Lecture 1: Introduction

Welcome to Data Science

Prof: Glen Berseth

6nd of September 2022

(some of the material inspired from slides from Zico Kolter, Golnoosh Faradi, Kris Sankaran and Jhelum Chakravorty)

Outline

1. Course Logistics and Safety
2. What is data science?
   a. Go through an example of a data science problem
3. Takeaways from this course
4. Resources

Course Logistic

First Step: Webpage of the course

- Static page of the course:
    - https://fracturedplane.notion.site/Data-Science-Course-IFT6758B-3cf2090ca067441f9325aad99f062646
- Prerequisite page:
    - https://fracturedplane.notion.site/Prerequisite-page-f5affd9c556d4260bf13439867c669e6
- Course Schedule
    - https://fracturedplane.notion.site/b2b681f229c945c0b9aa87849ac8f7b5?v=fa3b9d2f96a442aea454192ac7c233f4

Second Step: Piazza

- **Discussion and Questions** on Piazza [link here].

  https://piazza.com/class/l7gssyk5qh8n7

- **We will not use Studium!**
- Make sure you are registered to Piazza! Otherwise fill [this form]
- Make sure to check frequently Piazza.

Third Step: Survey

- Fill the survey [link here]
- **Super important!**
  - This will help us focus the course to your background.
- I will leave you more time for the break to fill the form!

Discussion and Questions

- Ask most of your questions on Piazza!
- In case of emergency (personal matter):
  - **Head TA:** Pravish Sainath --- pravishsainath@gmail.com
  - **Head of project TA:** David Dobre --- david.dobre@umontreal.ca
  - **TAs:**
    - Arka Mukherjee --- arka.mukherjee.1@umontreal.ca
    - Pavithra Parthasarathy-Rajasekar -- pavithra.parthasarathy@umontreal.ca
    - Yifan Bai -- yifan.bai@umontreal.ca

Course Logistic: Gradescope

- Weekly quizzes on **Gradescope**.
- Coding homework will be evaluated on Gradescope.
- Quizz 0 next week! (to practice how to submit answers)

Course Logistic

The instructor's lectures are on

- Tuesday at 11:30-12:20 ET -- Pavillon Marguerite-d'Youville 2010_559A
- Thursday at 16:30-18:20 ET (Including a 10 mins break btw 17:20 and 17:30) -- Pav. 3200 Jean Brillant Room B-2305

Weekly labs led by a TA on

- Tuesday at 12:30 -14:20 ET (Location TBA)

Coding homeworks will be based on the labs!

Course Logistics: Grading

- Final Project (3 milestones) **(35%)**
- Midterm + Final (written exams)
  - Midterm: Week of October 25th **(15%)**
  - Final December 14th (4:00 to 7:00) **(25%)**
- Weekly programing Homeworks (Gradescope)-- 10 of them 2.5% each -- **(25%)**

The Final Project

Let’s Break the ice!

The Final Project

- NHL data publicly available.
- Getting the data from an API.
- Some data visualization.
- Some basic ML stuffs.

Challenge: Non-standard domain -> real world data

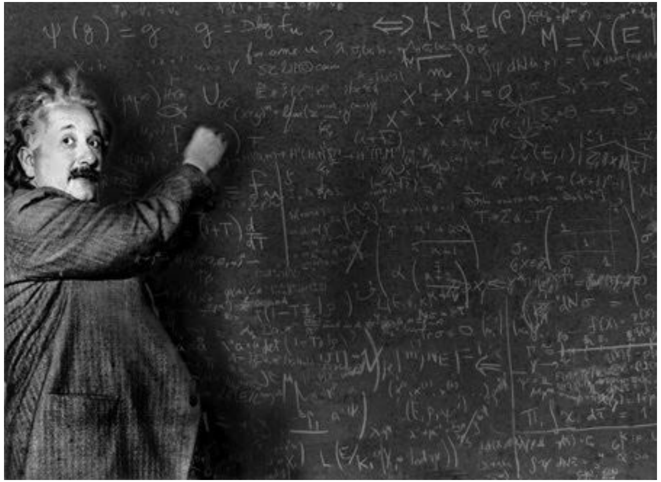
Goal: Production like pipeline

What is Data Science?

Data Science is not pure Machine Learing (IFT6390)

Pure ML is like Pure Physics

Asking: What is the grand unified theory of the universe?



Data Science is like engineering

Asking: How can we build a robot that looks like Einstein robot

It is not pure ML

The empirical side of ML

- Not about pure theory
- How to process data
- How to scale up our models
- How to make good use of our hardware

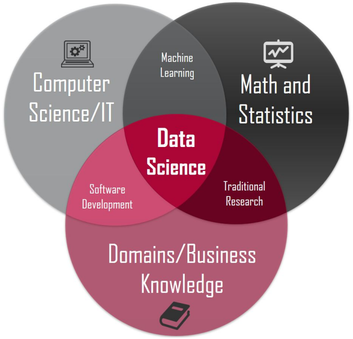
- **About the science of answering questions with data**

A possible Definition of Data Science

> “Data science is the application of **computational** and **statistical** techniques to address or gain insight into some problem in the **real world**”

Source: Zico Kolter’s course on Data Science

How do we get insight from data?



Start with the **Data**

- What data is available, do we need more, **pre-processing data**,

Analyze the data

- Which algorithm/model, **how do I know it is working**, how to improve it

Scale the **System**

- How to use hardware, **distribute experiments**, version control, avoid errors

Need an example

Go through a basic example as part of the first lecture

Food ordering presentation

Start with a fixed time series

What to predict if a person ate pancakes today

Via recorded mouse movements

Can use many different methods to process the data including unsupervised

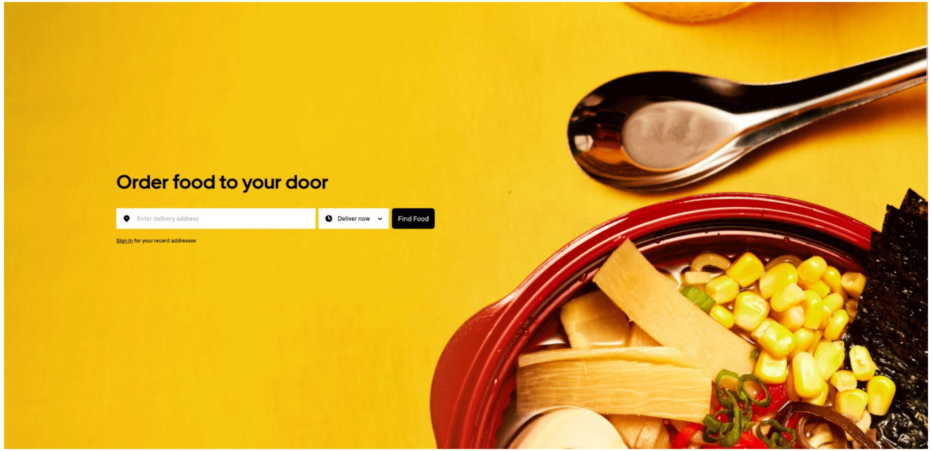
The target task can also change to predict if they will eat pancakes

Pretend the data is from one of those food ordering websites, Uber eats, std, etc. Then we have what the person ended up ordering.

Now we can even add location information some of the time.

Food ordering problem

- What food should we suggest to the client?
- **Data**:
  - Mouse movement
  - Searches
- **Analysis**:
  - What should we predict?
  - How can we train a model to predict it?
  - How do we know the model is working?
- **Scale**:
  - We have 10,000 people use the site/hour
  - What hardware should we use?
  - How can we make it easy to run this code in the cloud?

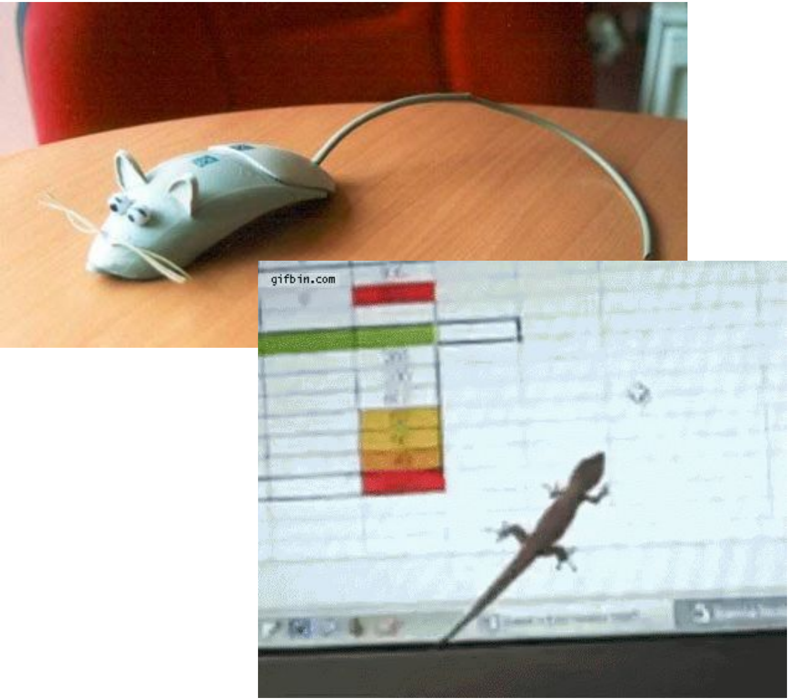


Order food to your door

Sign In for your recent addresses

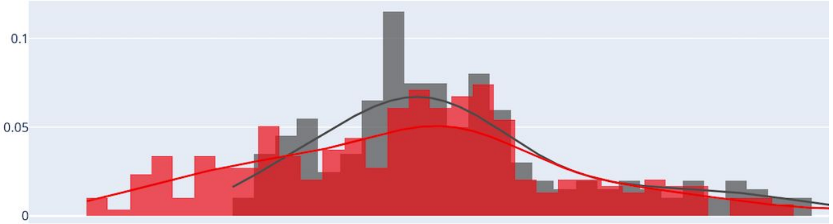
Data: What is available, what should be used?

- Mouse movement (x)
  - Sequence of points in 2d
    - [123,34], [124,34], [124, 33], …
  - Plus clicks at specific times
    - [r, 1.2s], [l, 2.3s], …
  - Is this enough?
- Feedback (y)
  - Food selection result
- Is this enough?
  - Want to predict what they will order.
  - Add time of day, etc.



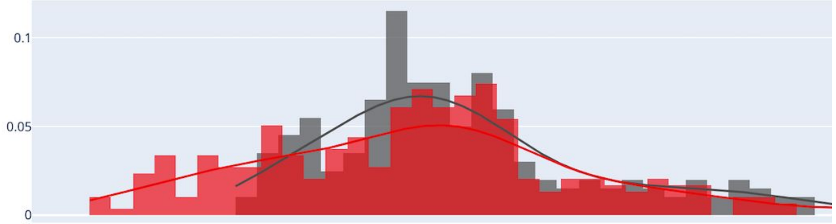
Analysis: Train a Model, What type of Model?

- What type is the input? (x)
  - Sequence of points in 2d
    - [123,34], [124,34], [124, 33], …
  - Plus clicks at specific times
    - [r, 1.2s], [l, 2.3s], …
  - Output y is a type of food.
- Feedback (y)
  - Food selection result
  - Multi class prediction
- Possible Problems
  - Classification, Regression
  - Clustering, Dimensionality reduction
  - Which one?



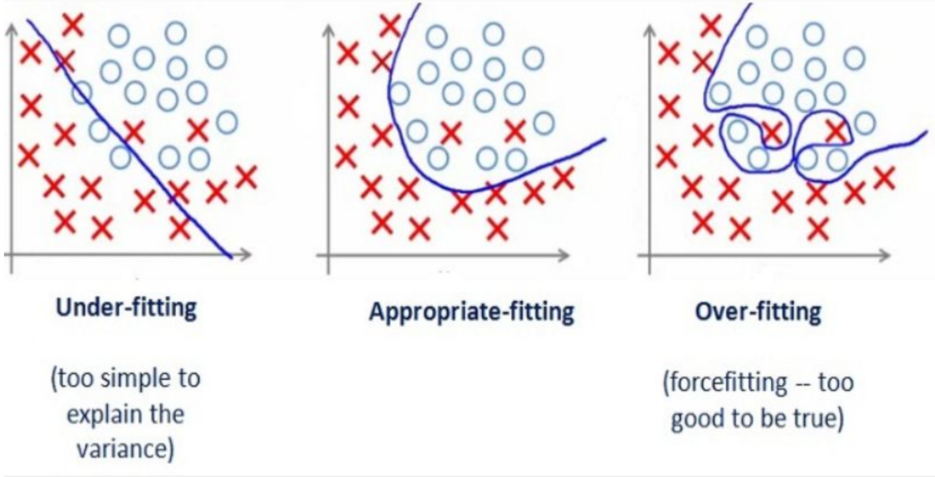
Analysis: Which Classification Algorithm Works Best?

- What type of Model for Classification?
- Support Vector machine
  - Theoretically sound
  - Does not scale well to large input sizes
- Linear Model
  - Is food choice a linear function of mouse movement?
  - No
- Random Forest
  - Does not scale well
- Neural Network
  - Sometimes unstable to train
  - Takes long time to train



Underfitting/Overfitting: Is the Model fitting the data well?

- Underfitting
  - Linear models have little representation
- Overfitting
  - Model has learned special cases
  - These special cases are not part real dist
- How to know the fit is appropriate
  - Leave out some recorded data for *testing*
  - Cross Validation
- How to get a better fit
  - Regularization
  - Occam's razor: simplest model is the best



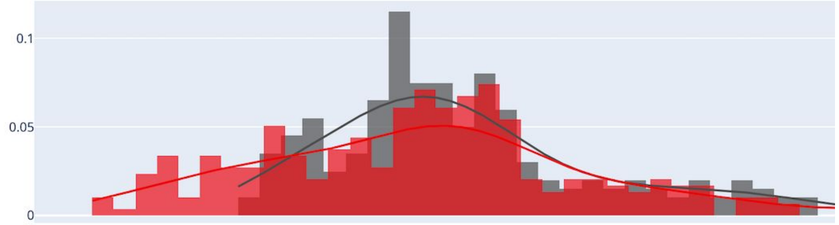
Under-fitting

(too simple to explain the variance)

Appropriate-fitting

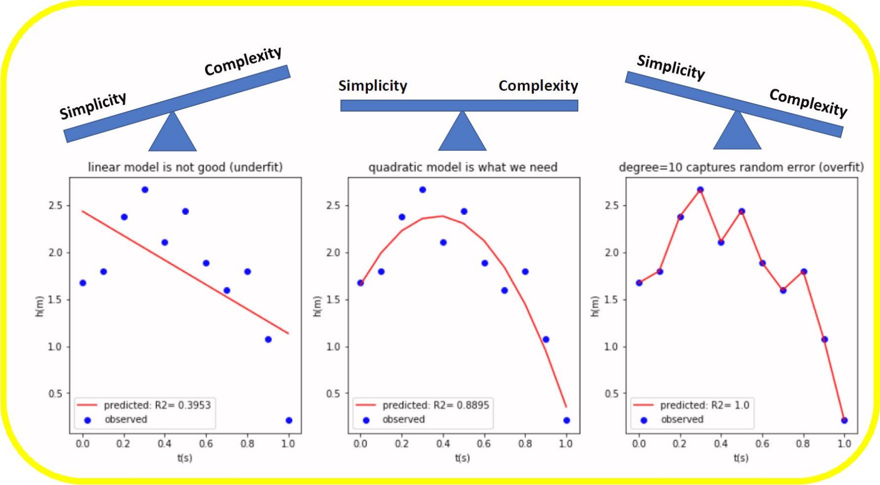
Over-fitting

(forcefitting -- too good to be true)



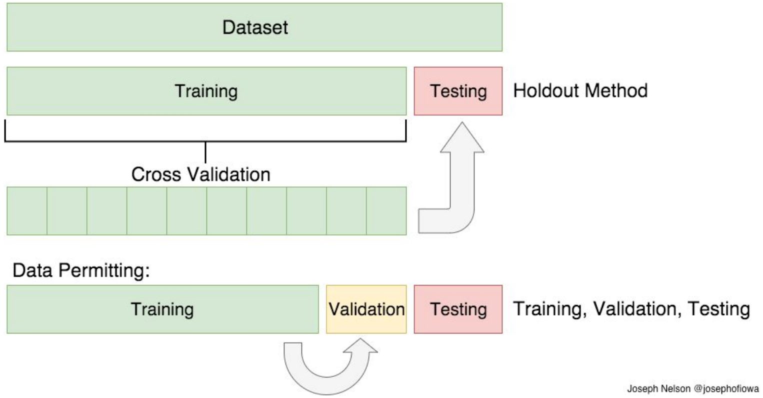
Model Complexity: And how to Measure it

- Decision Trees
  - Simple trees
- Deep networks
  - Number of parameters
  - Number of layers
- Random Forests
  - Decision combinations



Pre Processing and Model Tuning

- **Is the model performing well enough?**
  - We can often get better performance.
- **Pre-process the data**
  - How to split data for test/validation
  - Do we have enough data to split?
- **Model Tuning**:
  - All models have hyper parameters
  - Number of trees, clusters, parameters
  - Try different values of these

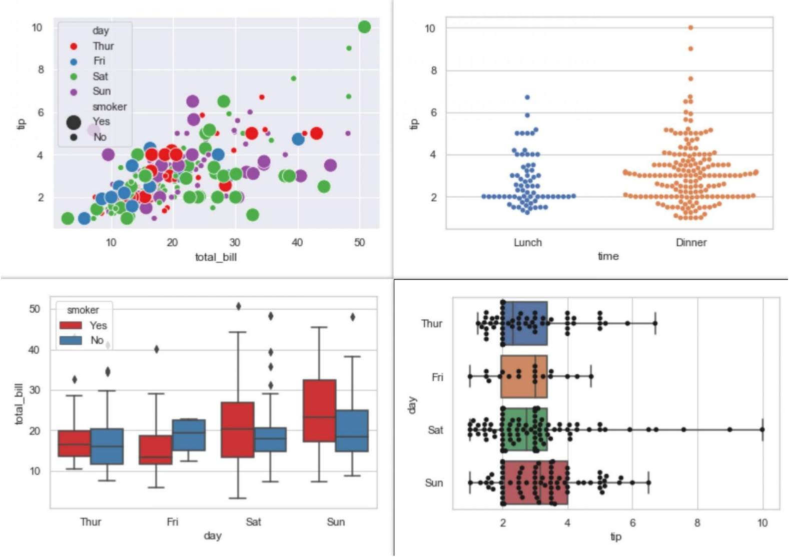


Joseph Nelson @josephofiowa

Data Visualization: Best method to display results

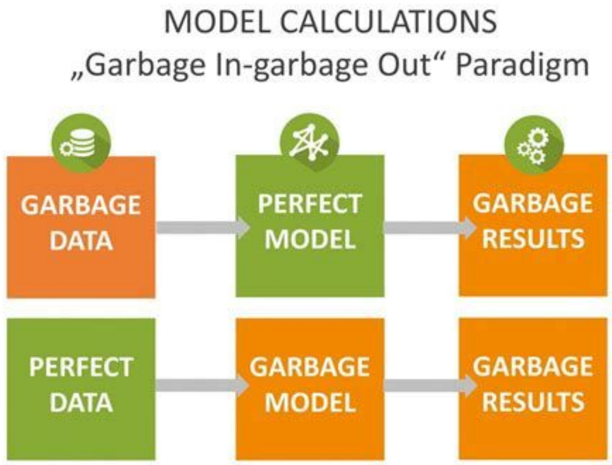
- How to show ourselves and others the model is working

- **Exploration:**
  - Are their other trends in the data?
  - How to find them?
- **Presentation**:
  - Best way to display the exploration?
  - Want a compact visualization with little noise
- **Advanced visualizations:**
  - **Multi-Modal visualization**
  - **Interactive visualizations**
  - **Traps of Visualization**



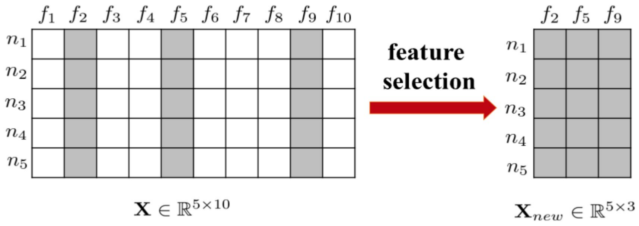
Data Preprocessing and Augmentation

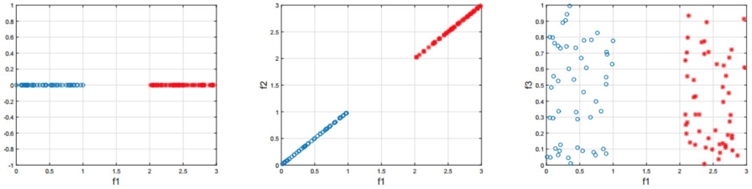
- How can we prepare the data to make the models job easier
- **Preprocessing**:
  - Garbage in garbage out
  - Process data to remove garbage
- **Augmentation:**
  - Scaling data
  - Others methos????



Feature Engineering: Selecting better features

- Not all features are *Equal*, use the better ones
  - Recal meal prediction problem
- **What is a feature**:
  - X = [123,34], [**124,34**], [124, 33]
  - Maybe the last 10 values in X work best
- **Selection**:
  - Reduce input size
  - Possibly use more simple learning model

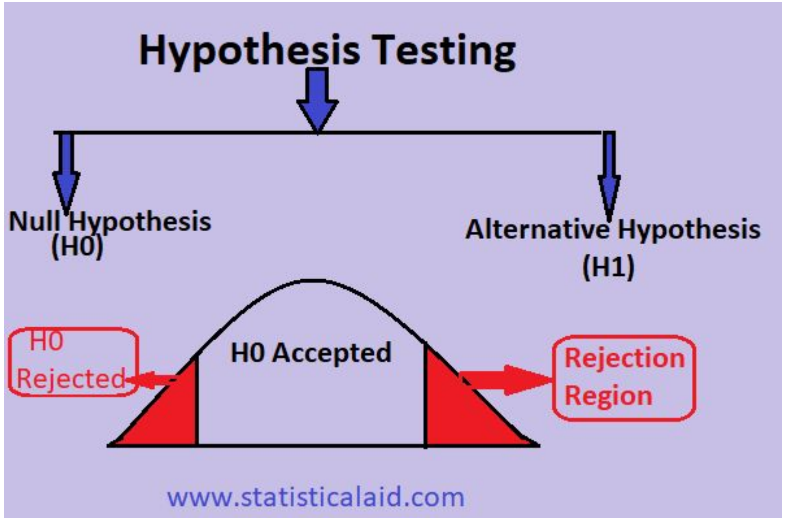

f_1 f_2 f_3 f_4 f_5 f_6 f_7 f_8 f_9 f_{10}
n_1 n_2 n_3 n_4 n_5
feature selection
f_2 f_5 f_9
$\mathbf{X} \in \mathbb{R}^{5\times10}$
$\mathbf{X}_{new} \in \mathbb{R}^{5\times3}$



(a) relevant feature f_1 (b) redundant feature f_2 (c) irrelevant feature f_3

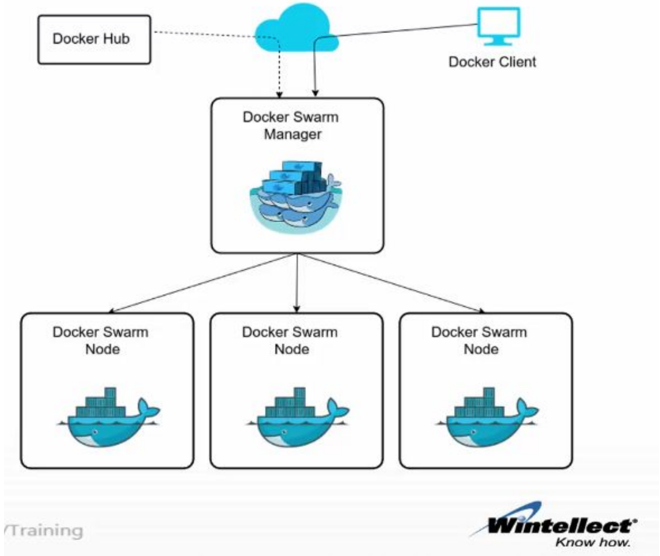
Hypothesis testing: Be confident in your model

- How confident are we in our results?
- **Hypothesis Testing**
  - Statistical test for likelihood of true
  - Vs null hypothesis
- **Methods**
  - Direct simulation
  - Shuffling
  - Bootstrapping
  - Cross Validation



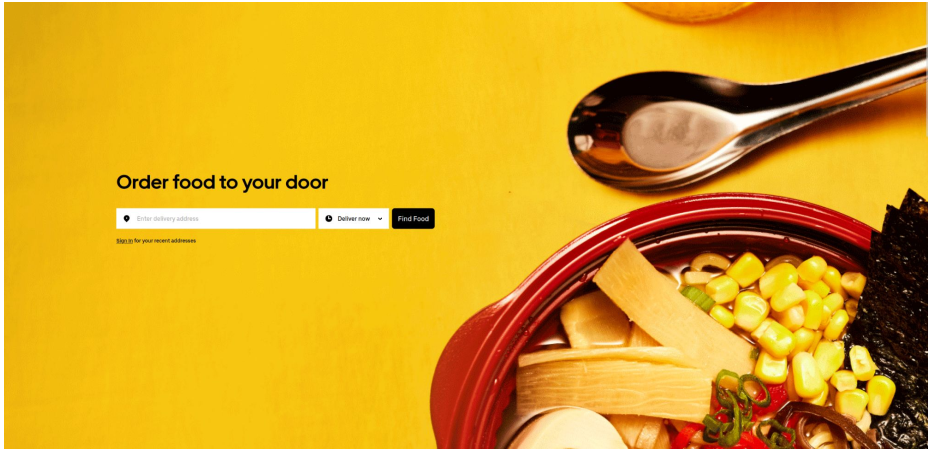
Virtualization with Docker: Make your code portable

- We need to deploy our code on the company servers
  - How to make this easy?
- **Virtual Machines (VM)**:
  - Simulate an entire new OS on top of an existing one
  - Copy this setup across computers
- **Docker**:
  - An ecosystem of lightweight VMs
  - Easy to share with other and deploy
- Useful for HP tuning
  - Replicate your code across multiple computers



Feature Selection and Outlier Detection

- What food should we suggest?
- **Data**:
  - Mouse movement
  - Searches
- **Analysis**:
  - What should we predict?
  - How can we train a model to predict it?
  - How do we know the model is working?
- **Scale**:
  - We have 10,000 people use the site/hour
  - What hardware should we use?
  - How can we make it easy to run this code in the cloud?



Order food to your door

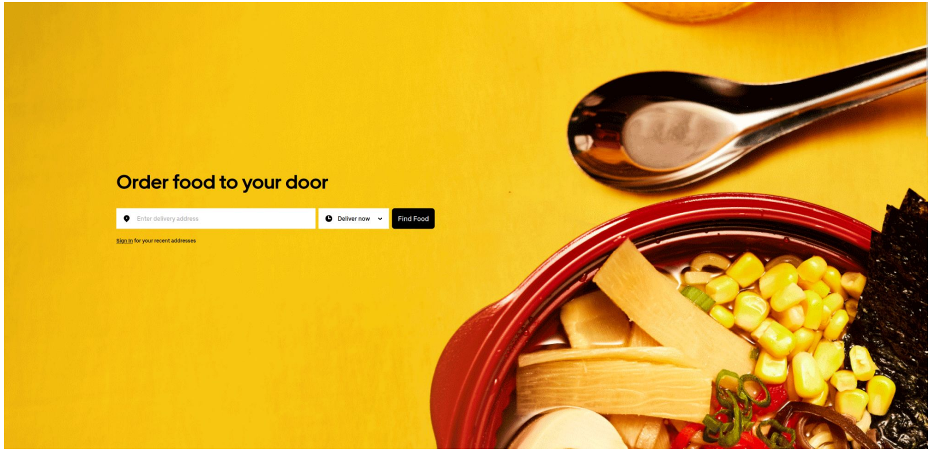
Sign In for your recent addresses

Scaling: Experiment Tracking

- When working at scale you need to organize at scale
- **What is Scale**:
  - Your experiments are in the cloud
  - You data is in the cloud
  - You need to analyze the results
- **Experiment tracking**:
  - Check on experiments while they are running
  - Perform additional exploration and analysis
  - Reproduce an experiment?????

Unsupervised Learning

- What food should we suggest?
- **Data**:
  - Mouse movement
  - Searches
- **Analysis**:
  - What should we predict?
  - How can we train a model to predict it?
  - How do we know the model is working?
- **Scale**:
  - We have 10,000 people use the site/hour
  - What hardware should we use?
  - How can we make it easy to run this code in the cloud?



Order food to your door

Sign in for your recent addresses

Algorithmic Bias

- How do we know the predictions are fair?
  - Really is the data fair?
- **Bias**:
  - Data is collected by people
  - Certain groups appear more often in the data
  - People with more $$$ will use online ordering system more often
- **Fairness**:
  - Understand the effects of these issues
  - Make you model and data more fair
  - Avoid amplifying historical bias

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

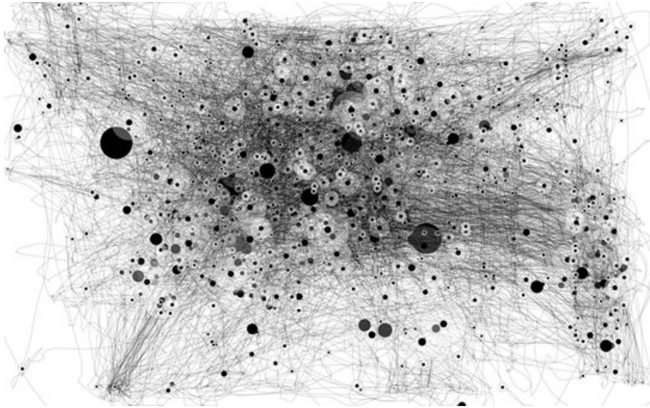
May 23, 2016

Hardware and Distributed learning

- What hardware should we buy to run our algorithm?
  - Need to justify this to the boss
- **Hardware**:
  - Image processing?
  - How many queries / minute?
  - Can we optimize to save $$
- **Distribute**:
  - Which cloud computer to use?
  - Depends on model assumptions

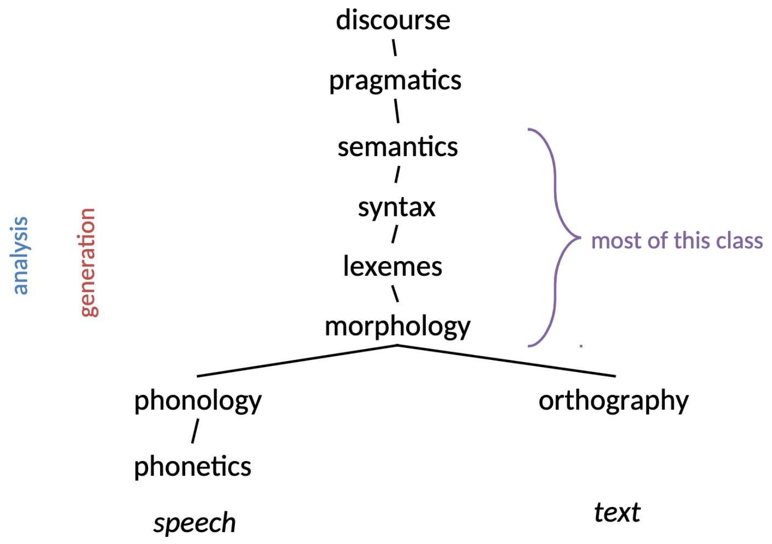
Image Data: Change of representation?

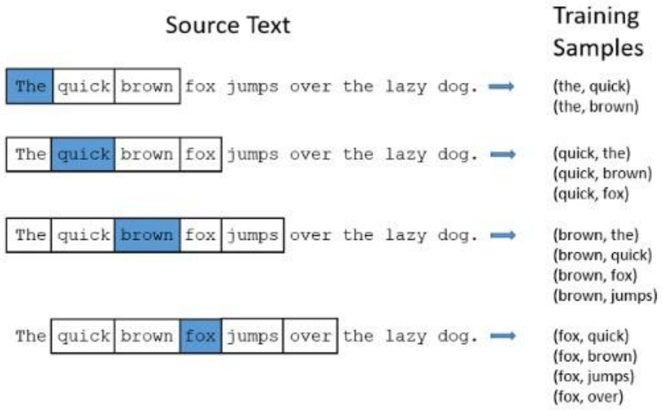
- What is the best way to process the data?
  - As text/image/graph?
- **Converting Data**:
  - Image processing is powerful
  - Convert data to images.
    - [123,34], [124,34], [124, 33] -> image
- **Other Representations**:
  - Many things can be converted to *images*
  - Images take up space on disk
  - Trade off on processing time vs disk space



Text Data

- Specificity of text data
    - standard tasks
    - pre-processing
    - difficulties
        - syntactic ambiguity
        - importance of context
        - Polysemy
- Probabilistic Models
    - NGrams
    - Word embeddings:
        - SVD Word2Vec
        - GloVe

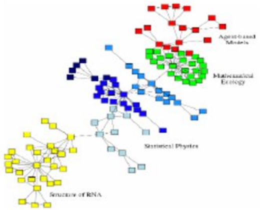




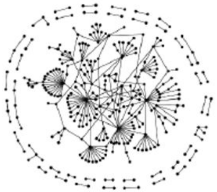
Graph Data

- Specificity of Graphs
  - Standard tasks
  - representations
  - Adjacency matrix, Laplacian matrix
- Node embeddings
  - By node similarity
  - Random walks

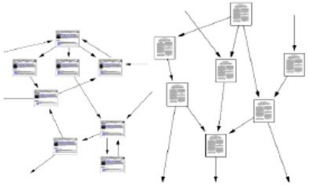
Social networks



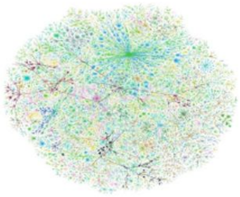
Economic networks



Biomedical networks



Information networks:
Web & citations



Internet

Networks of neurons

Analysis: More Complex Models for Text Data

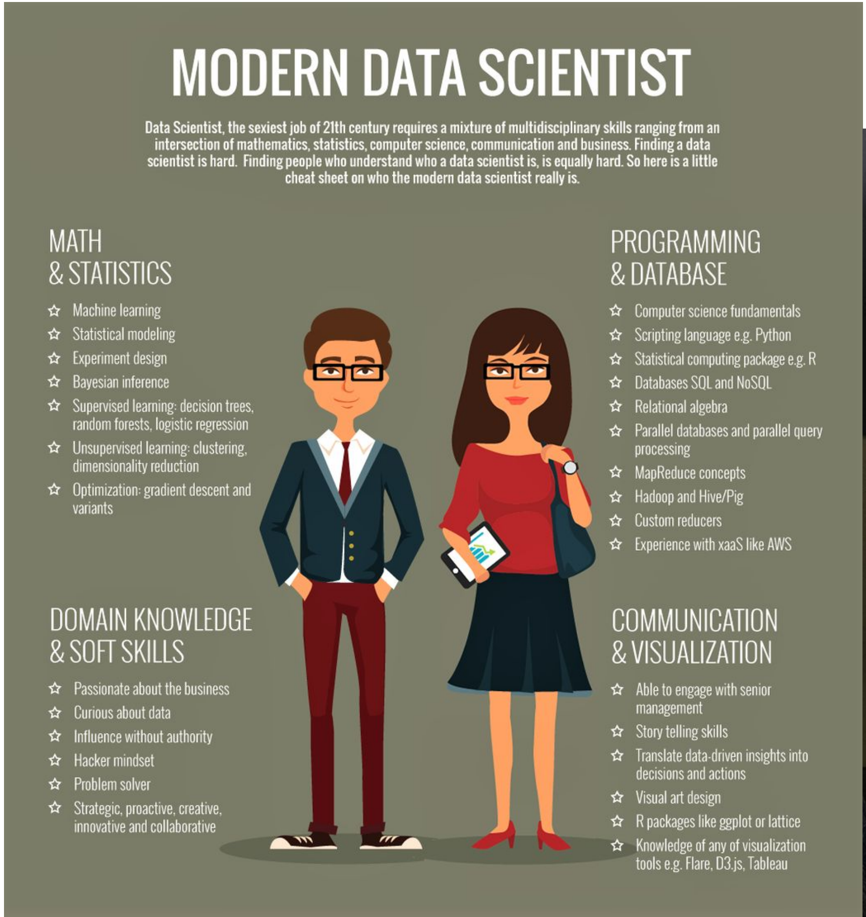
- What food should we suggest?
- **Data**:
  - Mouse movement
  - Searches
- **Analysis**:
  - What should we predict?
  - How can we train a model to predict it?
  - How do we know the model is working?
- **Scale**:
  - We have 10,000 people use the site/hour
  - What hardware should we use?
  - How can we make it easy to run this code in the cloud?

Digging deaper

At the end of the lecture note that we missed many details of these problems

1.

Data science require diverse skills



MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21th century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- ☆ Machine learning
- ☆ Statistical modeling
- ☆ Experiment design
- ☆ Bayesian inference
- ☆ Supervised learning: decision trees, random forests, logistic regression
- ☆ Unsupervised learning: clustering, dimensionality reduction
- ☆ Optimization: gradient descent and variants

PROGRAMMING & DATABASE

- ☆ Computer science fundamentals
- ☆ Scripting language e.g. Python
- ☆ Statistical computing package e.g. R
- ☆ Databases SQL and NoSQL
- ☆ Relational algebra
- ☆ Parallel databases and parallel query processing
- ☆ MapReduce concepts
- ☆ Hadoop and Hive/Pig
- ☆ Custom reducers
- ☆ Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

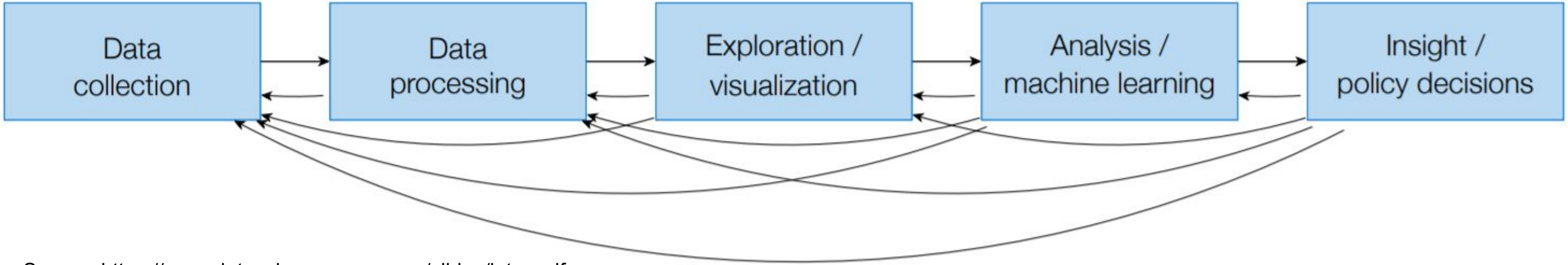
- ☆ Passionate about the business
- ☆ Curious about data
- ☆ Influence without authority
- ☆ Hacker mindset
- ☆ Problem solver
- ☆ Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- ☆ Able to engage with senior management
- ☆ Story telling skills
- ☆ Translate data-driven insights into decisions and actions
- ☆ Visual art design
- ☆ R packages like ggplot or lattice
- ☆ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

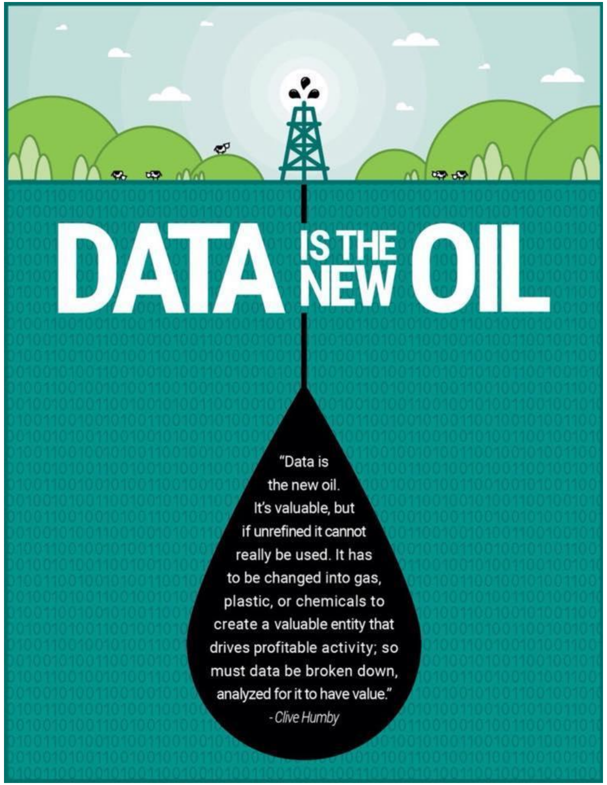
Algorithmic Justice League

Back to what is Data science


Data collection
Data processing
Exploration / visualization
Analysis / machine learning
Insight / policy decisions

Source: https://www.datasciencecourse.org/slides/intro.pdf

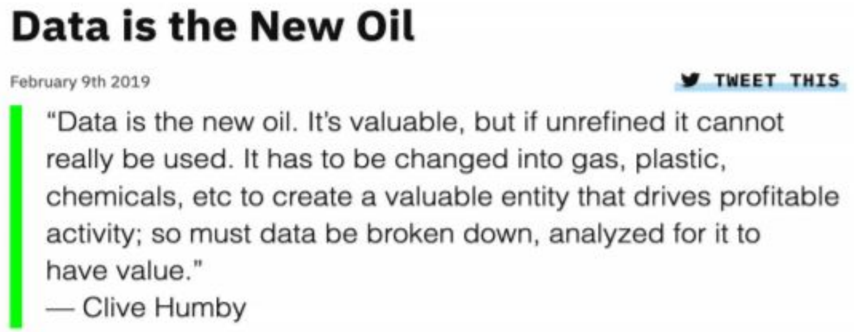
Data science is **not** the new oil



"Data is the new oil. It's valuable, but if unrefined it cannot really be used. It has to be changed into gas, plastic, or chemicals to create a valuable entity that drives profitable activity; so must data be broken down, analyzed for it to have value."
- Clive Humby

Unlike oil, *when dealing with data, it is far from clear how exactly to turn that data into profits*.

Why data is not new oil



Data is the New Oil

February 9th 2019

> "Data is the new oil. It's valuable, but if unrefined it cannot really be used. It has to be changed into gas, plastic, chemicals, etc to create a valuable entity that drives profitable activity; so must data be broken down, analyzed for it to have value."
> — Clive Humby

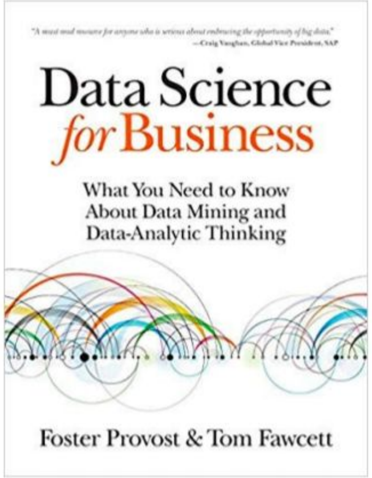
Data are Desserts!

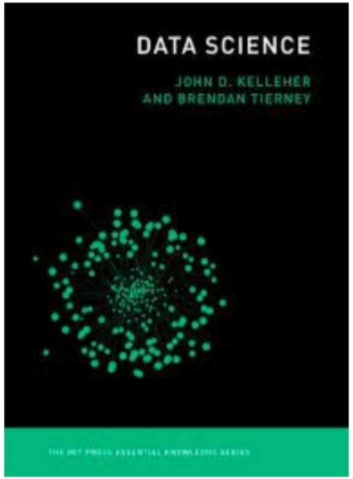
1. Data are the result of deliberate human intervention
2. Data are varied across domains
3. Data are varied within domains

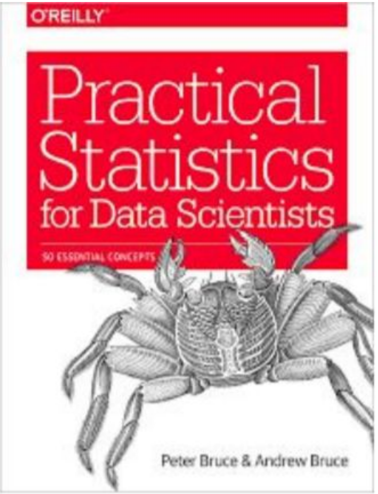
After this is course content

Useful Resources

- Books: Statistical Learning, Python Handbook, Introduction to Data Science
- Online courses: freeCodeCamp , Harvard CS109
- Harvard’s Data Science Review


Data Science for Business
What You Need to Know About Data Mining and Data-Analytic Thinking
Foster Provost & Tom Fawcett




O'REILLY
Practical Statistics for Data Scientists
50 ESSENTIAL CONCEPTS
Peter Bruce & Andrew Bruce

O'REILLY
Python Data Science Handbook
ESSENTIAL TOOLS FOR WORKING WITH DATA
Jake VanderPlas

Resources

1. Some content from (https://ucbrise.github.io/cs294-ai-sys-sp22/,
2. https://fullstackdeeplearning.com/spring2021/,
3. https://stanford-cs329s.github.io/).