CMPS 4450

Data Mining and Visualization

Dr. Chengwei Lei
CEECS
California State University, Bakersfield
Data Scientists are in high demand
Also in academia

**WHITE HOUSE TO UNIVERSITIES: WE NEED MORE DATA SCIENTISTS**

NEW YORK UNIVERSITY, UNIVERSITY OF CALIFORNIA-BERKELEY, AND THE UNIVERSITY OF WASHINGTON ARE LAUNCHING A $37.8 MILLION PROJECT TO BOOST THE NUMBERS OF AMERICAN DATA SCIENTISTS.

BY MEAL UNIVERSITY

It’s official: America needs more data scientists. This week, a $37.8 million project...
Big Data, Big Paycheck

Median salary for analytics professionals and those specifically within data science, by level of experience.

- **Up to 3 years**
  - Analytics professionals: $65,000
  - Data scientists: $80,000
- **4 to 8 years**
  - Data scientists: $85,000
  - Analytics professionals: $120,000
- **9+ years**
  - Analytics professionals: $115,000
  - Data scientists: $150,000

Note: Data do not include managers. Source: Burtch Works. The Wall Street Journal.
Traditional Hypothesis Driven Research

Hypothesis → Design → Experiment → Data → Data analysis → Result
Data Driven Science

No Prior Hypothesis
New Science of Data
Demand will outpace supply
Data Scientist Job Trend in last 3 years

Job postings: 0.151%
Jobseeker interest: 0.074%

Source: indeed.com
Data Science: Why all the Excitement?

e.g.,
Google Flu Trends:

Detecting outbreaks two weeks ahead of CDC data

New models are estimating which cities are most at risk for spread of the Ebola virus.
Why the all the Excitement?

**elections2012**

Numbers nerd Nate Silver's forecasts prove all right on election night
FiveThirtyEight blogger predicted the outcome in all 50 states, assuming Barack Obama's Florida victory is confirmed

_Luke Harding_
[guardian.co.uk](http://guardian.co.uk), Wednesday 7 November 2012 10.45 EST
...that was just one of several ways that Mr. Obama’s campaign operations, some unnoticed by Mr. Romney’s aides in Boston, helped save the president’s candidacy. In Chicago, the campaign recruited a team of behavioral scientists to build an extraordinarily sophisticated database

...that allowed the Obama campaign not only to alter the very nature of the electorate, making it younger and less white, but also to create a portrait of shifting voter allegiances. The power of this operation stunned Mr. Romney’s aides on election night, as they saw voters they never even knew existed turn out in places like Osceola County, Fla.


The White House Names Dr. DJ Patil as the First U.S. Chief Data Scientist, Feb. 18th 2015
The unreasonable effectiveness of Deep Learning (CNNs)

2012 Imagenet challenge:
Classify 1 million images into 1000 classes.
The unreasonable effectiveness of Deep Learning (CNNs)

Performance of deep learning systems over time:

Krizhevsky, Sutskever, and Hinton, NIPS 2012
Where does data come from?
“Big Data” Sources

It’s All Happening Online
Every:
- Click
- Ad impression
- Billing event
- Fast Forward, pause,
- Server request
- Transaction
- Network message
- Fault
...

User Generated (Web & Mobile)
...

Internet of Things / M2M

Health/Scientific Computing

Baseline information
Cost of genome sequencing compared with Moore’s law for computers

Source: Broad Institute
Graph Data

Lots of interesting data has a graph structure:
- Social networks
- Communication networks
- Computer Networks
- Road networks
- Citations
- Collaborations/Relationships
- …

Some of these graphs can get quite large (e.g., Facebook user graph)
There's certainly a lot of it!

Data produced each year

1 Zettabyte
1 Exabyte
1 Petabyte

Human brain's capacity

1 Petabyte == 1000 TB
1 TB = 1000 GB


References


5 EB
161 EB
800 EB
1.8 ZB
8.0 ZB

120 PB
60 PB
14 PB

100-years of HD video + audio

5 EB
161 EB
800 EB
1.8 ZB
8.0 ZB

Data, data everywhere…

1 TB = 1000 GB

(1/How much data can the human brain store) 2002
5 EB

- How much info


(logarithmic scale)
“Data Science” an Emerging Field

O’Reilly Radar report, 2011
Data Science – A Definition

Data Science is the science which uses computer science, statistics and machine learning, visualization and human-computer interactions to collect, clean, integrate, analyze, visualize, interact with data to create data products.
Goal of Data Science

Turn data into data products.
How to use data?

- Data => exploratory analysis => knowledge models => product / decision making
- Data => predictive models => evaluate / interpret => product / decision making
Data Scientist’s Practice

- Digging Around in Data
- Clean, prep
- Hypothesize
- Model
- Evaluate
- Interpret
- Large Scale Exploitation
Example data science applications

- **Marketing**: predict the characteristics of high life time value (LTV) customers, which can be used to support customer segmentation, identify upsell opportunities, and support other marking initiatives.

- **Logistics**: forecast how many of which things you need and where will we need them, which enables learn inventory and prevents out of stock situations.

- **Healthcare**: analyze survival statistics for different patient attributes (age, blood type, gender, etc.) and treatments; predict risk of re-admittance based on patient attributes, medical history, etc.
More Examples

- Transaction Databases → Recommender systems (NetFlix), Fraud Detection (Security and Privacy)

- Wireless Sensor Data → Smart Home, Real-time Monitoring, Internet of Things

- Text Data, Social Media Data → Product Review and Consumer Satisfaction (Facebook, Twitter, LinkedIn), E-discovery

- Software Log Data → Automatic Trouble Shooting (Splunk)

- Genotype and Phenotype Data → Epic, 23andme, Patient-Centered Care, Personalized Medicine
Data Science – One Definition

Hacking Skills

Math & Statistics Knowledge

Machine Learning

Data Science

Substantive Expertise

Danger Zone!

Traditional Research
Why “Danger Zone?”

Ronny Kohavi* keynote at KDD 2015

- People are incredibly clever at explaining “very surprising results”. Unfortunately most very surprising results are caused by data pipeline errors.

- Beware “HiPPOs” (Highest Paid-Person’s Opinion)

* General Manager for Microsoft’s Analysis and Experimentation Team
What’s Hard about Data Science

- Overcoming assumptions
- Making ad-hoc explanations of data patterns
- Overgeneralizing
- Communication
- Not checking enough (validate models, data pipeline integrity, etc.)
- Using statistical tests correctly
- Prototype → Production transitions
- Data pipeline complexity (who do you ask?)
Data Science concerns

New Gumshoes Go Deep With Data

A two-foot-long Lego model of the Imperial Star Destroyer from 'Star Wars' perches on the center table in the Hello Kitty-themed boardroom. Elsewhere in the building, taped-together cardboard boxes are piled to the ceiling. Amid the jumble of Care Bears, soda cans and packs of playing cards for 'Magic: The Gathering' are camping tents and sleeping bags. This isn’t kindergarten. It’s “homesteading” week at Palantir, the “big-data” company that’s the talk of Silicon Valley.

Once a year, employees (over 700 of them, 75% engineers) ward off corporate culture by living like this: a reminder of their rise from apartment start-up. Palantir occupies a growing number of buildings in downtown Palo Alto, and world-wise. It was recently valued at $4 billion.

Palantirians are the new Googlers. What search algorithms were to the 1990s, big data is today: a game changer. Imagine statistical analysis on steroids. Now multiply that.

Data Science: Buyer Beware

Any field of study followed by the word “science”, so goes the old wheeze, is not really a science, including computer science, climate science, police science, and investment science. And then there is the
Data Makes Everything Clearer?

Epidemiological modeling of online social network dynamics
John Cannarella¹, Joshua A. Spechler¹,*
1 Department of Mechanical and Aerospace Engineering, Princeton University, Princeton, NJ, USA
* E-mail: Corresponding spechler@princeton.edu

Abstract

The last decade has seen the rise of immense online social networks (OSNs) such as MySpace and Facebook. In this paper we use epidemiological models to explain user adoption and abandonment of OSNs, where adoption is analogous to infection and abandonment is analogous to recovery. We modify the traditional SIR model of disease spread by incorporating infectious recovery dynamics such that contact between a recovered and infected member of the population is required for recovery. The proposed infectious recovery SIR model (irSIR model) is validated using publicly available Google search query data for “MySpace” as a case study of an OSN that has exhibited both adoption and abandonment phases. The irSIR model is then applied to search query data for “Facebook,” which is just beginning to show the onset of an abandonment phase. Extrapolating the best fit model into the future predicts a rapid decline in Facebook activity in the next few years.
Data Makes Everything Clearer?

Searches for “MySpace”

Searches for “Facebook”

Figure 3: Data for search query “Myspace” with best fit (a) SIR and (b) InSIR models overlaid. The search query data are normalized such that the maximum data point corresponds to a value of 100.
Data Makes Everything Clearer?

In keeping with the scientific principle “correlation equals causation,” our research unequivocally demonstrated that Princeton may be in danger of disappearing entirely. Looking at page likes on Facebook, we find the following alarming trend:

and based on Princeton search trends:

“This trend suggests that Princeton will have only half its current enrollment by 2018, and by 2021 it will have no students at all,...

http://techcrunch.com/2014/01/23/facebook-losing-users-princeton-losing-credibility/