Foundations of Data Science
Lecture 1, Module 1
Fall 2022

Rumi Chunara, PhD

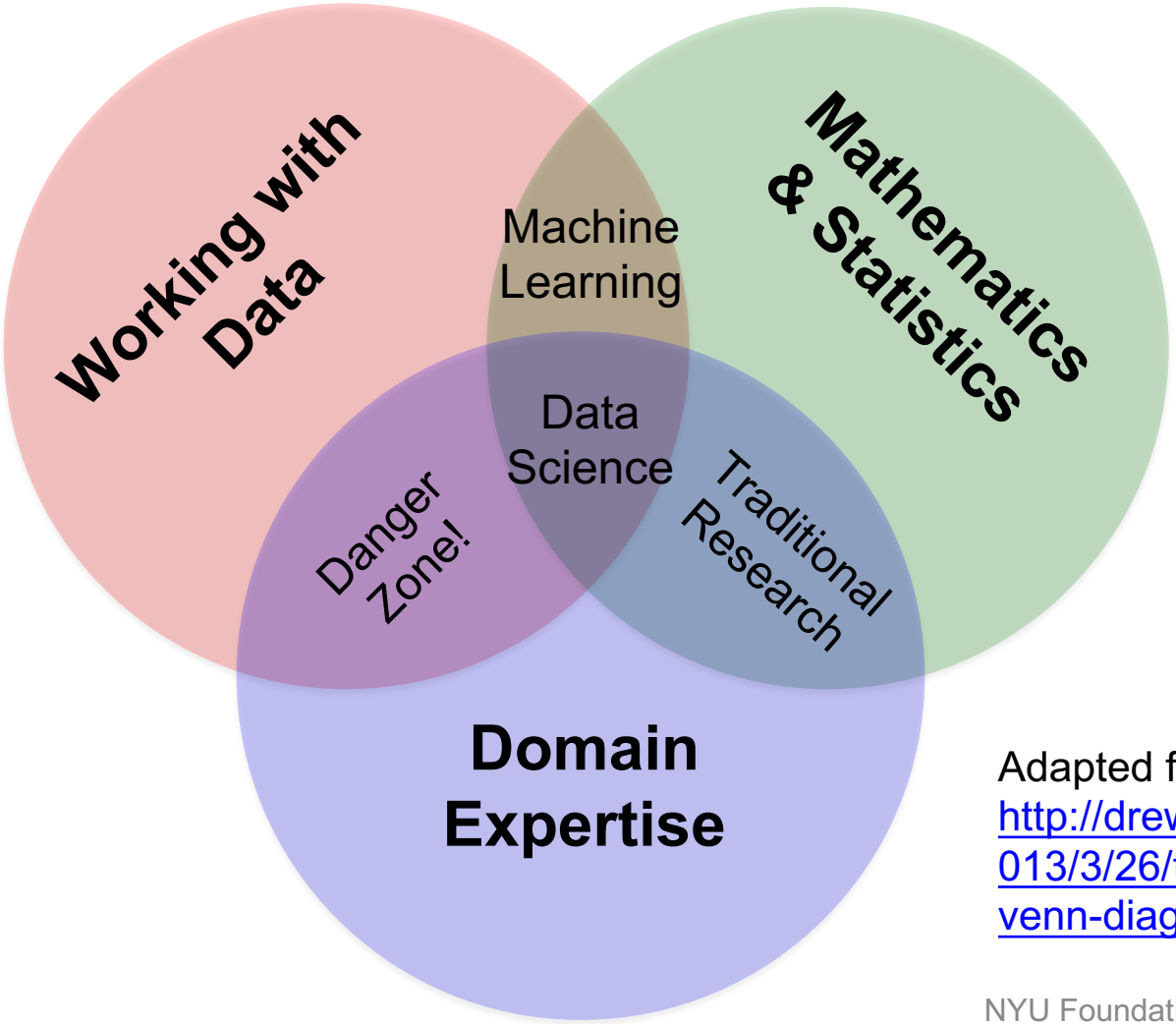
Fine Print: these slides are, and always will be a work in progress. The material presented herein is original, inspired, or borrowed from others' work. Where possible, attribution and acknowledgement will be made to content's original source. Do not distribute without the instructor's permission.

What is Data Science?

WIKIPEDIA
The Free Encyclopedia

*"**Data science** is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data, and apply knowledge and actionable insights from data across a broad range of application domains."*

Data Science – One Definition



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

What is the *Science* of Data Science?

"The key word in 'Data Science' is not Data, it is Science." **- Jeff Leek**

<u>The scientific method:</u> evaluating the merit of a hypothesis with rigorous empirical testing

in other words...

1. Ask a question.
2. Do background research.
3. Construct a hypothesis.
4. Test your hypothesis by doing an experiment.
5. Analyze your data and draw a conclusion.
6. Communicate your results.
7. *Rinse and repeat a few of the steps above as needed.*

...But Data Science is also "Art"

Outside of modeling competitions, well-posed problem or clean dataset are rarely the case...

The art comes in:

- Translating problems into the language of Data Science
- Formulating reasonable hypotheses
- Developing an intuition for good vs. bad data, good vs. bad models.
- Abstracting problems to identify similarities

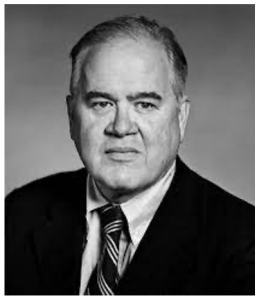
Is Data Science a new field?

https://trends.google.com/trends/explore?date=all&q=data%20science

- *Data analysis* (1960s)
- *Data mining* (1990s)
- *Knowledge discovery* (1990s)

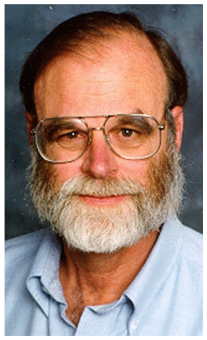
...

Data Analysis Timeline

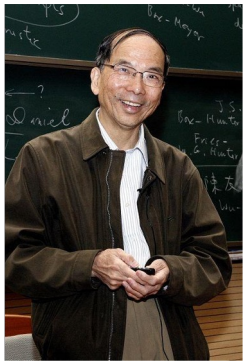


John Tukey
The Future of Data Analysis
(1962)
Exploratory Data Analysis
(1977)

Howard Dresner
Proponent of the term 'Business Intelligence', whose definition is similar to 'Data Science'
(1989)



Jim Gray
Proponent of the 'Fourth Paradigm'
(Book on the topic by Tony Hey - 2007)



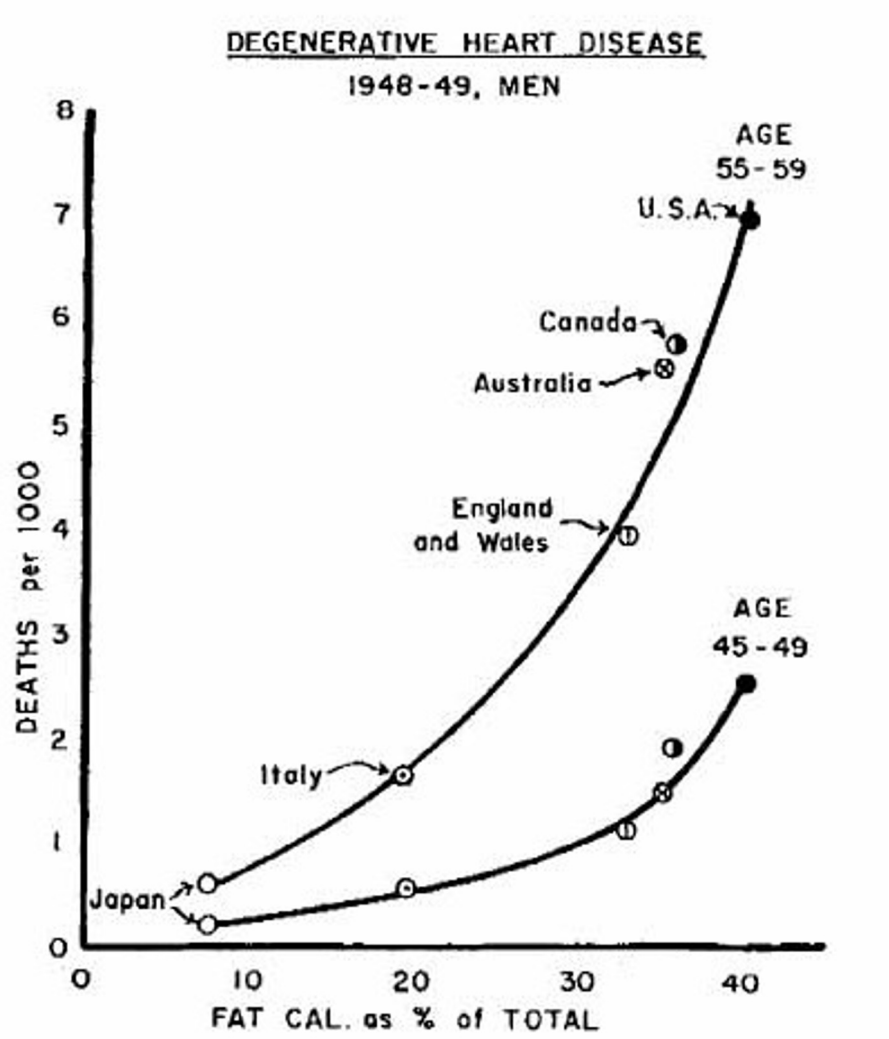
Chien-Fu Jeff Wu
Name 'Data Science' as an alternative to 'Statistics'
(1985)

Leo Breiman
Statistical Modeling: The Two Cultures
(2001)

50 years of datascience recap:
http://courses.csail.mit.edu/18.337/2015/docs/50YearsDataScience.pdf

Data Helps Answer Questions

- ***Does fat correlate with coronary heart disease?***

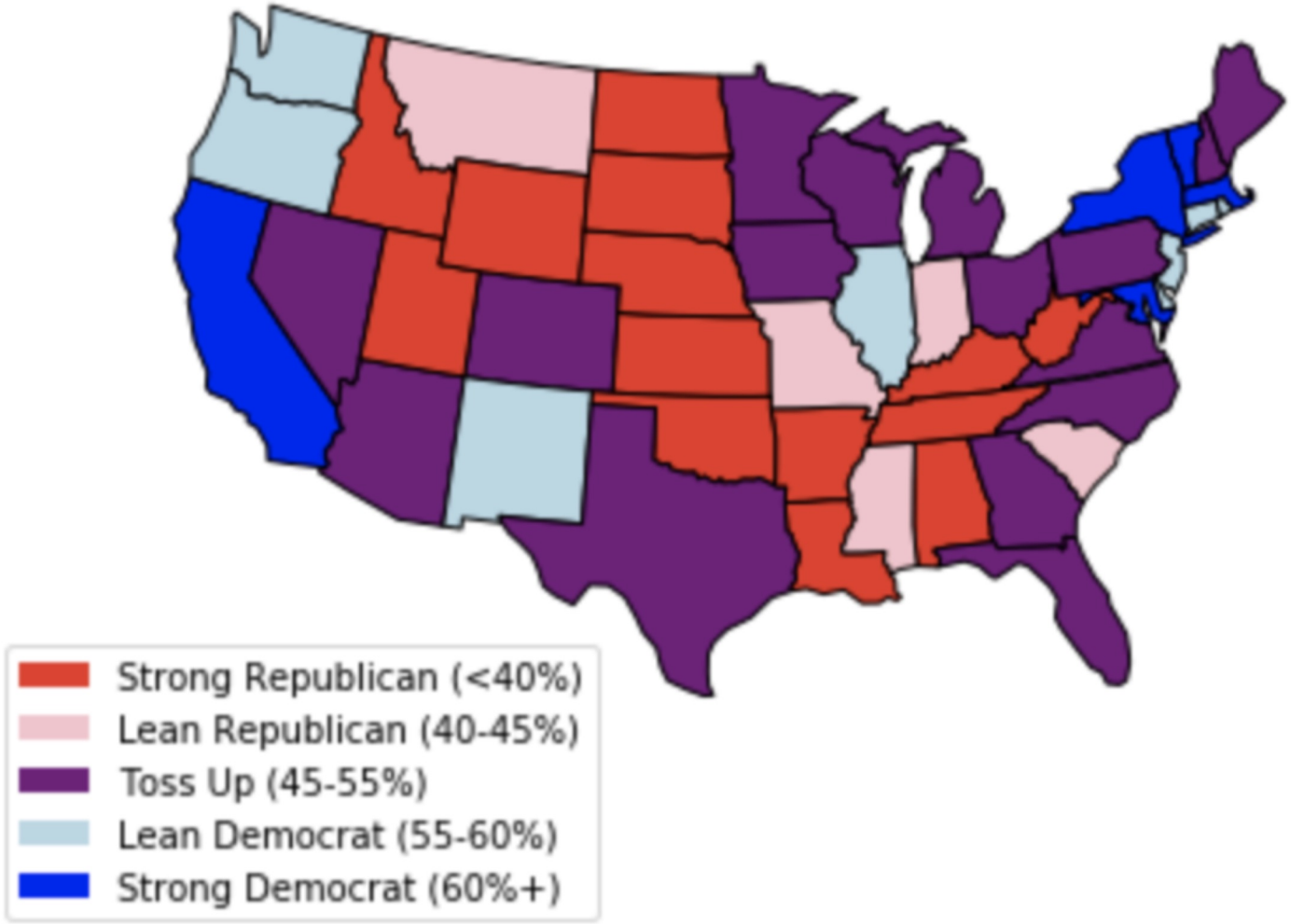


- *Seven Countries Study* (**1950s**)
- 13,000 subjects total, 5-40 years follow-up.

Ancel Keys

Data Helps Answer Questions

- ***Who is going to win the US 2020 elections?***



https://towardsdatascience.com/what-will-happen-in-the-2020-election-e6efb0a7b2ad

Statistical models and advanced analytics were used in the 2016 political advertising campaigns by both Hillary and Trump. Even Barack Obama had a strong data analytics team to support his campaigns back in 2008. The 2020 US elections were no different. Advanced analytics has helped political parties to go beyond just basic polling research. Here are some ways in which Biden's team leveraged analytics for his campaigns:

Adapted from https://www.crunchmetrics.ai/blog/how-joe-bidens-team-leveraged-data-analytics-in-the-us-elections/

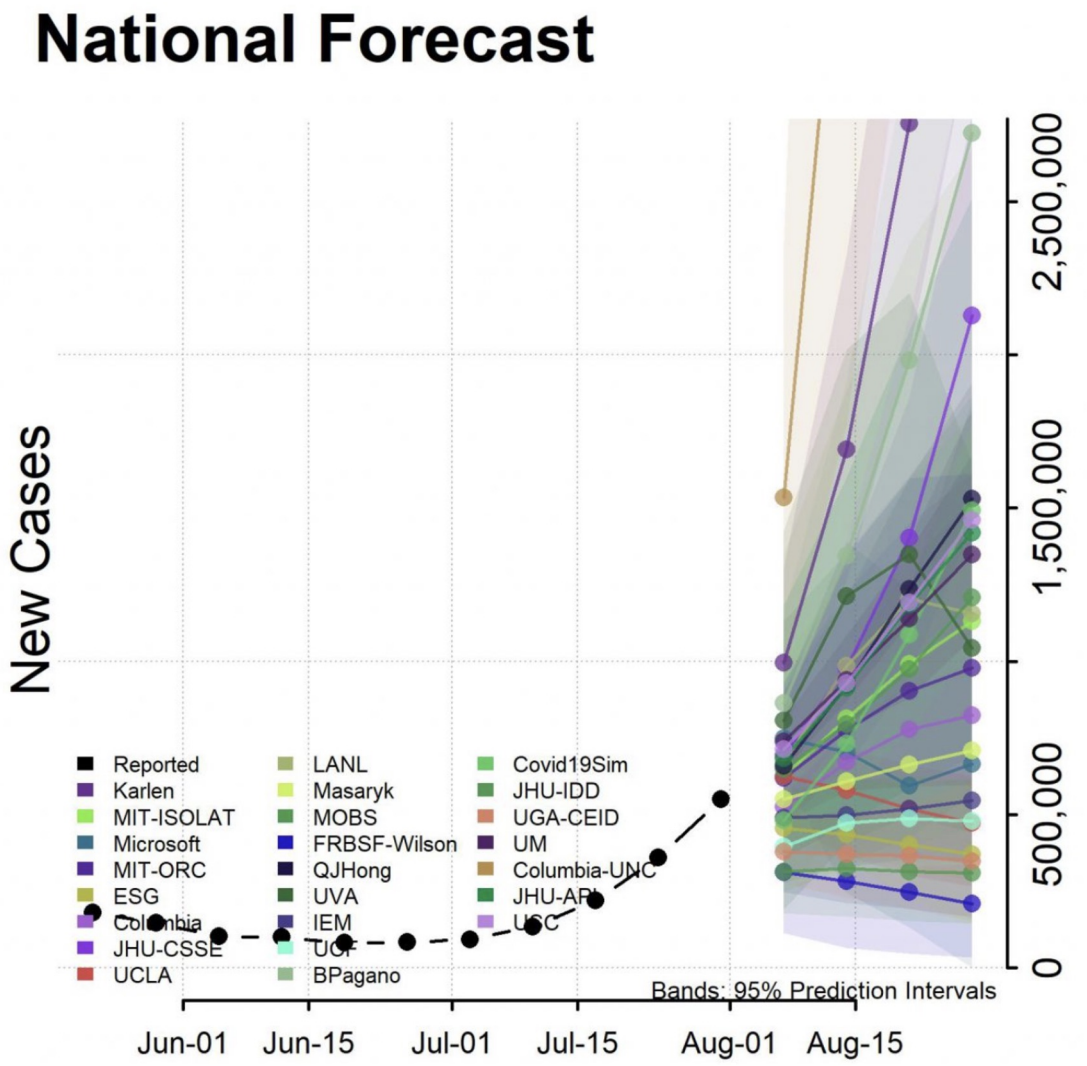
Statistical models and advanced analytics were used in the 2016 political advertising campaigns by both Hillary and Trump. Even Barack Obama had a strong data analytics team to support his campaigns back in 2008. The 2020 US elections were no different. Advanced analytics has helped political parties to go beyond just basic polling research. Here are some ways in which Biden's team leveraged analytics for his campaigns:

1. *Analysis of complex behavioral data to gauge voter sentiment*
2. *Social media sentiment analysis*
3. *Microtargeting to attract floating voters*
4. *Data analytics for digital Ad strategies*

Adapted from https://www.crunchmetrics.ai/blog/how-joe-bidens-team-leveraged-data-analytics-in-the-us-elections/

Data Helps Answer Questions

- ***How many COVID-19 cases will be reported this week?***



https://www.cdc.gov/coronavirus/2019-ncov/science/forecasting/forecasts-cases.html

Different models and model ensembles based on different datasets and using different methodologies.

Actual reported cases in **black**.

Data Helps Answer Questions

- **Descriptive Statistical Questions**
  - *"What are the genetic features of one group?"*
- **Hypothesis Testing**
  - *"Do different fats impact health differently?"*
- **Segmentation/Classification**
  - *"What are the food preferences of customers?"*
- **Prediction**
  - *"On what week will temperatures peak in NYC?"*

Data Science:
Difficulties and Challenges

What is Hard about Data Science

- Overcoming assumptions
- Making ad-hoc explanations of data patterns
- Overgeneralizing
- Communication
- Lack of verification (validate models, data pipeline integrity, etc.)
- Using statistical tests correctly
- Prototypes and production transitions
- Data pipeline complexity (who do you ask?)
- Picking the right questions

More Data Brings New Challenges

- **Data Storage**
  - It may cost more money, require more careful backup and replication policies etc.
- **Data Processing**
  - It may require distributed, parallel, or simply more efficient algorithms that are not as simple to understand or implement
- **Data Filtering**
  - Not all data is relevant
  - Finding “the needle in the haystack” can be hard
- **Data Privacy**
  - Too much public available data makes it hard to obtain truly anonymous data

etc...

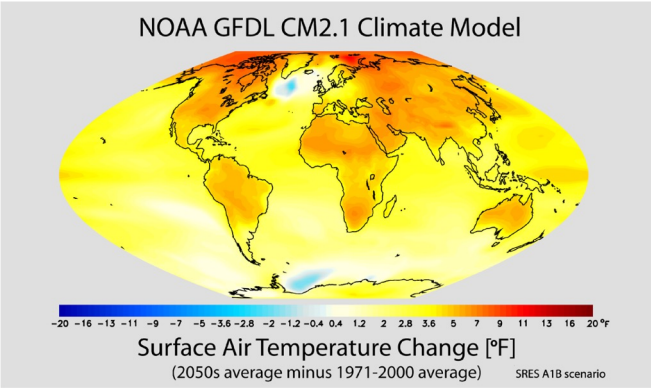
Data Science:
What it is NOT

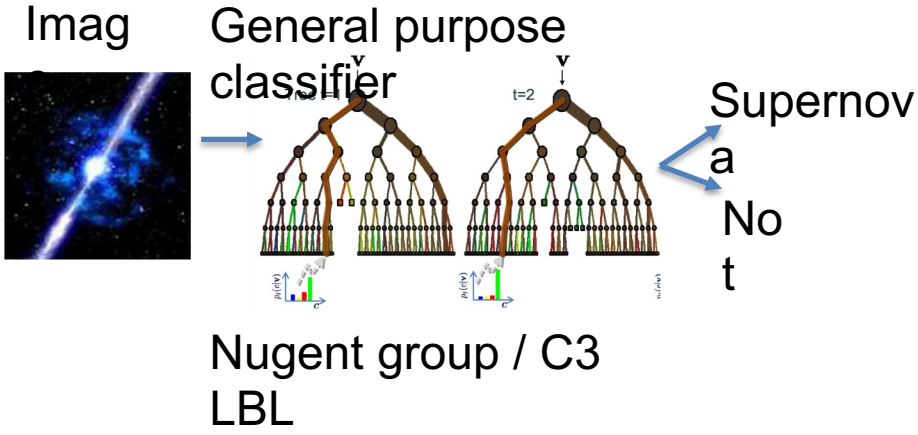
Databases vs. Data Science

	Databases	Data Science
Data Value	“Precious”	“Cheap”
Data Volume	Modest	Massive
Priorities	Consistency, Error recovery, Auditability	Speed, Availability, Query richness
Structured	Strongly (Schema)	Weakly or none (Text)
Properties	Transactions, ACID*	CAP* theorem (2/3), eventual consistency
Realizations	MySQL, PostgreSQ (relational), sometimes NoSQL (non-relational)	Sometimes NoSQL, ElasticSearch, Lucene, Apache River, Apache Hadoop, Apache Spark, Hbase, Cassandra

These are just a few examples of key differences!

Scientific Computing vs. Data Science





Scientific Modeling
Physics-based models
Problem-Structured
Mostly deterministic, precise
Run on Supercomputer or High-end Computing Cluster

Data-Driven Approach
Data and inference engine replaces model
Structure not related to problem
Statistical models handle randomness
Run on cheaper computer Clusters (EC2)

Machine Learning vs. Data Science

Machine Learning	Data Science
Develop new (individual) algorithms	Explore existing algorithms, build and tune hybrids
Prove mathematical properties of algorithms	Understand empirical properties of algorithms
Improve/validate on a few, relatively clean, small datasets	Develop/use tools that can handle massive datasets
(More often) Publish papers	(More often) Build tools, solve practical problems

Doing Data Science

Doing Data Science

- How to learn Data Science
- Why learn Data Science?
- What is a Data Scientist?
- Data Science Workflow

How to Learn Data Science

- Masters programs
  - New Data Science institutes created or repurposed – NYU, Columbia, UW, etc.
  - New degree programs, courses, bootcamps
    - e.g., at Berkeley: Stats, I-School, CS, Astronomy…
- Work in the field
- Learn on your own with online resources
  - Many books available online
  - Kaggle competitions

Some recent Data Science competitions

kaggle

Results

chaii - Hindi and Tamil Question Answering
Identify the answer to questions found in Indian language passages
Research · Code Competition · 3 months to go

Google Landmark Retrieval 2021
Given an image, can you find all of the same landmarks in a dataset?
Research · Code Competition · 2 months to go

Google Landmark Recognition 2021
Label famous, and not-so-famous, landmarks in images
Research · Code Competition · 2 months to go

NFL Health & Safety - Helmet Assignment
Segment and label helmets in video footage
Featured · Code Competition · 3 months to go

LearnPlatform COVID-19 Impact on Digital Learning
Use digital learning data to analyze the impact of COVID-19 on student learning
Analytics · 2 months to go

Tabular Playground Series - Aug 2021
Practice your ML skills on this approachable dataset!
Playground · 20 days to go

RSNA-MICCAI Brain Tumor Radiogenomic Classification
Predict the status of a genetic biomarker important for brain cancer treatment
Featured · Code Competition · 2 months to go

Why Learn Data Science?

- Growing demand for data specialists
  - News, marketing, and transportation companies
  - Hospitals
  - City agencies
  - Agriculture

Even restaurants could use Data Science!

Data Science in Restaurants: Best Practices and Benefits

Restaurants can—and should—be collecting and analyzing customer data in real-time.

OUTSIDE INSIGHTS | AUGUST 2019 | TONY TONG

Why Learn Data Science?

- *"… the sexy job in the next 10 years will be statisticians."* - Hal Varian, Google Chief Economist (2012)

- *"The U.S. will need 140,000-190,000 predictive analysts and 1.5 million managers/analysts by 2018."* - McKinsey Global Institute (2011)

What is a Data Scientist?

Zvi
@nivertech

Follow

"Data Scientist" is a Data Analyst who lives in California.

Reply Retweet Favorite More

RETWEETS	FAVORITES
140	40

9:55 PM - 14 Mar 2012

Josh Wills
@josh_wills

Follow

Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician.

Reply Retweet Favorite More

RETWEETS 907 FAVORITES 418

12:55 PM - 3 May 2012

Javier Nogales
@fjnogales

Follow

Data Scientist (2/2): person who is worse at statistics than any statistician and worse at software engineering than any software engineer

RETWEET 1

FAVORITES 5

9:08 AM - 27 Jan 2014

The resume: skills

Through the lens of a Data Scientist Job Description

PhD is a proxy for:
- **experience**
- **research ability**
- **technical expertise**

Requirements

- Ph.D. in a relevant technical field, or 4+ years experience in a
- Comfort manipulating and analyzing complex, high-volume, high-dimensionality data from varying sources
- Ability to communicate complex quantitative analysis and analytic approaches in a clear, precise, and actionable manner
- Fluency with scripting languages such as Python, Ruby, or PHP
- Familiarity with relational databases and SQL-like query languages
- Expert knowledge of a scientific computing language such as R, Python, or Julia
- Experience working with data-distributed query tools a plus (Hadoop, Hive, Presto, etc.)

The resume: skills

Through the lens of a Data Scientist Job Description

Requirements

- Ph.D. in a relevant technical field, or 4+ years experience in a
- Comfort manipulating and analyzing complex, high-volume, high-dimensionality data from varying sources
- Ability to communicate complex quantitative analysis and analytic approaches in a clear, precise, and actionable manner
- Fluency with scripting languages such as Python, Ruby, or PHP
- Familiarity with relational databases and SQL-like query languages
- Expert knowledge of a scientific computing language such as R, Python, or Julia
- Experience working with data-distributed query tools a plus (Hadoop, Hive, Presto, etc.)

… more recently, people from other fields (journalism, neuroscience etc) are also becoming data scientists (MSc, work experience etc).

The resume: skills

Through the lens of a Data Scientist Job Description

Requirements

- Ph.D. in a relevant technical field, or 4+ years experience in a rel
- Comfort manipulating and analyzing complex, high-volume, high-dimensionality data from varying sources
- Ability to communicate complex quantitative analysis and analytic approaches in a clear, precise, and actionable manner
- Fluency with scripting languages such as Python, Ruby, or PHP
- Familiarity with relational databases and SQL-like query languages
- Expert knowledge of a scientific computing language such as R, Python, or Julia
- Experience working with data-distributed query tools a plus (Hadoop, Hive, Presto, etc.)

You can't be a Data Scientist if you can't handle data…

The resume: skills

Through the lens of a Data Scientist Job Description

Requirements

- Ph.D. in a relevant technical field, or 4+ years experience in a relevant role
- Comfort manipulating and analyzing complex, high-volume, high-dimensionality data from varying sources
- Ability to communicate complex quantitative analysis and analytic approaches in a clear, precise, and actionable manner
- Fluency with scripting languages such as Python, Ruby, or PHP
- Familiarity with relational databases and SQL-like query languages
- Expert knowledge of a scientific computing language such as R, Python, or Julia
- Experience working with data-distributed query tools a plus (Hadoop, Hive, Presto, etc.)

This is essentially the goal of this course

The hard skills

Through the lens of a Data Scientist Job Description

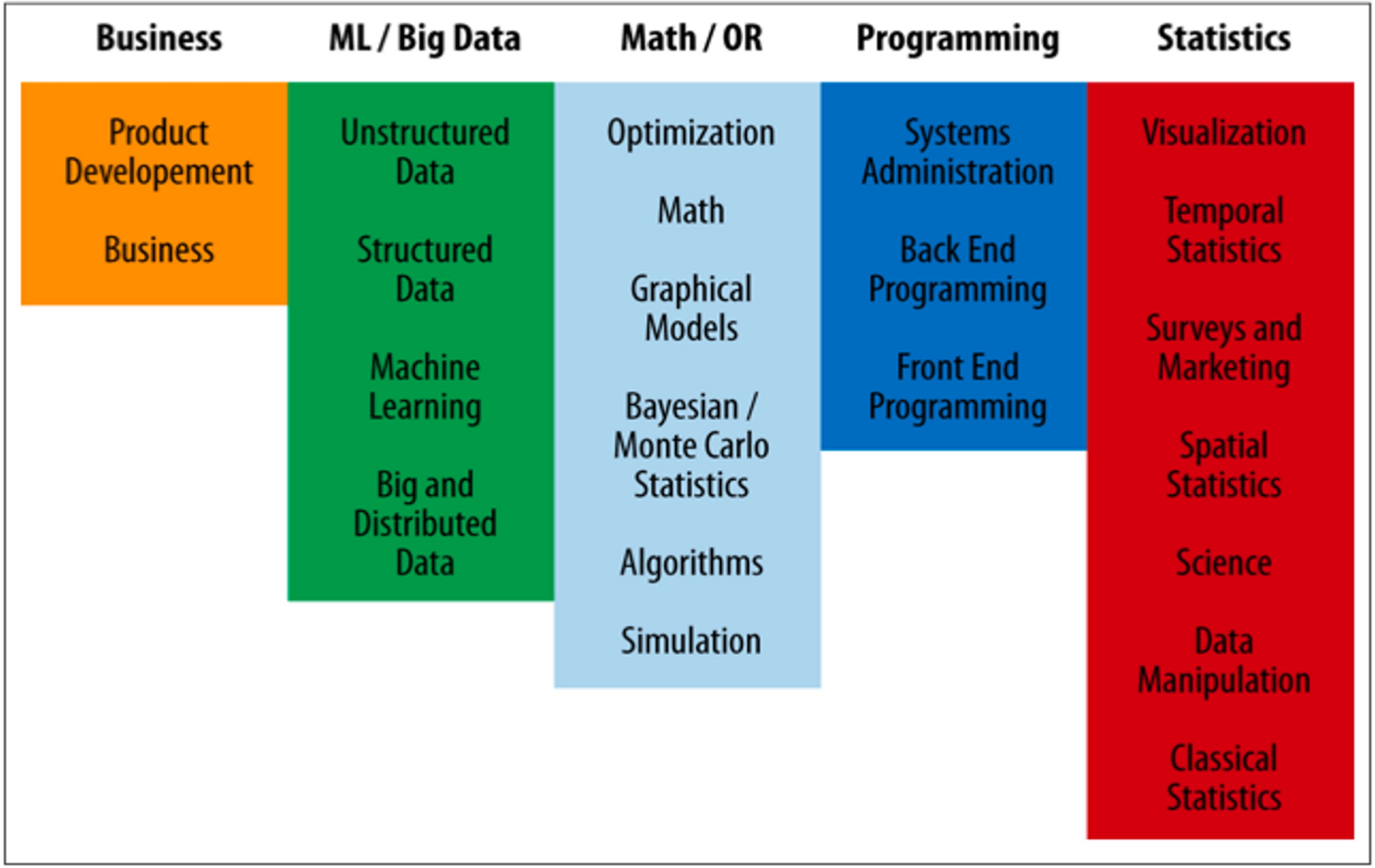
Requirements

- Ph.D. in a relevant technical field, or 4+ years experience in a relevant role
- Fluency with scripting languages such as Python, Ruby, or PHP
- Familiarity with relational databases and SQL-like query languages
- Expert knowledge of a scientific computing language such as R, Python, or Julia
- Experience working with data-distributed query tools a plus (Hadoop, Hive, Presto, etc.)

Important: a scripting language, SQL and a scientific computing language. You will get hands-on experience with some of this in this course, and you should *definitely* develop these skills further outside of this course

Range of Data Science Skills

- **Note that they sometimes intersect**



“Data Scientists are people with some mix of **coding and statistical skills** who work on **making data useful** in various ways.”

Data Scientist Type A (for Analysis):

- Primarily concerned with **making sense of data** or working with it in a fairly **static** way.

- Similar to a statistician, but knows all the **practical details of working with data** that aren't taught in statistics:
  - data cleaning
  - dealing with large data sets
  - visualization
  - domain knowledge

https://www.quora.com/What-is-data-science/answer/Michael-Hochster

“Data Scientists are people with some mix of **coding and statistical skills** who work on **making data useful** in various ways.”

Data Scientist Type B (for Building):

- Some statistical background, but **strong coder or software engineer**.

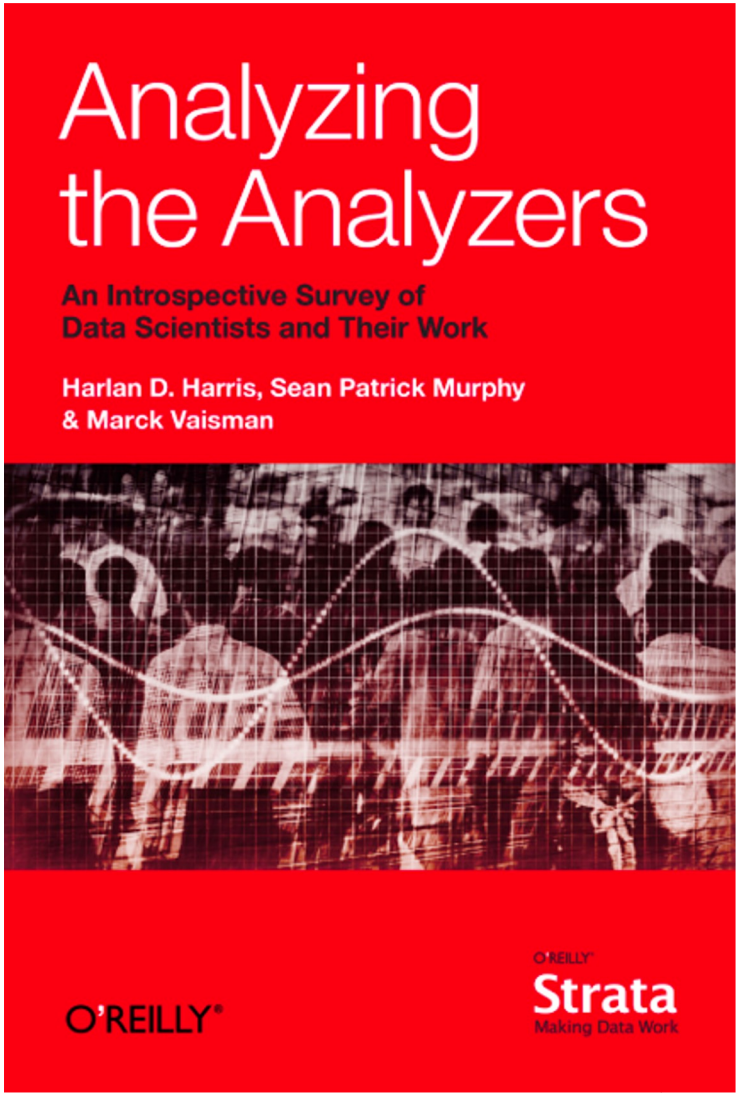
- Primarily concerned with **using data “in production”**:
  - building models that interact with users (by giving recommendations, for example).

Our course is focused primarily on **Type A**.

Data to Define Data Science!

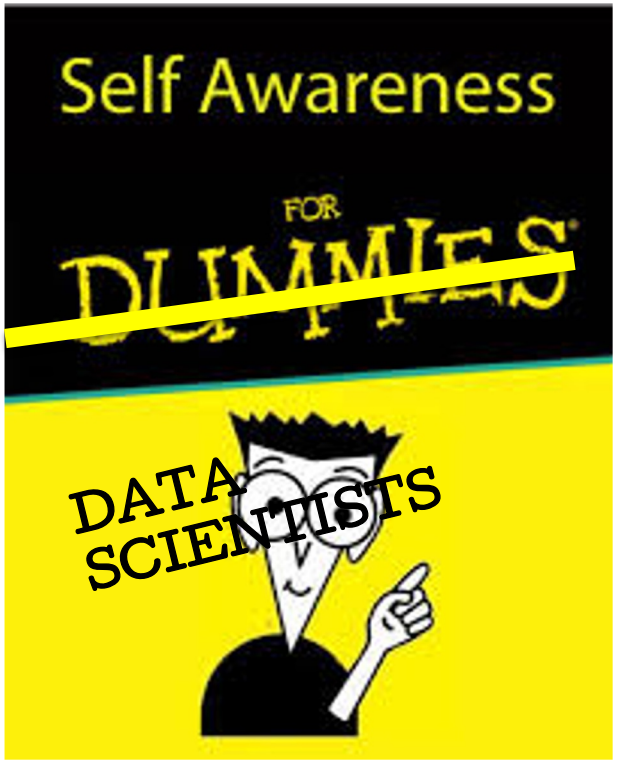
There is no ‘one-size-fits-all’ type of data scientist.

Luckily, people are using data science to define data science.



Figuring Out Your Interests

- **You don't have to fit into one bucket, but you should know where you stand and what you like**



- Personal skills development
- Choosing the right job (your future boss might not know what a data scientist is, or should be)

The Field Today

- Many mature off-the shelf tools for analysis exist
- Skills in **problem-solving**, **collaborating** and **communicating** are needed

What We will Learn

With this course we want to emphasize the *soft* skills of data science:

- *Art*: Abstract and intuitive thinking
- *Science*: process

We'll cover necessary Data Science tools, but with the goal of applying them towards analytic problem solving.

Data Science Workflow

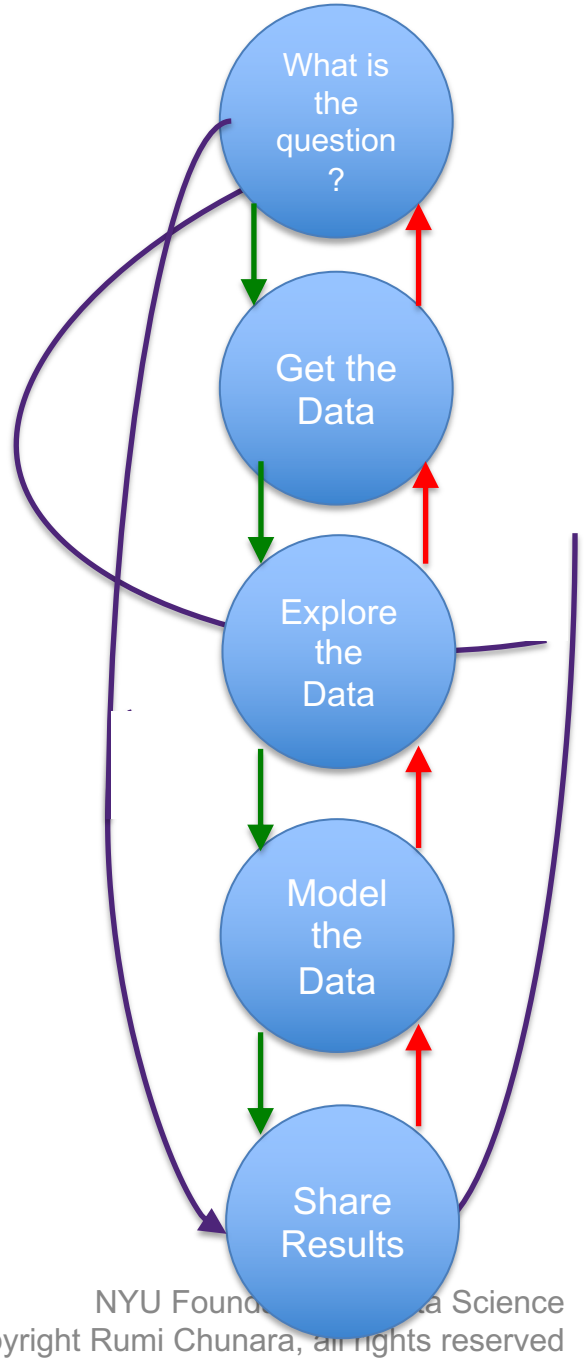
What is the scientific **goal**?
What would you do if you had all the **data**?
What do you want to **predict** or **estimate**?

How was the data **sampled**?
Which data is **relevant**?
Are there **privacy** issues?

Visualize the data
Are there **anomalies**?
Are there **patterns**?

Build a model
Fit the model
Validate the model

What did we **learn**?
Do the results make **sense**?
Can we tell a **story**?



Predicting Neonatal Infection

Problem: Children born prematurely are at high risk of developing infections, many of which are not detected until after the baby is sick

Goal: Detect subtle patterns and features in the data that predict infection before it occurs

Data: 16 vital signs such as heart rate, respiration rate, blood pressure, etc…

Impact: Model is able to predict the onset of infection 24 hours before the traditional symptoms of infection appear

Case Study: http://www.amazon.com/Big-Data-Revolution-Transform-Think/dp/0544002695