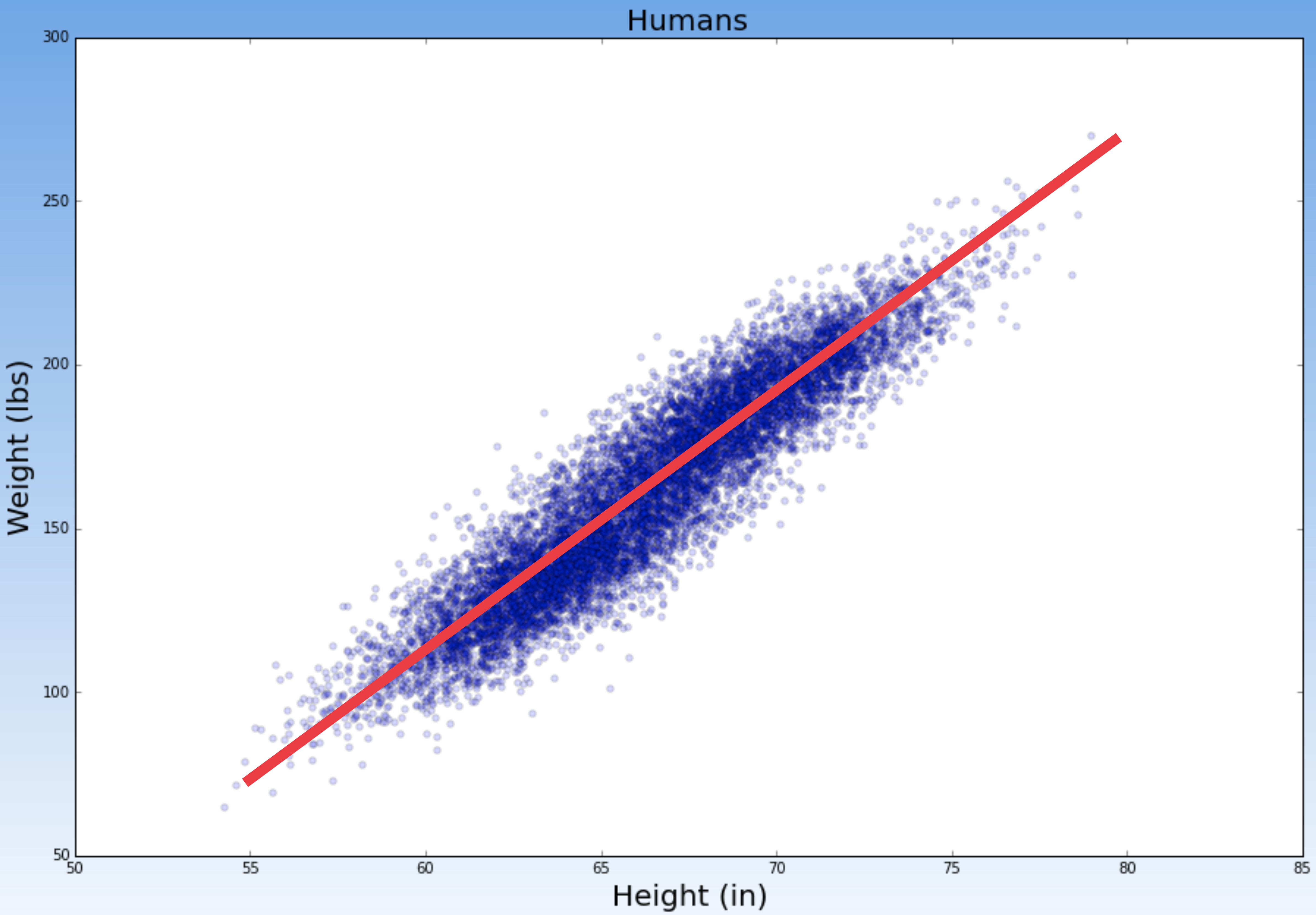
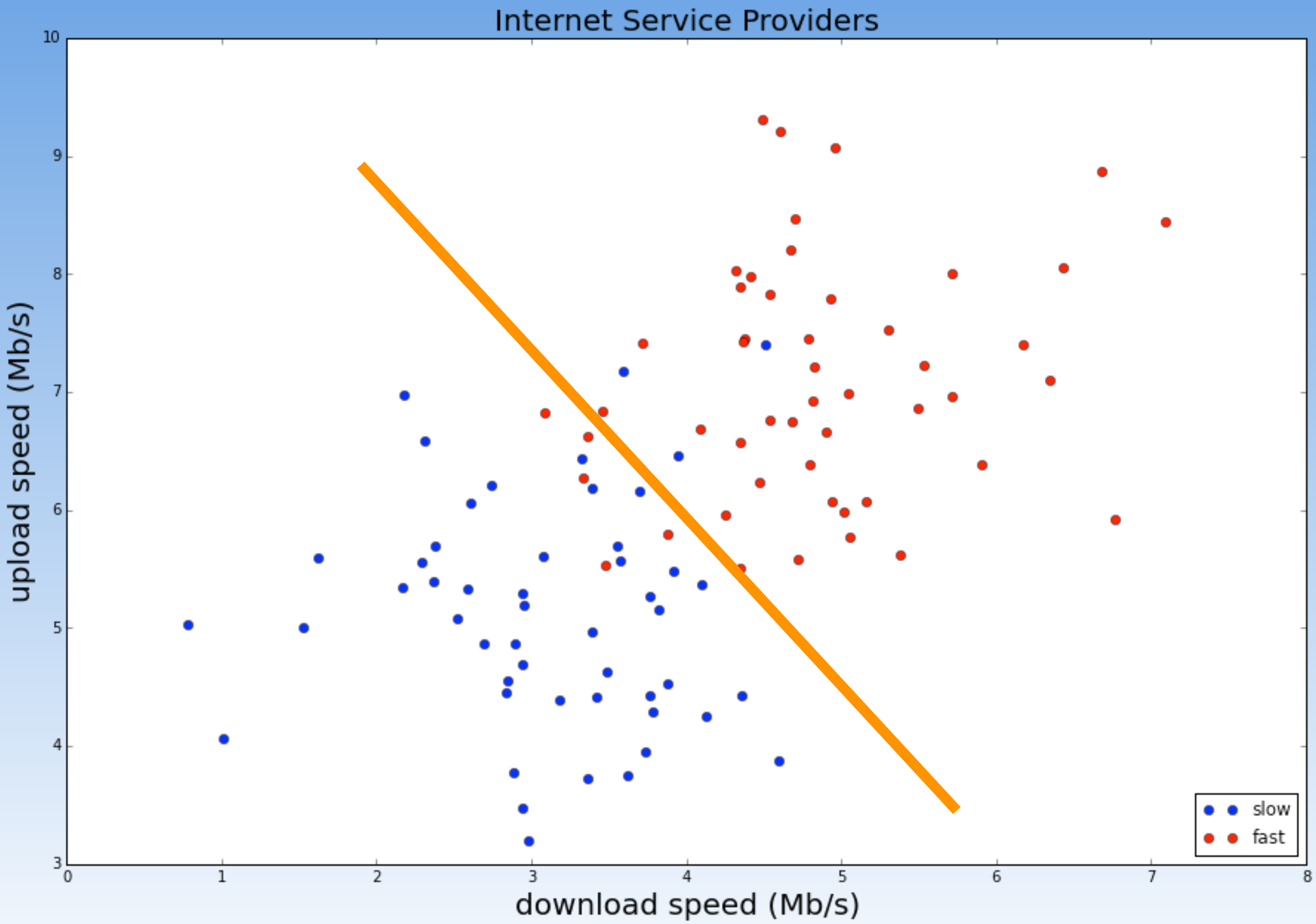
DataWknds.

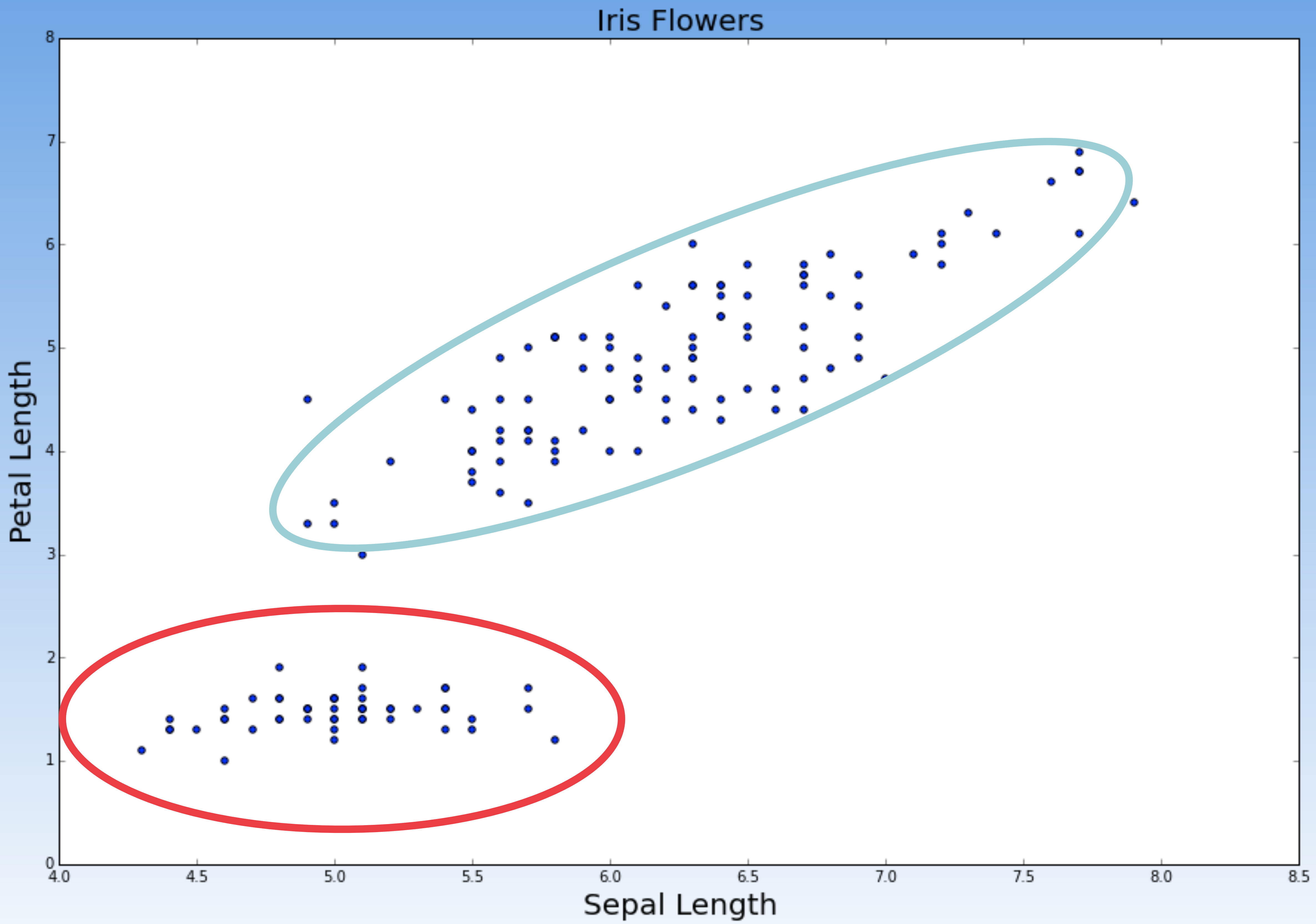
WELCOME BACK!


Humans
Weight (lbs)
Height (in)
300
250
200
150
100
50
50
55
60
65
70
75
80
85

Catalit LLC


Internet Service Providers
upload speed (Mb/s)
download speed (Mb/s)
10
9
8
7
6
5
4
3
0
1
2
3
4
5
6
7
8
● ● slow
● ● fast

Catalit LLC


Iris Flowers
Petal Length
Sepal Length
0
1
2
3
4
5
6
7
8
4.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
8.5

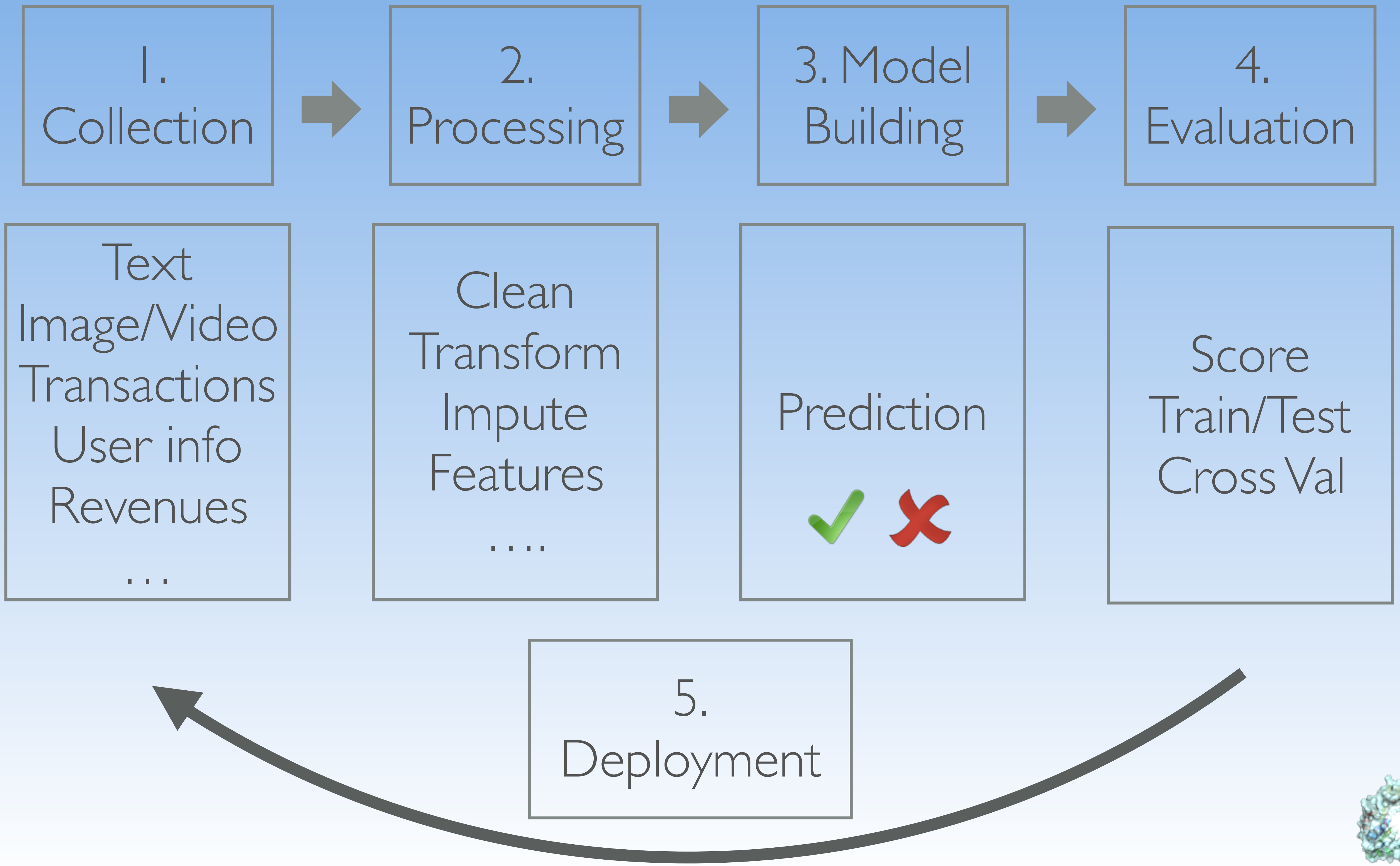
Catalit LLC

COMBINED

	CONTINUOUS	CATEGORICAL
SUPERVISED	REGRESSION	CLASSIFICATION
UNSUPERVISED	DIMENSION REDUCTION	CLUSTERING

Exploration

ML STEPS



1. Collection

2. Processing

3. Model Building

4. Evaluation

Text
Image/Video
Transactions
User info
Revenues
...

Clean
Transform
Impute
Features
....

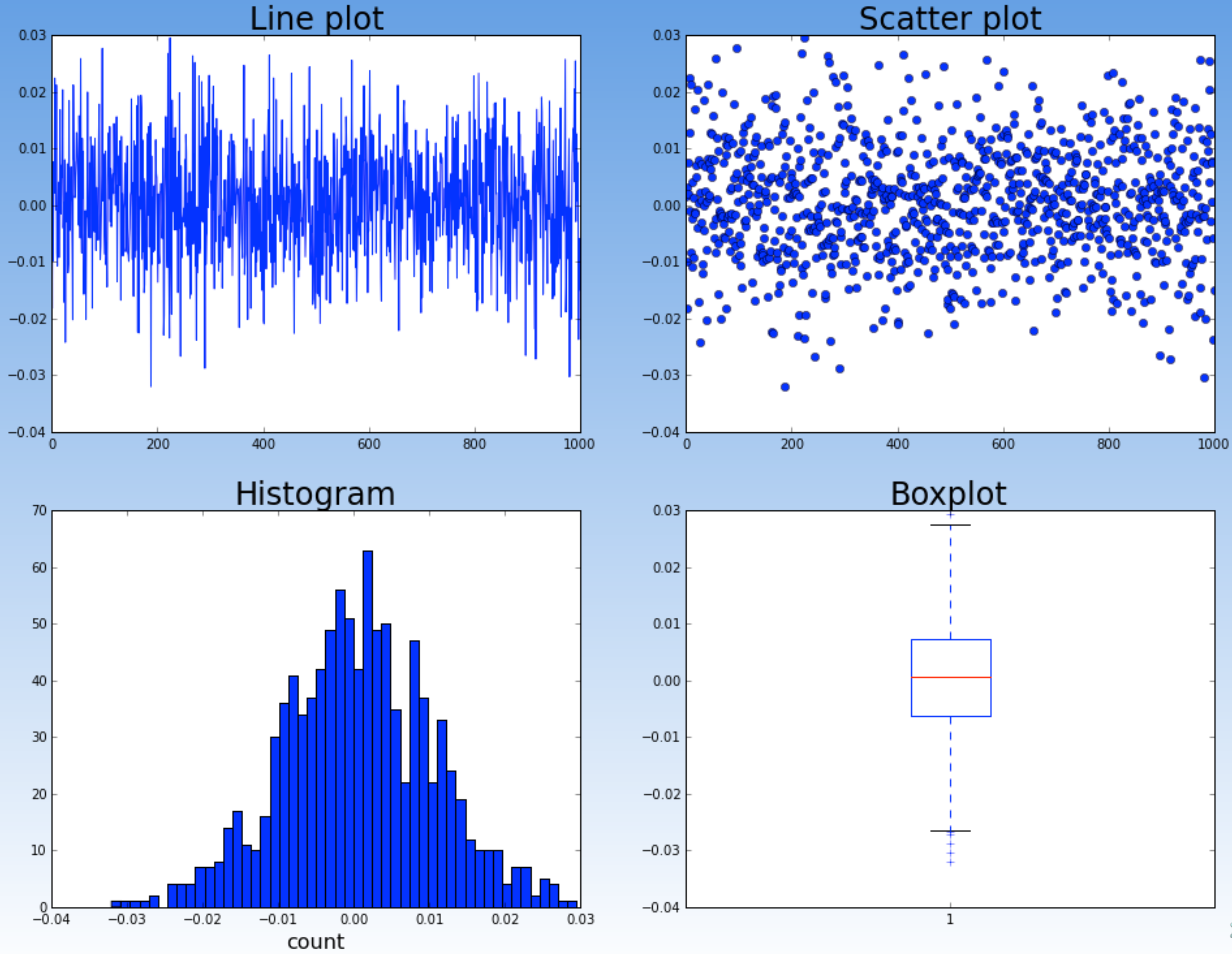
Prediction

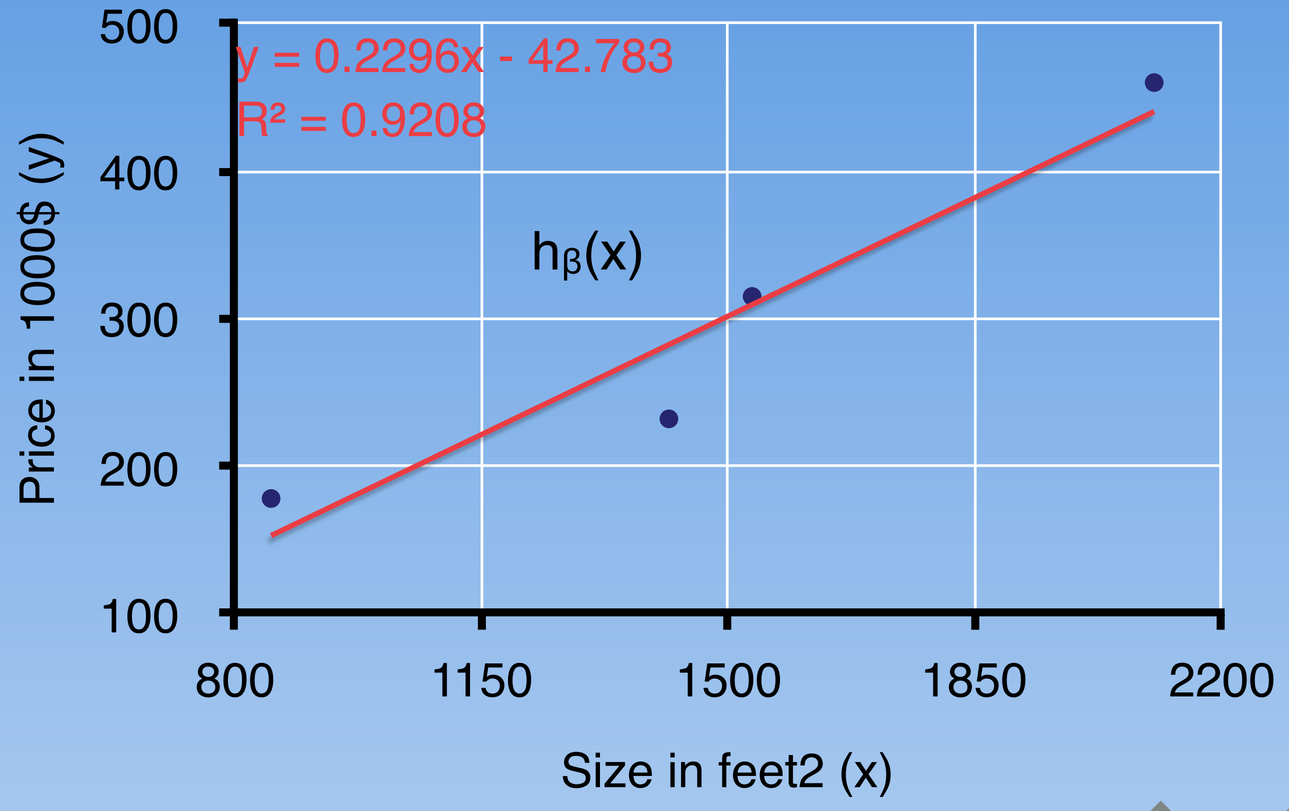
✓ ✗

Score
Train/Test
Cross Val

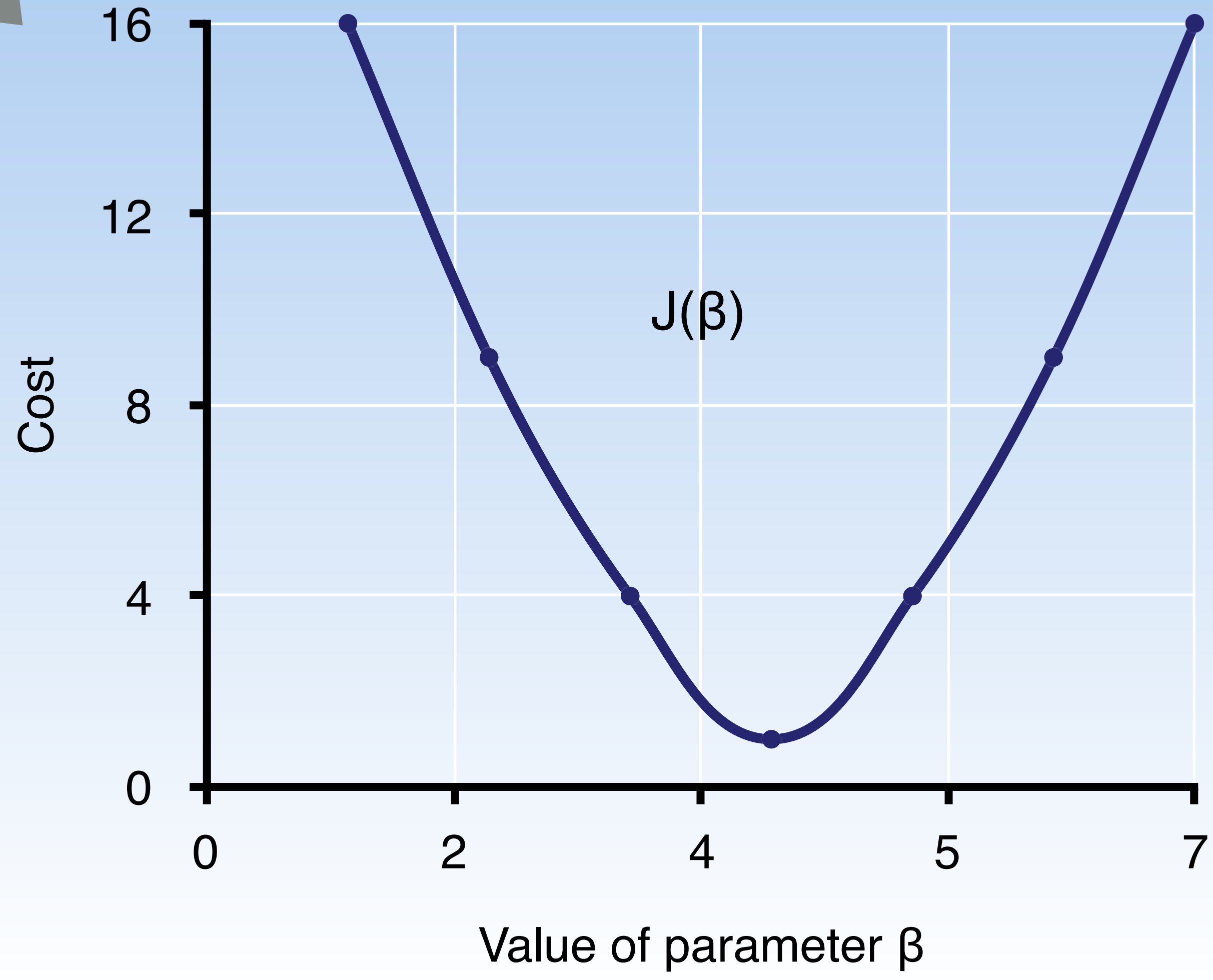
5. Deployment

Catalit LLC


Line plot
0.03
0.02
0.01
0.00
−0.01
−0.02
−0.03
−0.04
0
200
400
600
800
1000
Scatter plot
0.03
0.02
0.01
0.00
−0.01
−0.02
−0.03
−0.04
0
200
400
600
800
1000
Histogram
70
60
50
40
30
20
10
0
−0.04
−0.03
−0.02
−0.01
0.00
0.01
0.02
0.03
count
Boxplot
0.03
0.02
0.01
0.00
−0.01
−0.02
−0.03
−0.04
1



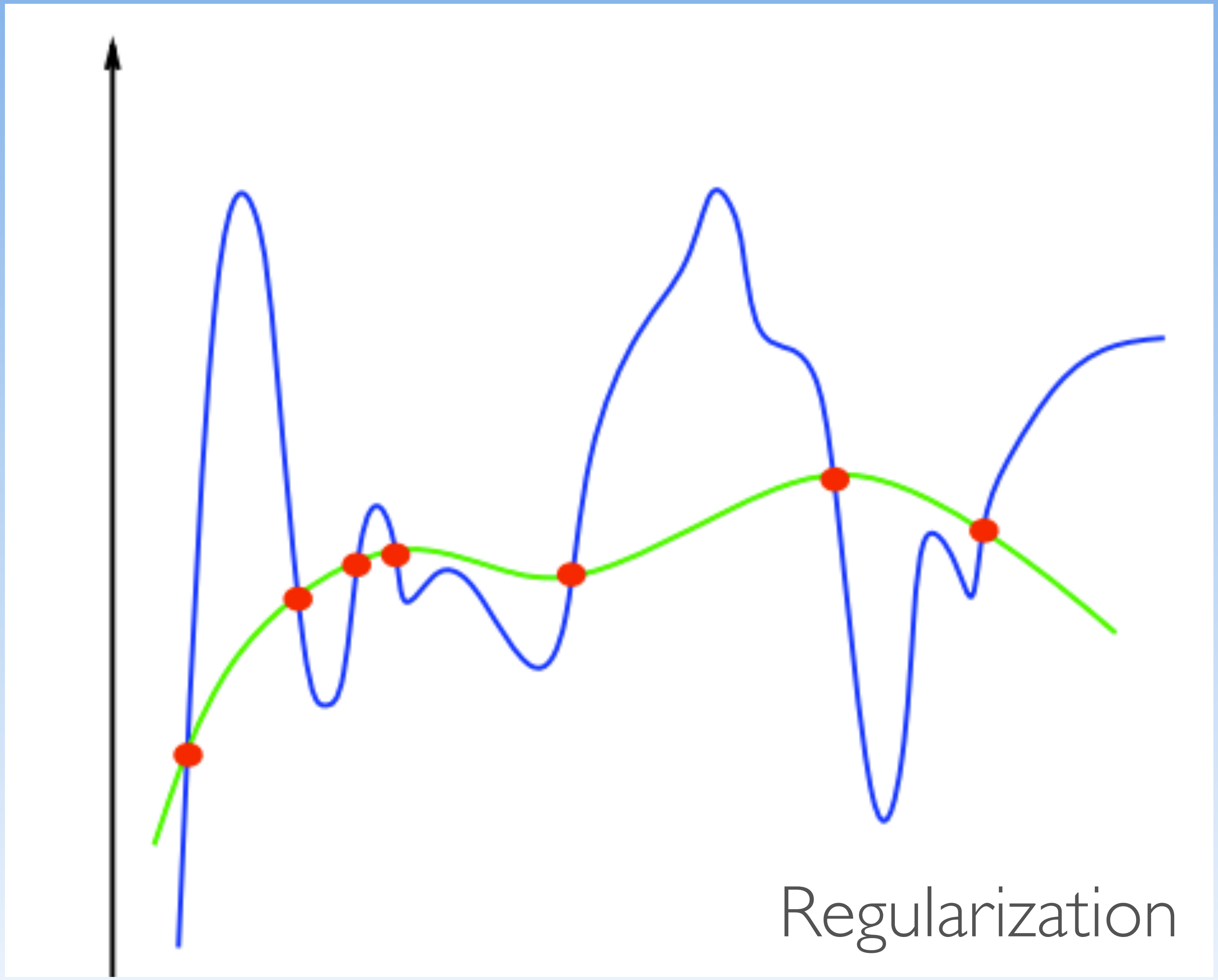
Define Hypothesis
Define Cost

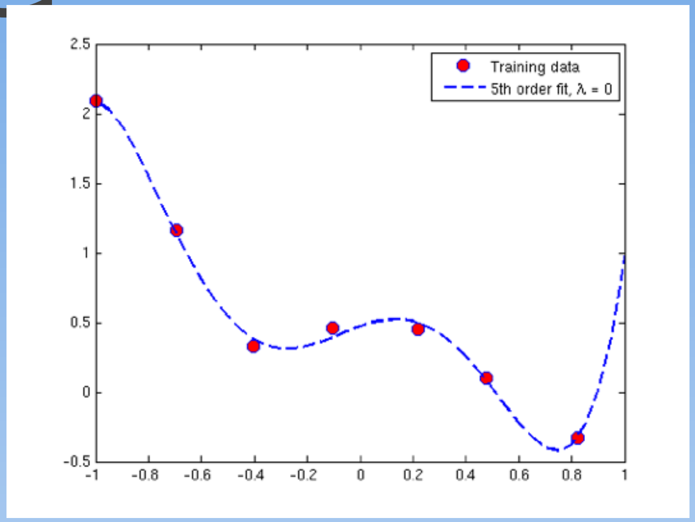


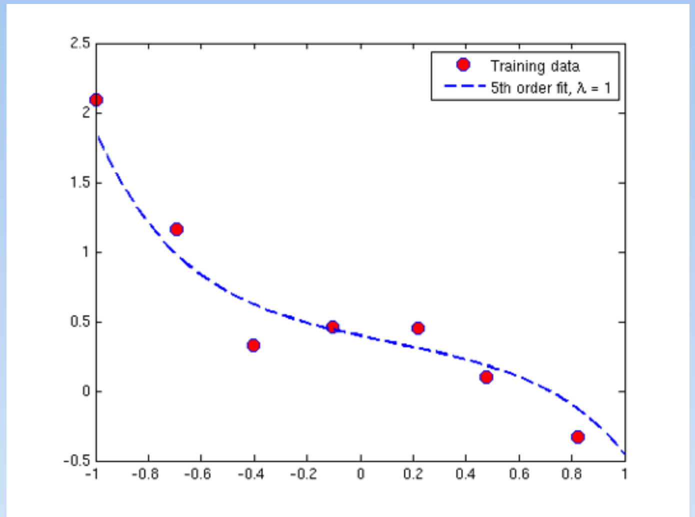
Minimize Cost

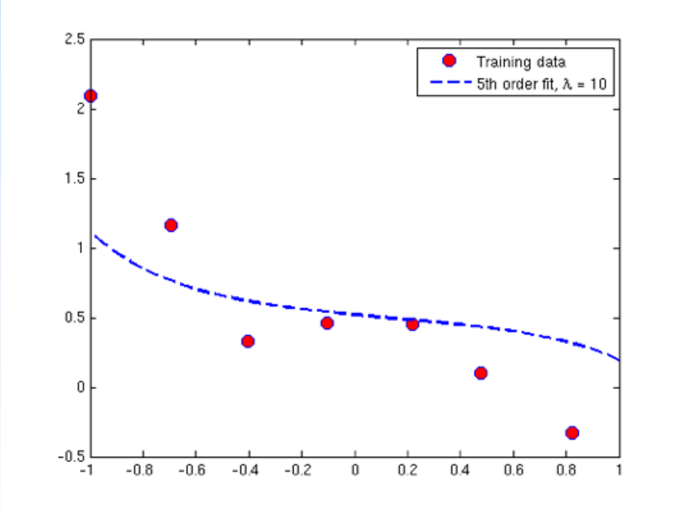
Catalit LLC

OVERFITTING



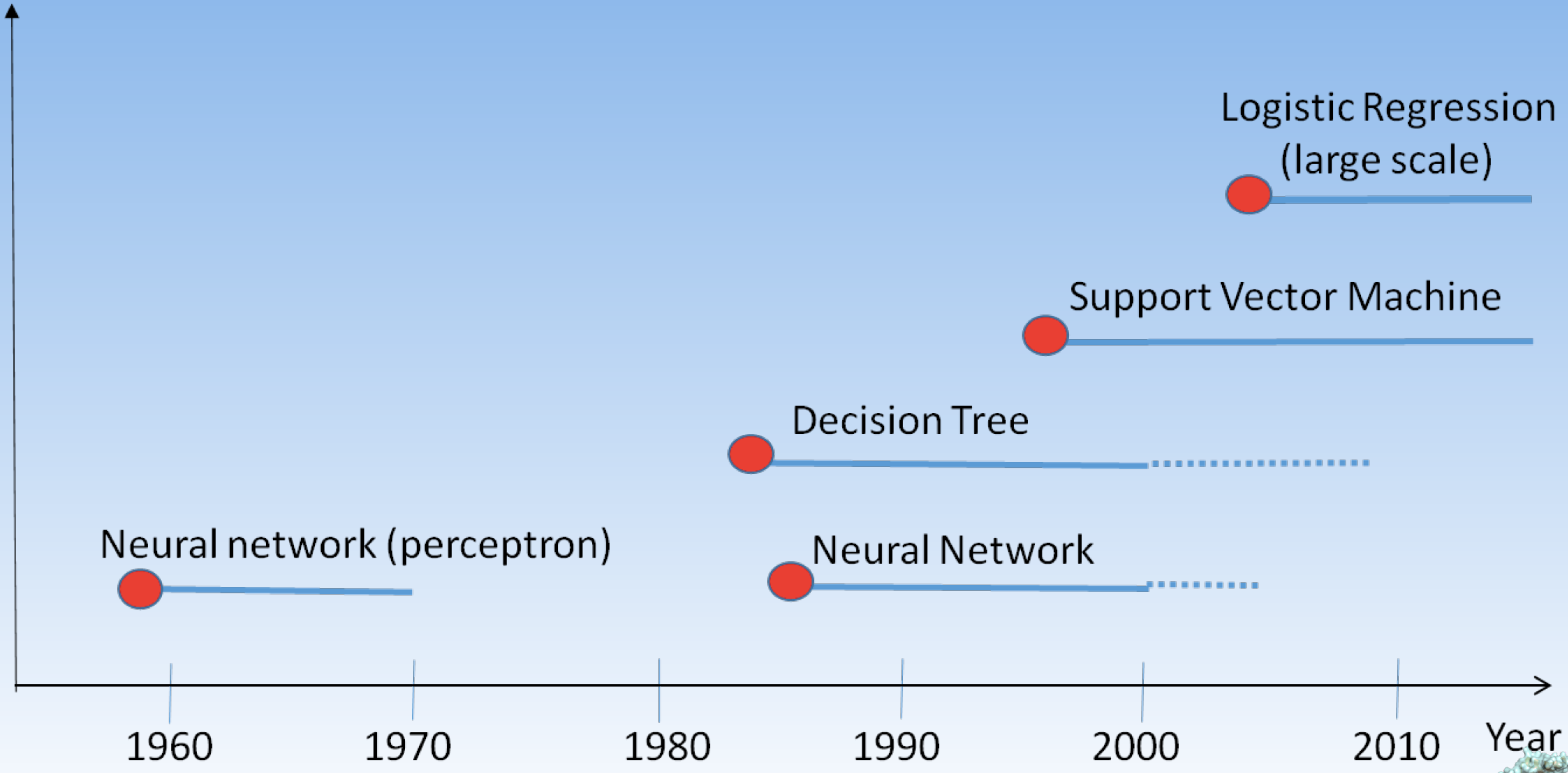






http://www.mit.edu/~9.520/spring12/slides/class02/class02.pdf

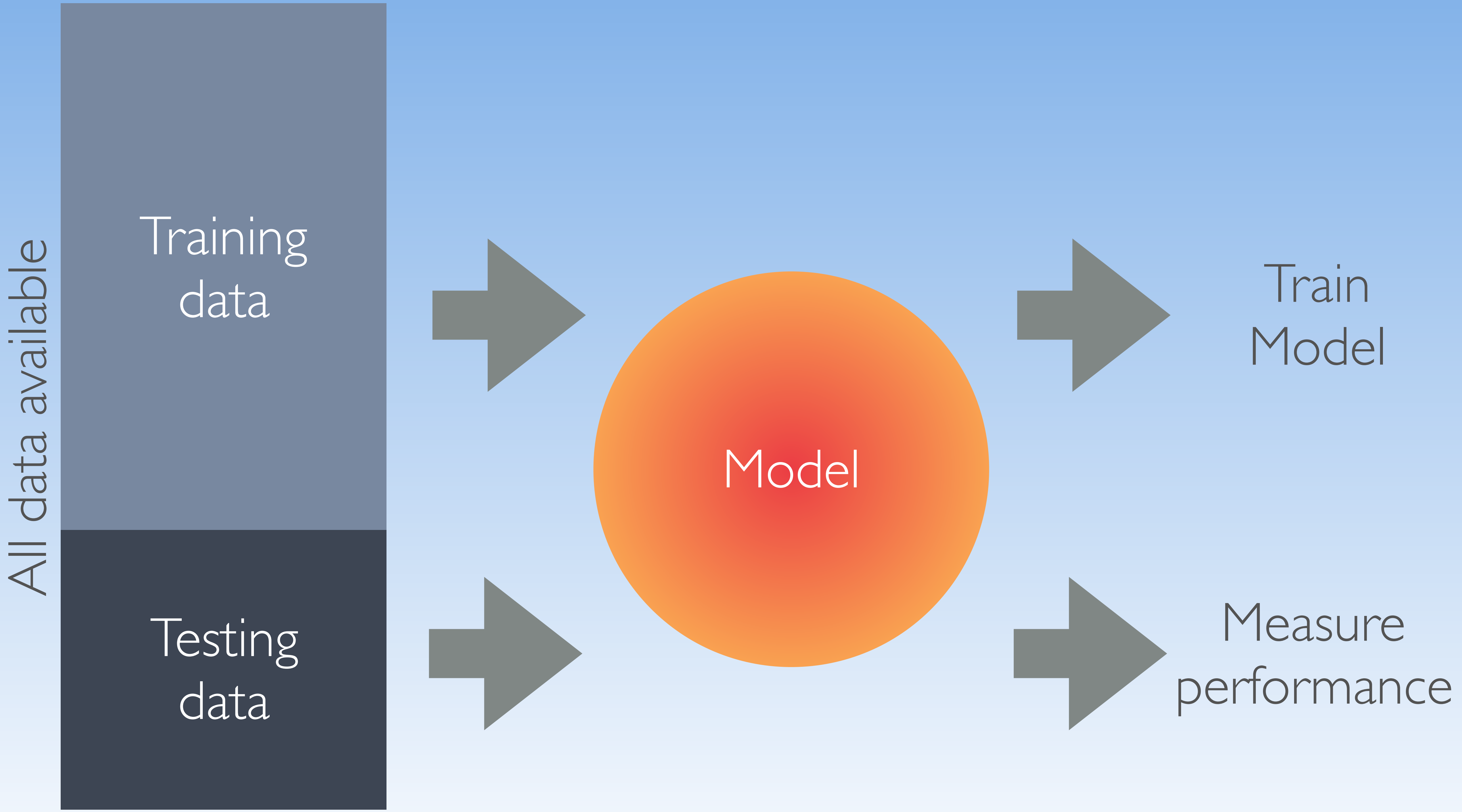
HISTORY OF SUPERVISED LEARNING



http://www.aboutdm.com/2013/04/history-of-machine-learning.html

Catalit LLC

TRAIN - TEST SPLIT



All data available

Training data

Testing data

Model

Train Model

Measure performance

Catalit LLC

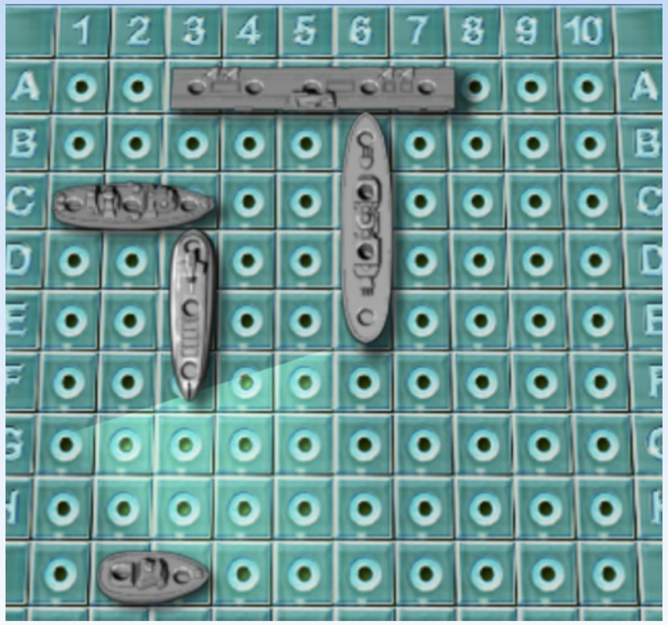
PRECISION - RECALL & ACCURACY

- **Precision:** When test is positive, how often is prediction correct?
  - TP / test yes
- **Recall:** When actual value is positive, how often is prediction correct?
  - TP / actual yes
- **Accuracy:** Overall, how often is it correct?
  - (TP + TN) / total

	Condition *Positive*	*Condition* *Negative*
Test *Positive*	*TRUE* *POSITIVE*	*FALSE* *POSITIVE*
Test *Negative*	*FALSE* *NEGATIVE*	*TRUE* *NEGATIVE*

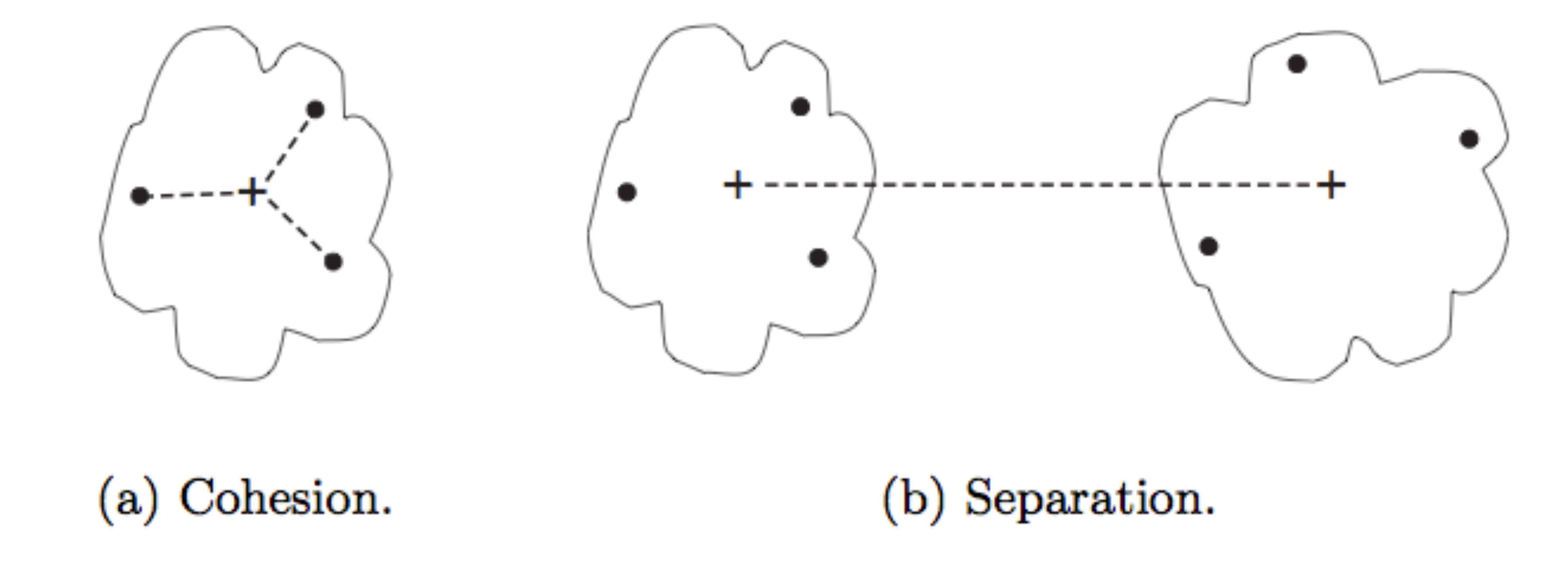
DISTANCE & SIMILARITY

	Age	Gender	Annual Salary	Months in residence	Months in job	Current Debt
Client 1	23	M	$30,000	36	12	$5,000
Client 2	30	F	$45,000	12	12	$1,000
Client 3	19	M	$15,000	3	1	$10,000



Catalit LLC

CLUSTER VALIDATION



(a) Cohesion.

(b) Separation.

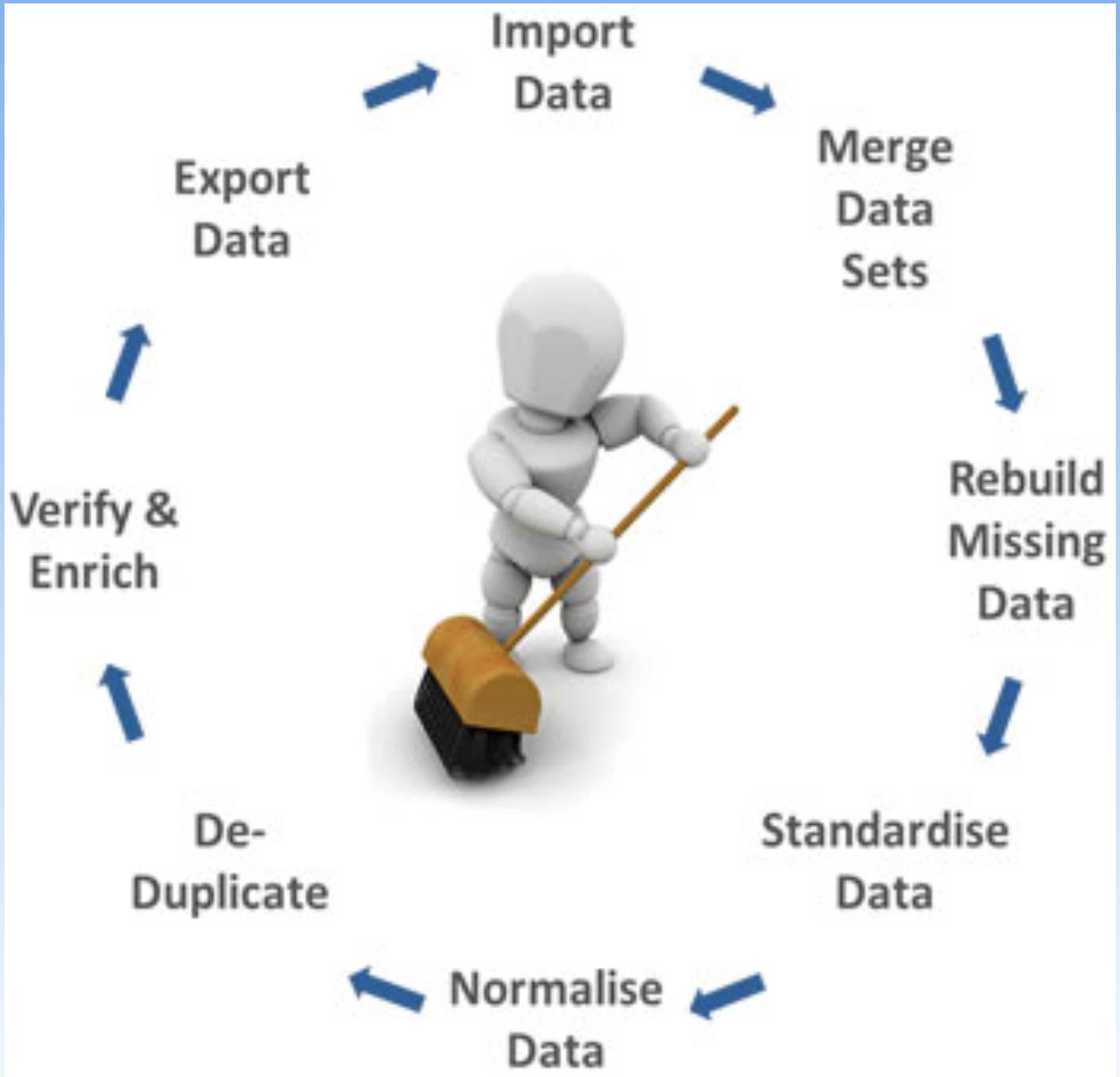
source: http://www-users.cs.umn.edu/~kumar/dmbook/ch8.pdf

Catalit LLC

DataWknds.

DATA CLEANING

DATA CLEANING



Other examples include:

Remove inconsistencies
Data type harmonization
Typos correction
Formatting (eg. timestamps)
Sorting

https://www.linkedin.com/pulse/big-data-101-cleaning-script-fennel-aurora

TIME SPENT

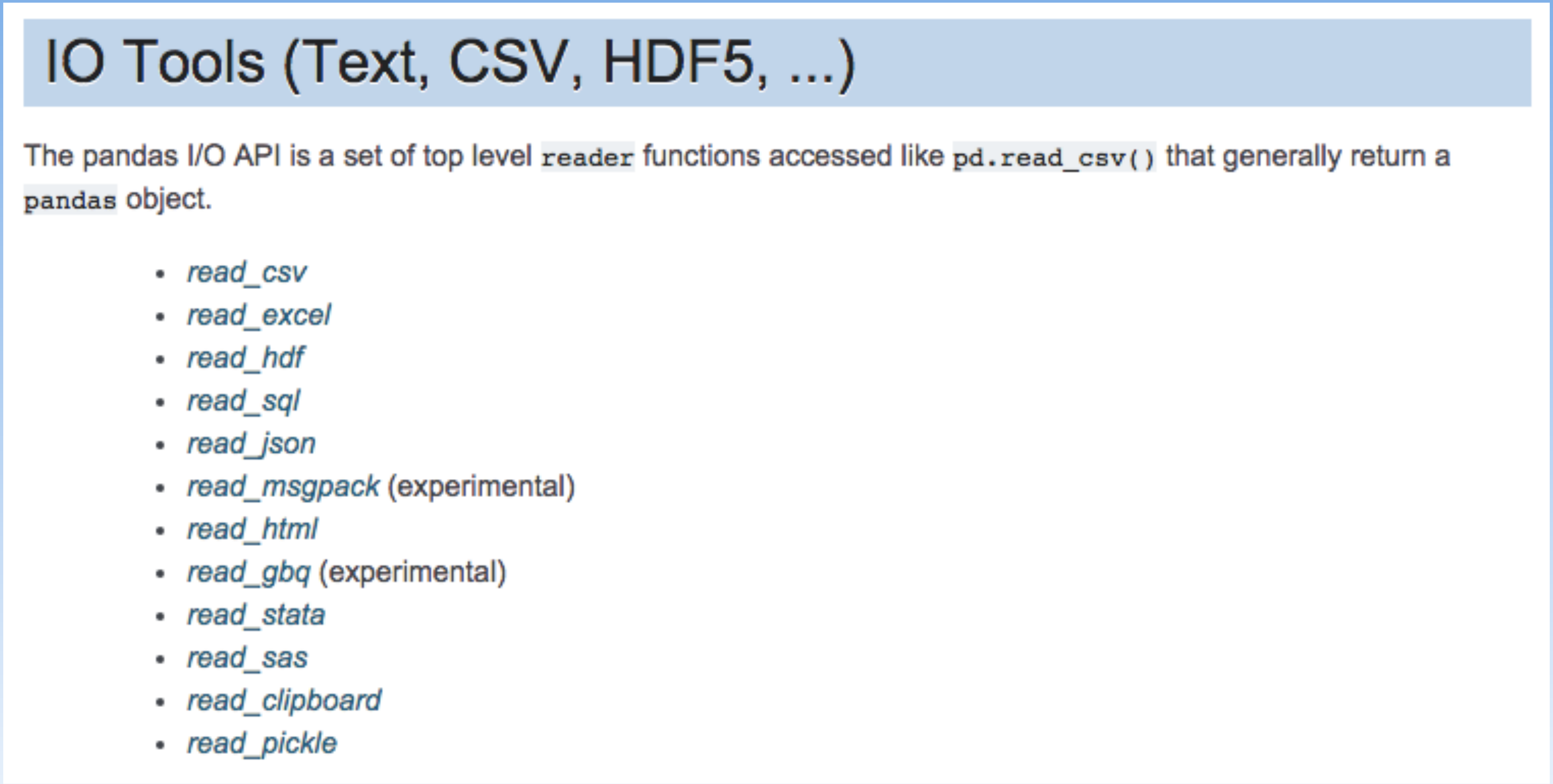
Obtain Data
Scrub Data
Explore

80%

Model Algorithms
iNterpret Results

20%

IMPORT DATA



IO Tools (Text, CSV, HDF5, ...)

The pandas I/O API is a set of top level `reader` functions accessed like `pd.read_csv()` that generally return a `pandas` object.

- *read_csv*
- *read_excel*
- *read_hdf*
- *read_sql*
- *read_json*
- *read_msgpack* (experimental)
- *read_html*
- *read_gbq* (experimental)
- *read_stata*
- *read_sas*
- *read_clipboard*
- *read_pickle*

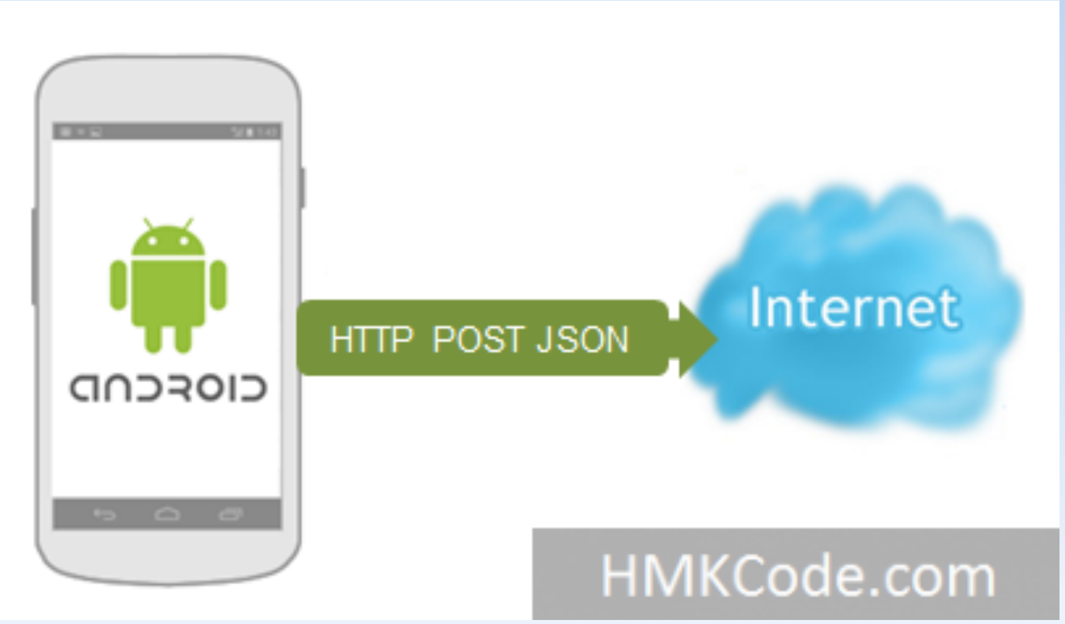
http://pandas.pydata.org/pandas-docs/stable/io.html

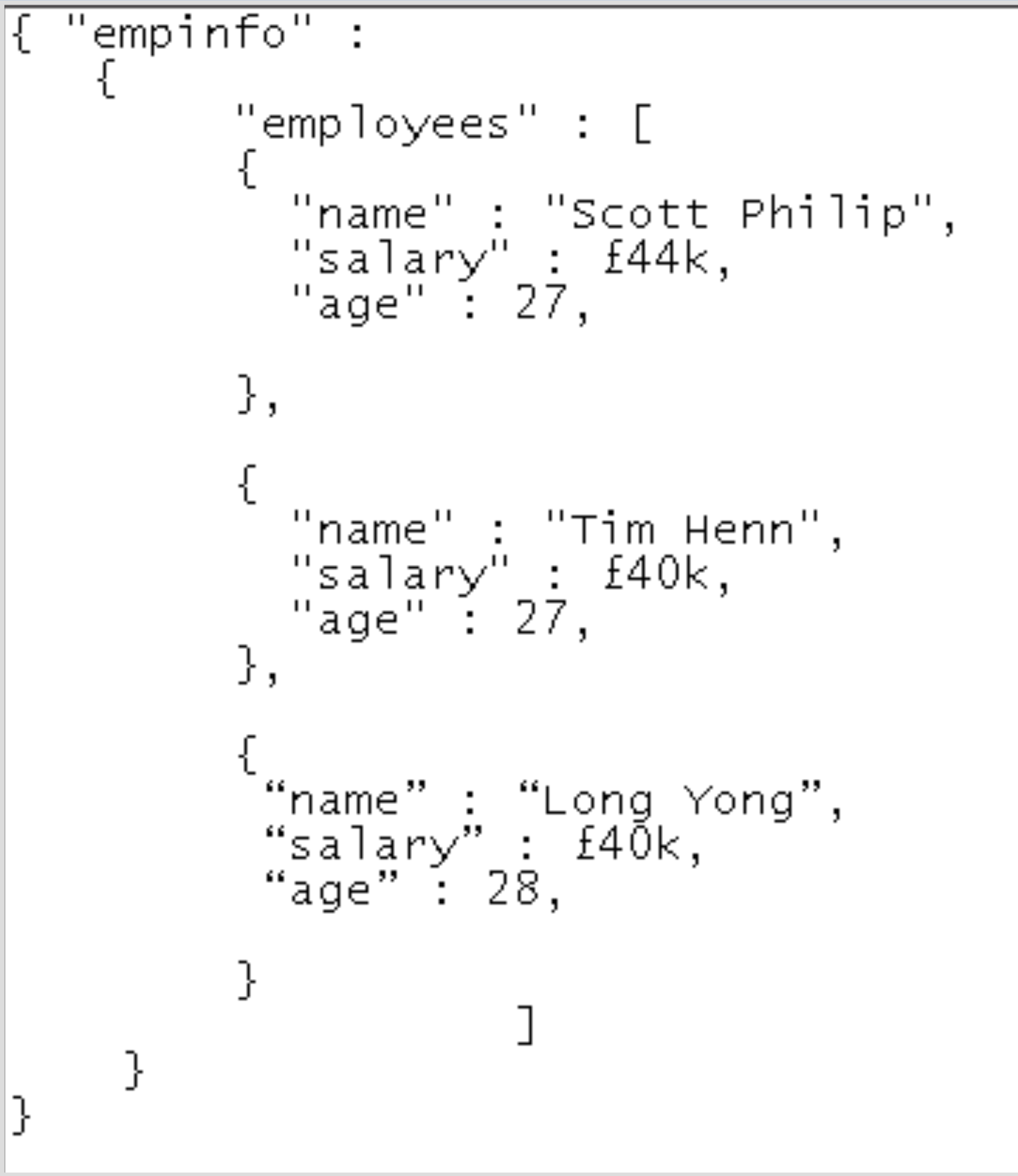
JSON

JSON (JavaScript Object Notation) is:

a lightweight data-interchange format

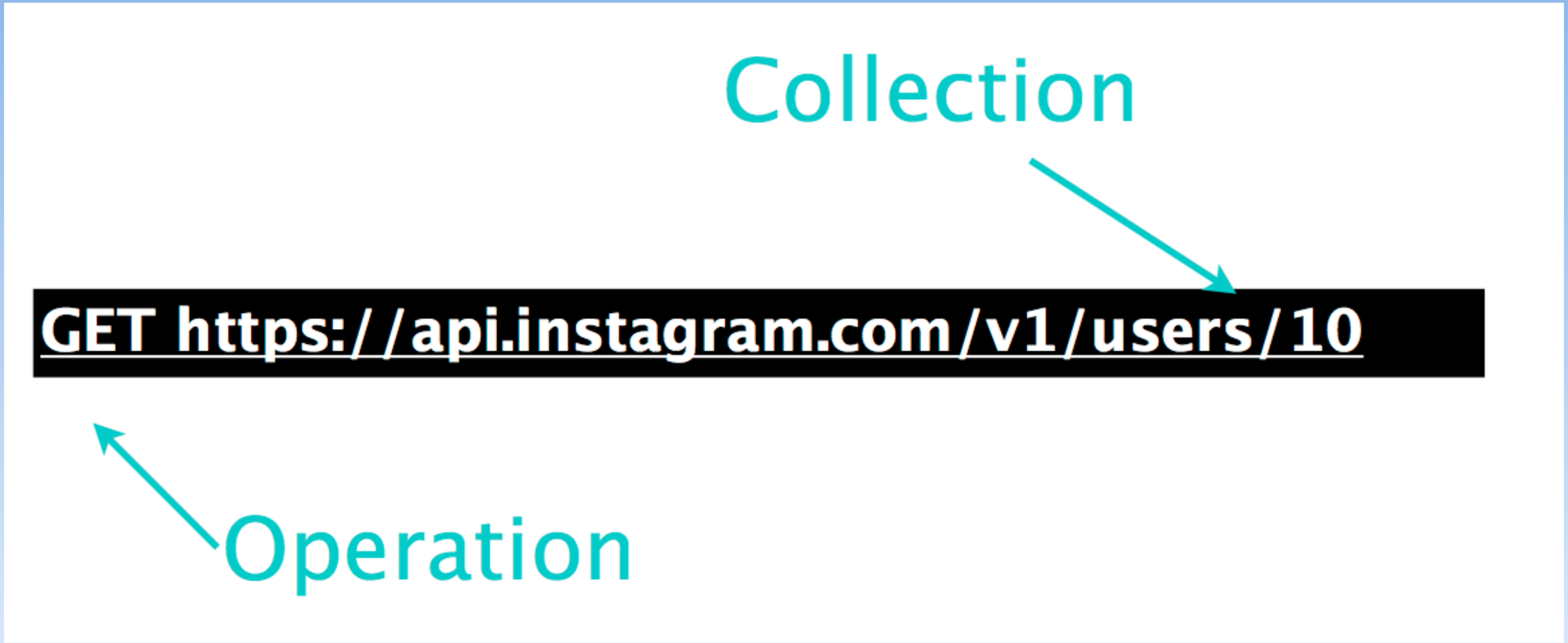
a string





```
{ "empinfo" :
    {
          "employees" : [
          {
            "name" : "Scott Philip",
            "salary" : £44k,
            "age" : 27,

          },

          {
            "name" : "Tim Henn",
            "salary" : £40k,
            "age" : 27,
          },

          {
           “name” : “Long Yong”,
           “salary” : £40k,
           “age” : 28,

          }
                    ]
     }
}
```

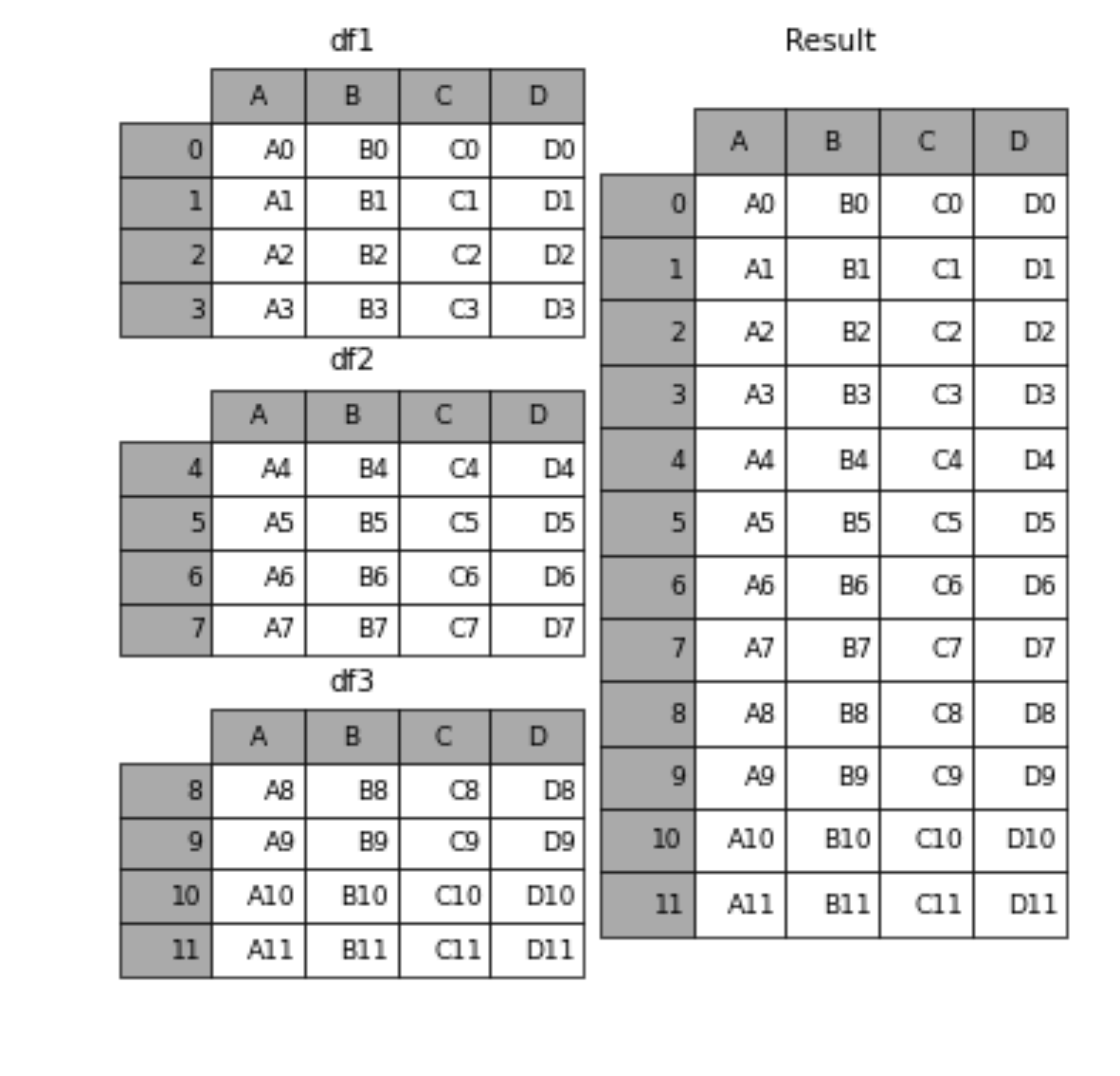
API



```
GET https://api.instagram.com/v1/users/10
```

http://www.pythonapi.com/

CONCATENATE DATA



df1

	A	B	C	D
0	A0	B0	C0	D0
1	A1	B1	C1	D1
2	A2	B2	C2	D2
3	A3	B3	C3	D3

df2

	A	B	C	D
4	A4	B4	C4	D4
5	A5	B5	C5	D5
6	A6	B6	C6	D6
7	A7	B7	C7	D7

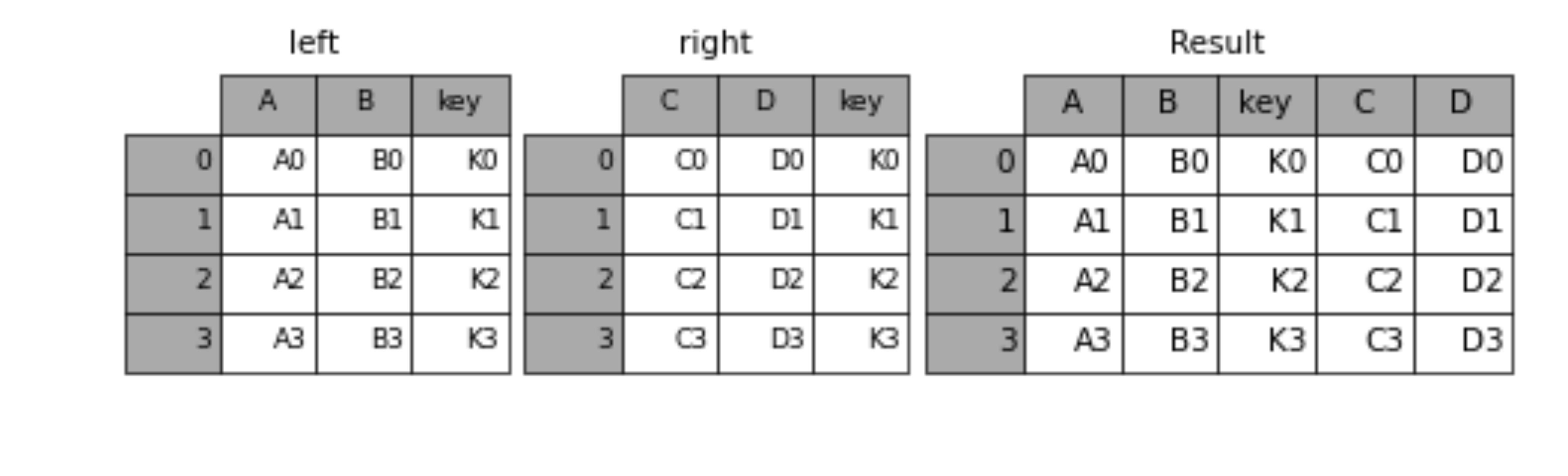
df3

	A	B	C	D
8	A8	B8	C8	D8
9	A9	B9	C9	D9
10	A10	B10	C10	D10
11	A11	B11	C11	D11

Result

	A	B	C	D
0	A0	B0	C0	D0
1	A1	B1	C1	D1
2	A2	B2	C2	D2
3	A3	B3	C3	D3
4	A4	B4	C4	D4
5	A5	B5	C5	D5
6	A6	B6	C6	D6
7	A7	B7	C7	D7
8	A8	B8	C8	D8
9	A9	B9	C9	D9
10	A10	B10	C10	D10
11	A11	B11	C11	D11

MERGE DATA



left

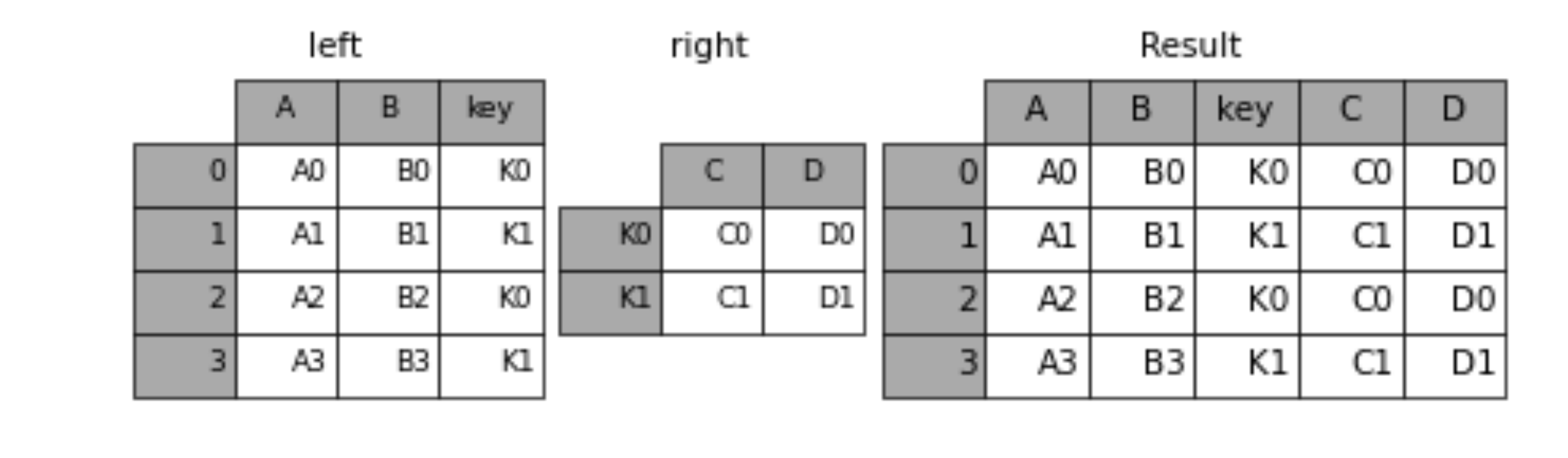
	A	B	key
0	A0	B0	K0
1	A1	B1	K1
2	A2	B2	K2
3	A3	B3	K3

right

	C	D	key
0	C0	D0	K0
1	C1	D1	K1
2	C2	D2	K2
3	C3	D3	K3

Result

	A	B	key	C	D
0	A0	B0	K0	C0	D0
1	A1	B1	K1	C1	D1
2	A2	B2	K2	C2	D2
3	A3	B3	K3	C3	D3



left

	A	B	key
0	A0	B0	K0
1	A1	B1	K1
2	A2	B2	K0
3	A3	B3	K1

right

	C	D
K0	C0	D0
K1	C1	D1

Result

	A	B	key	C	D
0	A0	B0	K0	C0	D0
1	A1	B1	K1	C1	D1
2	A2	B2	K0	C0	D0
3	A3	B3	K1	C1	D1

http://pandas.pydata.org/pandas-docs/stable/merging.html

REBUILD MISSING

- Missing at Random?

REBUILD MISSING

MCAR Missing completely at random	MAR Missing at random	MNAR Missing not at random
Missing value (y) neither depends on x nor y e.g.: some survey questions asked to fewer people	Missing value (y) depends on x, but not y e.g. Respondents in service occupations less likely to report income	The probability of a missing value depends on the variable that is missing e.g.: Respondents with high income less likely to report income

TECHNIQUES

- Imputation, Partial imputation
- Deletion, Partial deletion
- Analysis
- Interpolation

STANDARDIZATION

STANDARDIZATION

- Sep 12th, 2015   9/12/15   12-Sep-15
- USA   United States of America   EU   U.s.a.
- Mr   Mr.   Mister
- etc. etc.

NORMALIZATION

Catalit LLC

NORMALIZATION

- STANDARD
  - subtract mean
  - divide by std
- MINMAX
  - subtract min
  - divide by (max-min)

Catalit LLC

DE-DUPLICATE

FEATURES FROM TEXT

- Bag of Words approach:
  - Term Frequency (TF)
  - Inverse Document Frequency (IDF)

- NLP Approach
  - Stemming
  - Parts of Speech tagging
  - Named Entity Detection
  - Parsing

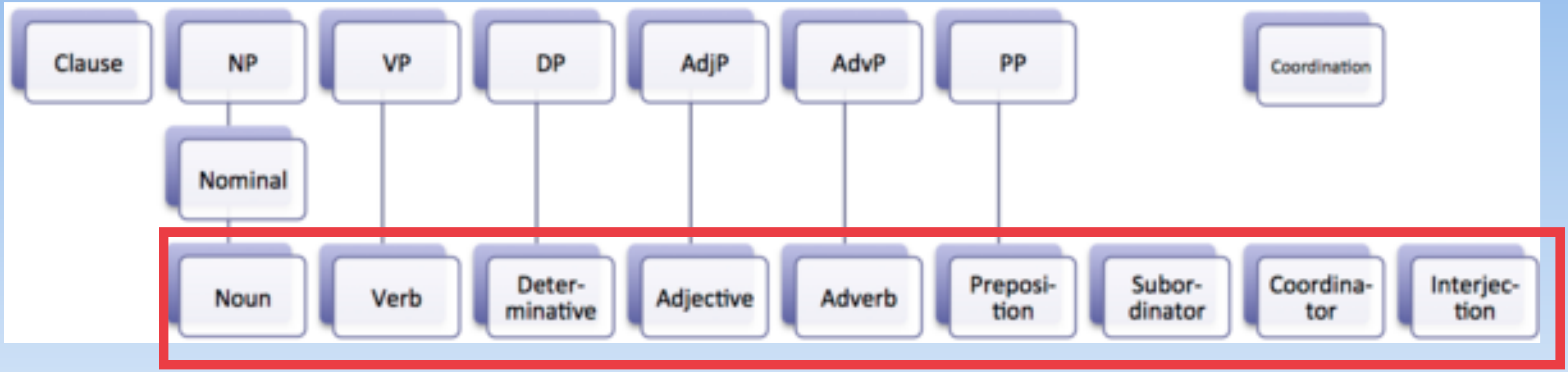
TFIDF

- Term frequency
  - Nterm/Nterms in document
- Document frequency
  - Ndocuments containing term/Ndocuments

STEMMING

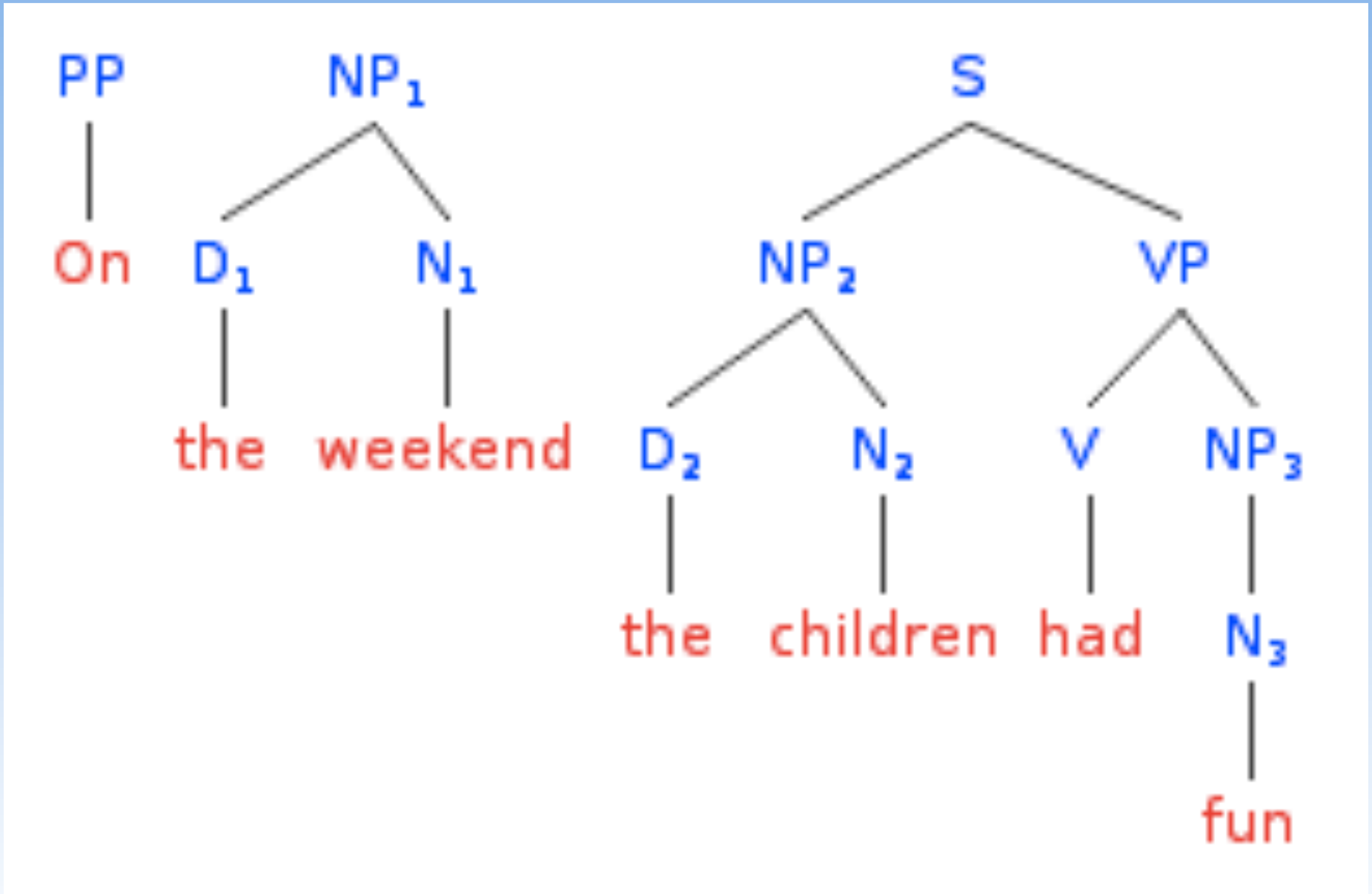
- science, scientist => scien
- swim, swimming, swimmer => swim
- Porter stemmer
- Very useful to reduce feature set size

PART OF SPEECH



Catalit LLC

PARSING


PP
NP_1
S
On
D_1
N_1
NP_2
VP
the
weekend
D_2
N_2
V
NP_3
the
children
had
N_3
fun

STANDARDIZATION

- you
  - u, ya, yo, yaaaa, yew, you, yoiu, youy, yoooooo, youz, yooouuuu
- Together
  - 2gether, tegetha, tgthr, togather, 2getha, 2gthr, togeter, togehter, t0gether, togeda, 2getter

LAB CLEANING + TEXT

Catalit LLC