Big Data in regional-scale mineral research: observations from multi-method studies and public/industry databases

Big Data em pesquisa mineral em escala regional: observações a partir de estudos multimétodos e bases de dados públicos/industriais

João Gabriel Motta, Ph.D., M.Sc.
Postdoctoral Research Felow
Camborne School of Mines - University of Exeter

Section 8. TECHNOLOGICAL INNOVATIONS IN PROSPECTIVE ANALYSIS OF MINERAL PROJECTS
Tuesday, November 29, 2022: 17:10-17:30

Big Data: "A wide-ranging field of research that deals with *large datasets*. A **key challenge** in big data is working out ***how to generate useful insights from the data*** without inappropriately compromising the privacy of the people to whom the data relates." [*https://www.turing.ac.uk/news/data-science-and-ai-glossary*]

The V's in Big Data *[https://www.oracle.com/uk/big-data/what-is-big-data/]*

Volume: The amount of data matters. With big data, you'll have to process **high volumes** of low-density, unstructured data.

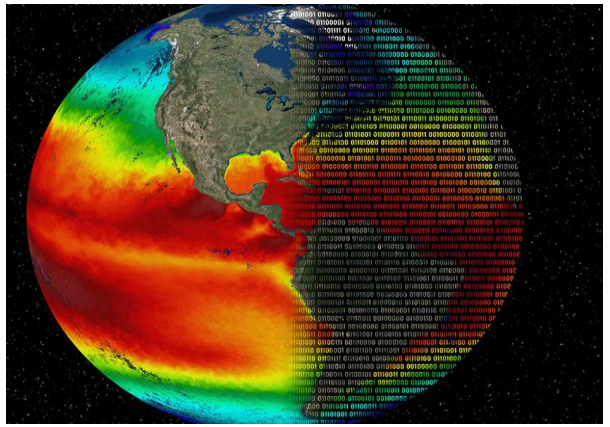
Velocity: Velocity is the fast rate at which data **is received and (perhaps) acted on.** Normally, the highest velocity of data streams directly into memory versus being written to disk.

Variety: Variety refers to the many **types of data that are available**. Traditional data types were structured and fit neatly in a relational database.

Value: Data has **intrinsic value**. But it's of no use until that value is **discovered**.

Veracity: How **truthful** is your data—and how much can you **rely** on it?

Legacy vs. **historical** data: obsolete formats and not accessible promptly vs. not maintained and not easy to update



[https://eos.org/editors-vox/analyzing-big-earth-data-progress-challenges-opportunities]

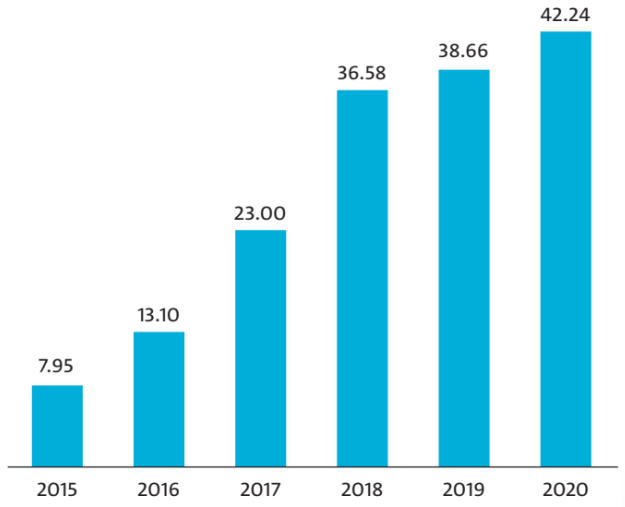


Figure 2. Private investment in artificial intelligence companies worldwide (billions of US dollars).

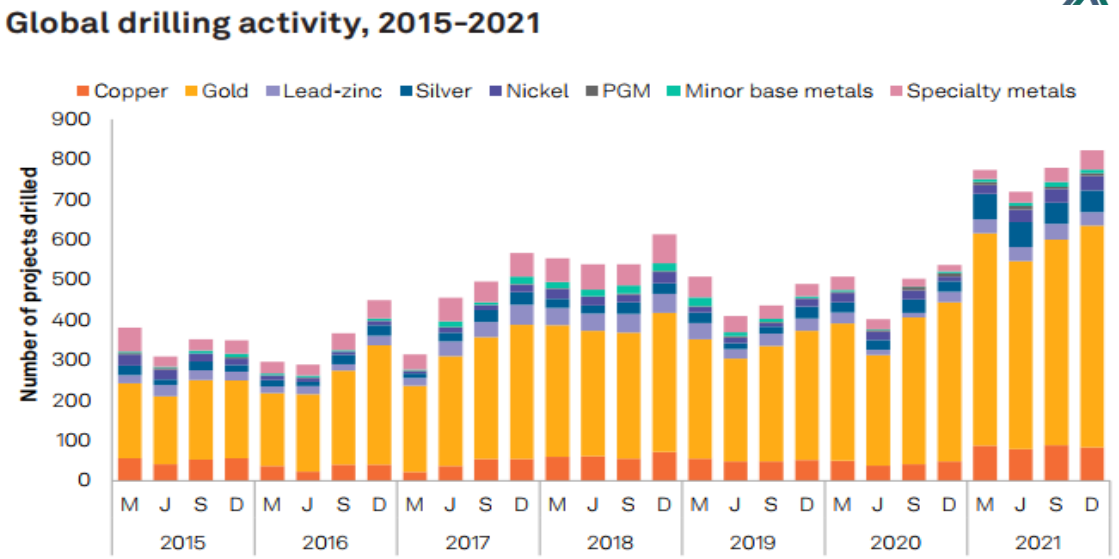
CSIRO (2022) – Report - Artificial intelligence for science

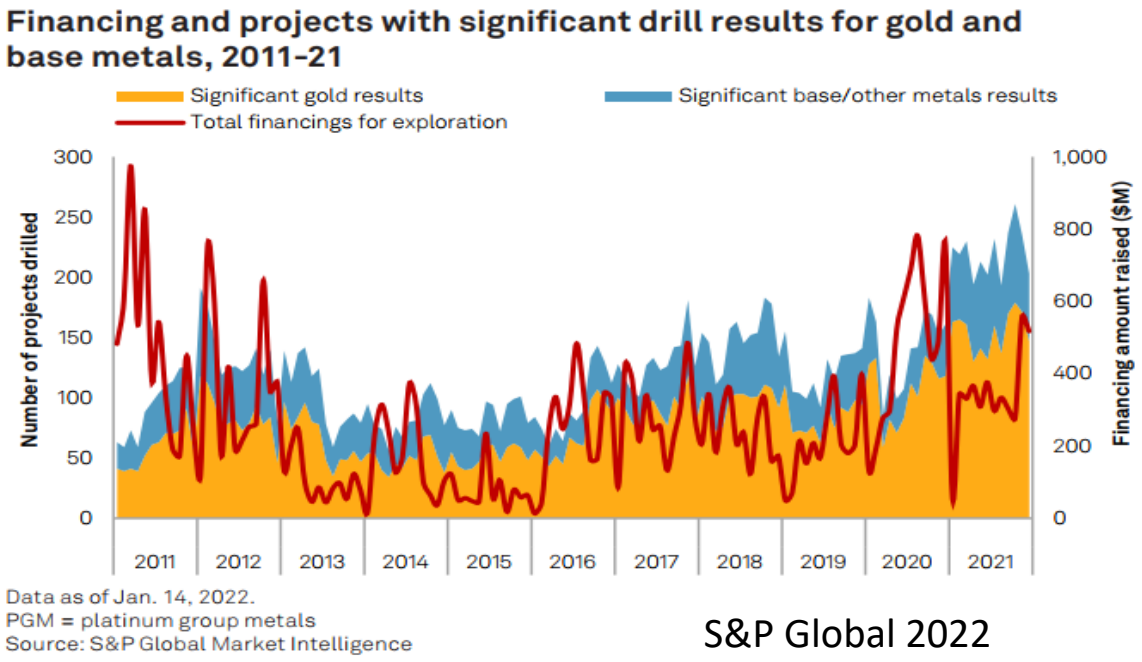
In the context of mineral/geological exploration

We are **walking towards** Big data in mineral/geological exploration

- Quite often we don't really deal with **BIG DATA** in **mineral exploration** as for the definition **(***V's are lacking***)**
- Problems and uncertainties often render (historical/new) data **not useful**, or at least **reduce the number of useful entries/variables** to critical levels. **Not Big Data anymore.**

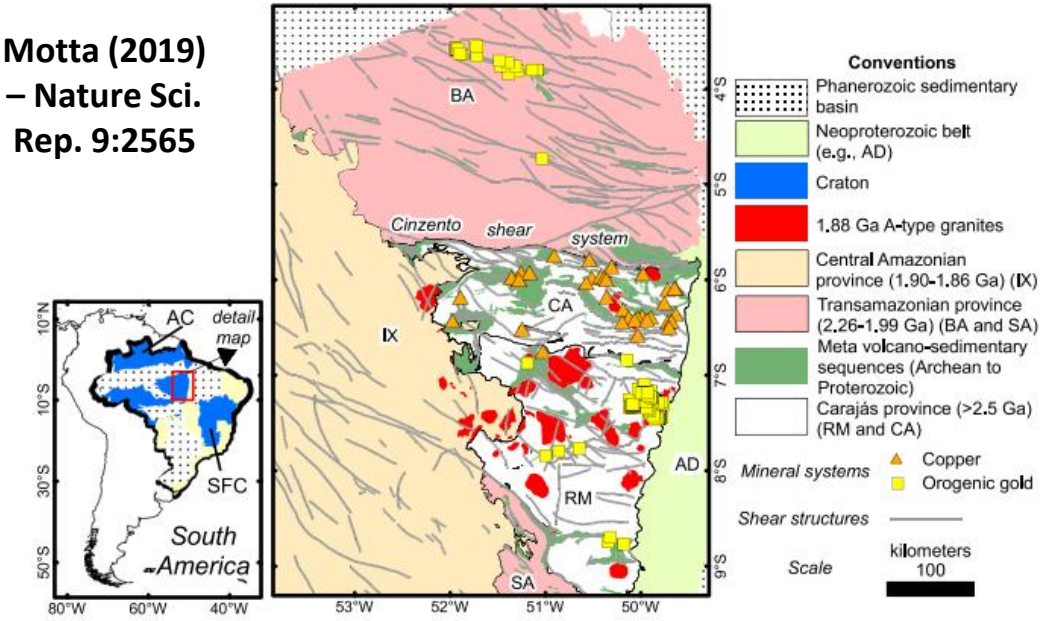
Industry is drilling and exploring more … is this **following** the V's in/for Big data?

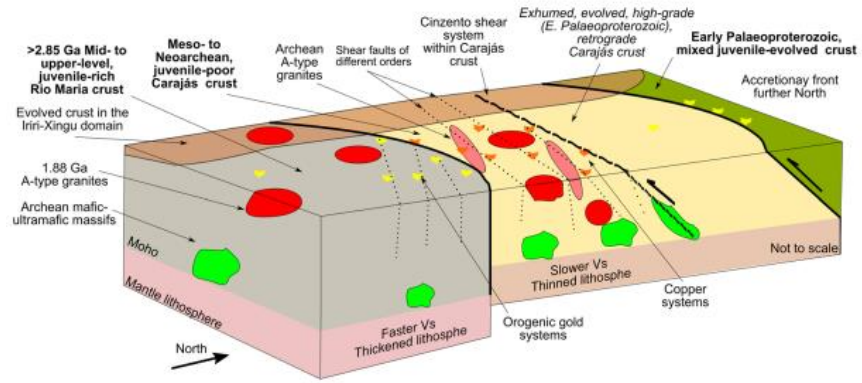


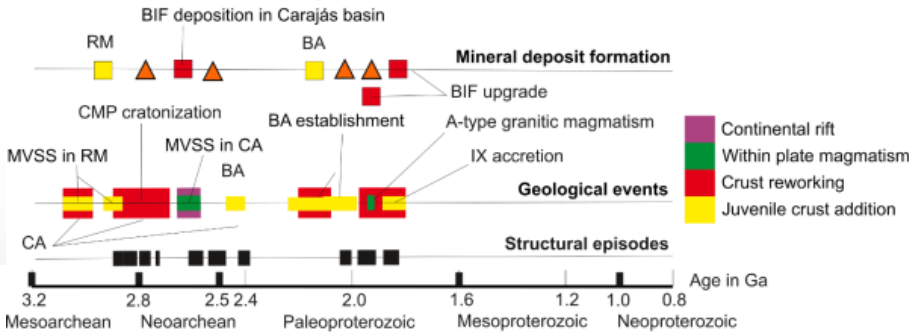


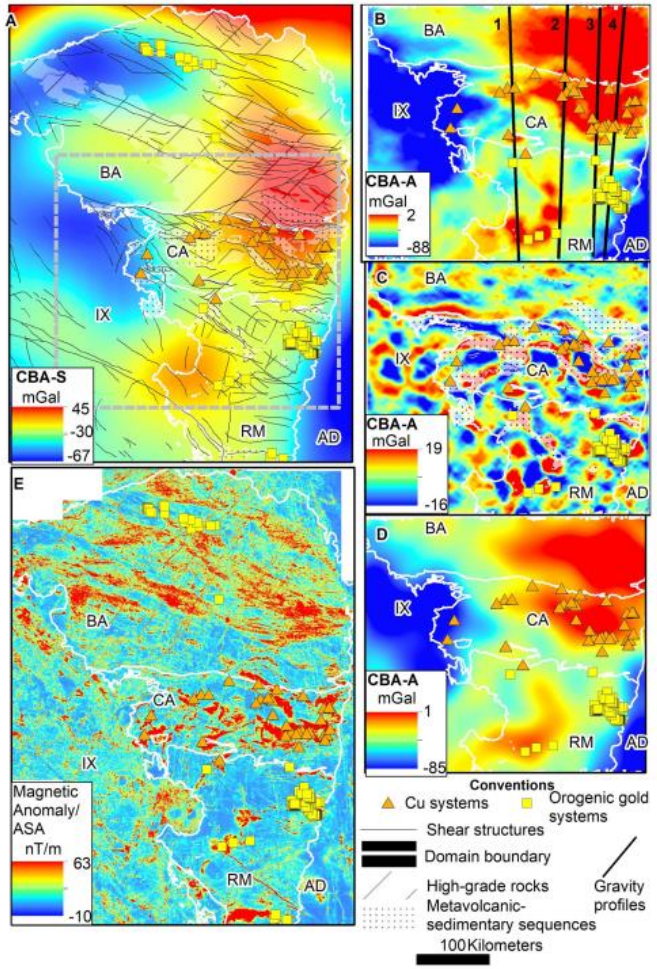
S&P Global 2022

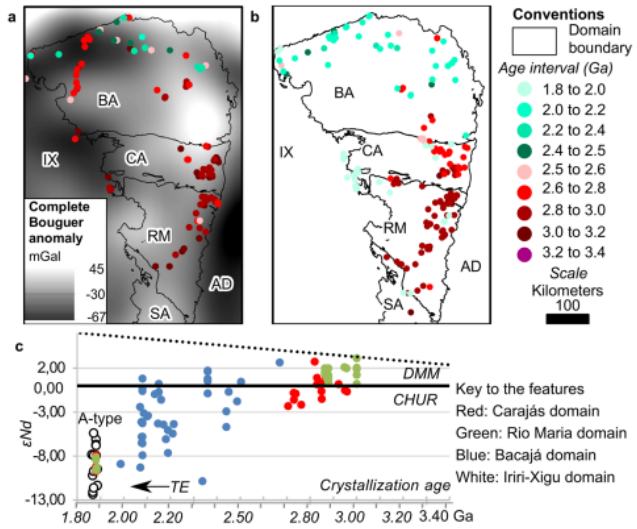
Motta (2019) – Nature Sci. Rep. 9:2565

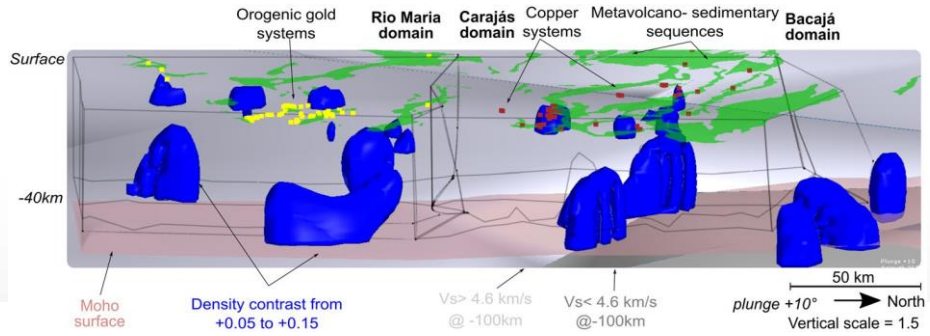


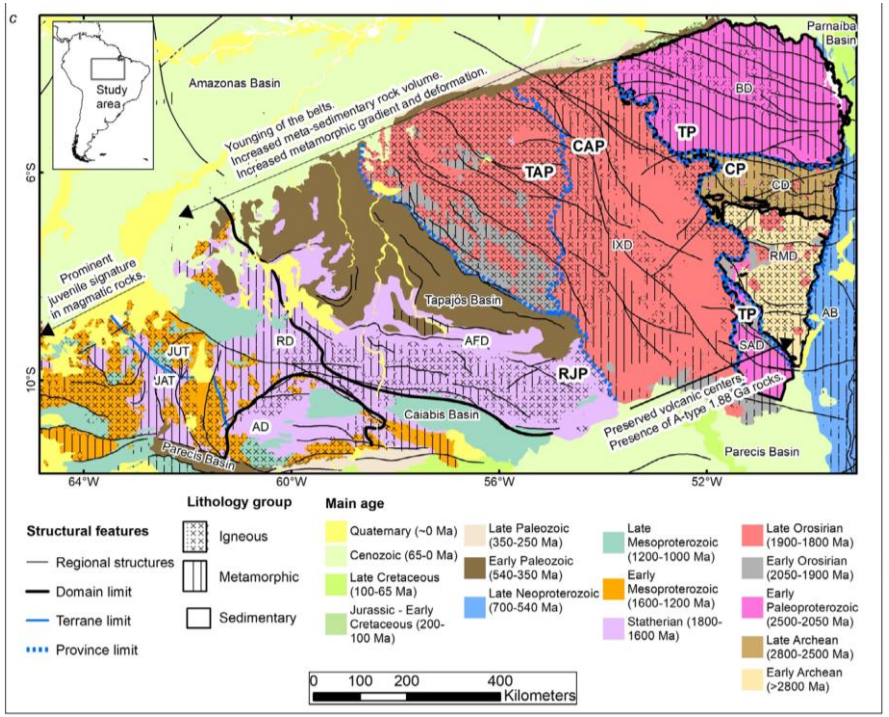


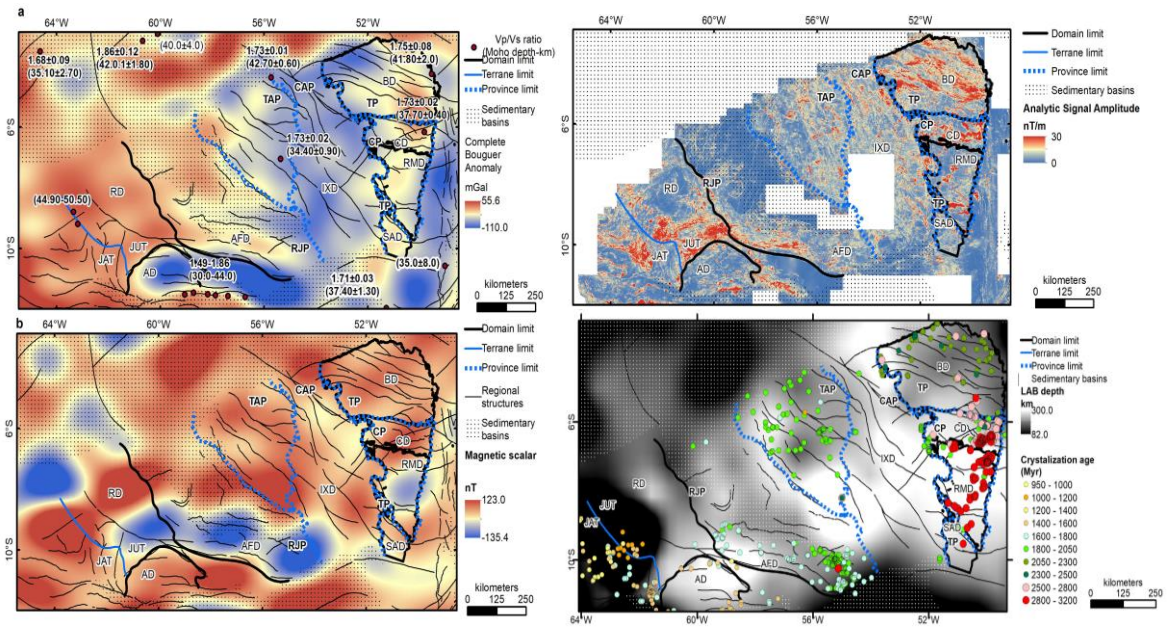




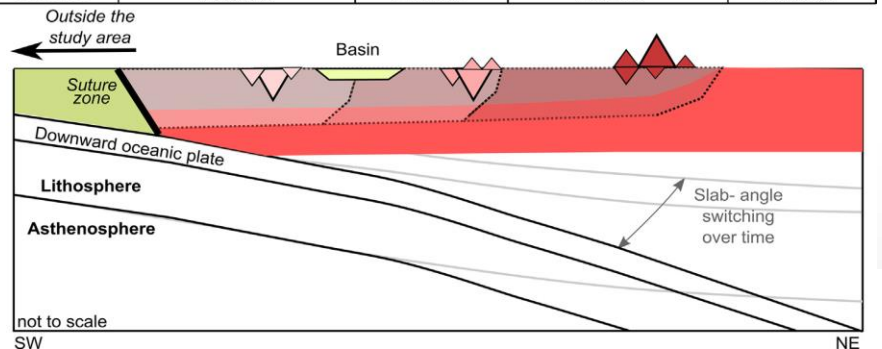


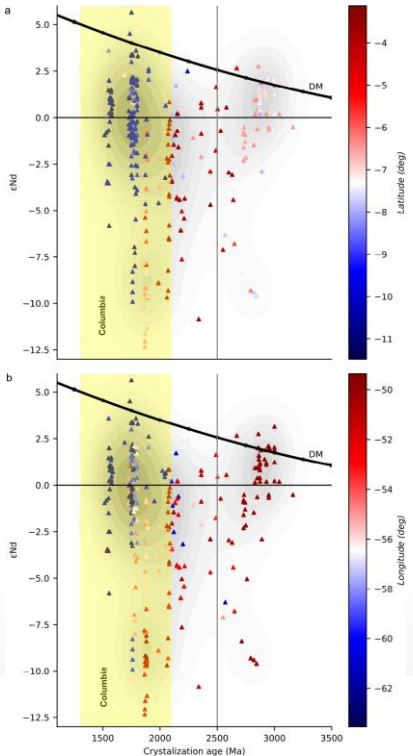


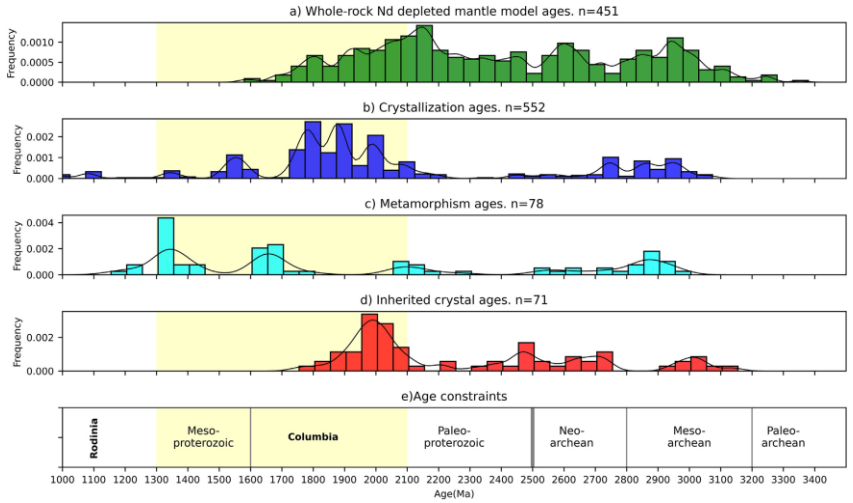




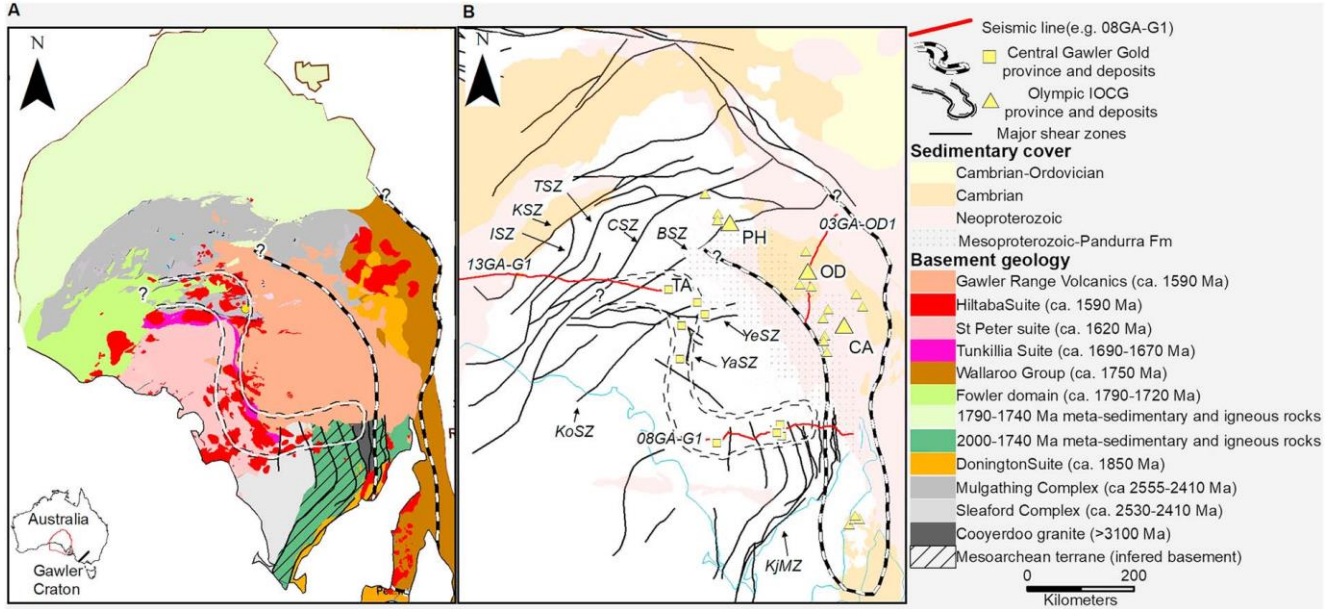
Role	Subducting plate	Uplifted retrowedge			Stationary retrowedge
Position to the subduction	Accretionary front	Supra-subduction zone Accretionary belt			Distal back-arc to intracontinental settings
Charac-teristics	*Younging towards the west* *Increasing deformation* *Increasing supra-crustal rock volume and juvenile material* *Increasing metamorphic grade*				Relatively stable crust
Tectonic province	*Sunsás Province*	*Rondônia-Juruena Province*	*Tapajós-Parima Province*	*Central Amazon Province*	*Carajás Province*

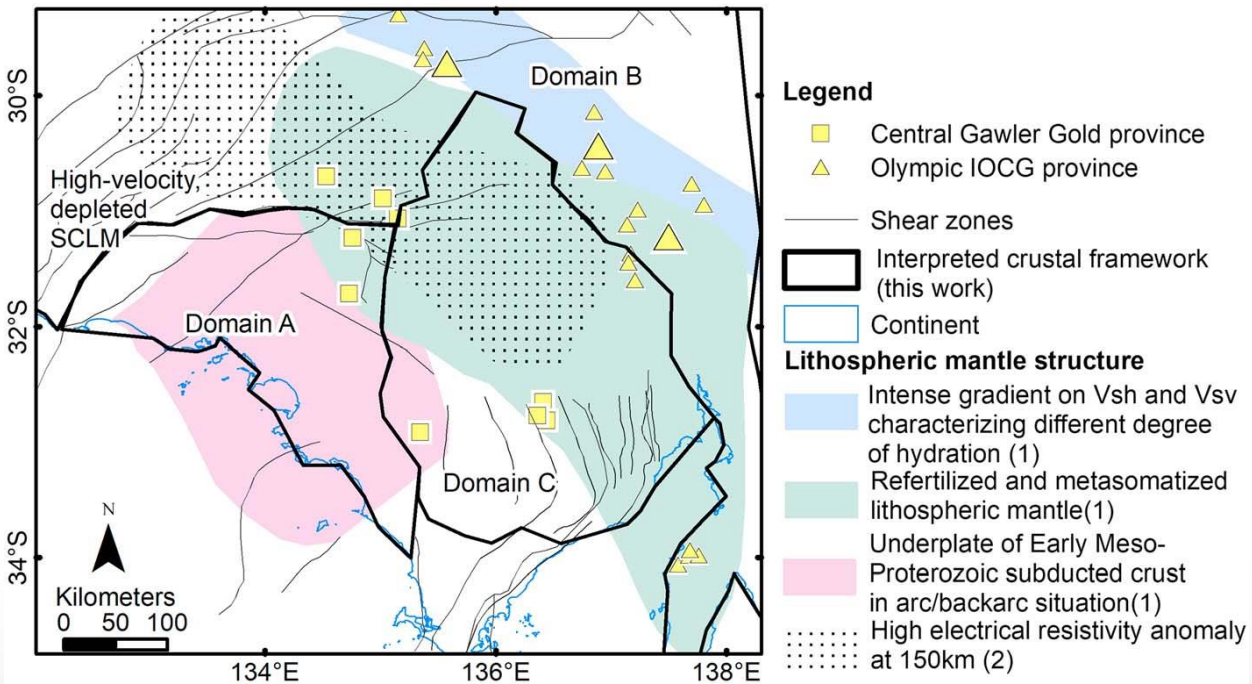


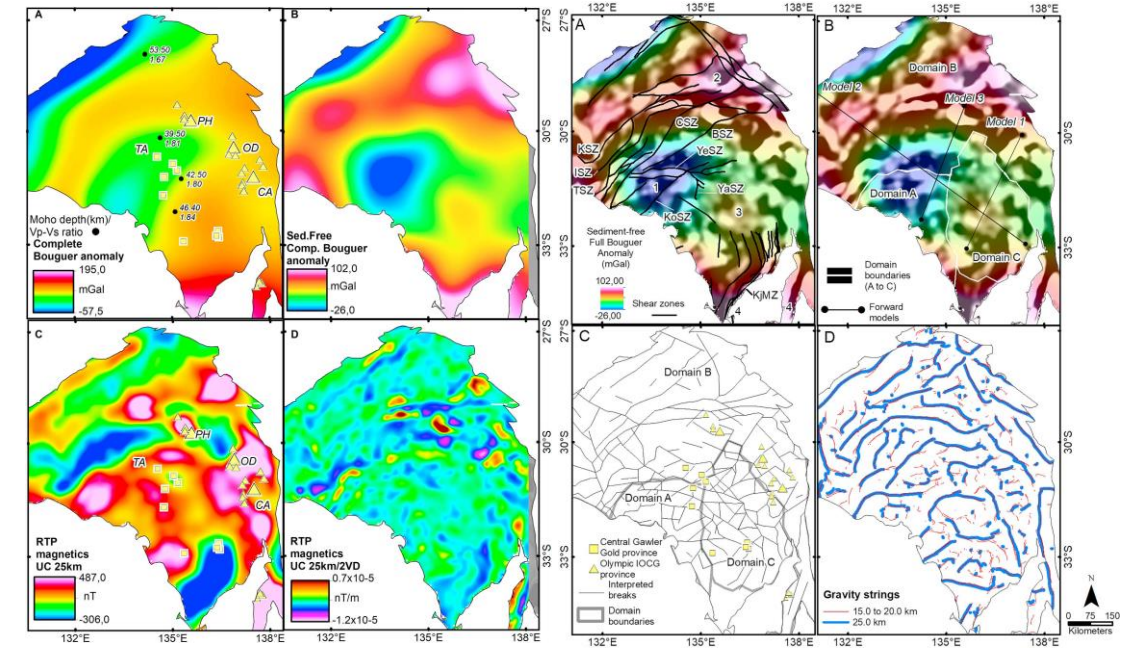


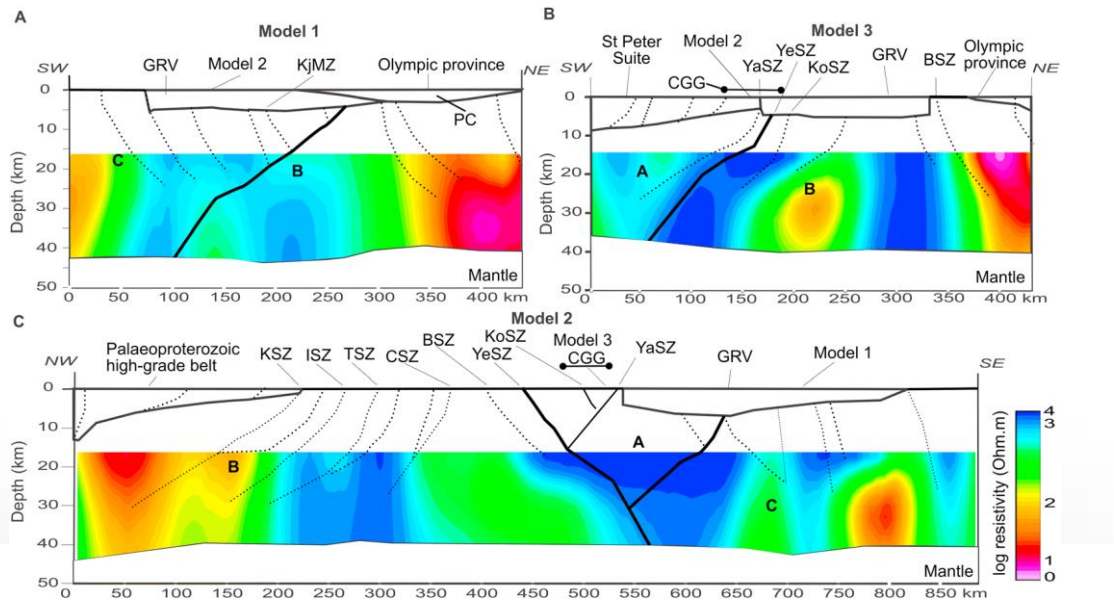


Motta (2022) – Geosc. Frontiers 13









Motta (2019) – J. Geoph. Res. 124 (3)

Example: From potential Big to 'regular' data

It was a compilation of proprietary (industry) historical **drilling** data.

- No documentation was provided, although quite simple in structure
- ALL sheets
  - Plethora of NaNs/Nulls
  - Missing units on headers
  - Unit conversion necessary (a problem if it is not noticed)
- Survey sheet
  - IDs differ non-systematically from collar
  - (some) Missing ID, location flags, dip, azimuth, reading depth
- Lithology and Stratigraphy
  - No keys provided
  - Inconsistent naming across geographic areas
- Two collar sheets provided
  - inconsistent/different fields
  - Different number of entries (varying in thousands)
  - entry ID varies between sheets
  - Missing key info (e.g., X, Y)

Selected collar table:

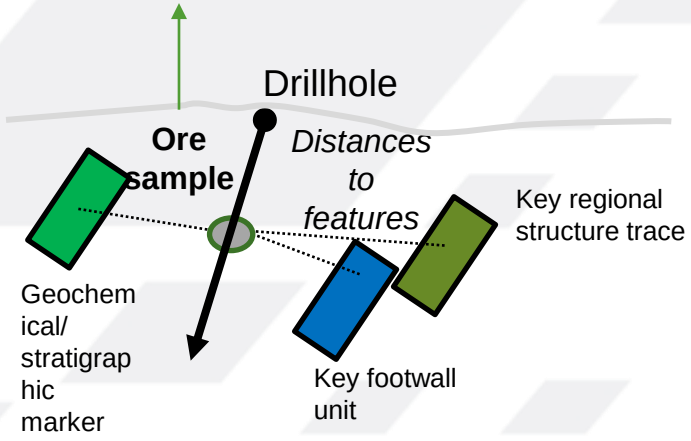
- started with **>4,500 DHs**
- Ends up with **~1,200 DHs** that can be used without imputation and guess-work

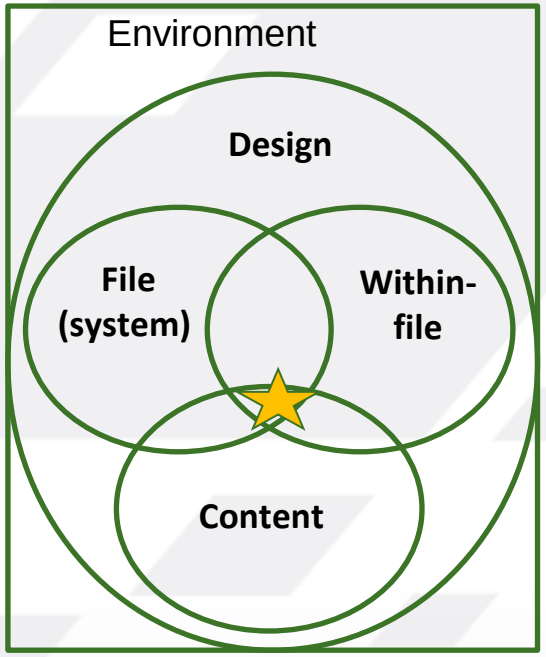
- Assay sheet
  - Format was not friendly for file size
  - >30 quantitative fields
  - Fields with potential relationships (e.g., imputation)
  - Entries missing key info
  - Some attributes are ambiguous/uncertain
  - Fields contemplating 'sum' attributes
    - What is summed in here?
    - Analytical Sum vs. user-created sum?
- Started with **more than 750,000 entries**
- Two sample groups identified
  - 3 attributes: **>113 k entries (DH ~ 1,200)**
  - 10 attributes: **>32k entries (~160 DHs)**

Final scene:

- **Poor documentation and delivery can drastically reduce reliability**
- **Took >2 months to 'break' data and understand deficiencies**
- **Took >2 months to devise mitigation plan**

Has: Litho, Strat., AH, HSI, Geochem, Phys. Prop….





Pathologies in (historical, potentially big) exploration data and its uses

Describe the issues found to pave way to mitigation.

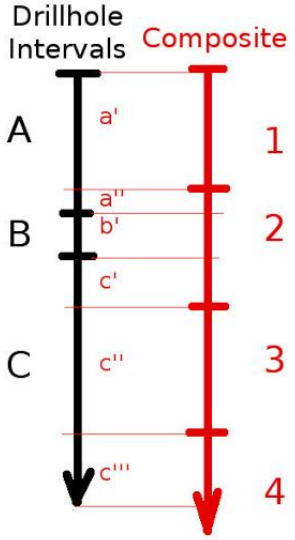
Five-level approach:

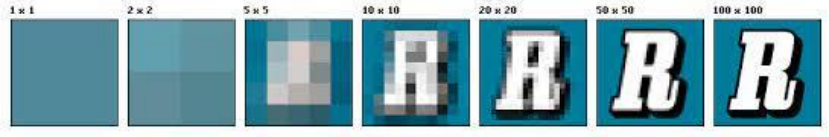
- **i) Survey Design**: the original intent to acquire the data was not the same as the current interest.
- **ii) File system:** issues on how data is organized (macro-level), stored and to be accessed across platforms/interfaces. *File System , Software and Hardware*
- **iii) Within-file**: structural/framework issues within individual files
- **iv) Content**: issues affecting actual data content, its potential interpretation and understanding
- **v) Development and interpretation environment**: where and how the (Big)data is being processed/analyzed to turn into information/value; how results are validated against benchmarks

Synergy between the (big) data set, the origin of the data, how it is stored, what it is represented and where actions are being taken.

Pathologies: Survey design level – *on data acquisition, philosophy and target*

- Data acquired to solve *one specific problem/commodity/geo-body* may not be necessarily helpful for other challenges.
- Data sheets that are not compatible in terms of *detection limits* and *spatial resolution.*
- Sampling only the *immediate ore environment* (and forget that it forms due to complex relationships that may lie outside of it)
- Sometimes… you *can't really get the data you want* at all (¤ $ £ € R$ ¥, land access, tenements) so you get what you can
- Technical standards and detection limits





Pathologies: Development and interpretation environment - The software/hardware computing ecosystem used to crunch/process/transform to explore and get information from the (big) data.

Development:

Combining what the *humans need* from data to how it *can be represented* by the computer environment.

Velocity? Architecture? Financial costs? Energy costs?

Interpretation:

What are the benchmarks used to constrain the interpretation of the Big Data solution?

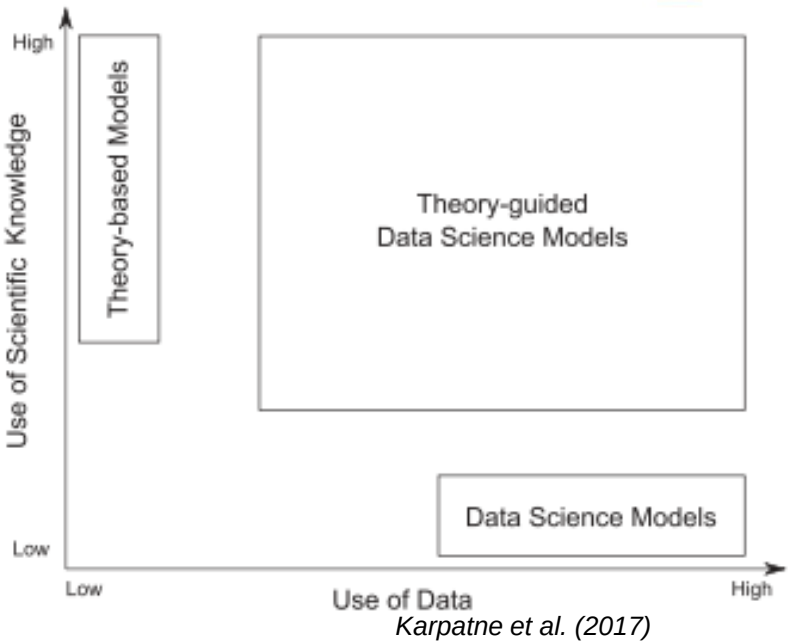
Classic Geochem vectors vs agnostic data exploration?

Compare BD/ML/AI outputs to mathematical models of geological processes is key

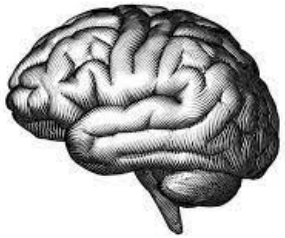
Quite verbose…

Local application setup; CPU, GPU; Distributed computing, Parallel computing; Cloud systems; Python, Julia, R; Spark, Hadoop; AWS Amazon, Collab; [data] silos, warehouses

Information theory; Theoretical/Mathematical modelling (e.g., petrology); Geological proxies (e.g., Alteration Indexes, Cu/(Cu+Ni), Pt/Pd)



Karpatne et al. (2017)



Wrap-up

- Acknowledge, understand (and respect) **limitations** of the perceived existing *pathologies [survey design, file system, within-file, content, environment]* on the (big?) data set and ecosystem

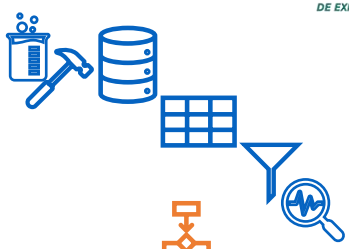
- Mitigate problems first **to achieve** better interpretations (intelligence delivery) and performance.
- **Troubleshooting** can be a … pain
- Make sure you have **enough** to **run** experiments/analysis (solve the problem).
- Make sure your ML approach is **aligned with geological concepts**.
- ML-models should provide relationships that are **feasible on the realm of geology** (how samples behave compared to geological benchmarks)

- In case **documentation** was not provided… **create and update it**.
- Document the **decisions** you make to restrict/flesh out the (big?) data

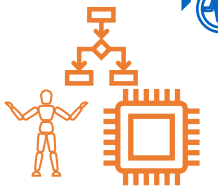
- Upcycling of legacy/historical data is **here and now**.
- Secure diligent preservation of **data being acquired at present.**

- **Learn from other business areas**
- **Project Management Institute (**PMI): **Data Management Practices**

- **Rest assured… DS/ML/AI wont make geologists obsolete… but they need to change a bit**



The data



The algorithm & model

The interpreter and domain knowledge

Garbage in

Garbage out

Thank you!

Obrigado!

Questions?

Comments?

Impressions from industry members.

Impressions from **users** of data-intensive techniques.

j.g.motta@exeter.ac.uk
jgmotta@gmail.com

João Motta

@jggsci

#SIMEXMIN2022