

Foundations of Data Science

DS 3001

Data Science Program

Department of Computer Science

Worcester Polytechnic Institute

Instructor: Prof. Kyumin Lee

Course Objectives

- Introduce
 - the theoretical foundations, algorithms, and methods of deriving valuable insights from data
- Study
 - big data management and processing techniques, data analytics, statistical methods and models, data visualization, and etc

Goal of the Class

- Define and explain the key concepts and models relevant to data science.
- Design, implement, and evaluate the core algorithms underlying an end-to-end data science workflow.
 - E.g., the experimental design, data collection, mining, analysis and presentation of information derived from large datasets.
- Apply "best practices" in data science, including facility with modern tools (e.g., Hadoop/Spark).

Class Topics

- Data Exploration
- Data Preprocessing
- Mining and Analytics
- Visualization
- Evaluation
- MapReduce
- Recommender Systems
- ...

Course Structure and Administrivia

Course Information

- Instructor
 - Kyumin Lee
 - kmlee@wpi.edu
 - Zoom: <https://wpi.zoom.us/j/6735453923>
 - Office hours: T: 9:30-10:30am, W: 4:00-5:00pm, or by appointment
- TA
 - Jianjun Luo
 - jluo@wpi.edu
 - Zoom: <https://wpi.zoom.us/j/8546407944>
 - Office hours: M, R and F: 10:00-11:00am, or by appointment
- Class hours:
 - 2:00-3:50pm TF
 - Zoom: <https://wpi.zoom.us/j/331089671>

Course Information

- Course web page
 - <http://web.cs.wpi.edu/~kmlee/ds3001/>
- Course schedule page
 - <http://web.cs.wpi.edu/~kmlee/ds3001/schedule.htm>
- Canvas
 - <https://canvas.wpi.edu/>

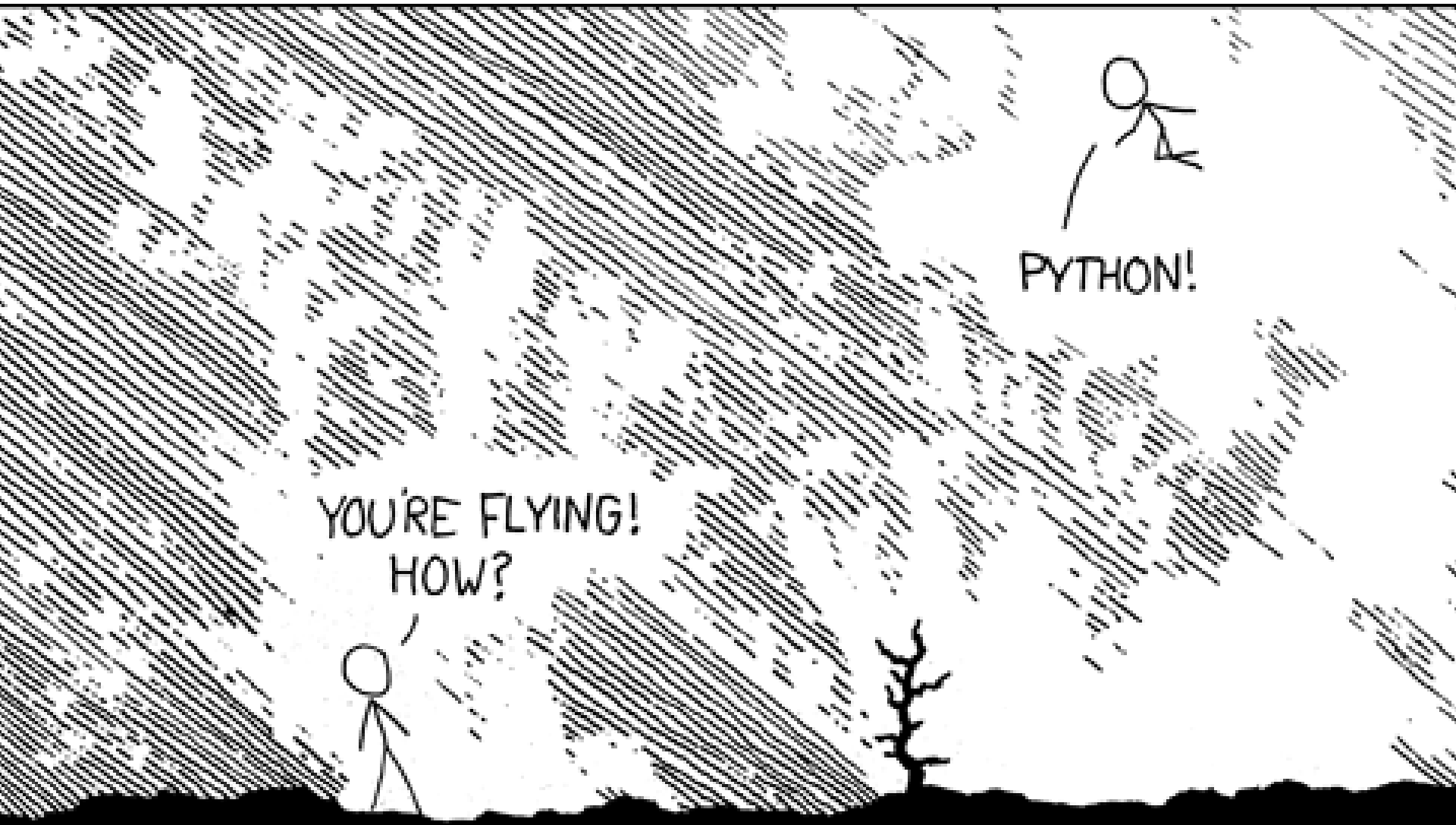
Course Materials

- No primary textbooks required
- But, we will refer to the following books:
 - (DMCT) [Data Mining: Concepts and Techniques](#), 3rd edition (2012). Jiawei Han and Micheline Kamber. Morgan Kaufmann.
 - (IDM) [Introduction to Data Mining](#) (2006). Pang-Ning Tan, Michael Steinbach and Vipin Kumar. Addison-Wesley.
 - (MMD) [Mining of Massive Datasets](#) (2020). Jure Leskovec, Anand Rajarman and Jeffrey D. Ullman. Cambridge University Press.
 - [Doing Data Science](#) (2013). Rachel Schutt, Cathy O'Neil. O'Reilly Media.
 - [Mining the Social Web](#), 3rd Edition (2019). Mikhail Klassens, Matthew A. Russell. O'Reilly Media.
 - [Data Science from Scratch](#), 2nd Edition (2019). Joel Grus. O'Reilly Media.
 - (Tufte) [The Visual Display of Quantitative Information](#) (2001) by Tufte.
 - (IIR) [Introduction to Information Retrieval](#), Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schutze, Cambridge University Press.
 - ...

Course Communication

- The lectures will be recorded and uploaded/linked to the Canvas page
- My lecture notes will be available under “Files” in the Canvas page
- I will email important announcements and post them to the website
- You may email me anytime ... but I only guarantee a response within two days

Official Language





Class Structure

- Lectures
 - By instructor -- I'll teach data science techniques
 - By us - Discussion and interaction in the class
- Your part
 - Homework
 - 4 assignments
 - Two Exams
 - Project
 - Proposal, execution, workshop presentation
- Participation
 - Ask good questions

Grading

- 40% (four) Assignments
- 30% Exams
- 30% Project

Assignments

Assignments

- 4 assignments
- Submit your solution to Canvas
 - You only use Canvas for submitting your assignments
- Late day policy: look at the syllabus

Exams

- The exams are closed book.
- You may bring one standard 8.5" by 11" piece of paper with any notes you think appropriate or significant (only use front page).

Project

The Project

- 3 or 4-person team
- Project idea:
 - Propose anything you wish
 - You are encouraged to talk to me
- **30% of your final grade!!**

Project Grading Criteria

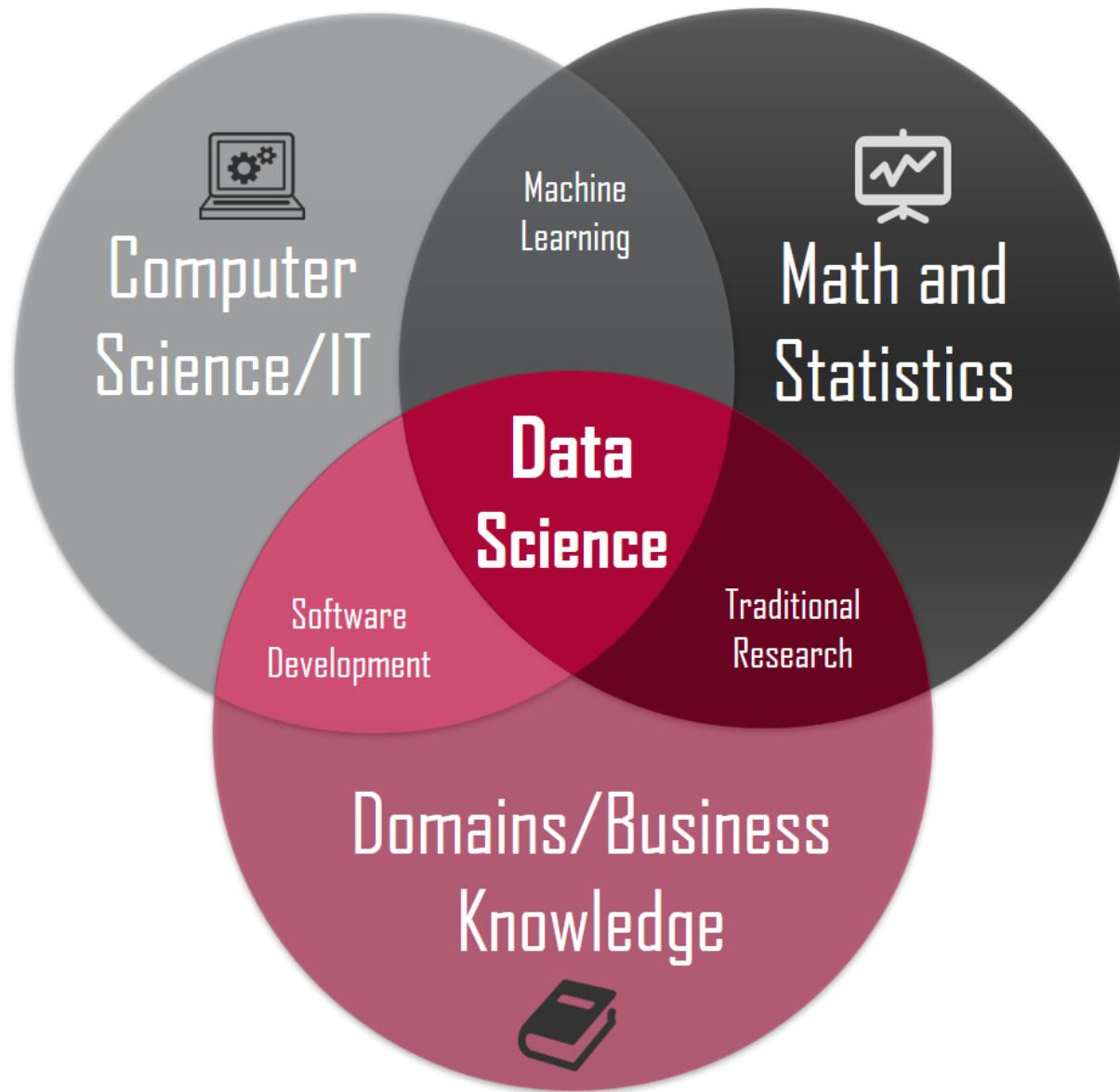
- Form a team by April 3
 - We will make teams and let you now.... But
 - You may share your preference by April 1
- [25%] Project Proposal: April 21
- [25%] Check Point: May 1
- [50%] Project Workshop: May 12

What's Next

What is Data Science?

Data Science...

- Data science is a "concept to unify statistics, data analysis, machine learning and their related methods" in order to "**understand** and **analyze** actual **phenomena**" with data.
- Data Science employs techniques and theories drawn from many fields within the context of **mathematics**, **statistics**, information science, and **computer science**.



Data Scientist Is the Best Job In America According Glassdoor's 2018 Rankings

Glassdoor 50 Best Jobs In America, 2018					
Ranking	Job	Median Base Salary	Job Score (5.0 scale)	Job Satisfaction (5.0 scale)	Job Openings
1	Data Scientist	\$ 110,000	4.8	4.2	4,524
2	DevOps Engineer	\$ 105,000	4.6	4	3,369
3	Marketing Manager	\$ 85,000	4.6	4	6,436
4	Occupational Therapist	\$ 74,000	4.5	4	11,903
5	HR Manager	\$ 85,000	4.5	3.9	4,458
6	Electrical Engineer	\$ 76,000	4.5	3.9	5,839
7	Strategy Manager	\$ 135,000	4.5	4.2	1,195
8	Mobile Developer	\$ 90,000	4.5	4.1	1,809
9	Product Manager	\$ 113,000	4.4	3.7	7,531
10	Manufacturing Engineer	\$ 72,000	4.4	4	4,241
11	Compliance Manager	\$ 96,000	4.4	4.3	1,222
12	Finance Manager	\$ 116,000	4.4	3.8	2,998
13	Risk Manager	\$ 97,000	4.4	4.2	1,209
14	Business Development Manager	\$ 75,000	4.4	3.9	4,060
15	Front End Engineer	\$ 100,000	4.4	4.2	1,222
16	Site Reliability Engineer	\$ 120,000	4.4	4.1	1,064
17	Mechanical Engineer	\$ 75,000	4.4	3.8	5,079
18	Analytics Manager	\$ 115,000	4.4	3.9	1,381
19	Tax Manager	\$ 110,000	4.4	3.7	3,309
20	Creative Manager	\$ 110,000	4.3	4.3	824
21	Software Engineer	\$ 102,500	4.3	3.6	29,187
22	Hardware Engineer	\$ 115,000	4.3	4.2	806
23	Corporate Recruiter	\$ 65,000	4.3	4.3	2,330
24	QA Manager	\$ 92,000	4.3	3.8	1,741
25	Physician Assistant	\$ 104,000	4.3	3.6	5,517
26	Database Administrator	\$ 94,000	4.3	3.8	2,370
27	UX Designer	\$ 90,000	4.3	3.8	1,963
28	Nursing Manager	\$ 84,660	4.3	3.7	4,209
29	Engagement Manager	\$ 115,000	4.3	3.7	2,169
30	Solutions Architect	\$ 125,000	4.2	3.6	3,325
31	Process Engineer	\$ 78,000	4.2	3.8	3,033

<https://www.forbes.com/sites/louiscolumbus/2018/01/29/data-scientist-is-the-best-job-in-america-according-glassdoors-2018-rankings/#5ff89ea65535>

in


data scientist

Jobs

Date Posted

Linked

Showing 17,077 results




Data Scientist - Data Scientist,

CyberCoders


Santa Monica, CA, US

MSc or PhD in a field like computer scienc

Strong skills in SQL, R, Python, SAS and rel

 1 alum works here

3 days ago



Data Analyst/ Data Scientist


enableit LLC

Florence-Graham, CA, US

Looking for a polished Data Scientist/Anal

Description. Expertise to perform analysis i

5 days ago




Data Scientist, Jr.


One Click Retail

Greater Salt Lake City Area

Strong foundation & understanding in fun

PCA. Experience with standard data scienc

3 days ago ·  Easy Apply




Data Scientist


BlueVine


Redwood City, CA, US

We are looking for a experienced Data Sci

efforts at modeling and researching consu

 1 alum works here

5 days ago ·  Easy Apply




Data Scientist

Progressive Leasing

Draper, UT, US

As a Data Scientist, you will blend analytics

result in massive improvements to the way

 2 alumni work here



Data Scientist

Confidential · Greater Boston Area

Posted 3 days ago · 1,456 views

Save

 Easy Apply



Job description

Data Scientist (Up to 165K Base)

Outstanding Company, Great Benefits with Amazing Culture!

Ideal candidates enjoy day-to-day data science problem-solving mixed with high levels of customer interaction.

Skills:

- 7+ years of engineering experience with python
- 3+ years of working experience in data science
- scikit-learn
- Java or C++
- Git
- Strong communication skills (English)

Experience with:

- Writing memory-efficient pre-processing pipelines
- Clustering and unsupervised learning
- Time series
- NLP

Bonus:

- Visualizing data with javascript
- Kaggle grandmaster, former top 10, or current top 100
- R, Java or C++
- Spark
- Geospatial modeling
- Core contributor of an open source machine learning or data science project

Contact the job poster

Kim Keys ^{2nd}

Matching Top Talent with Amazi...
Boston, Massachusetts

PREMIUM

Send InMail

Seniority Level

Associate

Industry

Information Technology and Services,
Computer Software

Employment Type

Full-time

Job Functions

Information Technology, Engineering

Is this Science?

Evolution of Sciences

- Before 1600, **empirical science**
- 1600-1950s, **theoretical science**
 - Each discipline has grown a *theoretical* component. Theoretical models often motivate experiments and generalize our understanding.
- 1950s-1990s, **computational science**
 - Over the last 50 years, most disciplines have grown a third, *computational* branch (e.g. empirical, theoretical, and computational ecology, or physics, or linguistics.)
 - Computational Science traditionally meant **simulation**. It grew out of our inability to find closed-form solutions for complex mathematical models.
- 1990-now, **data science**
 - The flood of data from new scientific instruments and simulations
 - The ability to economically store and manage petabytes of data online
 - The Internet and computing Grid that makes all these archives universally accessible
 - Scientific info. management, acquisition, organization, query, and visualization tasks scale almost linearly with data volumes. **Data Science** is a major new challenge!
- Jim Gray and Alex Szalay, *The World Wide Telescope: An Archetype for Online Science*, Comm. ACM, 45(11): 50-54, Nov. 2002

The key word in "Data Science" is not Data, it is Science

- Data science is only useful when the data are used to answer a question.
- It is much, much easier to say "My data are bigger than yours" or to say, "I can code in Hadoop, can you?" than to say, "I have this really hard question, can I answer it with my data?"
- The issue is that the hype around big data/data science will flame out if data science is only about "data" and not about "science". The long term impact of data science will be measured by the **scientific questions** we can answer with the data.

<http://simplystatistics.org/2013/12/12/the-key-word-in-data-science-is-not-data-it-is-science/>

Data Science:

Why should we care?

Data in Silicon Valley

<http://ed.ted.com/on/MYodUMYe>



Data in Wall Street



Data in Biomedical Research

consensus	AGAC.tcT...ca.A....gcTtATA.agAG..gAATTT.aAGGA.ACAC...ggaa.....gca...ccgCAGcgtAca.....tac.gtgAg...AT.cgAGtaccGgAT.gACGta.AAATT.AcCt.Tagaag.a....T.t...Aaga.gtct		
Rc_hemC	AGACATCTTTCCAAATTC--GCTTATAGAGAG--GTACTTTAAAGGAAACAC--AGAAGC--ACTTGGCCACC--CGCAGCGTACAAAAGG--ATG--GCACA--GATGCCAGTACCGGATTGACGTCACAAAT--ACCCCTAGAAGTAGA--GTTTGGGAAAGATGCTC	147	
Rc_RP502h	AGACATCTTTGCCAAATTC--GCTTATAGAGAG--GAATTTAAAGGAAACAC--GGAACG--CAACA--CCACAGCGGTACACTTAA--TAG--GTGAG--GATGCCGAGTACCGGATCGACGTCACAAAT--ACCTCTAGAAGCGAA--GTTTGGGAAAGATGCTC	144	
Rc_RR045h	AGACATCTTTGCCAAATTC--GCTTATAGAGAG--GAATTTAAAGGAGACAC--GAAACG--CAGCA--CCGCAGCGGTACACTTAA--TAG--GTGAG--GATGCCGAGTACCGGATCGACGTCACAAAT--ACCTCTAGAAGCGAA--GTTTGGGAAAGATGCTC	144	
Rc_RP545h	AGACATCTTTTAAACCA--GCTTATAGAGAG--GGATTTAAAGGAGACAC--GAAACA--CAGCA--CCGCAGCGGTACACTTAA--TAG--GTGAG--GATGCCGAGTACCGGATCGACGTCACAAAT--ACCTCTAGAAGCGAA--GTTTGGGAAAGATGCTC	144	
Rc_ubig	AGACACCTTTCCAAATTC--ACTTATAGAGAG--GAATTTGATAGGAAACAT--GCAGCA--CAGCA--CCGCAGCGGTACGCTTTG--TAG--GTGAG--GATGCCGAGTACCGGATCGACGTCACAAAT--ACCTCTAGAAGCGAA--GTTTGGGAAAGATGCTC	144	
Rc_ubih	AGACATCTTTGCCAAATTC--GCTTATAGAGAG--GAATTTGATAGGAGACAT--GAAACG--CAGCA--CCGCAGCGGTACAAAGC--TAG--GTGAG--GATGCCGAGTACCGGATCGACGTCACAAAT--ACCTCTAGAAGTAGAAGATTGGGAAAGATGCTC	147	
Rc_RP507h	AGACATCTTTGCCAAACCA--GCTTATAGAGAG--GAATTTAAAGGAGACAT--AGAATG--TAGCA--CCGCAGCGTACAAAAGG--TAG--TTGAG--GATGCCGAGTACCGGATCGACGTCACAAAT--ACCTCTAGAAGCGAA--GTTTGGGAAAGATGCTC	144	
Rc_RP167h	AGACATCTTTCCAAACCA--GCTTATAGAGAG--GAATTTAAAGGAGACAT--GAAACG--CAGCA--CTACAGCGTATATGGAC--TAG--TTGAG--GATGCCGAGTACCGGATCGACGTCACAAAT--ACCTCTAGAAGCGAA--GTTTGGGAAAGATGCTC	144	
Rc_mesJ	AGACATCTTTCCAAACCA--GTTTATAGAGAG--GAATTTAAAGGAGACAC--GAAACG--CAGCA--CCGCAGCGGTACACTTAA--TAG--GTGAG--GATGCCGAGTACCGGATCGACGTCACAAAT--ACCTCTAGAAGCGAA--GTTTGGGAAAGATGCTC	144	
Rc_era	AGACATCTTTGCCAAATTC--GCTTATAGAGAG--GAATTTAAAGGAGACACGGAAGCACT--TGCGC--CCGAG--CCGAG--TAG--GTGAG--GATGCCGAGTACCGGATCGACGTCACAAAT--ACCTCTAGAAGTAGA--GTTTGGGAAAGATGCTC	135	
Rc_orf1	AGACATCTTTCTAAACCA--GTTTATAGAGAG--GAATTTAAAGGAGATC--GAAACA--CAGCA--CCGCAGCGGTACACTTAA--TAG--GTGAG--GATGCCGAGTACCGGATCGACGTCACAAAT--ACCTCTAGAAGCGAA--GTTTGGGAAAGATGCTC	135	
Rc_gltX	ACACTTCTTGGCATCTC--CCTTATAGAGAG--GAATTTGAGGAAACAC--GAAACG--CAGCA--CCGCAGCGGTATATAGAC--TAG--GTGAG--GATGCCGAGTACCGGATCGACGTCACAAAT--ACCCCTAGAAGTAGA--GTTGCCGAGAAAGTTT	144	
Rc_mvN	CGACTTCT--GCATAACCTAGCTAATAAAGAG--AAATTTGAAGGAAACAC--GAA--CGCAGCA--CCGCAGCGGTACACTTAA--TAG--GTGAG--GATGCCGAGTACCGGATCGACGTCACAAAT--ACCCCTAGAAGCGAA--GTTTGGGAAAGATGCTC	144	
Rc_pcnB	AGATCTCT--TGCATAACAACTAATAAAGAG--GAATTTGAAGGAGACAC--GAA--CGCAGCA--CCGCAGCGGTACACTTAA--TAG--GTGAG--GATGCCGAGTACCGGATCGACGTCACAAAT--ACCCCTAGAAGCGAA--GTTTGGGAAAGATGCTC	144	
Rc_rlpA	AGACCTCT--TGCAGAACTGAAGCTAATAAAGC--AAATTTGAAGGAAACAC--GGA--TGTAGCTGCGAGCATACACTTAA--TAG--GTGAG--GATGCCGAGTACCGGATCGACGTCACAAAT--ACCTCTAGAAGCGAA--GTTTGGGAAAGATGCTC	143	
Rc_orf2	AGACTGCT--TACAGAACTGAAGCTAATAAAGAG--GCATTTGAAGGAGACAC--GAA--CGCGGACCCGCGAGTACACTTAA--TAG--GTGAG--GATGCCGAGTACCGGATCGACGTCACAAAT--ACCTCTAGAAGCGAA--GTTTGGGAAAGATGCTC	144	
Rc_kdtA	AGACTTCA--TGCAGAACTCGCTAATAAAGAG--GAATTTAAAGGAGACAC--TT--CACTGCGAAACCAACCTATACACTAA--TAG--GTGAG--GATGCCGAGTACCGGATCGACGTCACAAAT--ACCTCTAGAAGCGAA--GTTTGGGAAAGATGCTC	144	
Rc_orf3	AGACTCCT--TACATAACCTATCTTATAAAGAG--GAATTTGAAGGAGACAT--GAA--CGCAGCACCCGAGTACAAAAAG--TAA--GTGAG--ATCGGAGTATCGGATCGACGTCACAAAT--ACCTCTAGAAGCGAA--GTTTGGGAAAGATGCTC	143	
Rc_gmk	AGAGTGCT--TGCAAAATTC--GCTTATAGAGAG--GAATTTAAAGGAGACAT--GGAAGG--TAGCA--CCGACG--ACGTCCTAAAT--ACCTCTAGAAGTAGA--ACTTTCGAGAACTCTC	108	
Rc_rpe20	AGACTTCC--TGTAACACTTAGCTAATAAAGAG--GAATTTGTAGAGAGAC--GAA--CAGACGACCCGAGCGTACGCTTTG--TA--GTGAG--ATACGAGTAAACCGATCGACGTCACAAAT--ACCTCTAGAAGCGAA--TTATGC--AGAAAGTCTC	144	
Rc_rpe21	AGACTTCT--TGACTT--GCTAATAAAGAG--AAATTTGAAGGAGACAC--GAA--CGCAGAACCCGAGCGTACGCGAA--TAACGTGAGG--ATTCGAGTACCGGATCGACGTCACAAAT--ACCTCTAGAAGCGAA--TTATGC--AGAAAGTCTC	136	
Rc_rpe22	AAACTTTCTTAAATGACGTAGCTAATAAAGAG--AAATTTGAAGGAGACAC--GAA--CGCAGAACCCGAGCGTACGCGAA--TAACGTGAGG--ATTCGAGTACCGGATCGACGTCACAAAT--ACCTCTAGAAGCGAA--TTATTT--AGAAAGTCTC	146	
Rc_rpe23	AGACTTAT--TGCATAATATAGCTAATAAAGAG--GAATTTGAAGGAGACAC--GAAACACTTGGCACC--GCACGATACATATAAA--TAC--GTAGG--ATTCGAGTACCGGATCGACGTCACAAAT--ACCTCTAGAAGCGAA--TTATAC--AAAAAGTCTC	148	
Rc_rpe24	AGACTTCT--TGCATAACCTATCTTAAAGAG--AGACTTTGAAGGAGACAC--GAA--CGCAGAACCCGAGCGTACGCGAA--TAACGTGAGG--ATTCGAGTACCGGATCGACGTCACAAAT--ACCTCTAGAAGCGAA--TTATGC--AGAAAGTCTC	150	
Rc_rpe25	AGACTTGT--TGTGTAACTTATCTTAAAGAG--GAATTTGAAGGAGACAC--GGAAC--CGACGACCCGAGCGTACACTTAA--AC--GTAGG--ATTCGAGTACCGGATCGACGTCACAAAT--ACCTCTAGAAGCGAA--TTATAC--AGAAAGTCTC	143	
Rc_rpe26	AGACTTCT--TGCATAACCTATCTTAAAGAG--GAATTTGAAGGAGACAC--GGAAC--CATAGAACCCGAGCGTACCGAA--TAT--GTAGG--ATTCGAGTACCGGATCGACGTCACAAAT--ACCTCTAGAAGCGAA--TTATGC--AGAAAGTCTC	106	
Rc_rpe27	AGACTTAT--TGCATAACCTATCTTAAAGAG--GAATTTGAAGGAGACAC--GGA--CATAGAACCCGAGCGTACCGAA--TAT--GTAGG--ATTCGAGTACCGGATCGACGTCACAAAT--ACCTCTAGAAGCGAA--TTATGC--AGAAAGTCTC	142	
Rc_rpe28	AGGCTTCT--TGCATACCTATCTTAAAGAG--GAATTTGAAGGAGACAC--GGAACA--CAGCACTCGAGCGTACGTCACAAAT--TAT--GTAGG--ATTCGAGTACCGGATCGACGTCACAAAT--ACCTCTAGAAGCGAA--TTATGC--AGAAAGTCTC	144	
Rc_rpe29	AGGCTTCT--TGCACAAACGAGCTAATAAAGAG--GAATTTGAGGAGACAC--GAA--CAGCAACTCGAGTACGTCACAAAT--TAC--GTAGG--ATTCGAGTACCGGATCGACGTCACAAAT--ACCTCTAGAAGCGAA--TTGTGC--AAAAAGTCTC	144	
Rc_rpe30	AGACTTCT--TGCAGAACGAGCTAATAAAGAG--GAATTTGAGGAGACAC--GAA--CGCAGTATCGAGCTACACTTAA--TAG--GTAGG--ATTCGAGTACCGGATCGACGTCACAAAT--ACCTCTAGAAGCGAA--TTATGC--AGAAAGTCTC	144	
Rc_rpe31	AGACATTTTCCAAACCA--GCTTATAGAGAG--GAATTTGAAGGAGACAC--TTCCCC--TCGAA--CCGACAGTACAAAAGC--TAA--GTAG--GATTCGAGTACCGGATCGACGTCACAAAT--ACCTCTAGAAGCGAA--ATTCGAAAGATGCTC	144	
Rc_rpe32	AGACATCTTTCCGAAACCG--GCTTATAGAGAG--GAATTTGAAGGAGACAT--GGAATG--CAGAA--CCGACGCTATATA--CG--GAG--GTAG--GATTCGAGTACCGGATCGACGTCACAAAT--ACCTCTAGAAGTAGA--GTTTGGGAAAGATGCTC	143	
Rc_rpe33	AGACCTCTTTCCAAACCA--ACTTATAGAGAG--GAATTTAAAGGAGACAC--GGACA--CAGCA--CCACAGCATATAGAA--TAG--GTAG--GATTCGAGTACCGGATCGACGTCACAAAT--ACCTCTAGAAGCGAA--GTTTGGGAAAGATGCTC	144	
Rc_rpe34	AGACATCTTTCCAAA--A--ACTTATAGAGAG--GAATTTAAAGGAGACACGAAAGCACT--TGTC--CCGACGATACAAA--TAG--GTAG--GATTCGAGTACCGGATCGACGTCACAAAT--ACCTCTAGAAGCGAA--GTTTGGGAAAGATGCTC	145	
Rc_rpe35	AGACATCTTTCAGAACTC--GCTTATAGAGAG--GAATTTGAAGGAGACAC--GCCACG--CAGAA--CCGCAGCGTACATAGAG--TAG--GTAG--GATTCGAGTACCGGATCGACGTCACAAAT--ACCTCTAGAAGTAGA--GTTTGGGAAAGATGCTC	144	
Rc_rpe36	AGACCTCTTCCGAACTC--GCTTATAGAGAG--GAATTTGAAGGAGACAC--GGAACA--CAGAA--CCGCAGCATACAAAAG--TAG--GTAG--GATTCGAGTACCGGATCGACGTCACAAAT--ACCTCTAGAAGTAGA--GTTTCCGAAAGAGTCTC	144	
Rc_rpe37	AGACTTCTTCCGAAAC--GCTTATAGAGAG--GAATTTGTAGGAAACAC--CAGACA--CAACGAA--CCGACGCTACGTAGAA--TAG--GTAGGAGGATGTGAACCGGATCGACGTCACAAAT--ACCTCTAGAAGTAGA--GTTTCCGAAAGAGTCTC	149	
Rc_rpe38	AGACTTCTTTCGAACTC--GCTTATAGAGAG--GAATTTGAAGGAGACAT--GGAACG--CAGCA--CCGCAGCGTATACAAA--TAG--GTAG--GATTCGAGTACCGGATCGACGTCACAAAT--ACCTCTAGAAGTAGA--GTTCCGAAAGAGTCTC	144	
Rc_rpe39	AGACGCTCT--TACATAACGAGCTAATAAAGAG--AAATTTGAAGGAGACAC--GAA--CGCAGAACCCGAGCGTACACT--T--GAC--ATGAGG--ATTCGAGTACCGGATCGACGTCACAAAT--ACCTCTAGAAGTAGA--TTATGT--AGAAAGTCTC	141	
Rc_rpe40	AGCAATCTT--TACAAACGCTTGTATAAAGAG--GAATTTGAGGAGACAC--AGAAGC--CAGCA--CCGACG--ACGCAAAAT--ACCTCTAGAAGTAGA--GTATATGTAGGAACTCGC	108	
Rc_rpe41	AGGCTTCT--TGCATAAATAACTTATAGAGAG--GAATTTGAAGGAGACAC--AGAA--TGCAGAACCTGACGTAGTAGAC--CTAG--TTGAGG--ATTCGAGTACCGGATCGACGTCACAAAT--ACCTCTAGAAGTAGA--TTTTCG--AGGAGGCTC	143	
Rc_rpe42	AGACATCTTTCAGAACTC--GTTTATAGAGAG--GAATTTGAAGGAGACAC--AGAATG--CAGAA--TTCGAGTACCGGATCGACGTCACAAAT--ACCTCTAGAAGTAGA--GGCTATATCTTGTGA	142	
Rc_rpe43	AGAAATCT--TCCCAAGTA--GTTTATAGAGAG--GAATTTGAAGGAGACAC--GGAAG--CACTTGGCTCGCAGCATATATAGAC--TAG--GTAG--ATTCGAGTACCGGATCGACGTCACAAAT--ACCTCTAGAAGTAGA--TTATAC--AGAAAGTCTC	146	
Rc_rpe44	AAACTTAT--TGCATAAATACTTATAAAGAG--AAATTTGAAGGAGACAC--GGA--CGCAGAACCCGAGCGTACAAAAG--GAC--GTAG--ATTCGAGTACCGGATCGACGTCACAAAT--ACCTCTAGAAGTAGA--TTGGAA--GCCAAATATA	142	
Rp_RP474	AGATATTTTCTAAACCA--GCTTATAGAGAG--GAATTTAAAGGAGATAC--AATACG--TGGCA--CCACAGCGTATATAAAG--TAT--TTAAT--GATCCGATCTAGGCT-----CAACGTA--TCCCTAGAAGTAGC--GTTTGGGAAAGATGCTC	138	
Rp_coxB	AGATATTTGCTCAACAA--TCTTATAAAGATTAAAGTAGATCAAG--ATAACA--CCA--CA--GAATCACAAACG--TAT--GTAAT--AGCAGGATCTAGGTTGACGTCACAT--ACCTCTAAAAGTAGA--ATTATGGAAGATATAT	141	
Rp_RP68h	AAACTTTT--TGATAACATACTAATAAGAG--GAATTTAAAGGAGATAC--GGA--TGGCA--CGCTGCTGCTAGCAAAAC--TAT--GTATG--ATTCAAGTACCAATCAAGCTCTAAAT--AGCTTTAATAAGAG--TTATAC--ACAAATATCTC	144	
Rp_kdtA	GAATTTCA--TACAAATGCTGTTAATAAAGAG--GGATTTACAGACACAA--AACTACATCTTATAGACA--TGC--ATAGAA--ATGCAAGTCTA--ATGTATAAAT--ATCTTTAAGAAAG--TTATGC--ACACAGTCTC	129	
Rp_alr	AGAGCTCT--CGTCTAATCTGCTATATAGAGAG--GAATTTGAAGGAGACAC--AGCAGCTCTTATACC--ACAGCGTGTATAAAT--AAT--GTCCAG--ATGTTAGTGTGAGATGACCTCTAAAT--ACCTTTAAAAGAG--TTATAC--TGATAGTCTC	147	
Rp_RP545	AGATATCTCTCTAAACAA--ACTGATAGAAAC--GAATTTATCTACTGCG--AGAACT--GTA--TATAGAAATATCTAGCA--TAA--GA--GCTACTACATAACTT--ATAAAT--ACCTCTAAAAGAGCA--ATTGTGAATGTCTC	132	
Rp_lycC	ATACCTCTTCTGAAATCC--AGGTATATGGAAG--GAATTCGCACTAAATAT--GAAACA--CAGCA--TAAA--ATAGACA--TAG--GTAG--GATGCTAGTGGAT--CAACATATAAAT--ACCATAGATAGG--ATTACAAAACAGGTATC	135	
Rp_pyrG	CGACATATTTCCAAATCT--ACTATAGTGAA--GCATTTG--AATGC--CAGCGCT--T--CCACTGTATATACAACT--TAG--ATAG--GATTCGAAATCACTGTGTCTACAGCAAAAT--CCGATAGAAAGCA--TTTTGAAATATGCTC	135	
Rp_RP404	AGACATTTTCTAAACTA--ACATATAGCAAG--AAATTT--GAA--GCAAT--GCAATCTGAAGCA--TTT--G--CAACCACAGCT--ACCAAT--ACCATATAAATAG--ATTGAGGAATGCACTC	108	
Rp_rpe55	AGACTTCT--TGCATAAGTAGCTAATAAGCT--AAATTTGTAATAAACC--AGCA--C--ATAGTAGATACGAATA--AG--GTAG--ATACGAGTACTGGAATTGACATACAAAT--ACCTTTAGAAGAGG--CTATAG--AGGAAGCTC	135	
Rh_pola	AGGCTCTTTTCCAAATTC--GCTTATAGAGAG--GAATTTAAAGGAGACAC--GGAACG--CAGCA--CCGCAGCGTACAAAAGC--TAG--GTAG--GATGCCGAGTACCGGATCGACGTCACAAAT--ACCTCTAGAAGTAGA--GTTTGGGAAAGAGGCTC	144	
Rf_pola	AGACATCTTTCTAAACCA--GCTTATAGAGAG--GAATTTAAAGGAGACACGGAAGCACT--TGCCA--CCGCAGCGTACAAAAGC--TAG--GTAG--GATGCCGAGTACCGGATCGACGTCACAAAT--ACCTCTAGAAGCGAA--GTTTGGGAAAGATGCTC	147	
1.....10.....20.....30.....40.....50.....60.....70.....80.....90.....100.....110.....120.....130.....140.....150.....160.....			

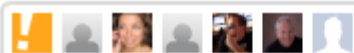
<https://www.youtube.com/watch?v=FzcTgrxMzZk>

BBC documentary

- The Age of Big Data
- Crime Prevention
- ...



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



306 comments, 167 called-out

[+ Comment Now](#)

[+ Follow Comments](#)

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 rampant — and loyal — buyers of all things pastel, plastic, and miniature. He talked to Target statistician Andrew Pole — before Target freaked out and cut off all communications — about the clues to a customer's impending bundle of joy. Target assigns every customer a Guest ID number, tied to their credit card, name, or email address that becomes a bucket that stores a history of everything they've bought and any demographic information Target has collected from them or bought from other sources.



Target has got you in its aim



All

Posts

People

Photos

Videos

Pages

Places

Groups

Apps

Filter Results

POSTS FROM

- ☒ Anyone
- ☐ You
- ☐ Your Friends and Groups
- ☐ Choose a Source...

POSTED IN GROUP

- ☒ Any group
- ☐ Your Groups
- ☐ Choose a Group...

TAGGED LOCATION

- ☒ Anywhere
- ☐ Choose a Location...

DATE POSTED

- ☒ Any date
- ☐ 2018
- ☐ 2017



Women who live in New York, New York and are Software engineer

**Evy Reds (Roo)**[Add Friend](#)

Software engineer at Facebook

Female

Lives in New York, New York · From Yonkers, New York

Studied Medicine at California State University, Northridge

**Alex Morgan**[Add Friend](#)

Software engineer at Facebook

Single · Female

Lives in New York, New York

Studies Muchas Cosas Wuuuuuuuu at Universidad Nacional Costa Rica

**Tina Ting-Chu Lin**[Add Friend](#)

Software engineer at Facebook

Female

Lives in New York, New York · From Hualian City

Studied Computer science at Columbia University '14

[See All](#)

Recommendation



Hi,
CUSTOMER SINCE 2014

YOUR ORDERS
1 recent order

TOP CATEGORIES FOR YOU
Prime Video
Movies & TV
Electronics

PRIME

Exclusive: Fire TV Stick with
Alexa Voice Remote \$24.99



VIDEO

Continue watching:
Lego Destruction



MUSIC

Amazon Music Unlimited: all
the music you love



MEET ALEXA

Voice control your world with
Echo & Alexa devices



AUDIBLE

Get hooked on audiobooks
We think you'll enjoy:



Deals recommended for you [See all deals](#)



\$98.40
\$179.99
Ends in 16:45:50



\$11.99 - \$179.99
Ends in 16:45:50



\$47.90 - \$67.50
Ends in 16:30:51



\$1,799.99
\$2,199.99
Ends in 16:40:51



\$39.99
\$49.99

Amazon Gift Cards



Millions of items,
no expiration.

>Shop now

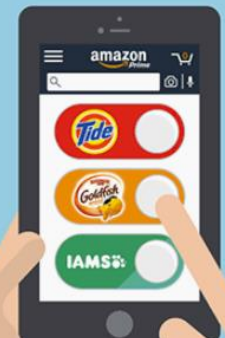
[Ad feedback](#)

Inspired by your shopping trends



New virtual Dash Buttons

Shortcuts to shop
your favorite products



The Data Science Process

Jeff Hammerbacher

1. Identify problem
2. Instrument data sources
3. Collect data
4. Prepare data (integrate, transform, clean, impute, filter, aggregate)
5. Build model
6. Evaluate model
7. Communicate results

Jim Gray

1. Capture
2. Curate
3. Communicate

This Class

Ask question: What data needs to be recorded? or collected?



Real World



Humans behaving
Biology
Finance
Internet
Medicine
Sociology
Olympics



Raw Data is
Collected / Recorded

email
logs
medical records
surveys
blood drawn
(microarray)
olympic records
NYT web pages
Social media data



Data is
Processed

pipelines
web scraping
cleaning
munging
joining
wrangling



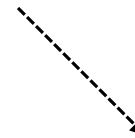
Data Set

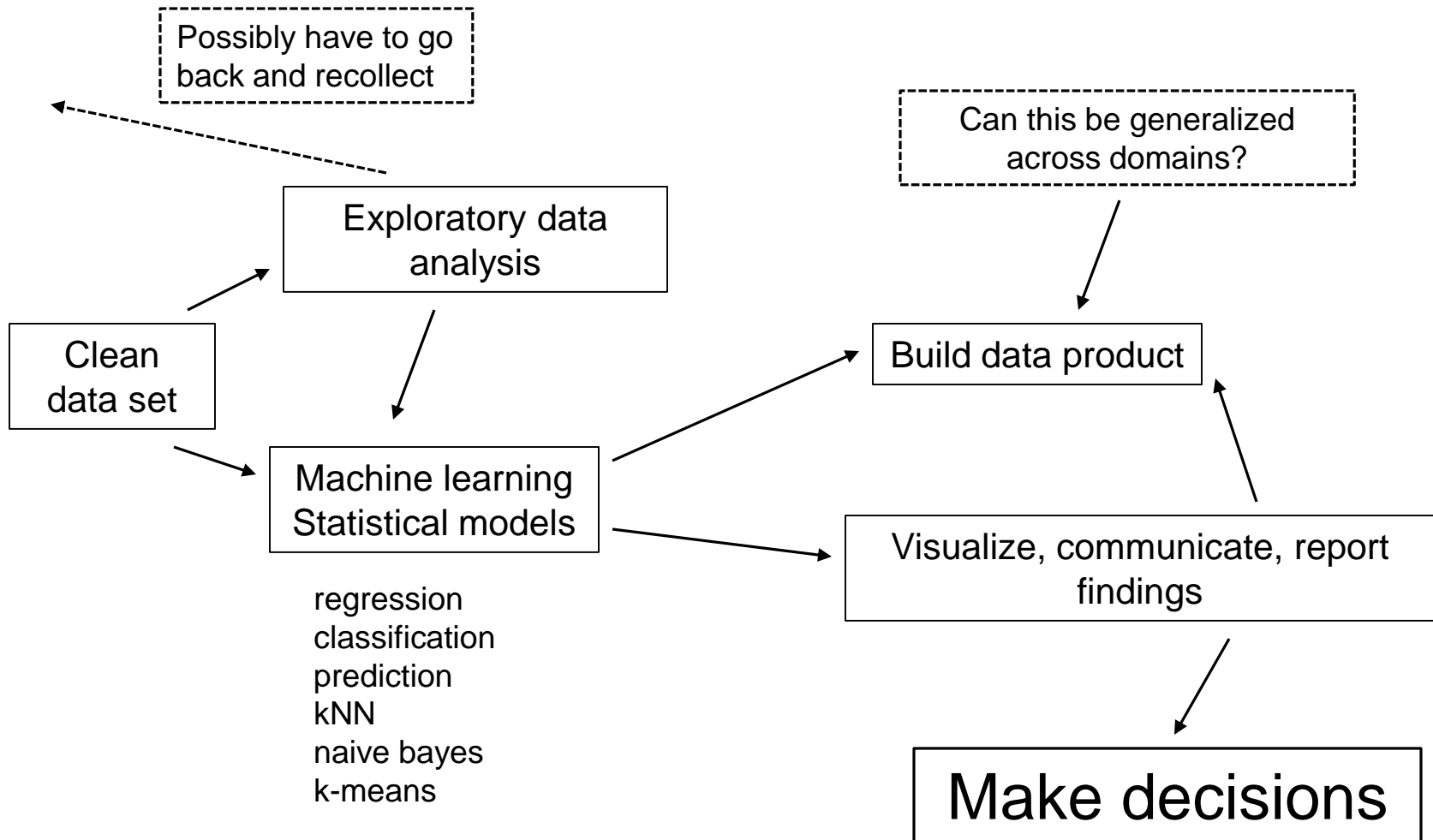
“clean” table

Why? What research question
am I going to answer?



What do I want it to look like?





Data Collection

1. Download Existing Datasets

- <https://www.kdnuggets.com/datasets/index.html>
- <https://github.com/caesar0301/awesome-public-datasets>
- <https://snap.stanford.edu/data/index.html>
- <https://www.kaggle.com/>
- <http://dumps.wikimedia.org/>
- <https://www.yelp.com/dataset/challenge>
- etc

2. Crawling Webpage and Process HTML

- HTML is all about how to display/show data, but not about giving you the data.
- Easy to download, but, hard to process
- Powerful, but it is your last choice of getting data from websites
- Follow rules: robots.txt

3. Use Web Application Programming Interfaces (APIs)

- Twitter: <https://developer.twitter.com/en/docs.html>
- Flickr: <https://www.flickr.com/services/api/>
- Google Maps: <https://developers.google.com/maps/documentation/>
- Facebook: <https://developers.facebook.com/docs/apis-and-sdks/>
- Foursquare: <https://developer.foursquare.com/>
- Airbnb: <https://www.airbnb.com/partner>
- Wikipedia API: https://www.mediawiki.org/wiki/API:Main_page
- Youtube API: <https://developers.google.com/youtube/>
- Halo API: <https://developer.haloapi.com/>

A list of Web APIs is in

<http://www.programmableweb.com/apis/directory>

Social Media Data

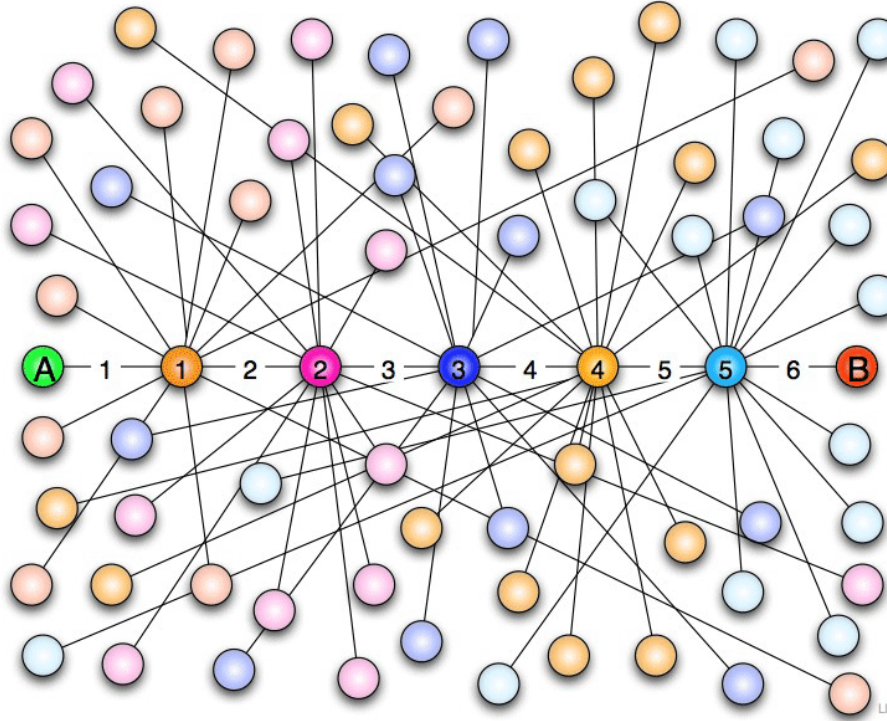
"six-degrees of separation"



296
letters
64 arrived



Stanley Milgram
Harvard University
1967



190 million
people



Jure Leskovec, Eric Horvitz
CMU, Microsoft
2007

Graph Data: Social Networks



Facebook social graph

4-degrees of separation [Backstrom-Boldi-Rosa-Ugander-Vigna, 2011]

April 23, 2013

Dow Jones Industrial Average ^DJI

14,695.42 +128.25 (0.88%)



The Associated Press

@AP

News, discussion and a behind-the-scenes look at the process from The Associated Press. Managed 24/7 by a team of editors based in NY: apne.ws/APStaff

Global · <http://www.ap.org>

50,187
TWEETS

7,012
FOLLOWING

1,904,925
FOLLOWERS



Following

Tweets All / No replies

AP

The Associated Press @AP

5m

Breaking: Two Explosions in the White House and Barack Obama is injured

Expand

AP

The Associated Press @AP

28m

Democratic Sen. Baucus, Finance committee chairman, says he won't run for re-election: apne.ws/10bIOFf -CJ

View summary



The Associated Press 
@AP



Following

Breaking: Two Explosions in the White House and Barack Obama is injured



Reply



Retweet



Favorite



More

3,146

RETWEETS

149

FAVORITES



1:07 PM

23 Apr 13

Hackers used the Associated Press Twitter account to send out a false tweet about an attack on the White House, briefly causing stock prices to tumble Tuesday.