The Hearts and Minds of Data Science

Cecilia R. Aragon
Dept. of Human Centered Design & Engineering
eScience Institute
University of Washington
aragon@uw.edu
What is data science?

Big data is like teenage sex:
everyone talks about it,
nobody really knows how to do it,
everyone thinks everyone else is doing it, so everyone claims they are doing it...

(Dan Ariely)
From data to knowledge to action

- The ability to extract knowledge from large, heterogeneous, noisy datasets – to move “from data to knowledge to action” – lies at the heart of 21st century discovery
- To remain at the forefront, researchers in all fields will need access to state-of-the-art data science methodologies and tools
- Data science is driven more by intellectual infrastructure (human skills) and software infrastructure (shared tools and services – digital capital) than by hardware
- These methodologies and tools will need to be designed to enable efficient discovery by the scientists who use them

Adapted from: Ed Lazowska et al., eScience Institute, UW
Moore/Sloan Data Science Environment: $37.8M Initiative at UCB, NYU, UW

Graphic by Ray Hong and eScience Institute, UW
DSE Goal

• Change the culture of universities to create a data science culture

• “we're a startup in the Eric Ries sense, a ‘human institution designed to deliver a new product or service under conditions of extreme uncertainty.’ Our outcomes are new forms of doing science in academic settings, and we're going against the grain of large institutional priors.” – Fernando Pérez

Six major thrusts

- Education and training
- Software tools and environments
- Career paths and alternative metrics
- Reproducibility and open science
- Working spaces and culture
- Ethnography and evaluation
What is ethnography and why use it?

• Qualitative field-based method of inquiry
• Can focus on the dynamics and uses of data and computation in context
• Ecological validity
• The determinants of culture are often invisible or unspoken
• Understanding *why* and not just *what*
Human centered data science
Two human issues

1. How do we build tools that facilitate data-driven science today?
2. How do we ensure that we have enough people with data science skills working in academia to produce the next generation of scientific discoveries?
Two human issues

1. How do we build tools that facilitate data-driven science today?

2. How do we ensure that we have enough people with data science skills working in academia to produce the next generation of scientific discoveries?
Why is this a human issue?
What is the rate-limiting step in data understanding?

(1 EB = 1,000,000,000,000,000,000 B = 10^{18} \text{ bytes} = \text{one quintillion bytes} = 100,000 \text{ Libraries of Congress})

What is the rate-limiting step in data understanding?

Processing power: Moore’s Law

Amount of data in the world
What is the rate-limiting step in data understanding?

Amount of data in the world

Processing power: Moore’s Law

Effective Processing Power: Amdahl’s Law
What is the rate-limiting step in data understanding?

Idea adapted from “Less is More” by Bill Buxton (2001)
Bridging the Data Science Gaps

- Hardware
- Software
- Parallel computing
- Machine learning

- Human centered data science
  - collaborative software, effective user interfaces
  - visualization
  - sociology, psychology, study of socio-technical systems
Collaboration Growth: “Hyperauthorship”

Maximum number of authors on a single paper, by year, 1992 to 2011

- Physics (3,221)
- Medicine (2,459)

Number of multi-author papers, 1998-2011

Collaboration is essential

• To understand vast and complex data sets
• To combine different skills
• To build interdisciplinary teams to solve the most challenging scientific, social, political, and human problems
But it’s not so easy to build collaboration

• Many barriers are social, not just technical
  – Charlotte Lee, “Barriers to Adoption of Collaboration Technologies,” presented at CHI 09 workshop
    • too little is known about dynamics of complex work teams
    • collaboration across disciplines is difficult (different languages, methods)
    • distributed work is difficult (different organizational structures and processes)
    • need to study how to foster productive collaborations
    • the human infrastructure of cyberinfrastructure
Data science involves solving both technical and social issues
Astrophysics example

1. New machine learning approach to transient search
2. Image processing for supernova detection
3. Collaboration building
4. Cross-cultural issues
Impact on Supernova Search

Imbalanced Data & Class Uncertainty

- Ratio of positives to negatives less than 1/10,000
- Many negatives in region of overlap
- Potentially mislabeled examples

Fast Algorithm

• Find good supernova candidates
  – R1 measures $||F_{-1}||$ vs. $||F_1||$ (elliptical vs. circular)
  – R2 measures power in higher terms vs. $||F_1||$ and $||F_{-1}||$

• Advantages of these R1 and R2:
  – Scale invariant (due to normalization by $F_1$)
  – Rotation invariant (only magnitudes used, no phase)
  – Fast! Only two transform terms, and they are complex conjugates ($F_1, F_{-1}$), so intermediate results are shared

• Performance depends on how fast we can obtain (for each point $z_i = x_i + iy_i$ on the contour) the primitives used to form $F_1$ and $F_{-1}$:

  $x_i \cos\left(\frac{2\pi i}{N}\right), \quad y_i \sin\left(\frac{2\pi i}{N}\right),$

  $x_i \sin\left(\frac{2\pi i}{N}\right), \quad y_i \cos\left(\frac{2\pi i}{N}\right)$

• In the end:
  – 41% reduction in false positives, no measurable increase in running time

Lightweight, context-linked tools for collaborative work

Measurable improvements from human centered approach

- Measurable improvements to transient search and data analysis in many areas.
  - Improved image processing algorithms:
    - Fourier contour analysis (41% reduction in false positives)
  - New machine learning algorithms for astronomical transient search:
    - support vector machines
    - boosted decision trees (90% reduction in false positives)
  - Improved control interfaces for scanning and vetting
    - approx. 70% speedup per image

- Search + Scanning: approx. 90% labor savings
  - from 6-8 people working ~4 hrs/day to 1 person working 1 hr/day
The sociotechnical ecosystem of science
Tools that fit within the sociotechnical ecosystem of science

- **IPython - Fernando Pérez**
  - base for interactive scientific environment – enables exploratory computing – browser-based notebooks (rich text, code, plots)

Tools that fit within the sociotechnical ecosystem of science

- mpld3, a toolkit for visualizing matplotlib graphics in-browser via d3 – Jake Vanderplas

http://jakevdp.github.io/blog/2014/01/10/d3-plugins-truly-interactive/
Research with Social Media Text Data

- Ranges along quantitative – qualitative spectrum
- Quantitative
  - Good: Large quantities of data, efficient
  - Bad: Relatively shallow, superficial
- Qualitative
  - Good: deep, explanatory conclusions
  - Bad: Small, focused samples, inefficient
- How to get best of both worlds?
  - Visual analytics, combining machine learning with visualization and qualitative research
Integrate machine learning and interactive visualization into a qualitative research workflow

- Conclusions based on large data sets
- Maintain context, nuance, depth
- Select tools based on transparency and understandability, not just efficiency
Human issues in data science

• **Ethics:** Lilly Irani and Six Silberman, UCSD, Turkopticon and crowdsourcing

• **Design:** What happens when you crowd-source design? Daniela Rosner, UW (systems that guarantee banal results?)

• **Data expectations:** What does data really mean across different fields? Gina Neff, B. Fiore-Silfvast
Human centered data science
Two human issues

1. How do we build tools that facilitate data-driven science today?

2. How do we ensure that we have enough people with data science skills working in academia to produce the next generation of scientific discoveries?
Interdisciplinary Data Science Skills

• Programming Foundations and Data Abstractions
• Software Engineering and Systems
• Databases and Data Management
• User-Centered Software Design; Design Foundations for Interactive Systems
• High-Performance Scientific Computing
• Scalable and Data-Intensive Computing in the Cloud
• Data Mining / Machine Learning / Statistics
• Data and Information Visualization

Human issues of data science in academia – Fernando Perez

- Academia favors individualism, hyper-specialization and novelty
- Collaborations and interdisciplinary work are less rewarded
- Writing software to enable new science is not rewarded (it’s time not spent writing papers)
- Scientists have been trained to think of computation as not “real science”
- Methodologists (e.g. computer scientists, statisticians) are not rewarded for creating applications
- The skills of data science are in tremendous demand in industry today = “big data brain drain” (see Jake Vanderplas blog)

The Big Data Brain Drain

• “the skills required to be a successful scientific researcher are increasingly indistinguishable from the skills required to be successful in industry.”

• “open, well-documented, and well-tested scientific code is essential not only to reproducibility in modern scientific research, but to the very progression of research itself.”

–Jake Vanderplas

http://jakevdp.github.io/blog/2013/10/26/big-data-brain-drain/
Human issues of data science in academia

• So what are we going to do about this?
• We know that data science skills are necessary for scientific discovery
• Who is going to teach these skills?
• Academic scientific disciplines have full schedules, tend not to teach computation and data skills or only via a course or two.
• And when people acquire data science skills, there is no job path in academia
Data science as an academic department

• Analogy: how the discipline of computer science was created
• No undergraduate CS degree at Caltech in the 1980s
The emerging discipline of computer science

• “Only failed mathematicians go into computer science.”
  – Unnamed mathematician, Caltech, 1980s
• “I wrote a program once. It wasn’t that hard.”
  – Unnamed physicist, Berkeley, 1990s
• “But... this programming... what if it’s all just a fad?”
  – Unnamed physics professor, U Penn, 2000s
Interdisciplinary departments

• Another exemplar: my own interdisciplinary department – Human Centered Design and Engineering
  – Tenure case does not involve only publishing in the “right” journals
  – Part of the case is that you present evidence of your impact (conferences, journals, altmetrics)
  – New department = unranked
  – But: fastest-growing department in the College of Engineering at UW
The hearts and minds of data science
Questions?

Cecilia R. Aragon
Human Centered Design & Engineering
University of Washington
aragon@uw.edu