Big Data – so what’s the big deal?

Jevin West
Information School, University of Washington
DataLab (MGH 310E)
jevinw@uw.edu
January 26, 2017
What is Data Science?
The world is awash in bullshit. Politicians are unconstrained by facts. Science is conducted by press release. So-called higher education often rewards bullshit over analytic thought. Startup culture has elevated bullshit to high art. Advertisers wink conspiratorially and invite us to join them in seeing through all the bullshit, then take advantage of our lowered guard to bombard us with second-order bullshit. The majority of administrative activity, whether in private business or the public sphere, often seems to be little more than a sophisticated exercise in the combinatorial reassembly of bullshit.
A Visual Introduction to Machine Learning

Translations: 中文 русский Français

In machine learning, computers apply statistical learning techniques to automatically identify patterns in data. These techniques can be used to make highly accurate predictions.

Keep scrolling. Using a data set about homes, we will create a machine learning model to distinguish homes in New York from homes in San Francisco.

http://www.r2d3.us/visual-intro-to-machine-learning-part-1/
Want to be a data scientist?
‘The Data Scientist’

Communication skills

Ethical Reasoning

Information/Data Management

Personnel Management

Interdisciplinary

Adaptable
Data Scientist

Hacking Skills

Math & Statistics Knowledge

Machine Learning

Danger Zone!

Data Science

Traditional Research

Substantive Expertise

Drew Conway, NYU
Examples of data science

“[Pole] ran test after test, analyzing the data, and before long some useful patterns emerged. Lotions, for example. Lots of people buy lotion, but one of Pole’s colleagues noticed that women on the baby registry were buying larger quantities of unscented lotion around the beginning of their second trimester. Another analyst noted that sometime in the first 20 weeks, pregnant women loaded up on supplements like calcium, magnesium and zinc. Many shoppers purchase soap and cotton balls, but when someone suddenly starts buying lots of scent-free soap and extra-big bags of cotton balls, in addition to hand sanitizers and washcloths, it signals they could be getting close to their delivery date.”

“My daughter got this in the mail!” he said. “She’s still in high school, and you’re sending her coupons for baby clothes and cribs? Are you trying to encourage her to get pregnant?”

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

Every time you go shopping, you share intimate details about your consumption patterns with retailers. And many of those retailers are studying those details to figure out what you like, what you need, and which coupons are most likely to make you happy. Target, for example, has figured out how to data-mine its way into your womb, to figure out whether you have a baby on the way long before you need to start buying diapers.

Charles Duhigg outlines in the New York Times how Target tries to hook parents-to-be at that crucial moment before they turn into high-value baby product shoppers.
Agenda

• What is data science?
• Cautionary Tales
• Data Science at UW and in Seattle
• Big data – why should you care?
• More cautionary Tales (Data and Society)
• Data Science, in action
• DataLab
• Data for Social Good
Google came under fire this week after its new Photos app categorized photos in one of the most racist ways possible. On June 28th, computer programmer Jacky Alciné found that the feature kept tagging pictures of him and his girlfriend as "gorillas." He tweeted at Google asking what kind of sample images the company had used that would allow such a terrible mistake to happen.
$74 million later, Mercer Mess is 2 seconds faster

by Matt Markovich | Monday, October 17th 2016
Universities are going big
DATA-DRIVEN DISCOVERY

Data Science Environments
Big Data at UW

- LSST
- CS (Farecast)
- Libraries (digital content)
- Oceanography
- Neuroscience
Education

Undergraduate
The eScience Institute has formed an inter-departmental working group on data science education, comprised of representatives from a large variety of departments, schools, and colleges. This working group is in the process of creating a template that describes what the University of Washington sees as the core of data science education at the undergraduate level.

LEARN MORE ABOUT UNDERGRADUATE PROGRAMS >

Masters
University of Washington’s new Master of Science in Data Science gives current and aspiring data science professionals the technical skills needed to extract insights from large, noisy and heterogeneous datasets – and the practical skills to make analytics easy to understand and use.

LEARN MORE ABOUT MASTERS PROGRAMS >
Data Science at the Information School

- Data Science Option (~ Spring 2016)
- INFO 370: Introduction to Data Science (Fall)
- INFO 371: Machine Learning (Spring)
- INFO 445: Advanced Database Design, Management, and Maintenance
- INFO 474: Interactive Data Visualization
Other Classes in iSchool

• INFX 551 (4 credits) – Fundamentals of Data Curation
• INFX 576 (4 credits) – Social Network Analysis
• INFO 470 (5 credits) – Research Methods
• INFX 573 (4 credits) – Introduction to Data Science
• INFX 574 (4 credits) – Core Methods in Data Science and Analytics
• INFX 575 (4 credits) – Advanced Methods in Data Science and Analytics
The Data Science Seminar is a university-wide effort bringing together thought-leading speakers and researchers across campus to discuss topics related to data analysis, visualization and applications to domain sciences. The seminar is typically held on Wednesdays 3:30-4:30pm. Unless otherwise noted, the location for Winter Quarter 2017 is Johnson 102.

All talks are free and open to the public.

Upcoming Speakers

JAN 11  Assessing Moderated Effects of Mobile Health Interventions on Behavior
Susan A. Murphy
*University of Michigan*

NOTE: Location is Johnson Hall 102

JAN 18  Adventures in Little Data
Paul Ginsparg
*Cornell University*

NOTE: Location is Johnson Hall 102

Previous Speakers (2016)

JAN 6  Geometric graph-based methods for high dimensional data
Andrea L. Bertozzi
*Professor, Department of Mathematics, UCLA*

JAN 27  Inferring Complex Behavioral Mechanisms in Difficult Places
Carter Butts
*Professor, Department of Sociology, UC Irvine*

FEB 24  Turning the Virtual Tables: Social Media, Opposition, and Government Responses in Russia and Venezuela
Joshua Tucker
What is big data?
Extreme Big Data: Beyond Zettabytes And Yottabytes

OracleVoice
Oct. 09 2013 - 10:00 AM

If we’re going to talk about big data—and it’s one of the most important discussions in business today—we better all speak the same language. In the hierarchy of big data, there are petabytes, exabytes, zettabytes, and yottabytes. After that, things get murky.

The challenge is only partly one of coming to agreement on the right words to describe what lies beyond a
“Yes, some of the best theorizing comes after collecting data because then you become aware of another reality…”

Data Exhaust: by-product of human activity

Examples: cell phone locations, purchase transactions, social media

Why big data?

- Cheaper sensors (climate research, astronomy, high energy physics, high-throughput gene sequencing, cell phones)
- Cheaper storage (4 TB, $168)
- People willing to share their personal information (Facebook, social media)
- Faster communication (internet, cell phones)
- Other reasons?
The Four A’s and V’s

- Architecture
- Acquisition
- Analysis
- Archiving
- Volume
- Velocity
- Variety
- Veracity
Why should you care about big data?

A shortage of 1.5 million jobs!
The following jobs matched your Microsoft job search agent that you set up at www.microsoft-careers.com.

Agent:
"Business Intelligence" OR "Data" OR "Data Analyst" AND Seattle
(Click to modify)

Job Matches:

Data Analyst - GFS 831264 Job - Seattle, WA, US
Software Development Engineer - Data Analytics Team Job - Redmond, WA, US
Data Mining Analyst, Senior - ADS - 817244 Job - Bellevue, WA, US
Program Manager II - Windows Job - Redmond, WA, US
Solution Manager, Office Finance, 821435 Job - Redmond, WA, US
Data Scientist, Sr Data Mining Job - Redmond, WA, US
Data Scientist, Sr Principal Data Mining/Hadoop Job - Redmond, WA, US
Software Development Engineer II- BING Job - Bellevue, WA, US
Senior Data Scientist, O365, 830416 Job - Redmond, WA, US
<table>
<thead>
<tr>
<th>Job Title</th>
<th>Avg. Salary</th>
<th>$110k</th>
<th>$150k</th>
<th>$190k</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Senior Data Scientist – Netflix</strong></td>
<td>n/a</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 Salary</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Data Scientist – Groupon</strong></td>
<td>n/a</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 Salary</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Data Scientist – EBH Enterprises</strong></td>
<td>$155,326</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 Salaries</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Data Scientist – Live Nation</strong></td>
<td>n/a</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 Salary</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Data Scientist – Nokia</strong></td>
<td>n/a</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 Salary</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Data Scientist – GREE International</strong></td>
<td>n/a</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 Salary</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Senior Data Scientist – LinkedIn</strong></td>
<td>$136,871</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 Salaries</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Senior Data Scientist – Tapjoy</strong></td>
<td>n/a</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 Salary</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Data Scientist – Gerson Lehrman Group</strong></td>
<td>n/a</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 Salary</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Data Scientist – Apollo Group</strong></td>
<td>$122,360</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 Salaries</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Concerns

• Privacy
• Overconfidence and Overfitting
• Correlation versus causation
• Who owns big data?
• What else?
Big Data is messy
They're Watching You at Work

What happens when Big Data meets human resources? The emerging practice of "people analytics" is already transforming how employers hire, fire, and promote.
New MIT algorithm rubs shoulders with human intuition in big data analysis

The Data Science Machine takes humans out of the equation. (REUTERS/Kacper Pempel)

Explore flu trends - United States

We've found that certain search terms are good indicators of flu activity. Google Flu Trends uses aggregated Google search data to estimate flu activity. Learn more »

United States > Washington

States | Cities (Experimental)
Correlation versus Causation

Now a ridiculous book!

- Spurious charts
- Fascinating factoids
- Commentary in the footnotes

Amazon | Barnes & Noble | Indie Bound
The inequality of who dies in car crashes

Motor vehicle crash deaths per 100 million vehicle miles traveled among people aged 25 or older. Shading indicates 95% confidence intervals.
Sampling
Big Data in action
If you had access to the personal calendars of 200 million people, what could you do with it? What products could you create?
Is there a secondary market for the data that companies are collecting?
Big data is about asking good questions
Social network analysis in Afghanistan
Joshua Blumenstock is studying the effect of unexpected shocks on network activity in Afghanistan with collaborators at Princeton, UW, UCLA, and UC Berkeley
Science of Science
RESEARCH AREAS

Scholarly Publishing  Mapping Science  Navigating Science  Sociology of Science

NEWS

23 Nov.  **JEVIN WEST ON MEGAJOURNALS IN THE CHRONICLE OF HIGHER EDUCATION**

Jevin West discusses the rise of the megajournal and our open access cost effectiveness tool in the Chronicle of Higher Education.

23 Nov.  **EIGENFACTOR TEAM PLACES SECOND IN MICROSOFT RESEARCH'S WSDM CUP**

The WSDM Cup Challenge asked teams to use 30GB of data from the Microsoft Academic Graph to rank the importance of a whole article set from a subset of the articles. Our Eigenfactor algorithm was ranked second.
Figure-Centric Search Engine

viziometrics.org

A project of the eScience Institute at the University of Washington
The Graph

The Microsoft Academic Graph is a heterogeneous graph containing scientific publication records, citation relationships between publications, as well as authors, institutions, journal and conference "venues," and fields of study.

The Data

This data is available as a set of zipped text files stored in Microsoft Azure blob storage and available via HTTP. The file size (zipped) is ~30GB and may be downloaded here.

The Challenge

The goal of the Ranker Challenge is to assess the query-independent importance of scholarly articles, using data from the Microsoft Academic Graph.
Mining the literature

The complexity of VEGF-VEGFR interactions lends itself to qBio + cBio=sBio

PIGF
- PIGF₁: Exons 1-5, 6, 7
- PIGF₂: Exons 1-5, 6, 7

VEGF-B
- VEGF-B₁₆₇: Exons 1-5, 6a, 7
- VEGF-B₁₈₆: Exons 1-5, 6a, 6b, 7

VEGF-A
- VEGF-A₁₂₁: Exons 1-5, 8
- VEGF-A₁₄₅: Exons 1-5, 6a, 8
- VEGF-A₁₆₅₈: Exons 1-5, 7, 8
- VEGF-A₁₆₅₈b: Exons 1-5, 7, 9
- VEGF-A₁₈₃: Exons 1-5, 6a, 7, 8
- VEGF-A₁₈₉: Exons 1-5, 6a, 7, 8
- VEGF-A₂₀₆: Exons 1-5, 6a, 6b, 7, 8

VEGF-C

VEGF-D

In collaboration with P.I. Imoukhuede, University of Illinois
Science of Science

Science is the greatest of human inventions. It has solved and continues to solve many of societies most pressing questions in human health, planetary wellness and economic viability. But one of Science's new challenges is the well being of Science itself. The reproducibility crisis, misaligned incentives and evaluations of scientists, literature overload, publication bias, and out-of-date publishing models are just a few of the maladies of Science. Turning the microscope on Science itself - the Science of Science - is the focus of my research.

How is this different than the sociology and history of science, science policy or scholarly communication? Overlaps exist and methods are borrowed from these established disciplines, but the difference is the scale and kind of data, the methods and tools from data science, and the amalgam of these disciplines under one roof. It is difficult to understand literature overload or the reproducibility crisis if one does not examine, in parallel, what drives scientists to publish, what technologies they use to disseminate their findings, and their established norms for publishing.
Why should you care about big data?

Jobs

Privacy
Enjoy the wave but be cautious...
Big Data involves people
“Data is increasingly digital air: the oxygen we breathe and the carbon dioxide that we exhale. It can be a source of both sustenance and pollution.” -- Dana Boyd

Data Science for Social Good
Jevin West

jevinw@uw.edu

@jevinwest

Website: jevinwest.org

Lab: datalab.ischool.uw.edu