The Pangeo Platform: Interactive Data Analytics for CESM

Kevin Paul

NCAR / CISL

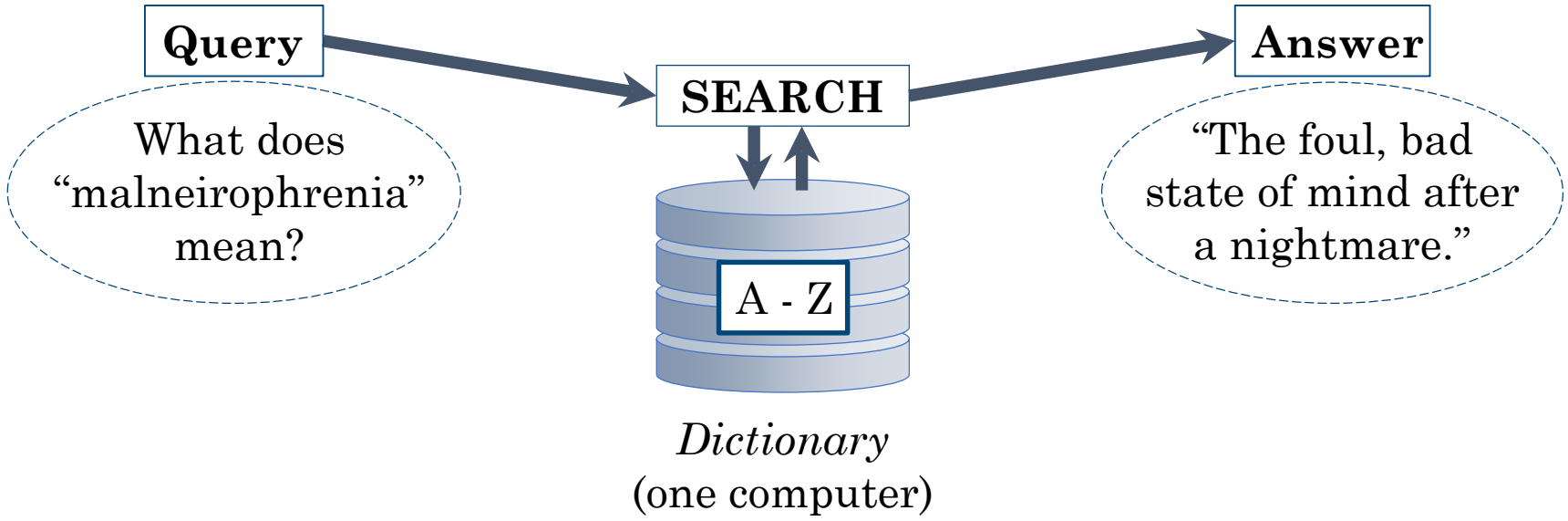
Data Analytics

- Analysis
  - Been doing this forever!
  - Increases *information density*
  - e.g., Reduce 20 PB of CMIP data to 100 1M papers
- Analytics?
  - Some say it's a synonym
  - Some say it implies the technology and methodology used in analysis
  - Comes from the business community ("Business analytics")

MapReduce

- Started by Google in their seminal 2004 paper

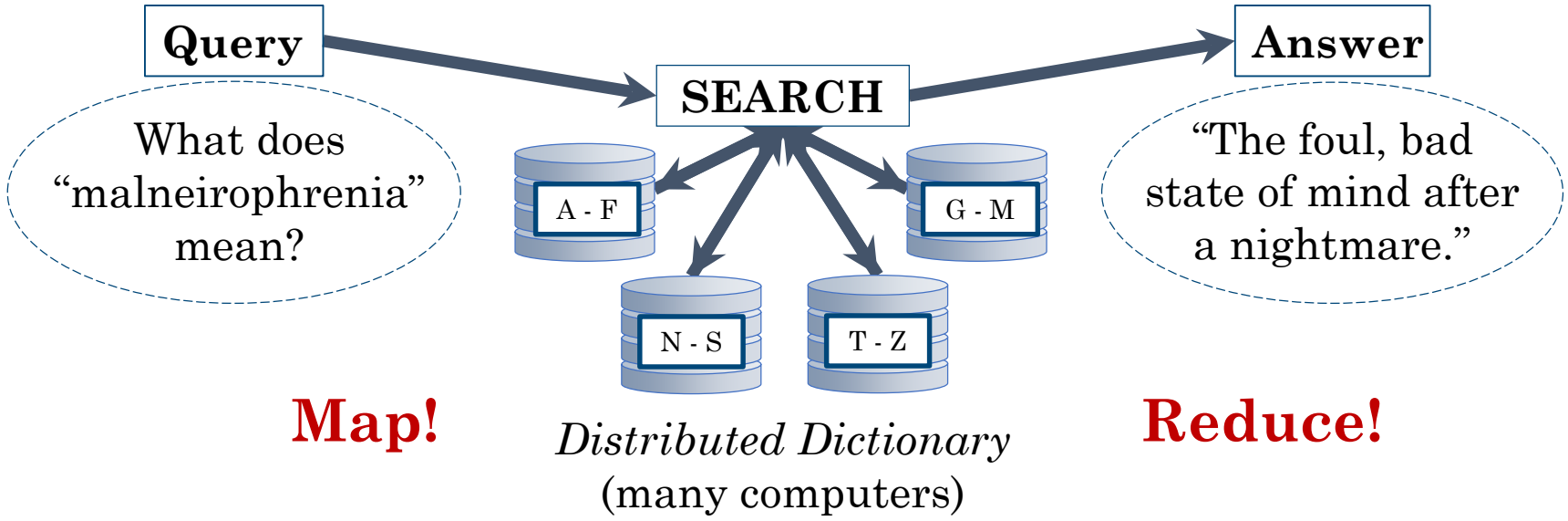
  Jeffrey Dean and Sanjay Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters." OSDI'04: Sixth Symposium on Operating System Design and Implementation (2004) 137-150.



MapReduce

- Started by Google in their seminal 2004 paper

  Jeffrey Dean and Sanjay Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters." OSDI'04: Sixth Symposium on Operating System Design and Implementation (2004) 137-150.



Beyond MapReduce…

- **2006:** Apache releases their own version of MapReduce (Hadoop)
- **2012:** Apache releases Spark to address limitations of MapReduce
  - More than just Map + Reduce
  - Provides a way of dealing with *distributed data objects*, without needing to know that the data is distributed at all!

```
d = DistributedDictionary()
query = d.find("malneirophrenia")
answer = query.compute()
```

Doesn't do anything until here!

  - Very useful for distributed databases (tables, spreadsheets, etc.)!
    - …Not so useful for Arrays!

Beyond MapReduce…

- **2014:** Matt Rocklin (Anaconda Inc.) makes first commit to **Dask**
  - "Spark in Python"
  - Expands on Spark's native database-like structures
    - Adds **Distributed Multidimensional Arrays**
- **2014:** Stephen Hoyer (Climate Corporation) starts **Xarray**
  - Provides an easy-to-use in-memory implementation of NetCDF-like data arrays
  - Builds on Dask Arrays
  - Add a lot of nice functionality to NetCDF-like data
    - Easy grouping data by month, season, etc.
    - Array access via coordinate values and array indices
    - Easily read/write from/to NetCDF

Beyond MapReduce…

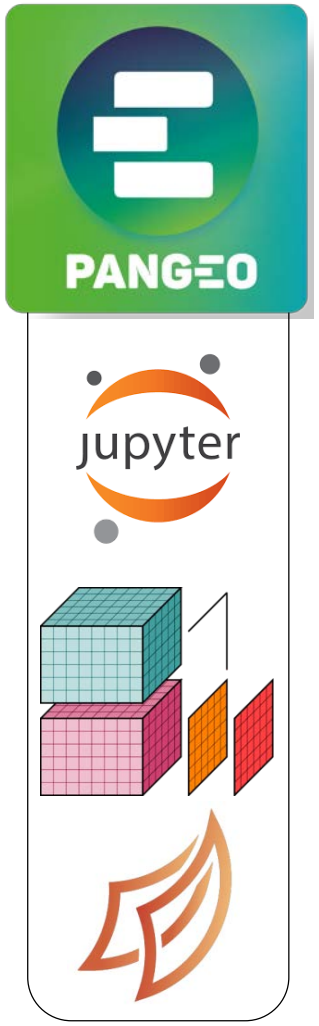
- **2016:** Ryan Abernathey (LDEO, Columbia U) organizes PyAOS workshop
  - Brought about 12 people interested in Dask+Xarray together
  - “How do we make Dask+Xarray+Jupyter work better?”
  - Branded our community with the name **Pangeo**
- **2017:** Ryan Abernathey leads writing of the NSF EarthCube grant
  - Awarded in the summer of 2017
  - Start Date: September 1, 2017

“Collaborative Proposal: EarthCube Integration: Pangeo: An Open Source Big Data Climate Science Platform”

- NSF-1740648: (LDEO, Columbia University)
  - PIs: Ryan Abernathey, Naomi Henderson, Richard Seager, Michael Tippett, Chiara Lepore
  - Sub-award to Matthew Rocklin (Anaconda, Inc.)
- NSF-1740633: (NCAR)
  - PIs: Kevin Paul, Joseph Hamman, Davide Del Vento, Ryan May (Unidata)
- Mission:
  - To improve the core functionality and integration of Xarray + Dask + Jupyter
  - To improve ease of deployment at HPC centers (e.g., NCAR) and in commercial cloud

So, what is “Pangeo”?

- It’s a **Community**:
  - Documentation, Tutorials, etc.: http://pangeo-data.org
  - Forum, Discussions, etc.: http://github.com/pangeo-data/pangeo/issues
  - Gitter Developer Chats: https://gitter.im/pangeo-data/Lobby
- It’s an NSF EarthCube **Project**:
  - NSF-1740648, NSF-1740633
- It’s a software **Stack** for data analysis:
  - Dask, Xarray, Jupyter Notebook/Lab/Hub
- It’s a **Platform** for data analysis at many locations
  - With deployments on GCP, AWS, Cheyenne and many others



PANGEO

"A community platform for Big Data geoscience"

---

jupyter — Interactive portable environments

xarray — Easy-to-use multidimensional arrays for targeted to the geosciences

DASK — Parallel "Big Data Analytics"

DEMO

http://pangeo.pydata.org