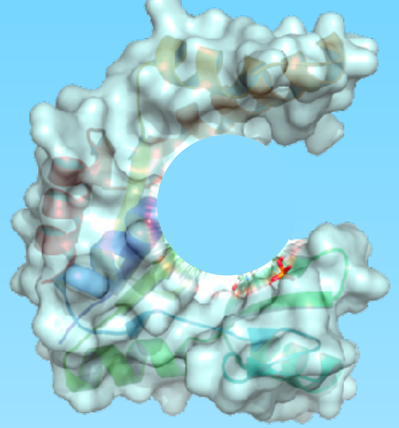
Data Weekends

Catalit LLC

ZERO TO DEEP LEARNING BOOTCAMP WELCOME!

Data Weekends



Catalit LLC

DAY 1

Zero to Deep Learning Bootcamp

QUICK CHECKS

- Setup (Anaconda)
- Skill Level
- Schedule
- Food
- Evenings (sign-up)
- Dinner tonight
- Slack
- Final project
- Questions?

Data Weekends

Catalit LLC

SCHEDULE

- 09:00 am Breakfast
- 09:30 am Class
- 11:00 am Break
- 11:15 am Class
- 12:45 pm Lunch
- 01:45 pm Class
- 03:15 pm Break
- 03:30 pm Class
- 05:00 pm End
- Evening activities

Data Weekends

Catalit LLC

ABOUT US

ZERO TO DEEP LEARNING
DEEP LEARNING FOR THE REST OF US

www.zerotodeeplearning.com

Data Weekends

www.dataweekends.com

CATALIT DATA SCIENCE

www.catalit.com

STARTUP BASECAMP

www.startupbasecamp.org

FINAL PROJECTS

- **By tonight: choose team mate**
- **By wednesday: decide topic, pitch**
- **Friday afternoon: work on project**

WHY HERE ?

WHY HERE ?

- Use Pandas, Scikit Learn, Keras and Tensorflow.
- Import, manipulate and visualize data.
- Apply Deep Learning algortihms to real problems.
- Develop an end-to-end Data Science project.
- Understand the fundamental theorethical concepts of Deep Learning.
- Learn how to use GPU and Cloud Computing services.
- See real-world applications of Machine Learning algorithms.

GOALS

- Intro Machine Learning
- Intro Deep Learning
- Intro Scikit-Learng, Keras, Tensorflow
- Identify use cases & Have fun

DAY 1

- **09:30 am Intro to ML**
- **11:15 am Regression**
- **12:45 pm Lunch**
- **01:45 pm Classification**
- **03:30 pm Clustering**

WHAT TO EXPECT

- Accelerated learning
- Interactive
- Fun / Discovery
- Take notes

Data Weekends

Catalit LLC

YOUR BRAIN

…A GREAT

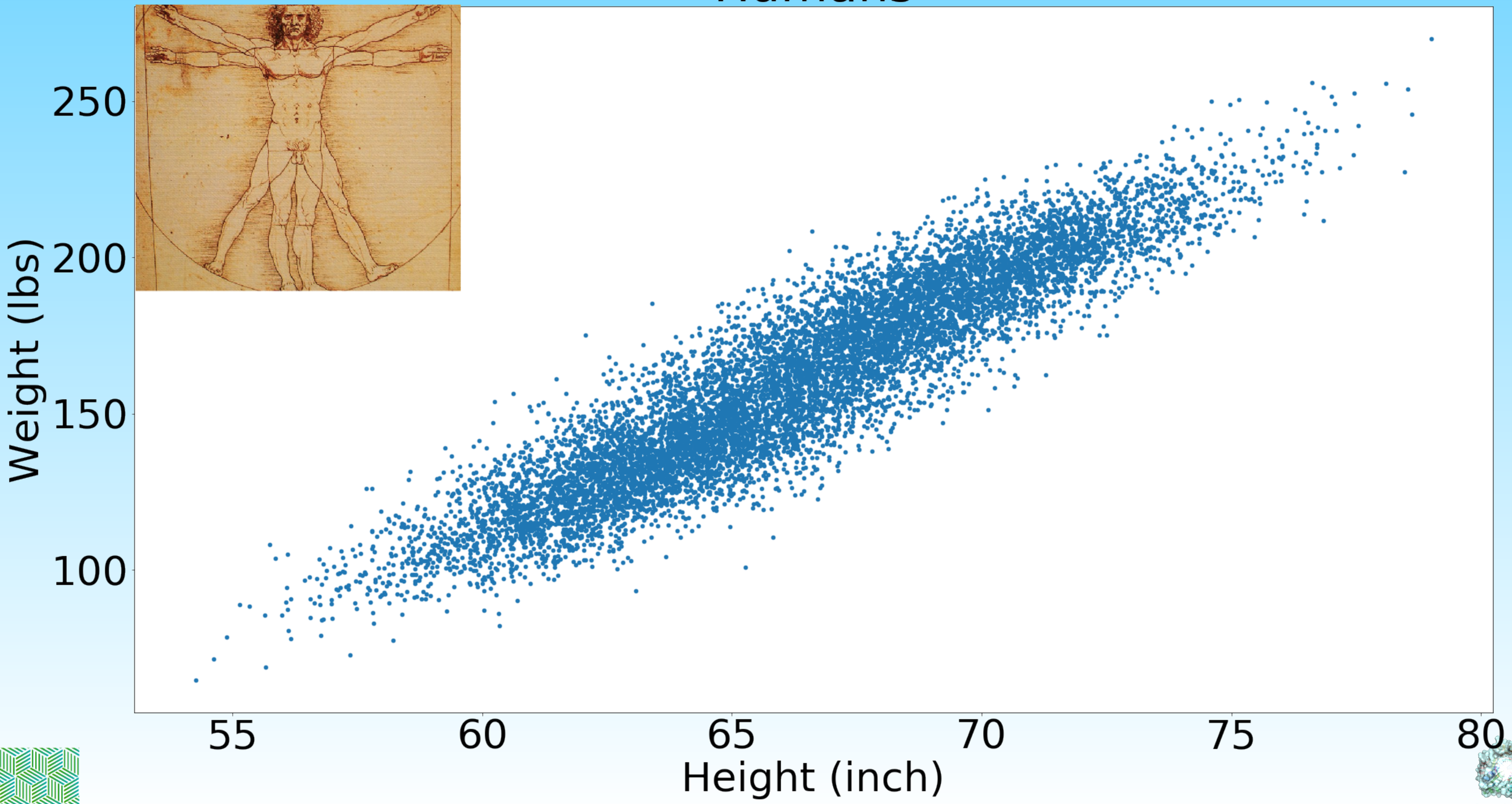
PATTERN
RECOGNIZER

http://scriptoriumdaily.com/

Data Weekends

Catalit LLC

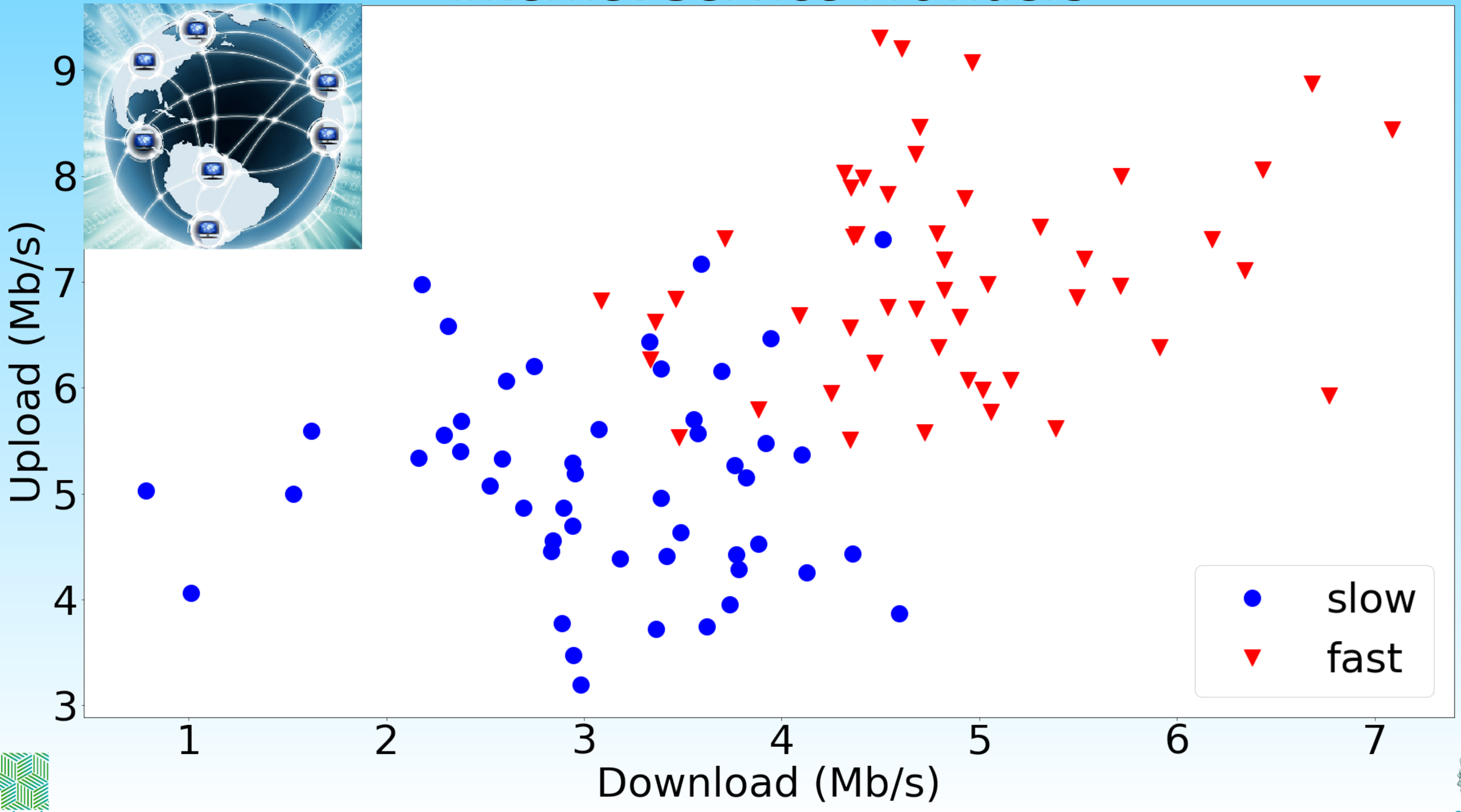
Humans



Data Weekends

Catalit LLC

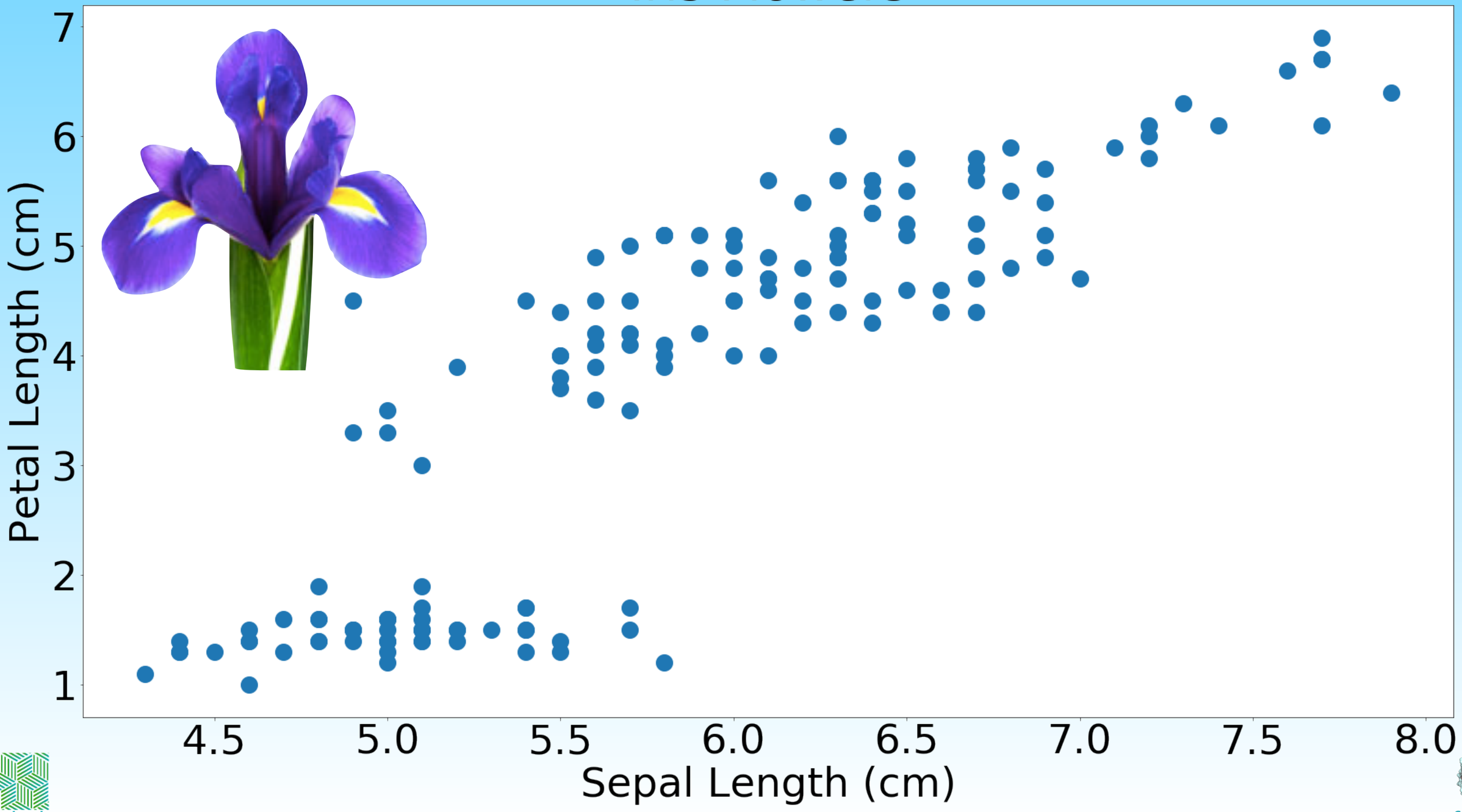
Internet Service Providers



Data Weekends

Catalit LLC

Iris Flowers



Data Weekends

Catalit LLC

REGRESSION
CLASSIFICATION
CLUSTERING

ML PROBLEMS

MACHINE LEARNING

"A field of study that gives computers the ability to learn without being explicitly programmed." (1959)

Arthur Samuel, AI pioneer
Source: Stanford

LEARNING

Learning is not memorizing. It is **generalizing** the conclusions to previously unseen examples

SUPERVISED LEARNING

http://www.realsafety.org/wp-content/uploads/2014/11/safety-supervisors-interaction.png

SUPERVISED LEARNING

	Feature 1	Feature 2	Label
1	-0.14	0.18	Cat
2	1.07	-0.75	Dog
3	-2.58	0.63	Cat
...	...	...	...

SUPERVISED LEARNING

Data Weekends

Catalit LLC

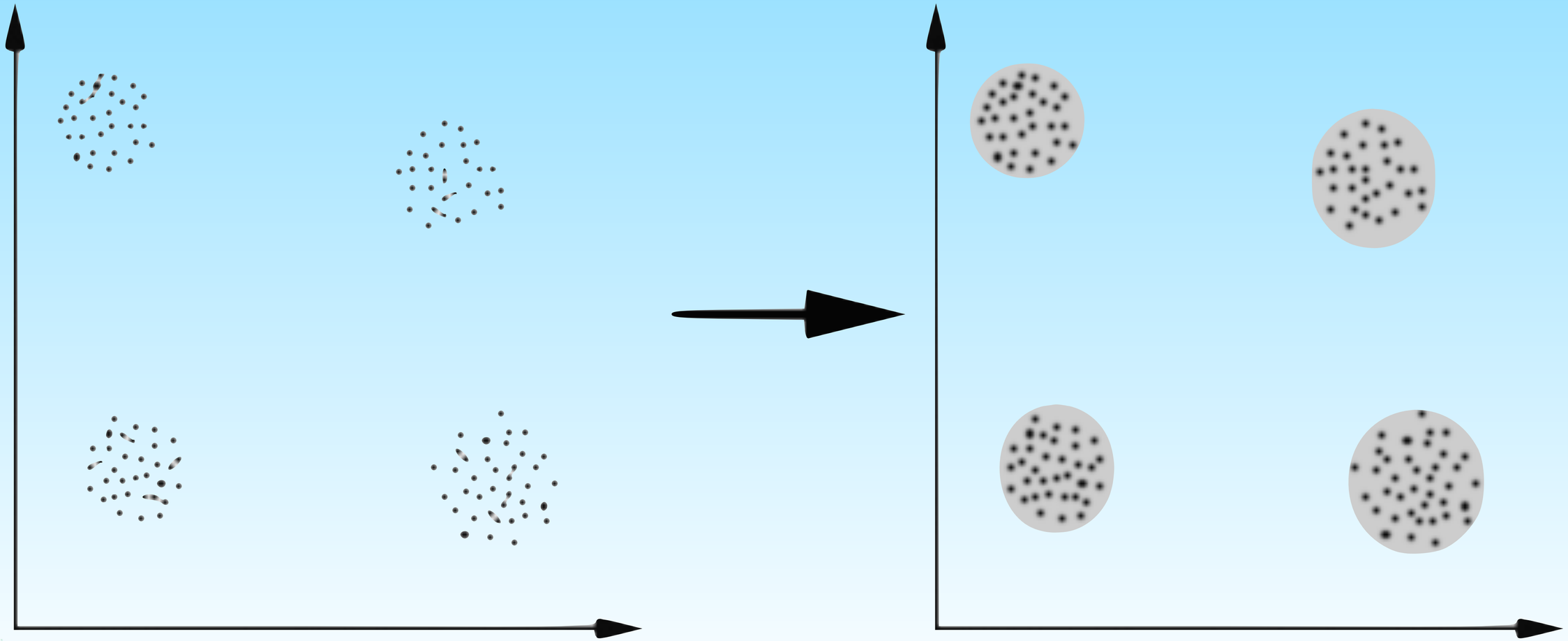
UNSUPERVISED LEARNING

http://blog.mwbookkeeping.co.uk/files/82/images/LEADER_2110.jpg

Data Weekends

Catalit LLC

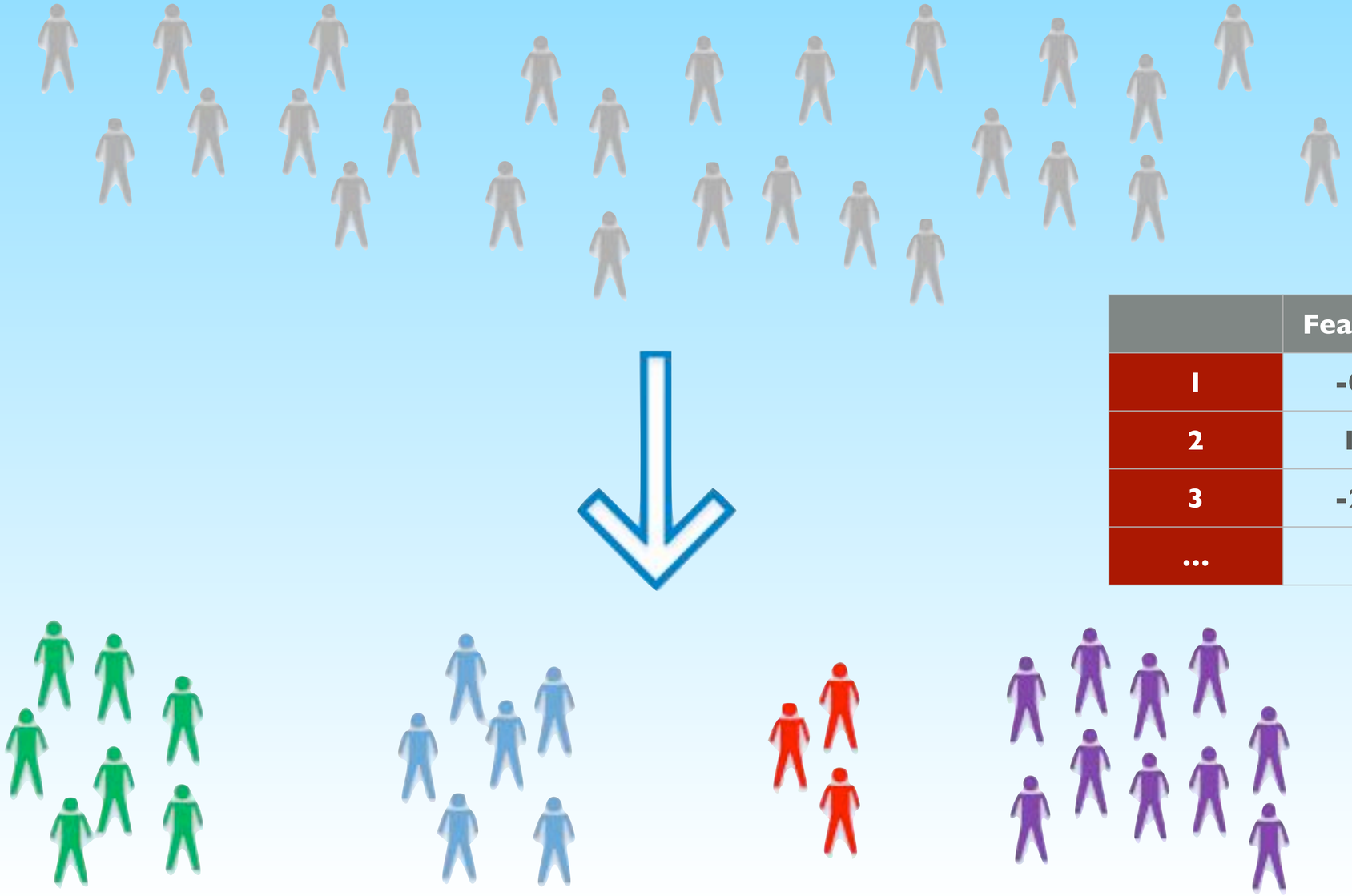
UNSUPERVISED LEARNING



Data Weekends

Catalit LLC

UNSUPERVISED LEARNING



	Feature 1	Feature 2	Label
1	-0.14	0.18	?
2	1.07	-0.75	?
3	-2.58	0.63	?
...	...	...	?

TABULAR DATA

TABULAR DATA

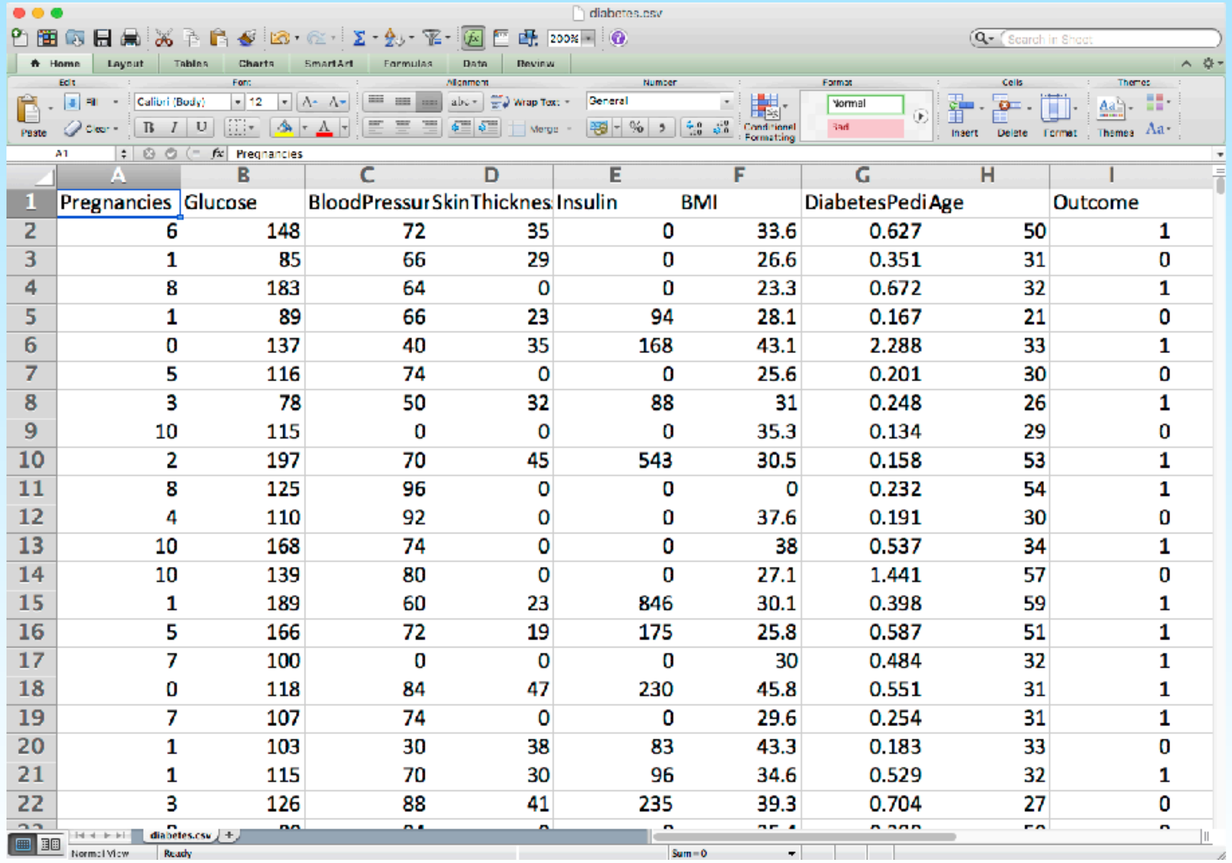
	Age	Gender	Annual Salary	Months in residence	Months in job	Current Debt	Paid off credit
Client 1	23	Male	$30,000	36	12	$5,000	Yes
Client 2	30	Female	$45,000	12	12	$1,000	Yes
Client 3	19	Male	$15,000	3	1	$10,000	No
Client 4	25	Male	$25,000	12	27	$15,000	?

ONE TABLE

- Excel
- CSV
- TSV
- …

Columns

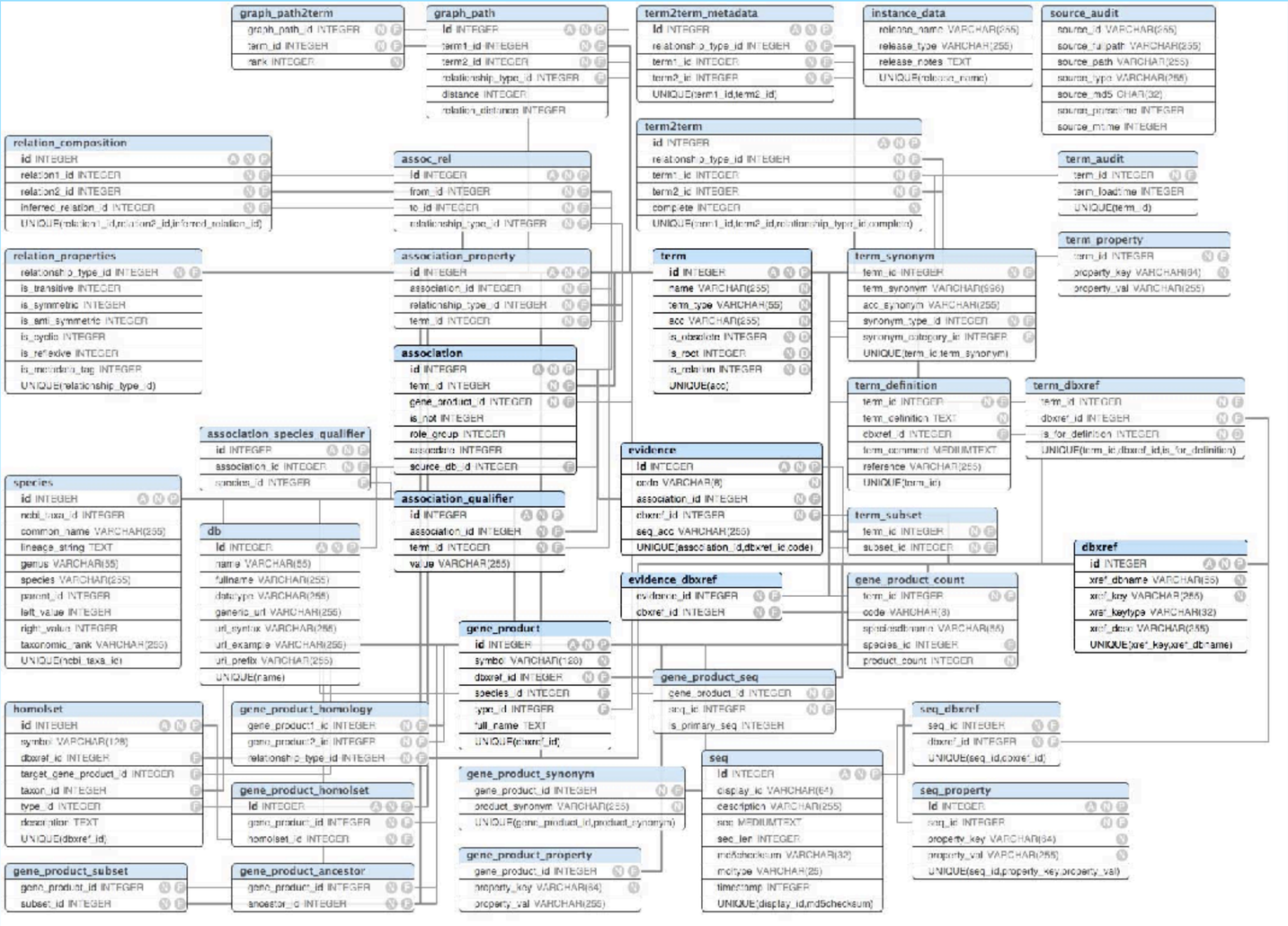
Rows



	A	B	C	D	E	F	G	H	I
1	Pregnancies	Glucose	BloodPressur	SkinThicknes	Insulin	BMI	DiabetesPedi	Age	Outcome
2	6	148	72	35	0	33.6	0.627	50	1
3	1	85	66	29	0	26.6	0.351	31	0
4	8	183	64	0	0	23.3	0.672	32	1
5	1	89	66	23	94	28.1	0.167	21	0
6	0	137	40	35	168	43.1	2.288	33	1
7	5	116	74	0	0	25.6	0.201	30	0
8	3	78	50	32	88	31	0.248	26	1
9	10	115	0	0	0	35.3	0.134	29	0
10	2	197	70	45	543	30.5	0.158	53	1
11	8	125	96	0	0	0	0.232	54	1
12	4	110	92	0	0	37.6	0.191	30	0
13	10	168	74	0	0	38	0.537	34	1
14	10	139	80	0	0	27.1	1.441	57	0
15	1	189	60	23	846	30.1	0.398	59	1
16	5	166	72	19	175	25.8	0.587	51	1
17	7	100	0	0	0	30	0.484	32	1
18	0	118	84	47	230	45.8	0.551	31	1
19	7	107	74	0	0	29.6	0.254	31	1
20	1	103	30	38	83	43.3	0.183	33	0
21	1	115	70	30	96	34.6	0.529	32	1
22	3	126	88	41	235	39.3	0.704	27	0

MANY TABLES

- Databases
- Collections of files
- …



http://www.geneontology.org/

TABULAR DATA

	Age	Gender	Annual Salary	Months in residence	Months in job	Current Debt	Paid off credit
Client 1	23	Male	$30,000	36	12	$5,000	Yes
Client 2	30	Female	$45,000	12	12	$1,000	Yes
Client 3					1	$10,000	No
Client 4	25	Male	$25,000	12	27	$15,000	?

Record OR Data Point

TABULAR DATA

	Age	Gender	Annual Salary	Months in residence	Months in job	Current Debt	Paid off credit
Client 1	23	Male	$30,000	36	12	$5,000	Yes
Client 2	30	Female	$45,000	12	12	$1,000	Yes
Client 3	19	Male	$15,000	3	1	$10,000	No
Client 4	25	Male	$25,000	12	27	$15,000	?

Features

Data Weekends

Catalit LLC

PREDICTION TYPE

PREDICTION TYPE

http://i.ytimg.com/vi/WX0hnuniLpI/maxresdefault.jpg

PREDICTION TYPE

CATEGORICAL	CONTINUOUS
Eye colors	Height of children
Courses at university	Weight of cars
Highest degree	Speed of the train
Gender	Temperature
Spam or not	Stock price

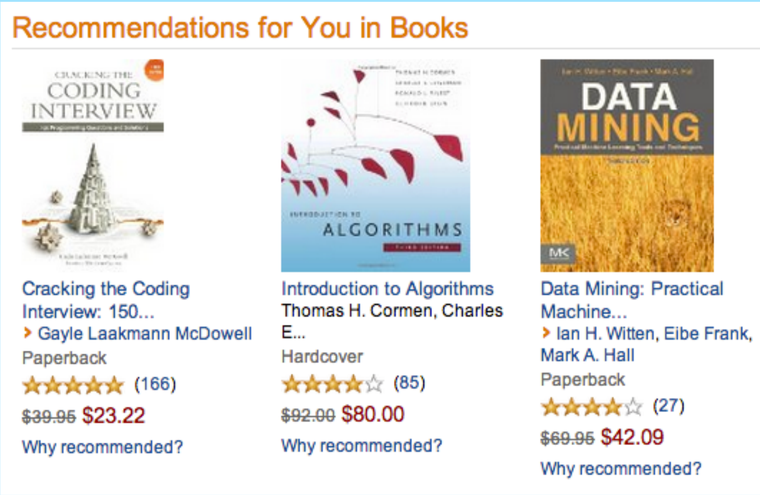
MAIN ML TECHNIQUES

	CATEGORICAL	CONTINUOUS
SUPERVISED	CLASSIFICATION	REGRESSION
UNSUPERVISED	CLUSTERING	

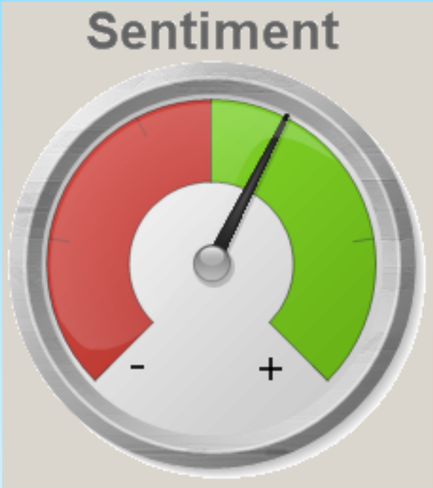
APPLICATIONS

APPLICATIONS

Book recommendation

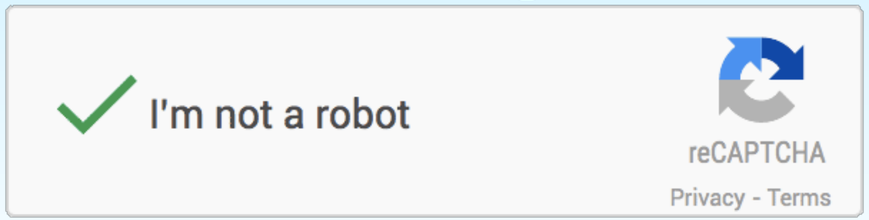


Sentiment Analysis



Heart Monitoring

Human recognition



Caption generation

"man in black shirt is playing guitar."

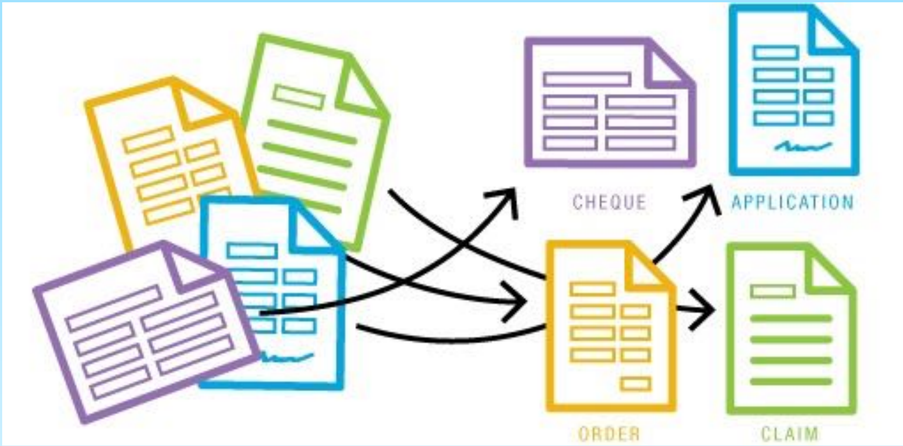
"construction worker in orange safety vest is working on road."

Data Weekends

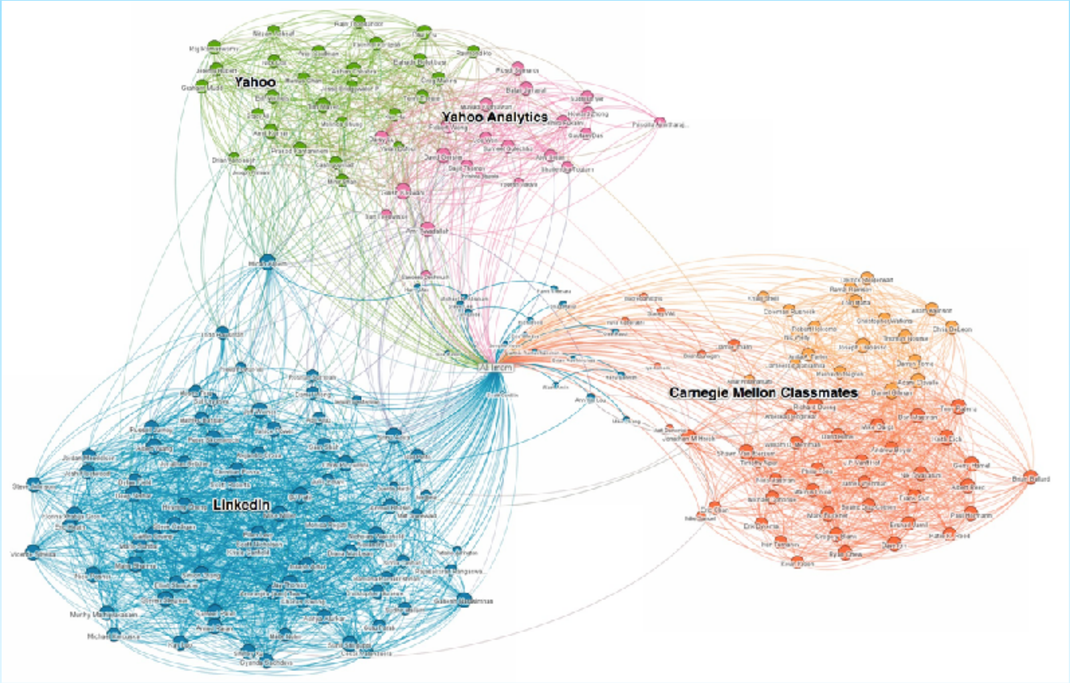
Catalit LLC

APPLICATIONS

Document classification



Social Network Analysis

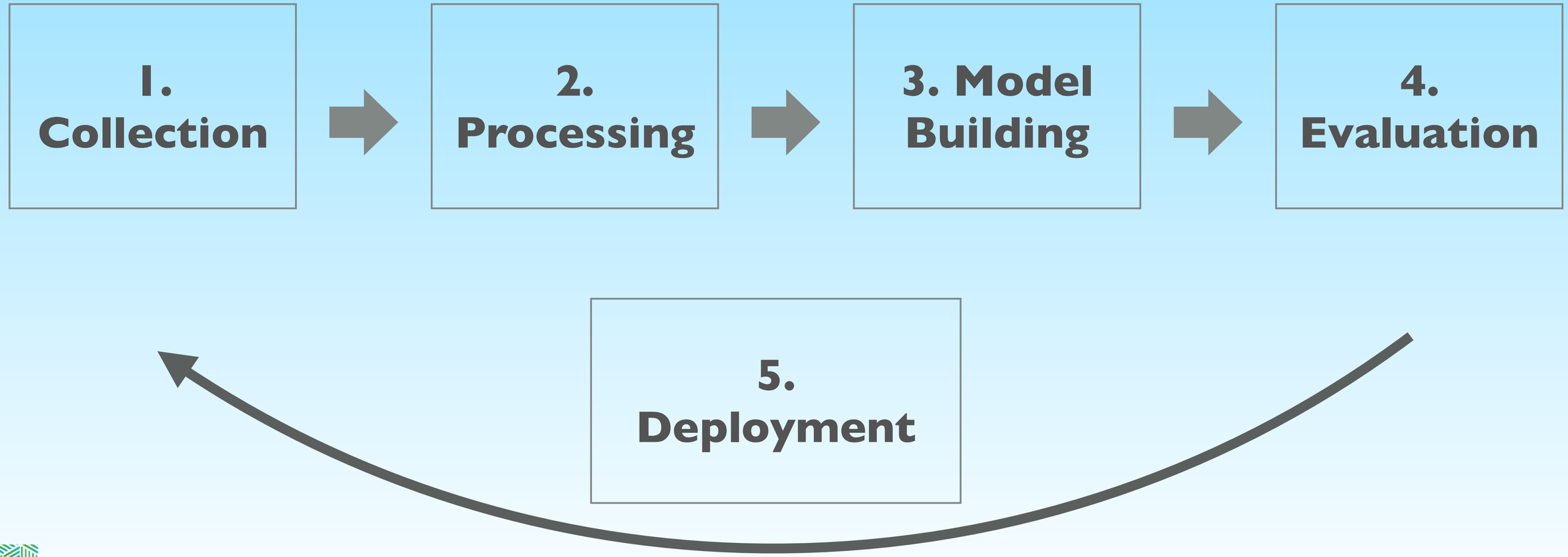


House price prediction

Data Weekends

Catalit LLC

ML STEPS



Data Weekends

Catalit LLC

DATA SCIENCE WITH PYTHON

Data Weekends

Catalit LLC

WHICH TOOL

KNIME

IPyl:

pandas

Augustus

open data

actian

Scalding

RStudio

{m}

Cascalog

julia

Spark

cascading

ML base

APACHE HBASE

ORACLE

BERKELEY DB

ADAPA®

mahout

WHICH TOOL

- Community? (to learn, to hire, to grow)
- Open source?
- Enterprise support?
- Versatile?

PYDATA ECHOSYSTEM

python™

NumPy

pandas

IP[y]: IPython Interactive Computing

jupyter

Anaconda

SciPy

scikits learn machine learning in Python

matplotlib

Bokeh

LAB 01 & LAB 02

- Install Anaconda or Miniconda
- Download https://goo.gl/xmABYJ
- Unzip to your computer
- Open terminal
- cd to unzipped folder
- conda env create
- jupyter notebook

Data Weekends

Catalit LLC

REGRESSION

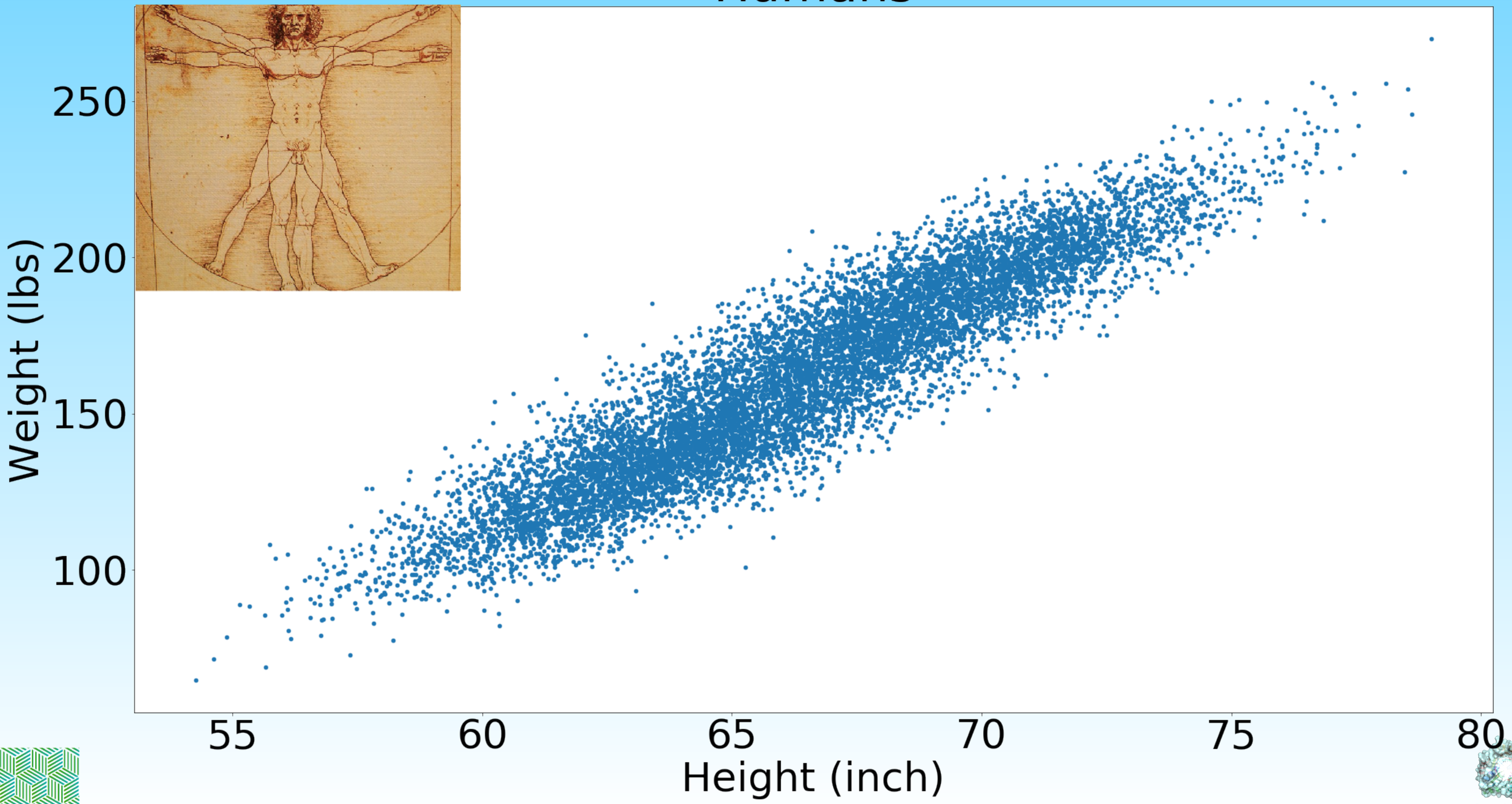
MAIN ML TECHNIQUES

	CATEGORICAL	CONTINUOUS
SUPERVISED	?	?
UNSUPERVISED	?	?

MAIN ML TECHNIQUES

	CATEGORICAL	CONTINUOUS
SUPERVISED	CLASSIFICATION	REGRESSION
UNSUPERVISED	CLUSTERING	

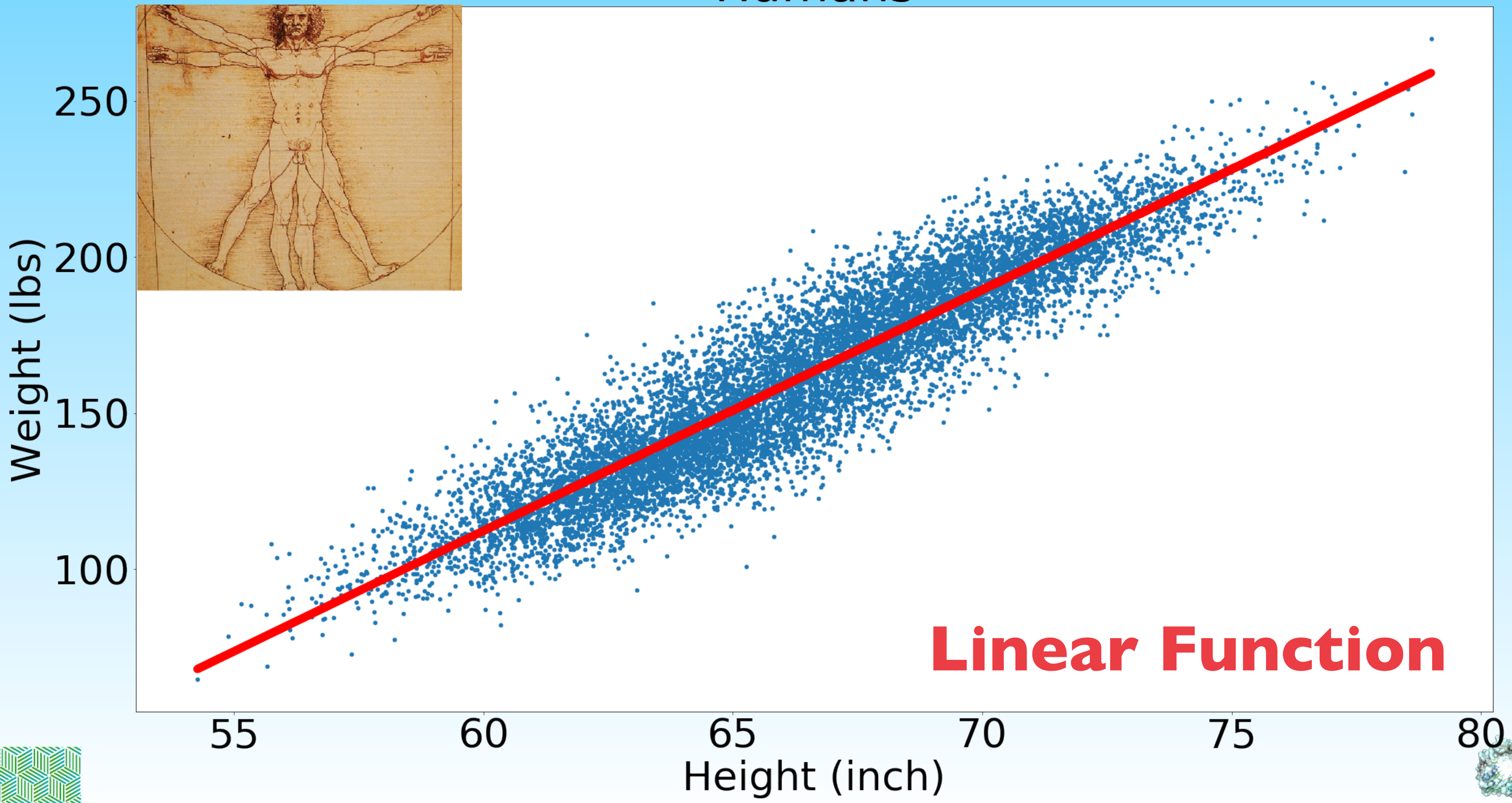
Humans



Data Weekends

Catalit LLC

Humans



Data Weekends

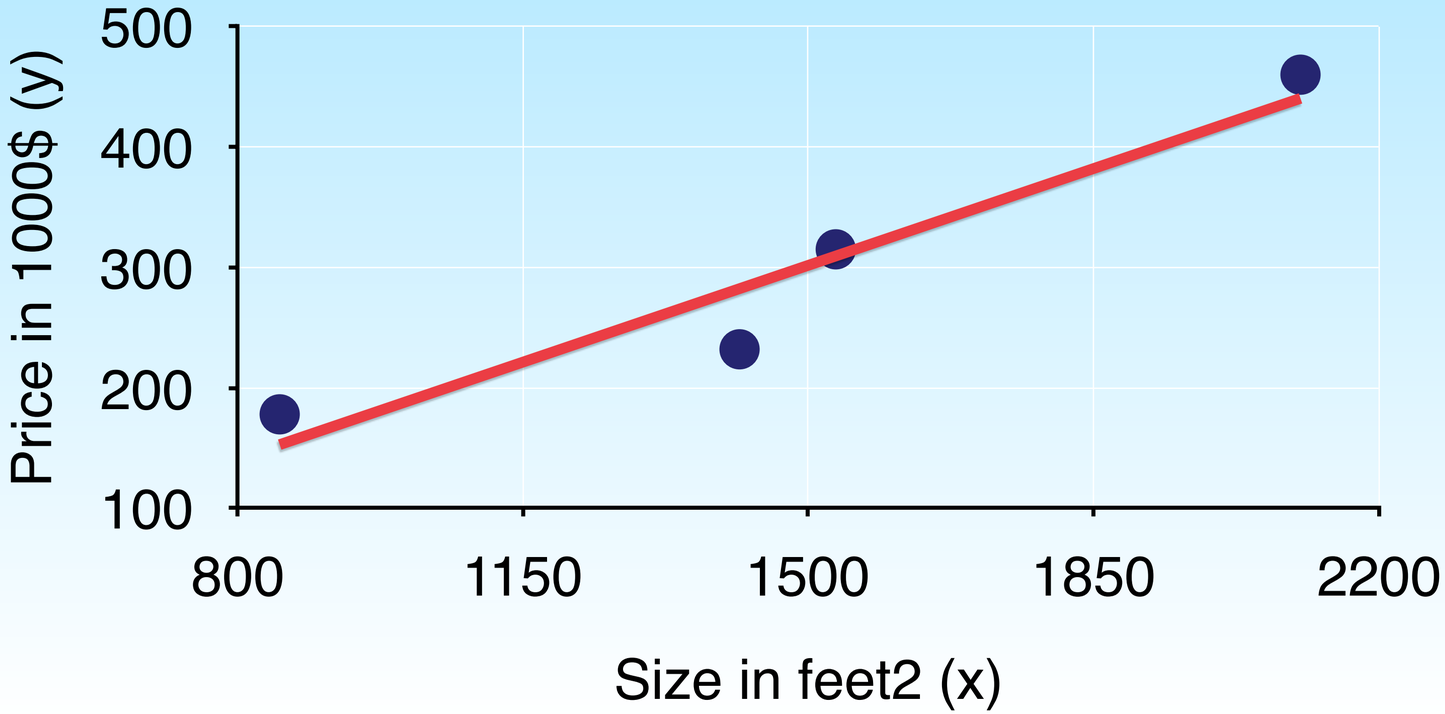
Catalit LLC

LINEAR REGRESSION

Size in feet² (x)	Price in 1000$ (y)
2104	460
1416	232
1534	315
852	178
…	…

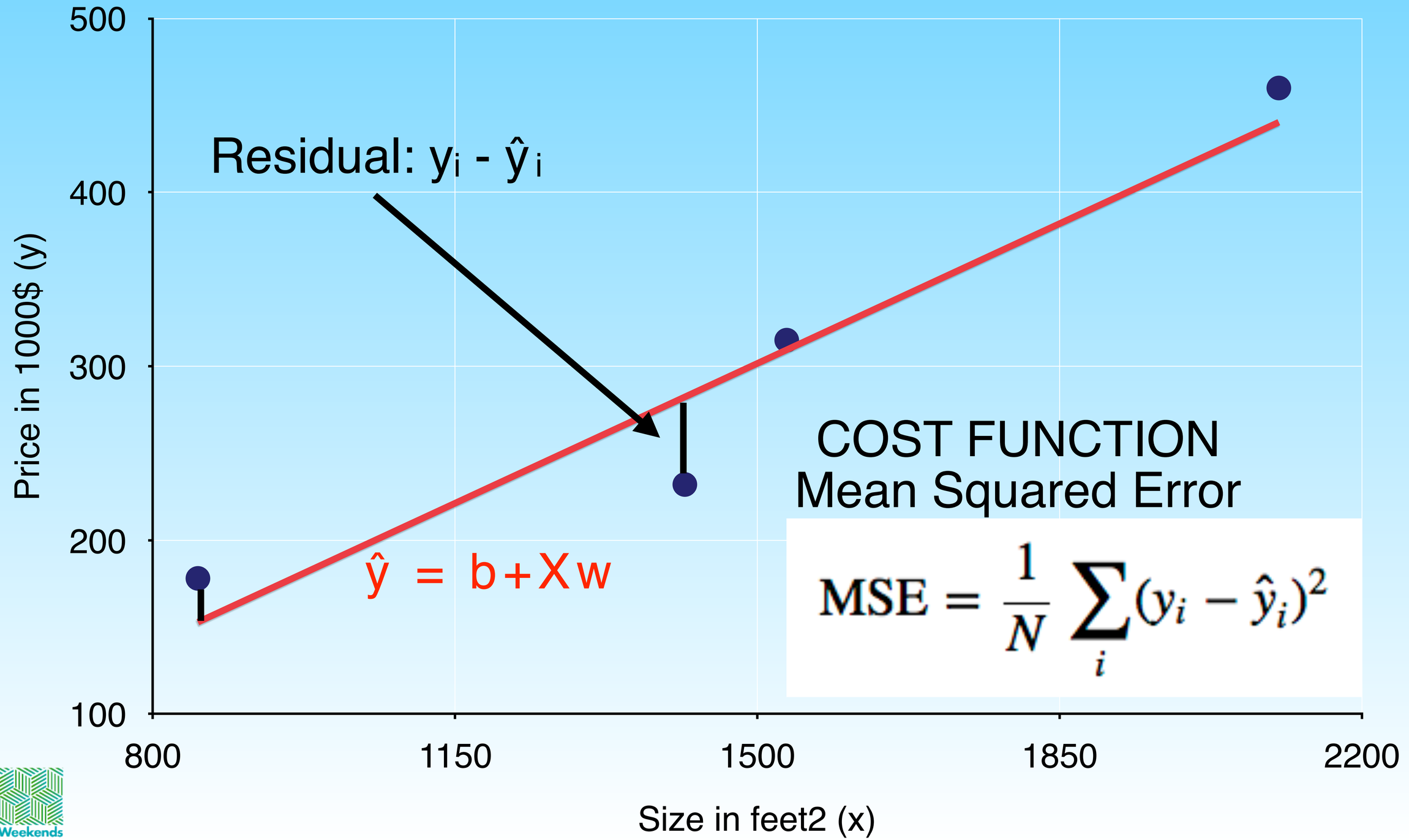
Hypothesis

$$\hat{y} = b + Xw$$



Data Weekends

Catalit LLC



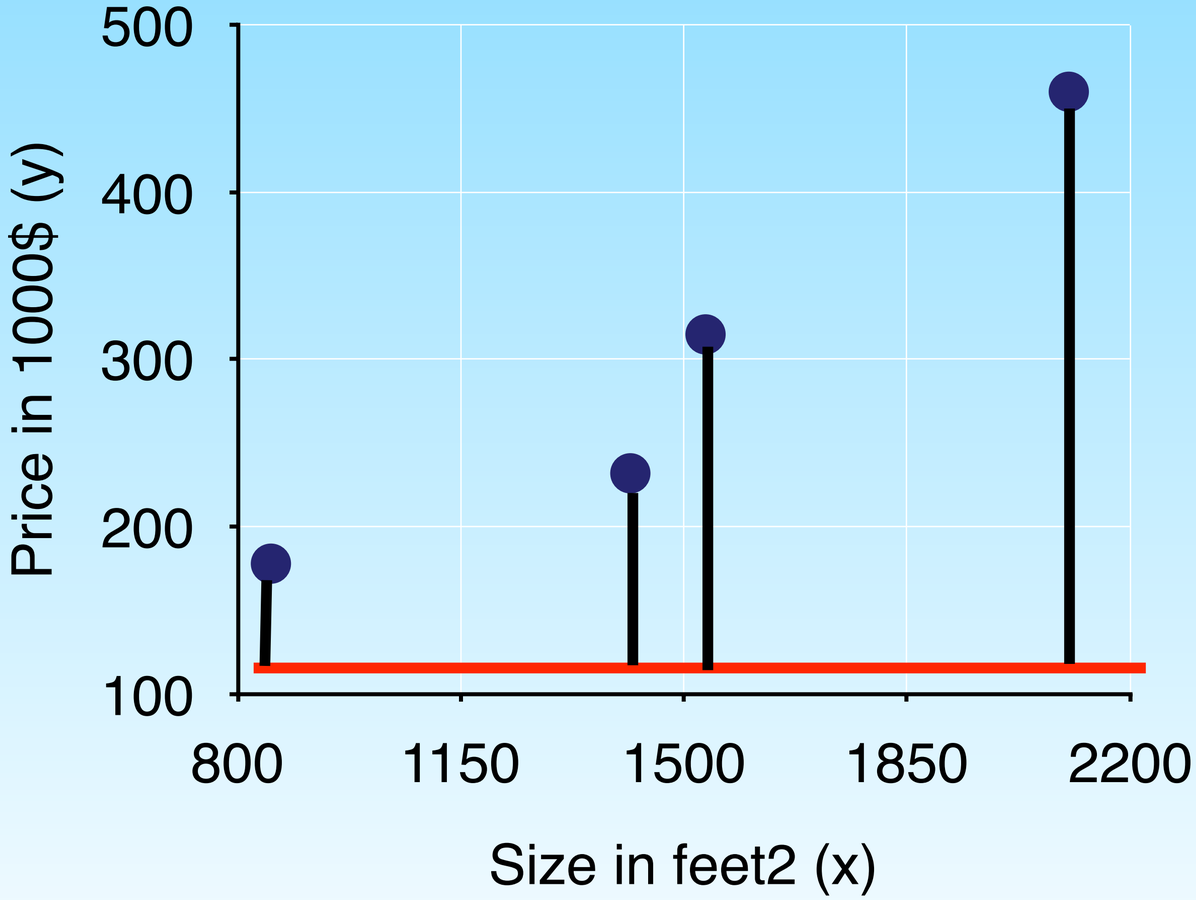
COST FUNCTION
Mean Squared Error

$$\mathrm{MSE} = \frac{1}{N}\sum_i (y_i - \hat{y}_i)^2$$

Data Weekends

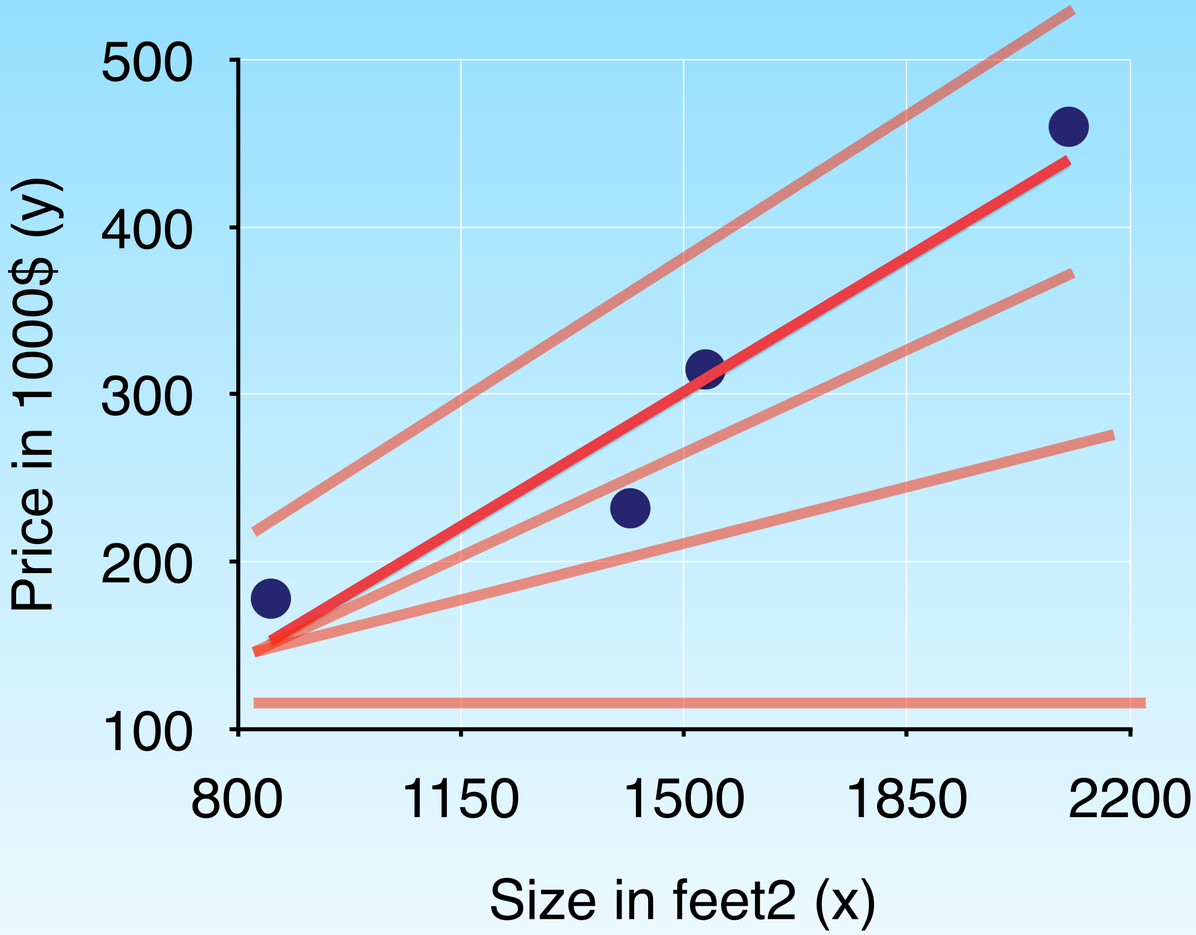
Catalit LLC

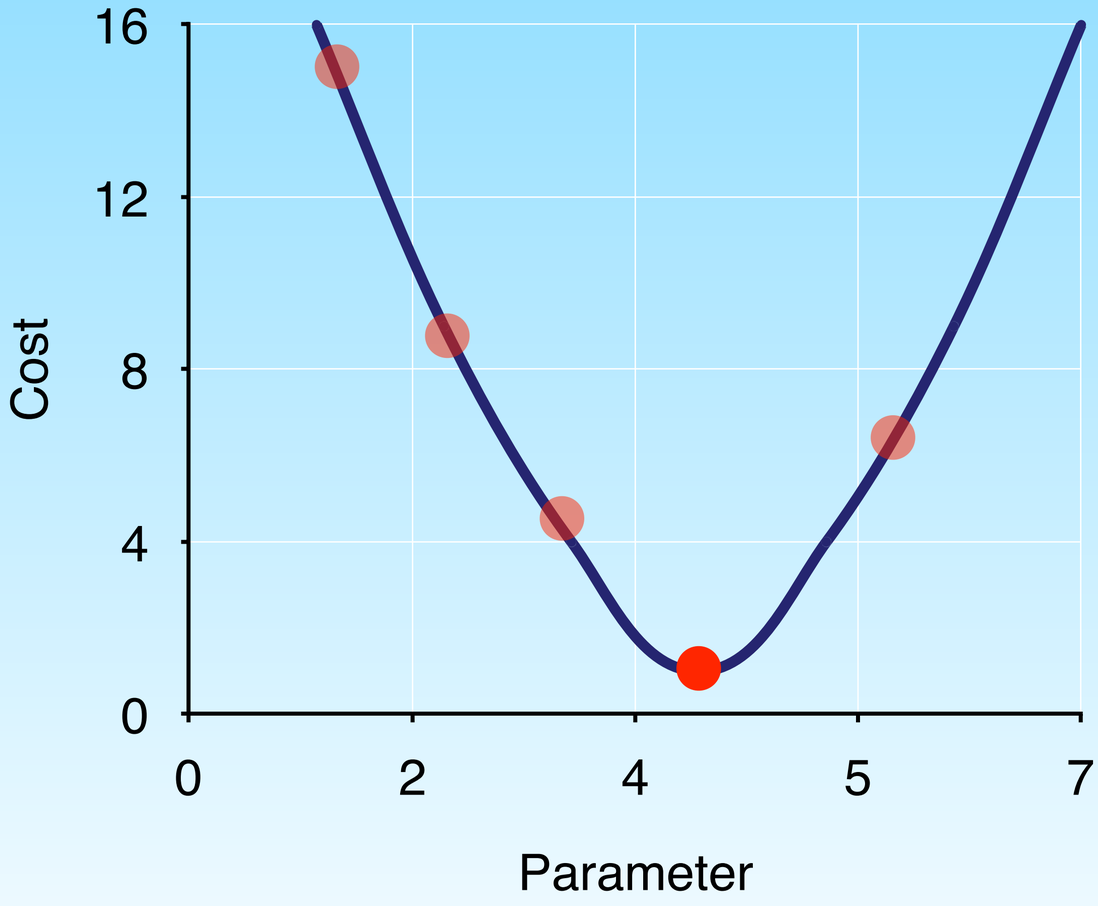
COST FUNCTION



- Fixed data
- Fixed hypothesis
- Cost depends on values of parameters

OPTIMIZATION

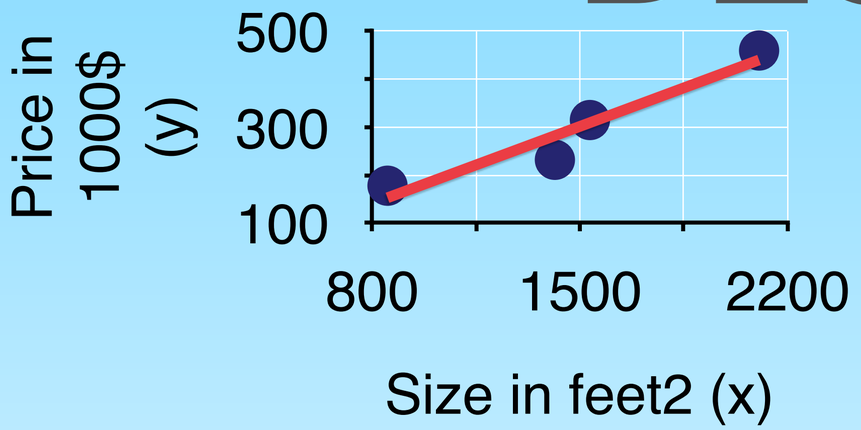




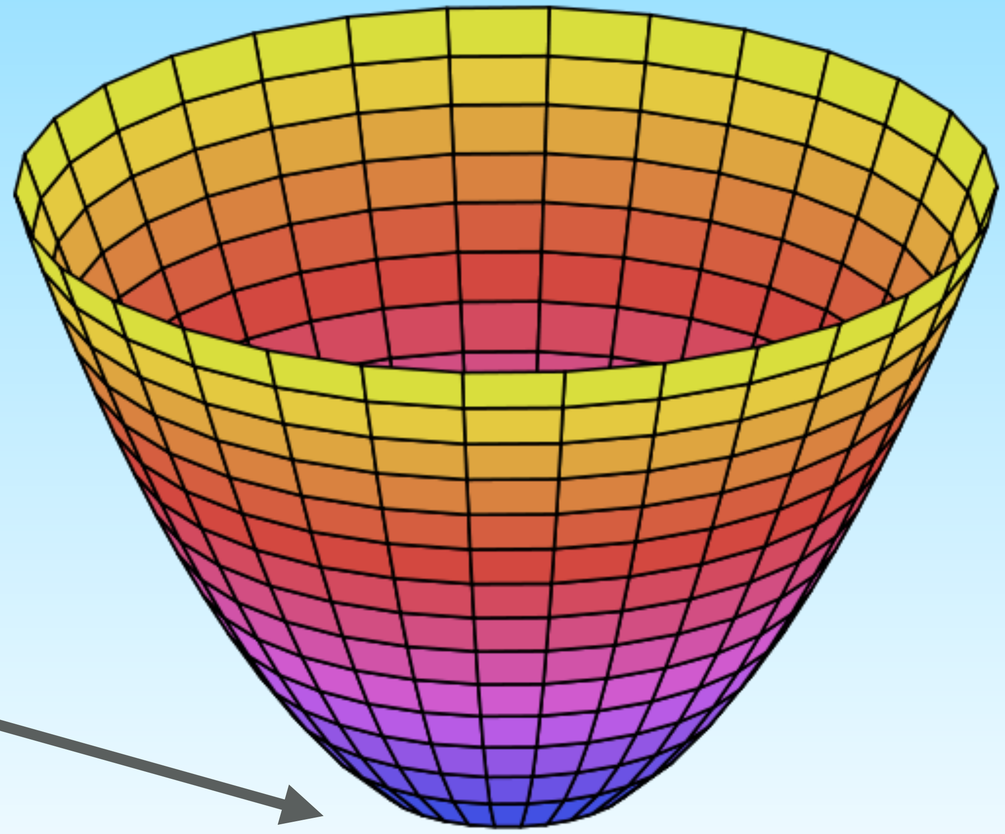
Data Weekends

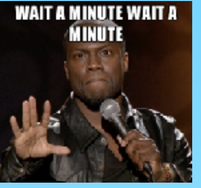
Catalit LLC

BEST MODEL



- Values of b and w that give MINIMUM COST





MANY FEATURES

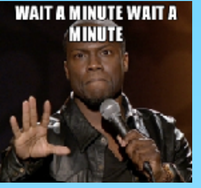
Size in feet² (x1)	Age (x2)	Number of rooms (x3)	…	Price in 1000$ (y)
2104	12	3	…	460
1416	4	2	…	232
1534	23	3	…	315
852	7	1	…	178
…	…	…	…	…

No problem!

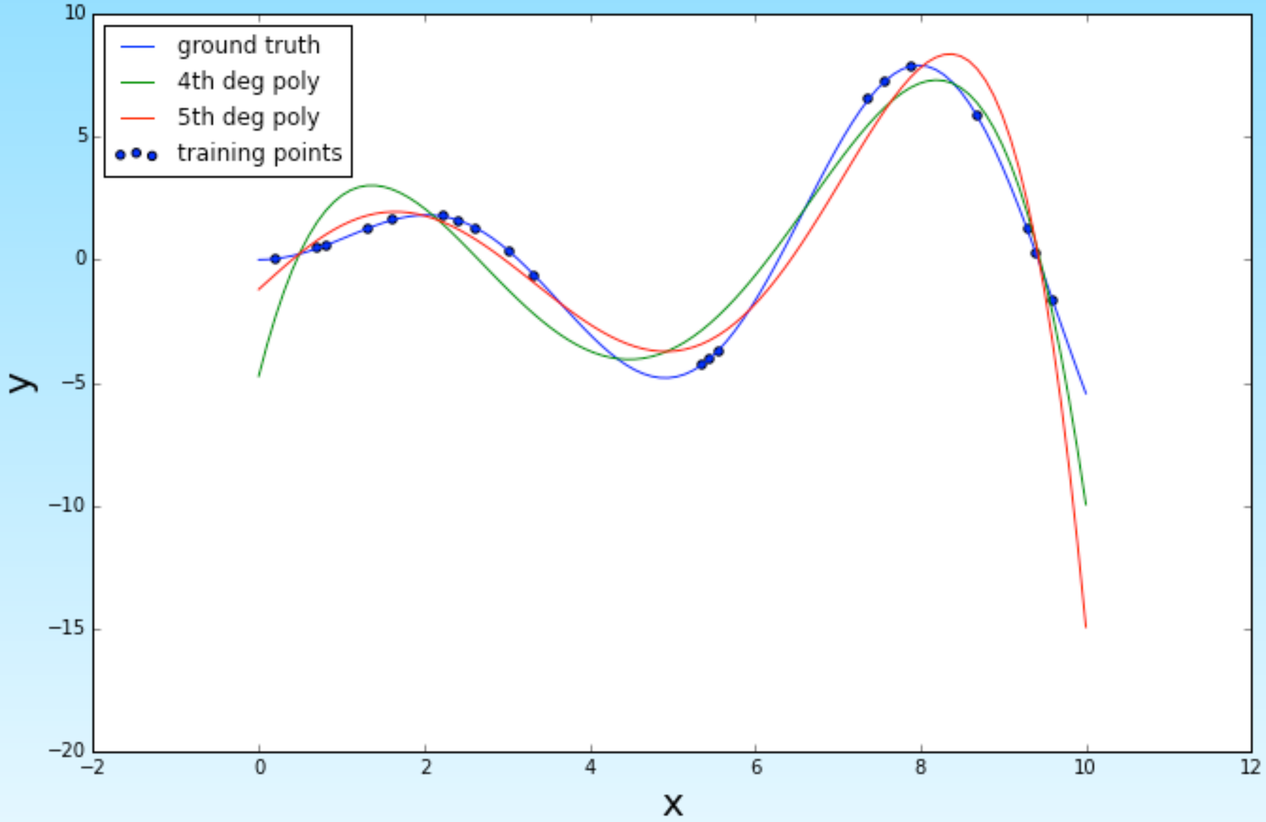
$$\hat{y} = b_0 + x_1 w_1 + x_2 w_2 + \ldots + x_n w_n$$

Data Weekends

Catalit LLC



NON LINEAR



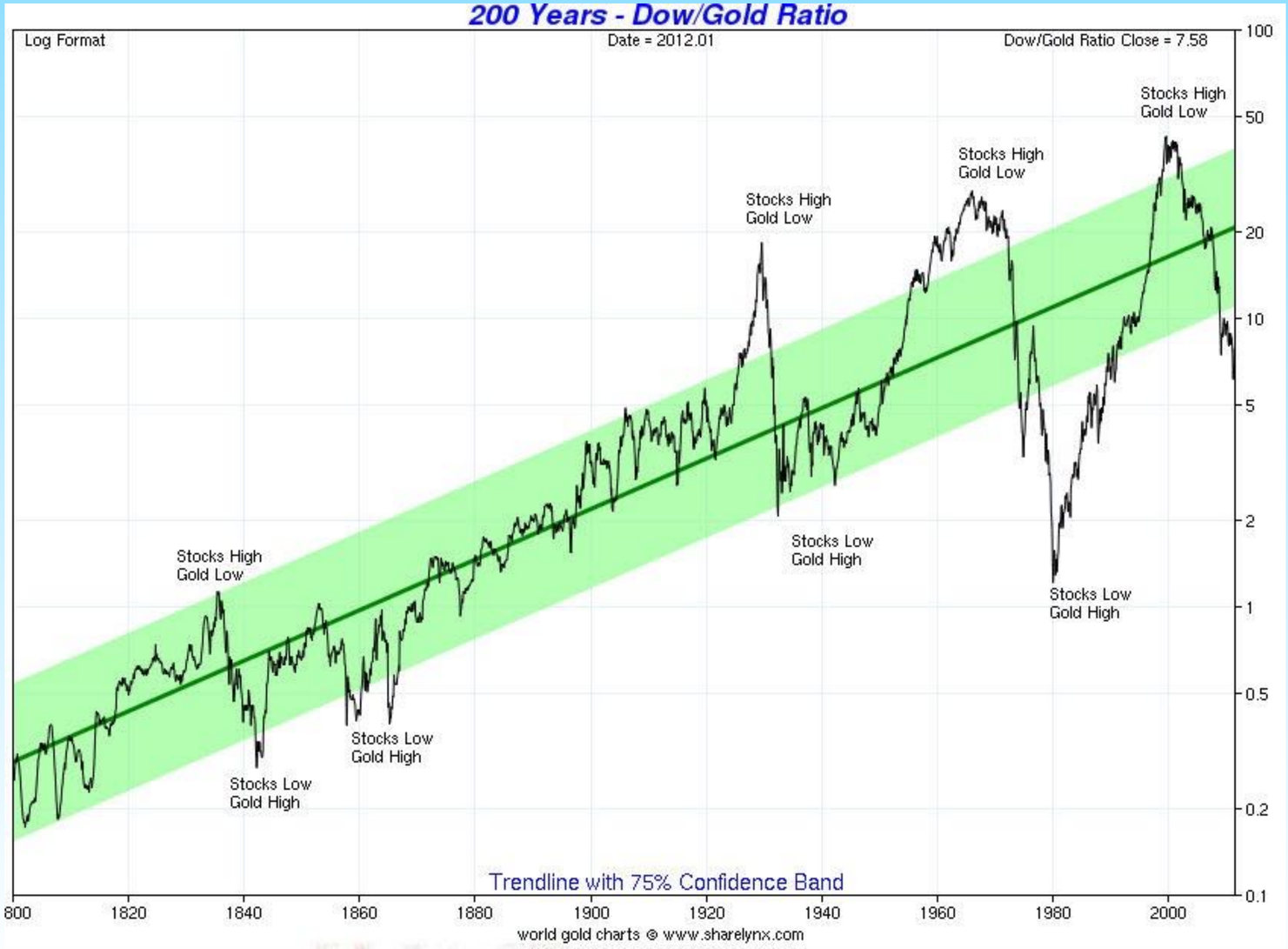
No problem!

$$\hat{y} = b_0 + x\,w_1 + f(x^2)w_2 + \ldots + f(x^n)w_n$$

Data Weekends

Catalit LLC

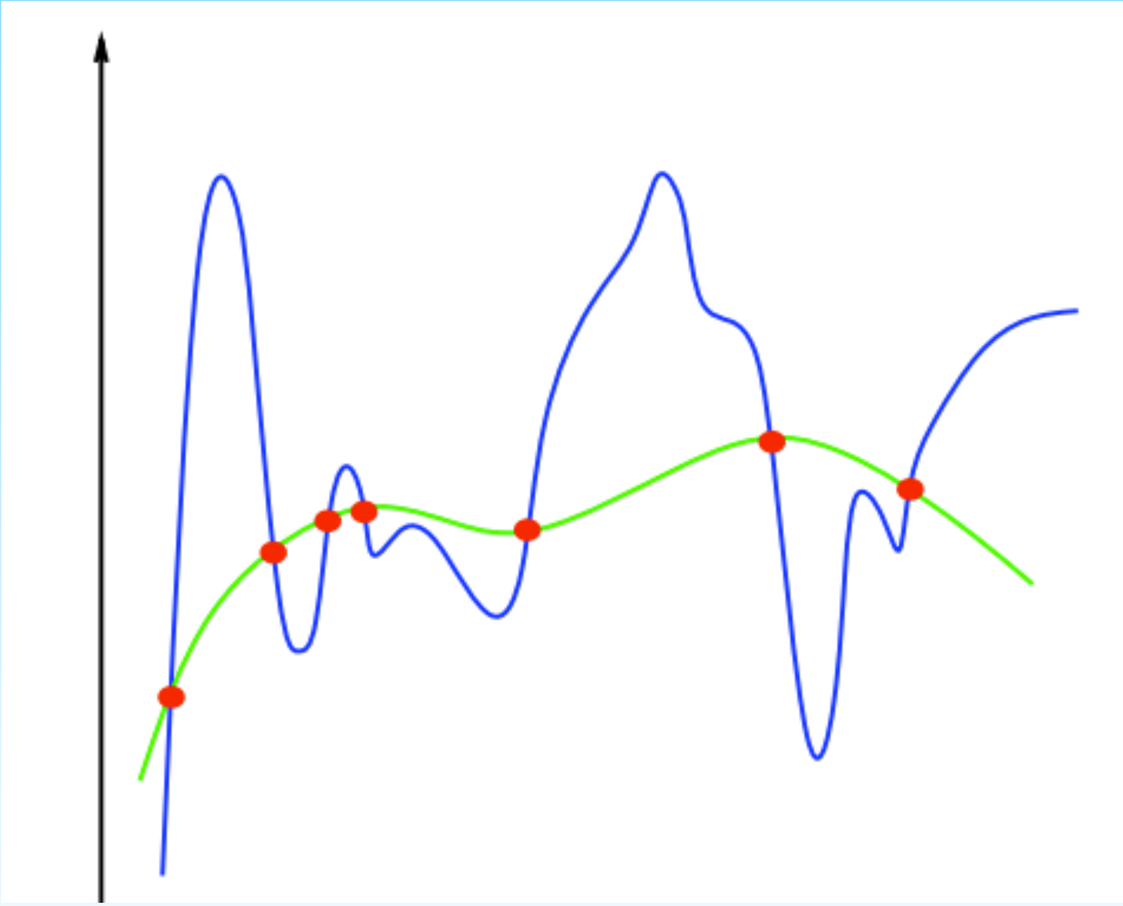
APPLICATIONS



Data Weekends

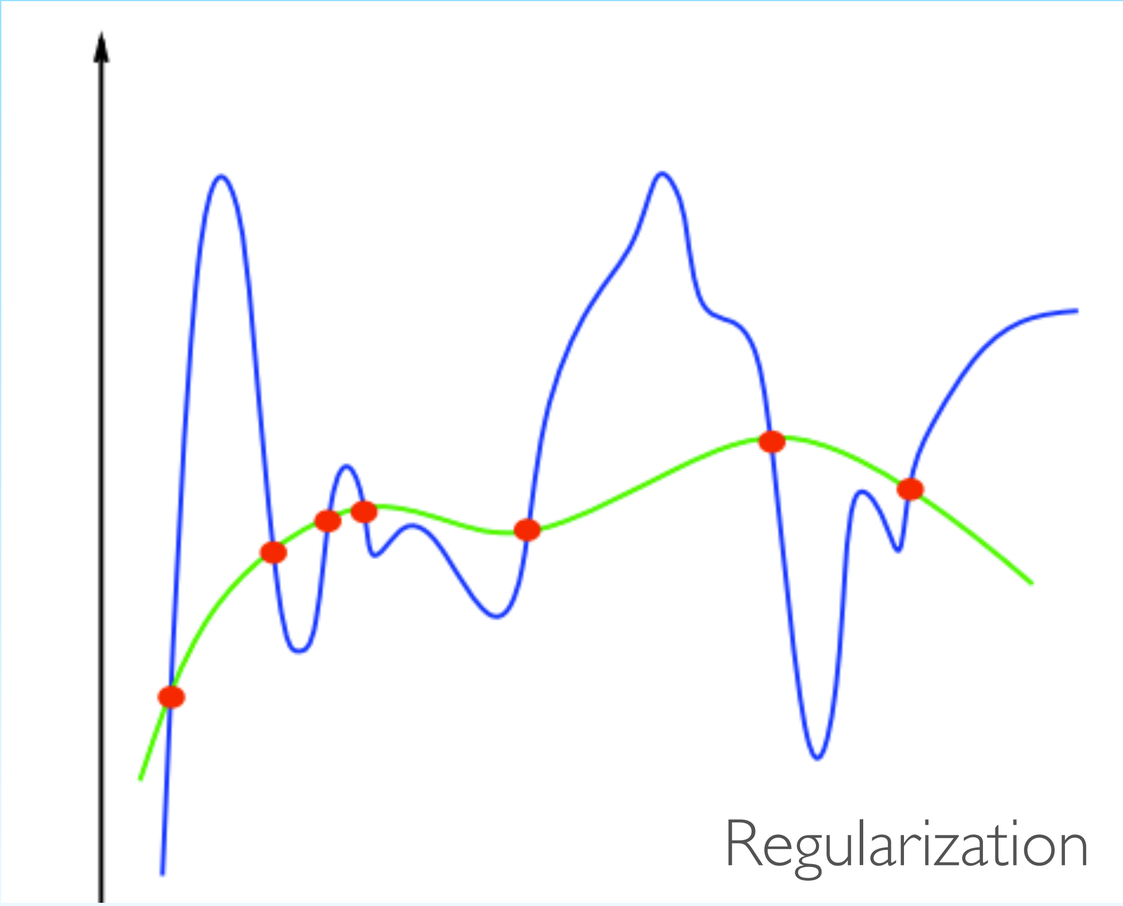
Catalit LLC

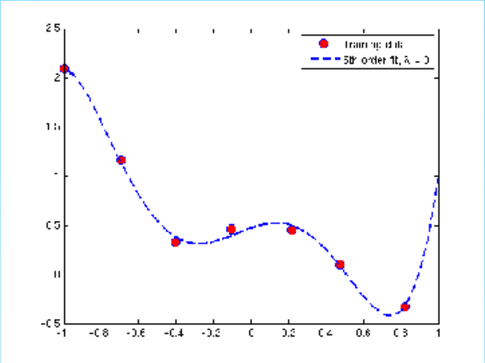
OVERFITTING

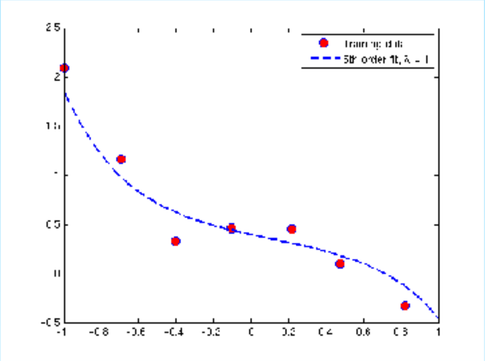


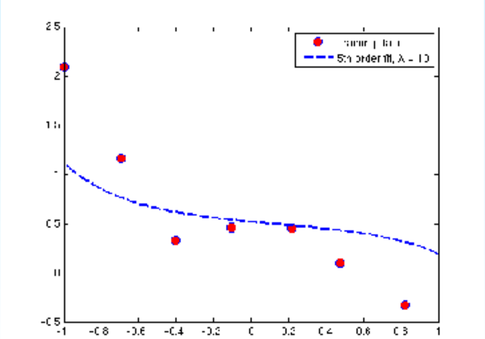
http://www.mit.edu/~9.520/spring12/slides/class02/class02.pdf

OVERFITTING





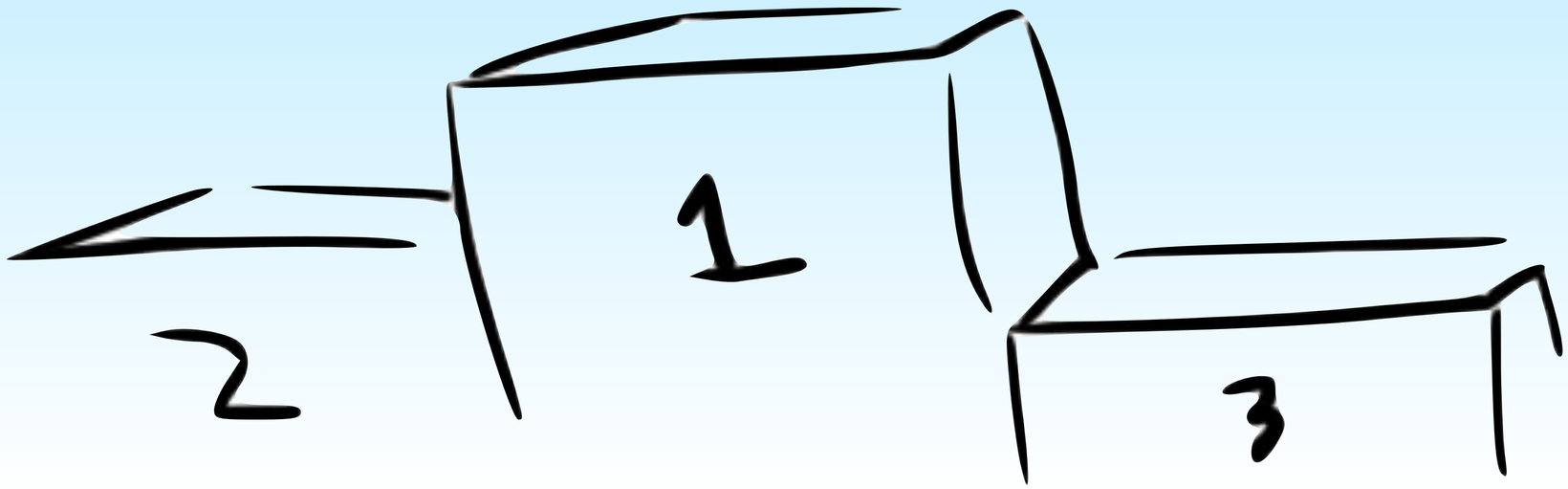




http://www.mit.edu/~9.520/spring12/slides/class02/class02.pdf

BASELINE

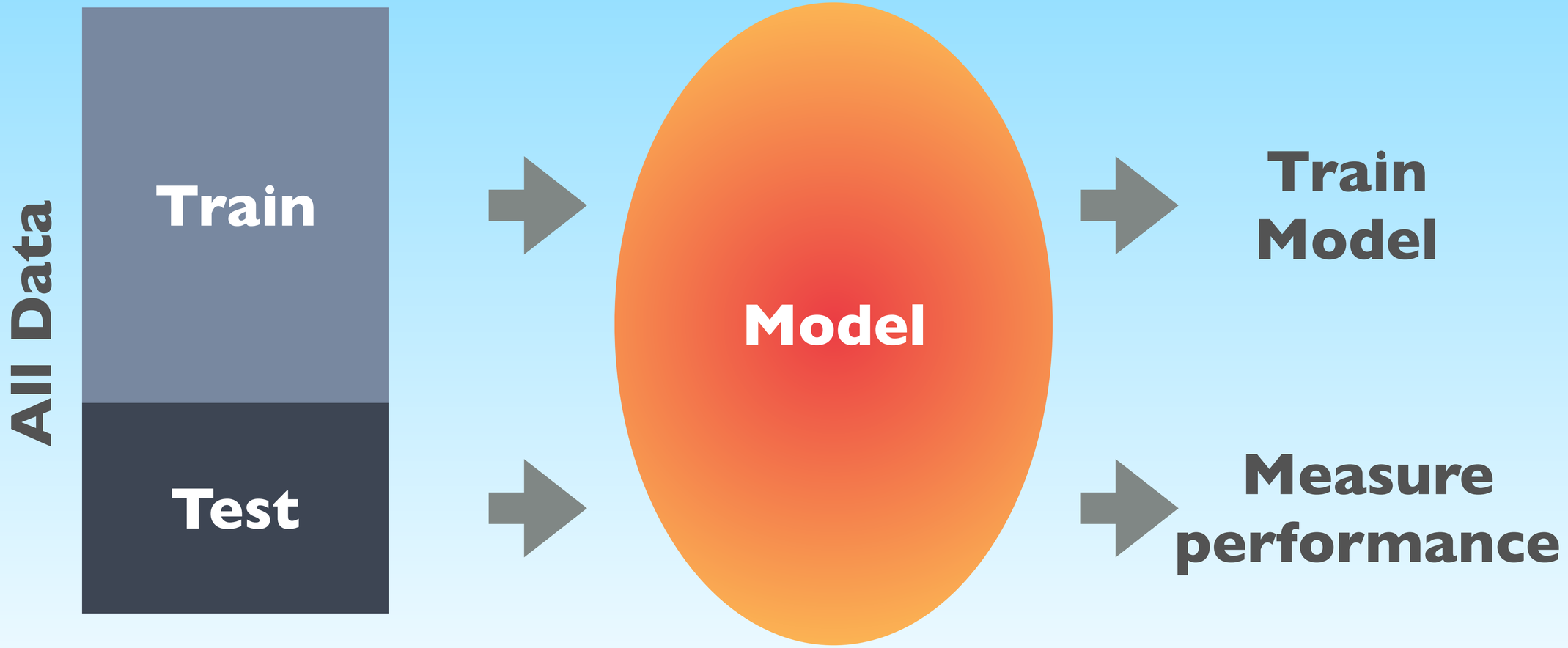
- Know the score of the simplest model
- Compare your score with simplest model score



Data Weekends

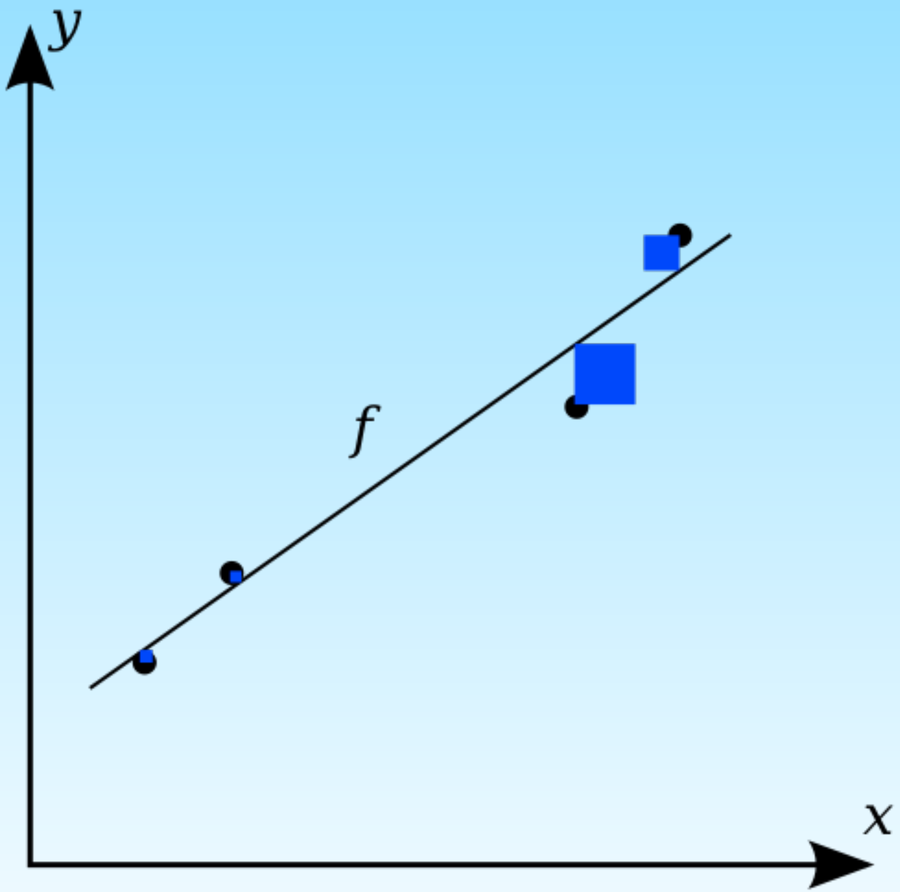
Catalit LLC

TRAIN / TEST



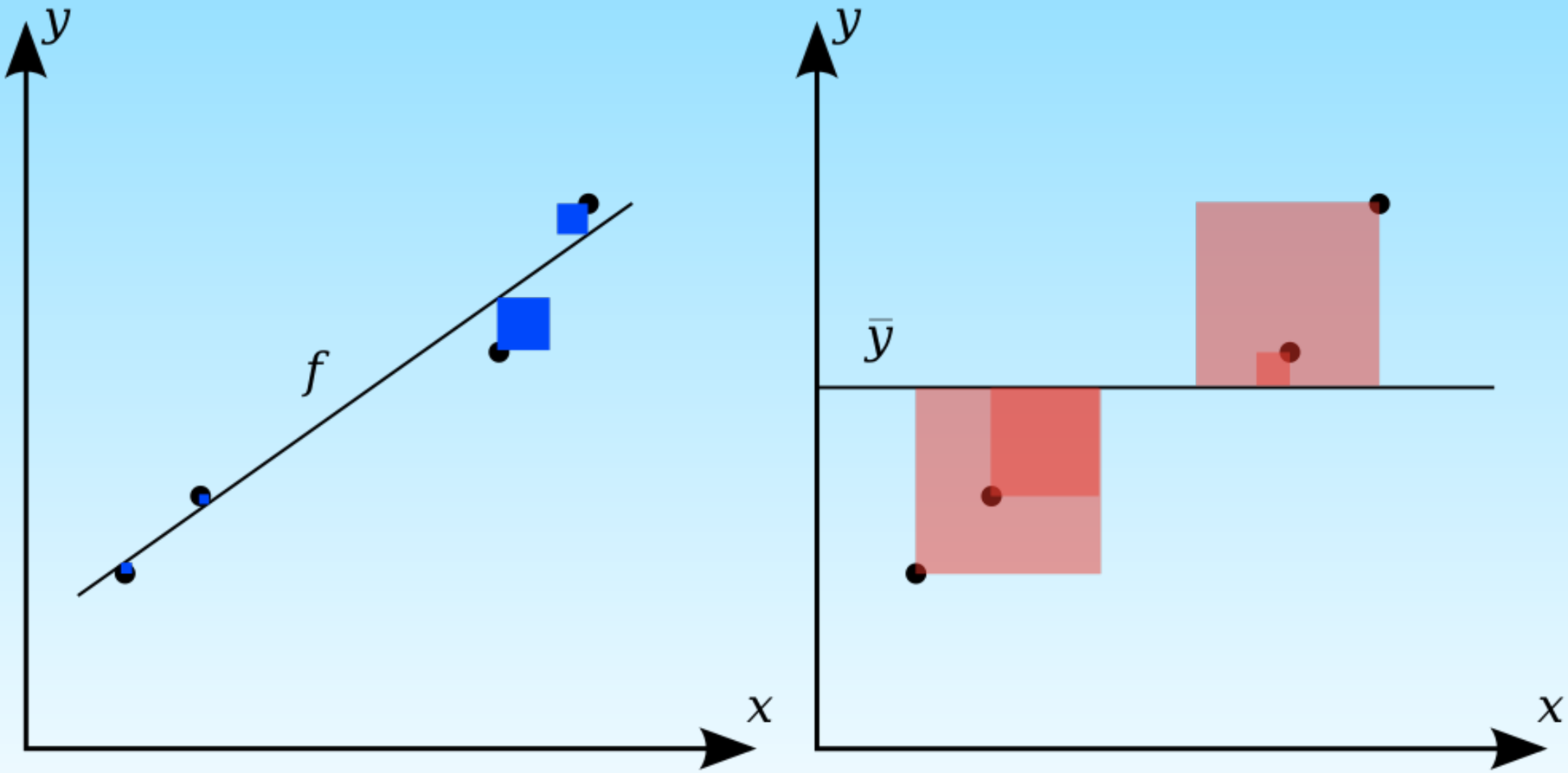
Typical values of test size: [10% - 50%]

R^2 SCORE



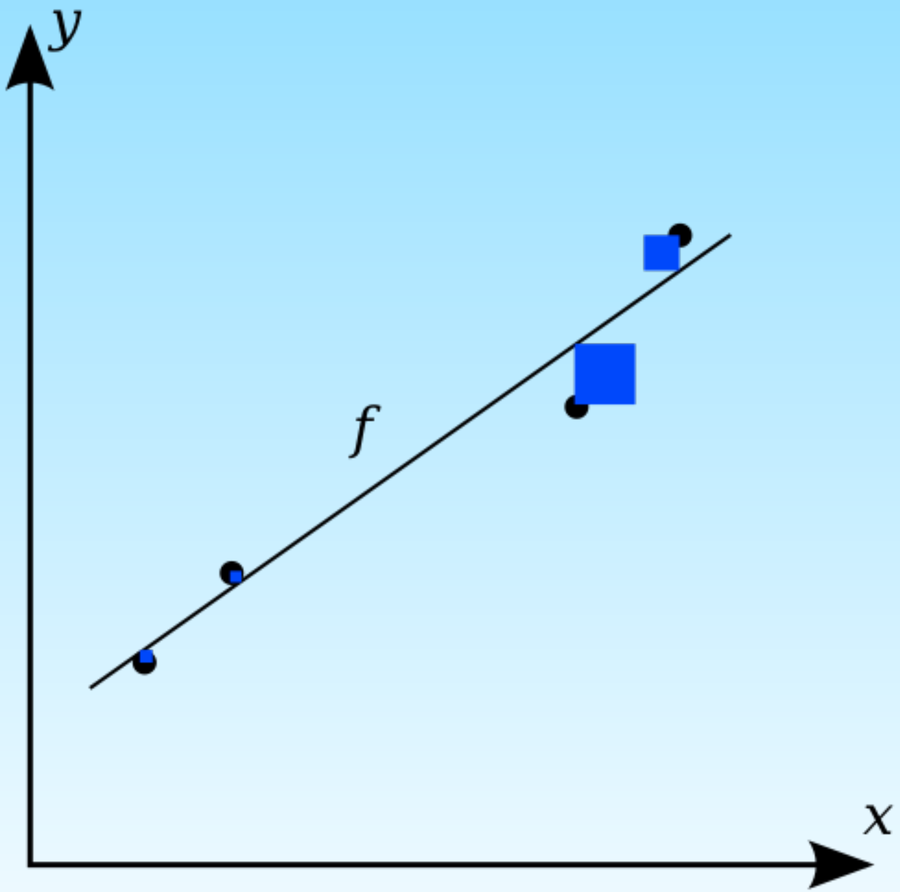
https://en.wikipedia.org/wiki/Coefficient_of_determination#/media/File:Coefficient_of_Determination.svg

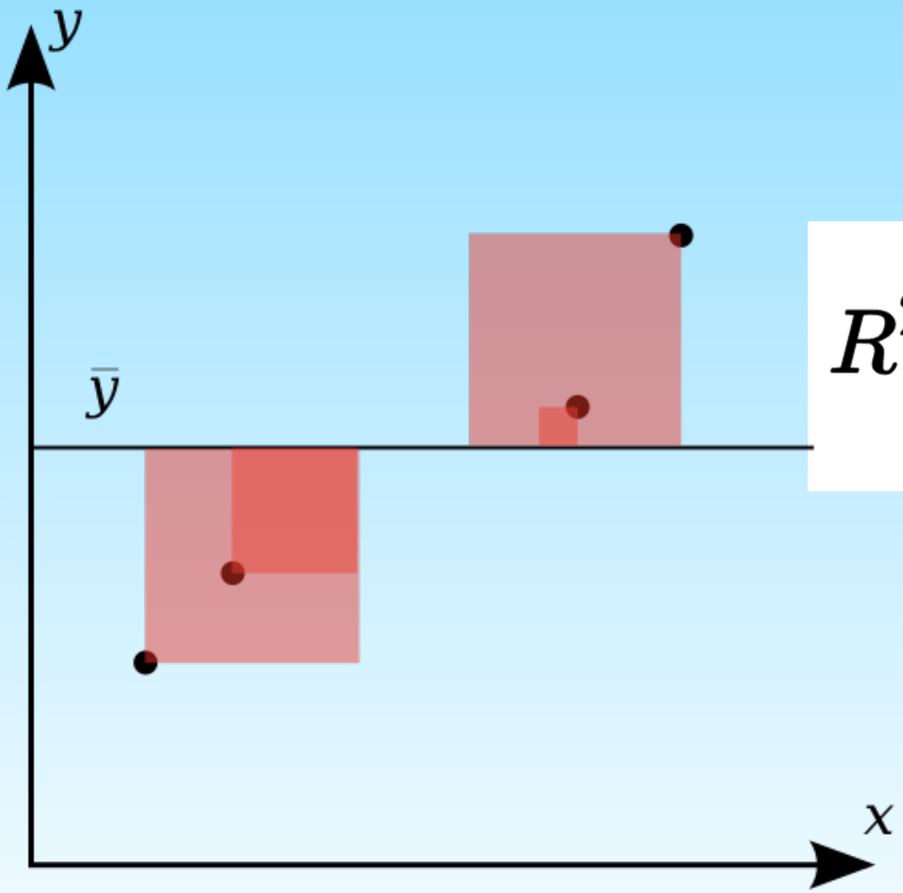
R^2 SCORE



https://en.wikipedia.org/wiki/Coefficient_of_determination#/media/File:Coefficient_of_Determination.svg

R^2 SCORE





Model

$$R^2 \equiv 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}.$$

https://en.wikipedia.org/wiki/Coefficient_of_determination#/media/File:Coefficient_of_Determination.svg

Data Weekends

Catalit LLC

LAB 03

Data Weekends

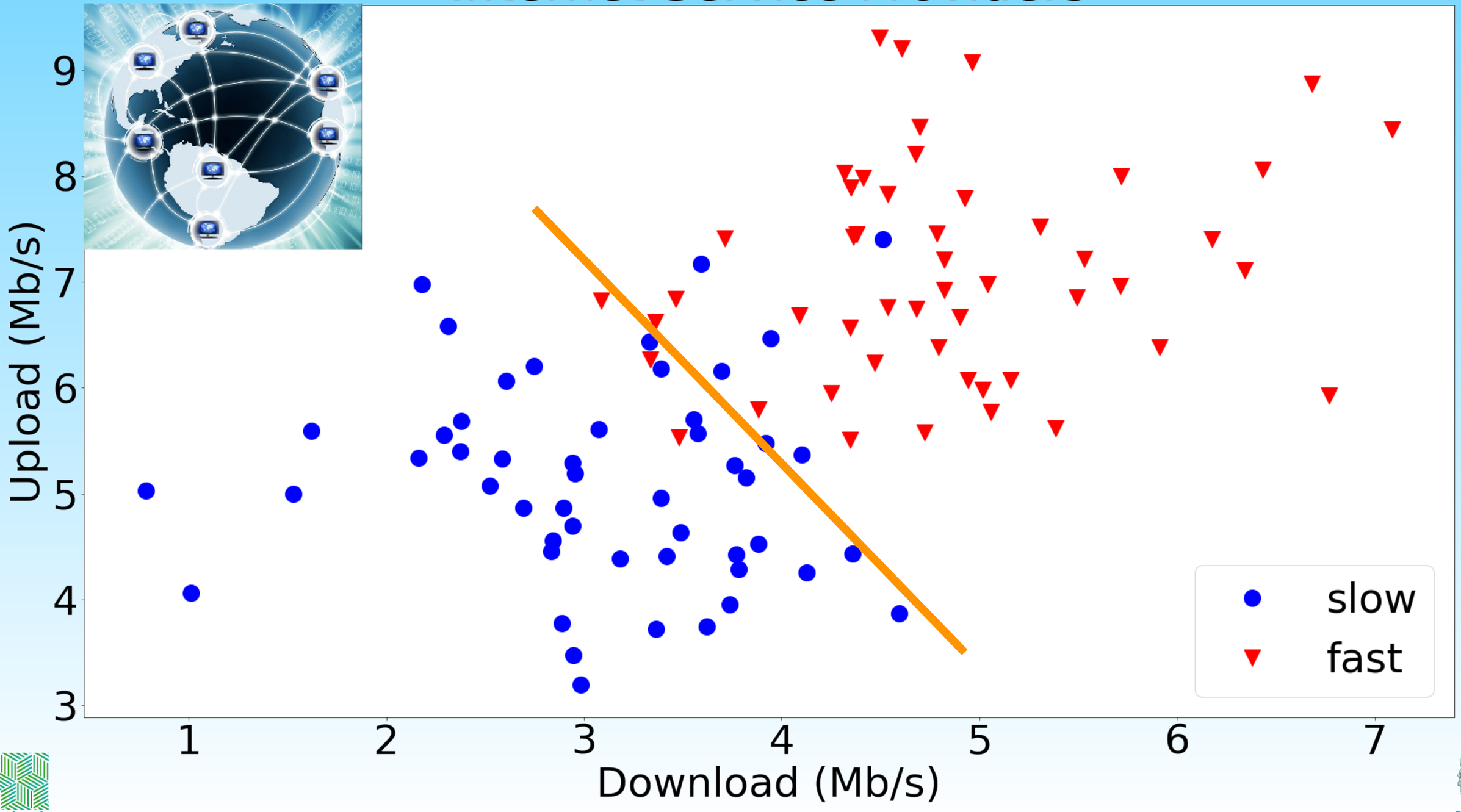
Catalit LLC

CLASSIFICATION

MAIN ML TECHNIQUES

	CATEGORICAL	CONTINUOUS
SUPERVISED	?	?
UNSUPERVISED	?	?

Internet Service Providers



Data Weekends

Catalit LLC

BINARY CLASSIFICATION

Features

Labels

Data Point

	Age	Gender	Annual Salary	Months in residence	Months in job	Current Debt	Paid off credit
Client 1	23	M	$30,000	36	12	$5,000	Yes
Client 2	30	F	$45,000	12	12	$1,000	Yes
Client 3	19	M	$15,000	3	1	$10,000	No

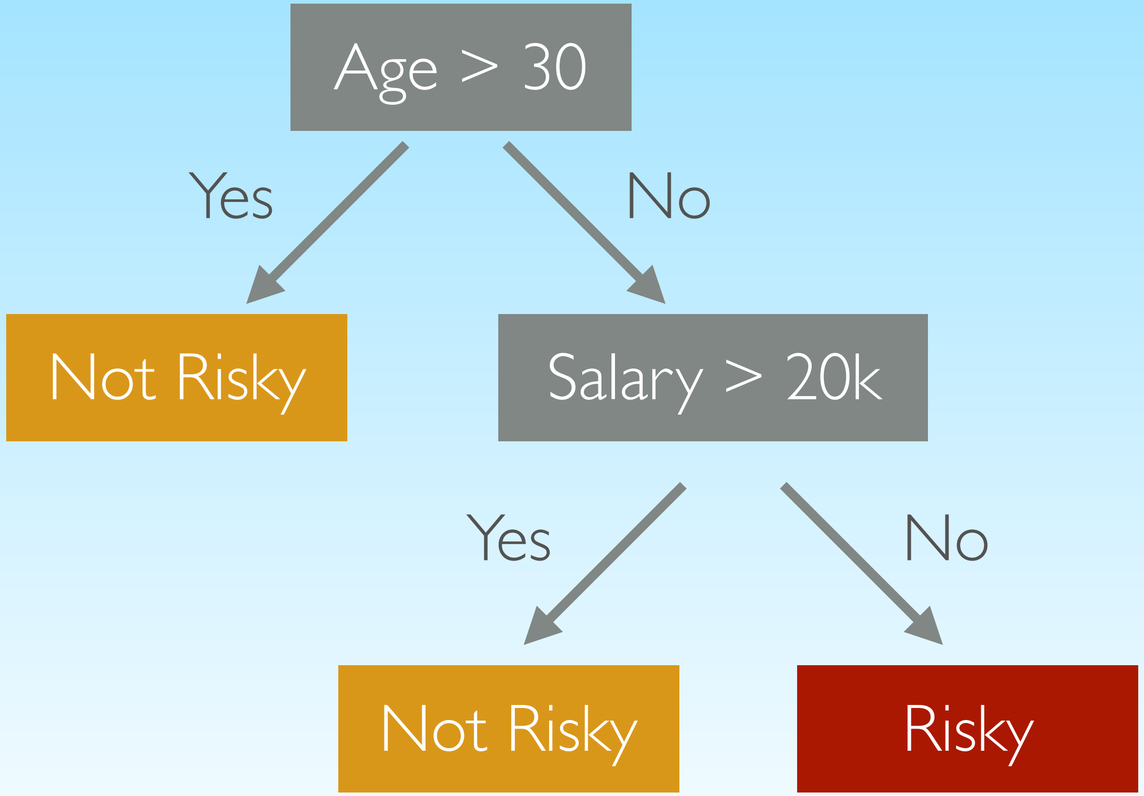
Data Weekends

Catalit LLC

CLASSIFIERS

DECISION TREE

- Advantages:
  - easy to interpret
  - fast prediction
  - rules based



DECISION TREE

Binary splits

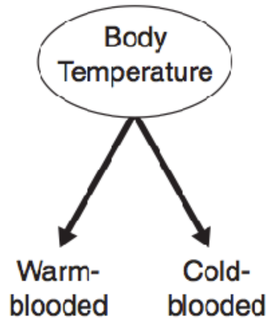
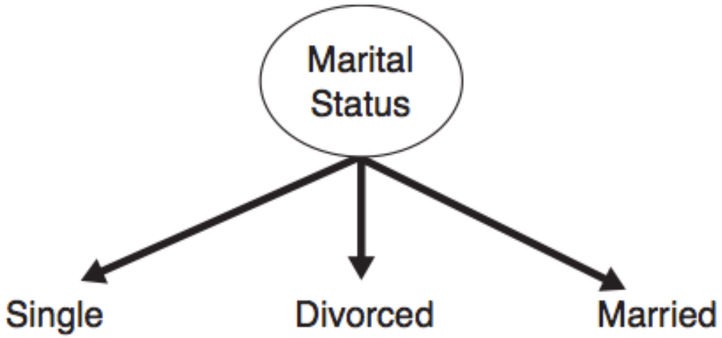


Figure 4.8. Test condition for binary attributes.

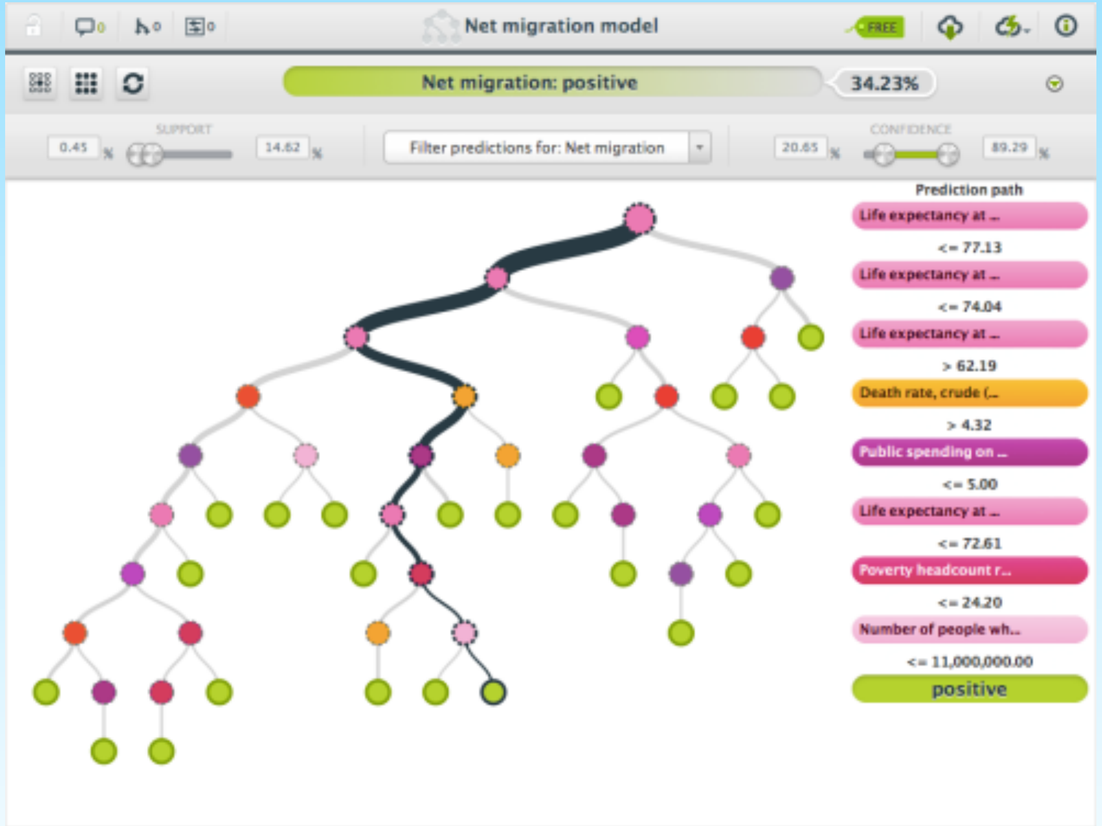
Multiway splits



(a) Multiway split

APPLICATIONS

- Lead scoring
- Customer churn
- User segmentation

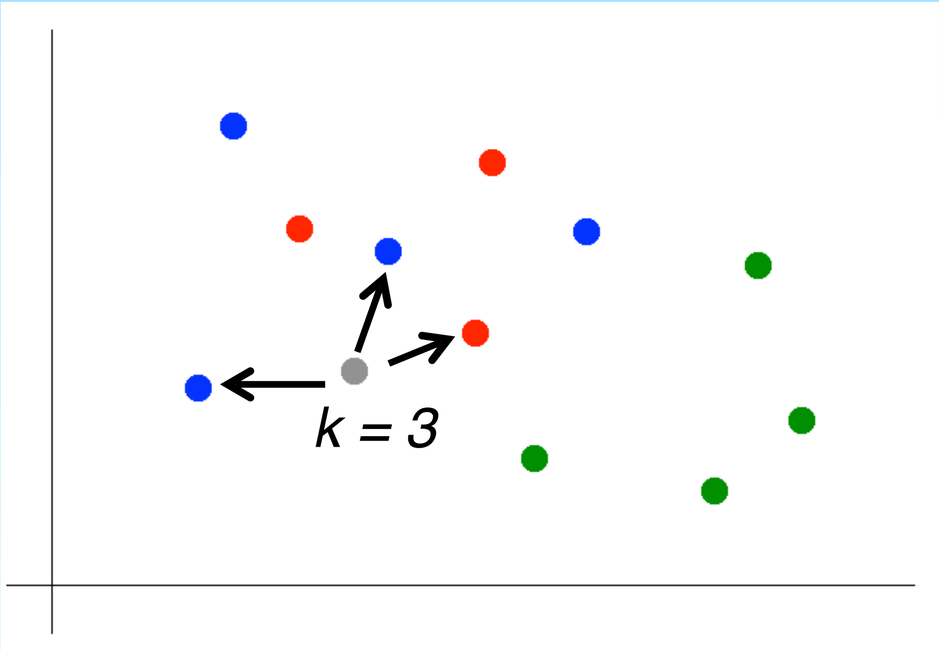


Data Weekends

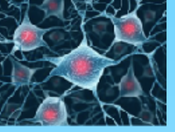
Catalit LLC

K-NEAREST NEIGHBORS

- Advantages:
  - easy to understand
  - any boundary

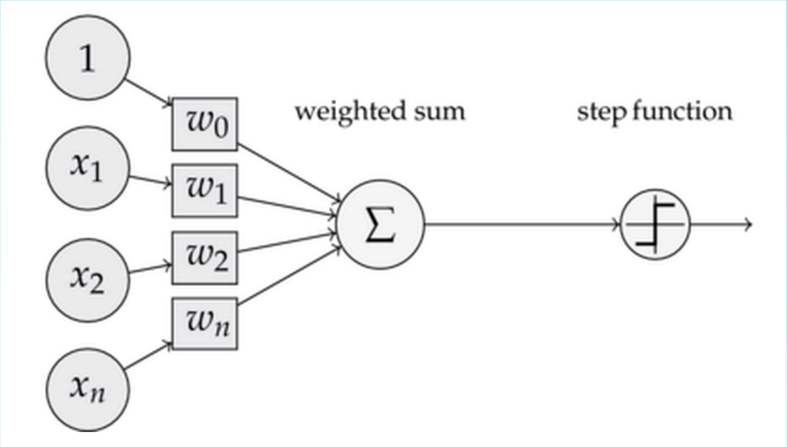


- GOAL: majority vote of nearby points

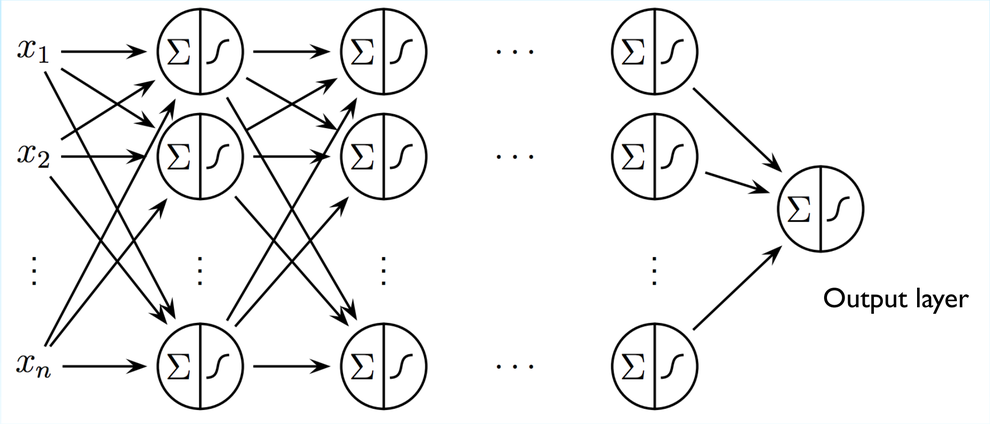


NEURAL NETWORKS

- Perceptron



- Multi-layer

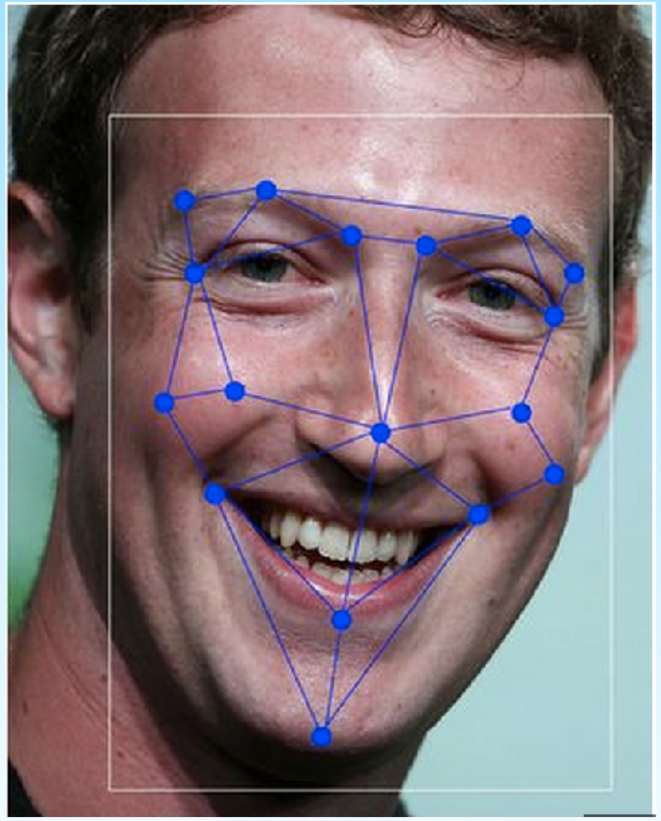


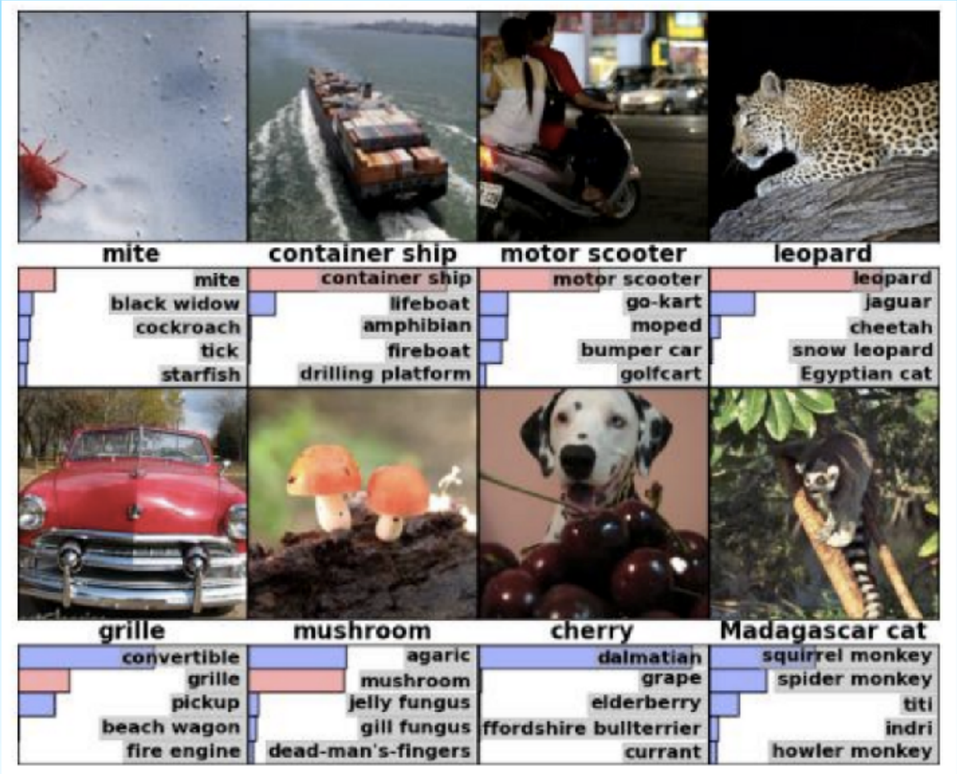
- GOAL: adjust weights to minimize output error

Data Weekends

Catalit LLC

APPLICATIONS OF NEURAL NETWORKS



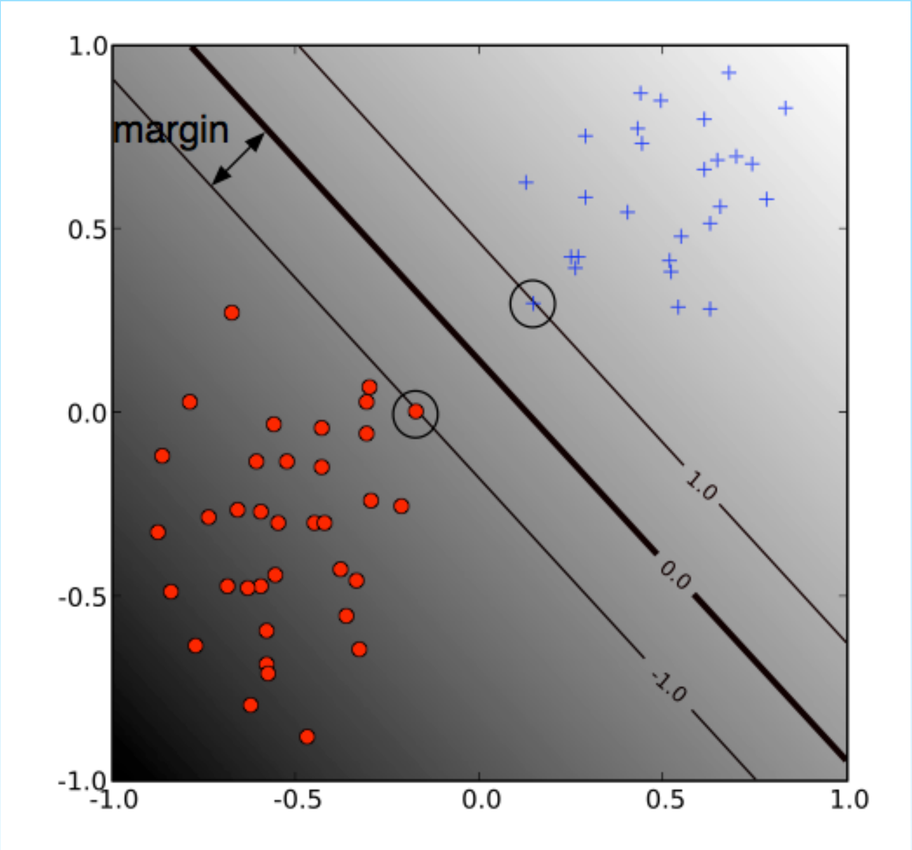


https://research.facebook.com/publications/480567225376225/deepface-closing-the-gap-to-human-level-performance-in-face-verification/

http://www.image-net.org/challenges/LSVRC/

SUPPORT VECTOR MACHINE

- Advantages:
  - powerful in non-linear boundaries
  - transform data to higher dimension



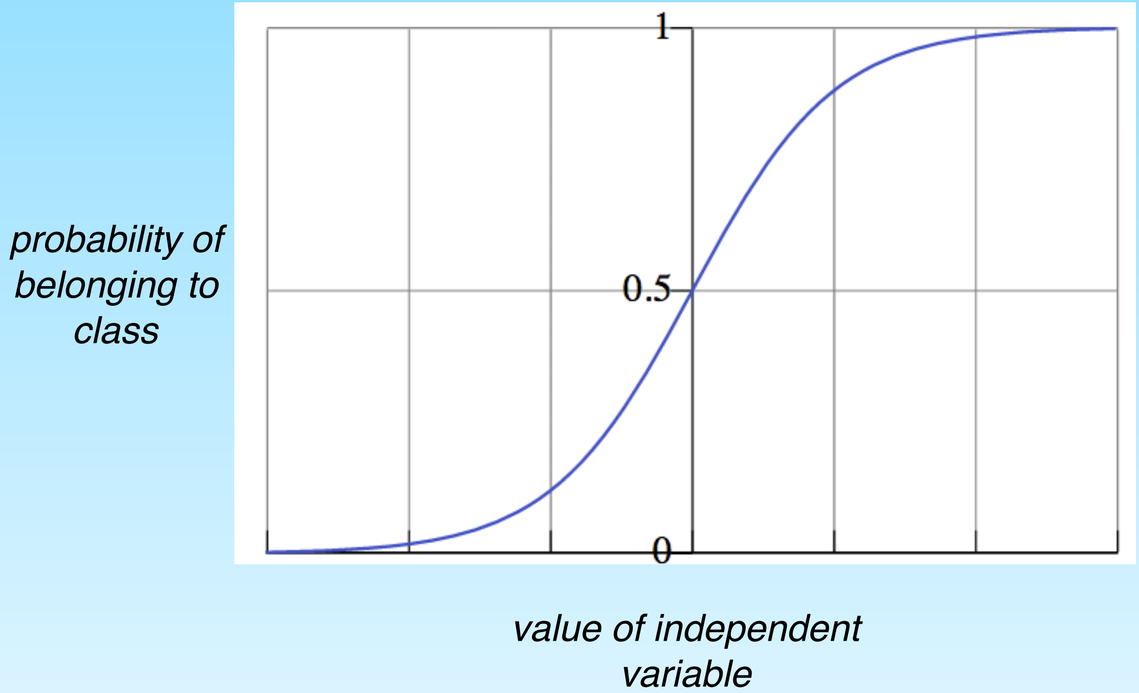
GOAL: Search for a separating hyperplane that maximizes margin

APPLICATIONS OF SVM

- Transmission
  - IP address —167.12.24.555
  - Sender URL -- one-spam.com
- Email header
  - From --"admin@one-spam.cpm"
  - To --"undisclosed"
  - cc
- Email Body
  - # of paragraphs
  - # words
- Email structure
  - # of attachments
  - # of links

LOGISTIC REGRESSION

- Advantages:
  - simple function
  - can be parallelized
  - large scale



GOAL: minimize squared error for probability

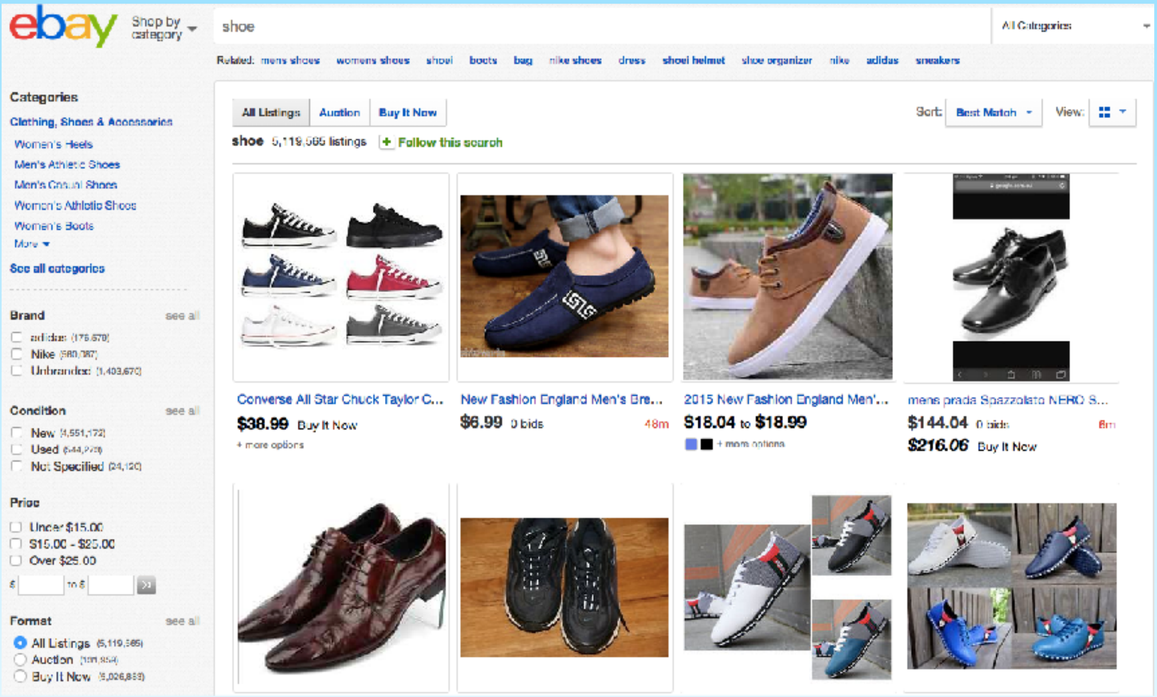
Data Weekends

Catalit LLC

APPLICATIONS

- Click prediction
  - Search ranking (web pages, products)
  - Online advertising
  - Recommendation

- The model
  - Output: Click/no click
  - Input features:
    - page content
    - search keyword
    - User information



Data Weekends

Catalit LLC

ENSEMBLE METHODS

Bagging
Boosting
Random Forest

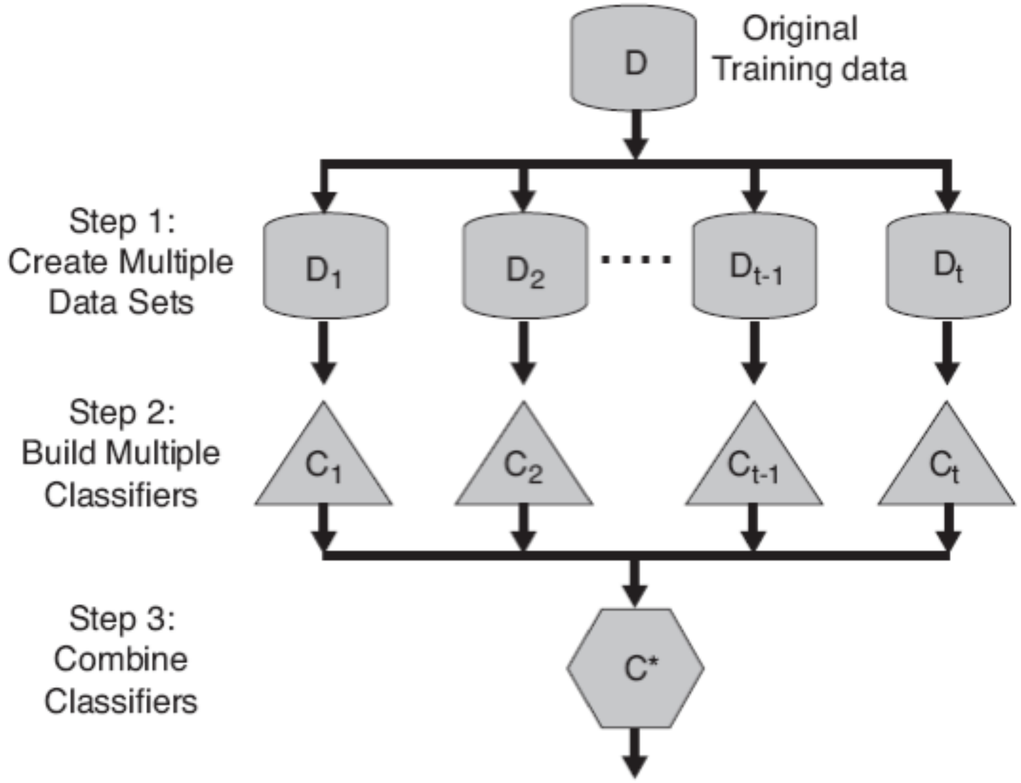
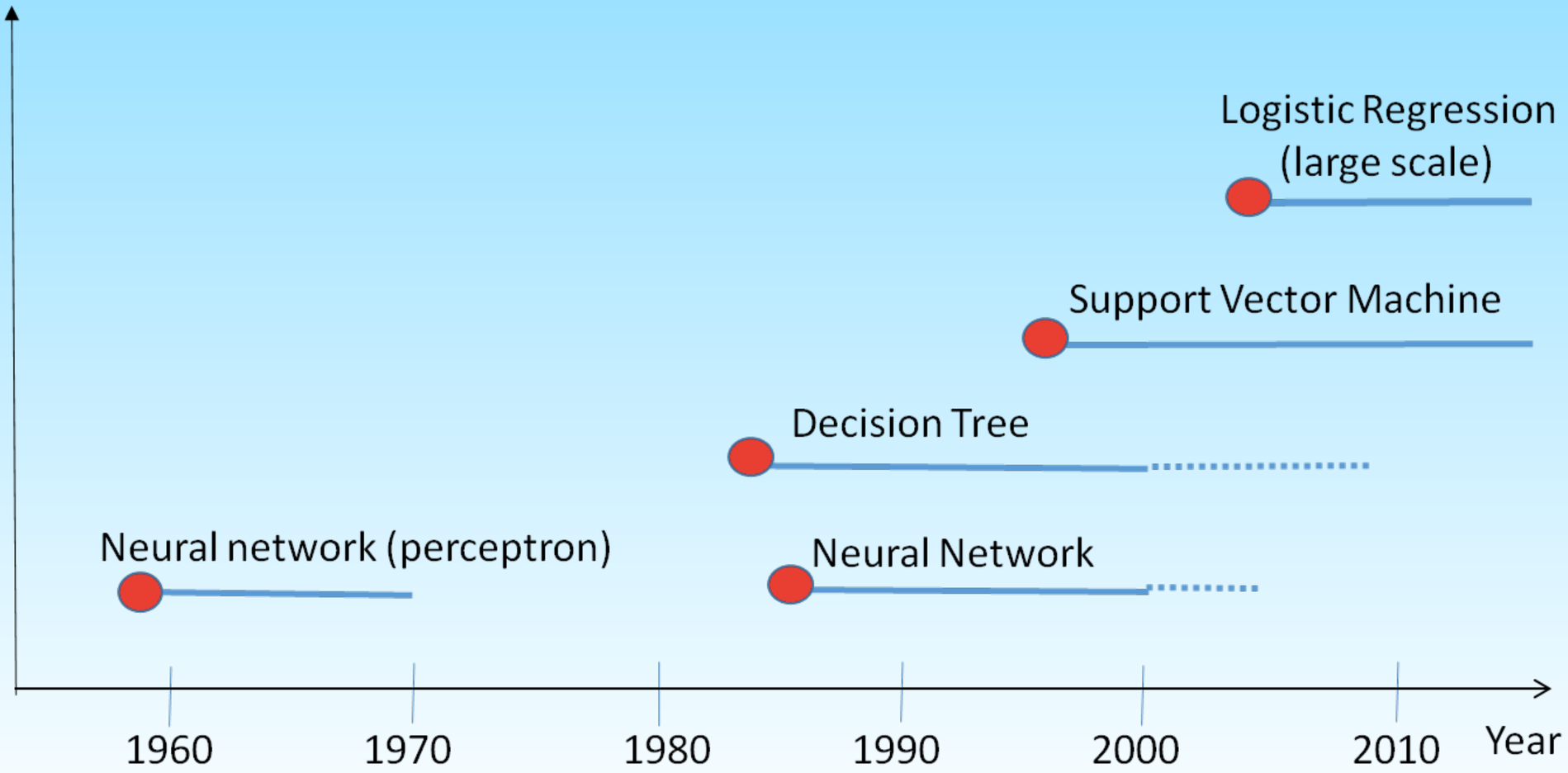


Figure 5.31. A logical view of the ensemble learning method.

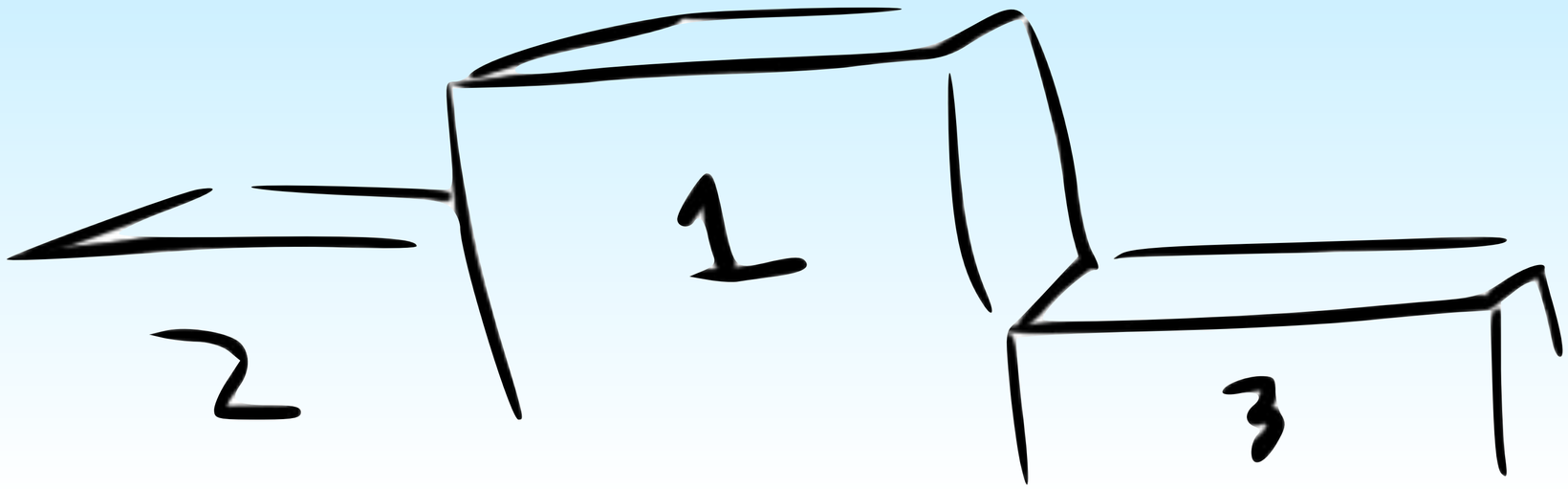
HISTORY OF SUPERVISED LEARNING



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

BASELINE

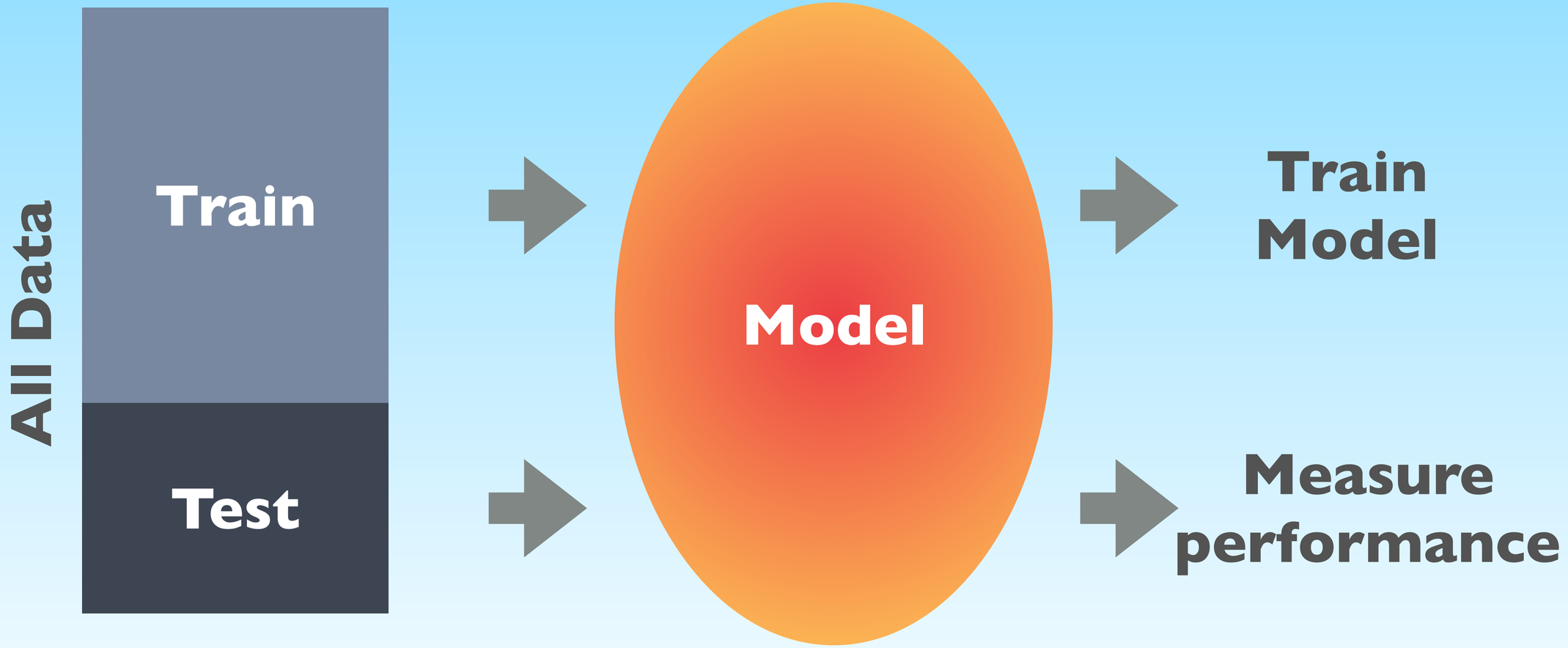
- Know the score of the simplest model
- Compare your score with simplest model score



Data Weekends

Catalit LLC

TRAIN / TEST



Typical values of test size: [10% - 50%]

CONFUSION MATRIX

	Test Negative	***Test Positive***
Condition Negative	***TRUE NEGATIVE***	***FALSE POSITIVE (Type I error)***
Condition Positive	***FALSE NEGATIVE (Type II error)***	***TRUE POSITIVE***

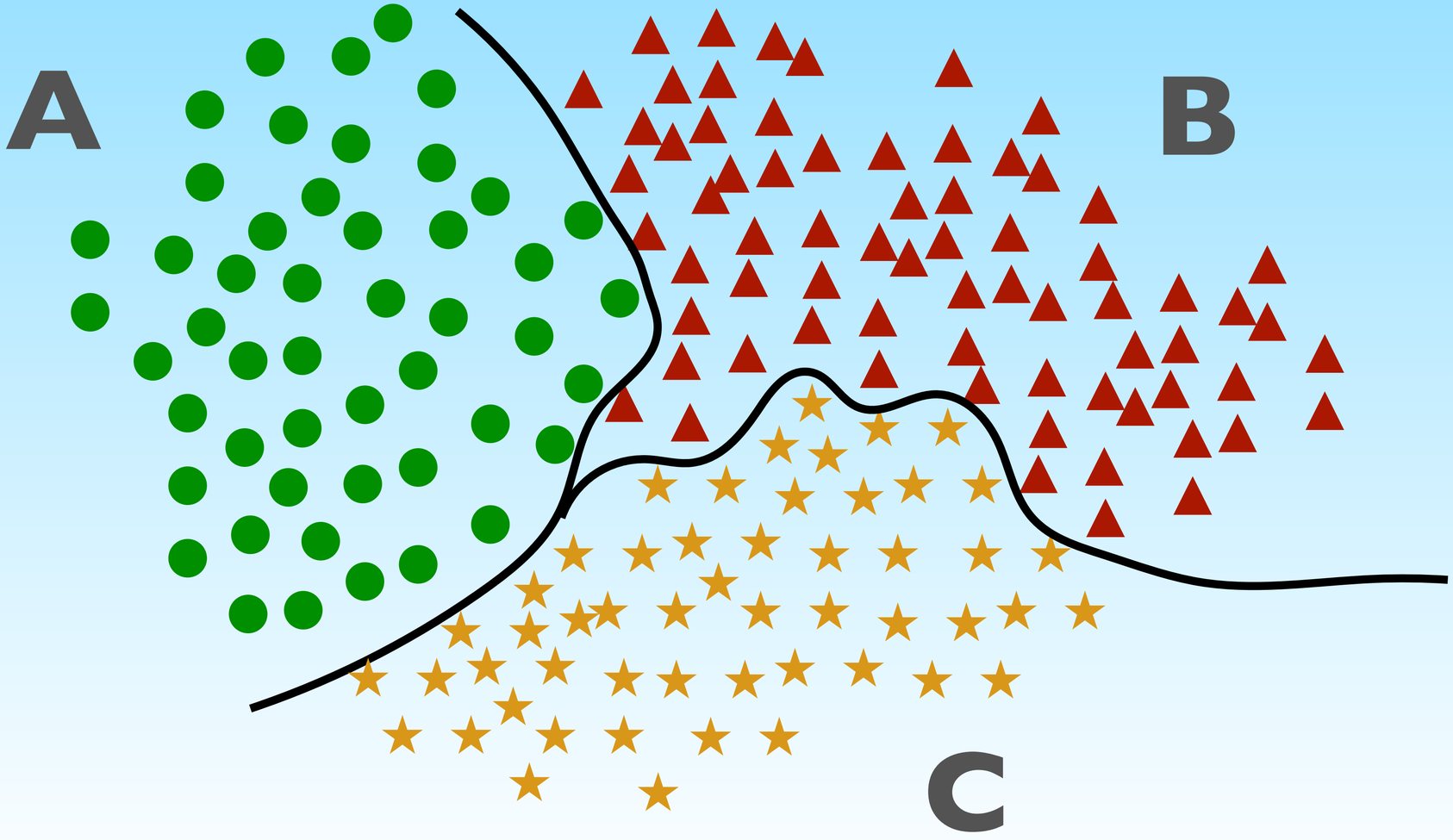
- **Accuracy:** Overall, how often is it correct?
  - (TP + TN) / total

CLASSIFICATION SCORES

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

	Test Negative	*Test Positive*
Condition Negative	*TRUE NEGATIVE*	*FALSE POSITIVE (Type I error)*
Condition Positive	*FALSE NEGATIVE (Type II error)*	*TRUE POSITIVE*

MULTI-CLASS



Label
A
A
B
C

MULTI CLASS

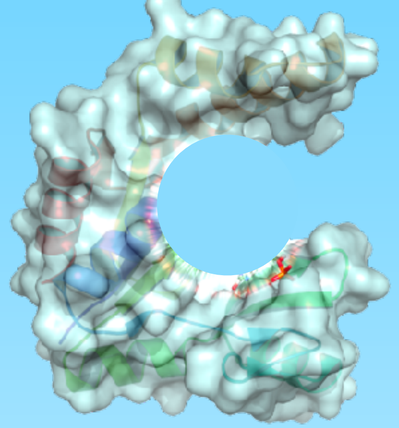
	Prediction A	Prediction B	Prediction C	Prediction D
Class A	84			
Class B		69	3	
Class C	2		78	
Class D				91

Data Weekends

Catalit LLC

LAB 04

Data Weekends



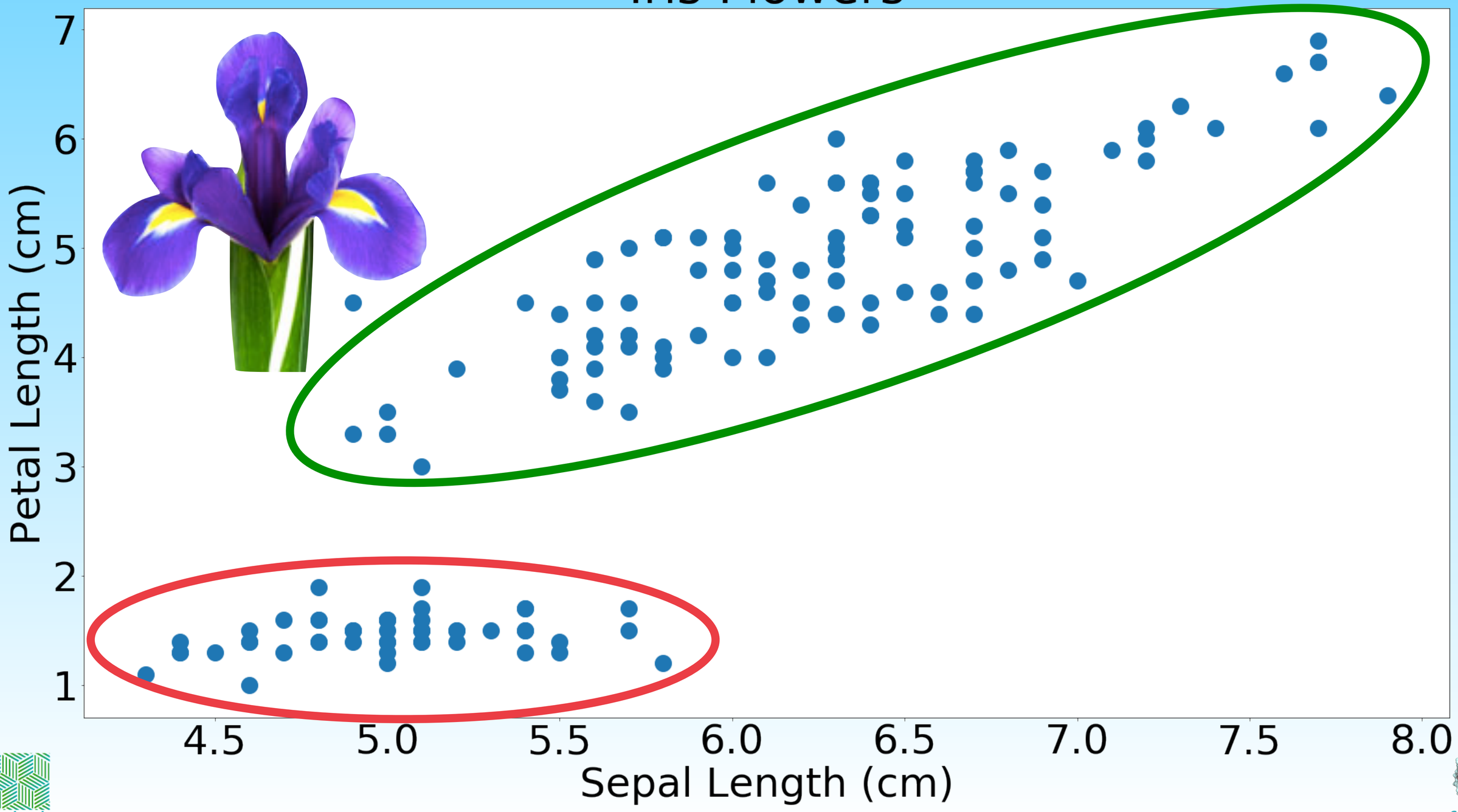
Catalit LLC

CLUSTERING

MAIN ML TECHNIQUES

	CATEGORICAL	CONTINUOUS
SUPERVISED	?	?
UNSUPERVISED	?	?

Iris Flowers



Data Weekends

Catalit LLC

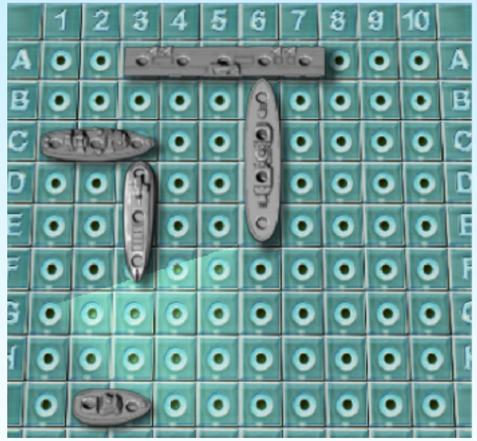
GOAL OF CLUSTERING

Data Weekends

Catalit LLC

DISTANCE & SIMILARITY

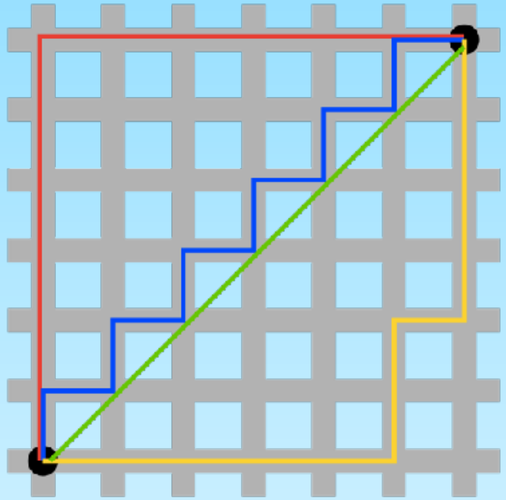
	Age	Gender	Annual Salary	Months in residence	Months in job	Current Debt
Client 1	23	M	$30,000	36	12	$5,000
Client 2	30	F	$45,000	12	12	$1,000
Client 3	19	M	$15,000	3	1	$10,000

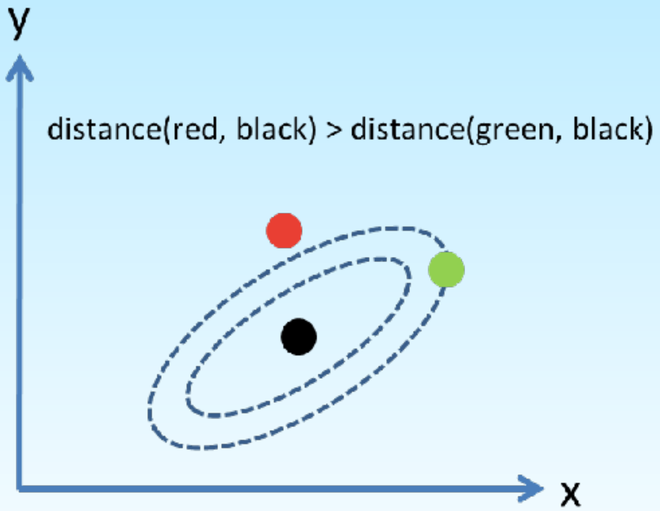


Data Weekends

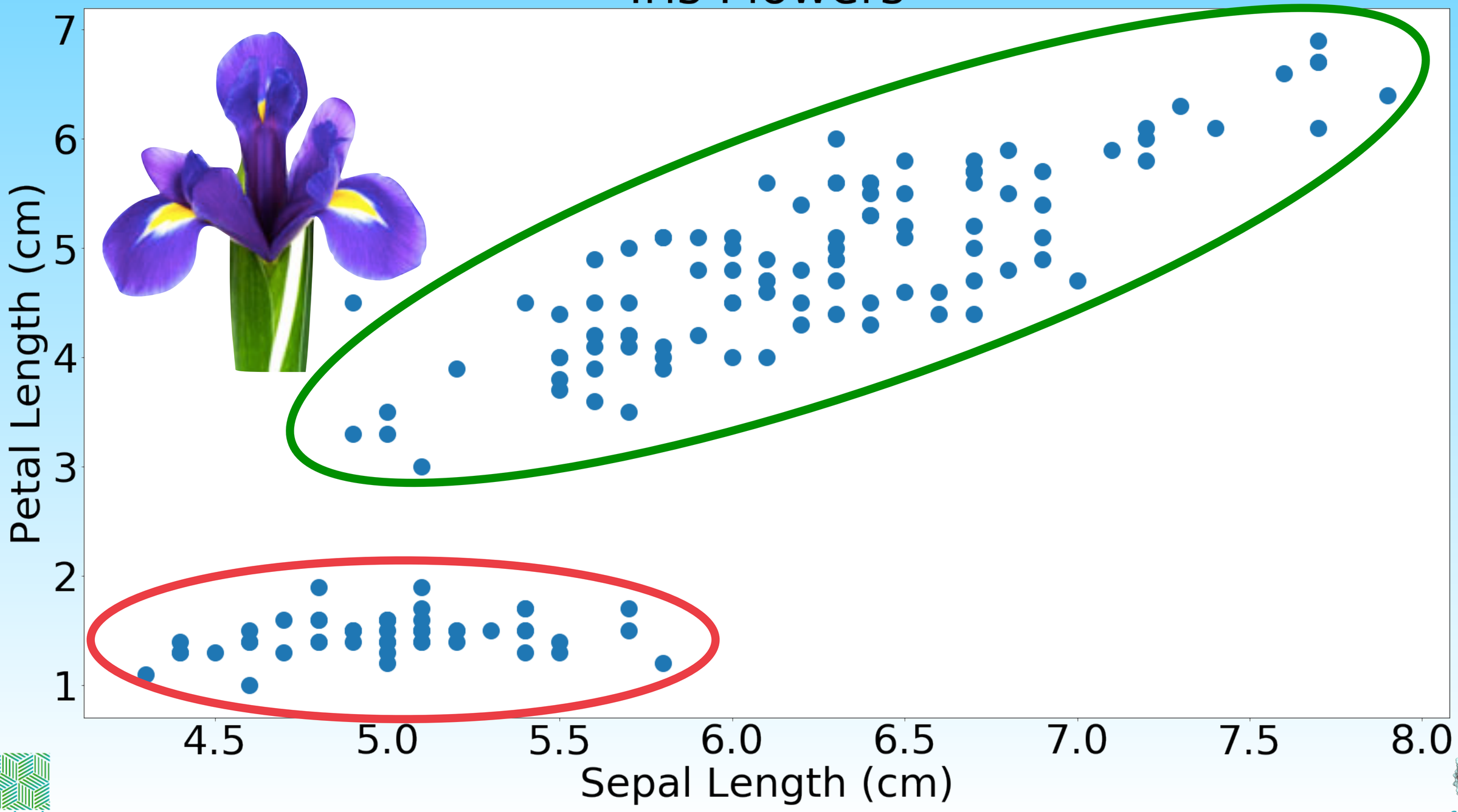
Catalit LLC

DISTANCE & SIMILARITY





Iris Flowers

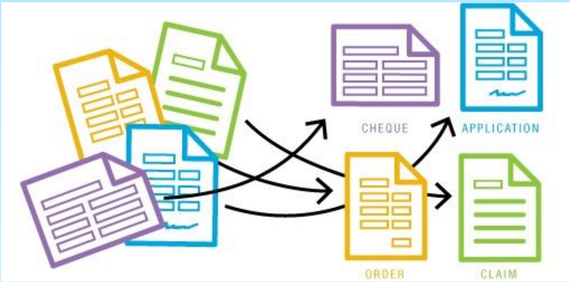


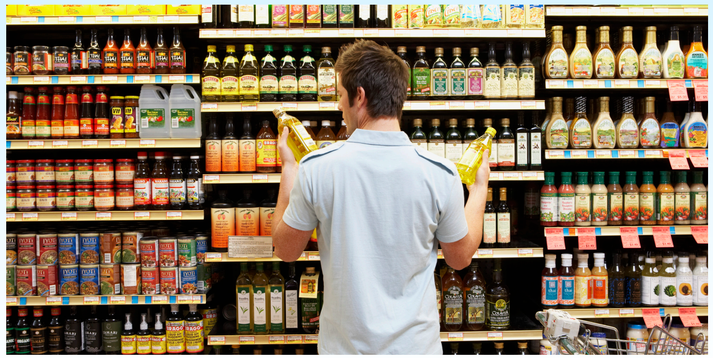
Data Weekends

Catalit LLC

APPLICATIONS

- Malware detection
- Intrusion detection
- Document clustering
- Product placing
- Customer segmentation
- Gene discovery

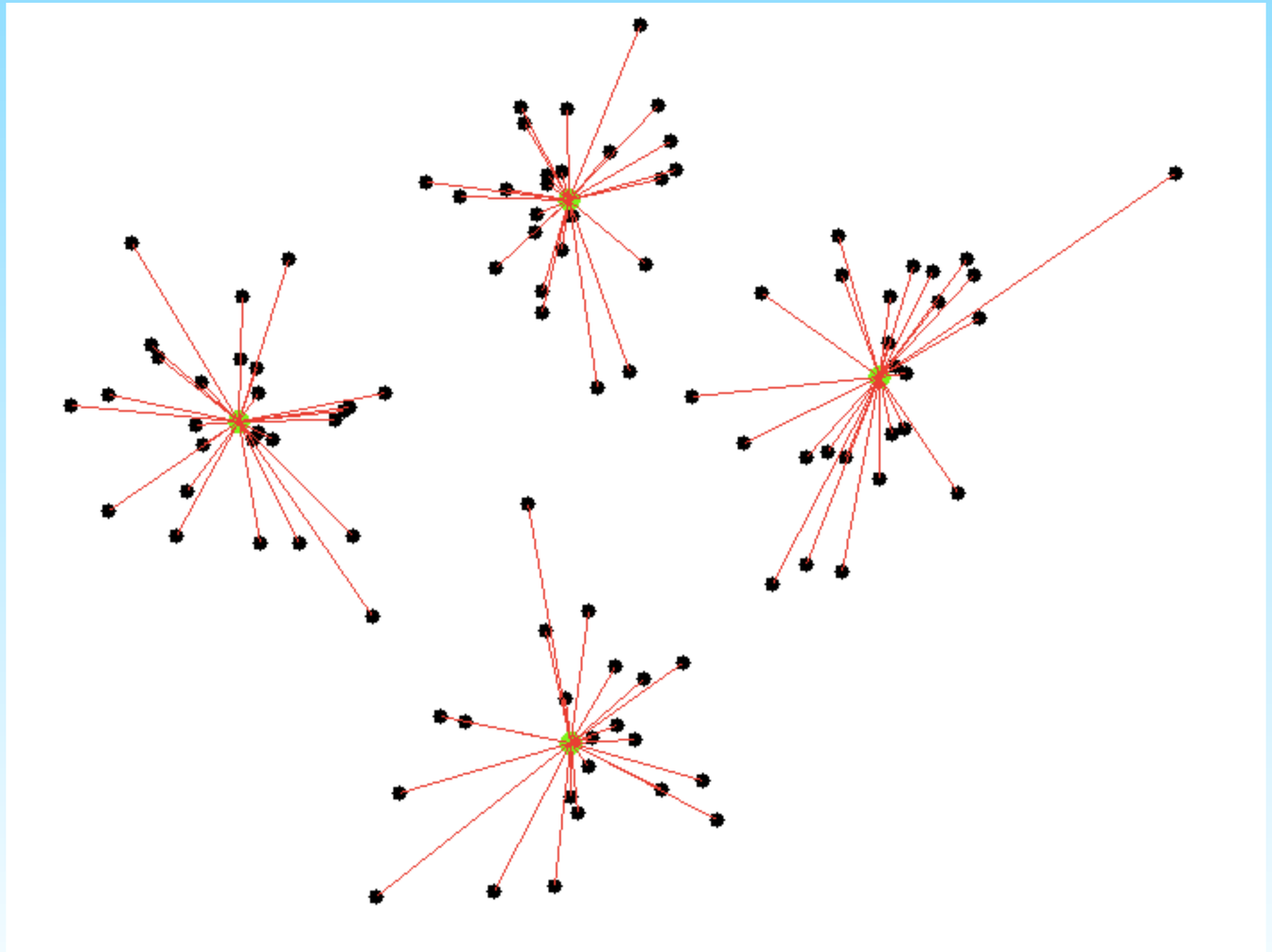




Data Weekends

Catalit LLC

K-MEANS

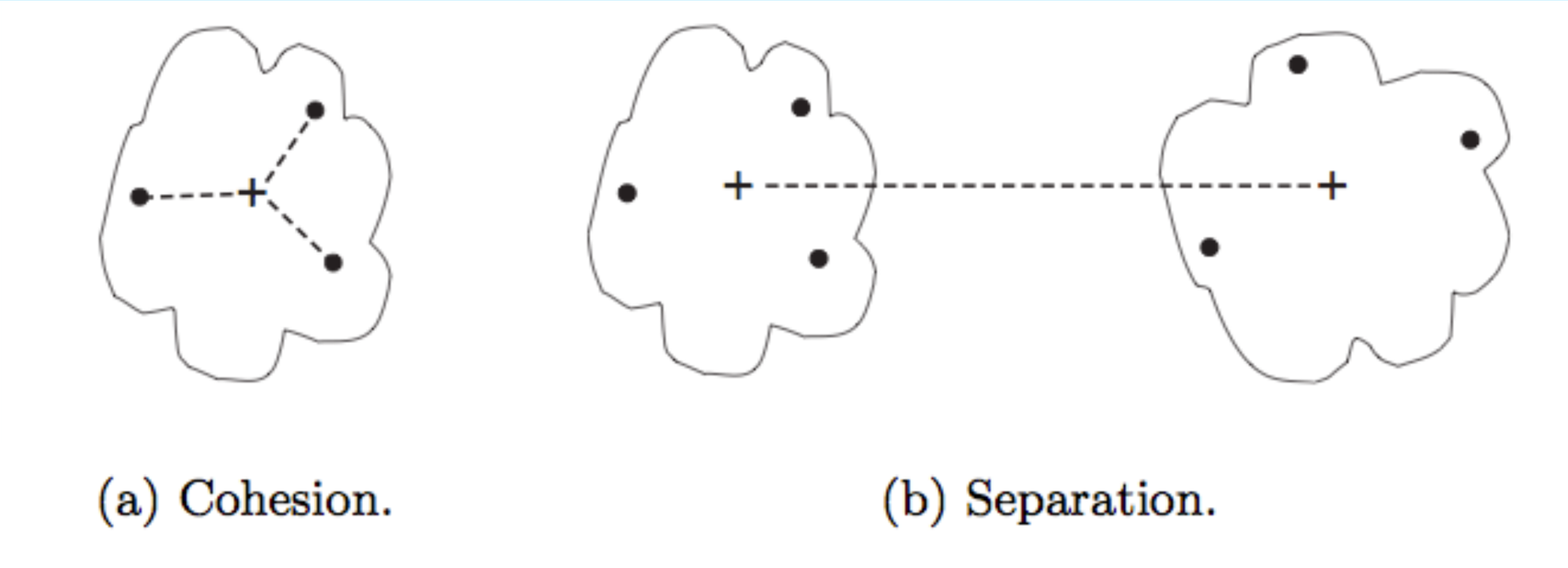


http://shabal.in/visuals/kmeans/1.html

Data Weekends

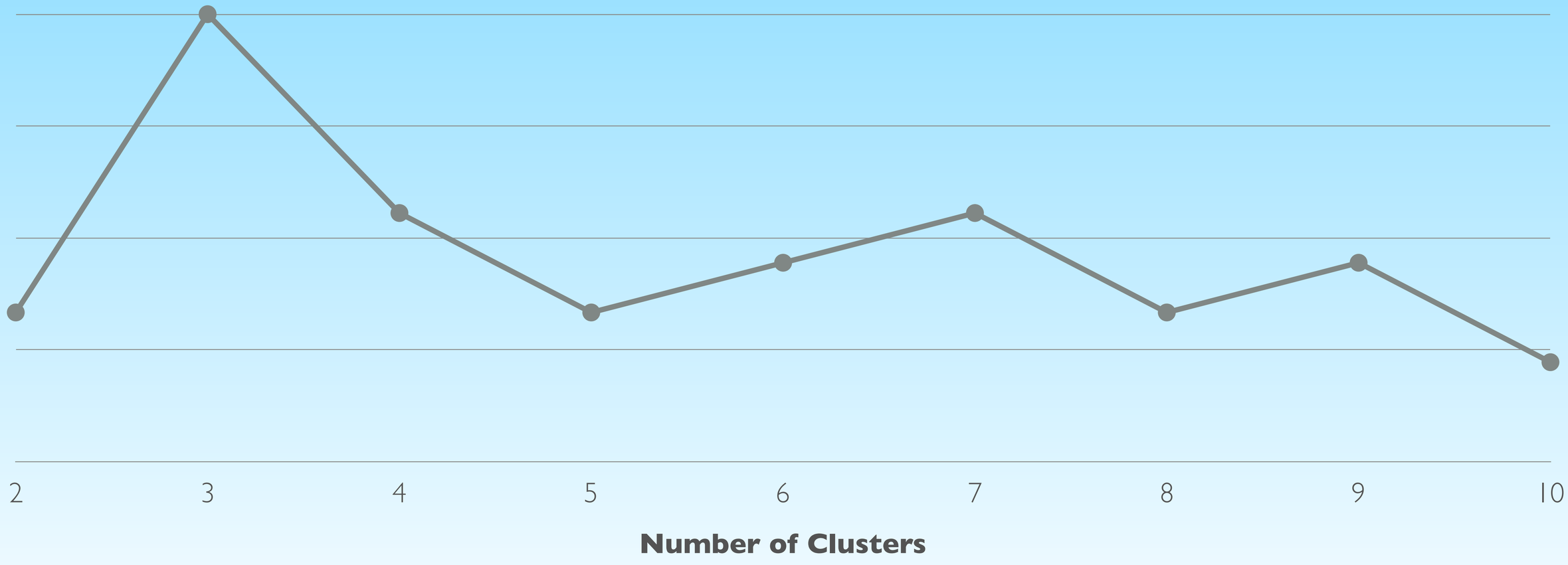
Catalit LLC

CLUSTER VALIDATION



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

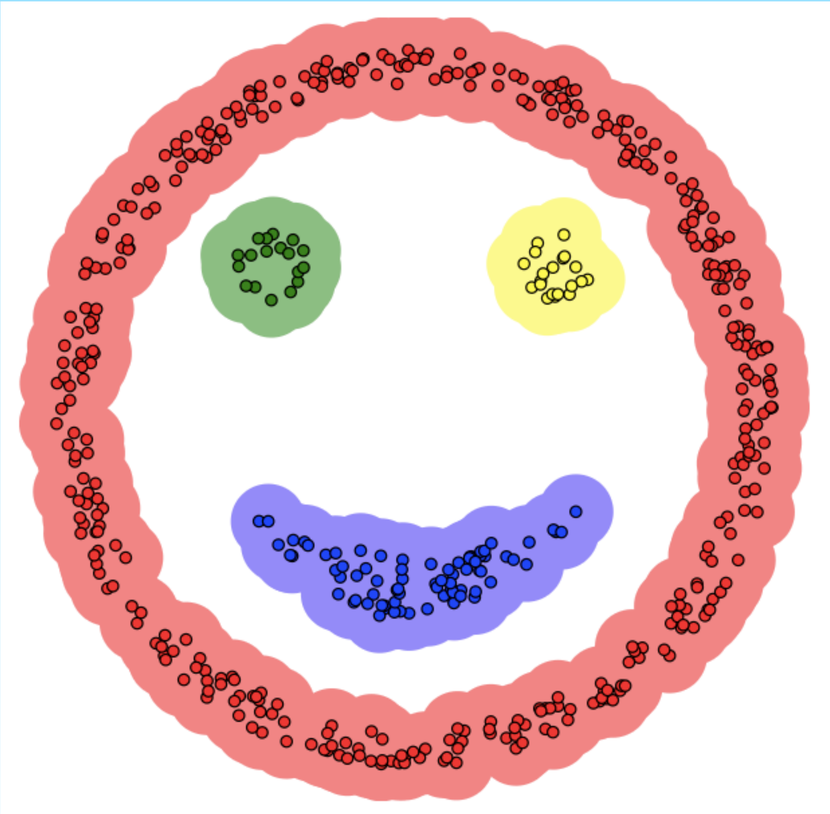
SILHOUETTE



Data Weekends

Catalit LLC

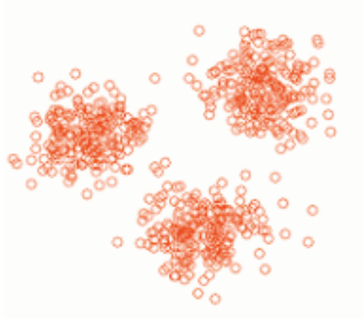
DBSCAN



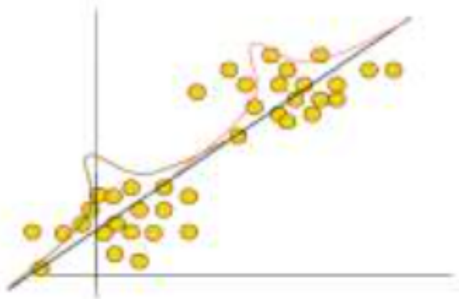
https://www.naftaliharris.com/blog/visualizing-dbscan-clustering/

METHODS FOR CLUSTERING

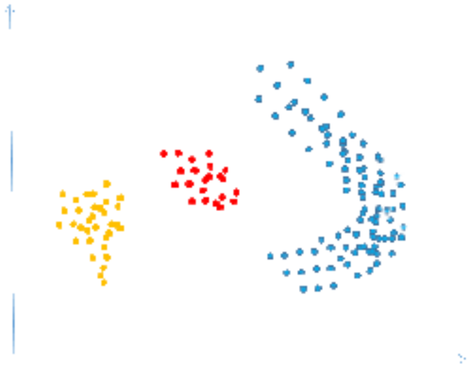
- K-means



> Expectation Maximization



- Density-based



> Hierarchical

LAB 05