# Info 2950 Fall 2014

| | Date | Lecture | Reading |
|---|---|---|---|
| Week 1 | Thu 8/29 | 1. Course introduction, set theory | Rosen ch 2.1-2.2 (online here)<br>Notes on set theory<br>(Ran out of time for ipython demo.) In the meantime, here are some instructions for installing python.<br>The python.org site has a tutorial, and there are other online resources, including the book Think Python. |
| Week 2 | Tue 9/3 | 2. Probability and counting | Began with ipython notebook demos: lec2a.ipynb (intro), lec2b.ipynb (sets)<br>Rosen ch 6 'Counting' (ch 5 in earlier editions)<br>Notes on probability<br>(above notes updated, and link to on-line chpts available on request) |
| | Thu 9/5 | 3. Probability and counting, cont'd. | Continued Notes on probability<br>Re "birthday" problem, see lec3.ipynb, and as well nytimes popular version (Strogatz).<br>Note Problem Set 1 due Thu 12 Sep, announced above |
| Week 3 | Tue 9/10 | 4. Conditional probability + Bayes' theorem | Rosen ch 7.1-7.2 ('discrete probability', ch 6 is older editions, link available on request)<br>Notes on cond'l prob and Bayes (now updated)<br>mentioned Wason selection task |
| | Thu 9/12 | 5. Bayes' theorem | Rosen ch 7.2, 7.3<br>continue above notes, mentioned Doomsday Argument, |
| Week 4 | Tue 9/17 | 6. Applications of Bayes' theorem | Rosen ch 7.3 plus notes (will be updated)<br>Discussed A plan for spam<br>Introduced The Unreasonable Effectiveness of Data (from within Cornell, click on "IEEE Xplore Subscribers" to get pdf -- **please read, it's fun**)<br>Went through lec5.ipynb<br>Note Problem Set 2 due Thu 26 Sep, announced above |
| | Thu 9/19 | 7. expected value, variance, binomial and normal distribution | A few more 'naive Bayes' resources: details on 'naive' Bayes, Bayesian spam filtering, and nytimes popular Bayes (Strogatz)<br>Rosen 7.4 plus continue notes. |
| Week 5 | Tue 9/24 | 8. more on variance and Bernoulli process | Rosen 7.4 plus finished notes.<br>Went through lec8.ipynb |
| | Thu 9/26 | 9. Gaussian (Normal) Distribution, Exponentials and logarithms, Zipf | notes on exponentials,<br>links for deviations from normal: Poincare bread and on-line dating<br>popular logarithms (Strogatz nytimes, includes video link)<br>Went through lec9.ipynb (uses politics/sports dataset from ps2#8) |
| Week 6 | Tue 10/1 | 10. Poisson distributions, graph theory | some notes on Poisson distribution, and real data: arxpoi.ipynb<br>start graph theory (Rosen ch 10.1-10.4, and these notes1) |
| | Thu 10/3 | 11. Graph theory: Eulerian and Hamiltonian circuits | Rosen ch 10.5, and notes2 (see also Hamiltonian circuit and Seven Bridges of Königsberg) |
| Week 7 | Tue 10/8 | 12. Trees, planarity, spanning trees, DAGs | notes3 (Rosen ch 10.7, 11.1)<br>notes4 (Rosen 11.3-11.4)<br>and proof of minimum spanning tree algorithms |
| | Thu 10/10 | 13. Applications of graphs | Covered example of Dijkstra algorithm (Rosen 10.6), and the graph partitioning algorithm described in pp 69-83 of Easley/Kleinberg Chpt 3 (not covered in info 2040), final result was fig 3.20 on p.81<br>Problem Set 3 due, ps3.8 extended to after break. |
| Week 8 | Tue 10/15 | fall break | |
| | Thu 10/17 | 14. Network statistics | Reviewed topics for midterm (enumerated below)<br>Reviewed Dijkstra algorithm (Rosen 10.6), and minimum spanning tree algorithms. Discussed random graphs and Poisson distributed degrees, then "Why your friends have more friends than you do" (Feld, 1991; slide, plus nytimes popular version [Strogatz]) |
| Week 9 | Tue 10/22 | 15. Midterm exam (in class, open book/computer but not open e-mail/texting) | Probability and counting (independent events, conditional probability), statistics (mean, variance), Bayes theorem (medical tests, lie detector, dishonest casino), Binary Classifiers (spam, text), Graphs and graph algorithms |
| | Thu 10/24 | 16. More on Network properties, standard deviation, p-value, and research statistics | Discussed "Unreliable Research", and xkcd on ps#5<br>Discussed some properties of conditional probabilities, recalled some properties of networks (clustering coefficient, degree distributions) for next assignment: ps4.8.ipynb<br>then some more notes on normal distributions and central limit theorem: see lec16.ipynb<br>started discussion of power laws (nytimes city math [Strogatz]) |
| Week 10 | Tue 10/29 | 17. More on power laws, the long tail | More comments on normal distributions and central limit theorem: see lec16.ipynb<br>Discussion of power laws, specifically the "long tail" (see also Wikipedia entry, and later book).<br>see lec17 slides |
| | Thu 10/31 | 18. finish power laws, start big data | generate and fit synthetic power law data: see powerlaw.ipynb,<br>for spell correct, see lec18 slides |
| Week 11 | Tue 11/5 | 19. More on big data algorithms | finished these lec18 slides, discussed The Anatomy of the Facebook Social Graph, and these notes on preferential attachment |
| | Thu 11/7 | 20. Markov Chains | Finished comments on preferential attachment notes, illustrated with this figure (see also pp 556-559 here).<br>Then some Markov chain notes1 and worked example<br>**Note:** Rest of Problem Set 4 due Tue 11 Nov in class |
| Week 12 | Tue 11/12 | 21. More on Markov Models | started with 2nd derivation of waiting time, interpreted multi-step Markov process as matrix multiplication, finished example from previous time, gave genomic example, then text example after "Mark V Shaney" and illustrated here: Markov.ipynb<br>(some other examples generated from trigram probability distributions) |
| | Thu 11/14 | 22. Stationary distributions and Page rank | first some Bernoulli->Normal notes, introduced Hidden Markov Models (see introductory sections from A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition (Rabiner, 1989), sections I, II, and III(b)),<br>and some Markov/PageRank notes |
| Week 13 | Tue 11/19 | 23. More Markov and hidden Markov | Markov/PageRank notes: stationary states, power method, relation to eigenvalues, paths with node omitted to determine expected number of steps, introduced hidden Markov Models, ... |
| | Thu 11/21 | 24. Modeling Markov data | Went over properties of Markov chain for prob 2, and also explained the discriminative procedure in prob 5.<br>Then covered lec24.ipynb: including the simulation of the Markov chains, power method for finding stationary states, random browser with teleportation, simulation of 'slightly crooked casino' HMM expectations for flipping multiple consecutive heads or rolling 6's, and the notes therein for problem 5 |
| Week 14 | Tue 11/26 | 25. HMMs, the Viterbi algorithm | Worked through example of Viterbi by hand,<br>mentioned example from Bursty and Hierarchical Structure in Streams,<br>Described implementation of algorithm in viterbi.ipynb, plus some other examples.<br>Finally, continued through roughly first 1.5 pages of these notes on information theory |
| | Thu 11/28 | Thanksgiving break | |
| Week 15 | Tue 12/3 | 26. Shannon info | Finished these notes on information theory (including code examples at the end), and discussed Problem Set 6 (due 6 Dec).<br>Started Pearson correlation in preparation for recommender systems (next time) |
| | Thu 12/5 | 27. | recommender systems, data science more generally, review for final exam |

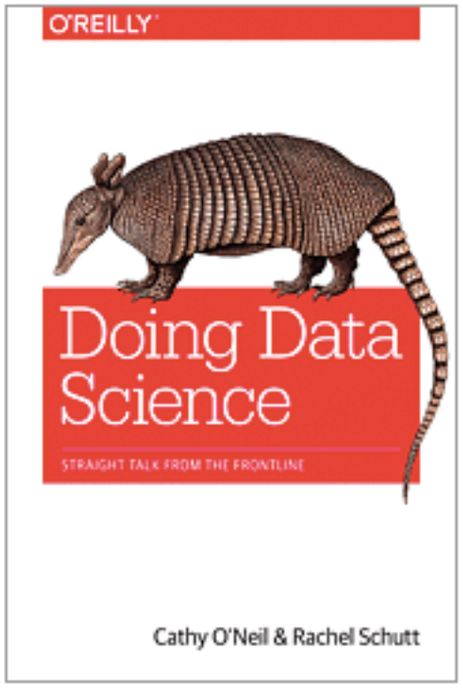
# 13 Dec 2pm-5pm Olin Hall 218

## Final Exam Topics

- Probabilility / Statistics
- Naive Bayes (classifier, inference, ...)
- Graphs, Networks
- Power Law Data
- Markov and other correlated data

Open book, computer, notebook, except email/IM

(Note: likely redo of Midterm problem 3; likely graph statistics; certain Markov)




Larger Cover

# Doing Data Science

## Straight Talk from the Frontline

By Cathy O'Neil, Rachel Schutt

Publisher: O'Reilly Media
Released: October 2013
Pages: 406

5.0

Read 3 Reviews | Write a Review

Now that people are aware that data can make the difference in an election or a business model, data science as an occupation is gaining ground. But how can you get started working in a wide-ranging, interdisciplinary field that's so clouded in hype? This insightful book, based on Columbia University's Introduction to Data...

Full description



Essential Links

# Peer Institutions …

CS109 Data Science

HARVARD
School of Engineering
and Applied Sciences

Home Piazza Syllabus Schedule Homework Readings Projects Resources


Learning from data in order to gain useful predictions and insights. This course introduces methods for five key facets of an investigation: data wrangling, cleaning, and sampling to get a suitable data set; data management to be able to access big data quickly and reliably; exploratory data analysis to generate hypotheses and intuition; prediction based on statistical methods such as regression and classification; and communication of results through visualization, stories, and interpretable summaries.

We will be using Python for all programming assignments and projects. All lectures will be posted here and should be available 24 hours after meeting time.

The course is also listed as AC209, STAT121, and E-109.

## Important Links

- Lecture videos
- Blackboard

1. Introduction: What Is Data Science?
2. Statistical Inference, Exploratory Data Analysis, and the Data Science Process
3. Algorithms
4. Spam Filters, Naive Bayes, and Wrangling
5. Logistic Regression
6. Time Stamps and Financial Modeling
7. Extracting Meaning from Data
8. Recommendation Engines: Building a User-Facing Data Product at Scale
9. Data Visualization and Fraud Detection
10. Social Networks and Data Journalism
11. Causality
12. Epidemiology
13. Lessons Learned from Data Competitions: Data Leakage and Model Evaluation
14. Data Engineering: MapReduce, Pregel, and Hadoop
15. The Students Speak
16. Next-Generation Data Scientists, Hubris, and Ethics
17. Index
18. Colophon

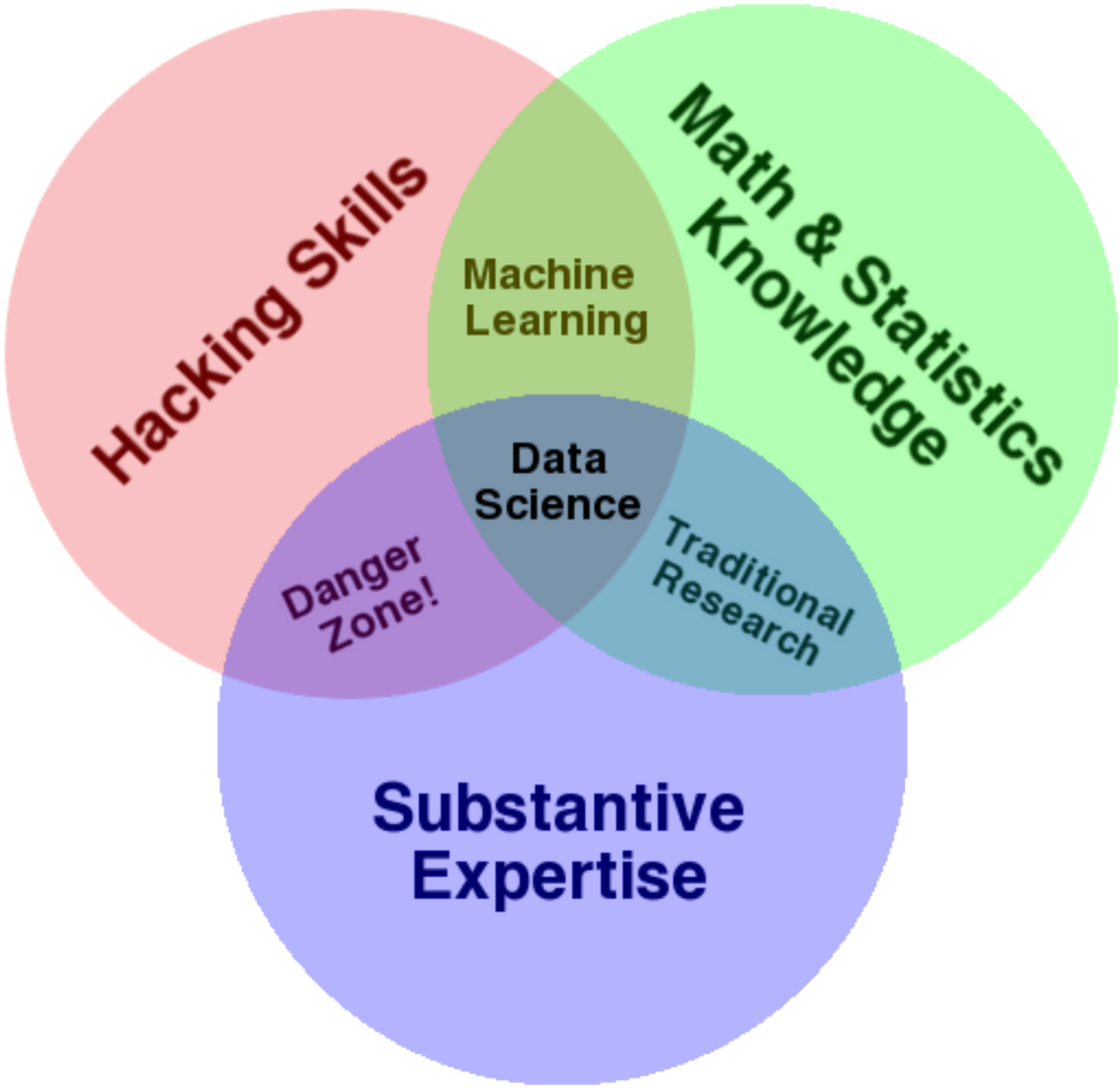
(assumes "prerequisites of linear algebra, some probability and statistics, and some experience coding in any language")

# Some notes from chapt 1 of "Doing Data Science"

Definitions lacking for most basic terminology:

- What is "Big Data"?
- What does "data science" mean?
- What is the relationship between Big Data and data science?
- Is data science the science of Big Data?
- Is data science only the stuff going on in companies like Google and Facebook and tech companies?
- Why do many people refer to Big Data as crossing disciplines (astronomy, finance, tech, etc.) and to data science as only taking place in tech?
- Just how big is big?

(terms so ambiguous, perhaps meaningless)




# Data Science Venn Diagram

(Drew Conway, Sep'10
Phd Pol.Sci. NYU '13)

# Data Scientist

Should be able to identify problems that can be solved with data and be well-versed in the tools of modeling and code

Interdisciplinary teams of people should include a data-savvy, quantitatively minded, coding-literate problem-solver

e.g. at Google: interdisciplinary teams of PhDs: statistician, social scientist, engineer, physicist, and computer scientist.

bring mix of skills: coding, software engineering, statistics, mathematics, machine learning, communication, visualization, exploratory data analysis, data sense, and intuition, plus expertise in social networks and the social space

[Courses in school need not be out of touch with reality ...]

Data Science has roots in many other disciplines:

- statistical inference
- algorithms
- statistical modeling
- machine learning
- experimental design
- optimization
- probability
- artificial intelligence
- data visualization
- exploratory data analysis

# In colloquial terms

Data science is the civil engineering of data,  requires a practical knowledge of tools and materials, coupled with a theoretical understanding of what's possible

- Statistics (traditional analysis familiar to statisticians)
- Data munging (parsing, scraping, and formatting data)
- Visualization (graphs, tools, etc.)

# Why us?  Why Now?

Massive amounts of data collected about many aspects of our lives, plus abundance of inexpensive computing power:
shopping, communicating, reading news, listening to music, searching for information, expressing opinions --- all tracked online

"datafication" of offline behavior has started as well, mirroring the revolution in collection of online data:
an enormous amount to learn about our individual and collective behavior

Not just Internet data, also finance, medical industry, pharmaceuticals, bioinformatics, social welfare, government, education, retail, ... .

A perceived growing influence of data in most sectors and most industries.

In some cases, the amount of data collected might be enough to be considered "big"

Browse the Web: passively (unintentionally) datafied through "cookies" and other tracking devices.

In a store, or on the street, datafied in other unintentional ways, via sensors, cameras, or Google glasses.

NSA?

# The Rise of Big Data

## How It's Changing the Way We Think About the World

*By Kenneth Neil Cukier and Viktor Mayer-Schoenberger*

FROM OUR MAY/JUNE 2013 ISSUE



Everyone knows that the Internet has changed how businesses operate, governments function, and people live. But a new, less visible technological trend is just as transformative: "big data." Big data starts with the fact that there is a lot more information floating around these days than ever before, and it is being put to extraordinary new uses. Big data is distinct from the Internet, although the Web makes it much easier to collect and share data. Big data is about more than just communication: the idea is that we can learn from a large body of information things that we could not comprehend when we used only smaller amounts.

In the third century BC, the Library of Alexandria was believed to house the sum of

# In industry context

The data itself, often in real time, becomes the building blocks of data products.

- On the Internet: Amazon recommendation systems, friend recommendations on Facebook, film and music recommendations, ...
- In finance: credit ratings, trading algorithms, and models
- In education: dynamic personalized learning and assessments (?)
- In government: policies based on data

Technology makes this possible: infrastructure for large-scale data processing, increased memory, and bandwidth, as well as a cultural acceptance of technology in the fabric of our lives.
(Wasn't true a decade ago.)

DJ Patil and Jeff Hammerbacher — then at LinkedIn and Facebook, respectively — coined the term "data scientist" in 2008.
So that is when "data scientist" emerged as a job title.
(Wikipedia finally gained an entry on data science in 2012.)

[But the basic idea also goes back further. In 2001, William Cleveland wrote a position paper about data science called "Data Science: An action plan to expand the field of statistics."]

Chief data scientist sets the data strategy of the company: everything from the engineering and infrastructure for collecting data and logging, to privacy concerns, to deciding what data will be user-facing, how data is going to be used to make decisions, and how built back into the product.

Once the skill set required to thrive at Google — working with a team on problems that required a hybrid skill set of stats and computer science paired with personal characteristics including curiosity and persistence — spread to other Silicon Valley tech companies, it required a new job title.

Once it became a pattern, it deserved a name.

And once it acquired a name, everyone and their mother wanted to be one.

It became even worse when Harvard Business Review declared data scientist to be the "Sexiest Job of the 21st Century" (Oct 2012)