Machine Learning Concepts

1.1

Some slides borrowed from the presentation of Szilard Pafka.

Zoltán Papp
Quant & Machine Learning
Practitioner/Teacher
https://www.linkedin.com/in/pzoltan/
@zozopp

Machine Learning

Why?

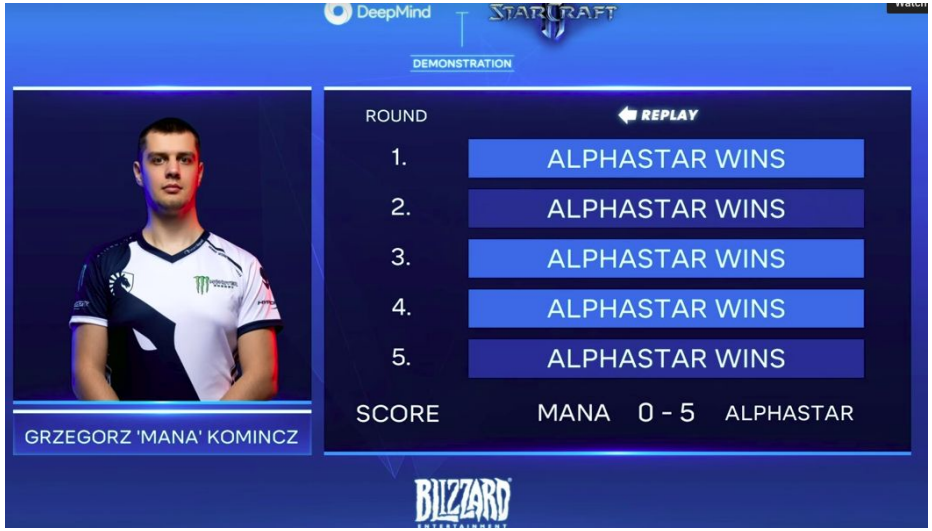
- Drive the car on a highway
- Park the car
- Autonomous driving expected in 2 years by Elon Musk

Why?

May, 2016: Google's AlphaGo Defeats Lee Sedol, Chinese Go Master

Why?

In December, 2018: Google's DeepMind Defeats TLO and MANA top ranking professional Starcraft players



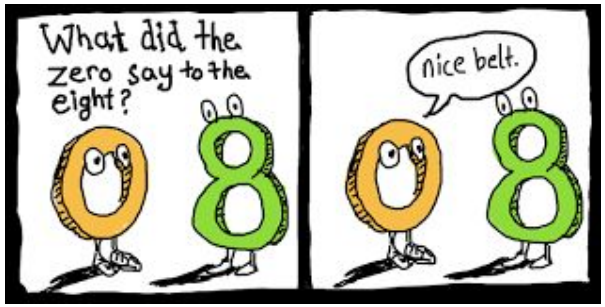
Why?


60
55
0
5
10


NEWS PAPER
LATEST
HEADLINES
www.shutterstock.com · 550233844


Calendar


amazon

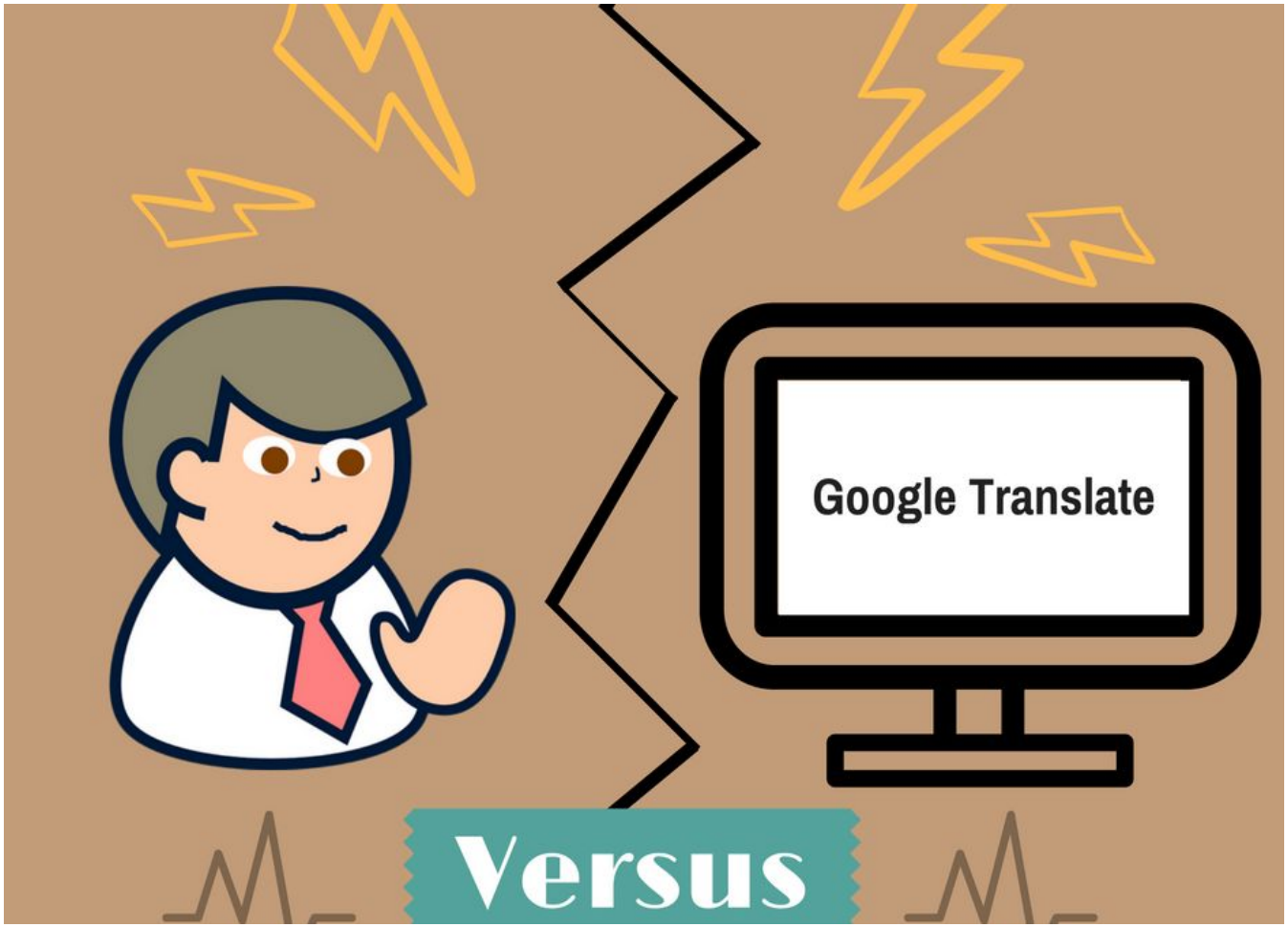

What did the zero say to the eight?
nice belt.

Amazon

Why?

Google Translate now serves 200 million people daily

A billion translations a day



What is needed for ML

"The rate at which we're generating data is rapidly outpacing our ability to analyze it"

Professor Patrick Wolfe, Executive Director of the University College of London's Big Data Institute

"The trick here is to turn these massive data streams from a liability into a strength."

+

Data

Computing Capacity

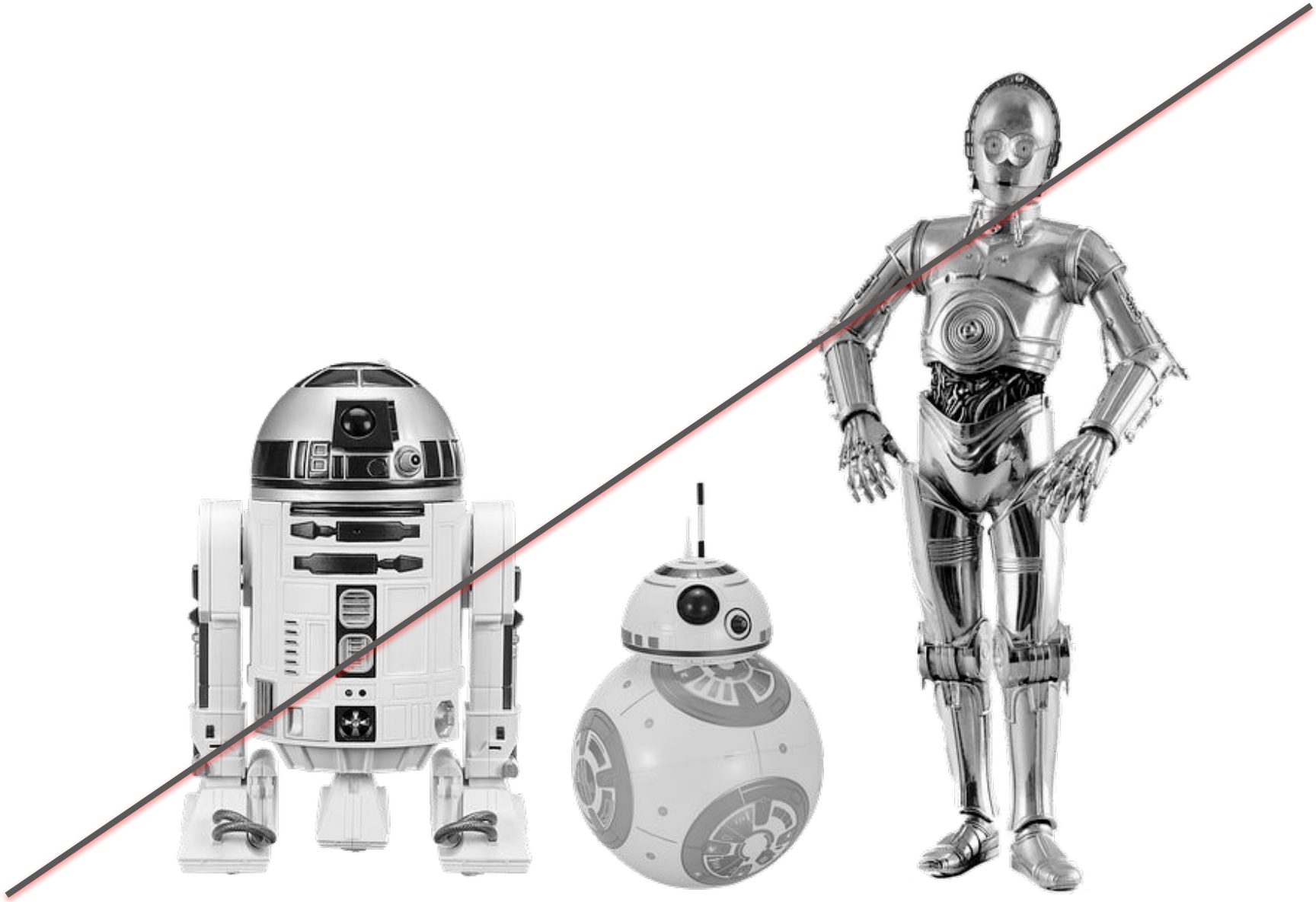

So What's
in it
for me?

What is your background?

What do you expect?

Is it General Artificial Intelligence?

It is not General Artificial Intelligence!



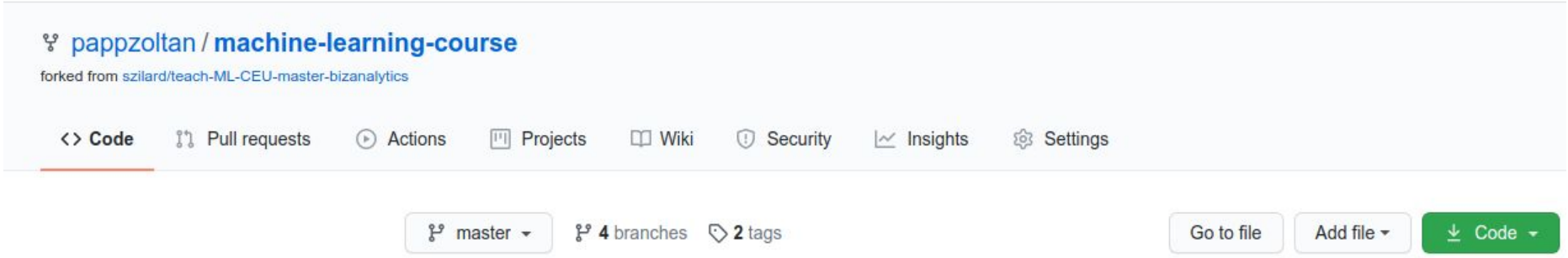
Course assignments and exam

Grading:

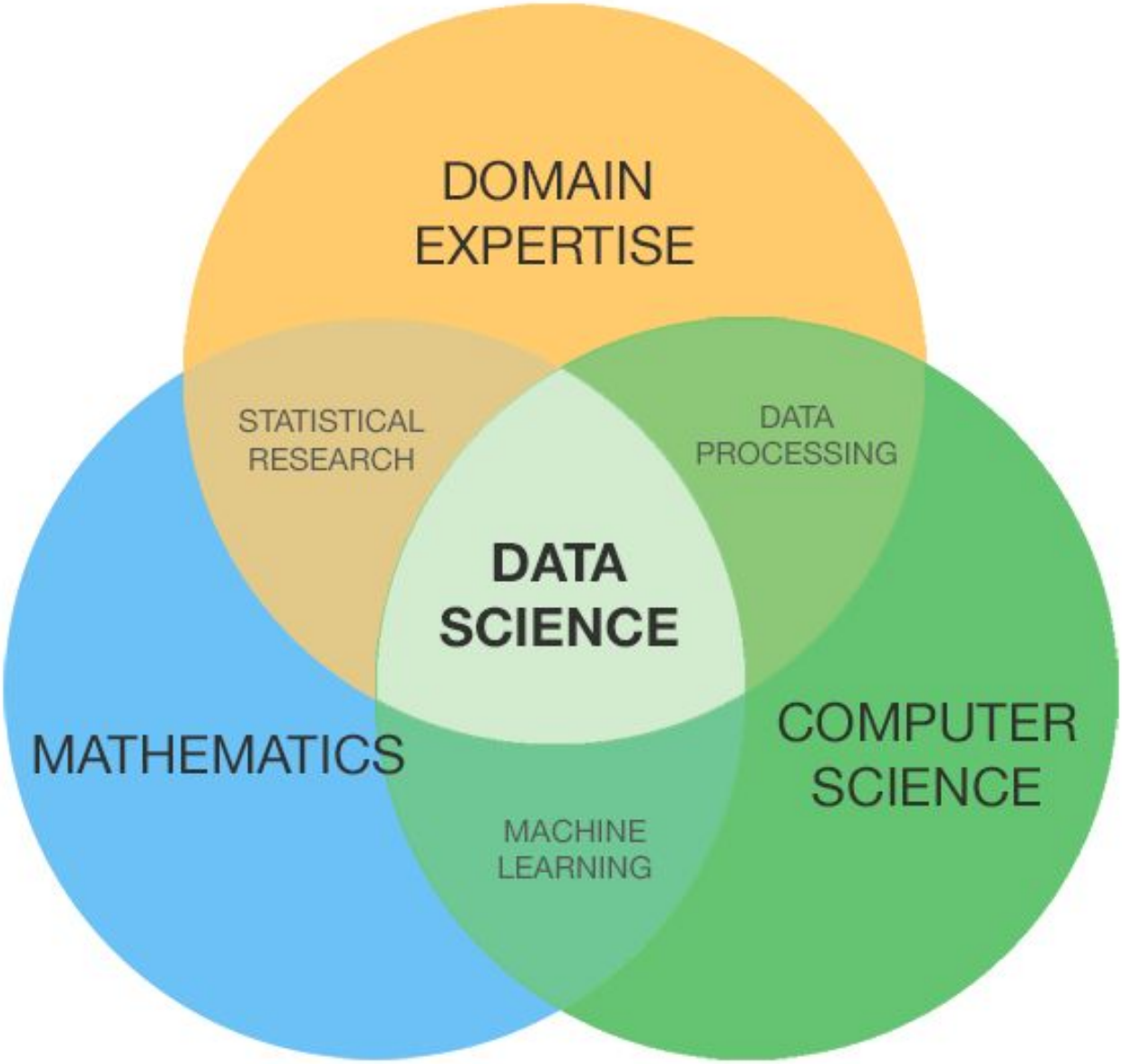
Exam and Assignment dates and deadlines:

Weekly assignment acceptance policy and achievable grades:

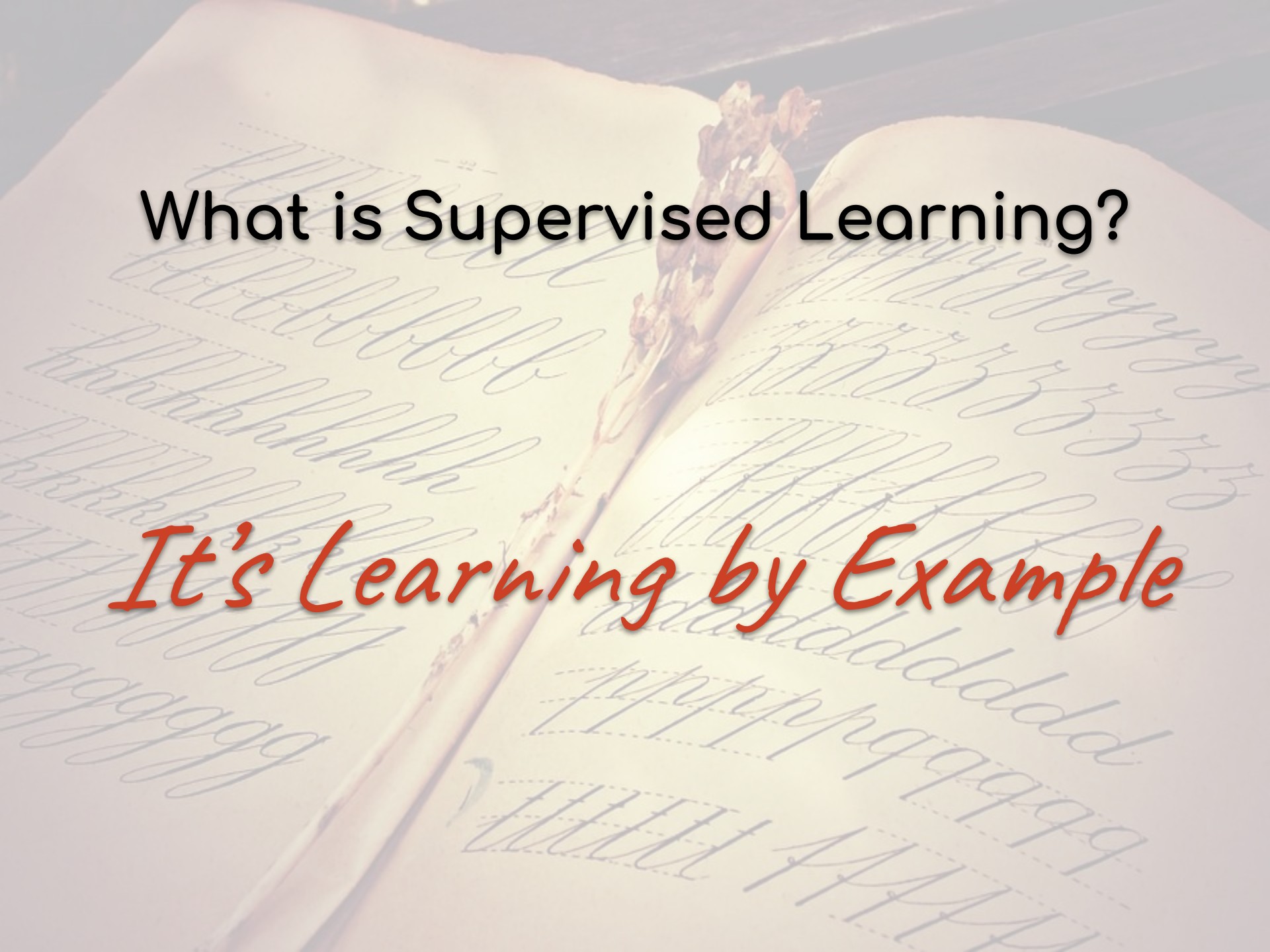
https://github.com/pappzoltan



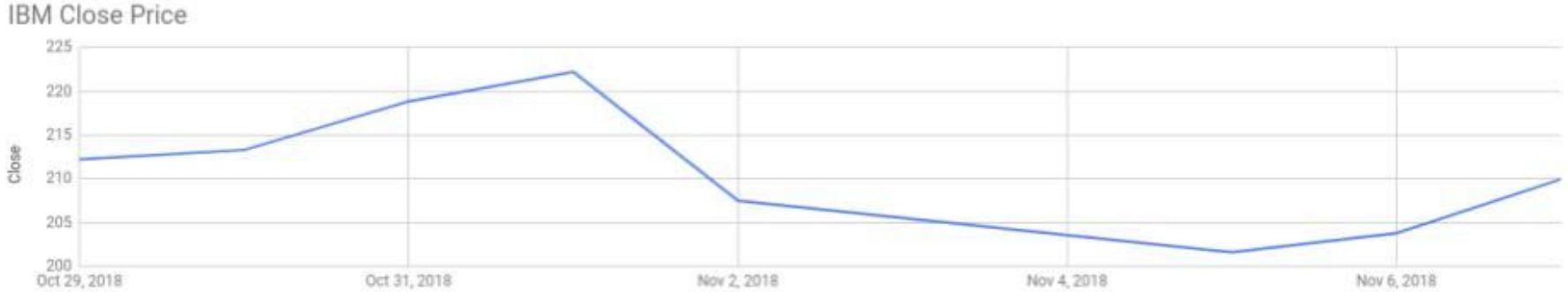
- Lectures
  a. Slides
  b. Code samples
  c. Data
- Labs
  a. Code samples
  b. Data



Source: Palmer, Shelly. Data Science for the C-Suite. New York: Digital Living Press, 2015. Print.


What is Supervised Learning?
It's Learning by Example

Financial panel data from Apple



Date	Open	High	Low	Close	Volume
Oct 29, 2018	219.19	219.69	206.09	212.24	45,935,500
Oct 30, 2018	211.15	215.18	209.27	213.3	36,660,000
Oct 31, 2018	216.88	220.45	216.62	218.86	38,358,900
Nov 01, 2018	219.05	222.36	216.81	222.22	58,323,200
Nov 02, 2018	209.55	213.65	205.43	207.48	91,328,700
Nov 05, 2018	204.3	204.39	198.17	201.59	66,163,700
Nov 06, 2018	201.92	204.72	201.69	203.77	31,882,900
Nov 07, 2018	205.97	210.06	204.13	209.95	33,424,400

Financial panel data from Apple

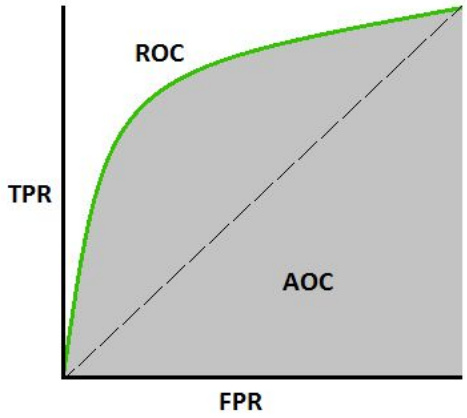
Date	Open	High	Low	Close	Volume
Oct 29, 2018	219.19	219.69	206.09	212.24	45,935,500
Oct 30, 2018	211.15	215.18	209.27	213.3	36,660,000
Oct 31, 2018	216.88	220.45	216.62	218.86	38,358,900
Nov 01, 2018	219.05	222.36	216.81	222.22	58,323,200
Nov 02, 2018	209.55	213.65	205.43	207.48	91,328,700
Nov 05, 2018	204.3	204.39	198.17	201.59	66,163,700
Nov 06, 2018	201.92	204.72	201.69	203.77	31,882,900
Nov 07, 2018	205.97	210.06	204.13	209.95	33,424,400

Date	Close	Previous Close	Previous Volume	Features
Oct 29, 2018	212.24	NaN	NaN	...
Oct 30, 2018	213.3	212.24	45,935,500	...
Oct 31, 2018	218.86	213.3	36,660,000	...
Nov 01, 2018	222.22	218.86	38,358,900	...
Nov 02, 2018	207.48	222.22	58,323,200	...
Nov 05, 2018	201.59	207.48	91,328,700	...
Nov 06, 2018	203.77	201.59	66,163,700	...
Nov 07, 2018	209.95	203.77	31,882,900	...

Classification? Regression? Performance

Classification

Is ig going up or down?



Classification Examples:

- Is mail a spam or not?
- Is it a cat on the image?
- Is it the same person on two pictures?

Regression

How much up or down?

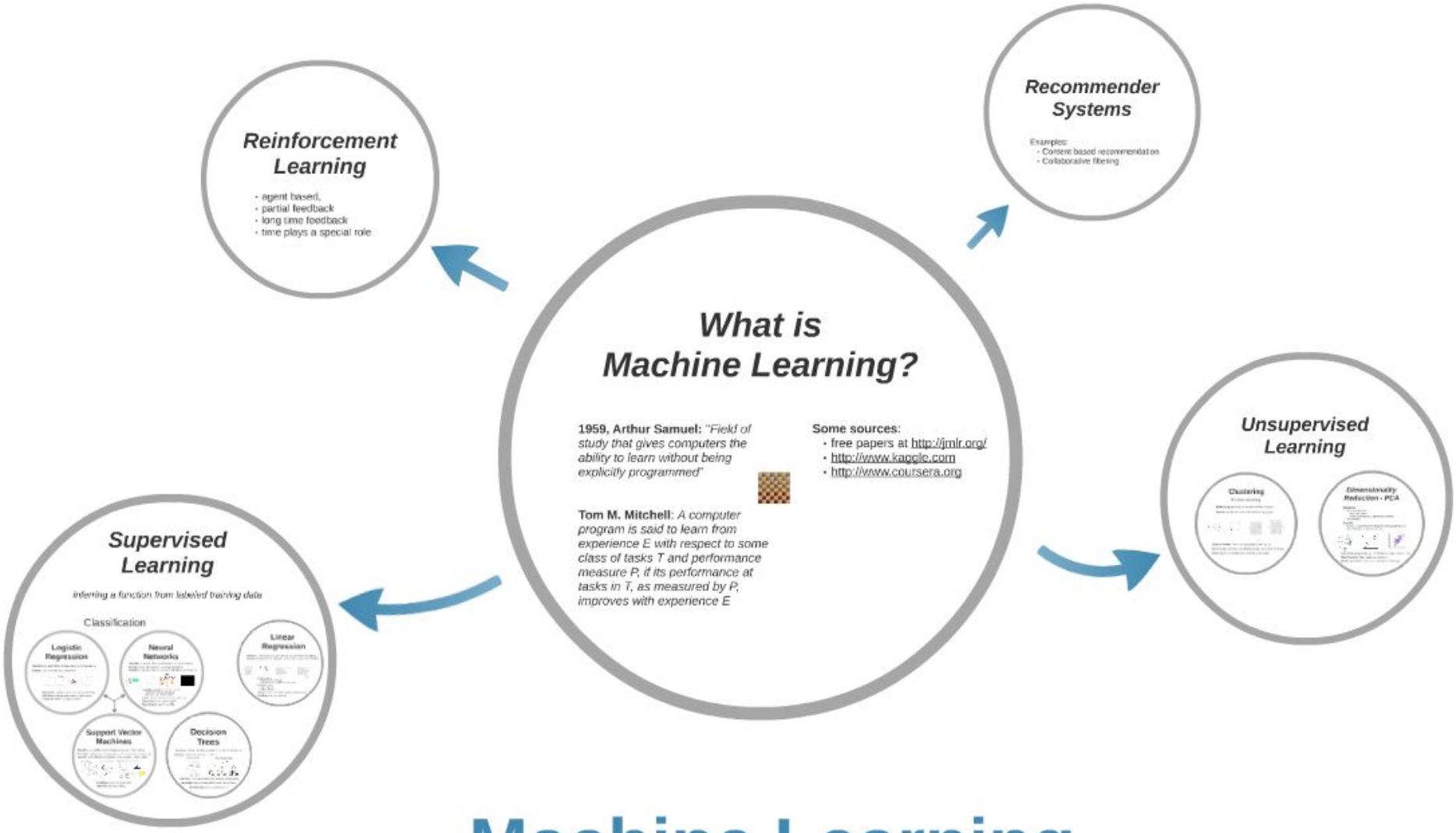
$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Regression Examples:

- Next day return for Apple?
- Temperature tomorrow?
- Expected bike sharing demand at a station tomorrow?

Let's see some supervised and unsupervised learning algorithms

https://prezi.com/06swcwazd0ai/machine-learning/



Machine Learning

Keywords to summarize topics

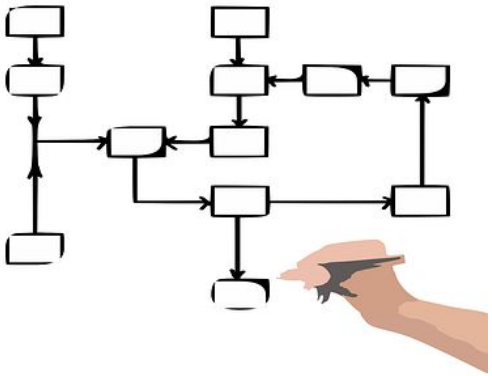
Supervised Learning
Data: X (n observations, p features), y (labels)
Regression, classification
Train/learn/fit f from data (model)
Score: for new x, get f(x)
Algos: LR, k-NN, DT, RF, GBM, NN/DL, SVM, NB...
Goal: max acc/min err new data
Metrics: MSE, AUC (ROC)
Bad: measure on train set. Need: test/CV
Hyperparameters, model capacity, overfitting
Regularization
Model selection
Hyperparameter search (grid, random)

Algos: LR, k-NN, DT, RF, GBM, NN/DL, SVM, NB...
ML Tools: R packages, glmnet, h2o, xgboost, lightgbm, keras, Tensorflow...

Unsupervised Learning

Best practices

Topics to cover



Data Mining Process

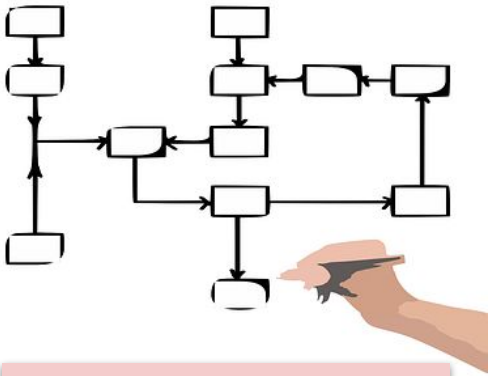
Exploratory Data Analysis

Data Visualization

Data Science Tools

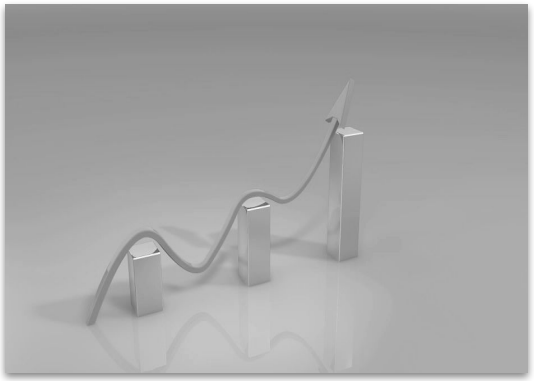
Reproducible Research

Topics to cover



Data Mining Process

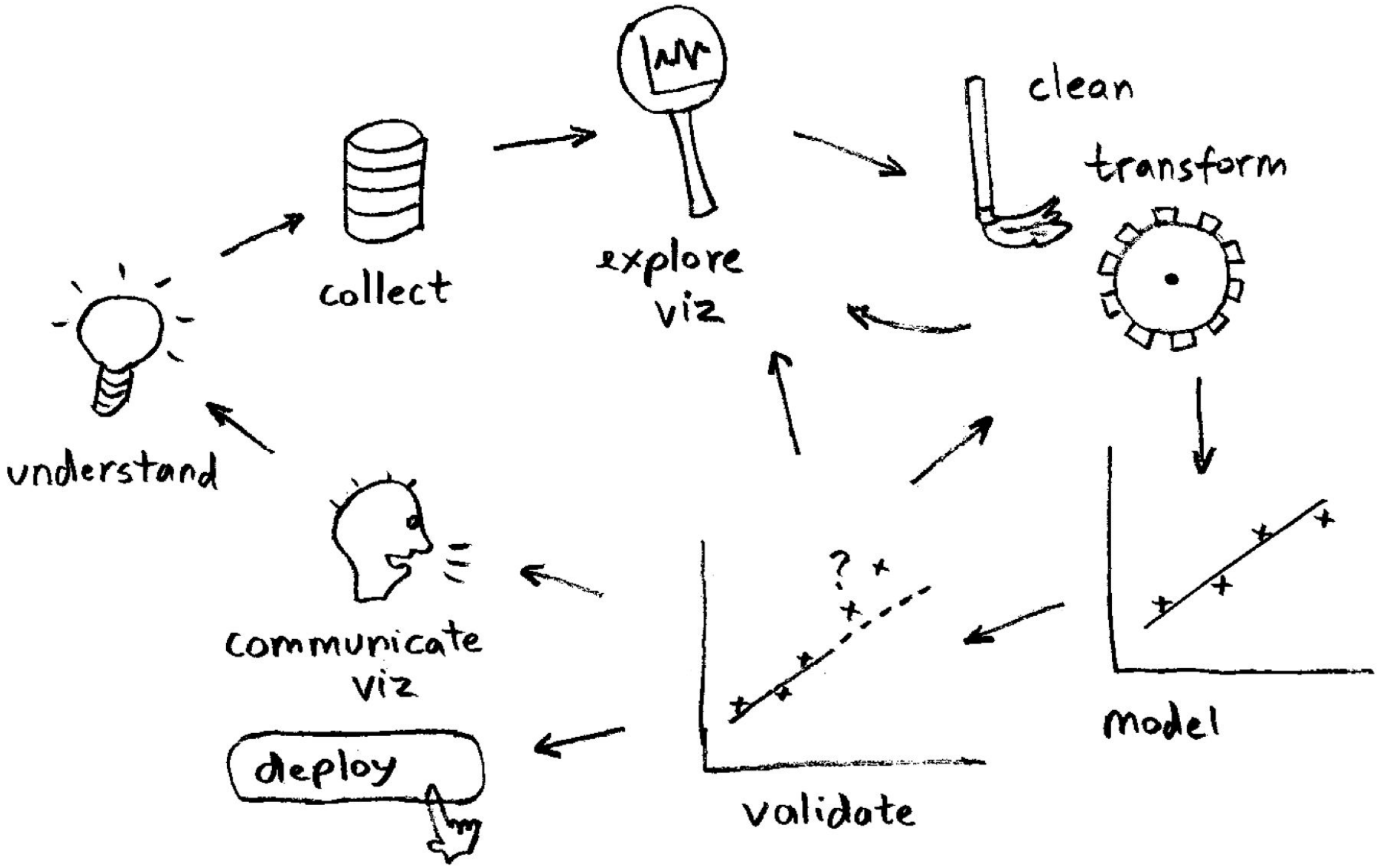
Exploratory Data Analysis



Data Visualization

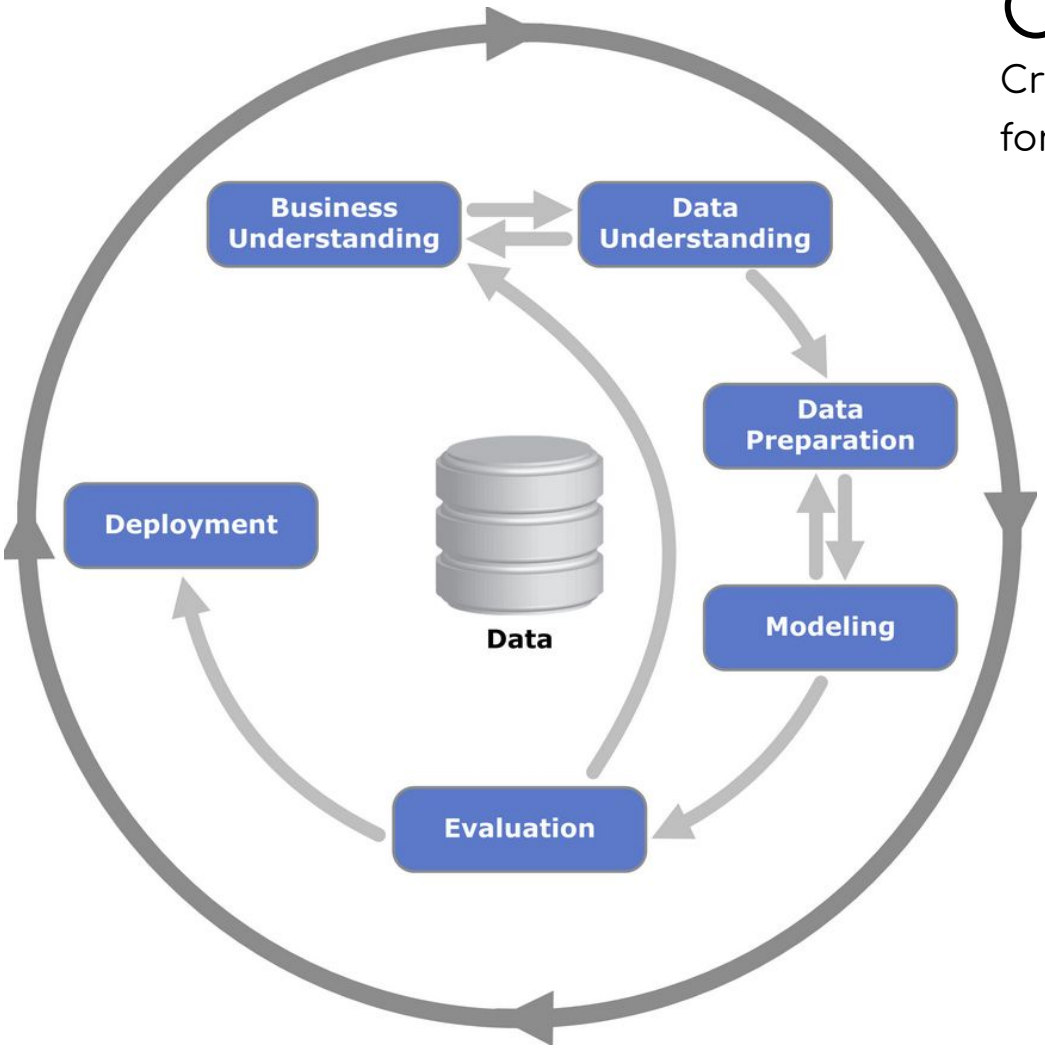
Data Science Tools

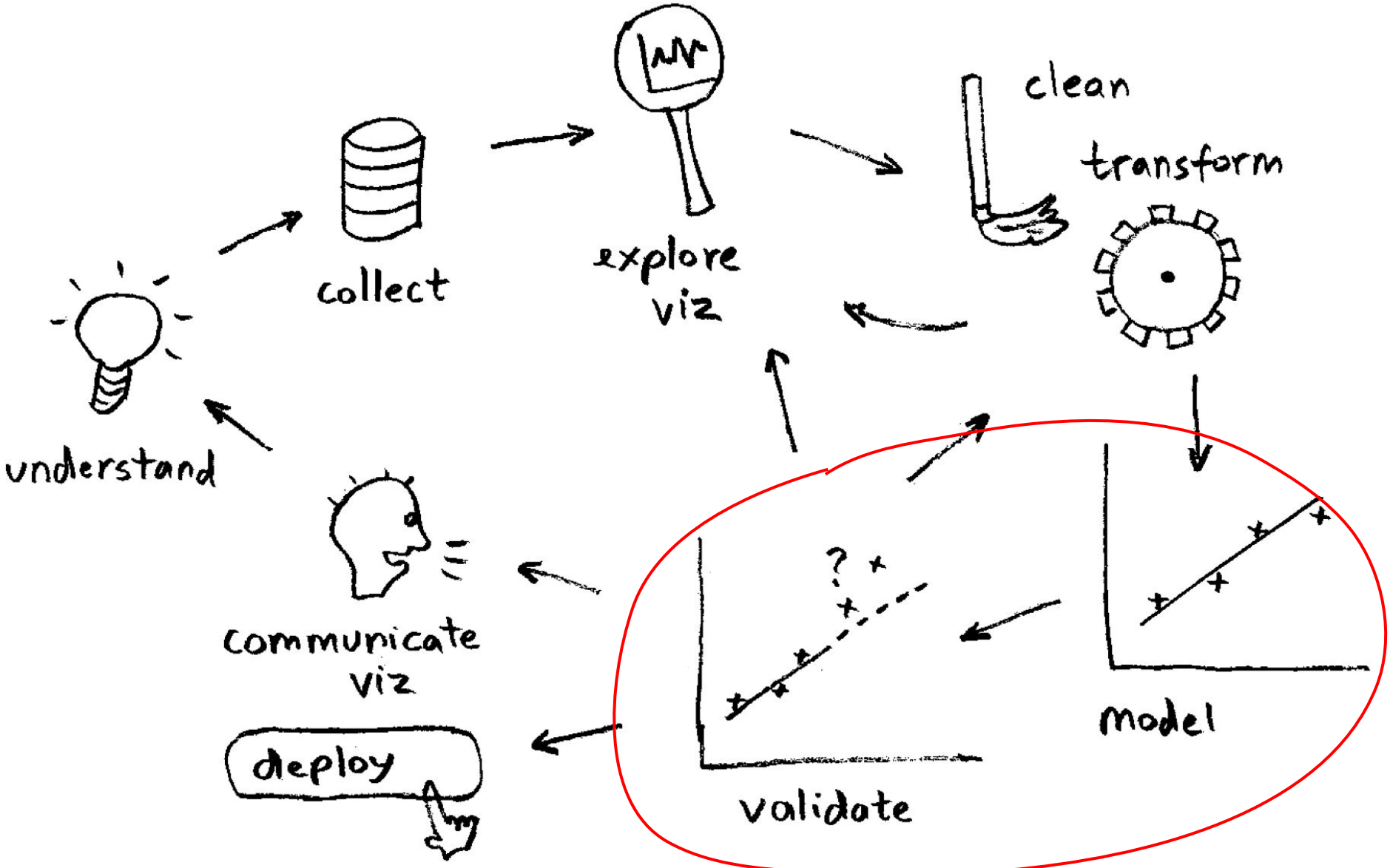
Reproducible Research

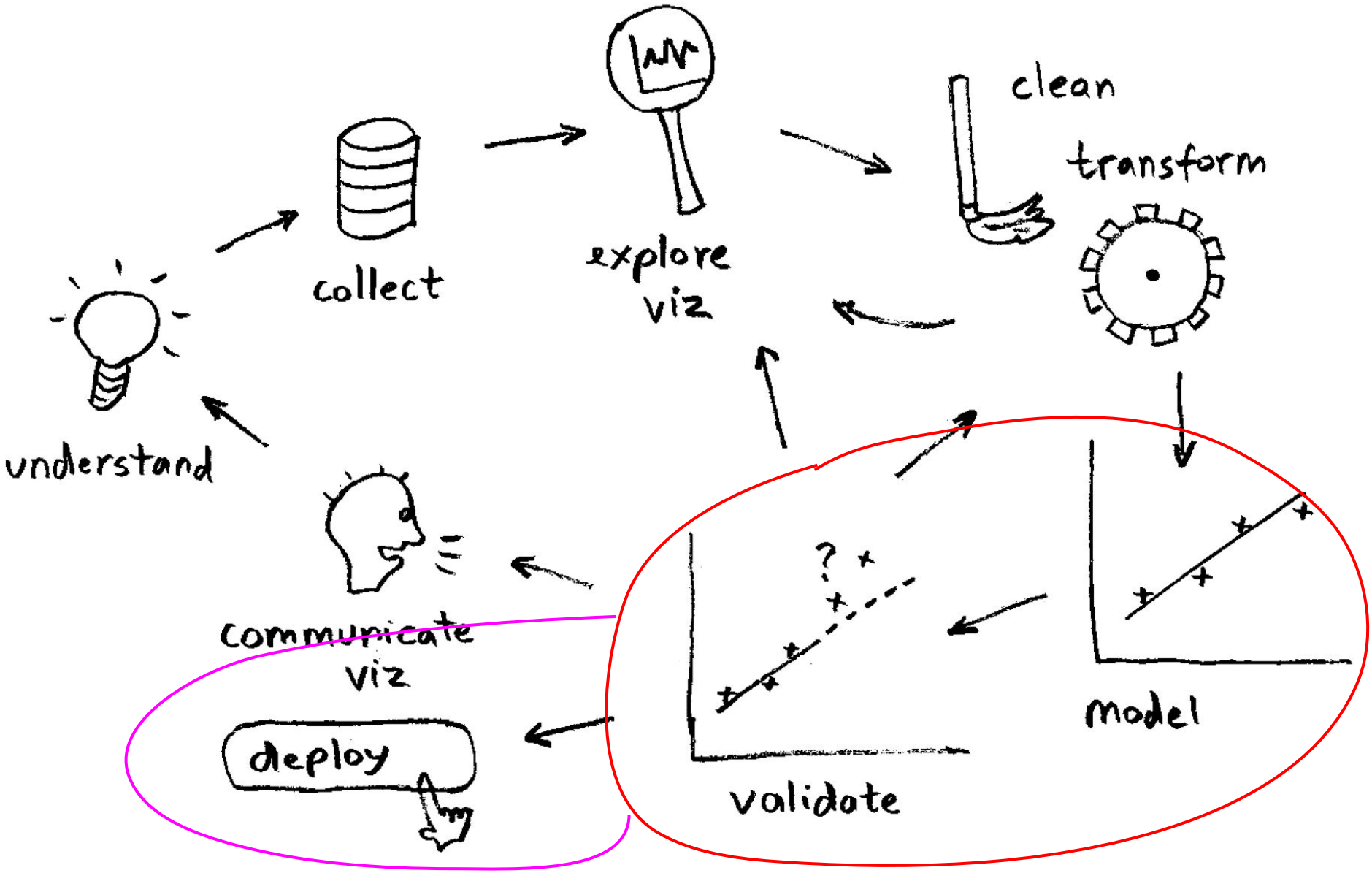

clean
transform
collect
explore
viz
understand
communicate
viz
deploy
validate
model

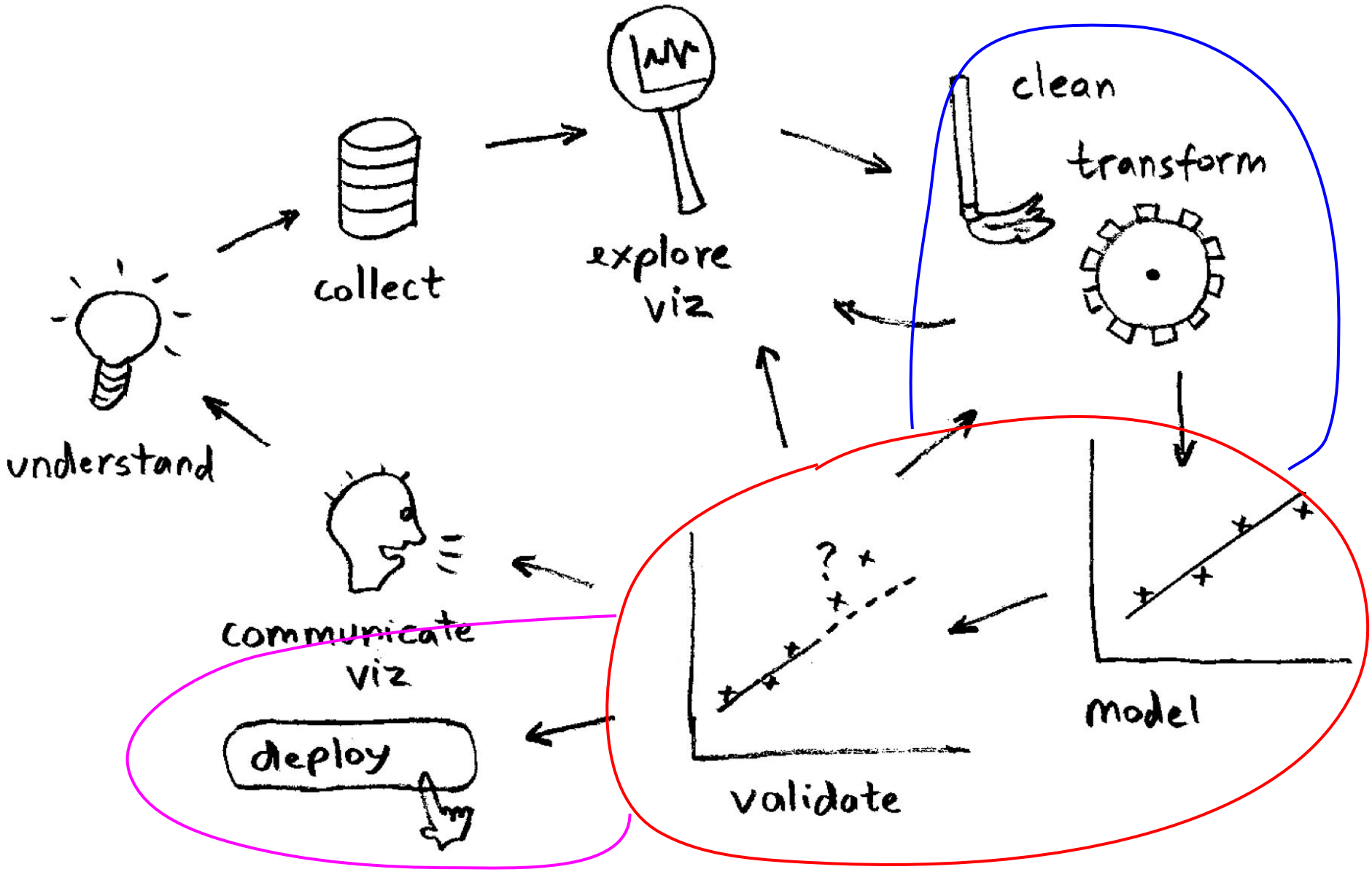
CRISP-DM, 1999

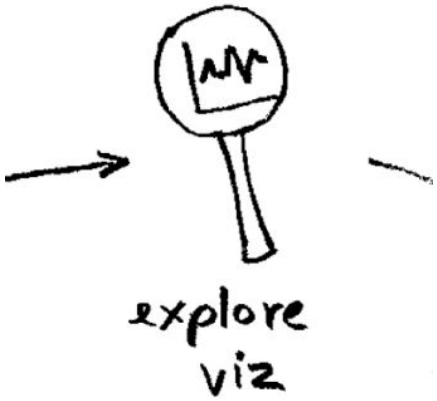
Cross-industry standard process for data mining

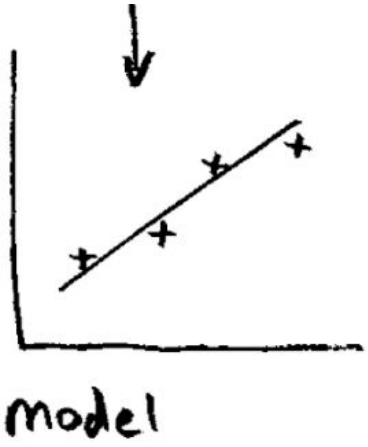



collect
explore
viz
clean
transform
understand
communicate
viz
deploy
validate
model


collect
explore
viz
clean
transform
understand
communicate
viz
deploy
validate
model


clean
transform
collect
explore
viz
understand
communicate
viz
deploy
validate
model


explore
viz


model

David Donoho: 50 years of Data Science

GDS1: Exploratory data analysis and data manipulation/preparation and cleaning

GDS2: Systems/databases

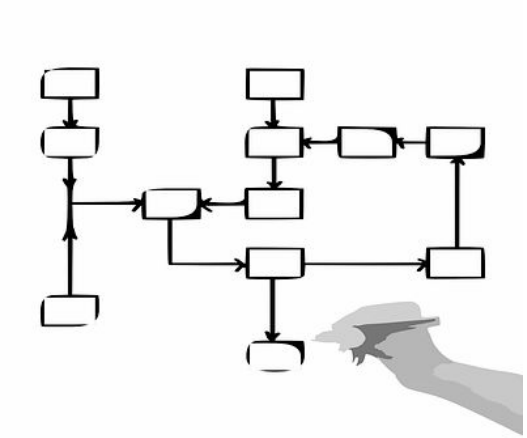
GDS3: Programming: R

GDS4: Data visualization

GDS5: Modeling (statistics/machine learning)

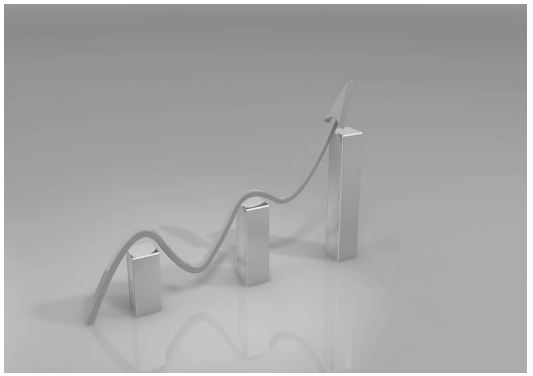
GDS6: Science of Data Science

Topics to cover



Data Mining Process

Exploratory Data Analysis



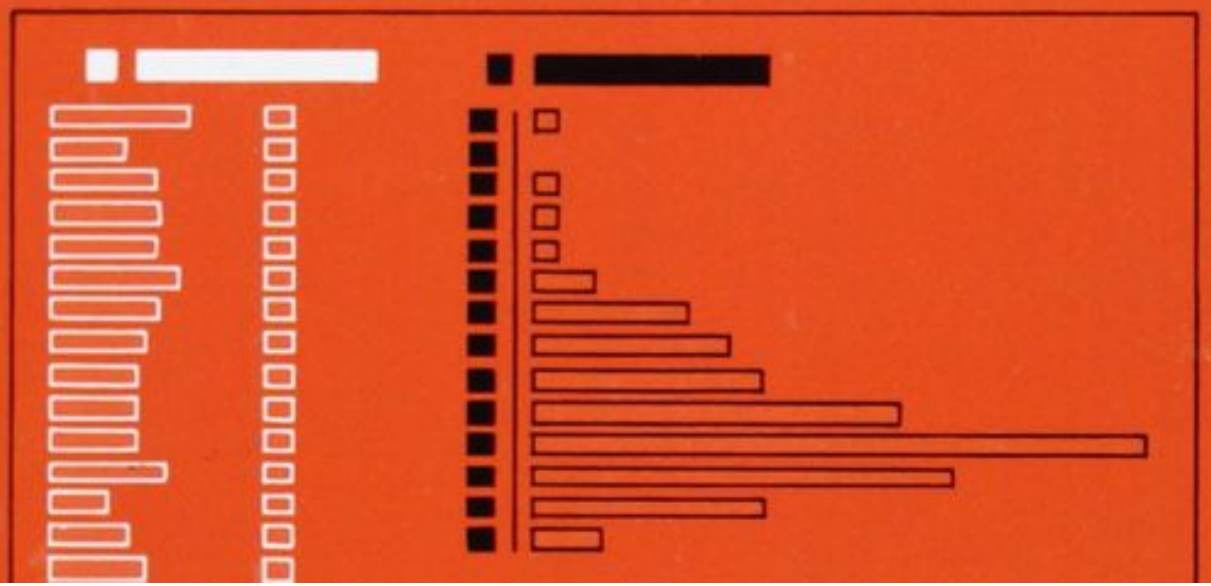
Data Visualization

Data Science Tools

Reproducible Research

John W. Tukey

EXPLORATORY DATA ANALYSIS



Big Data Borat
@BigDataBorat

Follow

In Data Science, 80% of time spent prepare data, 20% of time spent complain about need for prepare data.

Reply Retweet Favorite More

Data Scientist: The Sexiest Job of the 21st Century

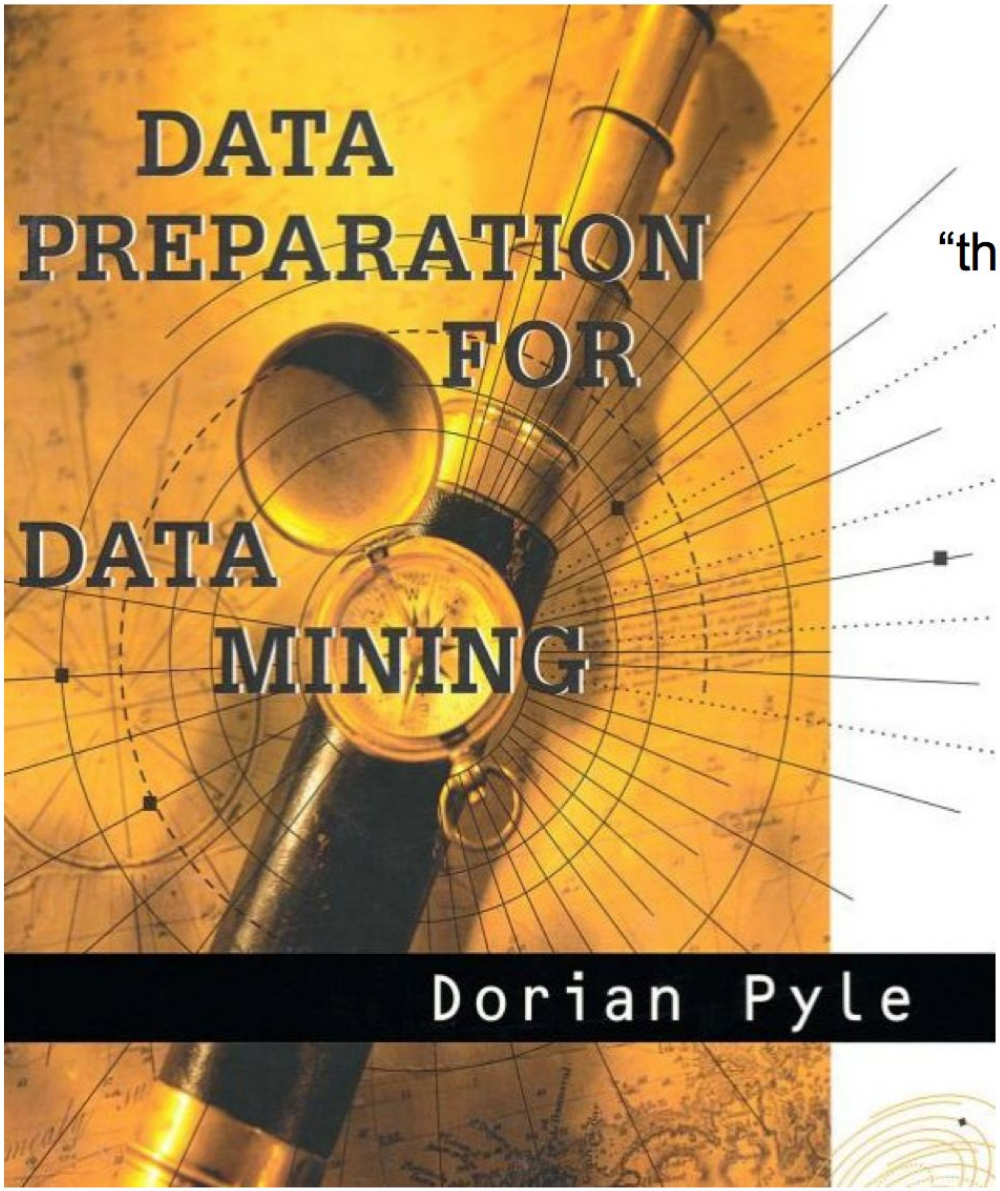
Harvard Business Review

Data Scientist: The Sexiest Job of the 21st Century

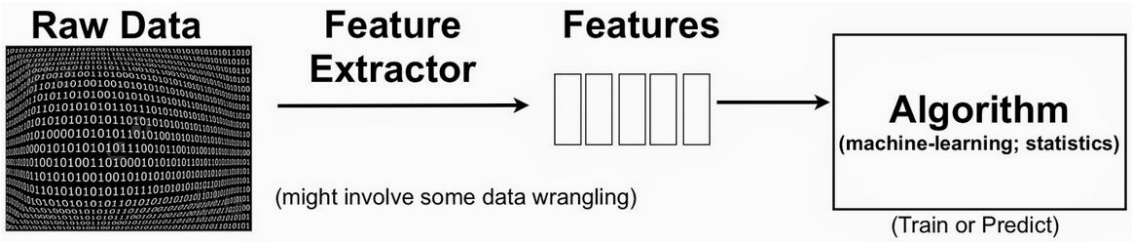
Harvard Business Review

For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights

The New York Times



"the data preparation process prepares both the data and the modeler"


Raw Data
Feature Extractor
Features
Algorithm
(machine-learning; statistics)
(might involve some data wrangling)
(Train or Predict)

Bike Sharing Demand

Data Fields

datetime - hourly date + timestamp

season - 1 = Q1, 2 = Q2, 3 = Q3, 4 = Q4

holiday - whether the day is considered a holiday

workingday - whether the day is neither a weekend nor holiday

weather

1. Clear, Few clouds, Partly cloudy, Partly cloudy
2. Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
3. Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
4. Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog

temp - temperature in Celsius

atemp - "feels like" temperature in Celsius

humidity - relative humidity

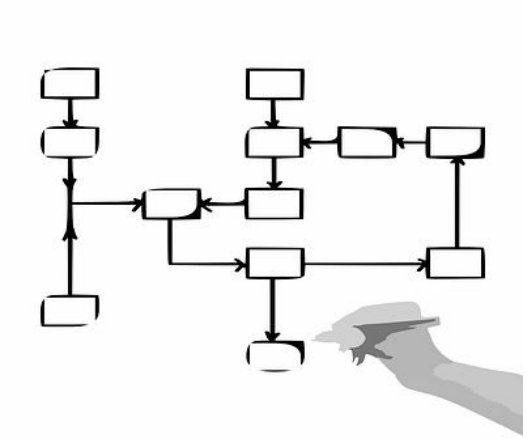
windspeed - wind speed

casual - number of non-registered user rentals initiated

registered - number of registered user rentals initiated

count - number of total rentals

Topics to cover



Data Mining Process

Exploratory Data Analysis

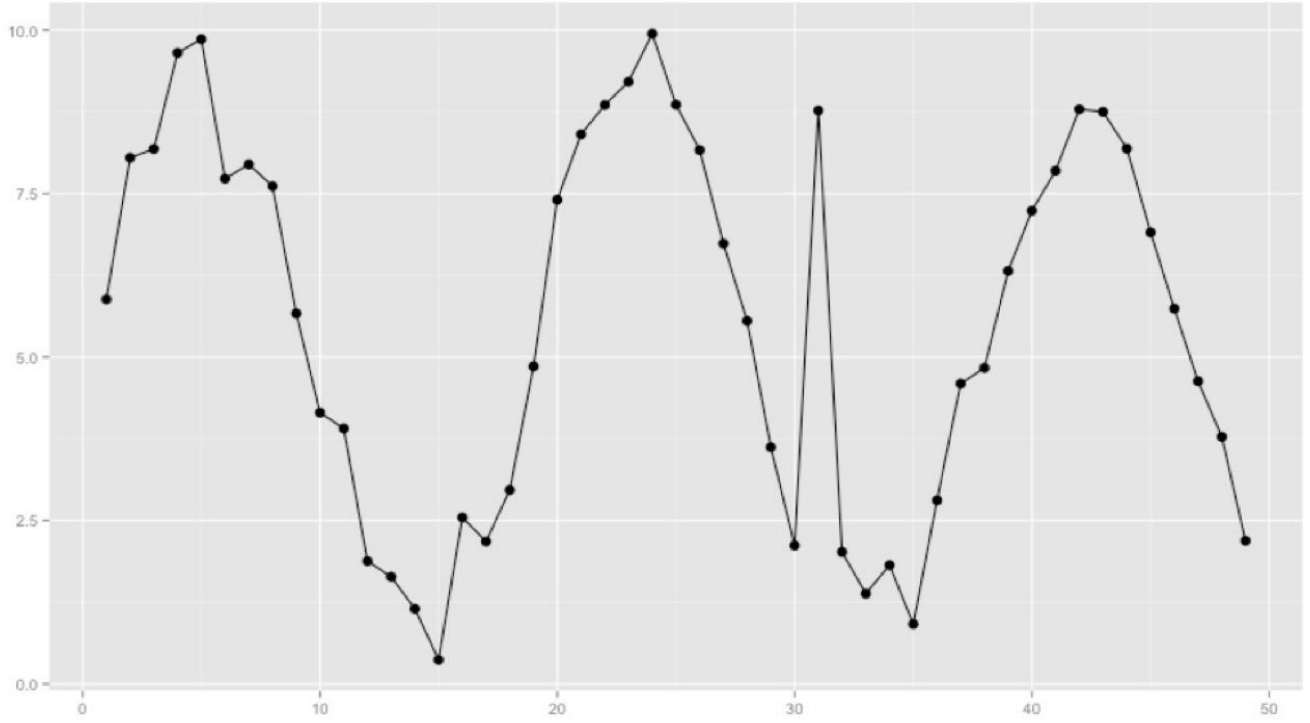
Data Visualization

Data Science Tools

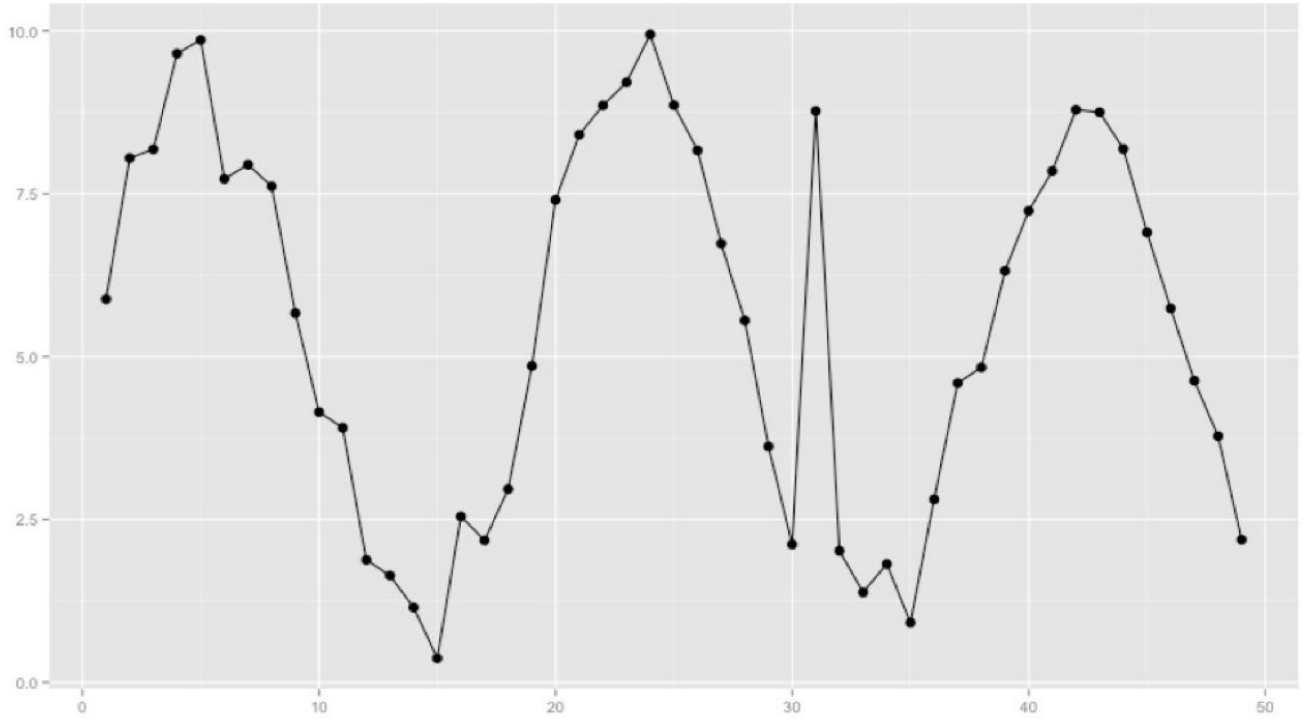
Reproducible Research

5.8839338 8.0500895 8.1838378 9.6537864 9.8625664 7.7283027 7.9485545
7.6139286 5.6673501 4.1509576 3.9085585 1.8794583 1.6390833 1.1494471
0.3701523 2.5463324 2.1793825 2.9664571 4.8563495 7.4056122 8.4070250
8.8555943 9.2110448 9.9459725 8.8605880 8.1672658 6.7366060 5.5535530
3.6201724 2.1181429 8.7715833 2.0190151 1.3814497 1.8169363 0.9166560
2.8093003 4.5931840 4.8333278 6.3170302 7.2390577 7.8509665 8.7915221
8.7507326 8.1899304 6.9060409 5.7413247 4.6312077 3.7803116 2.1903559

```
5.8839338 8.0500895 8.1838378 9.6537864 9.8625664 7.7283027 7.9485545
7.6139286 5.6673501 4.1509576 3.9085585 1.8794583 1.6390833 1.1494471
0.3701523 2.5463324 2.1793825 2.9664571 4.8563495 7.4056122 8.4070250
8.8555943 9.2110448 9.9459725 8.8605880 8.1672658 6.7366060 5.5535530
3.6201724 2.1181429 8.7715833 2.0190151 1.3814497 1.8169363 0.9166560
2.8093003 4.5931840 4.8333278 6.3170302 7.2390577 7.8509665 8.7915221
8.7507326 8.1899304 6.9060409 5.7413247 4.6312077 3.7803116 2.1903559
```



```
5.8839338 8.0500895 8.1838378 9.6537864 9.8625664 7.7283027 7.9485545
7.6139286 5.6673501 4.1509576 3.9085585 1.8794583 1.6390833 1.1494471
0.3701523 2.5463324 2.1793825 2.9664571 4.8563495 7.4056122 8.4070250
8.8555943 9.2110448 9.9459725 8.8605880 8.1672658 6.7366060 5.5535530
3.6201724 2.1181429 8.7715833 2.0190151 1.3814497 1.8169363 0.9166560
2.8093003 4.5931840 4.8333278 6.3170302 7.2390577 7.8509665 8.7915221
8.7507326 8.1899304 6.9060409 5.7413247 4.6312077 3.7803116 2.1903559
```

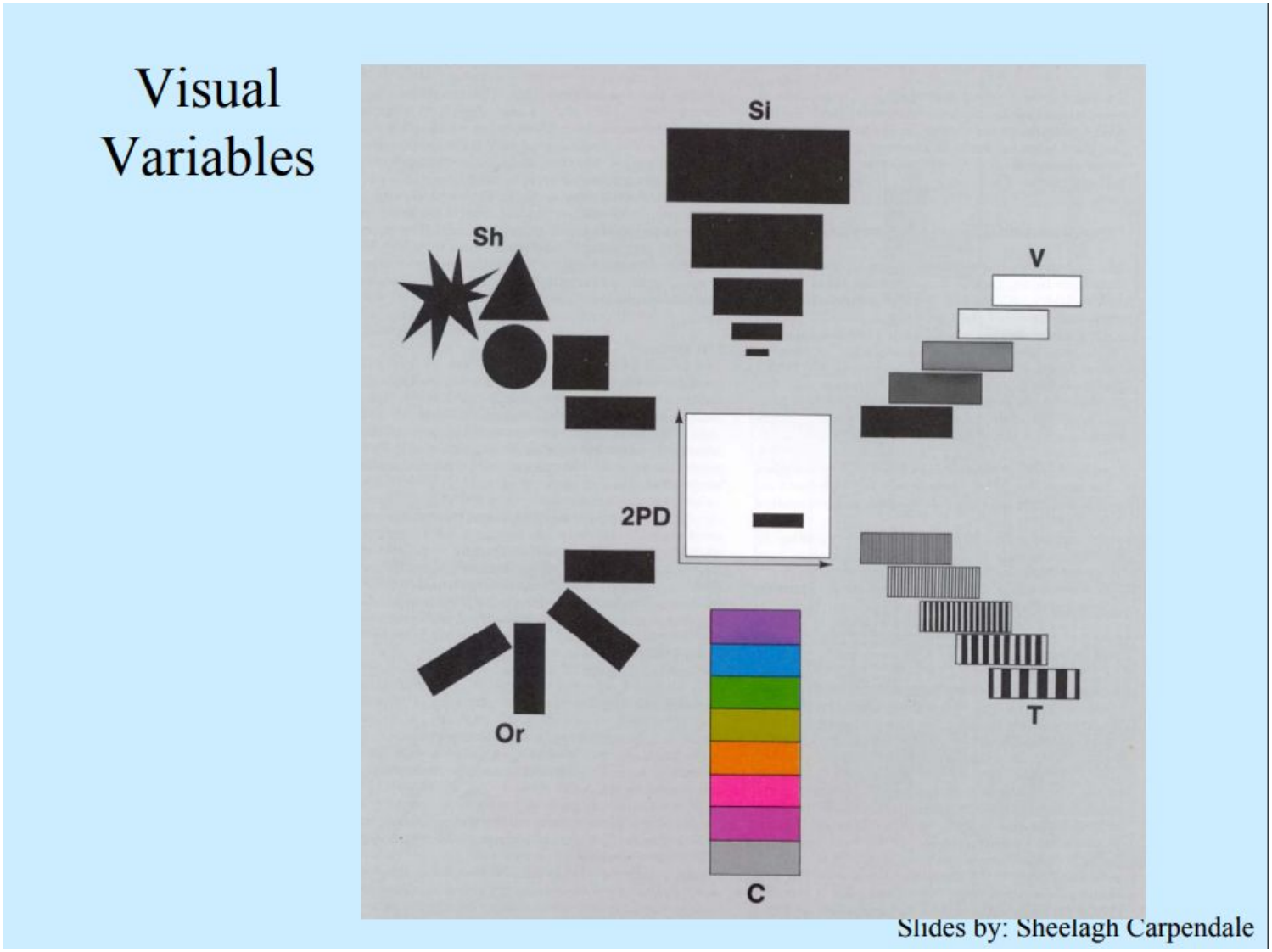


Visual Representation from Semiology of Graphics by J. Bertin

The Plane

- Points
  - *"A point represents a location on the plane that has no theoretical length or area. This signification is independent of the size and character of the mark which renders it visible."*
  - a location
  - marks that indicate points can vary in all visual variables
- Lines
  - *"A line signifies a phenomenon on the plane which has measurable length but no area. This signification is independent of the width and characteristics of the mark which renders it visible."*
  - a boundary, a route, a connection
- Areas
  - *"An area signifies something on the plane that has measurable size. This signification applies to the entire area covered by the visible mark."*
  - an area can change in position but not in size, shape or orientation without making the area itself have a different meaning

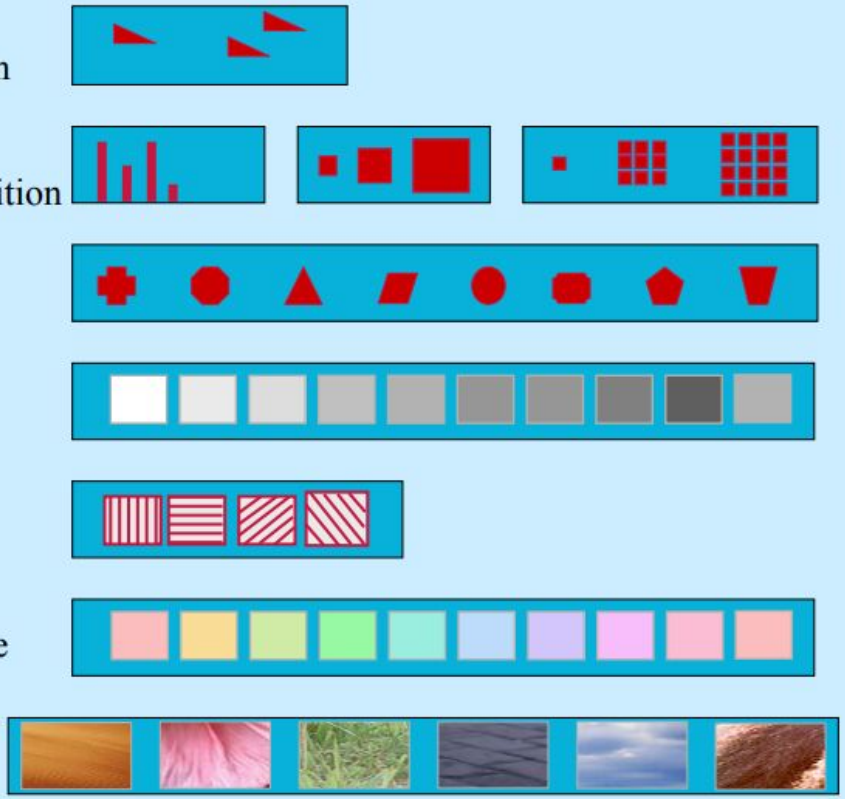
Visual Representation from Semiology of Graphics by J. Bertin



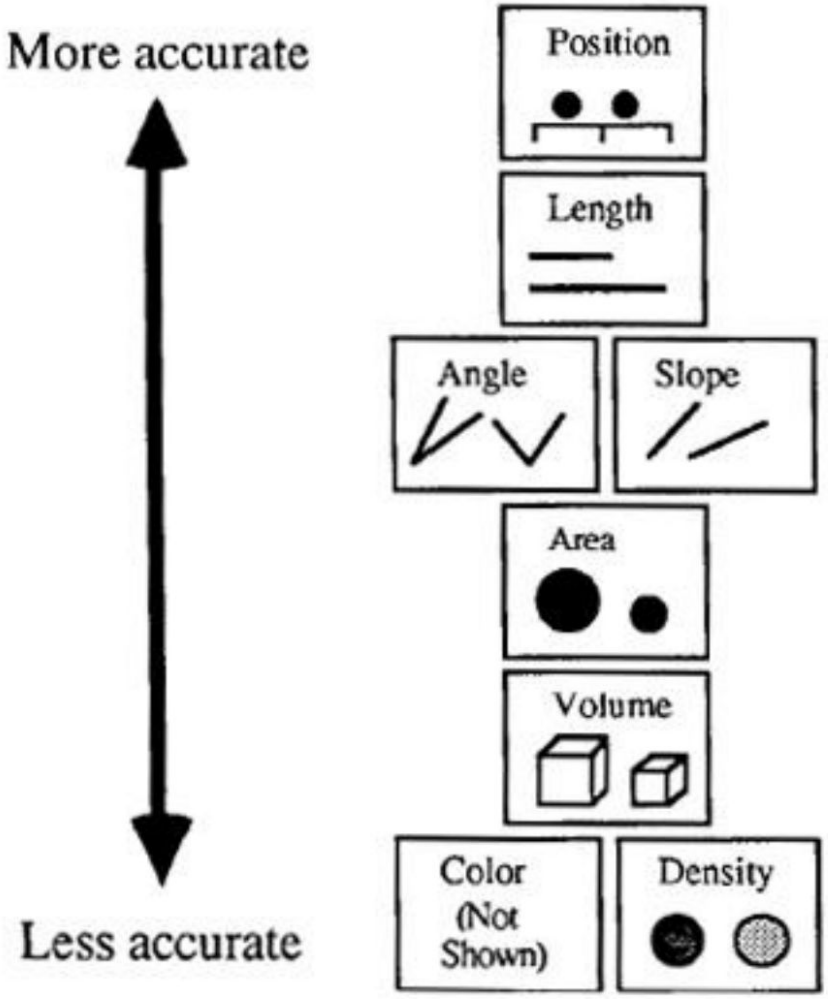
Visual Representation from Semiology of Graphics by J. Bertin

Visual Variables

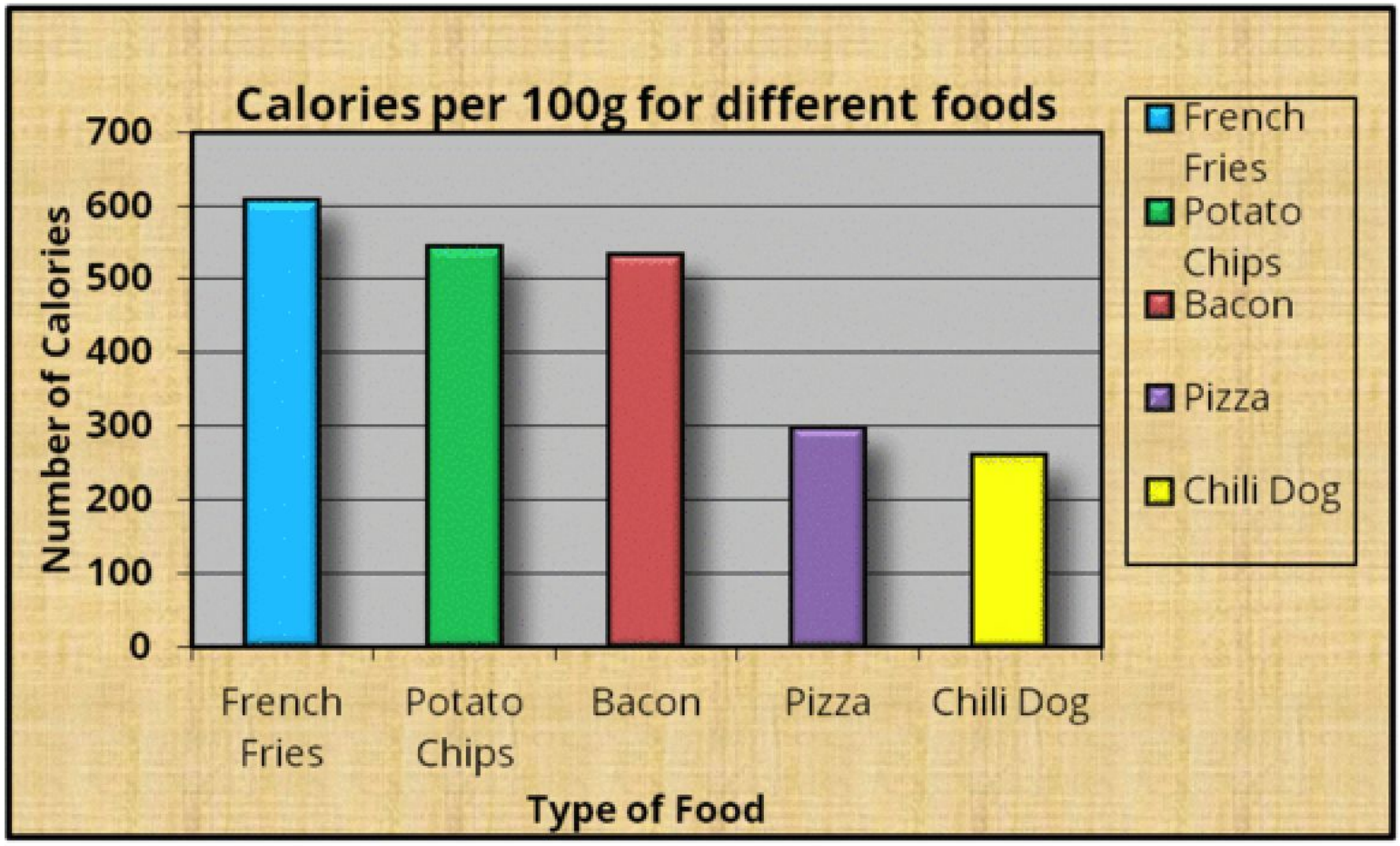
- **position**
  - changes in the x, y, (z) location
- **size**
  - change in length, area or repetition
- **shape**
  - infinite number of shapes
- **value**
  - changes from light to dark
- **orientation**
  - changes in alignment
- **colour**
  - changes in hue at a given value
- **texture**
  - variation in pattern
- **motion**

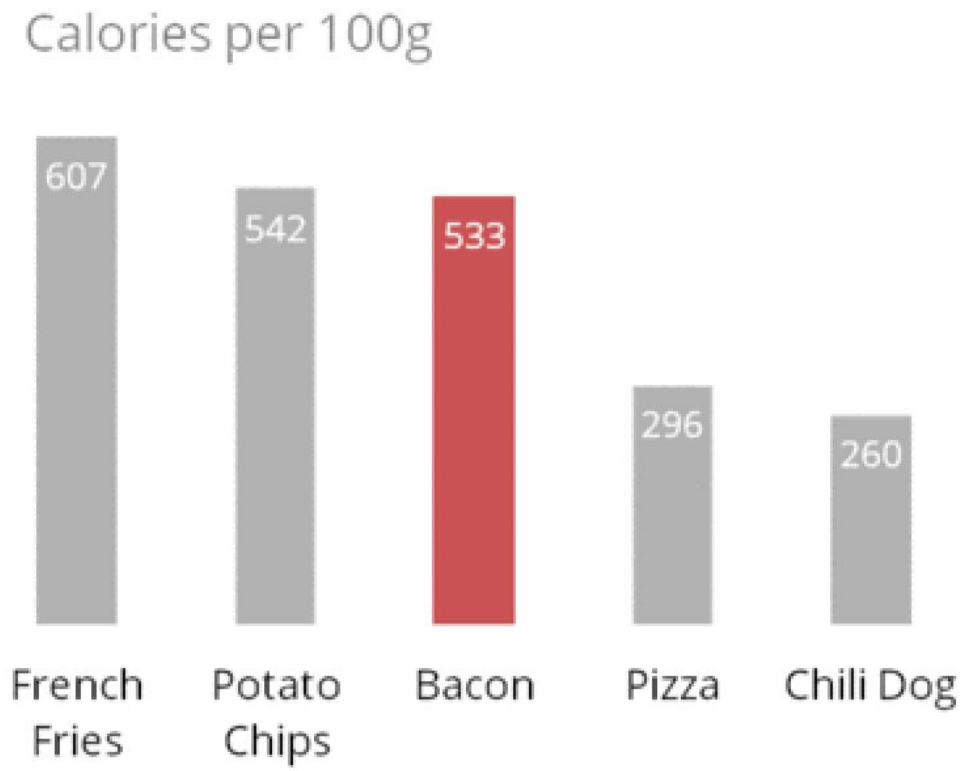


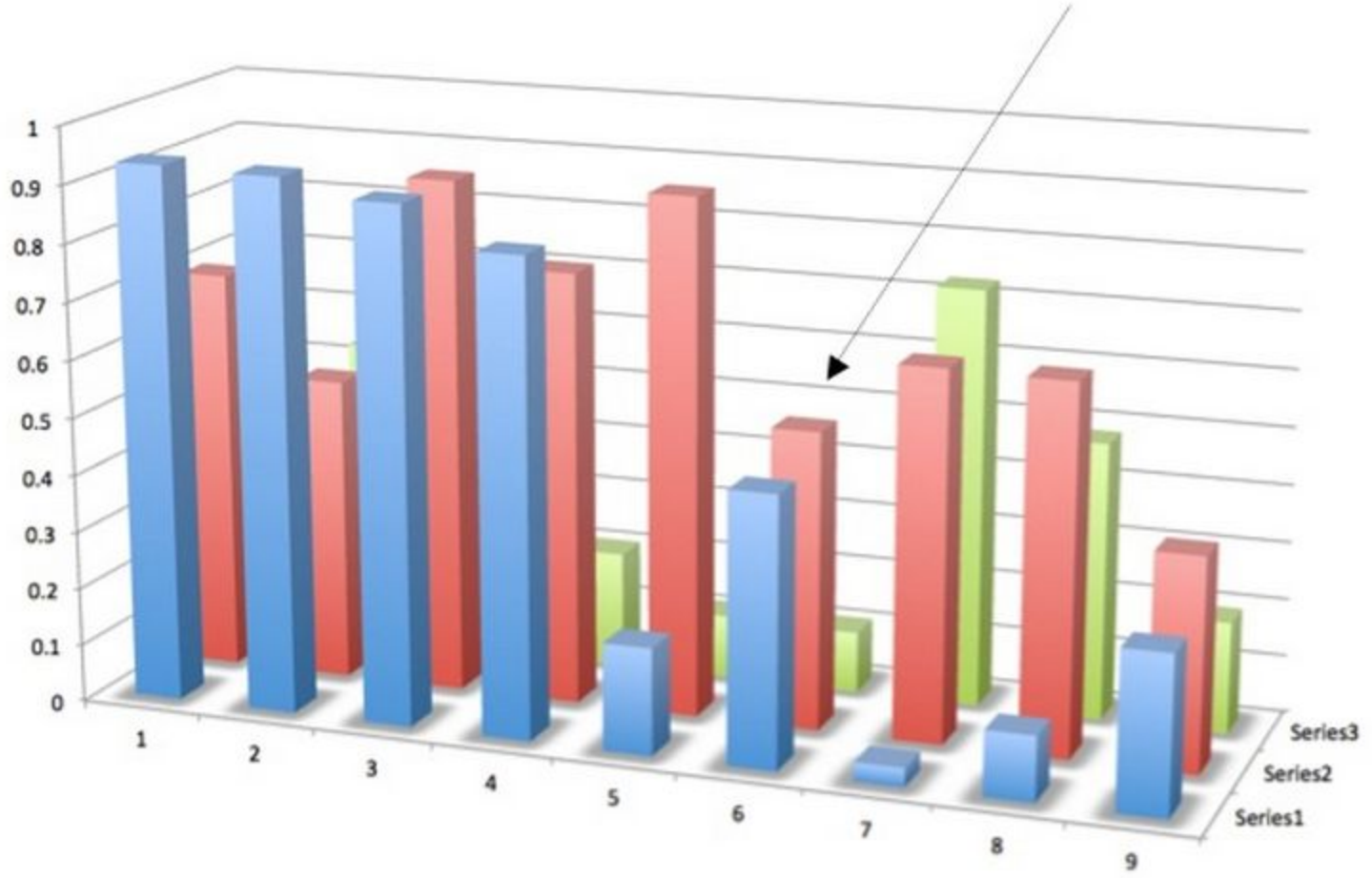
Slides by: Sheelagh Carpendale

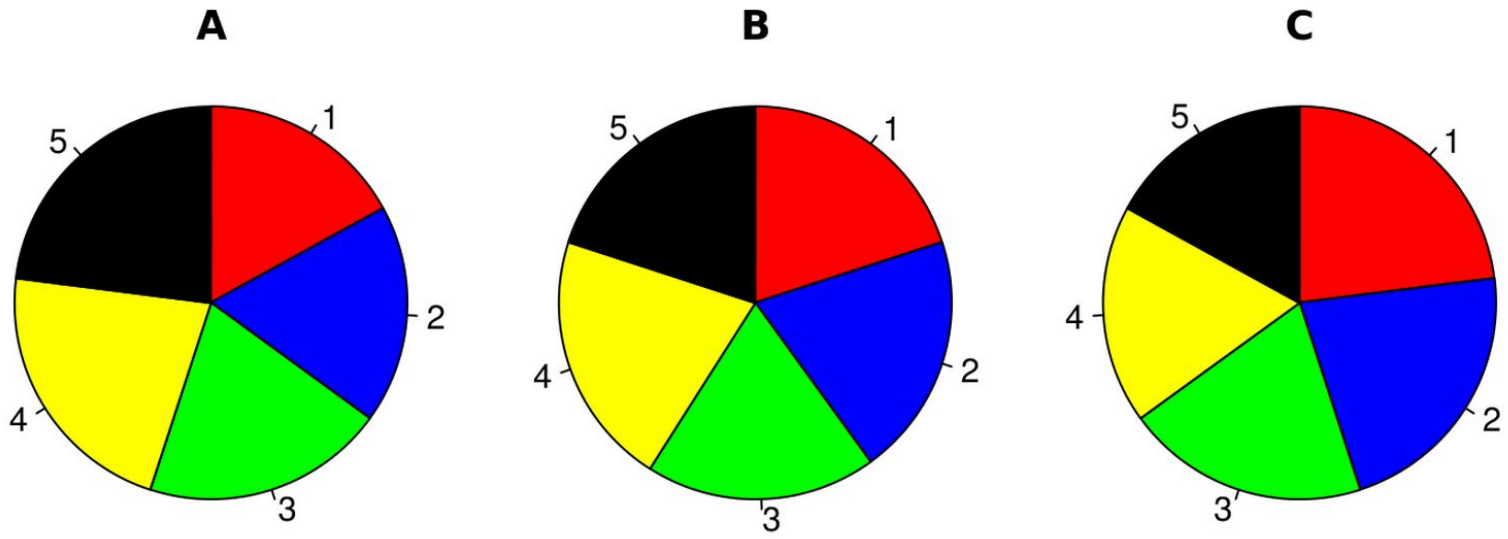

More accurate
Less accurate
Position
Length
Angle
Slope
Area
Volume
Color
(Not
Shown)
Density

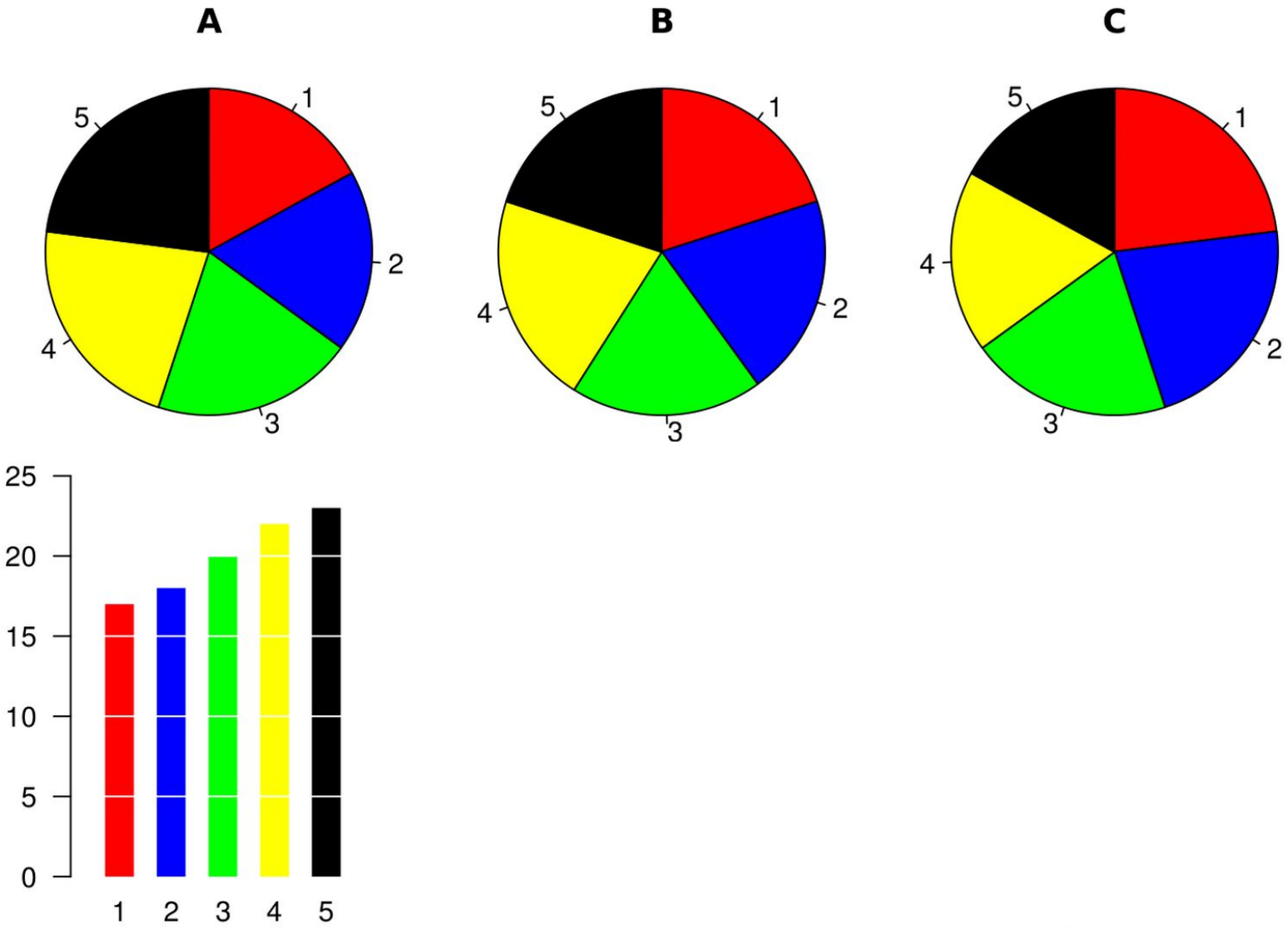
$$\text{Data-ink ratio} = \frac{\text{Data-ink}}{\text{Total ink used to print the graphic}}$$

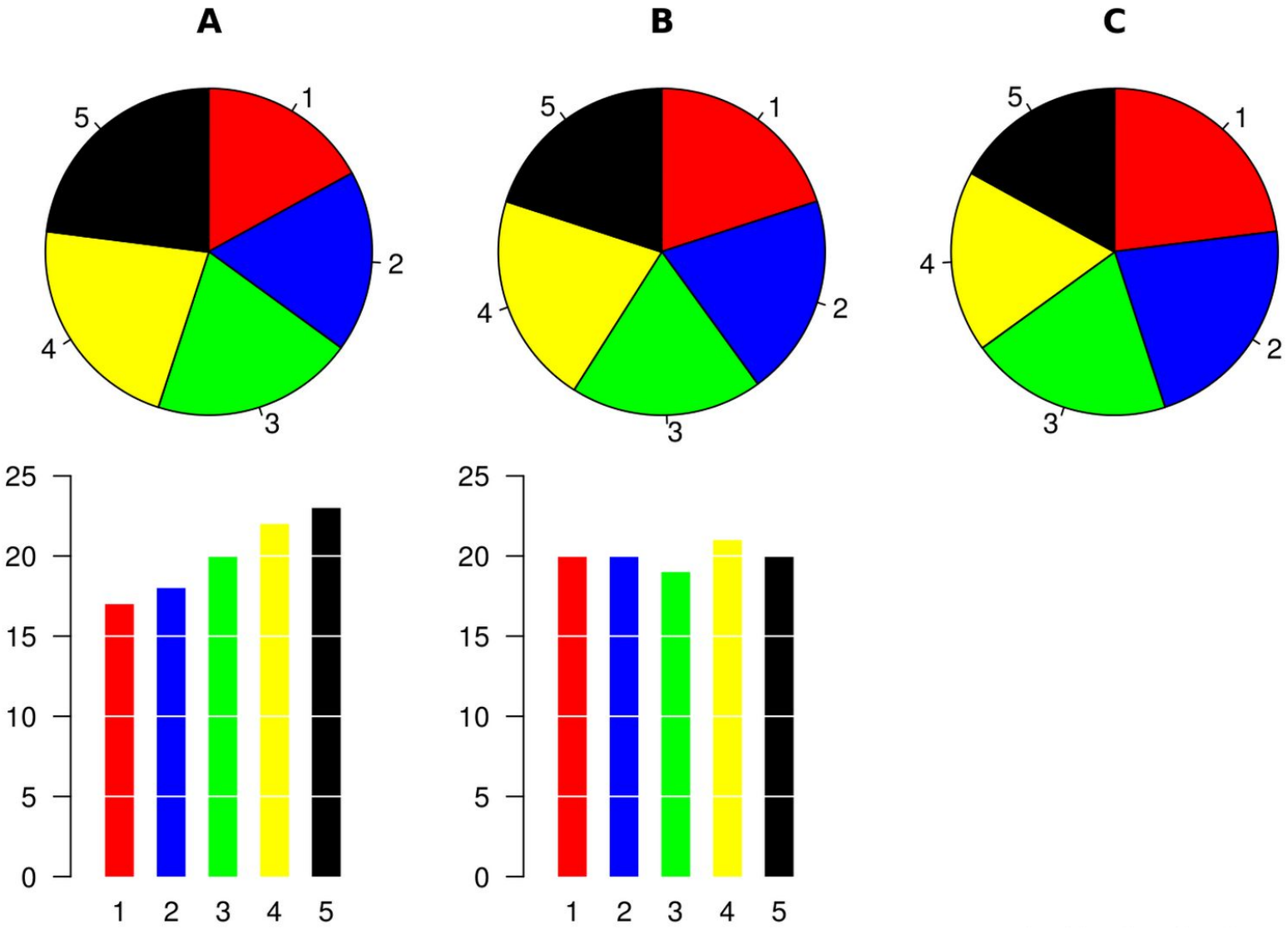



Calories per 100g
607
542
533
296
260
French Fries
Potato Chips
Bacon
Pizza
Chili Dog
Created by Darkhorse Analytics
www.darkhorseanalytics.com


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1
2
3
4
5
6
7
8
9
Series3
Series2
Series1


A
1
2
3
4
5
B
1
2
3
4
5
C
1
2
3
4
5


A
B
C
1
2
3
4
5
25
20
15
10
5
0


A
B
C
1
2
3
4
5
25
20
15
10
5
0

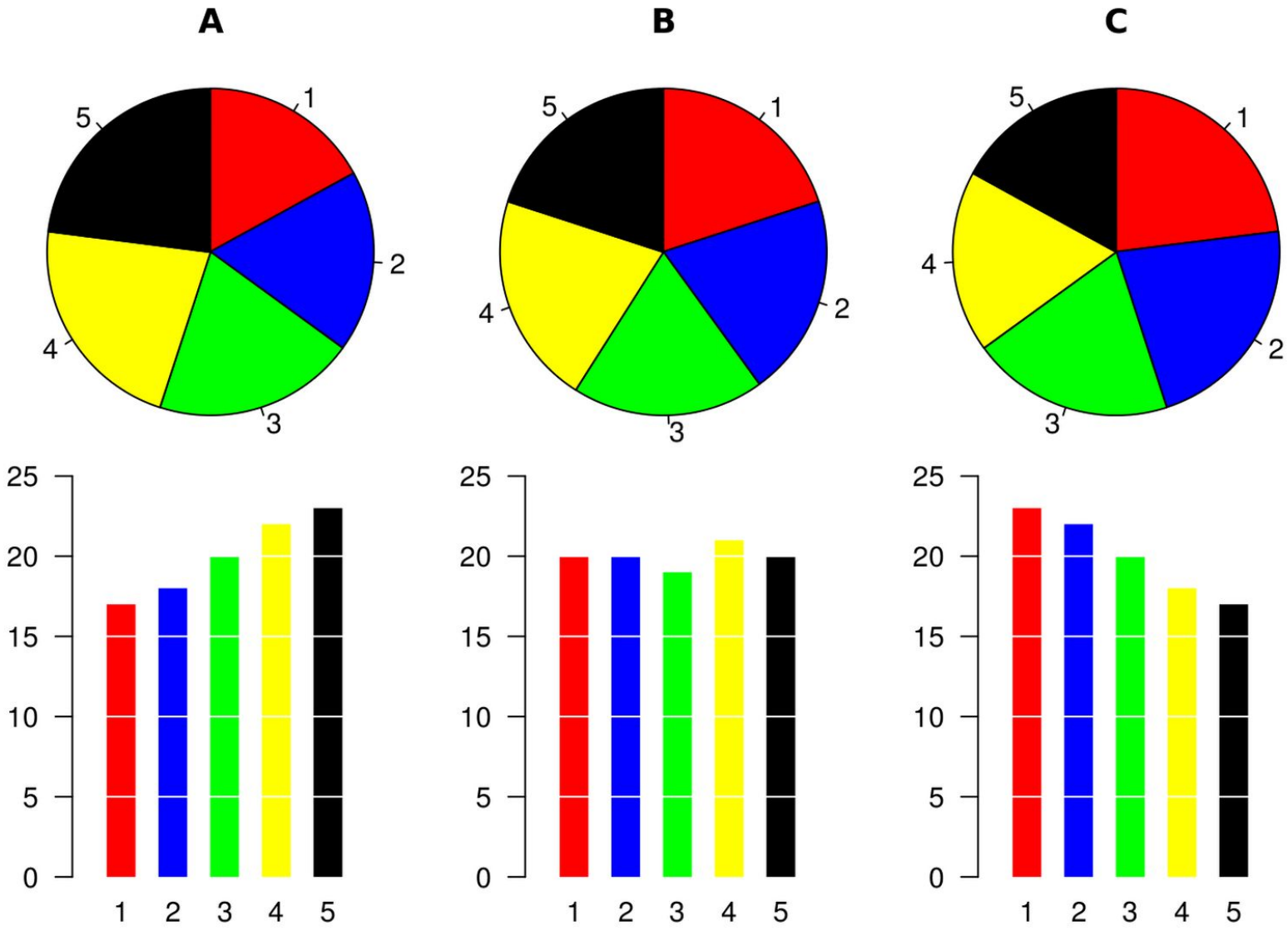
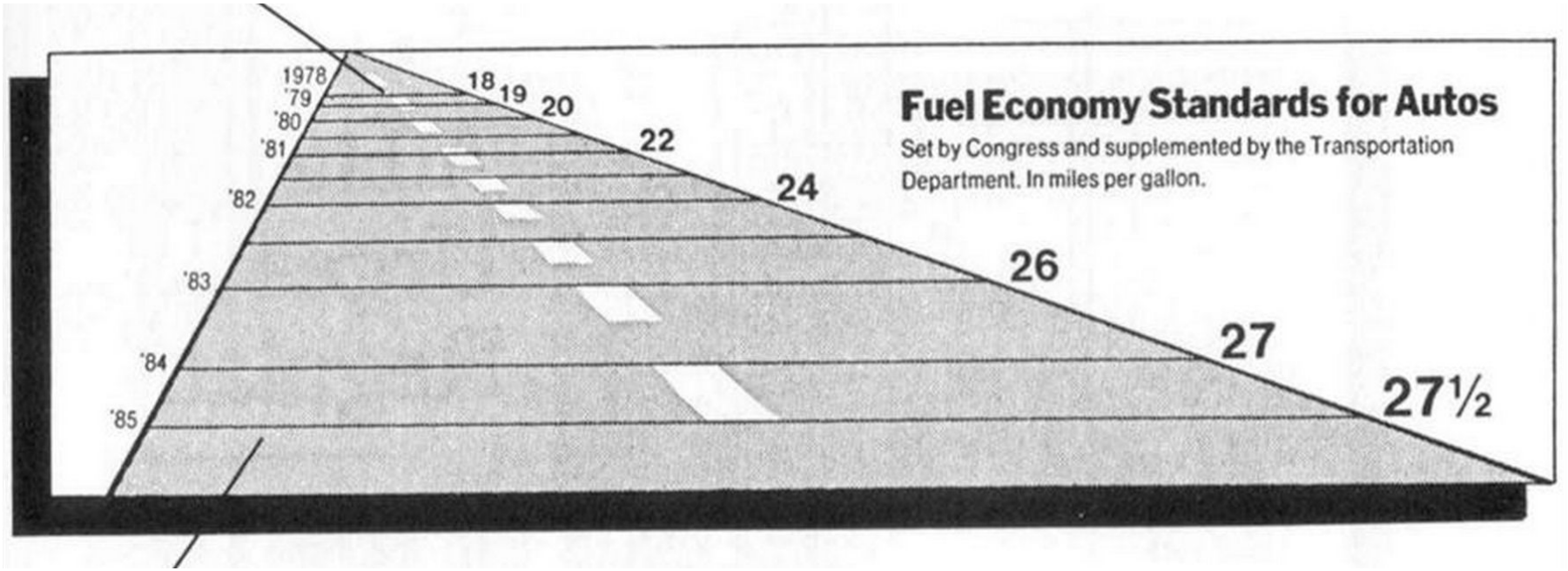
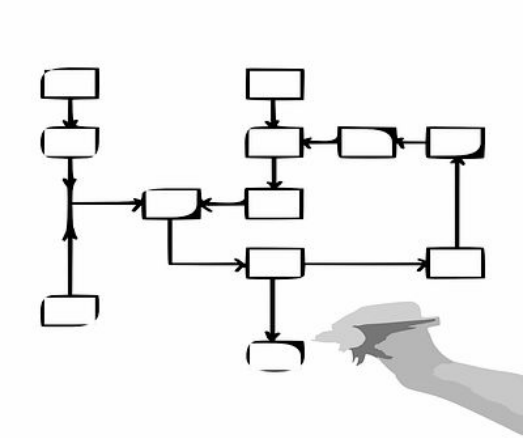

A
B
C
1
2
3
4
5
0
5
10
15
20
25

Chart junk


1978
'79
'80
'81
'82
'83
'84
'85
18
19
20
22
24
26
27
27½
Fuel Economy Standards for Autos
Set by Congress and supplemented by the Transportation Department. In miles per gallon.

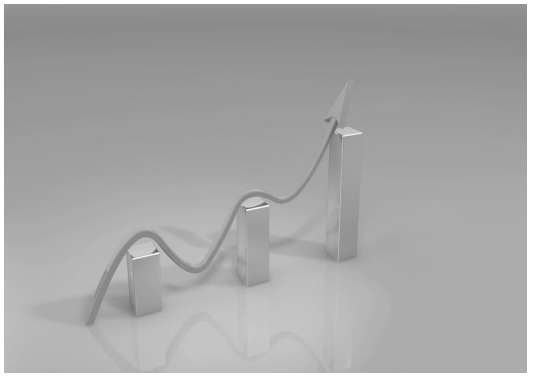

MONSTROUS COSTS
Total House and Senate campaign expenditures, in millions
$300
250
200
150
100
50
1972
'74
'76
'78
'80
'82 est.
VOTE

Topics to cover



Data Mining Process

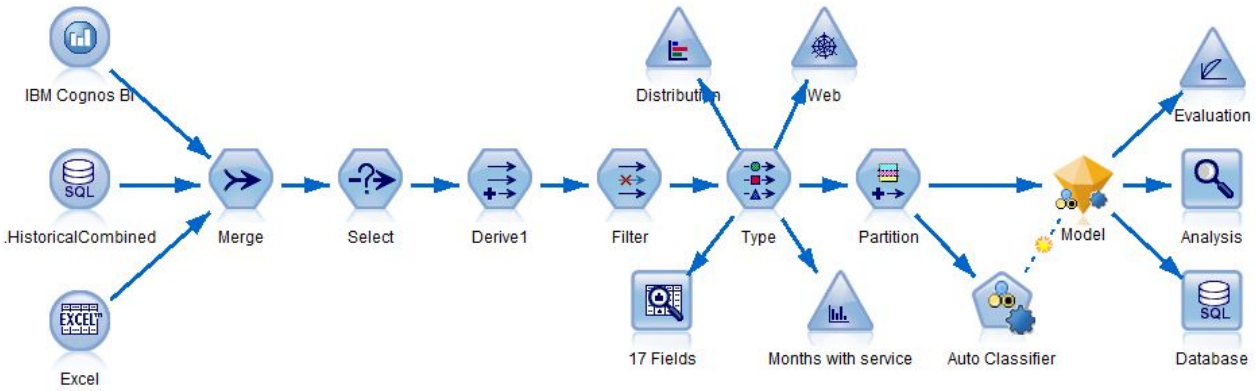
Exploratory Data Analysis



Data Visualization

Data Science Tools

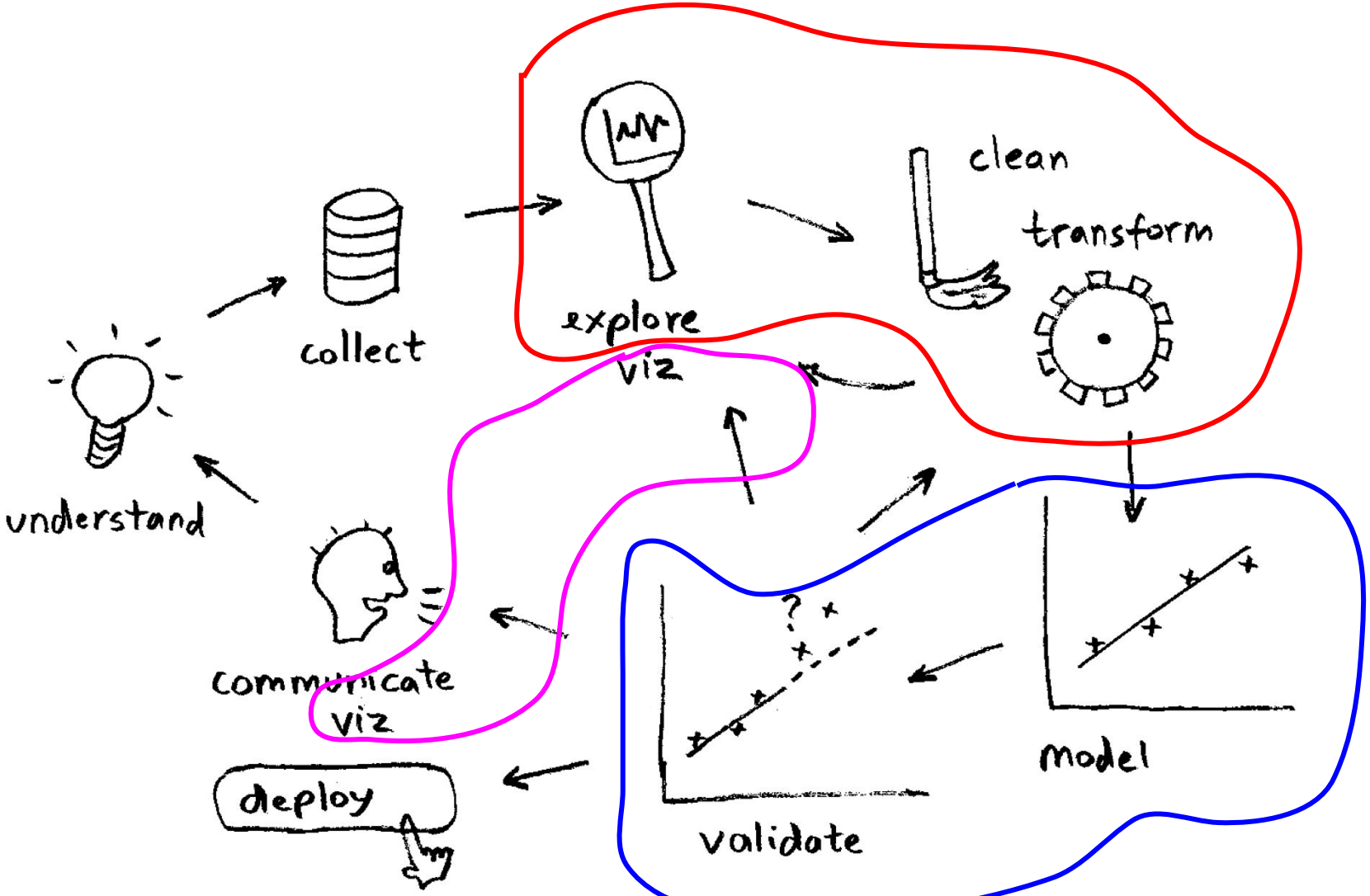
Reproducible Research


IBM Cognos BI
SQL
.HistoricalCombined
Excel
Merge
Select
Derive1
Filter
Distribution
Web
Type
Partition
17 Fields
Months with service
Auto Classifier
Model
Evaluation
Analysis
SQL
Database

Alice Zhao
@adashofdata

Follow

The Great Debate: If You Can't Code, You Can't Be a Data Scientist #TeamCode wins! #Strataconf


collect
explore
viz
clean
transform
understand
communicate
viz
deploy
validate
model

Tools I Use

knitr

RStudio

#!/bin/bash

MySQL

influxdb

PyData

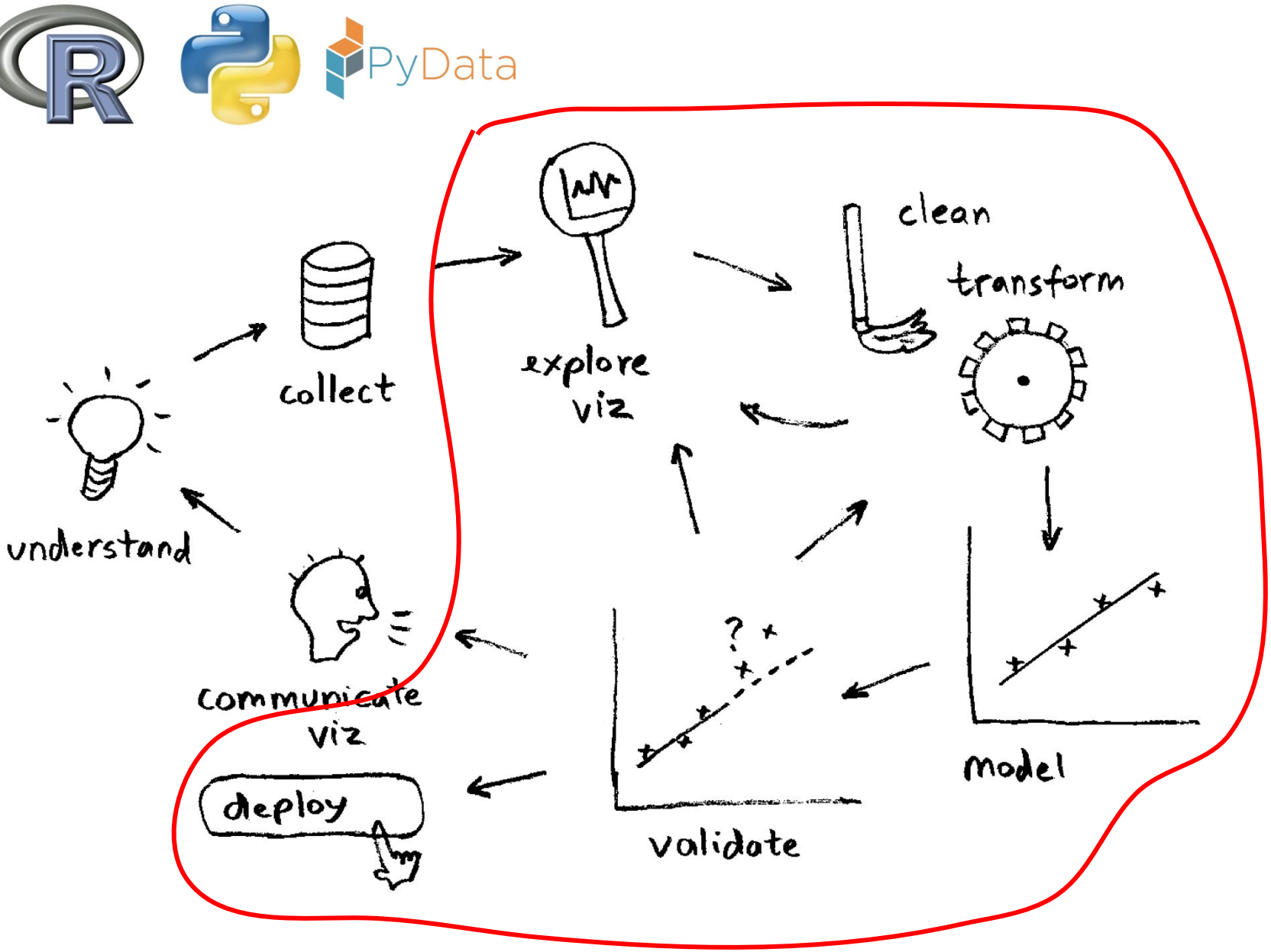
jupyter

kx

Tools Others Use (Survey)

LA Data Science/ML meetup, Apr 2014, 200 people

- Data munging: R 60%, Python 50%, SQL 40%, Hadoop (mostly Hive) 30%, Unix shell 20%, Excel 10% + Perl, Matlab, SAS, Impala, Pig, Shark...
- Visualization: R 40%,Python 30%, Tableau 10%, Javascript 10% + Matlab, Excel...
- Machine learning/modeling: R 30%, Python 30% + Vowpal Wabbit, Matlab, Mahout, SAS, SPSS...


PyData
collect
explore
viz
clean
transform
model
validate
communicate
viz
deploy
understand


```
Facet(gender, height).hist()
```

mplfacet

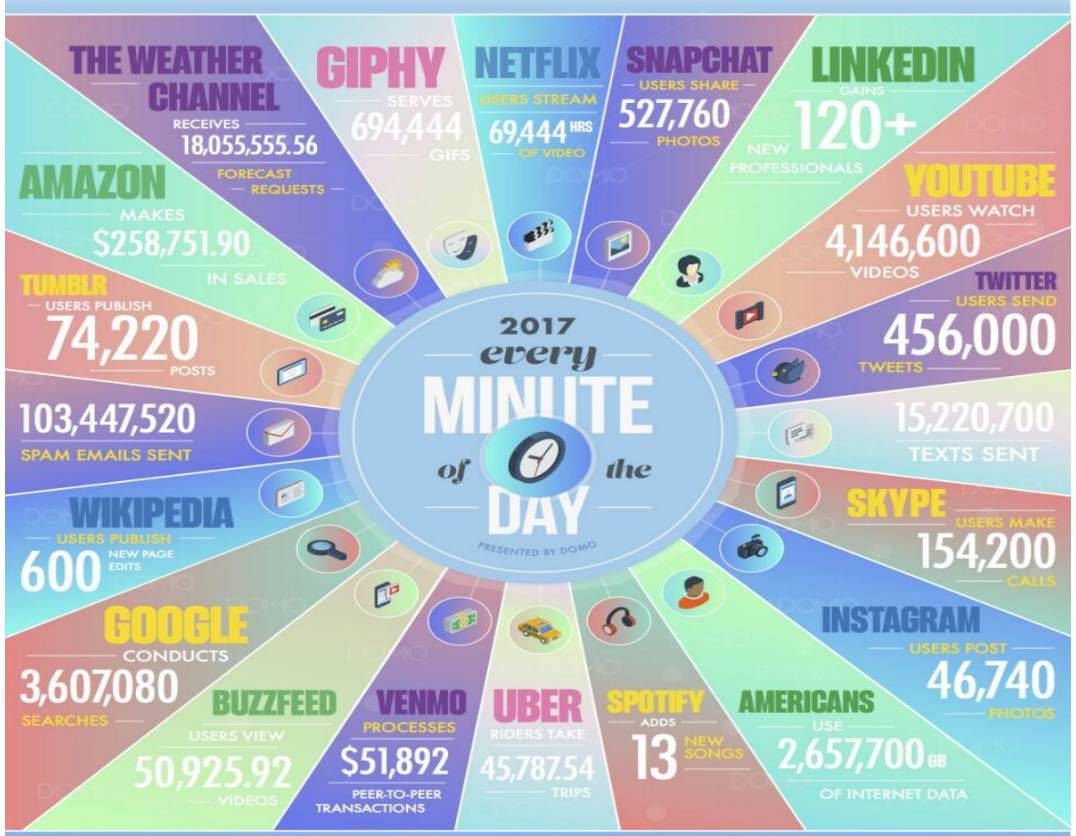
```
fig, axes = plt.subplots(nrows=1, ncols=2, sharex=True, sharey=True)

axes[0].hist(height[gender == 'Male'])
axes[0].set_title('Male')
axes[1].hist(height[gender == 'Female])
axes[1].set_title('Female')
```

Systems,
Databases
Big data

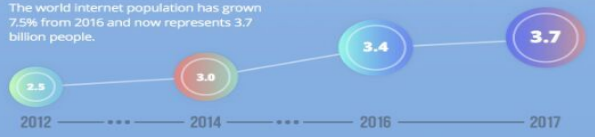
DATA NEVER SLEEPS 5.0

How much data is generated *every minute?*

90% of all data today was created in the last two years—that's 2.5 quintillion bytes of data per day. In our 5th edition of Data Never Sleeps, we bring you the latest stats on just how much data is being created in the digital sphere—and the numbers are staggering.



The world internet population has grown 7.5% from 2016 and now represents 3.7 billion people.

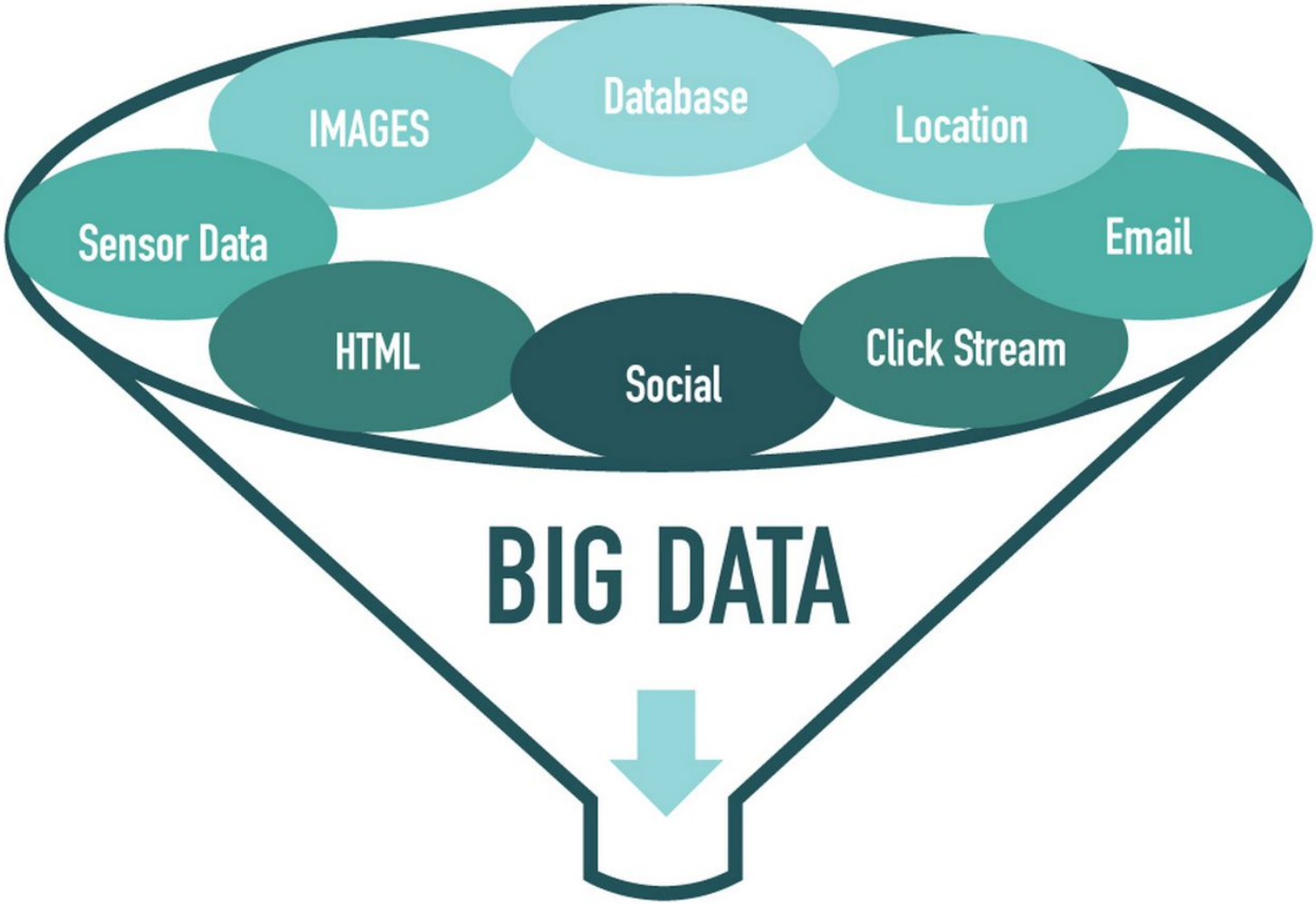


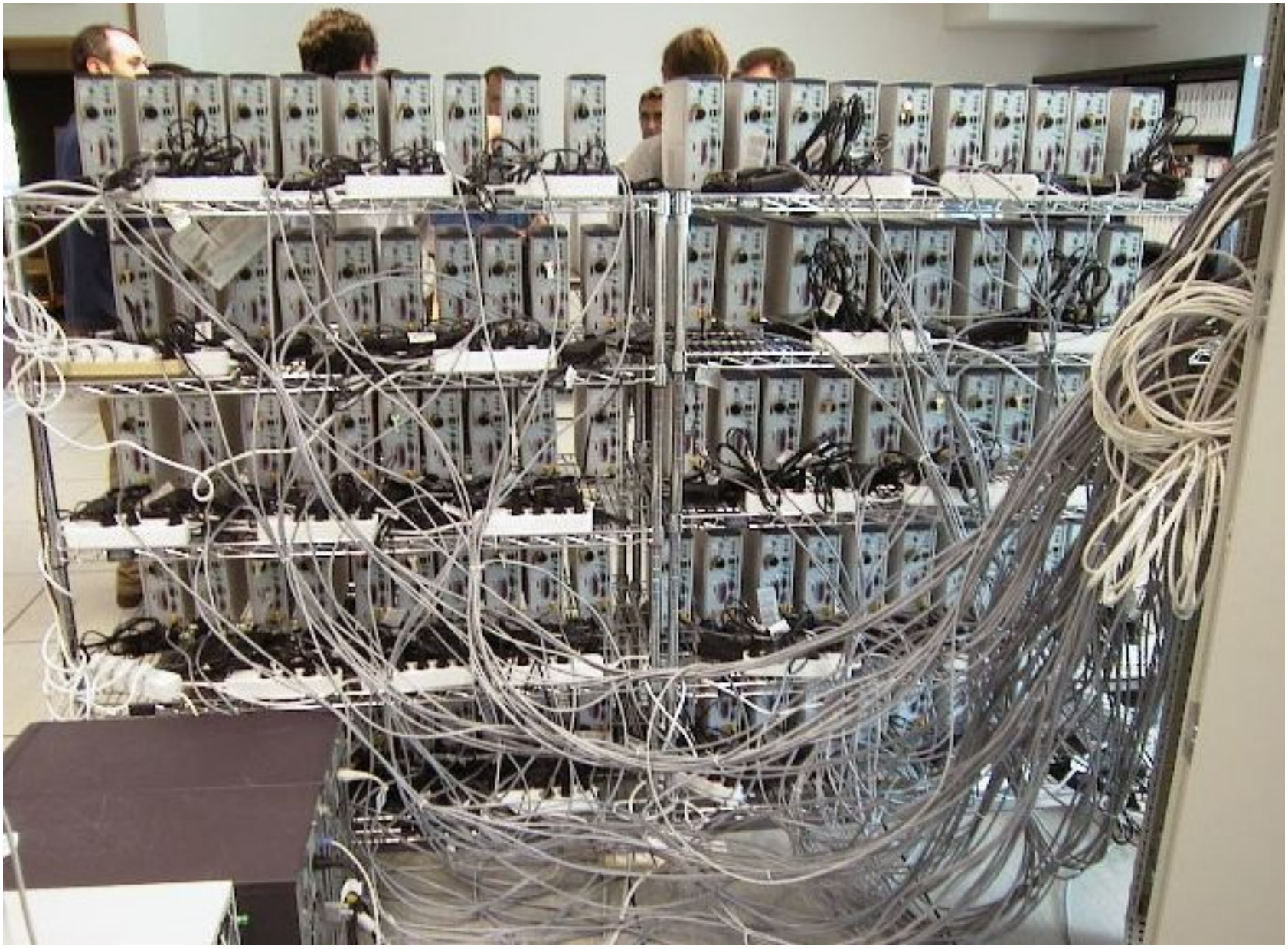
GLOBAL INTERNET POPULATION GROWTH 2012–2017
(IN BILLIONS)

With each click, swipe, share, and like, businesses are using data to make decisions about the future. Domo gives everyone in your business real-time access to data from virtually any data source in a single platform for smarter decision-making at any moment.

Learn more at domo.com

SOURCES: EXPANDEDRAMBLINGS.COM, WEARESOCIAL.COM, WIKIPEDIA, FORBES, ADWEEK.COM, FORTUNE.COM, BLOOMBERG.COM, ONEREACH.COM, IBM, BUZZFEED, INTERNET LIVE STATS, INTERNET WORLD STATS, BBC


IMAGES
Database
Location
Sensor Data
Email
HTML
Social
Click Stream
BIG DATA



Google Cloud Platform

aws

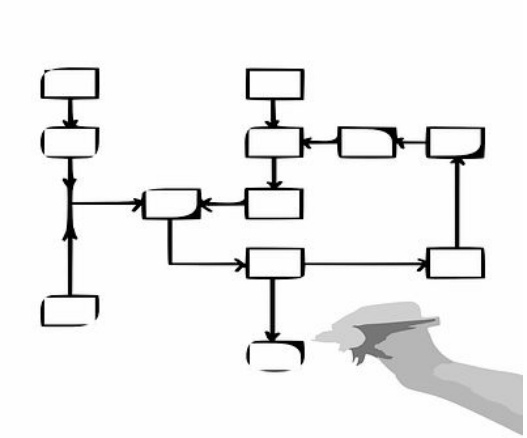
kx

APACHE Spark™

hadoop

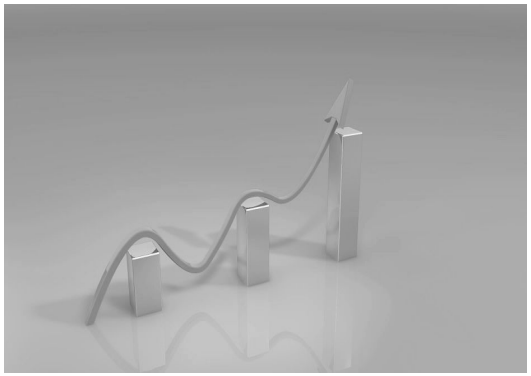

SAY BIG DATA
ONE MORE TIME

Topics to cover



Data Mining Process

Exploratory Data Analysis



Data Visualization

Data Science Tools

Reproducible Research

~/history – RStudio

diamondPricing.R* | diamonds

Source on Save | Run | Source

```
1  library(ggplot2)
2
3  View(diamonds)
4  summary(diamonds)
5  summary(diamonds$price)
6
7  qplot(carat, price, data=diamonds)
8
9  qplot(carat, price, data=diamonds, color=clarity,
10       xlab="Carat", ylab="Price",
11       main="Diamond Pricing") +
12         coord_cartesian(xlim=c(0, 3.5)) +
13         opts(plot.title=theme_text(size=23))
14
15
```

14:1 (Top Level) R Script

Console ~/

```
      x                y                z
 Min.   : 0.000   Min.   : 0.000   Min.   : 0.000
 1st Qu.: 4.710   1st Qu.: 4.720   1st Qu.: 2.910
 Median : 5.700   Median : 5.710   Median : 3.530
 Mean   : 5.731   Mean   : 5.735   Mean   : 3.539
 3rd Qu.: 6.540   3rd Qu.: 6.540   3rd Qu.: 4.040
 Max.   :10.740   Max.   :58.900   Max.   :31.800

> summary(diamonds$price)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    326     950    2401    3933    5324   18820
> qplot(carat, price, data=diamonds)
> qplot(carat, price, data=diamonds, color=clarity,
xlab="Carat", ylab="Price", main="Diamond Pricing") +
coord_cartesian(xlim=c(0, 3.5)) +
opts(plot.title=theme_text(size=23))
>
```

Workspace | History

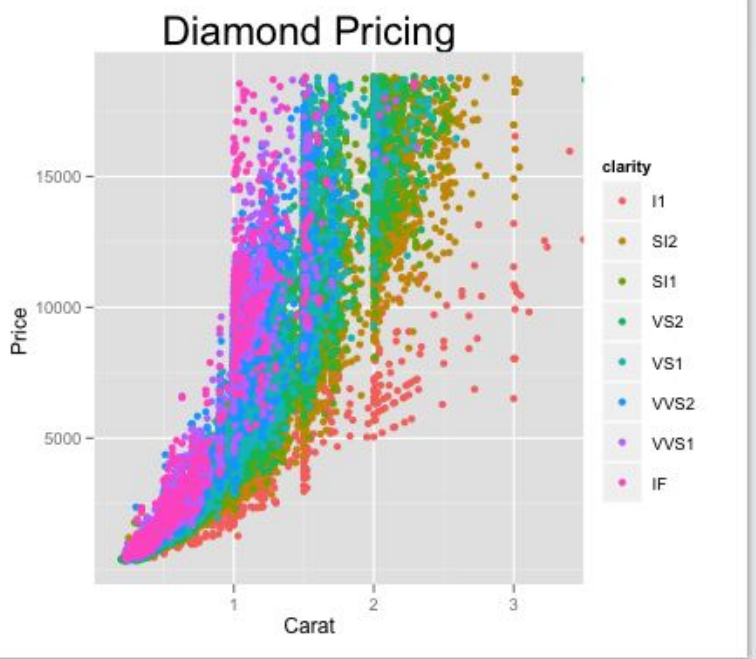
To Console | To Source | qplot | Done

Search results: qplot

```
qplot(carat, price, data=diamonds)                    2011-03-03 07:12
qplot(carat, price, data=diamonds,
    color=clarity, xlab="Carat",
    ylab="Price", main="Diamond Pricing") 2011-03-02 14:03
    + coord_cartesian(xlim=c(0, 3.5)) +
    opts(plot.title=theme_text(size=23))
qplot(carat, binwidth=0.05, data=diamonds) 2011-03-02 14:00
qplot(log(x), log(y)) + geom_smooth(method 2011-03-01 13:23
qplot(x, y, xlab="Year", ylab="Users in Mi 2011-03-01 13:20
qplot(x, y) + geom_smooth(method="loess")  2011-03-01 13:18
qplot(x, y)                                2011-03-01 13:17
```

Files | Plots | Packages | Help

Zoom | Export | Clear All



IPy IPython Dashboard | IPy spectrogram

127.0.0.1:8888/a5222740-848b-4ac1-b212-d732c9f8f78b

IP[y]: Notebook spectrogram Last sa

File Edit View Insert Cell Kernel Help

Markdown

Simple spectral analysis

An illustration of the Discrete Fourier Transform

$$X_k = \sum_{n=0}^{N-1} x_n e^{-\frac{2\pi i}{N} kn} \qquad k = 0, \ldots, N-1$$

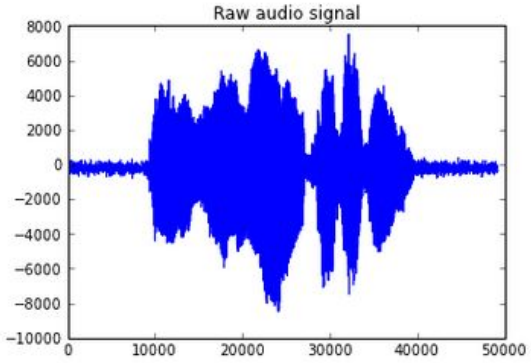
using windowing, to reveal the frequency content of a sound signal.

We begin by loading a datafile using SciPy's audio file support:

In [1]:
```
from scipy.io import wavfile
rate, x = wavfile.read('test_mono.wav')
```

And we can easily view its spectral structure using matplotlib's builtin specg

In [2]:
```
fig, (ax1, ax2) = plt.subplots(1, 2, figsize=(12,
ax1.plot(x); ax1.set_title('Raw audio signal')
ax2.specgram(x); ax2.set_title('Spectrogram');
```



RStudio

chunks.Rmd

Knit HTML | Chunks

```
1  R Code Chunks
2  ==========
3
4  With R Markdown, you can insert R code
   chunks including plots:
5
6  ```{r qplot, fig.width=4, fig.height=3,
   message=FALSE}
7  # quick summary and plot
8  library(ggplot2)
9  summary(cars)
10 qplot(speed, dist, data=cars) +
11       geom_smooth()
12 ```
13
```

RStudio: Preview HTML

Preview: ~/chunks.html | Save As | Publish

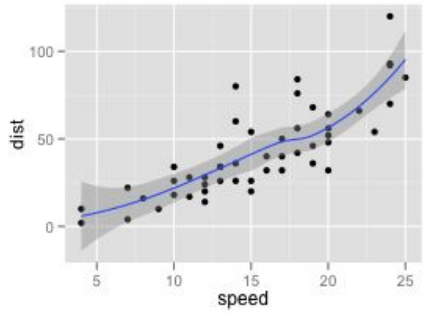
R Code Chunks

With R Markdown, you can insert R code chunks including plots:

```
# quick summary and plot
library(ggplot2)
summary(cars)
```

```
##      speed           dist
##  Min.   : 4.0   Min.   :  2
##  1st Qu.:12.0   1st Qu.: 26
##  Median :15.0   Median : 36
##  Mean   :15.4   Mean   : 43
##  3rd Qu.:19.0   3rd Qu.: 56
##  Max.   :25.0   Max.   :120
```

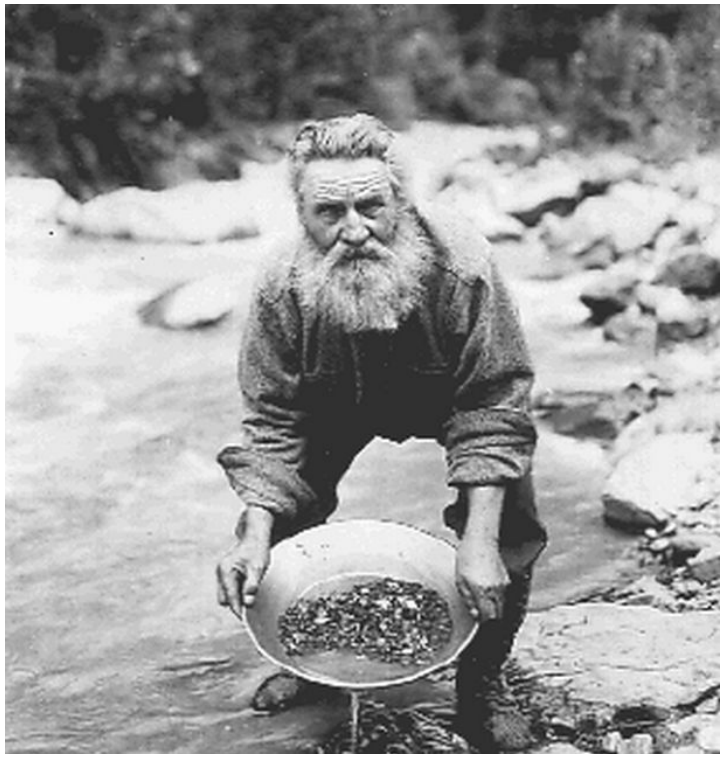
```
qplot(speed, dist, data = cars) + geom_smooth()
```



Query 1

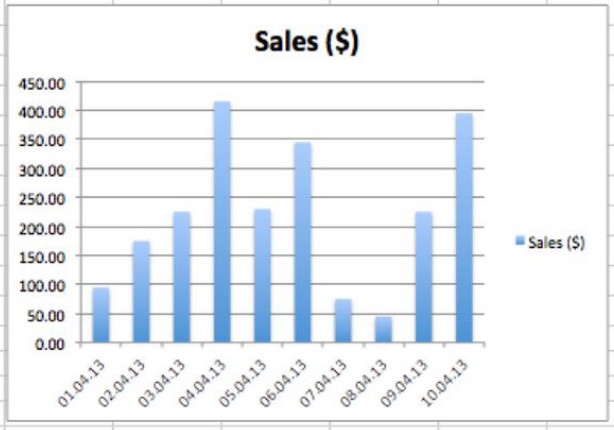
```
SELECT `actor`.`actor_id`,
    `actor`.`first_name`,
    `actor`.`last_name`,
    `actor`.`last_update`
FROM `sakila`.`actor`;

SELECT `film`.`film_id`,
    `film`.`title`,
    `film`.`description`,
    `film`.`release_year`,
    `film`.`language_id`,
    `film`.`original_language_id`,
    `film`.`rental_duration`,
    `film`.`rental_rate`,
    `film`.`length`,
    `film`.`replacement_cost`,
    `film`.`rating`,
    `film`.`special_features`,
```

Result Set Filter: Edit: Export/Import:



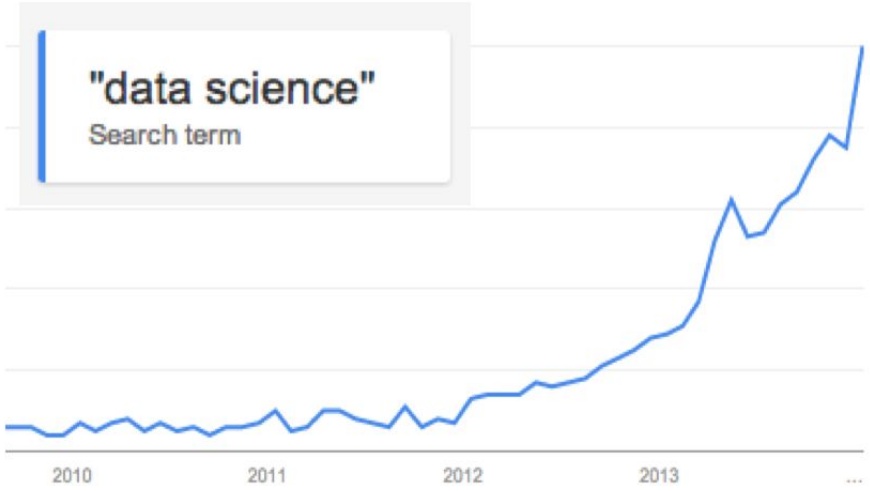
C11 fx A very good day

Date	Sales ($)	Comment
01.04.13	95.00	A good Monday
02.04.13	175.00	Looks better now
03.04.13	225.00	Even better
04.04.13	415.00	Ever high!
05.04.13	230.00	At the average
06.04.13	345.00	Above the average
07.04.13	75.00	Not so bad for a Sunday
08.04.13	45.00	A rather bad Monday
09.04.13	225.00	A very good Tuesday
10.04.13	395.00	A very good day



Data Science New?

Is Data Science New?



Data Science is a newly emerging field dedicated to analyzing and manipulating data to derive insights and build data products.

https://www.kaggle.com/wiki/WhatIsDataScience

Harvard Business Review 2012

Data Scientist: The Sexiest Job of the 21st Century

The United States alone faces a shortage of 140,000 to 190,000 people with deep analytical skills. [2011]

http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

Four major influences act on data analysis today:

1. The formal theory of statistics
2. Revolutionary developments in computers and display devices
3. The challenge, in many fields, of more and ever larger bodies of data
4. The accelerating emphasis on quantification in an ever wider variety of disciplines

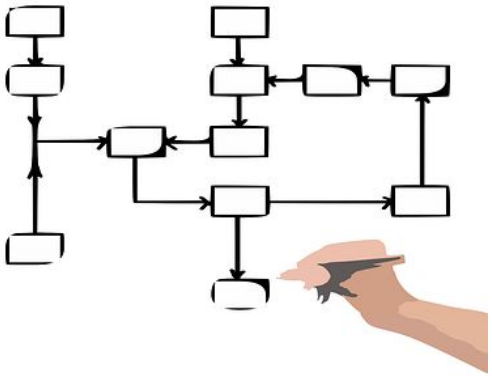
Four major influences act on data analysis today:

1. The formal theory of statistics
2. Revolutionary developments in computers and display devices
3. The challenge, in many fields, of more and ever larger bodies of data
4. The accelerating emphasis on quantification in an ever wider variety of disciplines

Tukey & Wilk, 1965

Tukey, J.W., & Wilk, M.B. (1965). Data analysis and statistics: techniques and approaches Reprinted in The Collected Works of John W. Tukey, Vol. V, Graphics 1965- 1985, 1-22 (1988)

Topics covered



Data Mining Process

Exploratory Data Analysis

Data Visualization

Data Science Tools

Reproducible Research