Introduction to Machine Learning

Goals

1. Understand the role of Machine Learning
2. Where Machine Learning fits into Information Technology strategies
3. Technical and business drivers
4. What it takes to be Data-Driven
5. Basic workflows for experimentation and deployment
6. Difference between Supervised and Unsupervised learning
7. Visualization strategies for understanding
8. How Machine Learning is being used at Salesforce
9. How Machine Learning can go wrong

Agenda

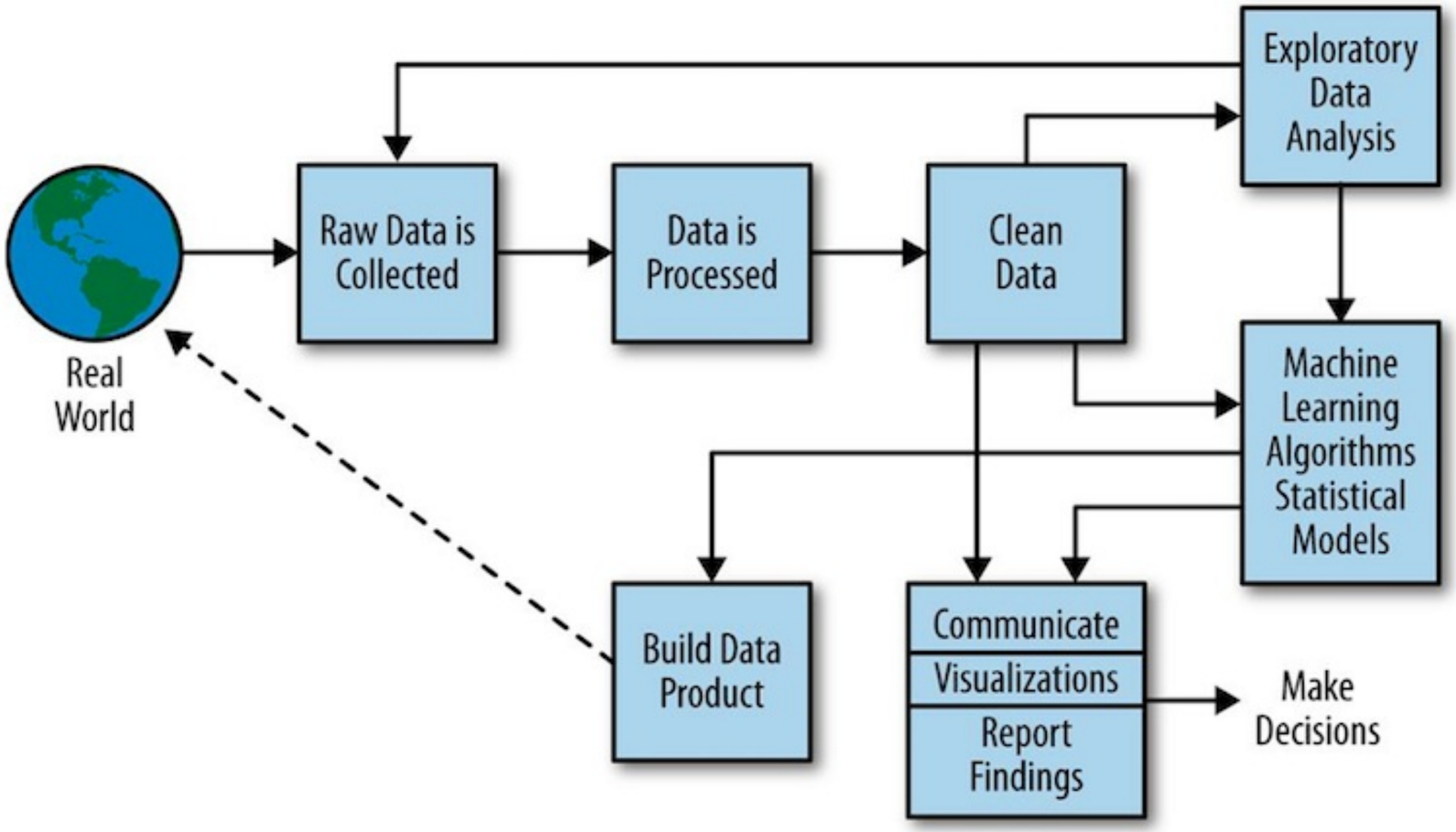
1. Introduction
2. Data and Data Processing
3. Data Sources
4. Data-Driven
5. Visualization
6. Data Science
7. Data-Directed
8. Infrastructure Demo

Introduction

Student Assessment

[Note] : We will do a quick assessment of the student background to help drive the course

Data Science Pipeline

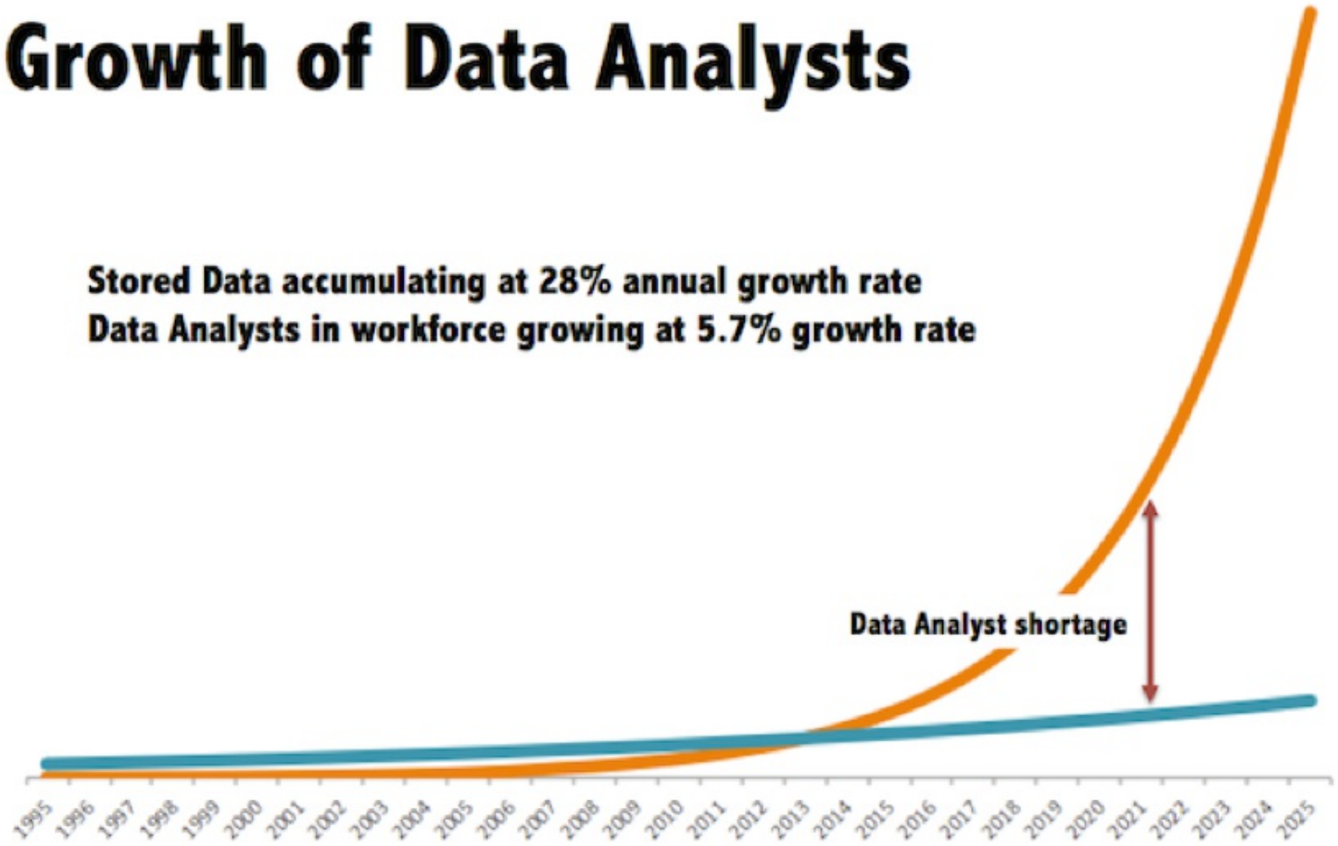


"Doing Data Science"

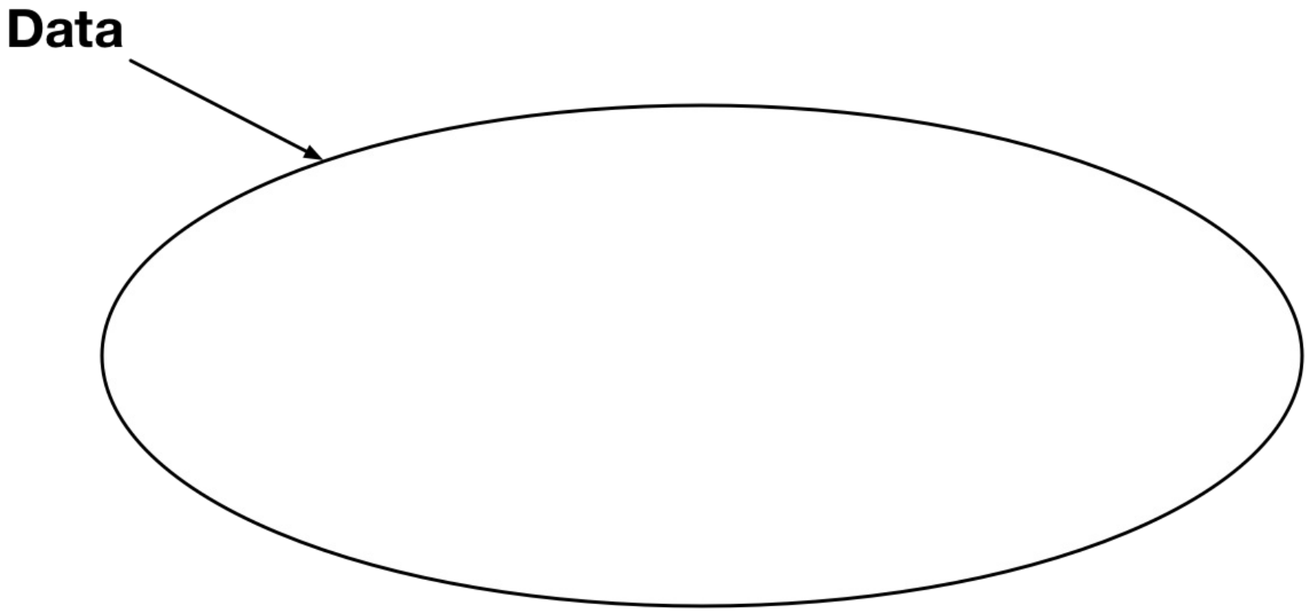
Data and Data Processing

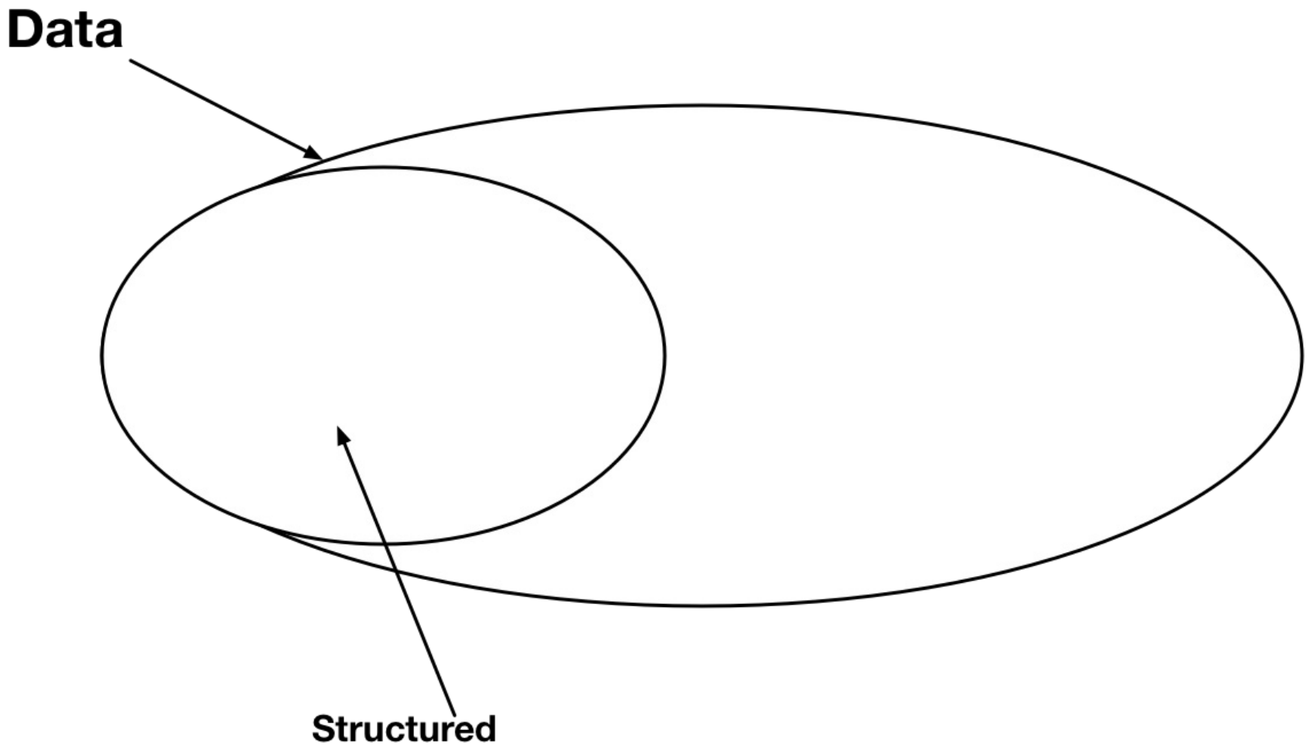
Growth of Data vs. Growth of Data Analysts

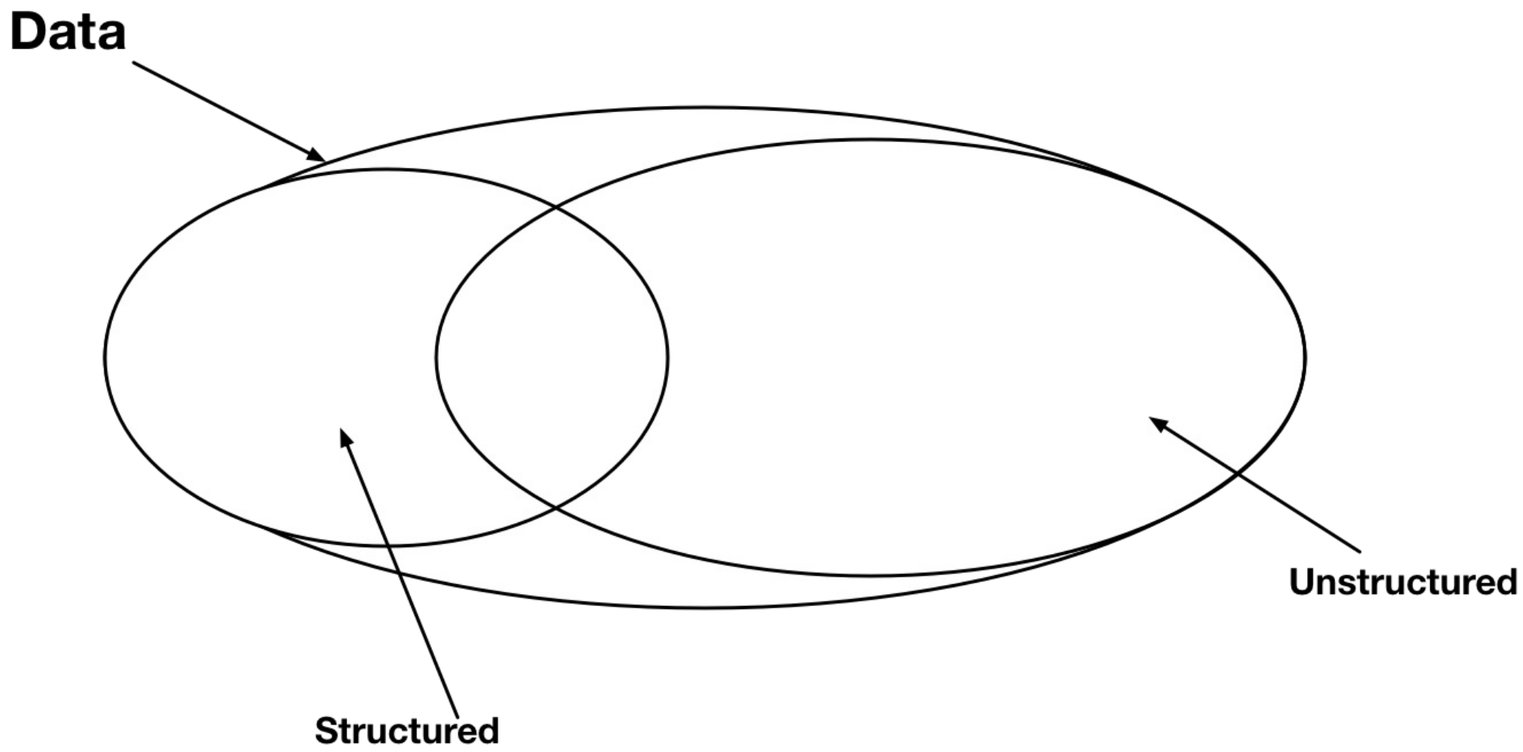
Stored Data accumulating at 28% annual growth rate
Data Analysts in workforce growing at 5.7% growth rate

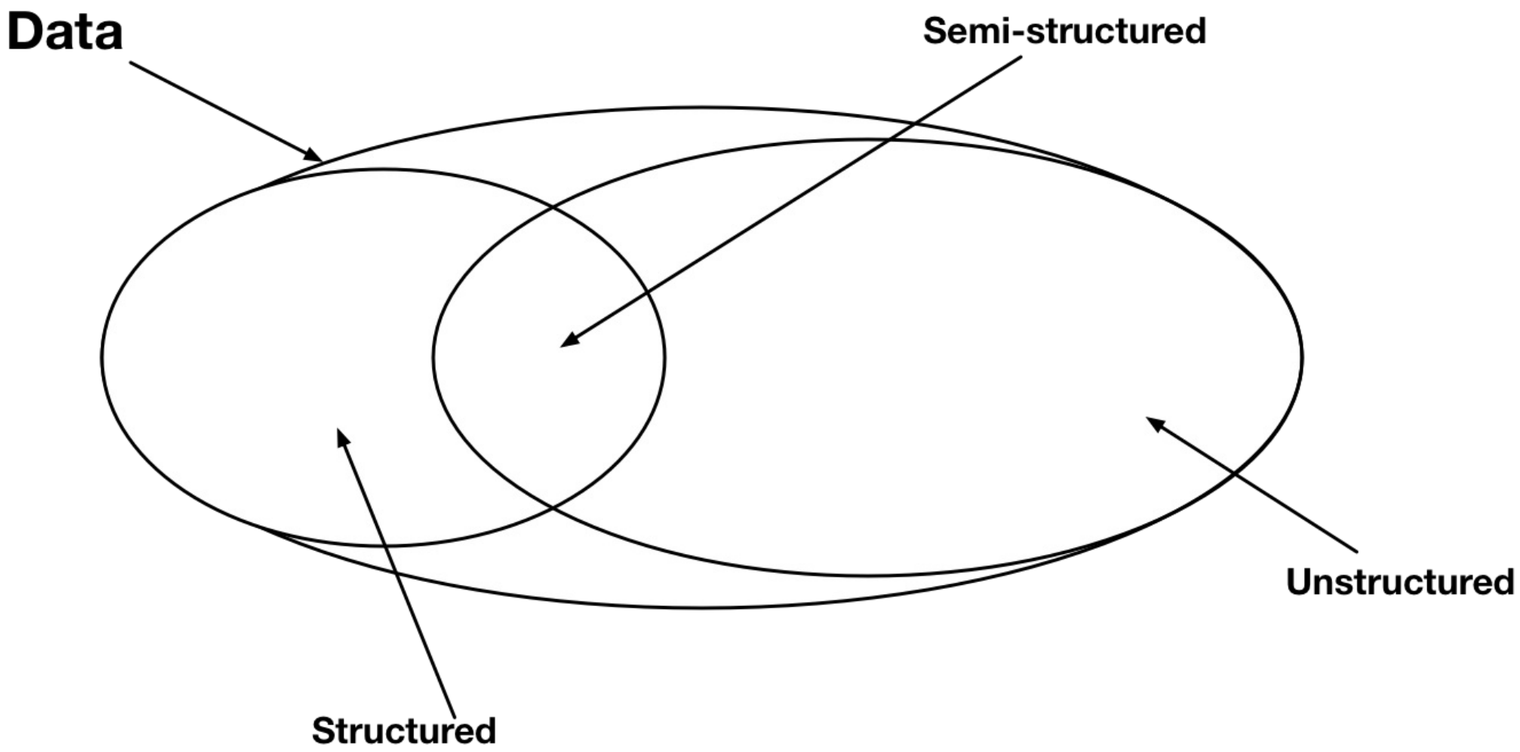


What is Data?


Data

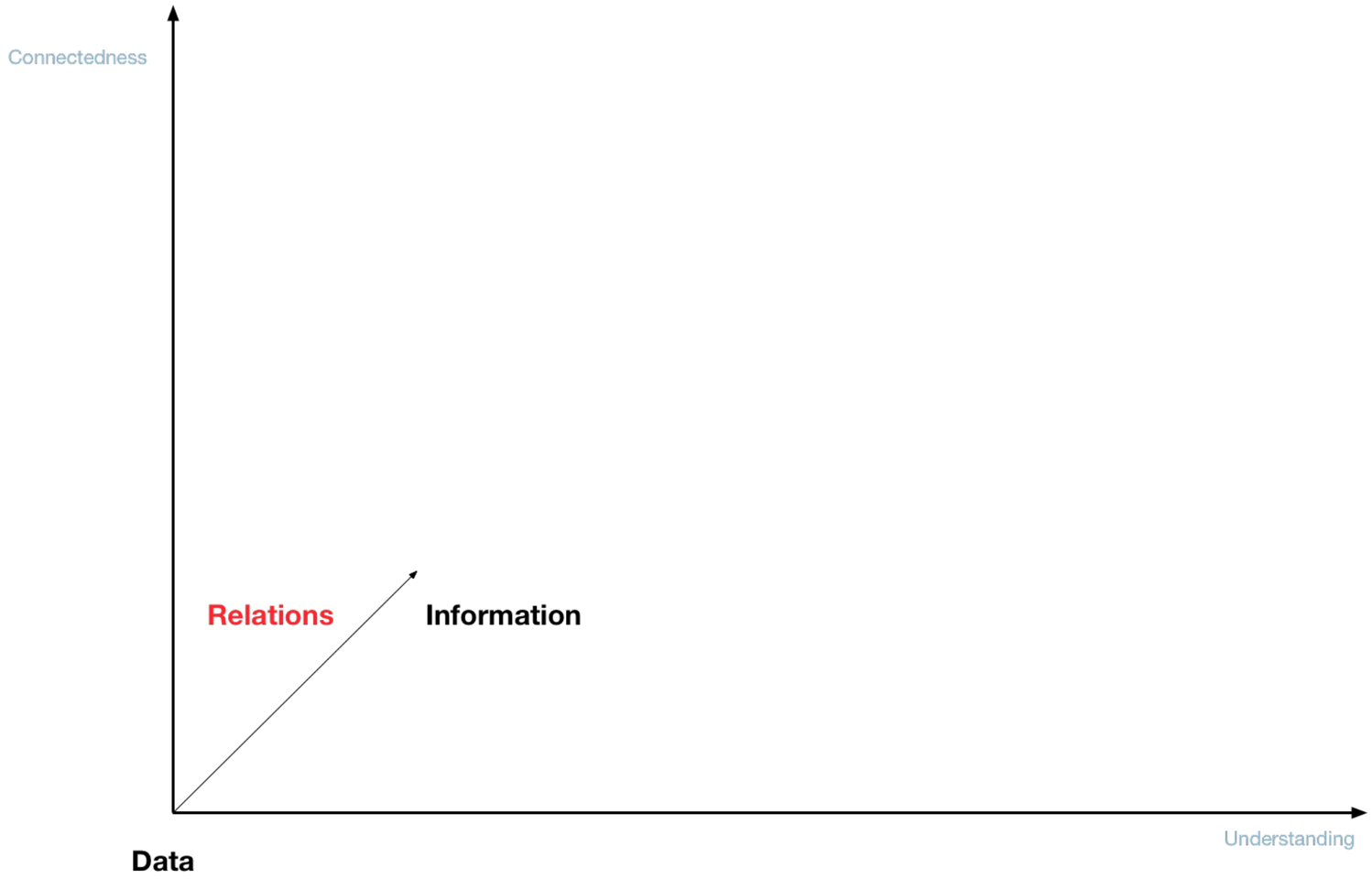

Data
Structured

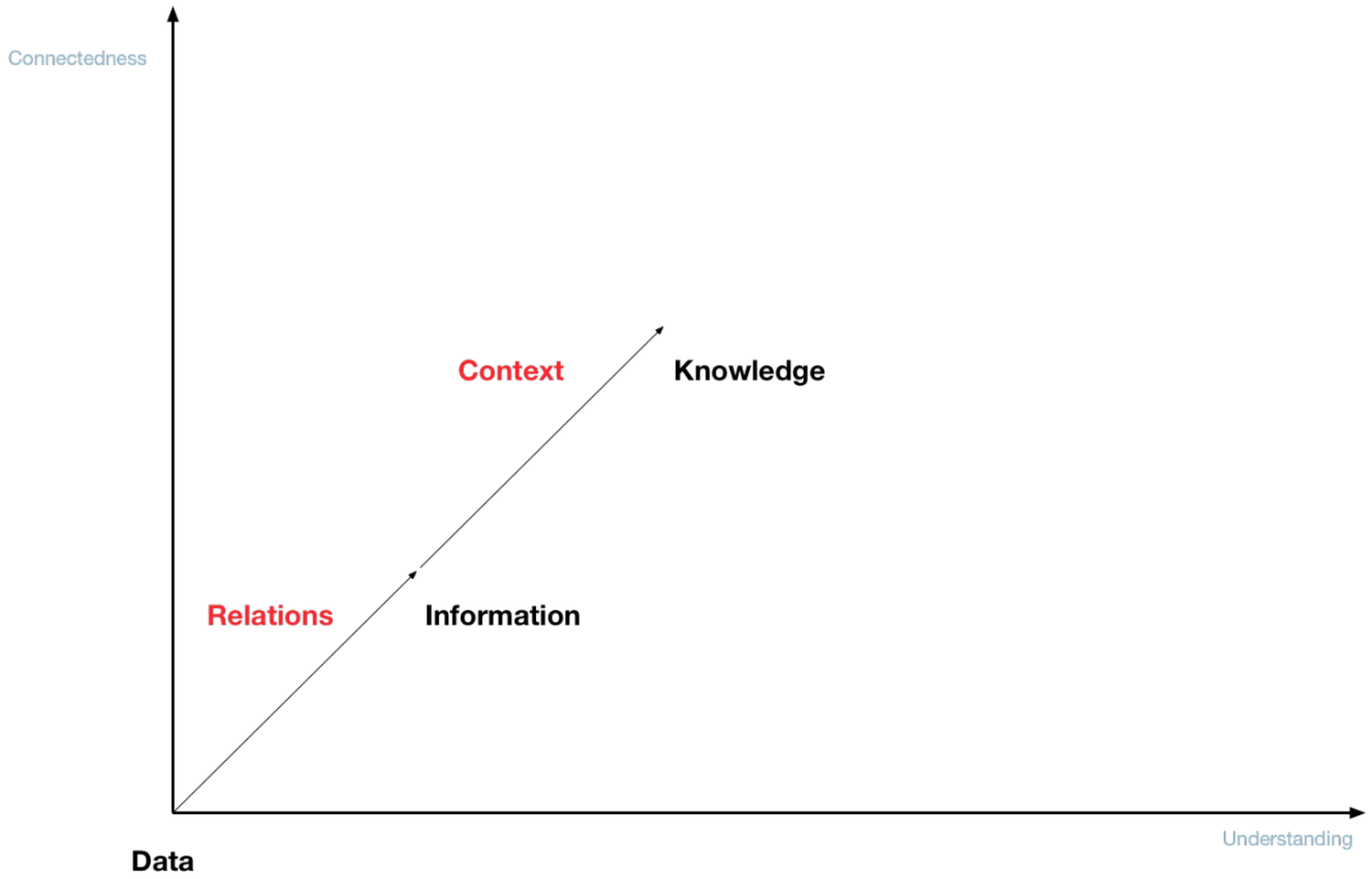

Data
Unstructured
Structured

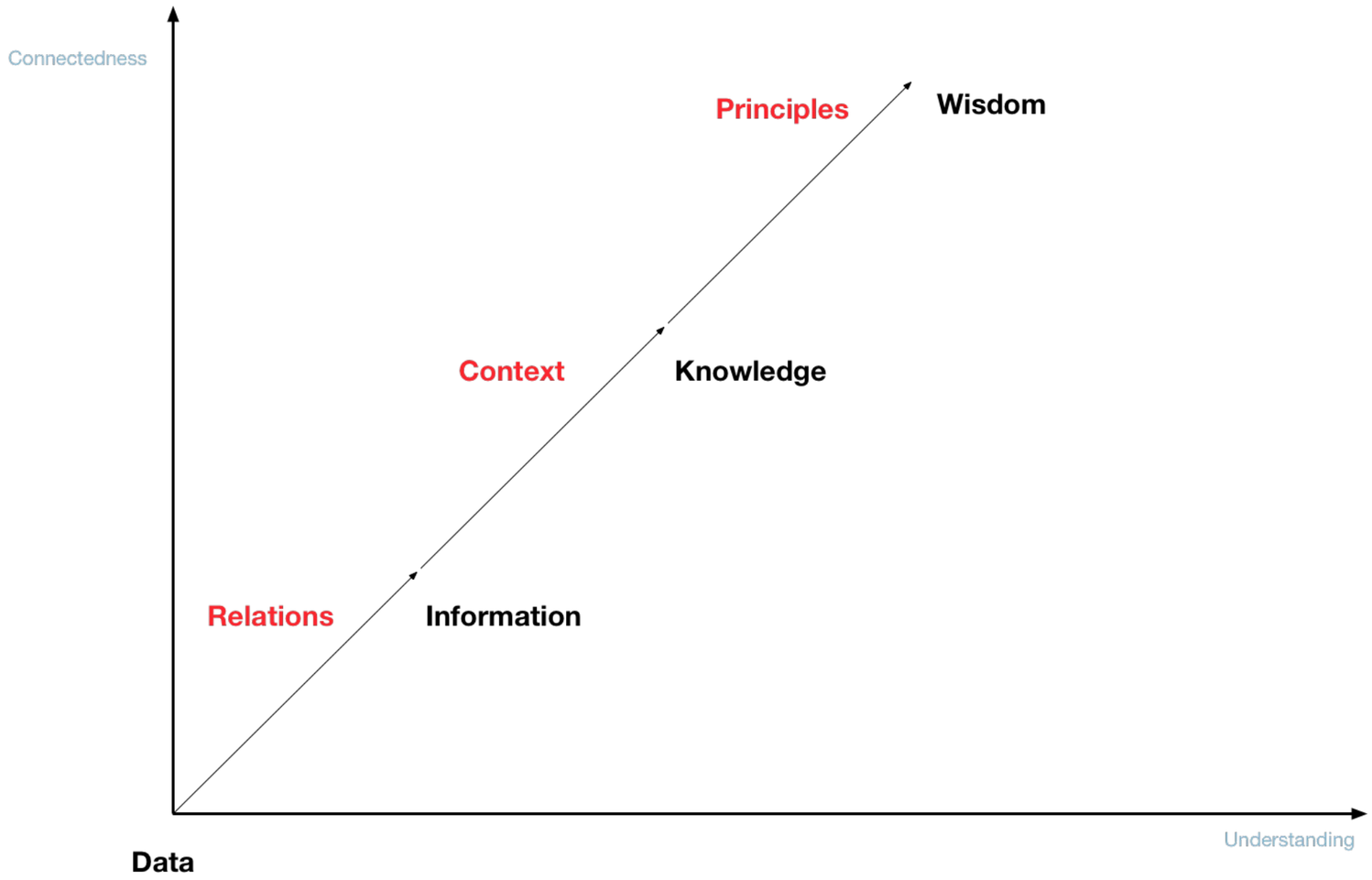

Data
Semi-structured
Unstructured
Structured


Connectedness
Understanding


Connectedness
Understanding
Data


Connectedness
Relations
Information
Data
Understanding


Connectedness
Context
Knowledge
Relations
Information
Data
Understanding


Connectedness
Principles
Wisdom
Context
Knowledge
Relations
Information
Data
Understanding

What is the Problem?

Wasting Time

https://tinyurl.com/ybwdq34u

Wasting Time

- Data professionals spend 60% of their time "getting to insight"

https://tinyurl.com/ybwdq34u

Wasting Time

- Data professionals spend 60% of their time "getting to insight"
  - 37% of this time is spent searching for data

https://tinyurl.com/ybwdq34u

Wasting Time

- Data professionals spend 60% of their time "getting to insight"
  - 37% of this time is spent searching for data
  - 36% of this time is spent preparing data

https://tinyurl.com/ybwdq34u

Wasting Time

- Data professionals spend 60% of their time "getting to insight"
  - 37% of this time is spent searching for data
  - 36% of this time is spent preparing data
  - Only 27% of that time is actual analysis

https://tinyurl.com/ybwdq34u

Wasting Time

- Data professionals spend 60% of their time "getting to insight"
  - 37% of this time is spent searching for data
  - 36% of this time is spent preparing data
  - Only 27% of that time is actual analysis
- 50% of the time is spent searching for or replicating existing data

https://tinyurl.com/ybwdq34u

Wasting Time

- Data professionals spend 60% of their time "getting to insight"
  - 37% of this time is spent searching for data
  - 36% of this time is spent preparing data
  - Only 27% of that time is actual analysis
- 50% of the time is spent searching for or replicating existing data
- 30-50% of organizations are not where they want to be

https://tinyurl.com/ybwdq34u

Wasting Time

- Data professionals spend 60% of their time "getting to insight"
  - 37% of this time is spent searching for data
  - 36% of this time is spent preparing data
  - Only 27% of that time is actual analysis
- 50% of the time is spent searching for or replicating existing data
- 30-50% of organizations are not where they want to be
- Costs to organizations annually

https://tinyurl.com/ybwdq34u

Wasting Time

- Data professionals spend 60% of their time "getting to insight"
  - 37% of this time is spent searching for data
  - 36% of this time is spent preparing data
  - Only 27% of that time is actual analysis
- 50% of the time is spent searching for or replicating existing data
- 30-50% of organizations are not where they want to be
- Costs to organizations annually
  - $1.7 million/100 employees in U.S.

https://tinyurl.com/ybwdq34u

Wasting Time

- Data professionals spend 60% of their time "getting to insight"
  - 37% of this time is spent searching for data
  - 36% of this time is spent preparing data
  - Only 27% of that time is actual analysis
- 50% of the time is spent searching for or replicating existing data
- 30-50% of organizations are not where they want to be
- Costs to organizations annually
  - $1.7 million/100 employees in U.S.
  - €1.1 million/100 employees in Europe

https://tinyurl.com/ybwdq34u

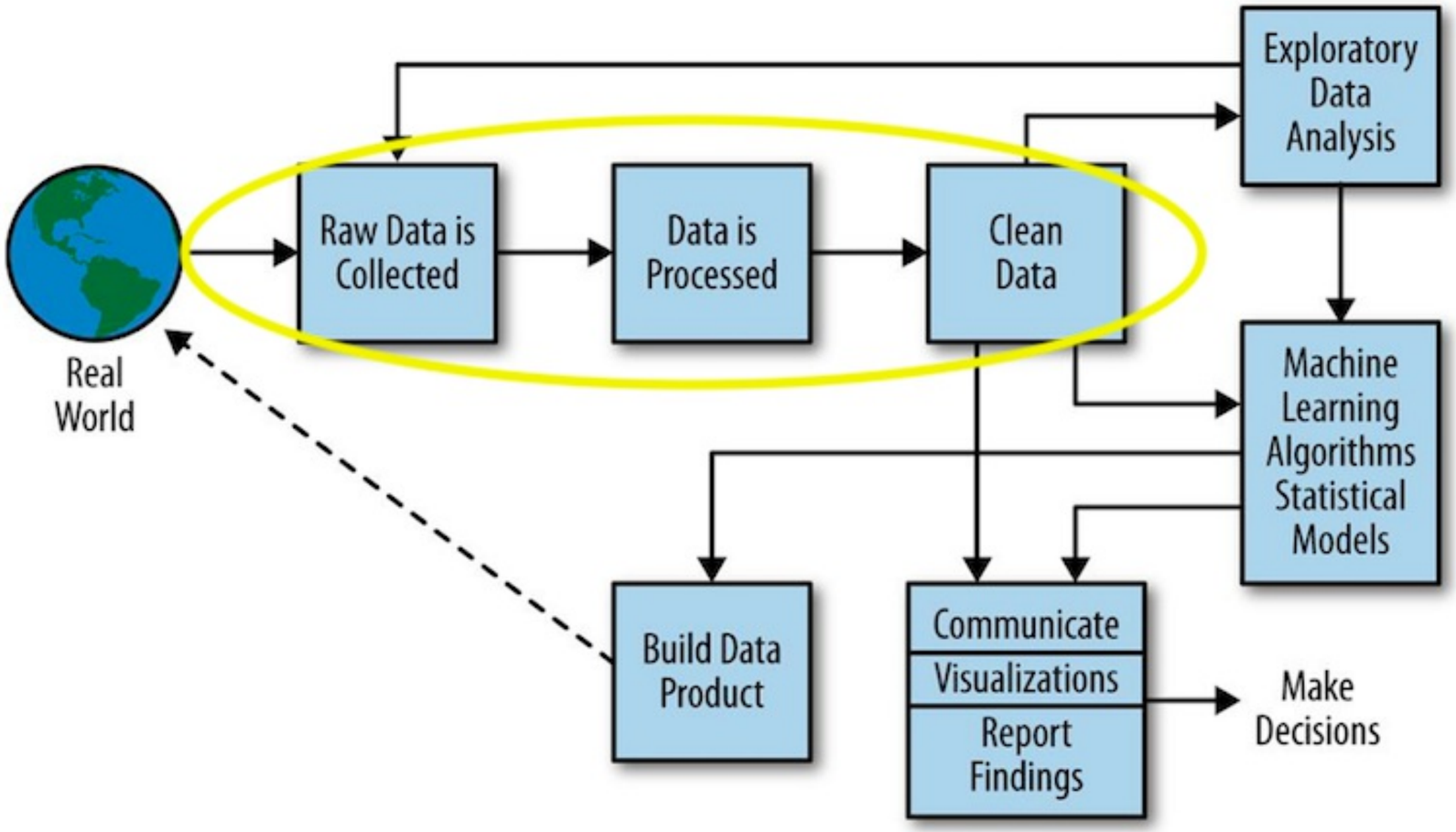
"It is evident that many professionals are not aware of what resources are available within data assets like data lakes, how to access the data, where it came from, or how to glean trusted insights..."

https://tinyurl.com/ybwdq34u

"Unless organizations make changes to their infrastructure now, and close the gaps on data discovery, integrity and cataloging, processes will only become more inefficient as data volume and variety continues to grow."

https://tinyurl.com/ybwdq34u

Data Science Pipeline



"Doing Data Science"


COMPLETE
-NESS
CONSIST-
ENCY
UNIQUE-
NESS
Data
Quality
Dimensions
ACCURACY
TIME-
LINESS
VALIDITY

Metadata

Metadata

- Data about data

Metadata

- Data about data
- Types of metadata

Metadata

- Data about data
- Types of metadata
  - Business metadata

Metadata

- Data about data
- Types of metadata
  - Business metadata
  - Technical metadata

Metadata

- Data about data
- Types of metadata
  - Business metadata
  - Technical metadata
  - Media metadata

Metadata

- Data about data
- Types of metadata
  - Business metadata
  - Technical metadata
  - Media metadata
  - Semantic metadata

Semi-Structured

```
<script type="application/ld+json">
  {"@context": "http://schema.org",
   "@type": "Organization",
   "url": "http://www.your-company-site.com",
   "contactPoint":
   [
      {
        "@type": "ContactPoint",
        "telephone": "+1-401-555-1212",
        "contactType": "customer service"
      }
    ]
  }
</script>
```

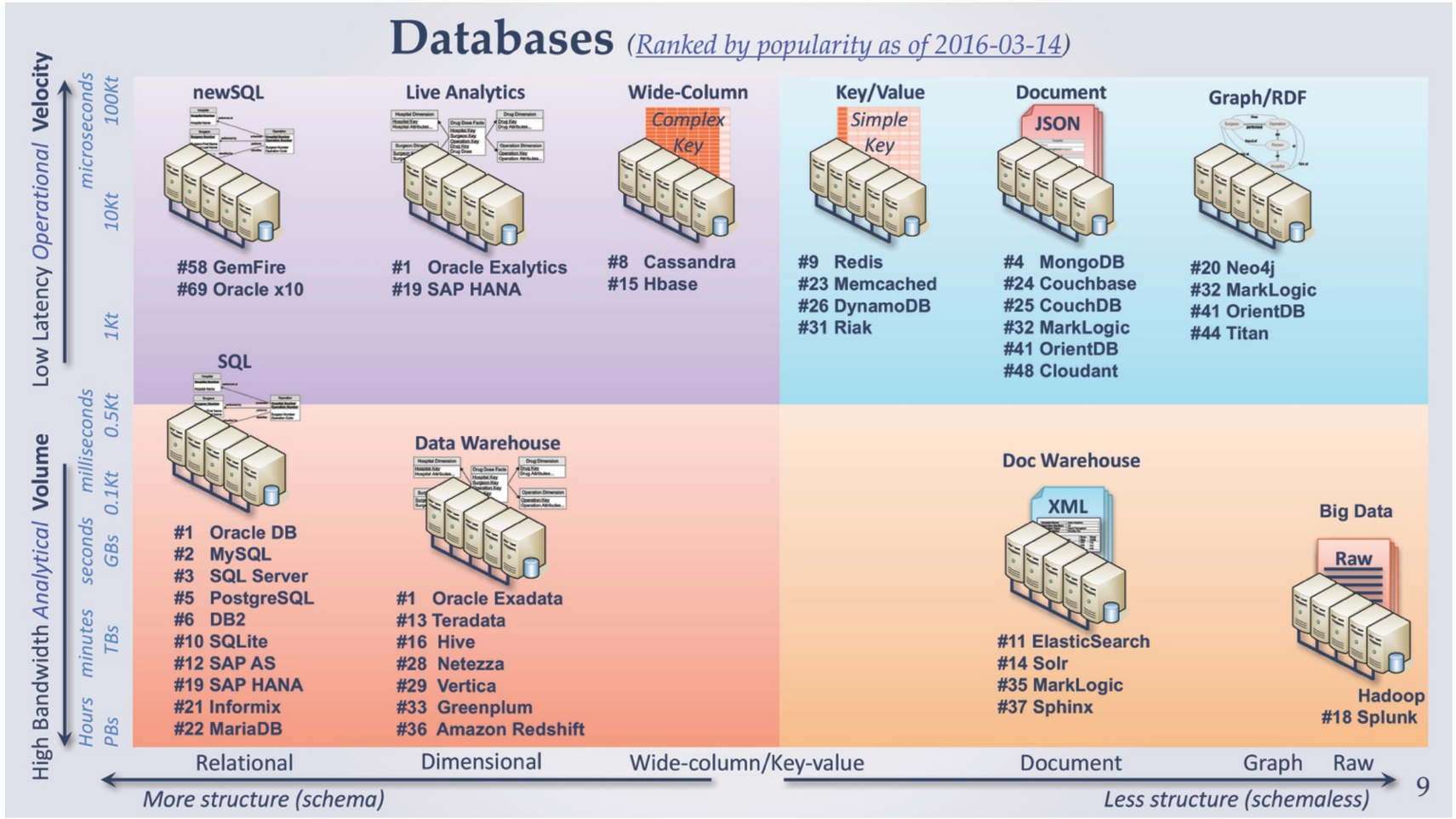
https://developers.google.com/gmail/markup/

https://developers.google.com/search/docs/dat
types/events

What did you do Saturday night?

Data Modeling Choices

- Relational
- Key-Value
- Column
- Document
- Graph


Databases (Ranked by popularity as of 2016-03-14)
Low Latency Operational Velocity
High Bandwidth Analytical Volume
microseconds 100Kt
10Kt
1Kt
milliseconds 0.5Kt
0.1Kt
seconds GBs
minutes TBs
Hours PBs
newSQL
#58 GemFire
#69 Oracle x10
Live Analytics
#1 Oracle Exalytics
#19 SAP HANA
Wide-Column
Complex Key
#8 Cassandra
#15 Hbase
Key/Value
Simple Key
#9 Redis
#23 Memcached
#26 DynamoDB
#31 Riak
Document
JSON
#4 MongoDB
#24 Couchbase
#25 CouchDB
#32 MarkLogic
#41 OrientDB
#48 Cloudant
Graph/RDF
#20 Neo4j
#32 MarkLogic
#41 OrientDB
#44 Titan
SQL
#1 Oracle DB
#2 MySQL
#3 SQL Server
#5 PostgreSQL
#6 DB2
#10 SQLite
#12 SAP AS
#19 SAP HANA
#21 Informix
#22 MariaDB
Data Warehouse
#1 Oracle Exadata
#13 Teradata
#16 Hive
#28 Netezza
#29 Vertica
#33 Greenplum
#36 Amazon Redshift
Doc Warehouse
XML
#11 ElasticSearch
#14 Solr
#35 MarkLogic
#37 Sphinx
Big Data
Raw
Hadoop
#18 Splunk
Relational
Dimensional
Wide-column/Key-value
Document
Graph
Raw
More structure (schema)
Less structure (schemaless)
9

RDBMS

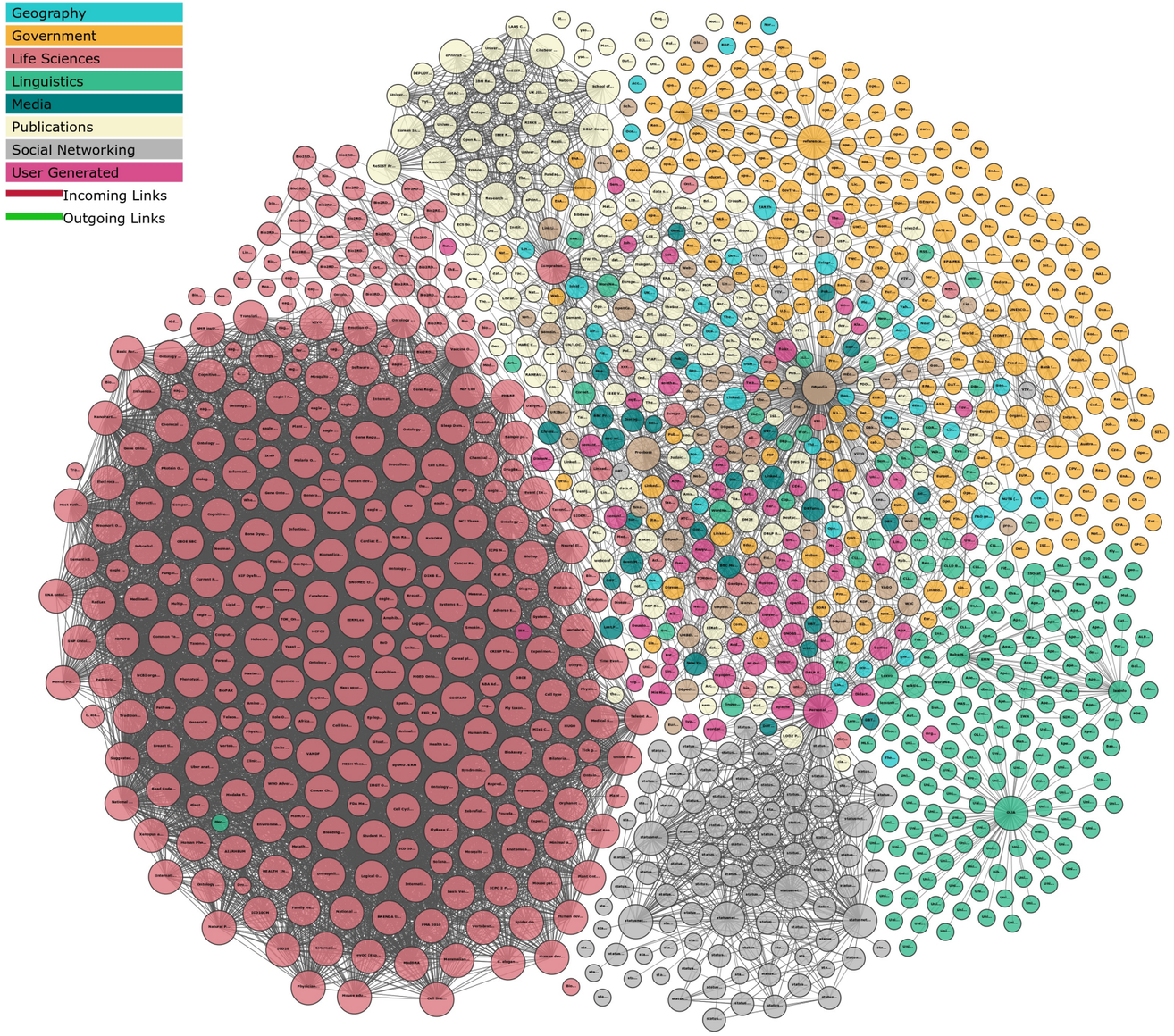
- Established technology, well-understood
- Works best for stable domains

Spreadsheets

- Conceptually aligned
- Easy tooling
- Familiarity
- Rapid experimentation
- Customization

NoSQL

- Volume
- Velocity
- Variety
- Veracity



CC BY-SA, http://lod-cloud.net

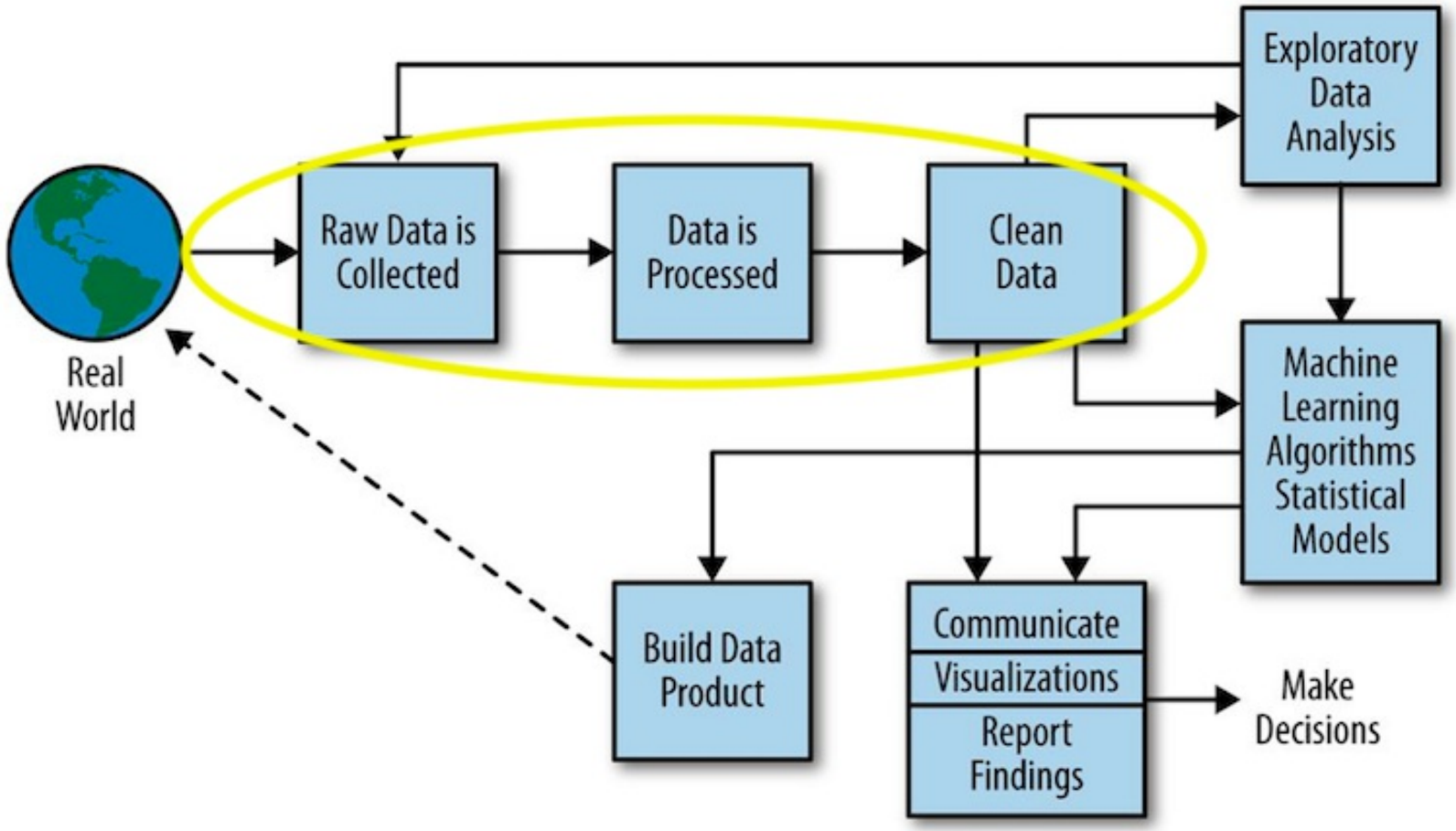
Exercise

Pod TTL Data

- Read CSV file data
- Read data from Postgres

Data-Driven

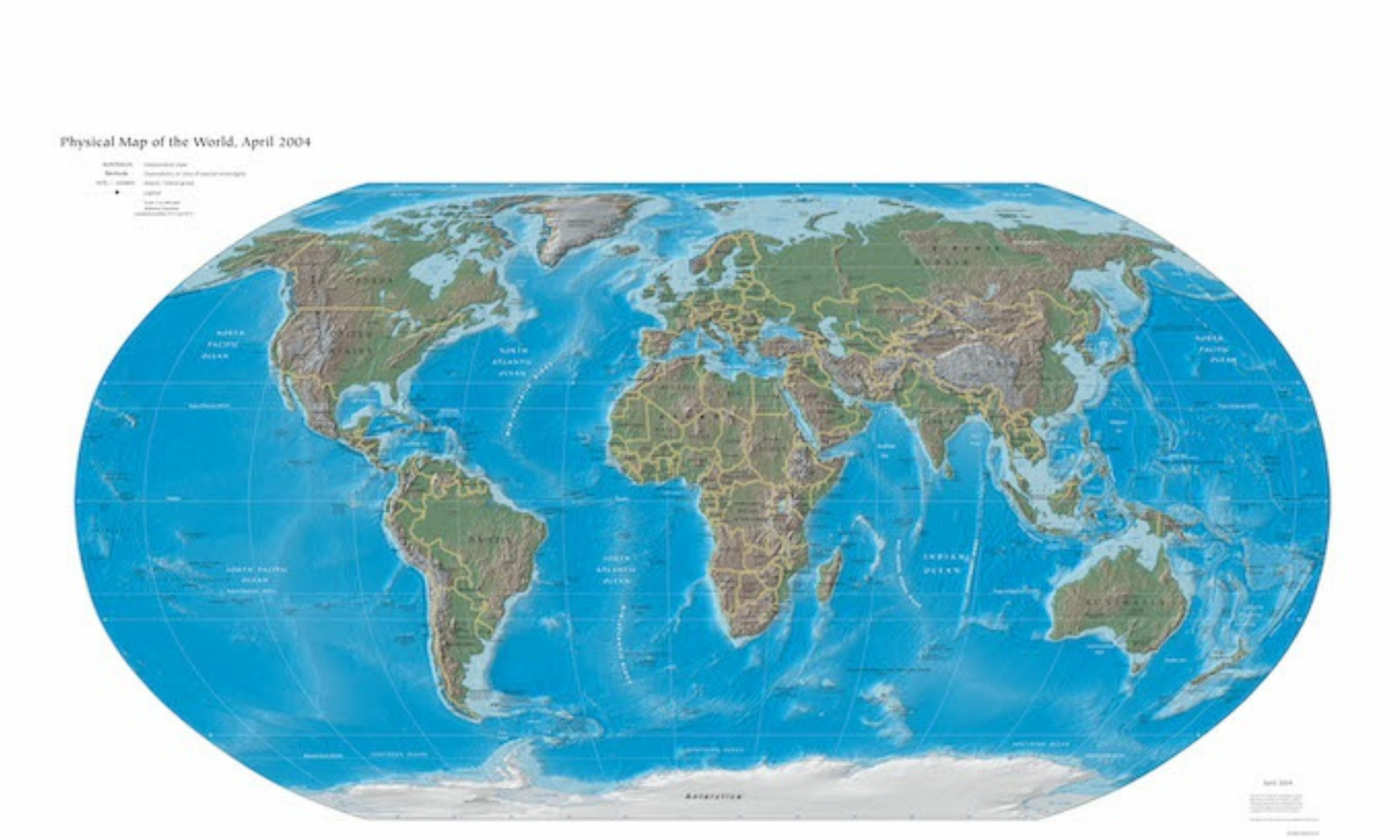
Data Science Pipeline



"Doing Data Science"

What was the degree with the highest average starting salary at the University of North Carolina in the 1980's?

Geography!



Let's Make Some Data

Let's Make Some Data

```
>>> ages = range(20,60)
```

Let's Make Some Data

```
>>> ages = range(20,60)
```

```
>>> random_ages = [random.choice(ages) for _ in range(100)]
```

Let's Make Some Data

```
>>> ages = range(20,60)
```

```
>>> random_ages = [random.choice(ages) for _ in range(100)]
```

```
>>> random_ages
[27, 21, 20, 22, 21, 27, 25, 52, 56, 43, 33, 28, 31, 30, 41, 29, 38, 52,
 48, 24, 50, 31, 29, 56, 45, 54, 37, 33, 22, 33, 23, 40, 37, 23, 28, 38,
 41, 34, 35, 40, 25, 57, 32, 48, 56, 54, 39, 25, 57, 24, 26, 52, 26, 35,
 57, 39, 26, 32, 50, 52, 54, 53, 30, 53, 56, 25, 43, 38, 53, 46, 51, 51,
 35, 38, 20, 52, 47, 53, 45, 27, 22, 26, 25, 47, 36, 52, 54, 58, 31, 42,
 46, 46, 43, 33, 35, 48, 49, 55, 36, 36]
```

Let's Make Some Data

```
>>> ages = range(20,60)
```

```
>>> random_ages = [random.choice(ages) for _ in range(100)]
```

```
>>> random_ages
[27, 21, 20, 22, 21, 27, 25, 52, 56, 43, 33, 28, 31, 30, 41, 29, 38, 52,
 48, 24, 50, 31, 29, 56, 45, 54, 37, 33, 22, 33, 23, 40, 37, 23, 28, 38,
 41, 34, 35, 40, 25, 57, 32, 48, 56, 54, 39, 25, 57, 24, 26, 52, 26, 35,
 57, 39, 26, 32, 50, 52, 54, 53, 30, 53, 56, 25, 43, 38, 53, 46, 51, 51,
 35, 38, 20, 52, 47, 53, 45, 27, 22, 26, 25, 47, 36, 52, 54, 58, 31, 42,
 46, 46, 43, 33, 35, 48, 49, 55, 36, 36]
```

```
>>> max(random_ages)
58
```

Let's Make Some Data

```
>>> ages = range(20,60)
```

```
>>> random_ages = [random.choice(ages) for _ in range(100)]
```

```
>>> random_ages
[27, 21, 20, 22, 21, 27, 25, 52, 56, 43, 33, 28, 31, 30, 41, 29, 38, 52,
 48, 24, 50, 31, 29, 56, 45, 54, 37, 33, 22, 33, 23, 40, 37, 23, 28, 38,
 41, 34, 35, 40, 25, 57, 32, 48, 56, 54, 39, 25, 57, 24, 26, 52, 26, 35,
 57, 39, 26, 32, 50, 52, 54, 53, 30, 53, 56, 25, 43, 38, 53, 46, 51, 51,
 35, 38, 20, 52, 47, 53, 45, 27, 22, 26, 25, 47, 36, 52, 54, 58, 31, 42,
 46, 46, 43, 33, 35, 48, 49, 55, 36, 36]
```

```
>>> max(random_ages)
58
```

```
>>> min(random_ages)
20
```

Range

Range

```
>>> def range(x):
...   return max(x) - min(x)
...
```

Range

```
>>> def range(x):
...   return max(x) - min(x)
...
```

```
>>> range(random_ages)
38
```

Range

```
>>> def range(x):
...   return max(x) - min(x)
...
```

```
>>> range(random_ages)
38
```

```
>>> nums = [10, 10, 100, 100]
>>> range(nums)
90
```

Range

```
>>> def range(x):
...   return max(x) - min(x)
...
```

```
>>> range(random_ages)
38
```

```
>>> nums = [10, 10, 100, 100]
>>> range(nums)
90
```

```
>>> nums = [10, 50, 50, 50, 50, 100]
>>> range(nums)
90
```

Range

```
>>> def range(x):
...   return max(x) - min(x)
...
```

```
>>> range(random_ages)
38
```

```
>>> nums = [10, 10, 100, 100]
>>> range(nums)
90
```

```
>>> nums = [10, 50, 50, 50, 50, 100]
>>> range(nums)
90
```

```
>>> numpy.ptp(random_ages)
38
```

Mean

Mean

```
>>> import numpy
>>> numpy.mean(random_ages)
38.990000000000002
```

Mean

```
>>> import numpy
>>> numpy.mean(random_ages)
38.990000000000002
```

```
>>> def mean(x):
...   return sum(x) / len(x)
...
```

Mean

```
>>> import numpy
>>> numpy.mean(random_ages)
38.990000000000002
```

```
>>> def mean(x):
...   return sum(x) / len(x)
...
```

```
>>> mean(random_ages)
38.99
```

Median

Median

```
>>> numpy.median(random_ages)
38.0
```

Median

```
>>> numpy.median(random_ages)
38.0
```

```
def median(x) :
  n = len(x)
  sorted_x = sorted(x)
  mid = n // 2

  if n % 2 == 0:
      return (sorted_x[mid - 1] + sorted_x[mid]) / 2
  else:
      return (sorted_x[mid])
```

Median

```
>>> numpy.median(random_ages)
38.0
```

```
def median(x) :
  n = len(x)
  sorted_x = sorted(x)
  mid = n // 2

  if n % 2 == 0:
      return (sorted_x[mid - 1] + sorted_x[mid]) / 2
  else:
      return (sorted_x[mid])
```

```
>>> median(random_ages)
38
```

Percentile

```
>>> numpy.percentile(random_ages, 33)
32.0
```

Percentile

```
>>> numpy.percentile(random_ages, 33)
32.0
```

```
>>> numpy.percentile(random_ages, 80)
52.0
```

Percentile

```
>>> numpy.percentile(random_ages, 33)
32.0
```

```
>>> numpy.percentile(random_ages, 80)
52.0
```

```
>>> numpy.percentile(random_ages, 50)
38.0
```

Interquartile Range (IQR)

Interquartile Range (IQR)

$$ IQR = Q_3 - Q_1 $$

Interquartile Range (IQR)

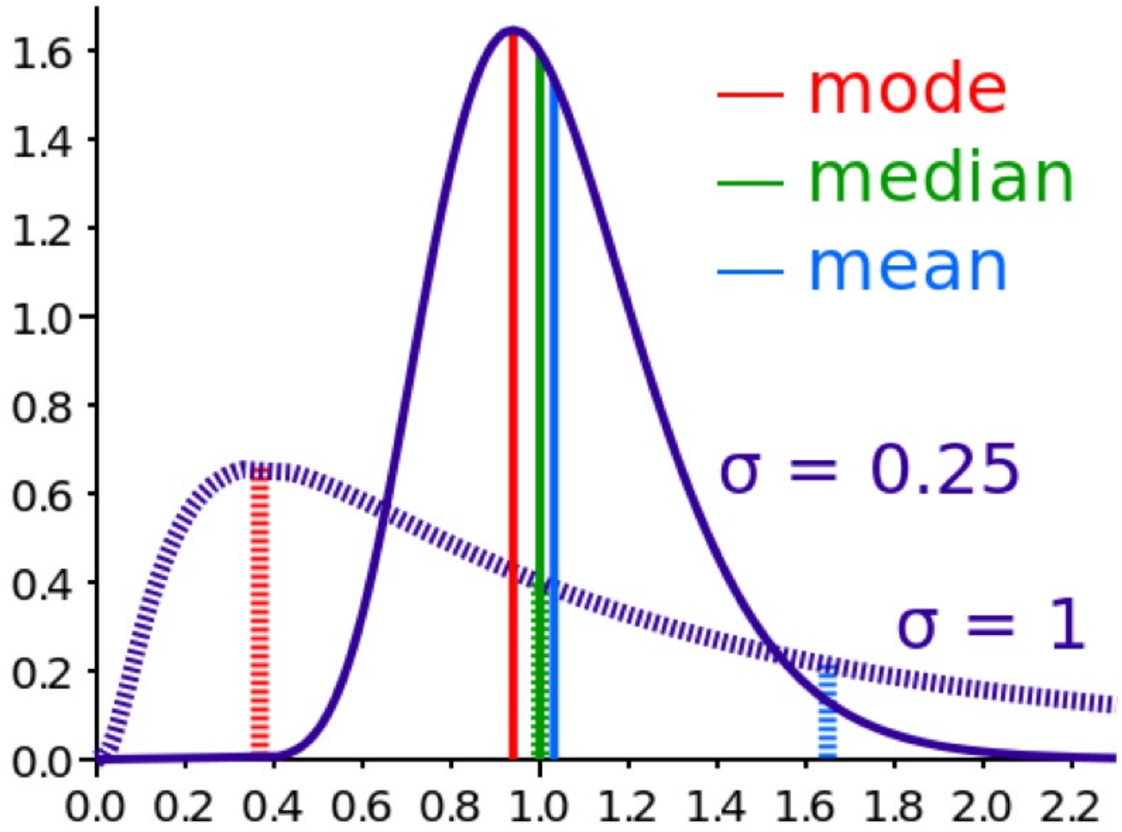
$$ IQR = Q_3 - Q_1 $$

```
>>> from scipy import stats
>>> stats.iqr(random_ages)
21.5
```

Mode

Mode

```
>>> from scipy import stats
>>> stats.mode(random_ages)
ModeResult(mode=array([52]), count=array([6]))
```


mode
median
mean
σ = 0.25
σ = 1
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0.0 0.2 0.4 0.6 0.8 1.0 1.2 1.4 1.6 1.8 2.0 2.2

Variance

Variance

```
$$Var(X) = \frac{1}{n} \sum_{i=1}^n ( a_i
- \bar x)^2$$
```

Variance

$$Var(X) = \frac{1}{n} \sum_{i=1}^n (a_i - \bar x)^2$$

```
>>> numpy.var(random_ages)
131.82989999999998
```

Standard Deviation

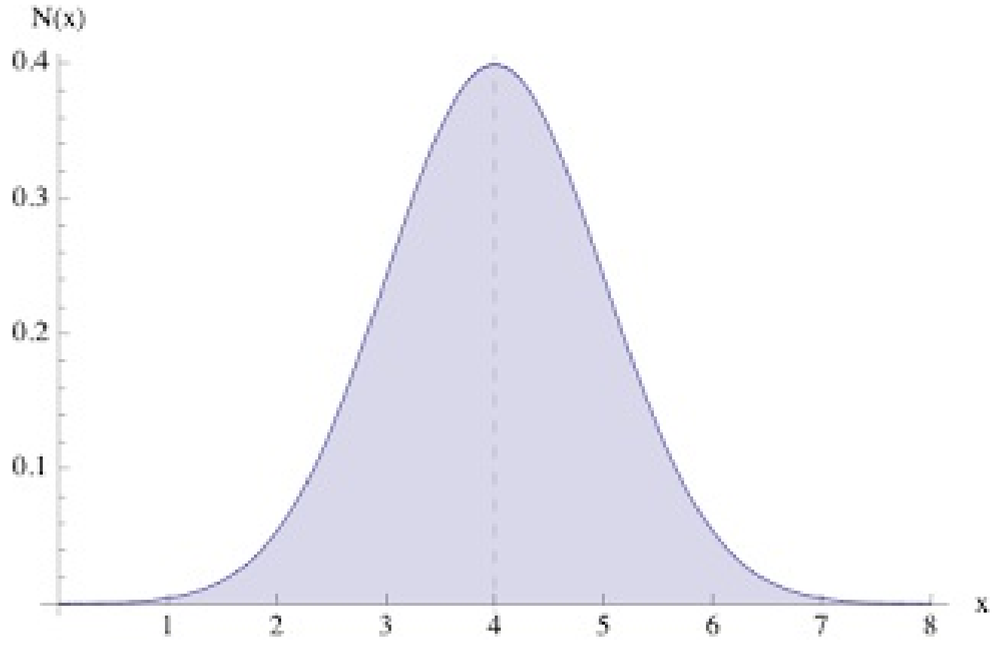
Standard Deviation

$$\sqrt {Var(X)}$$

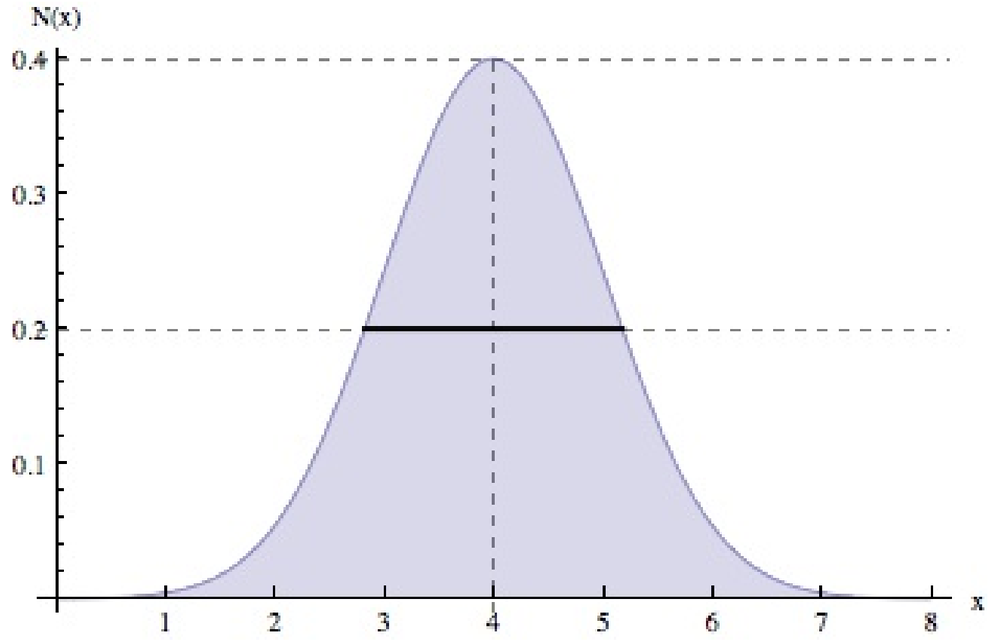
Standard Deviation

$$\sqrt {Var(X)}$$

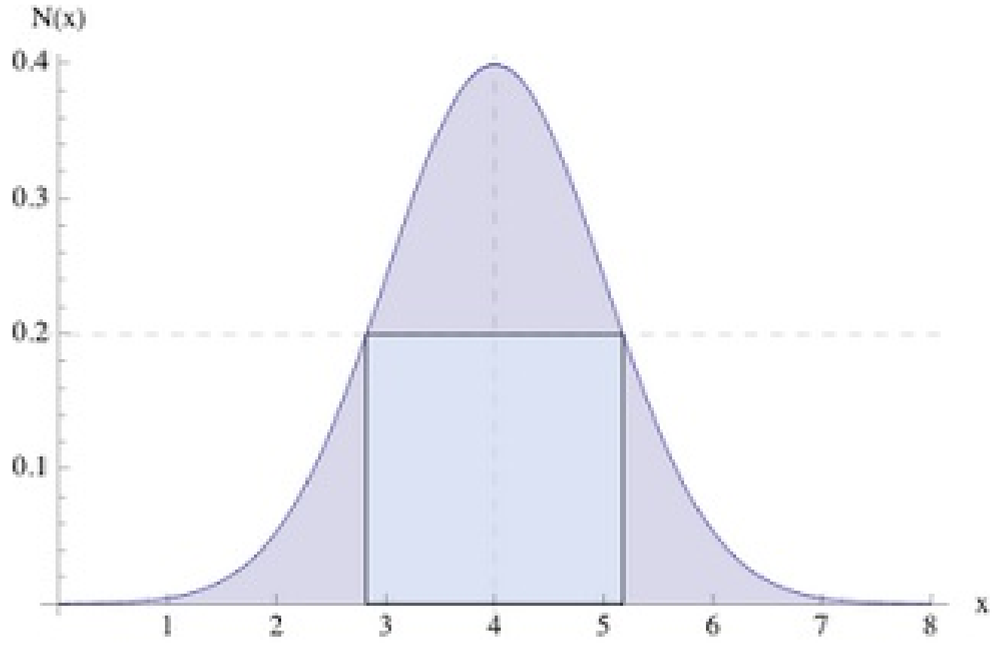
```
>>> numpy.std(random_ages)
11.481720254386969
```



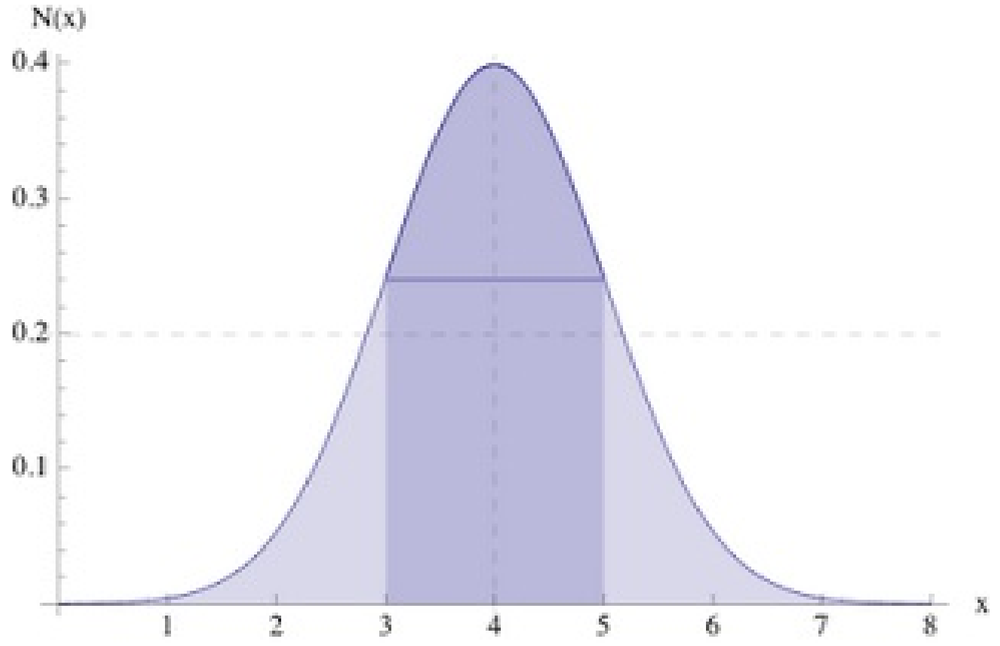
http://perfdynamics.blogspot.com/2013/01/visualizing-variance.html



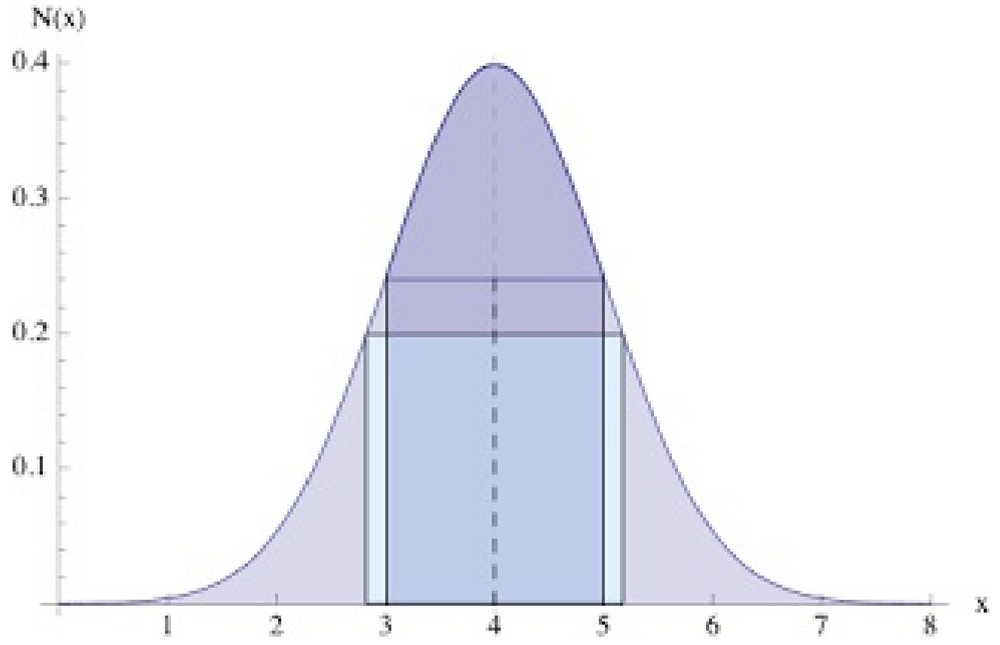
http://perfdynamics.blogspot.com/2013/01/visualizing-variance.html



http://perfdynamics.blogspot.com/2013/01/visualizing-variance.html

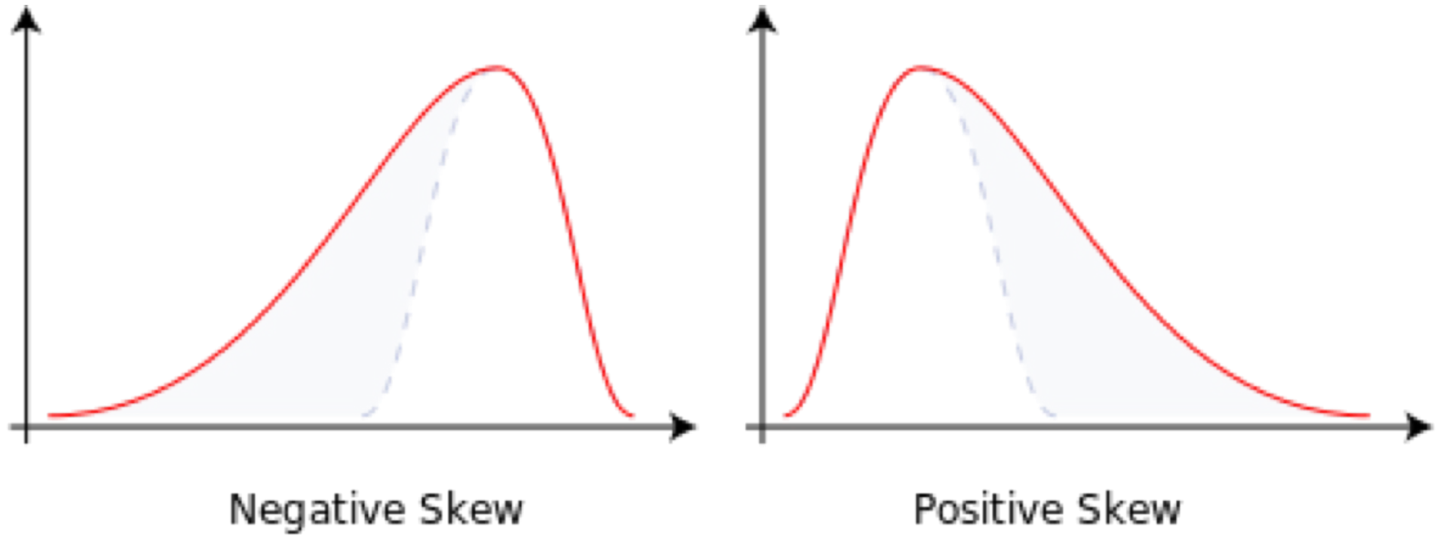


http://perfdynamics.blogspot.com/2013/01/visualizing-variance.html

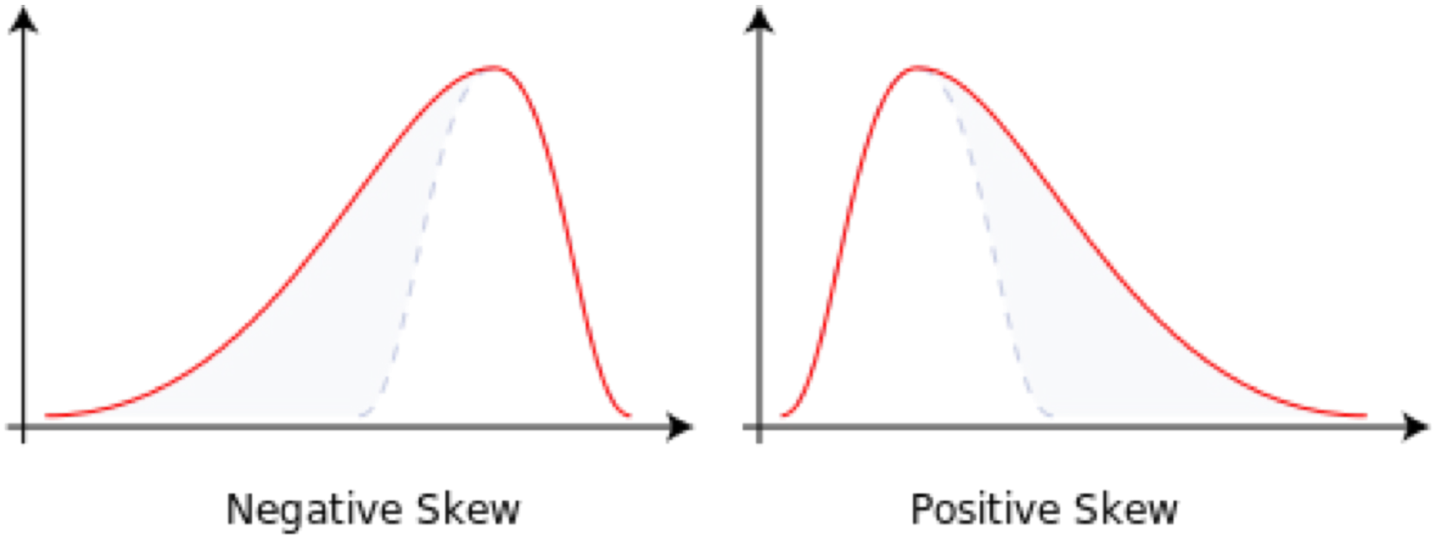


http://perfdynamics.blogspot.com/2013/01/visualizing-variance.html

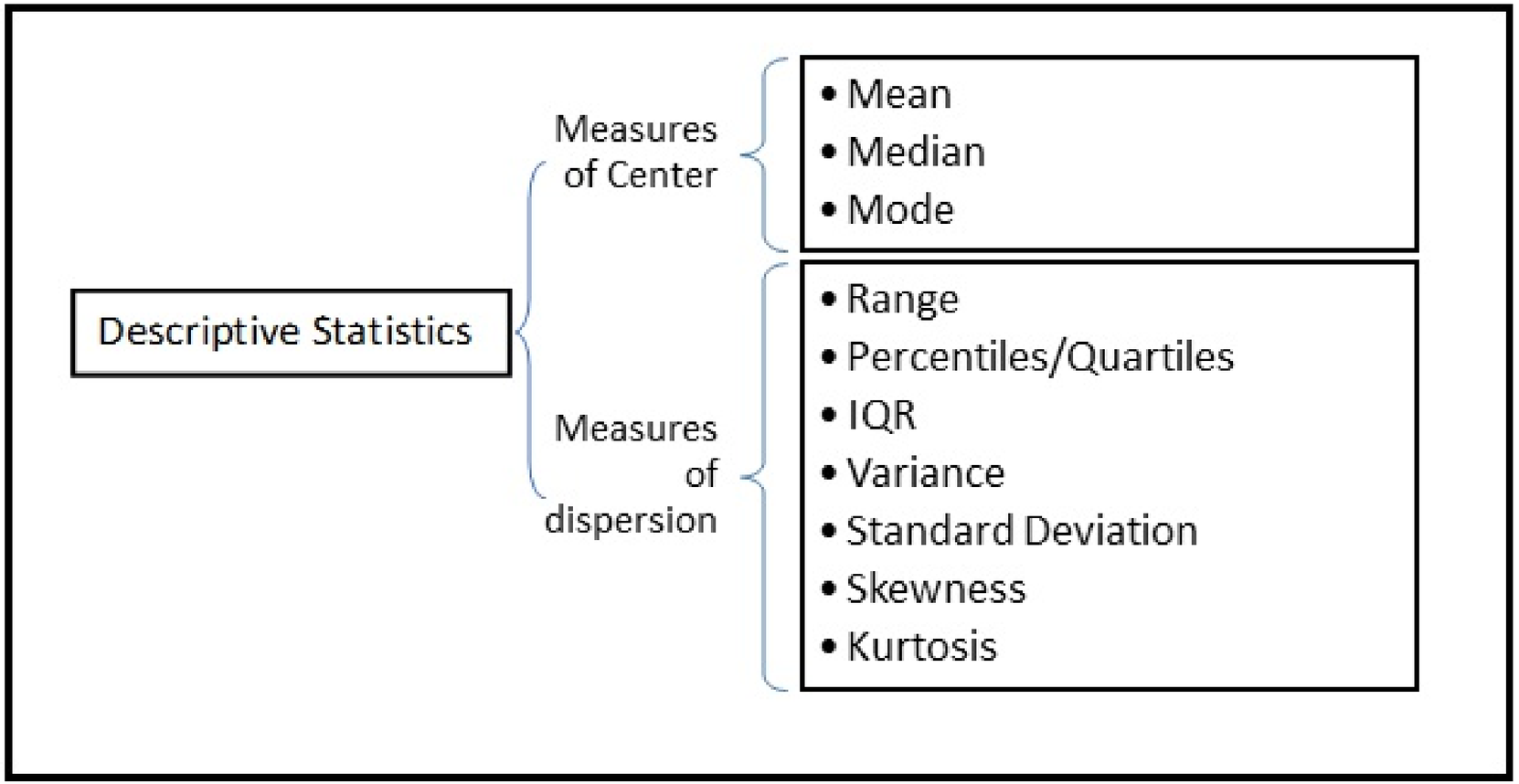
Skewness



Skewness




```
>>> stats.skew(random_ages)
0.037925216234465986
```



http://blog.hackerearth.com/descriptive-statistics-with-Python-NumPy

Covariance

Covariance

- Measure of joint variability

Covariance

- Measure of joint variability
- Zero covariance implies?

Covariance

- Measure of joint variability
- Zero covariance implies?
- Non-zero covariance implies?

"Correlation is not causation."

Anyone who has ever taken statistics

"Empirically observed covariation is a necessary but not sufficient condition for causality."

Edward Tufte

"Correlation is not causation but it sure is a hint."

Edward Tufte

Correlation/causation

Correlation/causation

- Could be unrelated

Correlation/causation

- Could be unrelated
- Reverse causation

Correlation/causation

- Could be unrelated
- Reverse causation
- Bi-directional relationships

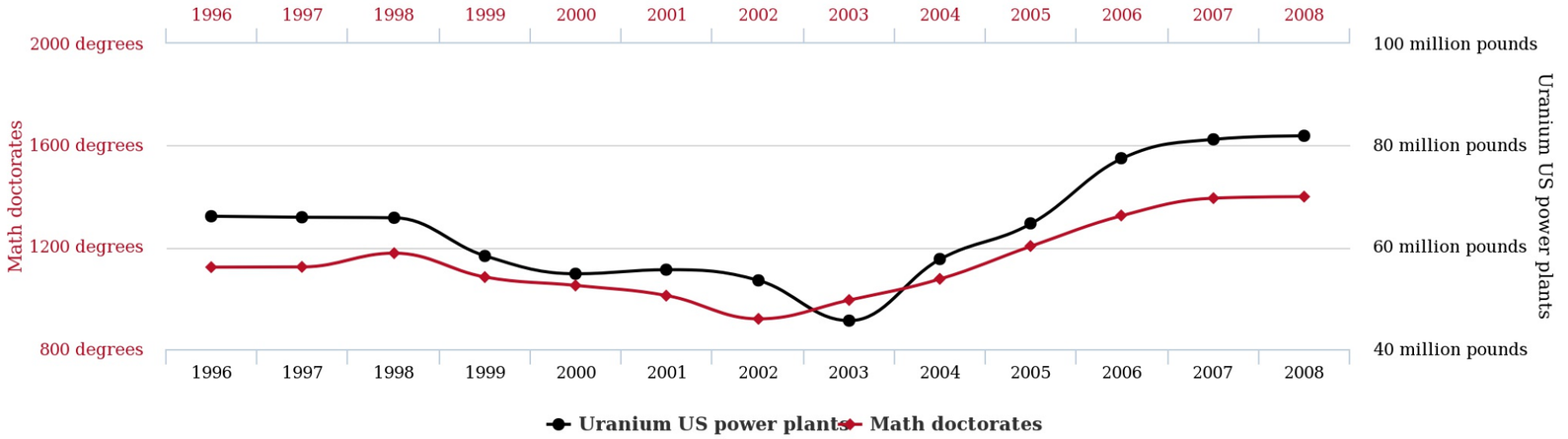
Correlation/causation

- Could be unrelated
- Reverse causation
- Bi-directional relationships
- Missed variable that explains it

Math doctorates awarded

correlates with

Uranium stored at US nuclear power plants

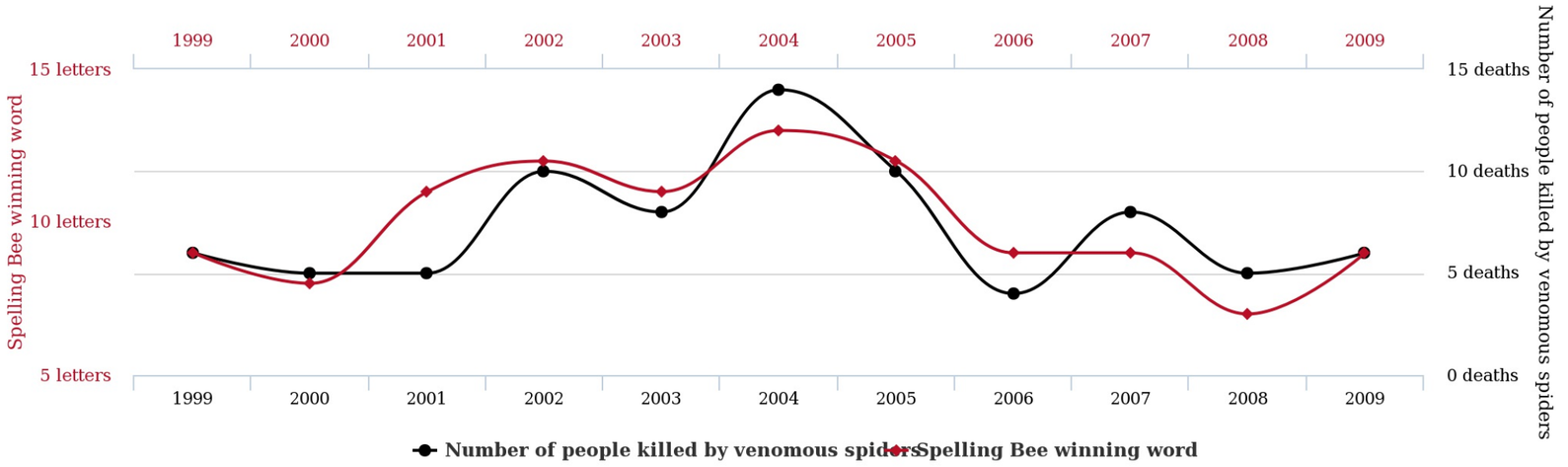


tylervigen.com

Letters in Winning Word of Scripps National Spelling Bee

correlates with

Number of people killed by venomous spiders

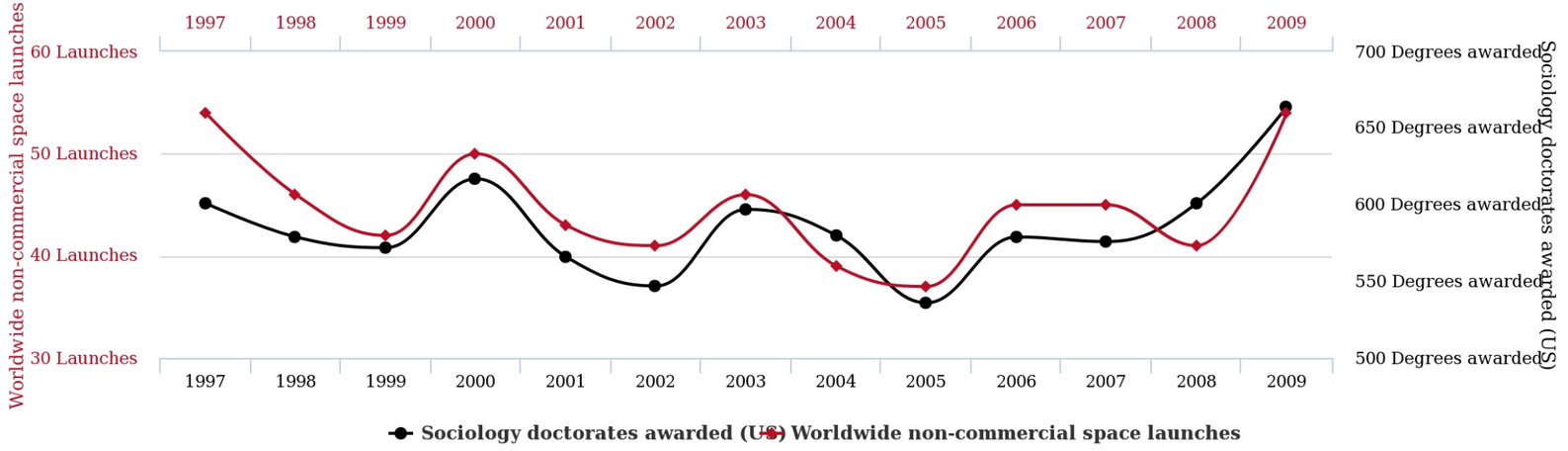


tylervigen.com

Worldwide non-commercial space launches

correlates with

Sociology doctorates awarded (US)

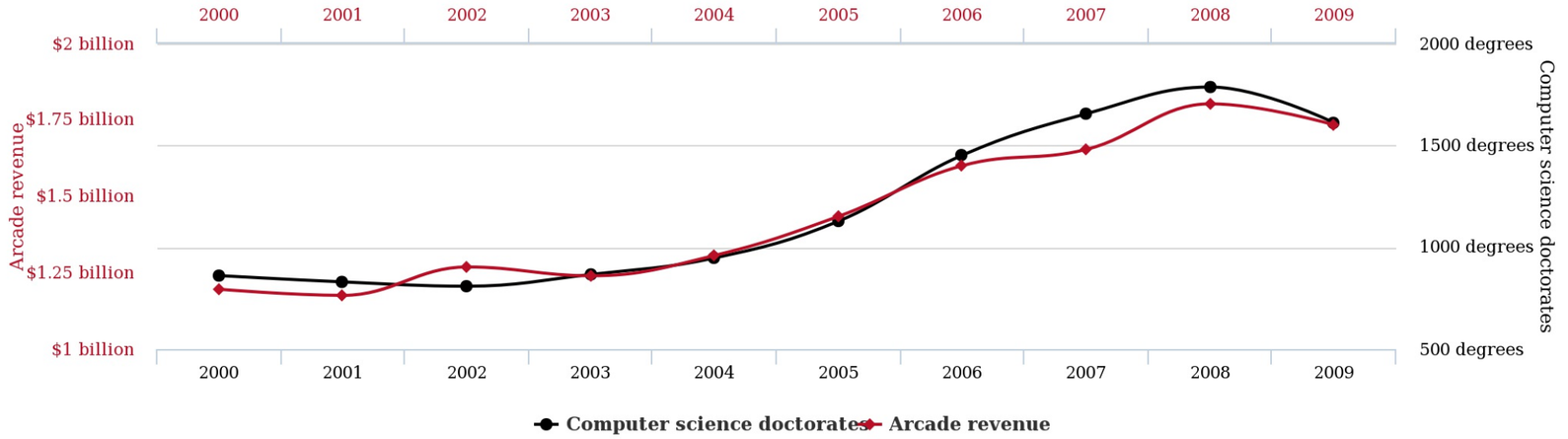


tylervigen.com

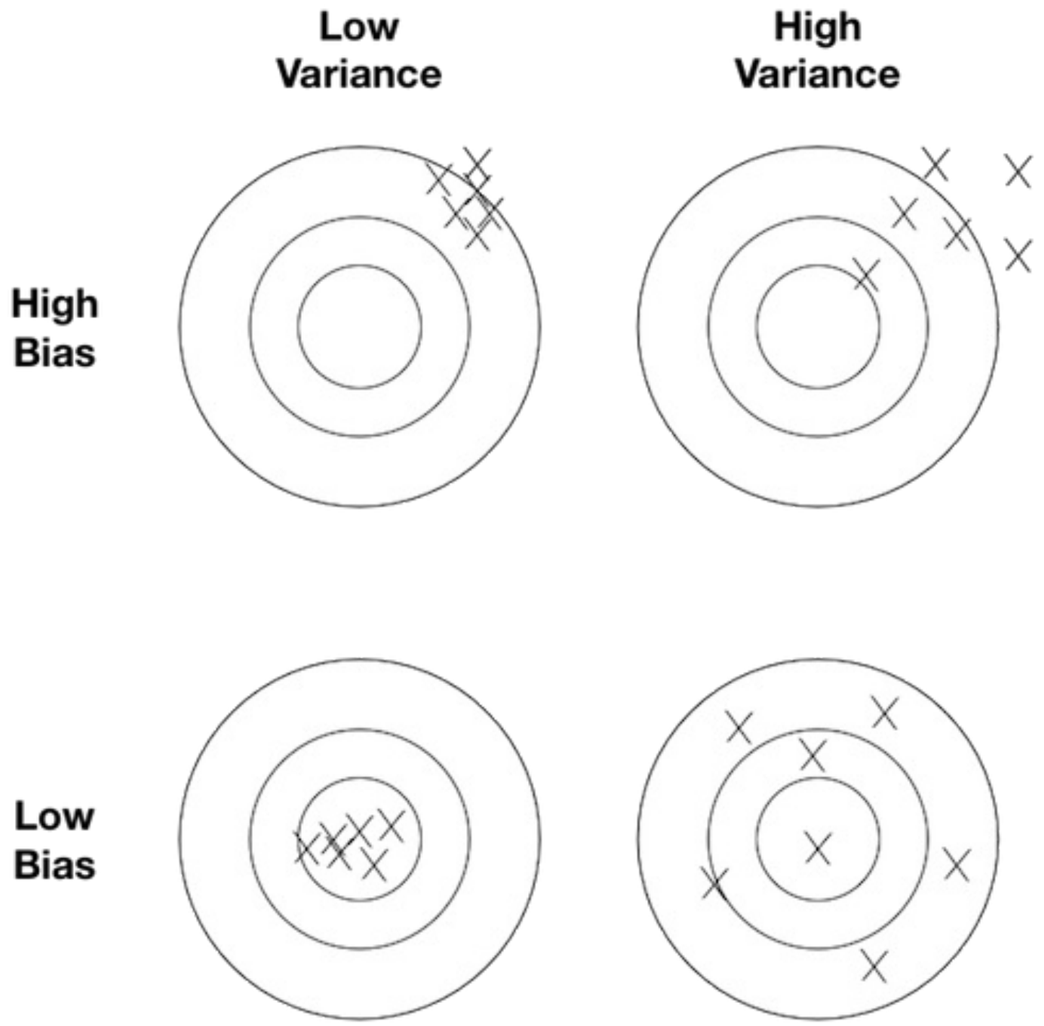
Total revenue generated by arcades

correlates with

Computer science doctorates awarded in the US



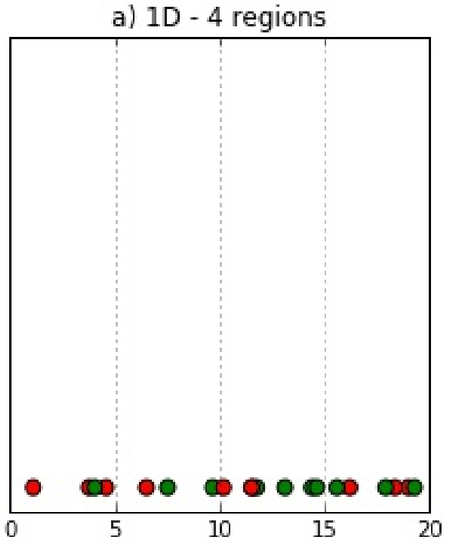
tylervigen.com


Low
Variance
High
Variance
High
Bias
Low
Bias

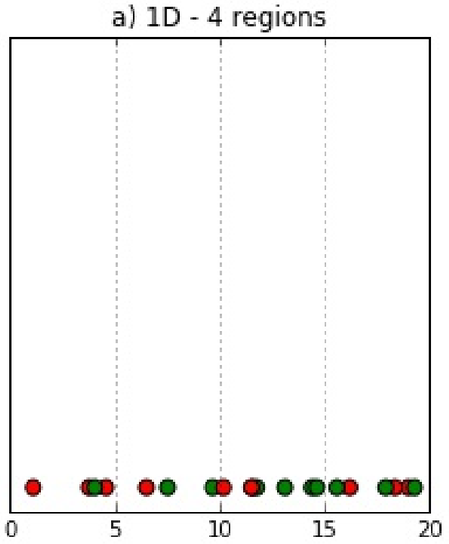
Hypothesis Testing

Precision, Recall and Accuracy

Curse of Dimensionality

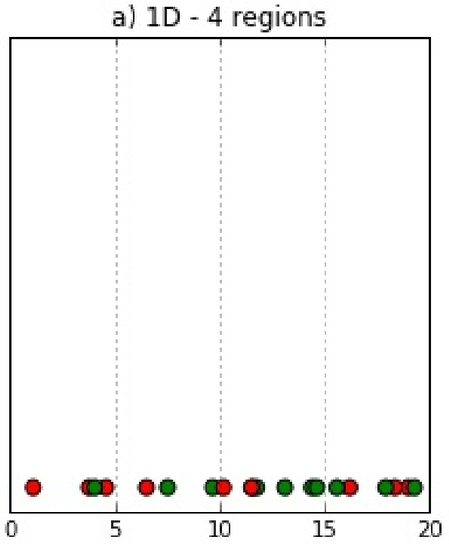


http://cleverowl.uk/2016/02/06/curse-of-dimensionality-explained/



$$ X^T = \{x_1,...,x_N\}, N = 204 $$

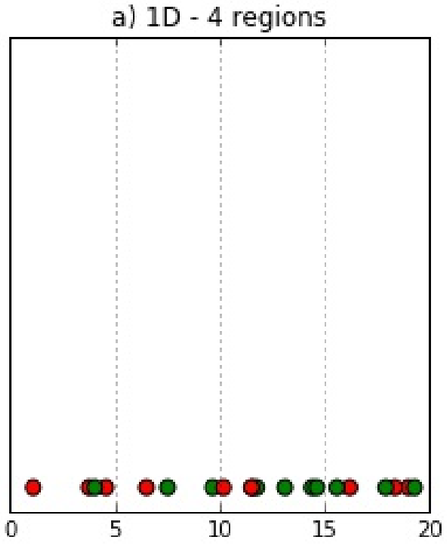
http://cleverowl.uk/2016/02/06/curse-of-dimensionality-explained/



$$ X^T = \{x_1,...,x_N\}, N = 204 $$

$$ T^T = \{t_1,...t_N\}, t \in \{g,r\}$$

http://cleverowl.uk/2016/02/06/curse-of-dimensionality-explained/

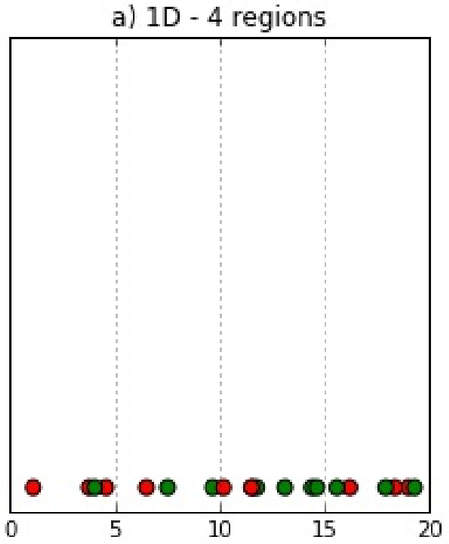


$$ X^T = \{x_1,...,x_N\}, N = 204 $$

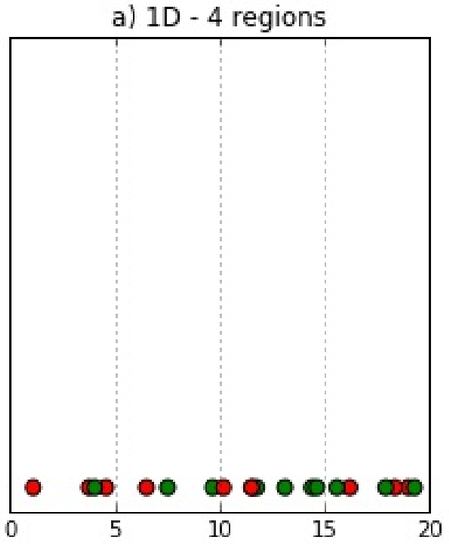
$$ T^T = \{t_1,...t_N\}, t \in \{g,r\}$$

$$ p(t_n = g|x_n)$$

http://cleverowl.uk/2016/02/06/curse-of-dimensionality-explained/

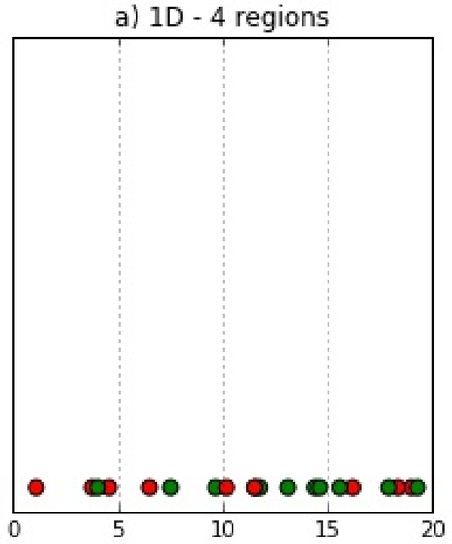


http://cleverowl.uk/2016/02/06/curse-of-dimensionality-explained/



$$ R_1, R_2, R_3, R_4 $$

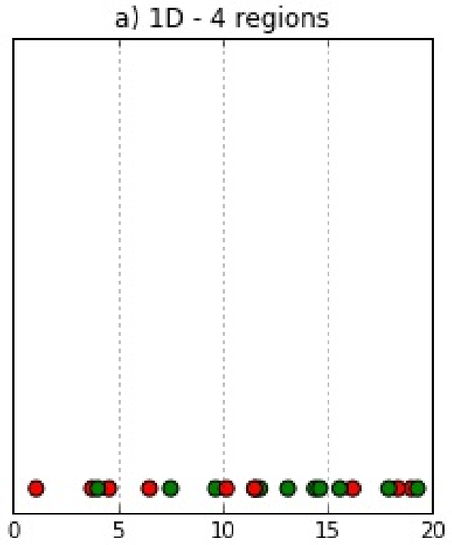
http://cleverowl.uk/2016/02/06/curse-of-dimensionality-explained/



$$ R_1, R_2, R_3, R_4 $$

$$ p(t_n = g | R_1) = \frac{1}{4} = 0.25$$

http://cleverowl.uk/2016/02/06/curse-of-dimensionality-explained/

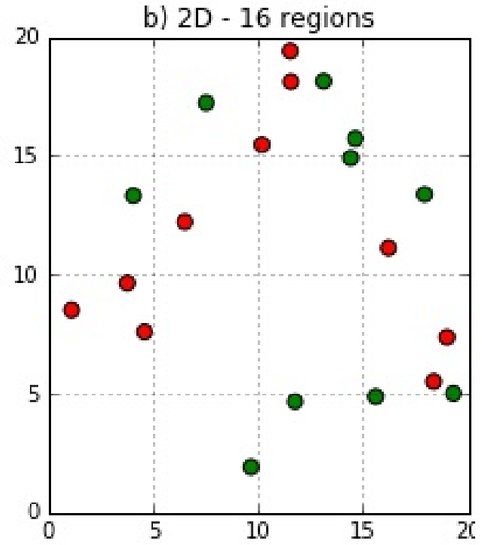


$$ R_1, R_2, R_3, R_4 $$

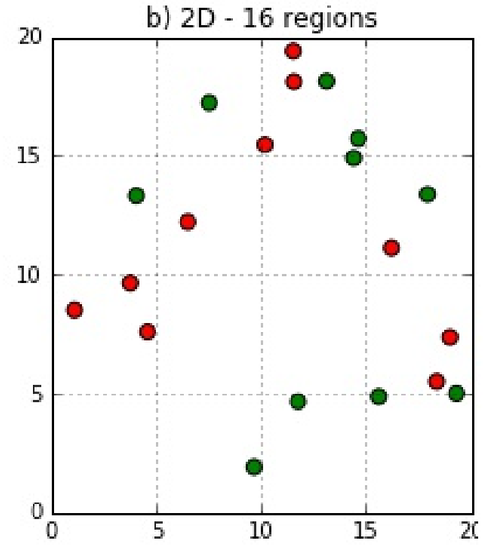
$$ p(t_n = g | R_1) = \frac{1}{4} = 0.25$$

$$ \frac{20}{4} = 5 \text{ obs per region}$$

http://cleverowl.uk/2016/02/06/curse-of-dimensionality-explained/

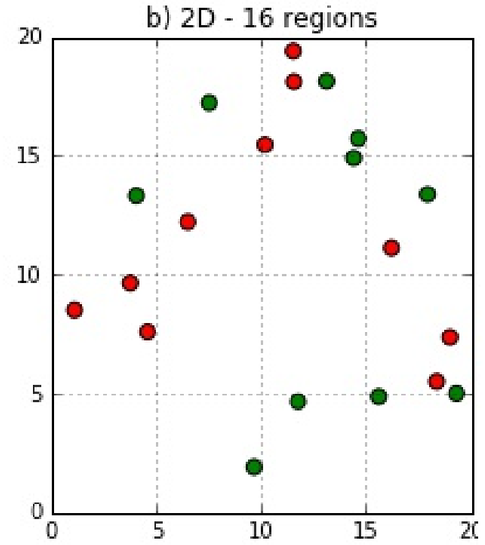


http://cleverowl.uk/2016/02/06/curse-of-dimensionality-explained/



$$ X^T = \{(x_{11},x_{12}),...,(x_{N1}, x_{N2})\} $$

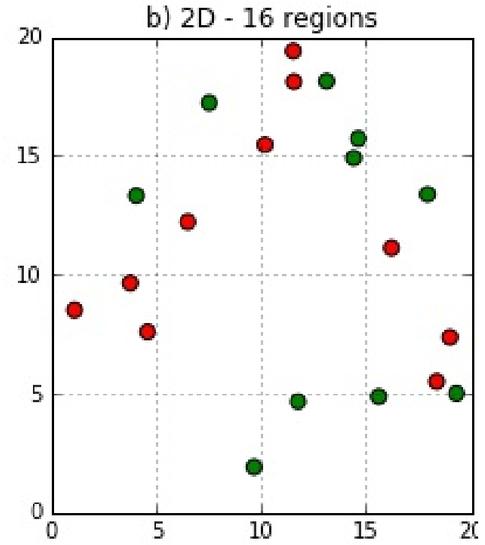
http://cleverowl.uk/2016/02/06/curse-of-dimensionality-explained/



$$ X^T = \{(x_{11},x_{12}),...,(x_{N1}, x_{N2})\} $$

$$ \frac{20}{16} = 1.25 \text{ obs per region}$$

http://cleverowl.uk/2016/02/06/curse-of-dimensionality-explained/

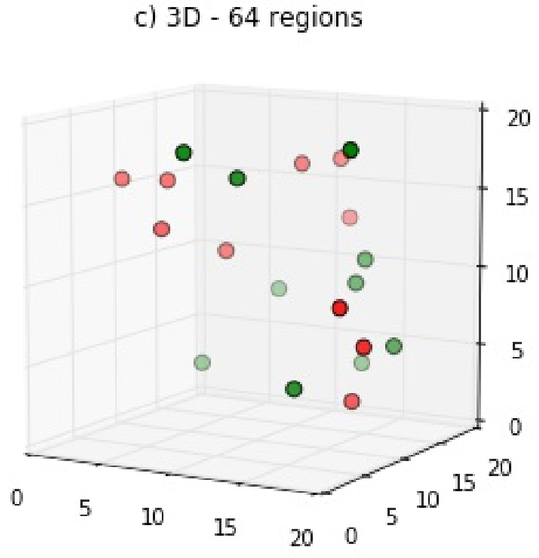


$$ X^T = \{(x_{11},x_{12}),...,(x_{N1}, x_{N2})\} $$

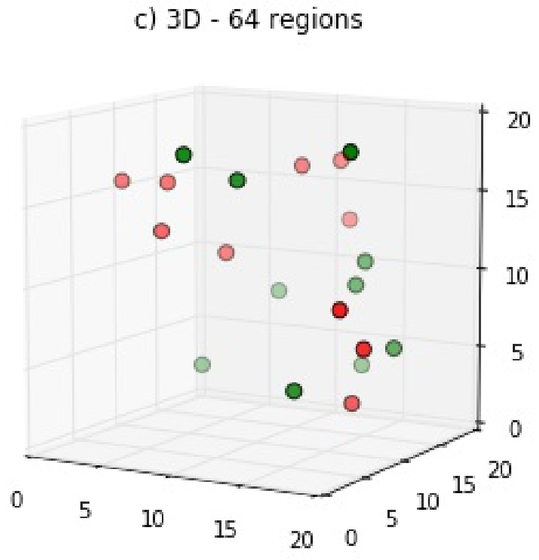
$$ \frac{20}{16} = 1.25 \text{ obs per region}$$

$$ \frac{20}{64} \approx 0.31 \text{ obs per region}$$

http://cleverowl.uk/2016/02/06/curse-of-dimensionality-explained/

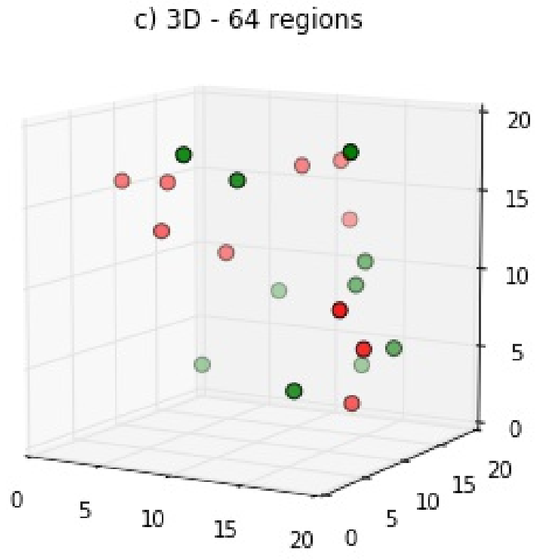


http://cleverowl.uk/2016/02/06/curse-of-dimensionality-explained/



$$ \text{Sampling density is proportional to } N^{\frac{1}{D}}$$

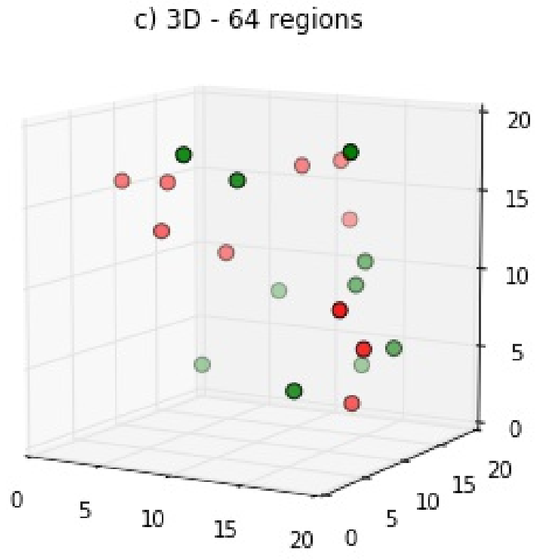
http://cleverowl.uk/2016/02/06/curse-of-dimensionality-explained/



$$ \text{Sampling density is proportional to } N^{\frac{1}{D}}$$

$$ 20^{\frac{1}{1}} = x^{\frac{1}{3}}$$

http://cleverowl.uk/2016/02/06/curse-of-dimensionality-explained/



$$ \text{Sampling density is proportional to } N^{\frac{1}{D}}$$

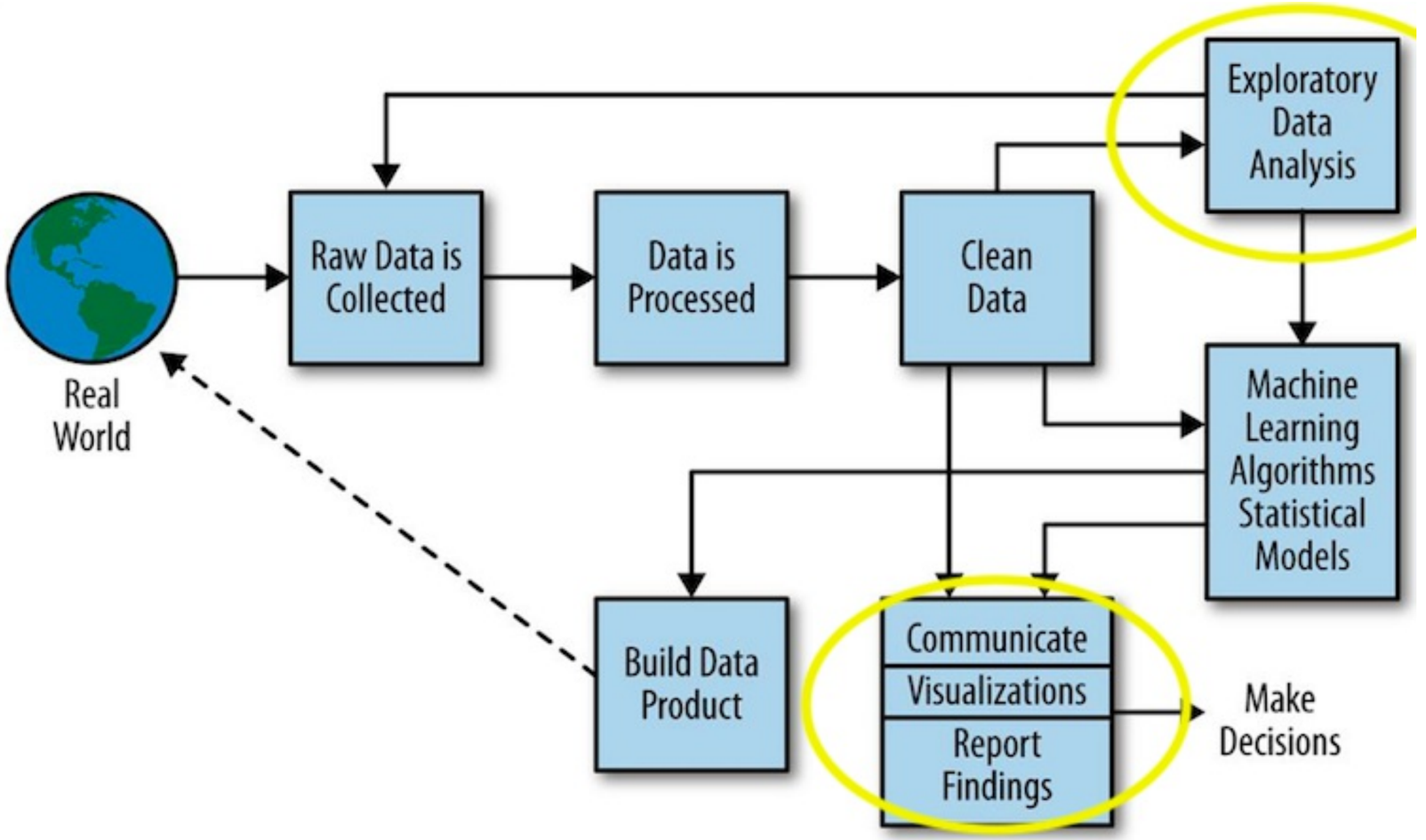
$$ 20^{\frac{1}{1}} = x^{\frac{1}{3}}$$

$$ x = 8000 $$

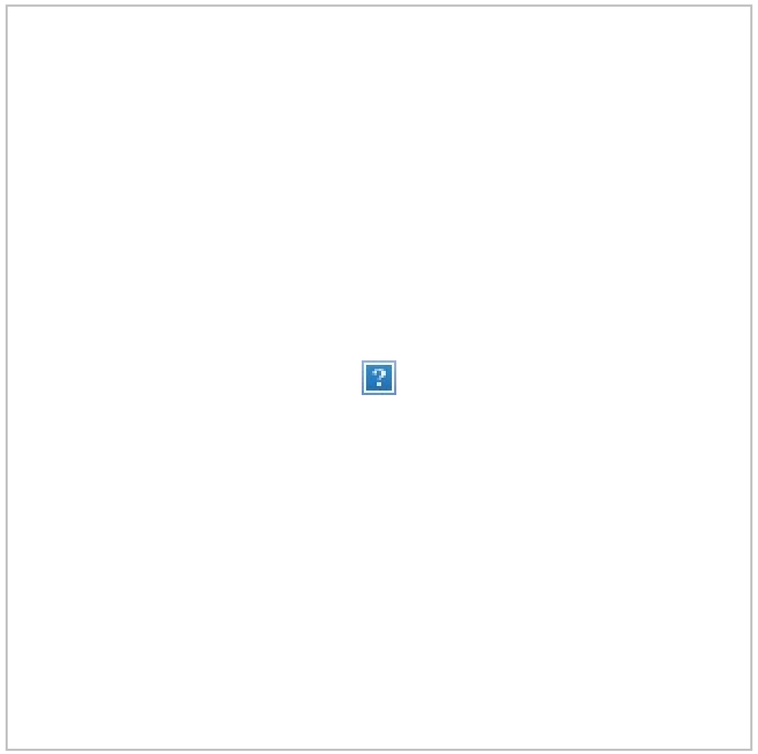
http://cleverowl.uk/2016/02/06/curse-of-dimensionality-explained/

Visualization

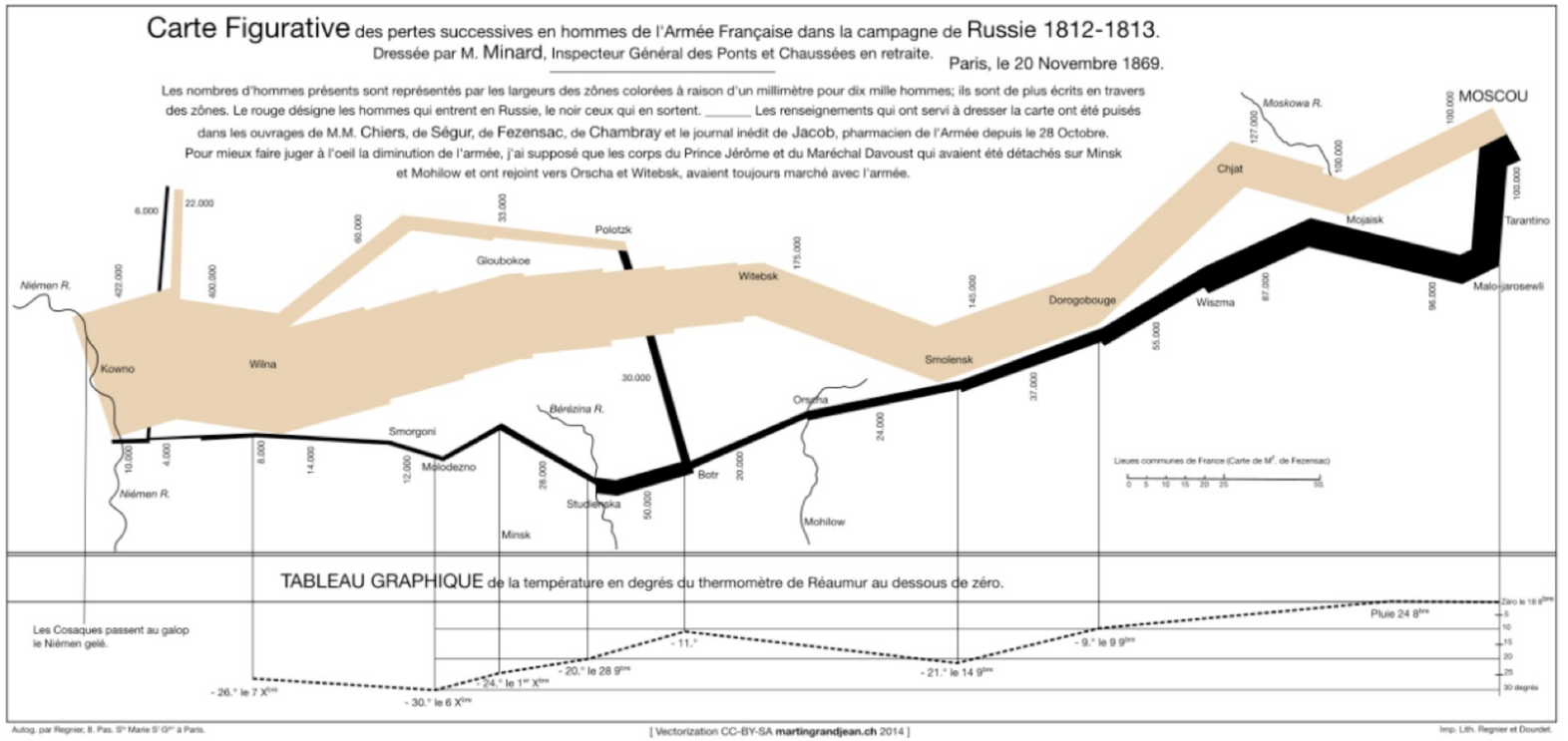
Data Science Pipeline



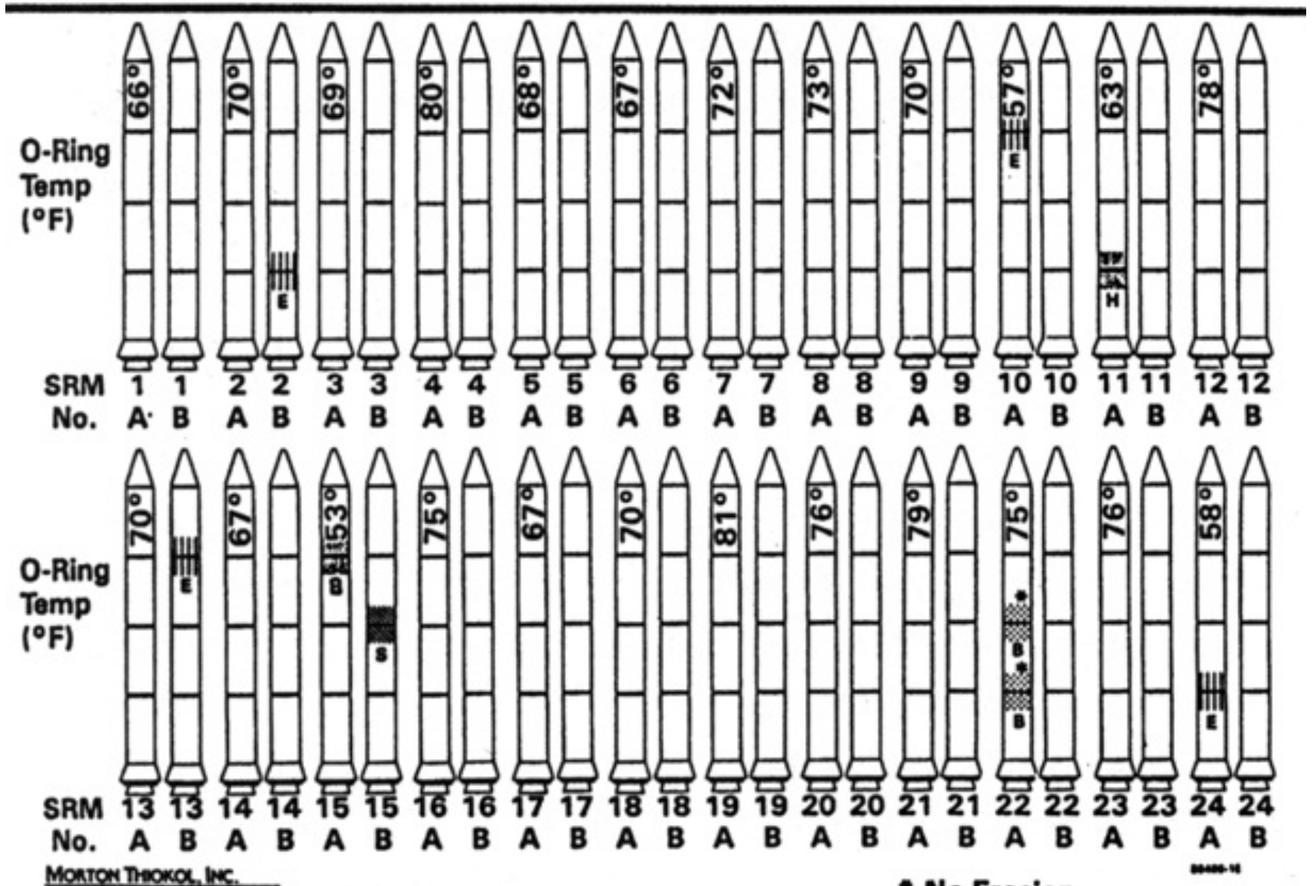
"Doing Data Science"



Edward Tufte


Carte Figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812-1813.
Dressée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite.
Paris, le 20 Novembre 1869.
Les nombres d'hommes présents sont représentés par les largeurs des zônes colorées à raison d'un millimètre pour dix mille hommes; ils sont de plus écrits en travers des zônes. Le rouge désigne les hommes qui entrent en Russie, le noir ceux qui en sortent. _____ Les renseignements qui ont servi à dresser la carte ont été puisés dans les ouvrages de M.M. Chiers, de Ségur, de Fezensac, de Chambray et le journal inédit de Jacob, pharmacien de l'Armée depuis le 28 Octobre.
Pour mieux faire juger à l'oeil la diminution de l'armée, j'ai supposé que les corps du Prince Jérôme et du Maréchal Davoust qui avaient été détachés sur Minsk et Mohilow et ont rejoint vers Orscha et Witebsk, avaient toujours marché avec l'armée.
Niémen R.
Kowno
Wilna
Gloubokoe
Polotzk
Witebsk
Smorgoni
Molodezno
Minsk
Studienska
Bérézina R.
Botr
Orscha
Mohilow
Smolensk
Dorogobouge
Moskowa R.
Chjat
Mojaisk
Wiszma
MOSCOU
Tarantino
Malo-jarosewli
Lieues communes de France (Carte de M.r de Fezensac)
TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.
Les Cosaques passent au galop le Niémen gelé.
- 26.° le 7 Xbre
- 30.° le 6 Xbre
- 24.° le 1er Xbre
- 20.° le 28 9bre
- 11.°
- 21.° le 14 9bre
- 9.° le 9 9bre
Pluie 24 8bre
Zéro le 18 8bre
Autog. par Regnier, 8. Pas. St.e Marie St. Gn.e à Paris.
[Vectorization CC-BY-SA martingrandjean.ch 2014]
Imp. Lith. Regnier et Dourdet.

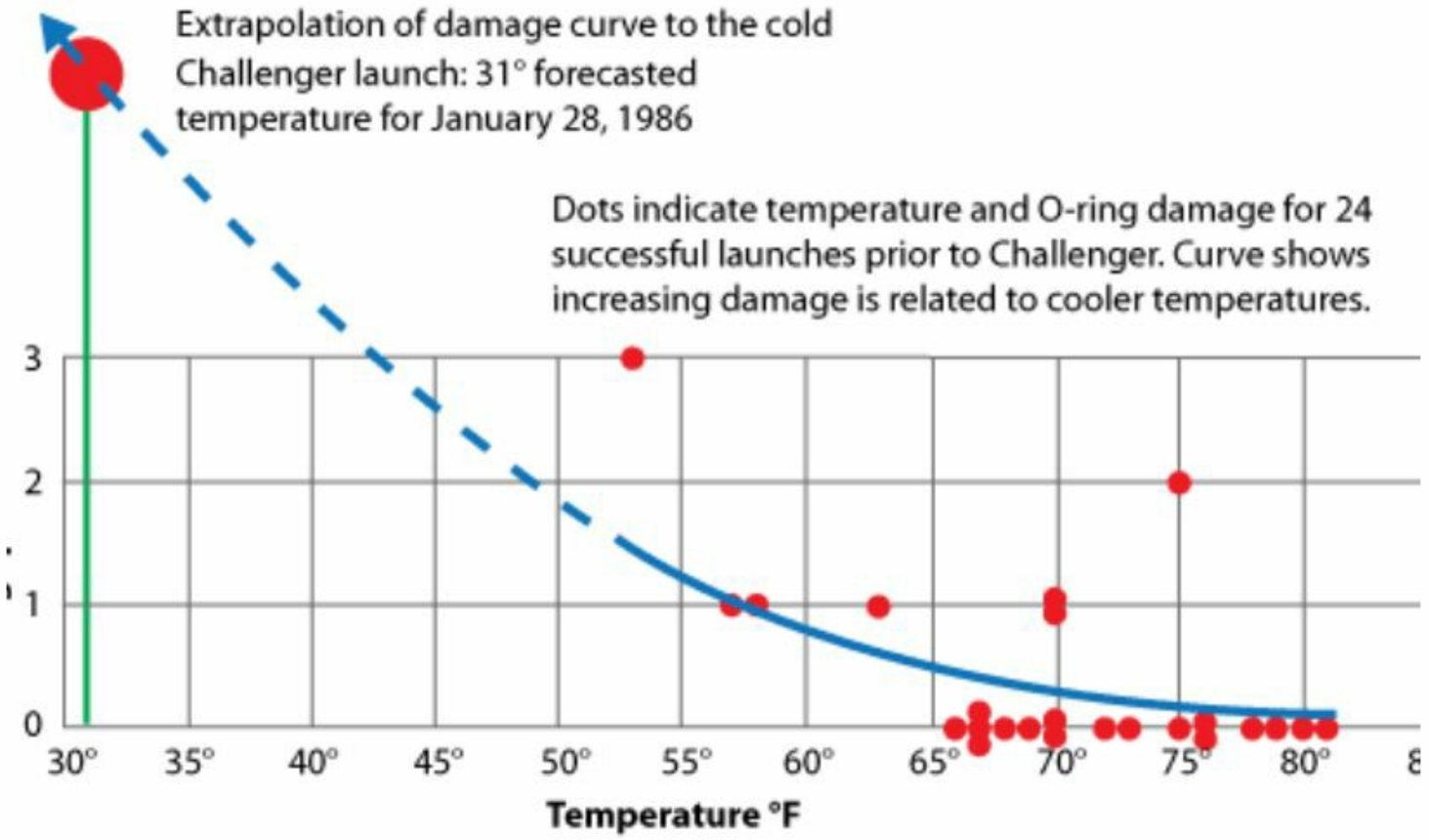
History of O-Ring Damage in Field Joints (Cont)



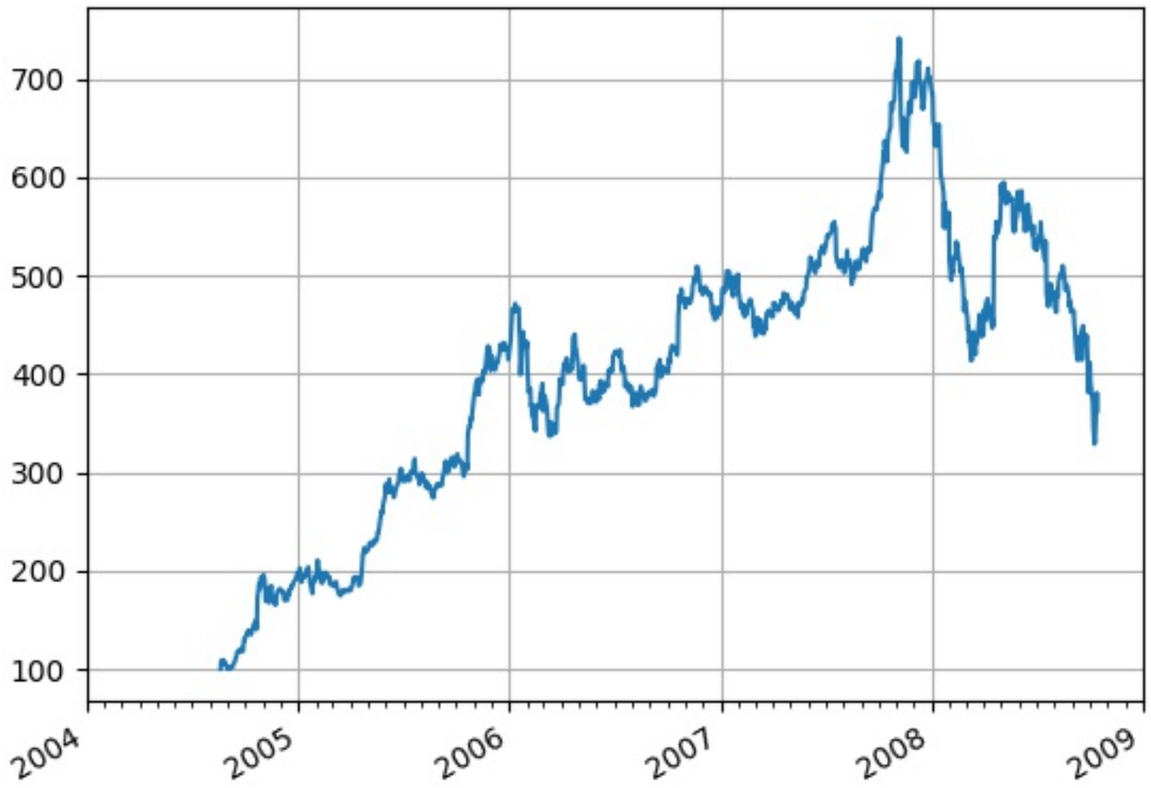
MORTON THIOKOL, INC.
Wasatch Operations

INFORMATION ON THIS PAGE WAS PREPARED TO SUPPORT AN ORAL PRESENTATION AND CANNOT BE CONSIDERED COMPLETE WITHOUT THE ORAL DISCUSSION

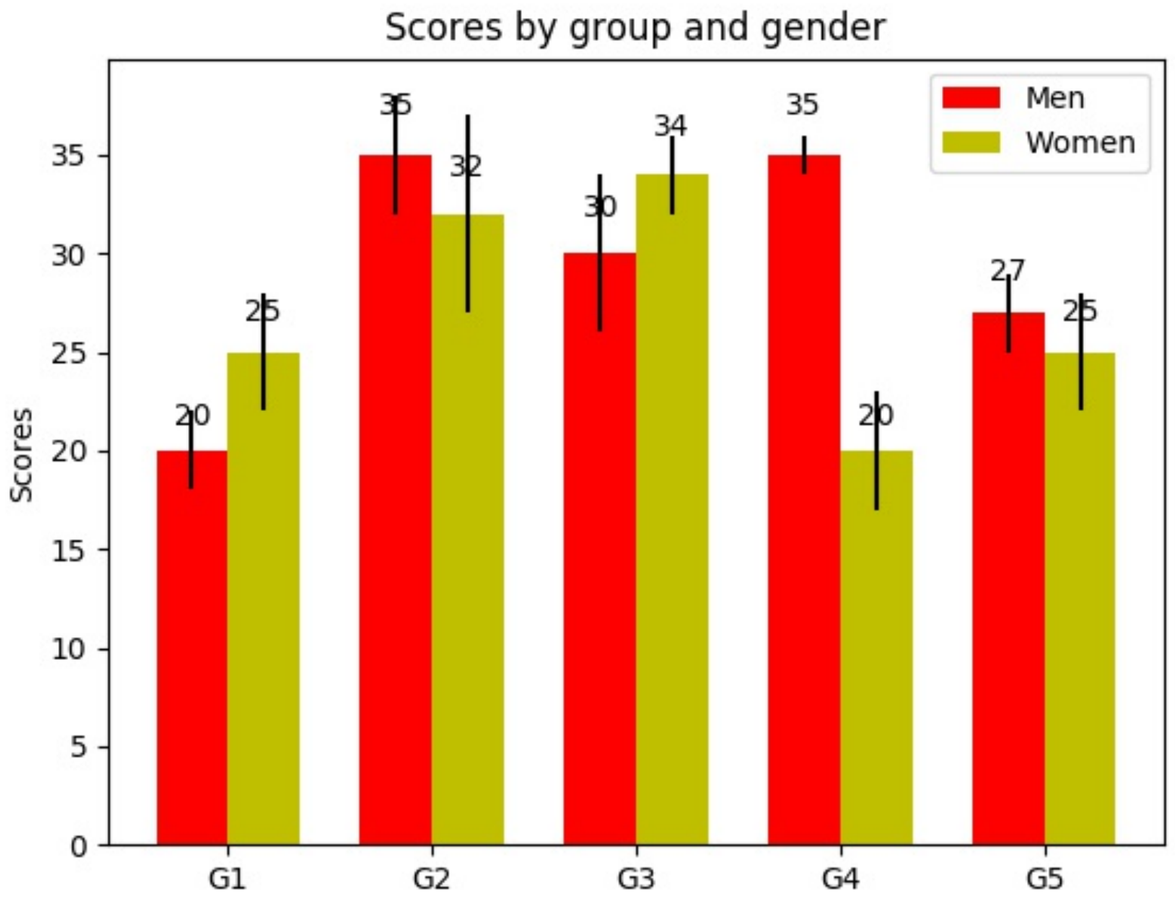
Redrawn (Tufte)



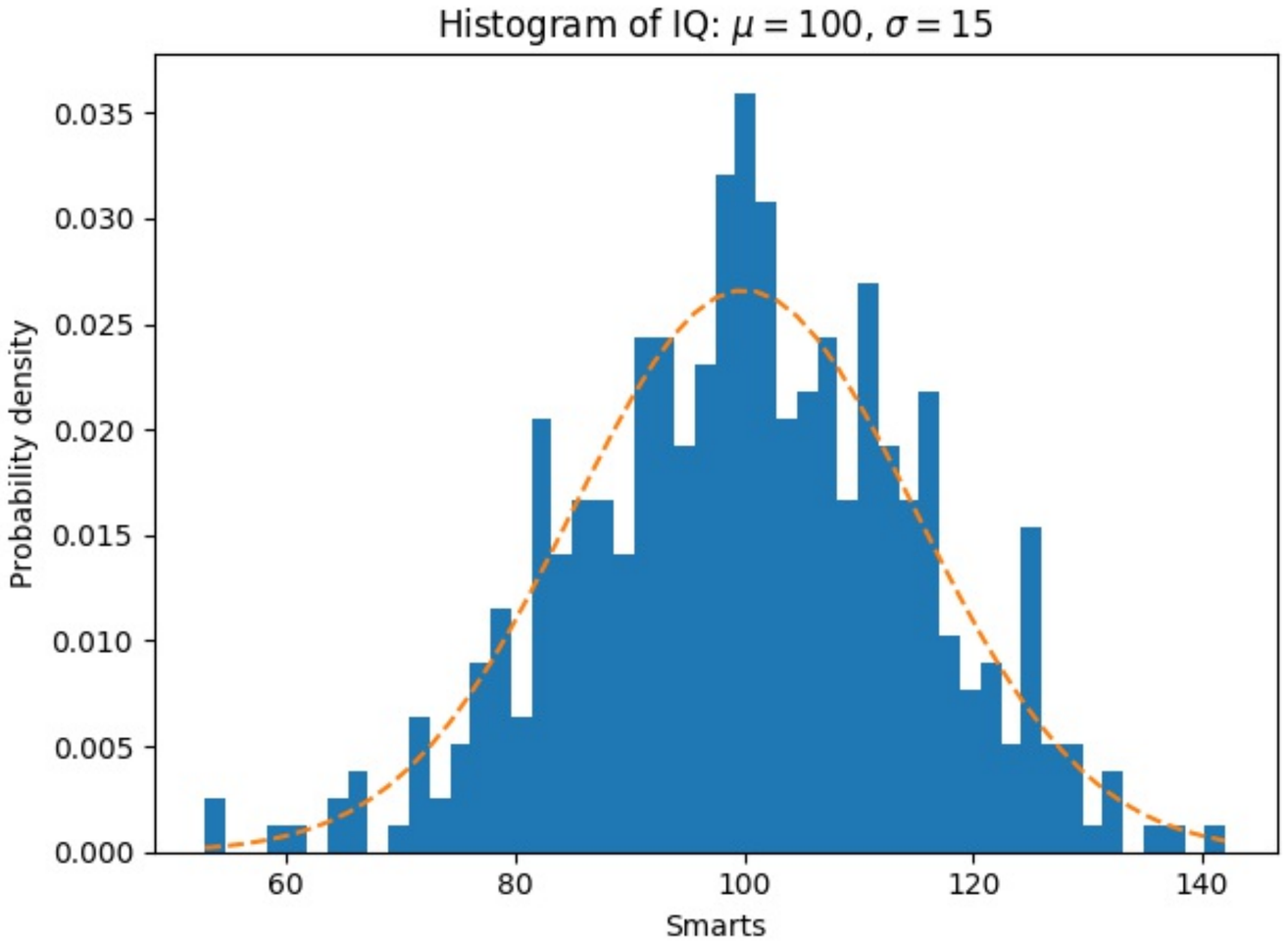
Timeseries data



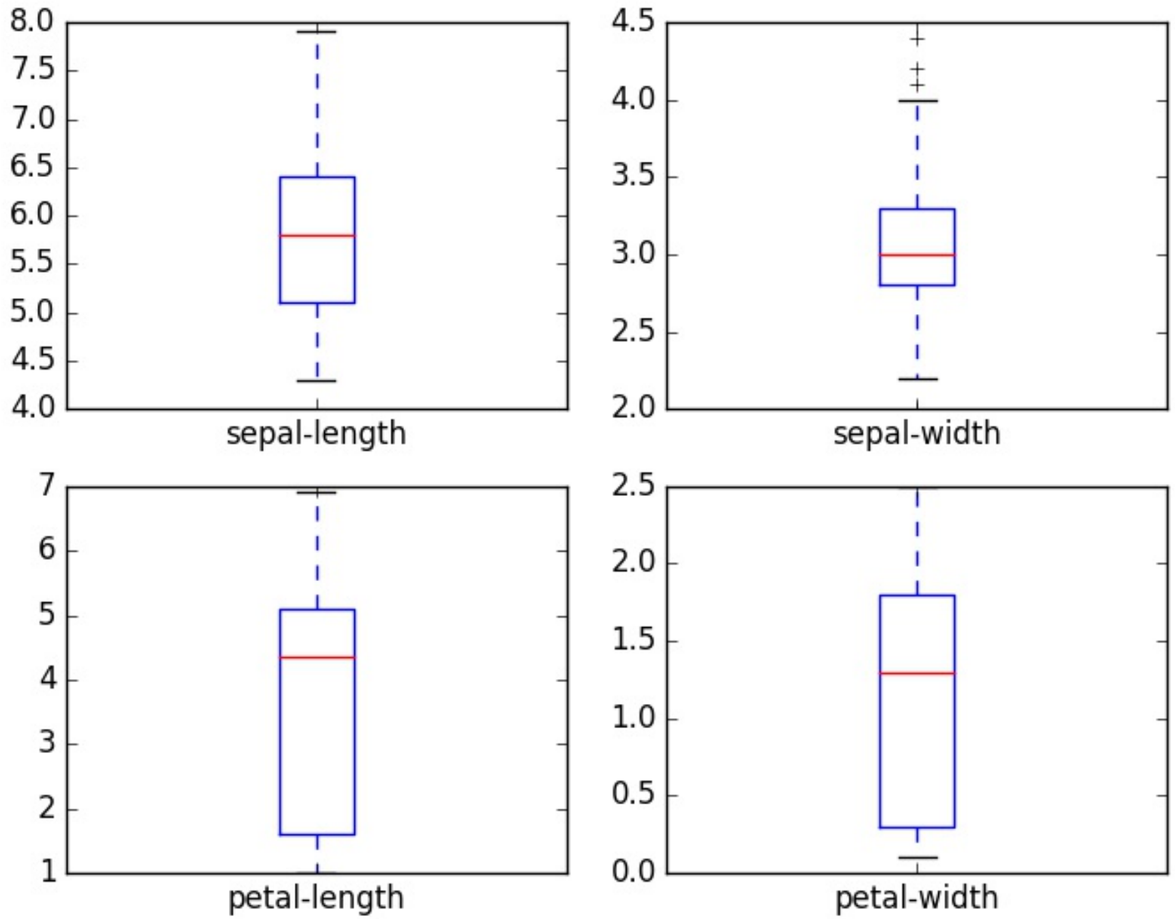
Barcharts


Scores by group and gender
Men
Women
Scores
0
5
10
15
20
25
30
35
20
25
35
32
30
34
35
20
27
25
G1
G2
G3
G4
G5

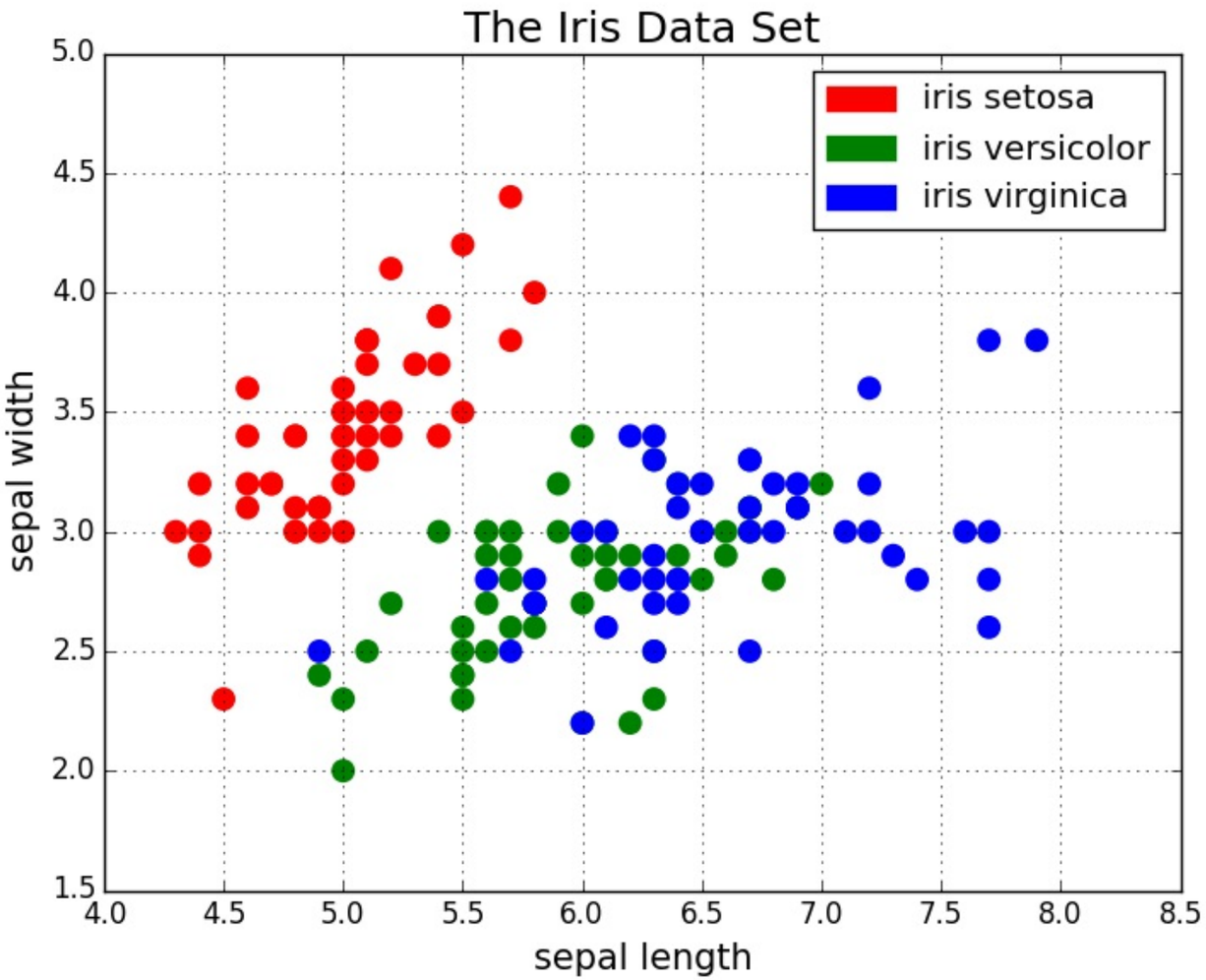
Histograms


Histogram of IQ: $\mu = 100$, $\sigma = 15$
Probability density
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
60
80
100
120
140
Smarts

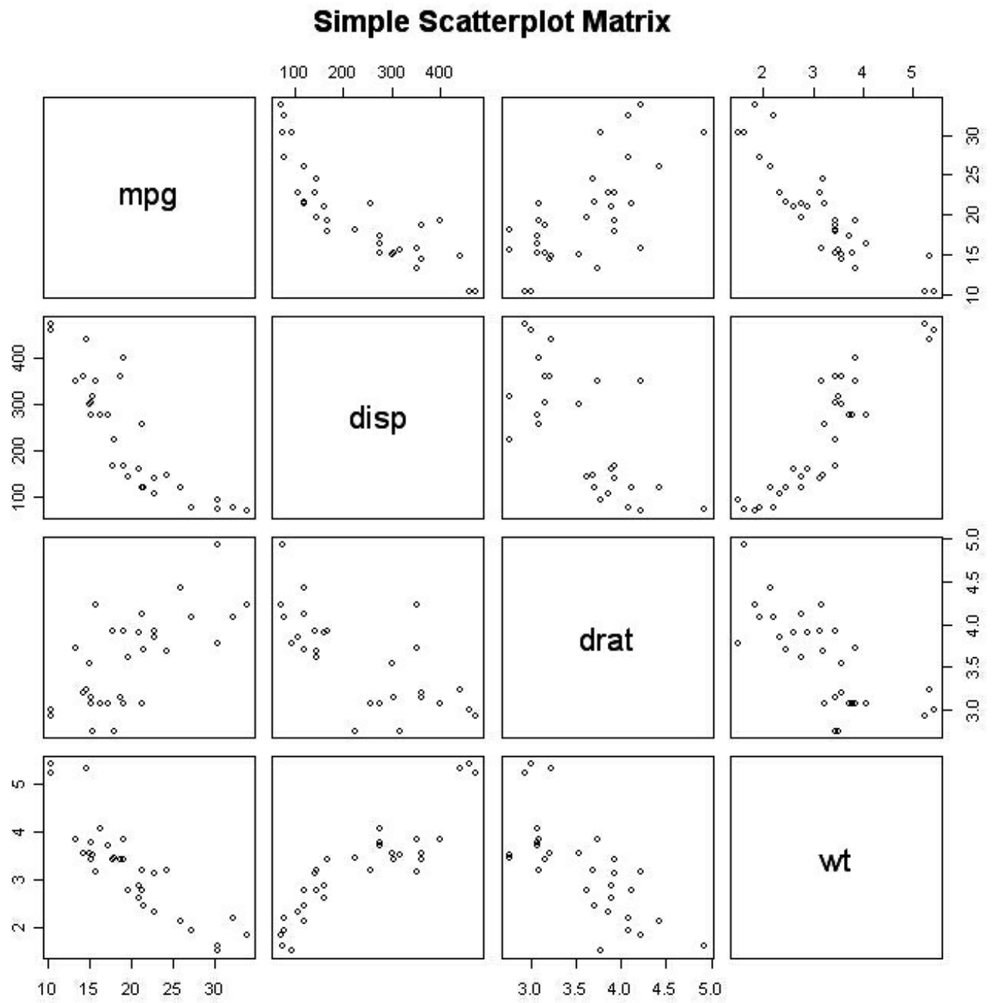
Box and Whiskers Plot



Scatter plots



Scatter plot matrices



Exercise

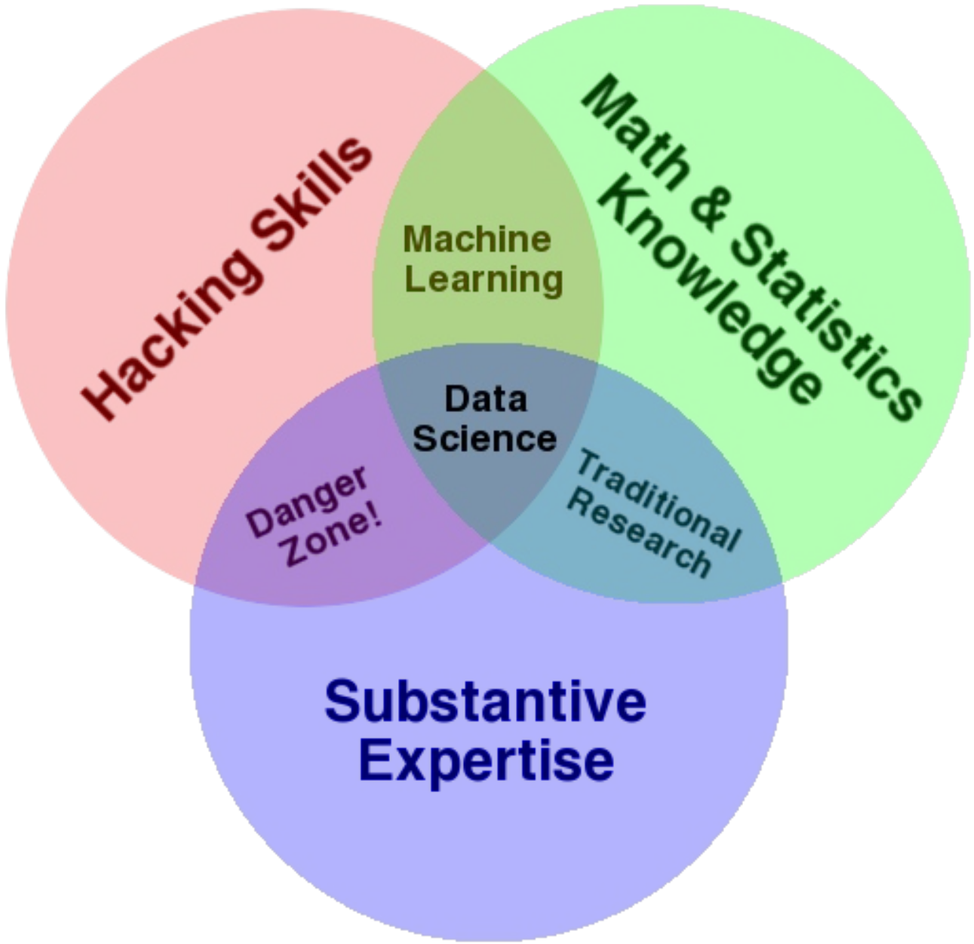
Pod TTL Visualization

- Timeseries
- Histogram of app cpu data
- Scatter plot of transactions vs app cpu data
- Pairs of variable scatter plots
- Running median smoothing example on minute level data for db cpu
- Transition from minute to hour

"Data scientist: n. person who is better at statistics than any software engineer and better at software engineering than any statistician."

Josh Wills

Data Science



"Drew Conway"

Why is the discipline so important?

"Long before worrying about how to convince others, you first have to understand what's happening yourself."

Andrew Gelman

"Naï¨ve realism, also known as direct realism or common sense realism, is a philosophy of mind rooted in a theory of perception that claims that the senses provide us with direct awareness of the external world."

http://en.wikipedia.org/wiki/Naï¨ve_realism

1951 Princeton/Dartmouth Football Game

1951 Princeton/Dartmouth Football Game

- Storied rivalry

1951 Princeton/Dartmouth Football Game

- Storied rivalry
- Princeton's star player's nose was broken

1951 Princeton/Dartmouth Football Game

- Storied rivalry
- Princeton's star player's nose was broken
- Another Princeton player retaliated and broke Dartmouth player's leg

1951 Princeton/Dartmouth Football Game

- Storied rivalry
- Princeton's star player's nose was broken
- Another Princeton player retaliated and broke Dartmouth player's leg
- Princeton won (13-0)

1951 Princeton/Dartmouth Football Game

- Storied rivalry
- Princeton's star player's nose was broken
- Another Princeton player retaliated and broke Dartmouth player's leg
- Princeton won (13-0)
- Both teams blamed the other side

1951 Princeton/Dartmouth Football Game

- Storied rivalry
- Princeton's star player's nose was broken
- Another Princeton player retaliated and broke Dartmouth player's leg
- Princeton won (13-0)
- Both teams blamed the other side
- Two versions of the Truth

"They Saw a Game"

"They Saw a Game"

- Albert Hastorf (Dartmouth) and Hadley Cantril (Princeton) showed the game again to students from both schools

"They Saw a Game"

- Albert Hastorf (Dartmouth) and Hadley Cantril (Princeton) showed the game again to students from both schools
- Asked them to notice infractions, penalties, fill out a questionnaire

"They Saw a Game"

- Albert Hastorf (Dartmouth) and Hadley Cantril (Princeton) showed the game again to students from both schools
- Asked them to notice infractions, penalties, fill out a questionnaire
- Princeton students 'saw' twice as many infractions by Dartmouth players than Dartmouth students did

"They Saw a Game"

- Albert Hastorf (Dartmouth) and Hadley Cantril (Princeton) showed the game again to students from both schools
- Asked them to notice infractions, penalties, fill out a questionnaire
- Princeton students 'saw' twice as many infractions by Dartmouth players than Dartmouth students did
- Dartmouth students saw a 'rough but fair' game

"In brief, the data here indicate that there is no such 'thing' as a 'game' existing 'out there' in its own right which people merely 'observe.' The game 'exists' for a person and is experienced by him only insofar as certain happenings have significances in terms of his purpose."

Hastorf and Cantril

"Everything that has ever happened to you has happened inside your skull."

David McRaney, "You Are Now Less Dumb"

Compare the Students

Compare the Students

- All male

Compare the Students

- All male
- Mostly similar ethnically and socioeconomically

Compare the Students

- All male
- Mostly similar ethnically and socioeconomically
- Geographically similar

Compare the Students

- All male
- Mostly similar ethnically and socioeconomically
- Geographically similar
- Similar in age

Compare the Students

- All male
- Mostly similar ethnically and socioeconomically
- Geographically similar
- Similar in age
- Similar basic cultural and religious beliefs

Compare the Students

- All male
- Mostly similar ethnically and socioeconomically
- Geographically similar
- Similar in age
- Similar basic cultural and religious beliefs
- Went to different schools

"It’s a real problem, though, when politicians, CEOs, and other people with the power to change the way the world works start bungling their arguments for or against things based on self-delusion generated by imperfect minds and senses."

David McRaney, "You Are Now Less Dumb"

"Narratives are meaning transmitters. They are history-preservation devices. They create and maintain cultures, and they forge identities that emerge out of the malleable, imperfect memories of life events."

David McRaney, "You Are Now Less Dumb"

"Your narrative bias makes it nearly impossible for you to really absorb the information from the outside world without arranging it into causes and effects."

David McRaney, "You Are Now Less Dumb"

"Your ancestors invented the scientific method because the common belief fallacy renders your default strategies for making sense of the world generally awful and prone to error."

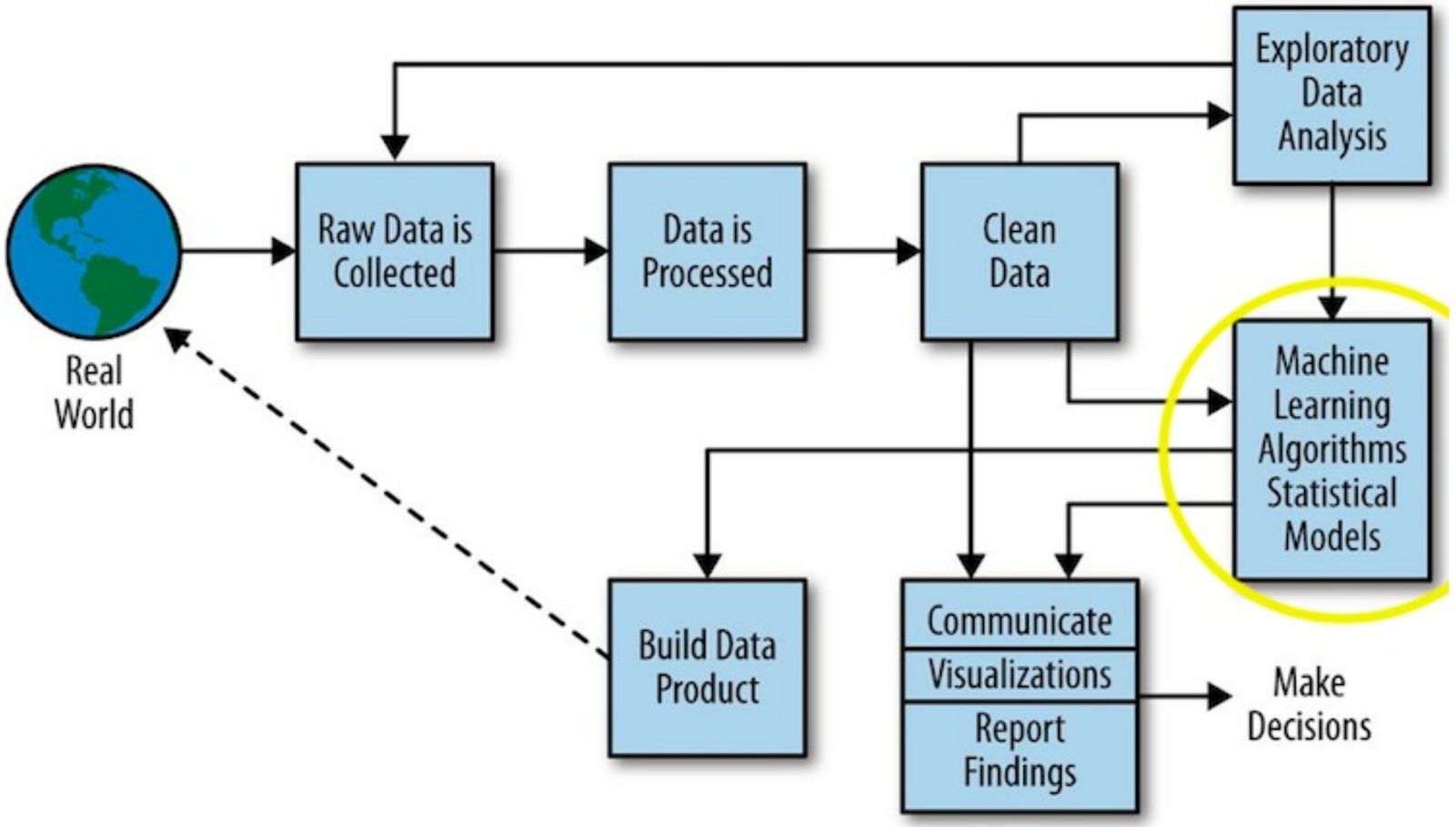
David McRaney, "You Are Now Less Dumb"

"Your natural tendency is to start from a conclusion and work backward to confirm your assumptions, but the scientific method drives down the wrong side of the road and tries to disconfirm your assumptions."

David McRaney, "You Are Now Less Dumb"

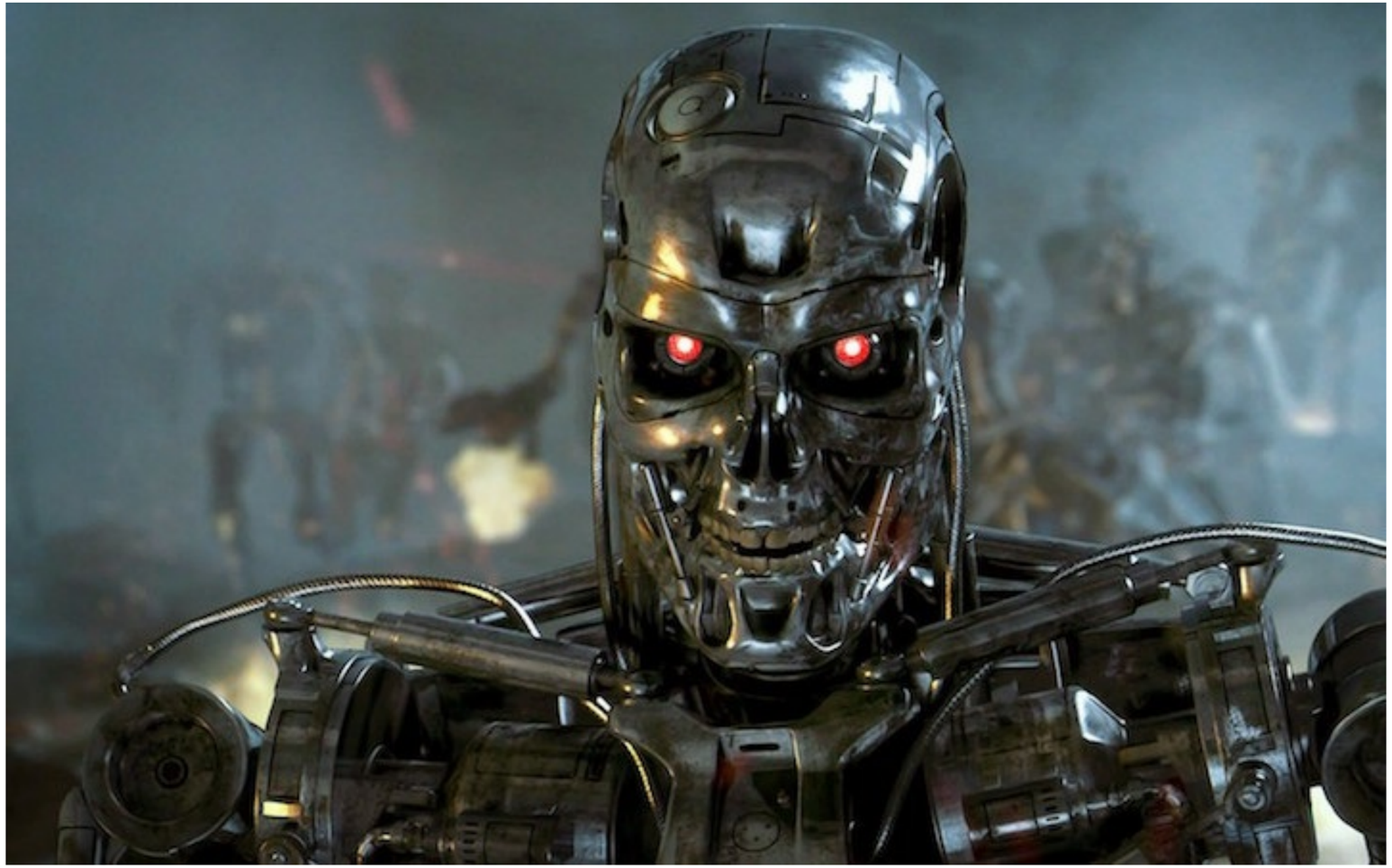
Data-Directed

Data Science Pipeline

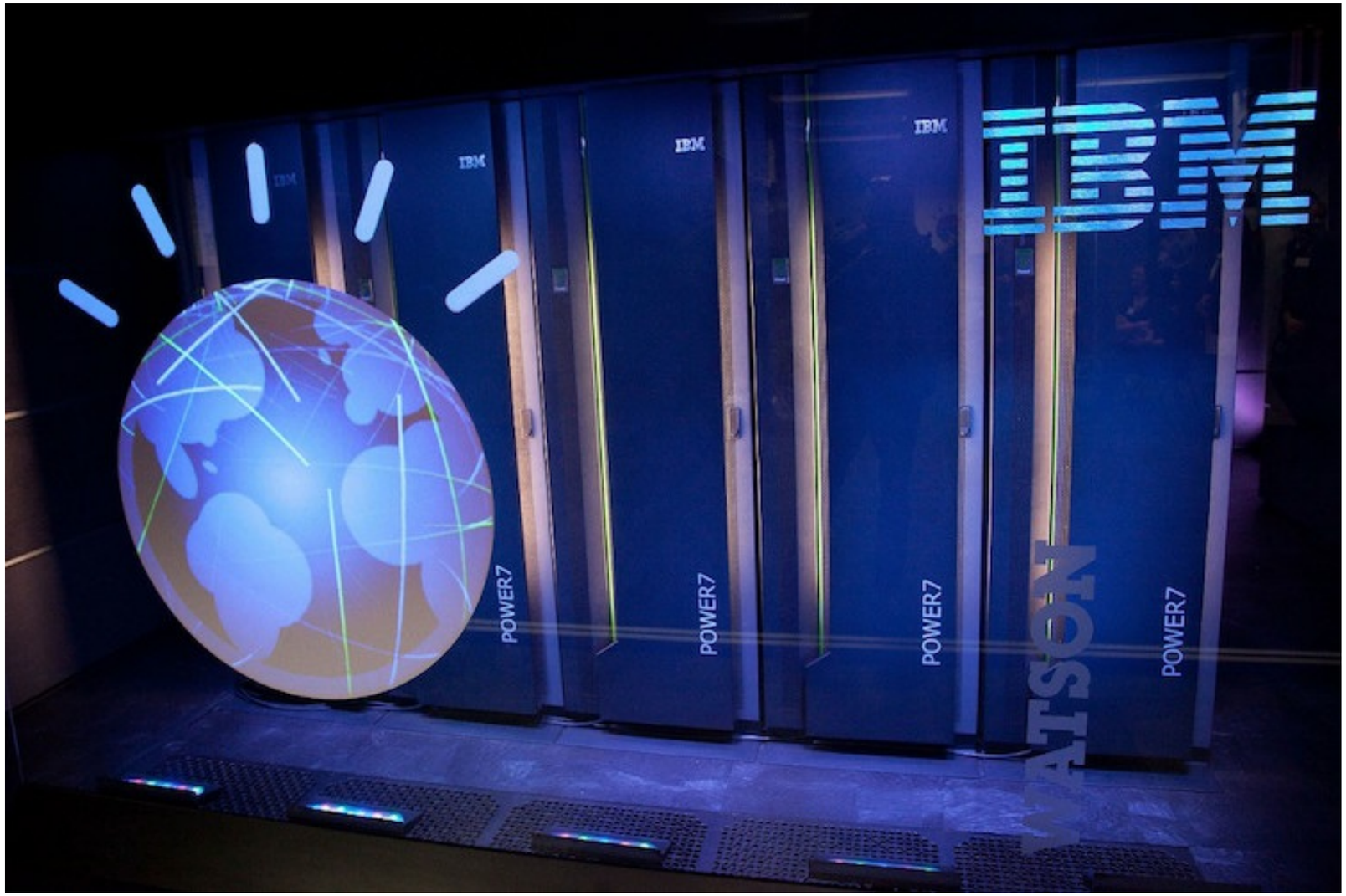


"Doing Data Science"


WOPR
WOPR



https://youtu.be/rVlhMGQgDkY?t=85


IBM
IBM
IBM
IBM
POWER7
POWER7
POWER7
POWER7
WATSON


304P
52 RP

"The term machine learning refers to the automated detection of meaningful patterns in data."

Source: Shavel-Shwartz and Ben-David, "Understanding Machine Learning: From Theory to Algorithms"

Advancement of Machine Learning

Advancement of Machine Learning

- Stock market prediction in the 1980s

Advancement of Machine Learning

- Stock market prediction in the 1980s
- Mining corporate databases in the 1990s (direct marketing, CRM, credit scoring, fraud detection)

Advancement of Machine Learning

- Stock market prediction in the 1980s
- Mining corporate databases in the 1990s (direct marketing, CRM, credit scoring, fraud detection)
- E-commerce (personalization, click analysis)

Advancement of Machine Learning

- Stock market prediction in the 1980s
- Mining corporate databases in the 1990s (direct marketing, CRM, credit scoring, fraud detection)
- E-commerce (personalization, click analysis)
- 9/11 brought interest to applying ML to fighting terror

Advancement of Machine Learning

- Stock market prediction in the 1980s
- Mining corporate databases in the 1990s (direct marketing, CRM, credit scoring, fraud detection)
- E-commerce (personalization, click analysis)
- 9/11 brought interest to applying ML to fighting terror
- Web 2.0 (social networks, sentiment analysis, etc.)

Advancement of Machine Learning

- Stock market prediction in the 1980s
- Mining corporate databases in the 1990s (direct marketing, CRM, credit scoring, fraud detection)
- E-commerce (personalization, click analysis)
- 9/11 brought interest to applying ML to fighting terror
- Web 2.0 (social networks, sentiment analysis, etc.)
- Science (molecular biologists and astronomers were early adopters)

Advancement of Machine Learning

- Stock market prediction in the 1980s
- Mining corporate databases in the 1990s (direct marketing, CRM, credit scoring, fraud detection)
- E-commerce (personalization, click analysis)
- 9/11 brought interest to applying ML to fighting terror
- Web 2.0 (social networks, sentiment analysis, etc.)
- Science (molecular biologists and astronomers were early adopters)
- Housing bust freed up a lot of talent

Advancement of Machine Learning

- Stock market prediction in the 1980s
- Mining corporate databases in the 1990s (direct marketing, CRM, credit scoring, fraud detection)
- E-commerce (personalization, click analysis)
- 9/11 brought interest to applying ML to fighting terror
- Web 2.0 (social networks, sentiment analysis, etc.)
- Science (molecular biologists and astronomers were early adopters)
- Housing bust freed up a lot of talent
- Big Data

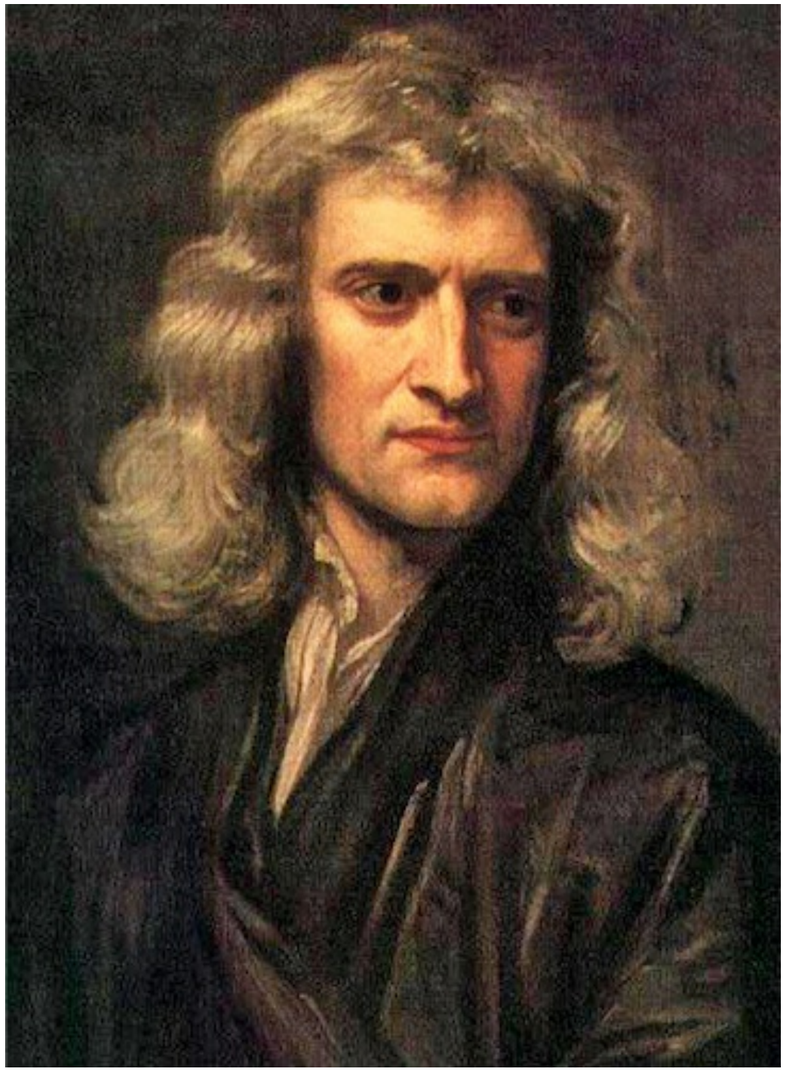
David Hume

Inductivist Turkey

CC BY 2.0, https://www.flickr.com/photos/29311691@N05/3577002803

Tycho Brahe

Johannes Kepler



Isaac Newton

Supervised Learning

Supervised Learning

$$\text{Given input variables x and output variable Y} $$

Supervised Learning

$$\text{Given input variables x and output variable Y} $$

$$ Y = f(x)$$

Supervised Learning

$$\text{Given input variables x and output variable Y} $$

$$ Y = f(x)$$

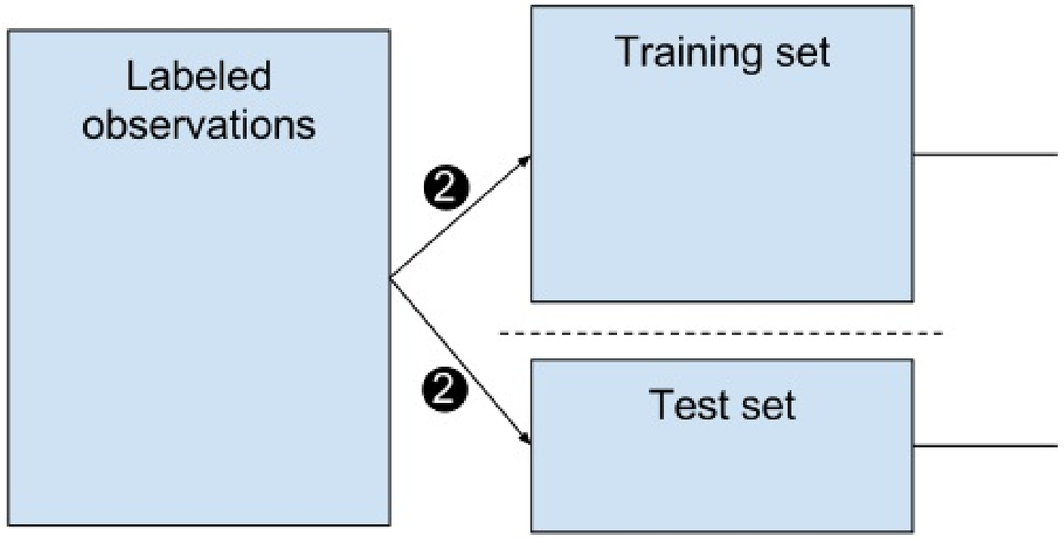
- Classification

Supervised Learning

$$\text{Given input variables x and output variable Y} $$

$$ Y = f(x)$$

- Classification
- Regression


Labeled observations
❷
Training set
❷
Test set

Unsupervised Learning

Unsupervised Learning

$$\text{Given only input variables x, find some underlying structure.} $$

Unsupervised Learning

$$\text{Given only input variables x, find some underlying structure.} $$

- Clustering

Unsupervised Learning

$$\text{Given only input variables x, find some underlying structure.} $$

- Clustering
- Association rules

Linear Regression

"Linear Regression: In statistics, linear regression is an approach for modeling the relationship between a scalar dependent variable y and one or more explanatory variables denoted X."

Source: https://en.wikipedia.org/wiki/Linear_regression

Ordinary Least Squares

Ordinary Least Squares

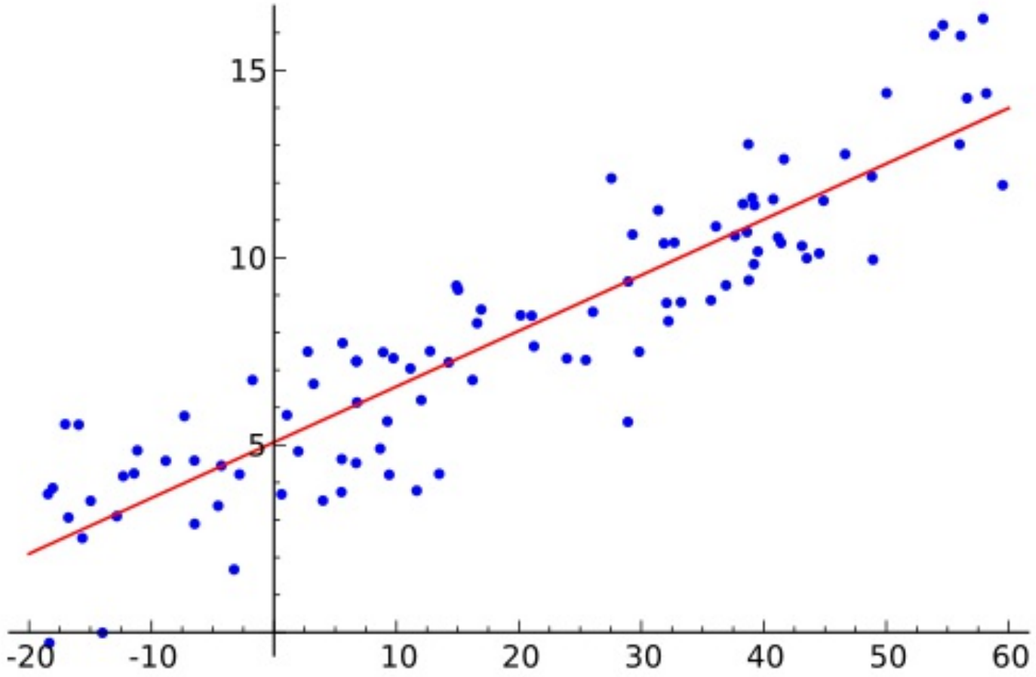
- Simple Linear Regression

Ordinary Least Squares

- Simple Linear Regression
- Fit a straight line through the observed points

Ordinary Least Squares

- Simple Linear Regression
- Fit a straight line through the observed points
- Minimizes the sum of square residuals of the model



https://commons.wikimedia.org/wiki/File:Linear_regression.svg

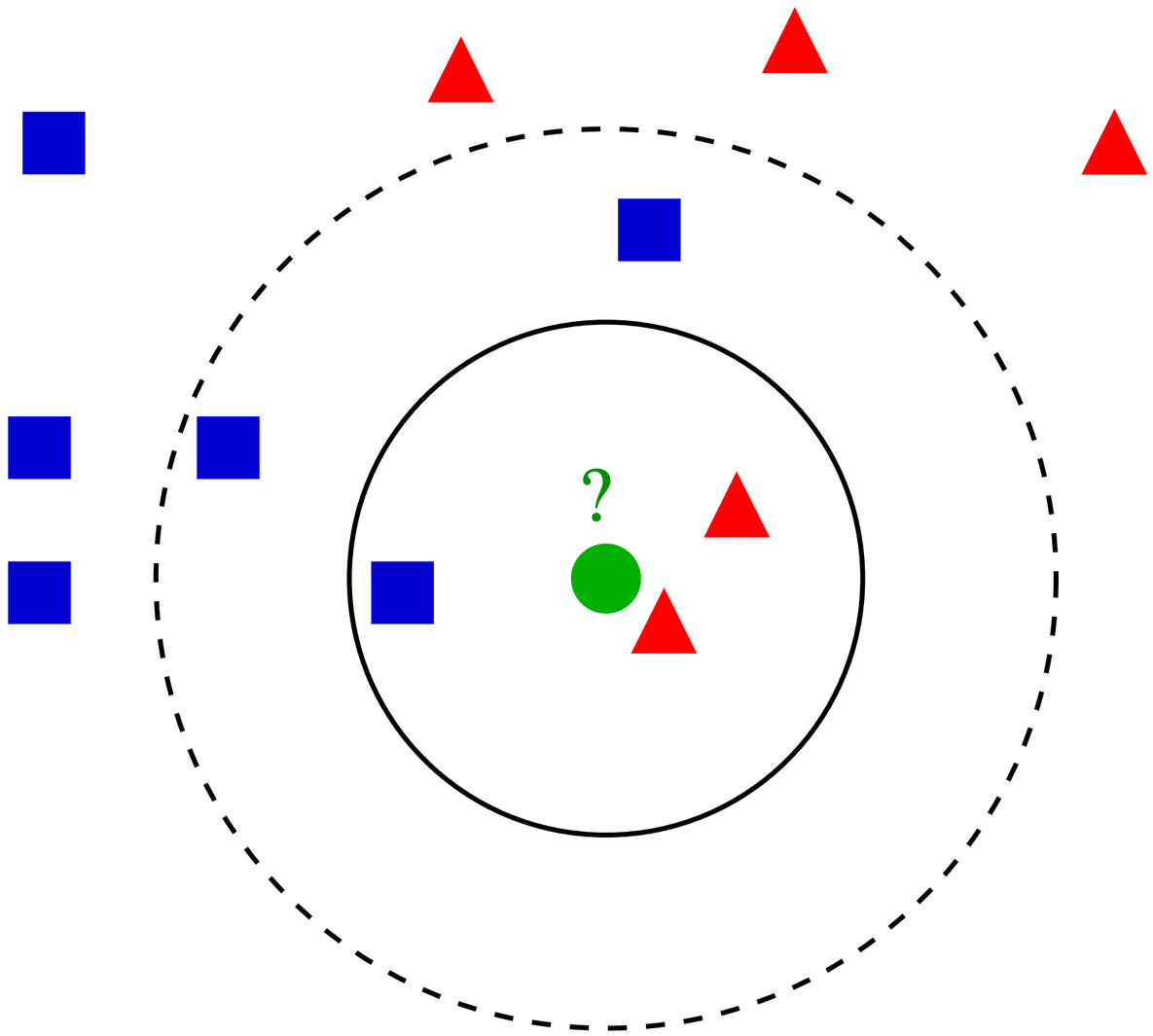
Linear Regression

Common approach for numeric data	Strong assumptions about the data
Can handle most modeling tasks	Model form must be specified in advance
Estimates the strength and size of the relationships among features and outcomes	Does not handle missing data
	Only numeric features
	Requires statistical knowledge to understand the model

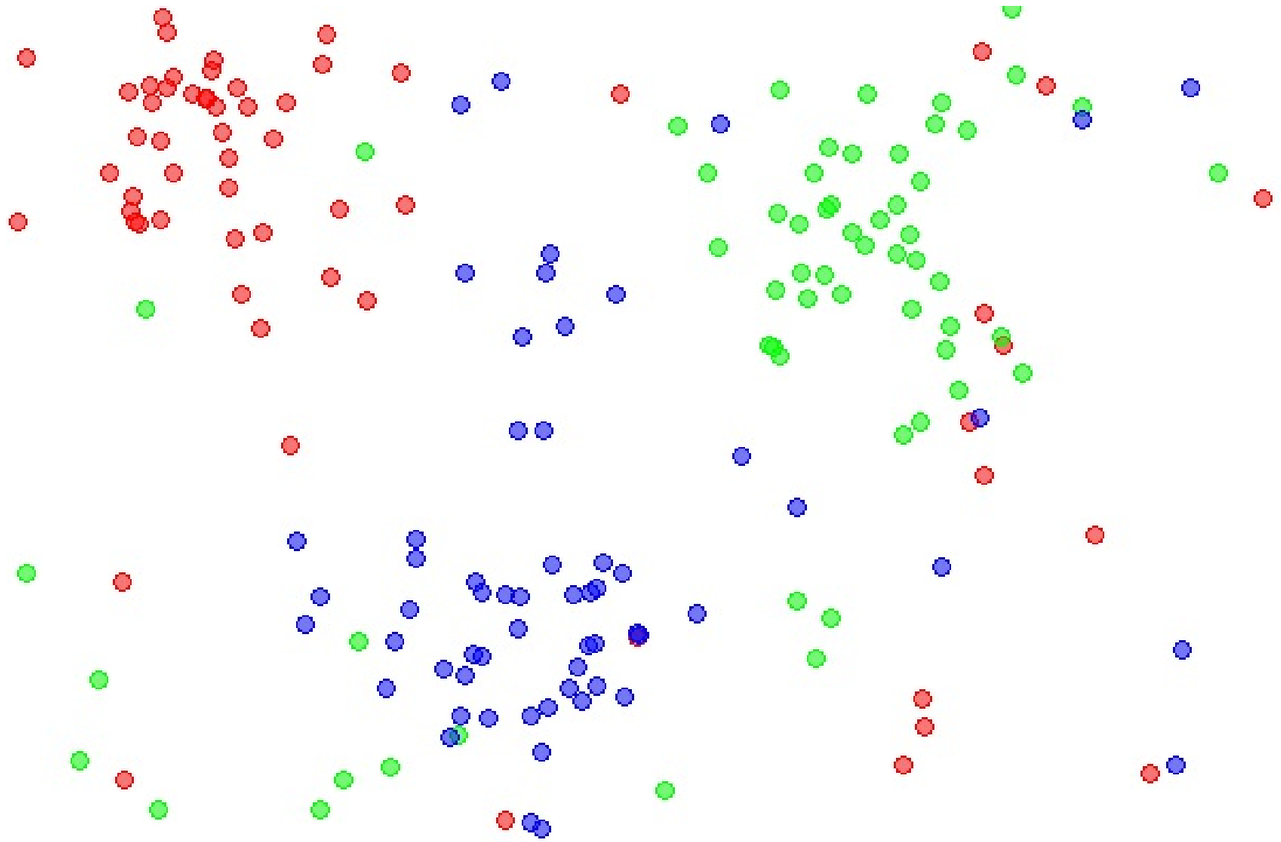
k-Nearest Neighbors

"k-Nearest Neighbor: a non-parametric method used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space. The output depends on whether k-NN is used for classification or regression."

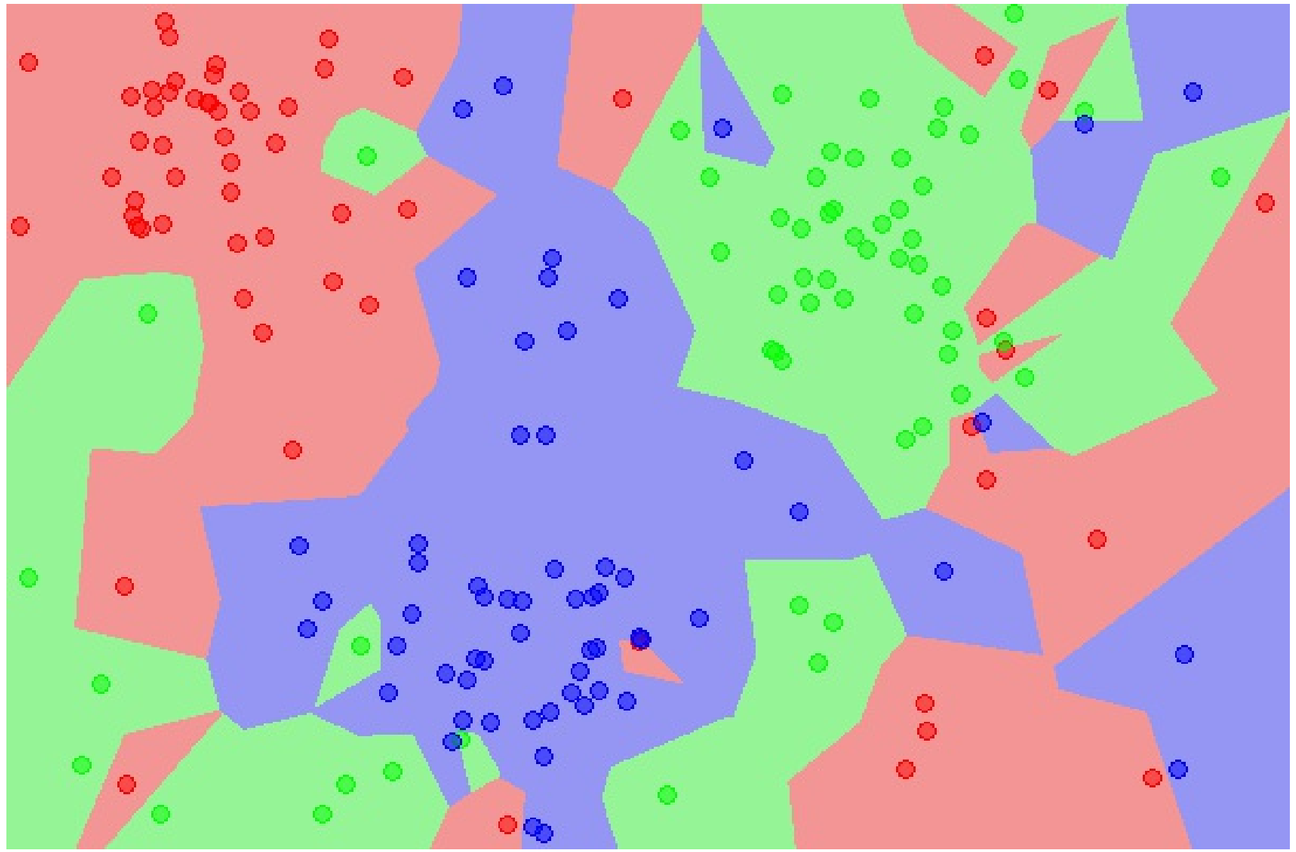
Source: http://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm



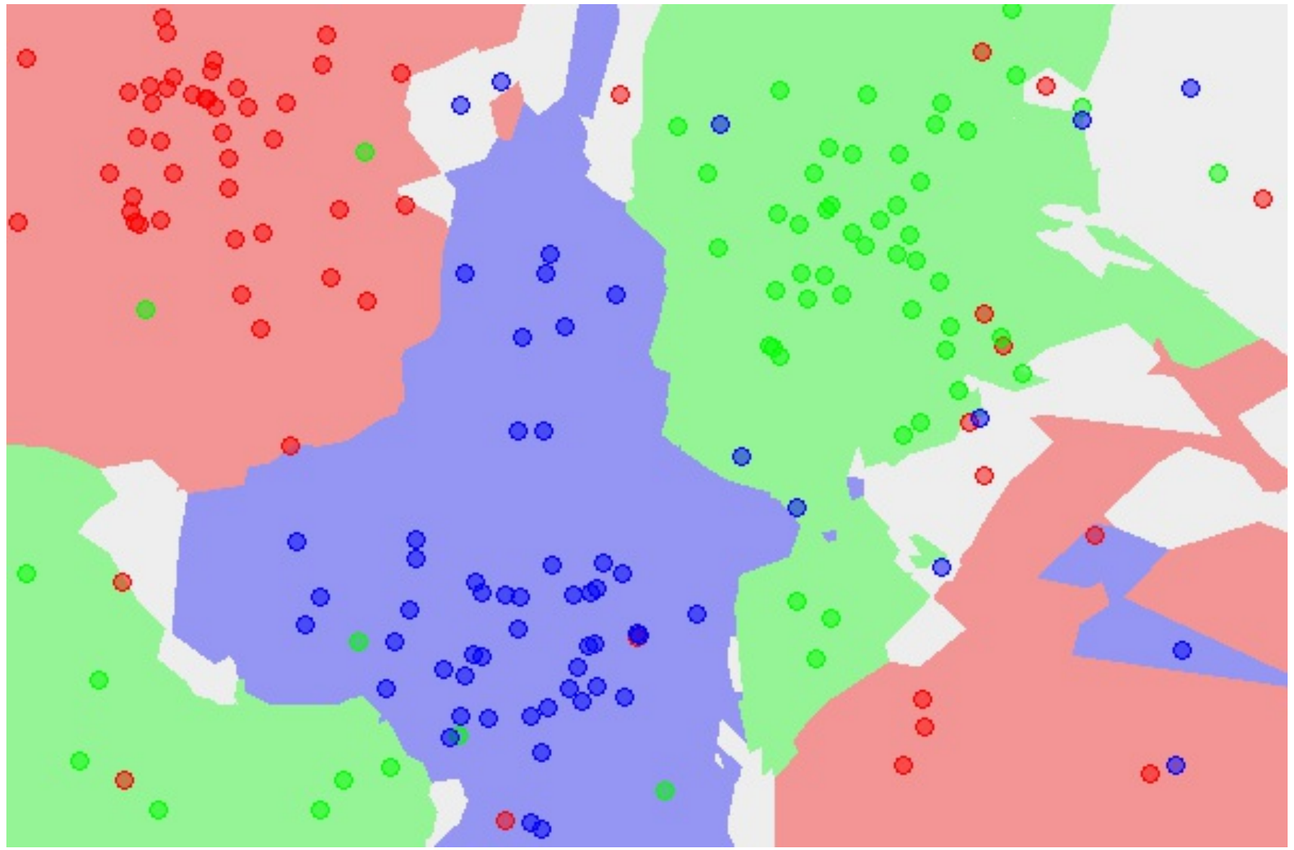
http://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm#/media/File:KnnClassification.svg



http://commons.wikimedia.org/wiki/File:Data3classes.png



http://commons.wikimedia.org/wiki/File:Map1NN.png



http://commons.wikimedia.org/wiki/File:Map5NN.png

Simple and effective	Does not produce a model
Makes no assumptions about the data distribution	Efficacy affected by choice of k
Fast training phase	Slow classification phase

Naive Bayes

Naive Bayes

Naive Bayes

- Family of algorithms to produce probabilistic classifiers based on Bayes Theorem

Naive Bayes

- Family of algorithms to produce probabilistic classifiers based on Bayes Theorem
- Requires relatively little training data

Naive Bayes

- Family of algorithms to produce probabilistic classifiers based on Bayes Theorem
- Requires relatively little training data
- Often used for text/document classification

Naive Bayes

- Family of algorithms to produce probabilistic classifiers based on Bayes Theorem
- Requires relatively little training data
- Often used for text/document classification
- Assumes independence of the features

```
$$P(A), P(B)$$
```

```
$$P(A), P(B)$$

$$P(A \cap B) = P(A) \cdot P(B)$$
```

```
$$P(A), P(B)$$

$$P(A \cap B) = P(A) \cdot P(B)$$

$$P(A | B) = \frac { P(A \cap B) }{ P(B)
}$$
```

```
$$P(A), P(B)$$

$$P(A \cap B) = P(A) \cdot P(B)$$

$$P(A | B) = \frac { P(A \cap B) }{ P(B)
}$$

$$P(A | B) = \frac { P(A \cap B) }{ P(B) }
= \frac{P(B|A)P(A)} {P(B)}$$
```

```
$$P(A), P(B)$$

$$P(A \cap B) = P(A) \cdot P(B)$$

$$P(A | B) = \frac { P(A \cap B) }{ P(B)
}$$

$$P(A | B) = \frac { P(A \cap B) }{ P(B) }
= \frac{P(B|A)P(A)} {P(B)}$$

$$P(spam|Viagra) = \frac { P(Viagra|spam)
\cdot P(spam) } { P(Viagra) }$$
```

Simple and effective	Assumption of the independence of features is usually wrong
Does well with noisy and missing data	Doesn't work well with lots of numeric features
Works well with arbitrary sizes of training data	Estimated probabilities aren't as reliable as the classifications
Easy to produce the estimated probability for predictions	

Decision Trees

Decision Trees

Decision Trees

- Tree structure-based classifier

Decision Trees

- Tree structure-based classifier
- Models relationships between features and outputs

Decision Trees

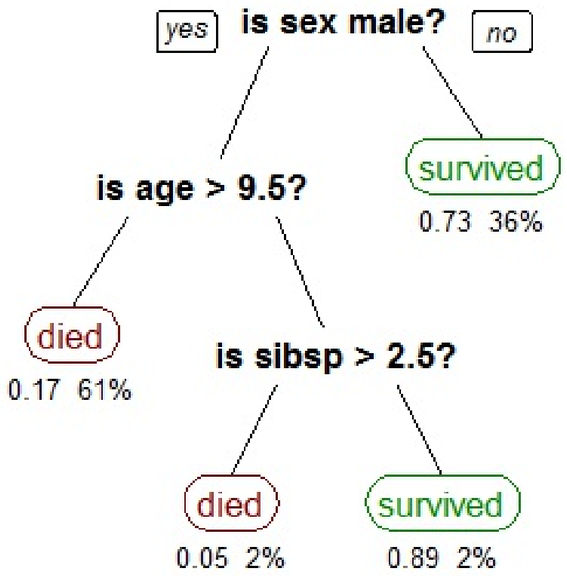
- Tree structure-based classifier
- Models relationships between features and outputs
- Easy to explain to users

Decision Trees

- Tree structure-based classifier
- Models relationships between features and outputs
- Easy to explain to users
- Can be turned into external representation beyond the classifier

Decision Trees

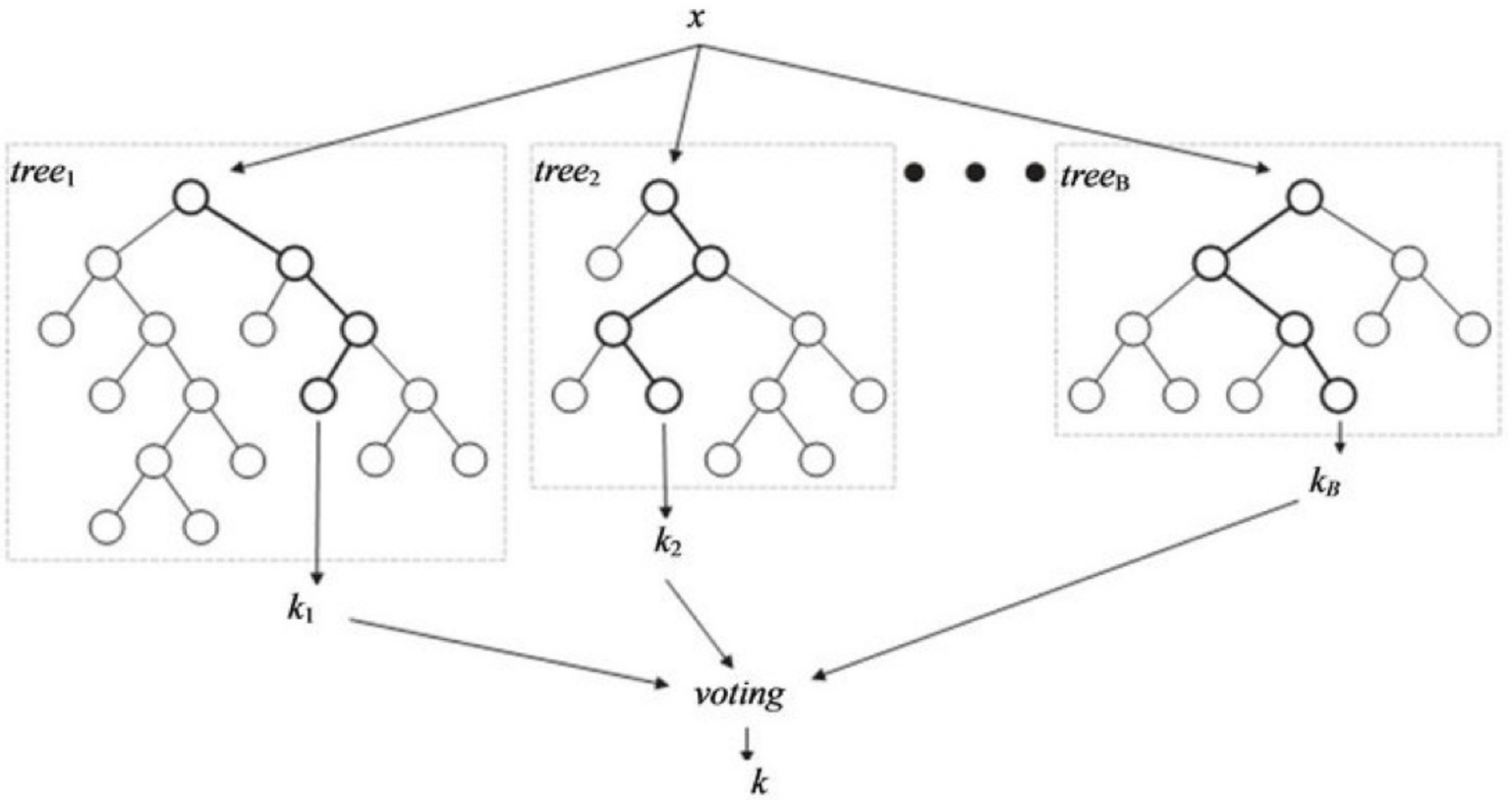
- Tree structure-based classifier
- Models relationships between features and outputs
- Easy to explain to users
- Can be turned into external representation beyond the classifier
- Supports both classification and regression



https://en.wikipedia.org/wiki/Decision_tree_learning

Useful classifier for most problems	Can be biased toward feature splits with several levels
Automated learning process	Easy to misfit the model
Supports numeric, nominal and missing data	Small changes in the training data can have been impact on decision logic
Works with large and small data sets	Large trees may be hard to interpret
Easily interpreted	
Efficient	

Random Forests

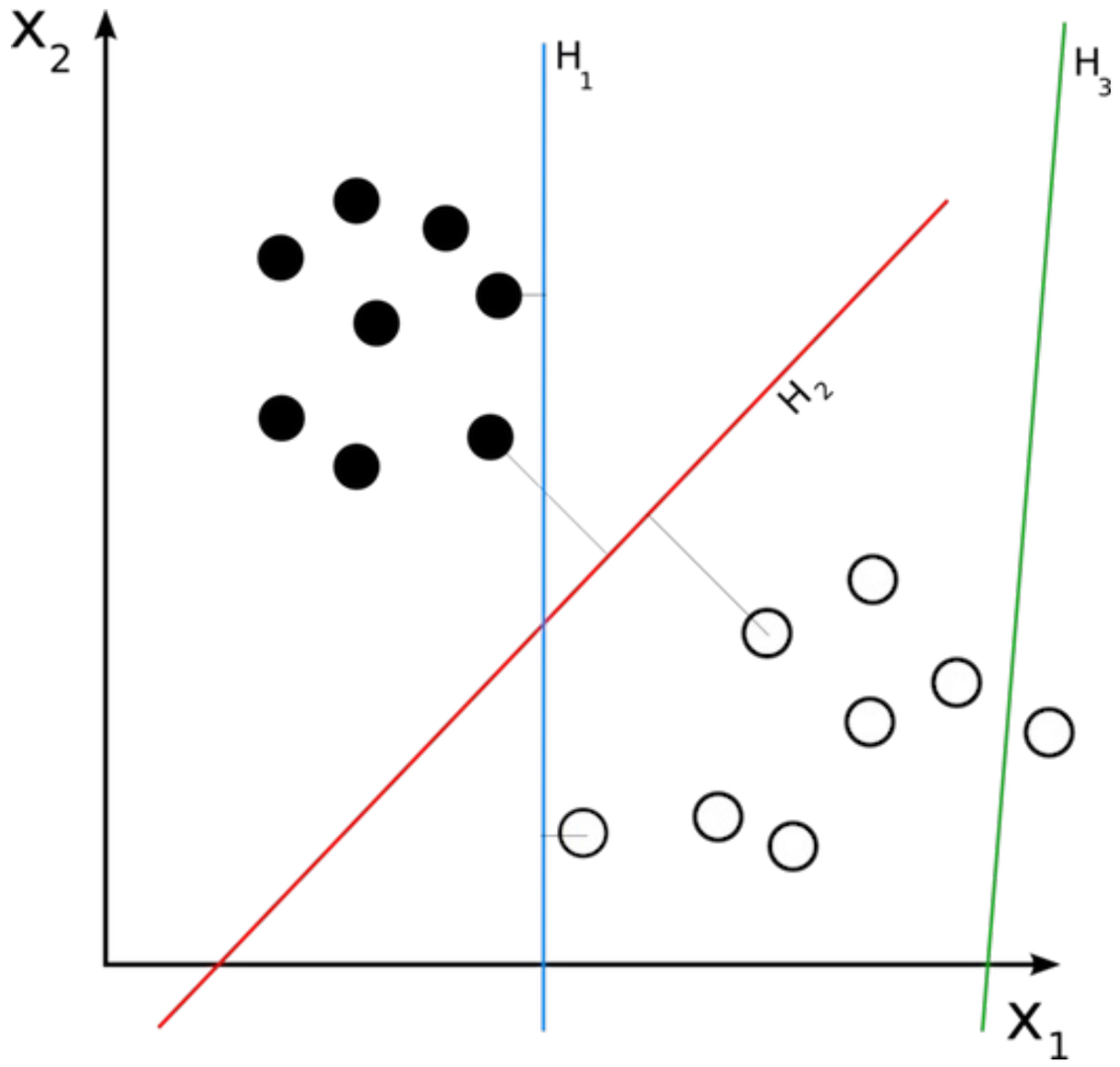


http://file.scirp.org/Html/6-9101686_31887.htm

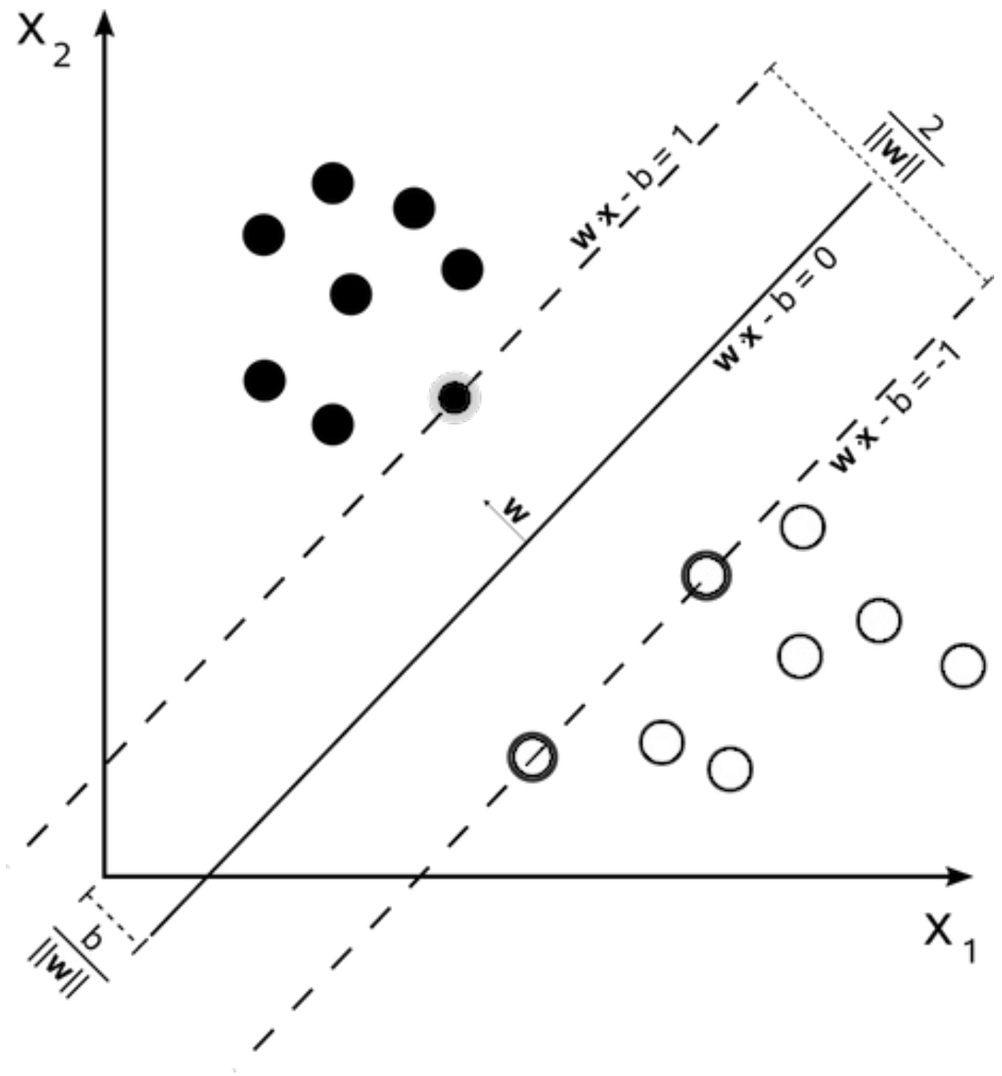
Support Vector Machines

"[a] support vector machine constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks"

Source: https://en.wikipedia.org/wiki/Support_vector_machine



https://en.wikipedia.org/wiki/Support_vector_machine



https://en.wikipedia.org/wiki/Support_vector_machine

Clustering

Anomaly Detection

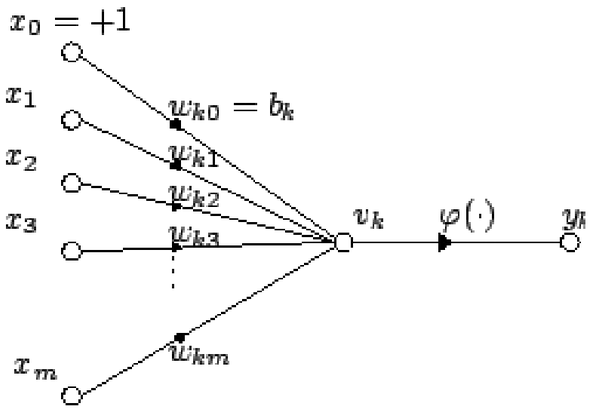
Modeling Neurons

Modeling Neurons

- Walter Pitts and Warren McCulloch in 1943

Modeling Neurons

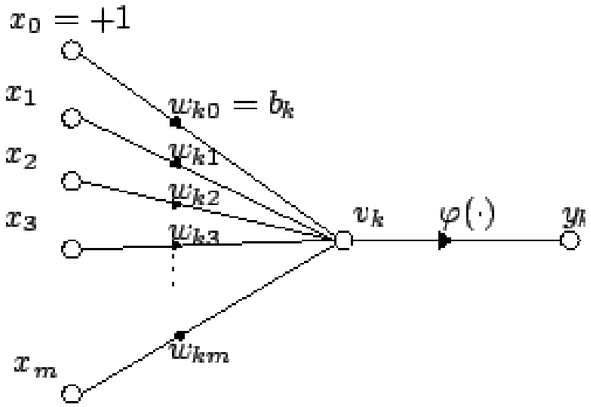
- Walter Pitts and Warren McCulloch in 1943



CC BY-SA 2.0, https://commons.wikimedia.org/w/index.php?curid=125222

Modeling Neurons

- Walter Pitts and Warren McCulloch in 1943




```
$$y_k = \phi \left( \sum_{j=0}^m
w_{kj}x_j\right)$$
```

CC BY-SA 2.0, https://commons.wikimedia.org/w/index.php?curid=125222

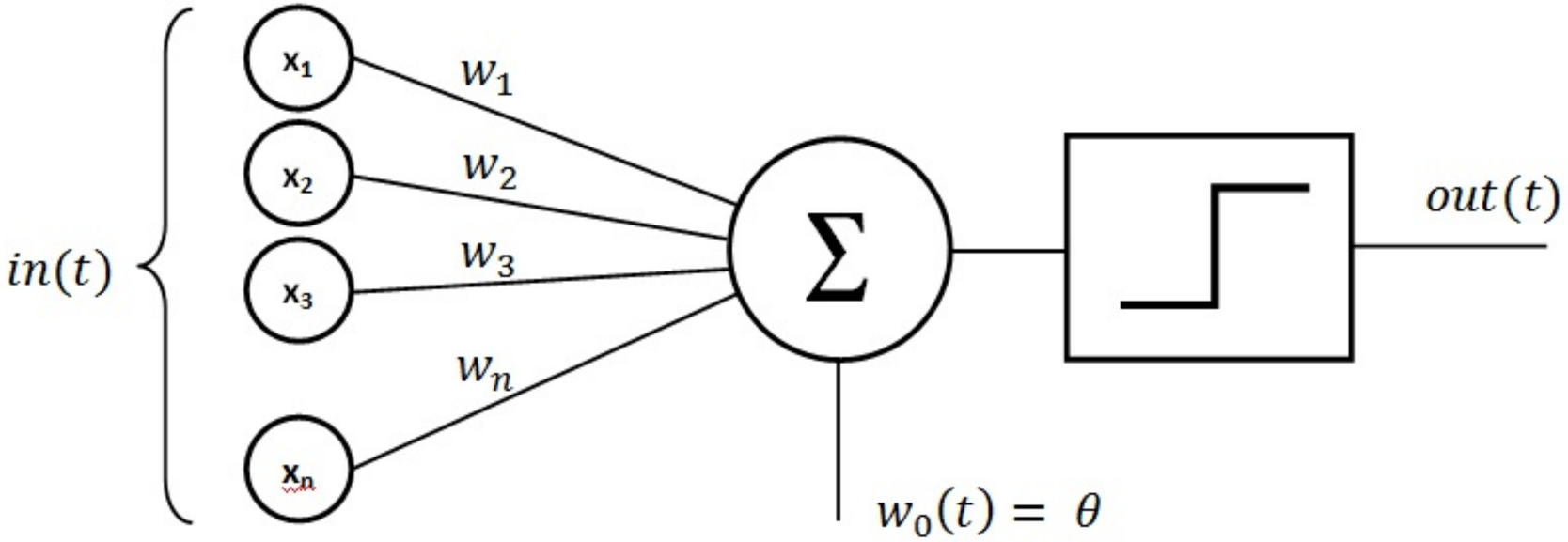
Perceptrons

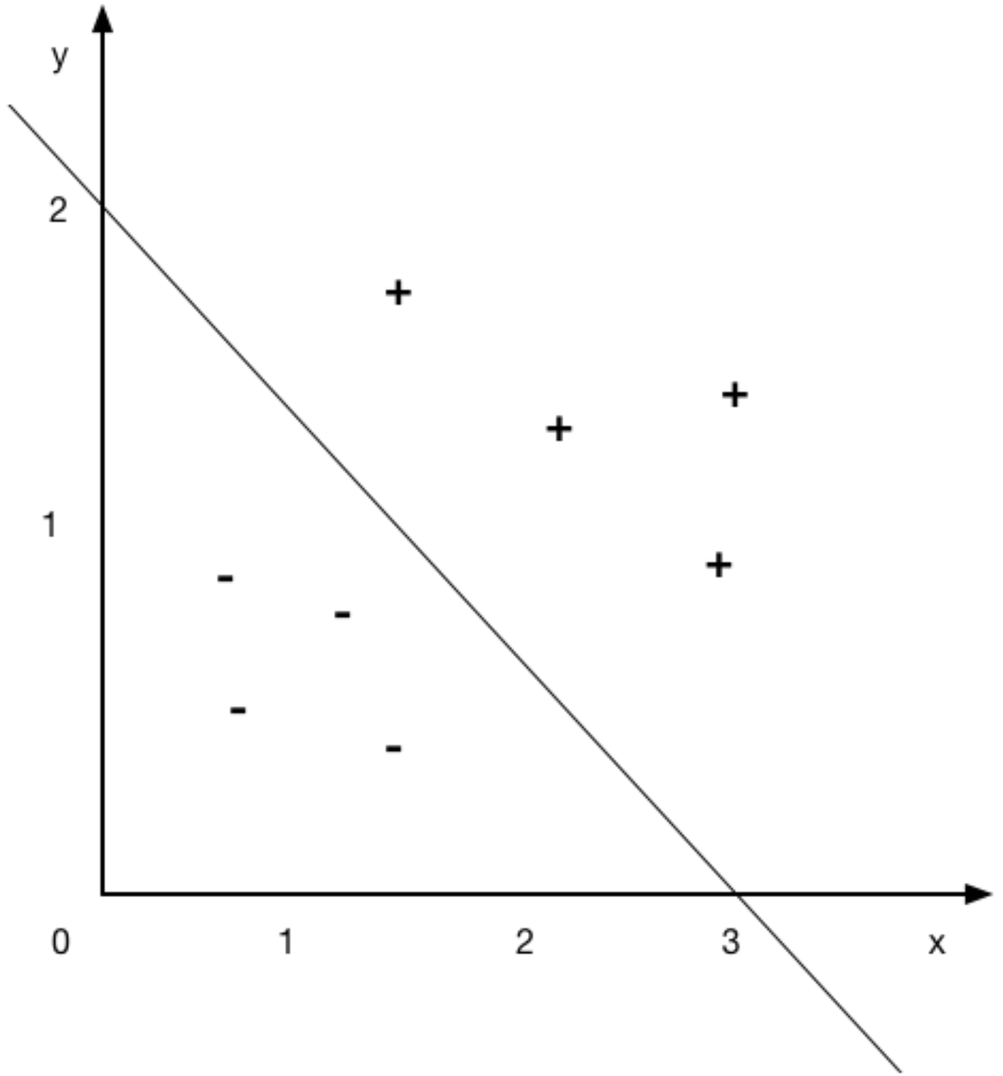
Perceptrons

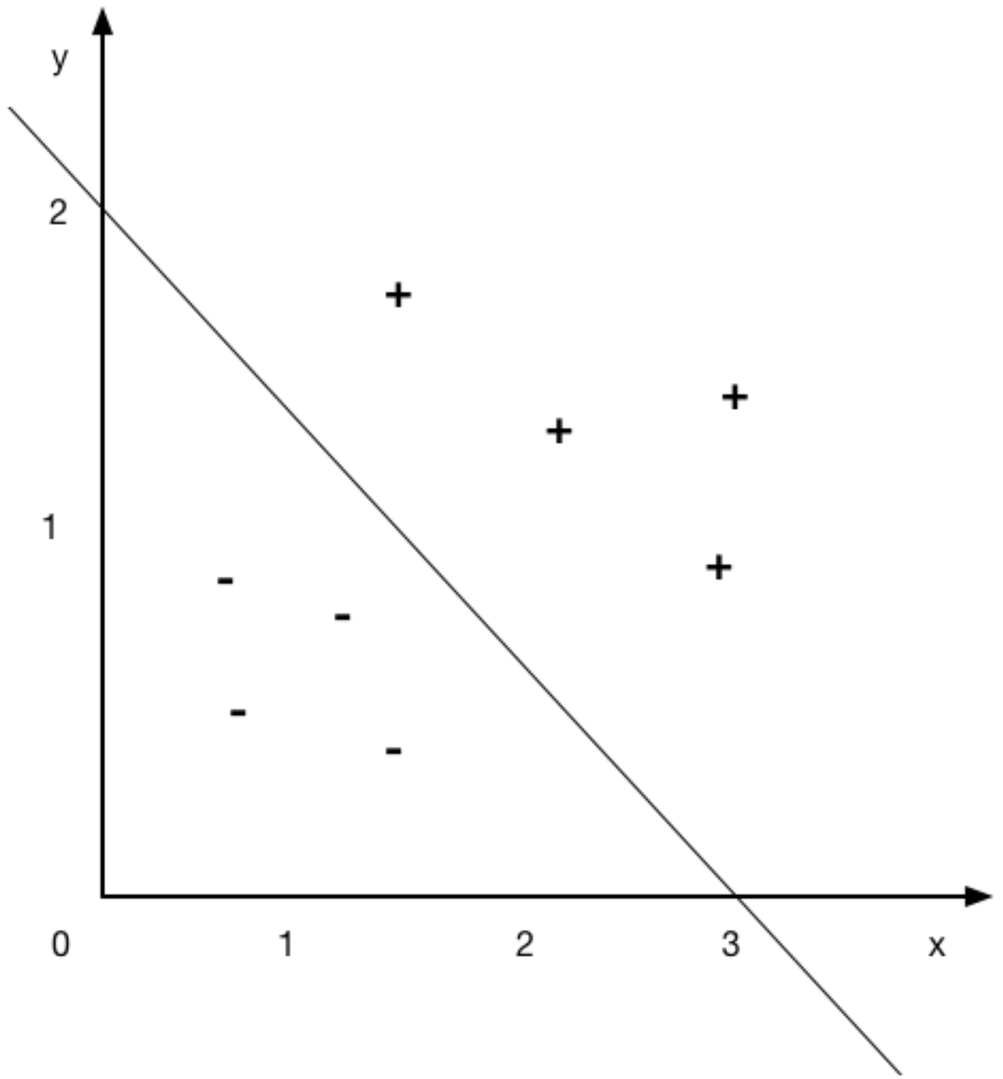
- Frank Rosenblatt in late 1950s

Perceptrons

- Frank Rosenblatt in late 1950s

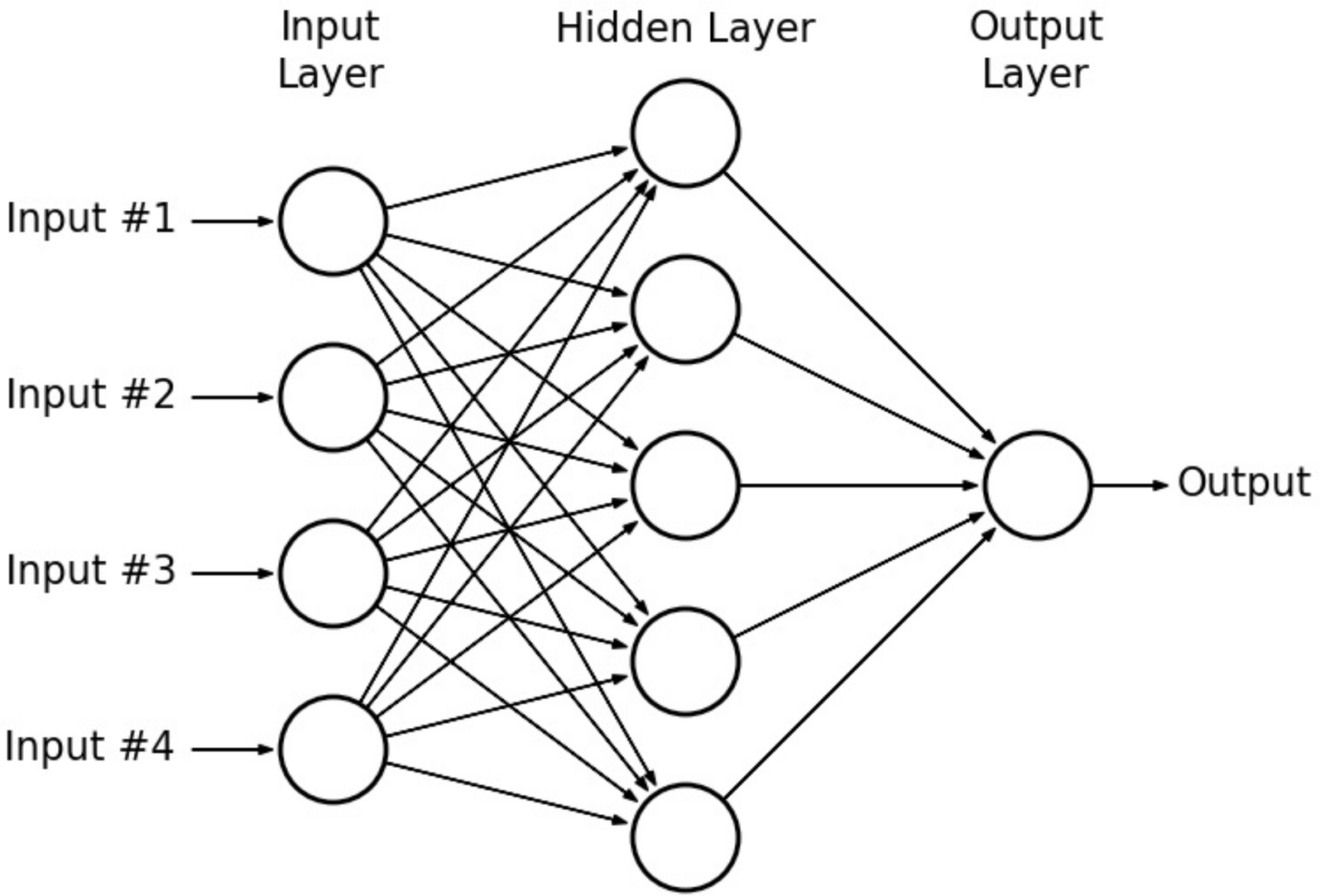



y
2
1
0
1
2
3
x
+
+
+
+
-
-
-
-



$$2x + 3y = 6$$


Old
Young
-
+
+
-
Male
Female


Input
Layer
Hidden Layer
Output
Layer
Input #1
Input #2
Input #3
Input #4
Output

Exercise

Infrastructure Demo

POD TTL models

- Linear regression - why it does not work
- TBATS
- Quantile Regression - not susceptible to utliers
- Prophet

POD TTL Model Selection/Goodness of Fit

- Testing, training and validation
- Model selection is conservative
- RMS, MAPE, AIC or new cost function depending on business use case_insensitive
- Visualization