UNIVERSITY of WASHINGTON

Data Science in Science

Bill Howe, PhD
Director of Research,
Scalable Data Analytics
University of Washington
eScience Institute

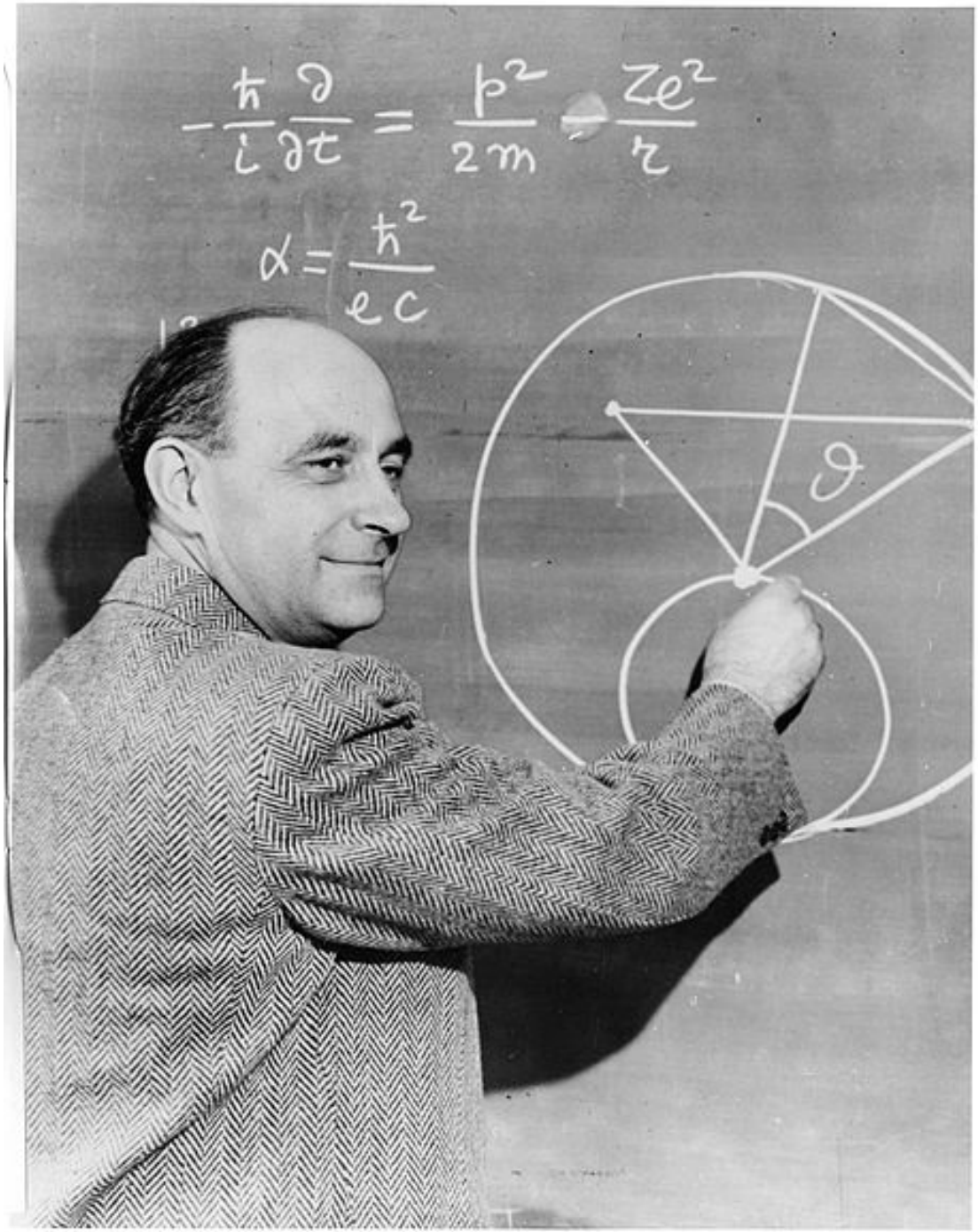
“eScience” = “Data Science”

public domain

Empirical

Theoretical

Computational



Empirical

Theoretical

Computational

Empirical

Theoretical

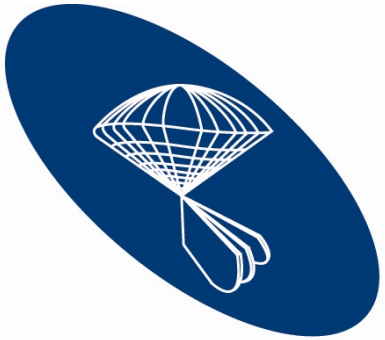
Computational

Empirical

Theoretical

Computational

eScience



SLOAN DIGITAL SKY SURVEY

Science is about asking questions

Traditionally: "Query the world"
Data acquisition activities coupled to a specific hypothesis

eScience: "Download the world"
Data acquired en masse in support of many hypotheses

The cost of data acquisition has dropped precipitously thanks to advances in technology

- **Astronomy: High-resolution, high-frequency sky surveys (SDSS, LSST, PanSTARRS)**
- **Life Sciences: lab automation, high-throughput sequencing,**
- **Oceanography: high-resolution models, cheap sensors, satellites**

The cost of finding, integrating, and analyzing data, then communicating results, is the new bottleneck

eScience is driven by *data* more than by computation

- Massive volumes of data from sensors and networks of sensors

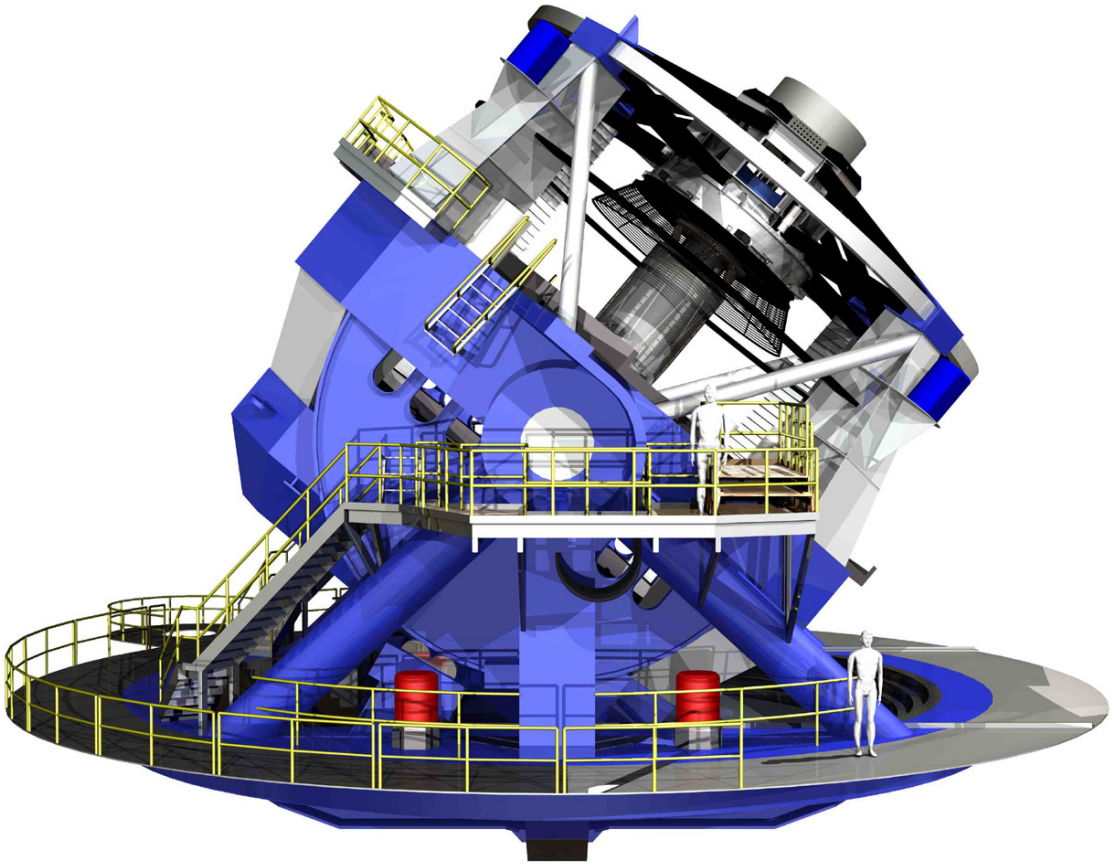
Apache Point telescope, SDSS

80TB of raw image data (80,000,000,000,000 bytes) over a 7 year period

Large Synoptic Survey Telescope (LSST)

40TB/day (an SDSS every two days), 100+PB in its 10-year lifetime

400mbps sustained data rate between Chile and NCSA



Illumina HiSeq 2000 Sequencer

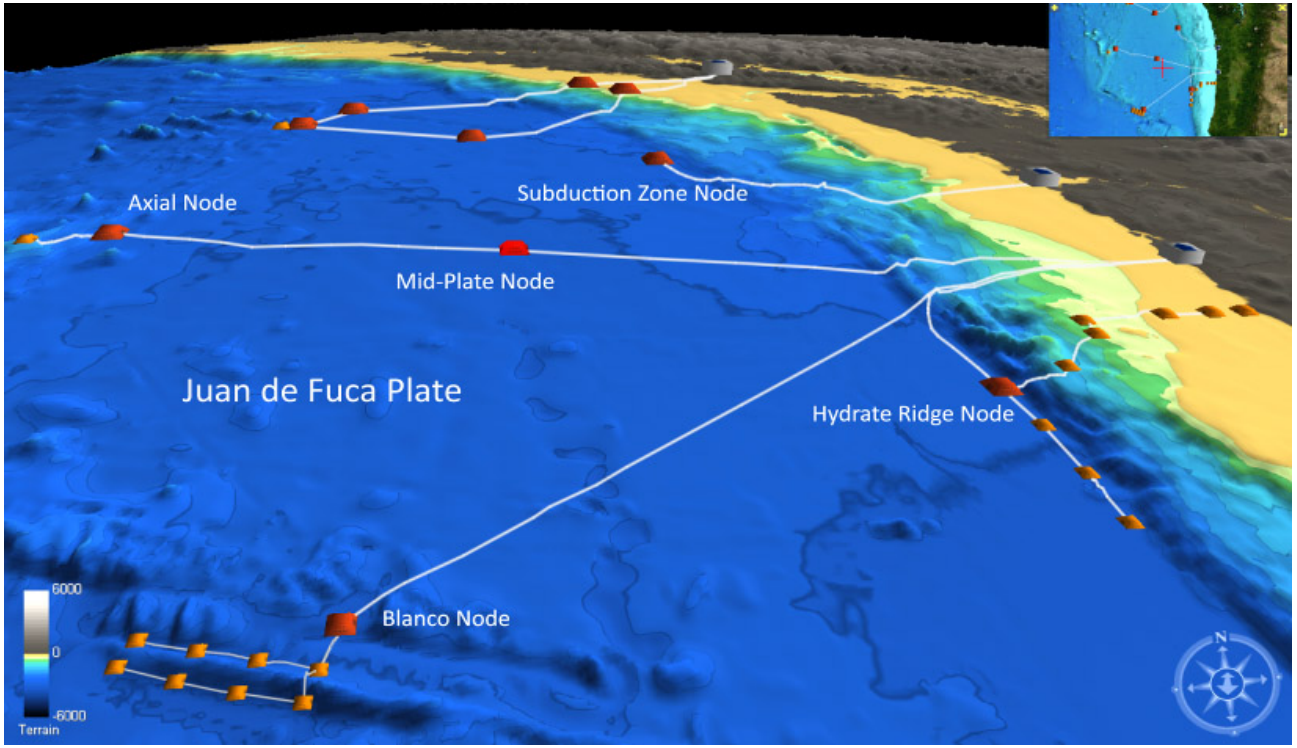
~1TB/day

Major labs have 25-100 of these machines

UNIVERSITY of WASHINGTON

Regional Scale Nodes of the NSF Ocean Observatories Initiative

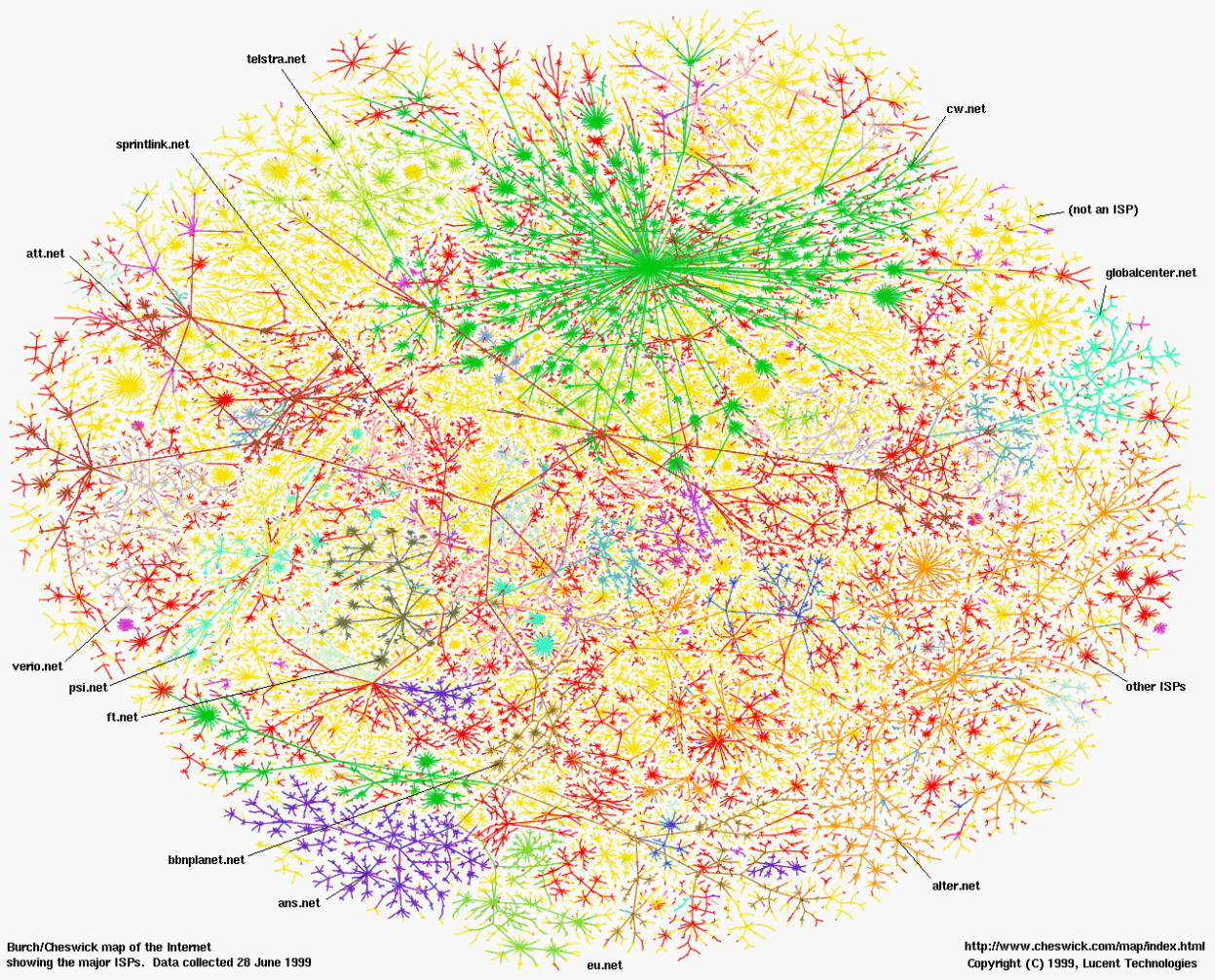
1000 km of fiber optic cable on the seafloor, connecting thousands of chemical, physical, and biological sensors



The Web

20+ billion web pages x 20KB = 400+TB

One computer can read 30-35 MB/sec from one disk => 4 months just to read the web



eScience is about the *analysis* of data

- The automated or semi-automated extraction of knowledge from massive volumes of data
  - There's simply too much of it to look at
  - But it's not just a matter of volume
- The Three V's of Big Data:
  - Volume: number of rows / objects / bytes
  - Variety: number of columns / dimensions / sources
  - Velocity: number of rows / bytes per unit time
- More V's:
  - *Veracity: Can we trust this data?*

Summary

- Science is in the midst of a generational shift from a data-poor enterprise to a data-rich enterprise
- Data analysis has replaced data acquisition as the new bottleneck to discovery

- What does this have to do with business?

Business is beginning to look a lot like science

- Acquire data aggressively and keep it around
- Hire data scientists
- Make empirical decisions