Agenda

1. What is Data Science
2. How does Data Science work (Theory)
3. What do Data Scientists actually do (Practice)
4. Lessons Learned

What is Data Science?

A Statistician that Lives in San Francisco?

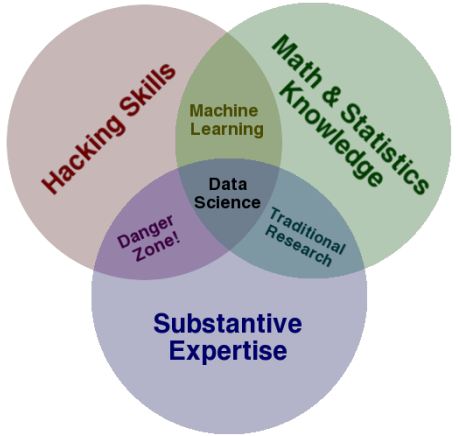
Data science is a young and rapidly evolving field

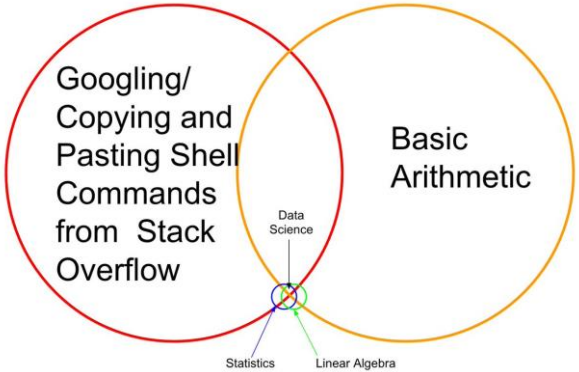
Lots of jargon and hype

- Big data, deep learning, predictive analytics, artificial intelligence

Lots of definitions

- “By ‘Data Science’ we mean almost everything that has something to do with data.” – Journal of 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.” – Wikipedia
- “A data scientist is someone who is better at statistics than any software engineer and better at software engineering than any statistician” – Josh Wills of Cloudera

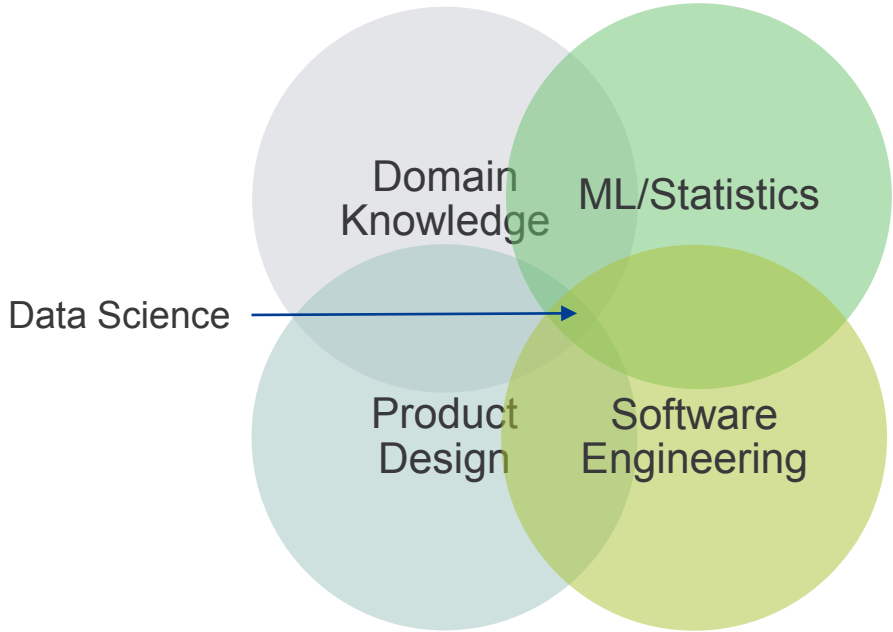




Building a Working Definition of Data Science

In my experience, data science usually involves a combination of two things:

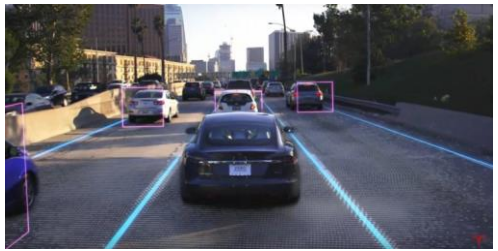
1. Deriving **new insight** from **complex and messy** data
2. The **design, development, and deployment** of useful **data products**



Some Examples of Data Science

Netflix: Personalized movie recommendations developed by analyzing what movies you watch

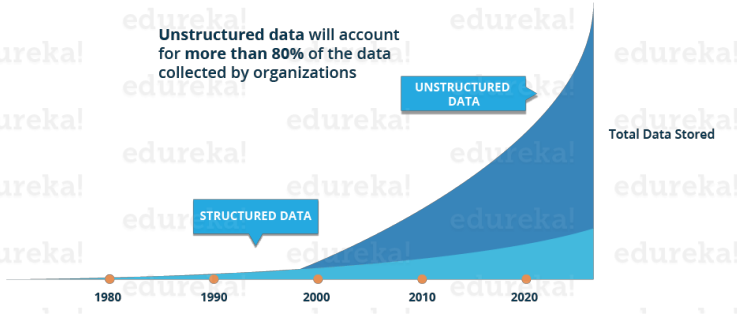
Tesla: Self driving technology developed by using driving telemetry data from vehicles on the road

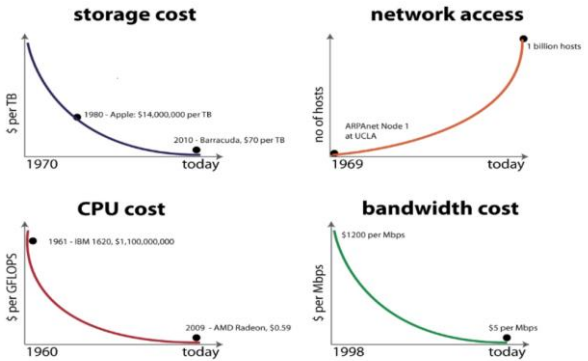


John Deere: Smarter farms enabled by crowd sourced information about seeds, climate, soil type, harvest technique, etc.

Why Now?

- Enterprises are accumulating massive quantities of complex and messy (unstructured) data about how customers are using their products
  - Support tickets, product telemetry, product reviews, etc.
- This data contains useful knowledge that can be applied to core business activities e.g. product development, customer support, and marketing
  - Key product failure modes
  - Best practices
  - Product usage
- Advances in computing, storage, networking, and machine learning have made it possible to algorithmically extract this knowledge at scale





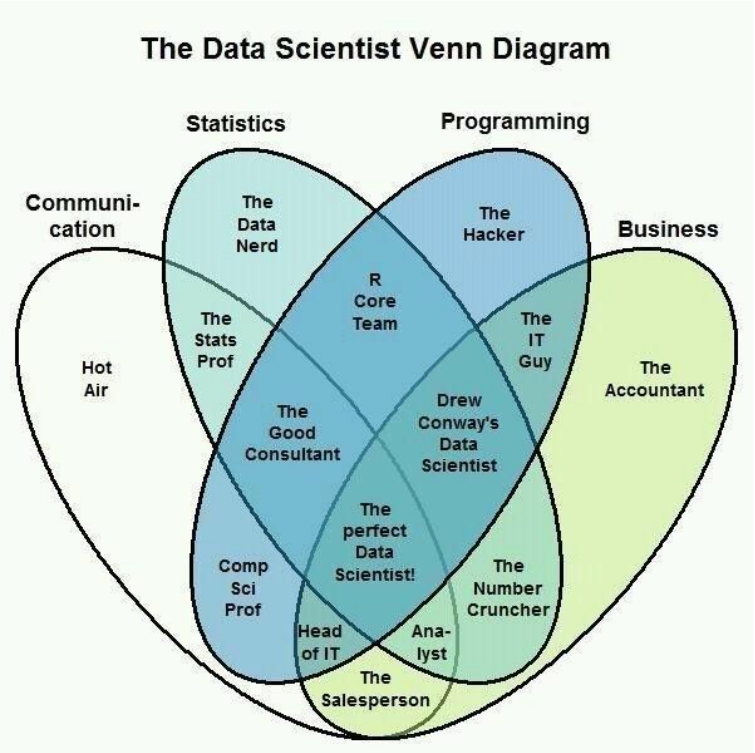
Who are these Data Wizards?

Data Scientists need to be able to single-handedly do at least prototype-level versions of all the steps needed to derive new insights or build data products

Requires a T shaped skill-set

- **Breadth**: Software Engineering, Business Acumen, Communication
- **Depth**: Machine learning/Statistics, Domain Knowledge

Breadth generally comes from experience



Why Should You Care?

Data scientist is currently ranked the best job in America, 4 years running (Glassdoor)

- High salaries
- High demand
- Interesting and interdisciplinary problems
- “Sexiest Job of the 21st century” (Harvard Business Review)

Even if you are not a data scientist your career will benefit from some level of data fluency

How Does Data Science Work?

Data Science 101

Two basic kinds of tasks:

1. **Descriptive:** Discovering useful patterns in given input data X (Unsupervised Learning)
   - Clustering Analysis (groupings)
   - Association Analysis (co-occurrence)
2. **Predictive**: Build a model that can predict output Y given input X based on training data (X,Y) (Supervised Learning)
   - Prediction (numerical Y)
   - Classification (categorical Y)

A Simple Example

Problem: Highway 17 is extremely dangerous, especially when the road is wet

1. Which groups of drivers are most likely to have an accident?

   We can answer this **descriptive task** using cluster analysis

   Methods: K-means, hierarchical clustering, DB-Scan

2. Is driver X likely to have an accident on a particular day?

   We can answer this **predictive task** using a classification model

   Methods: Naïve Bayes, Logistic Regression, Support Vector Machines, Decision Trees

The Data Science Process

Is driver X likely to have an accident on a particular day?

1. Problem Definition / Requirements
2. Data Collection
3. Data Pre-processing
4. Model Development
5. Model Evaluation
6. Model Deployment

Theory

Data Collection

Data-set: collection of instances (rows) across a set of attributes or features (columns)

- Categorical
- Numerical
- Text

Each instance represented as a feature vector and a target (X, Y)

$$X = (x_1, x_2, x_3, \dots, x_n)$$
$$Y = Accident$$

Non-numerical attributes must be encoded to numerical values

Attributes

Instances

Drive ID	Rainfall	Time of Day	Model	Make	Car HP	Accident	Notes
1	1.1”	3:00pm	Toyota	Camry	150.00	No	
2	0.1’	8:00am	Fordd	Mustang	400	Yes	“Flipped over center divider”
…							
10,000	3"	18:00	Subaru	WRX	250	No	

Data Pre-Processing

Real world data is often incomplete, noisy, and inconsistent

Four basic steps in cleaning:

1. **Data Cleaning**: filling in missing data, smoothing, removing outliers, removing duplicates
2. **Data Integration**: joining multiple data-sets, entity recognition
3. **Data Transformation**: normalization, binning
4. **Data Reduction**: attribute selection, dimensionality reduction, sampling

Drive ID	Rainfall	Time of Day	Model	Make	Car HP	Accident	Notes
1	1.1”	3:00pm	Toyota	Camry	150.00	No	
2	0.1’	8:00am	Fordd	Mustang	400	Yes	“Flipped over center divider”
…							
10,000	3"	18:00	Subaru	WRX	250	No	

Model Development

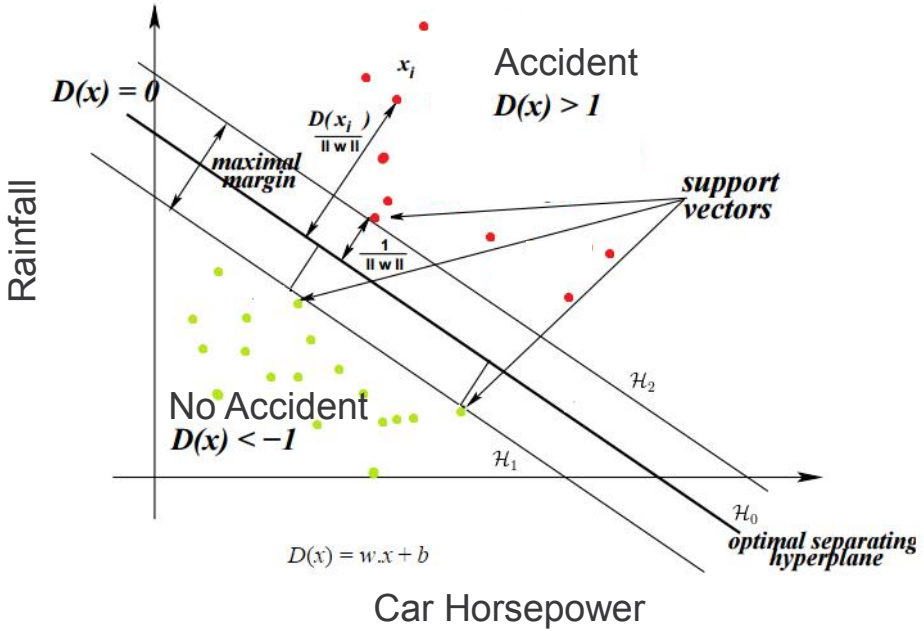
Create a model that can map input attributes (X) to target attribute (Y)

Lots of different ways to learn this mapping mathematically

- **Probability**: Naïve Bayes, Bayesian Networks
- **Hyperplane**: Logistic Regression, SVM, Neural Networks
- **Entropy**: J48, Random Forests

Selecting the best model is typically an iterative process that involves building and evaluating many different models

Support Vector Machines:
Maximum Margin Hyperplane



Model Evaluation

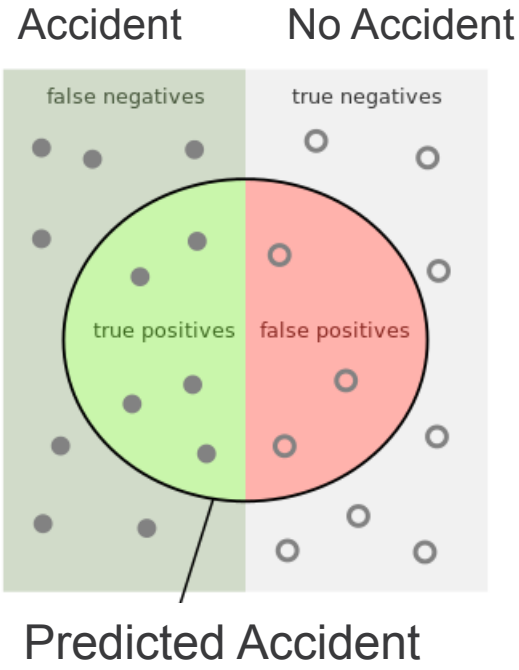
Accuracy is deceiving, e.g. we can get 99% accuracy by predicting no cars will have accidents

Two key ideas when evaluating classification models:

- **Recall**: percent of accidents predicted correctly, e.g. out of 100 accidents we predicted 70
- **Precision**: percent of correct predictions, e.g. out of 100 predictions we got 70 correct

Most data scientist use a combination of precision and recall when evaluating a model (F-Measure)

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

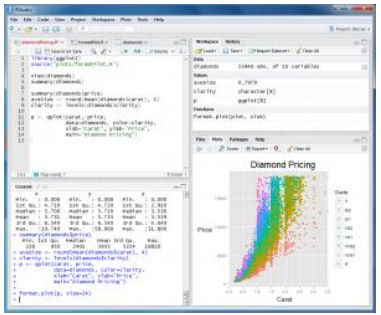


Data Science Toolkit

 mattturck.com/bigdata2018

FIRSTMARK
EARLY STAGE VENTURE CAPITAL

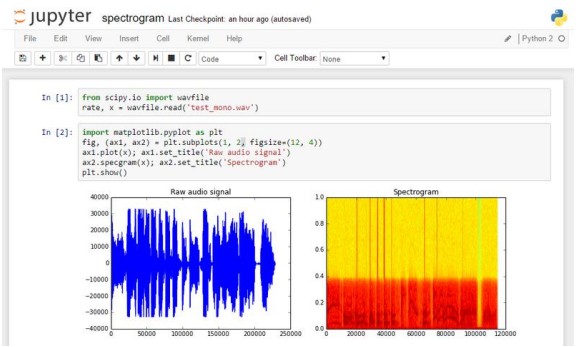
Data Science Toolkit - Simplified



Two main eco-systems:

- **R**: statistical programming language (Rstudio)
- **Python**: programming language with statistical and machine learning packages (Jupyter Notebooks)

They both have strength and weaknesses – learn them both



What Do Data Scientists Actually Do?

Life as a Data Scientist

Typically work on 2-3 projects at a time, each project in a different life-cycle (requirements, modeling, deployment, etc.)

Generally, only a small percentage (5-10%) of the time is actually spent doing the actual model development

Most of the time is spent:

1. Translating the business problem into a data science problem (this is the hard part of data science)
2. Cleaning up data (this is tedious part of data science)
3. Developing and deploying the data product (this is the frustrating, and rewarding, part of data science)

Lessons Learned

10 Years of Data Science Projects

Being a data scientist means being flexible, open minded, and ready to solve problems and embrace complexity

Solving open-ended problems is a mix of art and science

- **Art**: Data representation, encoding, alert thresholds – data scientists need to be comfortable in the domain in order to make these subjective decisions
- **Science**: Which learning algorithm to use, tuning the model to match user expectations, communicating model performance – data scientists need to understand the assumptions behind different machine learning approaches and the trade-offs involved

The machine learning component is generally a relatively small part of the solution, know when a model is “good-enough”

Important to engage end-users (stakeholders) throughout the entire process – users are much more likely to use a data product that they help create

How to Get Started

What I look for when hiring data scientists:

1. Strong analytic background and skills (machine learning, statistics, etc.)
2. Comfortable working with real-world data
3. Basic programming skills (Python, R, etc.)
4. Ability to perform independent research
5. Able to communicate ideas
6. Enthusiastic and excited about their work (capable of evangelizing their ideas)

Q&A