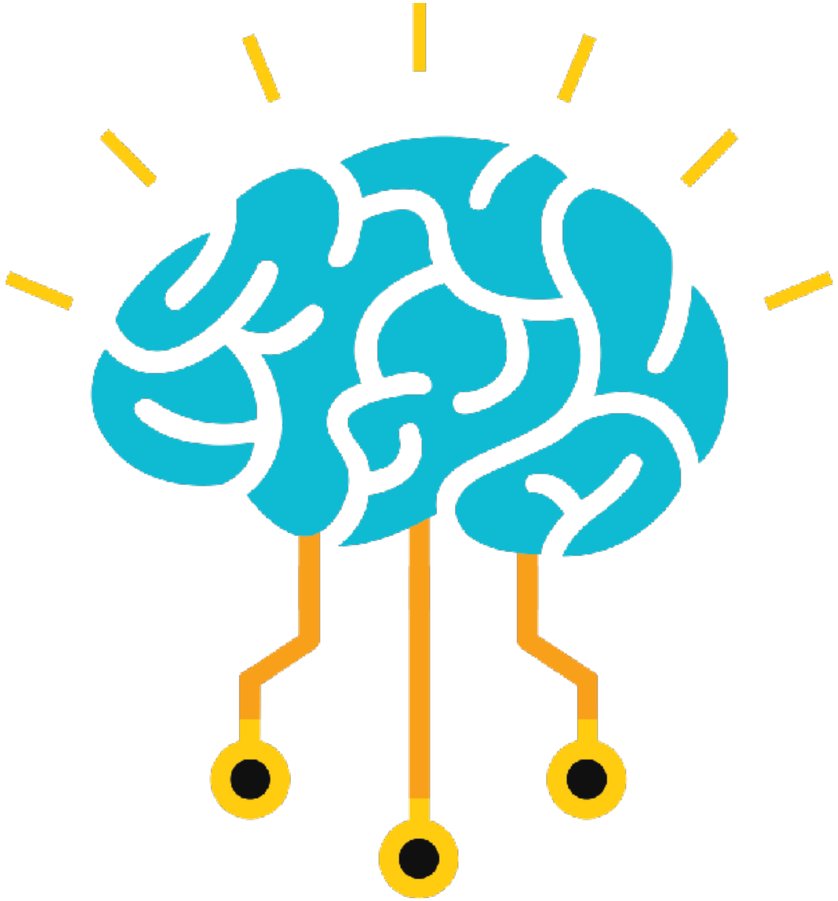
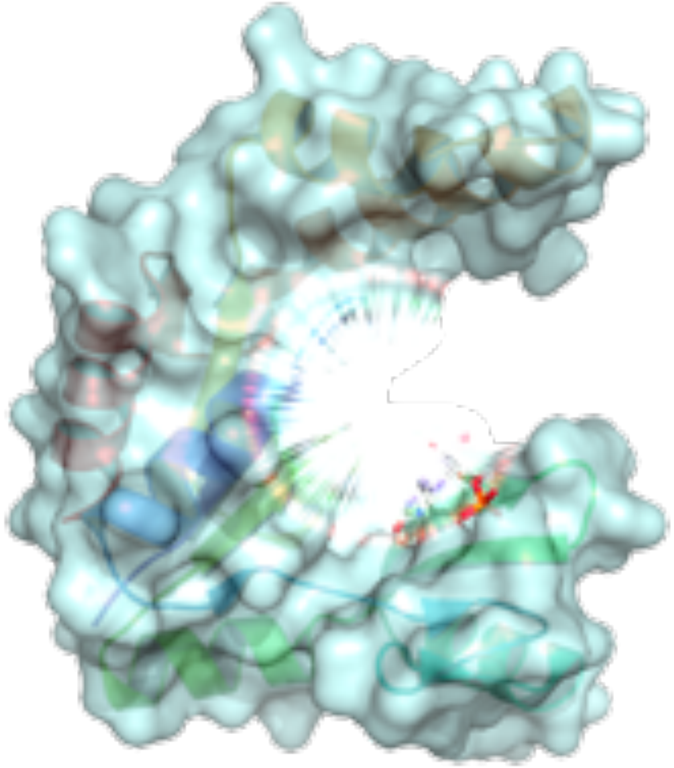
Intro to Machine Learning

Francesco Mosconi
Zero to Deep Learning / CATALIT LLC

Catalit LLC

Who are we





www.zerotodeeplearning.com

www.catalit.com

Catalit LLC

A few tips for today …

Ask questions

Catalit LLC

Interact with your classmates

Catalit LLC

Stay hydrated

Catalit LLC

Schedule

• **09:00 am Intro to ML**	45’ Lecture + 45’ Lab
• **10:30 am Break**	30’
• **11:00 am Regression**	45’ Lecture + 45’ Lab
• **12:30 pm Lunch**	60’
• **01:30 pm Classification**	45’ Lecture + 45’ Lab
• **03:00 pm Break**	30’
• **03:30 pm Unsupervised**	45’ Lecture + 45’ Lab
• **05:00 pm End**	

Catalit LLC

What to expect

I know machine learning !

Catalit LLC

What to expect

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

Catalit LLC

Machine Learning

Machine Learning

- Intuition, definition and workflow
- Supervised VS Unsupervised
- Applications
- Tabular data & Label types
- Why Python
- Lab

Catalit LLC

Your Brain

http://scriptoriumdai

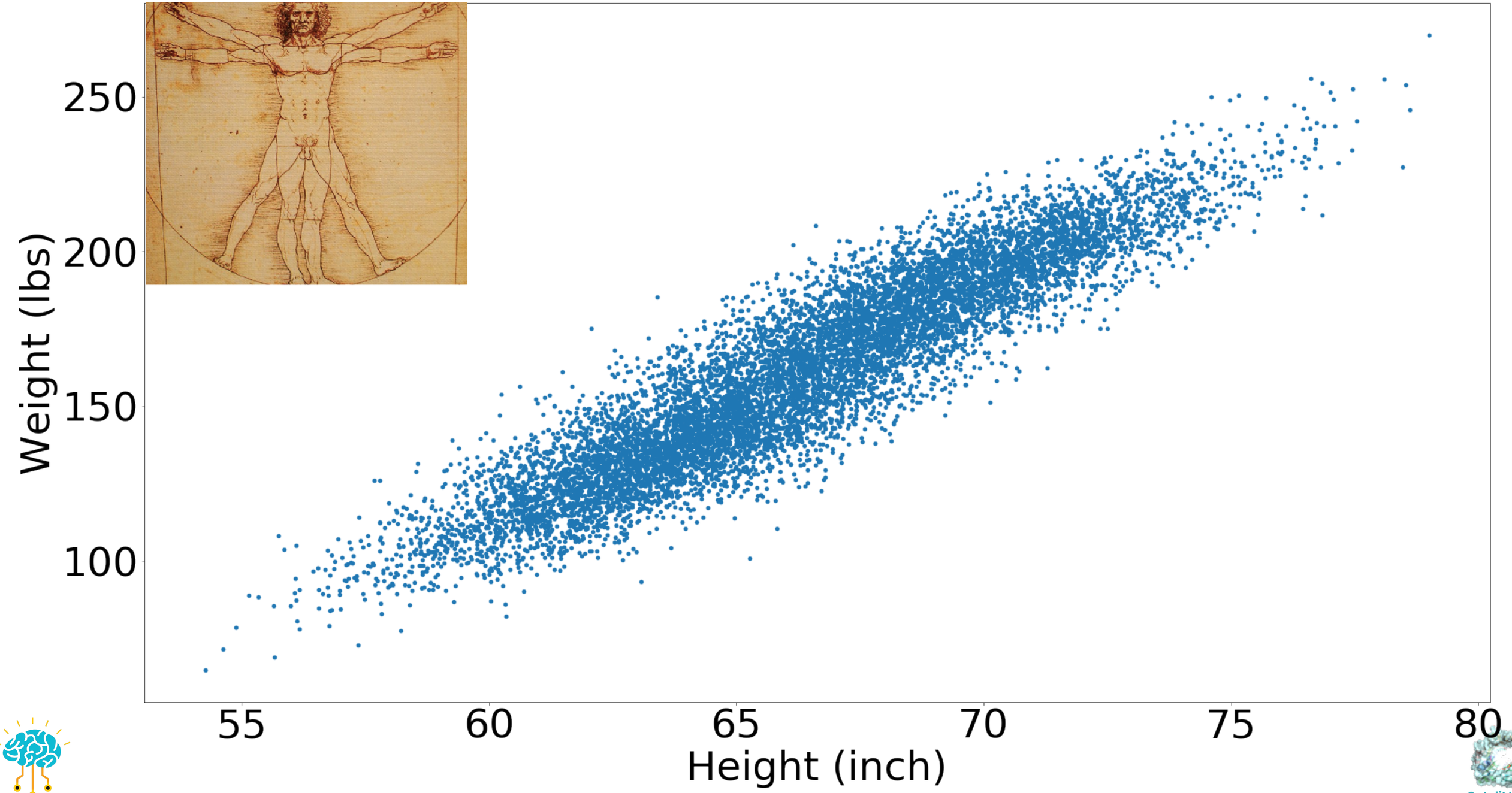
...A GREAT

PATTERN

RECOGNIZER

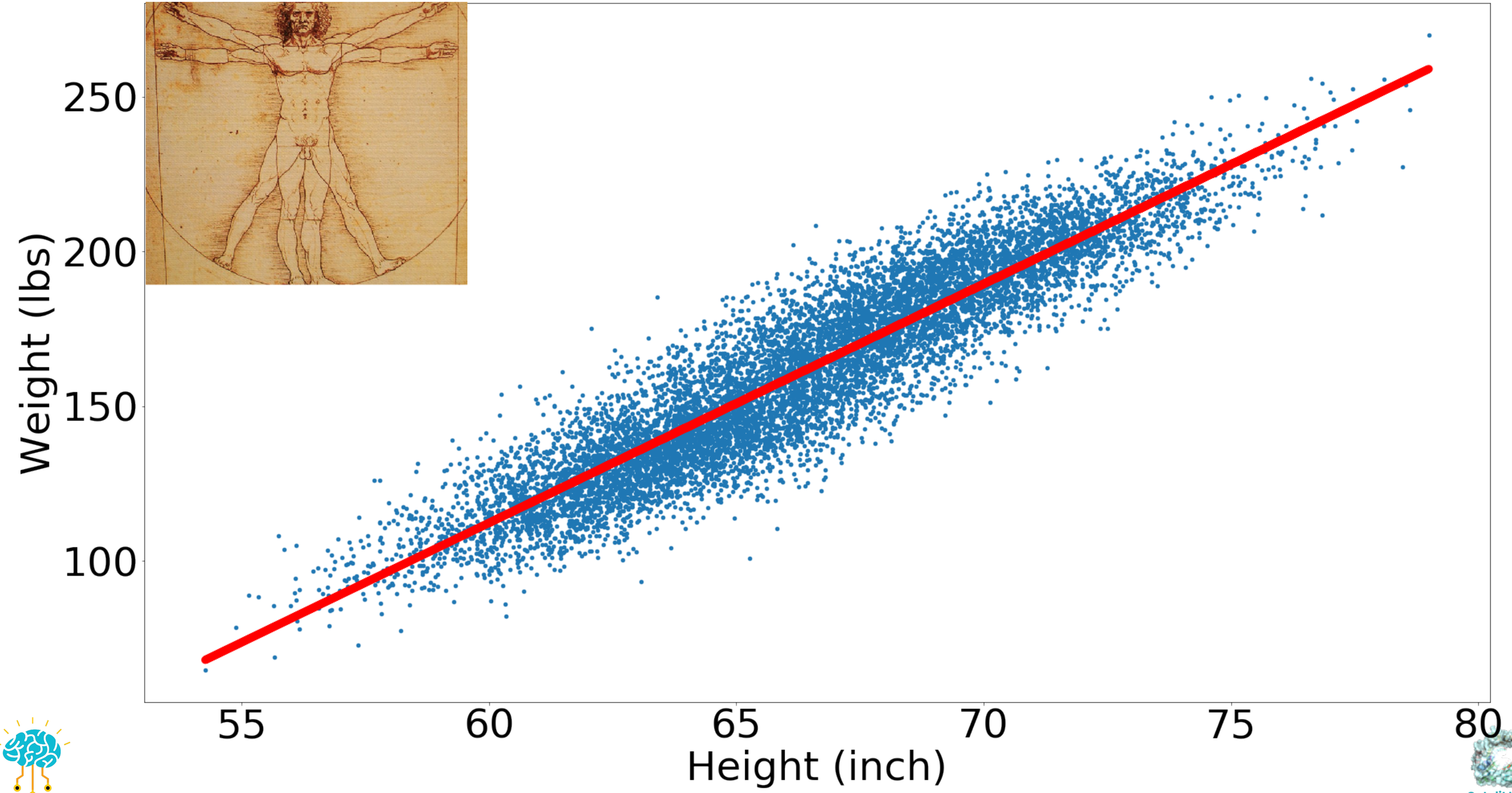
Catalit LLC

Humans



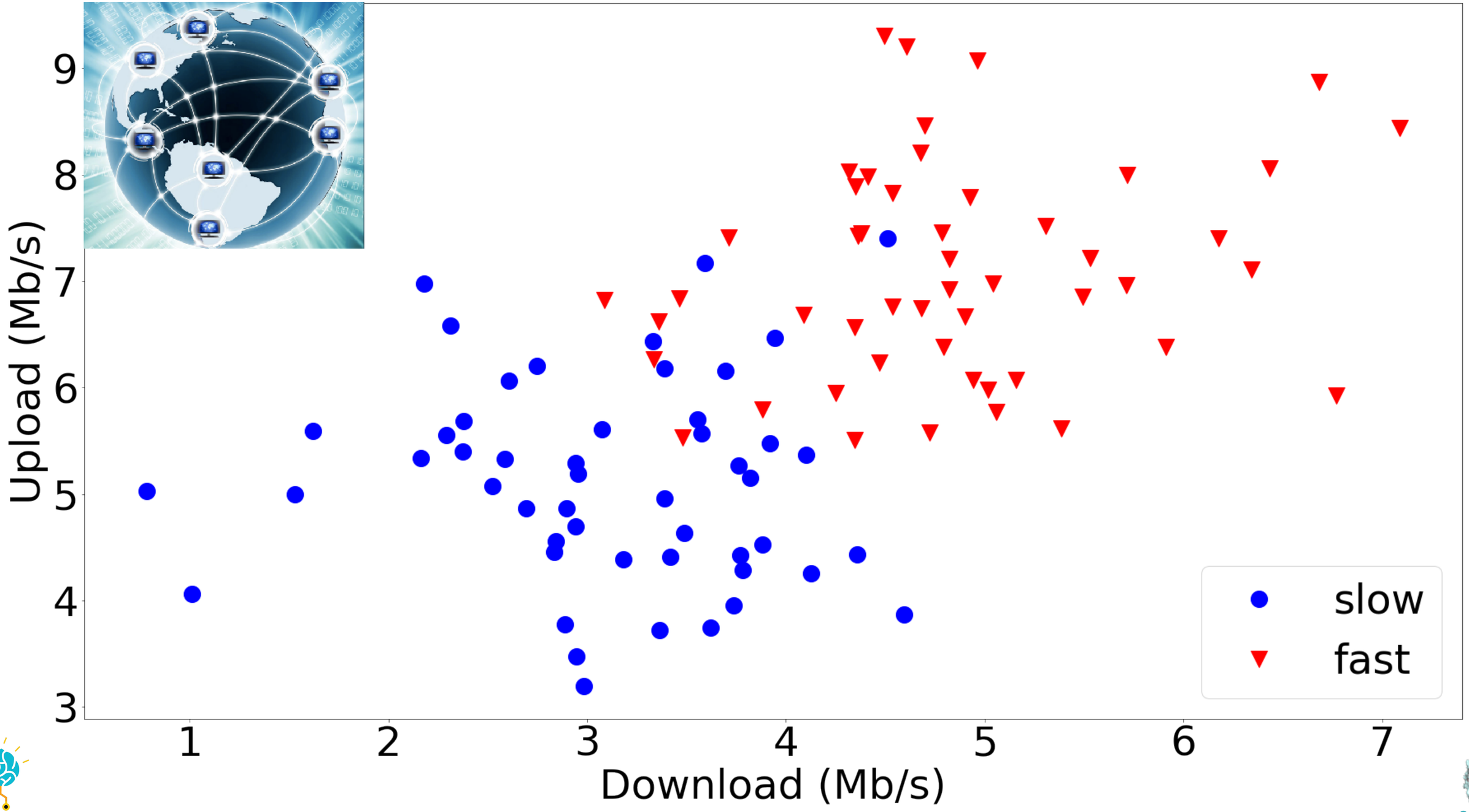
Catalit LLC

Humans



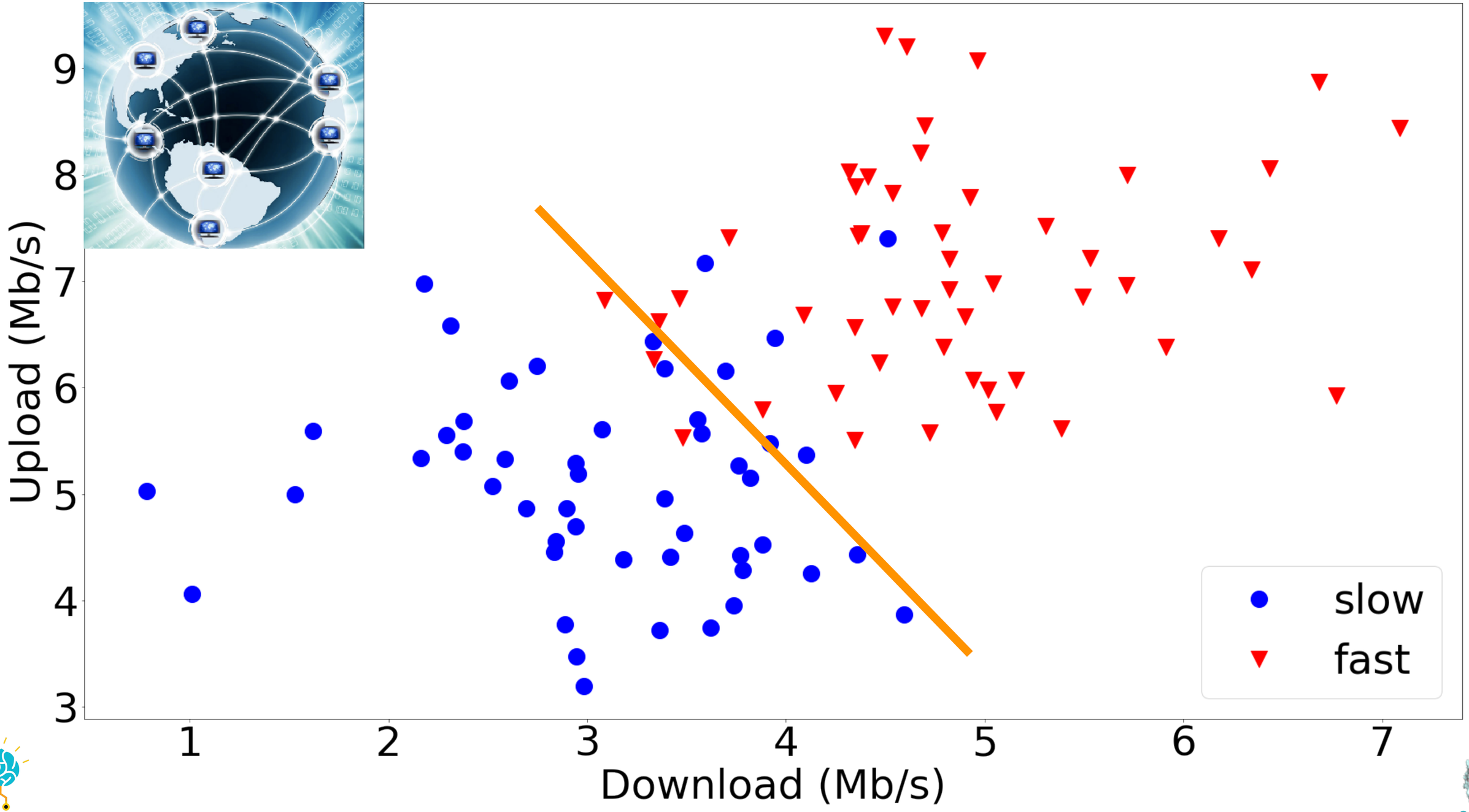
Catalit LLC

Internet Service Providers



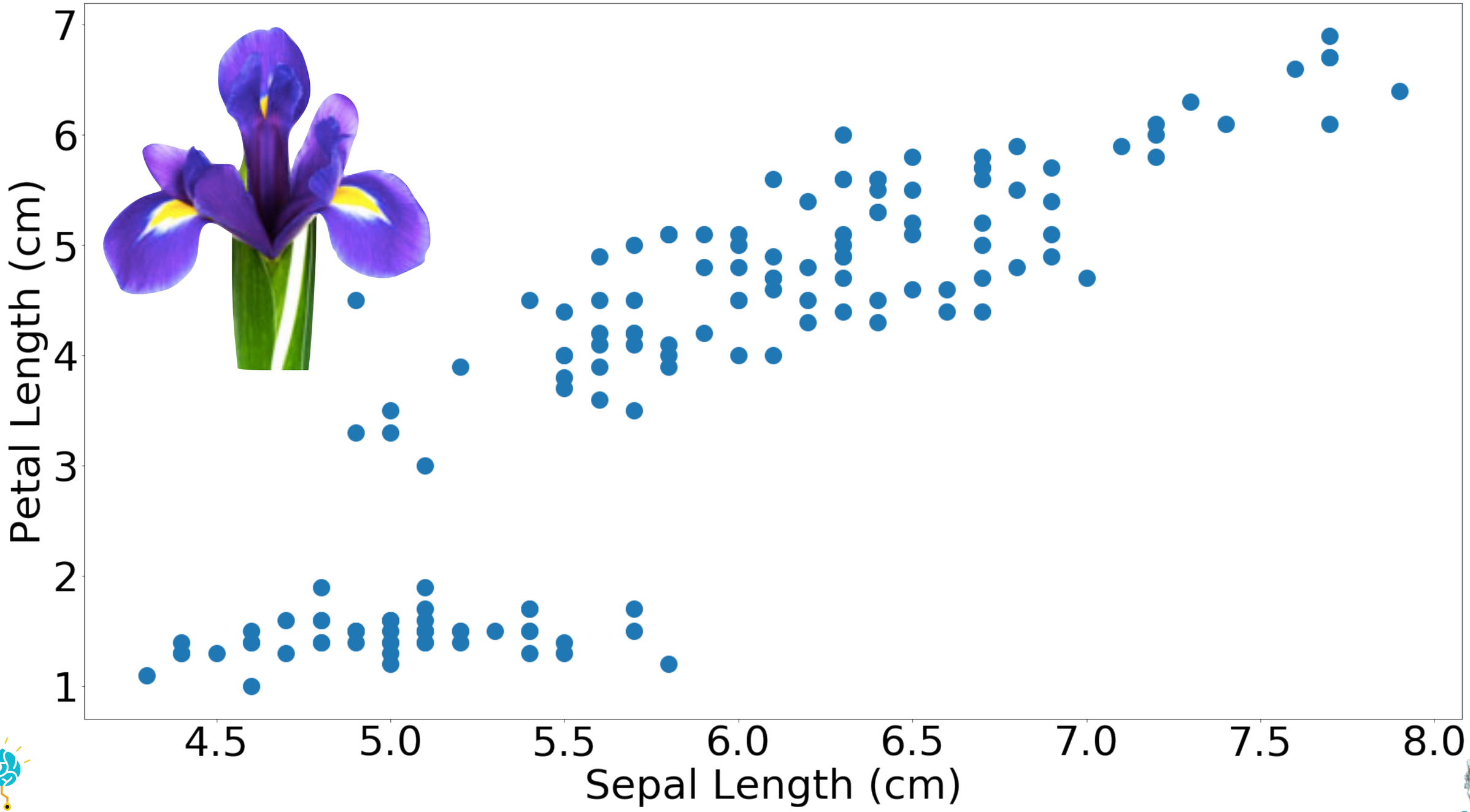
Catalit LLC

Internet Service Providers



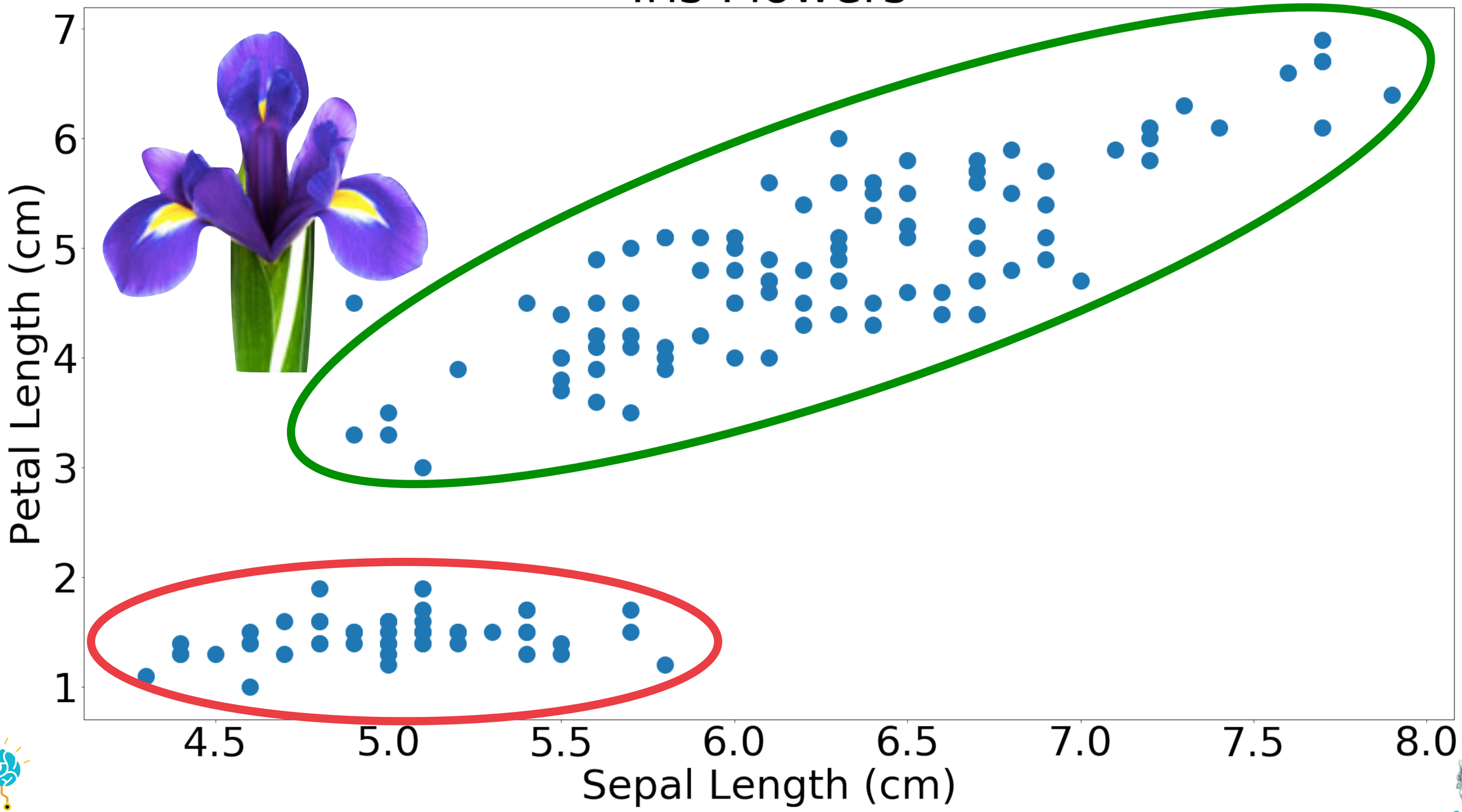
Catalit LLC

Iris Flowers



Catalit LLC

Iris Flowers



Catalit LLC

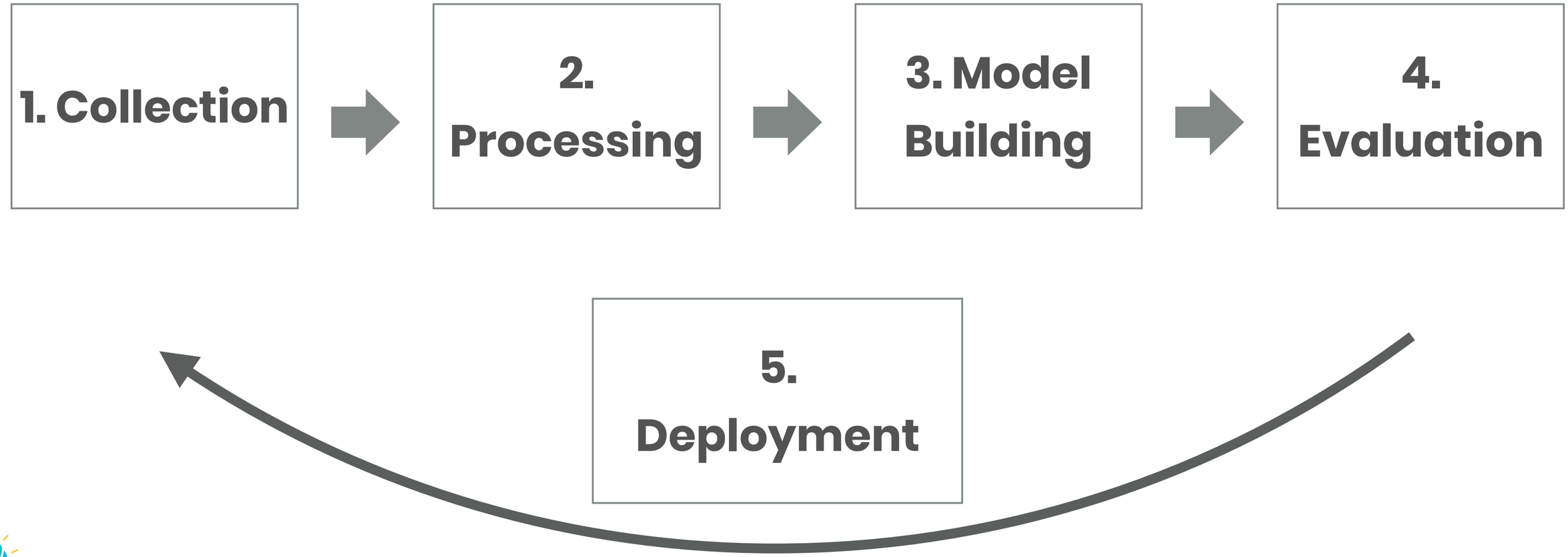
- Regression
- Classification
- Clustering

Machine Learning

- **"A field of study that gives computers the ability to learn without being explicitly programmed"** *(Arthur Samuel, 1959)*

- **"A new way of computer programming"** *(Google ML Dev, 2019)*

Machine Learning Process



Catalit LLC

Supervised VS Unsupervised

Supervised Learning

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

Catalit LLC

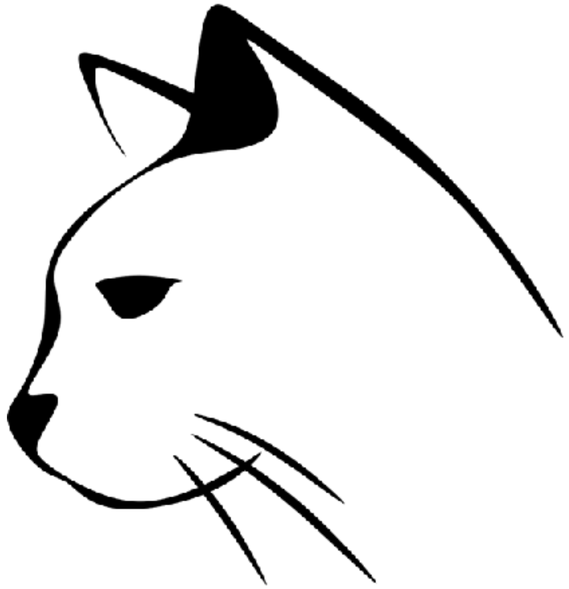
Supervised Learning

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

Catalit LLC

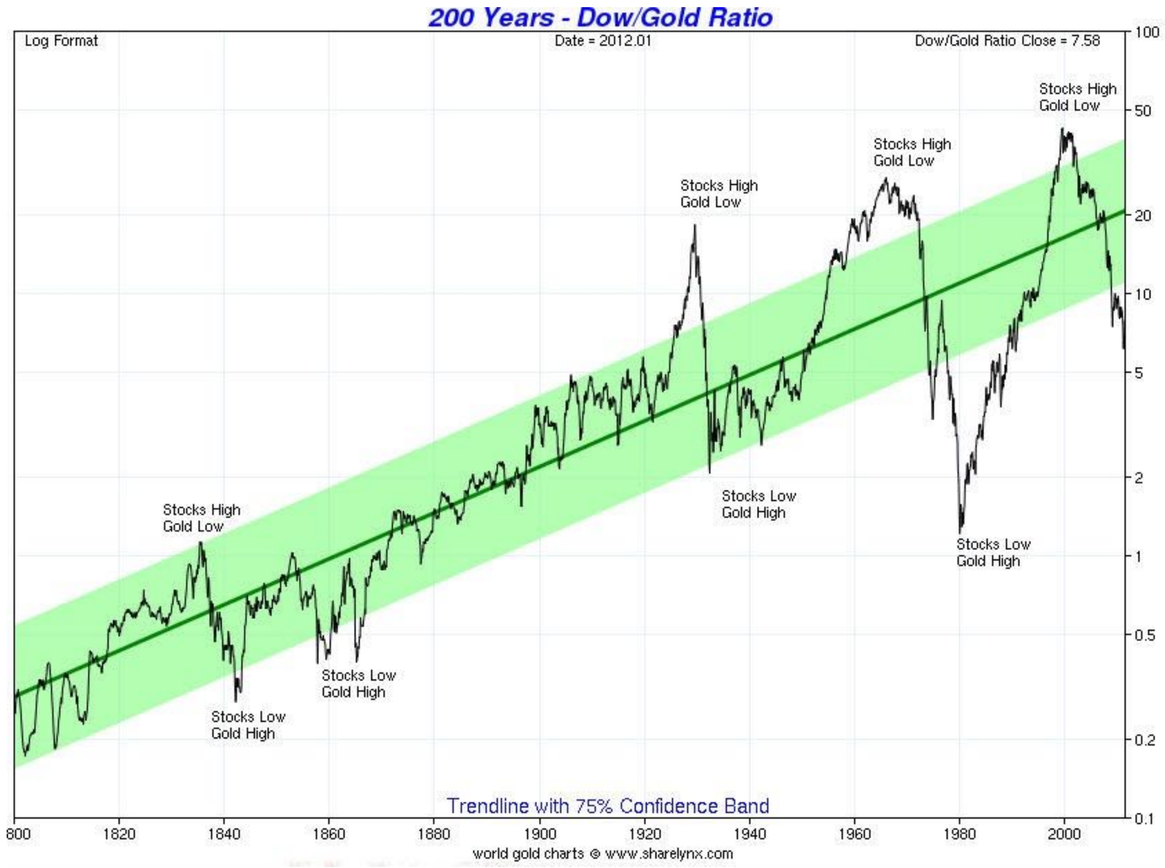
Supervised Learning

Generalizing the prediction to previously unseen examples



Catalit LLC

Supervised Learning



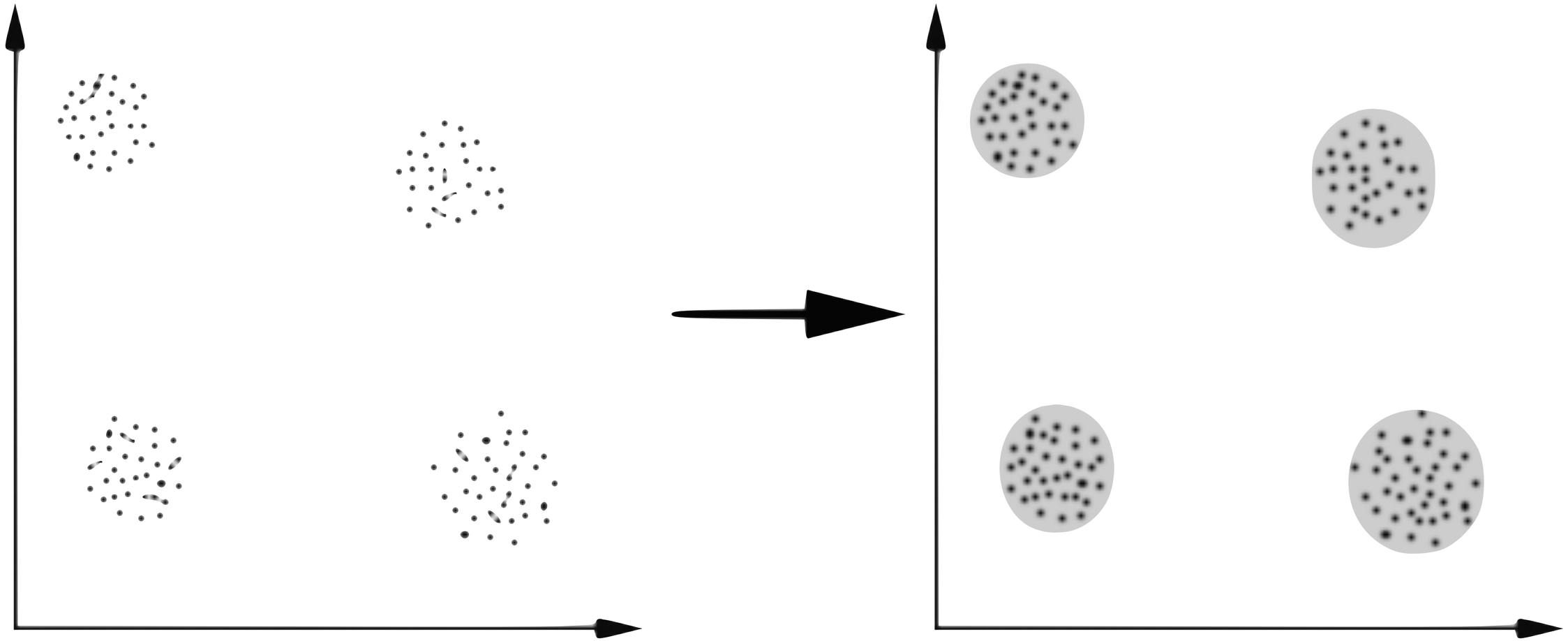
Catalit LLC

Unsupervised Learning

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

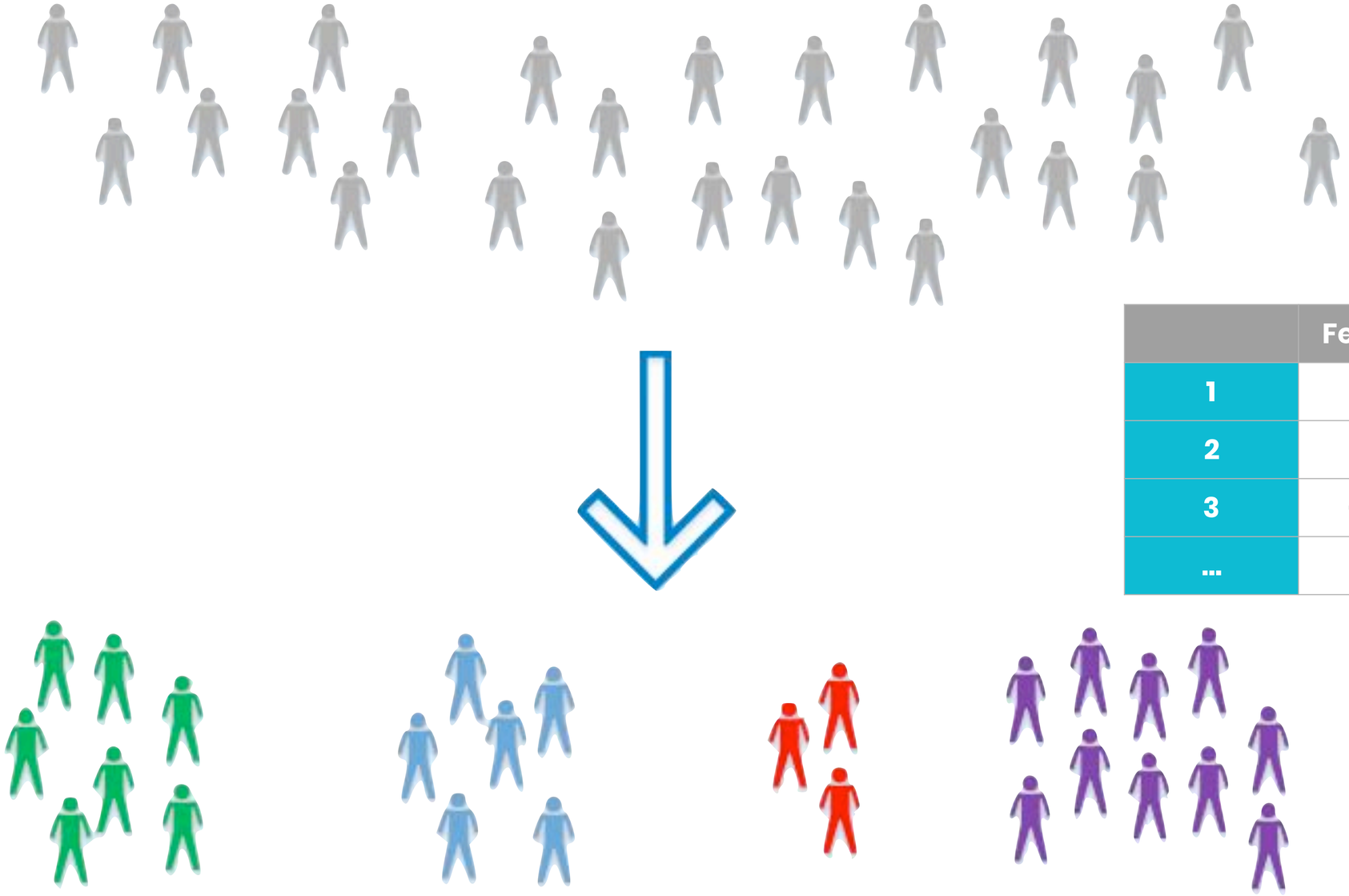
Catalit LLC

Clustering



Catalit LLC

Clustering



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

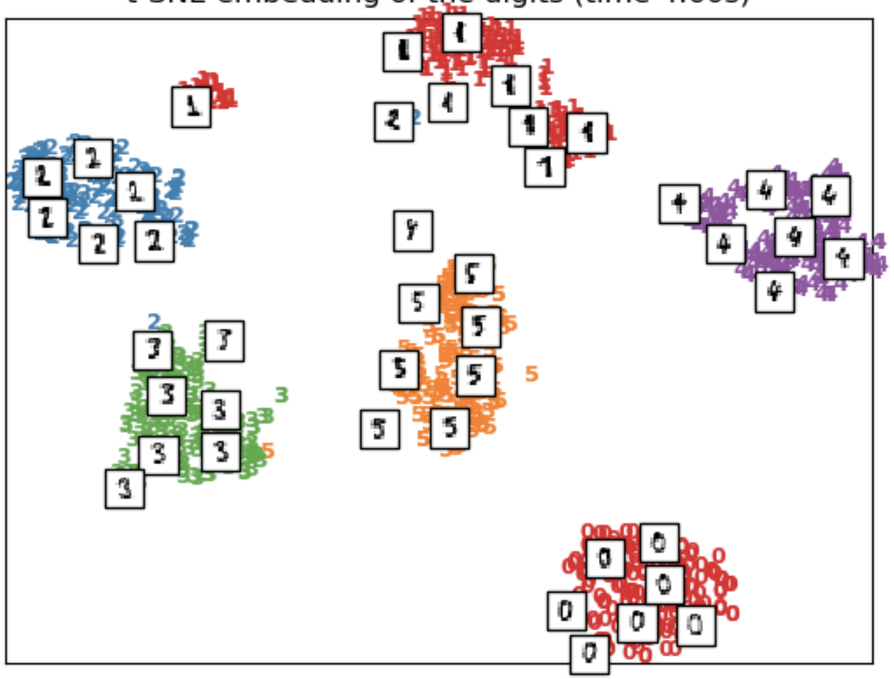
Catalit LLC

Dimensionality Reduction

- PCA
- LDA
- Isomap
- NMF
- t-SNE
- ....

A selection from the 64-dimensional digits dataset

t-SNE embedding of the digits (time 4.60s)

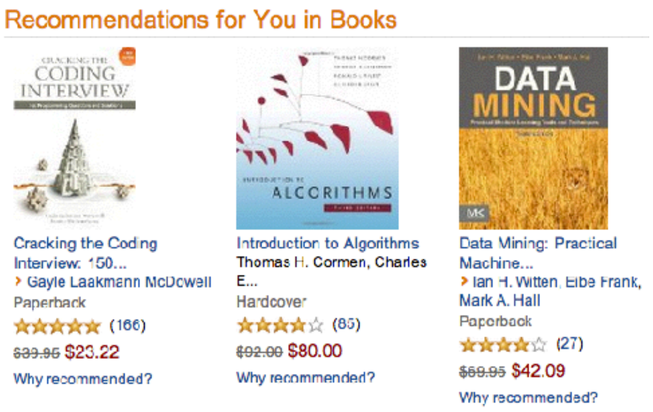


Catalit LLC

Applications

Applications

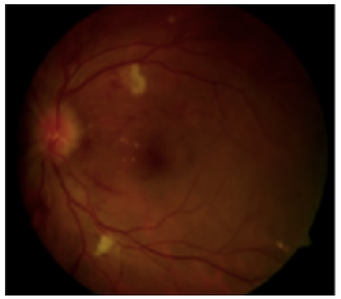
Book recommendation



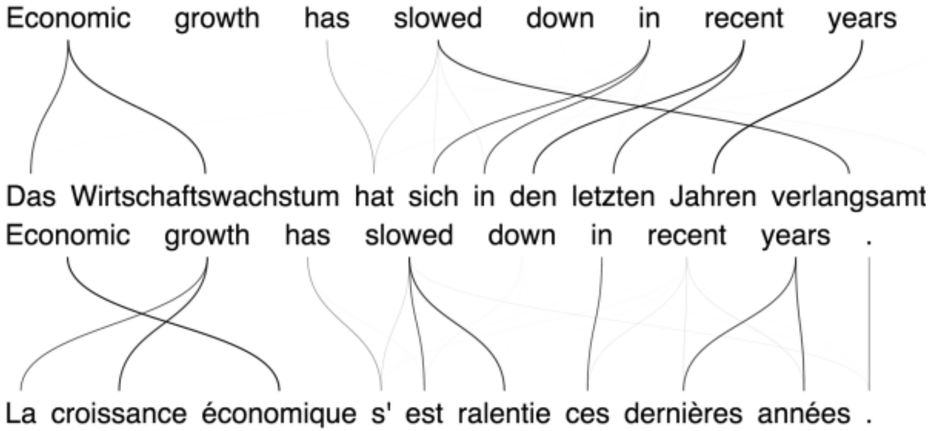
Sentiment Analysis

Image classification

Automated diagnosis



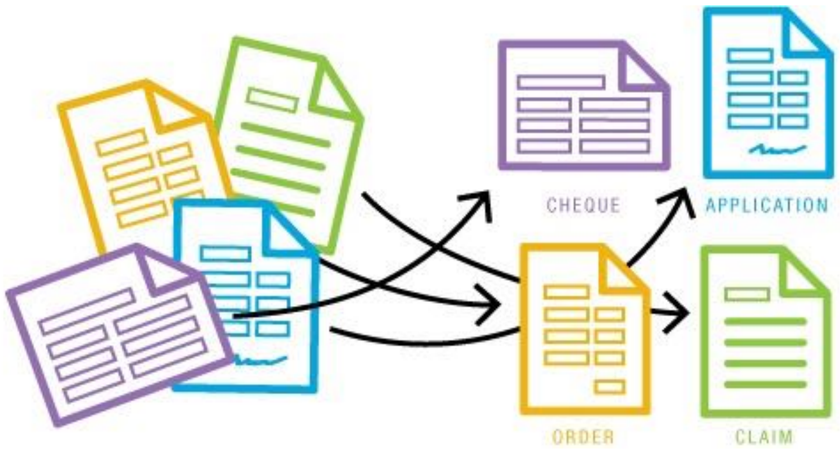
Machine translation



Catalit LLC

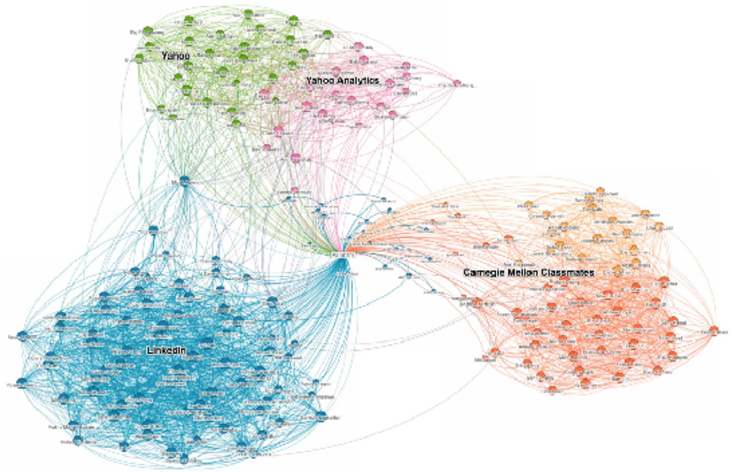
Applications

Document classification



House price prediction

Social Network Analysis



Malware / Fraud / Intrusion detection

Log Analysis

```
2018-02-22 00:14:39.906972 | Job console starting...
2018-02-22 00:15:07.534735 | PRE-RUN START: [trusted : git.openstack.org/openstack-infra/pro
e/pre.yaml@master]
2018-02-22 00:15:09.549712 |
2018-02-22 00:15:09.549988 | PLAY [localhost
2018-02-22 00:15:09.604921 |
2018-02-22 00:15:09.605190 | TASK [set-zuul-log-path-fact : Set log path for a change]
2018-02-22 00:15:09.683219 | localhost | ok
2018-02-22 00:15:09.717612 |
2018-02-22 00:15:09.717894 | TASK [set-zuul-log-path-fact : Set log path for a ref update]
2018-02-22 00:15:09.772548 | localhost | skipping: Conditional result was False
2018-02-22 00:15:09.796064 |
2018-02-22 00:15:09.796310 | TASK [set-zuul-log-path-fact : Set log path for a periodic job]
2018-02-22 00:15:09.851374 | localhost | skipping:
```

Catalit LLC

Tabular Data & Label Types

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	?

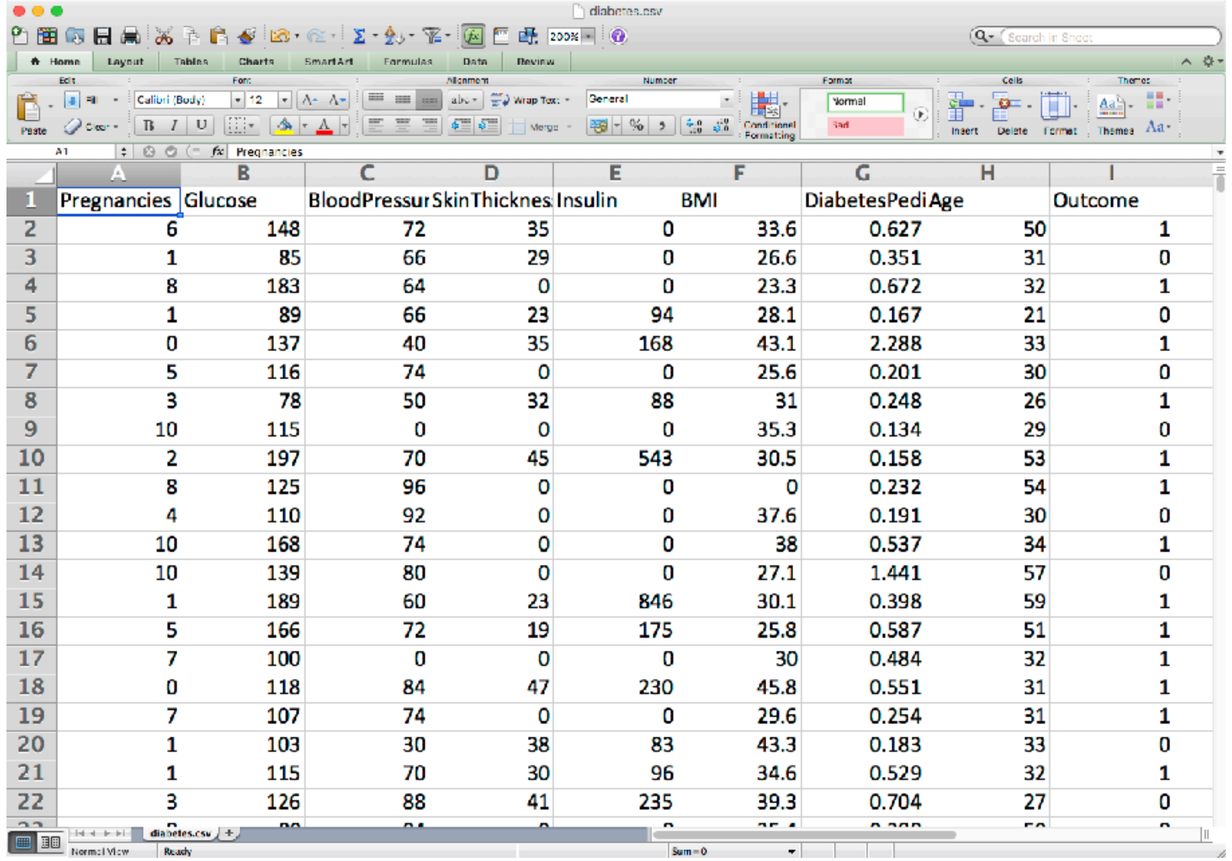
Catalit LLC

Tabular Data

- Excel
- CSV
- TSV
- Databases
- ....

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

Catalit LLC

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

Catalit LLC

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	$15,000	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

Catalit LLC

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	[illegible]	$1,000	Yes
Client 3	19	Male	$15,000	3	1	$10,000	No
Client 4	25	Male	$25,000	12	27	$15,000	?

Labels

Catalit LLC

Label types

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

Catalit LLC

Label types

CATEGORICAL	CONTINUOUS
Colors	Height
Job Titles	# downloads
Browser type	Debt
Fraud or not	Speed
Spam or not	Price

Catalit LLC

Summary of ML

	CATEGORICAL	CONTINUOUS
SUPERVISED	CLASSIFICATION	REGRESSION
UNSUPERVISED	CLUSTERING	DIM REDUCTION

Catalit LLC

Why Python

Why Python

Catalit LLC

Why Python

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

Why Python

python™

NumPy

SciPy

IP[y]: IPython
Interactive Computing

jupyter

Anaconda

pandas

matplotlib

scikits learn
machine learning in Python

K + TensorFlow

Catalit LLC

Lab 01 & 02

- **Jupyter**
  - interactive Python web application => our dev env
- **Pandas**
  - like Excel for python => data manipulation
- **Matplotlib & Seaborn**
  - Plots & charts => data visualization

Regression

Regression

- Intuition and definition
- Linear Regression with Gradient Descent
- Multivariate & Nonlinear
- Model Evaluation
- Overfitting & Regularization

Summary of ML

	CATEGORICAL	CONTINUOUS
SUPERVISED	?	?
UNSUPERVISED	?	?

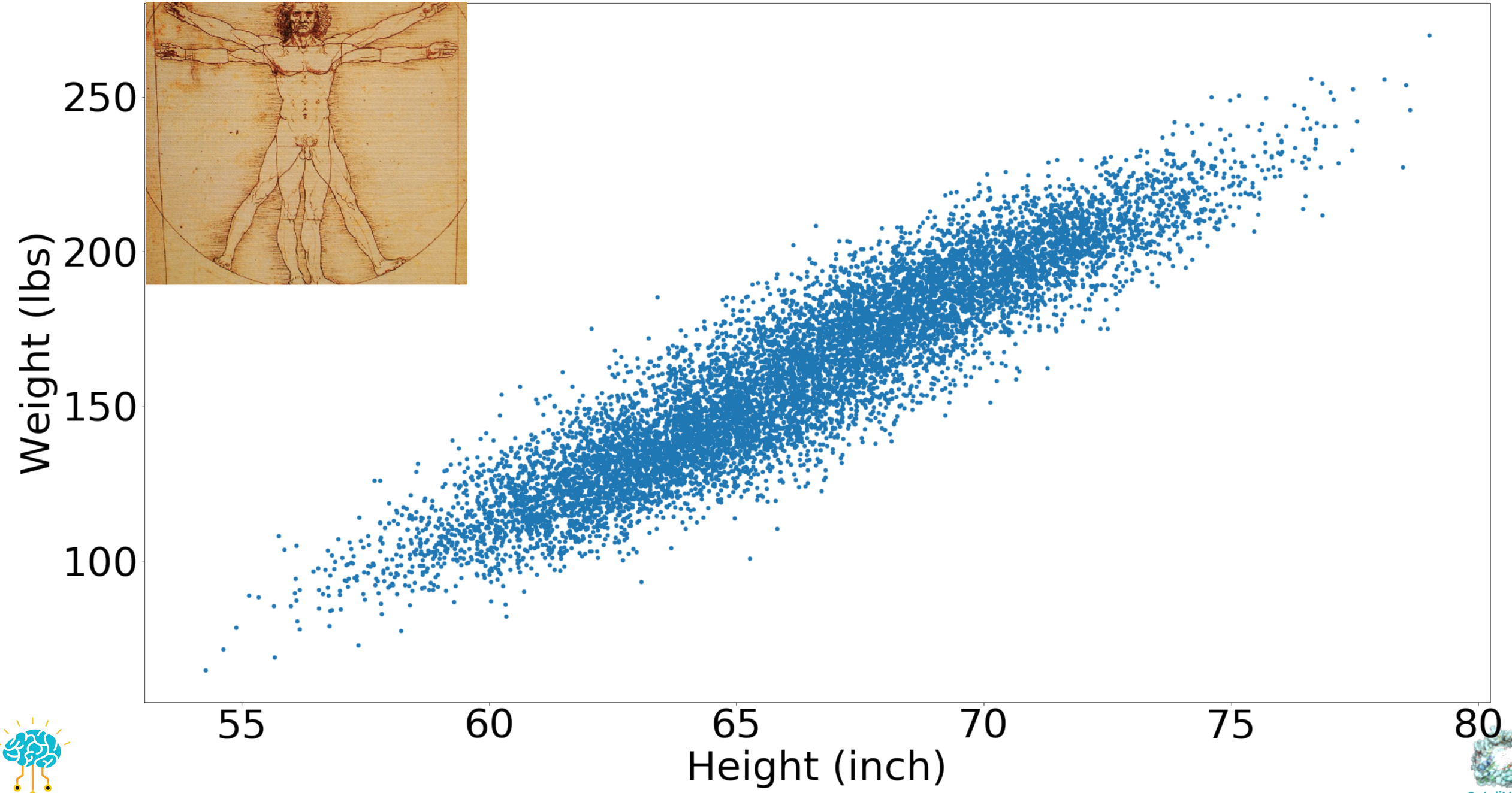
Catalit LLC

Summary of ML

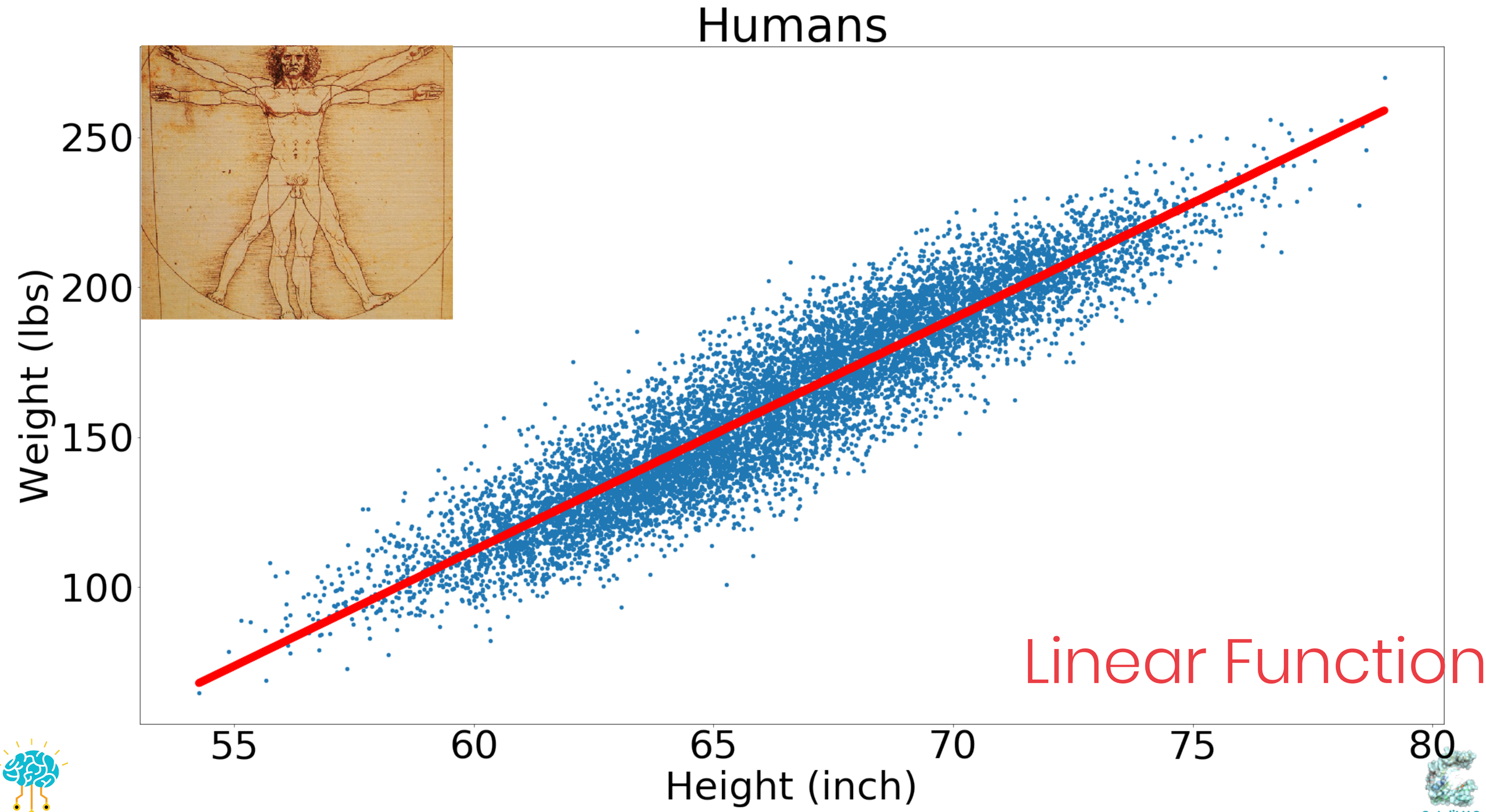
	CATEGORICAL	CONTINUOUS
SUPERVISED	CLASSIFICATION	REGRESSION
UNSUPERVISED	CLUSTERING	DIM REDUCTION

Catalit LLC

Humans



Catalit LLC

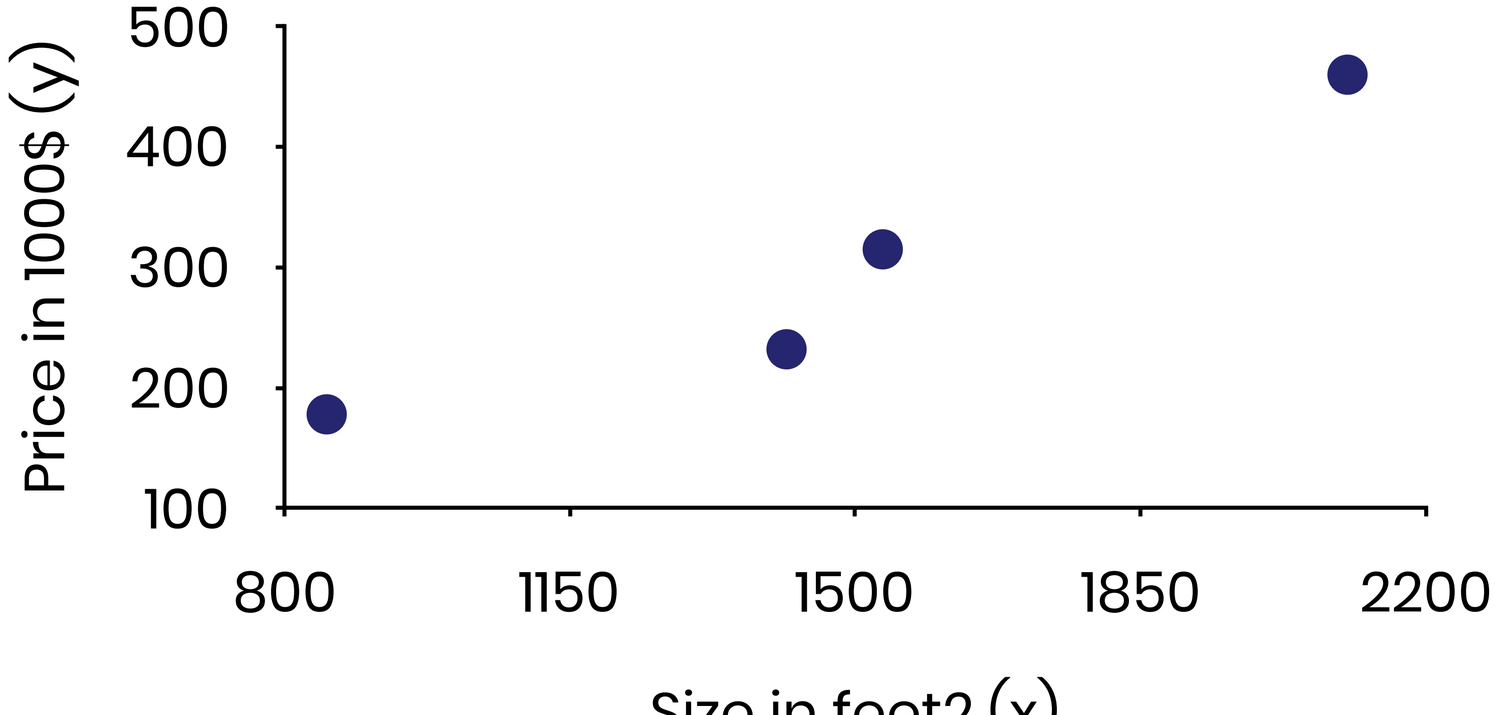

Humans
Weight (lbs)
250
200
150
100
55
60
65
70
75
80
Height (inch)
Linear Function

Catalit LLC

Linear Regression with Gradient Descent

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

y = func(X)

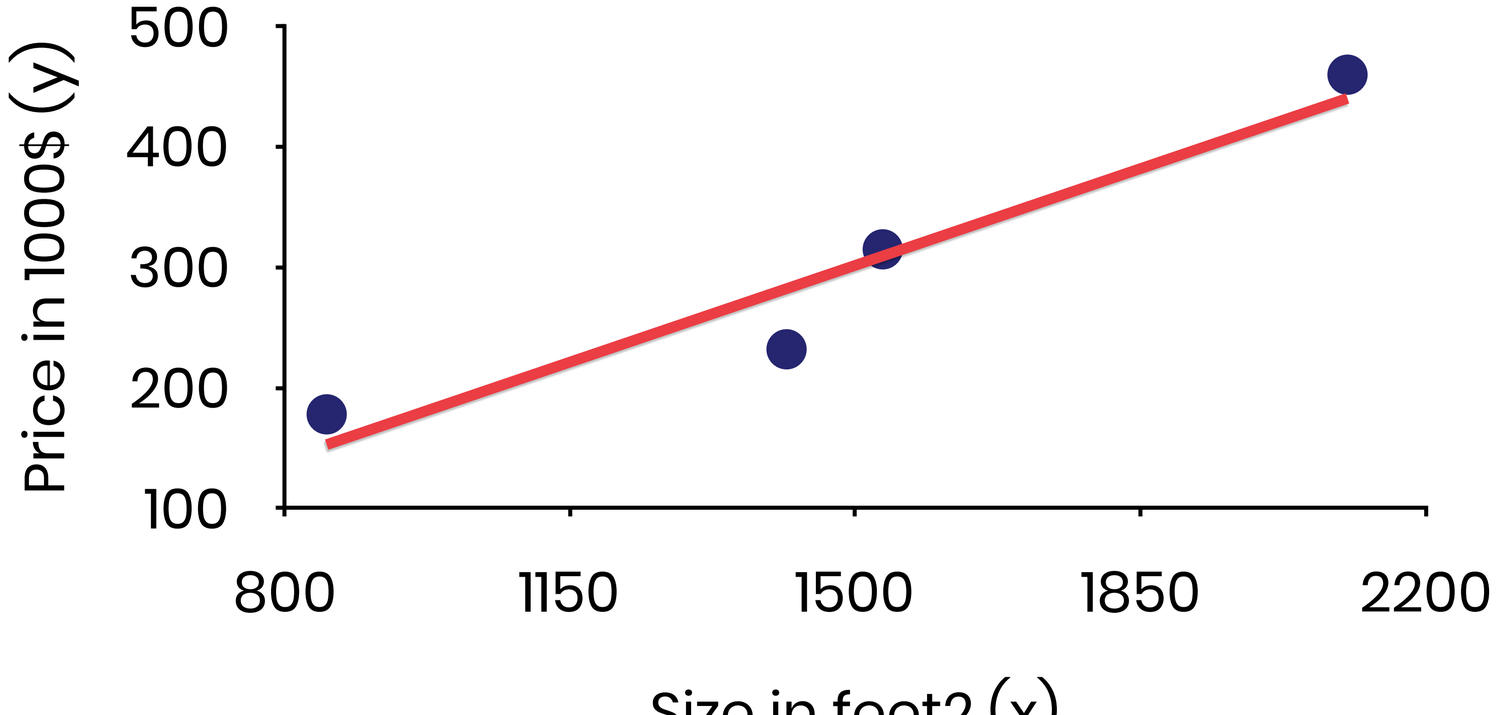


Catalit LLC

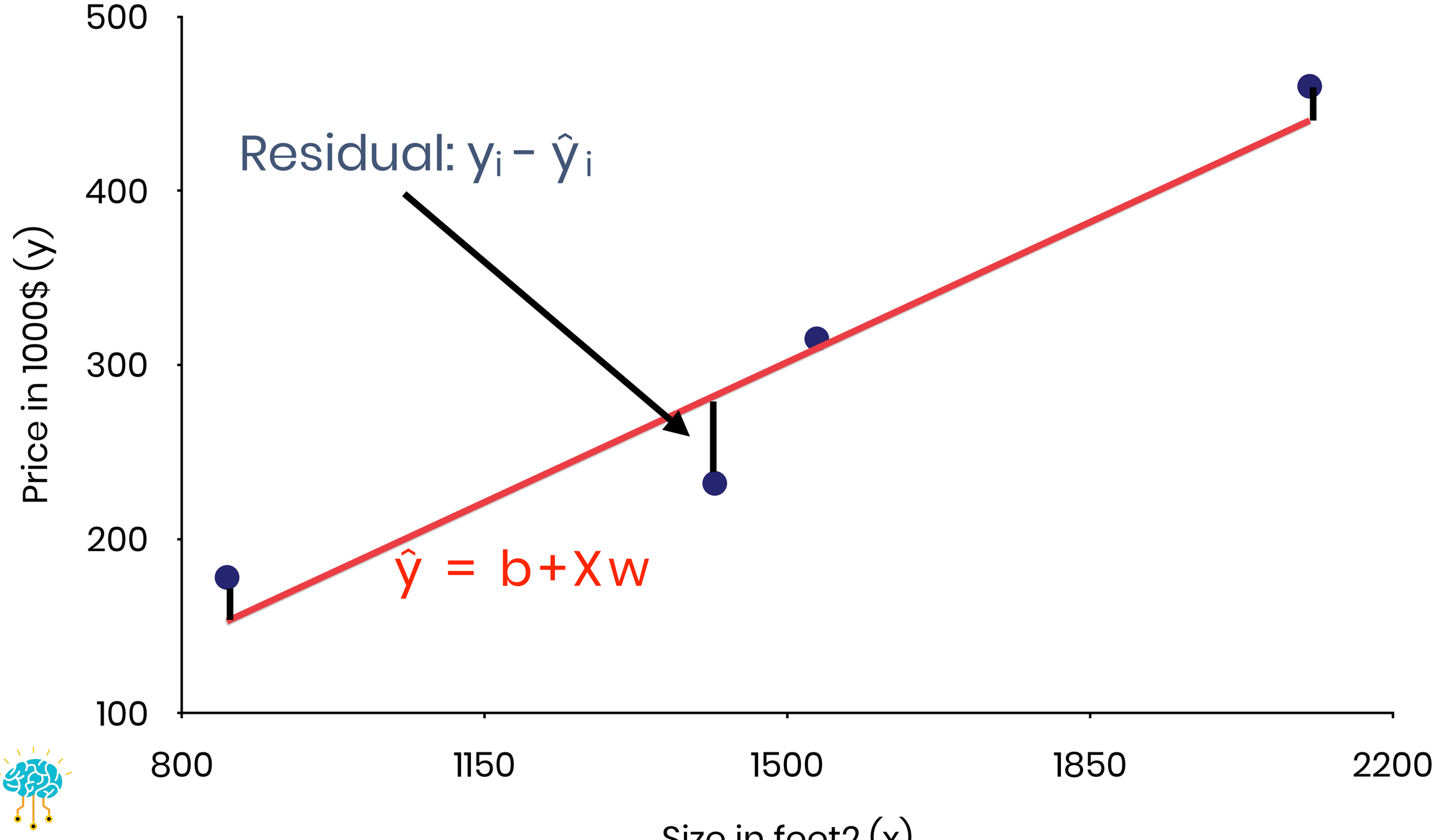
Size in feet² (x)	Price in 1000$ (y)
2104	460
1416	232
1534	315
852	178
...	...

Hypothesis

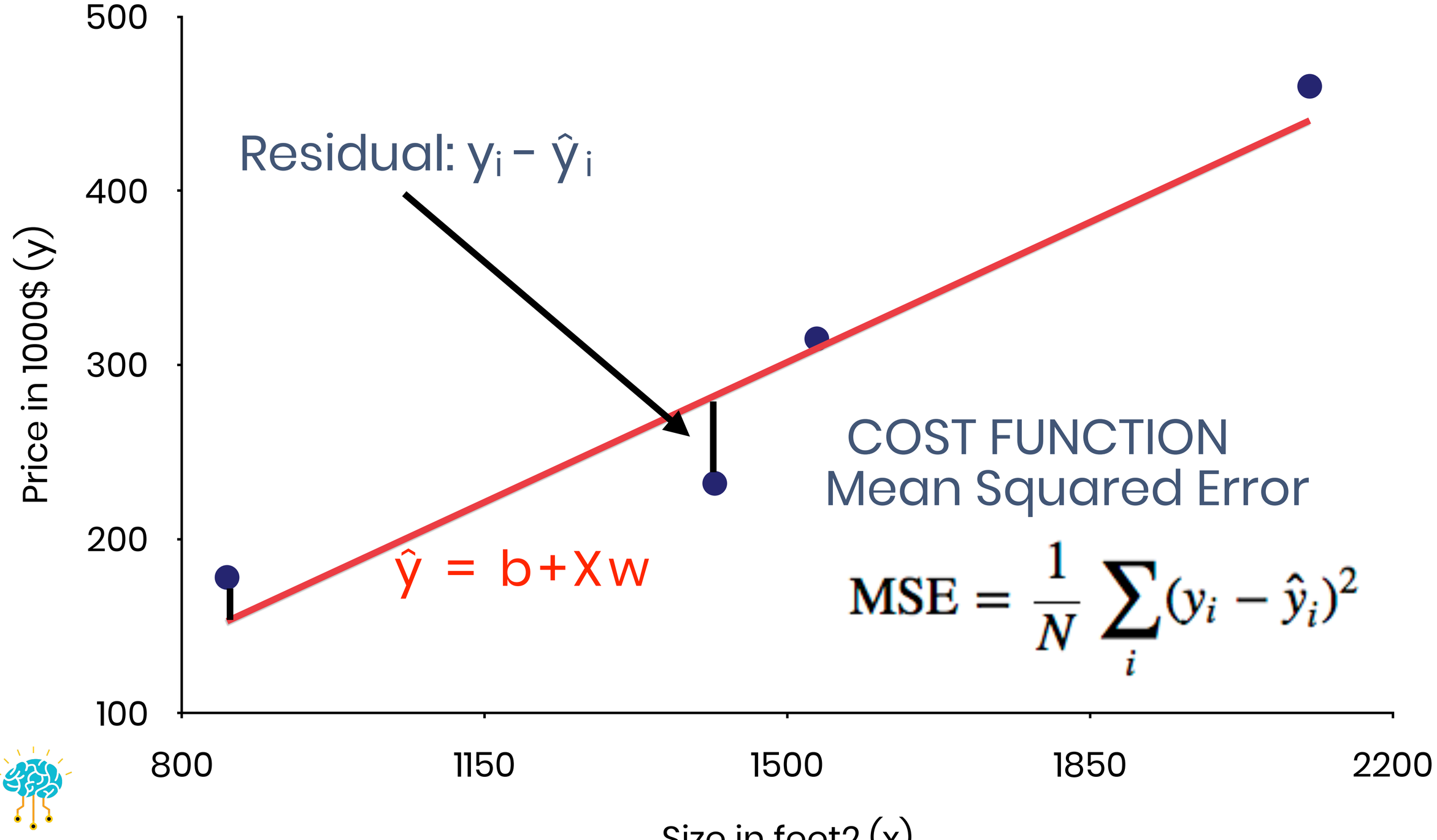
$$\hat{y} = b + Xw$$



Catalit LLC


Residual: $y_i - \hat{y}_i$
$\hat{y} = b + Xw$
500
400
300
200
100
Price in 1000$ (y)
800
1150
1500
1850
2200
Size in feet2 (x)

Catalit LLC

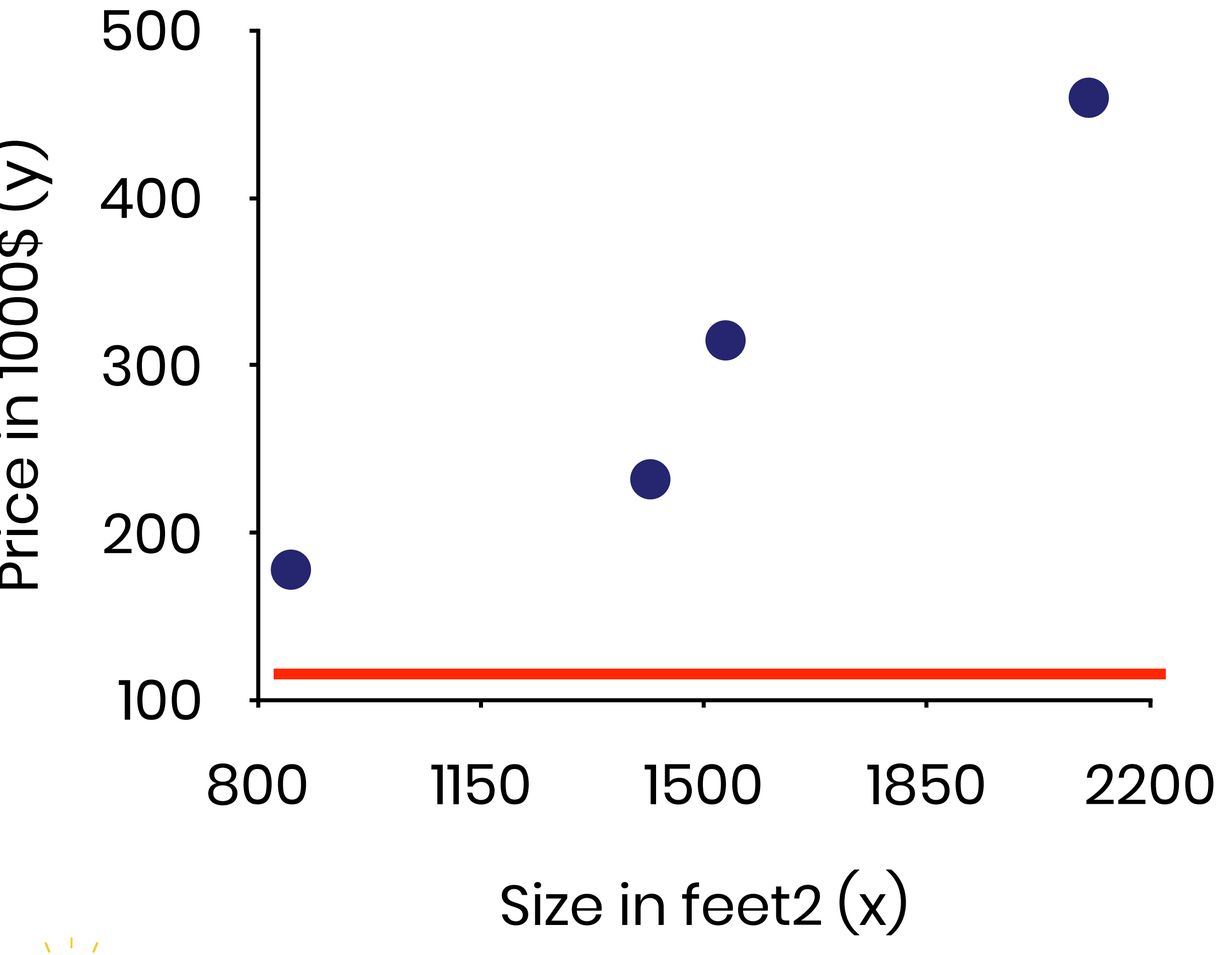


COST FUNCTION
Mean Squared Error

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

Catalit LLC

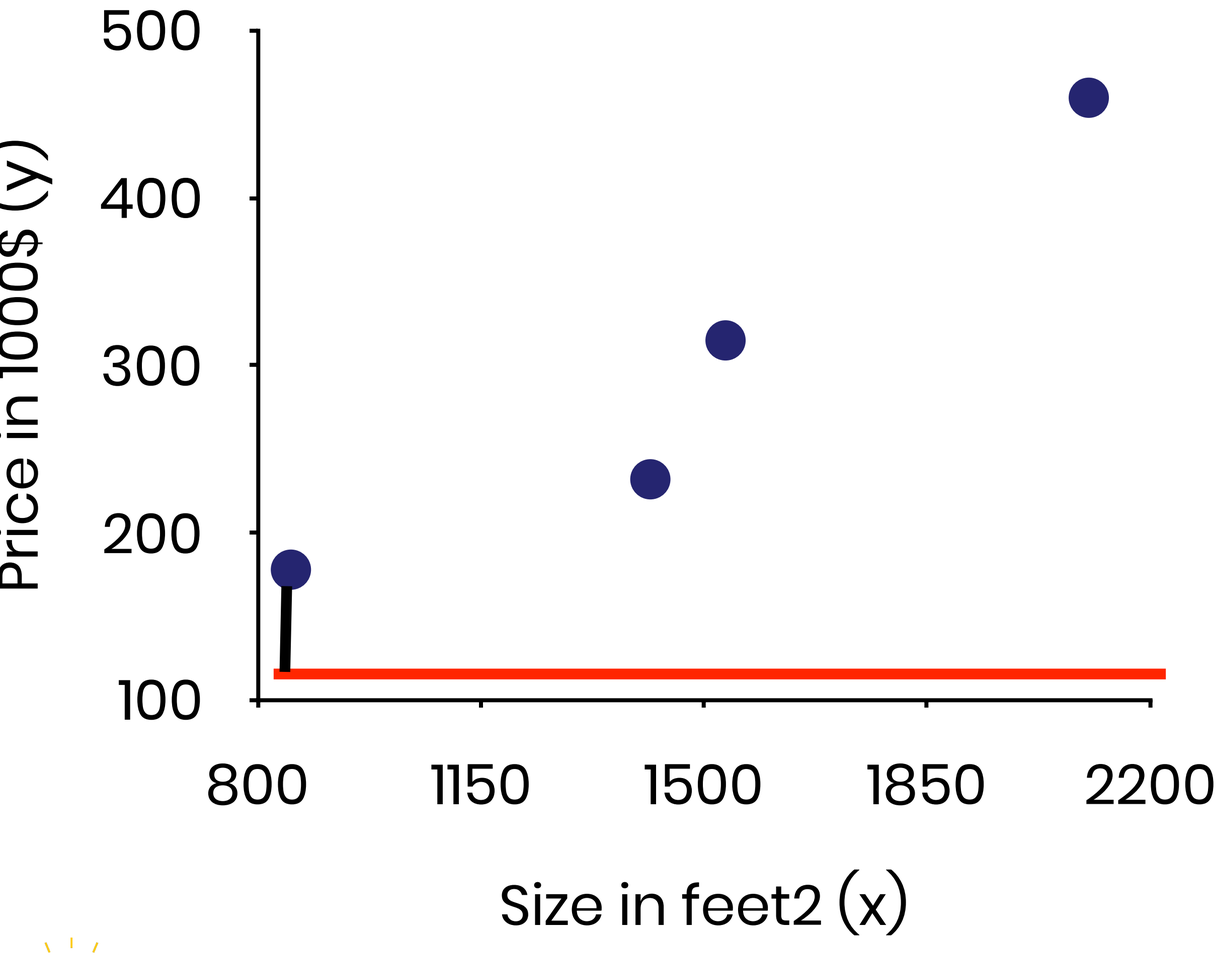
Cost function



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

Catalit LLC

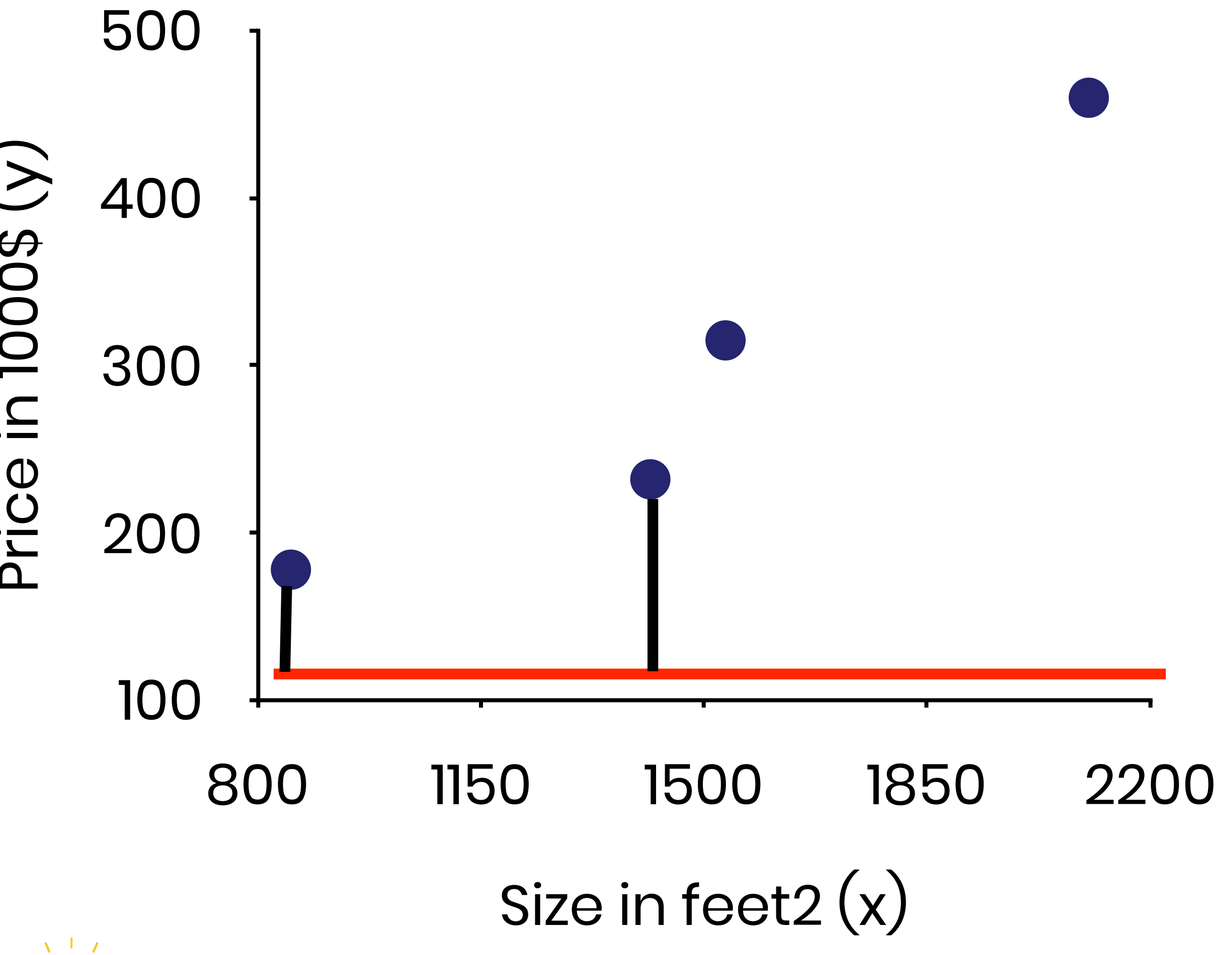
Cost function



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

Catalit LLC

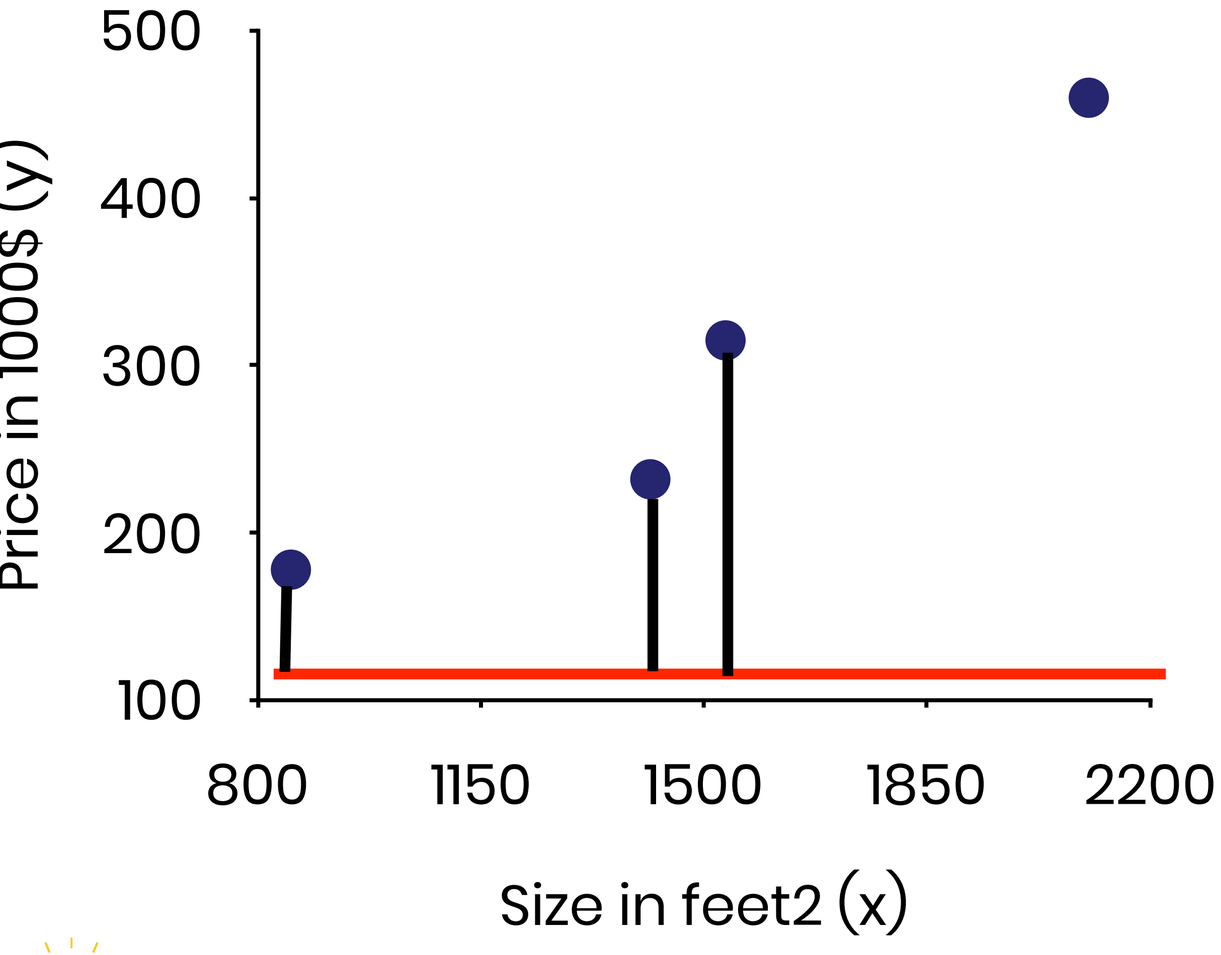
Cost function



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

Catalit LLC

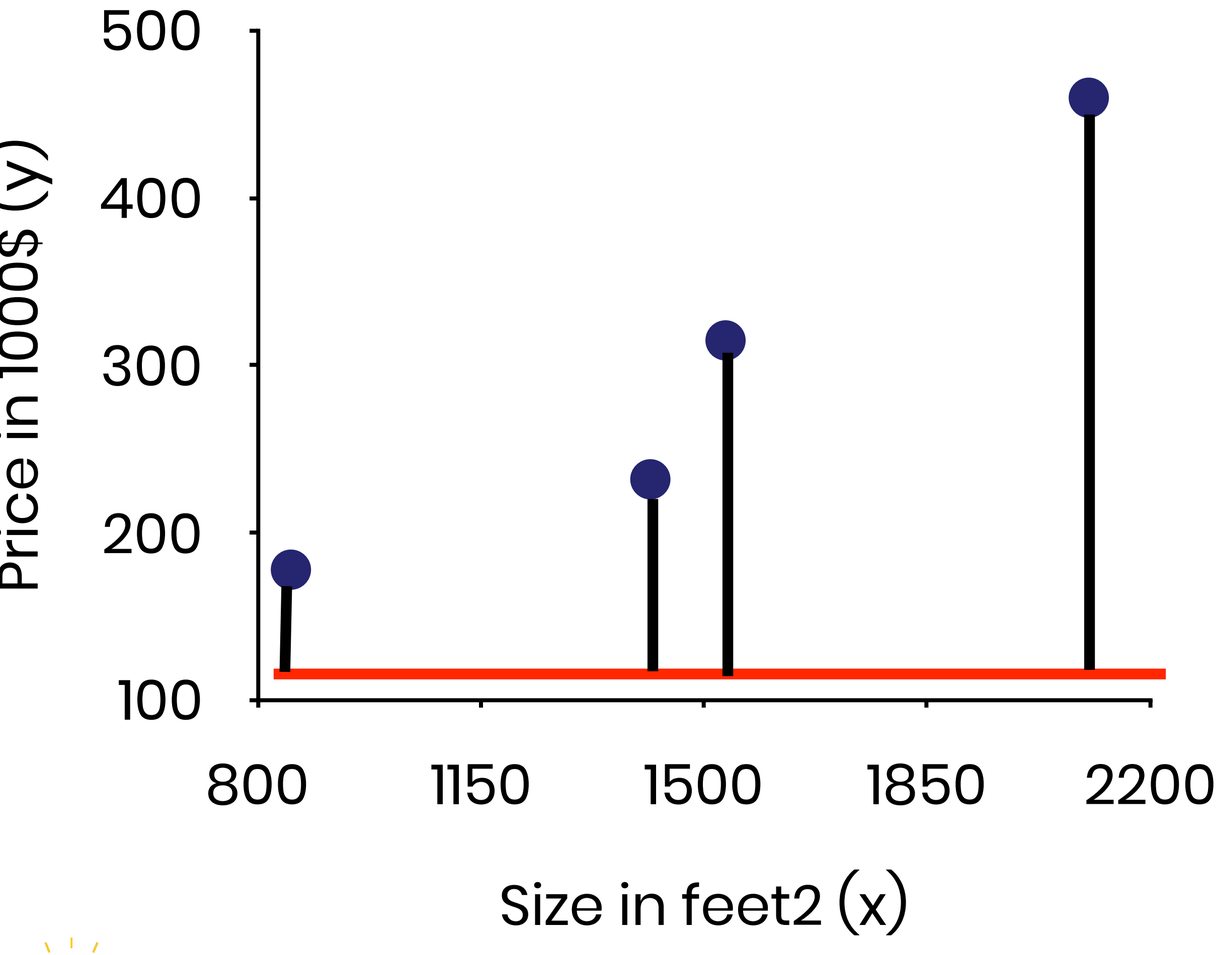
Cost function



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

Catalit LLC

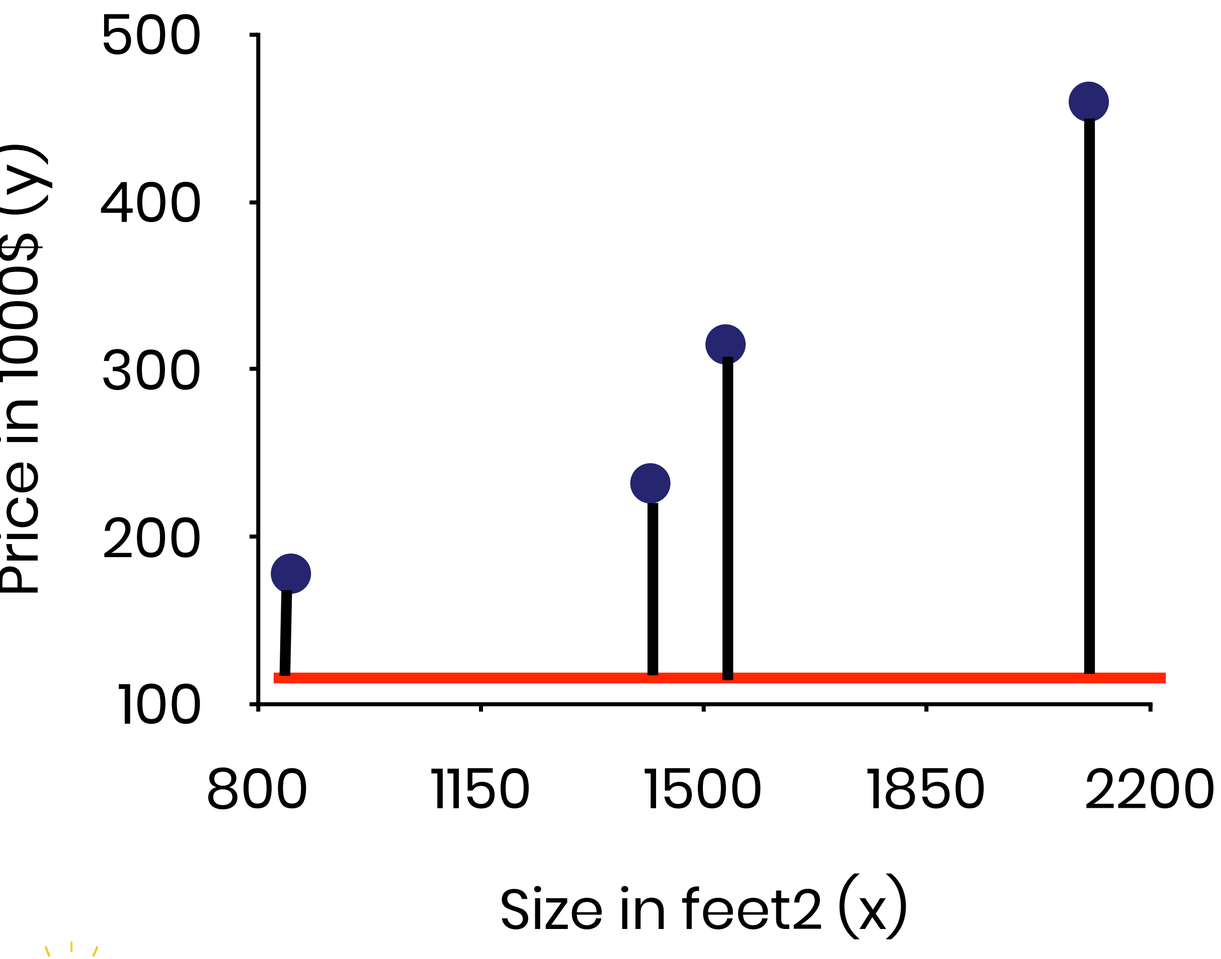
Cost function

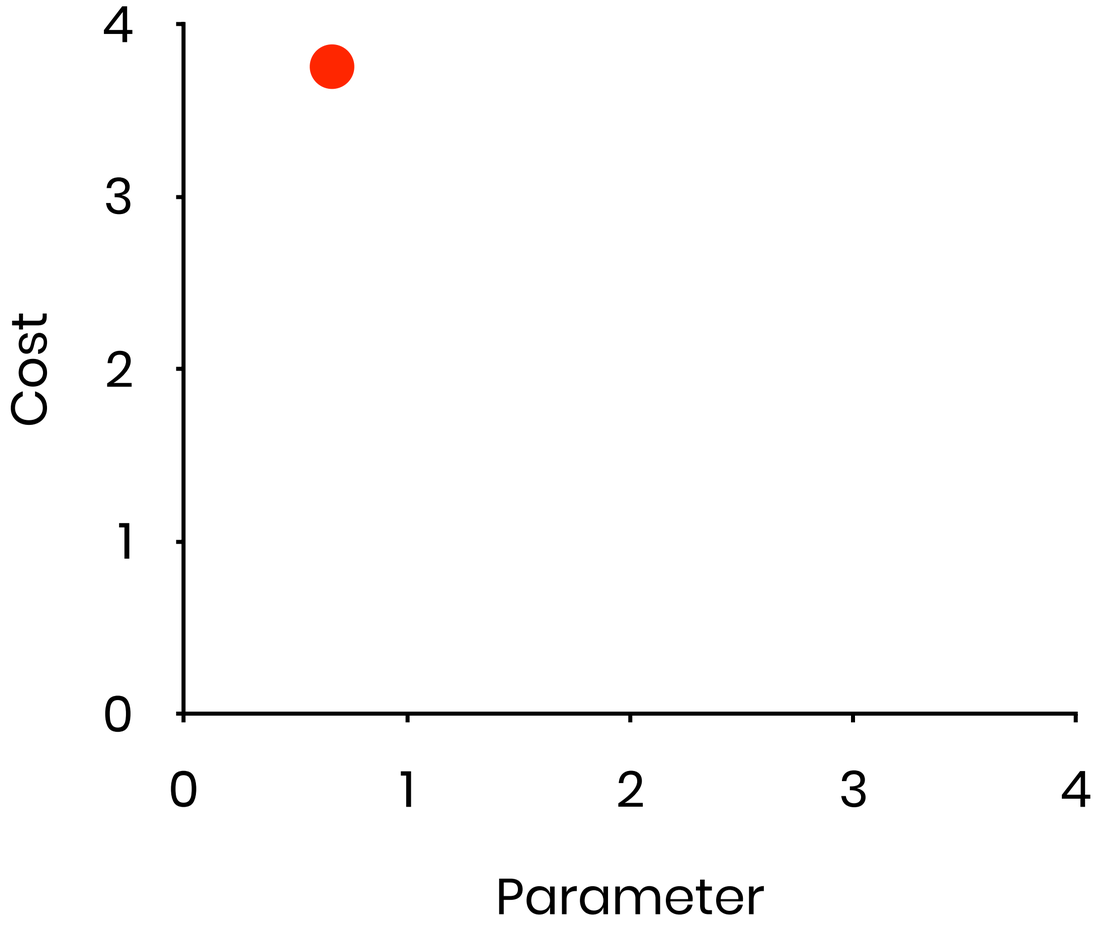


- Fixed data
- Fixed hypothesis
- Cost depends on values of parameters

Catalit LLC

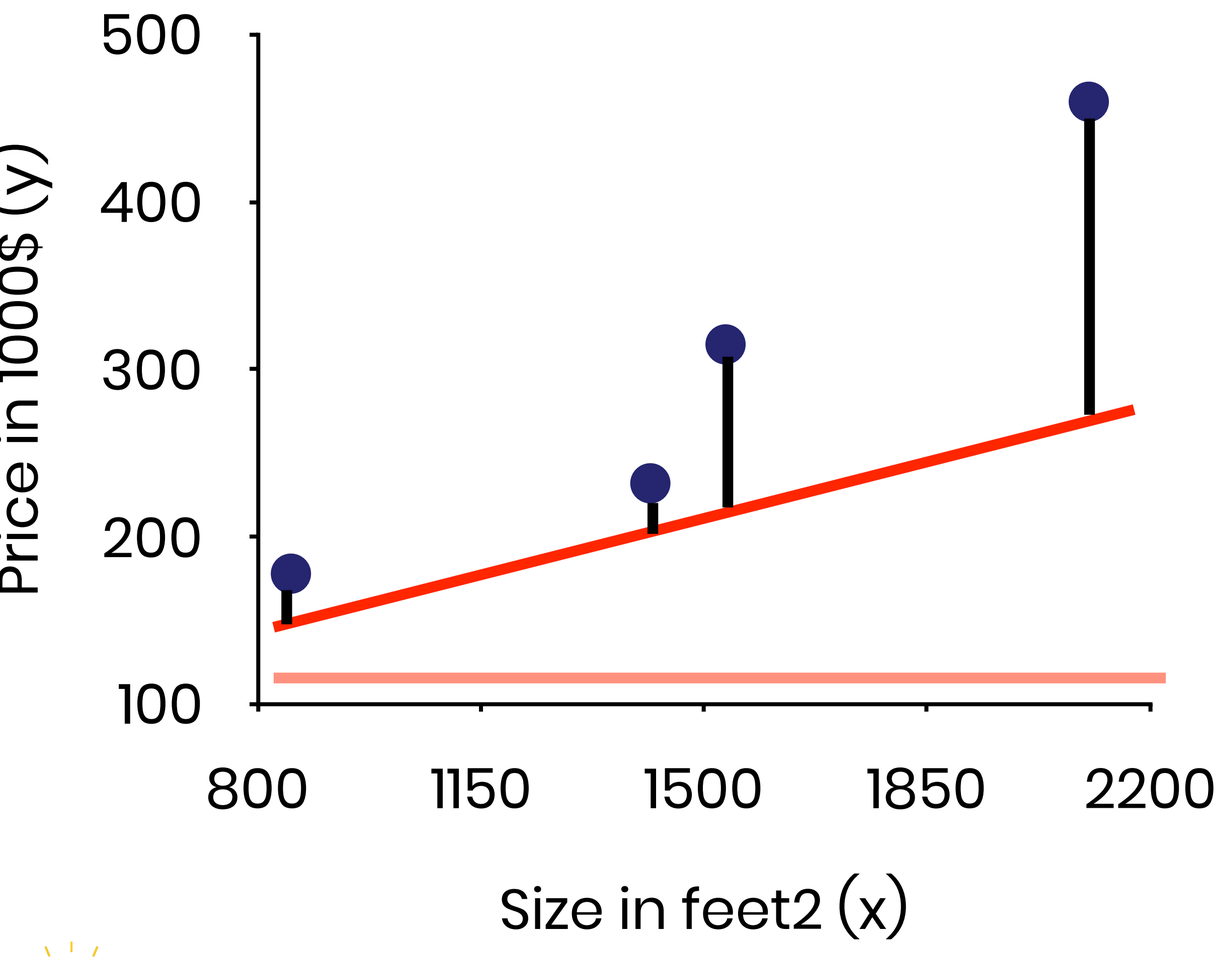
Cost function

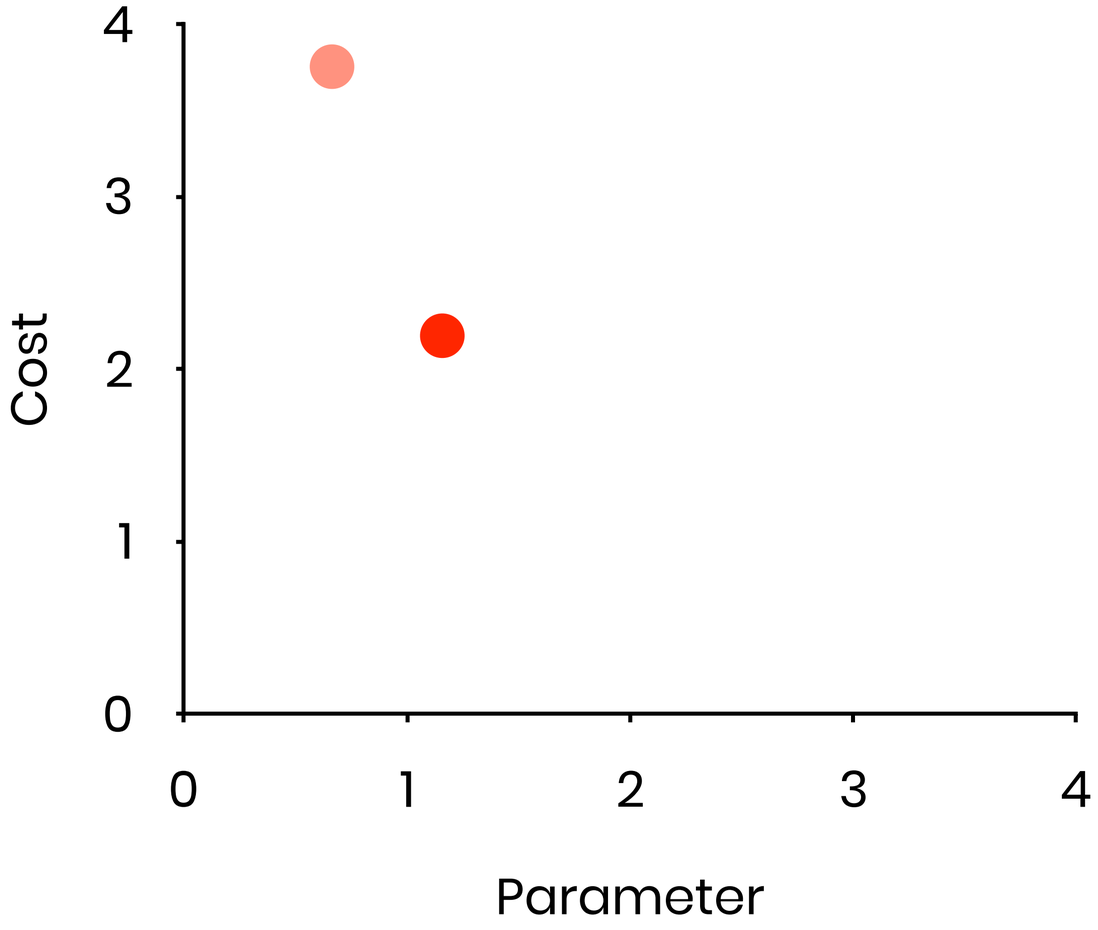




Catalit LLC

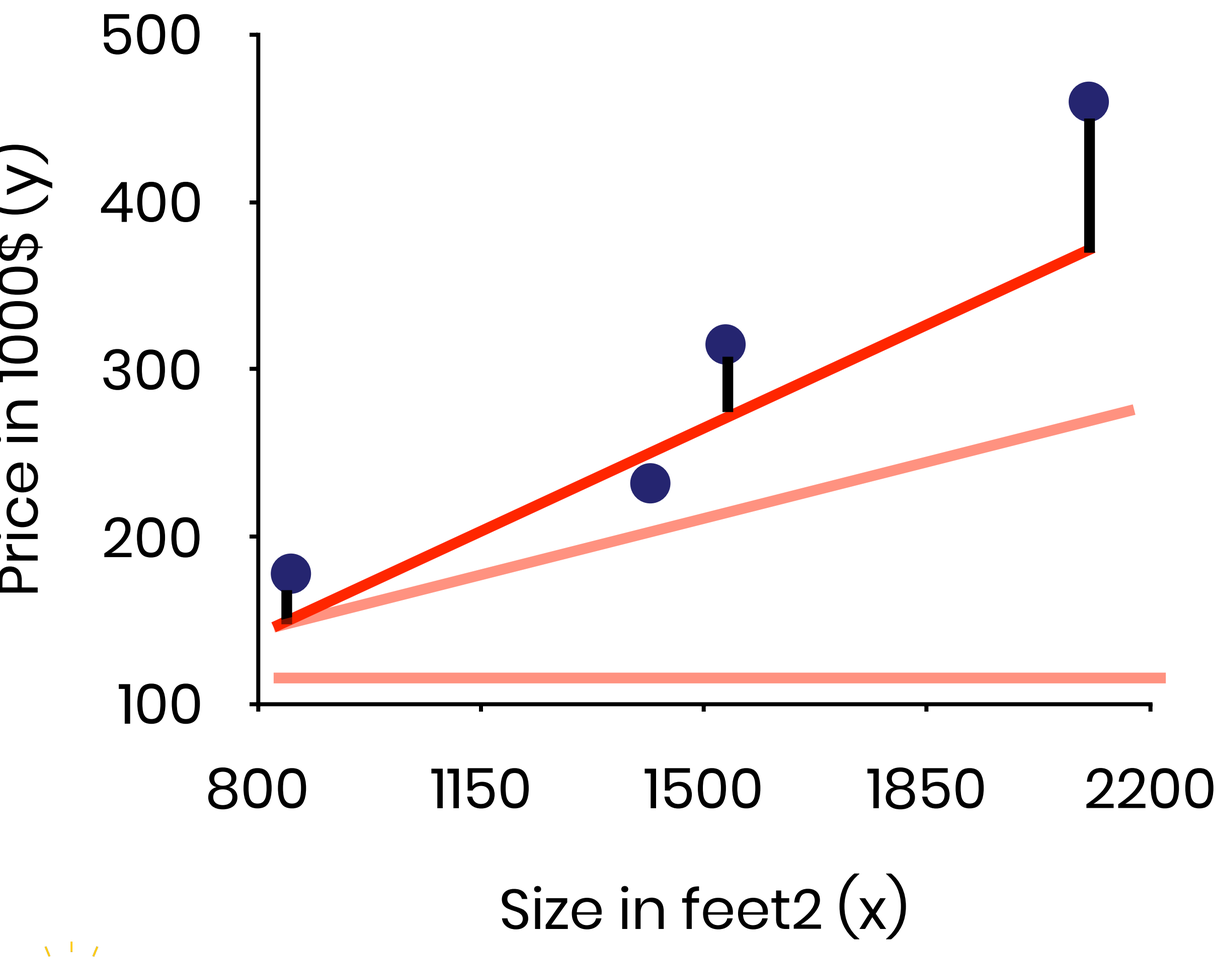
Optimization

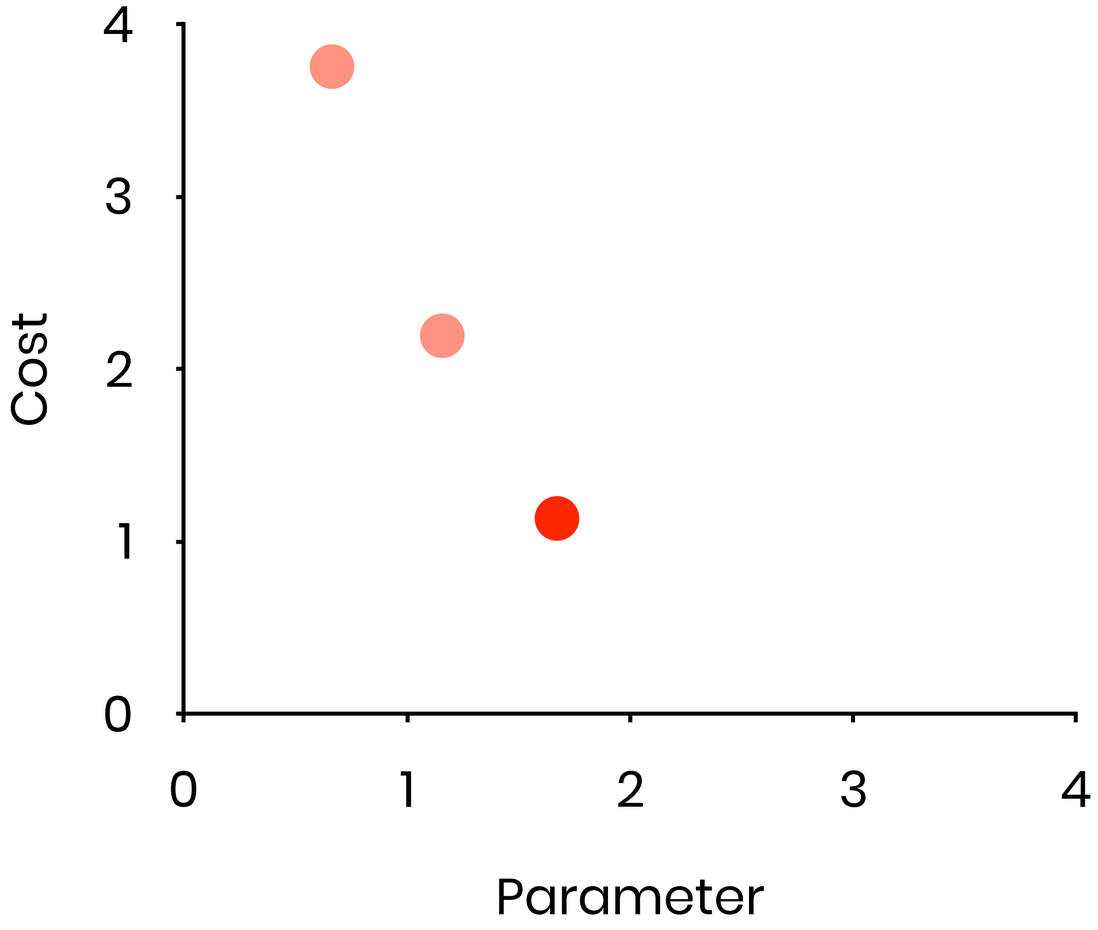




Catalit LLC

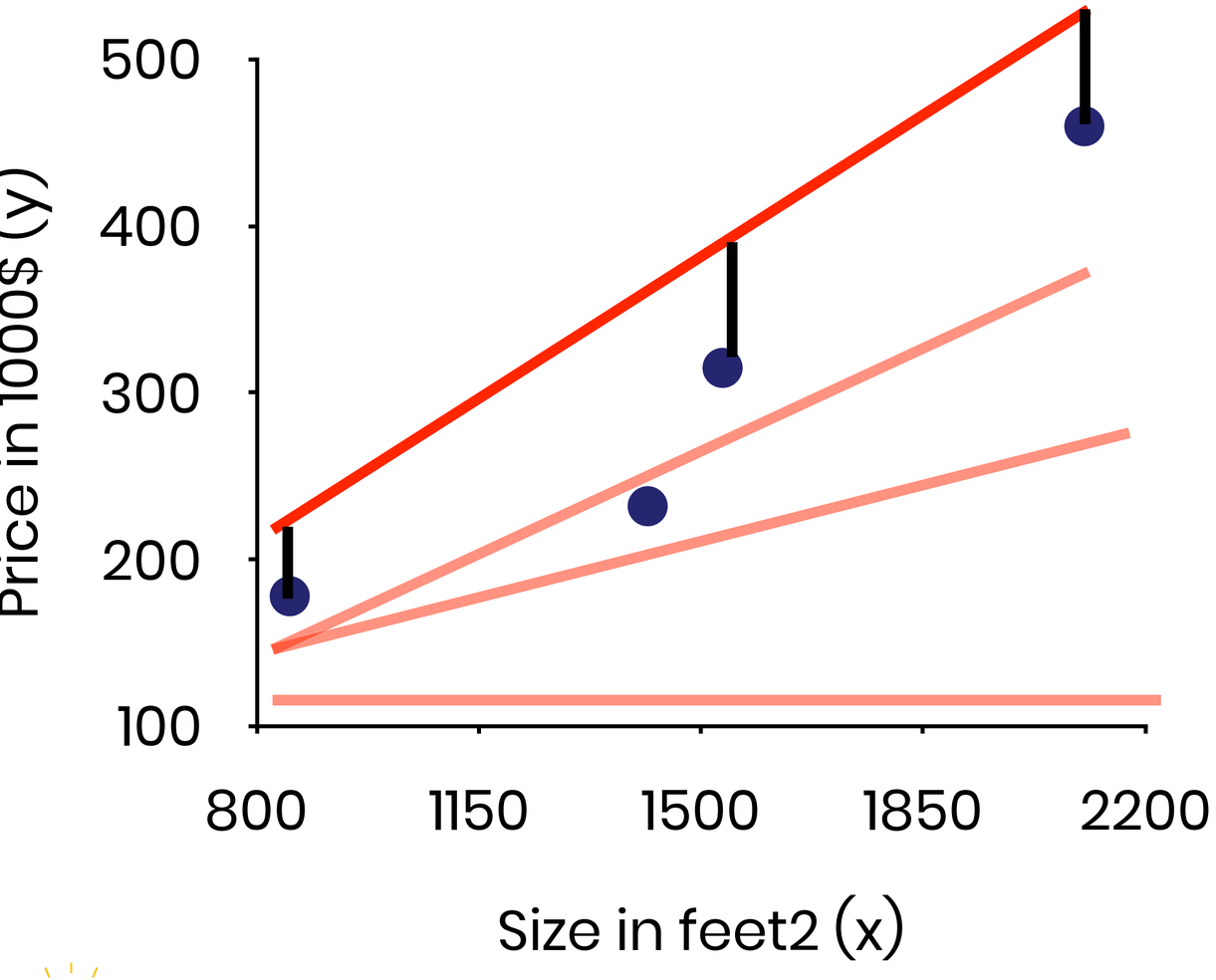
Optimization

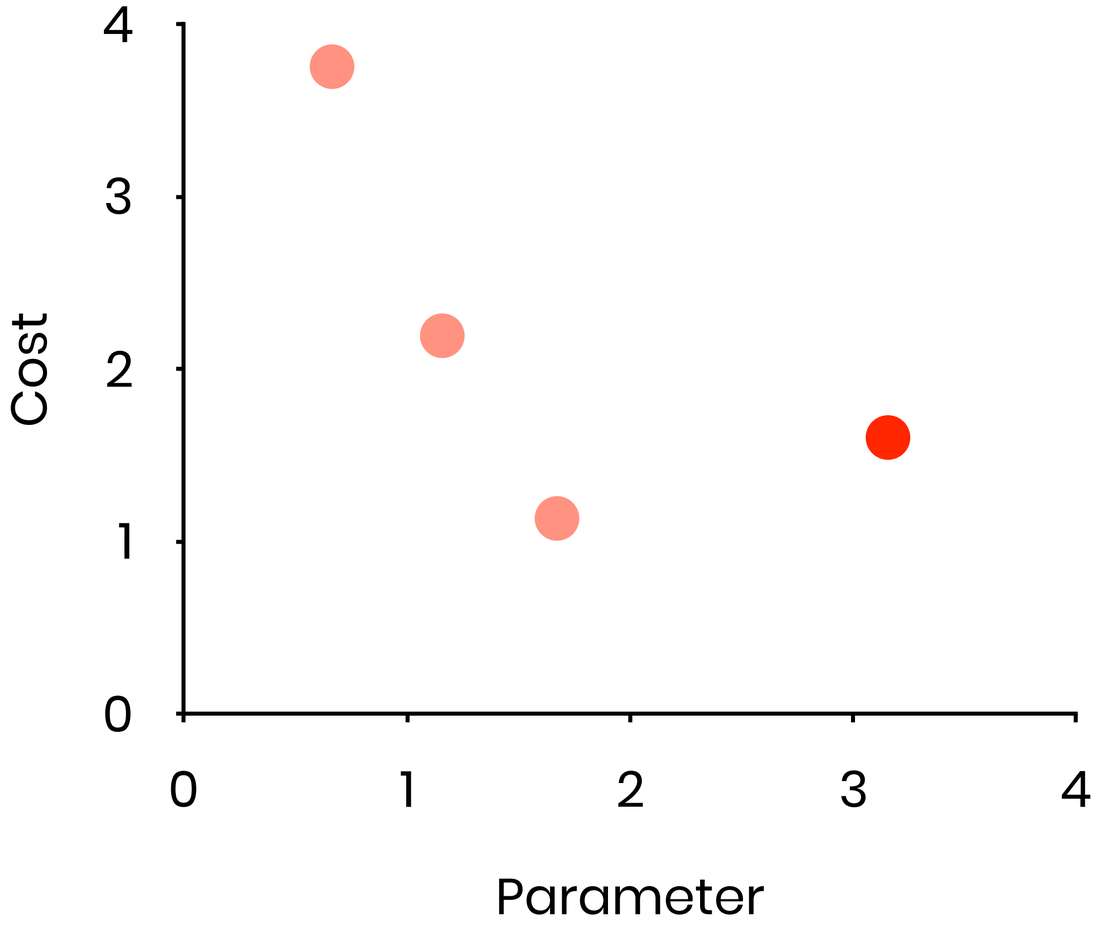




Catalit LLC

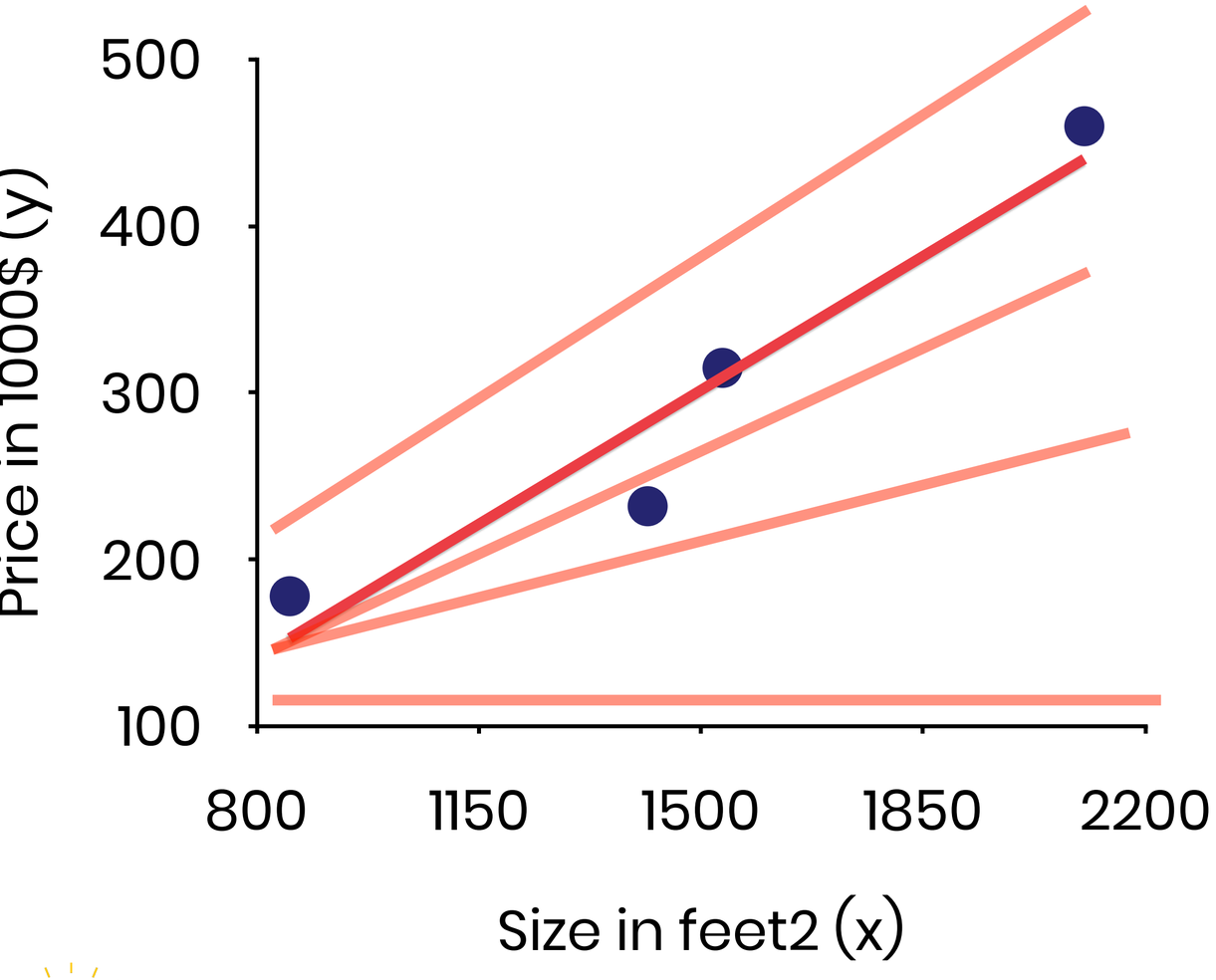
Optimization

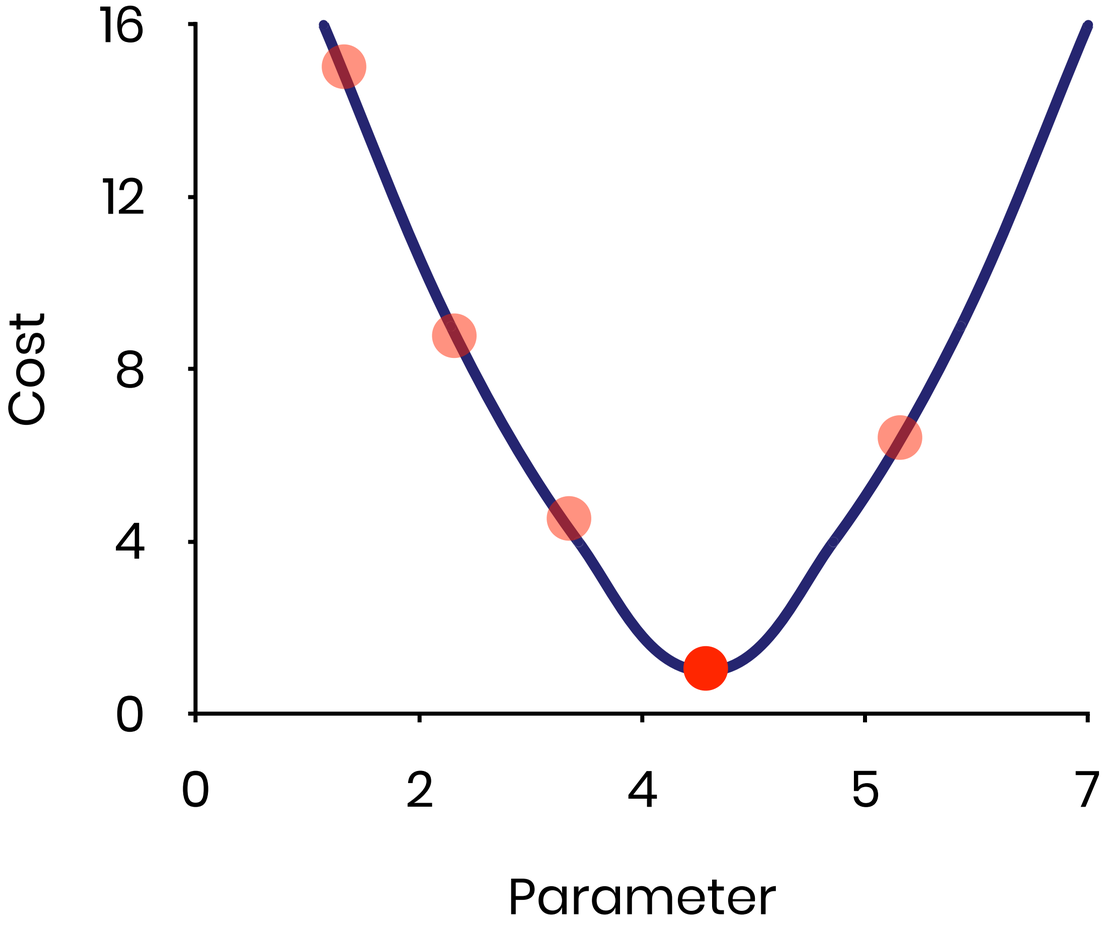




Catalit LLC

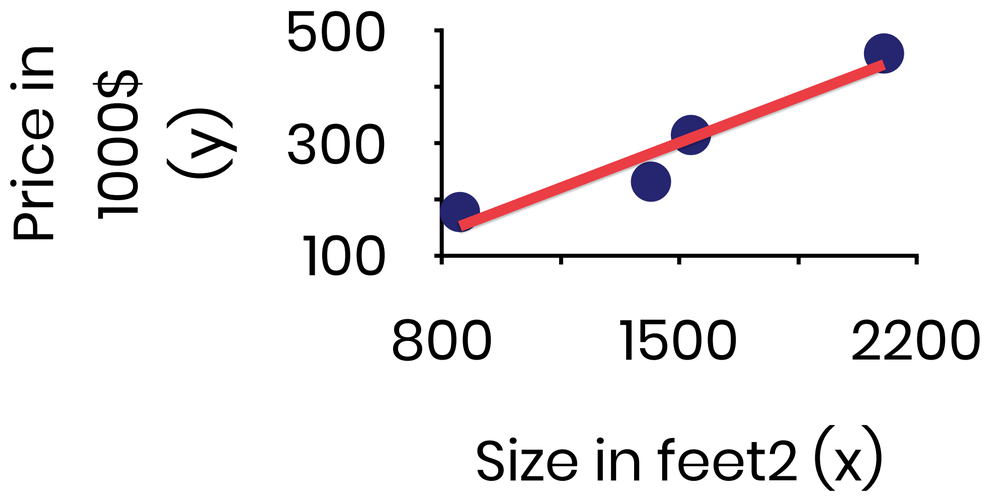
Optimization

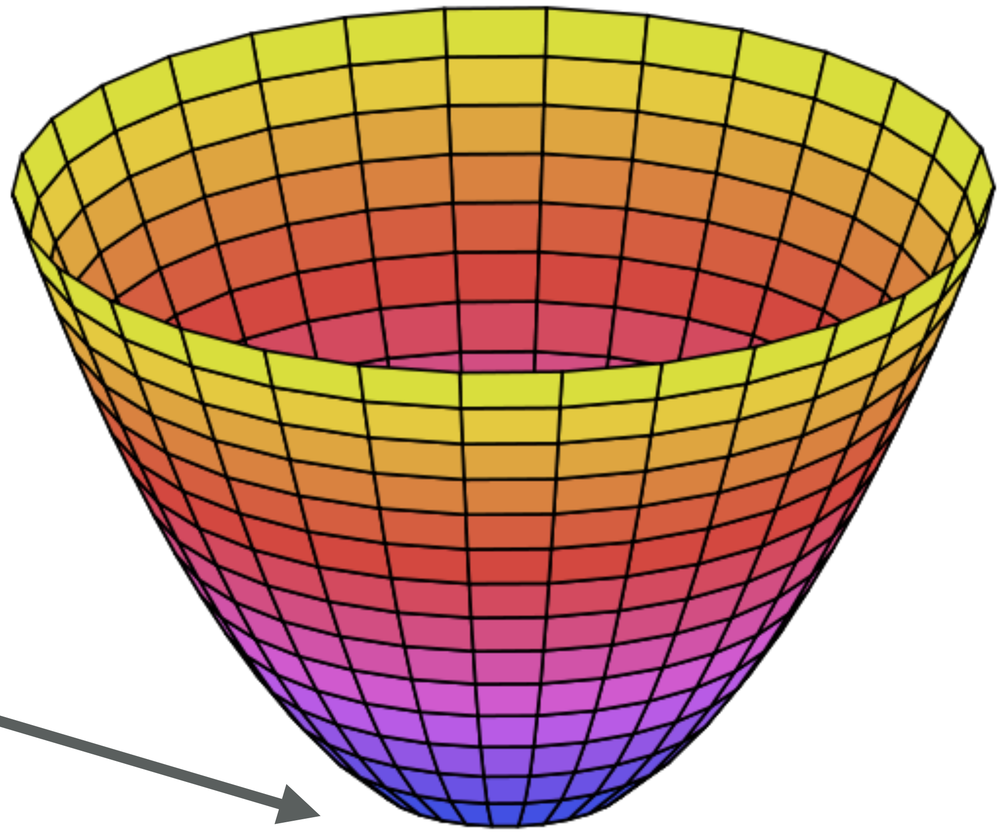




Catalit LLC

Best Model





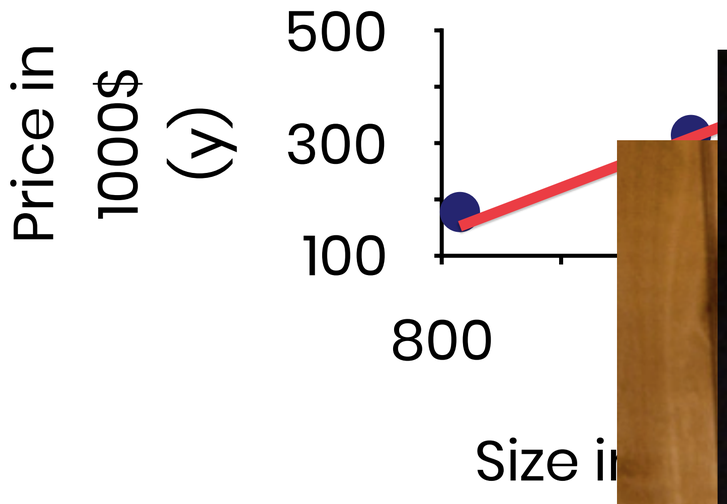
- Values of b and w that give MINIMUM COST

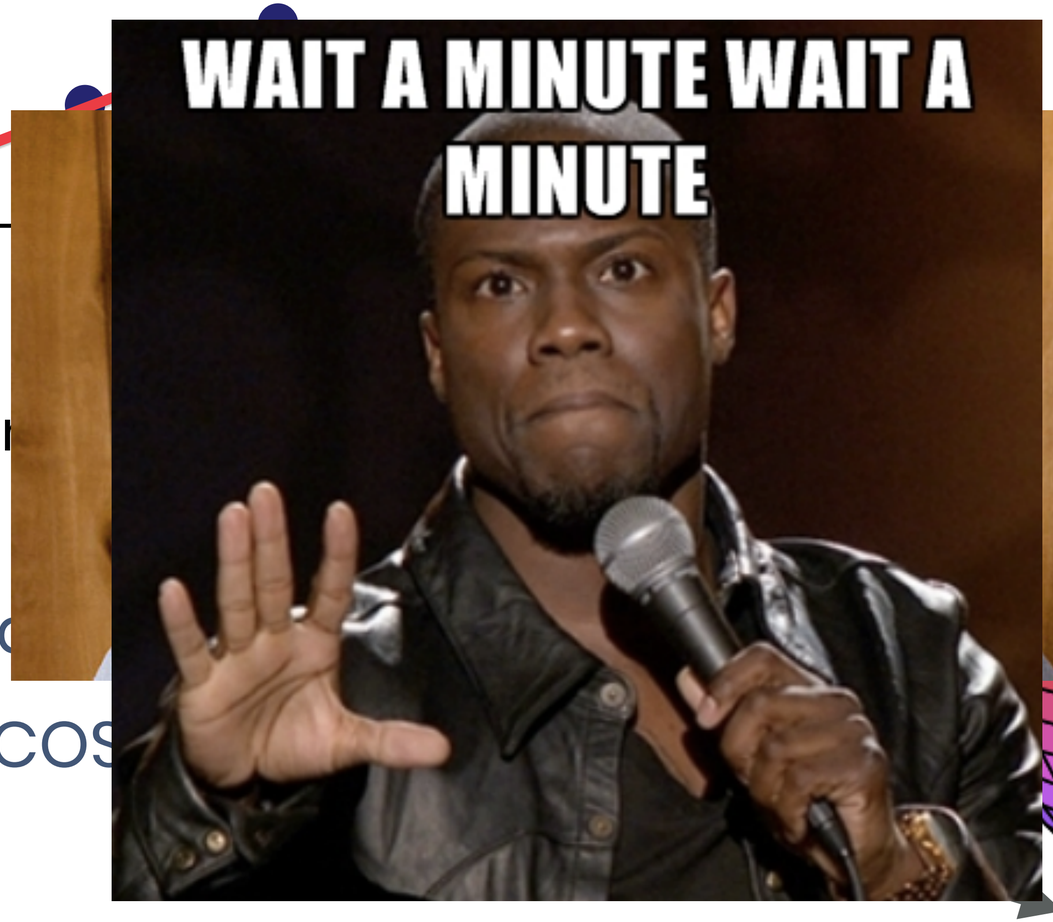
Best Model

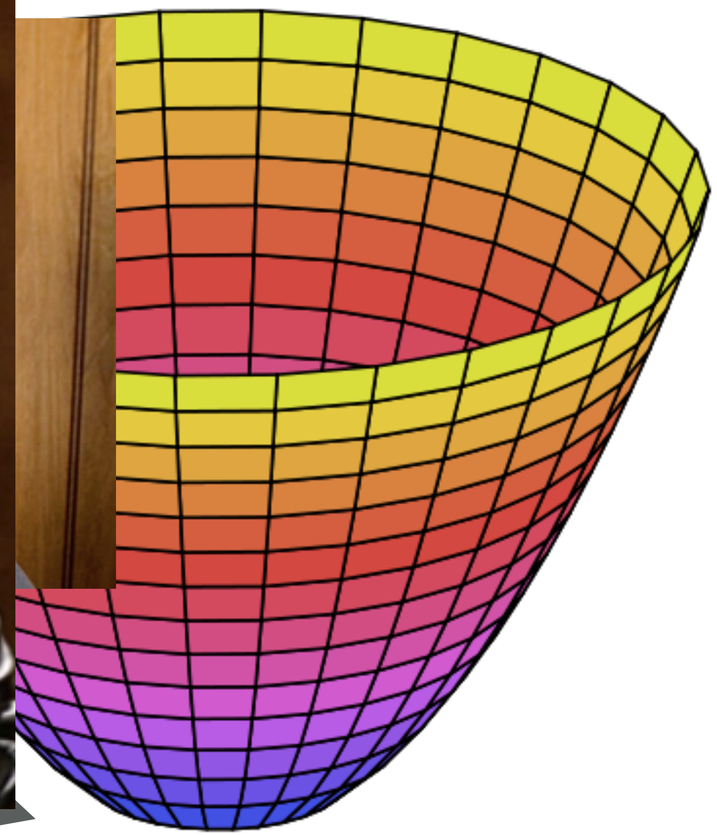
- Values of b and
  give MINIMUM COST

Catalit LLC

Best Model







- Values of b and
  give MINIMUM COS

Catalit LLC

Multivariate & Nonlinear

Multivariate

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
...	...	...	...	...

Catalit LLC

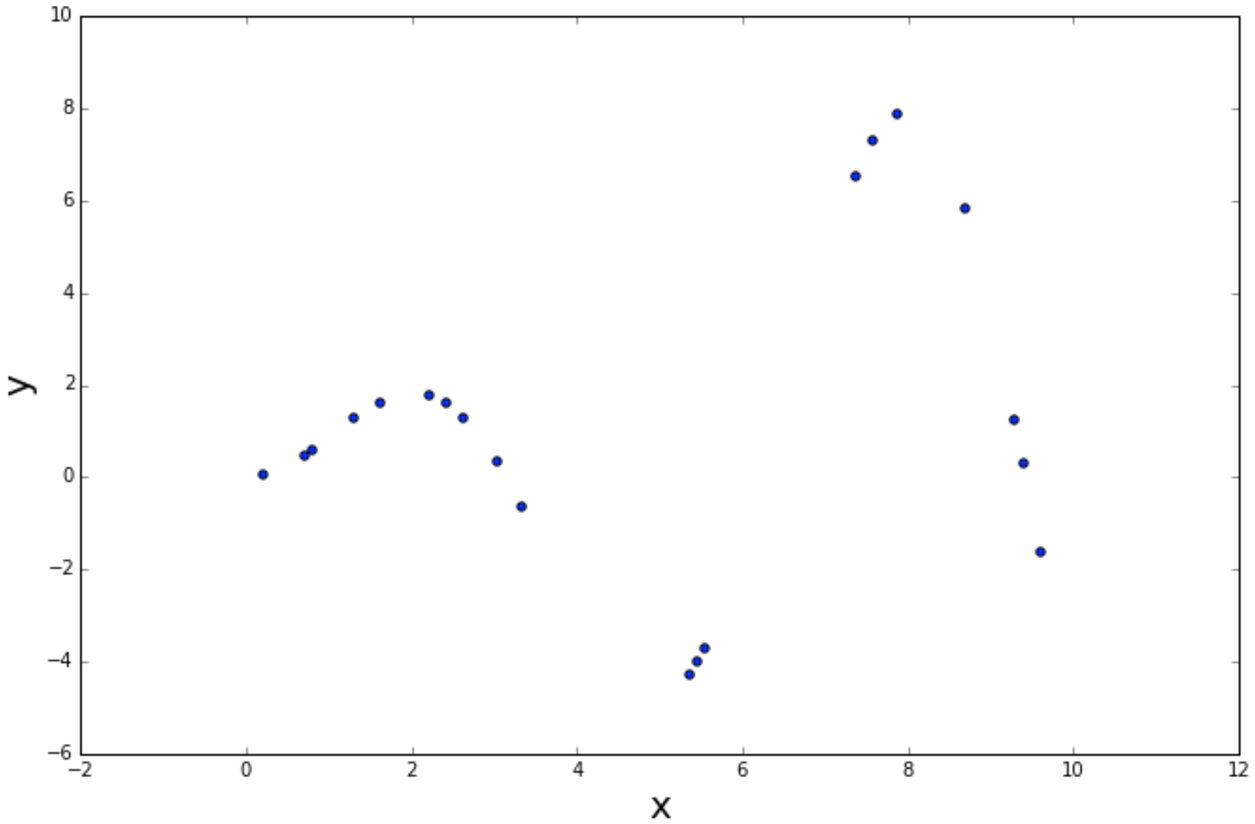
Multivariate

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
...	...	...	...	...

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

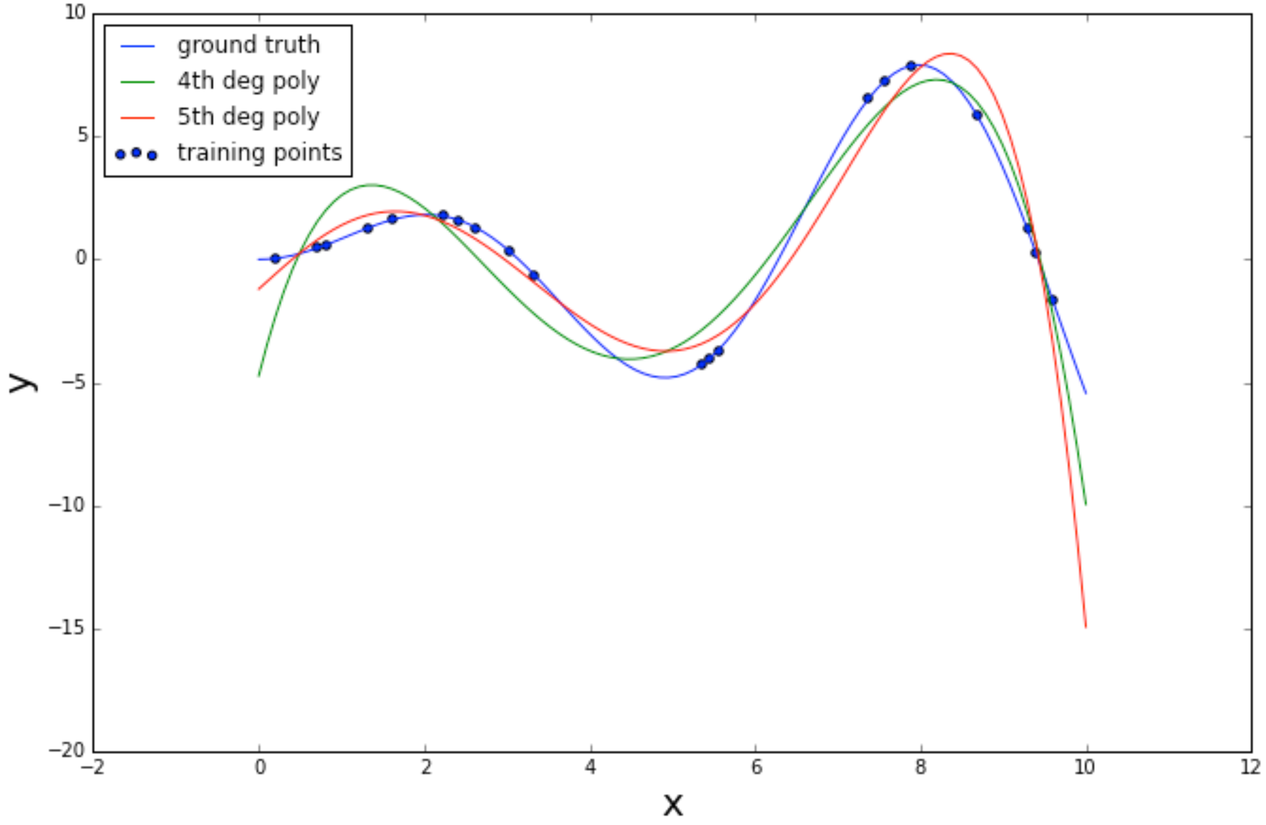
Catalit LLC

Nonlinear



Catalit LLC

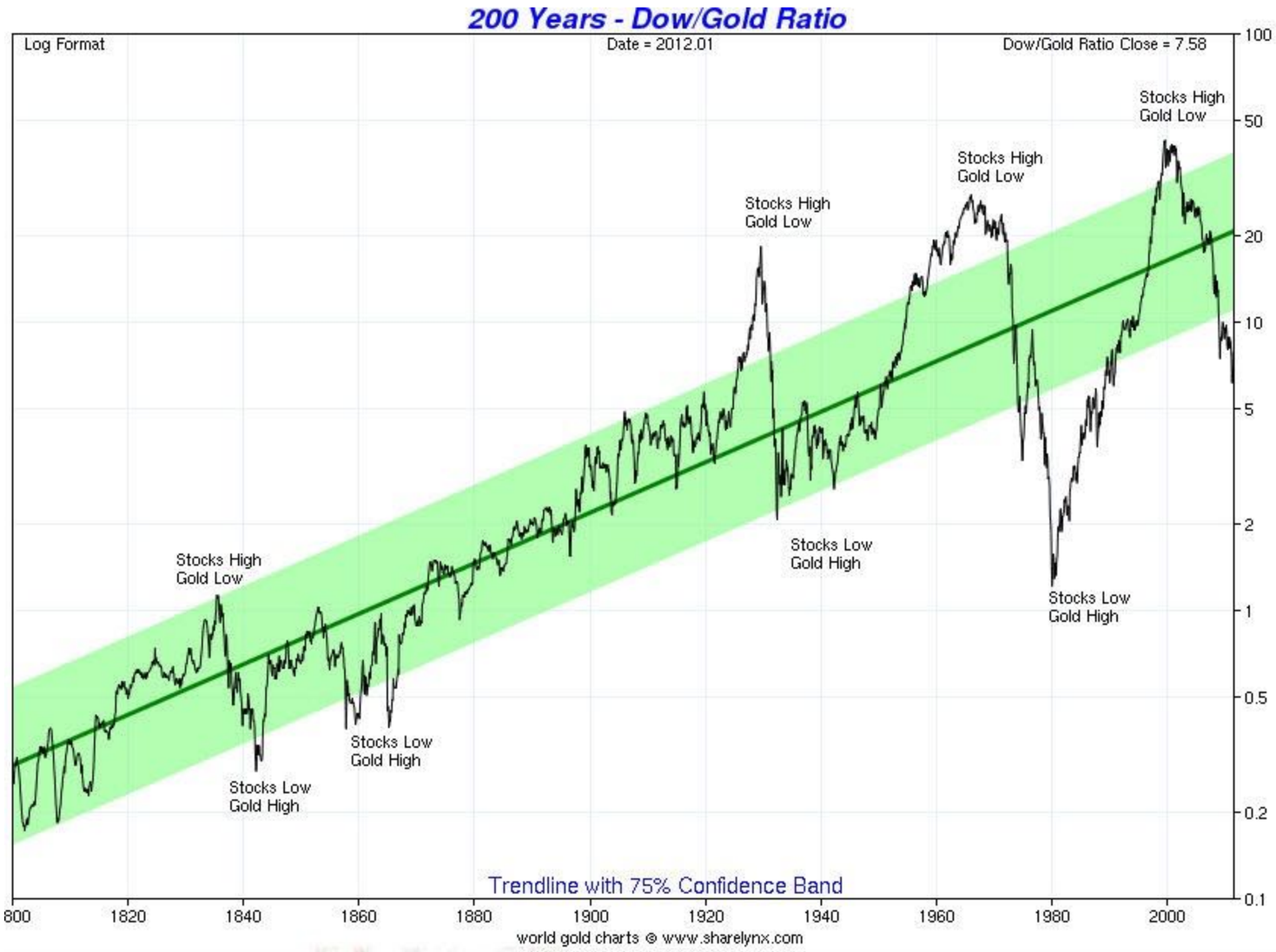
Nonlinear



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

Catalit LLC

Applications



Catalit LLC

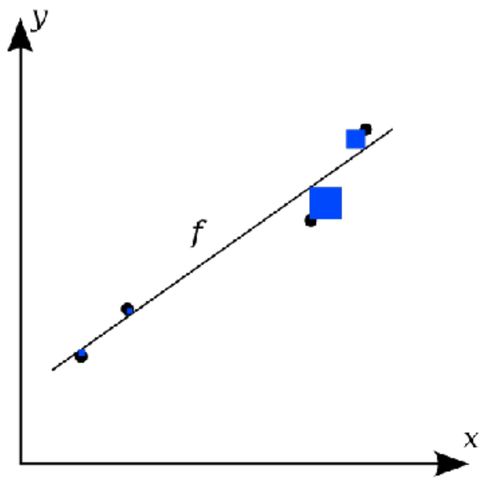
Model Evaluation

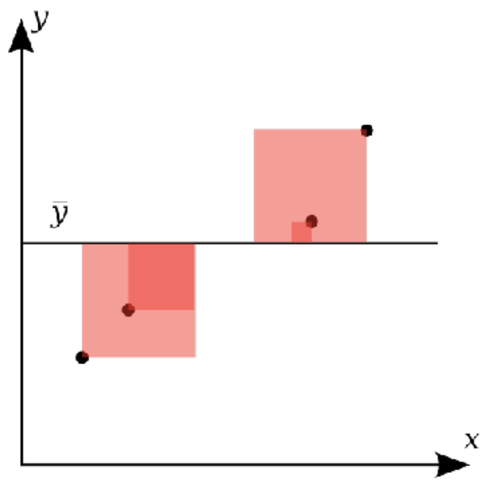
Model Evaluation

- Define Metric
- Know your baseline
- Train model
- Evaluate model on different data (to test generalization)

Define Metric: R^2 score

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

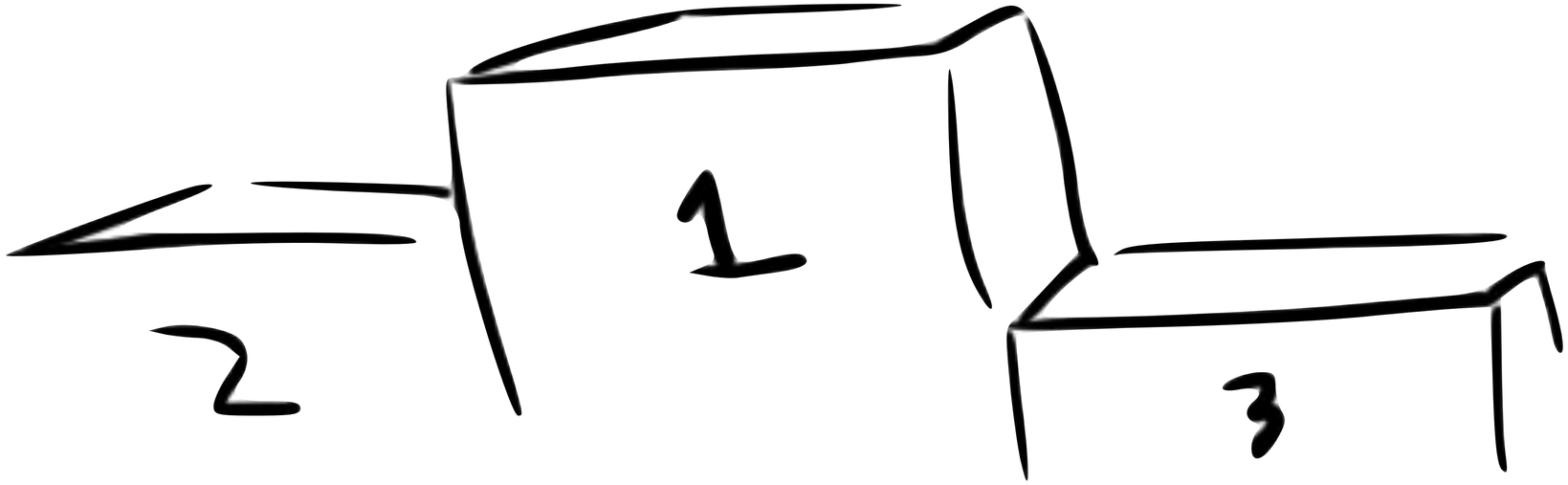




https://en.wikipedia.org/wiki/Coefficient_of_determination#/media/File:Coefficient_of_Determina

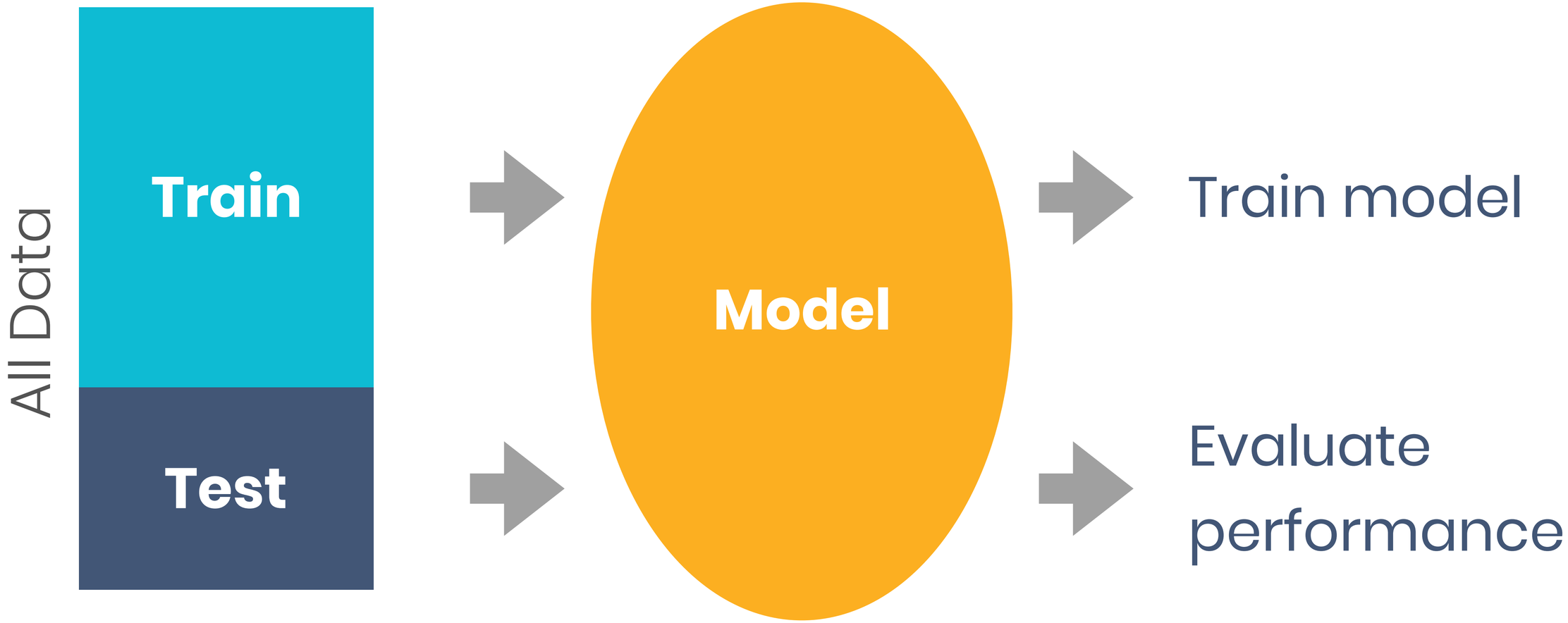
Baseline

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



Catalit LLC

Train / Test split

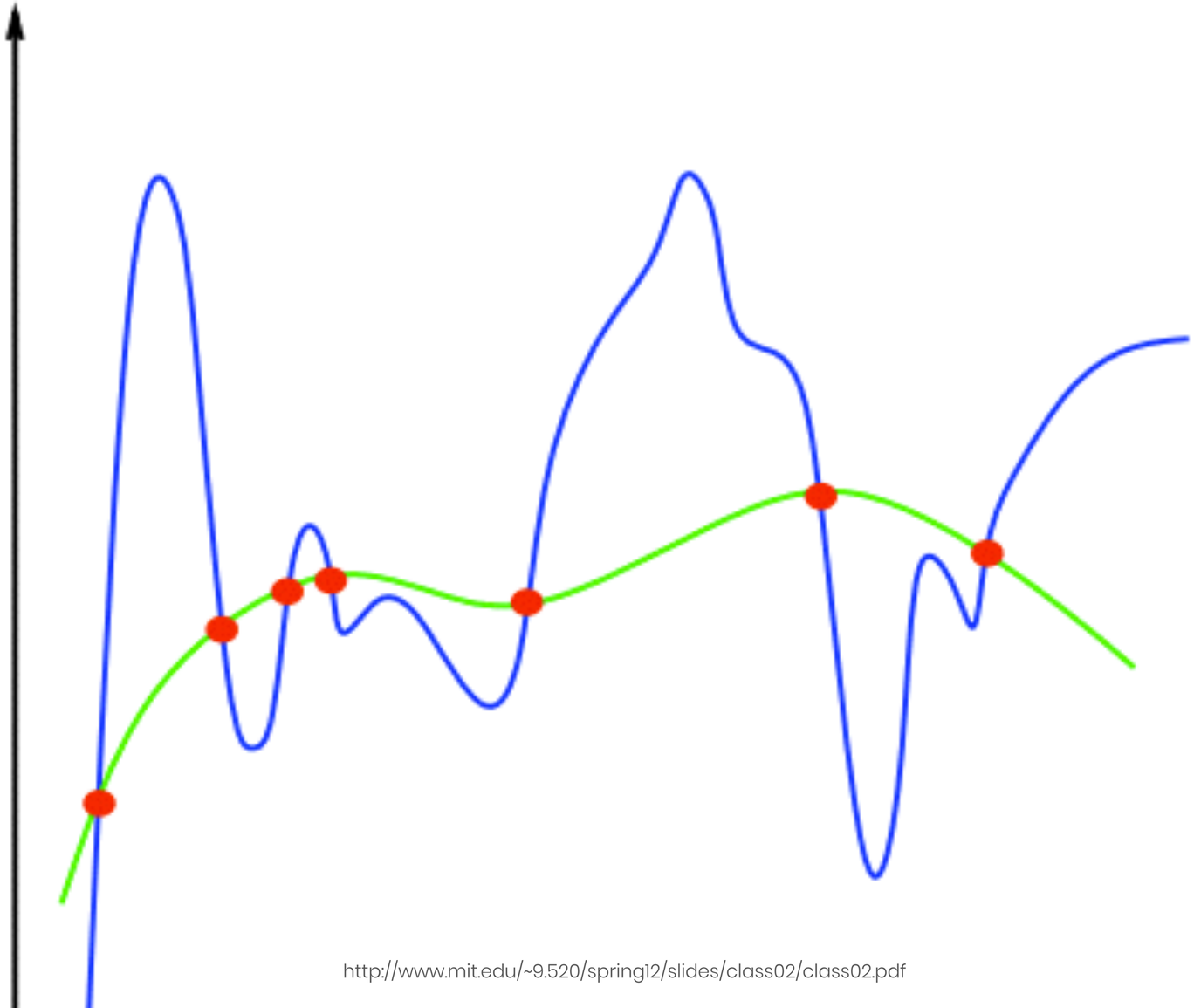


Typical values of test size: [10% - 50%]

Catalit LLC

Overfitting & Regularization

Overfitting



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

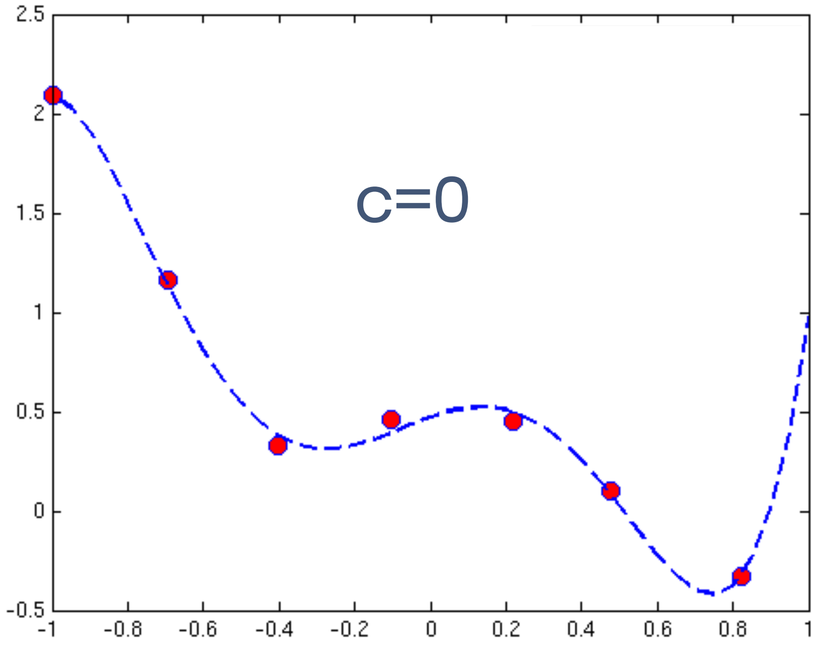
Catalit LLC

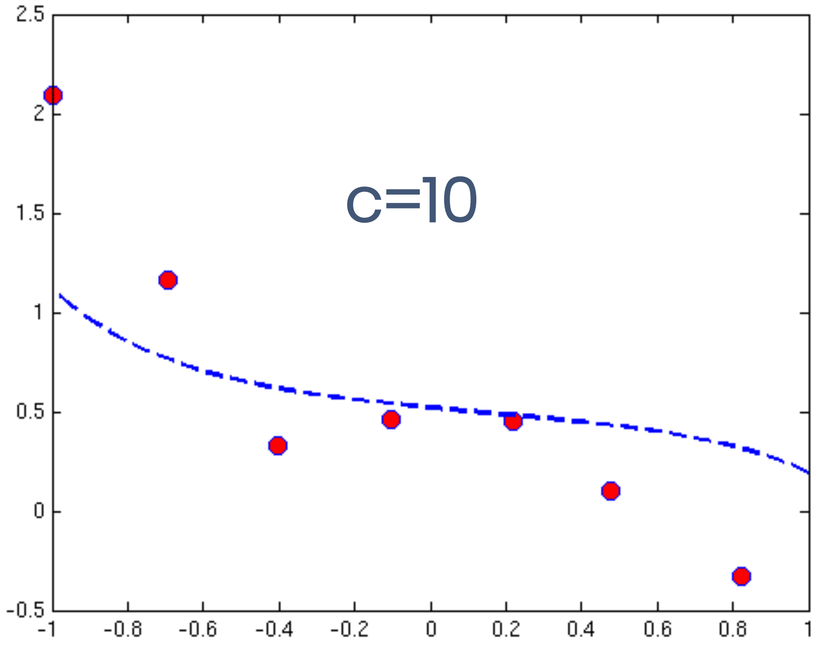
Regularization

Modified Loss => L + cR

Ridge: $R = \sum_i w_i^2$

Lasso: $R = \sum_i |w_i|$





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

Catalit LLC

Lab 03

- **Jupyter**
  - interactive Python web application => our dev env
- **Pandas**
  - like Excel for python => data manipulation
- **Matplotlib & Seaborn**
  - Plots & charts => data visualization
- **Scikit-Learn**
  - Machine Learning => model building, training & evaluation

Classification

Classification

- Intuition and definition
- Classifiers
- Ensemble methods
- Model Evaluation
- Multi-class
- Working with text data

Catalit LLC

Summary of ML

	CATEGORICAL	CONTINUOUS
SUPERVISED	?	?
UNSUPERVISED	?	?

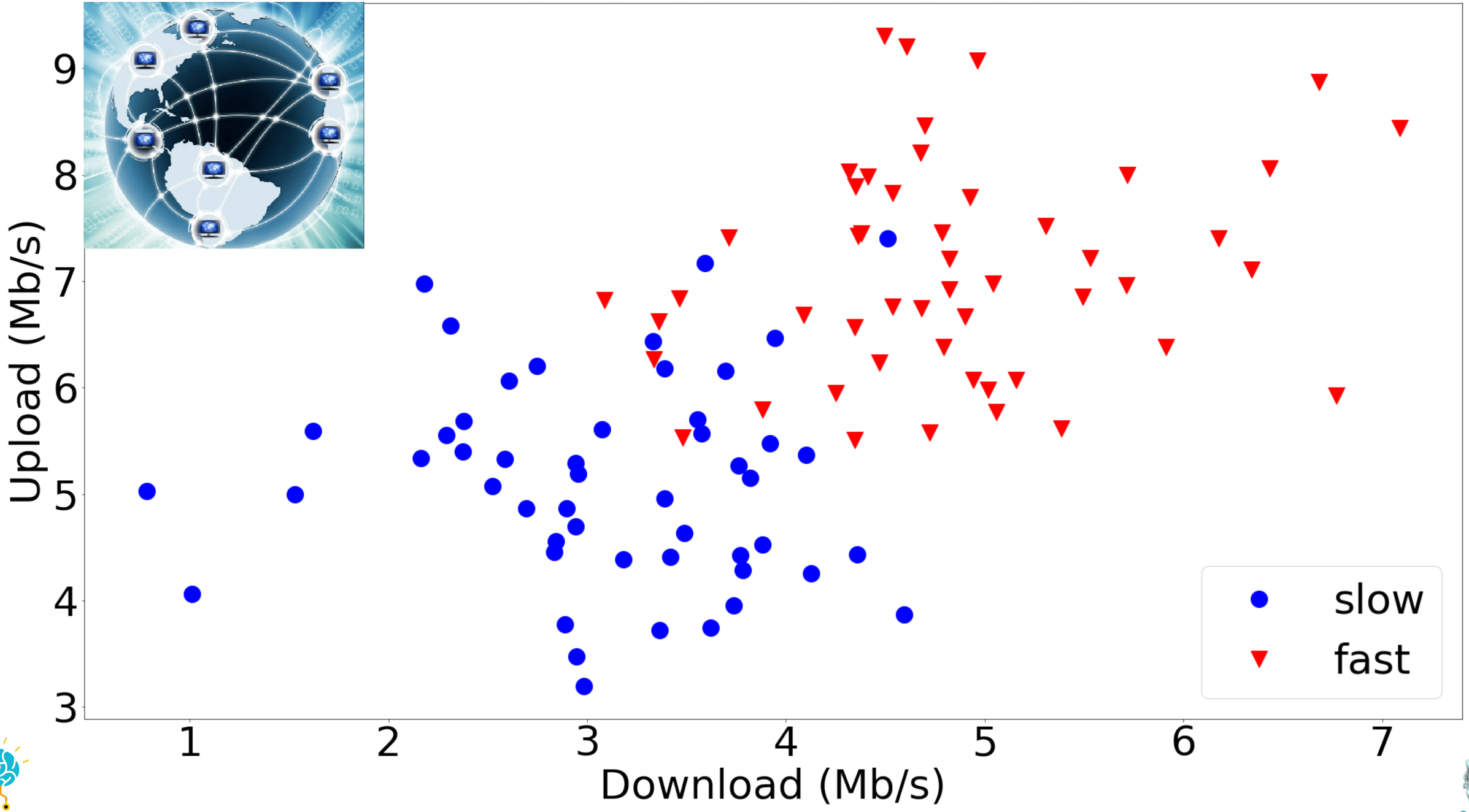
Catalit LLC

Summary of ML

	CATEGORICAL	CONTINUOUS
SUPERVISED	CLASSIFICATION	REGRESSION
UNSUPERVISED	CLUSTERING	DIM REDUCTION

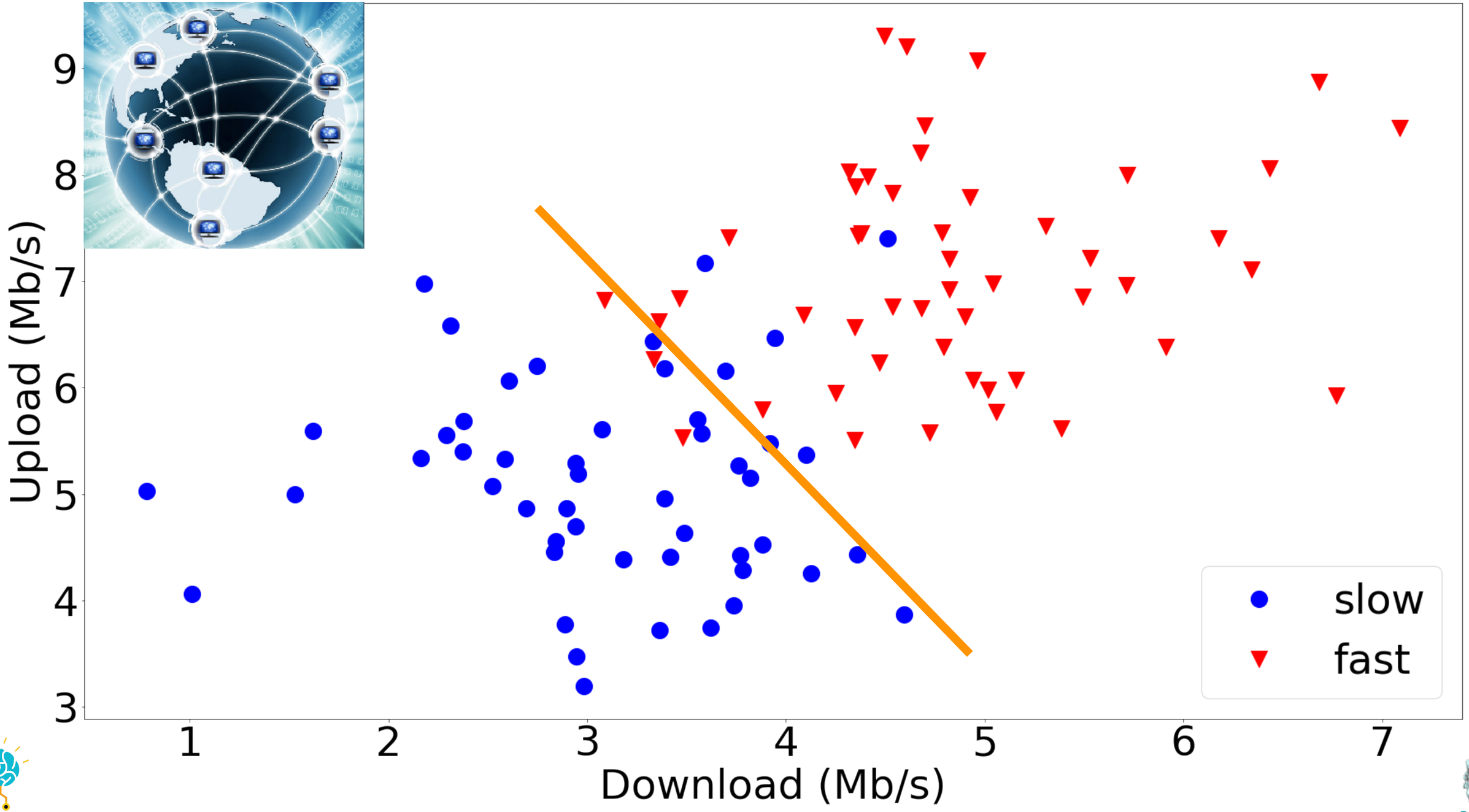
Catalit LLC

Internet Service Providers



Catalit LLC

Internet Service Providers



Catalit LLC

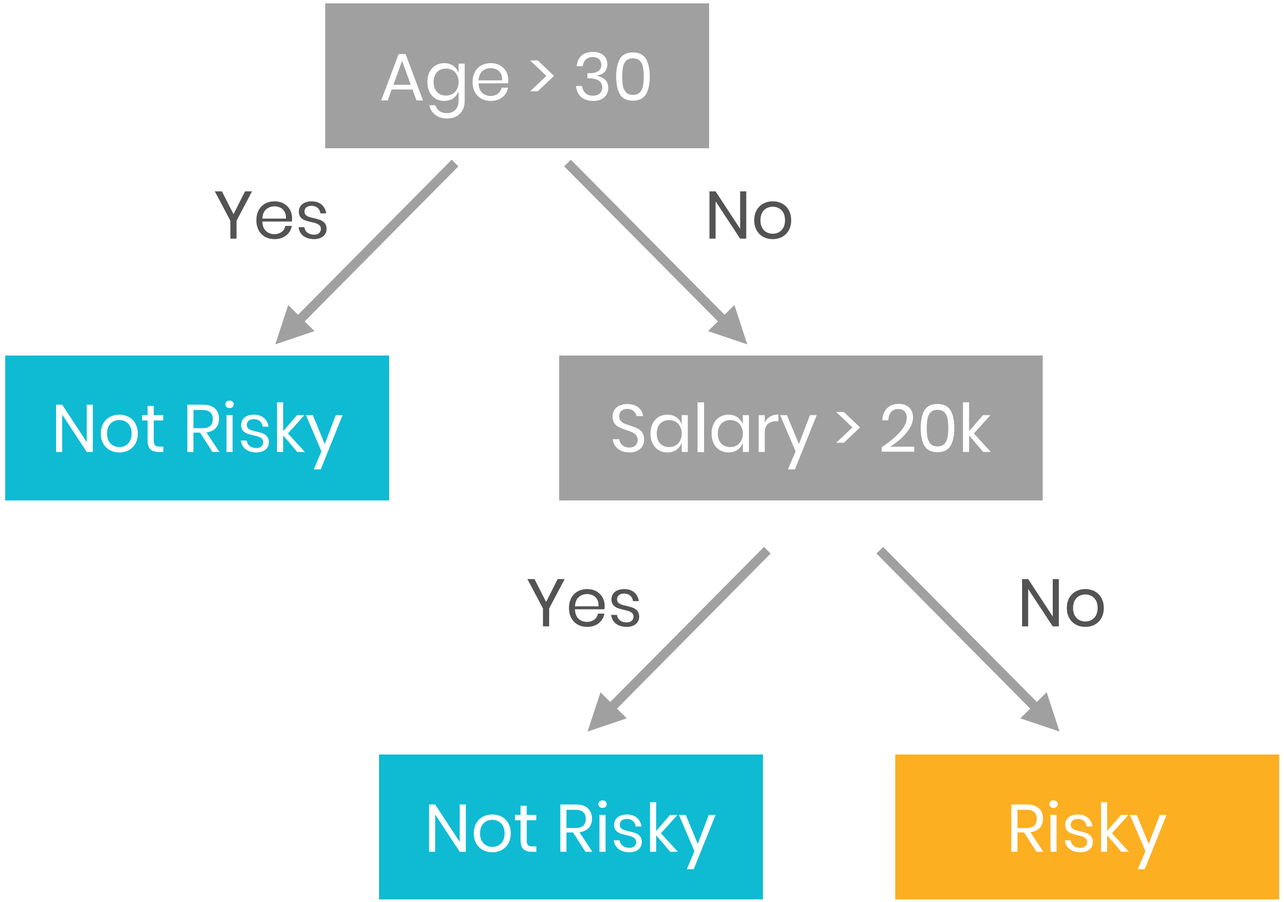
Binary classification

- Labels are only 1 or 0 (True/False)
  - Spam
  - Fraud
  - Pass/Fail
  - ....

Classifiers

Decision Tree

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

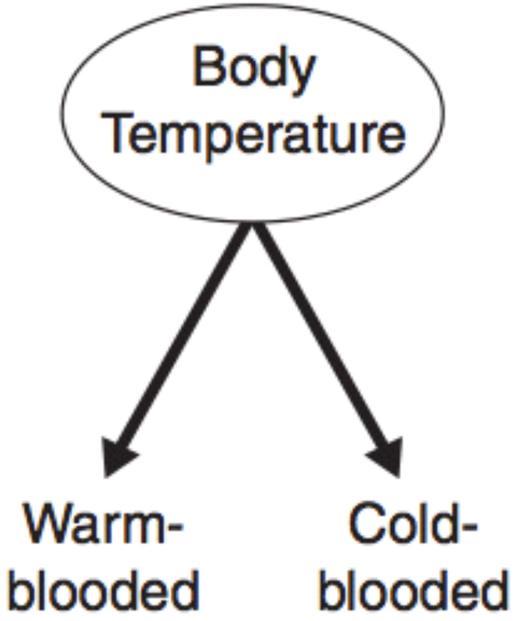


GOAL: split data with decision rules

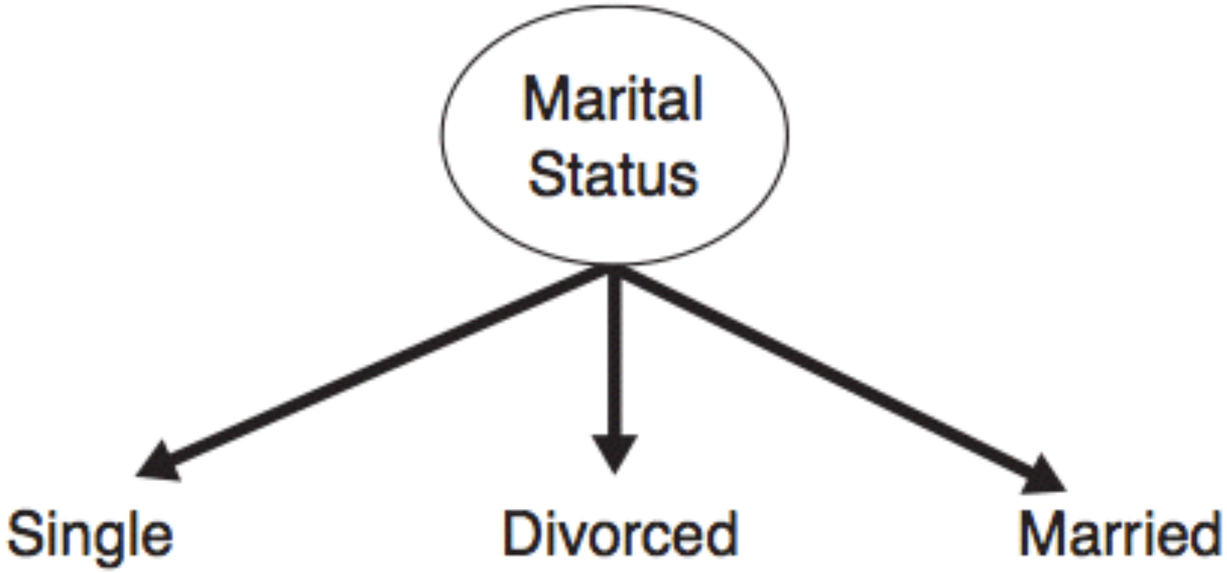
Catalit LLC

Decision Tree

Binary Split

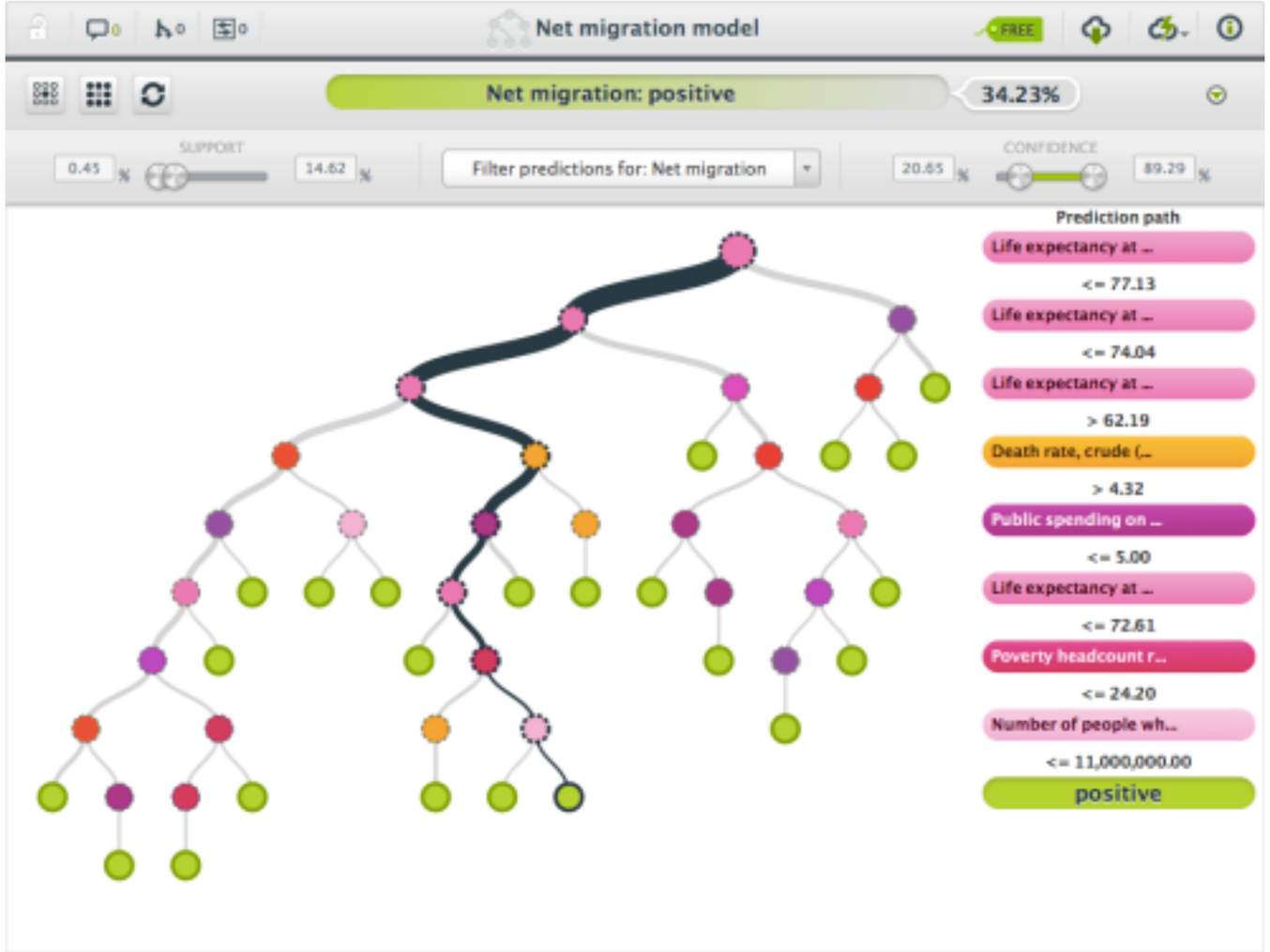


Multiway Split



Applications

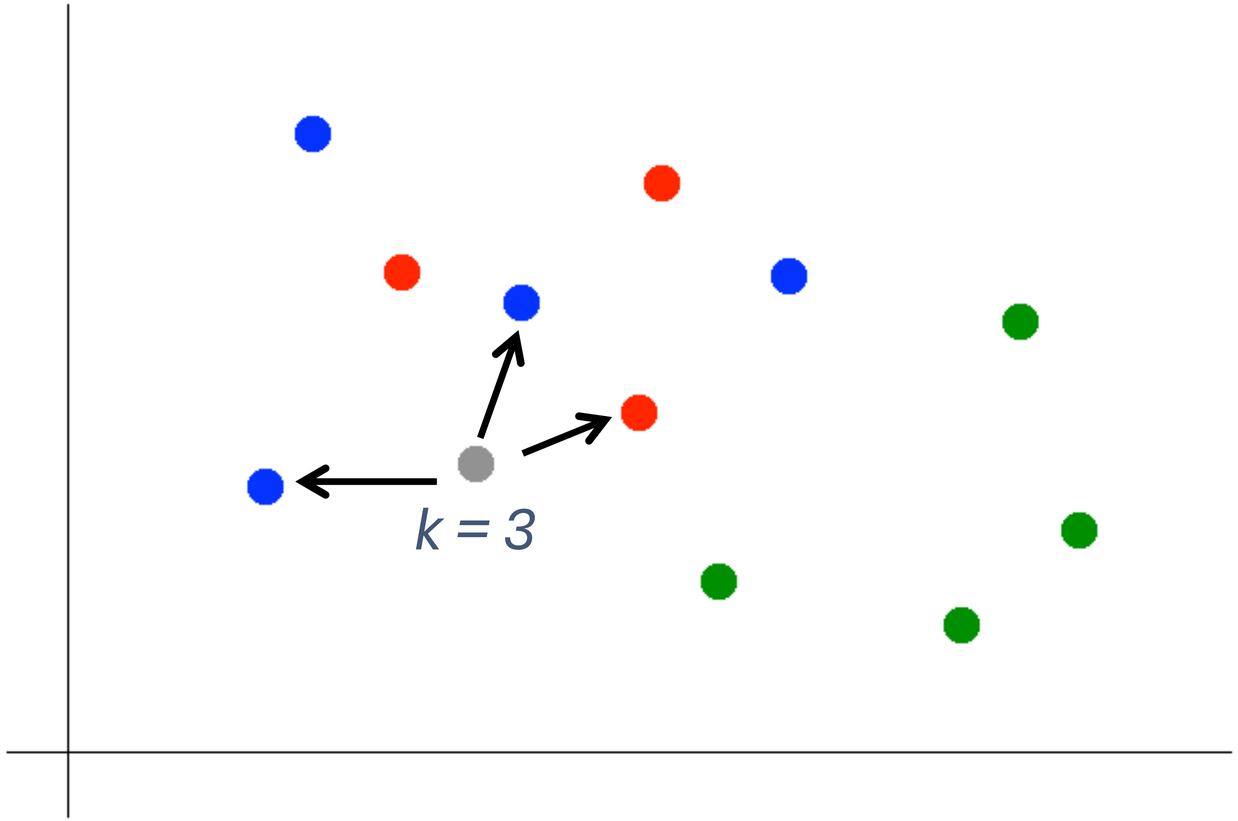
- Lead scoring
- Customer churn
- User segmentation



Catalit LLC

K-Nearest Neighbors

- Advantages:
  - easy to understand
  - any boundary



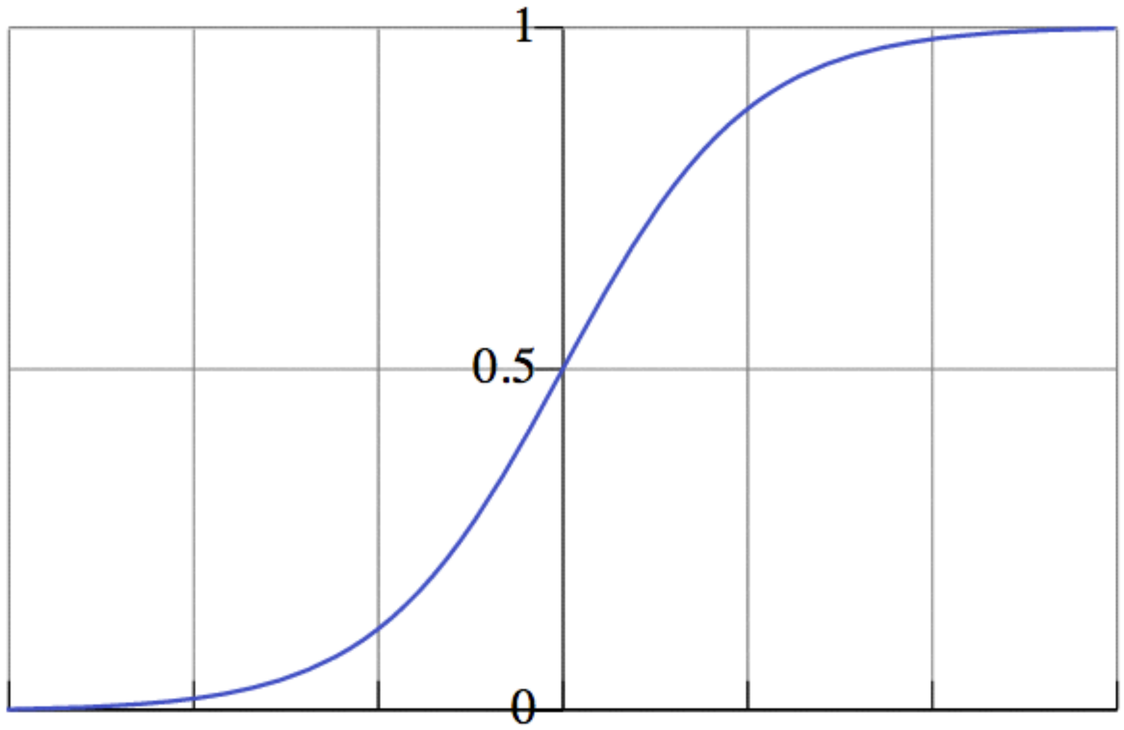
GOAL: majority vote of points nearby

Catalit LLC

Logistic Regression

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

probability of belonging to class



value of independent variable

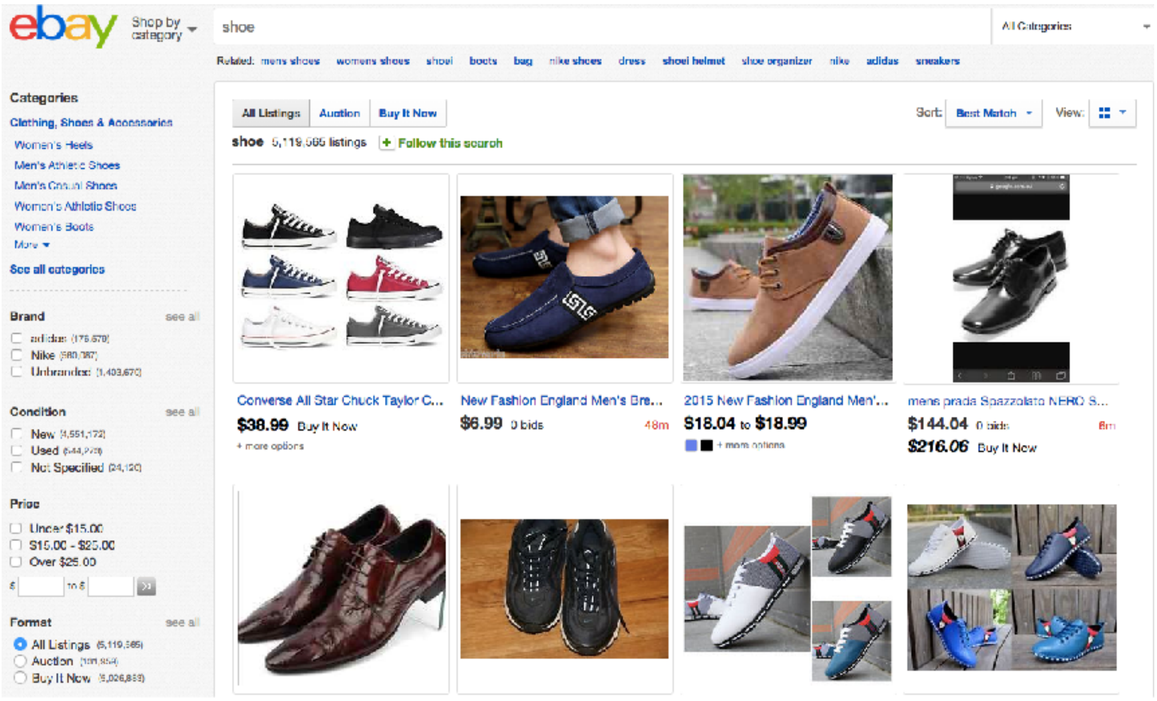
GOAL: minimize error in probability space

Catalit LLC

Applications

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

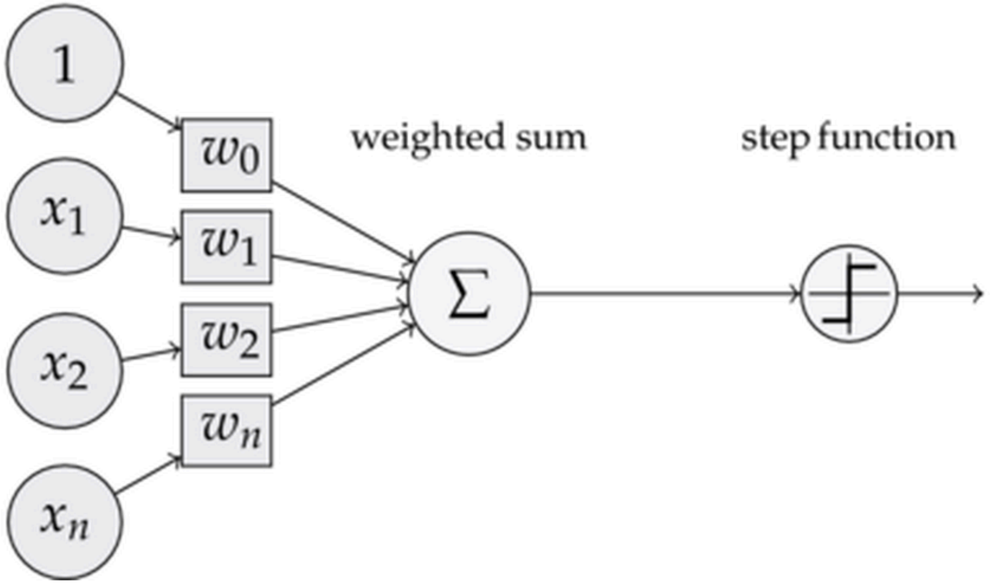
- The model
  - Output: Click/no click
  - Input features:
    - page content
    - search keyword
    - User information



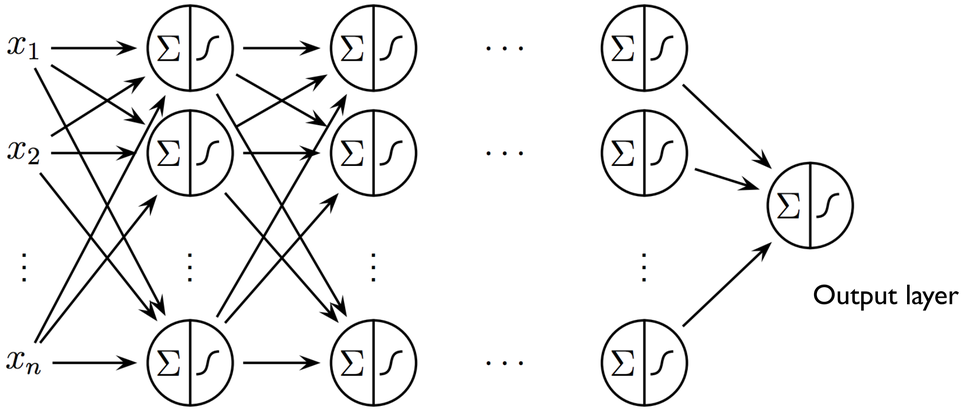
Catalit LLC

Neural Networks

- Perceptron



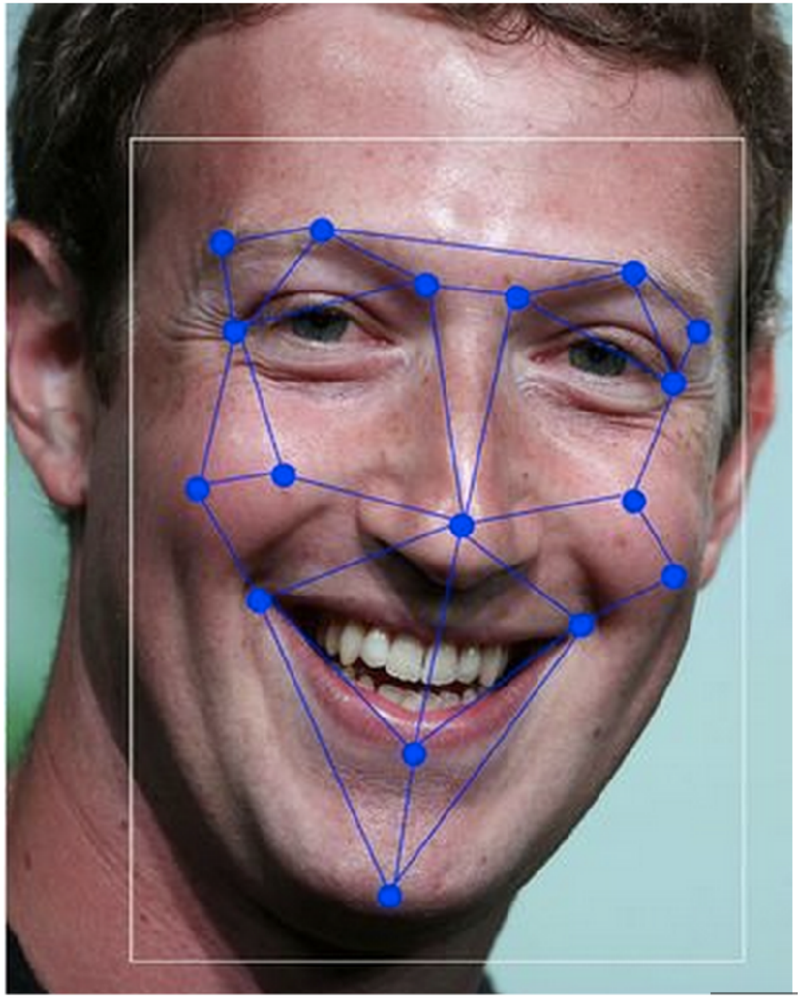
- Multi-layer

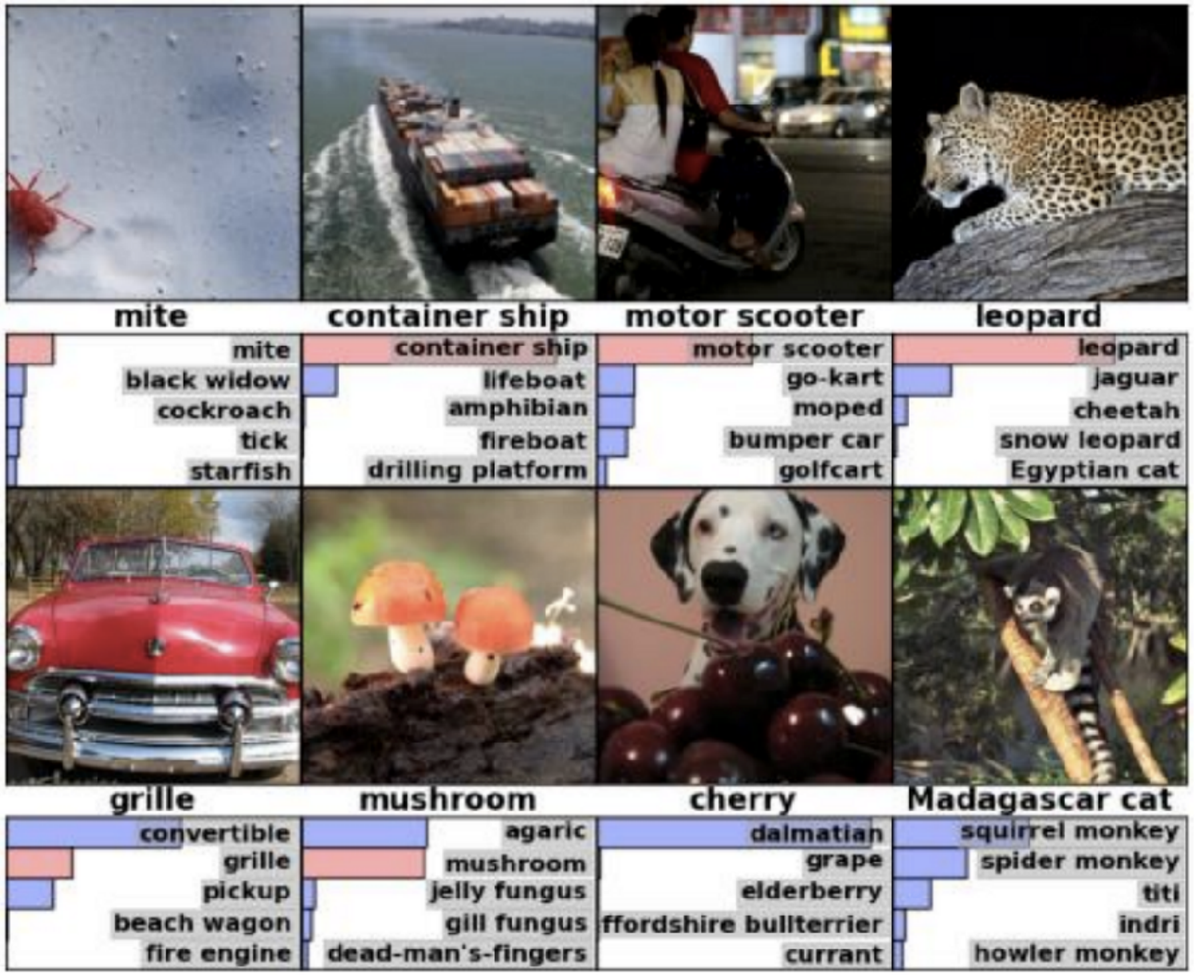


GOAL: adjust weights to minimize output error

Catalit LLC

Applications





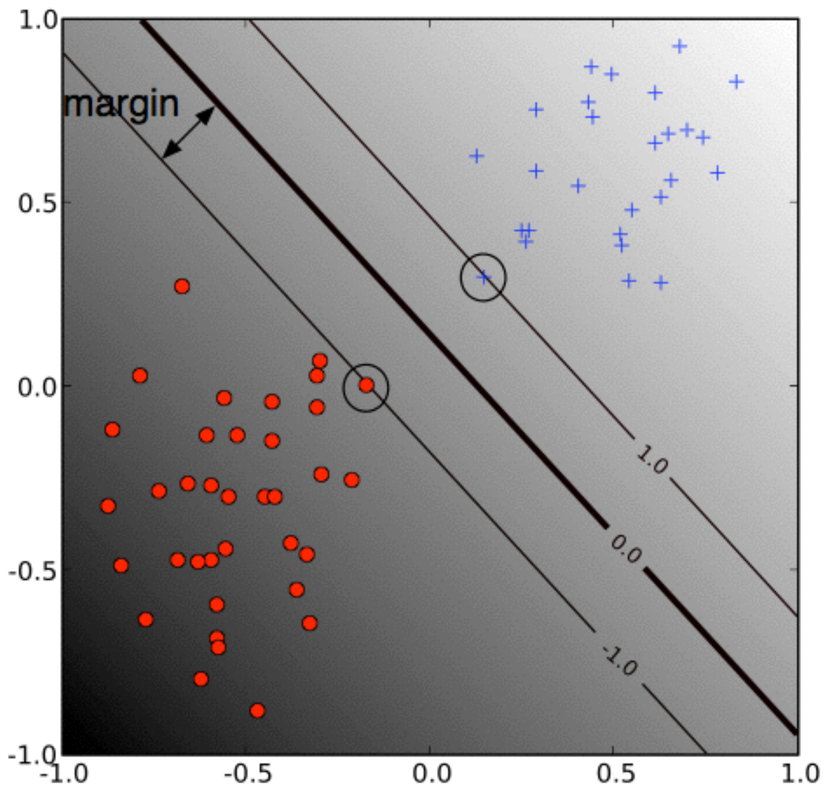
https://research.facebook.com/publications/480567225376225/deepface-closing-the-gap-to-human-level-performance-in-face-verification/

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

Catalit LLC

Support Vector Machine

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



GOAL: find maximally separating hyperplane

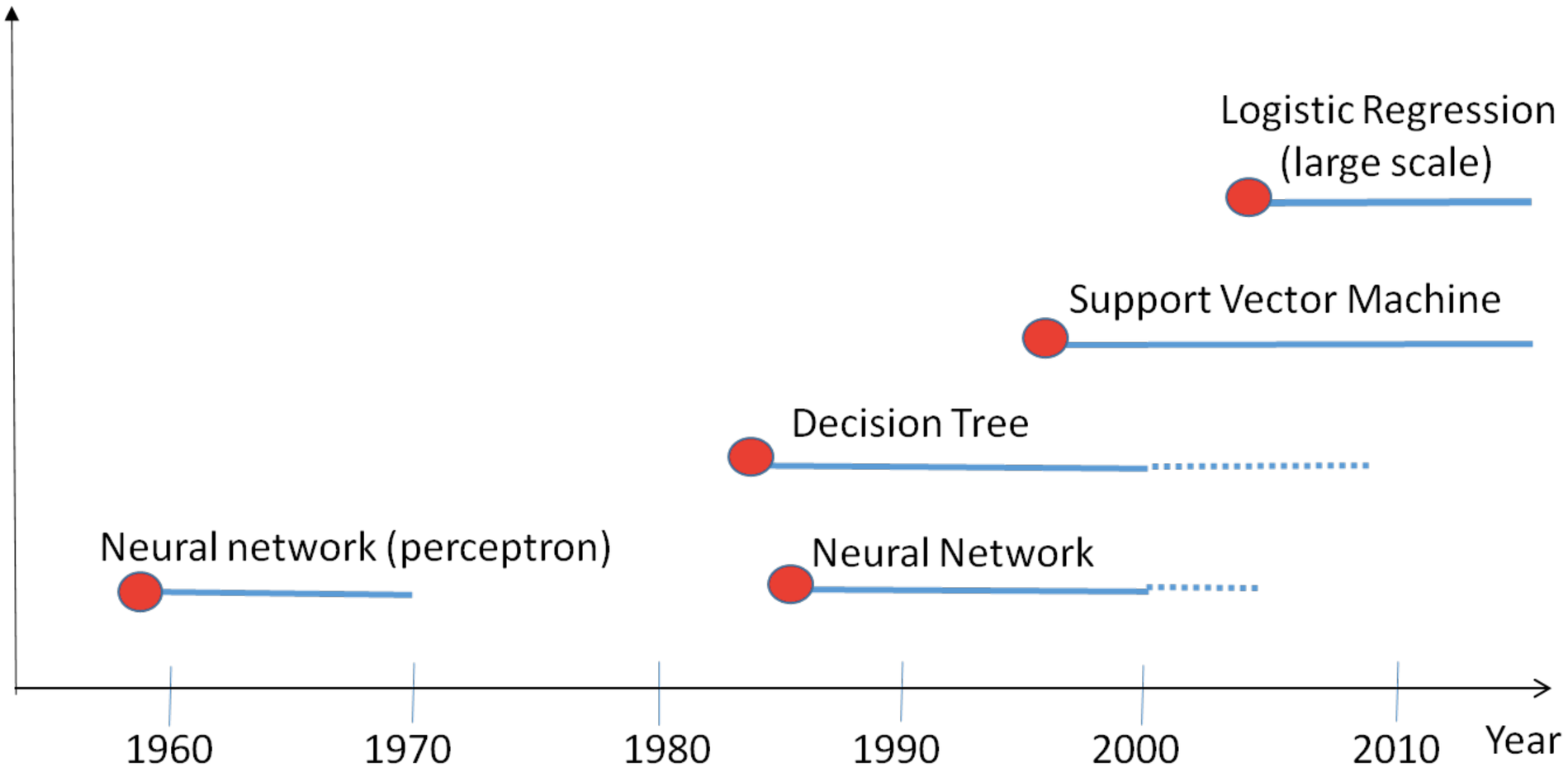
Catalit LLC

Applications

- 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

Catalit LLC

History of supervised learning

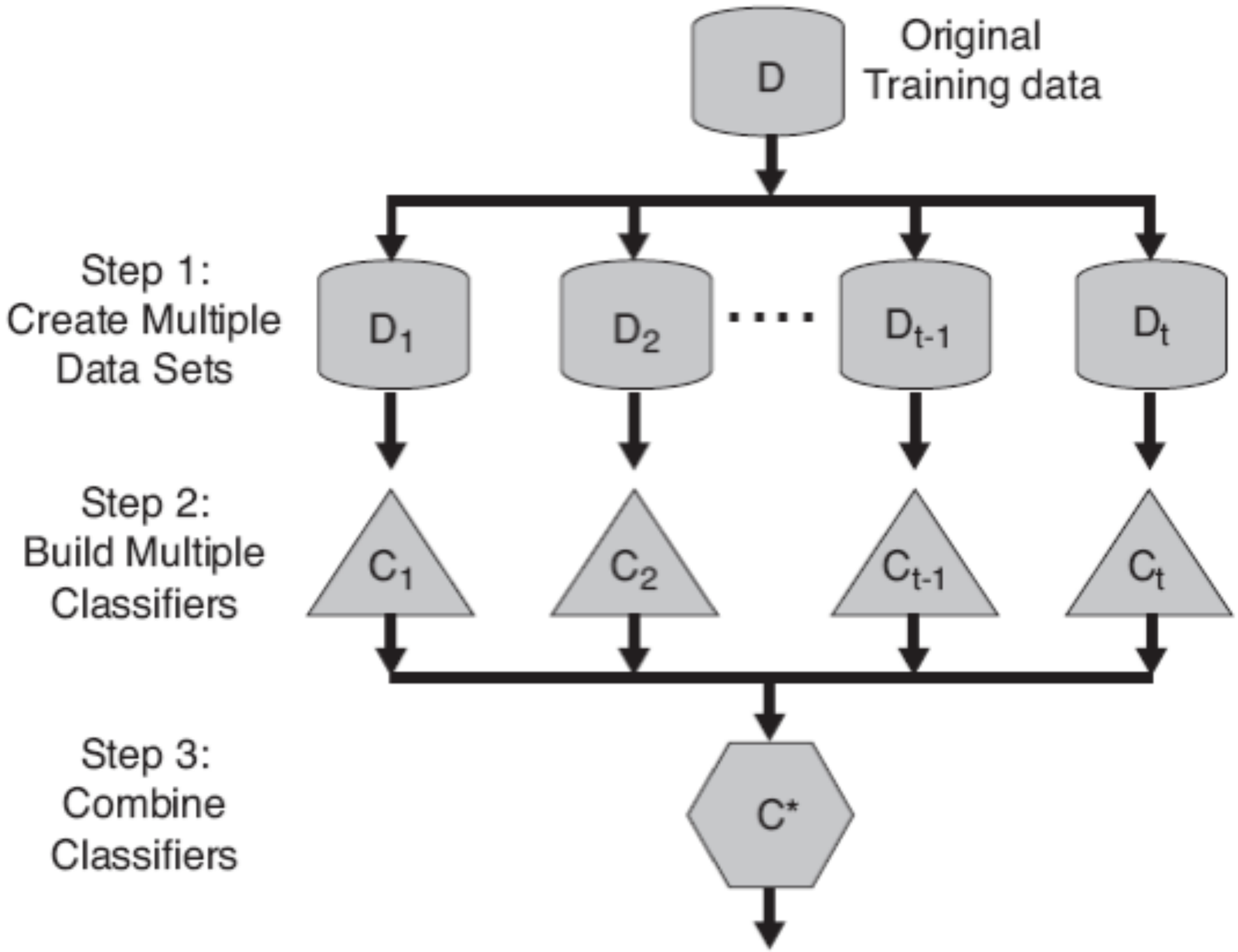


http://www.aboutdm.com/2013/04/history-of-machine-learning.html

Catalit LLC

Ensemble Methods

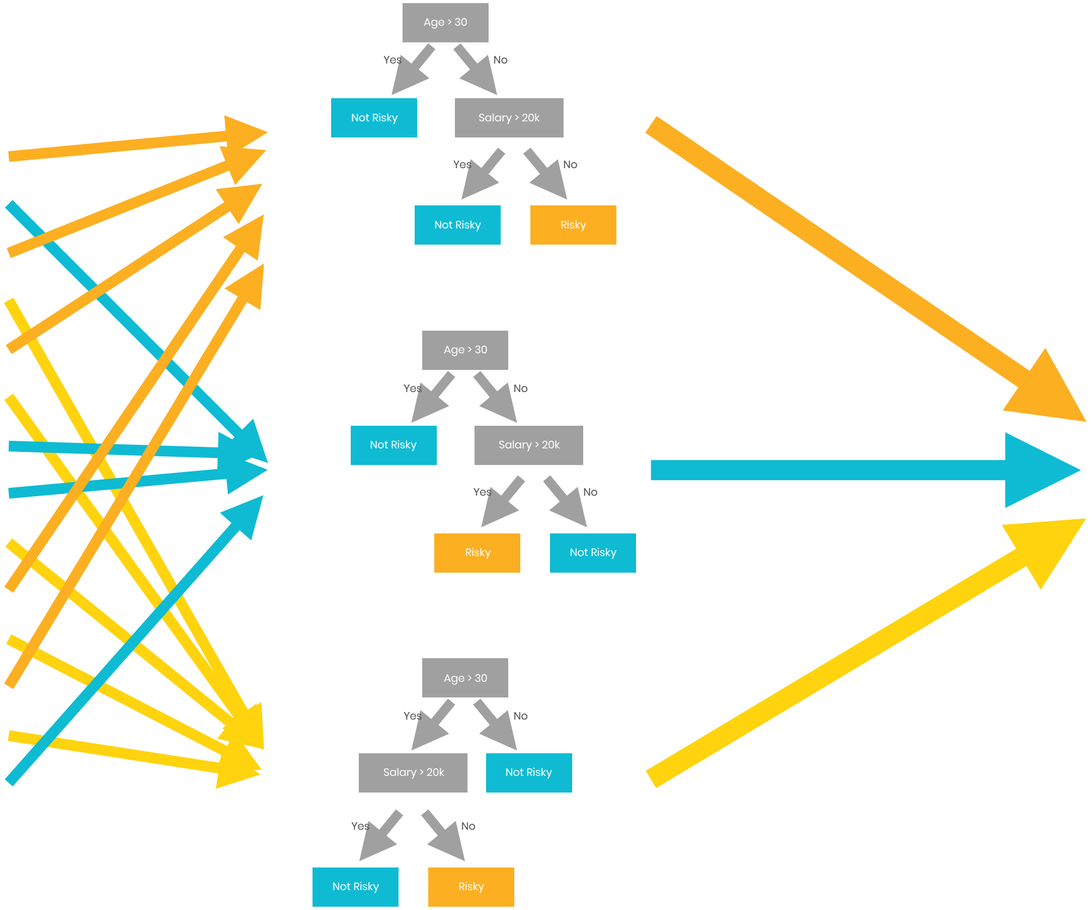
Ensemble Methods



Catalit LLC

Bagging

Client	Age	Gender	Annual Salary	Months in residence	Months in job	Current Debt	Paid off credit
1	23	Male	$30,000	36	12	$5,000	Yes
2	30	Female	$45,000	12	12	$1,000	Yes
3	19	Male	$15,000	3	1	$10,000	No
4	25	Male	$25,000	12	2	$15,000	Yes
5	32	Female	$45,000	12	3	$5,000	No
6	26	Male	$35,000	18	5	$1,000	No
7	25	Female	$25,000	7	12	$10,000	Yes
8	27	Male	$17,000	3	26	$15,000	Yes
9	28	Male	$25,000	4	39	$5,000	No
10	43	Female	$15,000	10	2	$1,000	Yes
11	42	Male	$25,000	10	4	$10,000	Yes
12	37	Female	$26,000	6	34	$15,000	Yes
13	33	Female	$29,000	9	11	$10,000	No
14	26	Female	$27,000	12	18	$15,000	Yes



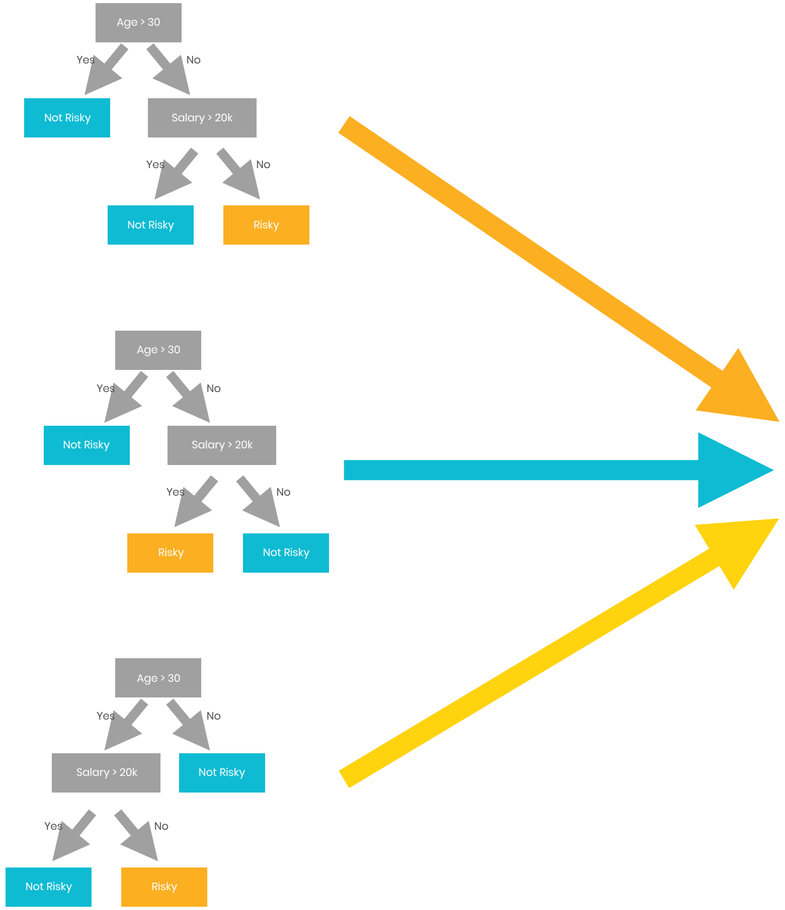
Prediction

GOAL: train copies of base model on samples of the data

Catalit LLC

Random Forest

Client	Age	Gender	Annual Salary	Months in residence	Months in job	Current Debt	Paid off credit
1	23	Male	$30,000	36	12	$5,000	Yes
2	30	Female	$45,000	12	12	$1,000	Yes
3	19	Male	$15,000	3	1	$10,000	No
4	25	Male	$25,000	12	2	$15,000	Yes
5	32	Female	$45,000	12	3	$5,000	No
6	26	Male	$35,000	18	5	$1,000	No
7	25	Female	$25,000	7	12	$10,000	Yes
8	27	Male	$17,000	3	26	$15,000	Yes
9	28	Male	$25,000	4	39	$5,000	No
10	43	Female	$15,000	10	2	$1,000	Yes
11	42	Male	$25,000	10	4	$10,000	Yes
12	37	Female	$26,000	6	34	$15,000	Yes
13	33	Female	$29,000	9	11	$10,000	No
14	26	Female	$27,000	12	18	$15,000	Yes



Prediction

GOAL: train trees on random subsets of the features

Catalit LLC

Gradient Boosting (XGBoost)

- Train sequence of trees, each learning from the mistakes of the preceding ones

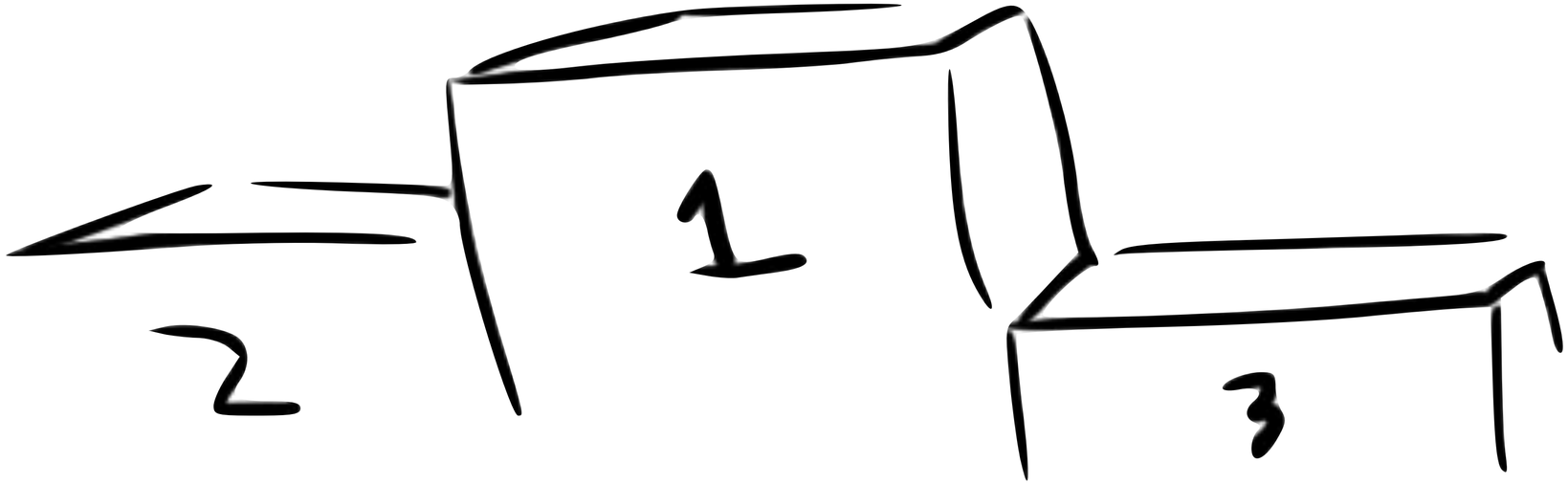
GOAL: add trees to fix hard-to-predict points

Catalit LLC

Model Evaluation

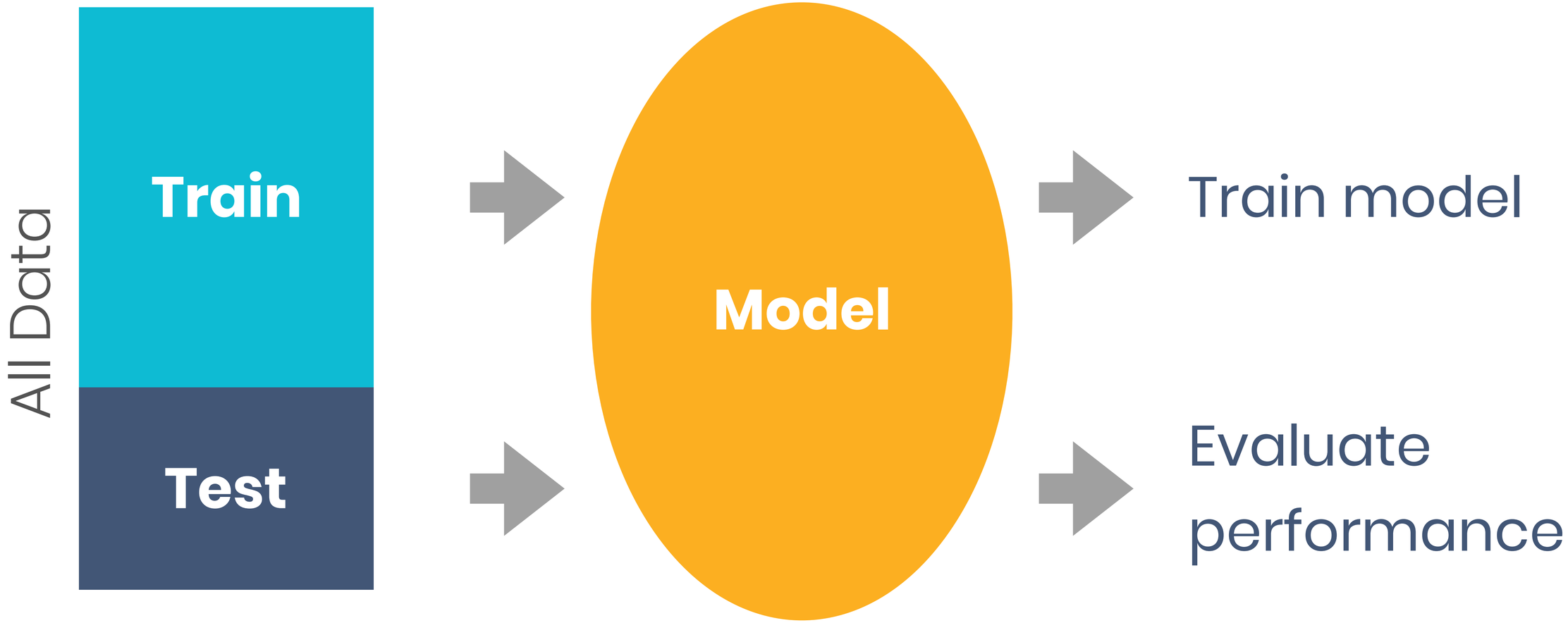
Baseline

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



Catalit LLC

Train / Test split



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

Catalit LLC

Confusion Matrix

	Pred 0	Pred 1
Label 0	True Negative (TN)	False Positive (FP / Type I err)
Label 1	False Negative (FN / Type II err)	True Positive (TP)

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

Catalit LLC

Classification Scores

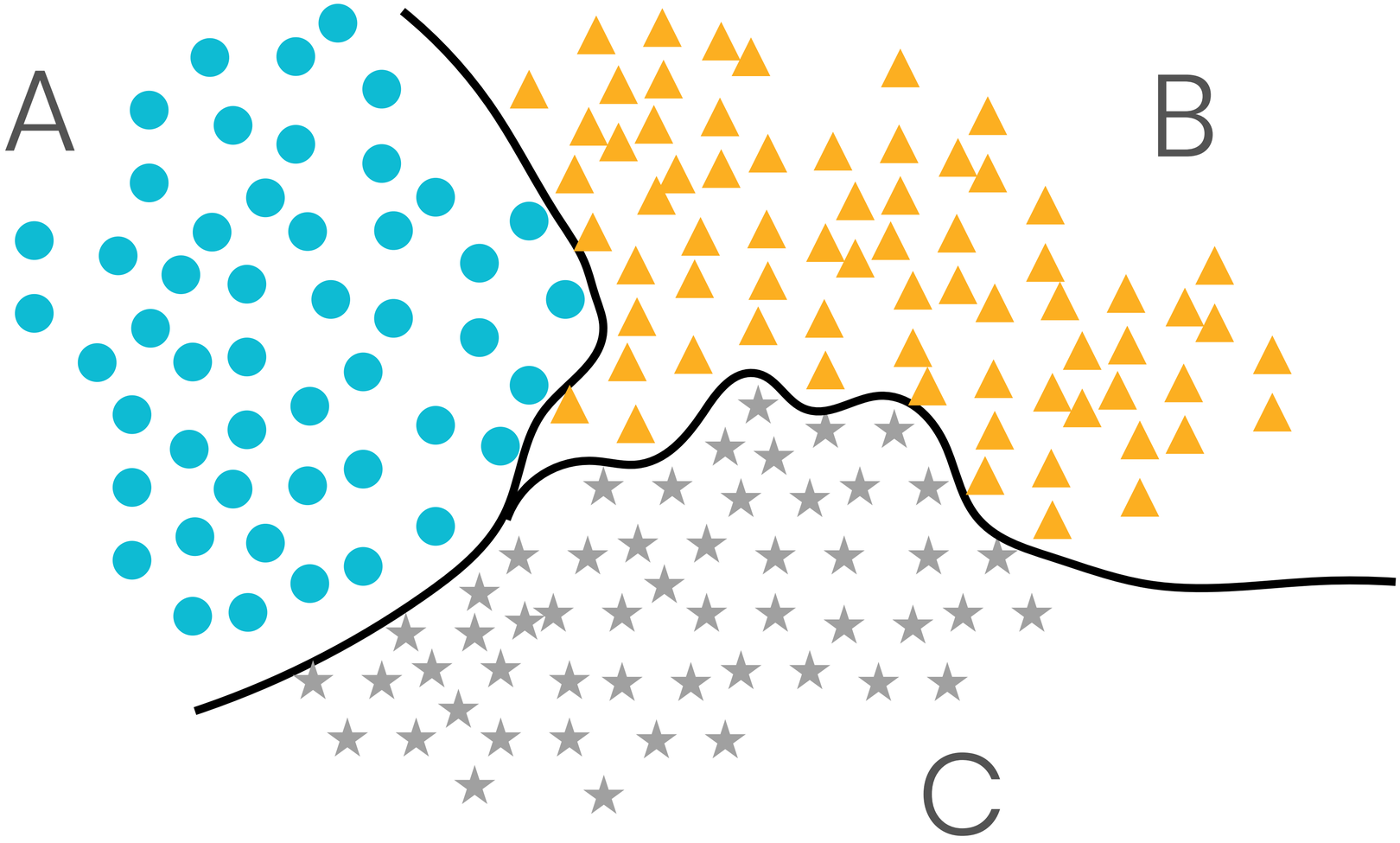
	Pred 0	Pred 1
Label 0	True Negative (TN)	False Positive (FP / Type I err)
Label 1	False Negative (FN / Type II err)	True Positive (TP)

- **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

Catalit LLC

Multi-class

Multi-class



Label
A
A
B
C

Catalit LLC

Multi-class

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

Catalit LLC

Working with text data

Working with text data

This is a regular message. Hello, how are you doing? → 0

Hello, win a free vacation to Hawaii, txt win to +1-800-SPAM → 1

Buy pills to live healthy and enriching life. Free samples ... → 1

Hello darling, did you remember to buy the cake? → 0

Catalit LLC

Keyword Features

	Hello	Buy	Win	Free	...	
This is a regular message. Hello, how are you doing?	1	0	0	0	...	0
Hello, win a free vacation to Hawaii, txt win to +1-800-SPAM	1	0	2	1	...	1
Buy pills to live healthy and enriching life. Free samples ...	0	1	0	1	...	1
Hello darling, did you remember to buy the cake?	1	1	0	1	...	0

Catalit LLC

Keyword Features

- User defined vocabularies, e.g. :
  - SPAM => free, buy, call, prize, ...
  - HAM => not, bye, and, see, ...

CountVectorizer

- Vocabulary is learnt from the text
- Count occurrences of each word
- 1000s of features
- Good for messages of fixed length (tweets, sms, error logs, ...)

TfidfVectorizer

- Term Frequency => # word / len(doc)
- Doc Frequency => # doc / len(corpus)
- Tfidf = Tf/Df

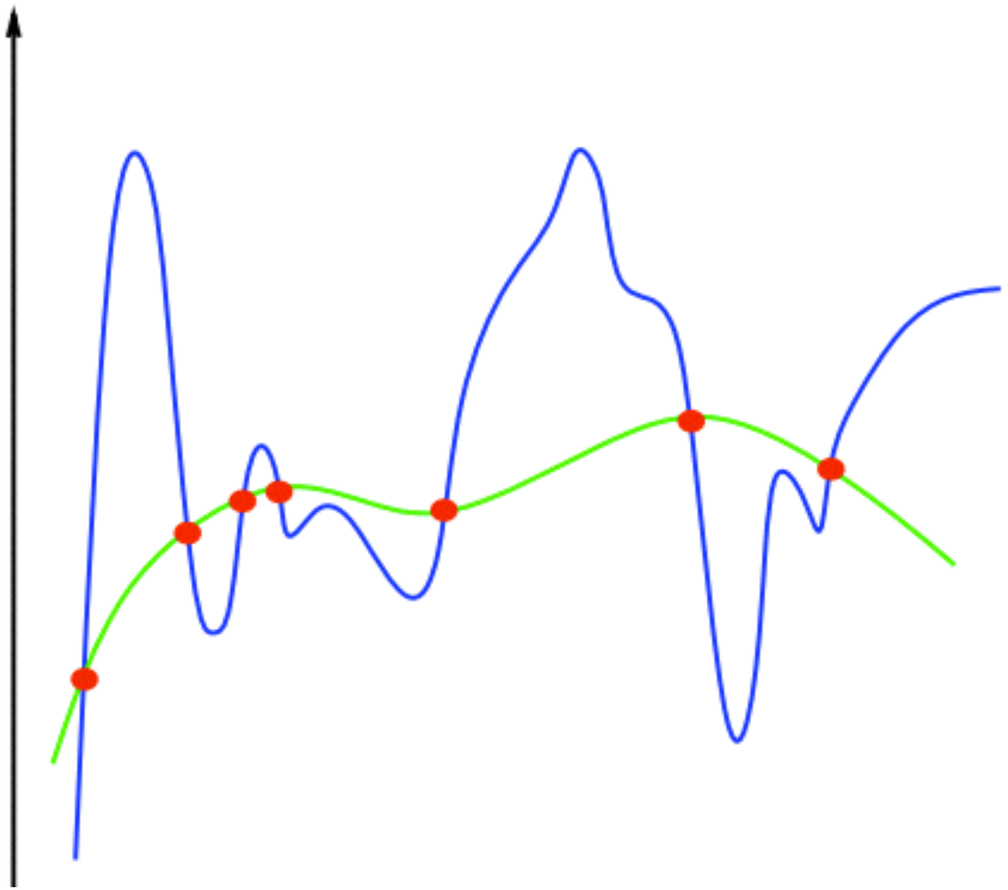
Recommended tf–idf weighting schemes

weighting scheme	document term weight	query term weight
1	$f_{t,d} \cdot \log \frac{N}{n_t}$	$\left(0.5 + 0.5 \frac{f_{t,q}}{\max_t f_{t,q}}\right) \cdot \log \frac{N}{n_t}$
2	$1 + \log f_{t,d}$	$\log\left(1 + \frac{N}{n_t}\right)$
3	$(1 + \log f_{t,d}) \cdot \log \frac{N}{n_t}$	$(1 + \log f_{t,q}) \cdot \log \frac{N}{n_t}$

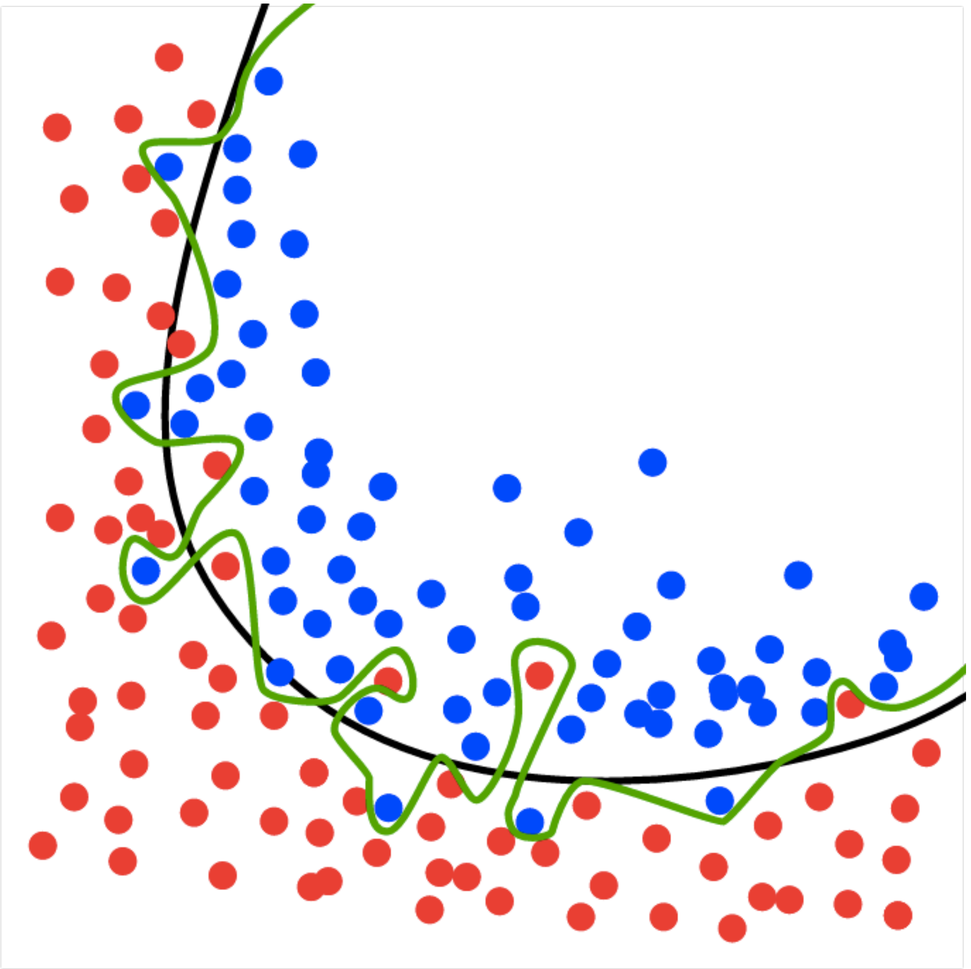
Hello	Buy	Win	Free	...
0.11	0	0	0	...
0.21	0	0.7	0.89	...
0	0.37	0	0.54	...
0.13	0.43	0	0.23	...

Improving performance

Overfitting



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



https://upload.wikimedia.org/wikipedia/commons/1/19/Overfitting.svg

Catalit LLC

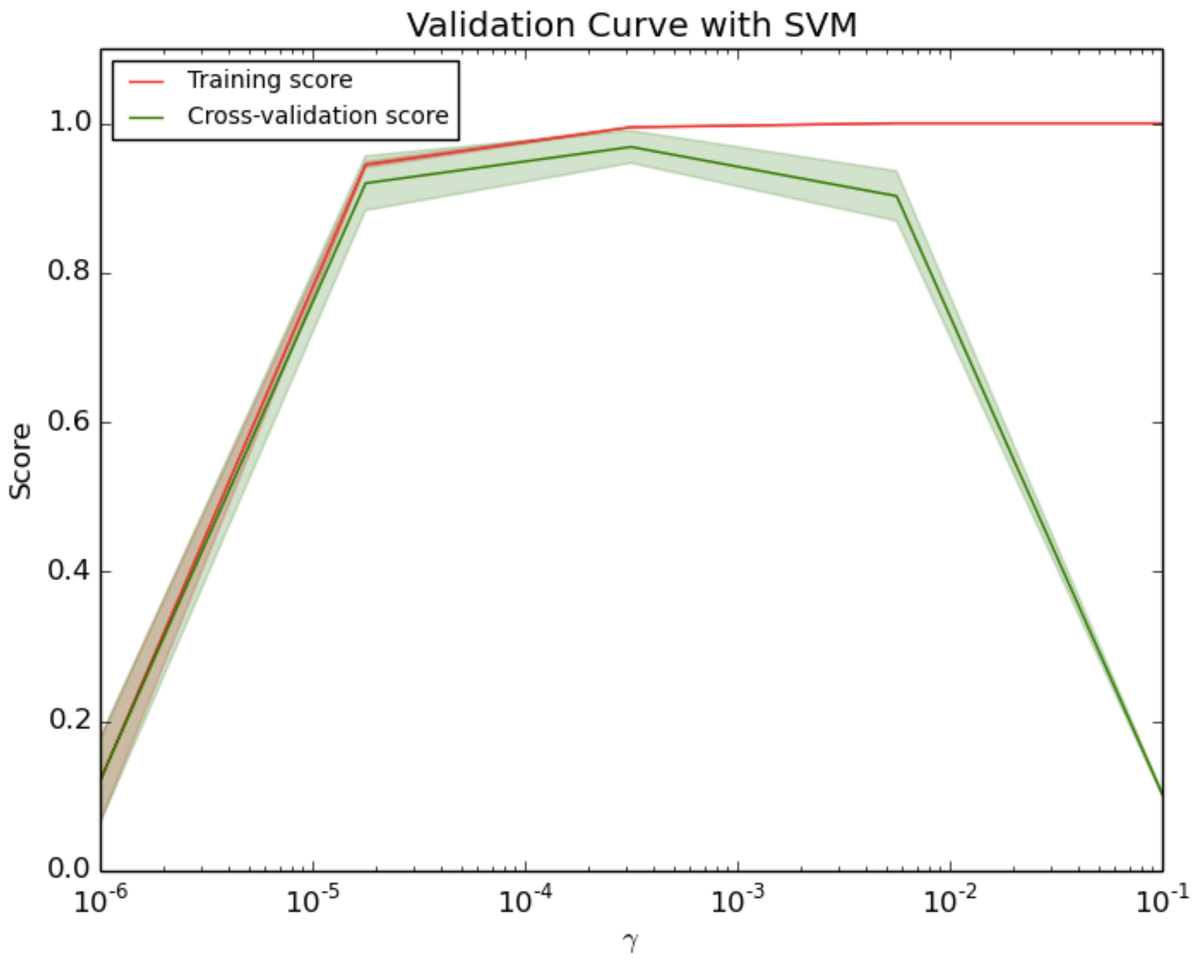
Cross Validation

Round 1	Round 2	Round 3	Round 4	Round 5
Test	Train	Train	Train	Train
Train	Test	Train	Train	Train
Train	Train	Test	Train	Train
Train	Train	Train	Test	Train
Train	Train	Train	Train	Test
93%	89%	91%	92%	88%

Accuracy = Average(Rounds)

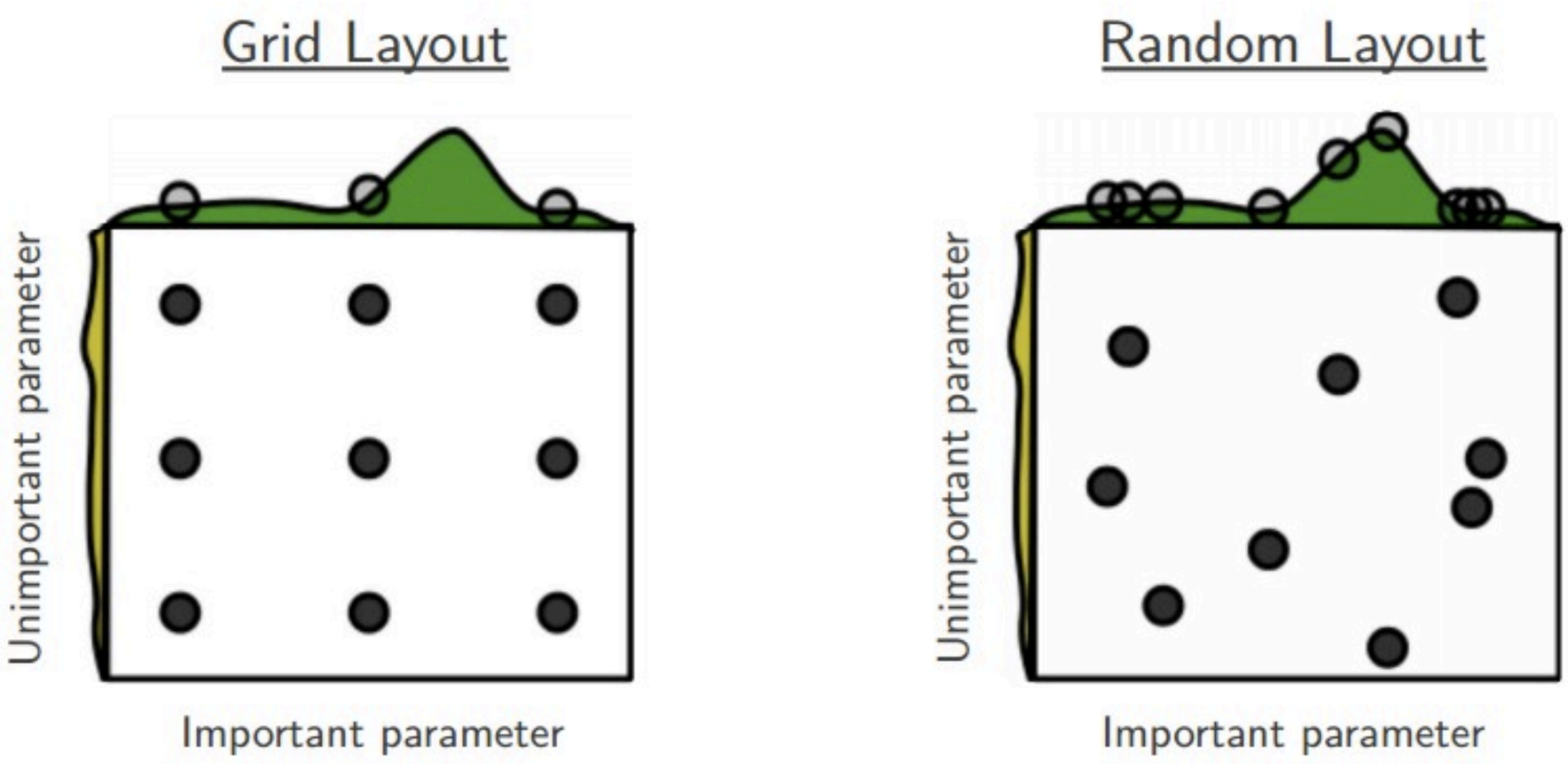
Catalit LLC

Validation curves



Catalit LLC

Random VS GRID

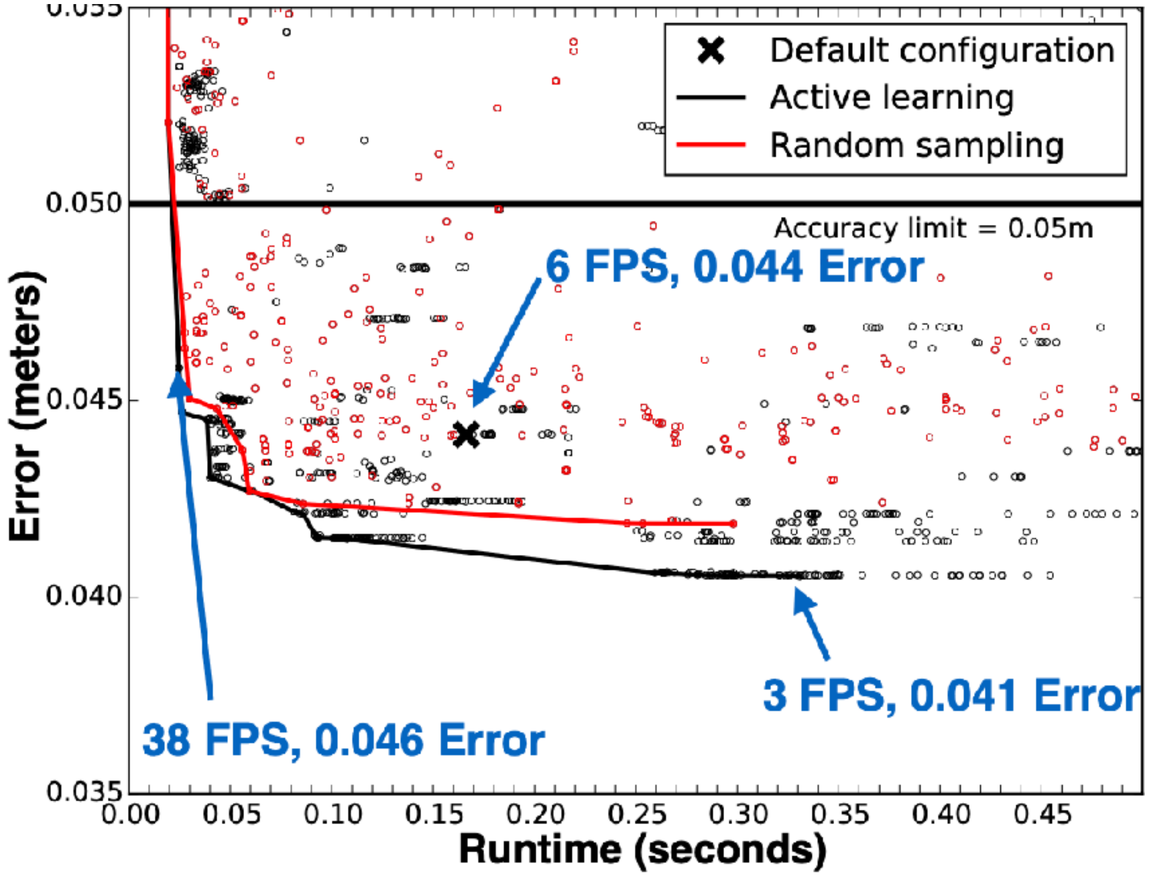


http://www.jmlr.org/papers/volume13/bergstra12a/bergstra12a.pdf

Catalit LLC

Bayesian Optimization

- Intelligent search of hyper-parameters
- Global optimization of **black-box functions**
- No need for derivatives
- Demonstrated on:
  - Hardware design
  - DBMS
  - Computer Vision
  - Robotics
  - AutoML



Ask me about it

Catalit LLC

Lab 04

- **Jupyter**
  - interactive Python web application => our dev env
- **Pandas**
  - like Excel for python => data manipulation
- **Matplotlib & Seaborn**
  - Plots & charts => data visualization
- **Scikit-Learn**
  - Machine Learning => model building, training & evaluation

Unsupervised Learning

Unsupervised Learning

- Clustering Intuition
- Distance & Similarity
- Clustering algorithms
- Model evaluation
- Dimensionality reduction
- Topic modeling

Catalit LLC

Summary of ML

	CATEGORICAL	CONTINUOUS
SUPERVISED	?	?
UNSUPERVISED	?	?

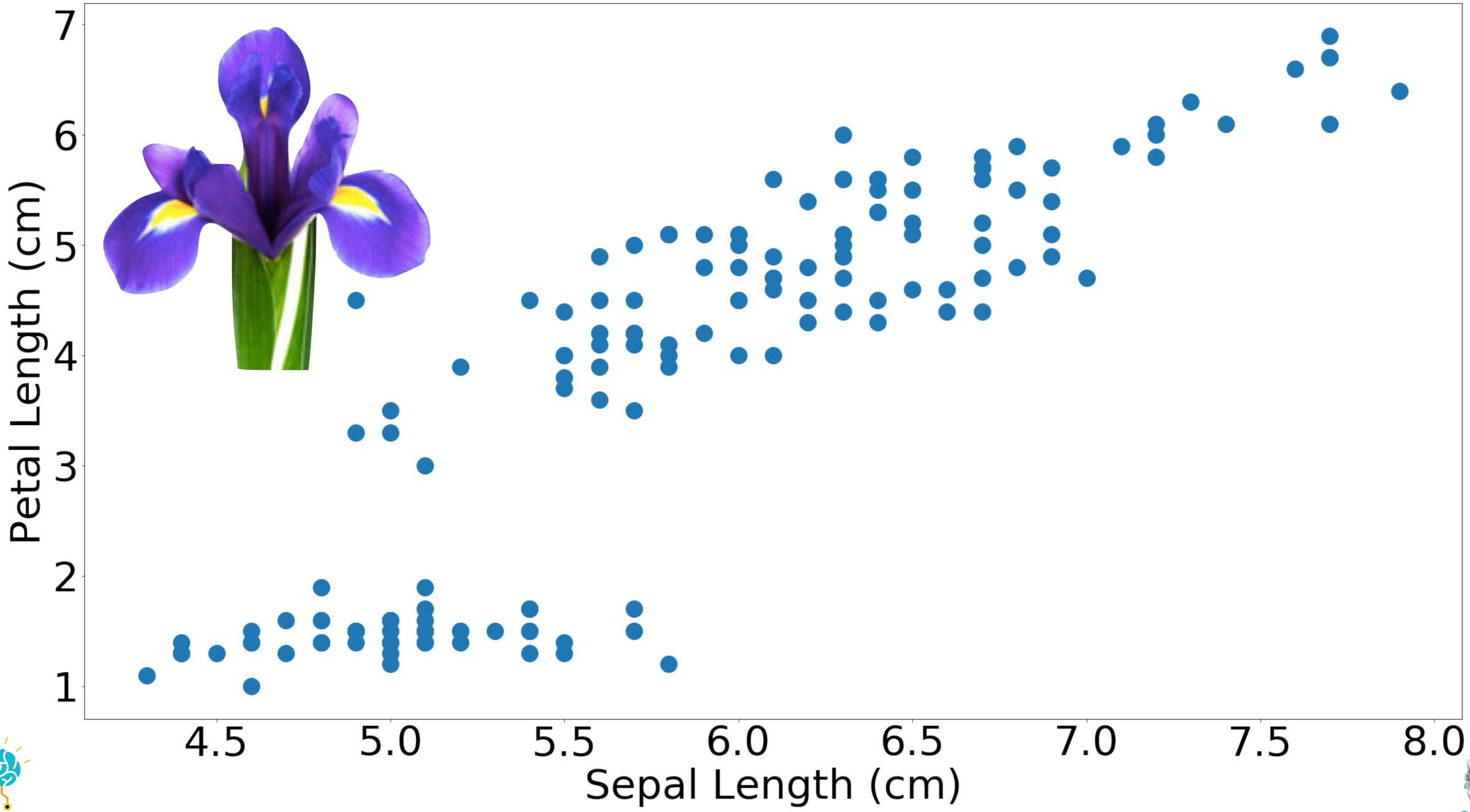
Catalit LLC

Summary of ML

	CATEGORICAL	CONTINUOUS
SUPERVISED	CLASSIFICATION	REGRESSION
UNSUPERVISED	CLUSTERING	DIM REDUCTION

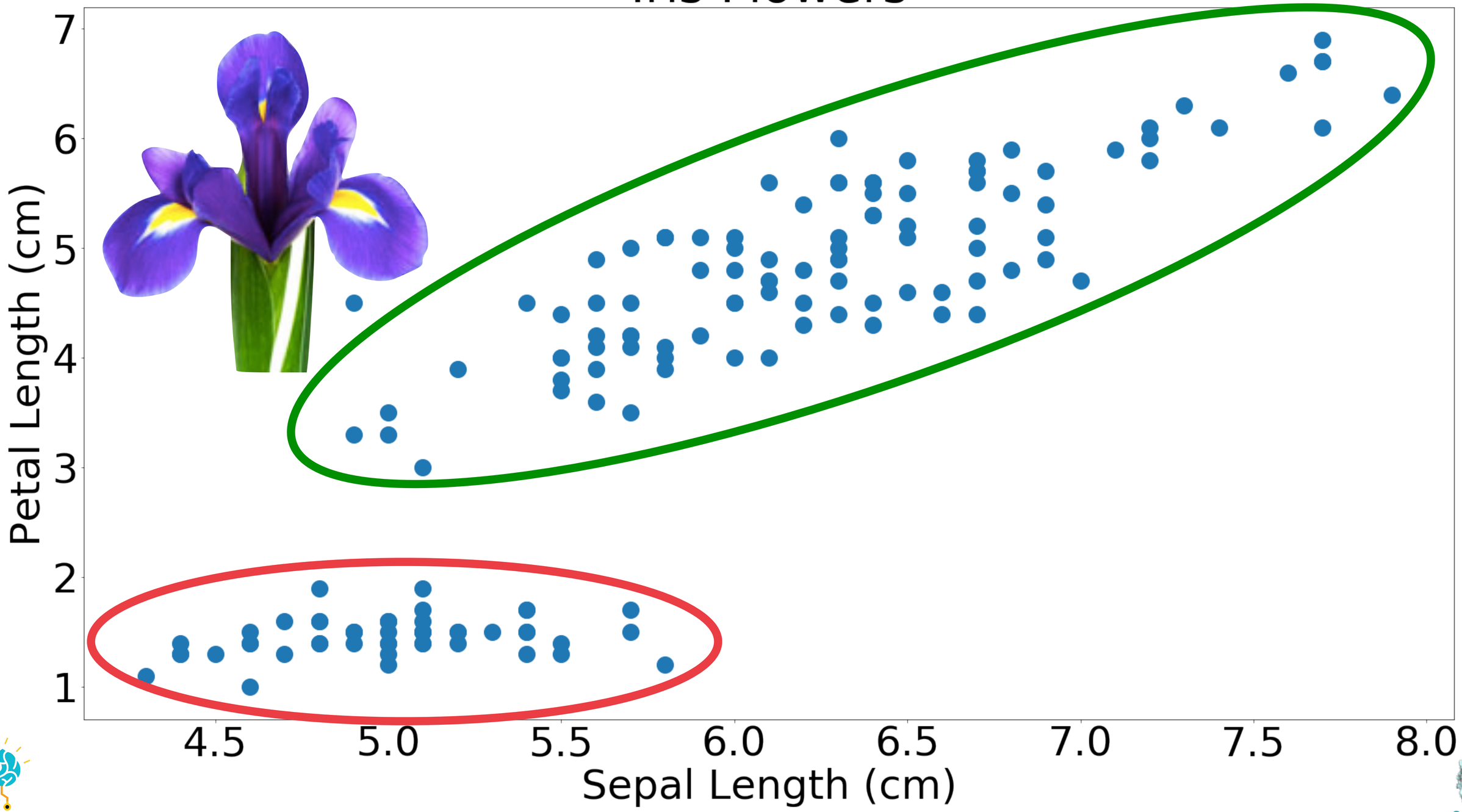
Catalit LLC

Iris Flowers



Catalit LLC

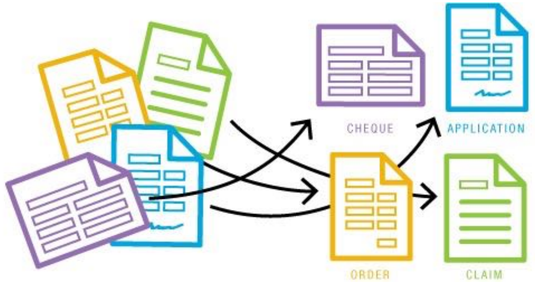
Iris Flowers

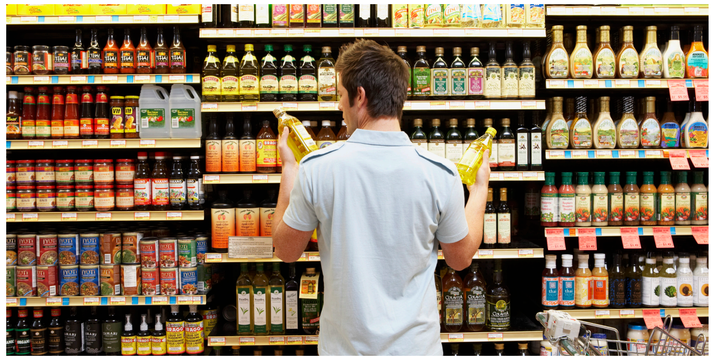


Catalit LLC

Applications

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





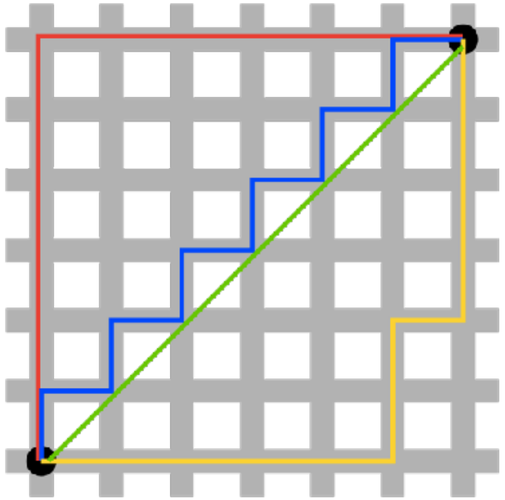
Catalit LLC

Distance & Similarity

Catalit LLC

Distance & Similarity

	Age	Gender	Annual Salary	Months in residence	Months in job	Current Debt
Client 1	23	Male	$30,000	36	12	$5,000
Client 2	30	Female	$45,000	12	12	$1,000
Client 3	19	Male	$15,000	3	1	$10,000
Client 4	25	Male	$25,000	12	27	$15,000



Catalit LLC

Distance & Similarity

- **Euclidean distance:** vector distance

$$d(\overline{v}, \overline{w}) = \sqrt{(\overline{v} - \overline{w})^2}$$

- **Cosine similarity:** normalized dot product

$$d(\overline{v}, \overline{w}) = \frac{\overline{v}.\overline{w}}{\|\overline{v}\| \, \|\overline{w}\|}$$

- **Jaccard score:** size(intersection) / size(union)

$$d(V, W) = \frac{|V \cap W|}{|V \cup W|}$$

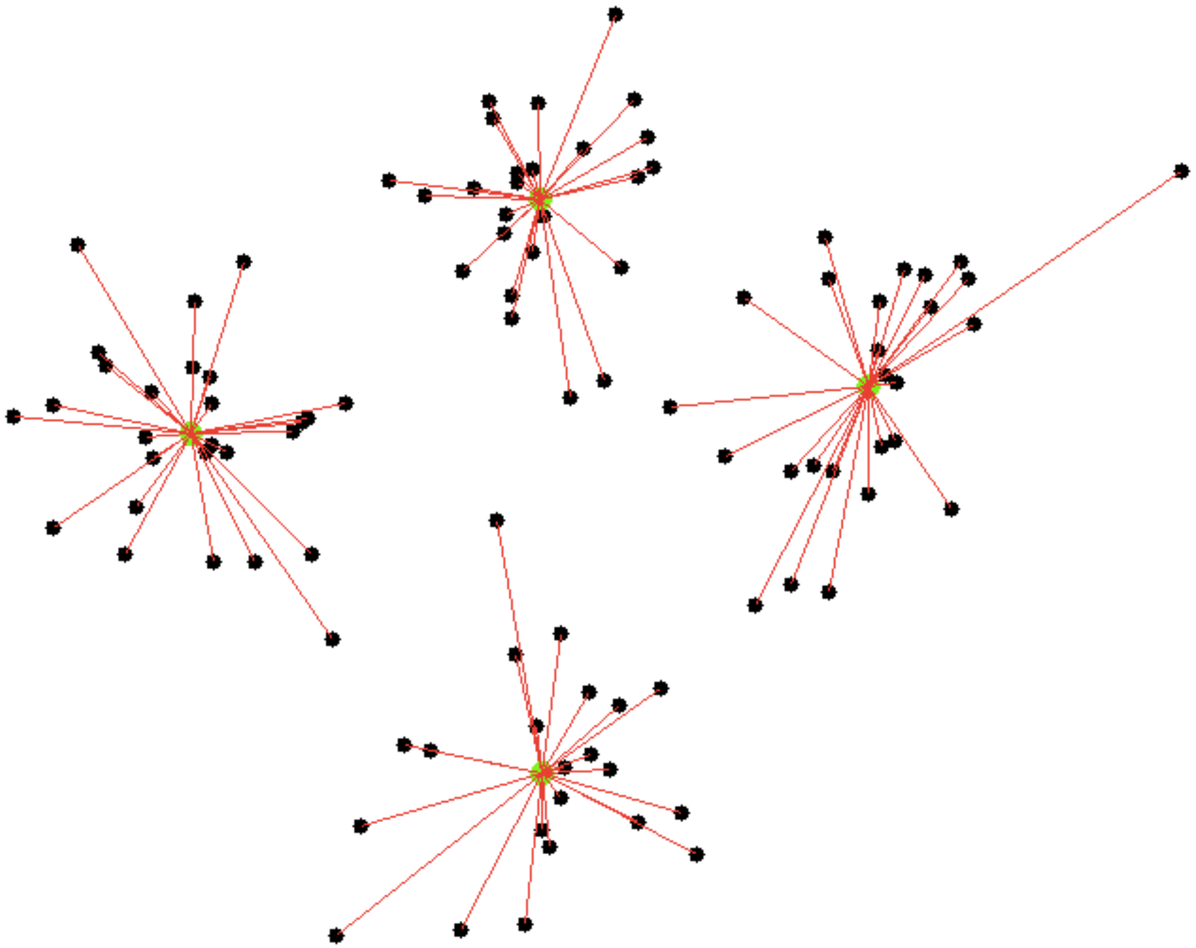
- **Levenshtein distance:** string edits

$$\mathrm{lev}_{a,b}(i,j) = \begin{cases} \max(i,j) & \text{if } \min(i,j) = 0, \\ \min \begin{cases} \mathrm{lev}_{a,b}(i-1,j) + 1 \\ \mathrm{lev}_{a,b}(i,j-1) + 1 \\ \mathrm{lev}_{a,b}(i-1,j-1) + 1_{(a_i \neq b_j)} \end{cases} & \text{otherwise.} \end{cases}$$

Catalit LLC

Clustering algorithms

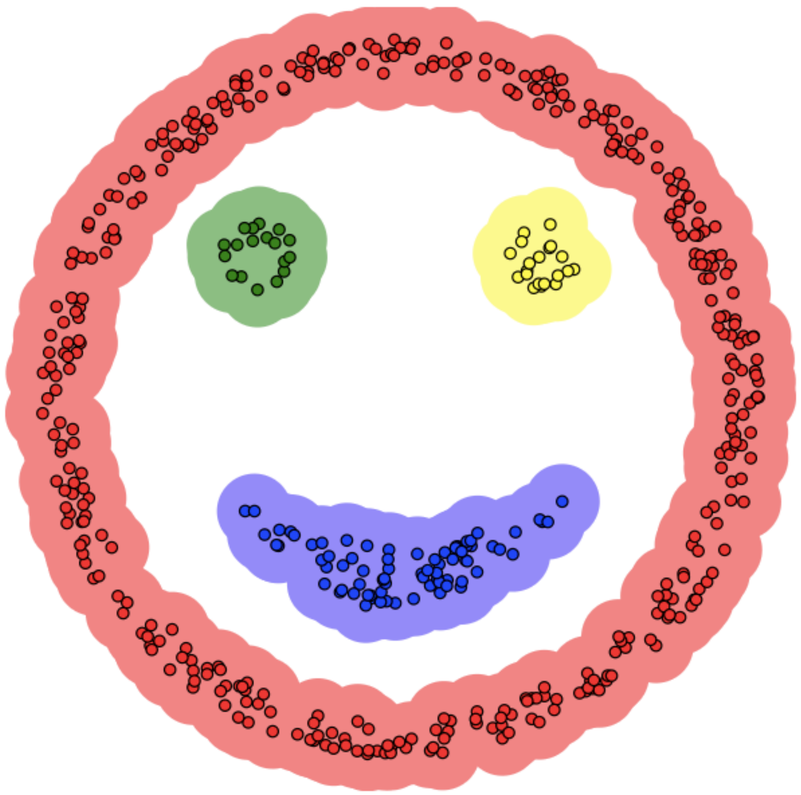
K-Means



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

Catalit LLC

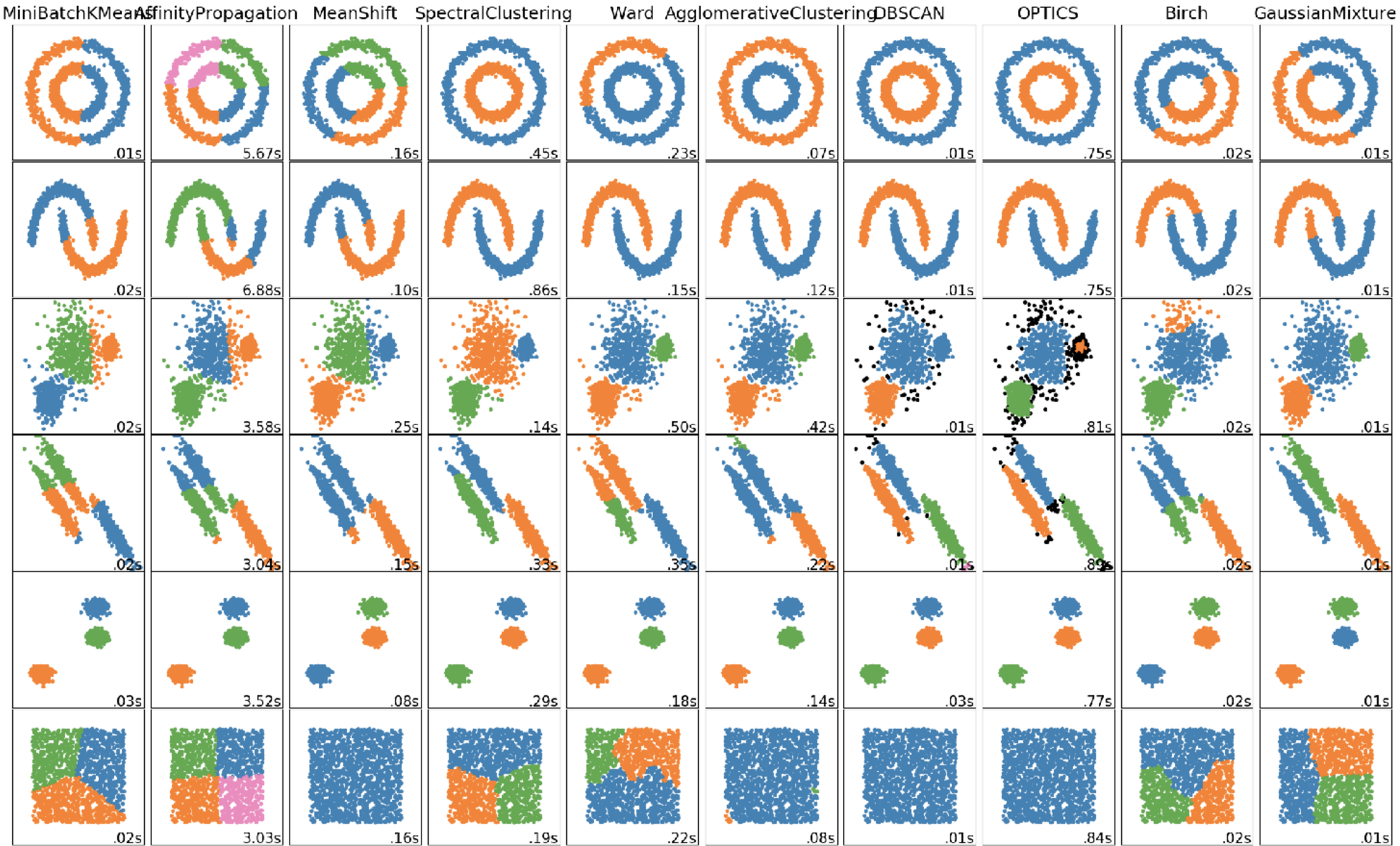
DBSCAN



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

Catalit LLC

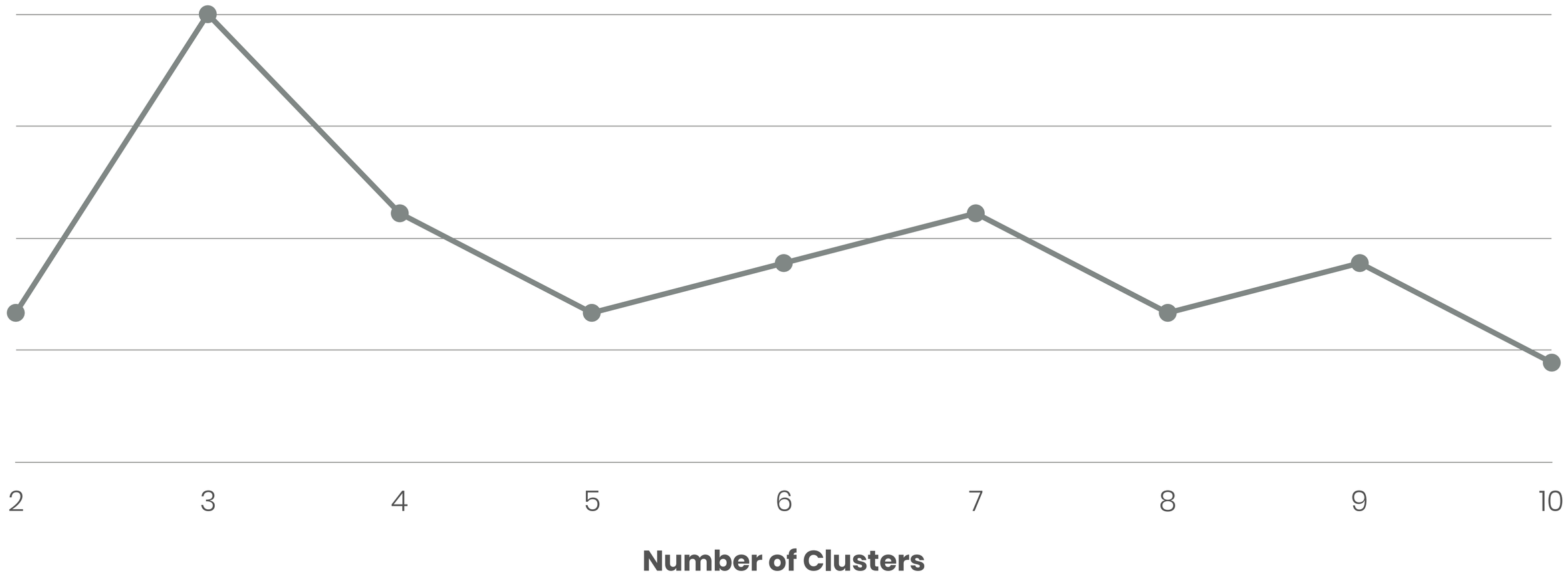
Clustering



Catalit LLC

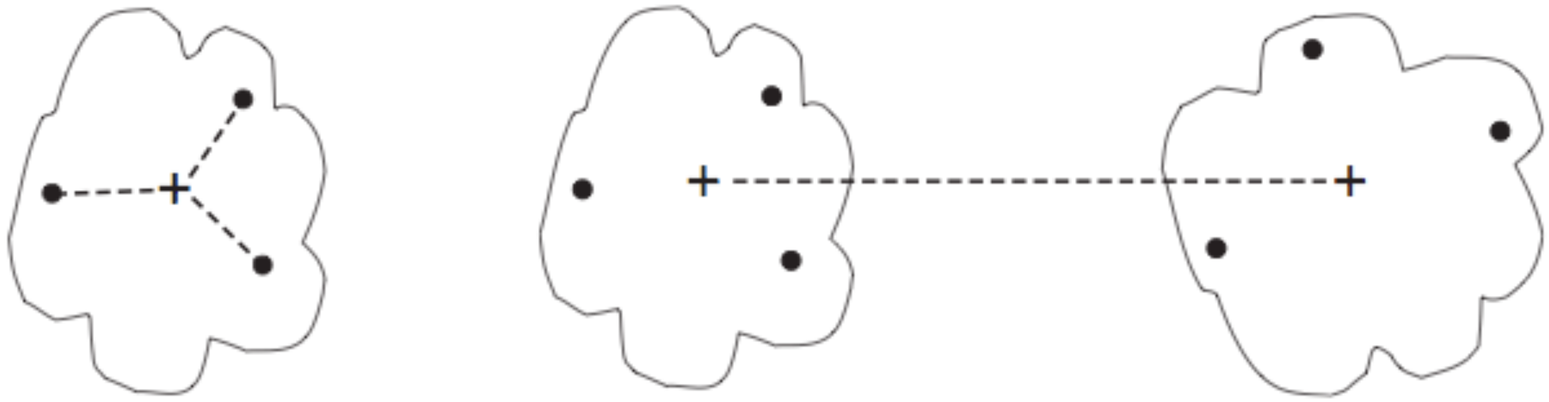
Model Evaluation

Silhouette score



Catalit LLC

Model Evaluation



(a) Cohesion. (b) Separation.

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

Catalit LLC

Dimensionality Reduction

Summary of ML

	CATEGORICAL	CONTINUOUS
SUPERVISED	?	?
UNSUPERVISED	?	?

Catalit LLC

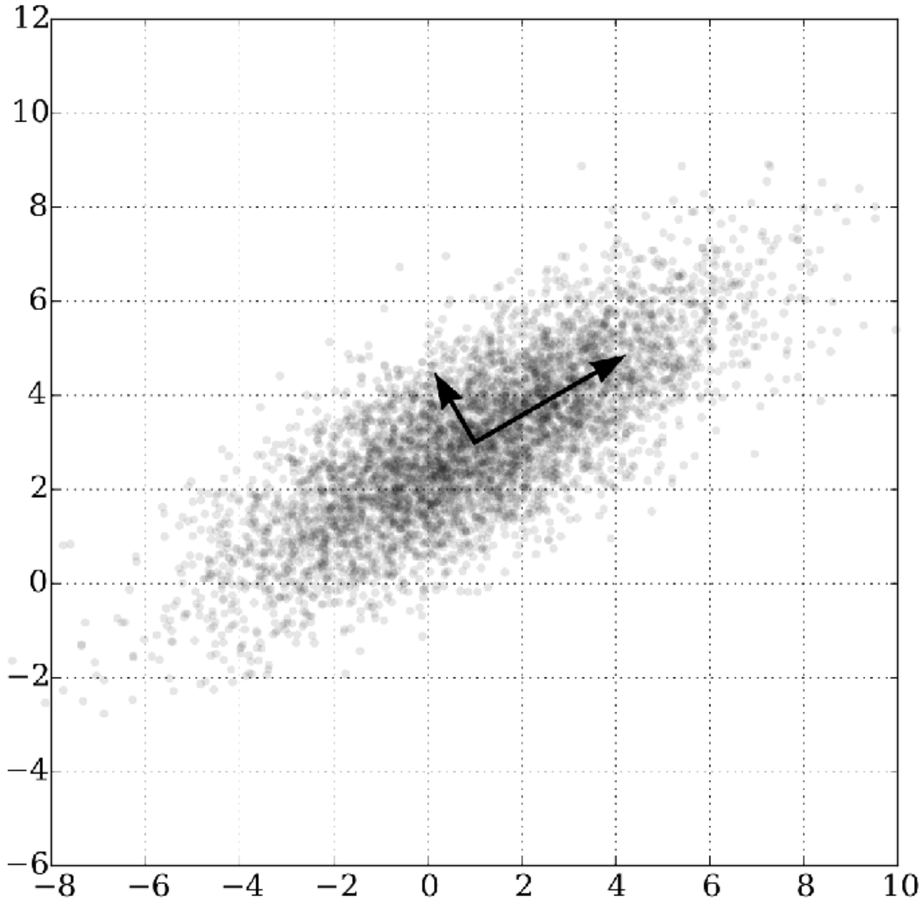
Summary of ML

	CATEGORICAL	CONTINUOUS
SUPERVISED	CLASSIFICATION	REGRESSION
UNSUPERVISED	CLUSTERING	DIM REDUCTION

Catalit LLC

Principal Components (PCA)

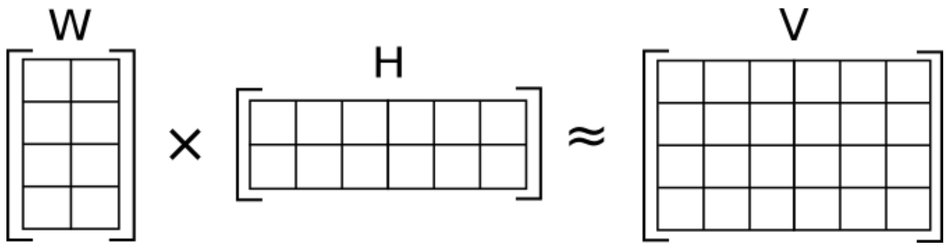
- Transform correlated variables to linearly uncorrelated variables
- First PC has largest variance



Catalit LLC

Non-negative Matrix Factorization (NMF)

- matrix V factorized into two matrices W and H
- All three matrices have no negative elements
- Used on term-frequency matrices



Catalit LLC

Lab 05

- **Jupyter**
  - interactive Python web application => our dev env
- **Pandas**
  - like Excel for python => data manipulation
- **Matplotlib & Seaborn**
  - Plots & charts => data visualization
- **Scikit-Learn**
  - Machine Learning => model building, training & evaluation

References - Docs

- Pandas: https://pandas.pydata.org/pandas-docs/stable/
- Matplotlib: https://matplotlib.org/3.1.1/contents.html
- Seaborn: https://seaborn.pydata.org/
- Scikit-Learn: https://scikit-learn.org/stable/documentation.html

References - Books

- Hands-on Machine Learning:
  - https://www.oreilly.com/library/view/hands-on-machine-learning/9781492032632/
  - https://github.com/ageron/handson-ml2
- Zero to Deep Learning:
  - https://book.zerotodeeplearning.com
  - https://github.com/zerotodeeplearning/ztdl-5-day-bootcamp
- An Introduction to Statistical Learning:
  - http://faculty.marshall.usc.edu/gareth-james/ISL/getbook.html
- Machine Learning: a Probabilistic Perspective
  - https://www.cs.ubc.ca/~murphyk/MLbook/
- Awesome Machine Learning:
  - https://github.com/josephmisiti/awesome-machine-learning/blob/master/books.md

References - Repos

- https://github.com/josephmisiti/awesome-machine-learning
- https://github.com/ujjwalkarn/Machine-Learning-Tutorials
- https://github.com/EthicalML/awesome-production-machine-learning