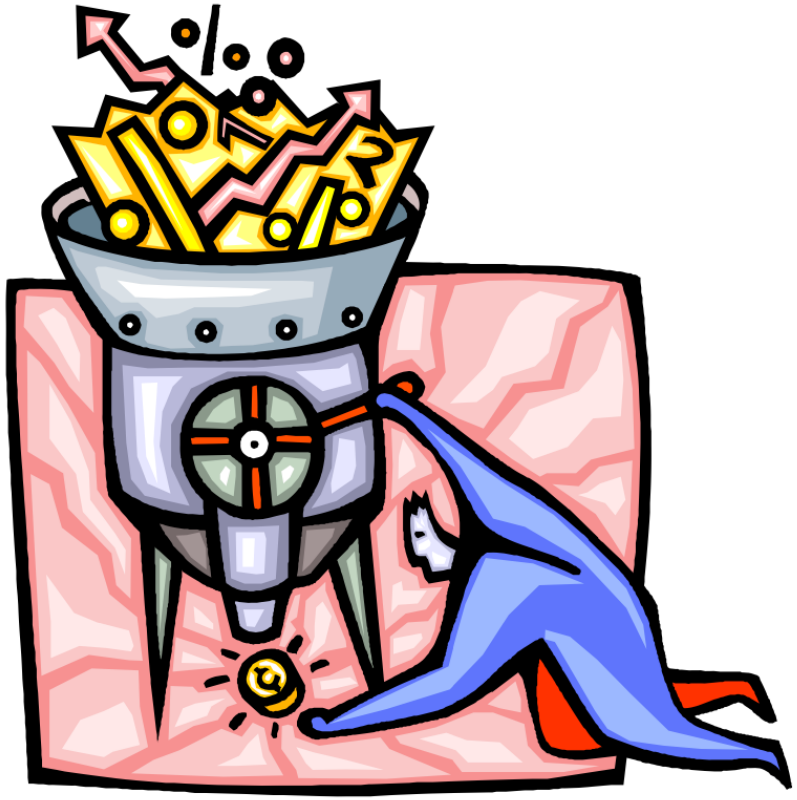


DATA 180

Introduction to Data Science

What is Data Science?

Data science is usually loosely defined as a collection of methods for managing and extracting information from data. It is an emerging field of study, with roots in statistics, economics, and computer science and draws heavily on techniques from these areas.

While not a comprehensive list, the most important topics in the field are generally considered to be:

- Data Wrangling
- Data Visualization
- Statistical (Machine) Learning
- Professional Ethics


Data Science
Statistics

Data Wrangling

The term data wrangling (also sometimes called data munging) typically refers to a set of techniques for manipulating data in preparation of the application of downstream data science techniques.

It can include:

- Data Retrieval, Aggregation, Merging, and Filtering
  - Web-scraping
- Data Cleaning
- Data Transformations

Some organizations employ *data engineers* whose primary job is to work upstream from data scientists to design, organize, and facilitate data pipelines.

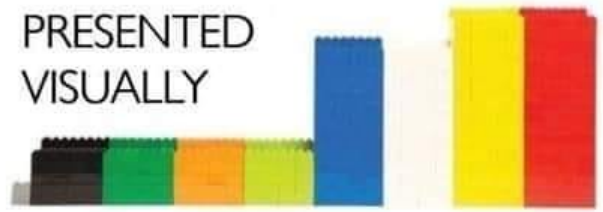
- How to store data?

DATA

SORTED

ARRANGED

PRESENTED
VISUALLY



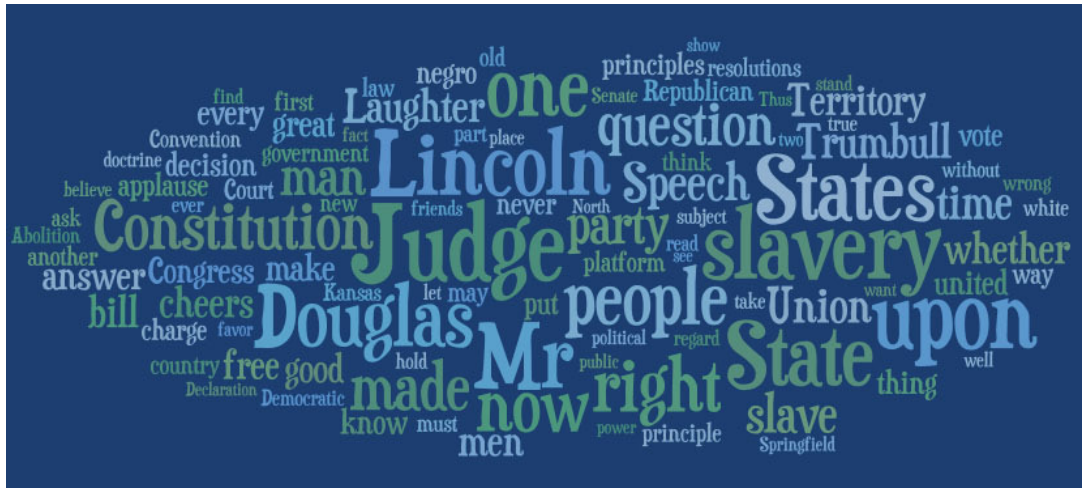
EXPLAINED
WITH A STORY

Data Visualization

Data visualization has been a staple in data analysis, with clear examples extending back into the 18th century (and earlier depending on the definition). Graphics activate sections of the brain in different ways than tables and charts, and visual processing is itself a rich and fascinating field.

The advent of modern computing and display technology has facilitated the production of visual expressions of data. In many cases graphics are now becoming dynamic and interactive.

"The greatest value of a picture is when it forces us to notice what we never expected to see." –J. Tukey

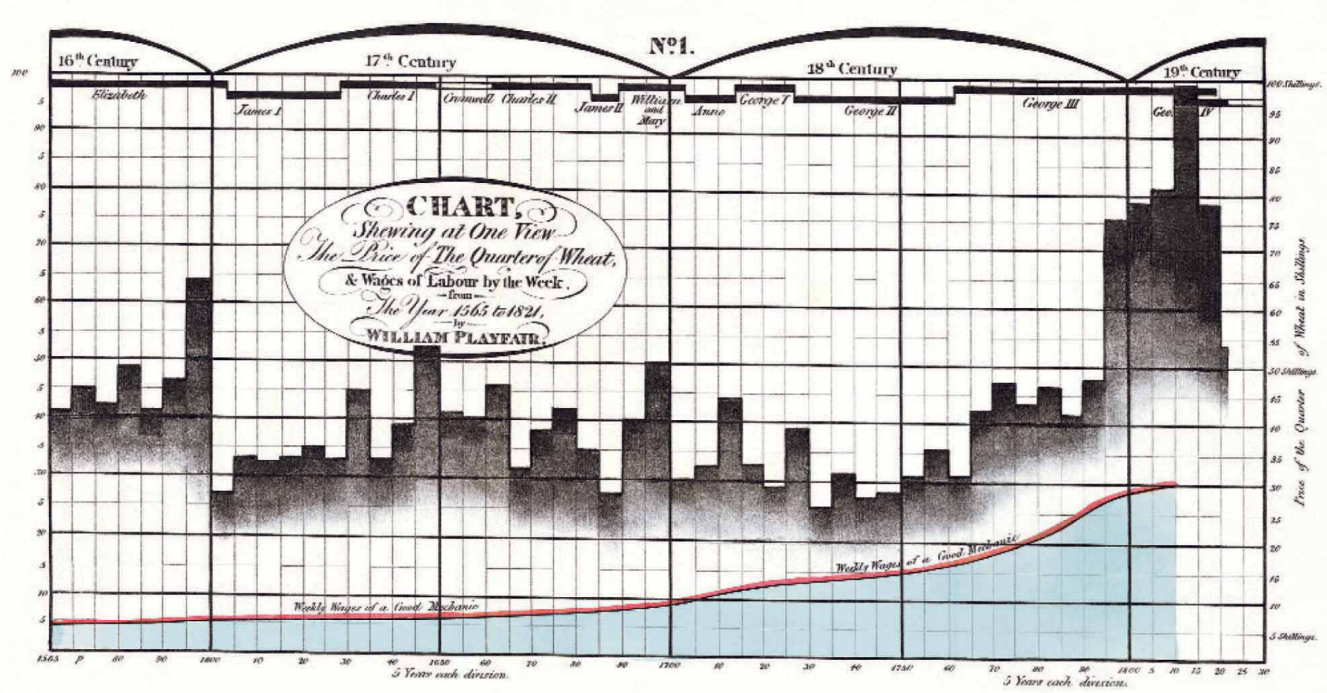


http://housedivided.dickinson.edu/sites/lincoln/clickable-word-clouds/

Playfair's last book addressed the question whether the price of wheat had increased relative to wages. In his *Letter on our agricultural distresses, their causes and remedies; accompanied with tables and copper-plate charts shewing and comparing the prices of wheat, bread and labour, from 1565 to 1821*, Playfair wrote:

> You have before you, my Lords and Gentlemen, a chart of the prices of wheat for 250 years, made from official returns; on the same plate I have traced a line representing, as nearly as I can, the wages of good mechanics, such as smiths, masons, and carpenters, in order to compare the proportion between them and the price of wheat at every different period. . . . the main fact deserving of consideration is, that never at any former period was wheat so cheap, in proportion to mechanical labour, as it is at the present time. . . . [pages 29-31]

Here Playfair plotted three parallel time-series: prices, wages, and the reigns of British kings and queens.



Overview of Statistical Learning

Broadly speaking, statistical learning falls into two categories:

- Supervised Learning
  - Observations are classified into *predictor* and *response* variables
  - Primary goal is model relationships between the predictor and response variables
  - Relationships can be used for *prediction* or for *inference*

- Unsupervised Learning
  - Each observation is a vector of variables, but they are not classified as predictors and responses
  - Primary goal is to **find** *structure* in the multivariable data set
  - One important technique is clustering, which seeks to find distinct groups of units or variables in the data set

Supervised Learning: `Boston` Example

Below is a portion of the `Boston` data set:

	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	medv
1	0.00632	18	2.31	0	0.538	6.575	65.2	4.0900	1	296	15.3	24.0
2	0.02731	0	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	21.6
3	0.02729	0	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	34.7
4	0.03237	0	2.18	0	0.458	6.998	45.8	6.0622	3	222	18.7	33.4
5	0.06905	0	2.18	0	0.458	7.147	54.2	6.0622	3	222	18.7	36.2
6	0.02985	0	2.18	0	0.458	6.430	58.7	6.0622	3	222	18.7	28.7

Bill Damon Flickr

Variables:

crim: per capita crime rate by town.
zn: proportion of residential land zoned for lots over 25,000 sq.ft.
indus: proportion of non-retail business acres per town.
chas: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).
nox: nitrogen oxides concentration (parts per 10 million).
rm: average number of rooms per dwelling.
age: proportion of owner-occupied units built prior to 1940.
dis: weighted mean of distances to five Boston employment centers.
rad: index of accessibility to radial highways.
tax: full-value property-tax rate per \$10,000.
ptratio: pupil-teacher ratio by town.
medv: median value of owner-occupied homes in \$1000s

Supervised Learning: `Boston` Example

We may wish to construct a model that predicts median value (`mdev`) based on the other variables in the data set.

This assigns `medv` as the *response variable* and the other variables as *predictor variables*.

- This model could be used to **predict** future (or unobserved) responses based on the predictors
  - If we built a new suburb with a set of predictors, what would the model predict to be the median value of homes in the new suburb?
- Some models can also be used for **inference** to understand the relationship between the response and the predictors.
  - If we reduced the pollution level in a particular suburb by half its current amount, what would the effect be on median values?
  - If we added access to the radial highways, what would the model estimate the effect to be the median value of homes in the suburb?

Unsupervised Learning: `MammalsMilk` Example

Below is a portion of the `MammalsMilk` data set:

Mammal	Water	Protein	Fat	Lactose	Ash
Horse	90.1	2.6	1	6.9	0.35
Orangutan	88.5	1.4	3.5	6	0.24
Monkey	88.4	2.2	2.7	6.4	0.18
Donkey	90.3	1.7	1.4	6.2	0.4
Hippo	90.4	0.6	4.5	4.4	0.1
Camel	87.7	3.5	3.4	4.8	0.71
Bison	86.9	4.8	1.7	5.7	0.9
Buffalo	82.1	5.9	7.9	4.7	0.78
Guinea Pig	81.9	7.4	7.2	2.7	0.85
Cat	81.6	10.1	6.3	4.4	0.75
Fox	81.6	6.6	5.9	4.9	0.93

Variables:

Water: percentage water
Protein: percentage of protein
Fat: percentage of fat
Lactose: percentage of lactose
Ash: percentage of ash (mineral content)

Each mammal produces a vector of milk components:

(Water, Protein, Fat, Lactose, Ash)

but we don't have a natural candidate for a response variable.

Unsupervised Learning: `MammalsMilk` Example

While we don't have a response variable we wish to predict, we can still ask questions about how the variables are related to each other or to the objects in our data set.

A natural question to ask is: Do the mammals in the data set form a natural collection of discrete groups based on the measurements from their milk?

For these groups (clusters) to be useful, we would expect

- Individual clusters should display some level of homogeneity
- Different clusters should display some level of separation

An important technique in unsupervised learning is *cluster analysis*, which provides a collections of methods for identifying, representing, and validating these internal groups.

Statistics

Many of the techniques in Data Science employ concepts from Descriptive and Inferential Statistics.

In order to properly contextualize these ideas, we will spend some time reviewing concepts from probability and the study of random variables.

We will also examine several numeric and graphical measures of important statistical properties including measures of

- Center
- Shape
- Spread
- Position


Always has been
Wait it's all statistics?
Data Science

Professional Ethics

- As our capabilities in data acquisition, retrieval, and processing increase, care needs to be exercised to assure that what institutions are doing with their data maintains ethical standards.
- Capabilities change at a rate that is faster than legislative and corporate bodies can react to, and this creates a large “ethics gap” that opens up when legal and ethical boundaries separate.
- It is important that the data analyst consider the ethical implications of the methods s/he employs and keep in mind that the things we can do are not necessarily in correspondence with the things we should do.
- The analyst should be considered as a stakeholder in the discussion and should be expected to conduct their work following ethical guidelines.

Cambridge Analytica

DATA 198: Philosophy of Data

DATA 198: Philosophy of Data

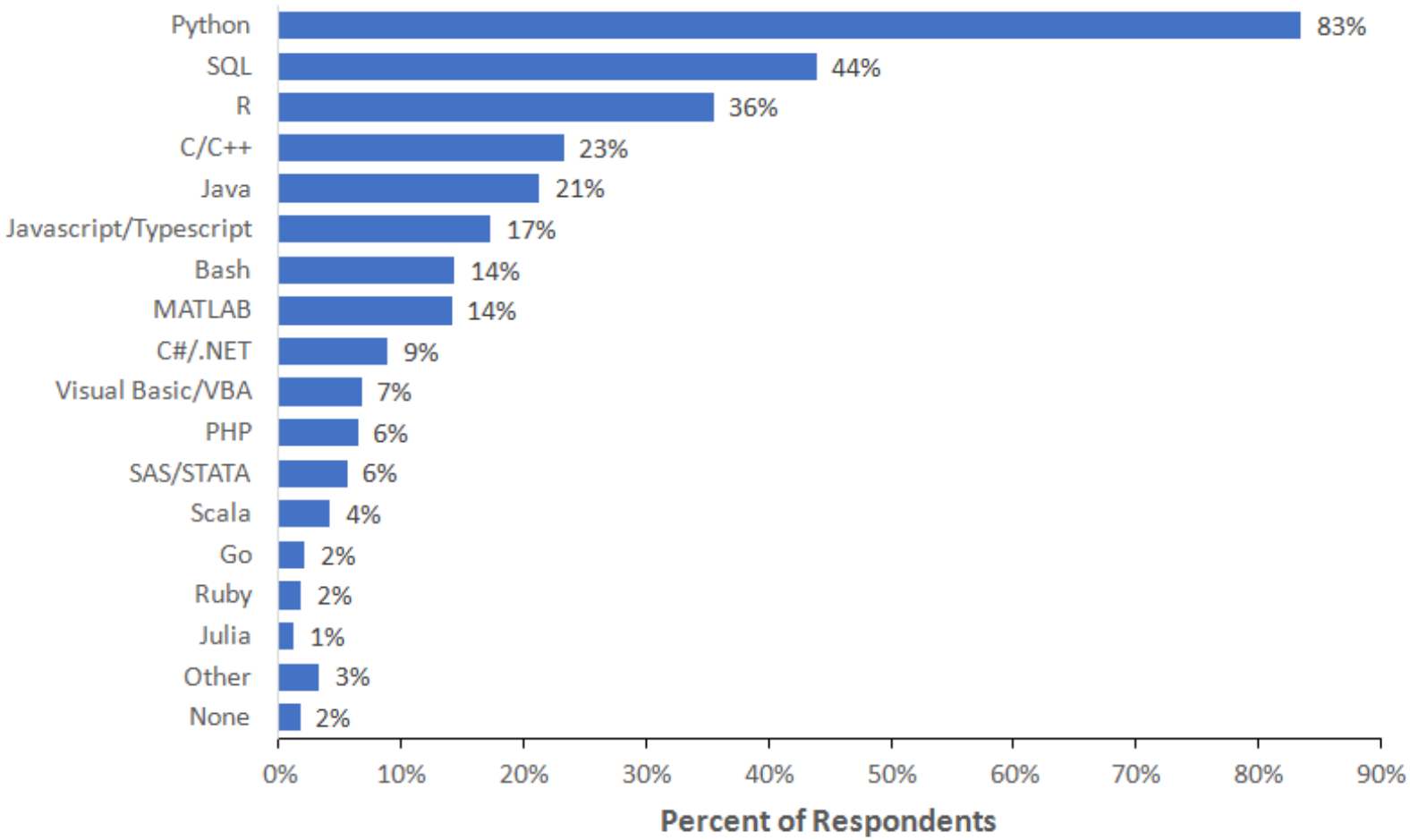
POKEITOUTWITHASTICK
HTTPS://TWITTER.COM/POKEITOUTWITHAS

R and R-Studio

- We will use the R Programming environment extensively during the course
- R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS.
- R has an enormous user base and a wealth of open-source resources.
- We will use R-Studio, which comes with base R + more
- References to objects in R will usually be presented in the `Courier New` font for readability.

https://cran.rstudio.com
https://www.rstudio.com/products/rstudio/download

What programming language do you use on a regular basis?



Note: Data are from the 2018 Kaggle Machine Learning and Data Science Survey. You can learn more about the study here: http://www.kaggle.com/kaggle/kaggle-survey-2018. A total of 18827 respondents answered the question.

BUSINESS BROADWAY
DATA SCIENCE | CUSTOMER ANALYTICS | MACHINE LEARNING

The answer: both!