

# APMA E4990.02: Introduction to Data Science In Industry

---

Instructor: Dorian Goldman

# Who am I?



- Dorian Goldman, Ph.D 2013
- Ph.D at Courant Institute NYU and Paris VI UPMC) (Calculus of Variations and Partial Differential Equations)  
- Instructor of mathematics at University of Cambridge (2013-2014) 
- Data Scientist - The New York Times 2014-2016 
- Data Scientist - Conde Nast 2016 - Current. 

# What are the goals of this class?

- Learn the rigorous mathematical foundation of machine learning as it relates to problems faced in realistic industry scenarios.
- Learn the tools used in industry, and how these algorithms and methods are used in practice (Python, Scikit-learn, SQL/Map Reduce, Github, Web scraping)
- Build your own Flask app which uses machine learning to solve a problem.

Examples include:

- Recommendation engine for Yelp/Netflix (or any other service).
- Taxi route time estimator.
- User engagement optimization tools (ie. ads, push notifications, etc)
- Music recommendation system.
- Stock/Investment tools.
- Anything you can think of which can use data to make predictions/suggestions.

# What do you need to be ready?

- A laptop (although not explicitly necessary), preferably running MacOS or Linux, but also not essential (you will be judged though). We will be working through algorithms in the class.
- Anaconda - Scientific Python package with IPython Notebook.  
<https://www.continuum.io/downloads>
- Github - If you don't have an account, please go to <https://github.com/>, and register an account. Send me an email at [d2991@columbia.edu](mailto:d2991@columbia.edu) with your Github account and I'll add you to the repo: [https://github.com/doriang102/Columbia\\_Data\\_Science](https://github.com/doriang102/Columbia_Data_Science)

# What is Data Science?

---

Introduction and Examples

# Outline

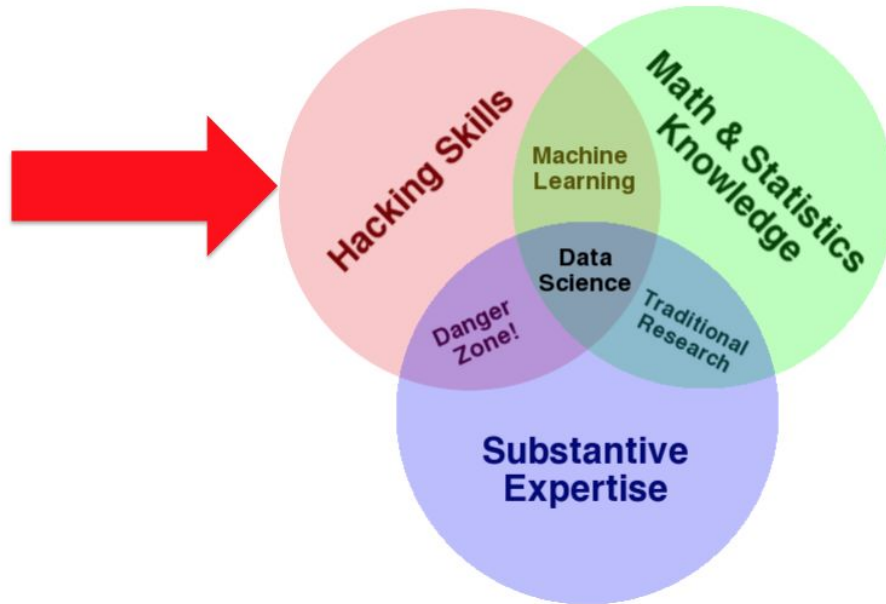
- What is Data Science? What are the skills needed?
- Examples from Industry. Amazon, Netflix, Booking.com, New York Times.
  - Predictive Learning (Supervised)
  - Descriptive Learning (Unsupervised)
  - Prescriptive Learning (Reinforcement)
- Why is Data Science Important?
- Overview of methods of Machine Learning.
- What will you learn in this course?
- What should you know before taking this course?

# What is Data Science?

- **Predictive (Supervised Learning):** The science of using data to predict an outcome (clicking, subscription, cancerous cells, user churn, virality, etc.)
- **Descriptive (Unsupervised Learning):** Using data to group items/users into categories (ie. Risky Customers, )
- **Prescriptive (Reinforcement Learning):** Optimizing action based on response variable (ie. who should receive a marketing email, based on sign ups from an experiment)
- **Exploratory:** Can we describe characteristics of items/users with particular attributes we are interested in? (ie. Are new users who sign up for the new york times mostly Democrats?)
- **Experimental:** Conduct experiments and interpret their outcome.

**Goal of this course:** Master the basics from a theoretical and practical viewpoint.

# What is Data Science?



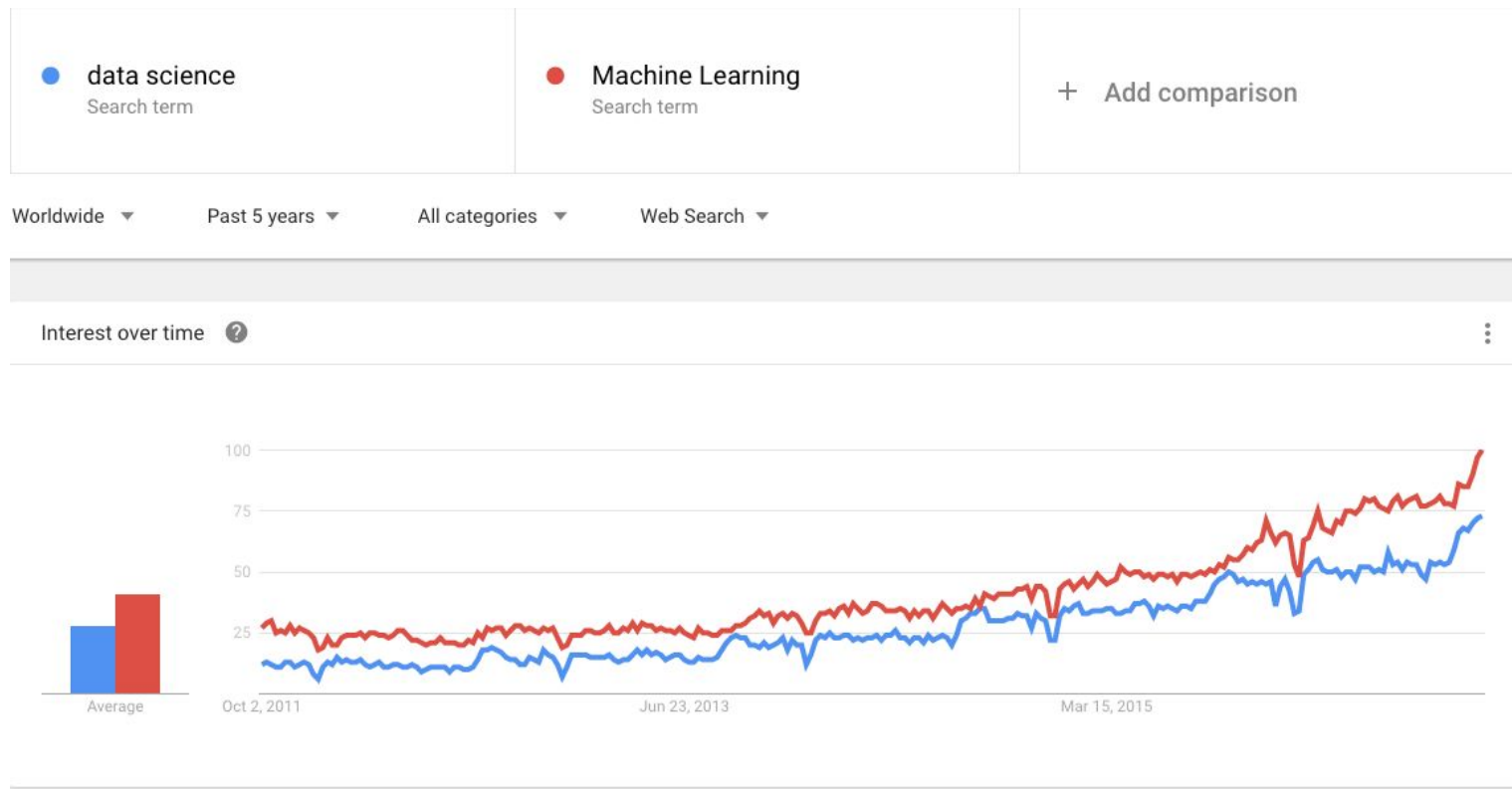
**Hacking Skills:** Comfort with Linux/Unix, networks, databases, working from the command line, debugging code.

**Math/Statistical Knowledge:** Need understanding probability, statistics, optimization methods to create and use models.

**Substantive Experience:** Need experience working with real data and business problems along with the problems that come along with them. Also need ability to communicate technical ideas to stakeholders.



# Interest in Data Science is Blowing Up



# Considered #1 Job by many rankings



## The Best Jobs of 2016: 1. Data Scientist

**2016 Jobs Rated Score:** 91

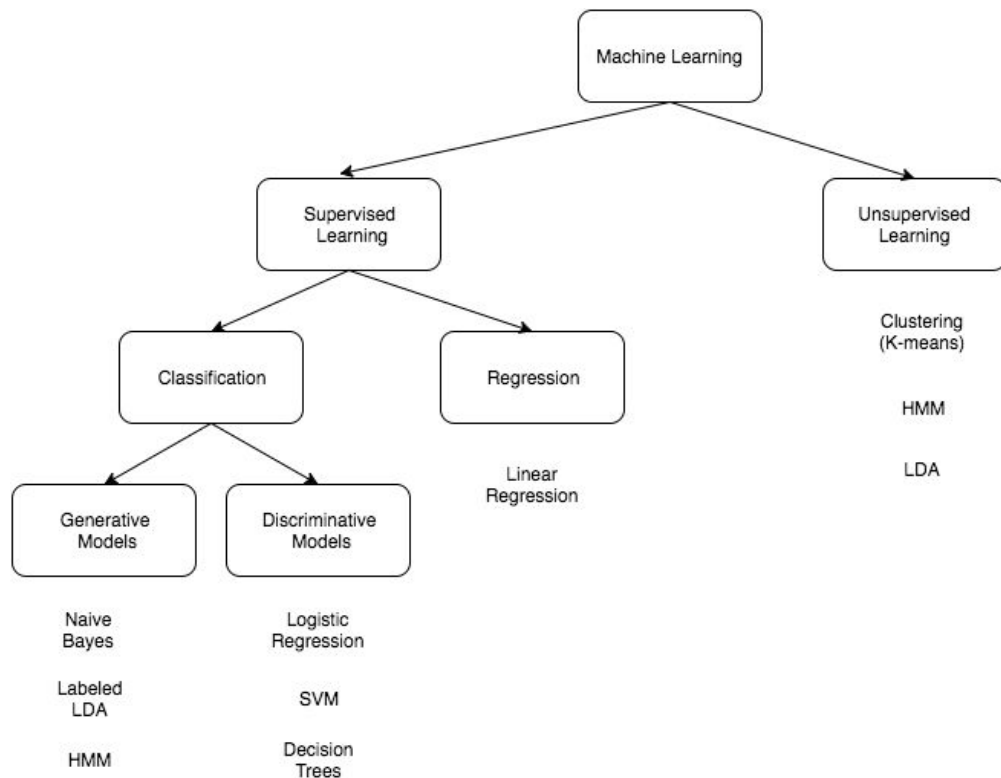
**Annual Median Salary:** \$128,240

**Growth Outlook :** 16%

Opportunities across a variety of fields make data scientist not just a high-growth job, but also one of the most lucrative tracked by the Jobs Rated report.

- Companies are willing to pay top dollar since Data Scientists can have a significant impact on revenue for a company.

# How do we break down machine learning algos?



# Predictive Learning

---

(Supervised)

# Predictive Learning - Summary

- Predictive learning attempts to learn a model from data **X** which predicts a variable **y** (ie. type of movie, number of views would be **X**, y is your rating).
- Learns from data which has ‘**correct**’ answers given data inputs - this is why it’s “**supervised**”.
- **Algorithms (you will learn):**
  - Linear Regression (Regression)/Logistic Regression(Classification).
  - Random Forest/Decision Tree Regression/Classification.
  - SVM (Support Vector Machines).
  - Naives Bayes (For Multiarmed Bandits for example)
  - Maximum Likelihood and Time Series Modeling (more advanced - fit data to a prior probability distribution).
  - Neural Nets (if time permits)

# Problem Motivation 1. Amazon.com

Can we predict how a user will rate an item? Why do we care?



Nikon COOLPIX S33 Waterproof Digital Camera (Blue)

by Nikon



582 customer reviews | 196 answered questions

#1 Best Seller

In Digital Point & Shoot Cameras

List Price: \$149.96

Price: **\$129.00** & FREE Shipping. Details

You Save: \$20.95 (14%)

In Stock.

Want it Friday, Sept. 30? Order within **19 hrs 2 mins** and choose **Same-Day Delivery** at checkout. Details

Ships from and sold by Amazon.com. Gift-wrap available.

Color: Blue



Style: Base

Accessory Bundle

Base

- Waterproof up to 33 feet deep; shockproof up to 5 feet; freezeproof down to 14° F
- 3x wide-angle NIKKOR glass zoom lens
- 13.2-MP CMOS sensor
- Full HD 1080p videos with stereo sound
- Oversized buttons and easy menus

- Can we predict how you would rate this item based on what we know about you?
- **Why do we care?**  
**Answer:** Will increase purchase rate and this can be measured in \$ via an experiment.
- **Good recommendations = \$\$.**

# Problem Motivation 2. Booking.com

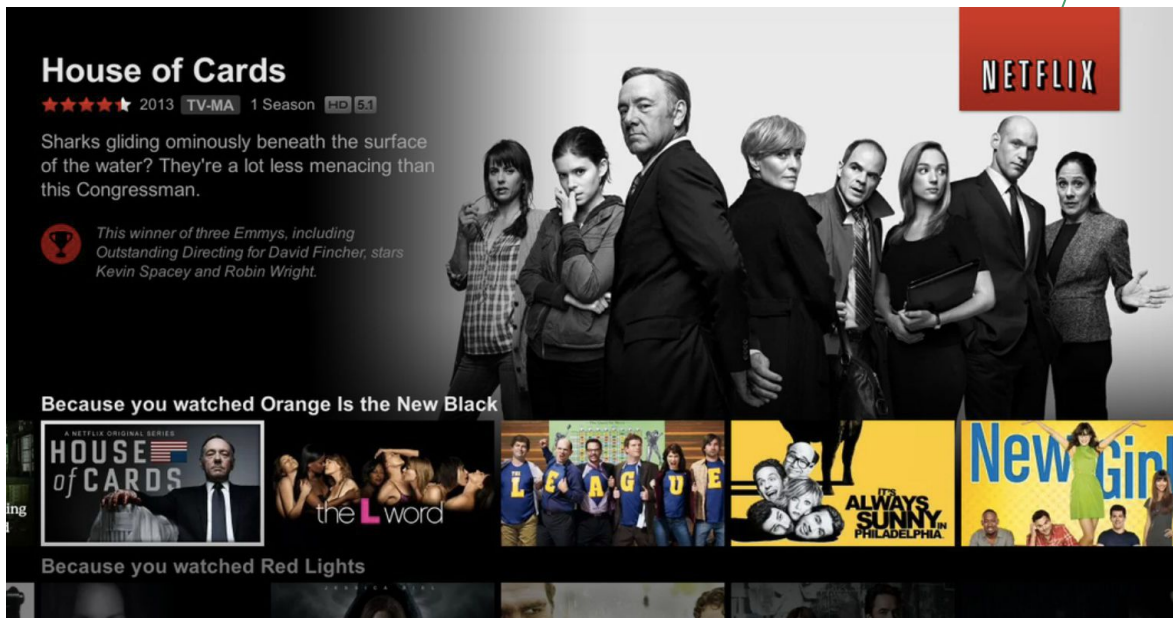
Can we predict that a hotel is likely to sell out soon? If so when? Why do we care?

The screenshot displays two hotel listings on the Booking.com interface. The top listing is for 'Dorsett Shepherd's Bush', featuring a '34% off' badge, a 'Value Deal' icon, a 4.5-star rating, and a 'Fabulous 8.6' score from 524 reviews. It highlights 'Popular now!' with 13 people looking at the hotel and a 'Latest booking: Less than 1 minute ago'. The bottom listing is for 'City Marquee Albert Serviced Apartments', showing a 'Value Deal' icon, a 4.0-star rating, and a 'Good 7.8' score from 85 reviews. A red-bordered notification box on this listing states: 'There are 2 people looking at these apartments. It's likely that these apartments will be sold out within the next 2 days.' Below this, it says 'Latest booking: 2 hours ago'. A green arrow originates from the main question text and points directly to this notification box. Both listings include a 'Book now' button and a 'Last chance! We have only 1 left on our site!' message.

- **Conversion:** Users may be **more inclined to purchase** if they are aware the room may sell out soon (improve purchase rate).
- **Retention:** Users may be **only interested in certain hotel options** and not be aware that they don't have the luxury of waiting - **this could upset customers** if the hotel sells out without warning (customer service).

# Problem Motivation 3. Netflix.com

Can we predict how you would rate a movie?













- **Engagement:** Users will be more engaged if movies they are likely to rate highly are shown to them first.
- **Retention:** Engaged customers are loyal customers, which means \$\$.



# Problem Motivation 4. Nytimes.com

Can we predict which articles you would like to read?

MOST EMAILED			MOST VIEWED			RECOMMENDED FOR YOU		
1.	THE OUTLAW OCEAN	A Renegade Trawler, Hunted for 10,000 Miles by Vigilantes						
2.		Campus Suicide and the Pressure of Perfection						
3.		As Tech Booms, Workers Turn to Coding for Career Change						
4.		Prison Worker Who Aided Escape Tells of Sex, Saw Blades and Deception						
5.		Under Oath, Donald Trump Shows His Raw Side						
6.		American Hunter Killed Cecil, Beloved Lion That Was Lured Out of Its Sanctuary						
7.		A Creature on the Loose Puts Milwaukee Residents on Edge						
8.		N.F.L. Upholds Tom Brady's Ban; Cellphone's Fate Helped Make the Call						
9.		Escalator Death in China Sets Off Furor Online						
10.	DAVID BROOKS	The Structure of Gratitude						

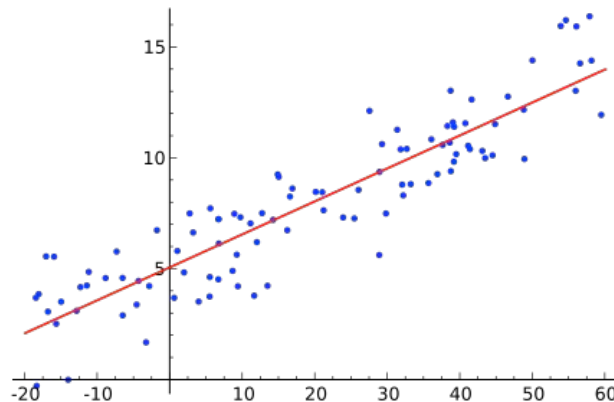
- **Engagement:** Users may be **more inclined to subscribe** if they are highly engaged with the content.
- **Retention:** Once again, engaged users are loyal users.

# Predictive Learning - Regression

Given a collection of points to learn from:  $(x_i, y_i)$

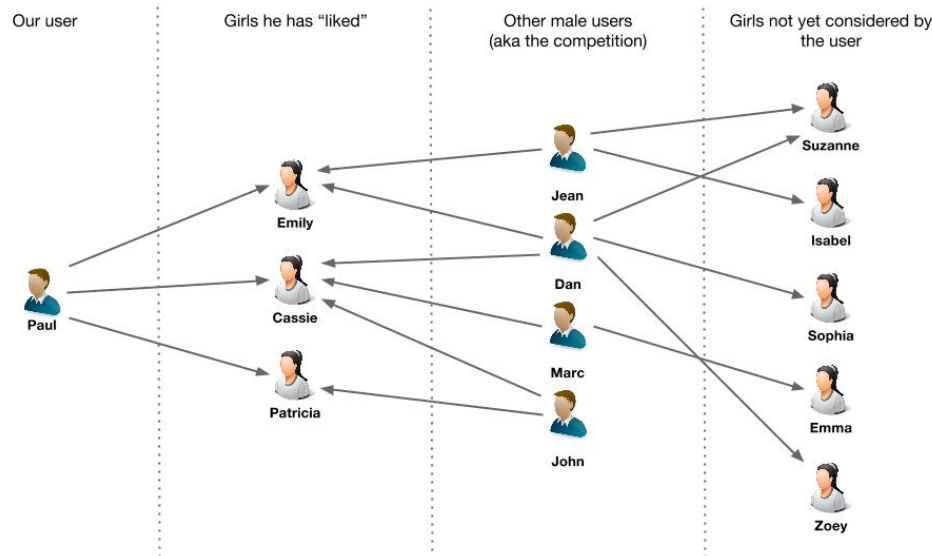
Can we find a function  $f : X \rightarrow Y$  minimizing the distance to the data.

$$\frac{1}{N} \sum_{i=1}^N |y_i - f(x_i)|^p$$



**All of predictive machine learning is based on discovering ways to find  $f$**  (although the norms we use will depend on the problem at hand, it won't always be this one).

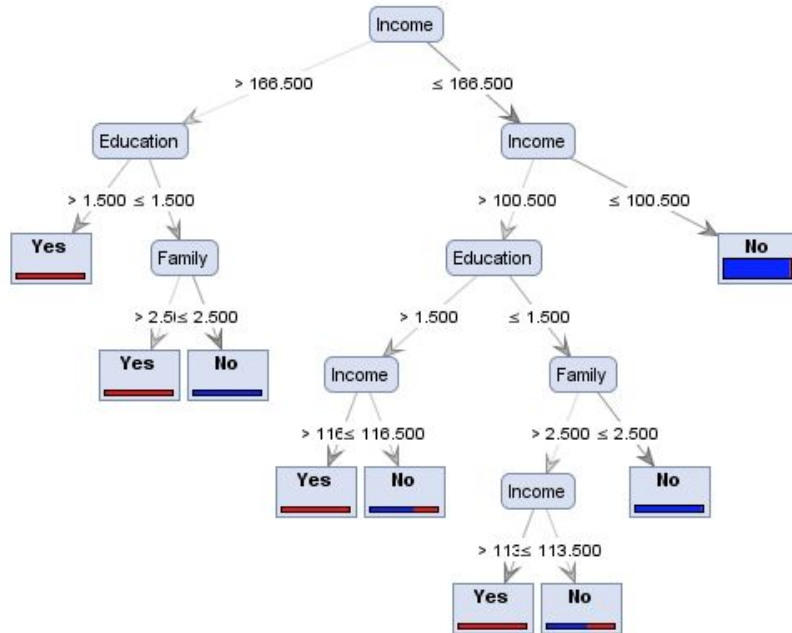
# Predictive Learning - Recommendation Engines



Guy  $j$  for Girl  $n$  has a propensity measured by a bipartite graph diffusion model

$$\pi(j, n) := \sum_{j', n'} p(n|j') p(j|n') q_{n'} = M^T G q_j,$$

# Predictive Learning - Decision Trees



Should this person receive a loan?  
A decision tree is another way of finding a “rule” which assigns user attributes to an outcome.

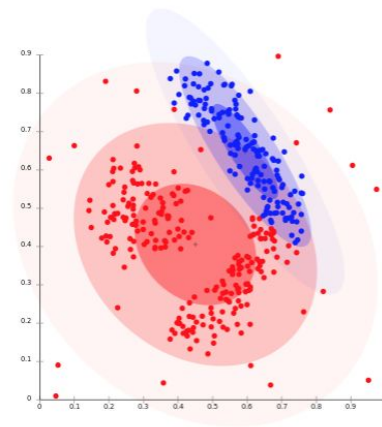
# Descriptive Learning

---

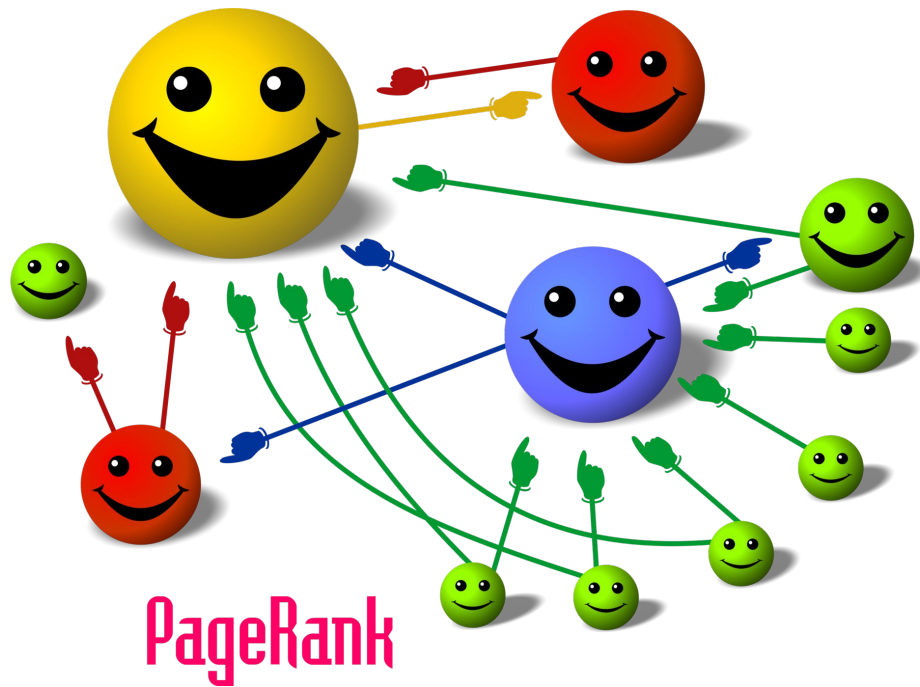
(Unsupervised)

# Descriptive Learning - Summary

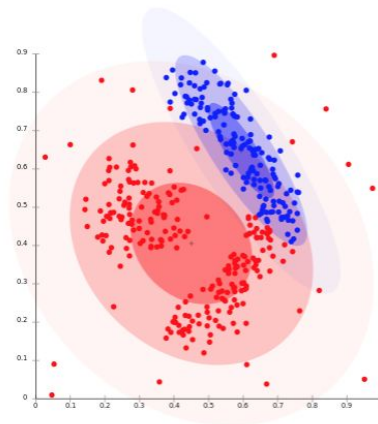
- Try to infer a hidden structure in the data without proper training examples (no teacher, hence 'unsupervised').
- I.e. Group certain customers into 'wealthy, educated' - no clear way of evaluating performance of model.
- This course will focus less on unsupervised learning.
- **Algorithms:**
  - K-means clustering.
  - Decision Tree Clustering.
  - SVM
  - Topic Models



# Problem Motivation 1. Google's PageRank



Cartoon illustrating the basic principle of PageRank. The size of each face is proportional to the total size of the other faces which are pointing to it. (Source: Wiki)



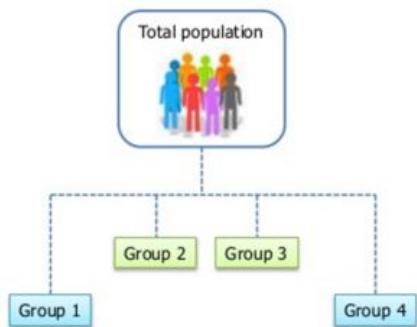
# Readerscope - An NyTimes.com example



- Can we identify topics in various countries and regions around the world?
- Can use this to improve engagement worldwide.



# Problem Motivation 2. NyTimes.com



- How to cluster subscribers into different categories?
- **Possible categories:** “Young, uneducated, low income”, “Wealthy, older, educated”, etc.
- Clustering algorithms may naturally find these categories, and this can inform marketing and sales strategy.
- **Possible features:** Age, income, location, reading history (online), number of complaints, subscription tenure, etc.

# Prescriptive Learning

---

Reinforcement Learning

# Prescriptive Learning - Summary

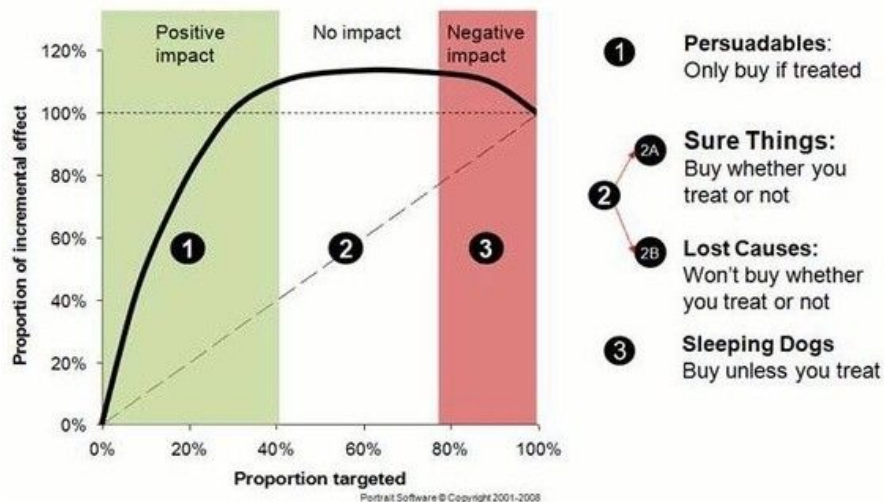
- Prescriptive learning attempts to find an optimal action  $a$  to maximize the expected reward/outcome (ie. who should we show this kind of ad to, or who should we send this marketing email to?). Uses **Bayesian methods**.
- A well defined metric is used to determine the performance of such a model (will be seen later).
- **Algorithms:**
  - Relies entirely on maximizing expectation of reward conditioned on user attributes and action.
  - Incredibly useful since it's actionable.
  - Can be “live” (multiarmed bandit) or from logged data (uplift modeling).
  - Easiest when we have an A/B (randomized controlled trial) where we can measure causal inference of an outcome conditioned on an action.

# Uplift Modeling - what is it?



How do we determine the right action to maximize our desired outcome?

- Not everyone should receive the same action.
- Maybe some users will be angered by an advertisement, and others will only sign up if you send them a particular ad.



# Uplift Modeling - Mathematical Formulation

- Given a collection of outcomes, features and actions, how do we find the optimal action that a given user should receive to maximize the expected outcome for the next iteration?
- Basic foundation of causal inference.

$$\max_{h: X \rightarrow A} \mathbb{E}_{\Omega}(y|x, a) \sim \frac{1}{NB(a_i|x_i)} \sum_{i=1}^N y_i \mathbf{1}(a_i = h(x_i))$$

(this formula will be explained/derived in later lectures)

$y$  – outcome

$x$  – item/user attribute

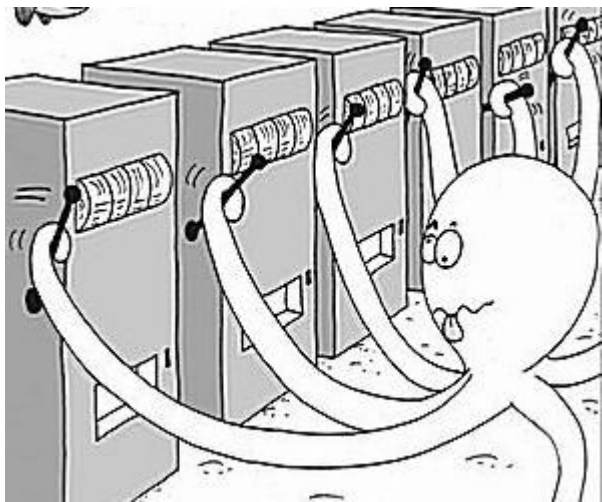
$a$  – action taken

$\Omega$  – distribution of rewards

The policy we choose is  $h$ .

# Multiarmed Bandits and Nytimes.com

For a new user with no data, how do we predict what you should see?



- If you were to play 5 slot machines, one of which was the best (but you don't know). How would you balance **exploration** with **exploitation**?
- [Multiarmed Bandits](#) do this in a rigorous mathematical fashion.

# Multiarmed Bandits - Mathematical Formulation

## Maximize Expected Reward

For each arm  $a \in A$  we define

$$Q(a, x, \theta) := \mathbb{E}(y|a, x, \theta). \quad (16)$$

For each arm  $a$  we have the distribution of rewards  $p(y|a, x, \theta)$  where the variable  $\theta$  represents the distribution over means of the possible arms (for example when  $y$  is Bernoulli,  $\theta$  represents the probability that you will get a 1 response from the arm).

## Update posterior via Bayes Rule

$$\begin{aligned} p_{D_t}(\theta) &= \frac{\prod_{t=1}^T p(y_t|\theta, a_t, x_t, \gamma) p(\theta|a_t, x_t, \gamma)}{p(y_t|a_t, x_t, \gamma)} \\ &= \frac{\prod_{t=1}^T p(y_t|\theta, a_t, x_t, \gamma) p(\theta|\gamma)}{p(y_t|a_t, x_t, \gamma)}. \end{aligned}$$

## Algorithm

- Observe context  $\{X_t\}$ .
- Compute  $P_{D_t}(\theta|a)$ , sample  $\theta_t$  from this distribution.
- Compute  $Q(a, x_t, \theta_t)$  for each  $a$ , then take  $\text{argmax}$  over  $a$ .
- Observe reward  $y_t$  from arm  $a_t = \text{argmax}_a Q(a, x, \theta_t)$ .
- Repeat.

Improve this slide.

What will you learn in this course?

---



# What will you learn in this course?

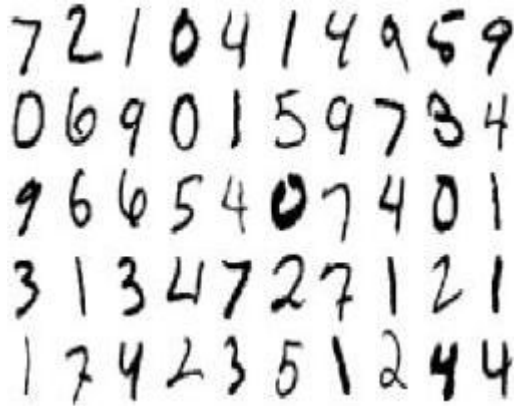
- Most common methods used in **Machine Learning** and how they are used in industry, more precisely, **how to build models from data**. Some examples we will cover are:
  - Recommendation engines (how do we deliver content meant for you?)
  - Predicting virality of content on the web
  - Time Series analysis - how do we predict what will happen at a future time based on the past? (Related to stock forecasting, paper distribution, etc).
- **Methods of Machine Learning:**
  - Regression, classification, decision trees, RF and clustering.
  - Model complexity and regularization (variance/bias tradeoff). Cross validation.
  - Graph Diffusion, collaborative filtering, random walks. Graphical models.
  - Bayesian statistics. Graphical Models. Expectation maximization.
  - Time Series Analysis. Autoregression. Poisson Regression, etc.
  - Much more!
- **Map Reduce/SQL and Data Engineering:** Will learn why and how we use distributed computing for processing data.
- **Build your own web app:** By the end of the class, the final project will be to build your own web app using any of the tools (or others!) covered in this class.

# How do we 'learn' from data?

---

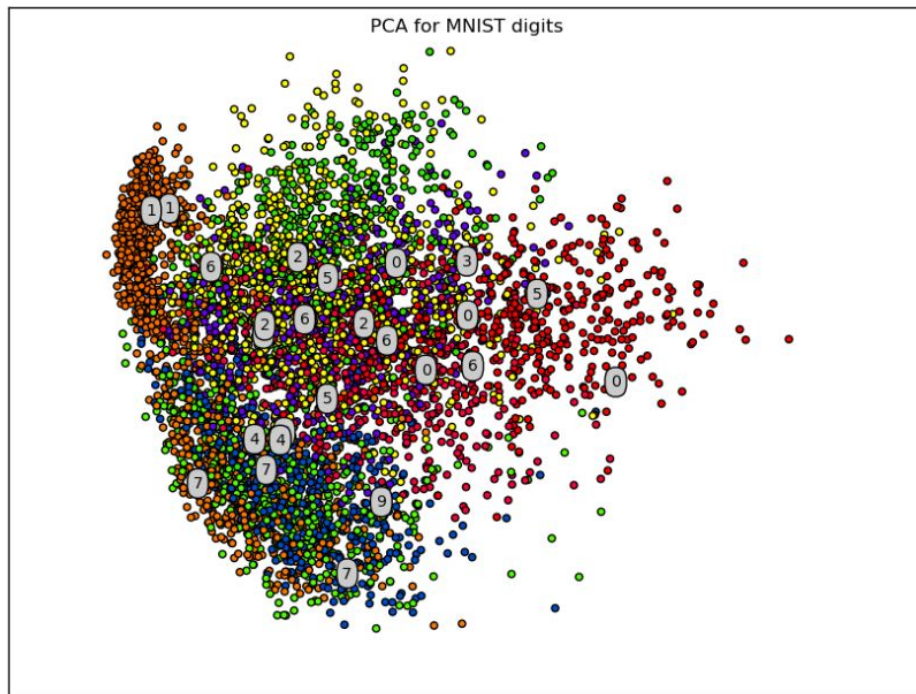
Is it even possible?

# Can we learn from data? MNIST dataset



- Can you recognize these digits? Why?
- You most likely have a notion of a '6' or '7' in your mind - what allows you to classify these? Mostly experience. How do we mimic 'experience' with data?
- Which models might we guess would work best here?
  - **Linear model** - have a unique feature for each pixel in an 8x8 image ( 64 features total).
  - **Nearest neighbors** - based on the above features, which images are I "closest" to?
  - **PCA** - reduce complexity to optimize for test set performance.
  - **Neural Nets** (will speak more of these later).

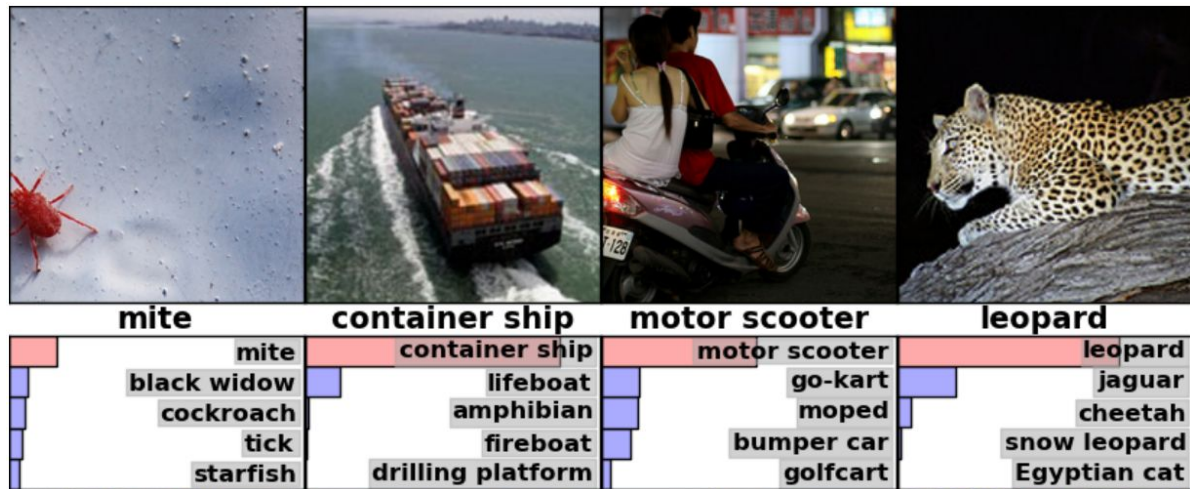
# Visualization of the top two components



- Here we see a clustering of the most 'significant' components extracted from the 64 pixels/features (*will be explained later via PCA*).
- A simple model would be to draw circles around the most dense regions, and cluster numbers based on whether or not they fall into those circles.



# Can we recognize more advanced images?



- Modern neural nets are able to classify objects better than humans can in some instances!
- Current model in tensor flow uses what is called a 'convolutional neural network' (to be explained later)

- Taken from Tensor Flow
- [https://www.tensorflow.org/tutorials/image\\_recognition/](https://www.tensorflow.org/tutorials/image_recognition/)

# Can we identify spam emails?



**SPAM**

**vs.**



**HAM**

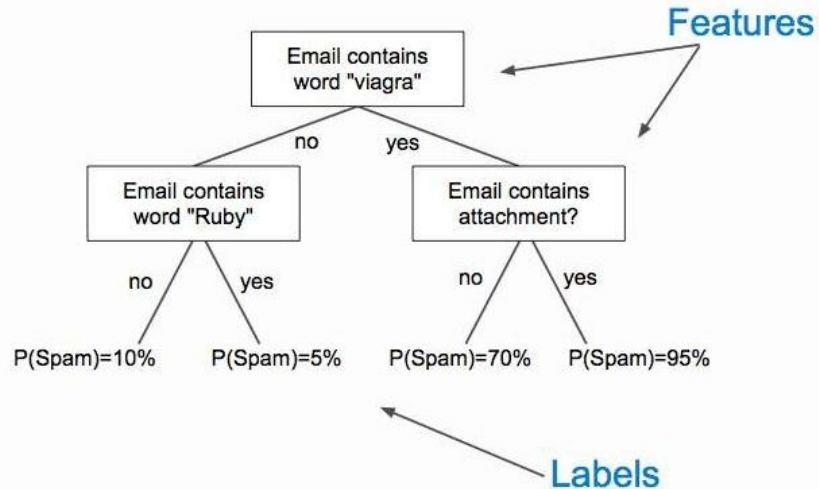
- Think carefully about what allows your brain to discern between the spam email and the non-spam email.
- Use of a title, Dr.
- Leaving a phone number - indicates lack of previous interaction potentially.
- Various words like developer, client, etc.
- Signing signature.

**Goal:** As a data scientist, can you turn these intuitive notions of 'spam' into well defined features to input into a classification model?

# A simple decision tree model for spam

## Algorithms:

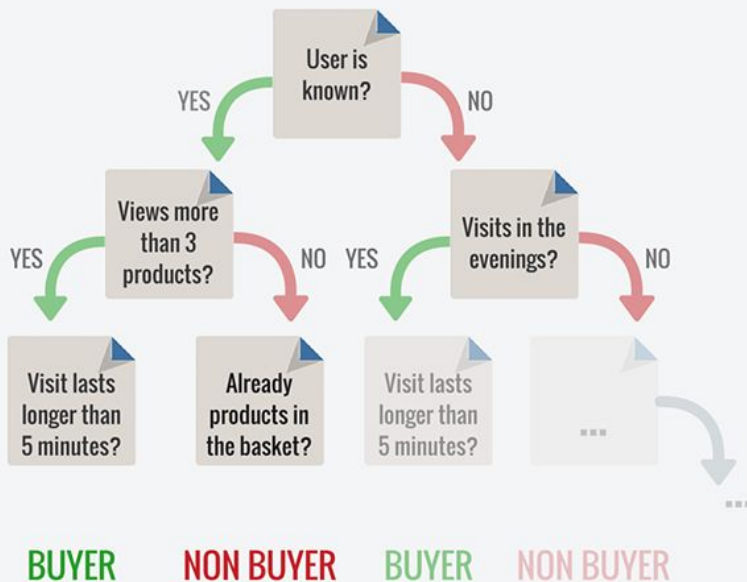
### Decision Tree Learning





# Buyers vs Non-Buyers

## Decision Tree: Buyers vs Non Buyers



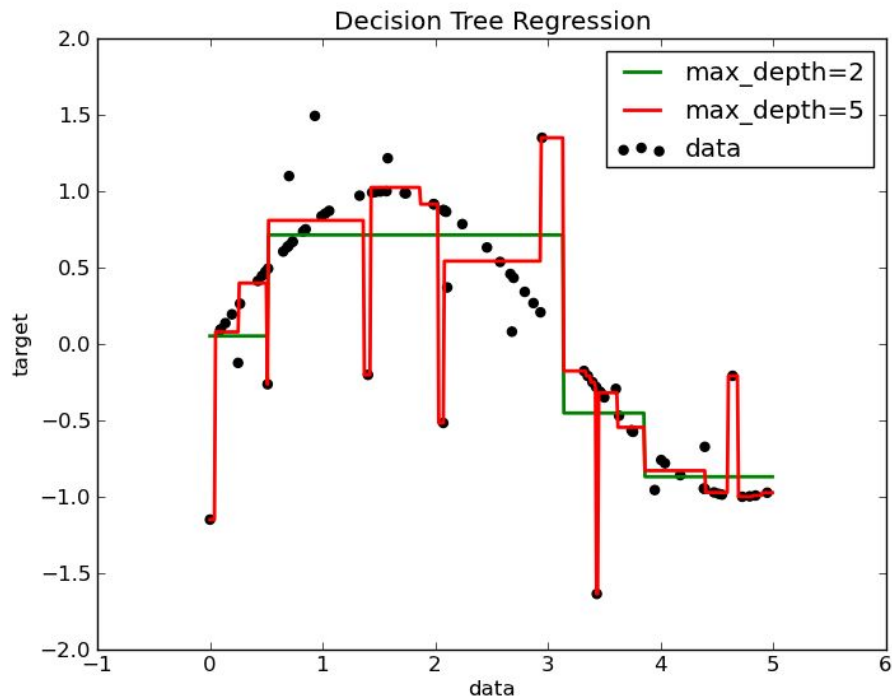
- But how deep should we make the tree?
- Is deeper better?

# How do we measure performance?

---

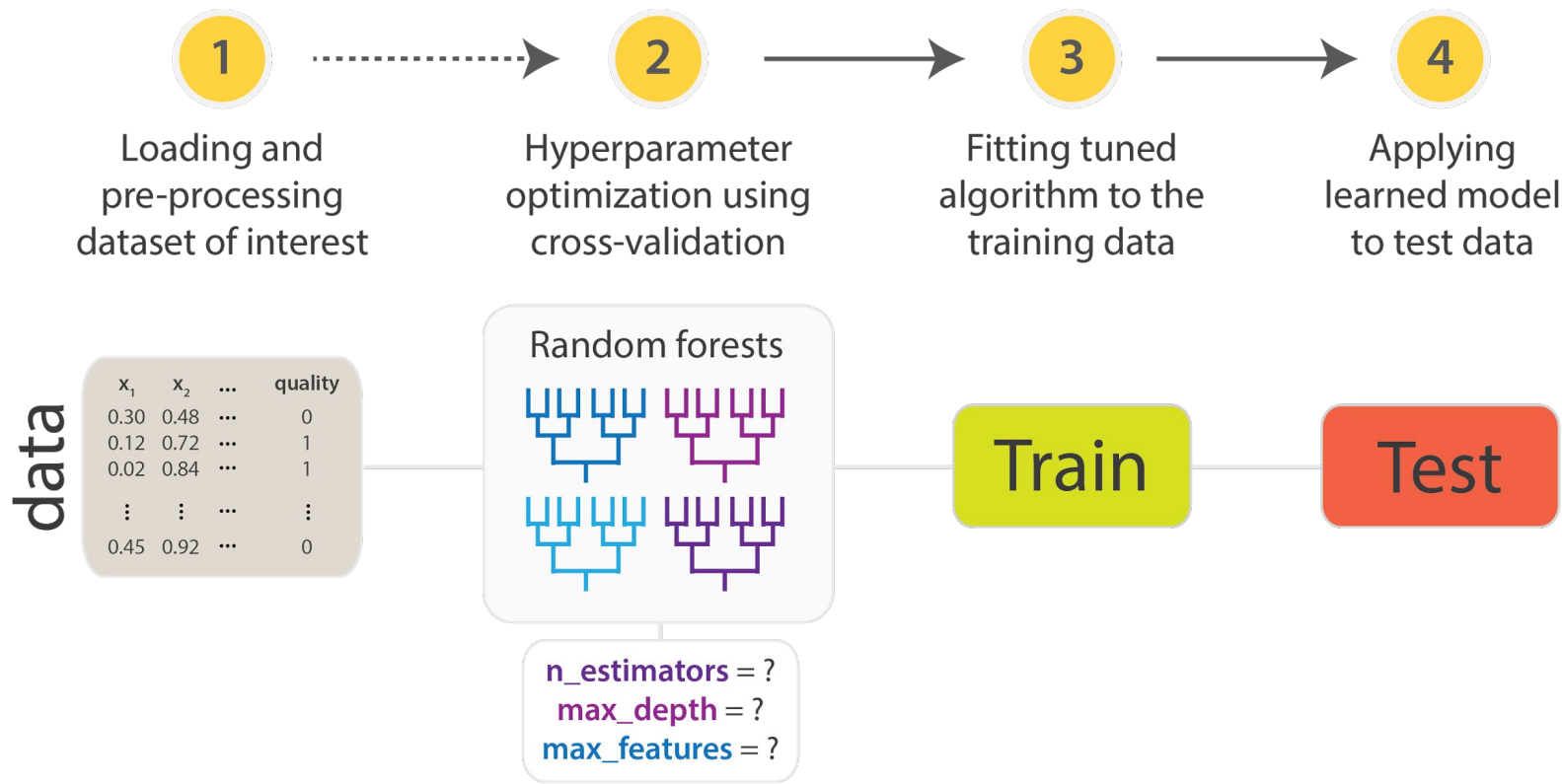
How complex of a model is ideal? What does ideal mean?

# ● How deep is too deep?

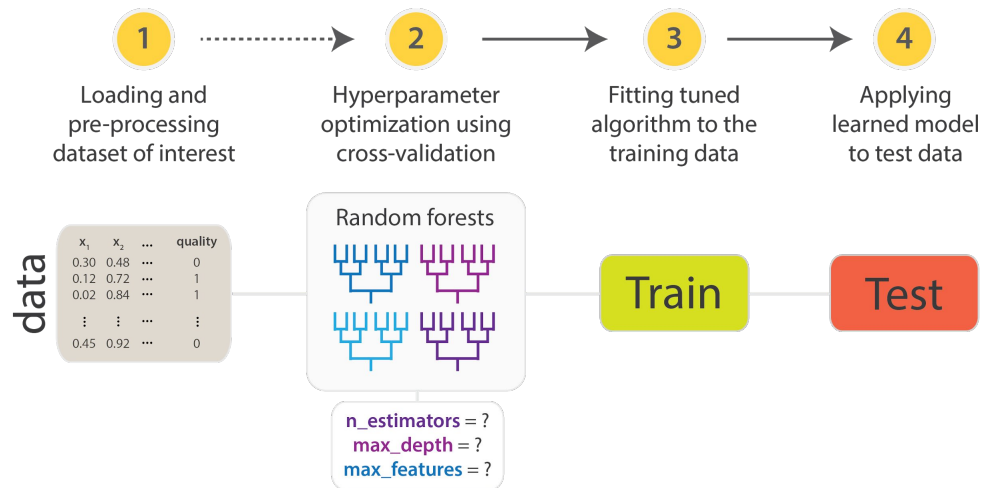


- **Not deep enough**, and your model will have **too much bias**.
- **Too deep** and your model will have **too much variance (overfitting)**.
- This is known as the **variance/bias tradeoff**.
- Your model will fit training data perfectly, but not fit the testing data.

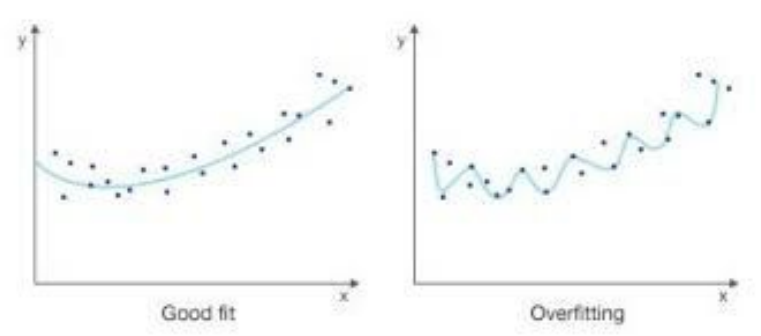
# What is the machine learning workflow?



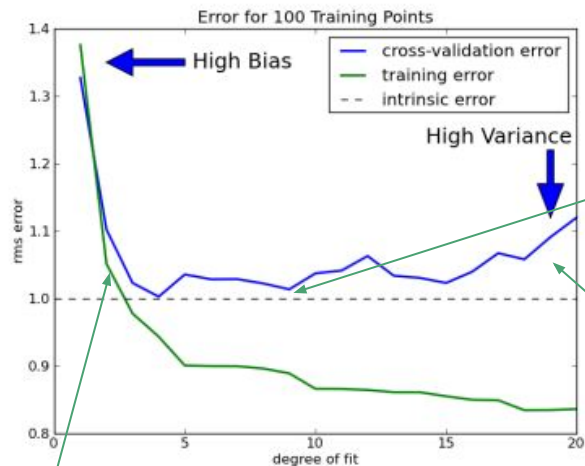
# How do we evaluate performance?



- We we train the model on different data than we evaluate it.
- The training data and testing data are sampled from similar distributions.
- We optimize the complexity of the model to prevent overfitting.

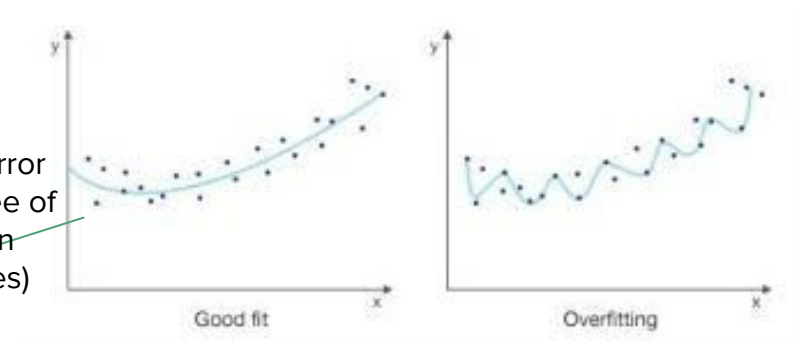


# What complexity is ideal?



Our model here is too simple, and has too high of a bias.

The lowest point here on testing error is our ideal degree of fit (depth of tree in previous examples)



We've gone too far here, since our testing error has started to increase. This means there is too much variance.

# References

**Main References:** These are references to deepen your understanding of material presented in lecture. The list is by no means exhaustive.

- Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, *An Introduction to Statistical Learning*, Springer 2013
- Trevor Hastie, Robert Tibshirani, Jerome Friedman, *Elements of Statistical Learning*, Springer 2013
- Christopher M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.
- Cameron Davidson-Pilon, *Bayesian Methods for Hackers*,  
<https://github.com/CamDavidsonPilon/Probabilistic-Programming-and-Bayesian-Methods-for-Hackers>

# Next Class

- Lecture 2: Introduction to Linear Regression
  - Homework discussion. Review of Linear Algebra.
  - Definition, Derivation and comparison of  $L_p$  norms. How does it affect the model?
  - Linear Regression and Derivation of Analytical Solution when  $p=2$ .
  - Model Training and Testing.
  - Gradient Descent, and Introduction to Convex Optimization.
  - Concrete Examples of Linear Regression and Gradient Descent in Python in an iPython Notebook.