Open Standards for Earth Science Data Platforms

Theo McCaie - UK Met Office, Infrastructure and Data Engineering Research Lead

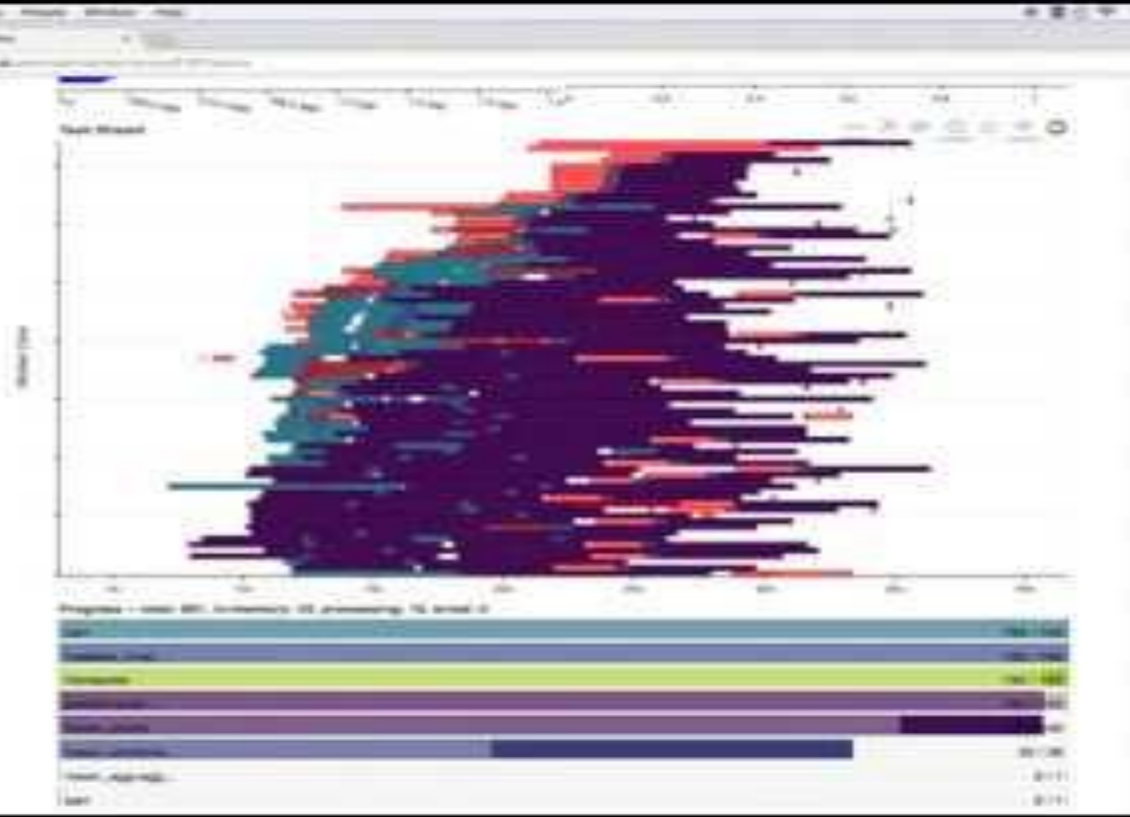
TLDR;

An open standards approach to data platforms will accelerate earth science.

PANGEO

A community platform for Big Data geoscience

Met Office | INFORMATICS LAB



PANGEO

Strengths:

- Multi-cloud
- Open Source
- Vibrant Community
- Elastic scalability
- Emerging consensus

Still developing:

- Billing opacity and control
- Maintenance overhead
- Ease of deployment
- No SLAs etc

Meanwhile...

Cloud vendor data science platforms

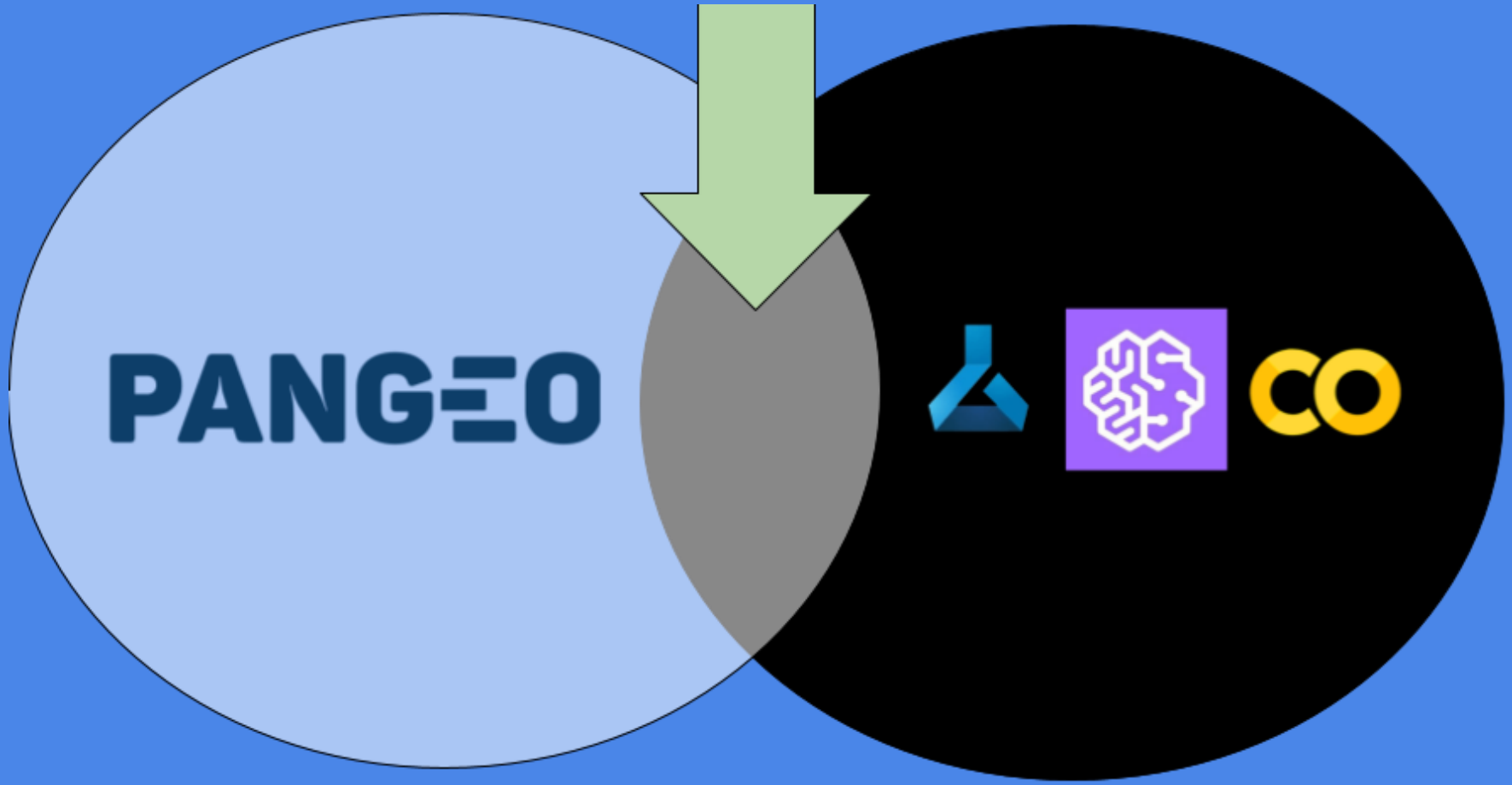
Strengths:

- Cost management
- SLAs
- Easy deployment
- TCO

At a cost of:

- Vendor lock in
- Difficulty working with datasets across multiple vendors
- Closed source

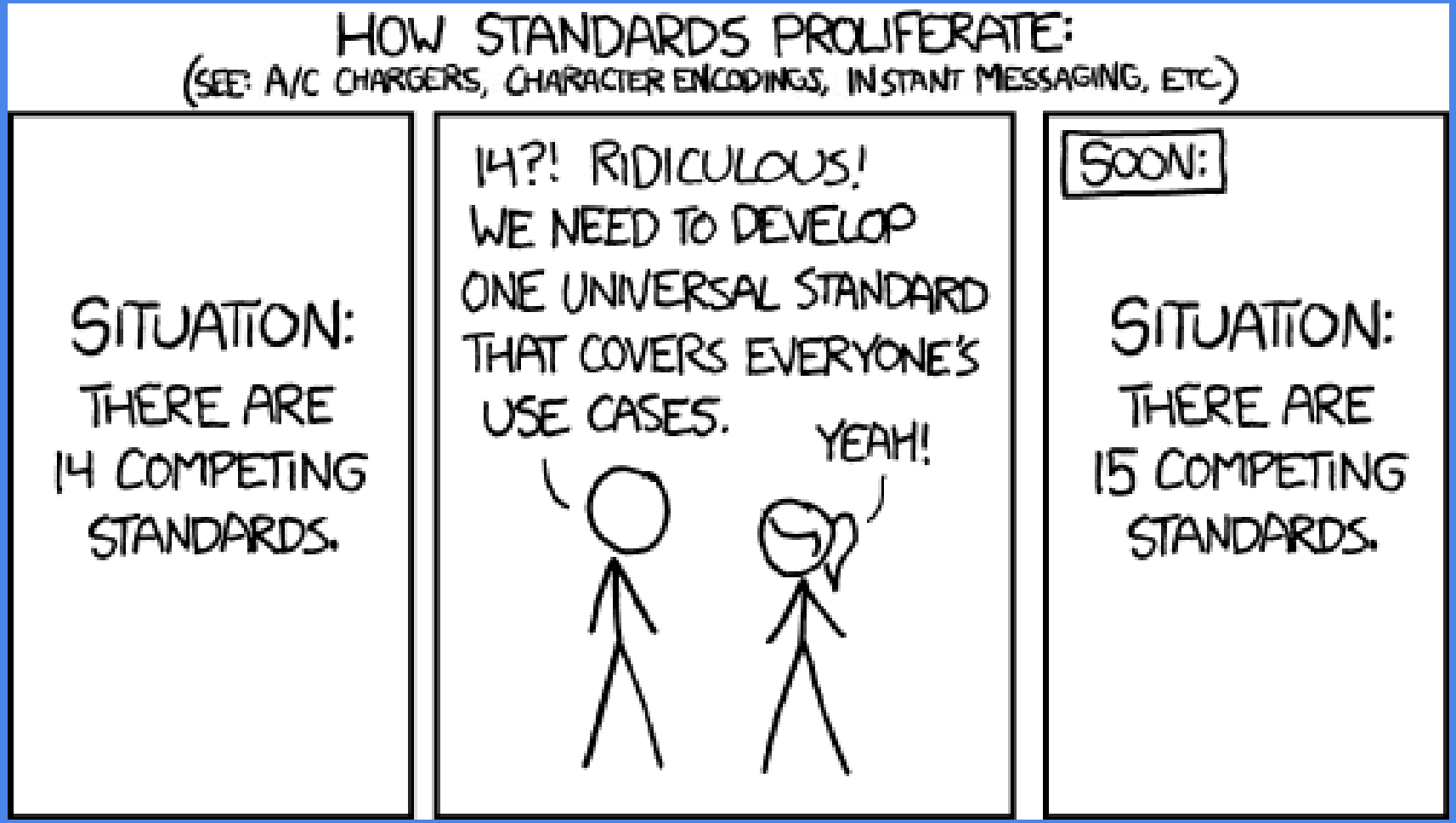
Brace for it...


PANGEO
CO

Open standards

- An agreed way of doing things that can be implemented by anyone
- Mutually beneficial for customers and suppliers
- Enable interoperability
- Examples:
  - Petrol pumps
  - Domestic electrics
  - The internet
  - Data science platforms???

Credit: https://www.bbc.co.uk/news/uk-27390466



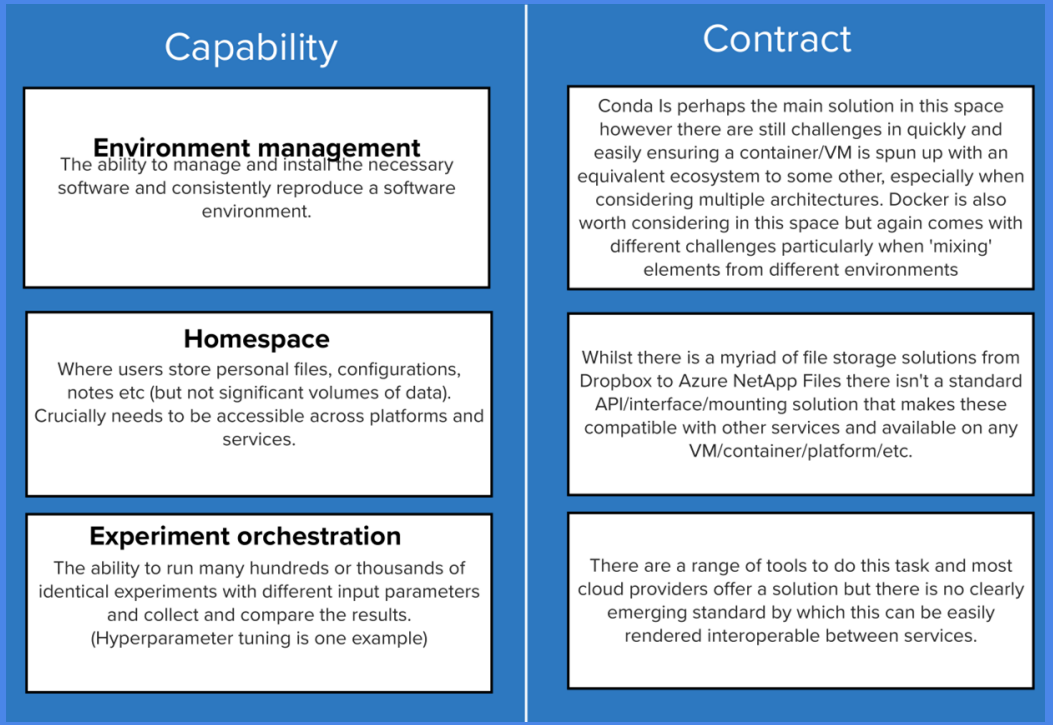
Credit: http://xkcd.com/927

Dominant design

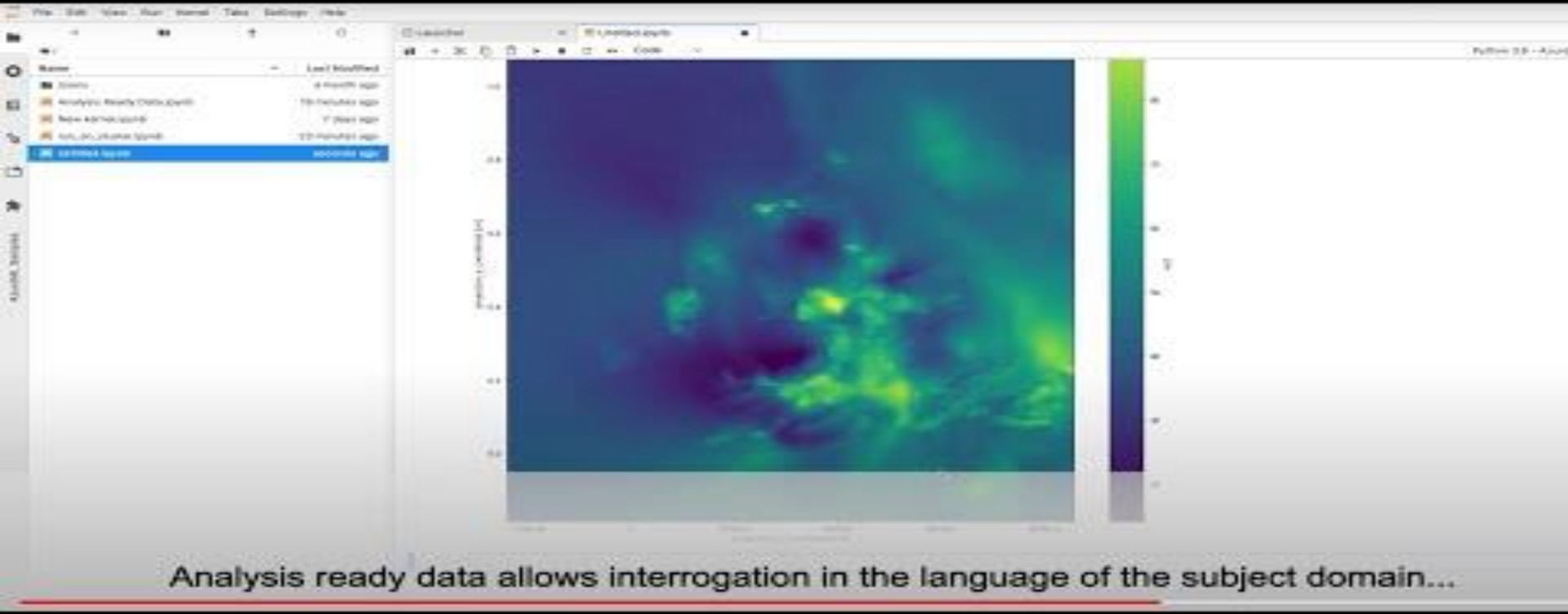
- Identify core capabilities required in the platform
- Find where there is consensus on how to achieve this
- Led by Pangeo community

Still fermenting

- For some capabilities it’s unclear what the dominant design will be
- In other cases we don’t know that the capability exists yet
- Keep watching this space and bring into the open standard fold as appropriate



From theory to practice...



Analysis ready data allows interrogation in the language of the subject domain...

Conclusions - open standards could:

- Democratise data access (you don’t need a platforms team to engage meaningfully)
- Speed up scientific delivery (focus on science not on tools, improve collaboration, consistent UX)
- Provide a common platform for research and operations (faster science to service?)
- Encourage healthy competition, dissuade aggressive lock in tactics
- But:
    - Requires buy-in from the vendors and community

Thanks, links, further reading...

- Grey new world - an article exploring the 'grey space' between open and closed source and the advantages of embracing this may offer to the geo-science community.
- Analysis ready data - a post explaining "Analysis Ready Data", it's importance and how despite rigorous data standards many institutions fail to produce it.
- What do we want from a dream data platform as a Service? - a series of user stories, personas and explanation describing what a "data science platform-as a Service" should look like.
- What should a geospatial data service look like? - a more detailed look at the challenges and solutions for powerful geospatial data services.
- Homespace, the missing as a service? - an argument that 'homespace' is a service that is missing from the current cloud ecosystem.
- aml-jupyterhub on GitHub - a JupyterHub spawner that targets Azure Machine Learning service as the Compute Instance Provider. As demoed in this poster.

- Using TileDB and Pangeo to Provide Access to Thousands of NetCDF Files as Analysis-Ready Data - a poster detailing on method for creating analysis ready data.

These slides:

shorturl.at/dkHK0

Joint Centre for Excellence in Environmental Intelligence

European Union
European Regional Development Fund

The Environmental Futures & Big Data impact lab

Azure

Met Office