ICOS big data camp

June 5-9, 2017 Co-sponsored by ICOS and MIDAS

Who is everybody?

- Executive producer:
 - ~ Teddy DeWitt
- Producers:
 - Jerry Davis, Cliff Lampe, Brian Noble, Jason Owen-Smith
- Code Concierges (CoCons):
 - ~ Nivi Karki, Ronnie Lee, Jeff Lockhart, Oskar Singer



What are we up to this week?

- Monday: overview, SQL, project group formation
- Tuesday: Python and its uses
- Wednesday: Python for human language; using APIs
- Thursday: Python for data analysis
- Friday: write <u>"Social capital asset pricing</u> model (SCAPM)" app for iPhone, sell to Facebook for \$10B, quit grad school



What does social life look like today?

Consultant running meeting on Google Hangouts

Real estate agent checking listings

Journalist applying for job



Student writing paper for class

Professor grading papers

Activist uploading files to Wikileaks



The job description for 90% of the people at the University of Michigan:

"Stare at a screen and type on a keyboard"



Computational social science: Making the links

From e-mails to social networks, the digital traces left by life in the modern world are transforming social science.

Jim Giles

22 August 2012 Corrected: 31 August 2012



Jon Kleinberg's early work was not for the mathematically faint of heart. His first publication¹, in 1992, was a computer-science paper with contents as dense as its title: 'On dynamic Voronoi diagrams and the minimum Hausdorff distance for point sets under Euclidean motion in the plane'.

That was before the World-Wide Web exploded across the planet, driven by millions of individual users making independent decisions about who and what to link to. And it was before Kleinberg began to study the vast array of digital by-products generated by life in the modern world, from e-mails, mobile phone calls and credit-card purchases to Internet searches and social networks. Today, as a computer scientist at Cornell University in Ithaca, New York, Kleinberg uses these data to write papers such as 'How bad is forming your own opinion?'² and 'You had me at hello: how phrasing affects memorability'³ — titles that would be at home in a social-science journal.

"I realized that computer science is not just about technology," he explains. "It is also a human topic."

Kleinberg is not alone. The emerging field of computational social science is attracting mathematically inclined scientists in ever-increasing numbers. This, in turn, is spurring the creation of academic departments and prompting companies such as the social-network giant Facebook, based in Menlo Park, California, to establish research teams to understand the structure of their networks and how information spreads across them.

"It's been really transformative," says Michael Macy, a social scientist at Cornell and one of 15 co-authors of a 2009

Related stories

- Facebook 'likes' the scientific method
- Sociology of science: Big data deserve a bigger audience
- · Social science: Open up

Thanks to ICTs, economics today is "roughly where astronomy was when the telescope was invented or where biology was when the microscope was invented." (Robert Shiller, certified smart guy)

HOW SHOULD THE PERVASIVE "MEDIATION" OF CONTEMPORARY SOCIAL LIFE AFFECT SOCIAL SCIENCE?



<u>Google Trends: the gateway</u> drug for big data



NEW INSIGHTS INTO TRADITIONAL TOPICS



Does racism influence voting?

Ehe New York Eimes Campaign Stops

STRONG OPINIONS ON THE 2012 ELECTION

JUNE 9, 2012, 5:46 PM

How Racist Are We? Ask Google

By SETH STEPHENS-DAVIDOWITZ

Barack Obama won 52.9 percent of the popular vote in 2008 and 365 electoral votes, 95 more than he needed. Many naturally concluded that prejudice was not a major factor against a black presidential candidate in modern America. My research, a comparison of Americans' Google searches and their voting patterns, found otherwise. If my results are correct, racial animus cost Mr. Obama many more votes than we may have realized.

Quantifying the effects of racial prejudice on voting is notoriously problematic. Few people admit bias in surveys. So I used a new tool, Google Insights, which tells researchers how often words are searched in different parts of the United States.

Can we really quantify racial prejudice in different parts of the country based solely on how often certain words are used on Google? Not perfectly, but remarkably well. Google, aggregating information from billions of searches, has an uncanny ability to reveal meaningful social patterns. "God" is Googled more often in the Bible Belt, "Lakers" in Los Angeles.

The conditions under which people use Google - online, most likely alone, not participating in an official survey - are ideal for capturing what they are really thinking and feeling. You may have typed things into Google that you would hesitate to admit in polite company. I certainly have. The majority of Americans have as well: we Google the word "porn" more often than the word "weather."

And many Americans use Google to find racially charged material. I performed the somewhat unpleasant task of ranking states and media markets in the United States based on the proportion of their Google searches that included the word "nigger(s)." This word was included in roughly the same number of Google searches as terms like "Lakers," "Daily Show," "migraine" and "economist."

Racially Charged Web Searches and Voting

In 2008, Barack Obama performed much worse than expected in areas with the greatest frequency of racially charged Google searches. The analysis looked at Web searches in 200 United States media markets.



Send Feedback

Does racism influence voting?



MICHIGAN M ROSS SCHOOL OF BUSINESS

NEW INSIGHTS INTO NEW TOPICS



If only someone would come up with a way to gather horrifyingly intrusive personal information online...





The Data That Turned the World Upside Down

HG HANNES GRASSEGGER & MIKAEL KROGERUS Jan 28 2017, 9:15am

ump during the presidential campaign. Image: Gage Skidmore

Psychologist Michal Kosinski developed a method to analyze people in minute detail based on their Facebook activity. Did a similar tool help propel Donald Trump to victory? Two reporters from Zurich-based Das Magazin went data-gathering.

Donald

An earlier version of this story appeared in <u>Das Magazin</u> in December.

 \equiv

On November 9 at around 8.30 AM., Michal Kosinski woke up in the Hotel Sunnehus in Zurich. The 34-year-old researcher had come to give a lecture at the Swiss Federal Institute of Technology (ETH) about the dan**MOTHERBOARD** e digital revolution. Kosinski gives regular lectures on this topic all over the world. He is a leading expert in psychometrics, a data-driven sub-branch of psychology. When he turned on the TV that morning, he saw that the bombshell had exploded: contrary to forecasts by all leading statisticians, Donald J. Trump had been elected president of the United States.

ICTs and social movements



10

One Facebook post





Who "dates" whom in an Ohio high school







Are Tinder and Grindr actually field experiments created by a rogue epidemiologist at the School of Public Health?*

*Note: if you do not know what Tinder and Grindr are, DO NOT GOOGLE THEM!



Surprising sources of network data

January 26, 2011 12:05 PM PST

Senator proposes mobile-privacy legislation

by Declan McCullagh

132

Federal law needs to be updated to halt the common police practice of tracking the whereabouts of Americans' mobile devices without a search warrant, a Democratic senator said today.

<u>Ron Wyden</u>, an Oregon Democrat, said it was time for Congress to put an end to this privacy-intrusive practice, which the Obama Justice Department has <u>sought to defend in court</u>.

In an <u>luncheon speech</u> at the libertarian <u>Cato Institute</u> in Washington, D.C., Wyden said his staff was drafting legislation to restore "the balance necessary to protect individual rights" by requiring police to obtain a search warrant signed by a judge before obtaining location information.

Even though police are tapping into the locations of mobile phones thousands of times a year, the legal ground rules remain hazy, and courts <u>have been divided</u> on the constitutionality and legality of the controversial practice. In September, the first federal appeals court to <u>rule on the legality</u> indicated that no search warrant was needed, but sent the case back to a district judge for further proceedings.



Sen. Wyden (right) tells Cato Institute audience that tracking cell phones is as privacy-invasive as searching someone's home.

MICHIGAN M ROSS SCHOOL OF BUSINESS

THE WALL STREET JOURNAL.

WSJ.com

Apple, Google Collect User Data

By JULIA ANGWIN And JENNIFER VALENTINO-DEVRIES



WSJ.com Senior Technology Editor Julia Angw in reports Apple's iPhone and Google's Android regularly transmit user location data back to those companies, based on data analyzed by The Wall Street Journal.

<u>Apple</u> Inc.'s iPhones and <u>Google</u> Inc.'s Android smartphones regularly transmit their locations back to Apple and Google, respectively, according to data and documents analyzed by The Wall Street Journal—intensifying concerns over privacy and the widening trade in personal data.

An office like yours...

- Location tracking data were collected over 71 days from 40 tags
- Hatched area is shadow area where signal is unreachable
- Red Dots denote occupied workstations



Mapped signals

- The tag generates signals when the tag is moving, and goes to sleep mode when there's no movement
- The recorded signals are below (total 35 million records)
- The recorded signal has the information of [tag id, x, y, t]



Space utilization by each person

- His office is in XX area. He reports to the director so many dots in front of the director's secretary
- He leads two team and often talks with one of the team's manager



- Her workstation is obvious
- She uses the copy machine often
- She works closely with her team members



Identified interactions



- Total 10377 interactions are identified.
 - ~ 220 interactions/day
 - ~ 11 interactions/day/person.

The Technology Helping Repressive Regimes Spy



iStockphoto.com Software allows repressive governments to track and monitor the movements of dissidents. December 14, 2011

text size A A A

As protesters in the Middle East use social media to organize and communicate, the regimes they're battling are using sophisticated technology to intercept their emails, text messages and cellphone calls.

On Wednesday's *Fresh Air*, journalist Ben Elgin talks about a Bloomberg News series, "Wired for Repression," which details how Western companies are selling surveillance technology to regimes including Iran, Syria, Bahrain and Tunisia.

Those regimes have then used the information

obtained from those technologies to torture protesters and dissidents, Elgin tells *Fresh Air* contributor Dave Davies.

"[One Iranian engineer] became caught up in the protest movements after the election of 2009 and he was arrested. He was beaten and put into prison and interrogated 14 times over 50 days," Elgin says. "During these interrogations, not only was he presented with [his] text message transcripts; he was presented with a very sophisticated diagram of who he had called, and then who those people had called. And he was interrogated on every connection within his network of contacts."

The engineer had worked for Ericsson AB, where he had helped install the systems that would later be used in his interrogations.

"The damage that can be done was suddenly very clear to him," Elgin says. "And that led him to want him to talk [to me]. [He] has since fled [Iran], which has made it easier for him."

January 29, 2013, 7:29 PM ET IBM Security Tool Can Flag 'Disgruntled Employees'

A new <u>International Business Machines</u> Corp. security tool uses Big Data to help CIOs detect internal and external security threats in new ways—and can even scan email and social media to flag apparently "disgruntled" employees who might be inclined to reveal company secrets, according to Sandy Bird, chief technology officer of IBM's security systems division.

The new tool, called IBM Security Intelligence with Big Data, is designed to crunch decades worth of emails, financial transactions and website traffic, to detect patterns of security threats and fraud. Beyond its more conventional threat prevention applications, the new platform, based on Hadoop, a framework that processes data-intensive queries across clusters of computers, will allow CIOs to conduct sentiment analysis on employee emails to determine which employees are likely to leak company data, Mr. Bird said. That capability will look at the difference between how an employee talks about work with a colleague and how that employee discusses work on public social media platforms, flagging workers who may be nursing grudges and are more likely to divulge company information. "By analyzing email you can say this guy is a disgruntled employee and the chance that he would be leaking data would be greater," Mr. Bird said of IBM's new tool.

For example, a company could analyze employee emails that express a positive sentiment to a manager at work, but detect "when he's talking to a peer or someone outside the company, the sentiment comes out a little different," Mr. Bird said. Such a pattern, combined with other factors, could cause an employee to be flagged for more investigation by an IT team. Sentiment analysis works by parsing patterns in words and phrases that signify whether the intent behind a message is likely positive, negative or neutral.

The platform also helps companies protect against hacker attacks and fraud by allowing security personnel to look for patterns in past attacks — like the time or location of attempted intrusions, and the applications that have been targeted.

As corporate fears about data leakage and hacker attacks rise, CIOs are being called on to quickly defend against intrusions of increasing sophistication. More and more companies are using Big Data to discover the pattern of security lapses as they struggle

A deep philosophical point:

A web page does not exist until you perceive it. (Whoah)





Interdisciplinary Committee on **O**rganizational **S**tudies

Enhancing the University of Michigan's strength as a world center for research and scholarship on organizations.

Login

Activities

- Lectures
- Research Support
- The ICOS Course

About

- ICOS
- Our Website

Community

- Groups
- Courses of Interest
- Other Communities
- People Directory
- Related Events

Latest posts from other organization studies blogs

orgtheory.net

- just your average minimalist finnish altfolk group
- new article on insider activism by briscoe and gupta
- bad reporting on bad science

Places the ICOS codirectors find fun

- Paul Krugman's blog
- Planet Money Blog: NPR

Upcoming lecture

Arne Kalleberg, University of North Carolina, Sociology

Institutions, Precarious Work and Inequality

Friday, September 30, 2016 1:30 - 3:00 PM Room R1220, Ross School of Business Introduced by Lindsey Cameron, Management & Organizations



Most Recent Lecture

- Gretchen Spreitzer, Lyndon Garrett, & Pete Bacevice, University of Michigan
- Can coworking enable thriving at work? Insights for the new world of work

Spotlight Lecture

- Part I: Relational Perspectives and Organizations
- Four Ways to Organize Any Social Interaction

Popular Tags

corporate governance social movements creativity design healthcare Sensemaking leadership collaboration organization theory economic sociology innovation social networks more tags

4:41 PM

Interdisciplinary Committee on Organizational Studies (ICOS)

100

M Inbox (5) - gfdavis@umich 🗙 🛛 🚾 University of Michigan - C 🗙 🗸 🖉 What's Wrong With Inegu 🗙 📄 view-source:icos.umich.ed 🔕 🔠 Surface Pro Print Screen 📔 🗙 view-source:icos.umich.edu C 👖 Apps 💿 The Enspiral Network 🛛 👧 Allied Media Project... 📄 Epic Innovation 🕜 Hatch · Tell stories ... 🍸 Taxes - Neighborho... 🦌 Databases, Tables &... 🔀 Charlie App | Make ... 🔞 SEC.gov | Company ... » Dther bookmarks <meta name= google-site-verification Content= inkqbxgzyop-zonbriniwonykrgngnbboky-4gpbj/0</pre> </head> 49 50 <body> 51 52 <div id="top"> <div id="logo"> </div> 54 <div id="block-menu-menu-login">Login 55 56 <div id="block-menu-menu-register"></div> <form action="/" accept-charset="UTF-8" method="post" id="search-theme-form"> 58 <div><div id="search" class="container-inline"> <div class="form-item" id="edit-search-theme-form-1-wrapper"> <label for="edit-search-theme-form-1"> </label> 61 <input type="text" maxlength="128" name="search theme form" id="edit-search-theme-form-1" size="15" value="" title="Enter the terms you wish to search for." class="form-text" /> 62 63 </div> 64 <input type="submit" name="op" id="edit-submit" value="Search" class="form-submit" /> 65 <input type="hidden" name="form build id" id="form-twrTD 8RV 0wM6eANRTuLgrru3J05qqM4eyISGEIS10" value="form-twrTD 8RV 0wM6eANRTuLgrru3J05qqM4eyISGEIS10" /> 66 <input type="hidden" name="form id" id="edit-search-theme-form" value="search theme form" /> 67 </div> 68 69 </div></form> 70 </div> 72 <div id="menu"> 73 </div> 74 75 </div> 76 <div id="main-container"> <div id="sidebar-left"> <div class="block block-menu" id="block-menu-menu-activities"> <h2 class="title">Activities</h2> <div class="content">lectures 81 82 Research Support 83 The ICOS Course 84 </div> </div> 85 <div class="block block-menu" id="block-menu-menu-about"> <h2_class="title">About</h2> 87 <div class="content">ICOS 89 Our Website 90 </div> </div> 91 <div class="block block-menu" id="block-menu-menu-community"> 92 <h2 class="title">Communitv</h2> 93 <div class="content">Groups 94 95 Courses of Interest 97 People Directory 98 Related Events 99 </div> 100 <div class="block block-views" id="block-views-Feed block-block 1"> 101 <h2 class="title">Latest posts from other organization studies blogs</h2> 102 <div class="content"><div class="view view-Feed-block view-id-Feed block view-displav-id-block 1 view-dom-id-4"> 103

Some big data questions

- Where do I get "big data"? Is there some secret handshake I need?
- What does it look like?
- How do I make gigabytes of words and numbers into something meaningful?
- If I can't learn to do everything I need about big data in a week, where can I go next?



How big is big data?

- Visit your favorite website (e.g., <u>www.umich.edu</u>)
- Right-click and "View page source"
- Wait, what *is* all this stuff?
- Search for <u>http</u>
- Is there some convenient way to search through all this junk online, copy it, and drop it into a database for future use? (Will the site's owner get mad?)
- Is there an easier way to just download all this stuff in bulk?



A method and three tools to start

- The method: learning in groups (cf. "agile software development")
- The tools:
 - SQL: how to manipulate those databases underlying what you see on the Web
 - Python: a pretty good open-source programming language
 - ~ APIs: how to get them to talk to you



BE NOT AFRAID: LESSONS FROM "COMPUTER SCIENCE"

Plagiarized from the estimable Prof. Brian Noble



We're All Charlatans

- Computer science: not a science
 - ~ Few "natural laws" because it is a human construction
 - ~ Exception: "This sentence is false." (1/3 of EECS 376)
- Software engineering: not an engineering discipline
 - ~ Engineering: static/dynamic modeling, safety margins, etc.
 - ~ Software: "Recovery-oriented computing" (1/5 of EECS 582)
- A culture of decentralized collaborative tinkering
- Facebook: likely the most successful company run this way



MOVE FAST AND BREAK THINGS

Facebook Rule #1

Don't be afraid to make a mistake

- Everyone makes mistakes!
 - I (Brian Noble) make programming mistakes all the time
 - Students who actually do things make mistakes as well
 - ~ Professional staff at Facebook do too (obviously!)
- Fundamental to the process
 - ~ These are formal languages (vs. natural)
 - Mortals aren't inherently great at this





STAY FOCUSED AND KEEP SHIPPING

Facebook Rule #2

Don't Wait to Find Your Mistakes

- Build a little, test a little
 - ~ "You keep using that word. I do not think it means what you think it means."
 - ~ --Inigo Montoya
- You have an important advantage!
 - ~ CS students believe they are **really good** at this
 - ~ But, **no one** is really good at this, just shades of bad





DONE IS BETTER THAN PERFECT

Facebook Rule #3

Never Fly Solo

- Two people per keyboard, always
 - Everyone is bad at this, but in different ways
 - Only one of you needs to see the problem
- Trade hands-on-keyboard frequently
 - ~ It's tempting to let one person "do the work"
 - ~ You lose much of the benefit this way
- Talk about what you are doing as you do it
 - ~ Forces you to reveal hidden assumptions
 - Catch some mistakes even before you make them





FORTUNE FAVORS THE BOLD



Practical tips

- There are no new problems under the sun
 - ~ Check Google
 - ~ Ask your physical neighbors
 - ~ Ask your virtual neighbors
- Steal, do not invent!
 - ~ Large community with a strong culture of sharing
 - ~ Before writing something, see if someone else has
- Keep versions of things around: your Lab Notebook
 - ~ Explains how you got there
 - ~ In case you have to "go backwards"



Final exams are graded! But just to add some excitement to the process, after completing them, I ran my script to compute totals WHICH CLOBBERED THE GRADED COPIES OF ALL THE EXAMS (I kid you not -- I rarely use caps, but you can imagine my horror). Thank you Apple Time Machine -- it backed up while I was out retrieving daughter and I lost only 2 scores on one problem (and I had actually captured the scores). Had I not had time machine, I fear that things would have gotten very ugly.

Like · Comment · Share · 53 minutes ago · 🤽





WHAT WOULD YOU DO IF YOU WERE NOT AFRAID?

Facebook Rule #5

A Few Caveats

- You can do almost anything, but should you?
 - Intellectual property restrictions on code
 - ~ Terms of Service restrictions on data providers
 - ~ Lots of personally-identifiable information (IRB)
- Computers allow you to make bigger mistakes more quickly
 - ~ What is "science" vs. "stuff I saw somewhere"
 - Our group brought campus-wide storage to its knees
- Get a sense for how this work is received elsewhere
 - ~ Check with advisor(s)



Find one interesting true thing to say about your group's topic by one week from Thursday afternoon, and explain how you got there



QUESTIONS SO FAR?





Your Group Task

1. Use the techniques you are practicing here to collaboratively demonstrate **one plausibly true thing** about a topic that interests you.

2. Reflect on the process of demonstrating that thing

3. Present your finding, your process and the fruits of your reflection to the group **on THURSDAY 06/11**



Over the next week we expect you to

- Form a group (to be done this afternoon)
- Articulate a topic or question of shared interest
- Identify and gather relevant data
- Parse data and insert it into a sql database you design, pay attention to linking variables
- Run queries or other analyses on your data to demonstrate your one true thing
- Prepare a presentation that describes your question, your process, your findings, and what doing this taught you about working with "big data"
- Have fun



SOME EXAMPLES OF TRUE (TRUTHY) THINGS





Men and women review books using different language (scraped and topic modeled data from Goodreads.com)

Another True Thing





Your	Personal	Formal
Data	Texts	Texts
3.22	11.4	4.2
9.17 💥	9.5	8
1.86	2.7	2.6
1.18	2.6	1.6
6.07	7.8	5.4
7.04	5	7.2
19.84	13.1	19.6
	Your Data 3.22 9.17 1.86 1.18 6.07 7.04 19.84	Your Personal Data Texts 3.22 11.4 9.17 9.5 1.86 2.7 1.18 2.6 6.07 7.8 7.04 5 19.84 13.1

The text you submitted was 201,626 words in length.

LIWC Dimension	Your
LIWC Dimension	Data
Self-references (I, me, my)	0.86
Social words	4.56
Positive emotions	1.76
Negative emotions	0.97
Overall cognitive words	4.7
Articles (a, an, the)	8.49
Big words (> 6 letters)	26.64

The text you submitted was 66,726 words in length.

Fox News and the New York Times evince different sentiments in discussions of climate change



Big ten college Facebook posts mostly talk about stuff other than academics

- >Here's who is in our group
- ➢Our motivating question was...
- ≻We tried to answer this by...
- We had to completely change direction when we discovered that...
- ≻Here is our fact:
- Here is how we got there and what we learned along the way

