Careers in eScience

Bill Howe

T - shaped

thanks to Alex Szalay

Π- shaped

thanks to Alex Szalay

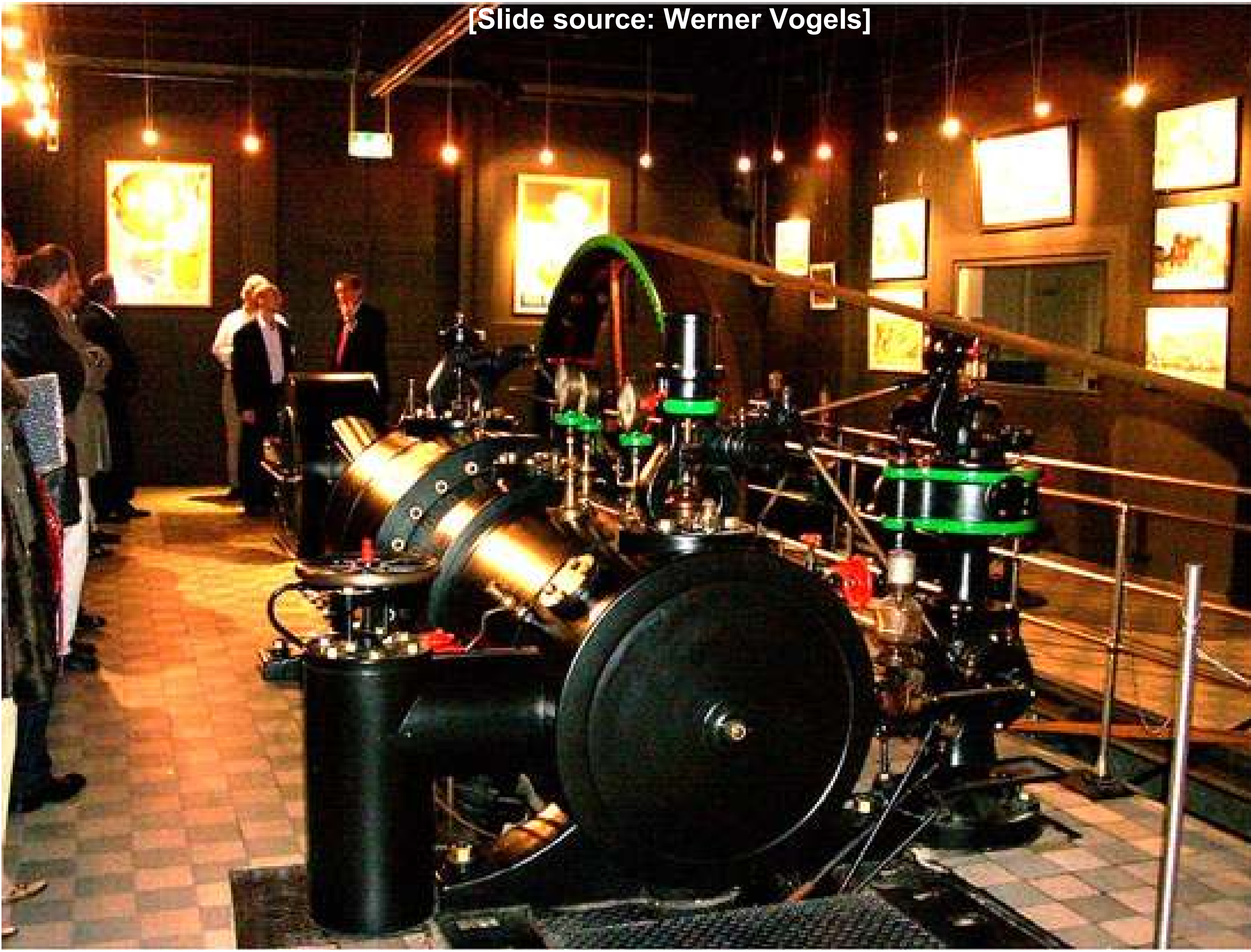
What to work on

- Maier’s (other) maxim:
    - “Better to design for 1 application than 0 applications”
- Paul Graham / Y-combinator:
    - “Make something people want”
- Paraphrasing Natassa Ailamaki:
    - “It takes 18 months of working with scientists to make progress”
- On large-scale:
    - If the GB/node ratio is less than 4, you’re not large-scale, and you should stop using Hadoop
    - Exception: you have happy users
- If you’re using synthetic data or toy problems, no one cares
    - Researchers at your university actually need your help – find them

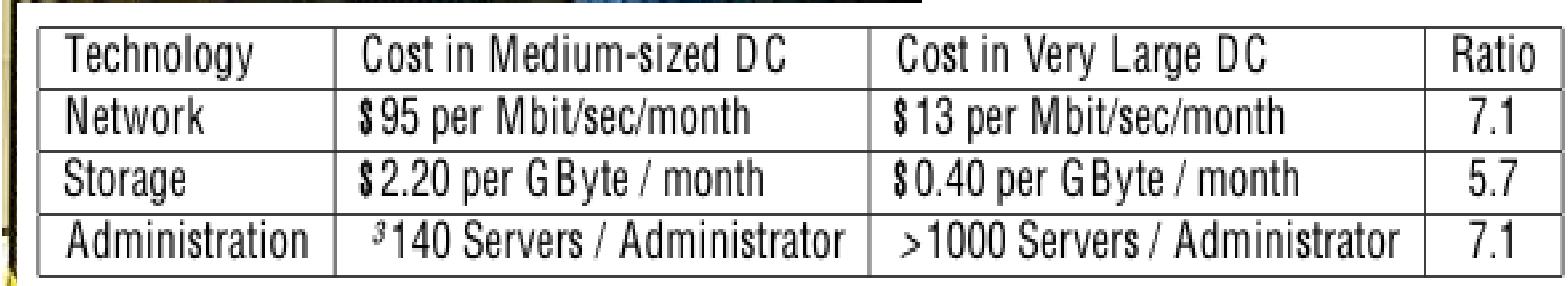
Data Management Plan Goal

All project data
online and queryable
now, by anyone

[Slide source: Werner Vogels]

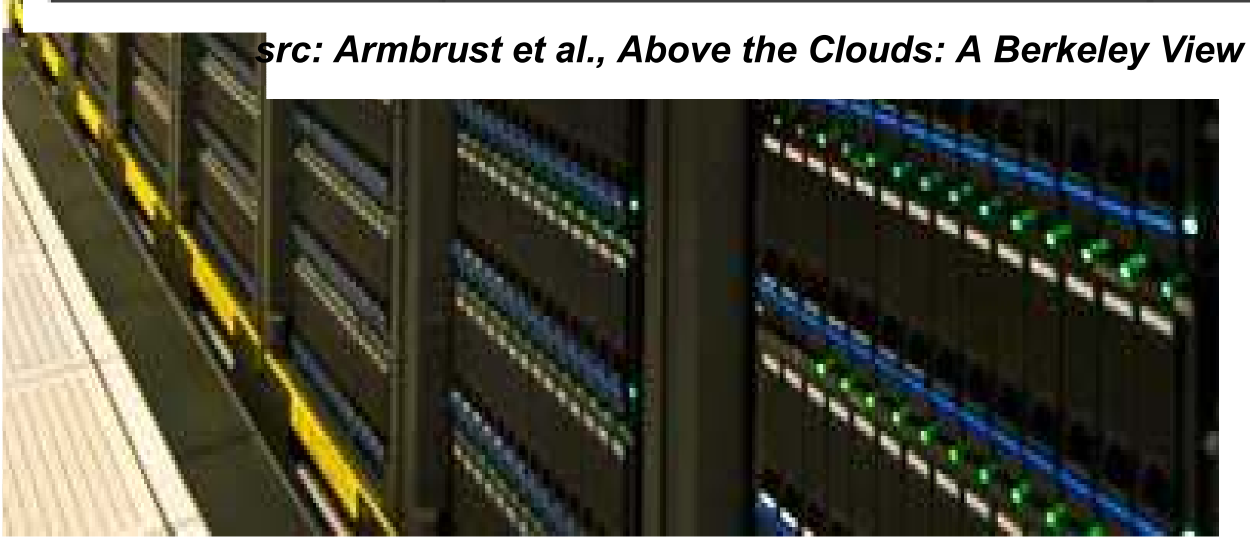


Economies of Scale



Technology	Cost in Medium-sized DC	Cost in Very Large DC	Ratio
Network	\$95 per Mbit/sec/month	\$13 per Mbit/sec/month	7.1
Storage	\$2.20 per GByte / month	\$0.40 per GByte / month	5.7
Administration	^{3}140 Servers / Administrator	>1000 Servers / Administrator	7.1

src: Armbrust et al., Above the Clouds: A Berkeley View of Cloud Computing, 2009



Cloud Growth

“Every day, Amazon buys enough computing resources to run the entire Amazon.com infrastructure as of 2001”
-- James Hamilton, Amazon, Inc., SIGMOD 2011 keynote

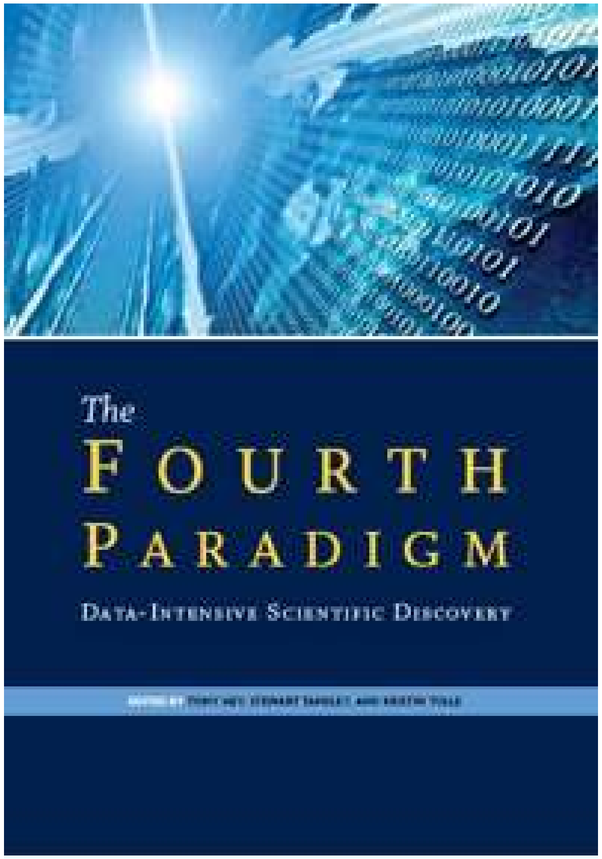
Fix the funny money

- Computing equipment incurs no indirect costs
  - “Capital Expenditures”
  - Power, cooling, administration?
- “Services” are charged full indirect cost load
  - Ex: 54% at UW; 100% at Stanford
- So every dollar spent on Amazon costs the PI $1.54
- Every dollar spent on equipment costs the PI $1.00, but also costs the university ~$1.00

Any Questions?

How is eScience different than computational science?

- Theory
- Experiment
- Computational Science
- Data-intensive science



Curricula

- Should we offer an eScience certificate?
- Should we offer an eScience Master’s degree?
- Should there (eventually) be an eScience Phd program?

eScience Courses

- Magda Balazinska, Bill HoweCS599c: Scientific Data Management, Spring 2010, University of Washington
- Laura Bright, Bill HoweCS410/510: Scientific Data Management, Summer 2006, Portland State University
- Syracuse University, eScience, Data Science