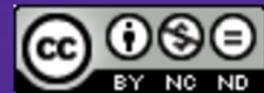


DS-UA 112

Introduction to Data Science

Week 1: Lecture 1

Overview - Solving Problems with Data





Center for
Data Science

What is data? What
characterizes data?

DS-UA 112

Introduction to Data Science

Week 1: Lecture 1

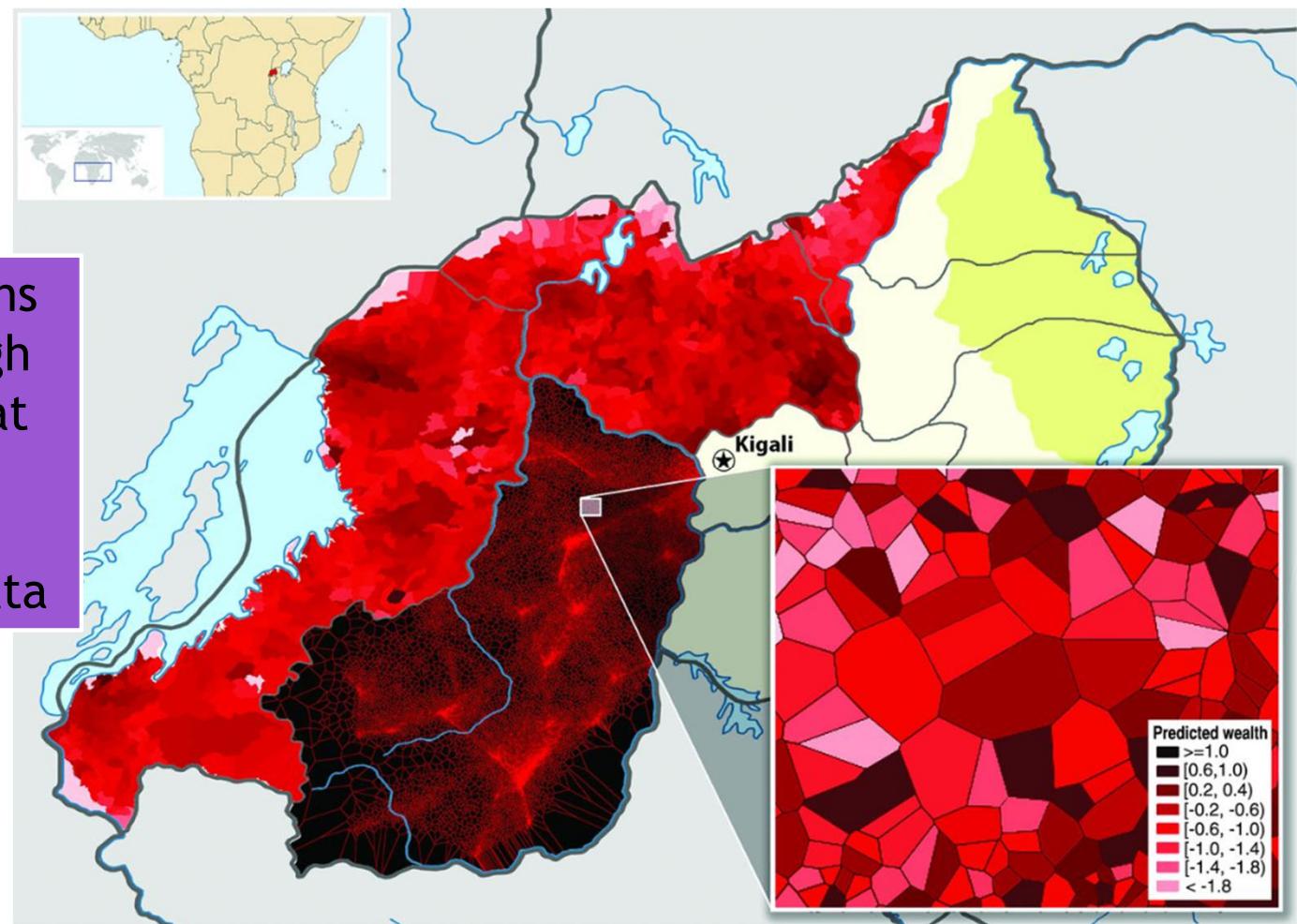
Overview - Solving Problems with Data

Adapted from Nolan, Denero, and Salganik



What is data?

Connecting questions and answers through research design that incorporates information determined from data



What is data?

REPORT

Predicting poverty and wealth from mobile phone metadata

Joshua Blumenstock^{1,*}, Gabriel Cadamuro², Robert On³

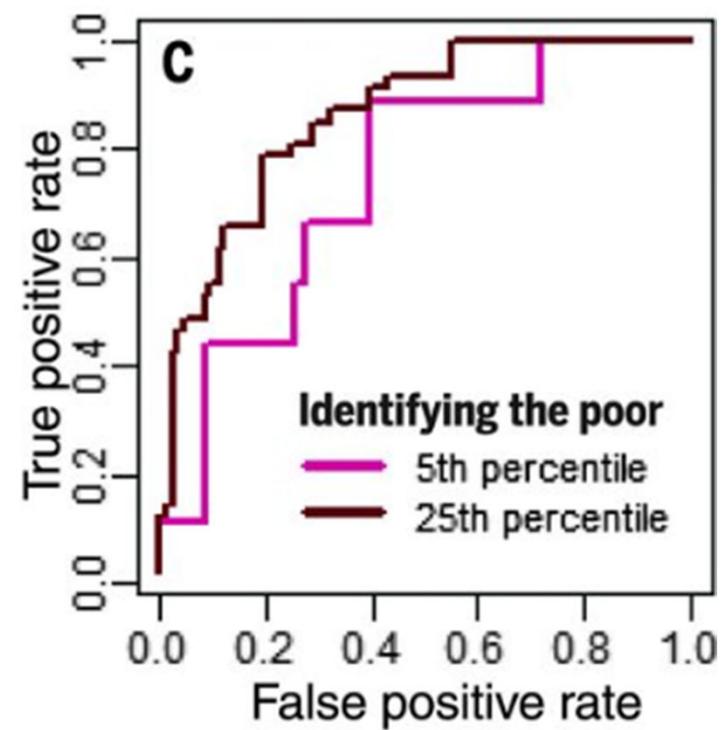
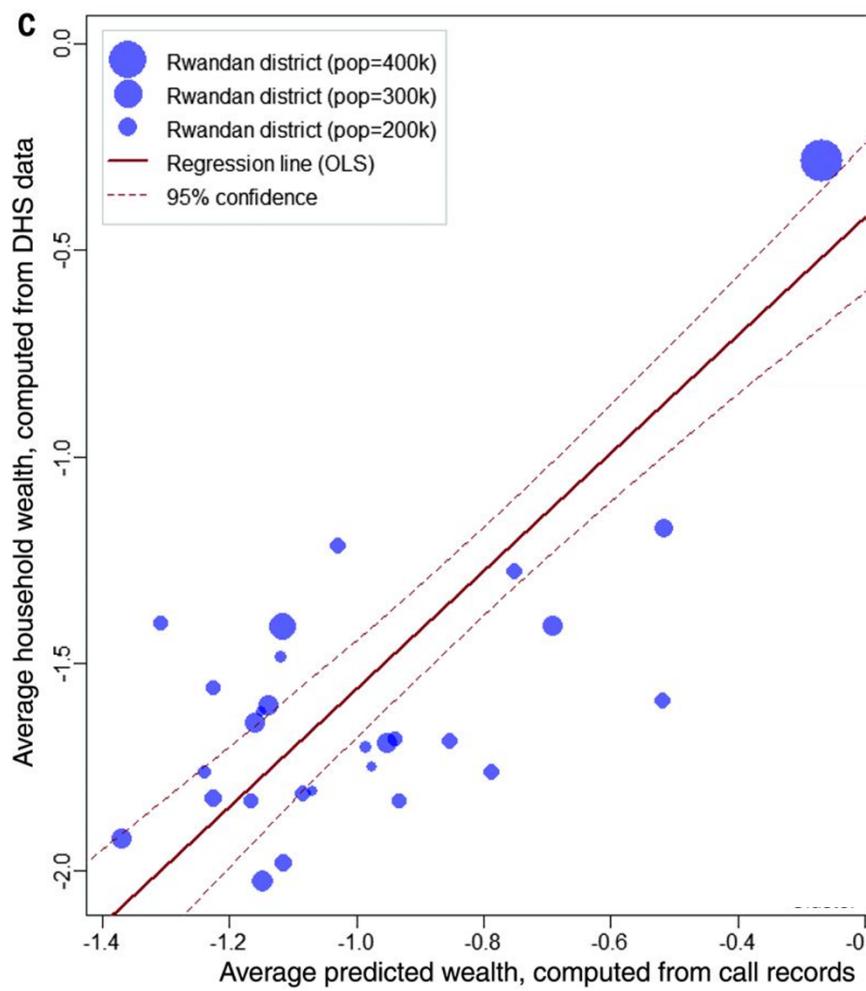
+ See all authors and affiliations

Science 27 Nov 2015:
Vol. 350, Issue 6264, pp. 1073-1076

Abstract

Accurate and timely estimates of population characteristics are a critical input to social and economic research and policy. In industrialized economies, novel sources of data are enabling new approaches to demographic profiling, but in developing countries, fewer sources of big data exist. We show that an individual's past history of mobile phone use can be used to infer his or her socioeconomic status. Furthermore, we demonstrate that the predicted attributes of millions of individuals can, in turn, accurately reconstruct the distribution of wealth of an entire nation or to infer the asset distribution of microregions composed of just a few households. In resource-constrained environments where censuses and household surveys are rare, this approach creates an option for gathering localized and timely information at a fraction of the cost of traditional methods.

What is data?



What is data?

Private traits and attributes are predictable from digital records of human behavior

Michal Kosinski, David Stillwell, and Thore Graepel

PNAS April 9, 2013 110 (15) 5802-5805;

Abstract

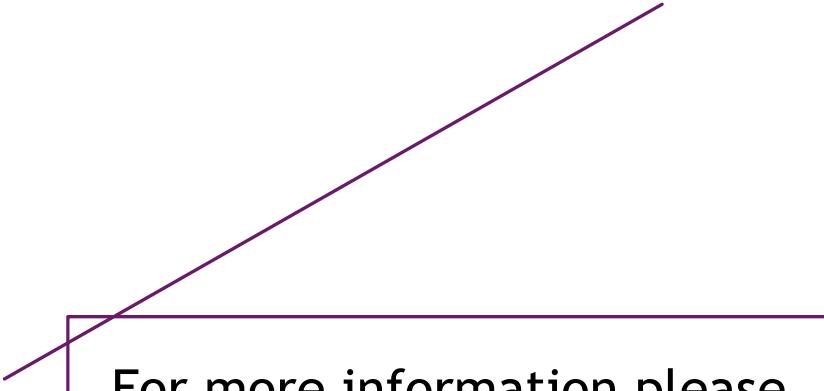
We show that easily accessible digital records of behavior, Facebook Likes, can be used to automatically and accurately predict a range of highly sensitive personal attributes including: sexual orientation, ethnicity, religious and political views, personality traits, intelligence, happiness, use of addictive substances, parental separation, age, and gender. The analysis

What is data?

- ▶ What would these groups gather from the study?
 - ▶ Social Scientists
 - ▶ Computer Scientists
 - ▶ Business Managers
 - ▶ Activists
 - ▶ Policy Makers

Logistics

For more information please refer to the syllabus along with
<https://wp.nyu.edu/idss20/>

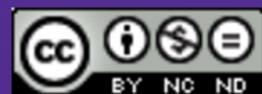


A word cloud graphic featuring various terms related to data science and learning. The words are in different sizes and shades of purple, with larger, darker words being the most prominent. The words include: applied, algorithm, don't, interest, understanding, deep, field, statistics, learning, clean, program, lot, idea, model, fun, set, expect, gain, work, skill, job, code, world, tool, large, hope, good, project, ds, knowledge, real, python, science, application, method, basic, hand, class, making, practical, analyze, experience, library, help, create, expand, actual.



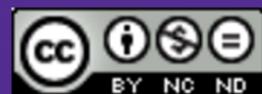
Logistics

- ▶ Instructor: Chris
 - ▶ Office Hours:
Thursdays 11-12PM
at 60 Fifth Avenue,
Room 650
 - ▶ Section Leaders:
Shreyas and Ashwin
 - ▶ Graders: Ruoyu,
Sarthak, Peeyush and
Serkan



Logistics

- ▶ NYU Classes
 - ▶ Weekly Agenda, Zoom Conference, Syllabus
 - ▶ JupyterHub
 - ▶ Class Materials, Submission Labs/Homework/Projects
 - ▶ Piazza
 - ▶ Announcements, Discussion
 - ▶ Gradescope
 - ▶ Submission
 - Homework/Projects,
 - Retrieve Exams



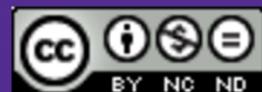
Logistics

- ▶ NYU Classes
 - ▶ Weekly Agenda, Zoom Conference, Syllabus
- ▶ JupyterHub
 - ▶ Class Materials, Submission Labs/Homework/Projects
- ▶ Piazza
 - ▶ Announcements, Discussion
- ▶ Gradescope
 - ▶ Submission Homework/Projects, Retrieve Exams



A word cloud graphic centered on data science, with words like learning, learn, data, science, python, application, etc., in various sizes and shades of purple and grey.

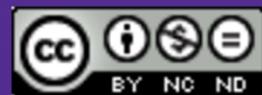
Please complete Survey 0
linked to the Weekly
Agenda on NYU Classes



Announcements

- ▶ Homework 40%
- ▶ Labs 15%
- ▶ Projects 20%
- ▶ Exams
 - ▶ Midterm 10%
 - ▶ Final 15%
- ▶ Participation
- ▶ Office Hours, Piazza

A word cloud centered on the term "data science". Other words include applied, algorithm, don't, interest, understanding, deep, field, statistics, learning, clean, program, model, fun, set, expect, gain, work, lot, idea, skill, job, code, world, tool, large, hope, project, ds, knowledge, real, python, basic, application, method, class, making, practical, analyze, experience, library, help, create, expand, actual.



Announcements

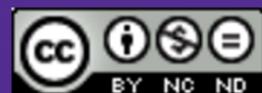
- ▶ Homework 40%
- ▶ Labs 15%
- ▶ Projects 20%
- ▶ Exams
 - ▶ Midterm 10%
 - ▶ Final 15%
- ▶ Participation
 - ▶ Office Hours, Piazza

A word cloud centered on the term "data science". Other words include applied, algorithm, don't, interest, understanding, deep, field, statistics, learning, clean, program, learn, lot, idea, skill, job, code, world, tool, large, hope, expect, gain, work, good, project, ds, knowledge, real, python, basic, application, method, class, making, practical, analyze, help, experience, library, create, expand, actual.



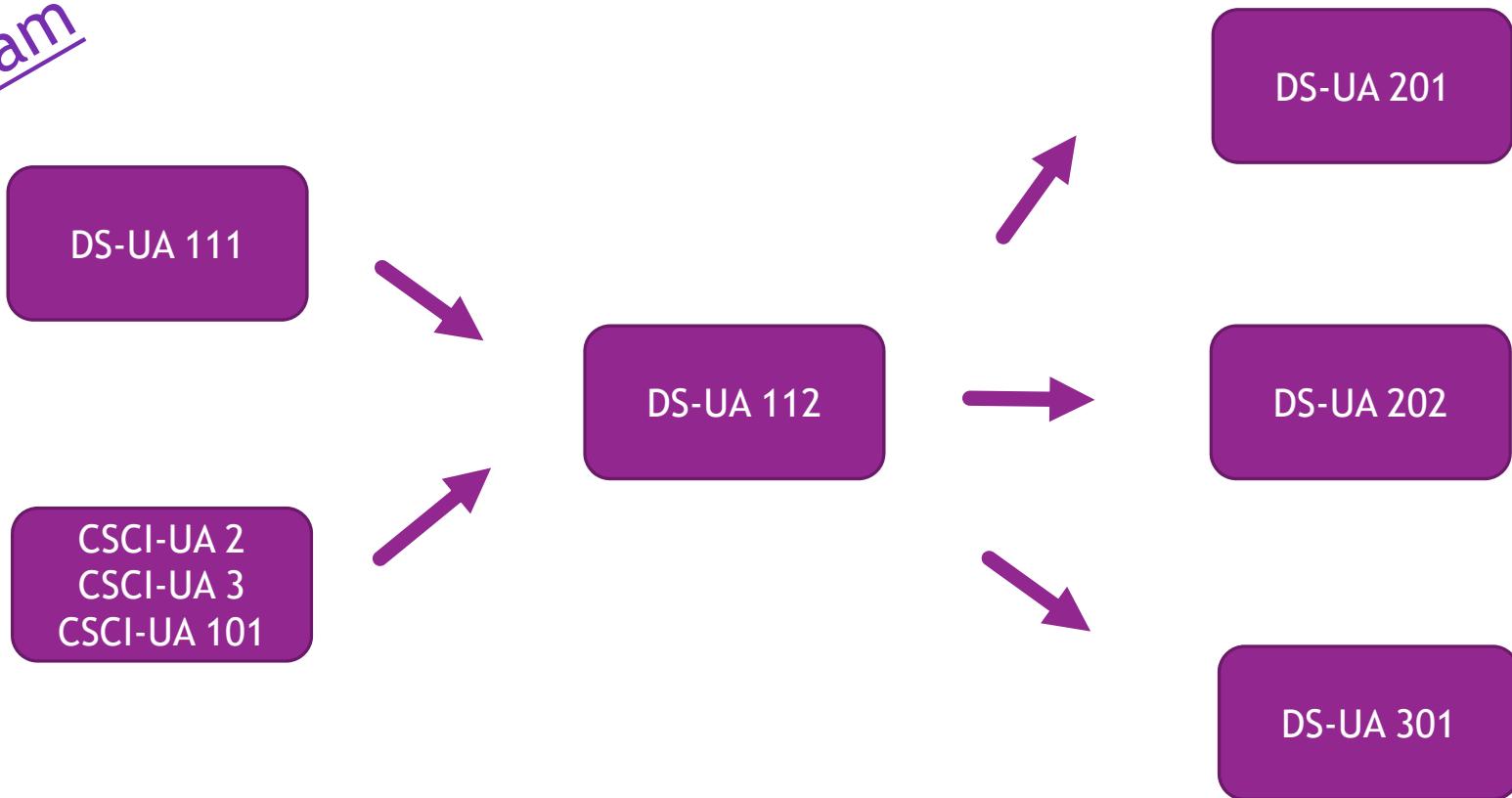
Logistics

- ▶ Part of the Major and Minor in Data Science
- ▶ Interdisciplinary course for students in the sciences and humanities
- ▶ Goals
 - ▶ Empower
 - ▶ Enable
 - ▶ Prepare
- ▶ Skills
 - ▶ Formulate
 - ▶ Process
 - ▶ Access
 - ▶ Visualize
 - ▶ Fit
 - ▶ Evaluate



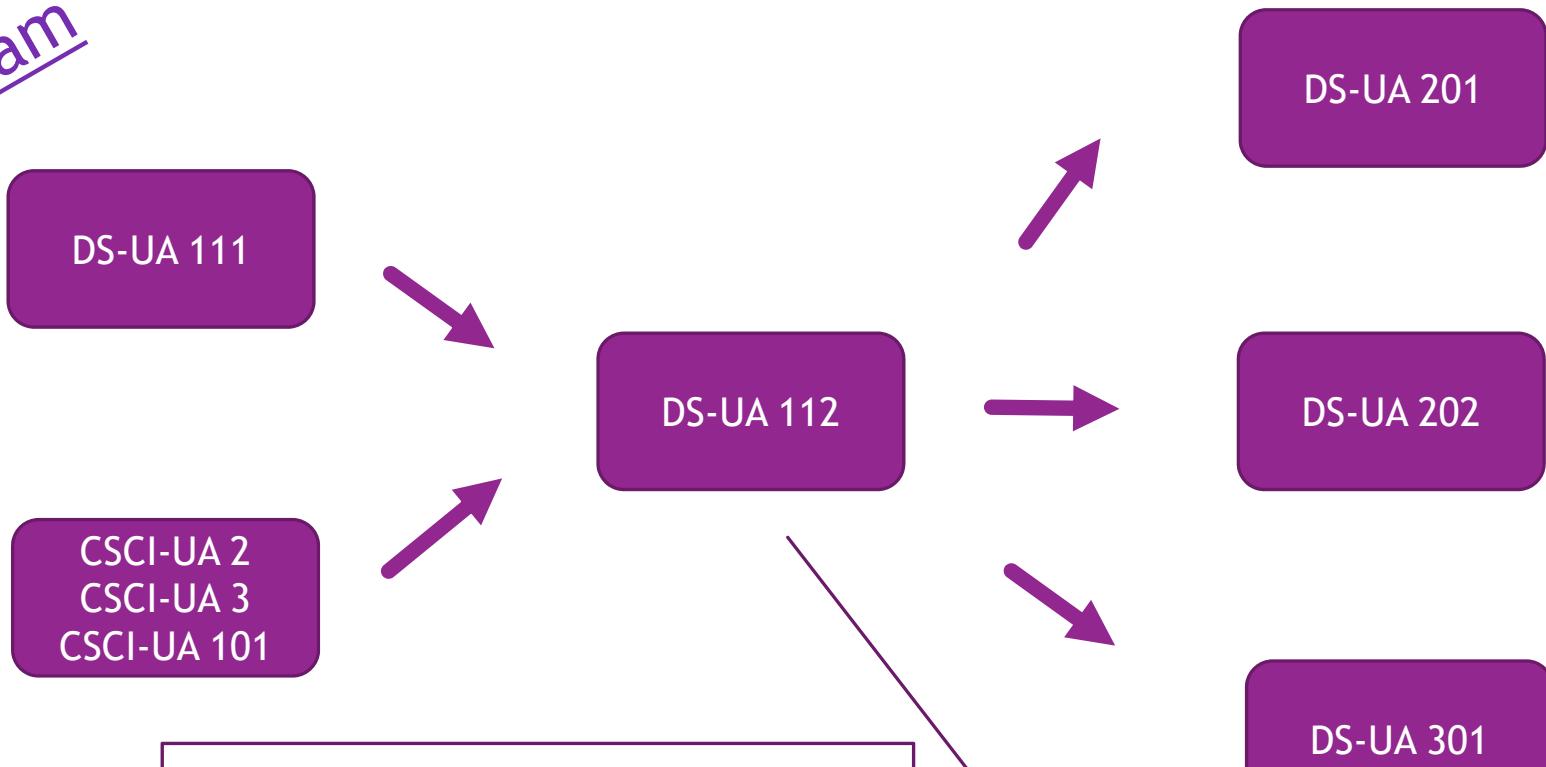
Logistics

DS Program

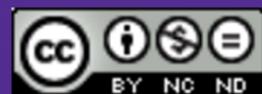


Logistics

DS Program



Access course textbook at
<https://cp71.github.io/textbook/intro>



Logistics

- ▶ DS-UA 111

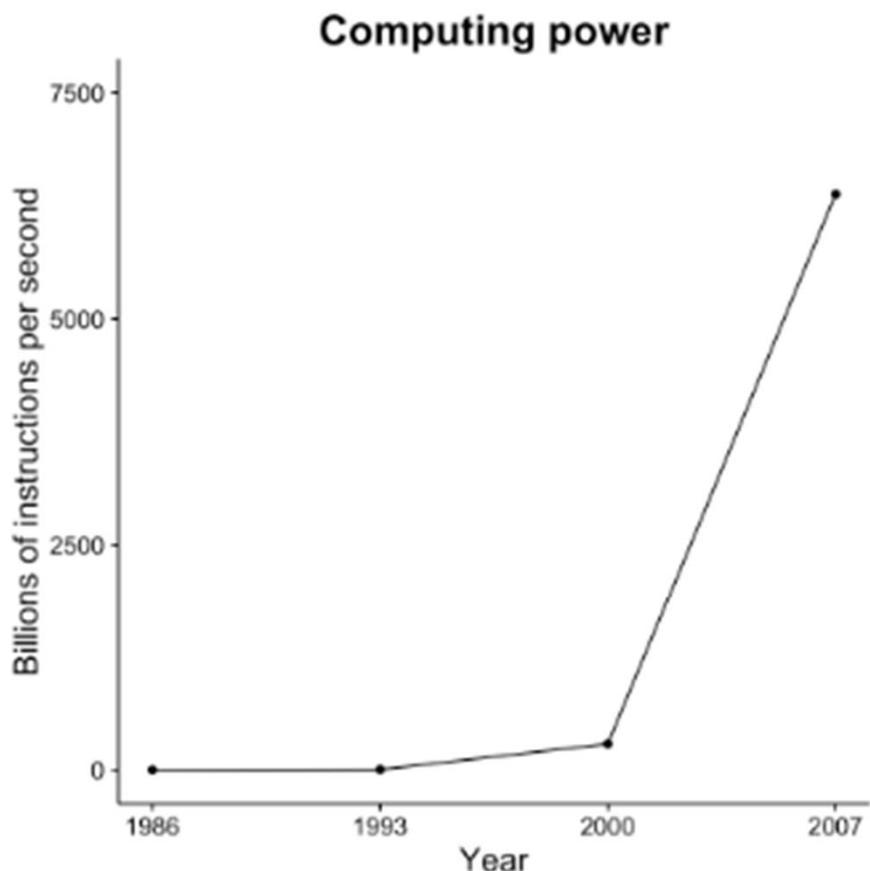
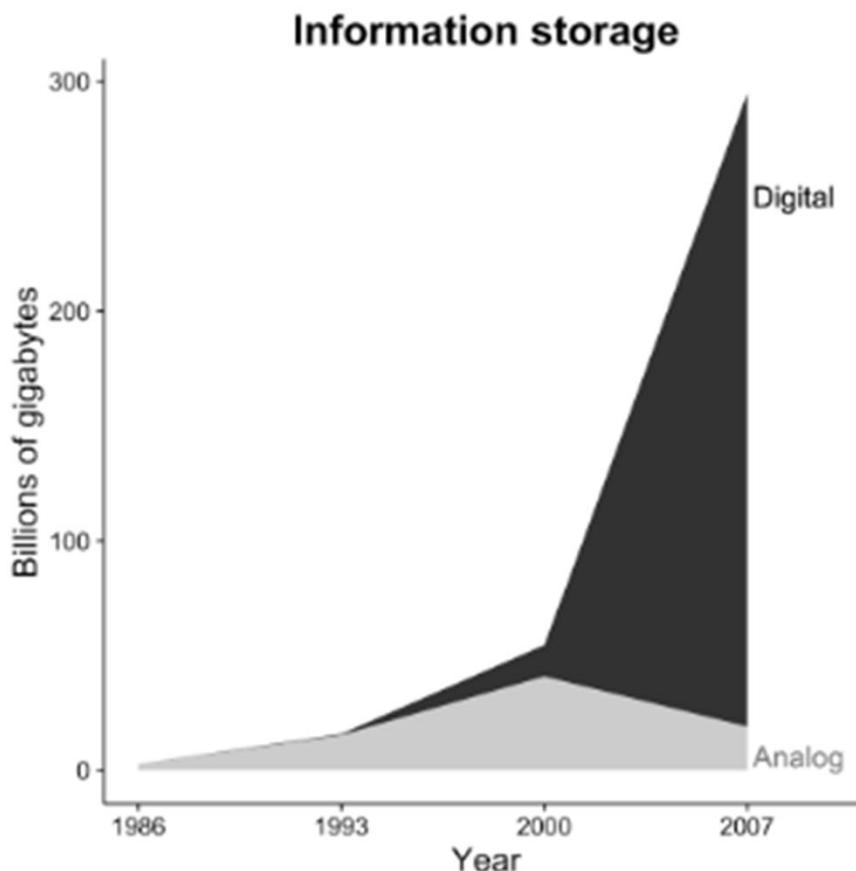
- ▶ Data has many properties including
 - ▶ Volume
 - ▶ Velocity
 - ▶ Variety
 - ▶ Veracity?

- ▶ CSCI-UA 2/3/101

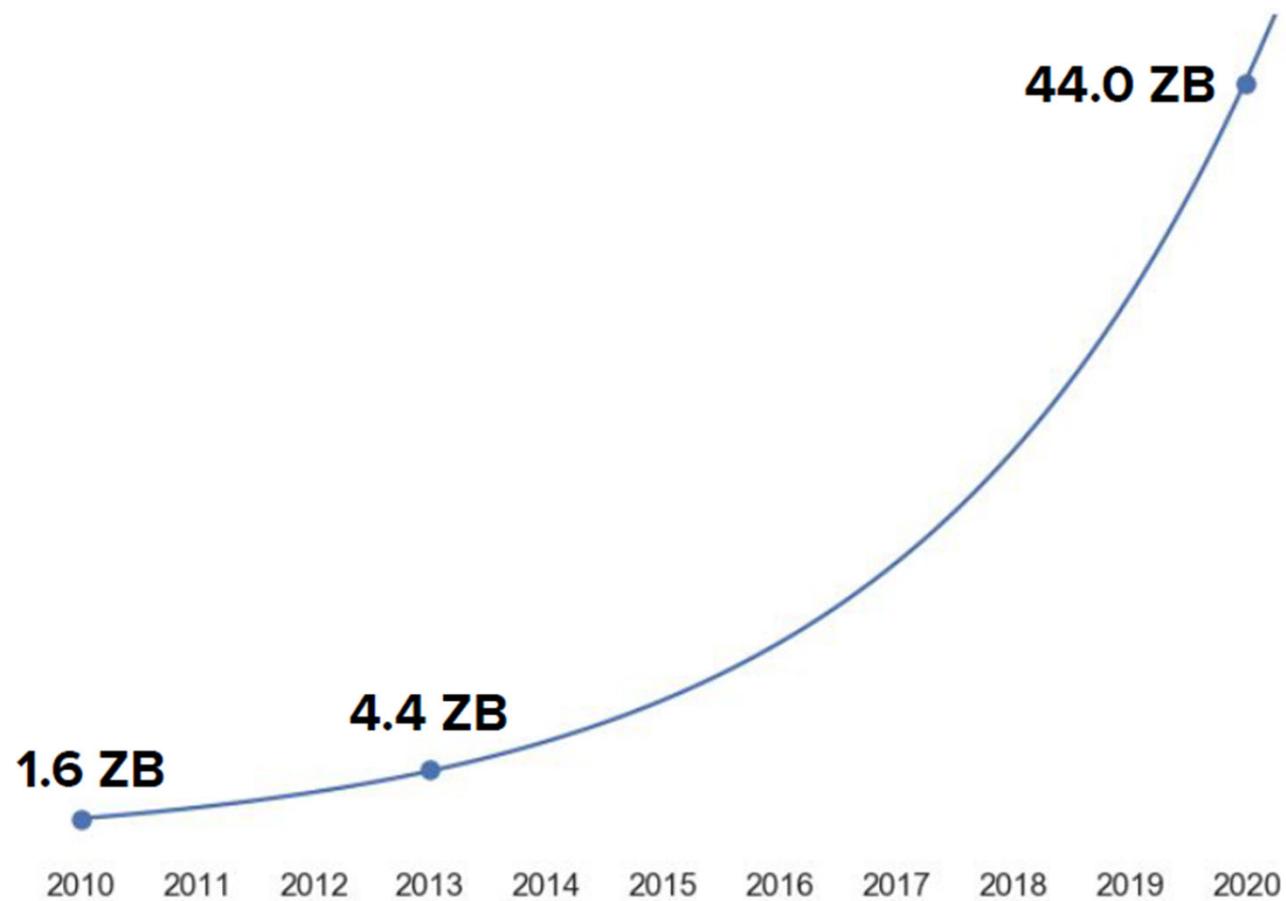
- ▶ Programming will be a tool to obtain information from data
- ▶ Python is one of many tools for data science.
- ▶ We will work backwards from problems to learn about the relevant tools



Exercise



Exercise



What characterizes data?

Inside the Fight to Save Alaska's 20 Native Languages from Dying Out

Alaskans are racing to collect their elders' knowledge of 20 native languages and design a pedagogy that breaches the generational divide.

By [Agnes Walton](#)

"[Our] corpus contains over 500 billion words, in English (361 billion), French (45 billion), Spanish (45 billion), German (37 billion), Chinese (13 billion), Russian (35 billion), and Hebrew (2 billion). The oldest works were published in the 1500s. The early decades are represented by only a few books per year, comprising several hundred thousand words. By 1800, the corpus grows to 98 million words per year; by 1900, 1.8 billion; and by 2000, 11 billion. The corpus cannot be read by a human. If you tried to read only English-language entries from the year 2000 alone, at the reasonable pace of 200 words/min, without interruptions for food or sleep, it would take 80 years. The sequence of letters is 1000 times longer than the human genome: If you wrote it out in a straight line, it would reach to the Moon and back 10 times over."

Volume

What characterizes data?



Volume

WIRED

NYC Now Knows More Than Ever About Your Uber and Lyft Trips

In 2007, New York City's Taxi and Limousine Commission, in a belated ... because along with those readers came GPS trackers that became a ... The ride-hail companies, though, remain wary of such data operations.

Jan 31, 2019



What characterizes data?



Memorandum

Date: January 24, 2020

To: THE NYU COMMUNITY

From: Carlo Ciotoli, MD, Associate Vice President for Student Health and Executive Director of the Student Health Center

Re: The Emergence of the Novel Coronavirus

Velocity

An AI Epidemiologist Sent the First Warnings of the Wuhan Virus

WIRED · 2 days ago



What characterizes data?

Reuters UK

Turkey escalates crackdown on dissent six years after Gezi protests

Turkey escalates crackdown on dissent six years after Gezi protests ... were inspired by the worldwide "Occupy" protests and Arab uprisings ...

Mar 18, 2019



Velocity

Participants	dataset in typical study		
Nonparticipants			ex-post panel in Budak and Watts (2015)
	Pre-Gezi (Jan 1, 2012 - May 28, 2013)	During Gezi (May 28, 2012 - Aug 1, 2013)	Post-Gezi (Aug 1, 2013 - Jan 1, 2014)

What characterizes data?

The cost of racial animus on a black candidate: Evidence using Google search data ★

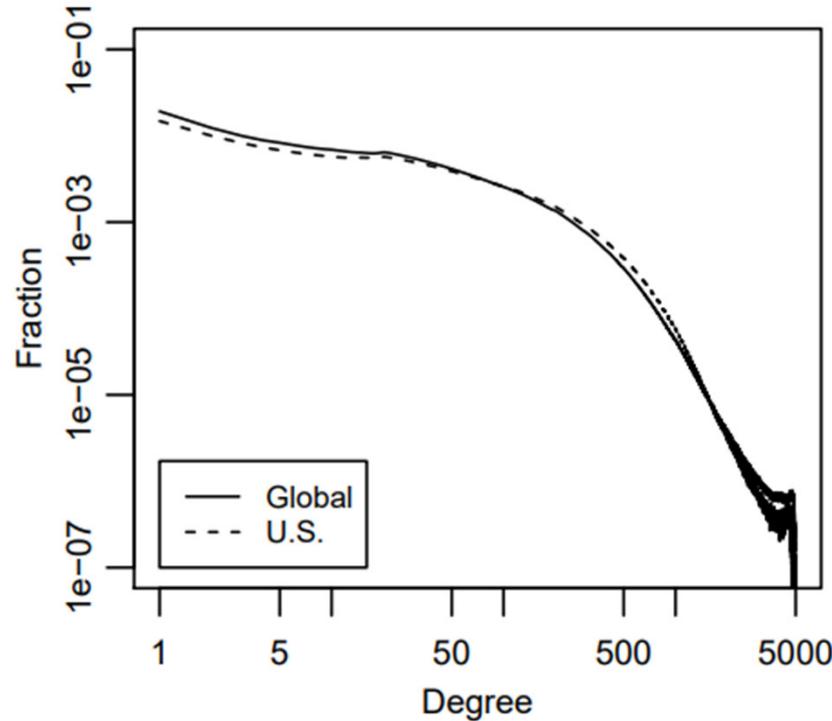
Seth Stephens-Davidowitz ↗

Highlights

- Google search data is used as a new measure of racial animus in the United States.
- In places where racial animus is highest, Obama did worse than comparable Democratic candidates.
- Racial animus cost Obama about 4 percentage points, giving his opponent the equivalent of a home state advantage.
- Analysis with Google searches suggest racism cost Obama substantially more votes than survey-based analysis.

Not Reactive

What characterizes data?



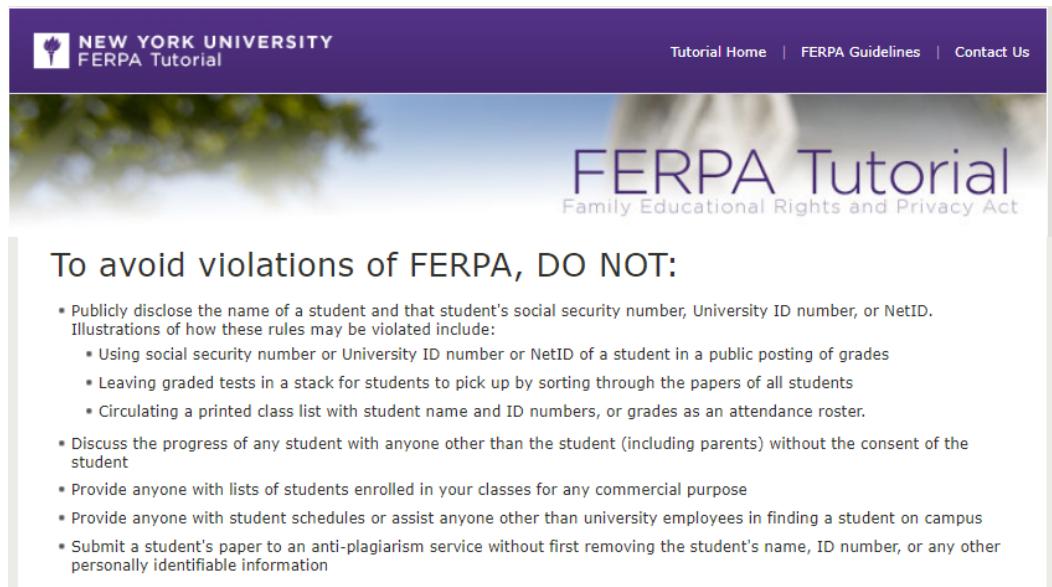
Not Reactive

for a small anomaly near 20 friends. This kink is due to forces within the Facebook product to encourage low friend count individuals in particular to gain more friends until they reach 20 friends. The distribution shows a clear cutoff at 5000 friends, a limit imposed by Facebook on the number of friends at the time

Questions

- ▶ Questions on Piazza?
- ▶ Question for You!

Should students have access to the class roster?



The screenshot shows the header of the NYU FERPA Tutorial website. The header features the NYU logo and the text "NEW YORK UNIVERSITY FERPA Tutorial". On the right side, there are links for "Tutorial Home", "FERPA Guidelines", and "Contact Us". Below the header is a large image of a tree. To the right of the tree, the text "FERPA Tutorial" and "Family Educational Rights and Privacy Act" is displayed. At the bottom left, there is a section titled "To avoid violations of FERPA, DO NOT:" followed by a list of prohibited actions.

To avoid violations of FERPA, DO NOT:

- Publicly disclose the name of a student and that student's social security number, University ID number, or NetID. Illustrations of how these rules may be violated include:
 - Using social security number or University ID number or NetID of a student in a public posting of grades
 - Leaving graded tests in a stack for students to pick up by sorting through the papers of all students
 - Circulating a printed class list with student name and ID numbers, or grades as an attendance roster.
- Discuss the progress of any student with anyone other than the student (including parents) without the consent of the student
- Provide anyone with lists of students enrolled in your classes for any commercial purpose
- Provide anyone with student schedules or assist anyone other than university employees in finding a student on campus
- Submit a student's paper to an anti-plagiarism service without first removing the student's name, ID number, or any other personally identifiable information

