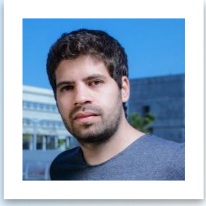
Introduction to Machine Learning

Lior Sidi & Noa Lubin

Meet us



Lior Sidi

- Machine Learning Team Lead at Wix
- NLP, RecSys, User Centric AI.

- Ex: Startup, Consultat, researcher
- BSc and MSc from BGU SISE.
- Scholar / Linkedin

Noa Lubin

- Machine Learning Team Lead at Diagnostic Robotics
- NLP, health, space

- Ex: NASA, Amazon, Elbit, IAI
- BSC EE Technion
- MSc CS NLP Bar Ilan
- Linkedin

Y-DATA
SCHOOL OF DATA SCIENCE

Yandex

What is ML?

Y-DATA SCHOOL OF DATA SCIENCE

Yandex

Traditional programing

Input + Logic = Output

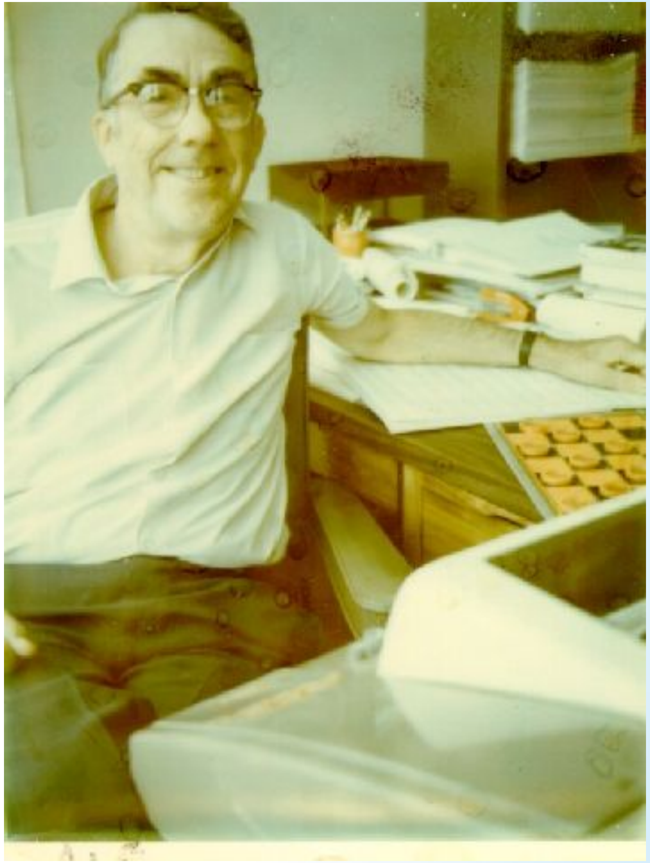
- Machines Follow Instruction
- Humans Learn From Experience

The Arthur Samuel's Checkers

"Machine Learning: Field of study that gives computers the ability to learn without being explicitly programmed."

— Arthur Samuel (1959)

Machine Learning is the ability to generalize from experience onto unseen example



Traditional programing

Input + Logic = Output

Machine Learning

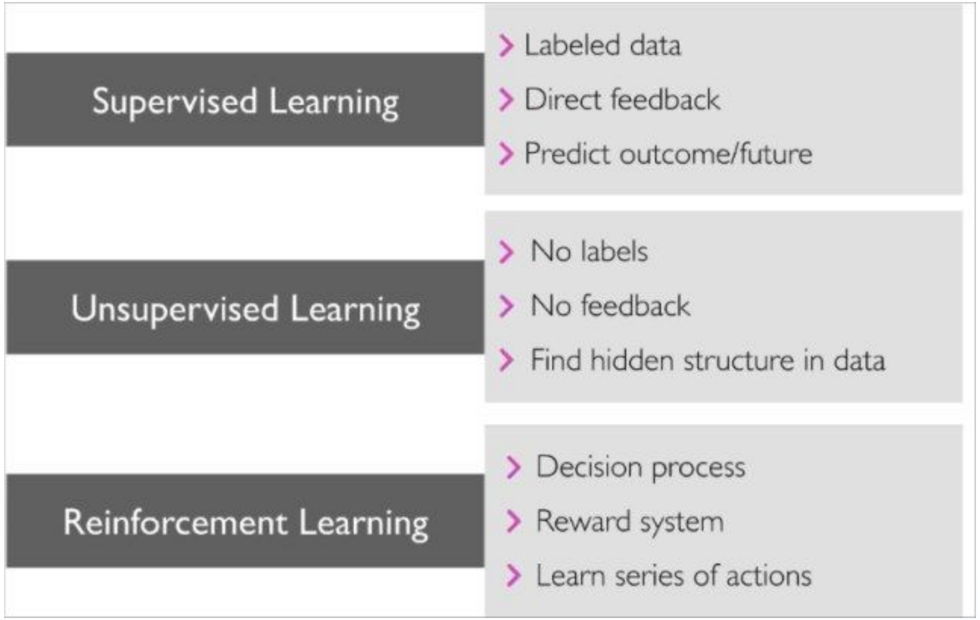
Input + Output = Logic

Learning Tasks

Y-DATA SCHOOL OF DATA SCIENCE

Yandex

Types of ML



[List of Machine Learning Algorithms] [3 different types of machine learning]

Unsupervised Learning

- No labels
- No feedback
- Find hidden structure in data

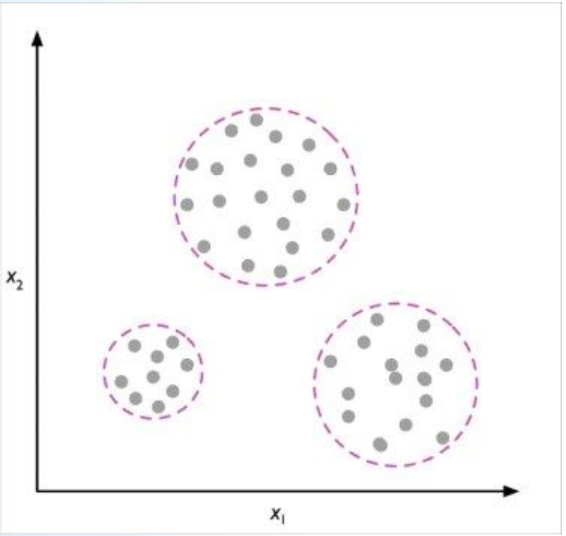
Y-DATA SCHOOL OF DATA SCIENCE

Yandex

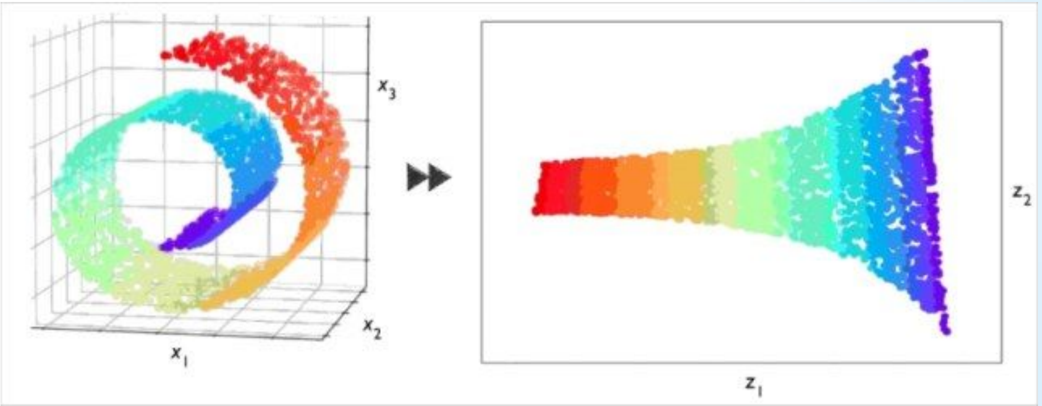
Unsupervised Learning

- No labels
- No feedback
- Find hidden structure in data

Clustering



Dimension Reduction

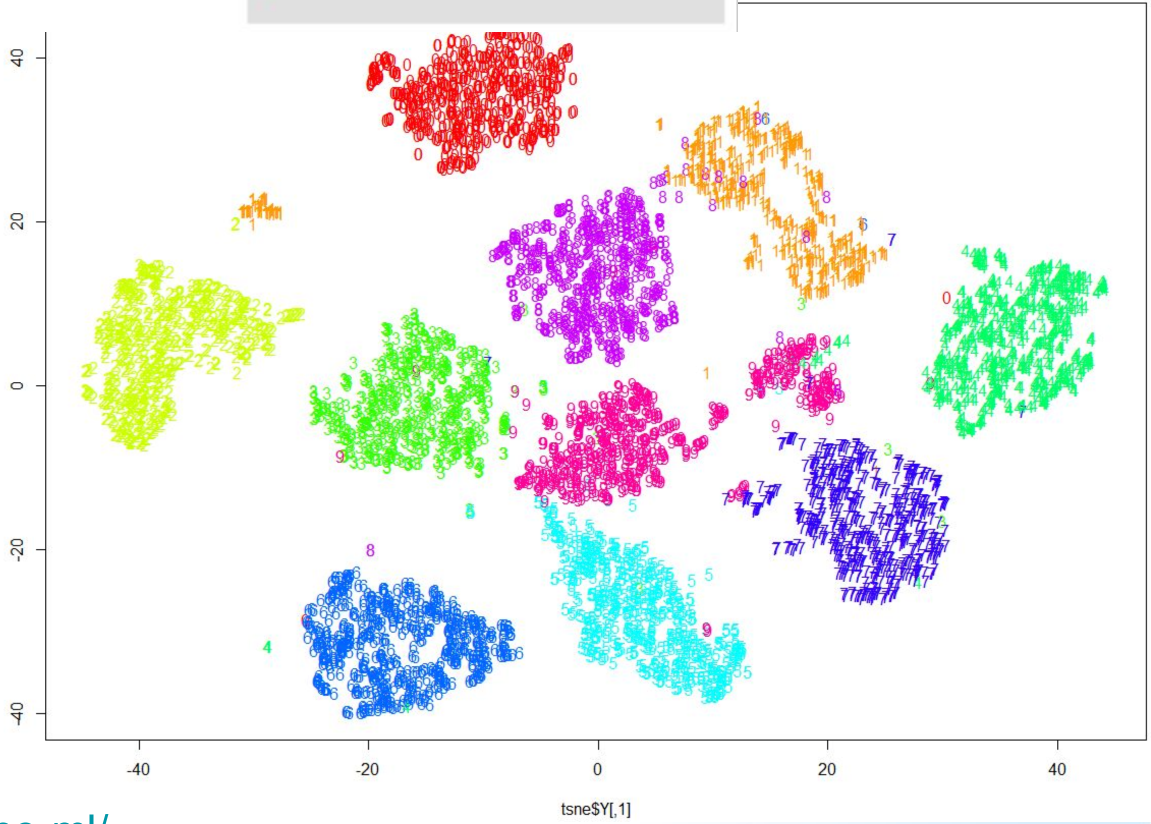


Y-DATA SCHOOL OF DATA SCIENCE

Yandex

Unsupervised Learning

- No labels
- No feedback
- Find hidden structure in data



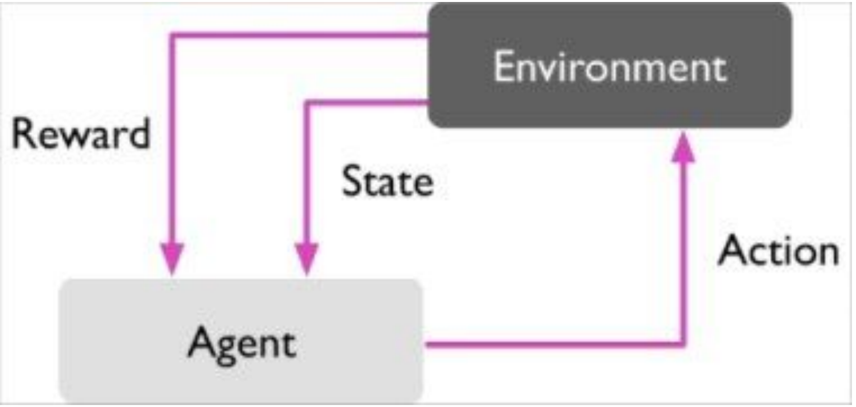
https://mark-borg.github.io/blog/2016/tsne-ml/

Y-DATA
SCHOOL OF DATA SCIENCE

Yandex

Reinforcement Learning

- Decision process
- Reward system
- Learn series of actions

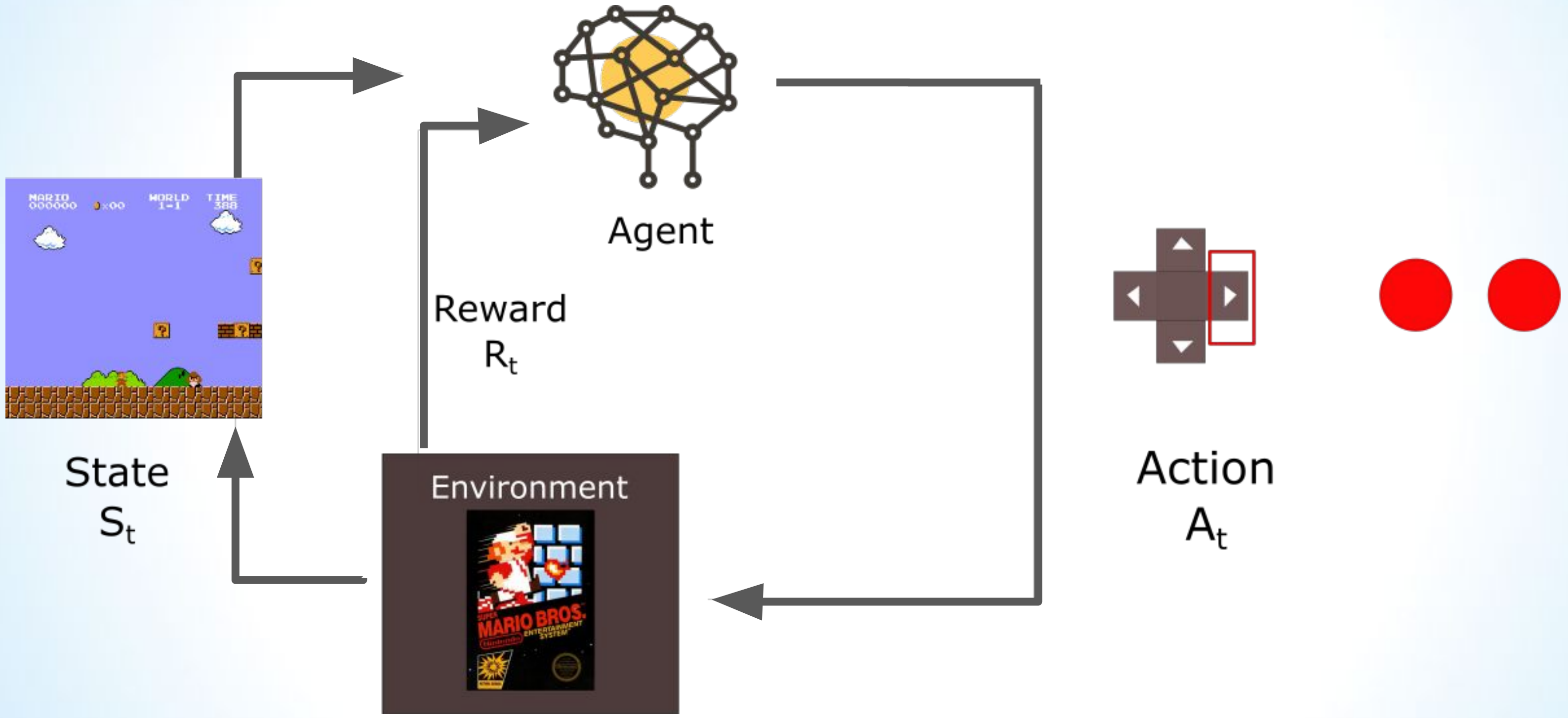


Y-DATA SCHOOL OF DATA SCIENCE

Yandex

Reinforcement Learning

- Decision process
- Reward system
- Learn series of actions



https://medium.freecodecamp.org/an-introduction-to-reinforcement-learning-4339519de419

Supervised Learning

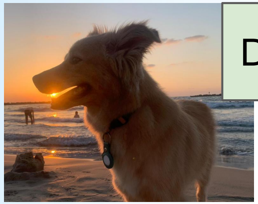
Y-DATA SCHOOL OF DATA SCIENCE

Yandex

Supervised Learning
- Labeled data
- Direct feedback
- Predict outcome/future

Binary Classification

Date + Label



Dog

Not a Dog

Step 1 - Training

Algorithm

Train/Fit a model

Step 2 - Prediction

Model

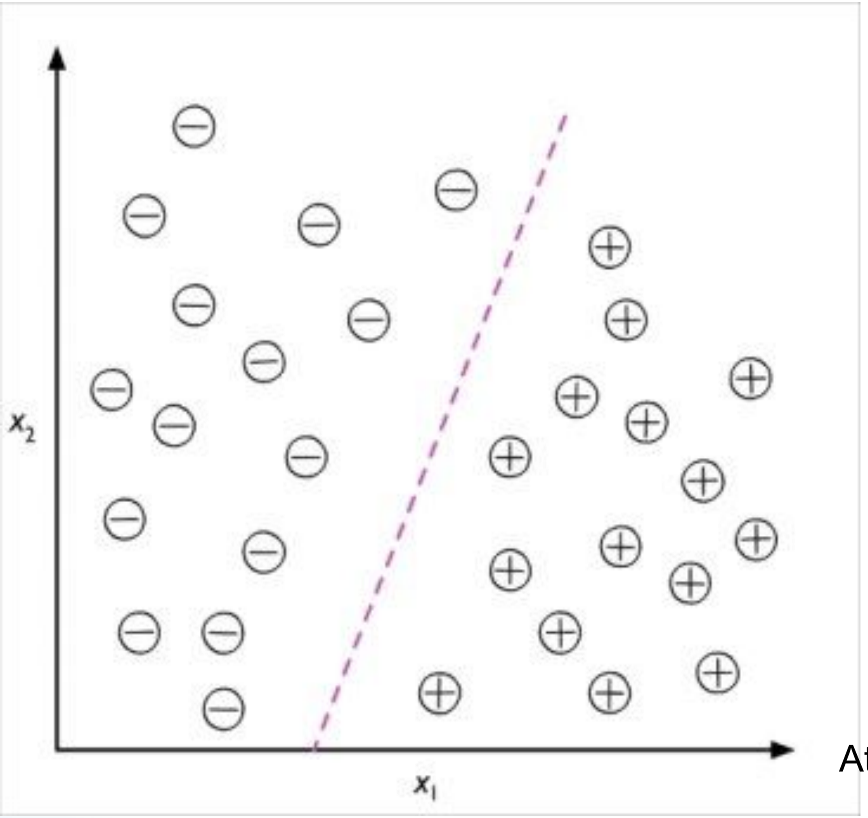
Predict a model

Dog

Unseen data

Binary Classification diagram

Attribute 2



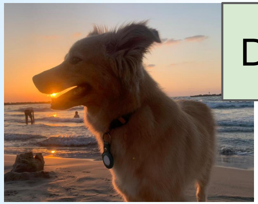
Attribute 1

Y-DATA
SCHOOL OF DATA SCIENCE

Yandex

Multi-Class Classification

Date + Label



Dog

Cat

Fish

Step 1 - Training

Algorithm

Train/Fit a model

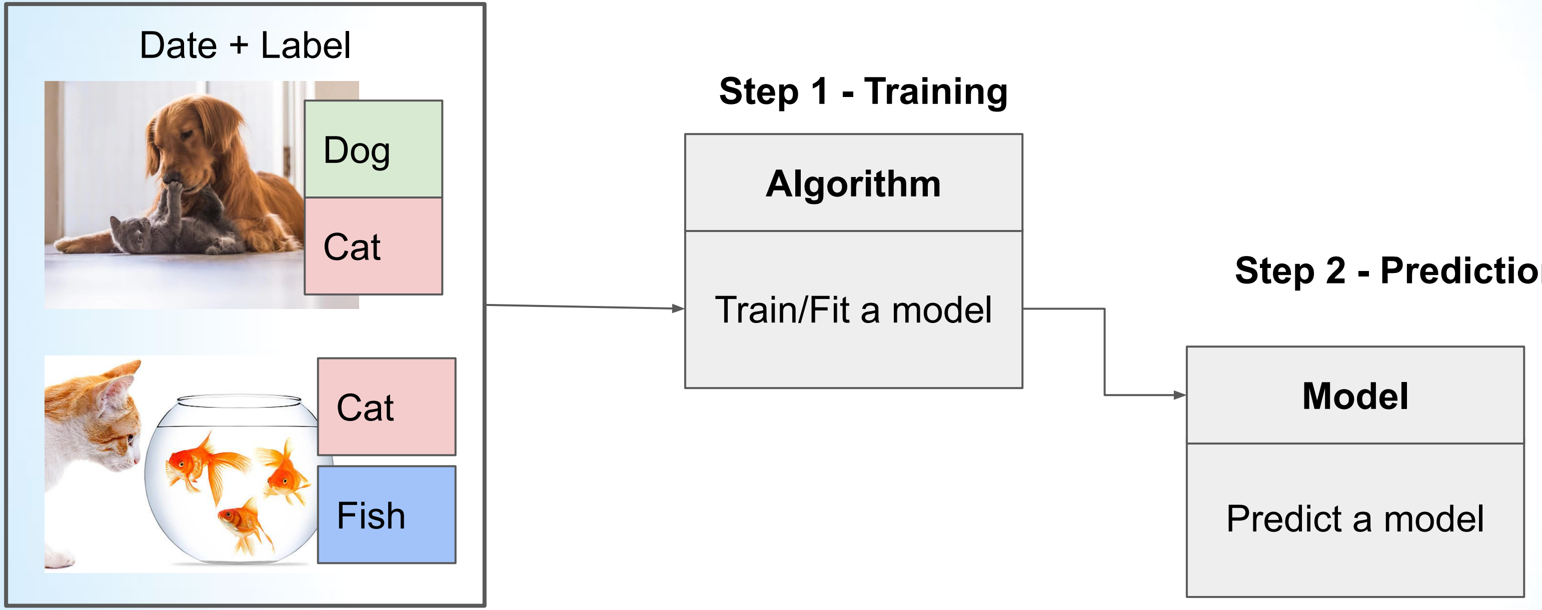
Step 2 - Prediction

Model

Predict a model

Y-DATA
SCHOOL OF DATA SCIENCE
Yandex

Multi-Label Classification



Y-DATA SCHOOL OF DATA SCIENCE

Yandex

Supervised Learning

- Labeled data
- Direct feedback
- Predict outcome/future

Regression

Date + Label

15 kg

3 kg

Step 1 - Training

Algorithm
Train/Fit a model

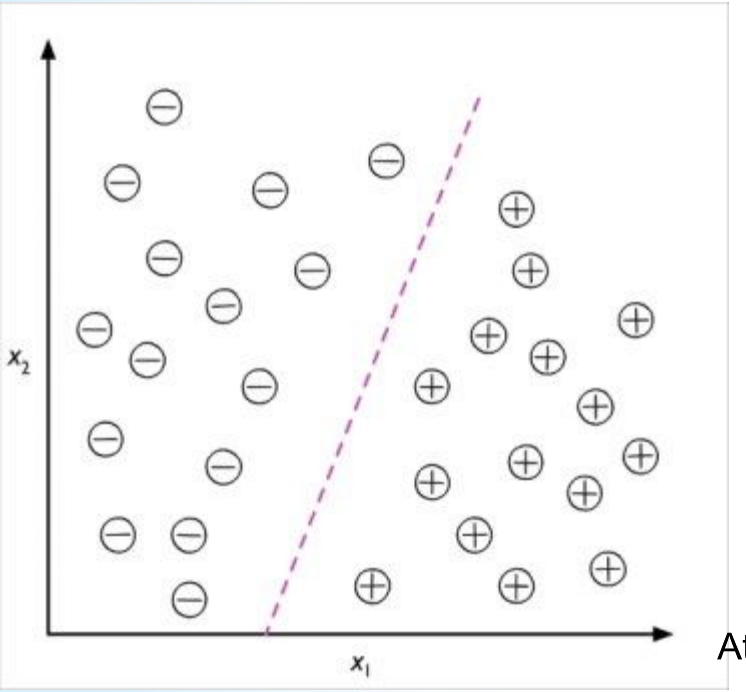
Step 2 - Prediction

Model
Predict a model

Regression Vs Classification

Classification

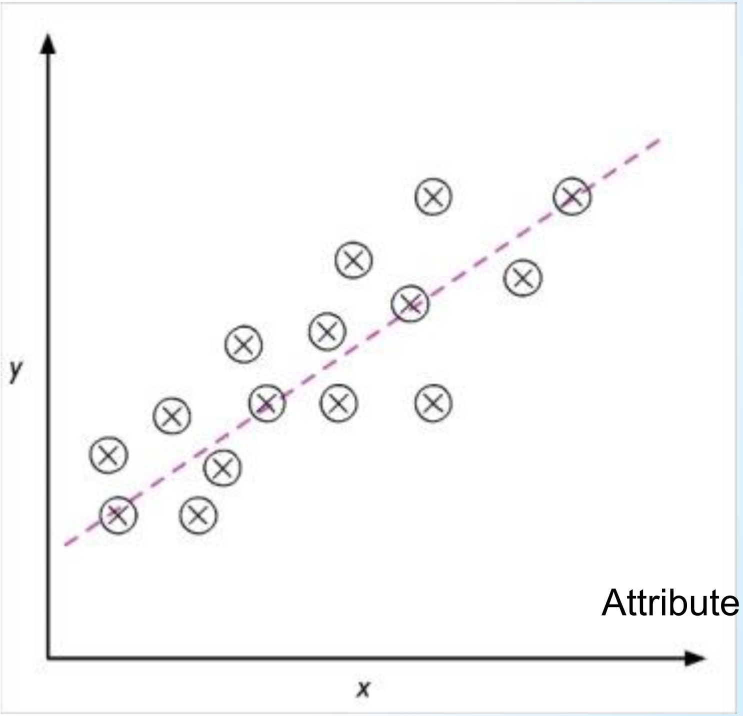
Attribute 2



Attribute 1

Regression

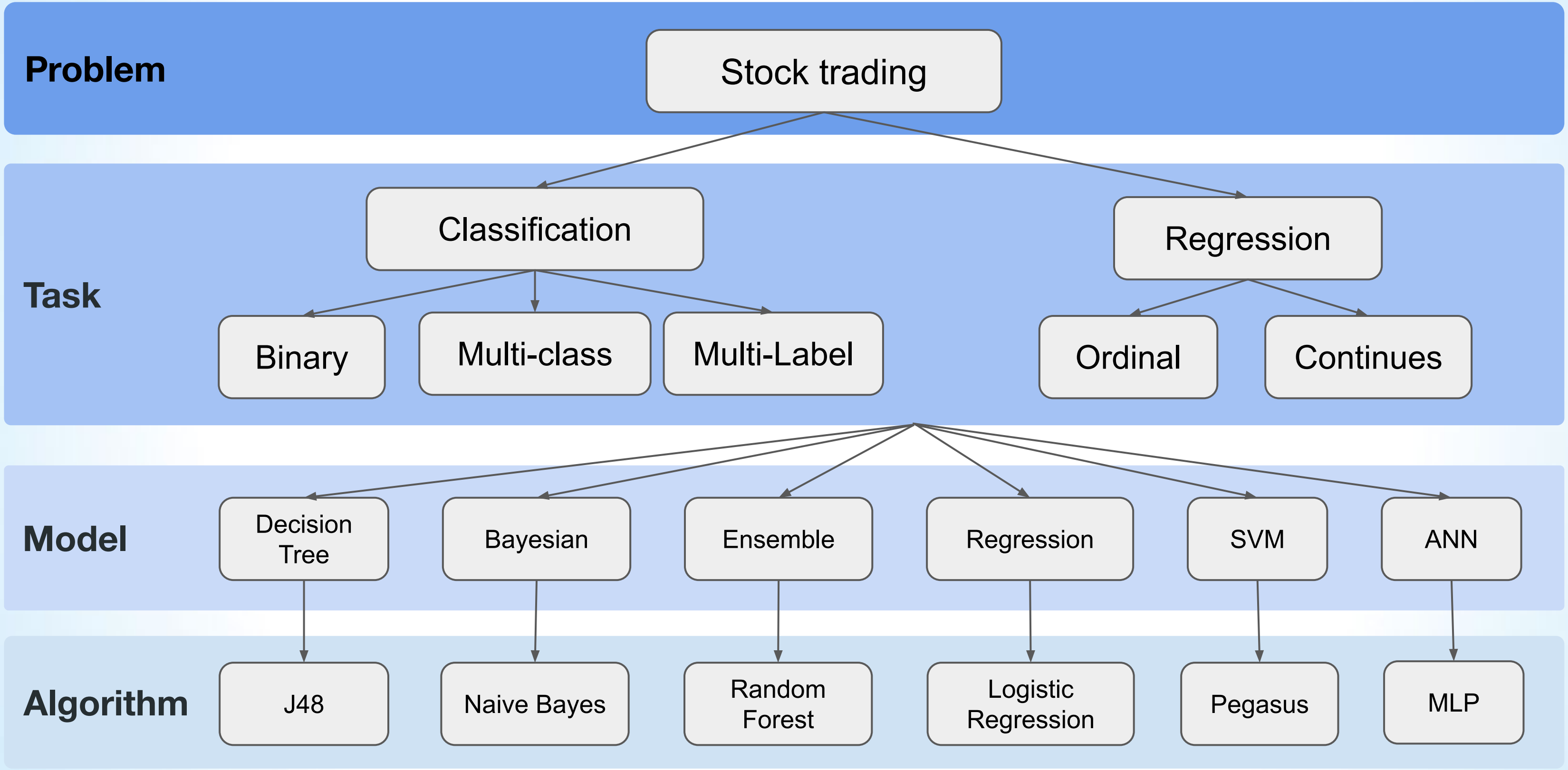
Label



Attribute


SUPERVISED LEARNING
SUPERVISED LEARNING, EVERYWHERE

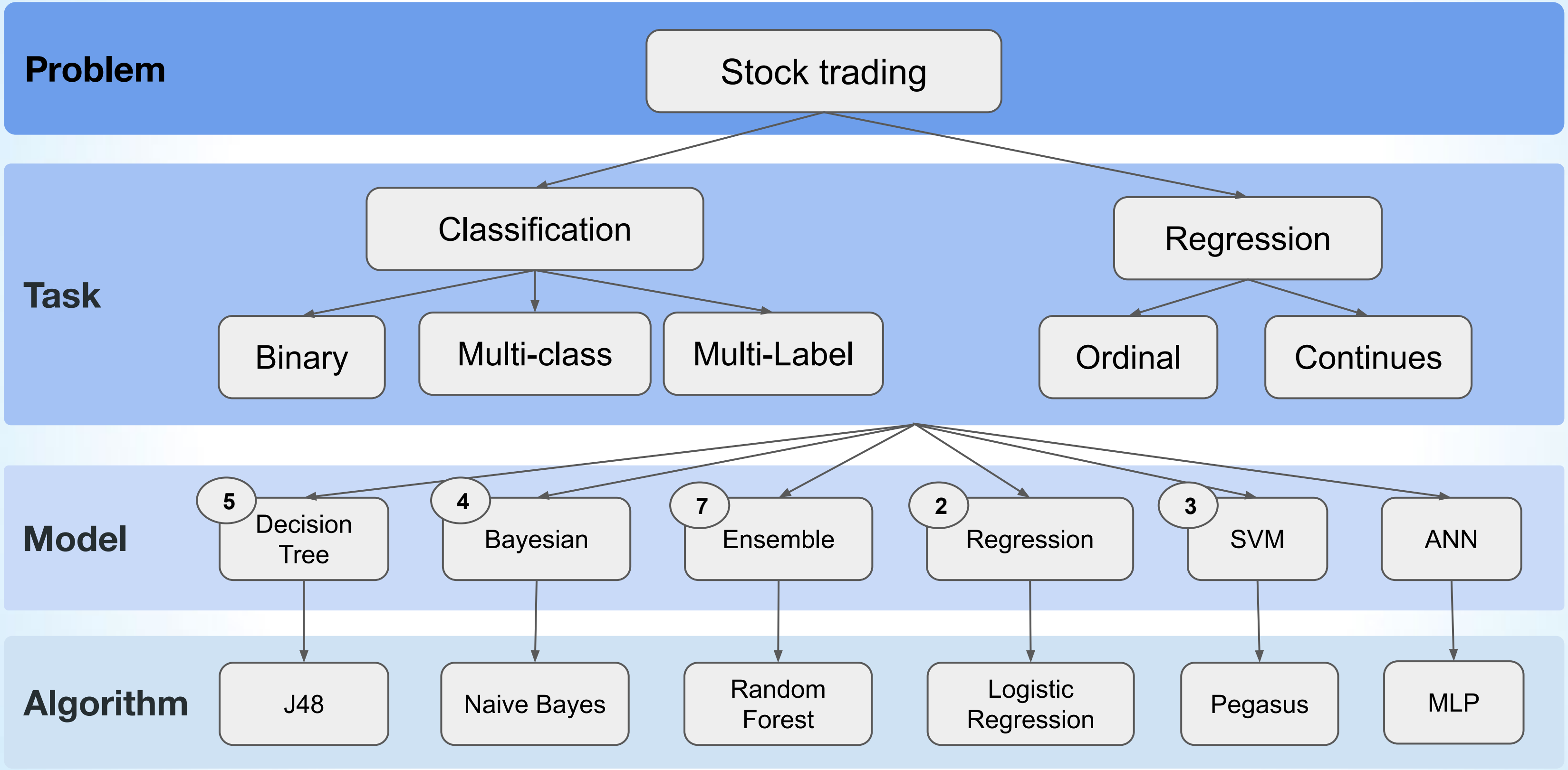
Supervised tasks



Y-DATA SCHOOL OF DATA SCIENCE

Yandex

Supervised tasks



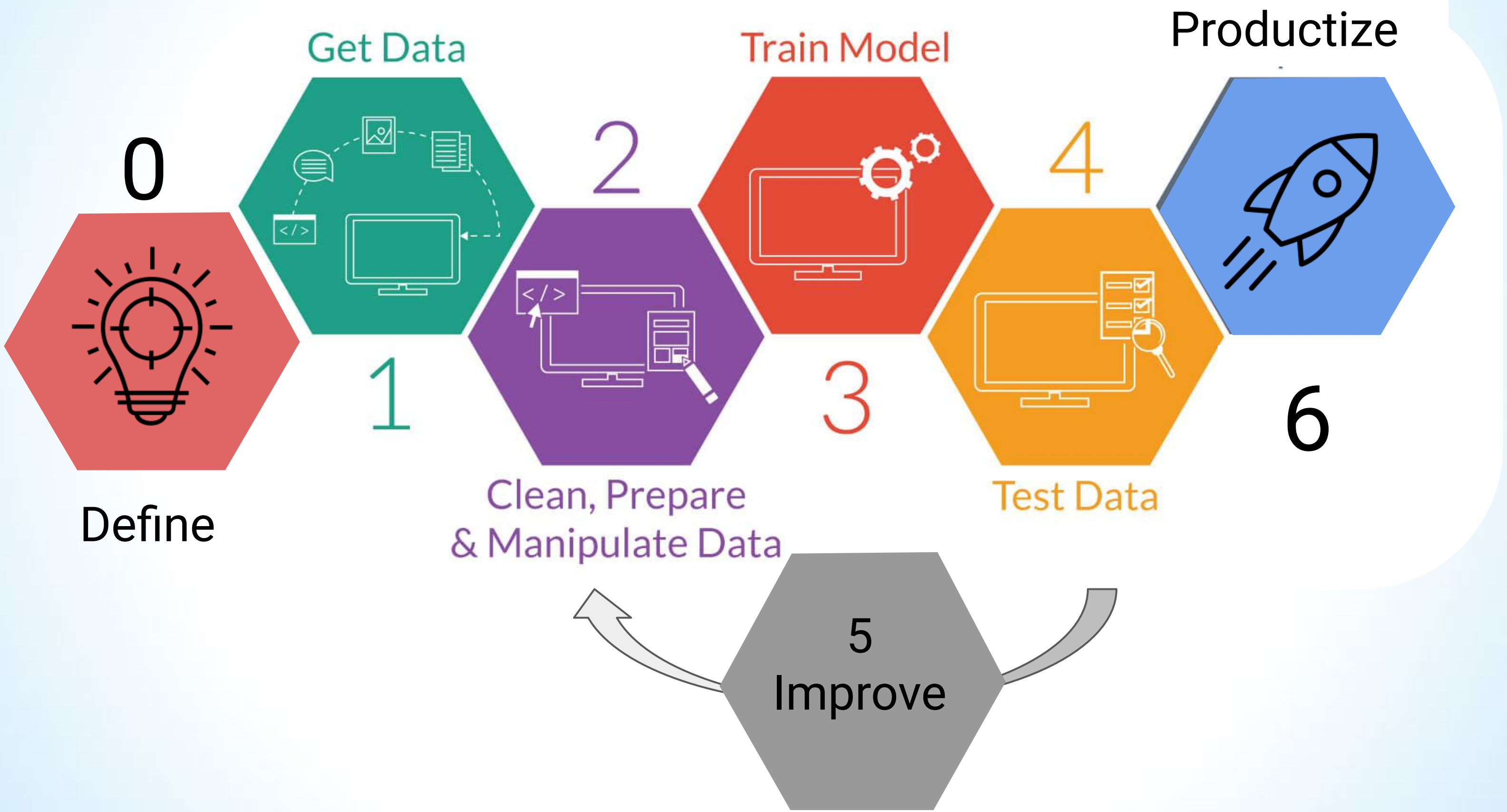
Syllabus

week	Topics
1 **Intro**	Intro to ML (Lior): supervised, unsupervsied, reinforcement Building model cycle: data, model, evaluation. Optimization and linear regression (Noa): Optimization: GD, early stopping, LR, SGD, regularization
2 **Linear & Logistic Regression**	Linear regression: Ordinary Linear Regression , L2/L1 regularization on OLS Logistic regression: Regression vs Classification, Cross entropy Binary/multi class logistic Regression Explainability
3 **SVM**	Hard and soft SVM Hinge loss Kernel trick

week	Topics
4 **Naive Bayes**	Refresher on the Bayes theorem Naive Bayes theory Gender detection/Spam detection use case (including preproseccing)
5 **Decision trees**	Decision Tree as a Greedy Method Optimization Criteria: Gini & Entropy+ L2 Depth, Leaves and other Hyper Parameters
6 **End2End ML**	Formulating a business problem - arranging data, data visualization, feature extraction, EDA, hyper parameter tuning,
7-8 **Ensemble**	Intro to Ensemble Methods - Aggregation - Bagging - Stacking - Boosting - Gradient Boosting

ML in Practice

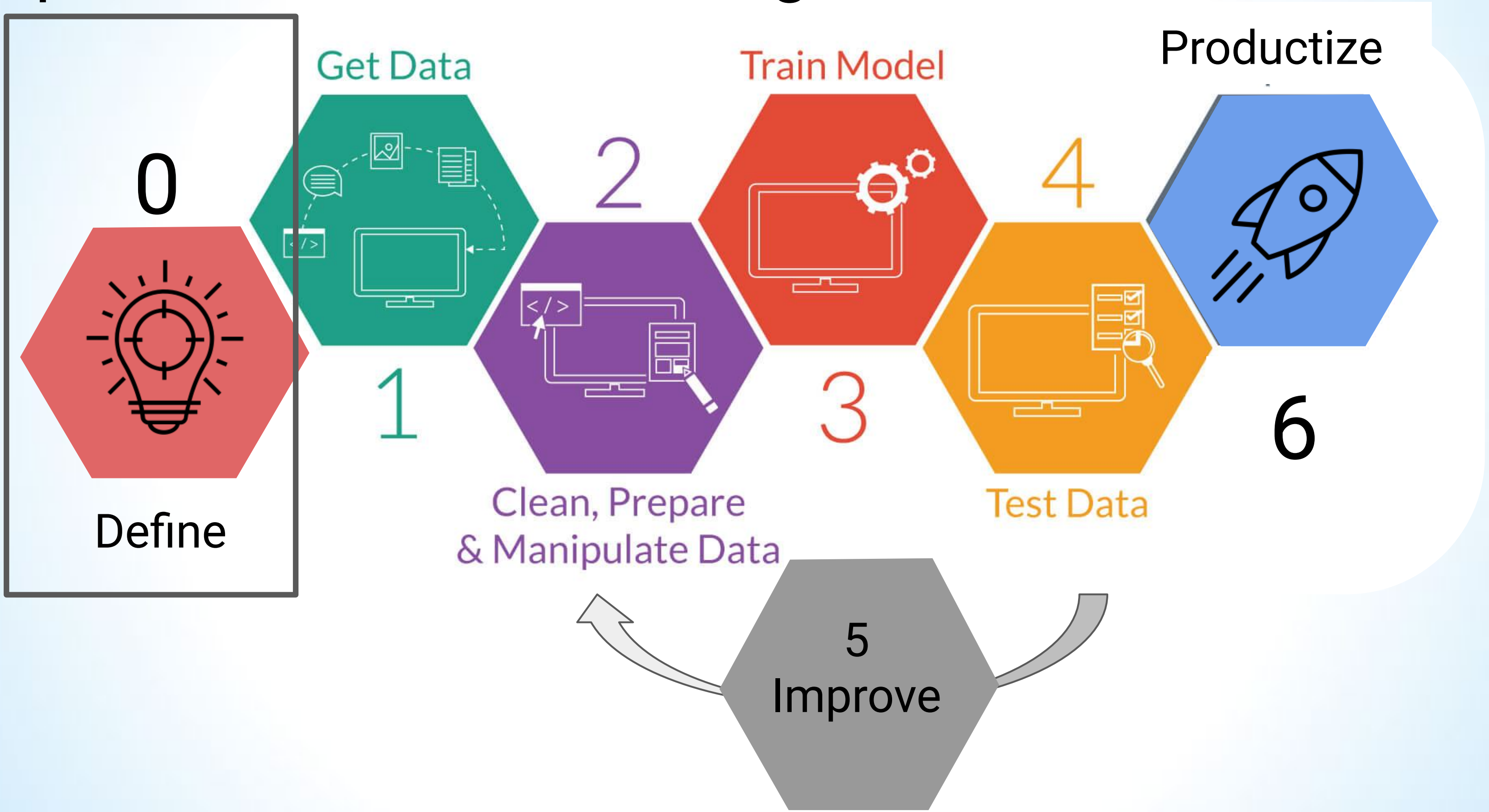
Steps to Predictive Modeling



Y-DATA
SCHOOL OF DATA SCIENCE

Yandex

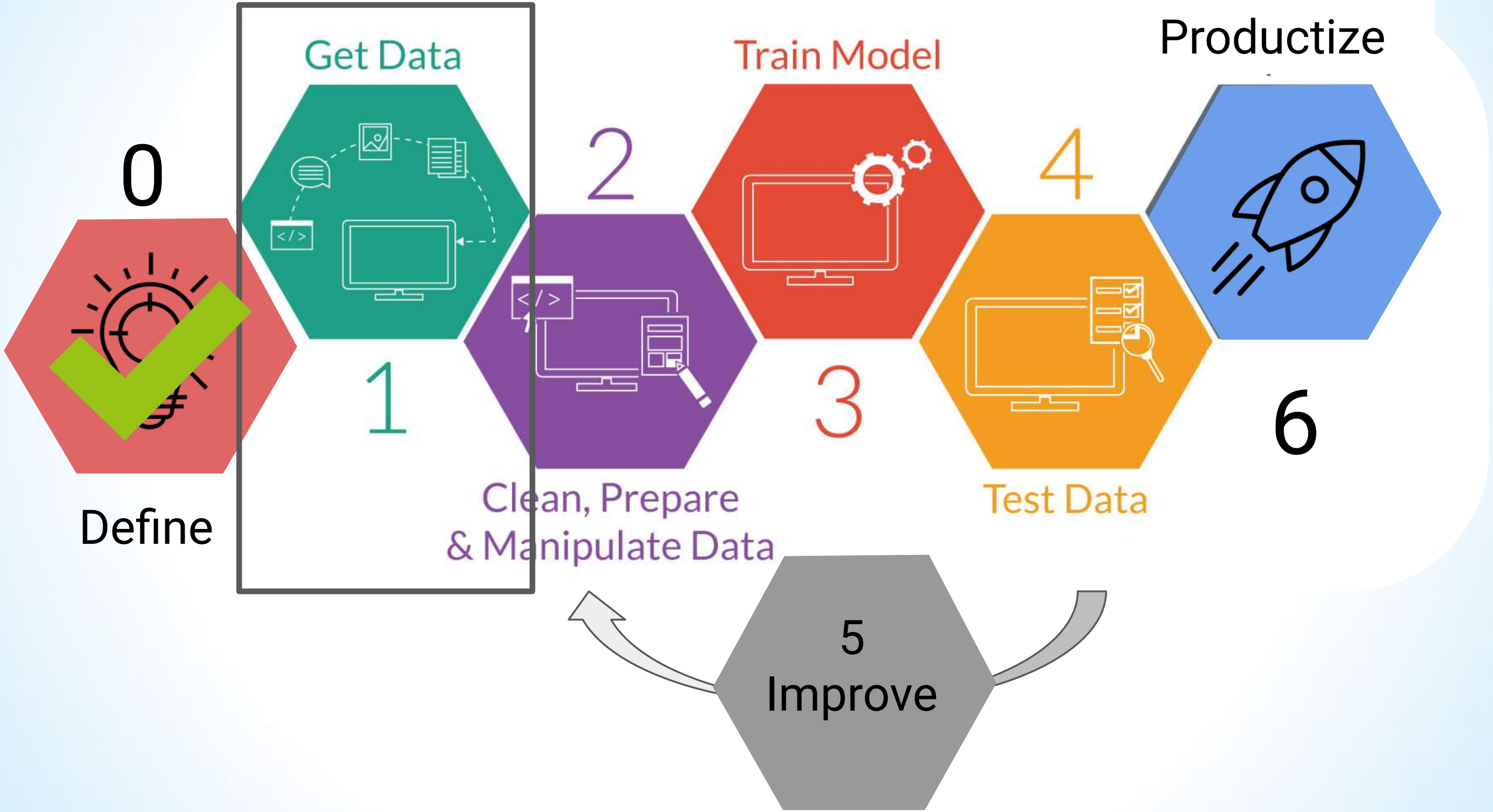
Steps to Predictive Modeling



Y-DATA SCHOOL OF DATA SCIENCE

Yandex

Steps to Predictive Modeling



Ground Truth

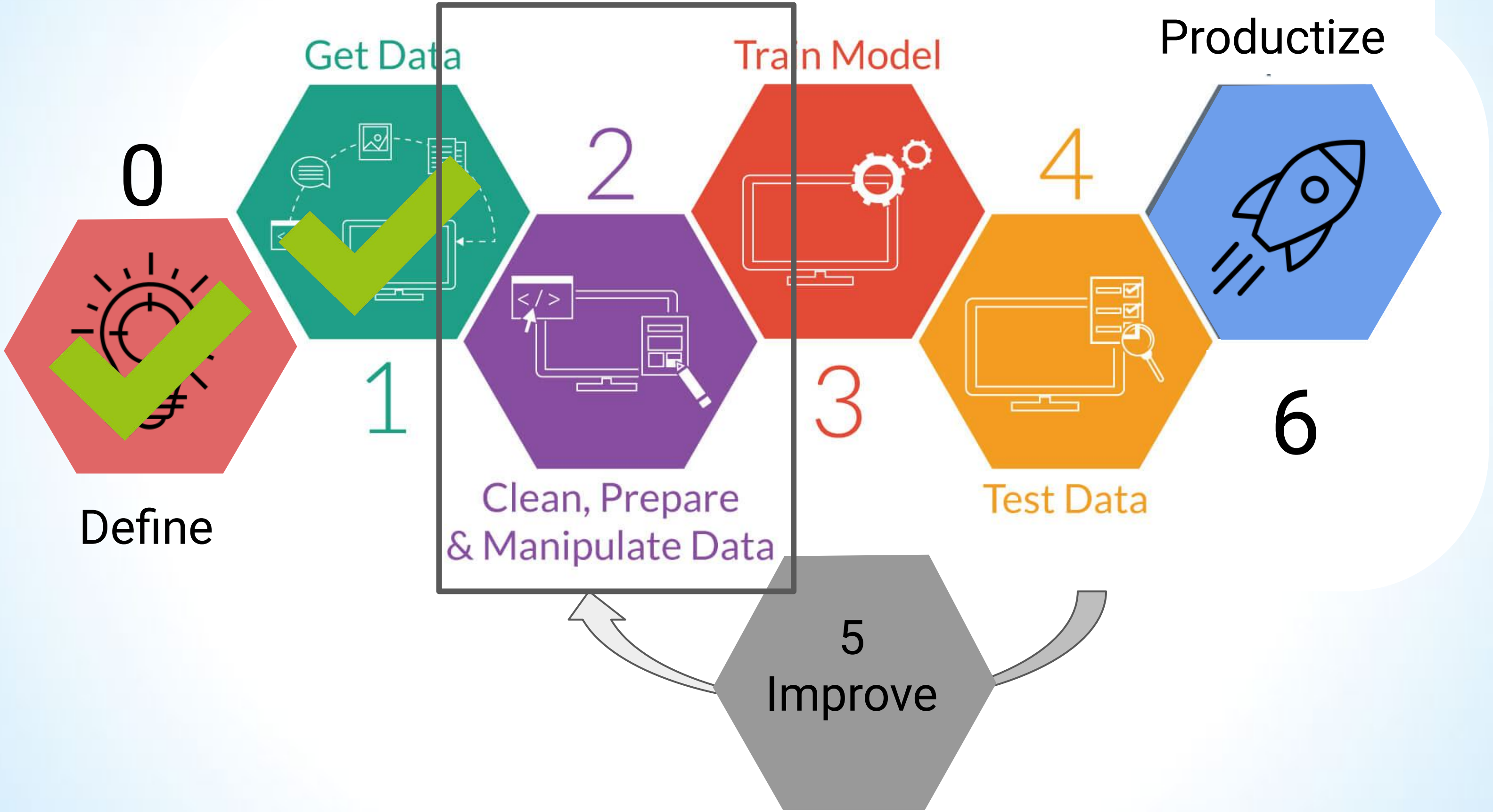
- **Product**
  - Stocks recommender for users.
- **Learning task**
  - Learn if a stock price will increase the next day
- **Data sources**
  - All stock of S&P 500
- **Labeling**
  - A stock with high revenue potential -> 4% price increment

Labeling function

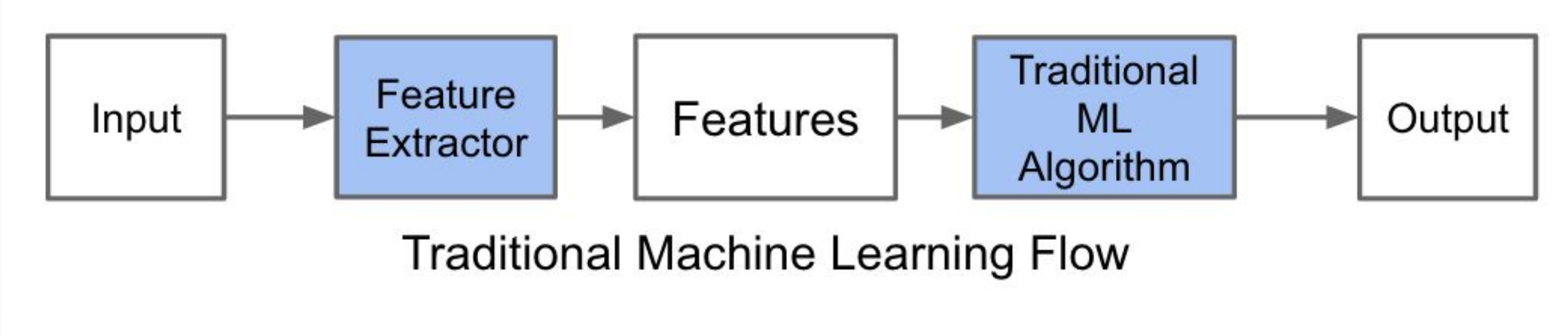
Date	Stock	Price
1/1/2020	TSLA	50
1/1/2020	AMZ	63
1/1/2020	APPL	42
2/1/2020	TSLA	55
2/1/2020	AMZ	60
2/1/2020	APPL	39
3/1/2020	TSLA	60

Date	Stock	Price	Label
1/1/2020	TSLA	50	
1/1/2020	AMZ	63	0
1/1/2020	APPL	42	0
2/1/2020	TSLA	55	1
2/1/2020	AMZ	60	0
2/1/2020	APPL	39	0
3/1/2020	TSLA	60	1

Steps to Predictive Modeling



Classical Learning Framework



http://adilmoujahid.com/posts/2016/06/introduction-deep-learning-python-caffe/

Y-DATA SCHOOL OF DATA SCIENCE

Modeling (Data & Learning)

Baseline

Only last price value:

- Last day increment
- Moving average

Improvements

Based on last year data predict next day performance

- Aggregate last week data
- Extract technical indicator features

Feature extraction - Data modeling

Date	Stock	Price	Label
1/1/2020	TSLA	50	1
1/1/2020	AMZ	63	0
1/1/2020	APPL	42	0
2/1/2020	TSLA	55	0
2/1/2020	AMZ	60	0
2/1/2020	APPL	39	0
3/1/2020	TSLA	60	1

Group by

Stock, week.

Days price increase	Price Chang e	RSI	Sector	Label
4	0.12	0.4	Auto	1

Feature extraction

Days price increase	Price Change	RSI	Sector	Label
4	0.12	0.4	Auto	1
2	0.35	0.7	Software	0
4	0.8	0.5	Energy	0
3	0.22	0.3	Materials	1
5	0.3	0.6	Health	0
1	0.1	0.3	Telco	1

Extract features

Days price increase norm	Price Change <0.3	RSI	Sector_Auto
4 / 7	0	0.4	1
2 / 7	1	0.7	0
4 / 7	1	0.5	0
3 / 7	0	0.3	0
7 / 7	1	0.6	0
1 / 7	0	0.3	0

Y-DATA SCHOOL OF DATA SCIENCE

Yandex

Feature Selection

Days price increase	Price Change	RSI	Sector	Label
4	0.12	0.4	Auto	1
2	0.35	0.7	Software	0
4	0.8	0.5	Energy	0
3	0.22	0.3	Materials	1
5	0.3	0.6	Health	0
1	0.1	0.3	Telco	1

Extract features

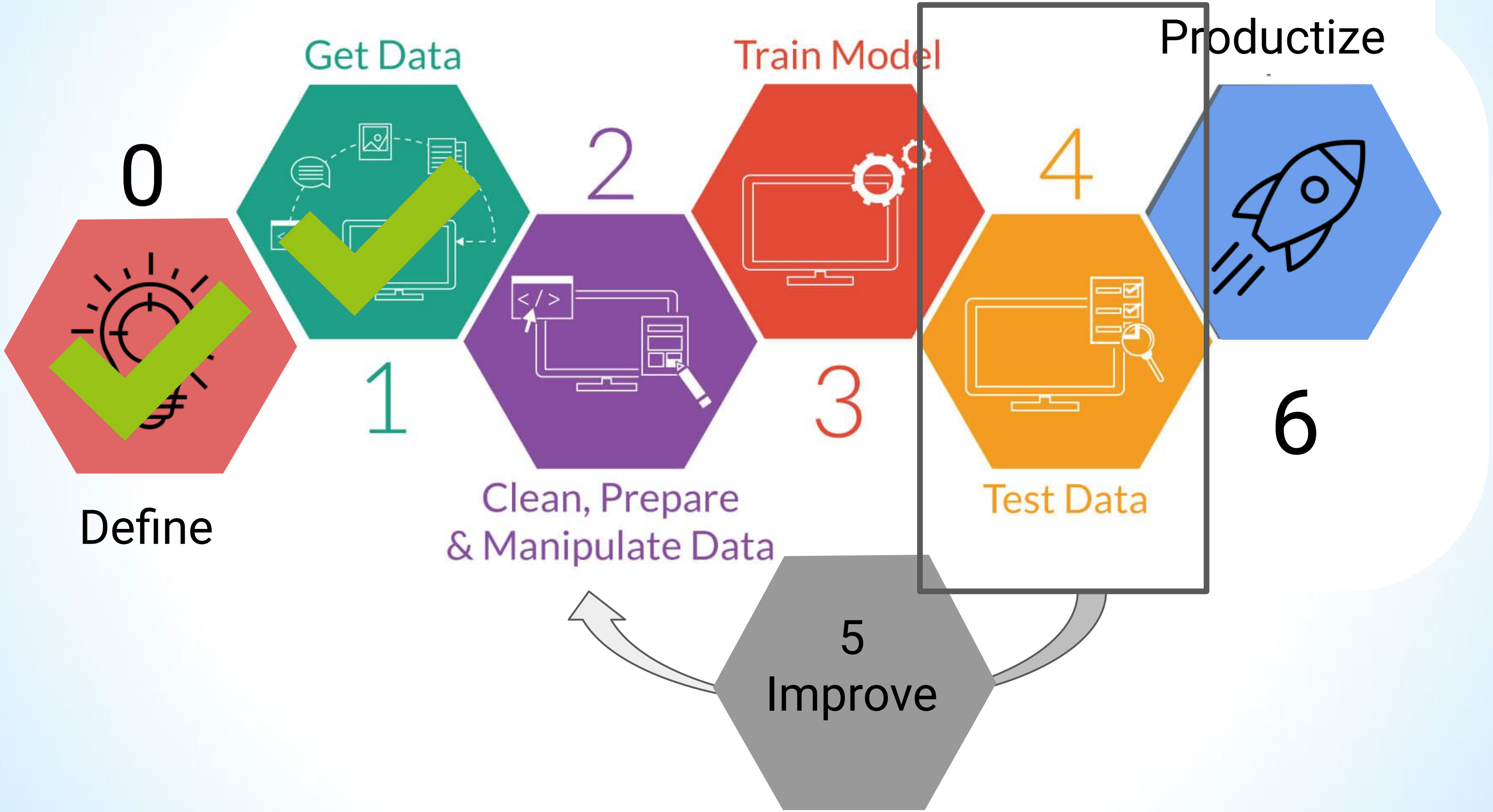
~~Days price increase norm~~	Price Change <0.3	RSI	Sector_ Auto
~~4 / 7~~	0	0.4	1
~~2 / 7~~	1	0.7	0
~~4 / 7~~	1	0.5	0
~~3 / 7~~	0	0.3	0
~~7 / 7~~	1	0.6	0
~~1 / 7~~	0	0.3	0

How to Estimate Model’s Performance

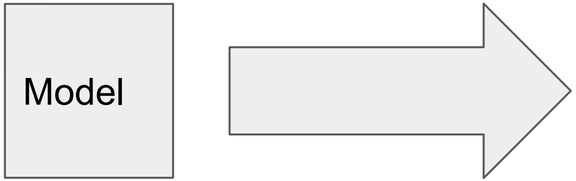
Y-DATA SCHOOL OF DATA SCIENCE

Yandex

Steps to Predictive Modeling



How To Evaluate?



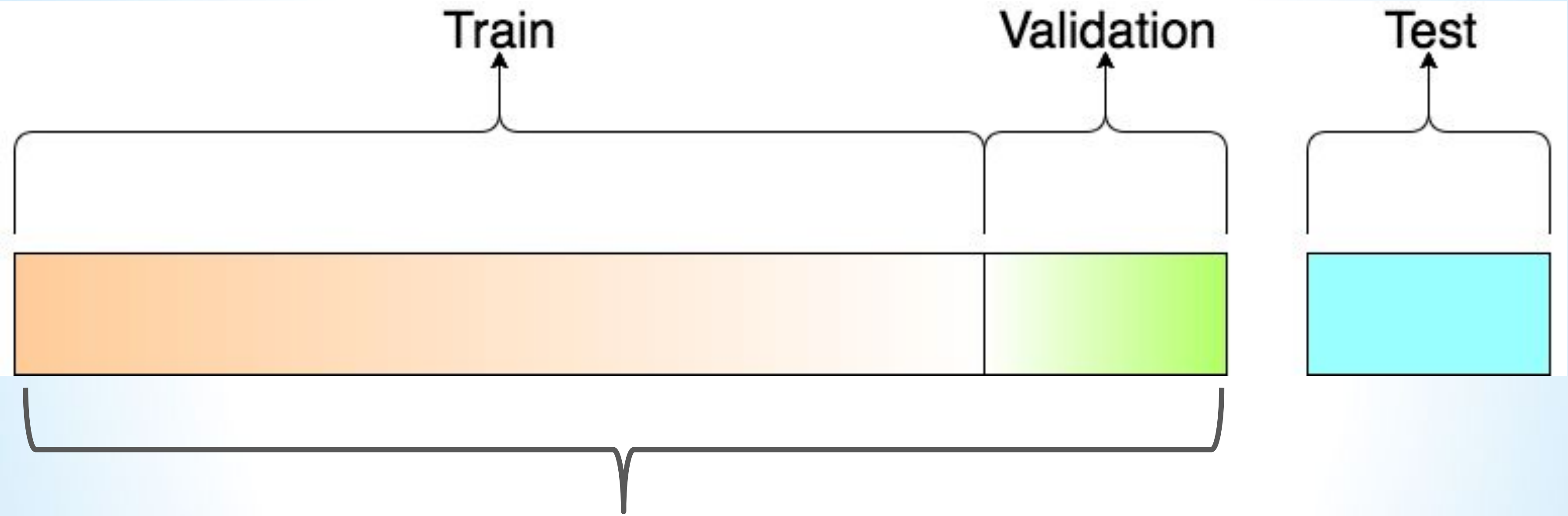
Tesla
- Buy Confidence 80%
- Not Buy Confidence 20%

Amazon
- Buy Confidence 49%
- Not Buy Confidence 51%

Y-DATA SCHOOL OF DATA SCIENCE

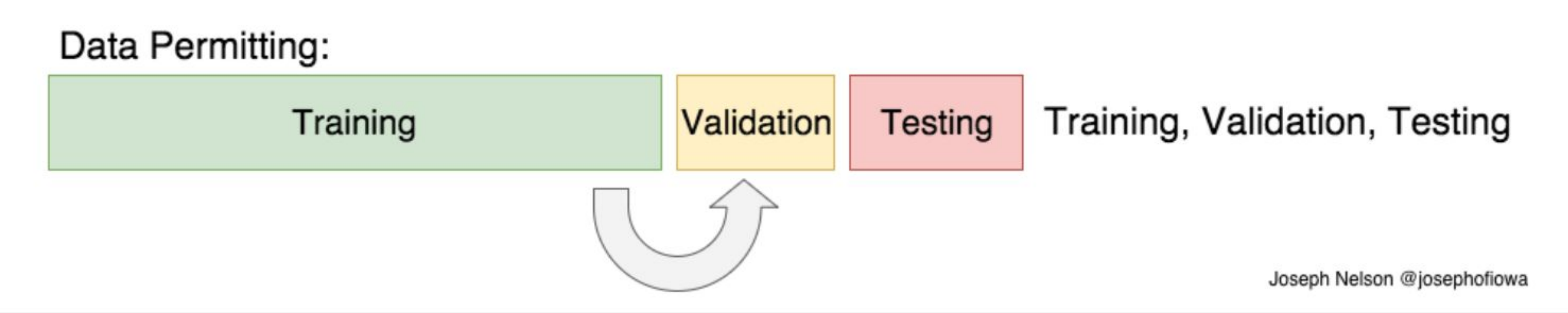
Yandex

Estimating Performance



E.g. Choose best model
Hyper parameter optimization

Estimating Performance - Data is Abundant

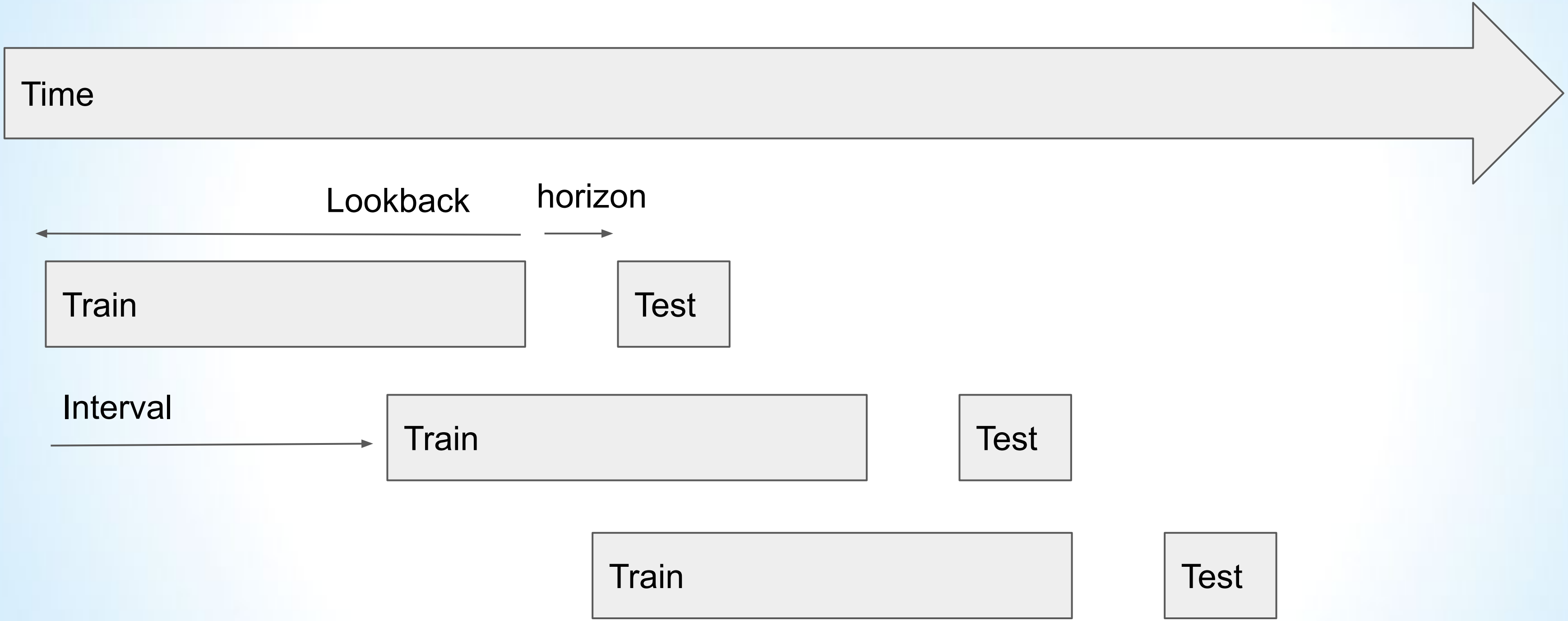


Datasets distribution: Training <> Validation == Test ~ Real world = Random

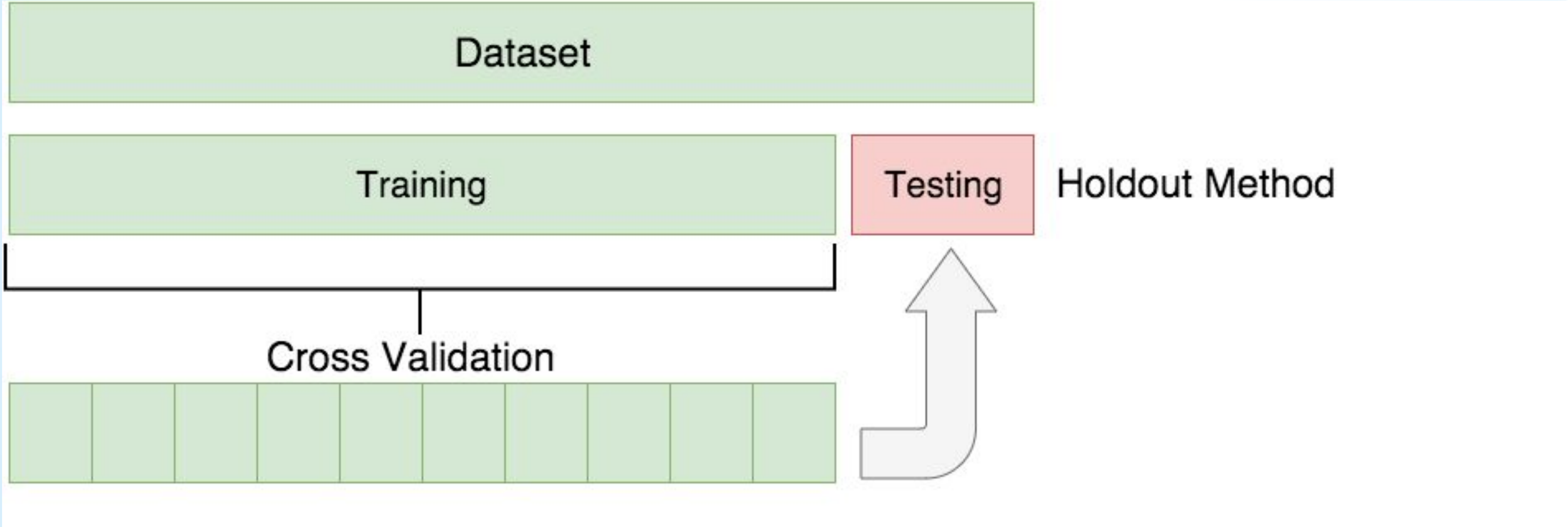
Validation used for Hypertuning and model Calibration

Testing used for final evaluation

Rolling window cross validation



Cross Validation



Classification Metrics

		Label		
		Condition Positive *(Buy)*	**Condition Negative** *(Don’t Buy)*	
Classifier	**Predict Positive** *(should buy)*			
	Predict Negative *(shouldn't buy)*			

Y-DATA SCHOOL OF DATA SCIENCE

Yandex

Classification Metrics

		Label	
		Condition Positive *(Buy)*	**Condition Negative** *(Don't Buy)*
Classifier	**Predict Positive** *(should buy)*	**True Positive** (TP) = 20	
	Predict Negative *(shouldn't buy)*		**True Negative** (TN) = 1820

Classification Metrics

		Label		
		Condition Positive *(Buy)*	**Condition Negative** *(Don't Buy)*	
Classifier	**Predict Positive** *(should buy)*	**True Positive** (TP) = 20	**False Positive** (FP) = 180	
	Predict Negative *(shouldn't buy)*	**False Negative** (FN) = 10	**True Negative** (TN) = 1820	

Classification Metrics

		Label		
		Condition Positive *(Buy)*	**Condition Negative** *(Don't Buy)*	
Classifier	**Predict Positive** *(should buy)*	**True Positive** (TP) = 20	**False Positive** (FP) = 180	**Positive predictive value | Precision** TP / (TP + FP) = 20 / (20 + 180) = **10%**
	Predict Negative *(shouldn't buy)*	**False Negative** (FN) = 10	**True Negative** (TN) = 1820	**Negative predictive value** TN / (FN + TN) = 1820 / (10 + 1820) ≈ **99.5%**

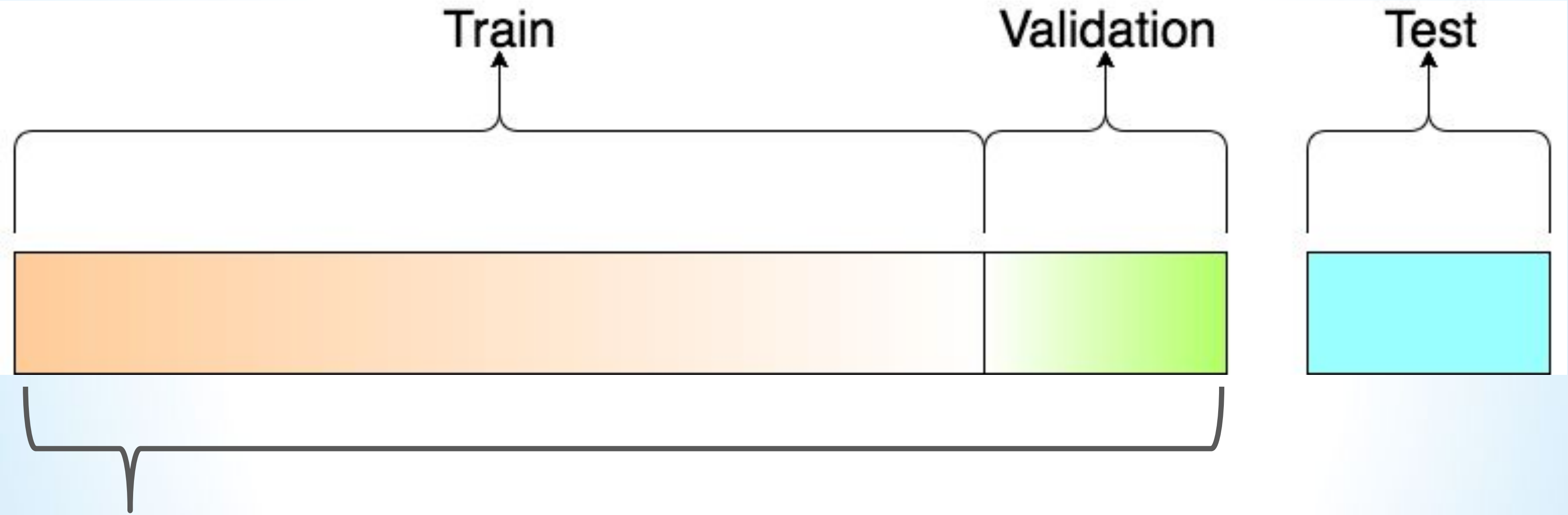
Classification Metrics

		Label		
		Condition Positive *(Buy)*	**Condition Negative** *(Don't Buy)*	
Classifier	**Predict Positive** *(should buy)*	**True Positive** (TP) = 20	**False Positive** (FP) = 180	**Positive predictive value | Precision** TP / (TP + FP) = 20 / (20 + 180) = **10%**
	Predict Negative *(shouldn't buy)*	**False Negative** (FN) = 10	**True Negative** (TN) = 1820	**Negative predictive value** TN / (FN + TN) = 1820 / (10 + 1820) ≈ **99.5%**
		True Positive Rate | Recall | Sensitivity TP / (TP + FN) = 20 / (20 + 10) ≈ 67%	**Specificity** TN / (FP + TN) = 1820 / (180 + 1820) = 91%	

Classification Metrics

		Label		
		Condition Positive *(Buy)*	**Condition Negative** *(Don't Buy)*	
Classifier	**Predict Positive** *(should buy)*	**True Positive** (TP) = 20	**False Positive** (FP) = 180	**Positive predictive value | Precision** TP / (TP + FP) = 20 / (20 + 180) = **10%**
	Predict Negative *(shouldn't buy)*	**False Negative** (FN) = 10	**True Negative** (TN) = 1820	**Negative predictive value** TN / (FN + TN) = 1820 / (10 + 1820) ≈ **99.5%**
		True Positive Rate | Recall | Sensitivity TP / (TP + FN) = 20 / (20 + 10) ≈ 67%	**Specificity** TN / (FP + TN) = 1820 / (180 + 1820) = 91%	**Accuracy** (TP + TN) / (TP + TN + FP + FN) **F1 score** (2*Precision*Recall) / (Precision + Recall)

Estimating Performance

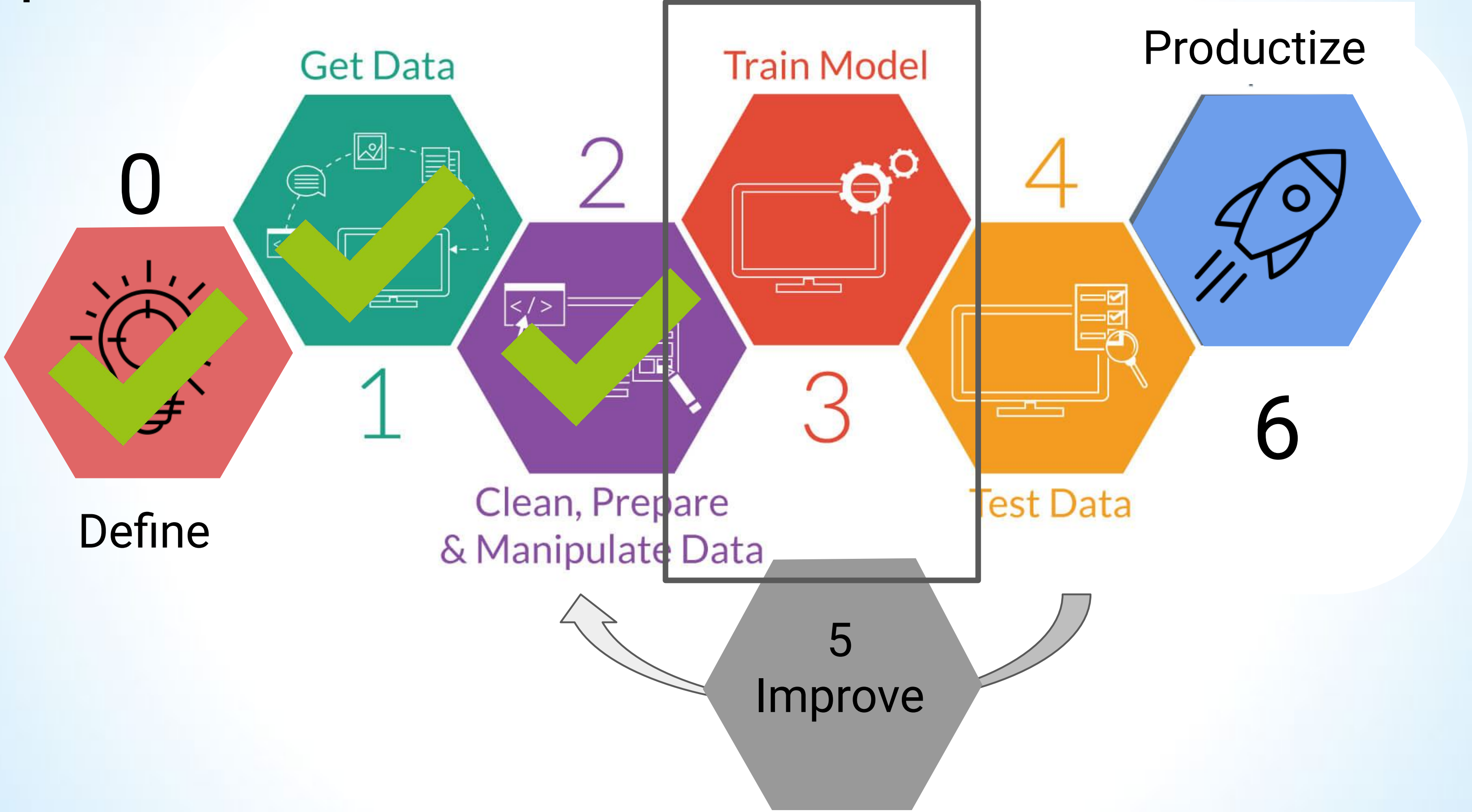


E.g. Choose best model
Hyper parameter optimization

How should we split the stocks data?
Why we need validation?
What is the test?

What types of ML Algorithms are there?

Steps to Predictive Modeling



Parametric vs. Non-parametric Models

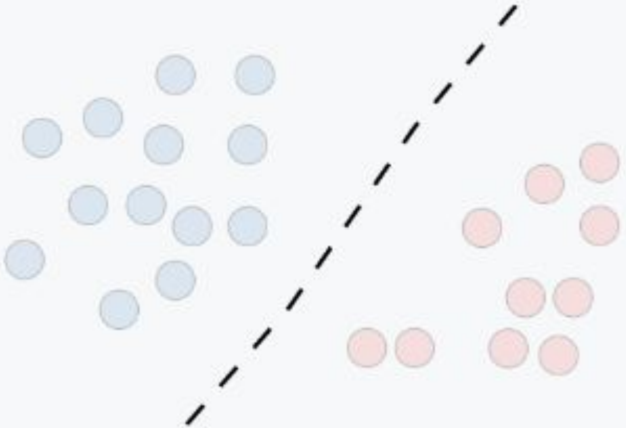
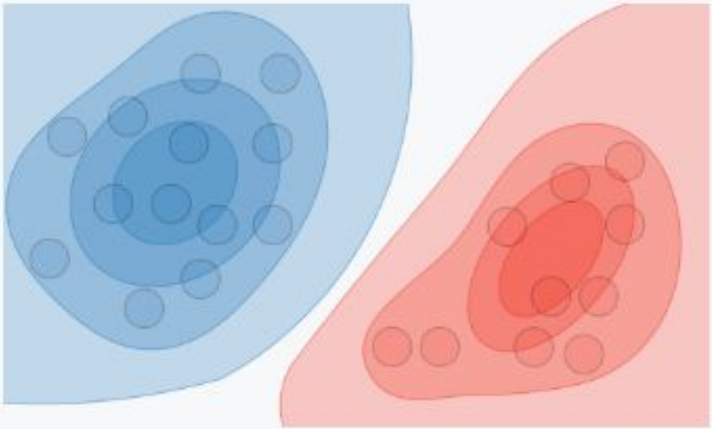
Almost all models for machine learning have "parameters" or "weights" that need to be learned.

Parametric Models	Nonparametric models
The number of parameters is constant, or independent of the number of training examples.	The number of parameters grows with the number of training examples.

Can you think of an example?

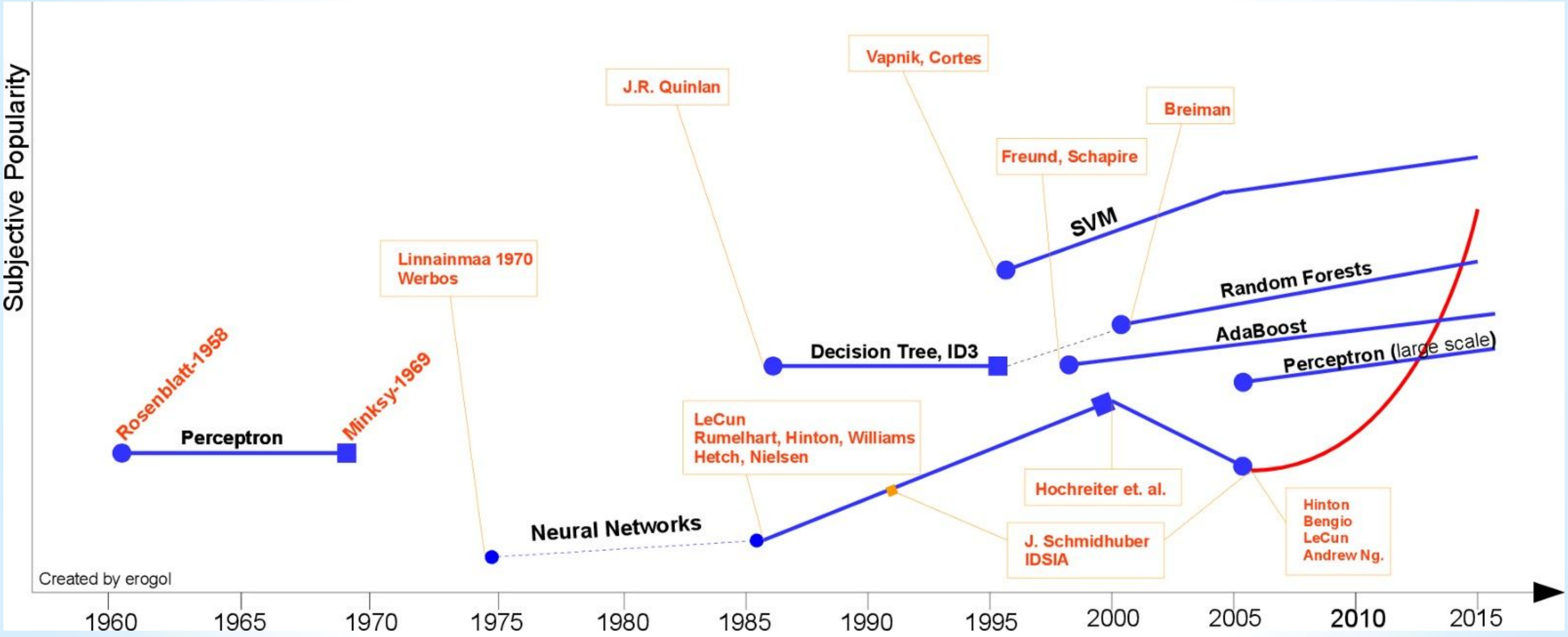
Can you think of an example for parametric and non-parametric method?

Generative vs. Discriminative

	Discriminative model	Generative model
Goal	Directly estimate $P(y\|x)$	Estimate $P(x\|y)$ to then deduce $P(y\|x)$
What's learned	Decision boundary	Probability distributions of the data
Illustration		
Examples	Regressions, SVMs	GDA, Naive Bayes

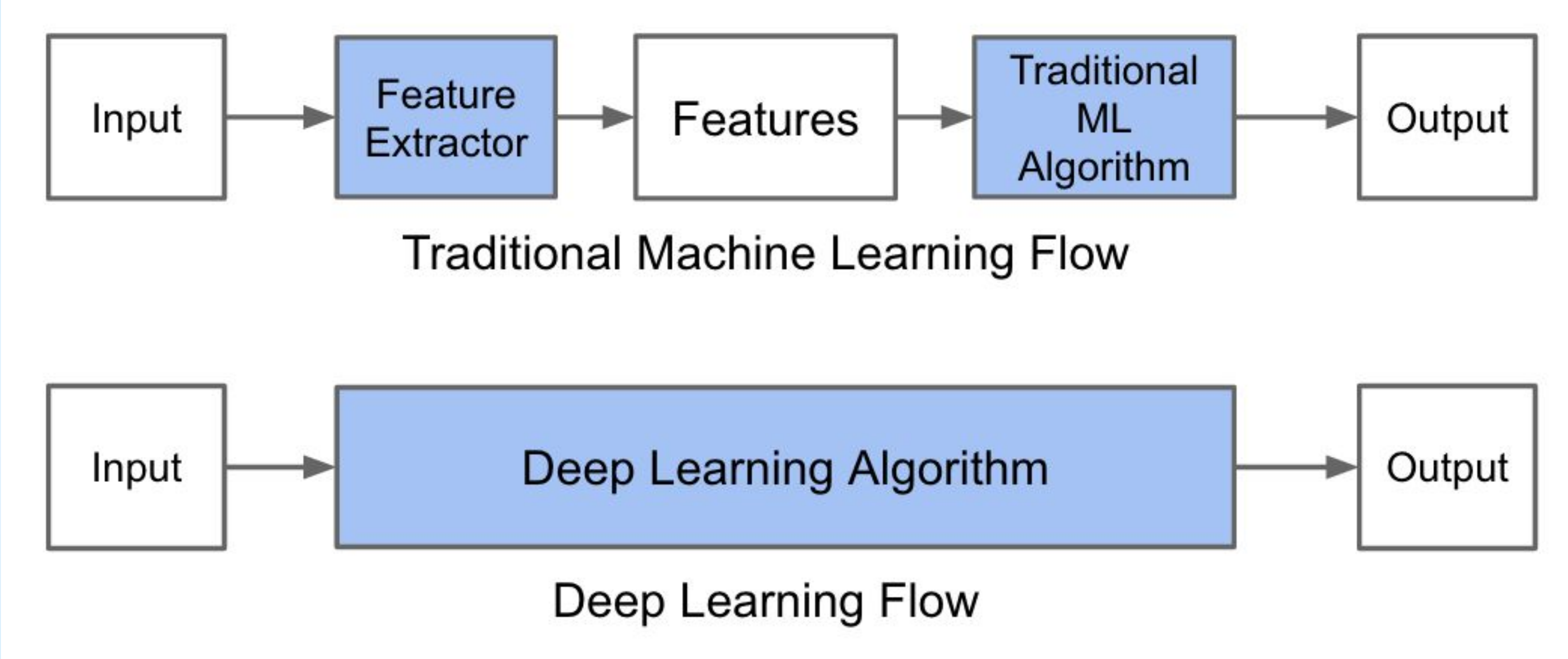
https://stackoverflow.com/questions/879432/what-is-the-difference-between-a-generative-and-discriminative-algorithm

The Brief History of Machine Learning



https://chatbotnewsdaily.com/since-the-initial-standpoint-of-science-technology-and-ai-scientists-following-blaise-pascal-and-804ac13d8151

Classical vs Deep Learning Framework



http://adilmoujahid.com/posts/2016/06/introduction-deep-learning-python-caffe/

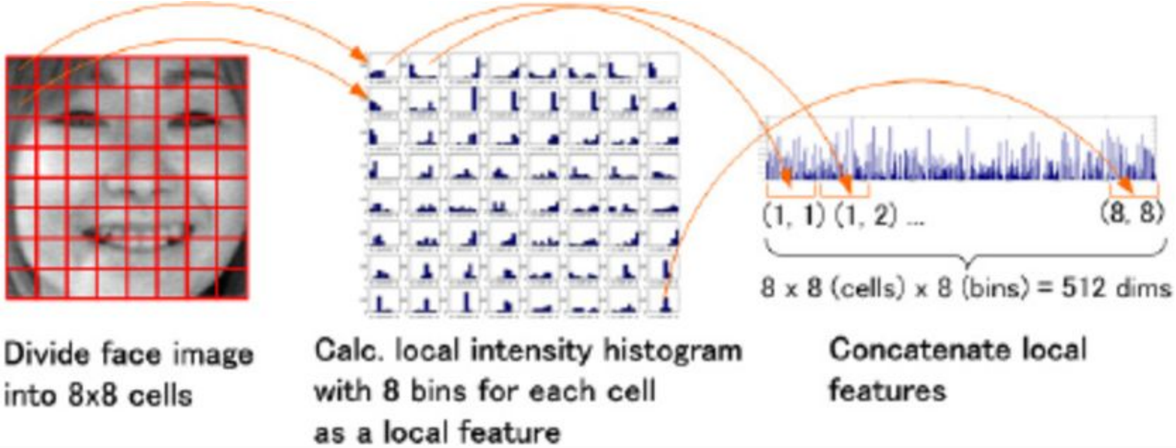
Y-DATA SCHOOL OF DATA SCIENCE

Yandex

Feature Extraction

“Algorithm which transforms raw data into numeric values which can be used as input to a learning algorithm. Usually helps with **reducing** and **fixing** dimensionality.”

e.g.



Shimada K., Matsukawa T., Noguchi Y., Kurita T. (2011) Appearance-Based Smile Intensity Estimation by Cascaded Support Vector Machines. In: Koch R., Huang F. (eds) Computer Vision – ACCV 2010 Workshops. ACCV 2010. Lecture Notes in Computer Science, vol 6468. Springer, Berlin, Heidelberg

Y-DATA
SCHOOL OF DATA SCIENCE

Some Realities on DL

Don't be fool by the hype

1. Can be beaten by GBT for tabular data (CatBoost, XGBoost, LightGBM).

2. No Feature Engineering - Yuppie.
   Yet… Network Architecture Search (NAS), Annoying GPU issues, Loss design, hours of training

3. Overkill sometimes and infeasible

Example: try to train a DL to predict if a number is even or odd…

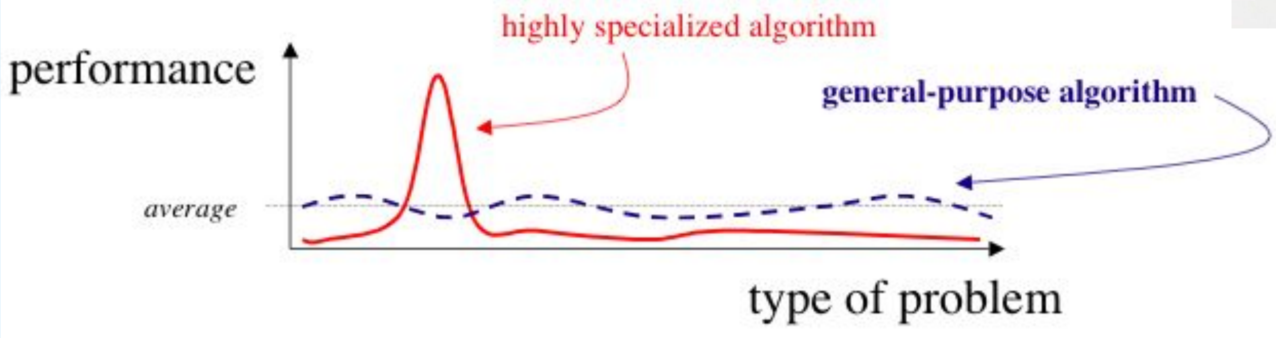
Y-DATA SCHOOL OF DATA SCIENCE

Yandex

No Free Lunch Theorem - Best Model Does Not Exists

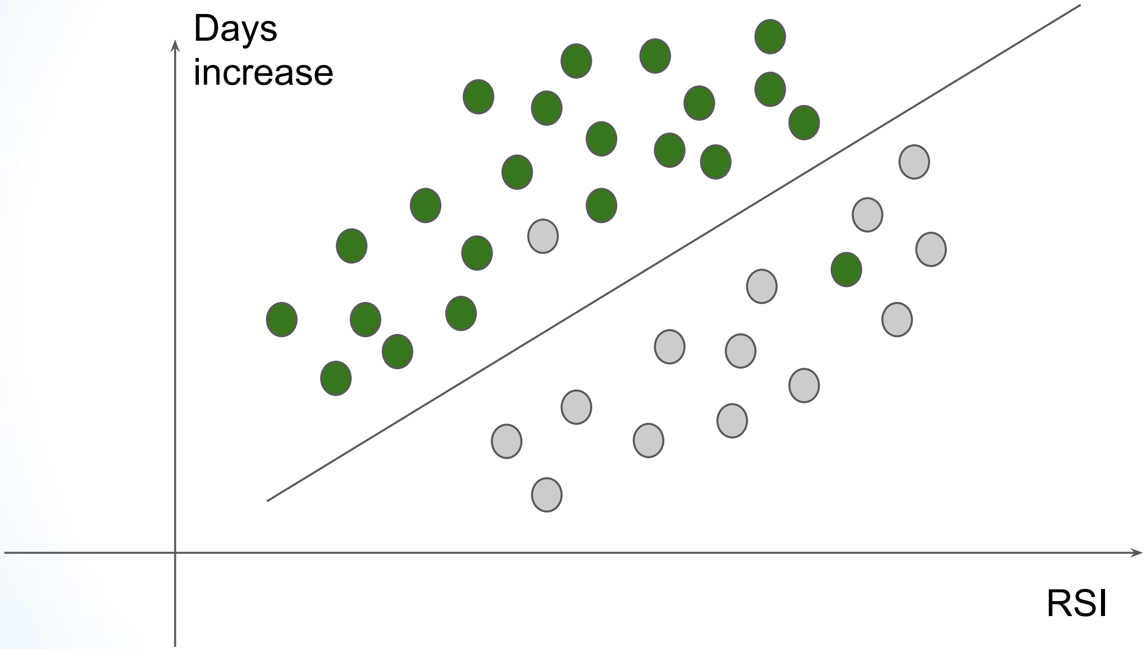
A superior black-box optimisation strategy, which is better than anything else for any kind of problem, is impossible.

Deep cannot be always better.



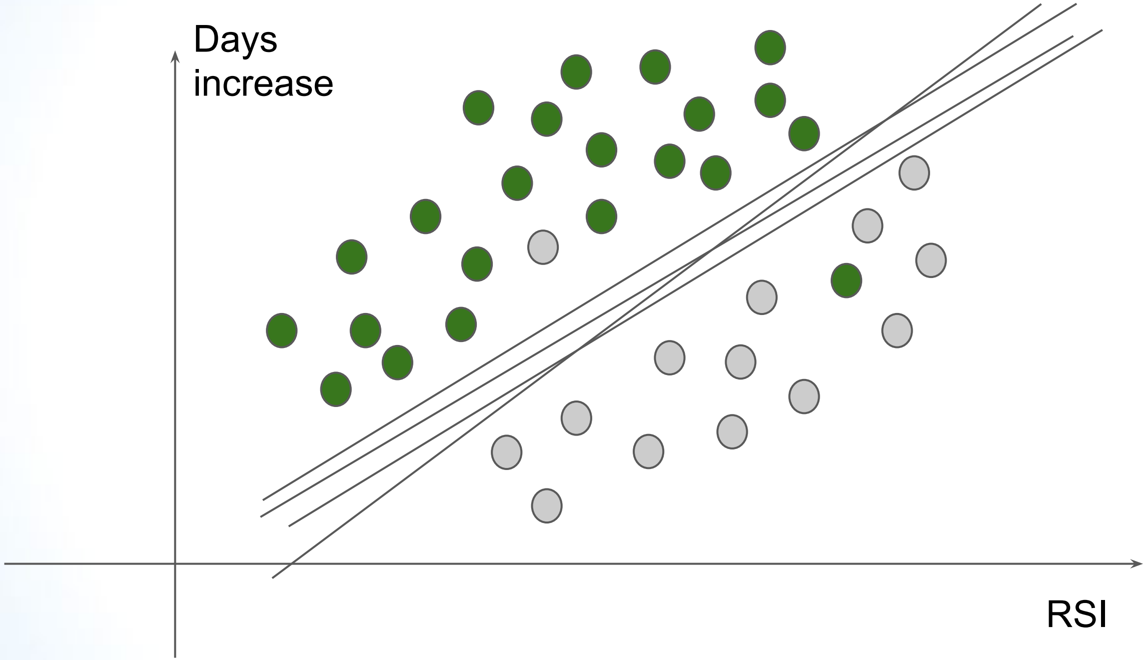
https://medium.com/@LeonFedden/the-no-free-lunch-theorem-62ae2c3ed10c

Modeling - Linear

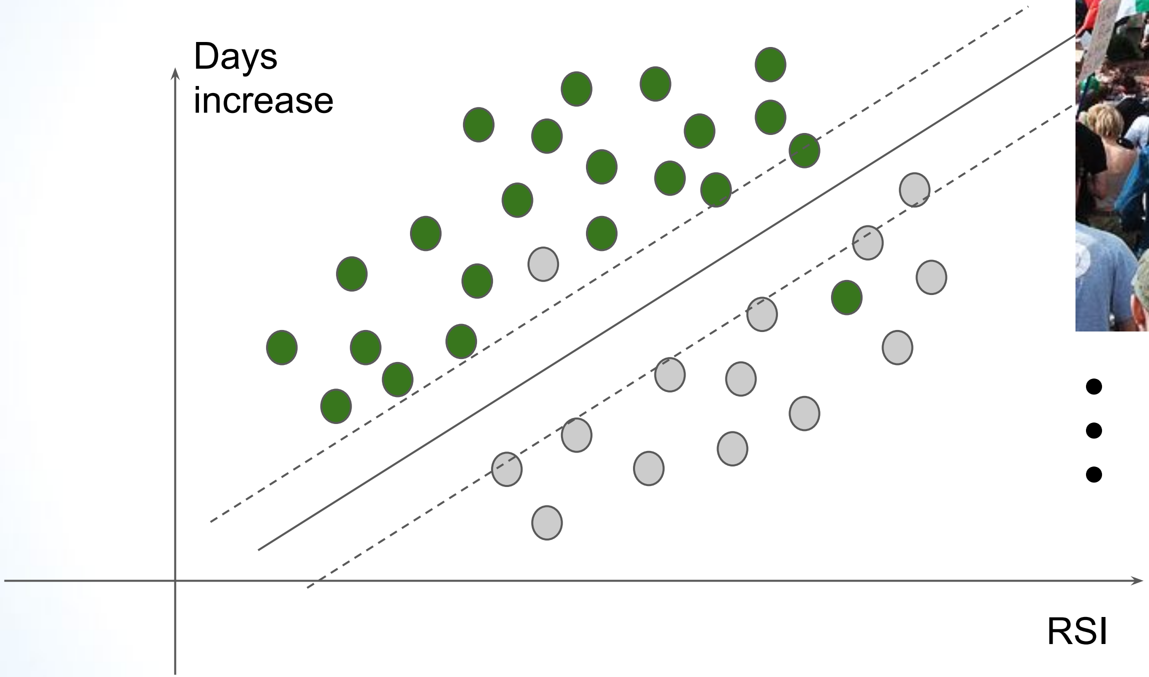


sign(w1*x1 + w2*x2)

Modeling - Linear

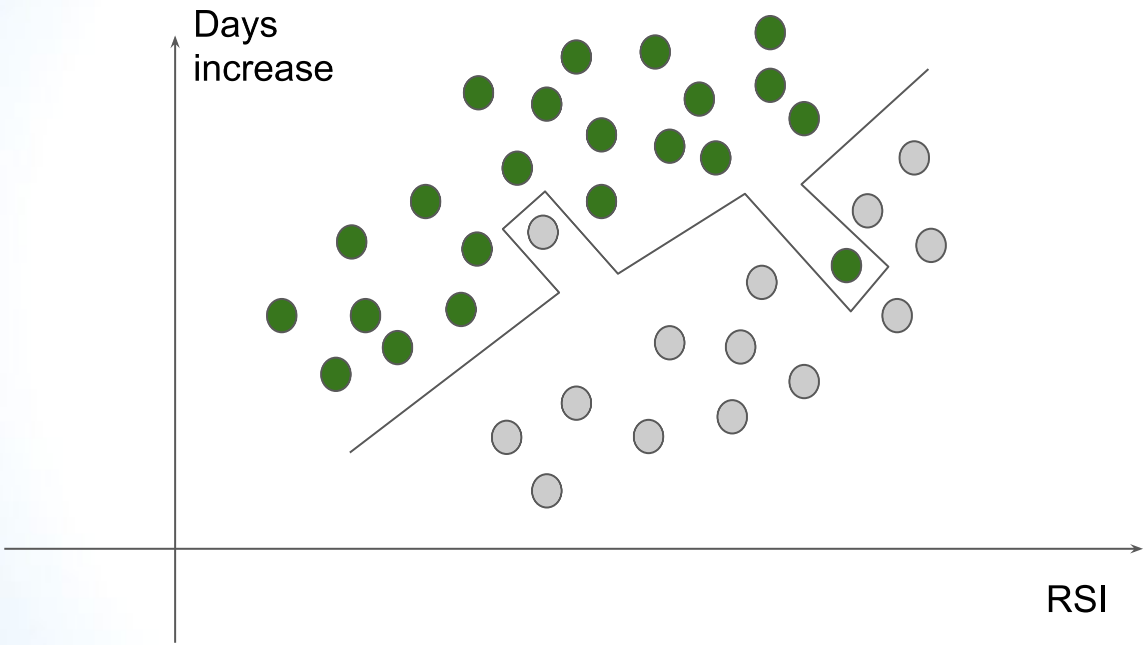


Modeling - Maximum Margin

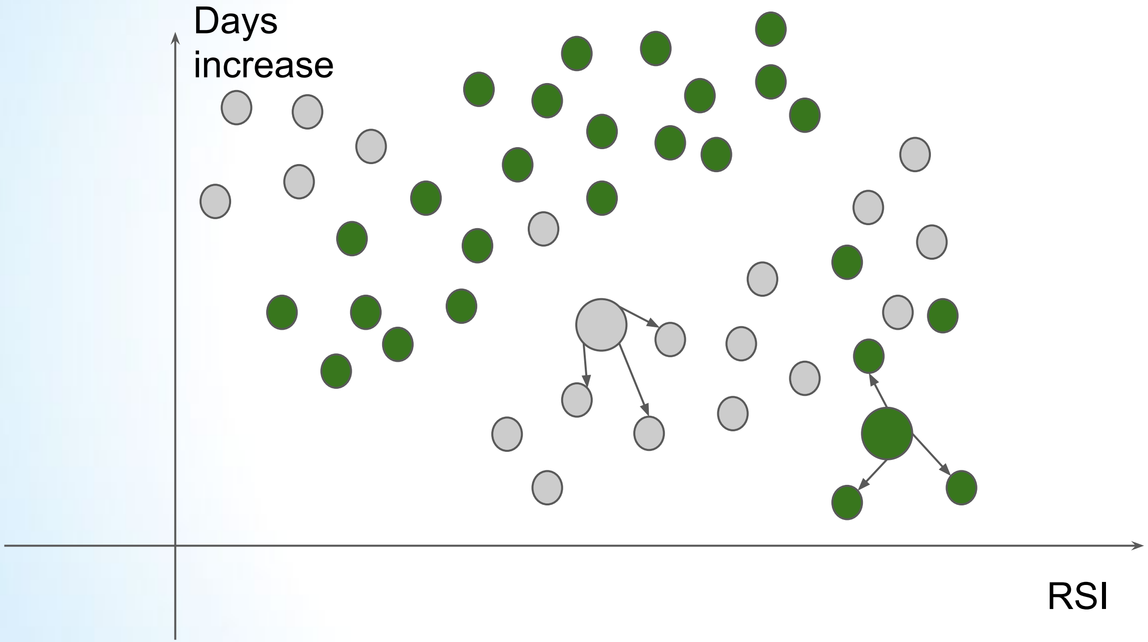


- A Linear classifier
- Maximize the margin
- Simple Linear SVM - LSVM

Modeling



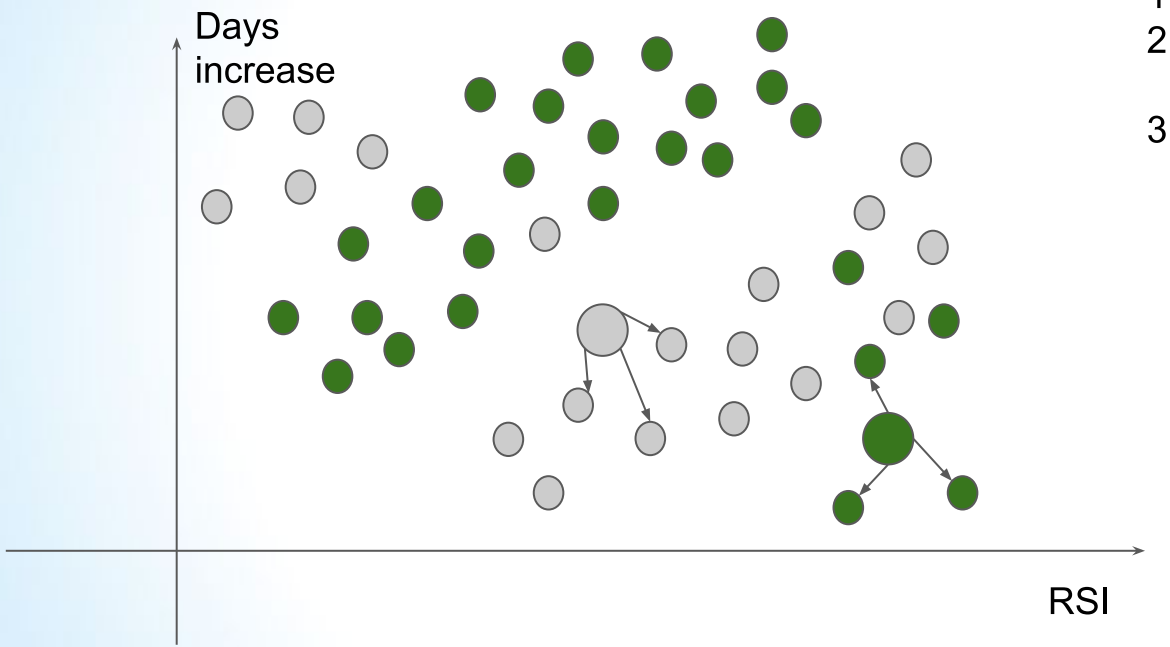
Modeling - K nearest neighbors



Y-DATA SCHOOL OF DATA SCIENCE

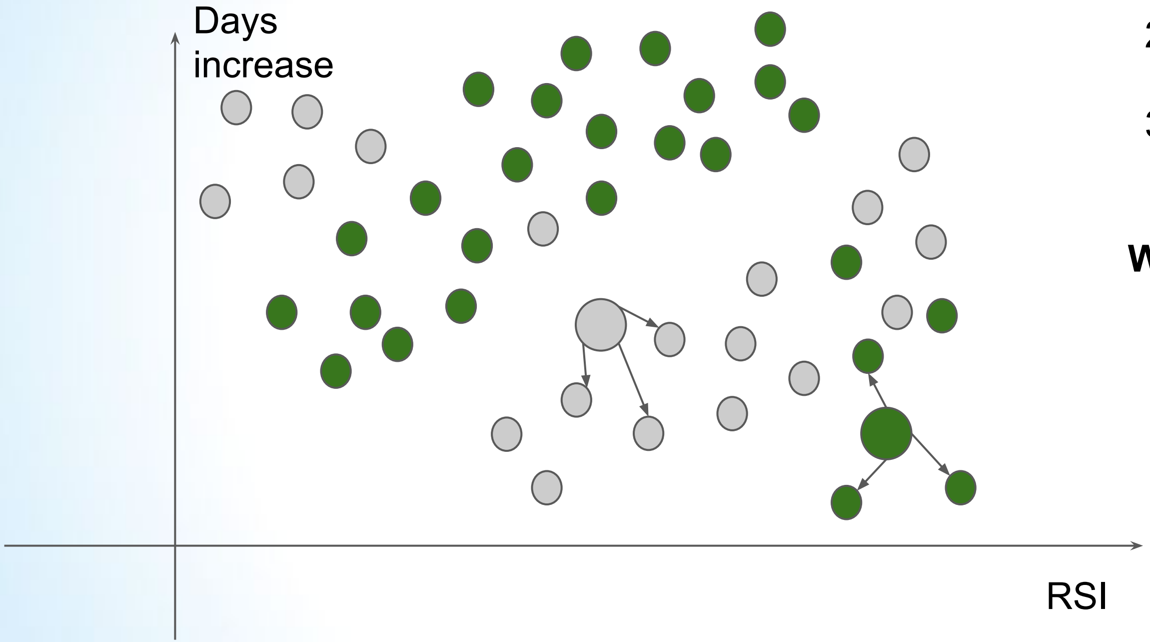
Yandex

Modeling - K nearest neighbors



1. To classify a new input vector x
2. Examine the k closest training data points to x
3. Assign the object to the most frequently occurring class

Modeling - K nearest neighbors

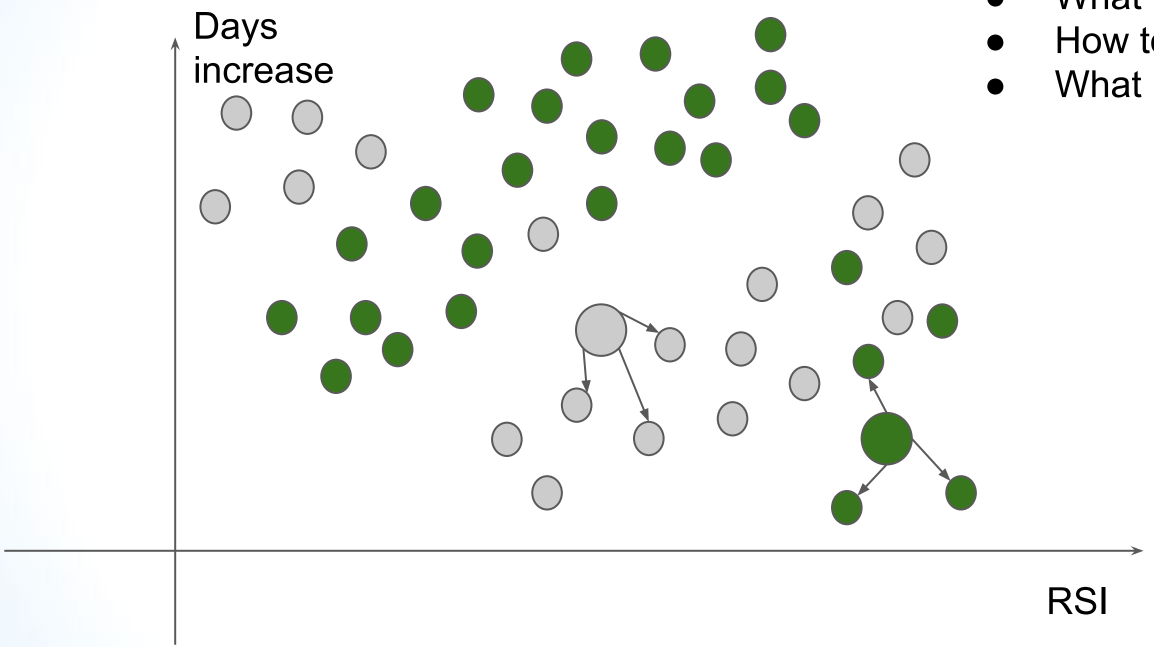


1. To classify a new input vector x
2. Examine the k closest training data points to x
3. Assign the object to the most frequently occurring class

What about?

- K is Odd vs Even K?
- How can we apply Voting?

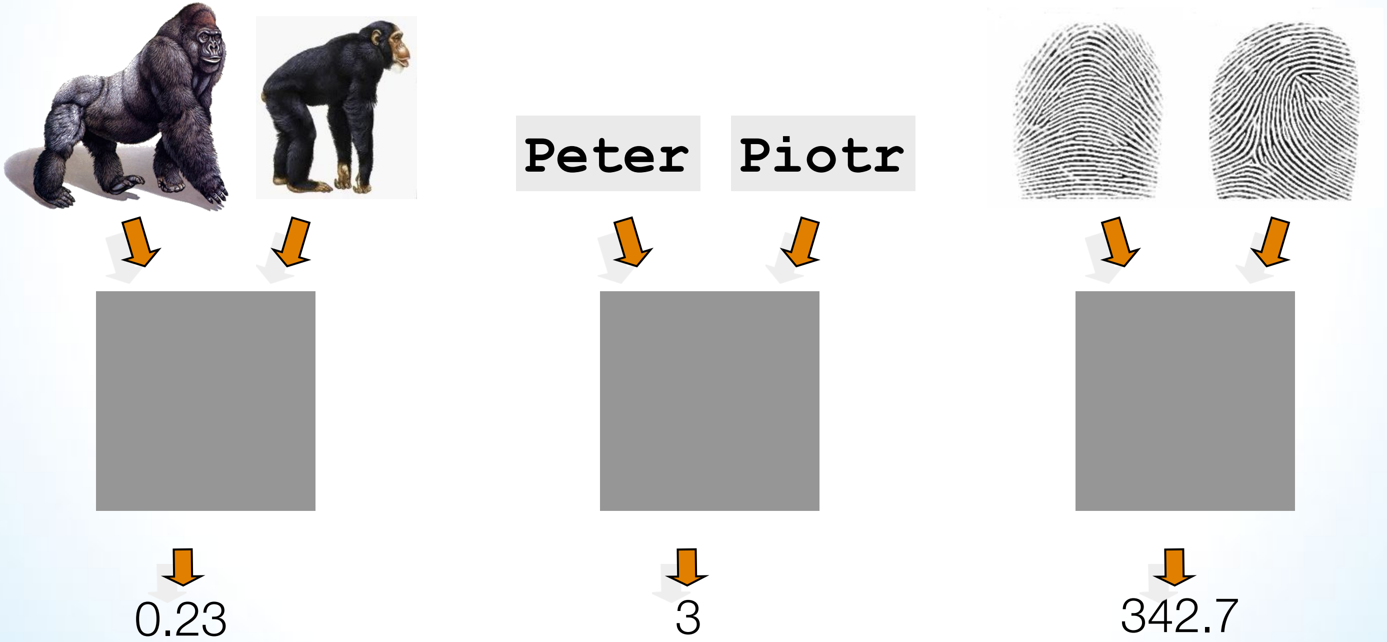
Modeling - K nearest neighbors



- What is the model?
- How to measure distance?
- What is the training error of 1nn?

Defining Distance Measures

Definition: Let O_1 and O_2 be two objects from the universe of possible objects. The distance (dissimilarity) between O_1 and O_2 is a real number denoted by $D(O_1,O_2)$

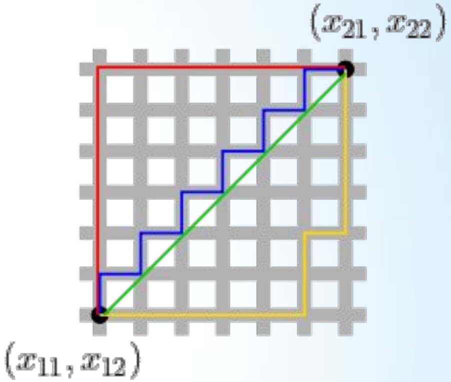


Distance function behavior

- $dis(x,y) \geq 0$
- $dis(x,y) = 0$ iff $x==y$
- $dis(x,y) = dis(y,x)$
- $dis(x, z) \leq dis(x, y) + dis(y, z)$

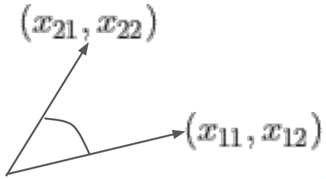
Distance Function

$$L1(X_1, X_2) = ManhattenDistance(\begin{bmatrix} x_{11} \\ x_{1i} \\ x_{1n} \end{bmatrix}, \begin{bmatrix} x_{21} \\ x_{2j} \\ x_{2n} \end{bmatrix}) = \sum_{i=1}^{n} |x_{1i} - x_{2i}|$$



$$L2(X_1, X_2) = EuclideanDistance(\begin{bmatrix} x_{11} \\ x_{1i} \\ x_{1n} \end{bmatrix}, \begin{bmatrix} x_{21} \\ x_{2i} \\ x_{2n} \end{bmatrix}) = \sqrt{\sum_{i=1}^{n} (x_{1i} - x_{2i})}$$

$$CosineSimilarity(X_1, X_2) = \frac{\sum_{i=1}^{n} (x_{1i} * x_{2i})}{\sqrt{\sum_{i=1}^{n} x_{1i}^2} * \sqrt{\sum_{i=1}^{n} x_{2i}^2}}$$



https://medium.com/datadriveninvestor/cosine-similarity-cosine-distance-6571387f9bf8

Using Euclidean Distance

Price Change <0.3	RSI	Sector_ Auto	Label
0	0.4	1	1
1	0.7	0	0
1	0.5	0	0
0	0.3	0	1
1	0.6	0	0
0	0.3	0	1

$$\sqrt{(0-1)^2+(0.4-0.7)^2+(1-0)^2}$$
$$=\sqrt{1+0.09+1}=1.44$$

$$\sqrt{(0-0)^2+(0.4-0.3)^2+(1-0)^2}$$
$$=\sqrt{0+0.01+1}=1.004$$

What do you think about the distance values?

Using Manhattan distance

Price Change <0.3	RSI	Sector_ Auto	Label
0	0.4	1	1
1	0.7	0	0
1	0.5	0	0
0	0.3	0	1
1	0.6	0	0
0	0.3	0	1

$$|0-1|+|0.4-0.7|+|1-0|$$
$$=1+0.3+1=2.3$$

$$|0-0|+|0.4-0.3|+|1-0|$$
$$=0+0.1+1=1.01$$

Any ideas about issues with using absolute ?

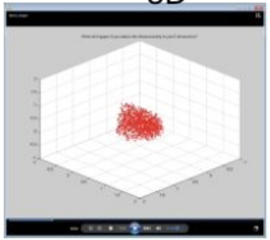
Y-DATA
SCHOOL OF DATA SCIENCE

Yandex

Yandex

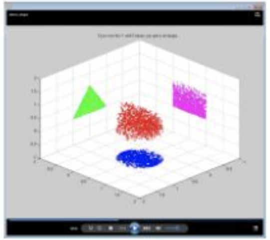
www.cs.gmu.edu/~jessica/DimReducDanger.html eamonn keogh

A cloud of points in 3D

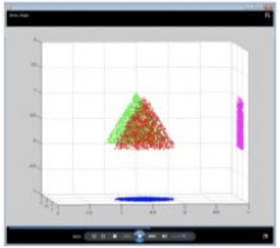


Can be projected into 2D

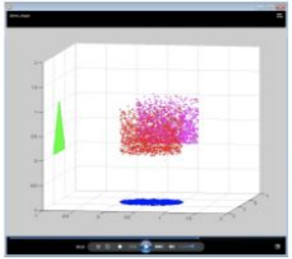
XY or XZ or YZ



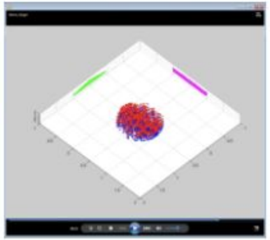
In 2D XZ we see a triangle



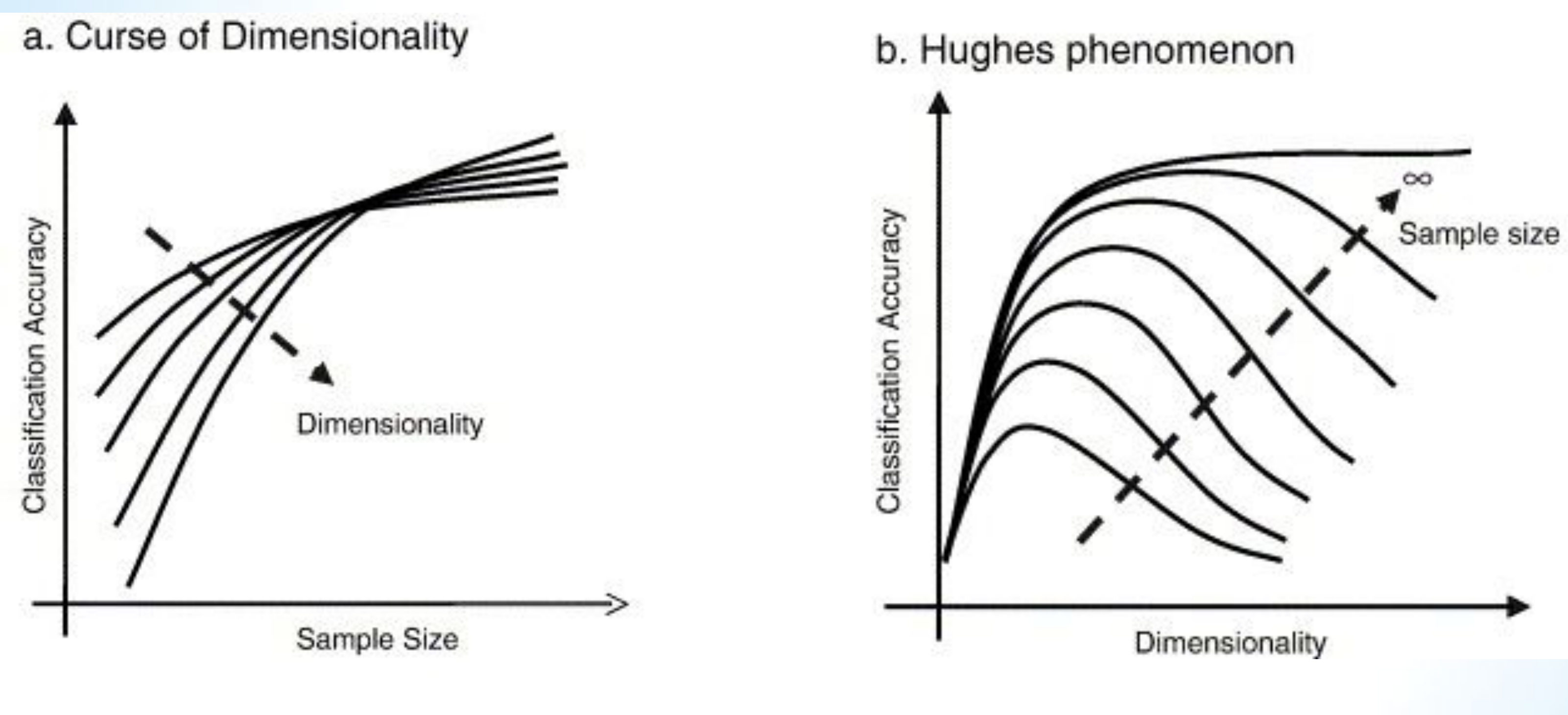
In 2D YZ we see a square



In 2D XY we see a circle



Hughes phenomenon (1968) (Peaking Paradox)



http://37steps.com/2322/hughes-phenomenon/

Y-DATA SCHOOL OF DATA SCIENCE

Yandex

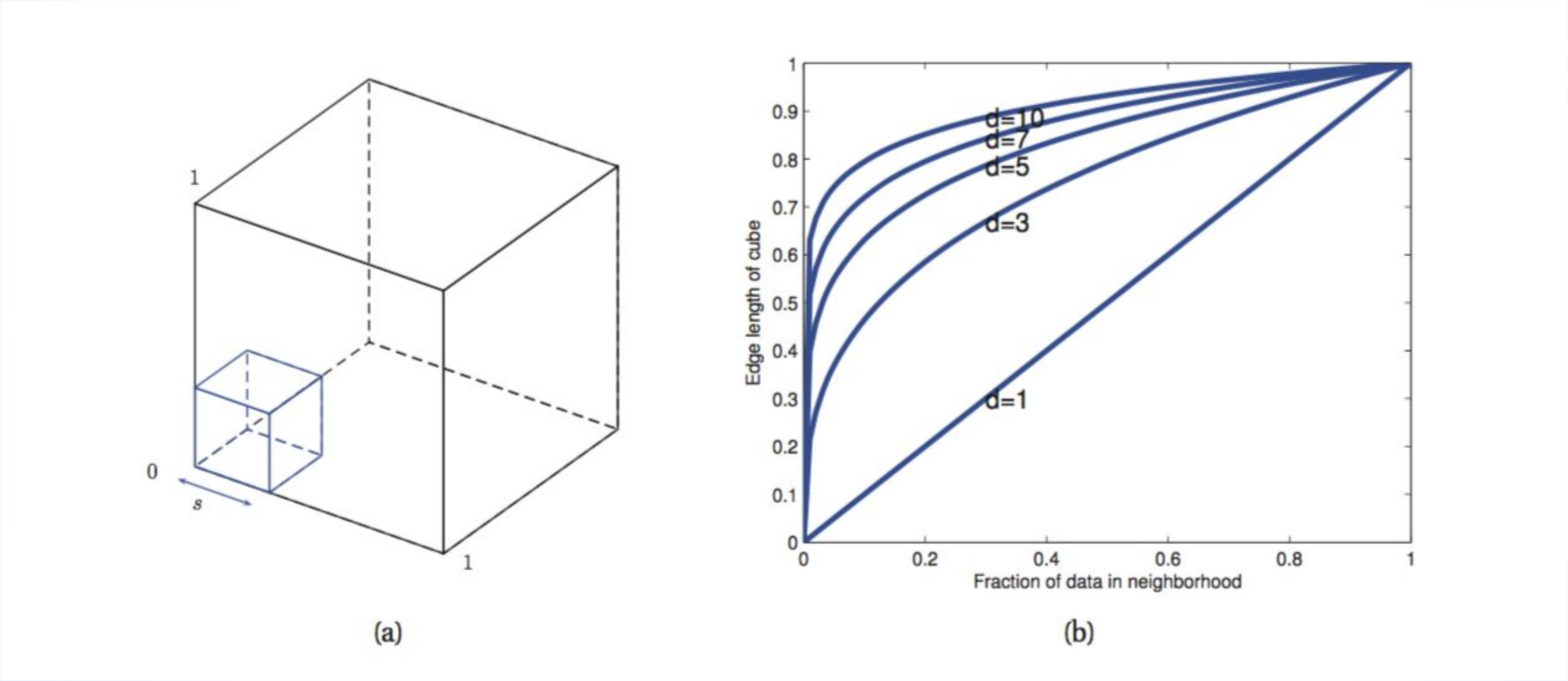
Attribute 1
1
0

Y-DATA SCHOOL OF DATA SCIENCE

Yandex

Attribute 1	Attribute 2	Attribute 3
1	1	1
0	1	1
1	0	1
0	0	1
1	1	0
0	1	0
1	0	0
0	0	0

Why Nearest Neighbours Fails in High Dimensions?

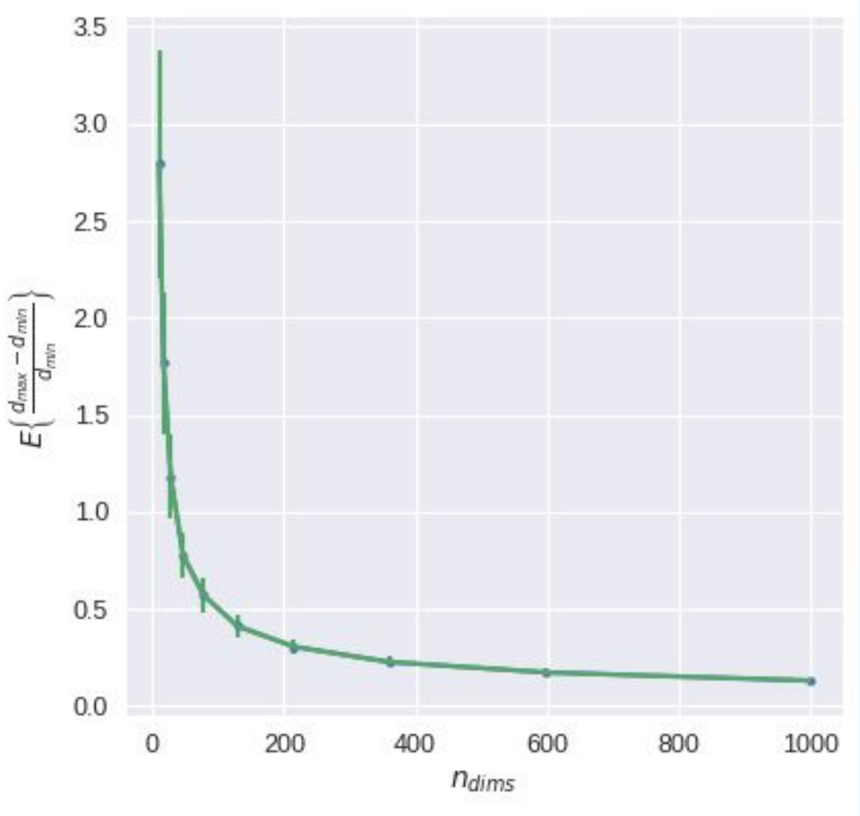


Kevin Murphy's book: Machine Learning - A probabilistic Perspective

Beyer et. al. Theorem

The difference between the maximum and minimum distances to a given query point does not increase as fast as the nearest distance to any point in high dimensional space.

This makes a proximity query meaningless and unstable because there is poor discrimination between the nearest and furthest neighbor.



KNN Best Practices

When to Consider

- Less than 20 attributes per instance
- Lots of training data

Advantages

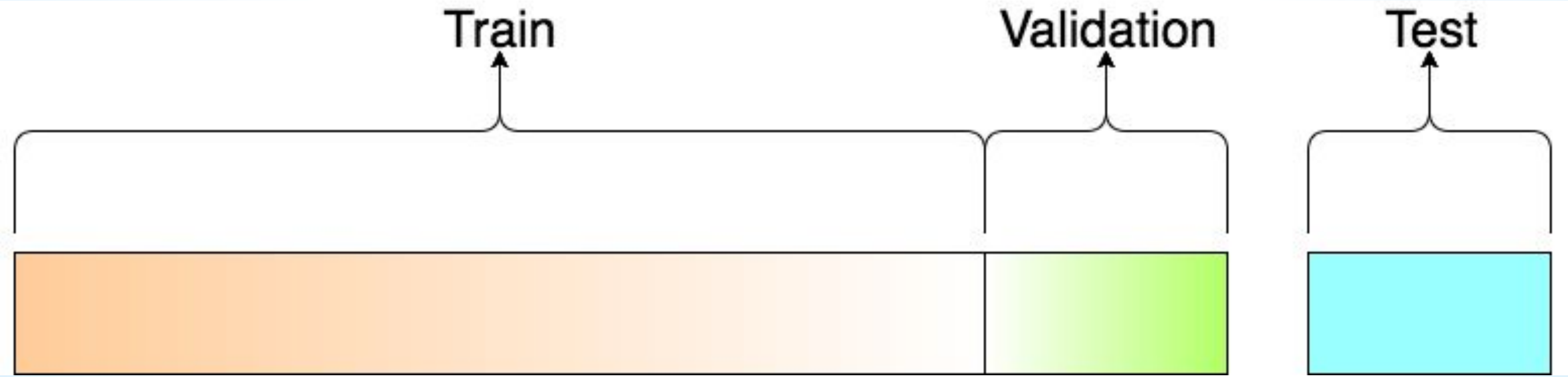
- Training is very fast
- Learn complex target functions
- Do not lose information

Disadvantages

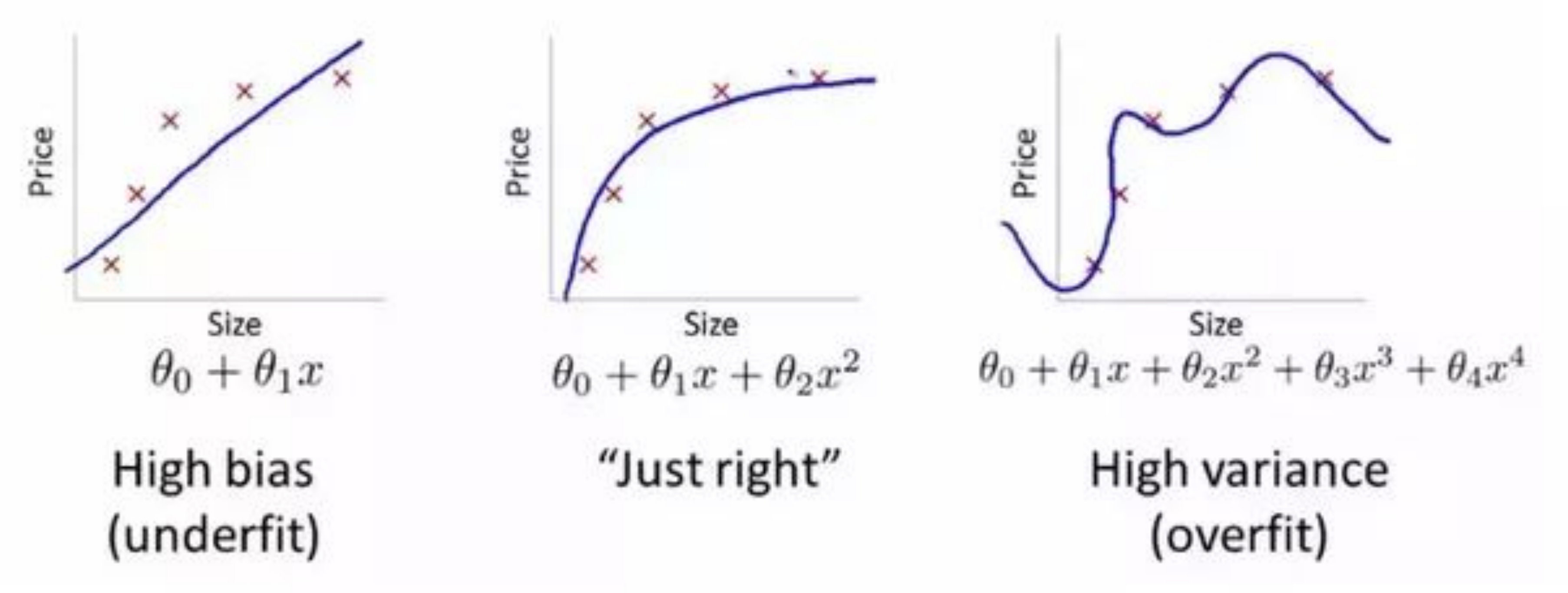
- Slow at query time
- Easily fooled by irrelevant attributes

How to evaluate our model performance?

Estimating Performance

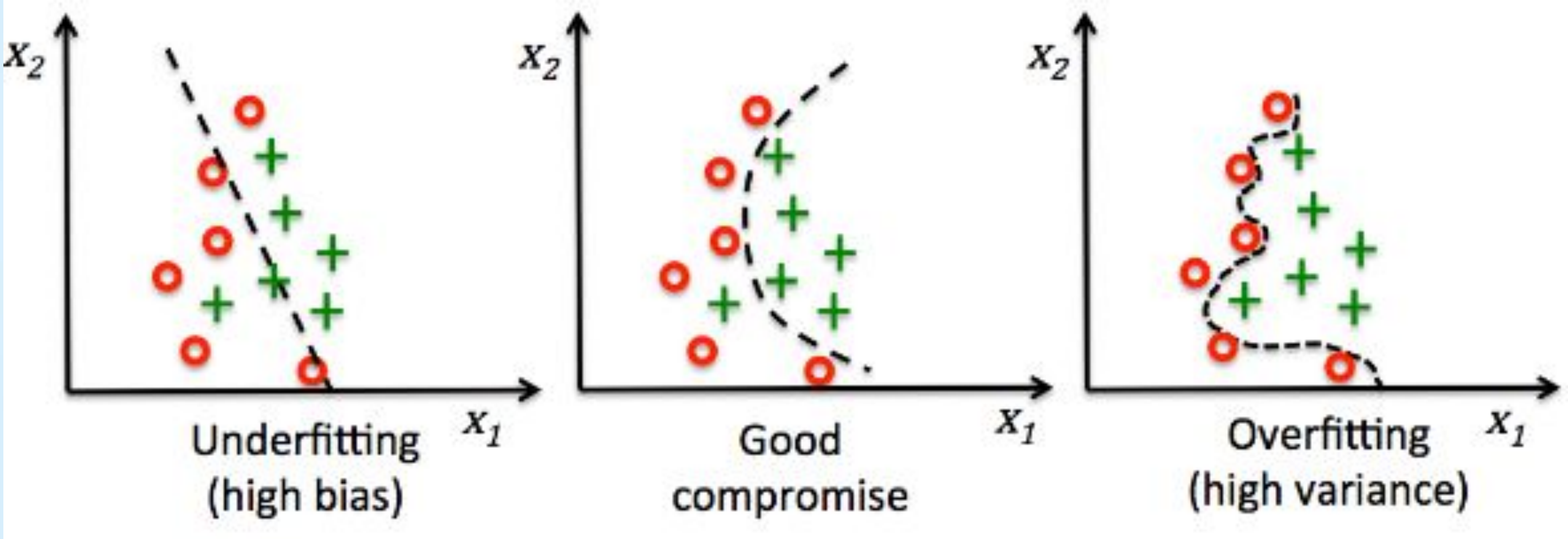


Bias Variance Tradeoff - Regression



https://www.quora.com/What-is-the-best-way-to-explain-the-bias-variance-trade-off-in-layman%E2%80%99s-terms

Bias Variance Tradeoff - Classification



Y-DATA SCHOOL OF DATA SCIENCE

Yandex

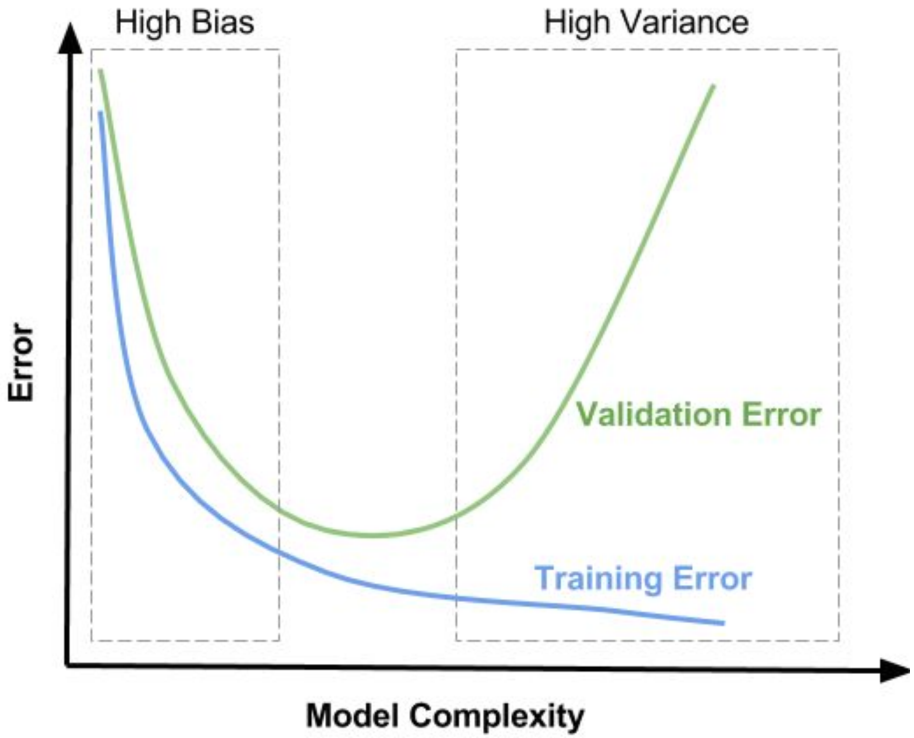
Total Error

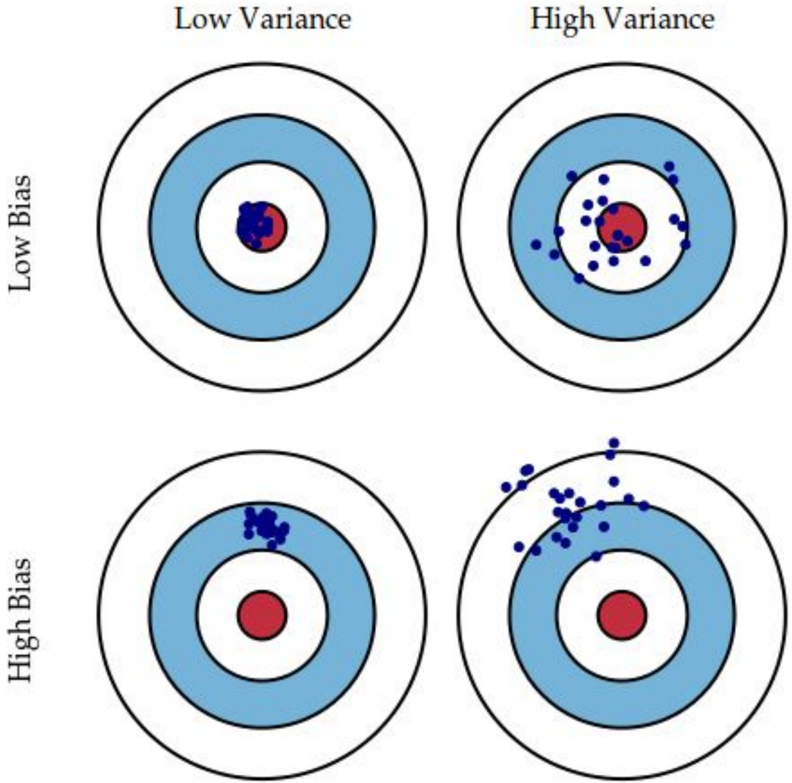
Assume a simple model $y = f(x) + \epsilon$, $\mathrm{E}(\epsilon) = 0$, $\mathrm{Var}(\epsilon) = \sigma_\epsilon^2$,

$$\begin{aligned}
\mathrm{Err}(x_0) &= \mathrm{E}[(y - h(x_0))^2 | \mathrm{X} = x_0] \\
&= \sigma_\epsilon^2 + [\mathrm{E}h(x_0) - f(x_0)]^2 + \mathrm{E}[h(x_0) - \mathrm{E}h(x_0)]^2 \\
&= \sigma_\epsilon^2 + \mathrm{Bias}^2(h(x_0)) + \mathrm{Var}(h(x_0)) \\
&= \text{Irreducible Error} + \text{Bias}^2 + \text{Variance}
\end{aligned}$$

Optional pencil and paper exercise: prove it in details

Bias Variance Tradeoff

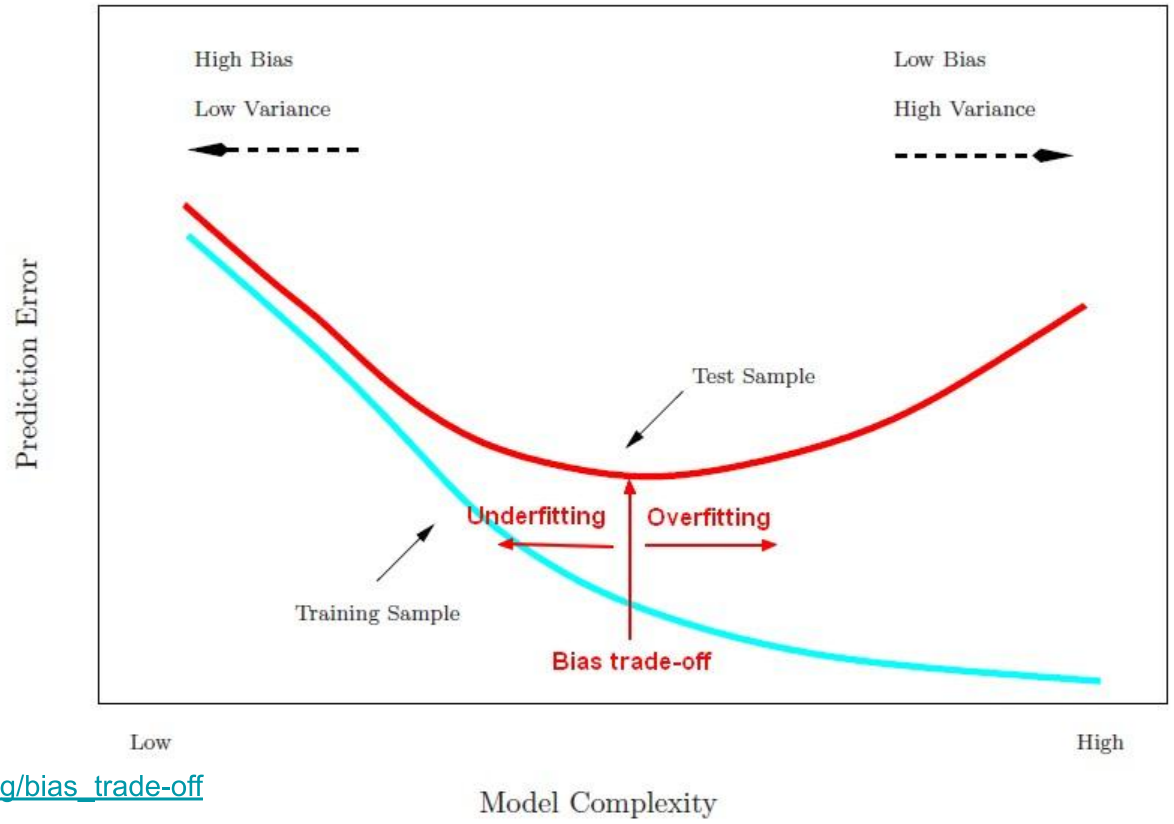




Y-DATA
SCHOOL OF DATA SCIENCE

Yandex

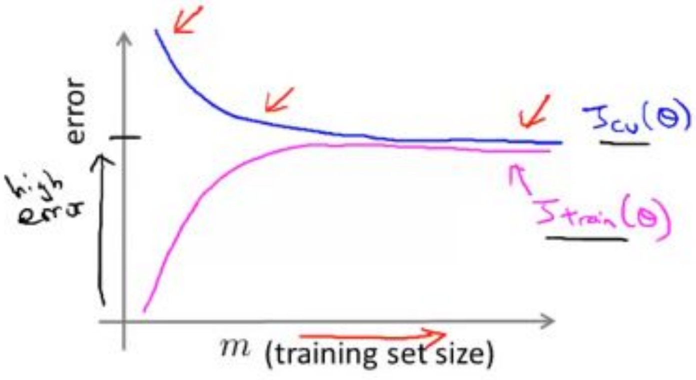
Over and Underfitting



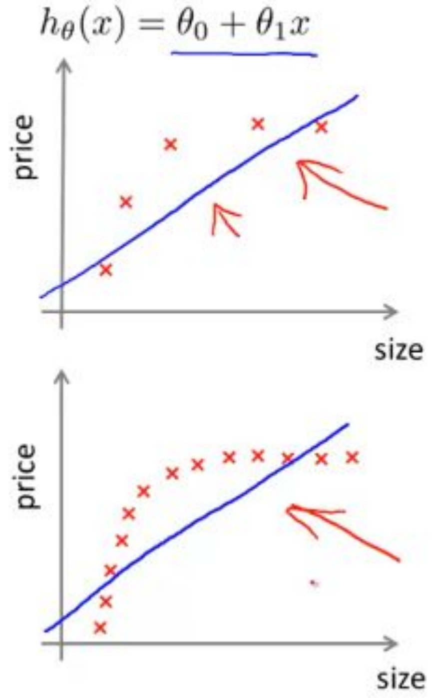
https://gerardnico.com/wiki/data_mining/bias_trade-off

Bias Variance Analysis - Learning Curve

High bias

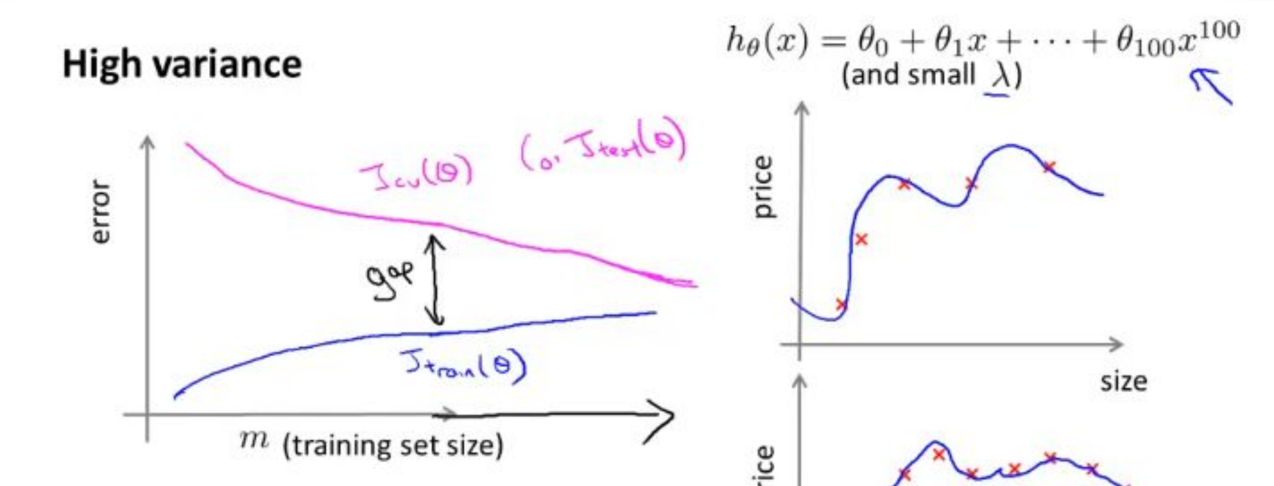


If a learning algorithm is suffering from high bias, getting more training data will not (by itself) help much.



https://medium.com/@aiii/machine-learning-diagnostics-b2256d78d51e

Bias Variance Analysis - Learning Curve



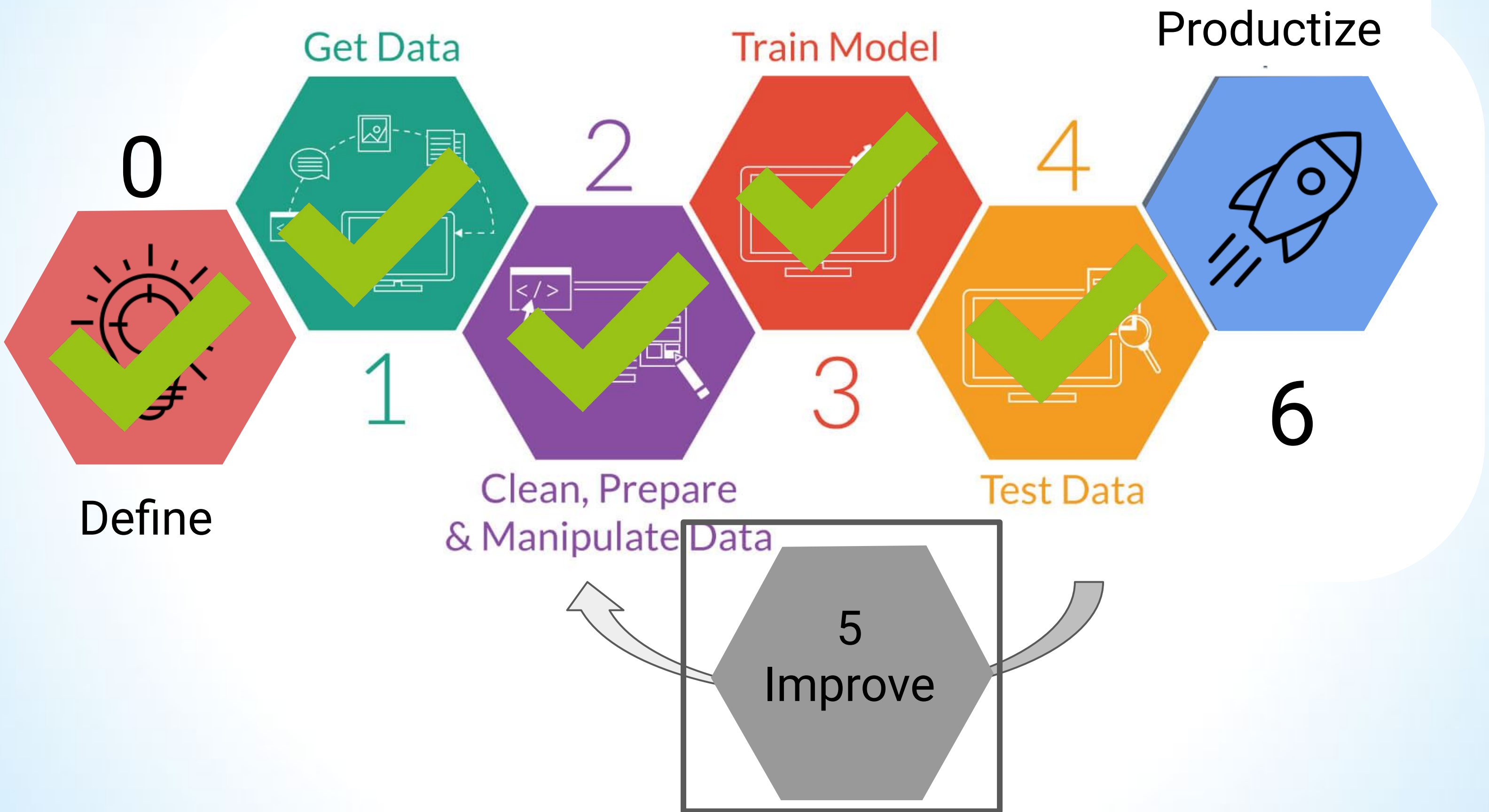
If a learning algorithm is suffering from high variance, getting more training data is likely to help.

https://medium.com/@aiii/machine-learning-diagnostics-b2256d78d51e

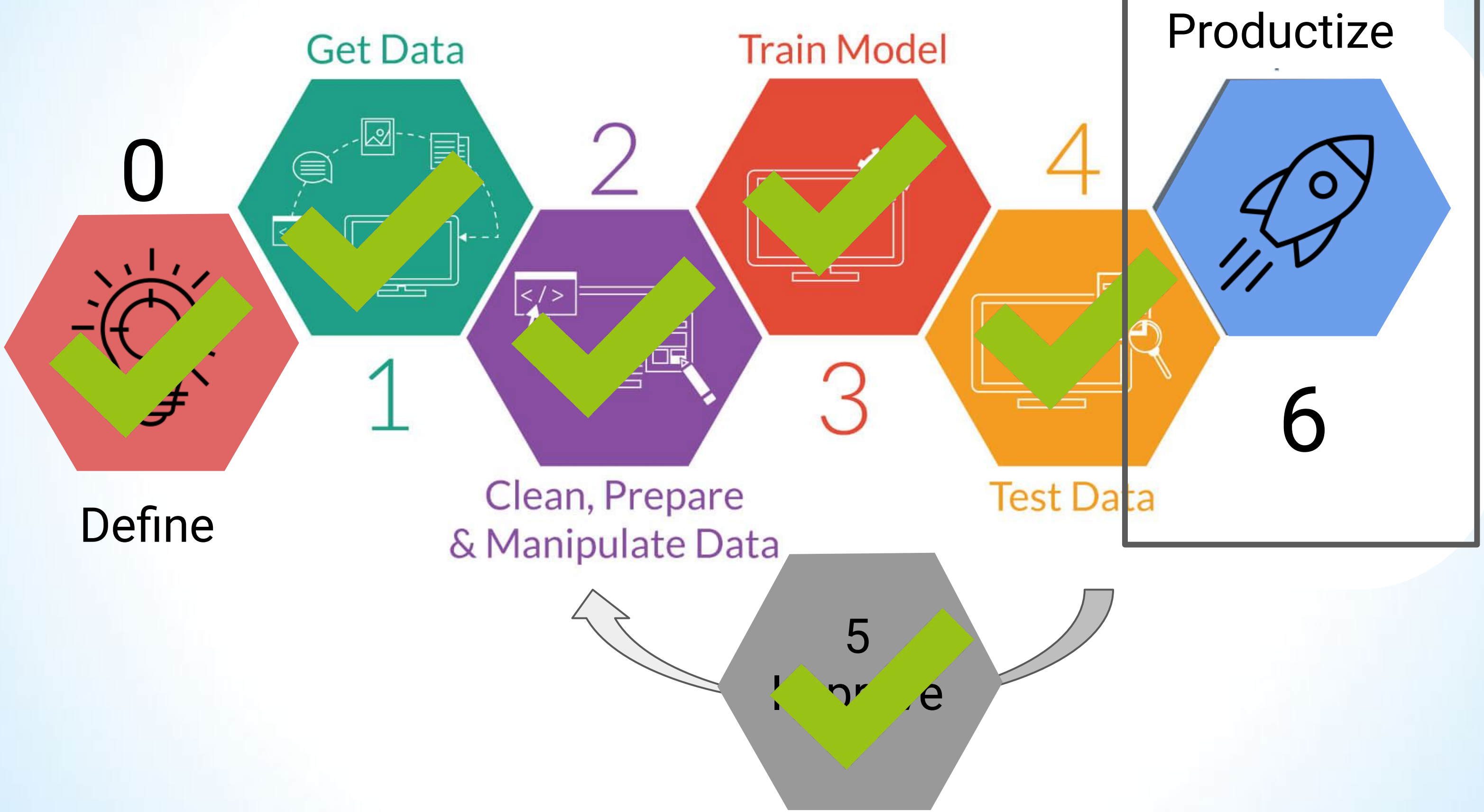
Review Homework

Part 1 - Implement k-Nearest Neighbours (KNN)
Part 2.1 - Learn and evaluate kNN algorithm on artificial data + Analyse the properties of KNN
Part 2.2 - Finding the optimal k
Part 2.3 - Using cross validation

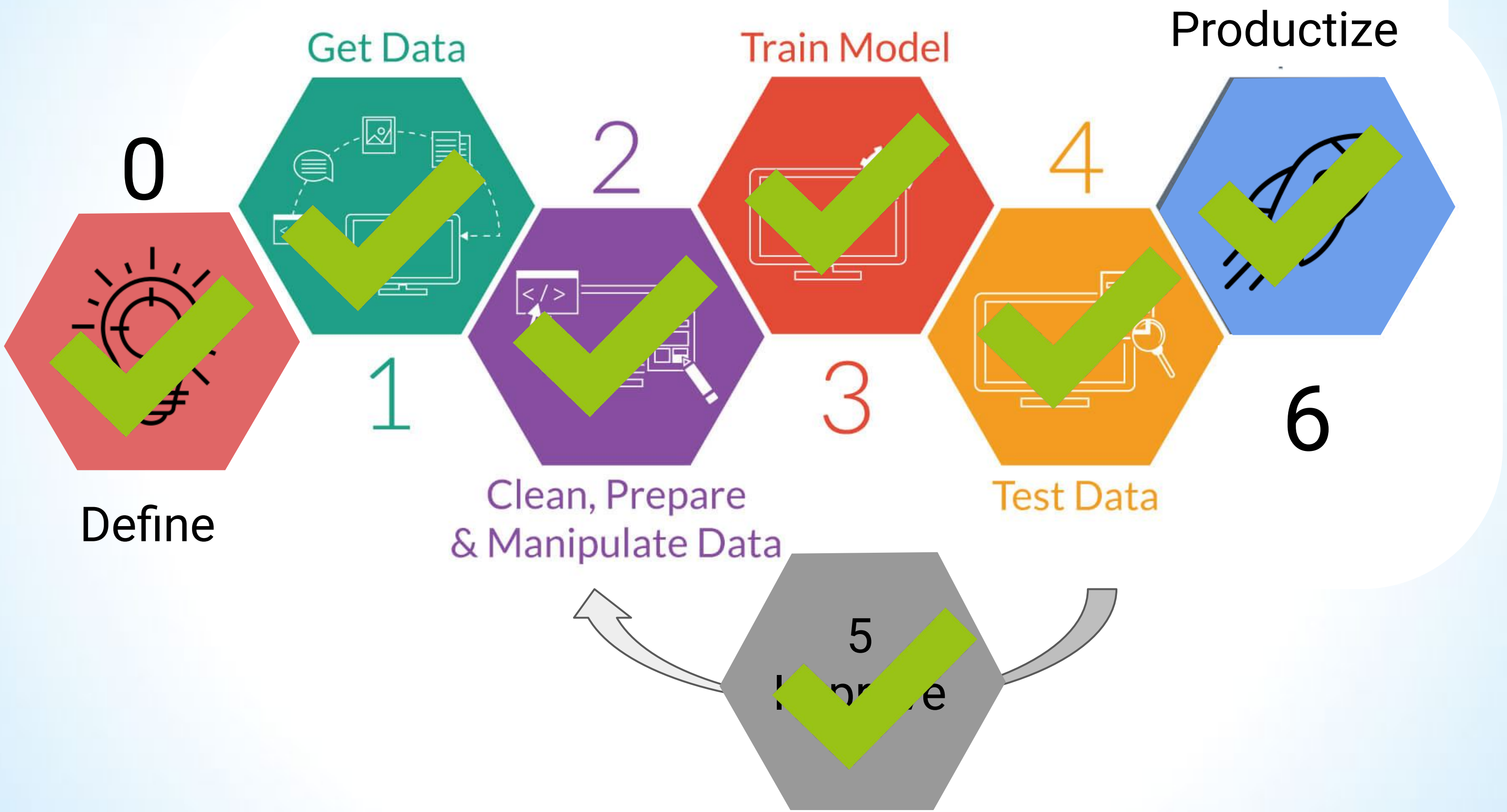
Steps to Predictive Modeling



Steps to Predictive Modeling



Steps to Predictive Modeling



TL,DR

- Traditional programming vs Machine Learning
- Learning vs Generalization
- Data Science is a practical profession
- There are many topics -> understand over memorize
- Build a Baseline model as soon as possible!
- Keep Asking questions
- There is no silver bullet / free lunch theory
-

Y-DATA SCHOOL OF DATA SCIENCE

Yandex

Reading Materials

1. A few useful things to know about machine learning.pdf
2. CIS 419:519 Introduction to Machine Learning.pdf
3. Empirical Risk Minimization.pdf
4. Confusion matrix
5. Cornell KNN intro
6. Introduction to Statistical Learning Theory.pdf
7. On the Surprising Behavior of Distance Metrics in High Dimensional Space.pdf
8. Statistical learning theory - a primer.pdf
9. Statistical Machine Learning- Introduction.pdf
10. 2012b A Geometrical Explanation of Stein Shrinkage.pdf
11. INADMISSIBILITY OF THE USUAL ESTIMATOR FOR THE MEAN OF MULTIVARIATE NORMAL DISTRIBUTION - STEIN.pdf
12. THE USE OF MULTIPLE MEASUREMENTS IN TAXONOMIC PROBLEMS - FISHER - 1936 - Annals of Eugenics - Wiley Online Library.pdf