Data Science in the Wild

Lecture 1: Introduction

Eran Toch

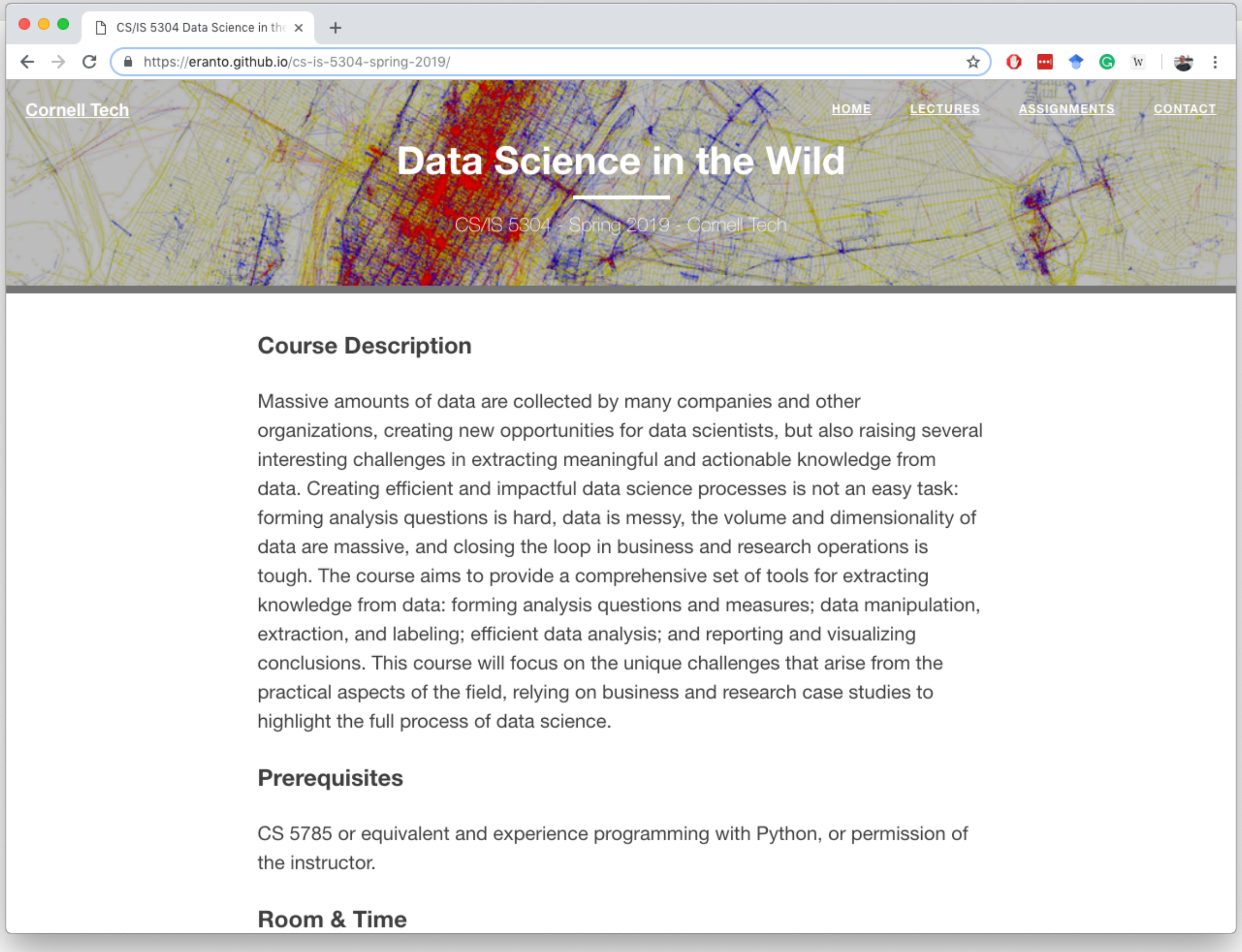
CORNELL TECH

Agenda

1. About the Course
2. The Data Explosion
3. Data Science Capabilities
4. The scientific method

Resources

- Website: https://eranto.github.io/cs5304-spring2019/
- Slack: wild-data-science.slack.com


CS/IS 5304 Data Science in th
https://eranto.github.io/cs-is-5304-spring-2019/
Cornell Tech
HOME
LECTURES
ASSIGNMENTS
CONTACT

Data Science in the Wild

CS/IS 5304 - Spring 2019 - Cornell Tech

Course Description

Massive amounts of data are collected by many companies and other organizations, creating new opportunities for data scientists, but also raising several interesting challenges in extracting meaningful and actionable knowledge from data. Creating efficient and impactful data science processes is not an easy task: forming analysis questions is hard, data is messy, the volume and dimensionality of data are massive, and closing the loop in business and research operations is tough. The course aims to provide a comprehensive set of tools for extracting knowledge from data: forming analysis questions and measures; data manipulation, extraction, and labeling; efficient data analysis; and reporting and visualizing conclusions. This course will focus on the unique challenges that arise from the practical aspects of the field, relying on business and research case studies to highlight the full process of data science.

Prerequisites

CS 5785 or equivalent and experience programming with Python, or permission of the instructor.

Room & Time

Prof. Eran Toch

Visiting associate Professor at Cornell Tech

Faculty, Tel Aviv University

etoch@cornell.edu

Twitter: @erant

http://toch.tau.ac.il

Mr. David Rimshnick

- Cornell OR alum, BS 2005, MEng 2006
  - Research on logistics problems (airline crew scheduling, vehicle routing)
- Spent career in data science and analytics in healthcare industry
  - ZS Associates
  - Novo Nordisk (biopharma company)
  - Pfizer (biopharma company)
- Currently Principal at Boston Consulting Group
  - Part of BCG Gamma, sub-organization devoted to advanced AI and ML applications

david.rimshnick@gmail.com

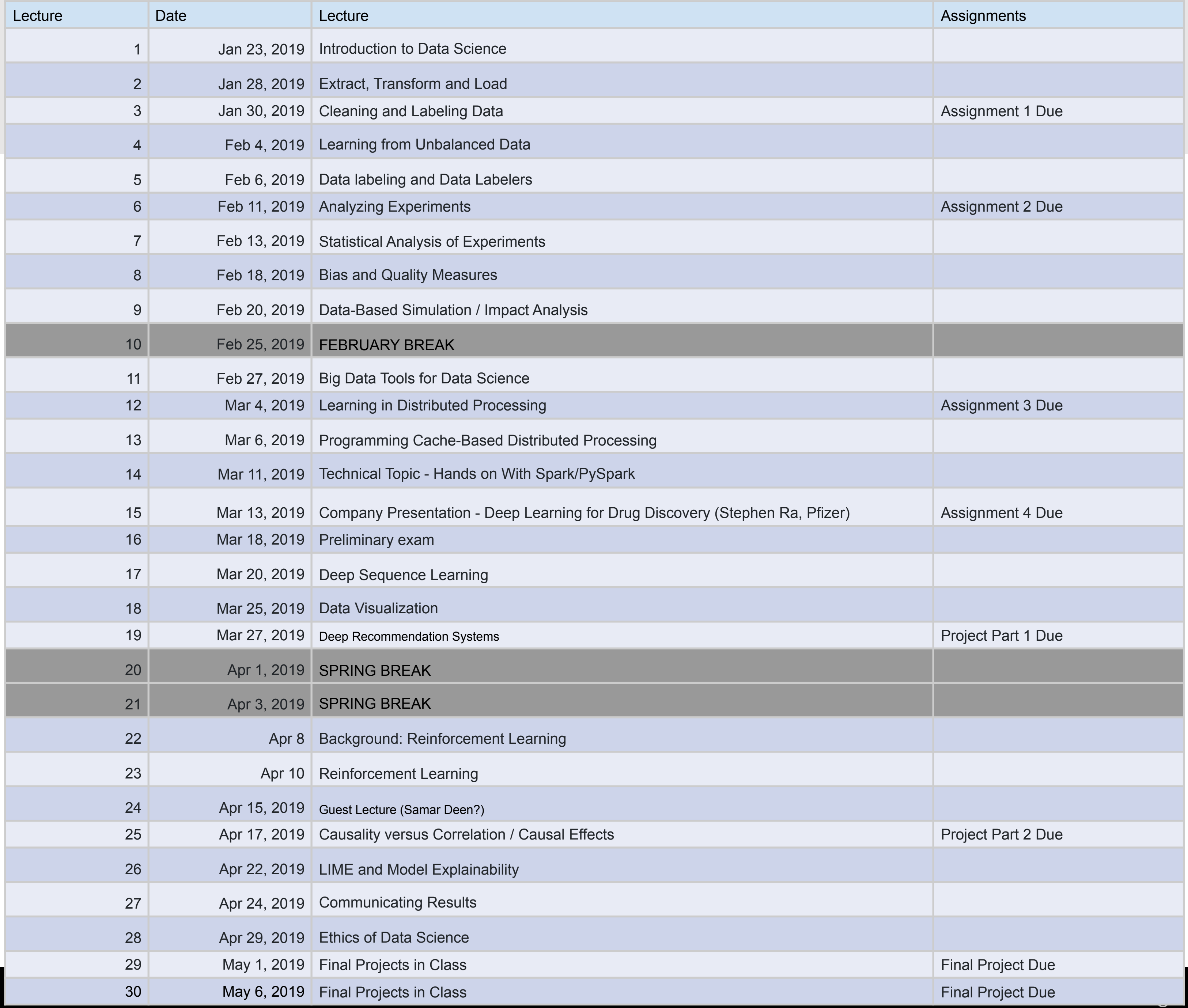
Team

- TA:
    - Zekun Hao

- Graders:
    - Summer Shi
    - Seye Bankole
    - Svava Kristinsdottir
    - Mohit Chawla

Timetable

Please let us know about absence days due to religious holidays



Lecture	Date	Lecture	Assignments
1	Jan 23, 2019	Introduction to Data Science	
2	Jan 28, 2019	Extract, Transform and Load	
3	Jan 30, 2019	Cleaning and Labeling Data	Assignment 1 Due
4	Feb 4, 2019	Learning from Unbalanced Data	
5	Feb 6, 2019	Data labeling and Data Labelers	
6	Feb 11, 2019	Analyzing Experiments	Assignment 2 Due
7	Feb 13, 2019	Statistical Analysis of Experiments	
8	Feb 18, 2019	Bias and Quality Measures	
9	Feb 20, 2019	Data-Based Simulation / Impact Analysis	
10	Feb 25, 2019	FEBRUARY BREAK	
11	Feb 27, 2019	Big Data Tools for Data Science	
12	Mar 4, 2019	Learning in Distributed Processing	Assignment 3 Due
13	Mar 6, 2019	Programming Cache-Based Distributed Processing	
14	Mar 11, 2019	Technical Topic - Hands on With Spark/PySpark	
15	Mar 13, 2019	Company Presentation - Deep Learning for Drug Discovery (Stephen Ra, Pfizer)	Assignment 4 Due
16	Mar 18, 2019	Preliminary exam	
17	Mar 20, 2019	Deep Sequence Learning	
18	Mar 25, 2019	Data Visualization	
19	Mar 27, 2019	Deep Recommendation Systems	Project Part 1 Due
20	Apr 1, 2019	SPRING BREAK	
21	Apr 3, 2019	SPRING BREAK	
22	Apr 8	Background: Reinforcement Learning	
23	Apr 10	Reinforcement Learning	
24	Apr 15, 2019	Guest Lecture (Samar Deen?)	
25	Apr 17, 2019	Causality versus Correlation / Causal Effects	Project Part 2 Due
26	Apr 22, 2019	LIME and Model Explainability	
27	Apr 24, 2019	Communicating Results	
28	Apr 29, 2019	Ethics of Data Science	
29	May 1, 2019	Final Projects in Class	Final Project Due
30	May 6, 2019	Final Projects in Class	Final Project Due

Grade Breakdown

- Home assignments (30%)
- Final project (30%)
- Preliminary exam (20%) - in class
- Final exam (20%) - take home

Assignments

- 4 home assignments
  - Each with programming and a written exercise
  - Each students has a total of one slip day
- The officially supported programming language is Python
  - But you are welcome to work on your assignments using other languages
- You can use well-known libraries but cite them.
- You are encouraged to work in groups of 2 students.

Bibliography

The books are not required for the course, but they can be of interest to students.

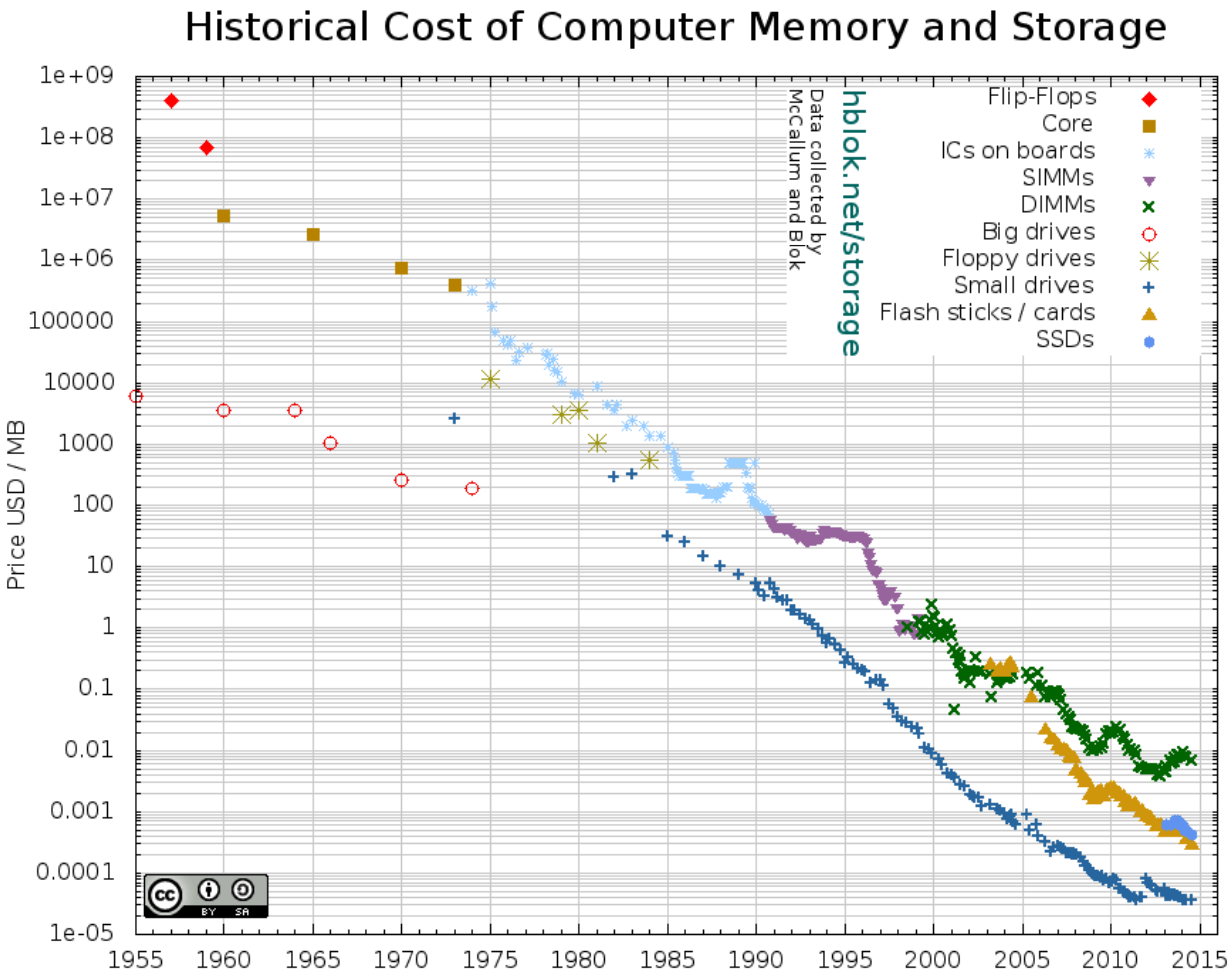
1. Foster Provost and Tom Fawcett, Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking, O'Reilly Media; 1st edition (2013)

2. Jake VanderPlas, Python Data Science Handbook, O'Reilly Media; 1 edition (2016) - Free book

3. Russell Jurney, Agile Data Science 2.0: Building Full-Stack Data Analytics Applications with Spark, O'Reilly Media; 1st edition (2017).

4. A. Rajaraman, J. Leskovec and J. Ullman, Mining of Massive Datasets, Cambridge University Press, 3rd version

Data Storage Prices



3.75 Megabyte

1 Terrabyte

How do we make decisions?

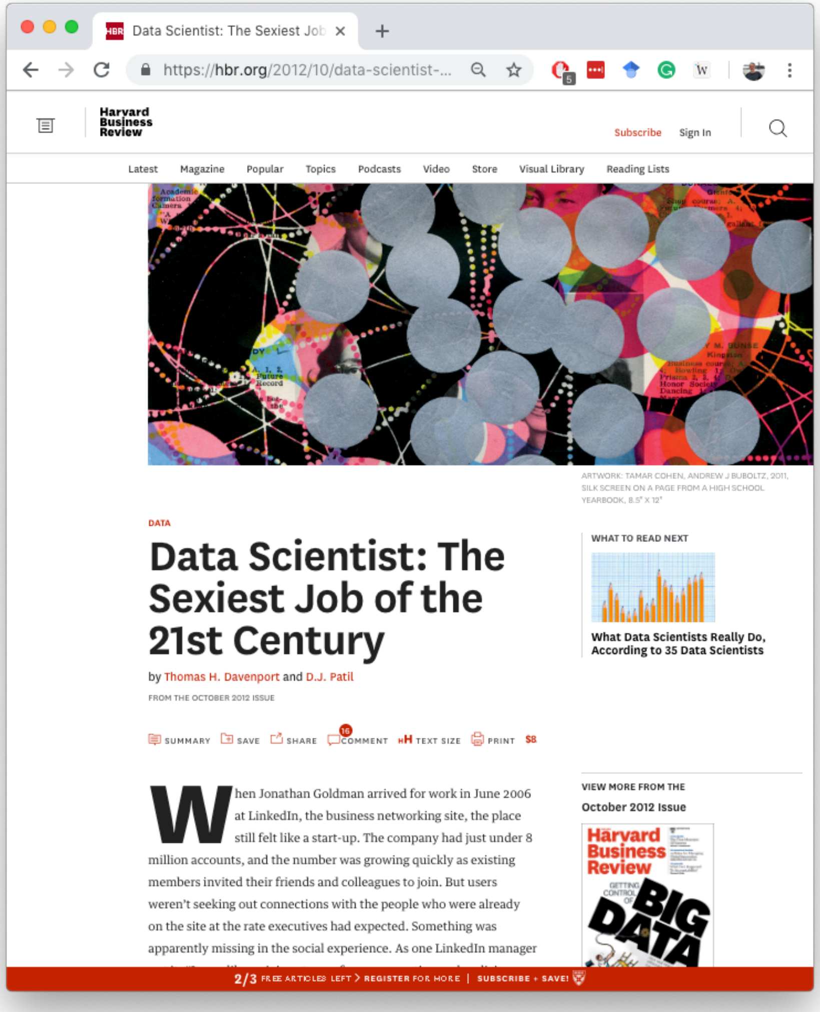
According to HiPPO

(highest paid person's opinion)

According to data

(Go see Moneyball)

Data Science as a Profession



ARTWORK: TAMAR COHEN, ANDREW J BUBOLTZ, 2011, SILK SCREEN ON A PAGE FROM A HIGH SCHOOL YEARBOOK, 8.5" X 12"

DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

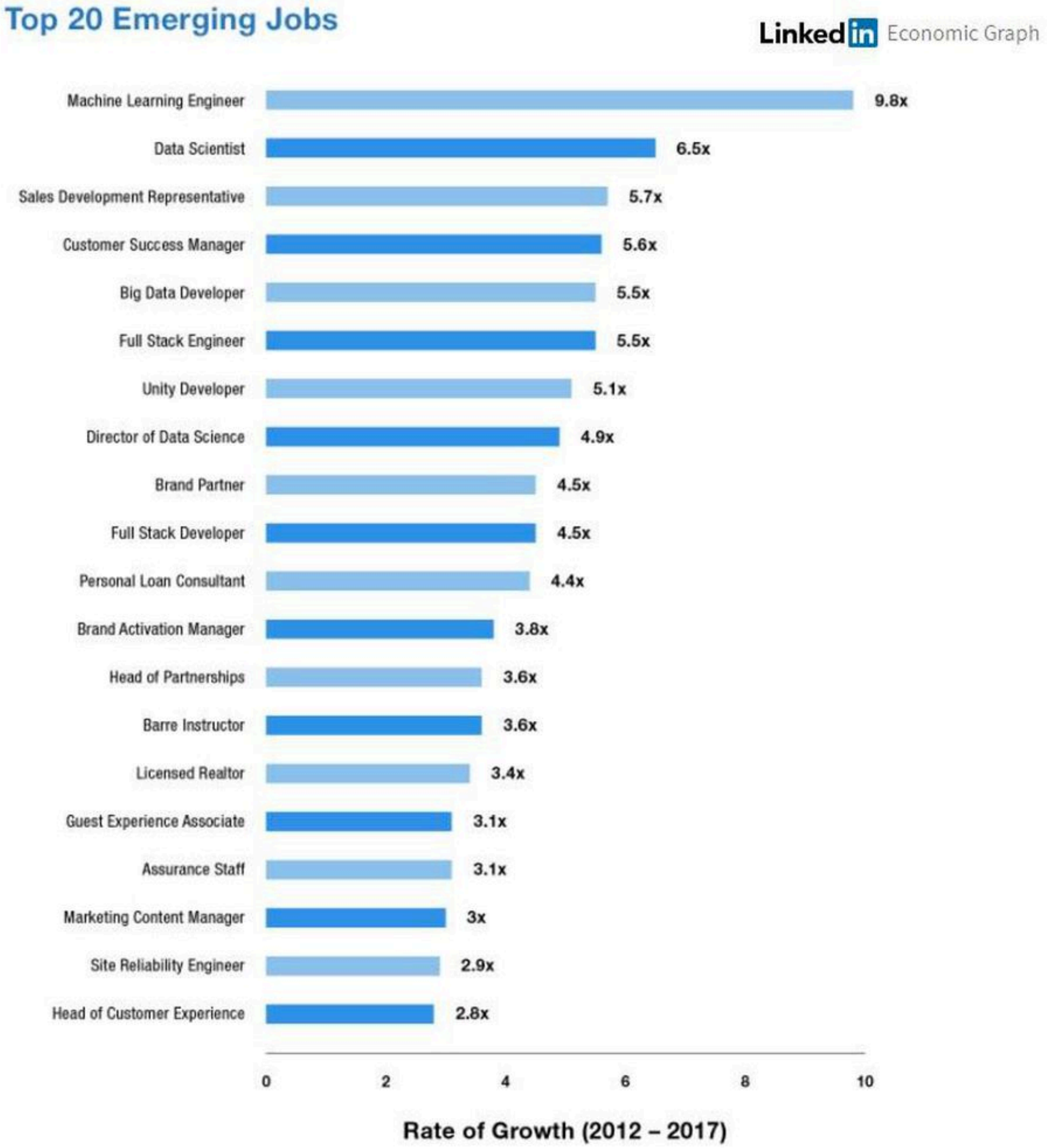
When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager

WHAT TO READ NEXT

What Data Scientists Really Do, According to 35 Data Scientists

VIEW MORE FROM THE

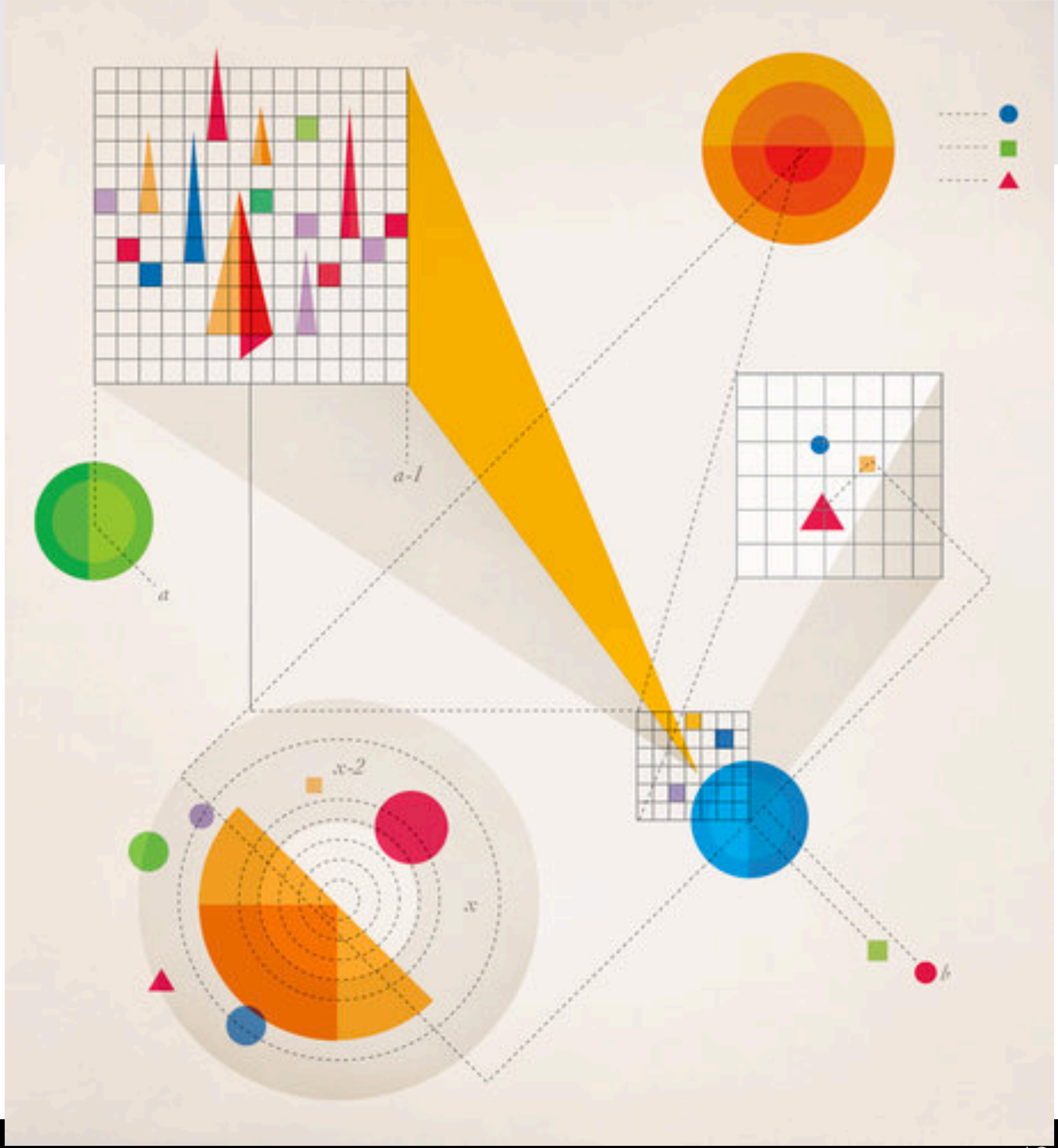
October 2012 Issue



Data-Literate

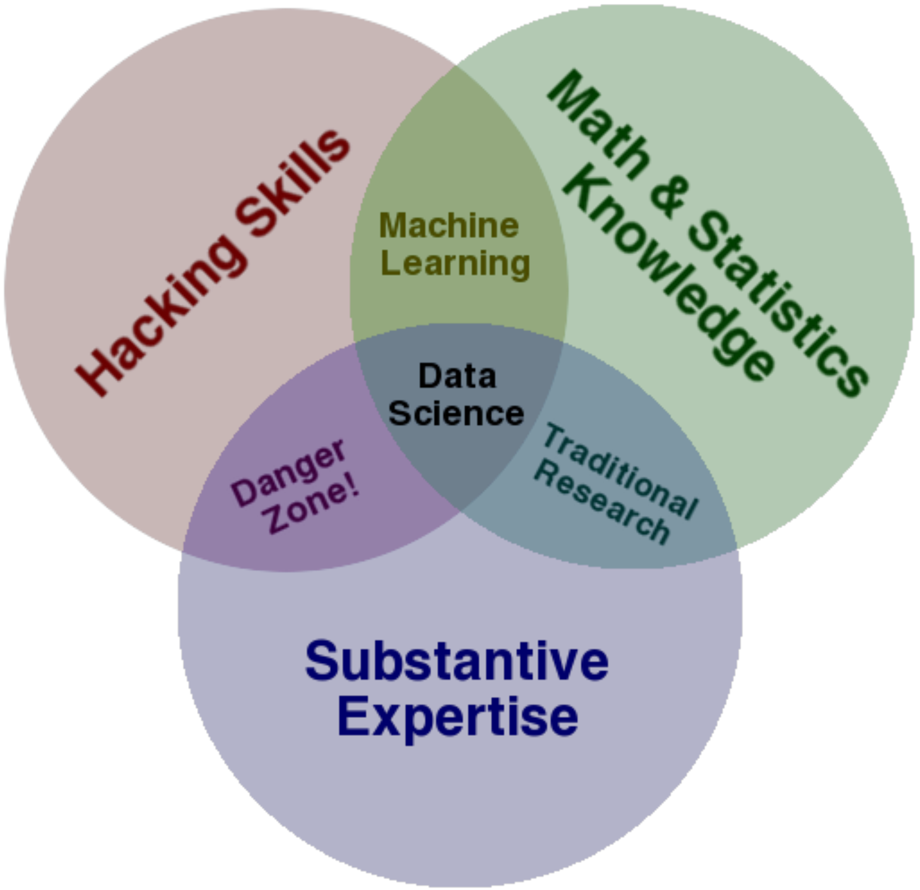
McKinsey Global Institute projected that the United States needs 140,000 to 190,000 more workers with “deep analytical” expertise and 1.5 million more data-literate managers, whether retrained or hired.

http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html



What is Data Science?

Data science is a professional approach to apply data engineering, statistics, and machine learning to solve problems in a scientific way

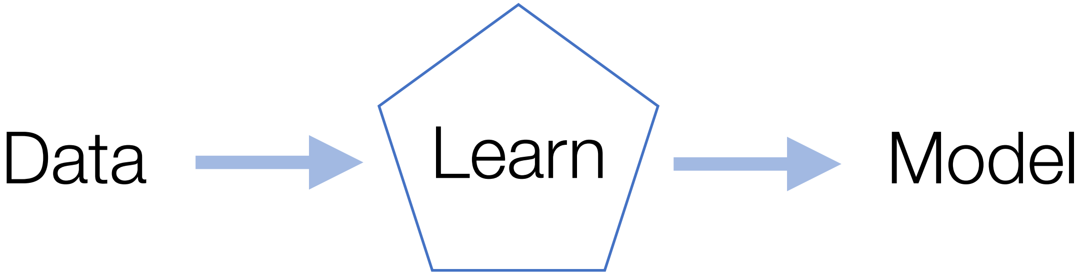


http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

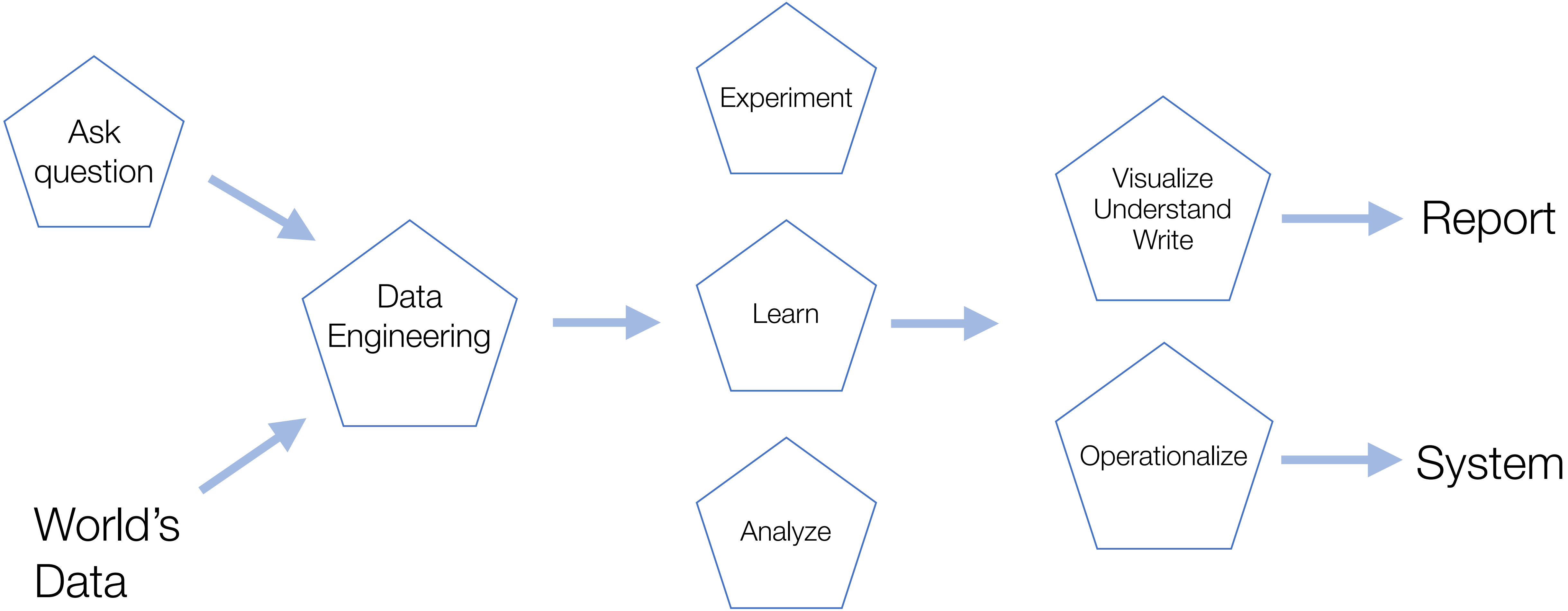
Buzz word hell

- Data science is a heavily criticized concept
- It is hard to distinguish it from science
- And from any type of data-intensive transaction

The Machine Learning Model



The Data Science Model



Ask question

World's Data

Data Engineering

Experiment

Learn

Analyze

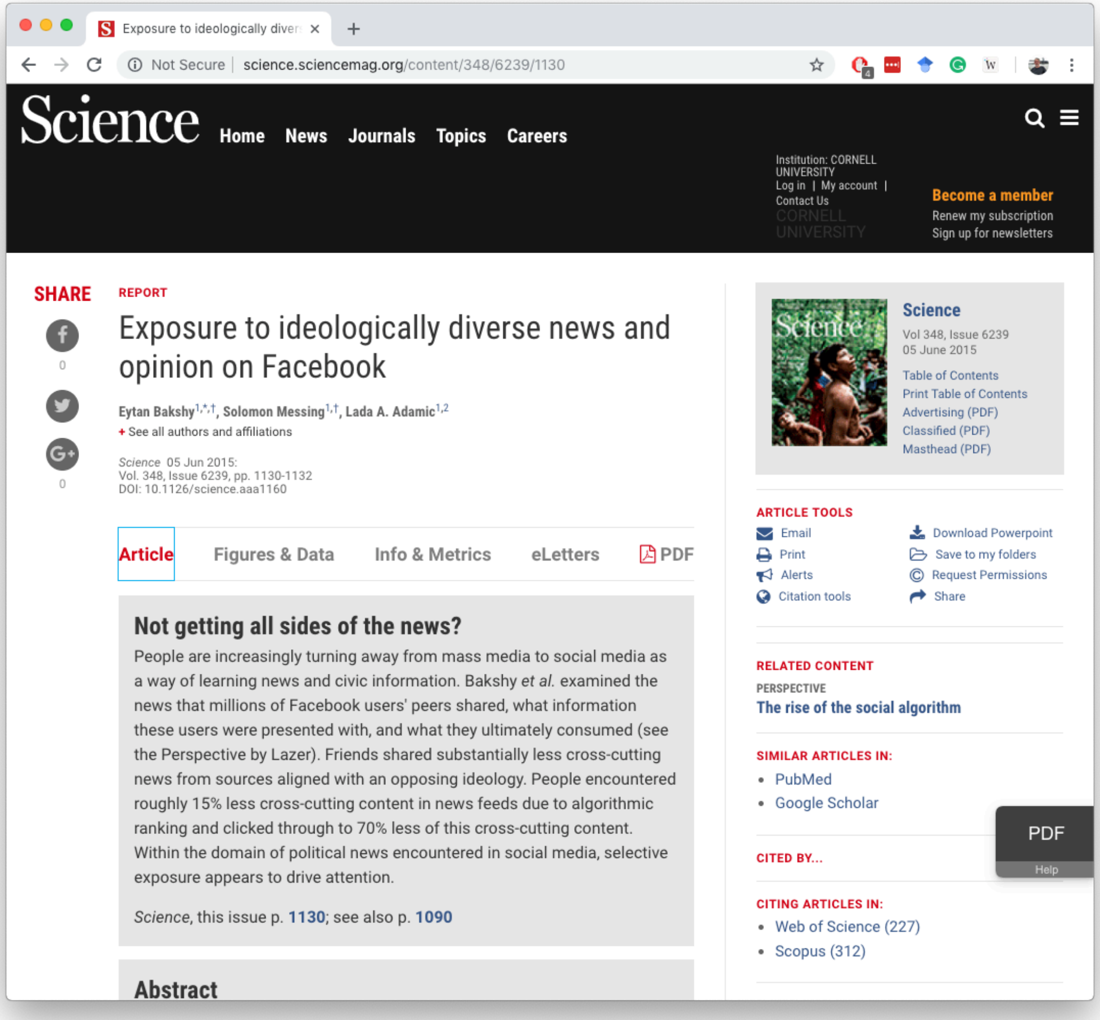
Visualize
Understand
Write

Operationalize

Report

System

Science



Exposure to ideologically diver ×
Not Secure | science.sciencemag.org/content/348/6239/1130
Science
Home News Journals Topics Careers
Institution: CORNELL UNIVERSITY
Log in | My account | Contact Us
CORNELL UNIVERSITY
Become a member
Renew my subscription
Sign up for newsletters
SHARE
0
0

REPORT

Exposure to ideologically diverse news and opinion on Facebook

Eytan Bakshy[1,*,†], Solomon Messing[1,†], Lada A. Adamic[1,2]

+ See all authors and affiliations

Science 05 Jun 2015:
Vol. 348, Issue 6239, pp. 1130-1132
DOI: 10.1126/science.aaa1160

Article Figures & Data Info & Metrics eLetters PDF

Not getting all sides of the news?

People are increasingly turning away from mass media to social media as a way of learning news and civic information. Bakshy *et al.* examined the news that millions of Facebook users' peers shared, what information these users were presented with, and what they ultimately consumed (see the Perspective by Lazer). Friends shared substantially less cross-cutting news from sources aligned with an opposing ideology. People encountered roughly 15% less cross-cutting content in news feeds due to algorithmic ranking and clicked through to 70% less of this cross-cutting content. Within the domain of political news encountered in social media, selective exposure appears to drive attention.

Science, this issue p. 1130; see also p. 1090

Abstract

Science
Vol 348, Issue 6239
05 June 2015

RELATED CONTENT

PERSPECTIVE

The rise of the social algorithm

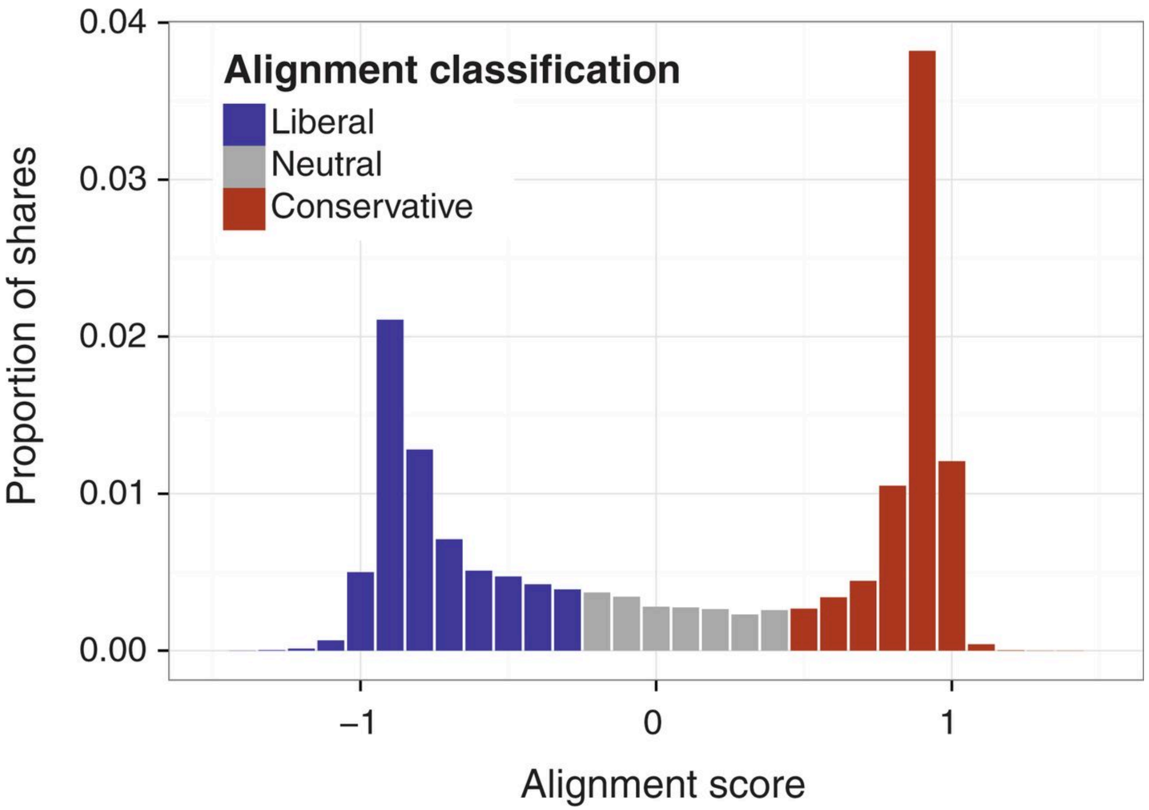
SIMILAR ARTICLES IN:

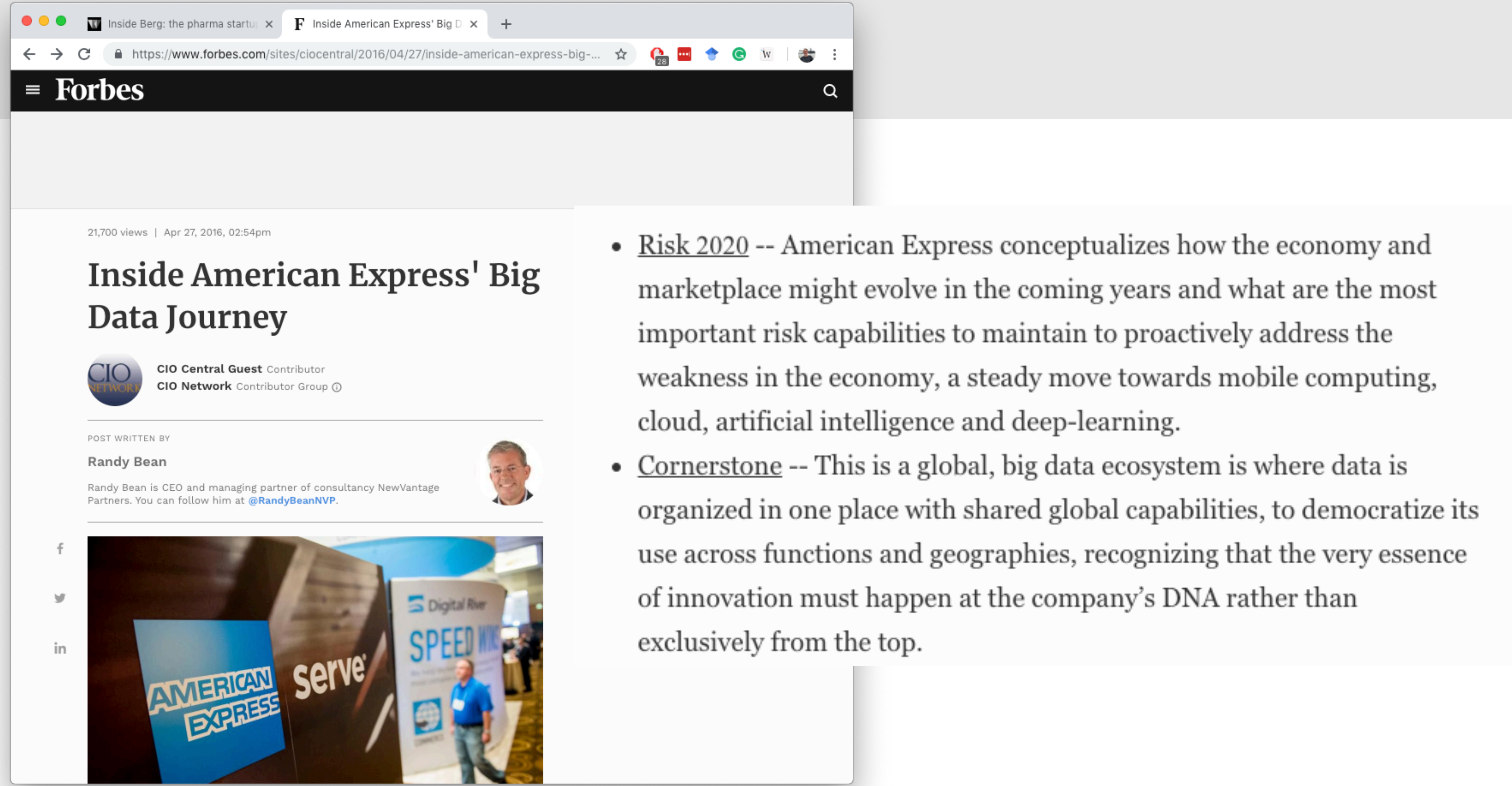
- PubMed
- Google Scholar

CITED BY...

CITING ARTICLES IN:

- Web of Science (227)
- Scopus (312)





Inside Berg: the pharma startu
Inside American Express' Big D
https://www.forbes.com/sites/ciocentral/2016/04/27/inside-american-express-big-...
Forbes

21,700 views | Apr 27, 2016, 02:54pm

Inside American Express' Big Data Journey

CIO Central Guest Contributor
CIO Network Contributor Group

POST WRITTEN BY

Randy Bean

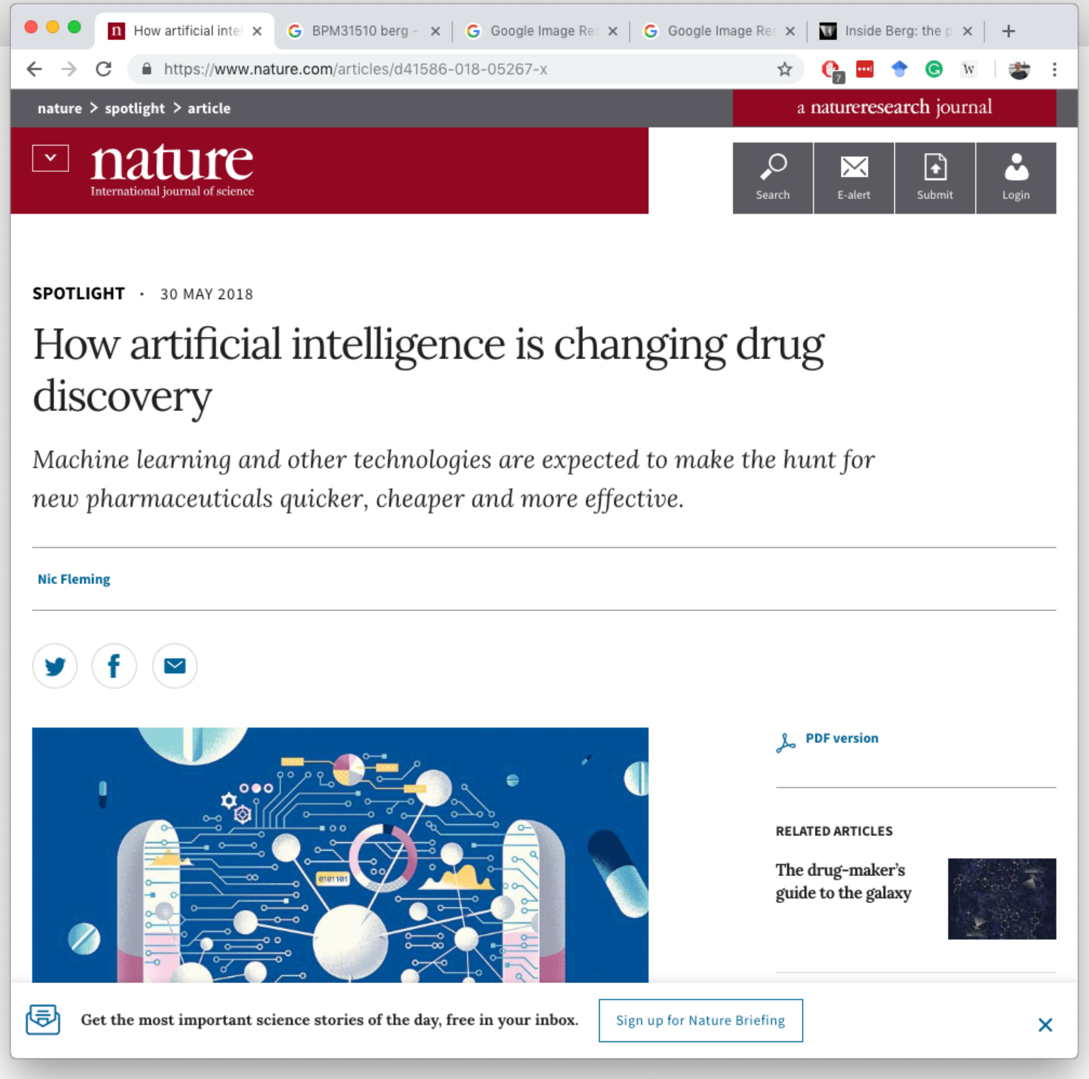
Randy Bean is CEO and managing partner of consultancy NewVantage Partners. You can follow him at @RandyBeanNVP.

AMERICAN EXPRESS
serve
Digital River
SPEED

- Risk 2020 -- American Express conceptualizes how the economy and marketplace might evolve in the coming years and what are the most important risk capabilities to maintain to proactively address the weakness in the economy, a steady move towards mobile computing, cloud, artificial intelligence and deep-learning.
- Cornerstone -- This is a global, big data ecosystem is where data is organized in one place with shared global capabilities, to democratize its use across functions and geographies, recognizing that the very essence of innovation must happen at the company's DNA rather than exclusively from the top.

Data Pharmaceuticals

For example, researchers at biotechnology company Berg, near Boston, Massachusetts, have developed a model to identify previously unknown cancer mechanisms using tests on more than 1,000 cancerous and healthy human cell samples. They modelled diseased human cells by varying the levels of sugar and oxygen the cells were exposed to, and then tracked their lipid, metabolite, enzyme and protein profiles. The group uses its AI platform to generate and analyse immense amounts of biological and outcomes data from patients to highlight key differences between diseased and healthy cells.



SPOTLIGHT · 30 MAY 2018

How artificial intelligence is changing drug discovery

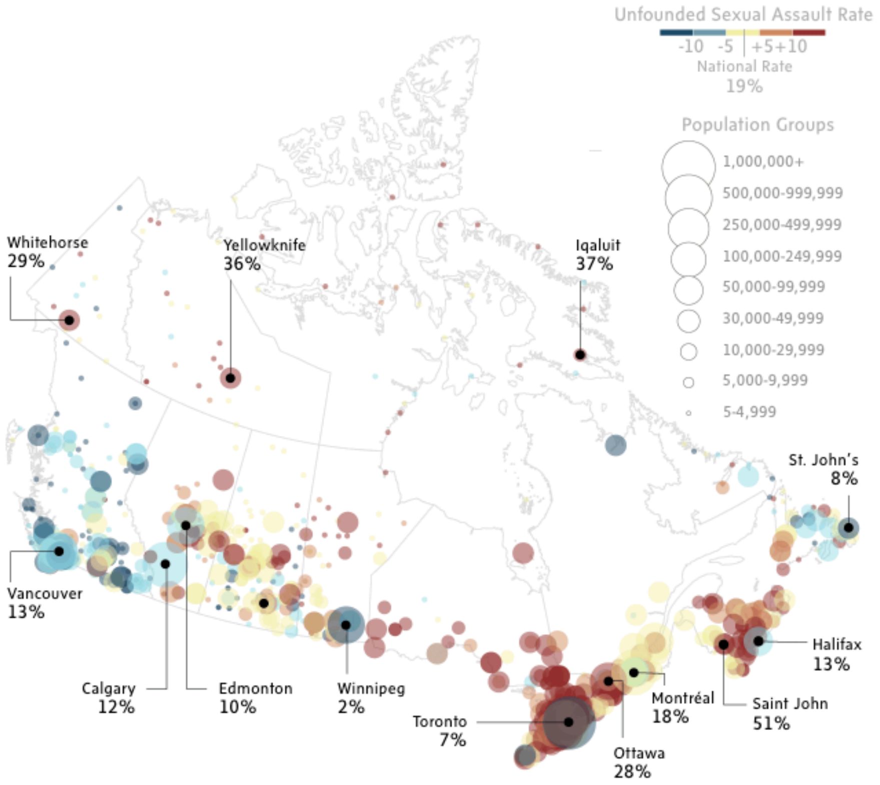
Machine learning and other technologies are expected to make the hunt for new pharmaceuticals quicker, cheaper and more effective.

Nic Fleming

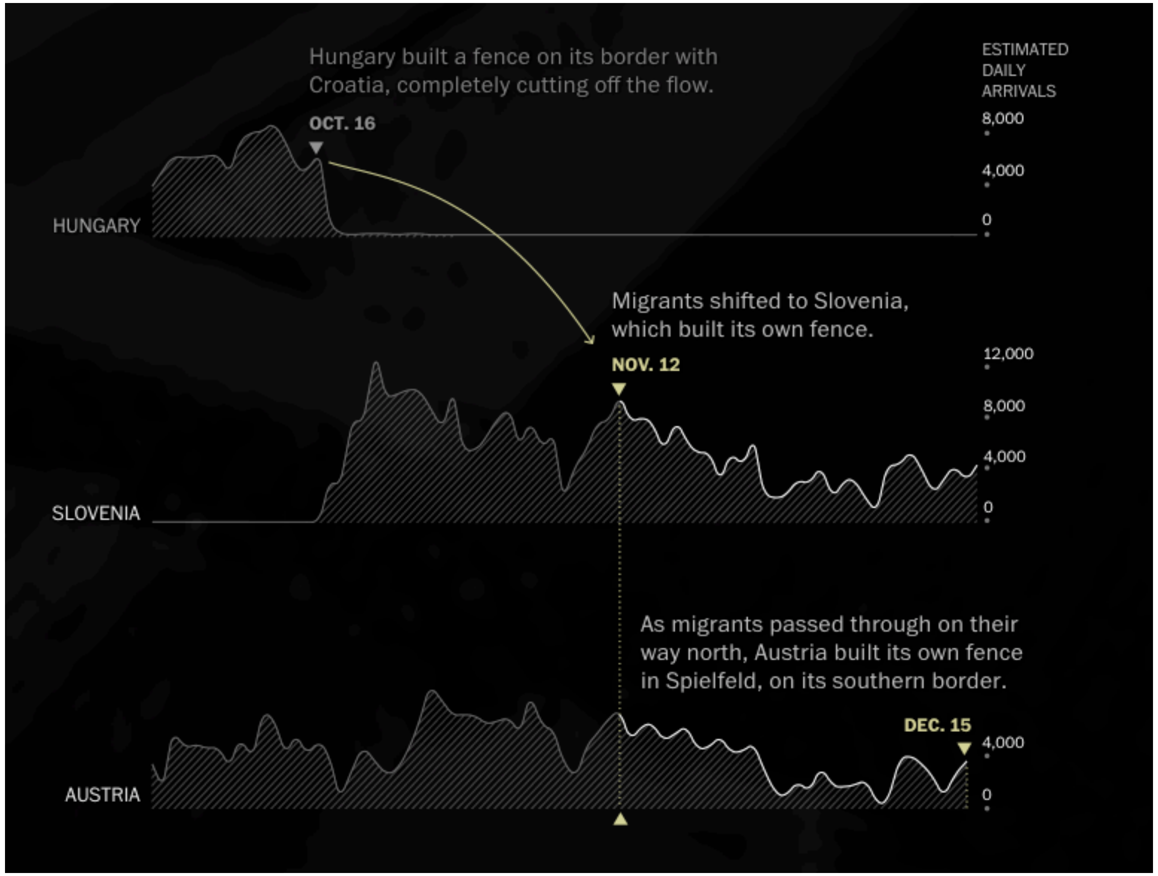
The drug-maker's guide to the galaxy

Get the most important science stories of the day, free in your inbox.

Journalism

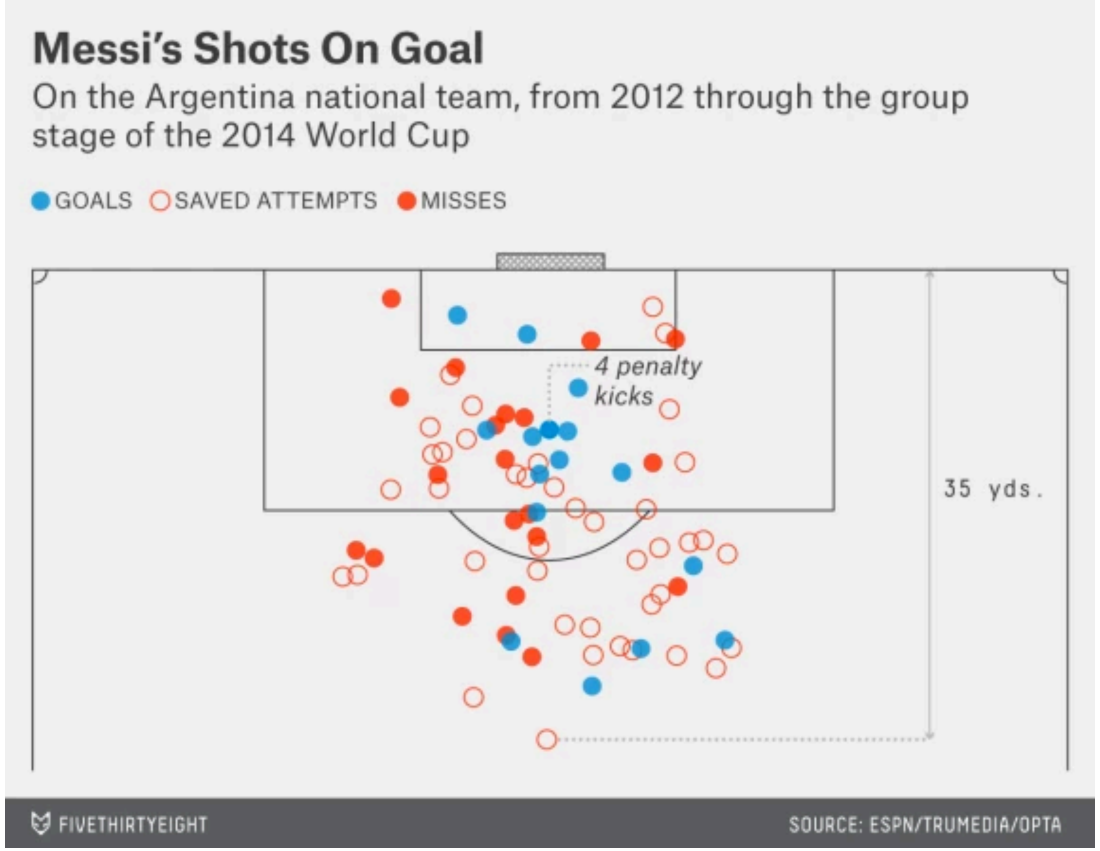


https://beta.theglobeandmail.com/news/investigations/unfounded-sexual-assault-canada-main/article33891309/

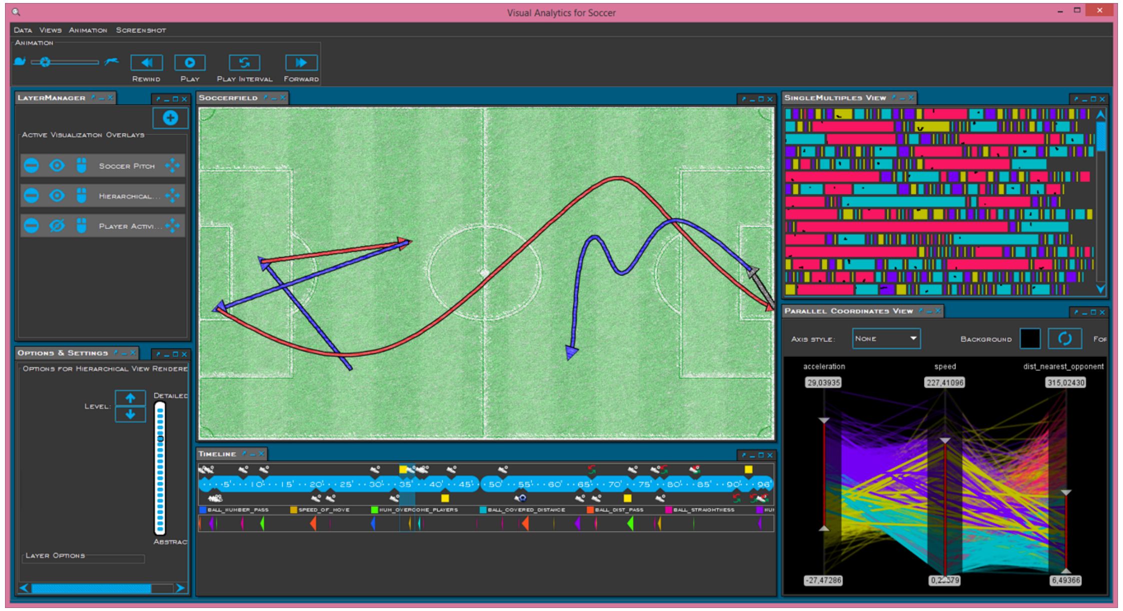


https://www.washingtonpost.com/graphics/world/border-barriers/europe-refugee-crisis-border-control/?noredirect=on

Sports

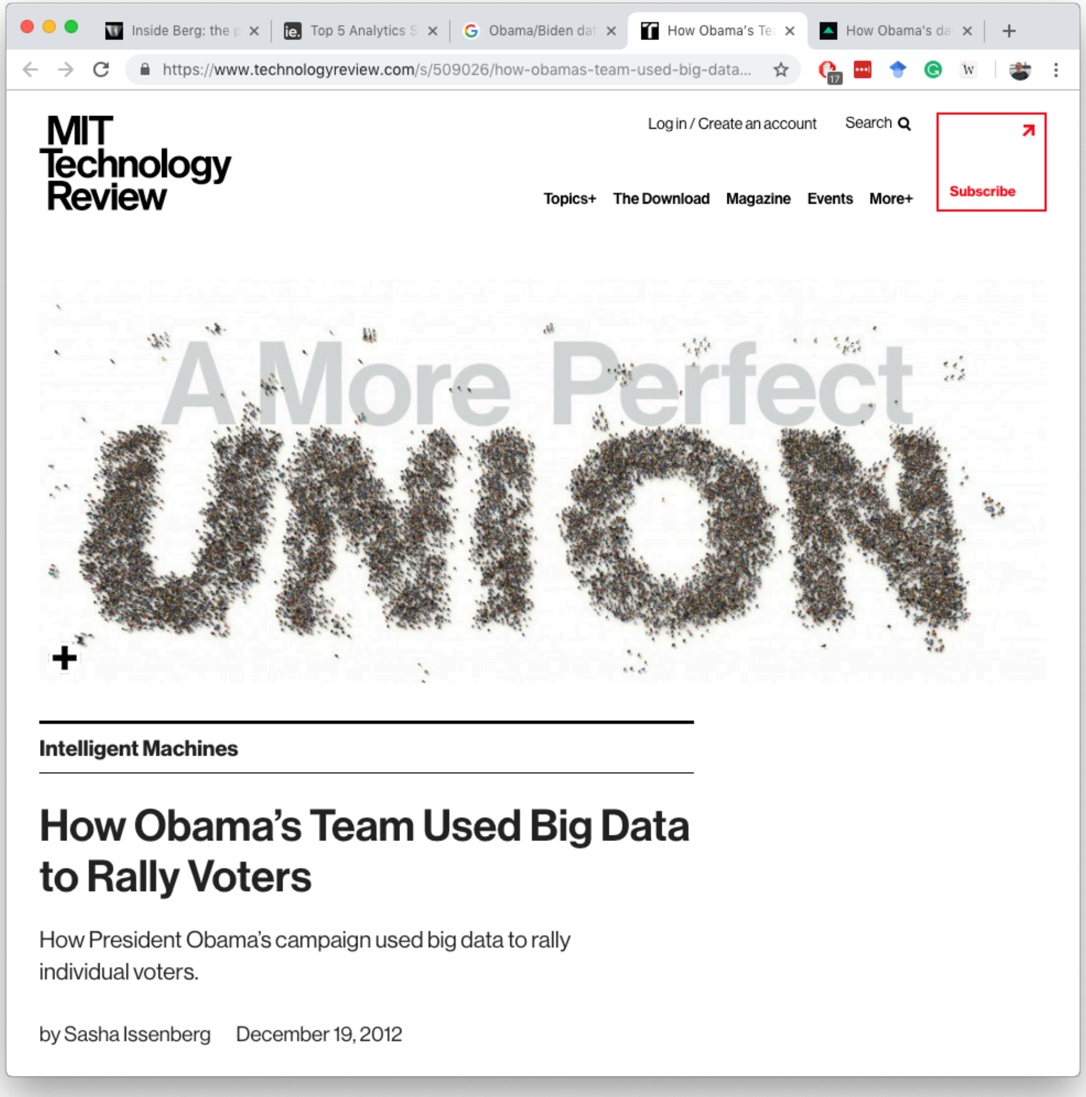


https://fivethirtyeight.com/features/lionel-messi-is-impossible/



https://www.janetzko.eu/project/soccer/

Politics

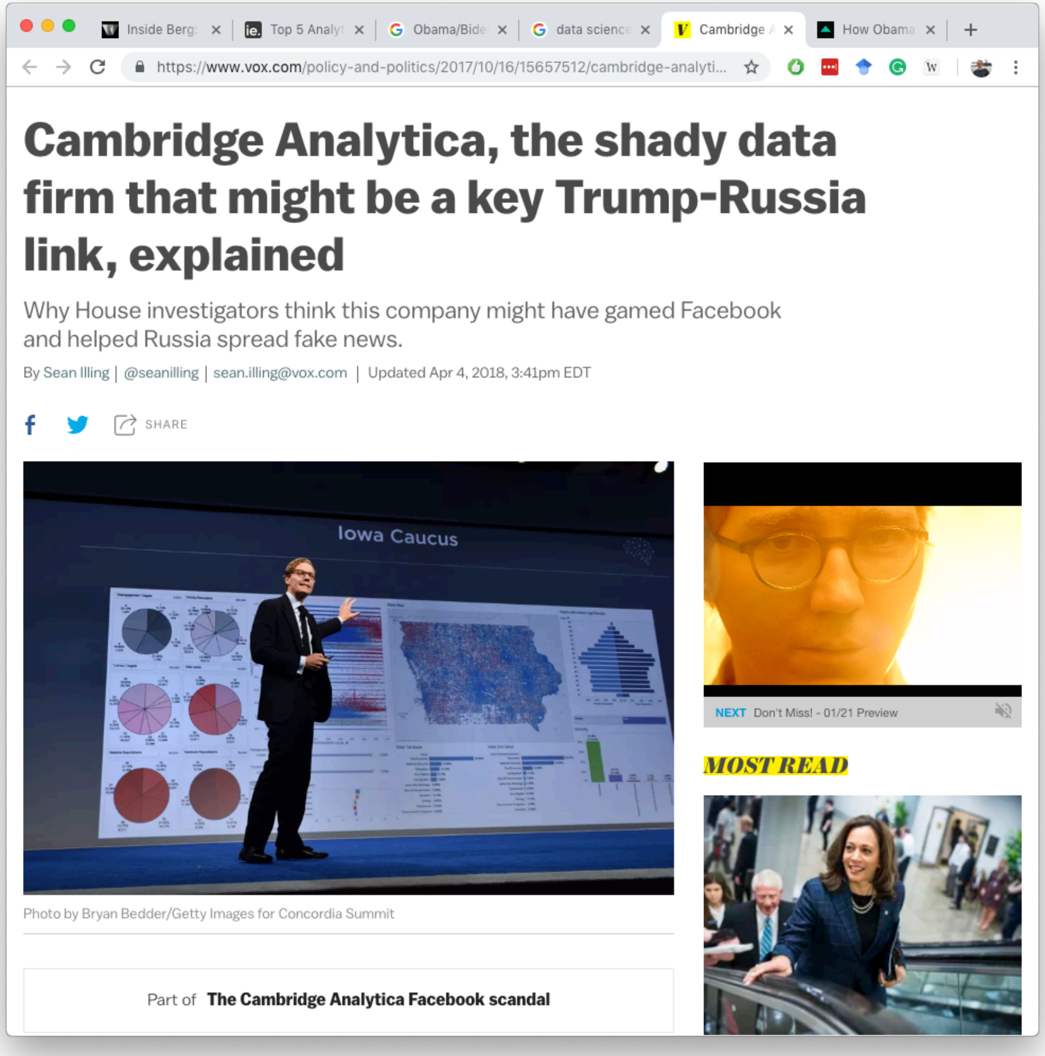


Intelligent Machines

How Obama's Team Used Big Data to Rally Voters

How President Obama's campaign used big data to rally individual voters.

by Sasha Issenberg December 19, 2012


Inside Berg
Top 5 Analy
Obama/Bide
data science
Cambridge
How Obama
https://www.vox.com/policy-and-politics/2017/10/16/15657512/cambridge-analyti...
SHARE
Iowa Caucus
NEXT Don't Miss! - 01/21 Preview
MOST READ

Cambridge Analytica, the shady data firm that might be a key Trump-Russia link, explained

Why House investigators think this company might have gamed Facebook and helped Russia spread fake news.

By Sean Illing | @seanilling | sean.illing@vox.com | Updated Apr 4, 2018, 3:41pm EDT

Photo by Bryan Bedder/Getty Images for Concordia Summit

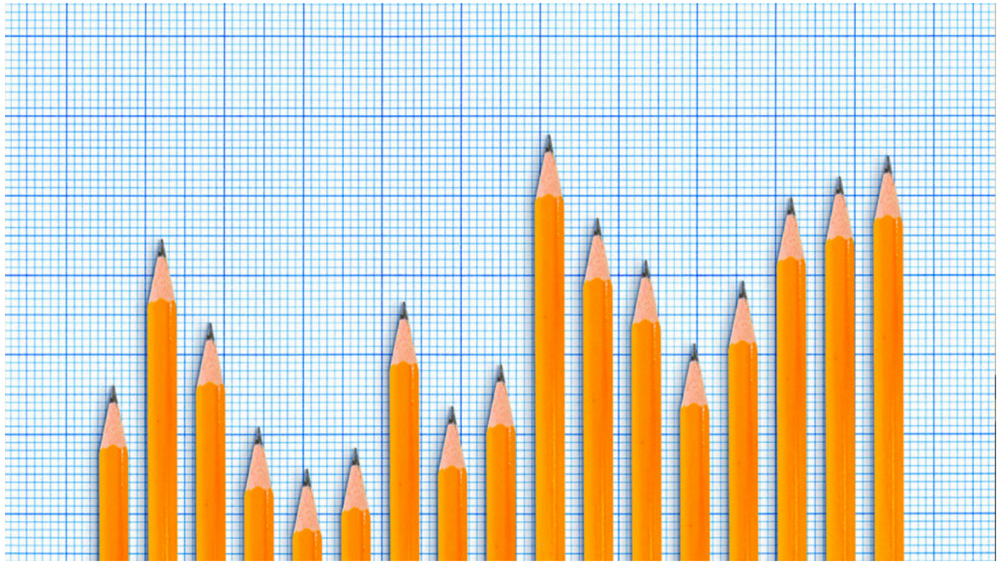
Part of **The Cambridge Analytica Facebook scandal**

Summary

- Data science overwhelms science, business, and civics
- The main challenges are not technical:
  - Asking good research questions
  - Applying the right tools
  - Creating data pipelines
  - Telling a story

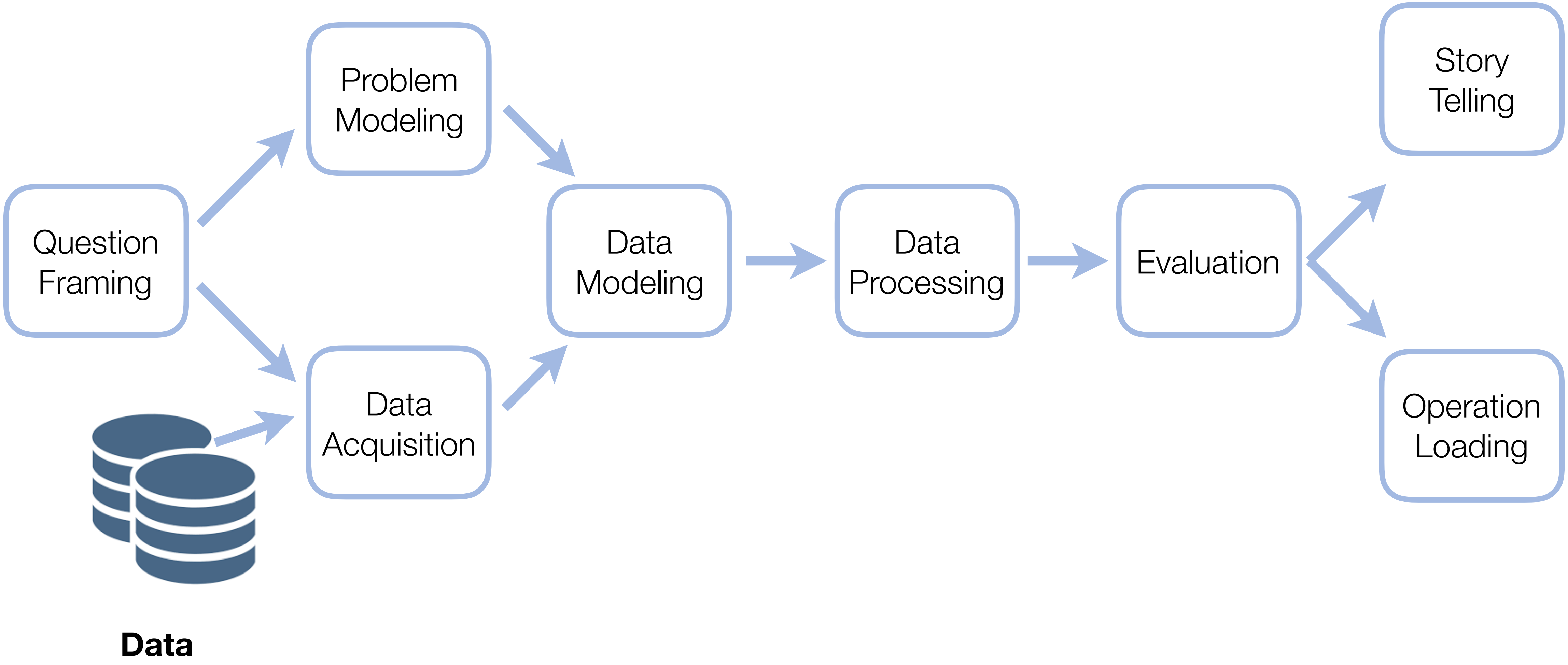
The Data Science Capabilities

1. Understand the data science process
2. Model problems and answer them with <u>real</u> data
3. Control the standard “toolbox” of data science methods
4. Analyze the quality of data science results
5. Know how to report, visualize, and discuss findings
6. Introduced to the societal challenges of data science



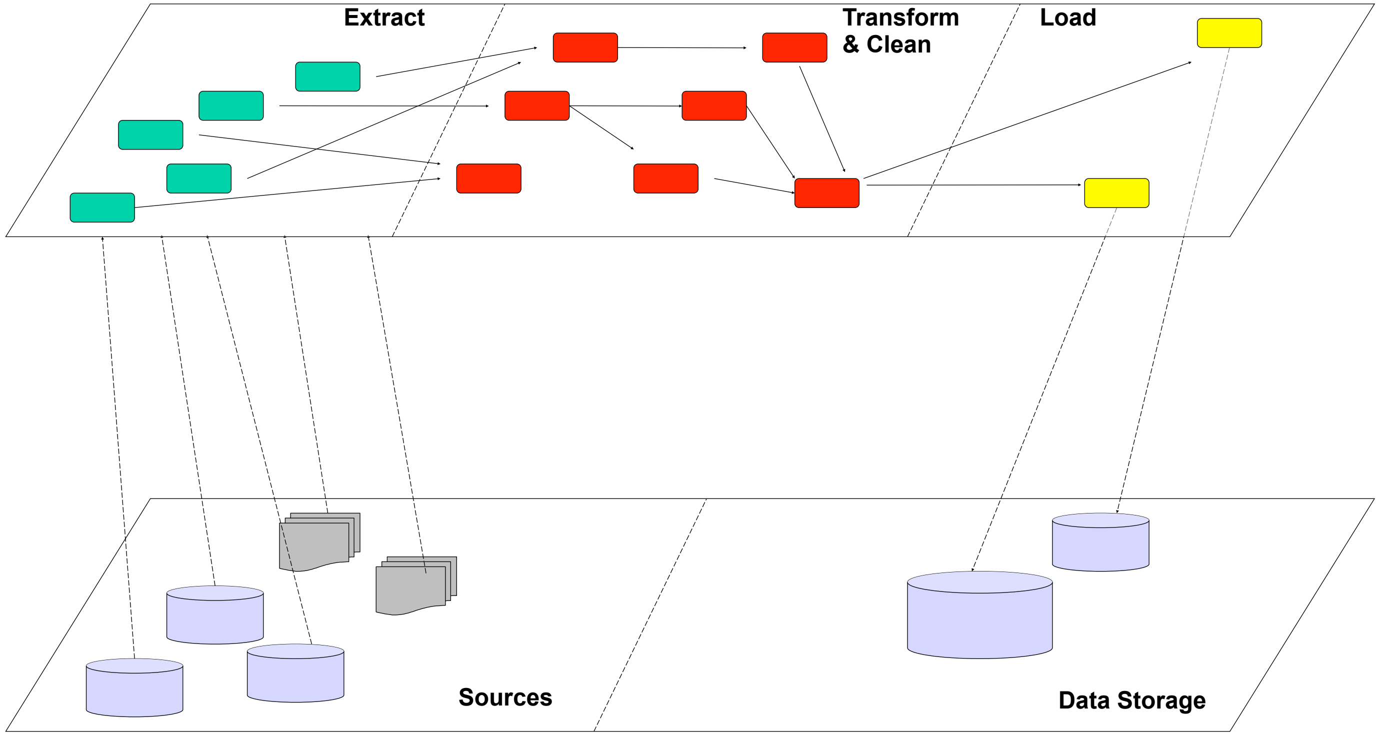
https://hbr.org/2018/08/what-data-scientists-really-do-according-to-35-data-scientists?referral=03758&cm_vc=rr_item_page.top_right

The Data Science Process



Data Engineering

ETL (Extract, Transform, and Load) is the process in which data is integrated and transferred from the operating systems to the data warehouse.



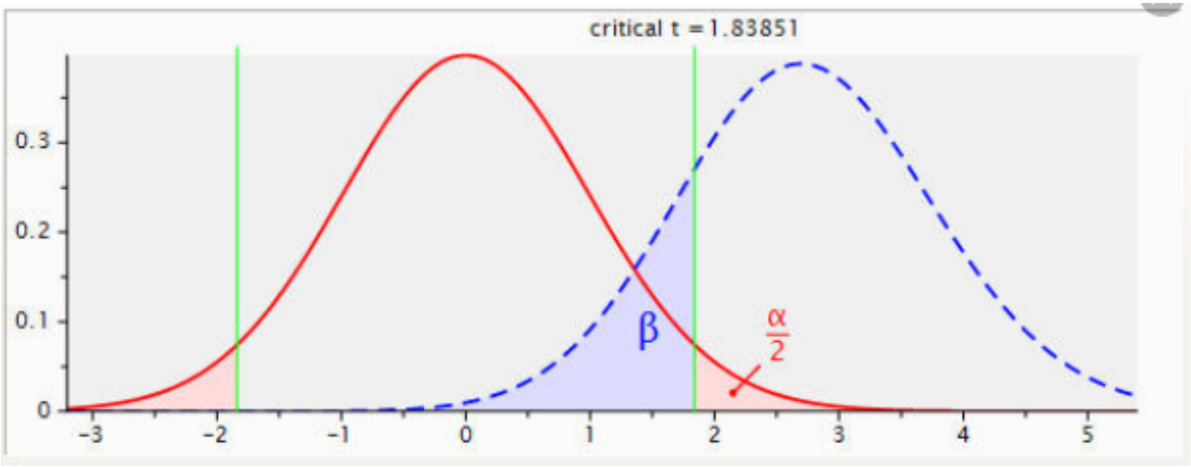
Data Staging Area

Big Data Storage and Processing

- How to manage massive amounts of data in a way which is optimized for analysis
- Learning general data warehousing models
- Post-rational technologies: based on distributed file systems and processing:
  - Hadoop
  - Hive
  - Spark

Experiments

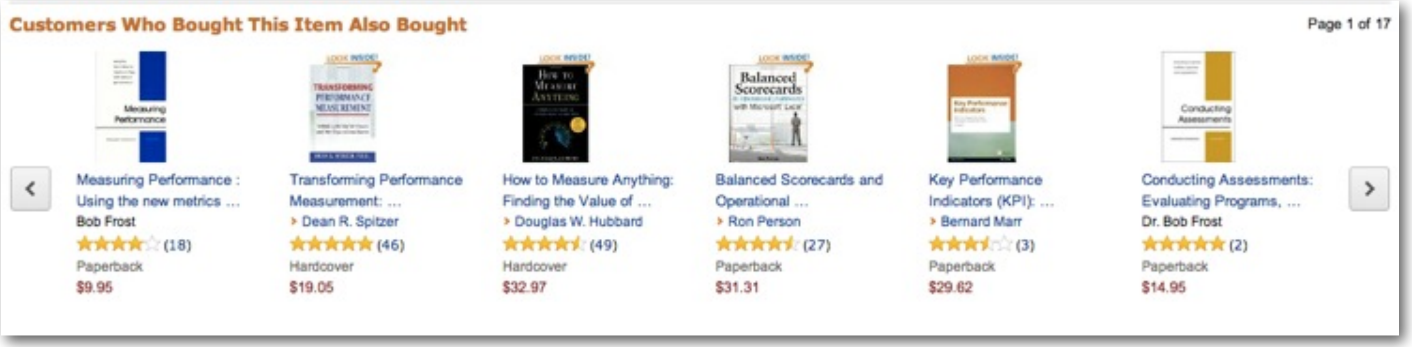
- Introduction to experiment design
- Parametric and non-parametric data modeling
- Statistical tests
- Running online experiments



The Interface with machine Learning

Understanding the interfaces with machine learning:

- Deep Sequence Learning
- Exploratory Data Analysis
- Reinforcement Learning



Customers Who Bought This Item Also Bought

Page 1 of 17

Measuring Performance : Using the new metrics ...
Bob Frost
★★★★☆ (18)
Paperback
$9.95

Transforming Performance Measurement: ...
› Dean R. Spitzer
★★★★★ (46)
Hardcover
$19.05

How to Measure Anything: Finding the Value of ...
› Douglas W. Hubbard
★★★★½ (49)
Hardcover
$32.97

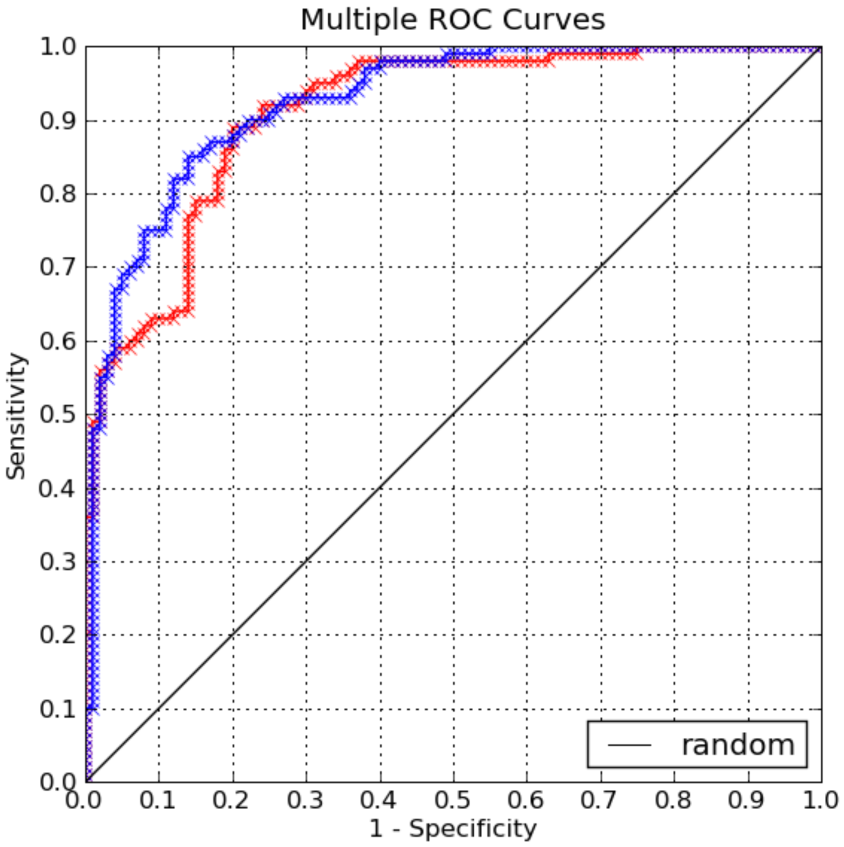
Balanced Scorecards and Operational ...
› Ron Person
★★★★½ (27)
Paperback
$31.31

Key Performance Indicators (KPI): ...
› Bernard Marr
★★★½☆ (3)
Paperback
$29.62

Conducting Assessments: Evaluating Programs, ...
Dr. Bob Frost
★★★★★ (2)
Paperback
$14.95

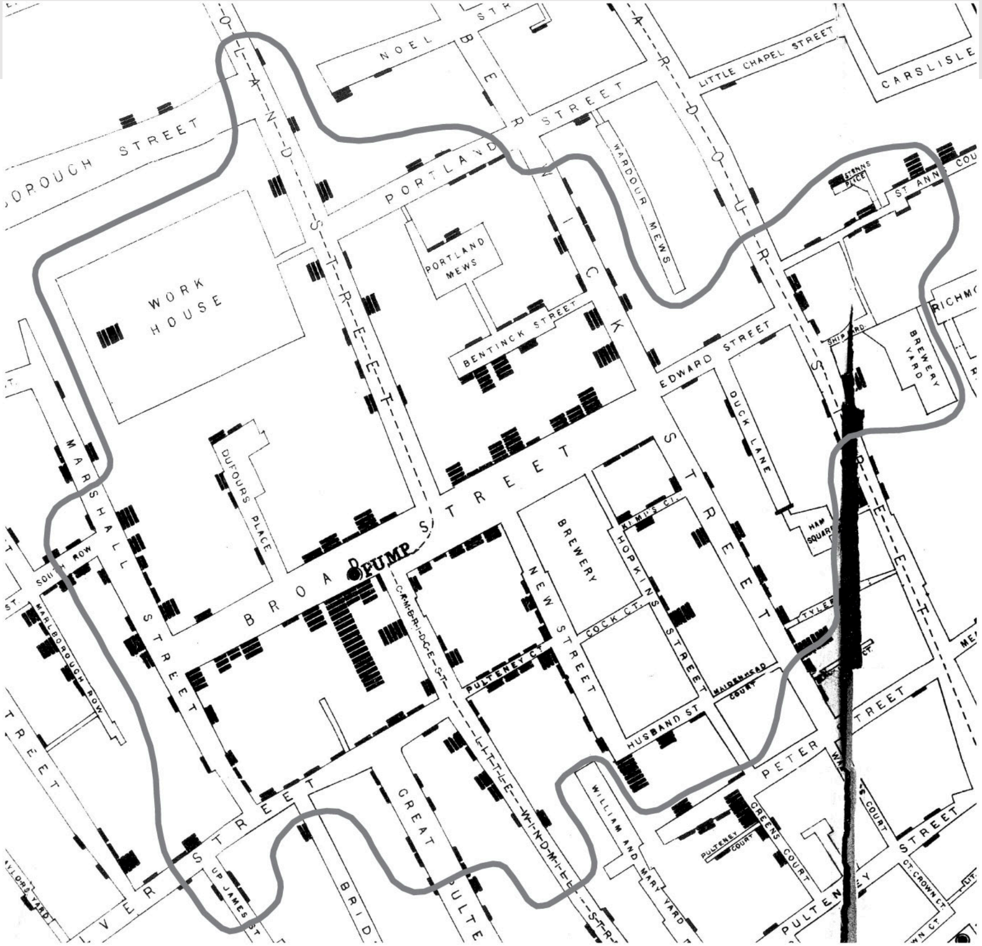
The Quality of Data Science

- How to evaluate the quality of data science models?
- Identifying bias
- Simulation
- Impact assessment
- Model explainability



Reporting

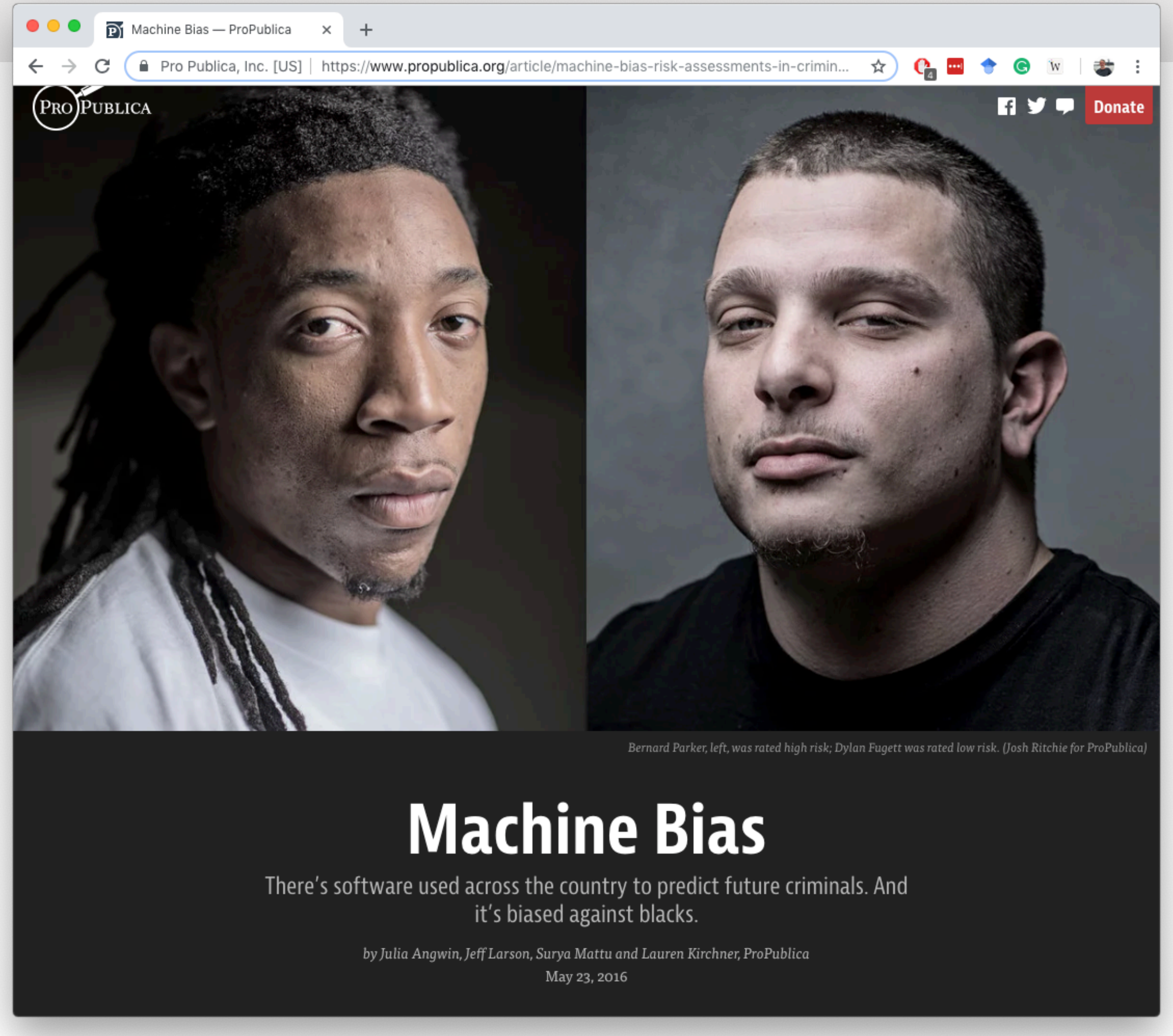
- Visualization methods
  - What makes a good data visualization?
- Reporting principles and communicating data
- Operationalizing data



John Snow's map of the 1854 Broad Street cholera epidemic

Ethics of Data Science

- Legal and ethical boundaries of data science
- Privacy
- Fairness


Machine Bias — ProPublica
Pro Publica, Inc. [US] | https://www.propublica.org/article/machine-bias-risk-assessments-in-crimin...
PRO PUBLICA
Donate

Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk. (Josh Ritchie for ProPublica)

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

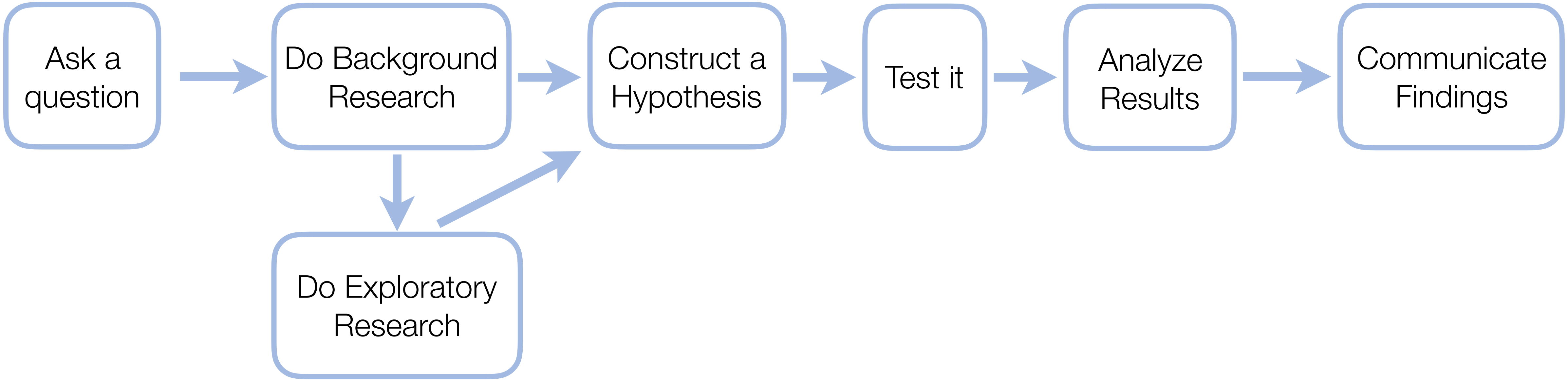
Summary

- Basic capabilities
- The course:
  - Website: https://eranto.github.io/cs5304-spring2019/
  - Slack: wild-data-science.slack.com
- The essence of the profession

What is the “science” in Data Science?

- Data science is more than an engineering practice
- It is a professional approach that strives to embed scientific principles in data tasks
- It includes:
    - Applying a scientific method
    - Adhering (to some extent) to scientific ethical code
    - And to its culture

The Data Science Process



Formulate a question

- Research questions should be:
  - Crunchy (either true or false)
  - Asking a question about something that can be observed: How, What, When, Who, Which, Why, or Where?
- Background research should make sure the questions should reflect the state of the art

Ask a question

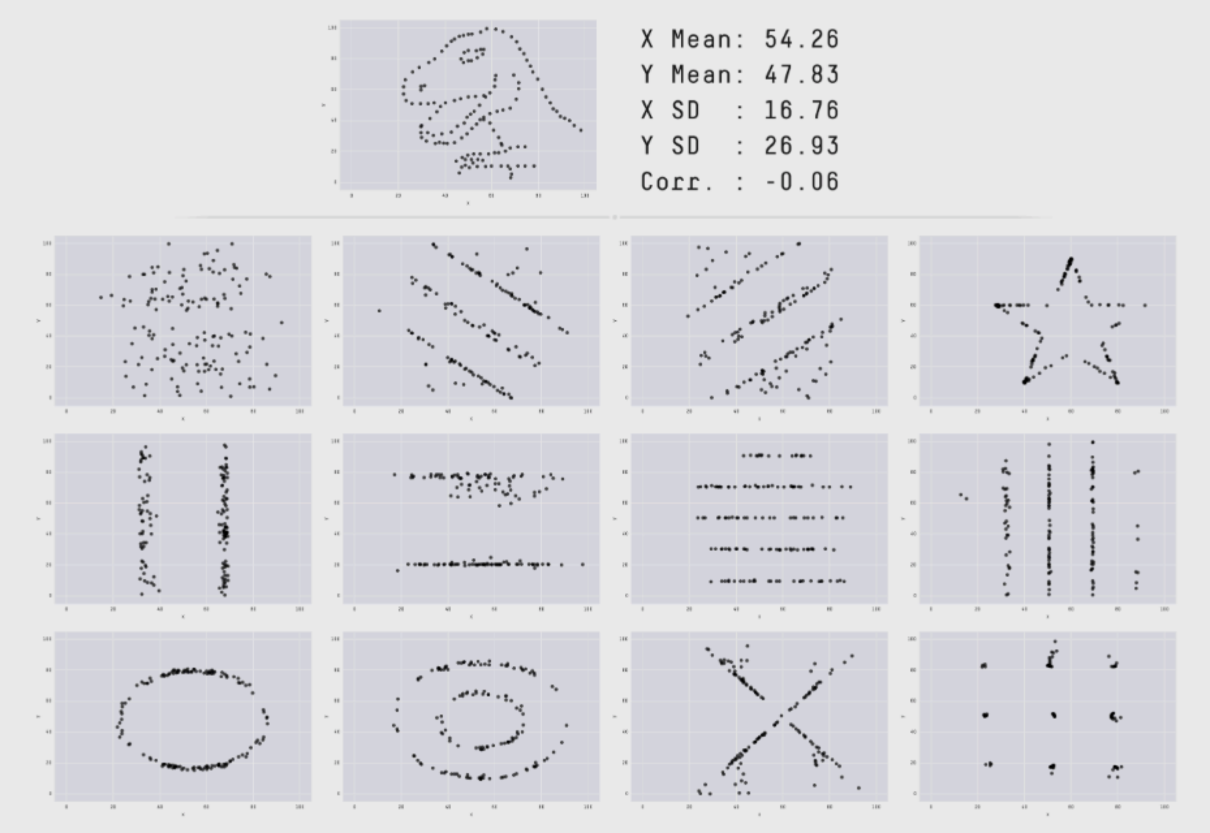
Hypotheses Making

- The scientific method asks for a clearly defined hypothesis:
  - an educated guess about how things work
- An exploratory data analysis can teach us about the data, but it is not enough
- We need to show that the prediction is accurate and thus the hypothesis is supported or not

Construct a Hypothesis

Levels of Modeling

- Classification and class probability estimation
- Regression ("value estimation")
- Similarity matching
- Clustering
- Association discovery
- Profiling
- Data reduction
- Casual modeling



https://www.autodeskresearch.com/publications/samestats

Analyzing Results

- Be ready to fail
  - Science is about taking some risks
- Analysis should lead to something bigger than just the current problem
  - In the academia, to the construction of theory
  - In practice, to the construction of generalizable business practices

Analyze Results

Communicating Findings

- Description of the hypotheses (so readers will know what had failed)
- Review of the state of the art
- Comprehensive description of the method
    - The standard is **reproducibility**
- Explanations of measures
- The actual findings, in a way that is both truthful and appropriate to the audience
- A discussion of the meaning of the findings

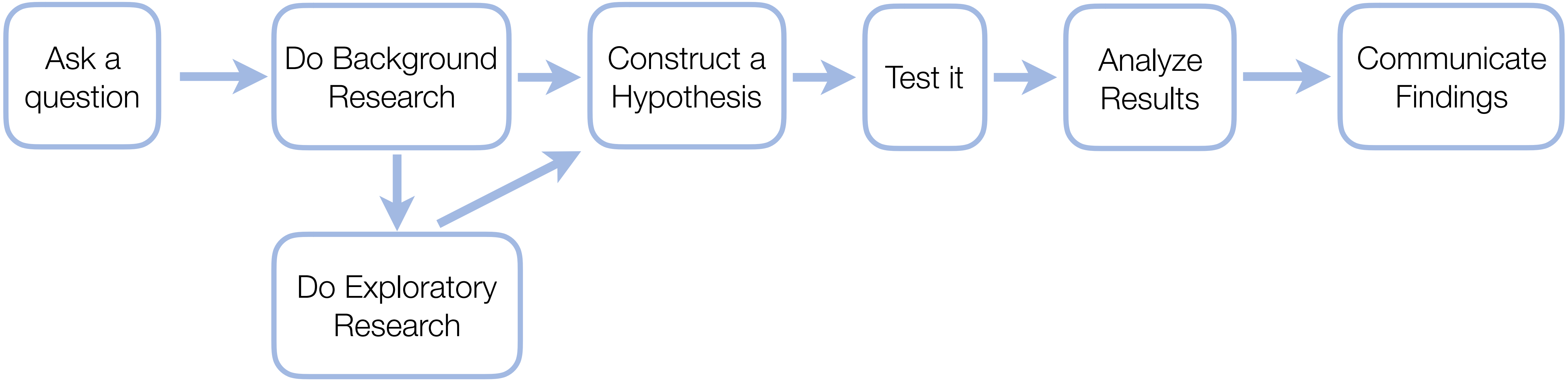
Communicate Findings

Thinking about the writing

- What is the problem?
- Why is it interesting and important?
- Why is it hard? (E.g., why do naive approaches fail?)
- Why hasn't it been solved before? (Or, what's wrong with previous proposed solutions? How does mine differ?)
- What are the key components of my approach and results?What are the limitations?

https://cs.stanford.edu/people/widom/

The Data Science Process


Ask a question
Do Background Research
Construct a Hypothesis
Test it
Analyze Results
Communicate Findings
Do Exploratory Research