COSC 3337 : Data Science I

N. Rizk

College of Natural and Applied Sciences

Department of Computer Science

University of Houston

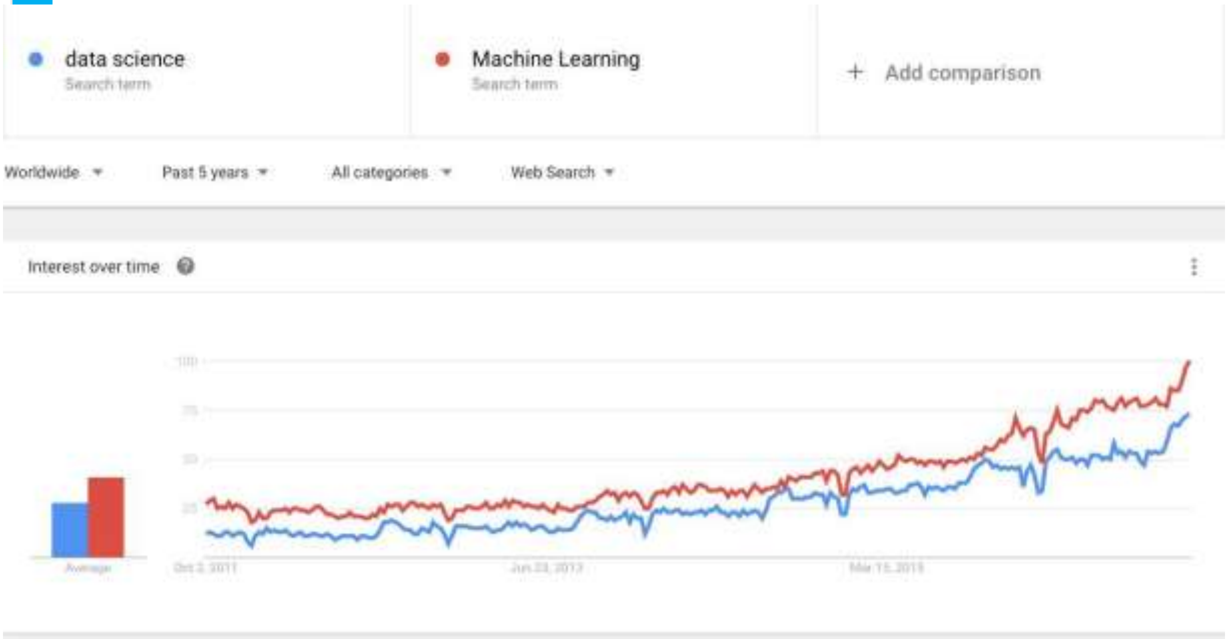
Overview

- 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?
- How do we learn from data? How do we measure performance?

What is Data Science?

- **Predictive (Supervised Learning):** The science of using data to predict an outcome (clicking, subscription, cancerous cells, price of a stock)
- **Descriptive (Unsupervised Learning):** Using data to group items/users into categories (ie. extract topics/categories from articles)
- **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.

Interest in Data Science is Blowing Up



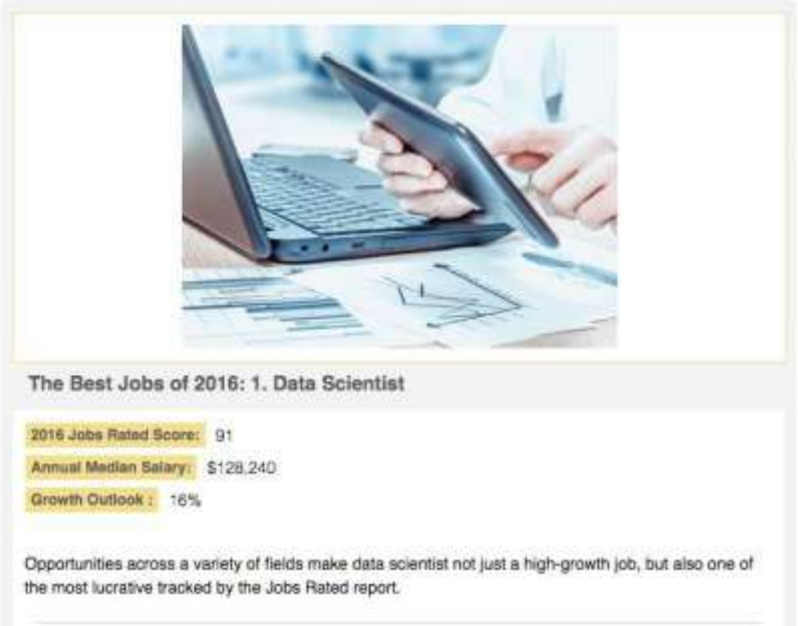
● data science
Search term

● Machine Learning
Search term

+ Add comparison

Worldwide ▾ Past 5 years ▾ All categories ▾ Web Search ▾

Interest over time



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.

- Intellectually rich landscape of problems in a relatively new field.
- Can save a company millions of dollars by implementing the right algorithm effectively, allowing us to have significant impact.

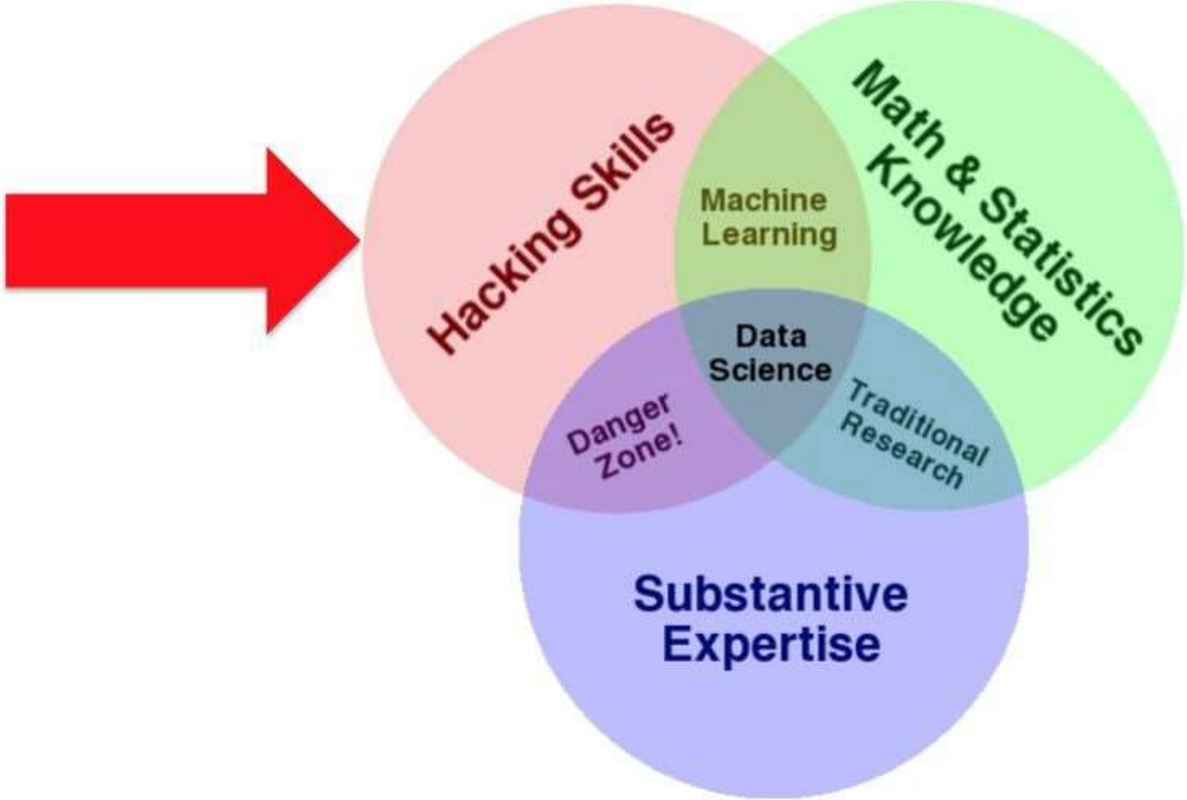
“"Anderson left Harvard before getting his PhD because he came to view the field much as Boykin does—as an intellectual pursuit of diminishing returns. But that’s not the case on the internet. “Implicit in ‘the internet’ is the scope, the coverage of it,” Anderson says. “It makes opportunities are much greater, but it also enriches the challenge space, the problem space. There is intellectual upside.”" - ***WIRED***

https://www.wired.com/2017/01/move-coders-physicists-will-soon-rule-silicon-valley/

But Data Science is losing its meaning

- Because of the popularity of data science, there are far too many “fake” data scientists.
- More and more candidates are graduating from data science masters programs without being able to answer simple questions about which model to use where.
- Don’t be a fake data scientist!

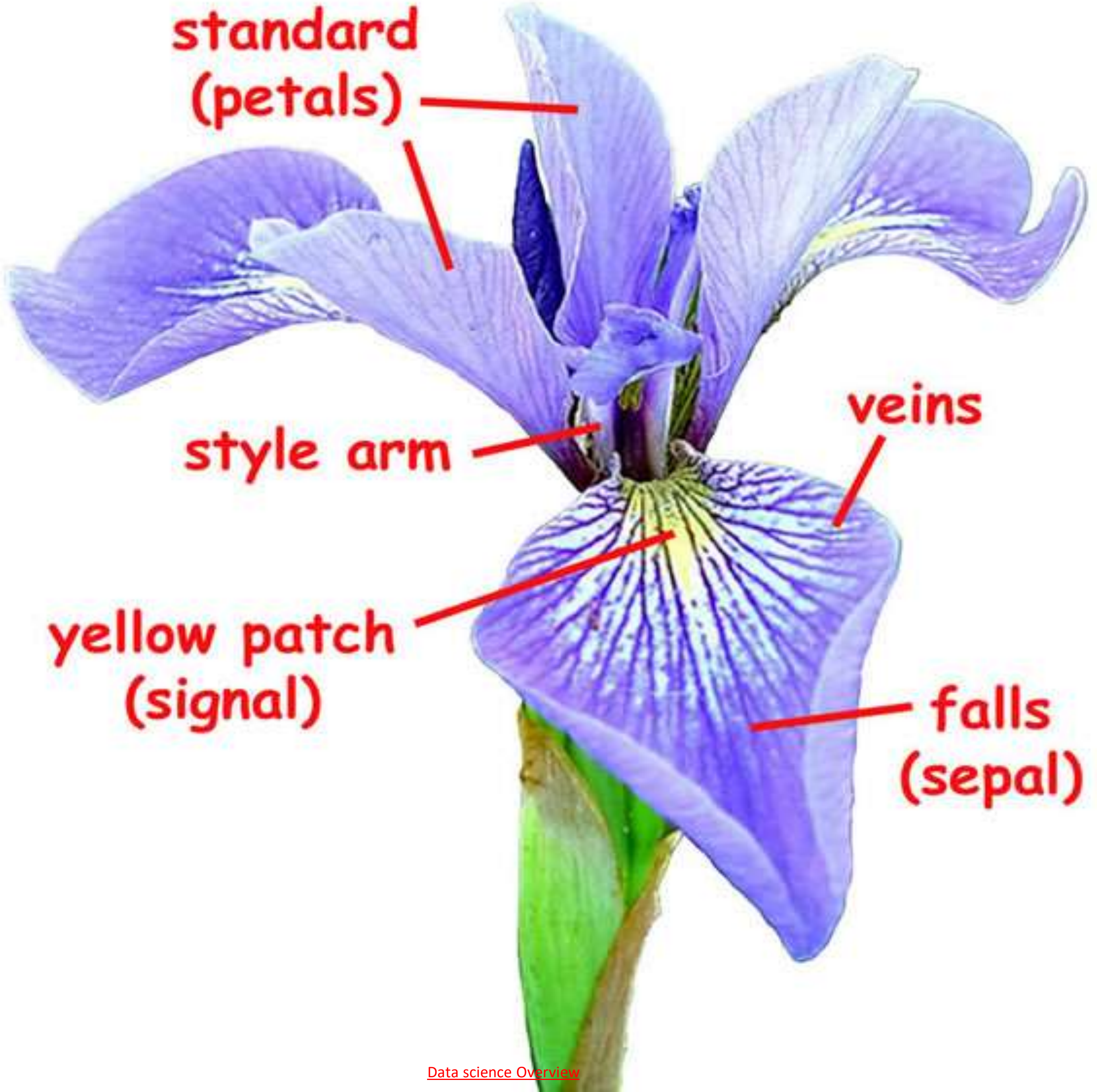
What is Data Science?



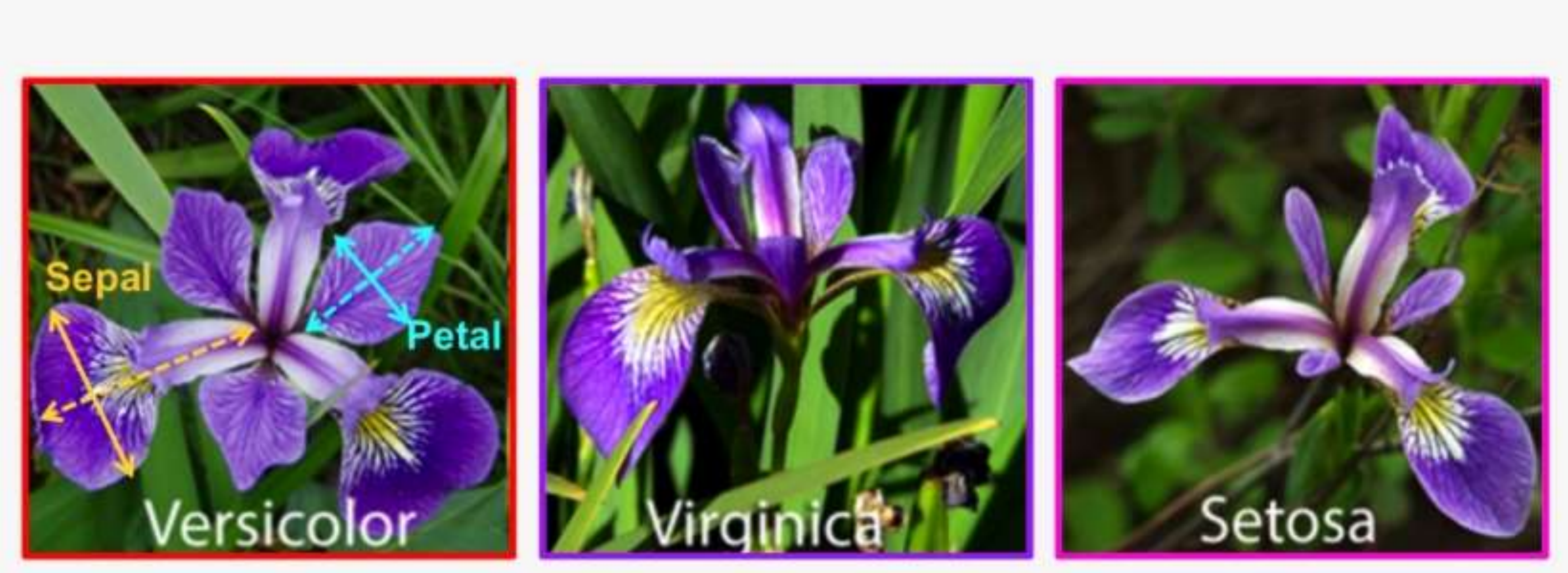
Math/Statistical Knowledge: Need understanding probability, statistics, optimization methods to create and use models.

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

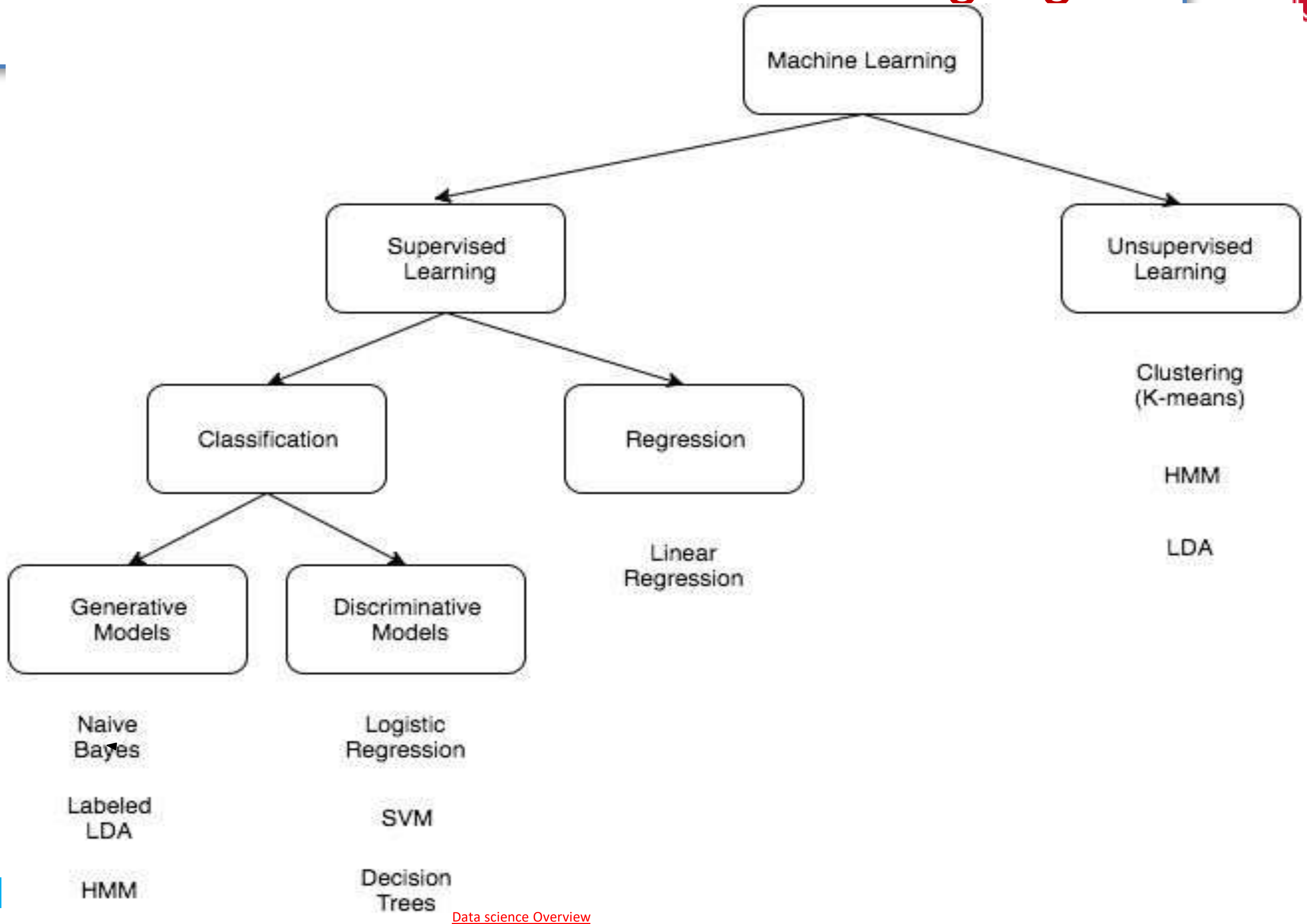
Substantive Experience: Need experience working with real data and business problems along with the problems that come along with them. Also need ability to communicate technical ideas to stakeholders.


standard
(petals)
style arm
veins
yellow patch
(signal)
falls
(sepal)

Different Kind of IRIS



How do we break down machine learning algorithms?



Machine Learning

Supervised Learning

Unsupervised Learning

Classification

Regression

Clustering (K-means)

HMM

LDA

Linear Regression

Generative Models

Discriminative Models

Naive Bayes

Labeled LDA

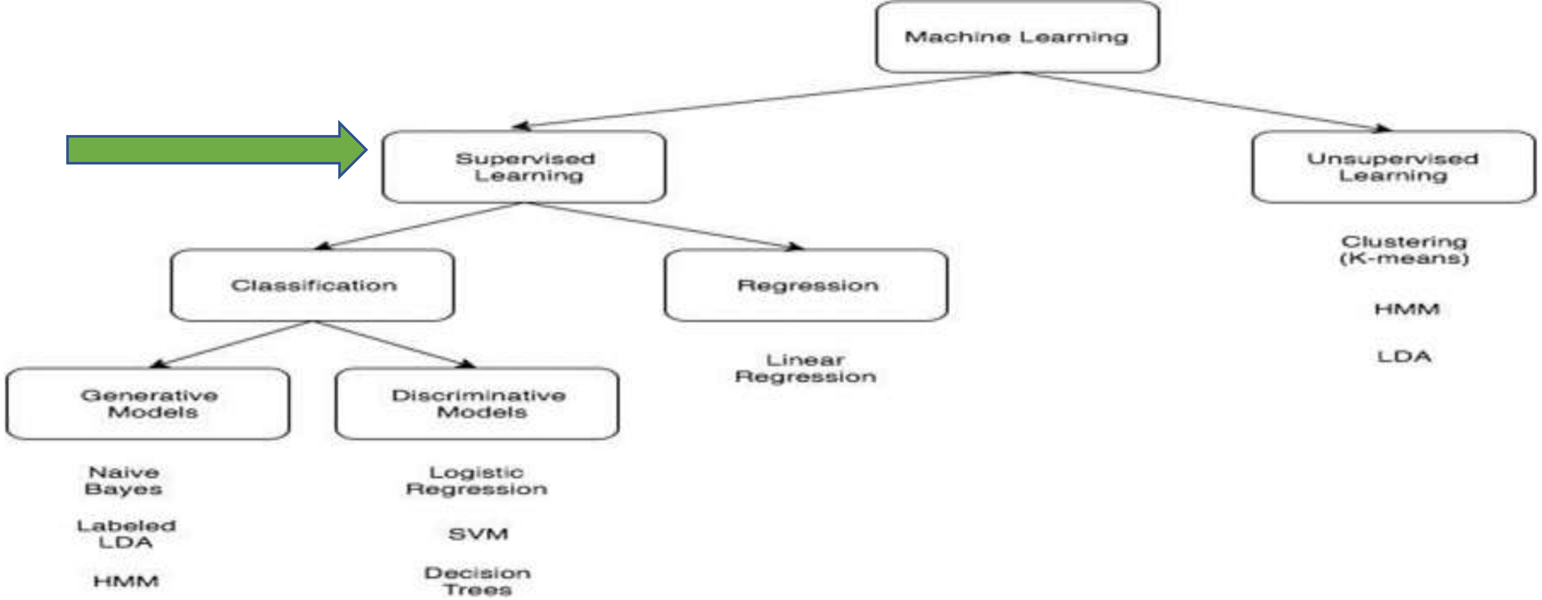
HMM

Logistic Regression

SVM

Decision Trees

Data science Overview



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"**.

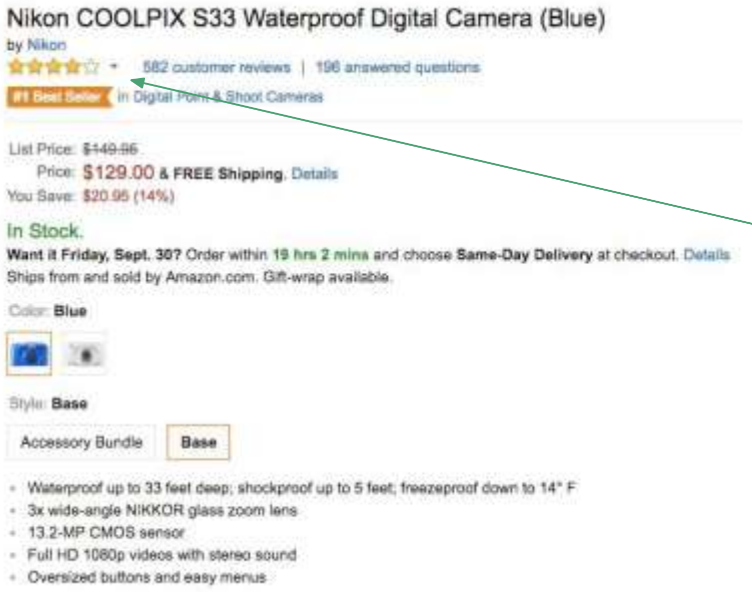
- **Algorithms:**
  - Linear Regression (Regression)/Logistic Regression(Classification).
  - Random Forest/Decision Trees/Gradient Boosting.
  - SVM (Support Vector Machines) and nonlinear kernels.
  - Recommendation Engines: Graph Diffusion, Matrix Factorization.
  - Gaussian Mixture models and Expectation Maximization.
  - Maximum Likelihood
  - Time Series Modeling
  - Neural Nets

Predictive Learning - Examples

(Supervised)

Amazon.com purchases

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: $20.96 (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**

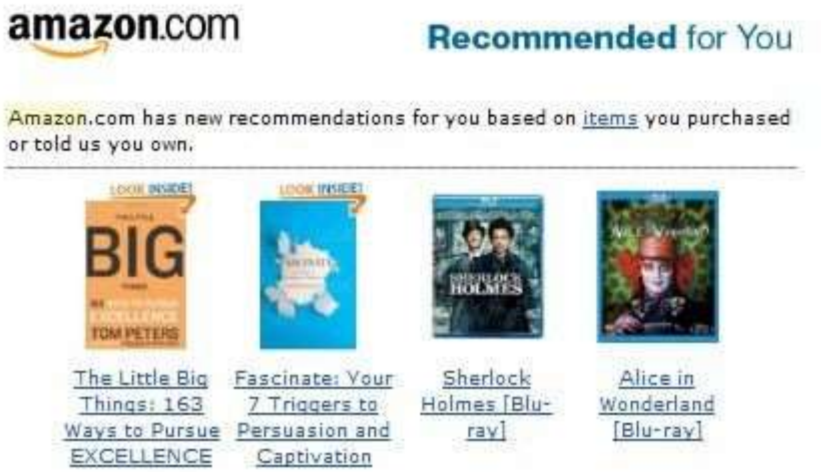
Accessory Bundle | **Base**

- 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 an experiment.
- **Good recommendations = $$.**

1. Generate the model, evaluate.
2. Run A/B test to measure performance or utility.
3. Learn from the model and improve.

How do we use our model? A/B Testing



amazon.com Recommended for You

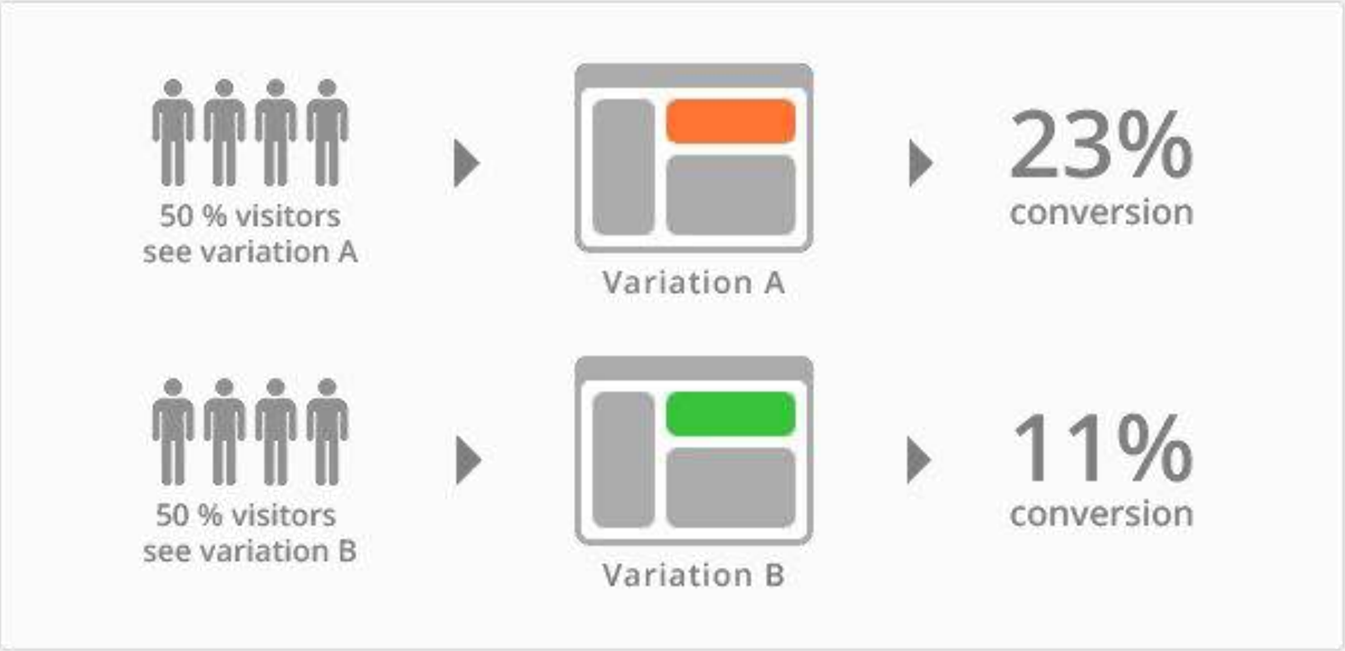
Amazon.com has new recommendations for you based on items you purchased or told us you own.

The Little Big Things: 163 Ways to Pursue EXCELLENCE

Fascinate: Your 7 Triggers to Persuasion and Captivation

Sherlock Holmes [Blu-ray]

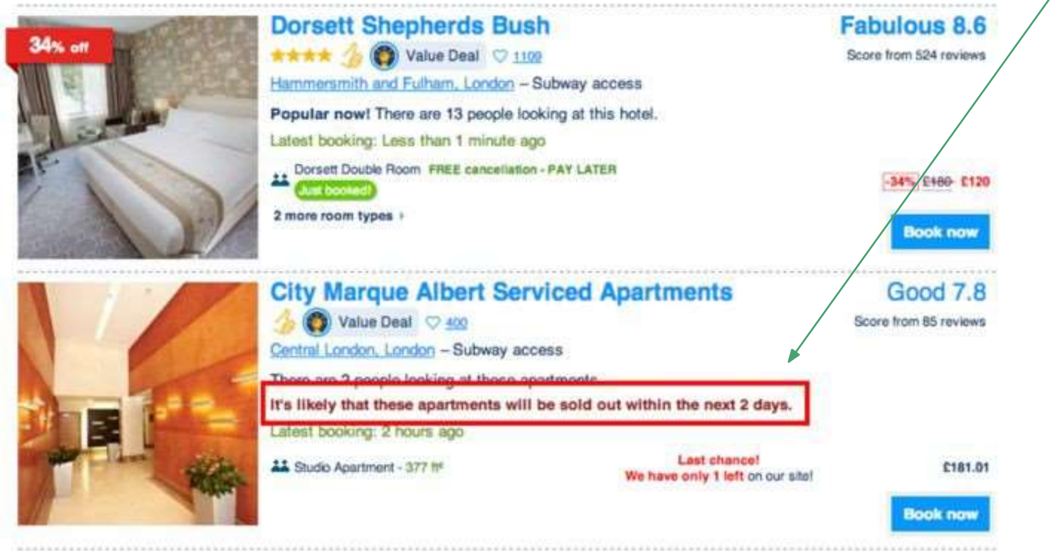
Alice in Wonderland [Blu-ray]



- Our model suggestions →
- Top items (BAU) →

Booking.com hotel bookings

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



Dorsett Shepherds Bush

★★★★ Value Deal ♡ 1109

Fabulous 8.6

Score from 524 reviews

Hammersmith and Fulham, London – Subway access

Popular now! There are 13 people looking at this hotel.

Latest booking: Less than 1 minute ago

Dorsett Double Room FREE cancellation - PAY LATER

Just booked!

2 more room types ›

-34% ~~£180~~ £120

City Marque Albert Serviced Apartments

Value Deal ♡ 400

Good 7.8

Score from 85 reviews

Central London, London – Subway access

There are 2 people looking at these apartments.

It's likely that these apartments will be sold out within the next 2 days.

Latest booking: 2 hours ago

Studio Apartment - 377 ft²

Last chance!
We have only 1 left on our site!

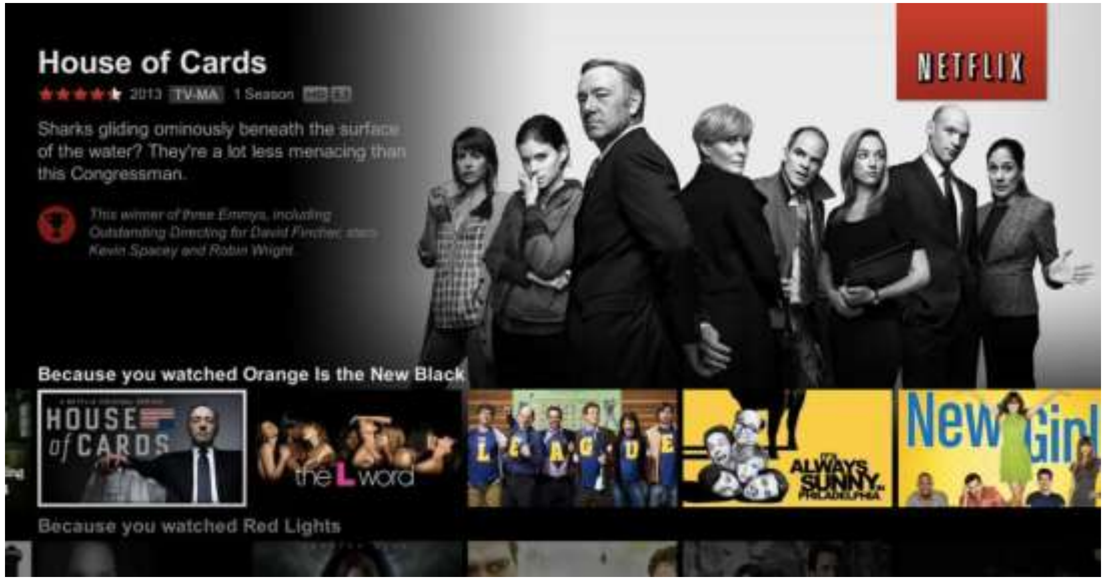
£181.01

- **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).

Netflix.com movie ratings

Can we predict how you would rate a movie?



NETFLIX
House of Cards
2013 TV-MA 1 Season

Sharks gliding ominously beneath the surface of the water? They're a lot less menacing than this Congressman.

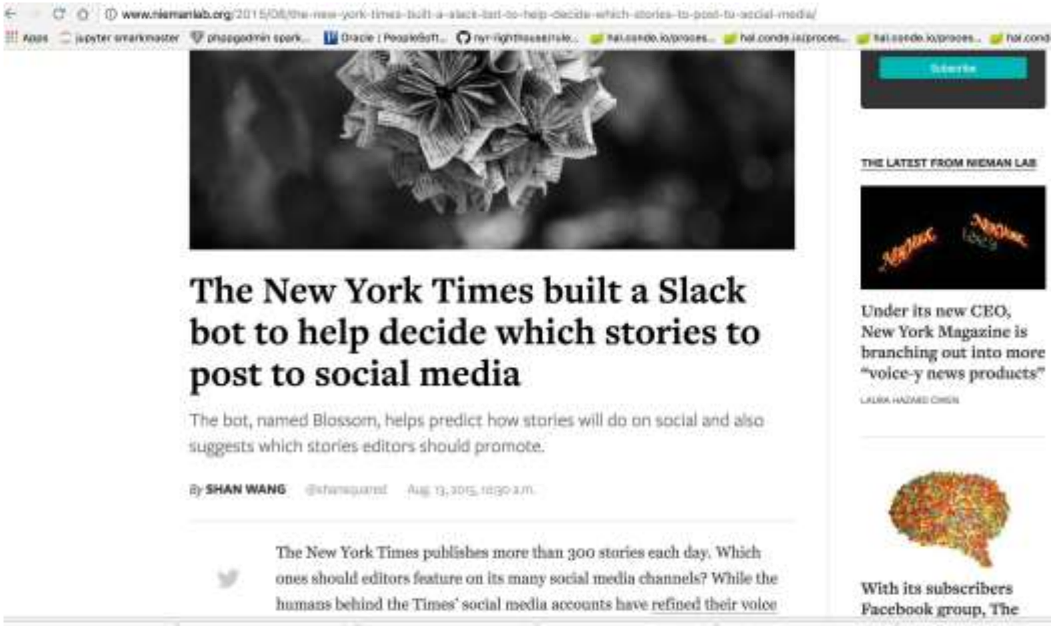
This winner of three Emmys, including Outstanding Directing for David Fincher, stars Kevin Spacey and Robin Wright.

Because you watched Orange Is the New Black
HOUSE of CARDS
the L word
LEAGUE
IT'S ALWAYS SUNNY IN PHILADELPHIA
New Girl
Because you watched Red Lights

- **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 $$.

Predicting viral content



The New York Times built a Slack bot to help decide which stories to post to social media

The bot, named Blossom, helps predict how stories will do on social and also suggests which stories editors should promote.

By SHAN WANG @shansquared Aug. 13, 2015, 10:30 a.m.

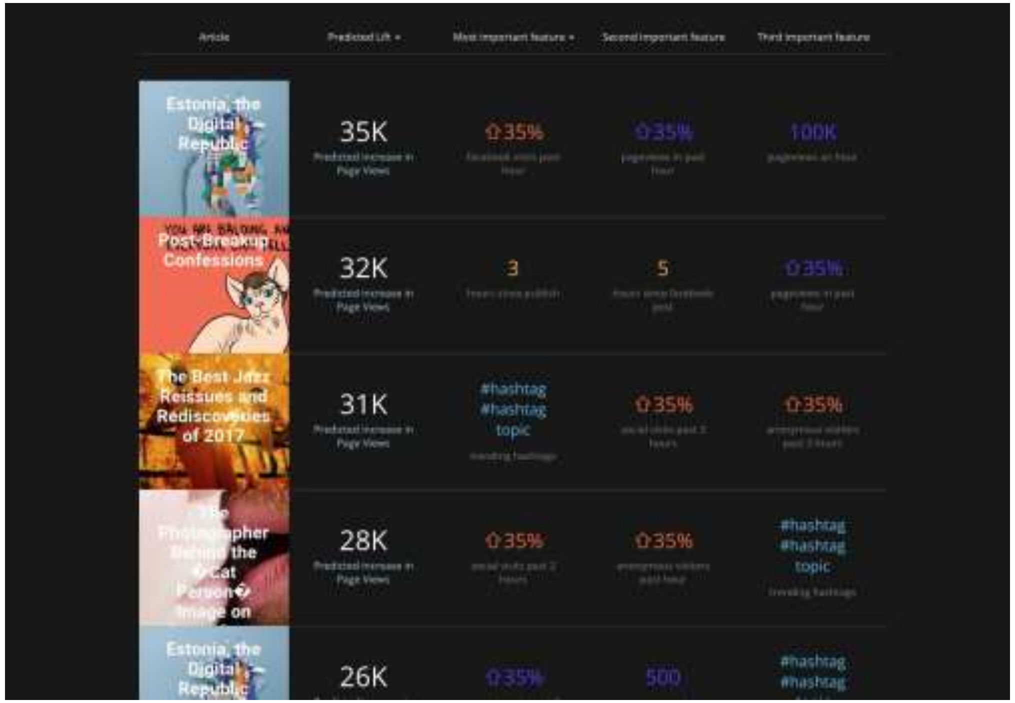
The New York Times publishes more than 300 stories each day. Which ones should editors feature on its many social media channels? While the humans behind the Times' social media accounts have refined their voice

THE LATEST FROM NIEMAN LAB

Under its new CEO, New York Magazine is branching out into more "voice-y news products"

LAURA HAZARD OWEN

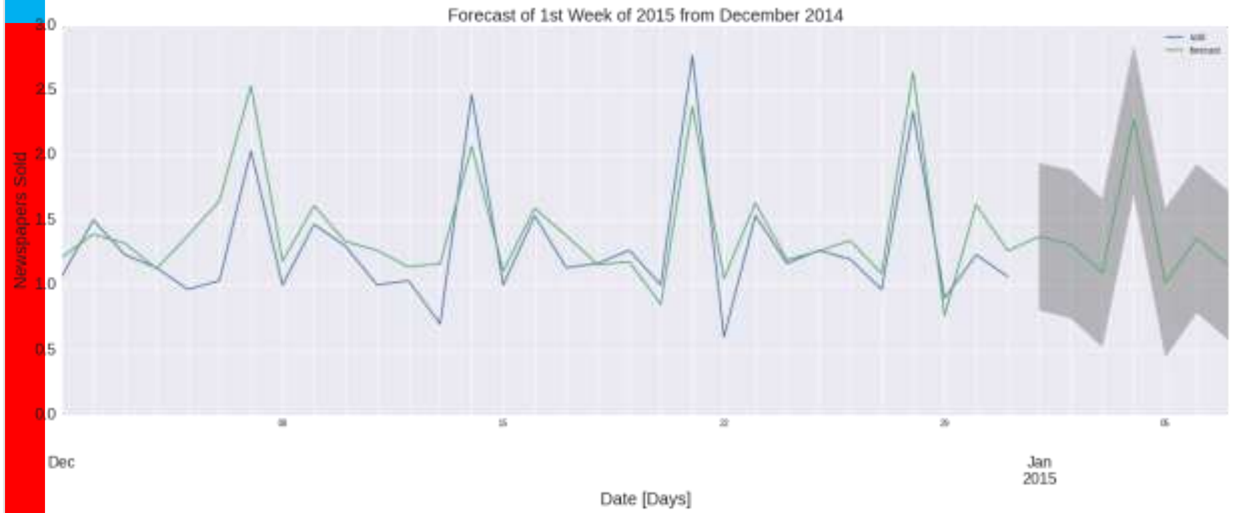
With its subscribers Facebook group, The



Article	Predicted Lift ▾	Most important feature ▾	Second important feature	Third important feature
Estonia, the Digital Republic	35K Predicted increase in Page Views	⇧35% facebook visits past hour	⇧35% pageviews in past hour	100K pageviews an hour
Post-Breakup Confessions	32K Predicted increase in Page Views	3 hours since publish	5 hours since facebook post	⇧35% pageviews in past hour
The Best Jazz Reissues and Rediscoveries of 2017	31K Predicted increase in Page Views	#hashtag #hashtag topic trending hashtags	⇧35% social visits past 2 hours	⇧35% anonymous visitors past 3 hours
The Photographer Behind the �Cat Person� Image on	28K Predicted increase in Page Views	⇧35% social visits past 2 hours	⇧35% anonymous visitors past hour	#hashtag #hashtag topic trending hashtags
Estonia, the Digital Republic	26K	⇧35%	500	#hashtag #hashtag

- Which content will go viral? (predictive)
- Where is the optimal place to post it? Twitter, Facebook? (prescriptive)

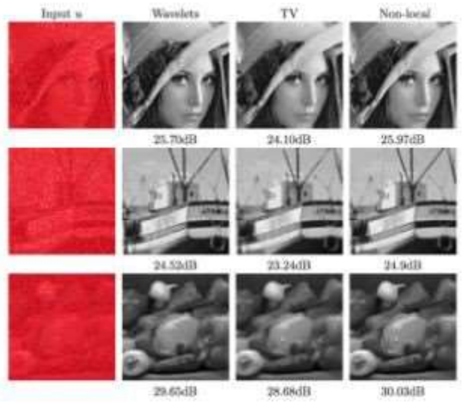
Optimizing paper distribution



- **Can we optimize profits by knowing how many papers to deliver to each Starbucks across America?**
- Answer: Yes!
- Problem involves profit optimization, time series regression, maximum likelihood methods and running live experiments to evaluate performance.

Modern Artificial Intelligence

Artificial Intelligence



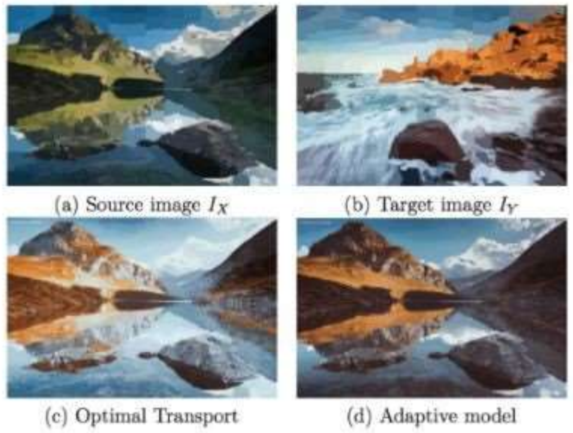

(a) Source image I_X
(b) Target image I_Y
(c) Optimal Transport
(d) Adaptive model

Image classification, segmentation, denoising, and many more applications! (Probably not at the level of this course but we will see!

Predictive Learning - Methods

(Supervised)

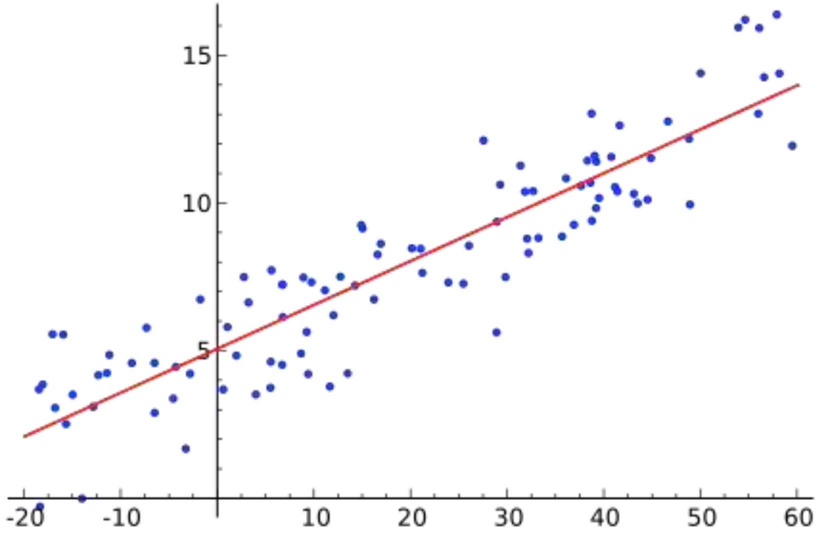
Predictive Learning - Linear Regression

Given a collection of points to learn from:

$$(x_i, y_i)$$

Can we find a function minimizing the distance to the data.

$$f : X \to Y$$



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

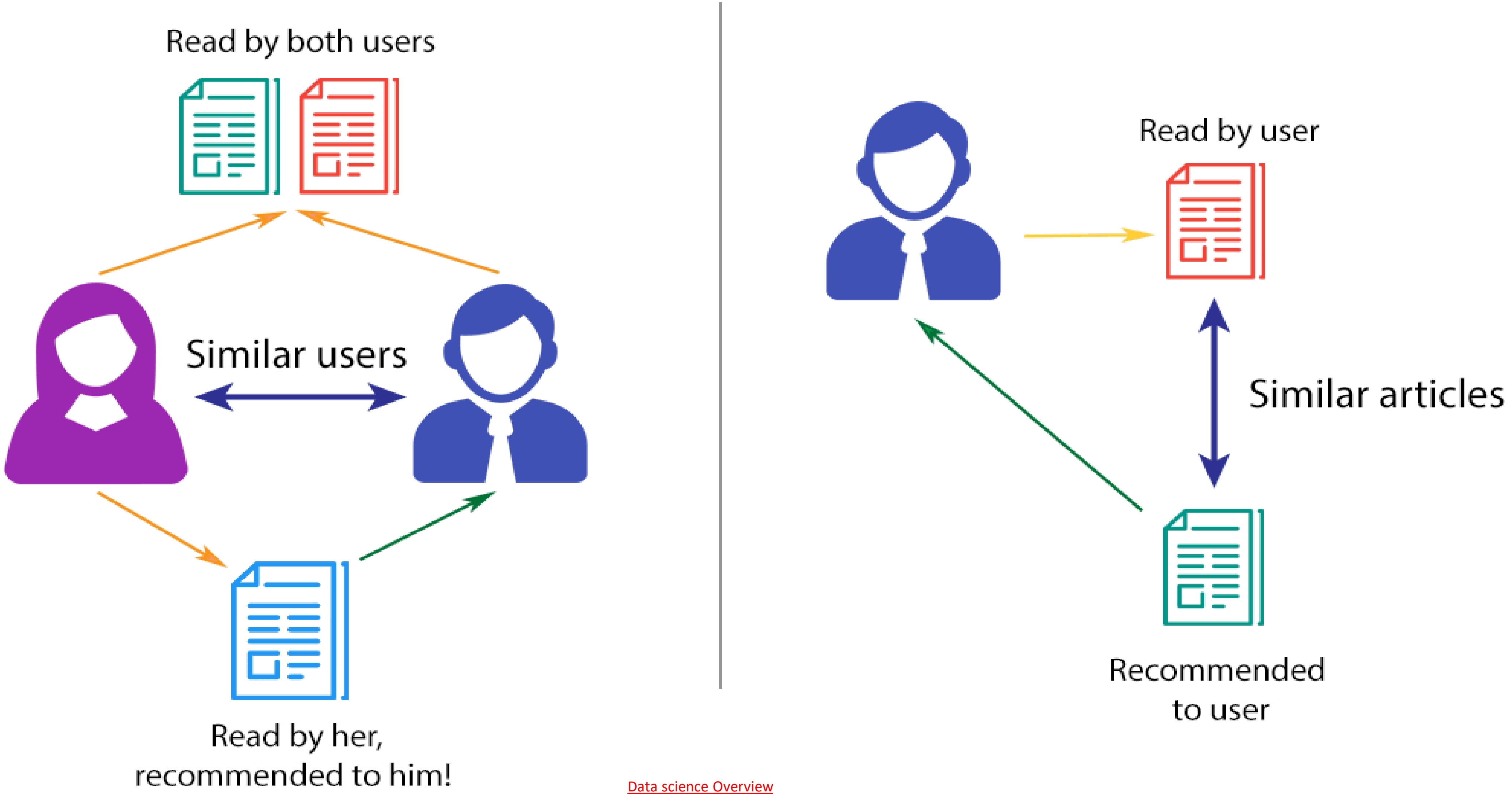
Linear: $f(x_i) = \beta \cdot x_i$

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

COLLABORATIVE FILTERING

CONTENT-BASED FILTERING



Read by both users

Similar users

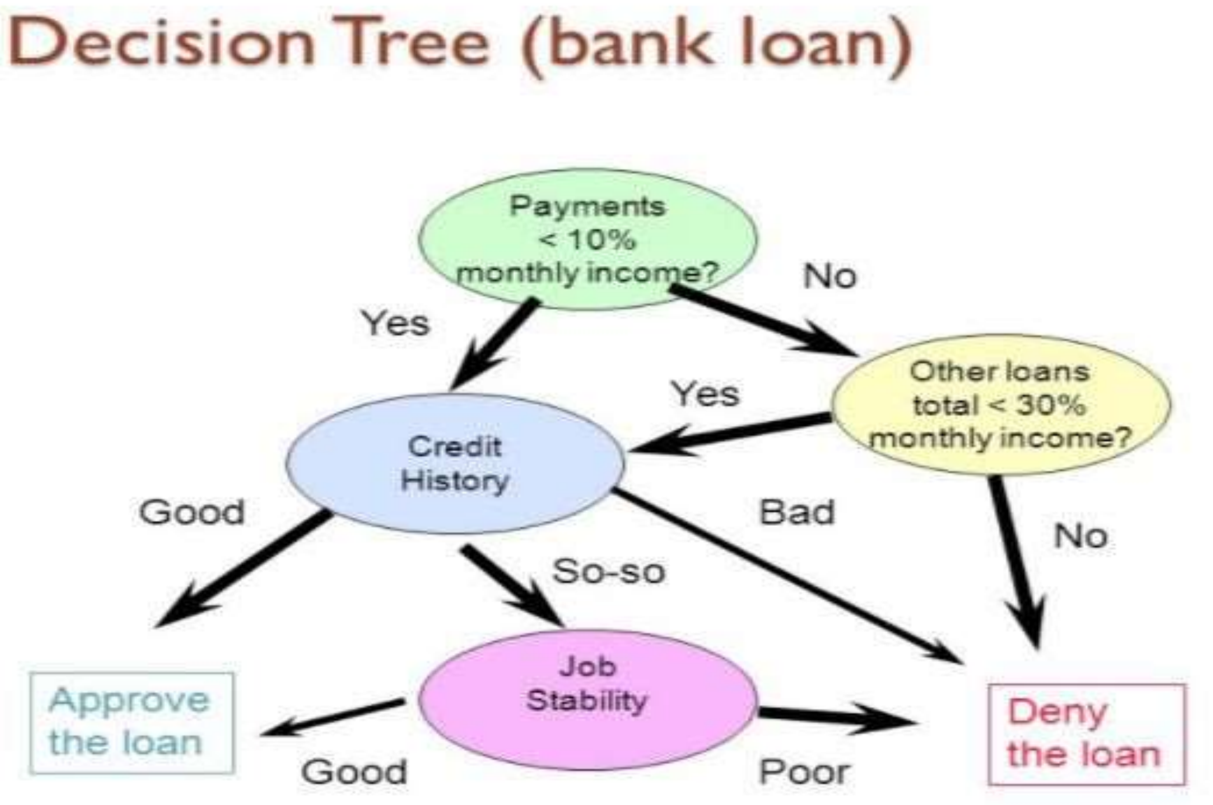
Read by her, recommended to him!

Read by user

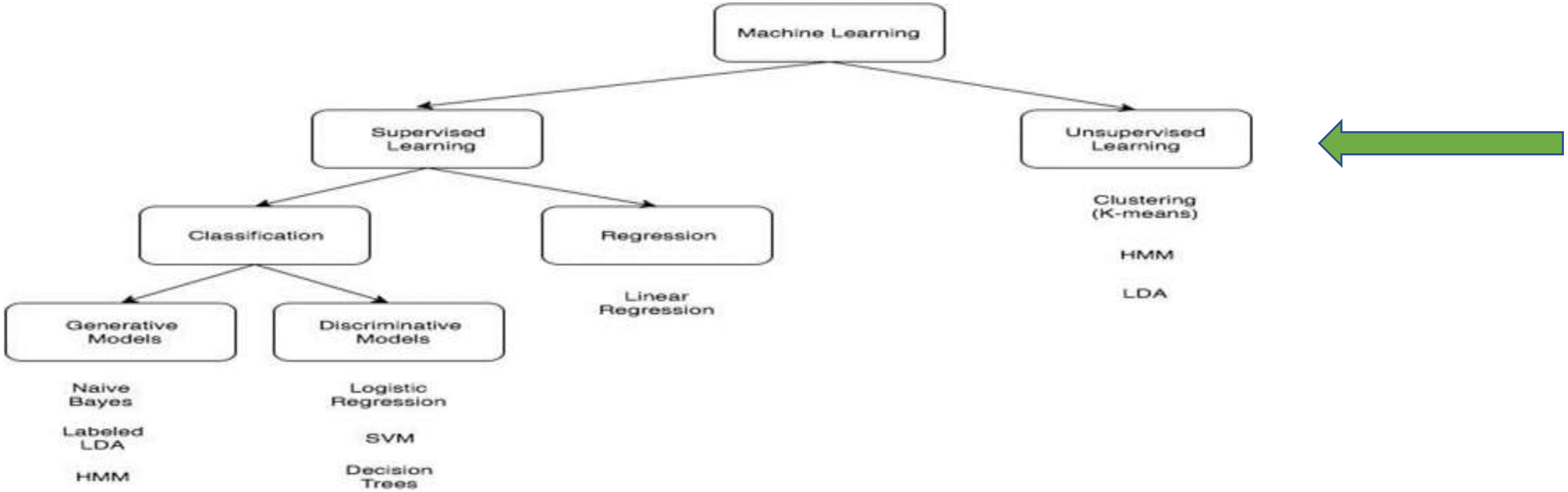
Similar articles

Recommended to user

Predictive Learning - Decision Trees


Decision Tree (bank loan)
Payments < 10% monthly income?
Yes
No
Other loans total < 30% monthly income?
Yes
No
Credit History
Good
So-so
Bad
Job Stability
Good
Poor
Approve the loan
Deny the loan

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



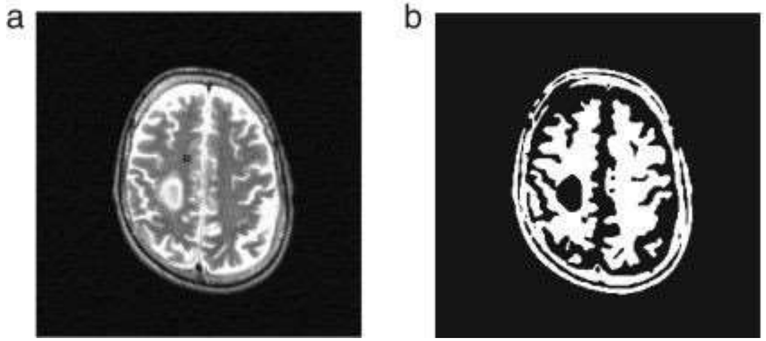
Descriptive Learning

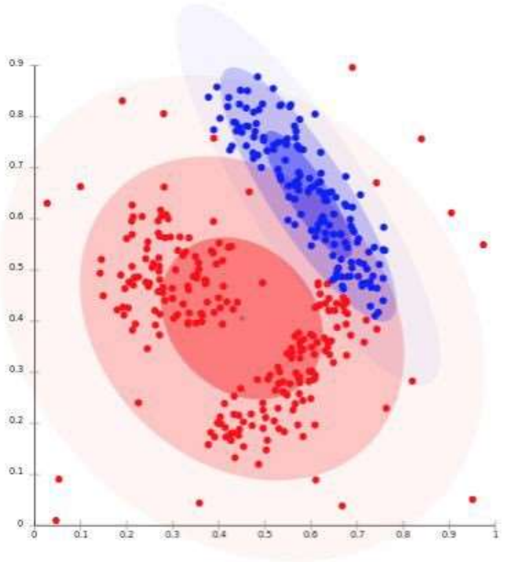
(Unsupervised)

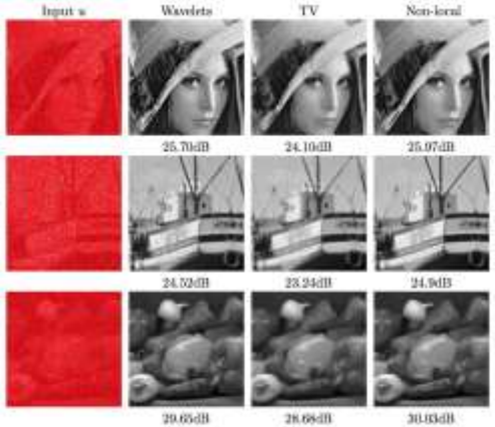
Descriptive Learning

- Try to infer a hidden structure in the data without proper training examples (no teacher, hence ‘unsupervised’).

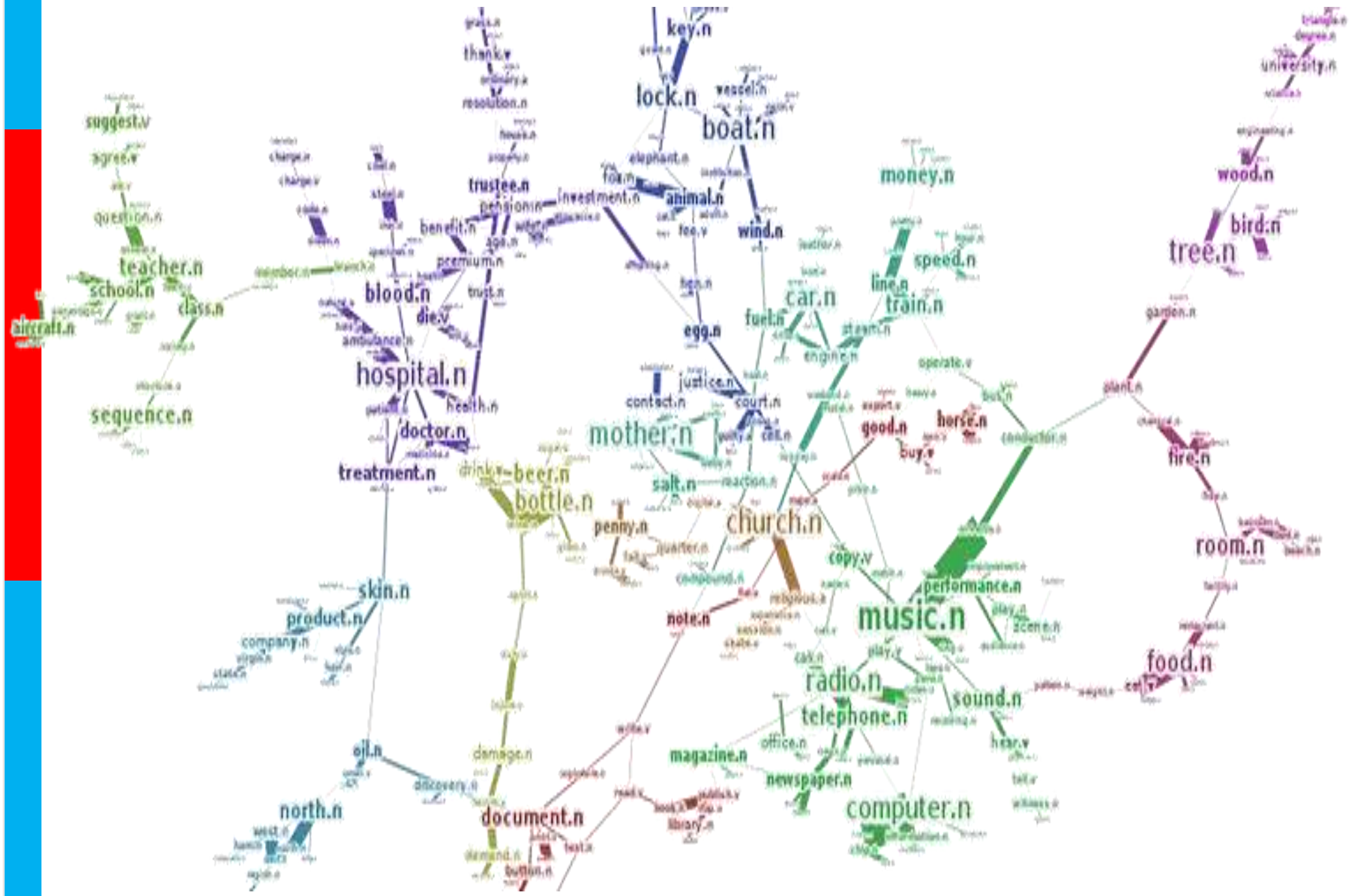
- **Algorithms:**
  - K-means clustering.
  - Decision Tree Clustering.
  - SVM
  - Topic Models (LDA, etc)
  - Gaussian Mixture Models







Topic Models

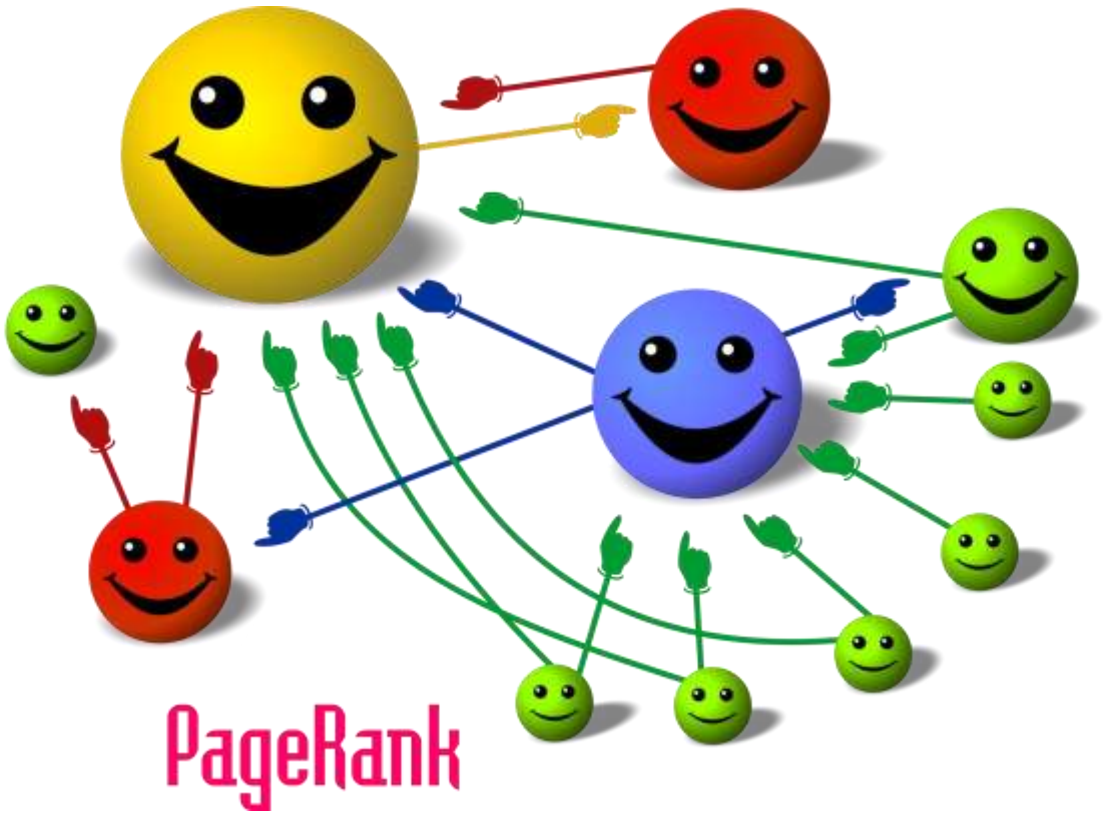


- Topic models attempt to cluster content into a finite collection of ‘topics’.

- The model creates topic categories by clustering words commonly occurring together into groups (roughly speaking).

- Reading/Writing behavior, while unsupervised, has tremendous predictive power for many algorithms in practice.

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)

Prescriptive Learning

Reinforcement Learning

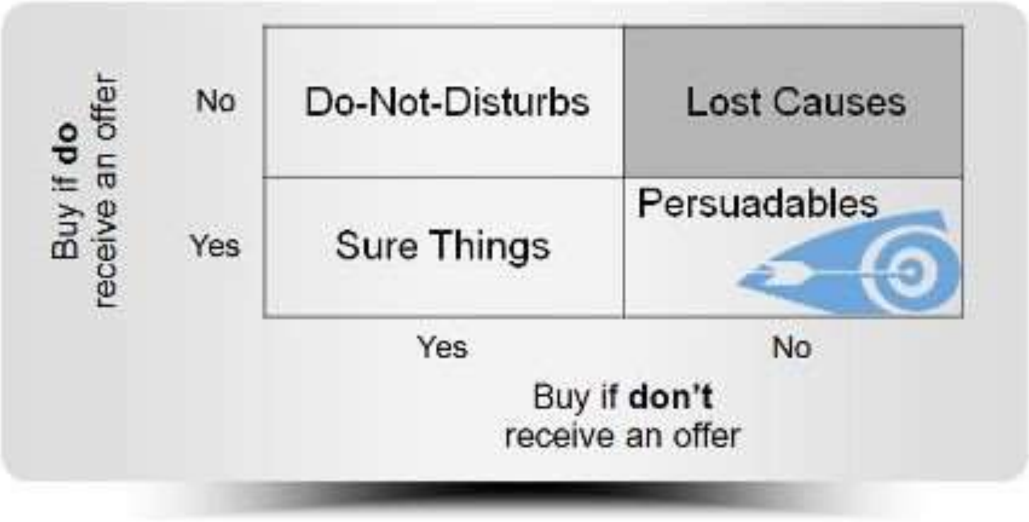
Prescriptive Learning

- Prescriptive learning attempts to find an **optimal action** 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?).

- A well defined metric is used to determine the performance of such a model.

- **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.

Example: Uplift Modeling



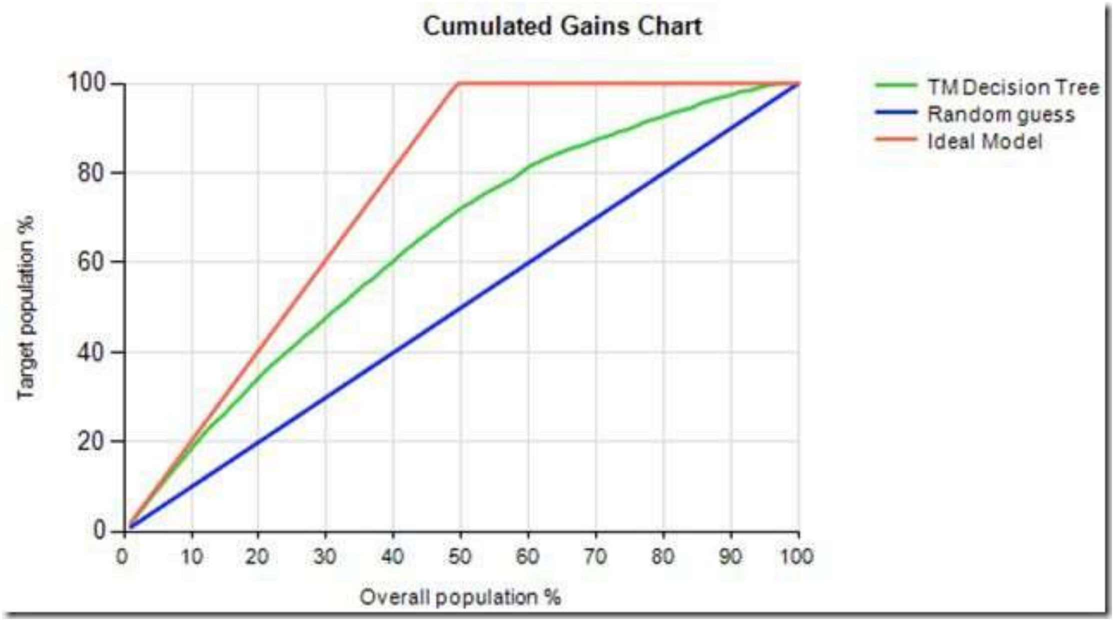
Buy if **do** receive an offer

No	Do-Not-Disturbs	Lost Causes
Yes	Sure Things	Persuadables
	Yes	No

Buy if **don't** receive an offer

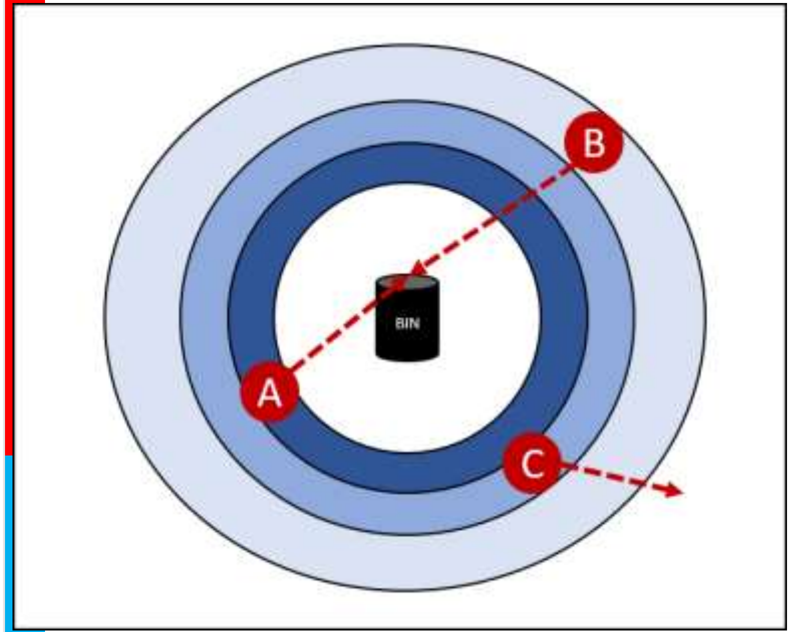
- Not everyone should receive the same action.
- Will users leave buy if they receive an offer?

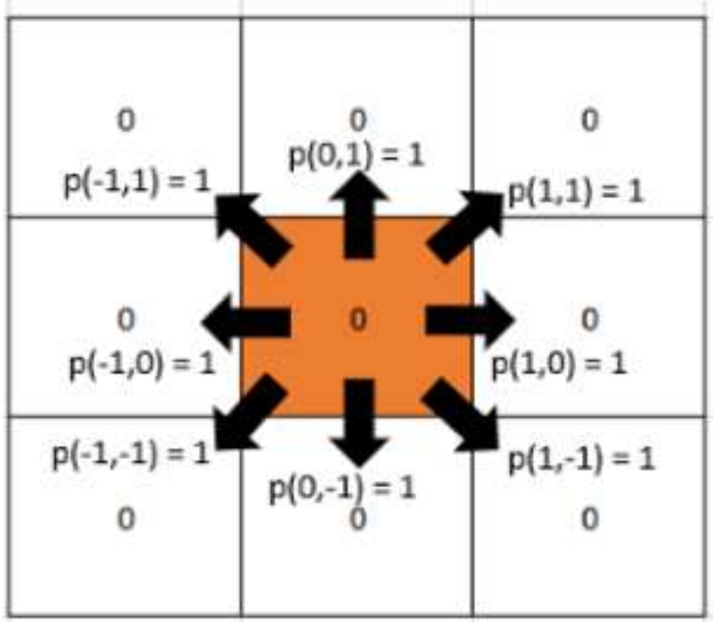
How do we determine the right action to maximize our desired outcome?

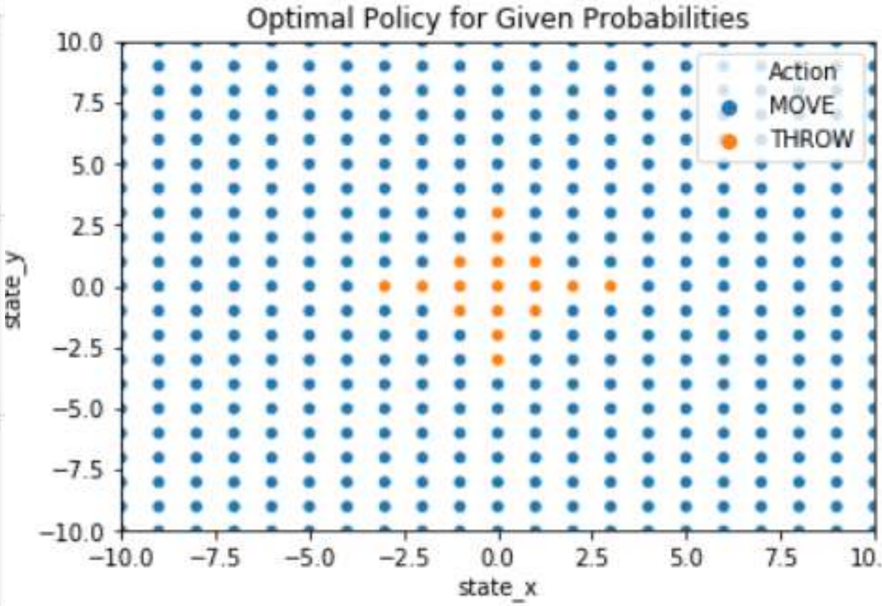


Percentage of people who bought

Example: Optimal Policy for Environment where the Probabilities are known

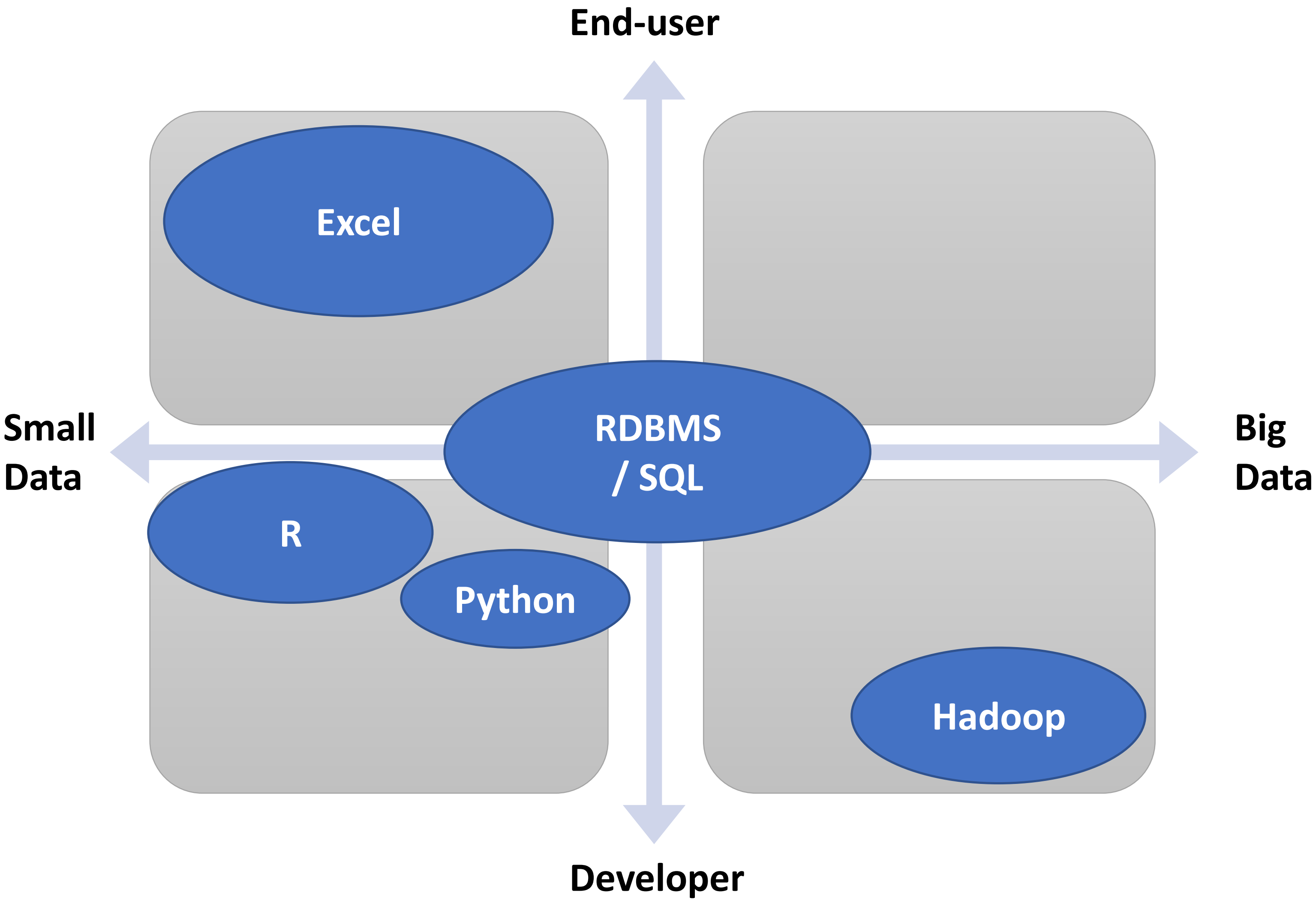






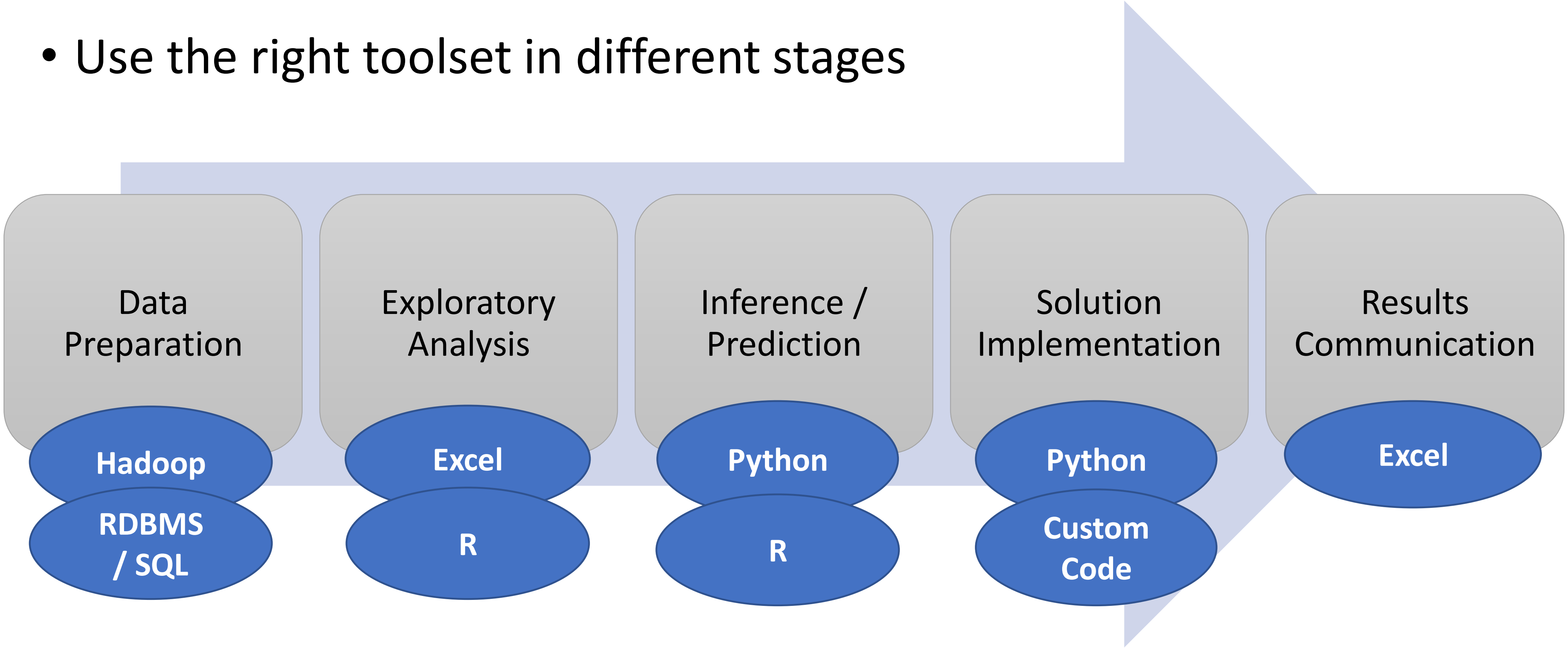
Tools?

Choosing Tools for Data Science



Chaining Tools for Data Science

- Use the right toolset in different stages



Make sure you check for quality issues!

Completeness

- Is the data representative of the problem space?
- Any missing observations / attributes?

Fidelity

- Do the measurements capture the reality?
- Any issues of bias or variance?

Consistency

- Are values follow data types specified?
- Do different attributes agree with each other?