   improve our understanding of the evolution of our planet, our solar system, and

1020   the role our planet's evolving geosphere played in the origin and proliferation of

1021   life. Key, synergistic aspects of the ongoing paradigm shift in mineralogy includes

1022   systematic efforts to collect and curate mineralogical information in data

1023   resources that enable open and widespread dissemination, and the use of those

1024   data to make scientific discoveries.

1025

1026

1027   b. Look around! (At other fields)

1028   As mentioned earlier in this paper, informatics methods have been followed,

1029   implemented, and improved upon in other fields over the past decades. The

1030   concept of "X-informatics" has also been around since its first conceptualization

1031   in 2007 (Gray and Szalay 2007; Hey et al. 2009), and over the past decade there

1032   has been a steady decline in researchers conducting informatics research in the

1033   silos of their respective fields. When planning for a new paradigm like mineral

1034   informatics, it is important to learn from successes and failures of more mature

1035   fields of informatics (Lord et al. 2004; Goble & Stevens 2008; Heberling et al. 2021)

1036   and modify the methods developed by past researchers to apply them to

1037   comprehensively address our needs as a community.

1038

1039   Over the last decade, there have been some efforts at collating various data

1040   resources in the geosciences and providing these data to researchers with minimal

1041   barriers and maximum interoperability. These efforts include Onegeology

1042   (Jackson 2010), Onegeochemistry (Wyborn et al. 2021; Chamberlain et al. 2021),

1043   and Onestratigraphy (Wang et al. 2021). The Onegeochemistry initiative also

1044   includes plans to develop best practices for FAIR geochemical data, governance

1045   models to ensure participation and trust, and a business model to ensure long-

1046   term   sustainability

1047   (https://www.earthchem.org/communities/onegeochemistry/).   Efforts   to

1048   improve the access, usage, and impact of mineral data resources can learn from

1049   the successes and challenges faced by such global initiatives. Developing a set of

1050   best practices and recommendations for creating, linking, and releasing mineral

1051   data would improve the mineral data landscape and make it easier for researchers

1052   to produce and use mineral data without too many barriers.

1053

1054   Just as increasing the findability, accessibility, interoperability, reusability, and

1055   other important aspects of mineral data management and stewardship, obtaining

1056   scientific insights from mineral data using data-driven methods are another key

1057   facet of mineral informatics. For this, too, we can look to and learn from the

1058   success and failures of other domains applying informatics methods to answer

1059   their research questions. We hope the research directions for informatics and

1060   other fields like mineralogy, planetary science, and other related fields using

1061   mineral data that have been documented in this paper act as an initial step

1062   towards the ultimate goal of systematizing data driven scientific exploration using

1063   mineral data.

1064

1065

1066   9. Conclusion

1067

1068   Mineralogy is facing new opportunities and challenges with the increased interest

1069   in and applications of data-driven methods. We believe the next paradigm for the

1070   field of mineralogy is that of mineral informatics. Mineral informatics focuses on

1071   deciphering the patterns and trends hidden in mineralogical, geochemical, and

1072   related data and using these patterns to answer scientific questions, thus making

1073   important new discoveries. In this paper, we show how the study of minerals is

1074   essential to improving our understanding of the evolution of our planet, our solar

1075   system, and more. We present a broad methodology for the study and use of

1076   mineral informatics methods (Figure 1) and document the needs of the field and

1077   important scientific questions that may be answered using mineral informatics.

1078   We reiterate the symbiotic relationship between data scientists and domain

1079   scientists (e.g., mineralogists, planetary scientists, biologists) to make continuous

1080   and sustainable scientific progress.

1081

1082   In summary, our vision for the next decade of mineralogical research is built upon

1083   the systematic and coordinated study of mineral data and of the data science

1084   methods used to gain scientific insights.

1085

1086   10. Acknowledgements

1087

those of the authors and do not necessarily reflect the views of the National Aeronautics and Space Administration.

References

- Akaogi, M., & Akimoto, S. (1977). Pyroxene-Garnet solid-solution equilibria in the systems Mg4Si4O12-Mg3Al2Si3O12 and Fe4Si4O12-Fe3Al2Si3O12 at high pressure and temperatures. PEPI, 15, 90–106.

- Anslow, C., Brosz, J., Maurer, F., & Boyes, M. (2016, February). Datathons: an experience report of data hackathons for data science education. In Proceedings of the 47th ACM Technical Symposium on Computing Science Education (pp. 615-620).

- Assante M, Candela L, Castelli D, Tani A (2016) Are scientific data repositories coping with research data publishing? Data Sci J 15:6. https://doi.org/10.5334/dsj-2016-006

- Bandy, Mark Chance and Jean A. Bandy (1955). De Natura Fossilium. New York: George Banta Publishing Company.

- Beneventano, D., & Bergamaschi, S. (2004). The MOMIS methodology for integrating heterogeneous data sources. In Building the information society (pp. 19-24). Springer, Boston, MA.

1115   ●   Borgman, C. L., Abelson, H., Dirks, L., Johnson, R., Koedinger, K. R., Linn, M. C., ...

1116   & Szalay, A. (2008). Fostering learning in the networked world: The cyberlearning

1117   opportunity and challenge. A 21st century agenda for the National Science

1118   Foundation.

1119   ●   Boujibar, A., Howell, S., Zhang, S., Hystad, G., Prabhu, A., Liu, N., Stephan, T.,

1120   Narkar, S., Eleish, A., Morrison, S.M., Hazen, R.M., and Nittler, L.R. (2021) Cluster

1121   analysis of presolar silicon carbide grains: Evaluation of their classification and

1122   astrophysical implications. Astrophysical Journal Letters, 907, L39 (14 pp).

1123   ●   Brack, A. (2013). Clay Minerals and the Origin of Life (pp. 507–521).

1124   https://doi.org/10.1016/B978-0-08-098258-8.00016-X

1125   ●   Bradley, D.C., 2011, Secular trends in the geologic record and the supercontinent

1126   cycle: Earth Science Reviews, v. 108, p. 16-33.

1127   ●   Bragg, W.H. and Bragg, W.L., 1913. The reflection of X-rays by crystals.

1128   Proceedings of the Royal Society of London. Series A, Containing Papers of a

1129   Mathematical and Physical Character, 88(605), pp.428-438.

1130   ●   Brase, J., 2009. DataCite-A global registration agency for research data. In

1131   Proceedings of The Fourth International Conference on Cooperation and

1132   Promotion of Information Resources in Science and Technology, Beijing, China.

1133   pp.257-261.

1134   ●   Brodaric, B. and Richard, S.M., 2020. The GeoScience Ontology. The 2020 AGU Fall

1135   Meeting, Virtual. Abstract IN030-07.

- Bullard T., Freudenthal J., Avagyan S., Kahr B. (2007) Test of Cairns-Smith's 'crystals-as-genes' hypothesis. Faraday Discussions, 136, 231-245.

- Bullock, M. A., & Grinspoon, D. H. (1996). The stability of climate on Venus. Journal of Geophysical Research: Planets, 101(E3), 7521–7529. https://doi.org/10.1029/95JE03862

- Burke, R., Felfernig, A., & Göker, M. H. (2011). Recommender systems: An overview. Ai Magazine, 32(3), 13-18.

- Cable, M. L., Runčevski, T., Maynard-Casely, H. E., Vu, T. H., & Hodyss, R. (2021). Titan in a Test Tube: Organic Co-crystals and Implications for Titan Mineralogy. Accounts of Chemical Research, 54(15), 3050–3059. https://doi.org/10.1021/acs.accounts.1c00250

- Cairns-Smith, A. G. (1990). Canto: Seven clues to the origin of life: A scientific detective story. Cambridge University Press.

- Cairns-Smith, A. G., & Hartman, H. (Eds.). (1986). Clay minerals and the origin of life. Cambridge University Press.

- Chamberlain, K. J., Lehnert, K. A., McIntosh, I. M., Morgan, D. J., & Wörner, G. (2021). Time to change the data culture in geochemistry. Nature Reviews Earth & Environment, 2(11), 737-739.

- Chiama K, Gabor M, Lupini I, Rutledge R, Nord JA, Zhang S, Boujibar A, Bullock ES, Walter MJ, Lehnert K, Spear F, Morrison SM, Hazen RM. ESMD- Garnet Dataset. Published (2022a) via Open Data Repository. https://doi.org/10.48484/camh-xy98

1158 ● Chiama K, Gabor M, Lupini I, Rutledge R, Nord JA, Zhang S, Boujibar A, Bullock ES,

1159   Walter MJ, Lehnert K, Spear F, Morrison SM, Hazen RM (2022b) The secret life of

1160   garnets: A comprehensive, standardized dataset of garnet geochemical analyses

1161   integrating localities and petrogenesis, Earth System Science Data (in prep)

1162 ● Chiama, K., Rutledge, R., Gabor, M., Lupini, I., Hazen, R. M., Zhang, S., et al. (2020):

1163   Garnet: a comprehensive, standardized, geochemical database incorporating

1164   locations and paragenesis. Geological Society of America Abstracts with Programs

1165   <doi:10.1130/abs/2020se-344505>

1166 ● Childers, S. E., Ciufo, S., & Lovley, D. R. (2002). Geobacter metallireducens accesses

1167   insoluble Fe(iii) oxide by chemotaxis. Nature, 416(6882), 767–769.

1168   https://doi.org/10.1038/416767a

1169 ● Cleland, C. E., Hazen, R. M., & Morrison, S. M. (2021). Historical natural kinds and

1170   mineralogy: Systematizing contingency in the context of necessity. Proceedings of

1171   the National Academy of Sciences, 118(1), e2015370118.

1172   https://doi.org/10.1073/pnas.2015370118

1173 ● Coates D.R. (1985) Mineral Resources. In: Geology and Society. Environmental

1174   Resource Management Series. Springer, Boston, MA.

1175   https://doi.org/10.1007/978-1-4613-2543-7_2

1176 ● Collen, M. F. (1986). Origins of medical informatics. Western Journal of Medicine,

1177   145(6), 778.

1178 ● Dana, J.D. and Dana, E.S., 1895. The System of Mineralogy of James Dwight Dana.

1179   1837-1868: Descriptive Mineralogy. J. Wiley & sons.

1180   ● De Sanctis, M. C., Ammannito, E., Capria, M. T., Tosi, F., Capaccioni, F., Zambon,

1181   F., … Turrini, D. (2012). Spectroscopic Characterization of Mineralogy and Its

1182   Diversity Across Vesta. Science, 336(6082), 697–700.

1183   https://doi.org/10.1126/science.1219270

1184   ● Derczynski, L., Ritter, A., Clark, S., & Bontcheva, K. (2013, September). Twitter part-

1185   of-speech tagging for all: Overcoming sparse and noisy data. In Proceedings of the

1186   international conference recent advances in natural language processing ranlp

1187   2013 (pp. 198-206).

1188   ● Dymshits, A. M., Bobrov, A. V., Bindi, L., Litvin, Y. A., Litasov, K. D., Shatskiy, A. F.,

1189   & Ohtani, E. (2013). Na-bearing majoritic garnet in the $Na_2MgSi_5O_{12}$-

1190 $Mg_3Al_2Si_3O_{12}$ join at 11-20GPa: Phase relations, structural peculiarities and solid

1191   solutions. Geochimica et Cosmochimica Acta, 105, 1–13.

1192   https://doi.org/10.1016/j.gca.2012.11.032

1193   ● Ehlmann, B. L., & Edwards, C. S. (2014). Mineralogy of the Martian Surface. Annual

1194   Review of Earth and Planetary Sciences, 42(1), 291–315.

1195   https://doi.org/10.1146/annurev-earth-060313-055024

1196   ● Fegley, B., Treiman, A. H., & Sharpton, V. L. (1992). Venus surface mineralogy:

1197   Observational and theoretical constraints. Proceedings of Lunar and Planetary

1198   Science, volume 22.

1199   ● Fischer, G., & Herrmann, T. (2011). Socio-technical systems: a meta-design

1200   perspective. International Journal of Sociotechnology and Knowledge

1201   Development (IJSKD), 3(1), 1-33.

- Fox, P. (2011, August). The rise of informatics as a research domain. In Proceedings of WIRADA Science Symposium, Melbourne, Australia (Vol. 15, pp. 125-131).

- Fox, P. and Hendler, J., 2014. The science of data science. Big Data, 2(2), pp.68-70.

- Fox, P., L. Gundersen, K. Lehnert, D. McGuinness, K. Sinha, and W. Snyder (2006), Toward broad community collaboration in geoinformatics, Eos Trans. AGU, 87(46), 513–513, doi:10.1029/2006EO460005.

- Fox, P., McGuinness, D. (2008). TWC Semantic Web Methodology. https://archive.tw.rpi.edu/media/latest/SemanticMethodologyPathwayPretty.png

- Fritz, S., Milligan, I., Ruest, N. and Lin, J. (2020), "Building community at distance: a datathon during COVID-19", Digital Library Perspectives, Vol. 36 No. 4, pp. 415-428. https://doi.org/10.1108/DLP-04-2020-0024

- Fürnkranz, J., & Flach, P. A. (2003). An analysis of rule evaluation metrics. In Proceedings of the 20th international conference on machine learning (ICML-03) (pp. 202-209).

- Fyfe, A., McDougall-Waters, J., & Moxham, N. (2015). 350 years of scientific periodicals. Notes and Records: the Royal Society journal of the history of science, 69(3), 227-239.

- Geng, X., 2016. Label distribution learning. IEEE Transactions on Knowledge and Data Engineering, 28(7), pp.1734-1748.

1222 ● Geng, X., Wang, Q. and Xia, Y., 2014, August. Facial age estimation by adaptive
1223   label distribution learning. In Pattern Recognition (ICPR), 2014 22nd International
1224   Conference on (pp. 4465-4470). IEEE.

1225 ● Geng, X., Yin, C. and Zhou, Z.H., 2013. Facial age estimation by learning from label
1226   distributions. IEEE transactions on pattern analysis and machine intelligence,
1227   35(10), pp.2401-2412.

1228 ● Gilmore, M., Treiman, A., Helbert, J., & Smrekar, S. (2017). Venus Surface
1229   Composition Constrained by Observation and Experiment. Space Science Reviews,
1230   212(3–4), 1511–1540. https://doi.org/10.1007/s11214-017-0370-8

1231 ● Glein, C. R., & Waite, J. H. (2020). The Carbonate Geochemistry of Enceladus'
1232   Ocean. Geophysical Research Letters, 47(3).
1233   https://doi.org/10.1029/2019GL085885

1234 ● Goble, C., & Stevens, R. (2008). State of the nation in data integration for
1235   bioinformatics. Journal of biomedical informatics, 41(5), 687-693.

1236 ● Golden, J.J. (2019) Mineral Evolution Database: Data Model for Mineral Age
1237   Associations. M.S. Thesis, University of Arizona, Tucson AZ.

1238 ● Golden, J.J., Pires, A.J., Hazen, R.M., Downs, R.T., Ralph, J. and Meyer, M. (2016)
1239   Building the Mineral Evolution Database: Implications for future big data analysis.
1240   Geological Society of America Abstracts with Programs, 28602.

1241 ● Gorlas, A., Jacquemot, P., Guigner, J. M., Gill, S., Forterre, P., & Guyot, F. (2018).
1242   Greigite nanocrystals produced by hyperthermophilic archaea of Thermococcales
1243   order. PLoS One 13, 1–10.

1244   ● Golugula, A., Lee, G., & Madabhushi, A. (2011, August). Evaluating feature

1245   selection strategies for high dimensional, small sample size datasets. In 2011

1246   Annual International conference of the IEEE engineering in medicine and biology

1247   society (pp. 949-952). IEEE.

1248   ● Gray J and Szalay A. (2007) eScience - A transformed scientific method. NRC-CSTB

1249   meeting,   Mountain   View,   CA.   Retrieved   from

1250   http://jimgray.azurewebsites.net/talks/NRC-CSTB_eScience.ppt.

1251   ● Greenland, S., Mansournia, M. A., & Altman, D. G. (2016). Sparse data bias: a

1252   problem hiding in plain sight. bmj, 352.

1253   ● Greenland, S., Schwartzbaum, J. A., & Finkle, W. D. (2000). Problems due to small

1254   samples and sparse data in conditional logistic regression analysis. American

1255   journal of epidemiology, 151(5), 531-539.

1256   ● Greenwell H.C., Coveney P.V. (2006) Layered Double Hydroxide Minerals as

1257   Possible Prebiotic Information Storage and Transfer Compounds. Origin of Life and

1258   Evolution of Biospheres, 36, 13–37.

1259   ● Gregory, D.D., Cracknell, M.J., Large, R.R., McGoldrick, P., Kuhn, S., Maslennikov,

1260   V.V., Baker, M.J., Fox, N., Belousov, I., Figueroa, M.C., Steadman, J.A., Fabris, A.J.,

1261   and Lyons, T.W. (2019) Distinguishing ore deposit type and barren sedimentary

1262   pyrite using laser ablation-inductively coupled plasma-mass spectrometry trace

1263   element data and statistical analysis of large data sets. Economic Geology, 114,

1264   771-786.

1265   ● Grew ES, Hystad G, Hazen RM, Krivovichev SV, Gorelova LA. How many boron

1266   minerals occur in Earth's upper crust?. American Mineralogist: Journal of Earth

1267   and Planetary Materials. 2017 Aug 1;102(8):1573-87.

1268   ● Hashimoto, G. L., & Abe, Y. (2005). Climate control on Venus: Comparison of the

1269   carbonate and pyrite models. Planetary and Space Science, 53(8), 839–848.

1270   https://doi.org/10.1016/j.pss.2005.01.005

1271   ● Hazen, R.M. (2019) An evolutionary system of mineralogy: Proposal for a

1272   classification based on natural kind clustering. American Mineralogist, 104, 810-

1273   816. DOI: 10.2138/am-2019-6709

1274   ● Hazen RM & Morrison SM (2021) On the paragenetic modes of minerals: A mineral

1275   evolution perspective, American Mineralogist (In Press)

1276   ● Hazen RM, Ferry JM. Mineral evolution: Mineralogy in the fourth dimension.

1277   Elements. 2010 Feb 1;6(1):9-12.

1278   ● Hazen, R.M., Morrison, S.M., Krivovichev, S.L., and Downs, R.T. (2022) Lumping

1279   and splitting: Toward a classification of mineral natural kinds. American

1280   Mineralogist, in press.

1281   ● Hazen RM, Hummer DR, Hystad G, Downs RT, Golden JJ. Carbon mineral ecology:

1282   Predicting the undiscovered minerals of carbon. American Mineralogist. 2016 Apr

1283   1;101(4):889-906.

1284   ● Hazen RM, Morrison SM, Prabhu A, Williams J (2021a) On the paragenetic modes

1285   of minerals: A mineral evolution perspective, Geological Society of America

1286   Abstracts with Programs. Vol 53, No. 6, 2021, doi: 10.1130/abs/2021AM-365916

1287   ● Hazen, R. M. (2005). Genesis: Rocks, Minerals, and the Geochemical Origin of Life.

1288   Elements, 1(3), 135–137. https://doi.org/10.2113/gselements.1.3.135

1289   ● Hazen, R. M. (2018). Titan mineralogy: A window on organic mineral evolution.

1290   American Mineralogist, 103(3), 341–342. https://doi.org/10.2138/am-2018-6407

1291   ● Hazen, R. M., & Morrison, S. M. (2020). An evolutionary system of mineralogy.

1292   Part I: Stellar mineralogy (&gt;13 to 4.6 Ga). American Mineralogist, 105(5),

1293   627–651. https://doi.org/10.2138/am-2020-7173

1294   ● Hazen, R. M., & Sholl, D. S. (2003). Chiral selection on inorganic crystalline

1295   surfaces. Nature Materials, 2(6), 367–374. https://doi.org/10.1038/nmat879

1296   ● Hazen, R. M., & Sverjensky, D. A. (2010). Mineral Surfaces, Geochemical

1297   Complexities, and the Origins of Life. Cold Spring Harbor Perspectives in Biology,

1298   2(5), a002162–a002162. https://doi.org/10.1101/cshperspect.a002162

1299   ● Hazen, R.M., Griffin, P.L., Carothers, J.M., Szostak, J.W., 2007. Functional

1300   information and the emergence of biocomplexity. Proc Natl Acad Sci USA 104,

1301   8574. https://doi.org/10.1073/pnas.0701744104

1302   ● Hazen, R. M., Downs, R. T., Eleish, A., Fox, P., Gagné, O. C., Golden, J. J., Grew, E.

1303   S., Hummer, D. R., Hystad, G., Krivovichev, S. V., Li, C., Liu, C., Ma, X., Morrison, S.

1304   M., Pan, F., Pires, A. J., Prabhu, A., Ralph, J., Runyon, S. E., & Zhong, H. (2019).

1305   Data-Driven Discovery in Mineralogy: Recent Advances in Data Resources,

1306   Analysis, and Visualization. In Engineering (Vol. 5, Issue 3, pp. 397–405). Elsevier

1307   BV. https://doi.org/10.1016/j.eng.2019.03.006

1308 ● Hazen, R. M., Hystad, G., Downs, R. T., Golden, J. J., Pires, A. J., and Grew, E. S.

1309   (2015). Earth's "missing" minerals. Am. Mineral. 100, 2344–2347. doi:

1310   10.2138/am-2015-5417

1311 ● Hazen, R.M., Liu, X.-M., Downs, R.T., Golden, J.J., Pires, A.J., Grew, E.S., Hystad, G.,

1312   Estrada, C., and Sverjensky, D.A. (2014) Mineral evolution: Episodic

1313   metallogenesis, the supercontinent cycle, and the coevolving geosphere and

1314   biosphere. Society of Economic Geologists Special Publication, 18, 1-15.

1315 ● Hazen, R. M., Morrison, S. M., & Prabhu, A. (2021b). An evolutionary system of

1316   mineralogy. Part III: Primary chondrule mineralogy (4566 to 4561 Ma). American

1317   Mineralogist, 106(3), 325-350.

1318 ● Hazen, R.M., and Morrison, S.M. (2022) On the paragenetic modes of minerals: A

1319   mineral evolution perspective. American Mineralogist, in press.

1320 ● Hazen, R. M., Papineau, D., Bleeker, W., Downs, R. T., Ferry, J. M., McCoy, T. J., …

1321   Yang, H. (2008). Mineral evolution. American Mineralogist, 93(11–12), 1693–

1322   1720. https://doi.org/10.2138/am.2008.2955

1323 ● Hazen, R.M., Grew, E.S., Downs, R.T., Golden, J. and Hystad, G., 2015. Mineral

1324   ecology: chance and necessity in the mineral diversity of terrestrial planets. The

1325   Canadian Mineralogist, 53(2), pp.295-324.

1326 ● Heberling, J. M., Miller, J. T., Noesgaard, D., Weingart, S. B., & Schigel, D. (2021).

1327   Data integration enables global biodiversity synthesis. Proceedings of the National

1328   Academy of Sciences, 118(6).

1329 ● Hey, T., Tansley, S., & Tolle, K. (2009). The Fourth Paradigm: Data-Intensive

1330 Scientific Discovery. The Fourth Paradigm: Data-Intensive Scientific Discovery.

1331 Microsoft Research. Retrieved from https://www.microsoft.com/en-

1332 us/research/publication/fourth-paradigm-data-intensive-scientific-discovery/

1333 ● Hindrichs, Eleazer, Lui, Williams, Nord, Gregory, Morrison, Hazen, Ostroverkhova

1334 (2022) Oxide spinel and data-driven discovery: A comprehensive mineralogical

1335 and geochemical data resource, incorporating composition, location, and

1336 paragenesis, Geological Society of America Abstracts with Programs. Vol. 54, No.

1337 4, 2022. doi: 10.1130/abs/2022NC-375662

1338 ● Hinkel, N. R., & Unterborn, C. T. (2018). The star–planet connection. I. Using stellar

1339 composition to observationally constrain planetary mineralogy for the 10 closest

1340 stars. The Astrophysical Journal, 853(1), 83.

1341 ● Hossin M & Sulaiman MN. (2015). A Review on Evaluation Metrics for Data

1342 Classification Evaluations. International Journal of Data Mining & Knowledge

1343 Management Process, 5(2), 01–11. https://doi.org/10.5121/ijdkp.2015.5201

1344 ● Hummer, D.R., Golden, J.J., Hystad, G., Downs, R.T., Eleish, A., Liu, C., Ralph, J.,

1345 Morrison, S.M., and Hazen, R.M. (2022) Evidence for the oxidation of Earth's crust

1346 from the evolution of manganese minerals. Nature Communications, in press.

1347 ● Hystad G, Morrison SM, Hazen RM. Statistical analysis of mineral evolution and

1348 mineral ecology: The current state and a vision for the future. Applied Computing

1349 and Geosciences. 2019 Oct 1;1:100005.

- Hystad, G., Downs, R.T. and Hazen, R.M., 2015. Mineral species frequency distribution conforms to a large number of rare events model: prediction of Earth's missing minerals. Mathematical Geosciences, 47(6), pp.647-661.

- Hystad, G., Boujibar, A., Liu, N., Nittler, L.R., and Hazen, R.M. (2021) Evaluation of the classification of presolar silicon carbide grains using consensus clustering with resampling methods: an assessment of the confidence of grain assignments. Monthly Notices of the Royal Astronomical Society, in press.

- Irifune, T. (1987). An experimental investigation of the pyroxene-garnet transformation in a pyrolite composition and its bearing on the constitution of the mantle. Physics of the Earth and Planetary Interiors, 45(4), 324–336. https://doi.org/10.1016/0031-9201(87)90040-9

- Jackson, I. (2010). OneGeology: improving access to geoscience globally. Earthwise, 26, 14-15.

- Jeff Gauthier, Antony T Vincent, Steve J Charette, Nicolas Derome, A brief history of bioinformatics, Briefings in Bioinformatics, Volume 20, Issue 6, November 2019, Pages 1981–1996, https://doi.org/10.1093/bib/bby063

- Katz, S. (1987). Estimation of probabilities from sparse data for the language model component of a speech recognizer. IEEE transactions on acoustics, speech, and signal processing, 35(3), 400-401.

- Kläs, M. (2018). Towards identifying and managing sources of uncertainty in AI and machine learning models-an overview. arXiv preprint arXiv:1811.11669.

1371   ● Kluyver T, Ragan-Kelley B, Perez F, Granger B, Bussonnier M, Frederic J, Kelley K,

1372   Hamrick J, Grout J, Corlay S, Ivanov P, Avila D, Abdalla S, Willing C, & Jupyter

1373   Development Team. (2016). Jupyter Notebooks - a publishing format for

1374   reproducible computational workflows. Stand Alone, 0(Positioning and Power in

1375   Academic Publishing: Players, Agents and Agendas), 87–90.

1376   https://doi.org/10.3233/978-1-61499-649-1-87

1377   ● Krivovichev, S.V. (2013) Structural complexity of minerals: information storage

1378   and processing in the mineral world. Mineralogical Magazine, 77, 275–326.

1379   https://doi.org/10.1180/minmag.2013.077.3.05

1380   ● Krivovichev, S.V. (2015) Structural complexity of minerals and mineral

1381   parageneses: information and its evolution in the mineral world. In: Danisi R,

1382   Armbruster T. Highlights in Mineralogical Crystallography. Berlin/Boston:Walter

1383   de Gruyter GmbH, pp 31-73.

1384   ● Krivovichev, S.V. (2016) Structural complexity and configurational entropy of

1385   crystalline solids. Acta Crystallographica, B72, 274-276.

1386   ● Krivovichev, S.V., Krivovichev, V.G., Hazen, R.M. (2018) Structural and chemical

1387   complexity of minerals: Correlations and time evolution. European Journal of

1388   Mineralogy, 30, 231–236. DOI: https://doi.org/10.1127/ejm/2018/0030-2694.

1389   ● Krivovichev, S.V., Yakovenchuk, V.N., Zhitova, E.S. (2012) Natural double layered

1390   hydroxides: structure, chemistry, and information storage capacity. Minerals as

1391   Advanced Materials II (Ed. S.V. Krivovichev). Springer-Verlag, Berlin Heidelberg.

1392   2012. pp. 87-91.

1393   ● L. Zhang, Y. Xie, L. Xidao and X. Zhang, "Multi-source heterogeneous data fusion,"

1394   2018 International Conference on Artificial Intelligence and Big Data (ICAIBD),

1395   2018, pp. 47-51, doi: 10.1109/ICAIBD.2018.8396165.

1396   ● Lafuente B, Downs R T, Yang H, Stone N (2015) The power of databases: the RRUFF

1397   project. In: Highlights in Mineralogical Crystallography, T Armbruster and R M

1398   Danisi, eds. Berlin, Germany, W. De Gruyter, pp 1-30.

1399   ● Large, R. R., Hazen, R. M., Morrison, S. M., Gregory, D. D., Steadman, J. A., &

1400   Mukherjee, I. (2022). Evidence that the GOE was a prolonged event with a peak

1401   around 1900 Ma. Geosystems and Geoenvironment, 1(2), 100036.

1402   ● Lehnert, K. A., Markey, K., Ji, P., Evans, C., & Zeigler, R. (2019, March). The

1403   Astromaterials Data System: Transforming Access to Planetary Sample Data. In

1404   Lunar and Planetary Science Conference (No. 2132, p. 2799).

1405   ● Lehnert, K., Su, Y., Langmuir, C. H., Sarbas, B., & Nohl, U. (2000). A global

1406   geochemical database structure for rocks. Geochemistry, Geophysics,

1407   Geosystems, 1(5). https://doi.org/10.1029/1999GC000026

1408   ● Liu C, Runyon SE, Knoll AH, Hazen RM. The same and not the same: Ore geology,

1409   mineralogy and geochemistry of Rodinia assembly versus other supercontinents.

1410   Earth Science Reviews. 2019 May 13.

1411   ● Liu, B., Wei, Y., Zhang, Y., & Yang, Q. (2017). Deep Neural Networks for High

1412   Dimension, Low Sample Size Data. In Proceedings of the Twenty-Sixth

1413   International Joint Conference on Artificial Intelligence. Twenty-Sixth

1414   International Joint Conference on Artificial Intelligence. International Joint

Conferences on Artificial Intelligence Organization. https://doi.org/10.24963/ijcai.2017/318

● Liu, X.-M., L. C. Kah, A. H. Knoll, H. Cui, C. Wang, A. Bekker, and R. M. Hazen, A persistently low level of atmospheric oxygen in Earth's middle age, Nature Communications, 12, 351, 2021.

● Lohr, S. (2012). The age of big data. New York Times, 11(2012).

● Lord, P., Bechhofer, S., Wilkinson, M. D., Schiltz, G., Gessler, D., Hull, D., ... & Stein, L. (2004, November). Applying semantic web services to bioinformatics: Experiences gained, lessons learnt. In International Semantic Web Conference (pp. 350-364). Springer, Berlin, Heidelberg.

● Ma X, Hummer D, Golden JJ, Fox PA, Hazen RM, Morrison SM, Downs RT, Madhikarmi BL, Wang C, Meyer MB (2017) Using Visual Exploratory Data Analysis to Facilitate Collaboration and Hypothesis Generation in Cross-Disciplinary Research. ISPRS International Journal of Geo-Information 6(11):368

● Ma, X. (2021). Data Science for Geoscience: Recent Progress and Future Trends from the Perspective of a Data Life Cycle. https://doi.org/10.31223/X55S4D

● Maynard-Casely, H. E., Cable, M. L., Malaska, M. J., Vu, T. H., Choukroun, M., & Hodyss, R. (2018). Prospects for mineralogy on Titan. American Mineralogist, 103(3), 343–349. https://doi.org/10.2138/am-2018-6259

● McGuinness KN, Klau GW, Morrison SM, Moore EK, Seipp J, Falkowski PG, Nanda V (2022) Evaluating mineral lattices as evolutionary proxies for metalloprotein evolution, Origins of Life and Evolution of Biospheres (In Review)

1437   ● McHenry K, Bobak M, Coakley K, Fils D, Gatzke L, Zhang B, Kooper R, Richard S,

1438   Valentine D, Zaslavsky I, Shepherd, A & Lingerfelt E., 2021. GeoCODES. EarthCube.

1439   https://geocodes.earthcube.org.

1440   ● Meadows, V. S., Reinhard, C. T., Arney, G. N., Parenteau, M. N., Schwieterman, E.

1441   W., Domagal-Goldman, S. D., … Grenfell, J. L. (2018). Exoplanet Biosignatures:

1442   Understanding Oxygen as a Biosignature in the Context of Its Environment.

1443   Astrobiology, 18(6), 630–662. https://doi.org/10.1089/ast.2017.1727

1444   ● Morrison SM, Downs RT, Blake DF, Prabhu A, Eleish A, Vaniman DT, Ming DW,

1445   Rampe EB, Bristow TF, Achilles CN, Chipera SJ, Yen AS, Morris RV, Treiman AH,

1446   Hazen RM, Sarrazin PC, Gellert R, Fendrich KV, Morookian JM, Farmer JD, Des

1447   Marais DJ, Craig PI (2018b) Relationships between unit-cell parameters and

1448   composition for rock-forming minerals on Earth, Mars, and other extraterrestrial

1449   bodies, American Mineralogist, 103(6), 848-856

1450   ● Morrison SM, Prabhu A, Eleish A, Narkar S, Fox P, Golden JJ, Downs RT, Perry S,

1451   Burns PC, Ralph J, Hazen RM (2022) Mineral Association Analysis: Predicting

1452   unknown mineral occurrences based on association rule learning (in prep)

1453   ● Morrison SM, Downs RT, Blake DF, Vaniman DT, Ming DW, Rampe EB, Bristow TF,

1454   Achilles CN, Chipera SJ, Yen AS, Morris RV, Treiman AH, Hazen RM, Sarrazin PC,

1455   Fendrich KV, Morookian JM, Farmer JD, Des Marais DJ, Craig PI (2018a) Crystal

1456   chemistry of martian minerals from Bradbury Landing through Naukluft Plateau,

1457   Gale crater, Mars, American Mineralogist, 103(6), 857-871

1458 ● Morrison SM, Hazen RM, Prabhu A, Williams J, Eleish A, Fox P (2021) Mineral

1459 network analysis: Exploring geological, geochemical, and biological patterns in

1460 mineralization via multidimensional analysis, Geological Society of America

1461 Abstracts with Programs. Vol 53, No. 6, 2021, doi: 10.1130/abs/2021AM-370437

1462 ● Morrison, S. M., Pan, F., Gagné, O. C., Prabhu, A., Eleish, A., Fox, P. A., ... & Hazen,

1463 R. (2018c). Predicting Multi-Component Mineral Compositions in Gale crater,

1464 Mars with Label Distribution Learning. In AGU Fall Meeting 2018. AGU.

1465 ● Morrison SM, Liu C, Eleish A, Prabhu A, Li C, Ralph J, Downs RT, Golden JJ, Fox P,

1466 Hummer DR, Meyer MB, and Hazen RM (2017) Network analysis of mineralogical

1467 systems. American Mineralogist 102. https://doi.org/10.2138/am-2017-

1468 6104ccbyncnd

1469 ● Morrison, S. M., Buongiorno, J., Downs, R. T., Eleish, A., Fox, P., Giovannelli, D.,

1470 Golden, J. J., Hummer, D. R., Hystad, G., Kellogg, L. H., Kreylos, O., Krivovichev, S.

1471 V., Liu, C., Merdith, A., Prabhu, A., Ralph, J., Runyon, S. E., Zahirovic, S., & Hazen,

1472 R. M. (2020). Exploring Carbon Mineral Systems: Recent Advances in C Mineral

1473 Evolution, Mineral Ecology, and Network Analysis. Frontiers in Earth Science, 8.

1474 https://doi.org/10.3389/feart.2020.0020

1475 ● Moore, E., Jelen, B., Giovannelli, D. et al. Metal availability and the expanding

1476 network of microbial metabolisms in the Archaean eon. Nature Geosci 10, 629–

1477 636 (2017). https://doi.org/10.1038/ngeo3006

1478 ● Murchie, S. L., Mustard, J. F., Ehlmann, B. L., Milliken, R. E., Bishop, J. L., McKeown,

1479 N. K., … Bibring, J.-P. (2009). A synthesis of Martian aqueous mineralogy after 1

Mars year of observations from the Mars Reconnaissance Orbiter. Journal of Geophysical Research, 114, E00D06. https://doi.org/10.1029/2009JE003342

- Murray H.H. (1995) Industrial Minerals—Key to Economic Development. In: Miller R.L., Escalante G., Reinemund J.A., Bergin M.J. (eds) Energy and Mineral Potential of the Central American-Caribbean Region. Circum-Pacific Council for Energy and Mineral Resources Earth Science Series, vol 16. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-79476-6_46

- Namur, O., & Charlier, B. (2017). Silicate mineralogy at the surface of Mercury. Nature Geoscience, 10(1), 9–13. https://doi.org/10.1038/ngeo2860

- Nance, R.D., Murphy, J.B., and Santosh, M., 2014, The supercontinent cycle: A retrospective essay: Gondwana Research, v. 25, p. 4-29.

- NASEM (National Academies of Sciences, Engineering, and Medicine), 2015. Enhancing the Effectiveness of Team Science. The National Academies Press, Washington, DC, 268pp. doi: 10.17226/19007.

- Nazabal, A., Olmos, P. M., Ghahramani, Z., & Valera, I. (2020). Handling incomplete heterogeneous data using vaes. Pattern Recognition, 107, 107501.

- Needham, Joseph (1986). Science and Civilization in China: Volume 3. Taipei: Caves Books, Ltd

- Nitschke, W., McGlynn, S. E., Milner-White, E. J., & Russell, M. J. (2013). On the antiquity of metalloenzymes and their substrates in bioenergetics. Biochimica et Biophysica Acta (BBA) - Bioenergetics, 1827(8–9), 871–881. https://doi.org/10.1016/j.bbabio.2013.02.008

1502   ● Novikov, Y., & Copley, S. D. (2013). Reactivity landscape of pyruvate under

1503   simulated hydrothermal vent conditions. Proceedings of the National Academy of

1504   Sciences, 110(33), 13283–13288. https://doi.org/10.1073/pnas.1304923110

1505   ● Noy, N., Burgess, M., Brickley, D., 2019. Google Dataset Search: Building a search

1506   engine for datasets in an open Web ecosystem. In: Proceedings of The 2019 World

1507   Wide Web Conference, San Francisco CA, pp.1365-1375.

1508   ● Postberg, F., Kempf, S., Hillier, J. K., Srama, R., Green, S. F., McBride, N., & Grün, E.

1509   (2008). The E-ring in the vicinity of Enceladus. In Icarus (Vol. 193, Issue 2, pp. 438–

1510   454). Elsevier BV. https://doi.org/10.1016/j.icarus.2007.09.001

1511   ● Prabhu A. (2018) Informatics. In: Schintler L., McNeely C. (eds) Encyclopedia of Big

1512   Data. Springer, Cham. https://doi.org/10.1007/978-3-319-32001-4_372-1

1513   ● Prabhu A., Fox P. (2021) Reproducible Workflow. In: Daya Sagar B., Cheng Q.,

1514   McKinley J., Agterberg F. (eds) Encyclopedia of Mathematical Geosciences.

1515   Encyclopedia   of   Earth   Sciences   Series.   Springer,   Cham.

1516   https://doi.org/10.1007/978-3-030-26050-7_277-1

1517   ● Prabhu, A, Morrison, SM, Eleish, A, et al. Global earth mineral inventory: A data

1518   legacy. Geosci. Data J. 2021a; 8: 74– 89. https://doi.org/10.1002/gdj3.106

1519   ● Prabhu, A., Morrison, S. M., & Giovannelli, D. (2021b, December). A new way to

1520   evaluate association rule mining methods and its applicability to mineral

1521   association   analysis.   In   AGU   Fall   Meeting   2021.   AGU.

1522   https://doi.org/10.1002/essoar.10509679.1

- Prabhu, A., Morrison, S. M., Eleish, A., Narkar, S., Fox, P. A., Golden, J. J., ... & Hazen, R. (2019, December). Predicting unknown mineral localities based on mineral associations. In AGU Fall Meeting 2019. AGU.

- Prettyman, T. H., Yamashita, N., Ammannito, E., Ehlmann, B. L., McSween, H. Y., Mittlefehldt, D. W., ... Russell, C. T. (2019). Elemental composition and mineralogy of Vesta and Ceres: Distribution and origins of hydrogen-bearing species. Icarus, 318, 42–55. https://doi.org/10.1016/j.icarus.2018.04.032

- Putirka, K. D., Dorn, C., Hinkel, N. R., & Unterborn, C. T. (2021). Compositional diversity of rocky exoplanets. Elements: An International Magazine of Mineralogy, Geochemistry, and Petrology, 17(4), 235-240.

- Ramachandran, R., Bugbee, K., & Murphy, K. (2021). From Open Data to Open Science. Earth and Space Science, 8(5). https://doi.org/10.1029/2020EA001562

- Rampe, E. B., Lapotre, M. G. A., Bristow, T. F., Arvidson, R. E., Morris, R. V., Achilles, C. N., Weitz, C., Blake, D. F., Ming, D. W., Morrison, S. M., Vaniman, D. T., Chipera, S. J., Downs, R. T., Grotzinger, J. P., Hazen, R. M., Peretyazhko, T. S., Sutter, B., Tu, V., Yen, A. S., ... Treiman, A. H. (2018). Sand Mineralogy Within the Bagnold Dunes, Gale Crater, as Observed In Situ and From Orbit. In Geophysical Research Letters (Vol. 45, Issue 18, pp. 9488–9497). American Geophysical Union (AGU). https://doi.org/10.1029/2018gl079073

- Ramsdell L S (1925) The crystal structures of some metallic sulfides American Mineralogist 10 281-304

1544 ● Rogers, K. L., Thomson, B. L., Colwell, F. S., Eleish, A., Fontaine, K. S., Fox, P. A., …

1545 & Twing, K. I. (2018, December). The Census of Deep Life: Metadata Then and

1546 Now. In AGU Fall Meeting Abstracts (Vol. 2018, pp. IN53C-0629).

1547 http://doi.org/10.13140/RG.2.2.16160.30720

1548 ● Russell, M. (2018). Green Rust: The Simple Organizing 'Seed' of All Life? Life, 8(3),

1549 35. https://doi.org/10.3390/life8030035

1550 ● Russell, M.J.; Hall, A. The emergence of life from iron monosulphide bubbles at a

1551 submarine hydrothermal redox and pH front. J. Geol. Soc. 1997, 154, 377–402.

1552 ● Sandve, G. K., Nekrutenko, A., Taylor, J., & Hovig, E. (2013). Ten Simple Rules for

1553 Reproducible Computational Research. PLoS Computational Biology, 9(10),

1554 e1003285. https://doi.org/10.1371/journal.pcbi.1003285

1555 ● Shah, K., Salunke, A., Dongare, S., & Antala, K. (2017, March). Recommender

1556 systems: An overview of different approaches to recommendations. In 2017

1557 International Conference on Innovations in Information, Embedded and

1558 Communication Systems (ICIIECS) (pp. 1-4). IEEE.

1559 ● Shannon, R. C., Lafuente, B., Shannon, R. D., Downs, R. T., & Fischer, R. X. (2017).

1560 Refractive indices of minerals and synthetic compounds. American Mineralogist:

1561 Journal of Earth and Planetary Materials, 102(9), 1906-1914.

1562 ● Shepperd, M., & Cartwright, M. (2001). Predicting with sparse data. IEEE

1563 Transactions on Software Engineering, 27(11), 987-998.

1564   ● Shi, L., Dong, H., Reguera, G., Beyenal, H., Lu, A., Liu, J., … & Fredrickson, J. K.

1565   (2016). Extracellular electron transfer mechanisms between microorganisms and

1566   minerals. Nature Reviews Microbiology, 14(10), 651-662.

1567   ● Sinha, A. K. (Ed.). (2006). Geoinformatics: data to knowledge (Vol. 397). Geological

1568   Society of America.

1569   ● Sinha, A. K., Malik, Z., Rezgui, A., Barnes, C. G., Lin, K., Heiken, G., … & Zimmerman,

1570   H. (2010). Geoinformatics: transforming data to knowledge for geosciences. GSA

1571   Today, 20(12), 4-10.

1572   ● Statnikov, A., Wang, L. & Aliferis, C.F. A comprehensive comparison of random

1573   forests and support vector machines for microarray-based cancer classification.

1574   BMC Bioinformatics 9, 319 (2008). https://doi.org/10.1186/1471-2105-9-319

1575   ● Strunz, H. and Tennyson, C., 1941. Mineralogische tabellen. Akademische

1576   Verlagsgesellschaft Becker & Erler Kom.-Ges..

1577   ● Sverjensky, D. A., & Lee, N. (2010). The Great Oxidation Event and Mineral

1578   Diversification. Elements, 6(1), 31–36. https://doi.org/10.2113/gselements.6.1.31

1579   ● Sweeting, M.J., Sutton, A.J., & Lambert, P.C. (2004). What to add to nothing? Use

1580   and avoidance of continuity corrections in meta-analysis of sparse data. Statistics

1581   in medicine, 23(9), 1351-1375.

1582   ● Tomašev N., Radovanović M. (2016) Clustering Evaluation in High-Dimensional

1583   Data. In: Celebi M., Aydin K. (eds) Unsupervised Learning Algorithms. Springer,

1584   Cham. https://doi.org/10.1007/978-3-319-24211-8_4

1585   ●   Treiman, A. H., & Bullock, M. A. (2012). Mineral reaction buffering of Venus'

1586   atmosphere: A thermochemical constraint and implications for Venus-like planets.

1587   Icarus, 217(2), 534–541. https://doi.org/10.1016/j.icarus.2011.08.019

1588   ●   Unterborn, C. T., & Panero, W. R. (2019). The pressure and temperature limits of

1589   likely rocky exoplanets. Journal of Geophysical Research: Planets, 124(7), 1704-

1590   1716.

1591   ●   Unterborn, C. T., Dismukes, E. E., & Panero, W. R. (2016). Scaling the Earth: a

1592   sensitivity analysis of terrestrial exoplanetary interior models. The Astrophysical

1593   Journal, 819(1), 32.

1594   ●   Uzuner, Ö. (2009). Recognizing obesity and comorbidities in sparse data. Journal

1595   of the American Medical Informatics Association, 16(4), 561-570.

1596   ●   Voice, P.J., Kowalewski, M., and Eriksson, K.A., 2011, Quantifying the timing and

1597   rate of crustal evolution: Global compilation of radiometrically dated detrital

1598   zircon grains. The Journal of Geology, v. 119, p. 109-126.

1599   ●   Wachter, S. Data protection in the age of big data. Nat Electron 2, 6–7 (2019).

1600   https://doi.org/10.1038/s41928-018-0193-y

1601   ●   Waite, J. H., Glein, C. R., Perryman, R. S., Teolis, B. D., Magee, B. A., Miller, G., …

1602   Bolton, S. J. (2017). Cassini finds molecular hydrogen in the Enceladus plume:

1603   Evidence for hydrothermal processes. Science, 356(6334), 155–159.

1604   https://doi.org/10.1126/science.aai8703

1605   ●   Wang, C., Hazen, R.M., Cheng, Q., Stephenson, M.H., Zhou, C., Fox, P., Shen, S.Z.,

1606   Oberhänsli, R., Hou, Z., Ma, X. and Feng, Z., 2021. The Deep-Time Digital Earth

1607   program: data-driven discovery in geosciences. National Science Review, 8(9),

1608   https://doi.org/10.1093/nsr/nwab027

1609 ● Wang, L. (2017). Heterogeneous data and big data analytics. Automatic Control

1610   and Information Sciences, 3(1), 8-15.

1611 ● Wiederhold G. (1999) Mediation to Deal with Heterogeneous Data Sources. In:

1612   Včkovski A., Brassel K.E., Schek HJ. (eds) Interoperating Geographic Information

1613   Systems. INTEROP 1999. Lecture Notes in Computer Science, vol 1580. Springer,

1614   Berlin, Heidelberg. https://doi.org/10.1007/10703121_1

1615 ● Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak,

1616   A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E., & Bouwman, J.,

1617   2016. The FAIR Guiding Principles for scientific data management and

1618   stewardship. Scientific data, 3(1), 1-9.

1619 ● Wise, A. F., & Shaffer, D. W. (2015). Why theory matters more than ever in the age

1620   of big data. Journal of Learning Analytics, 2(2), 5-13.

1621 ● Wyborn, L. A., Lehnert, K., & Klump, J. F. (2021, December). The Future of X-

1622   informatics Lies in Collaborative Convergence: An Exemplar from the Global

1623   OneGeochemistry Initiative. In AGU Fall Meeting 2021. AGU.

1624 ● Yang, H., Sun, H.J. and Downs, R.T., 2011. Hazenite, $KNaMg_2 (PO_4) 2· 14H_2O$, a

1625   new biologically related phosphate mineral, from Mono Lake, California, USA.

1626   American Mineralogist, 96(4), pp.675-681

1627 ● Young, R. A. (1993). The rietveld method (Vol. 5, pp. 1-38).

1628   ● Yu, S. (2016). Big privacy: Challenges and opportunities of privacy study in the age

1629   of big data. IEEE access, 4, 2751-2763.

1630   ● Zhang S, Morrison, S.M., Prabhu, A., Ma, C., Huang, F., Gregory, D., Large, R.R. and

1631   Hazen, R., 2019. Natural clustering of pyrite with implications for its formational

1632   environment. AGU FM, 2019, EP23D-2284.

1633   ● Zhao, D., Bartlett, S., & Yung, Y. L. (2020). Quantifying Mineral-Ligand Structural

1634   Similarities: Bridging the Geological World of Minerals with the Biological World

1635   of Enzymes. Life, 10(12), 338. https://doi.org/10.3390/life10120338

1636   ● Zhou J, Gandomi AH, Chen F, Holzinger A. Evaluating the Quality of Machine

1637   Learning Explanations: A Survey on Methods and Metrics. Electronics. 2021;

1638   10(5):593. https://doi.org/10.3390/electronics10050593

1639   ● Zolotov, M. Y. (2018). Gas–Solid Interactions on Venus and Other Solar System

1640   Bodies. Reviews in Mineralogy and Geochemistry, 84(1), 351–392.

1641   https://doi.org/10.2138/rmg.2018.84.10

1642