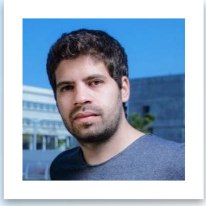
Introduction to Machine Learning

Lior Sidi & Noa Lubin

Y-DATA SCHOOL OF DATA SCIENCE

Yandex

Meet us



Lior Sidi

- Data Science Tech Lead at Wix
- **Ex:** AutoML startup, Consultant, Deutsche Telekom Labs.
- **Domains**: NLP, RecSys, Design, User Centric AI (B2C products).
- BSc and MSc from BGU SISE.
- Scholar / Linkedin

Noa Lubin

- Machine Learning Team Lead at Diagnostic Robotics
- **Ex**: NASA, Amazon, Elbit, IAI
- **Domains**: NLP, health, space
- BSC EE Technion
- MSc CS NLP Bar Ilan
- Linkedin

WIIFM?

“If you invent a breakthrough in AI, so machines can learn, that is worth 10 Microsofts”

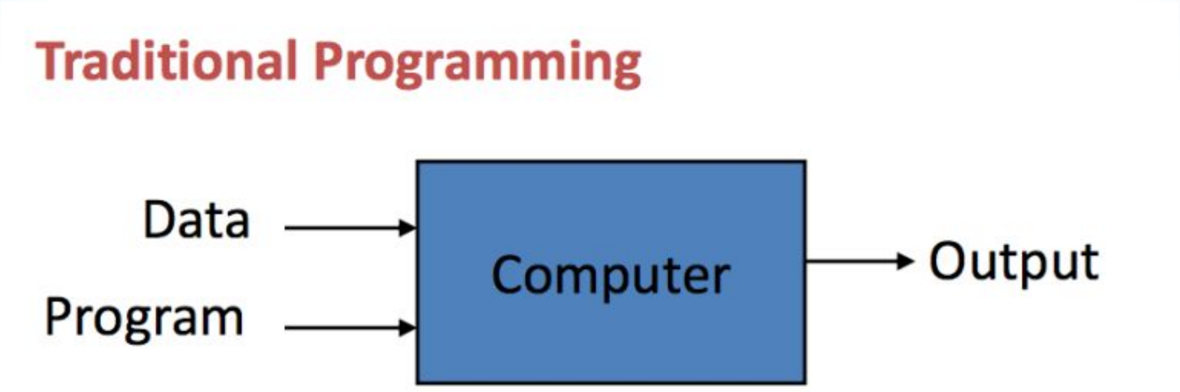
— Bill Gates, Former Chairman, Microsoft. 2004

General Outline

1. What is ML?
2. What are the types of ML Problems?
3. How is ML in Practice?
4. How to estimate Model Performance?
5. How to prepare data for ML?
6. What type of ML algorithms are there?
7. How to improve ML models?

1. What is ML?

How is Machine Learning Relates to Programming



- Machines Follow Instruction
- Humans Learn From Experience

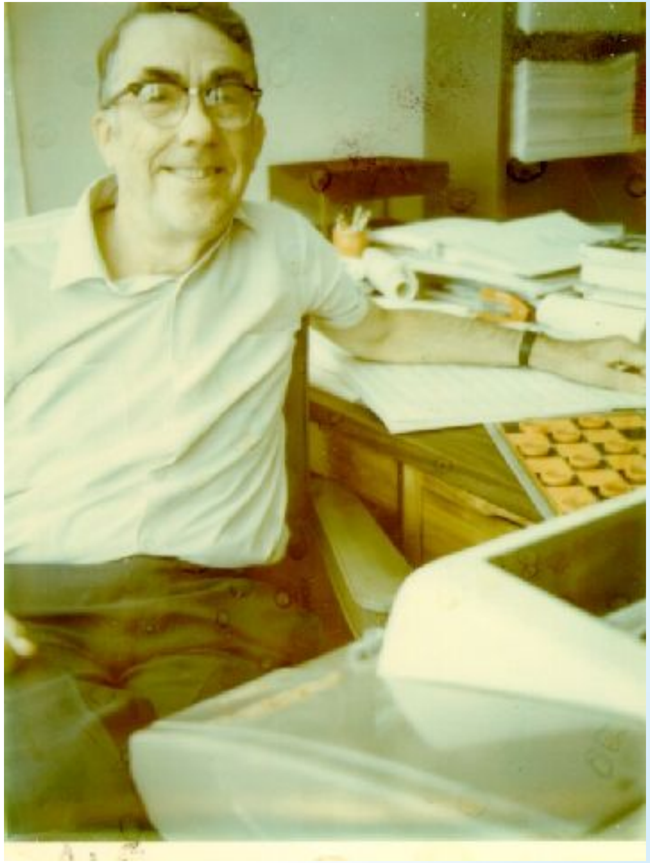
Y-DATA
SCHOOL OF DATA SCIENCE

Yandex

The Arthur Samuel’s Checkers

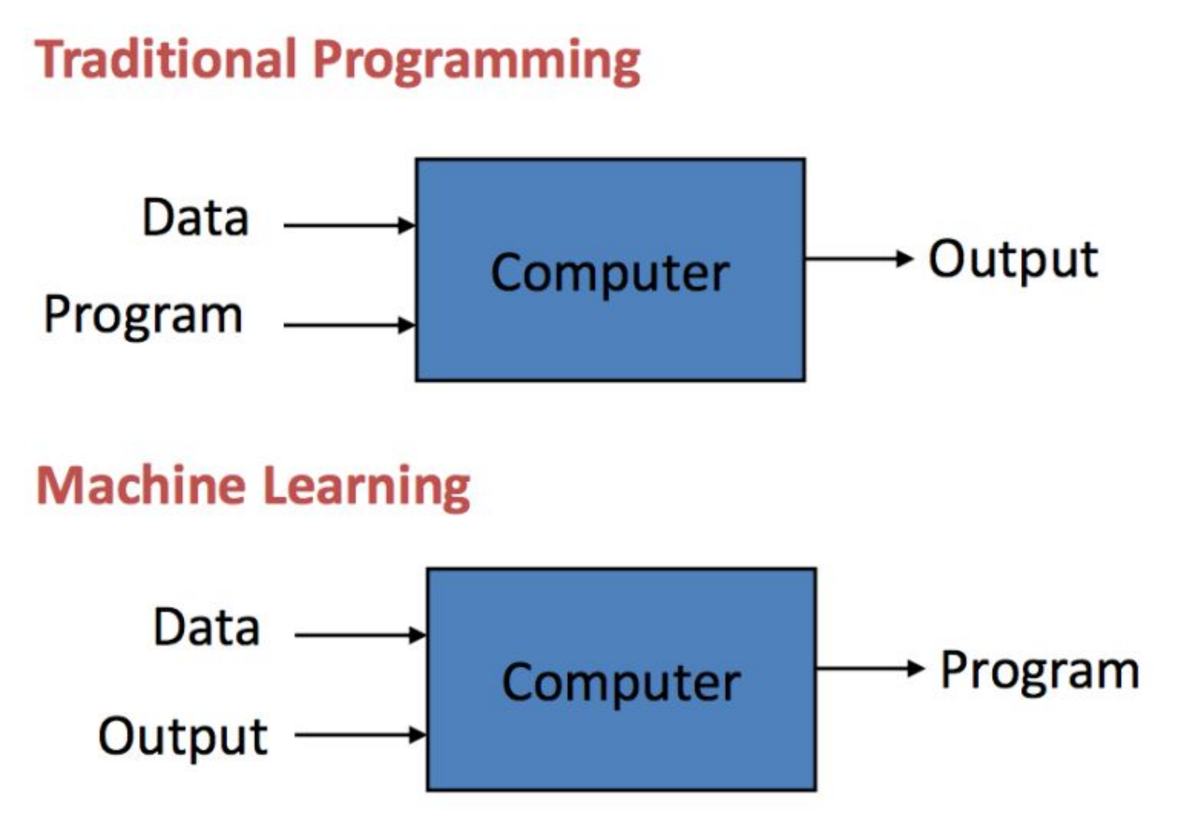
“Machine Learning: Field of study that gives computers the ability to learn without being explicitly programmed.”

— Arthur Samuel (1959)



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

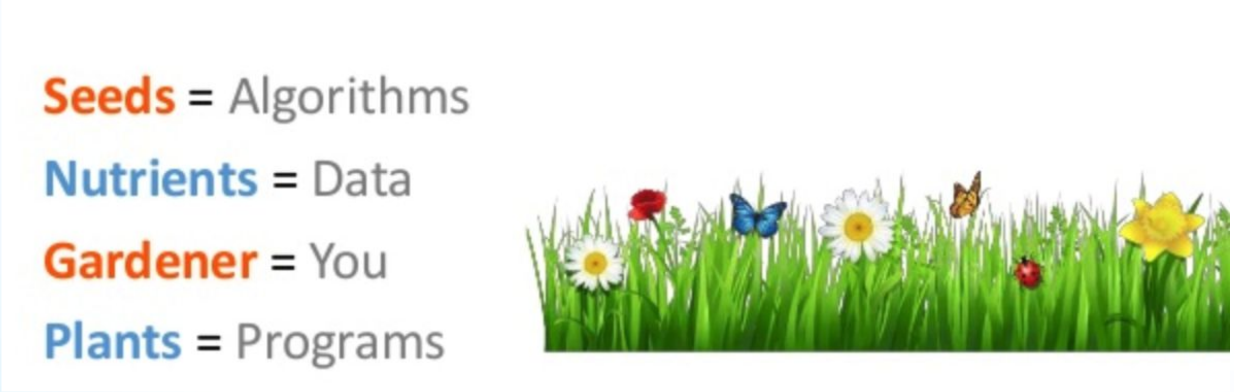
How is Machine Learning Relates to Programming



Agriculture Metaphor of ML

“Machine learning is like farming or gardening. Seeds is the algorithms, nutrients is the data, the gardner is you and plants is the programs.”

[A Study on Machine Learning: Overviews and Applications International conference on Recent Trends in Artificial Intelligence, IOT, Smart Cities & Applications (ICAISC-2020)]



Y-DATA SCHOOL OF DATA SCIENCE

Yandex

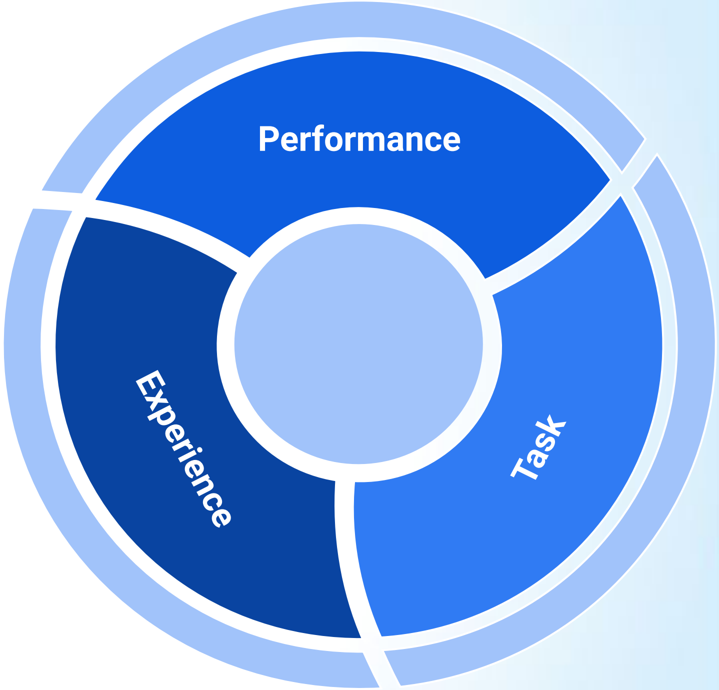
The PTE View on ML

Machine Learning is the study of algorithms that

- improve their **performance** (P)
- at some **task** (T)
- with **experience** (E).

A well-defined learning task is given by <P, T, E>.

e.g., predict house pricing (T), according to last year's data (E), minimizing the square error (P).



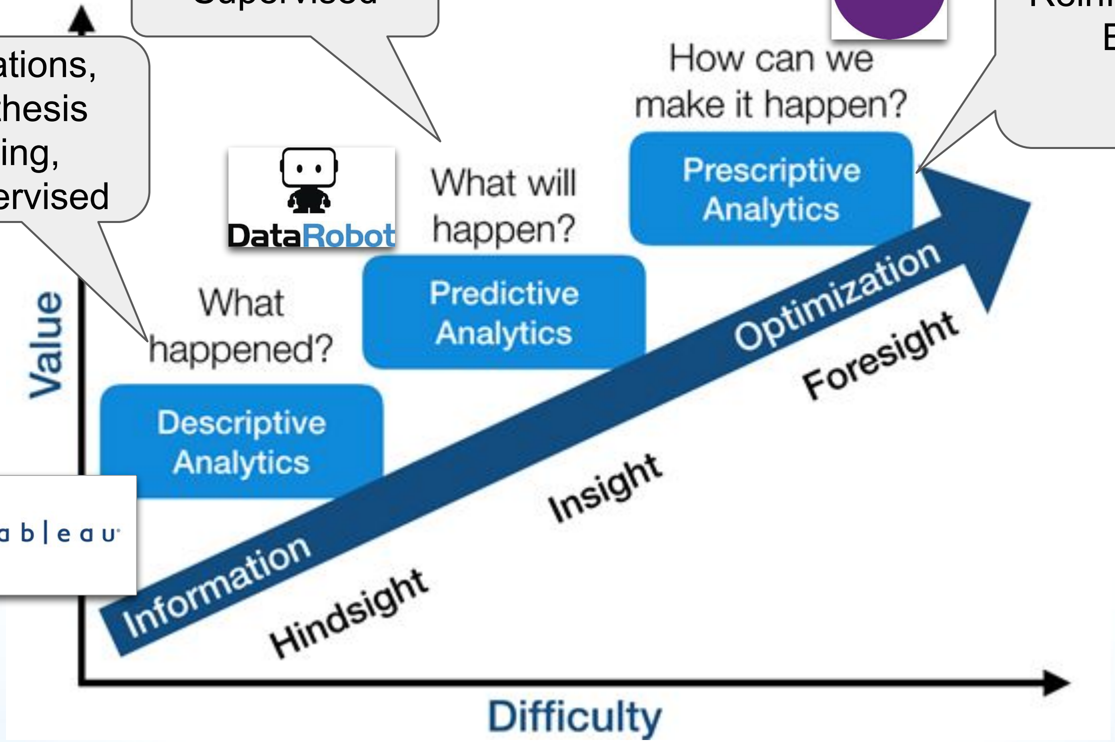
https://courses.cs.washington.edu/courses/cse473/04au/lectures/14-ml-intro.pdf

Descriptive vs. Predictive vs. Prescriptive

Estimations, Hypothesis Testing, Unsupervised

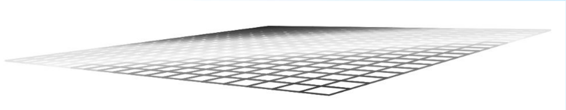
Supervised

Reinforcement Learning, Bandits, Causal Inference

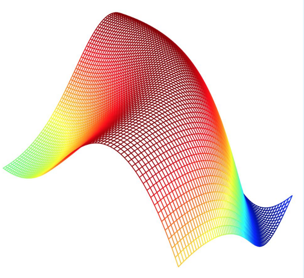


Key Elements of Machine Learning

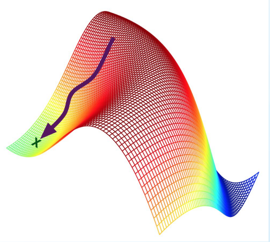
1. **Representation**: how to represent knowledge.
   The Model.



2. **Evaluation**: the way to evaluate candidate representations (programs/hypotheses).



3. **Optimization**: how to find the optimal representation in the evaluation terms.



[Domingos, Pedro. "A few useful things to know about machine learning." Communications of the ACM 55.10 (2012): 78-87.]

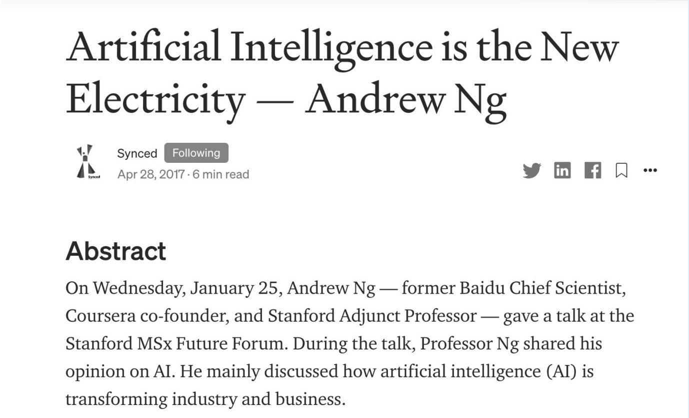
Y-DATA
SCHOOL OF DATA SCIENCE

Yandex

Application of Machine Learning

Sample applications of machine learning:

- The
- List
- Is
- Just
- Too
- Long



Artificial Intelligence is the New Electricity — Andrew Ng

Synced Following
Apr 28, 2017 · 6 min read

Abstract

On Wednesday, January 25, Andrew Ng — former Baidu Chief Scientist, Coursera co-founder, and Stanford Adjunct Professor — gave a talk at the Stanford MSx Future Forum. During the talk, Professor Ng shared his opinion on AI. He mainly discussed how artificial intelligence (AI) is transforming industry and business.

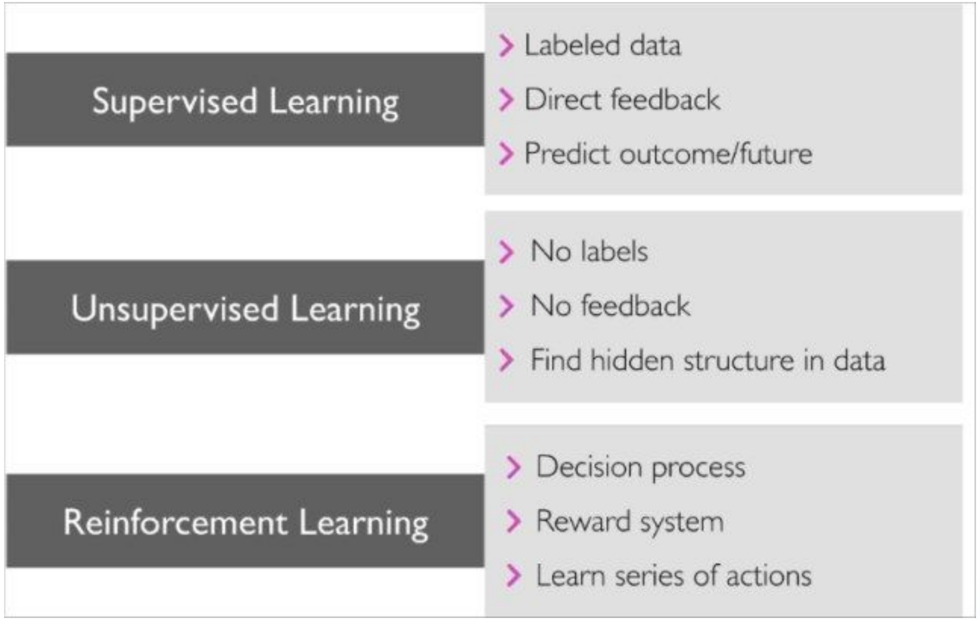
[Machine Learning is the new Electricity]

2. What are the types of ML Problems?

Y-DATA SCHOOL OF DATA SCIENCE

Yandex

Types of ML

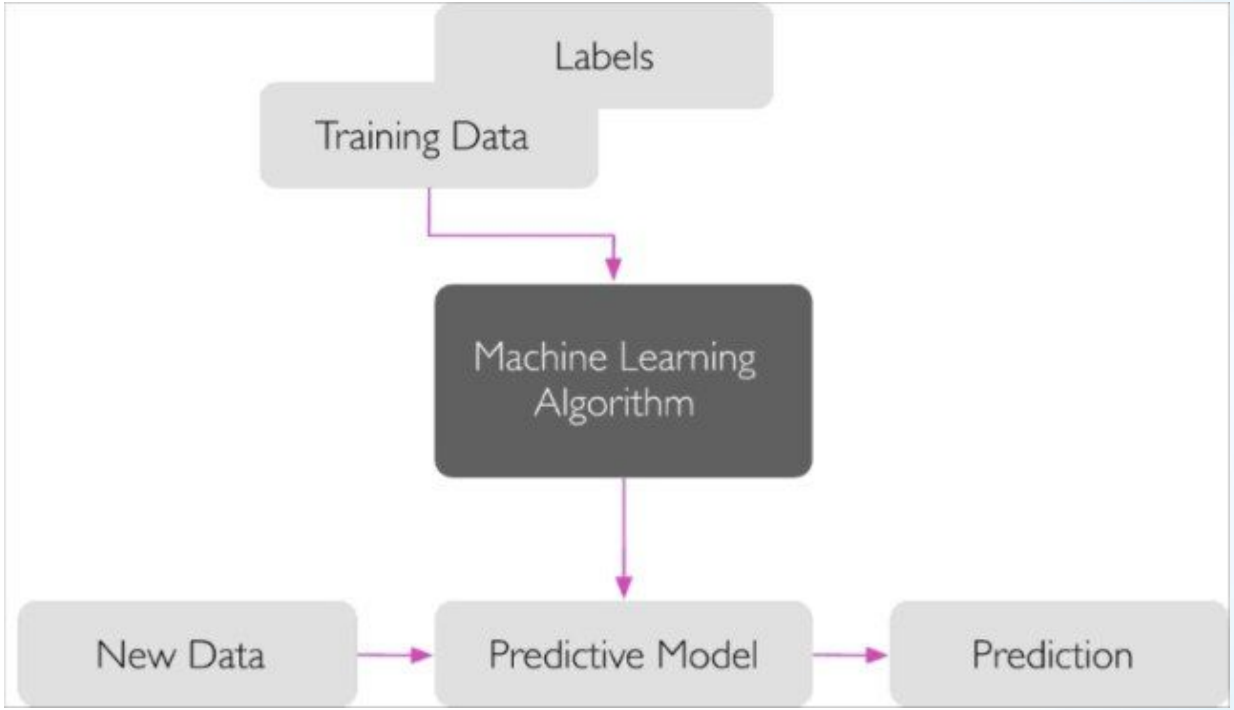


[List of Machine Learning Algorithms] [3 different types of machine learning]

Additional Types of Learning

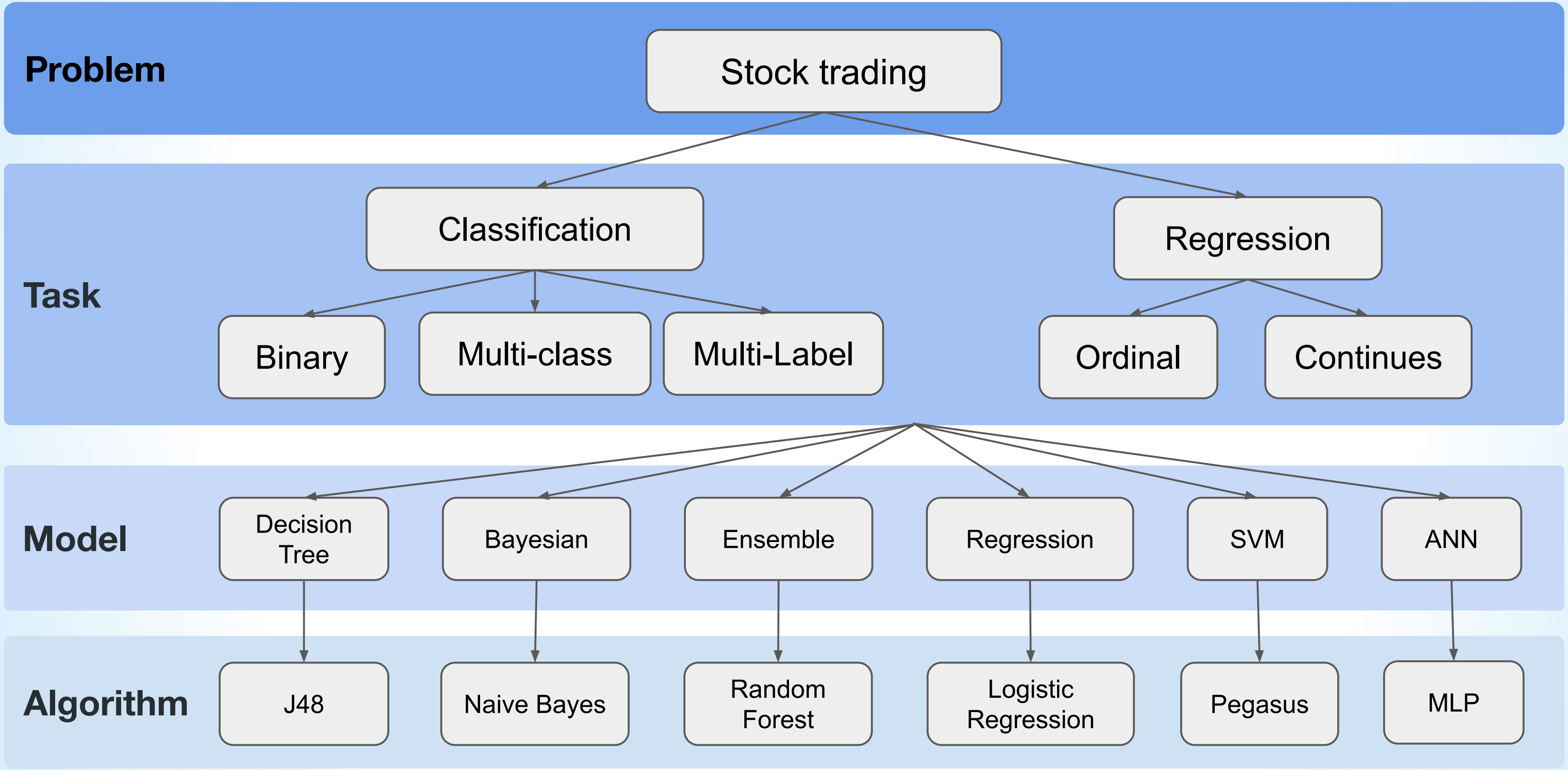
- **Semi-Supervised learning**
  Training data includes a few desired outputs.
- **Causal Inference Learning**
  Discover and estimate the causal relationship between variables

Supervised

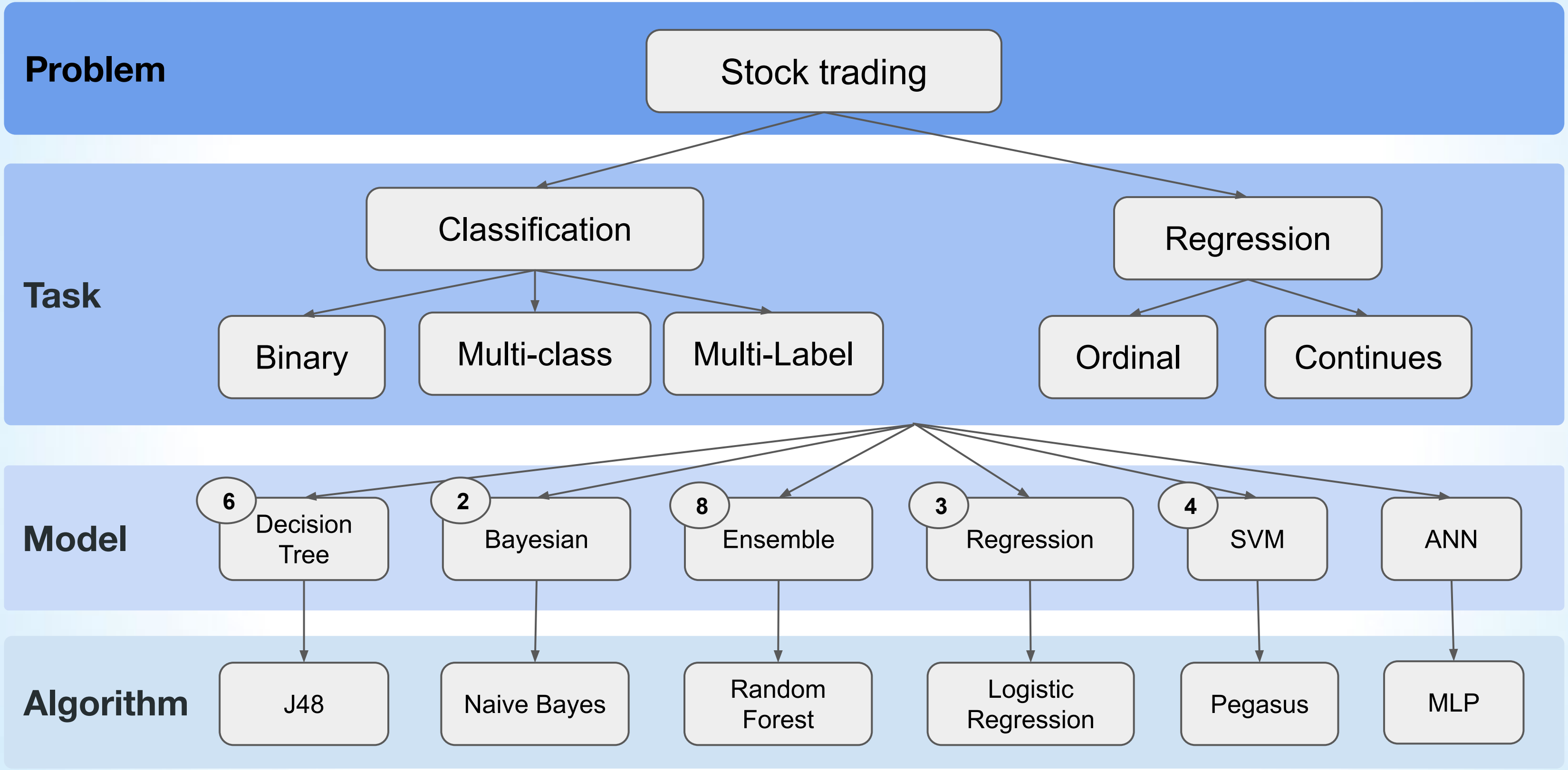


https://www.quora.com/What-is-supervised-learning

Supervised tasks



Supervised tasks



Y-DATA SCHOOL OF DATA SCIENCE

Yandex

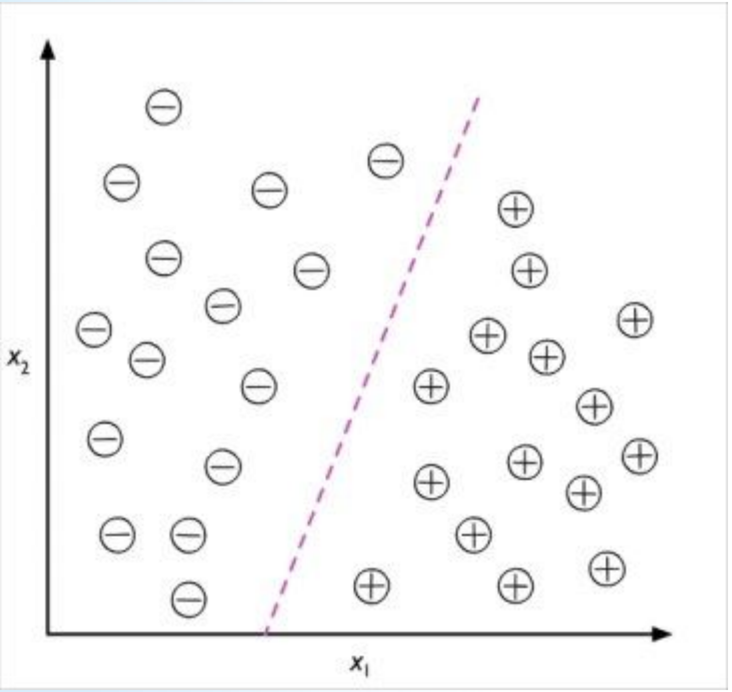
Syllabus

week	Topics
1 **Intro**	Introduction, definition of ML, types of ML, KNN
2 **Naive Bayes**	-Refresher on the Bayes theorem -Naive Bayes theory -Spam detection use case including preprocessing phase
3 **Linear Regression**	- Ordinary Linear Regression - Analytical solution and Gradient Descent solution - Common Pitfalls - weight is not importance - L2 regularization on OLS - L1 regularization on OLS
4 **Logistic Regression**	- Regression vs Classification - Classifier performance measurement - Binary Logistic Regression - Multi-Class classification

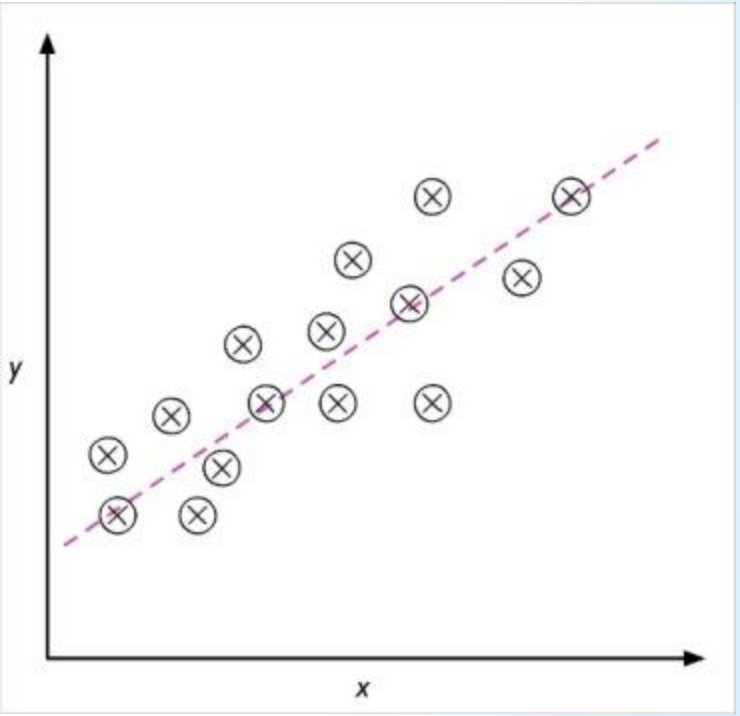
week	Topics
5 **SVM**	- Support vector machines
6 **Decision trees**	- Decision Tree as a Greedy Method - Optimization Criteria: Gini & Entropy+ L2 - Depth, Leaves and other Hyper Parameters
7 **End2End ML**	- Bias-variance tradeoff, validation set, cross-validation - Overfitting and underfitting - Regularization and hyperparameter tuning - Features selection
8-9 **Ensemble**	Intro to Ensemble Methods - Aggregation - Bagging - Stacking - Boosting - Gradient Boosting

Regression Vs Classification

Classification

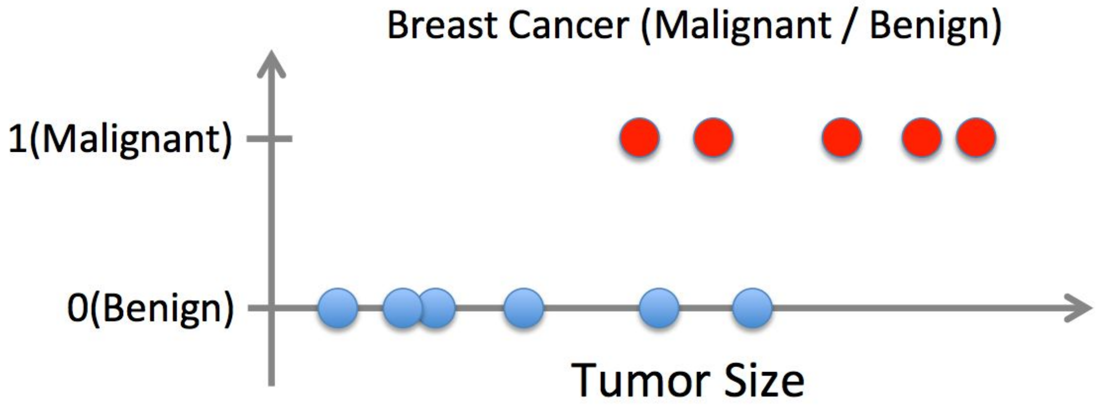


Regression



Supervised Classification

- Given $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
- Learn a function $f(x)$ to predict y given x
  - y is categorical == classification

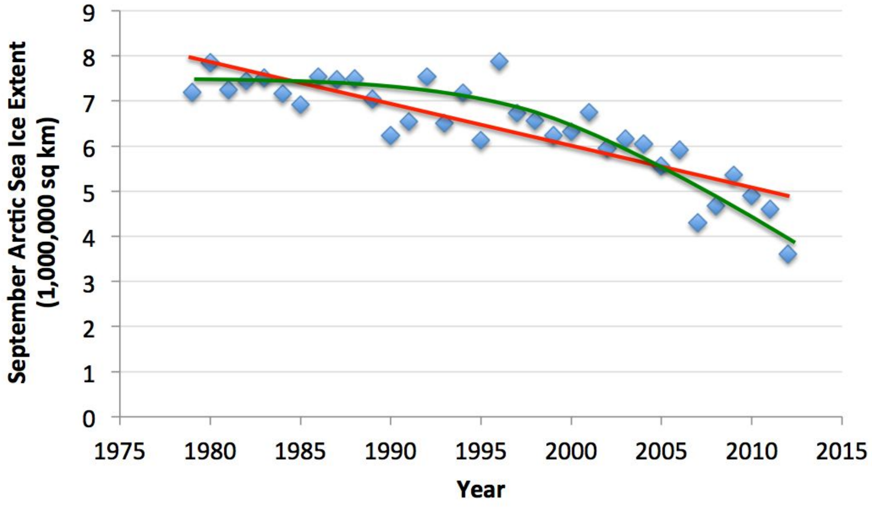


Y-DATA SCHOOL OF DATA SCIENCE

Yandex

Supervised Regression

- Given $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
- Learn a function $f(x)$ to predict y given x
  - y is real-valued == regression



Data from G. Witt. Journal of Statistics Education, Volume 21, Number 1 (2013)

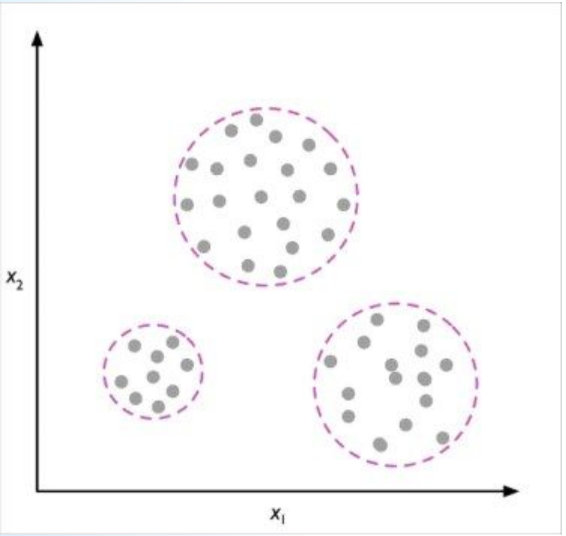
Type of Supervised Learning

- Output Data Type
  - Discrete - Classification
  - Continuous - Regression
  - Structured - Structure Prediction (e.g. sentence tagging)
- Classification Types:
  - Multi-class
  - Multi-label
- Regression Types:
  - Dimensions - Uni/Multivariate
  - Monotonic - Isotonic Regression
  - Loss Function - MSE, MAE
  - Regularization Type - Lasso, Ridge, etc
  - Percentile - Quantile Regression

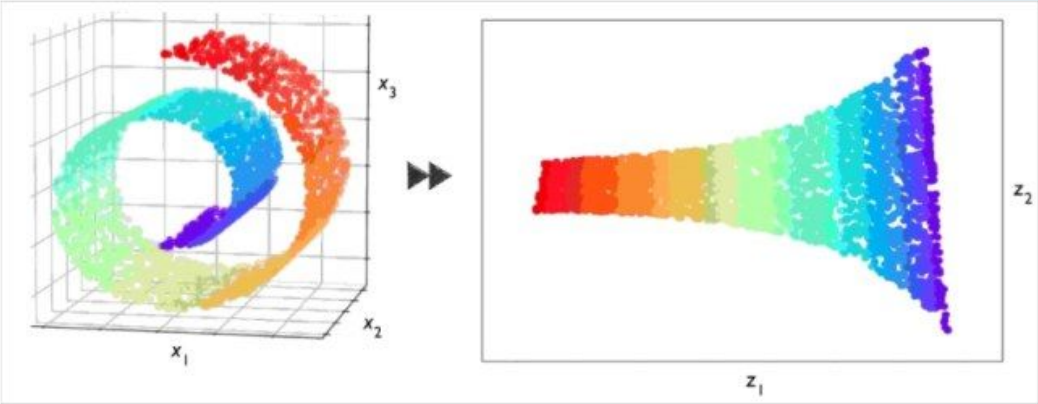
This	DT
is	VBZ
a	DT
tagged	JJ
sentence	NN
.	.

Unsupervised Learning

Clustering

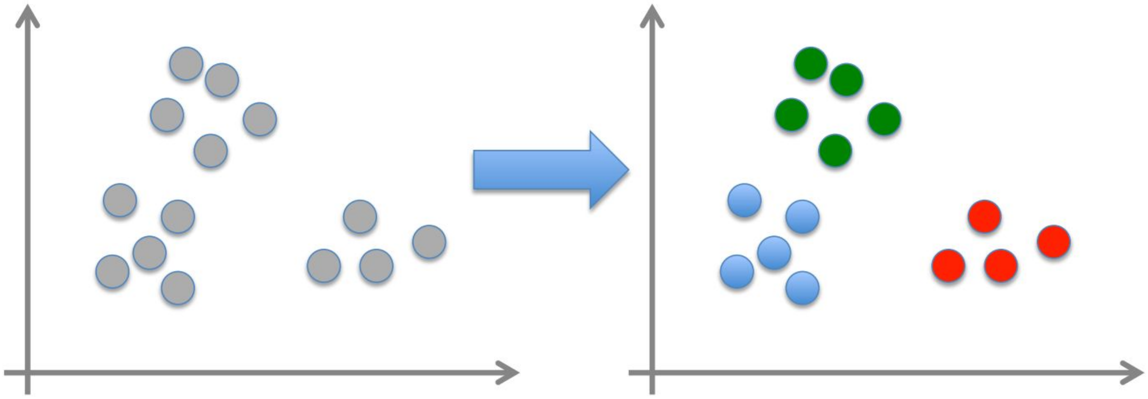


Dimension Reduction

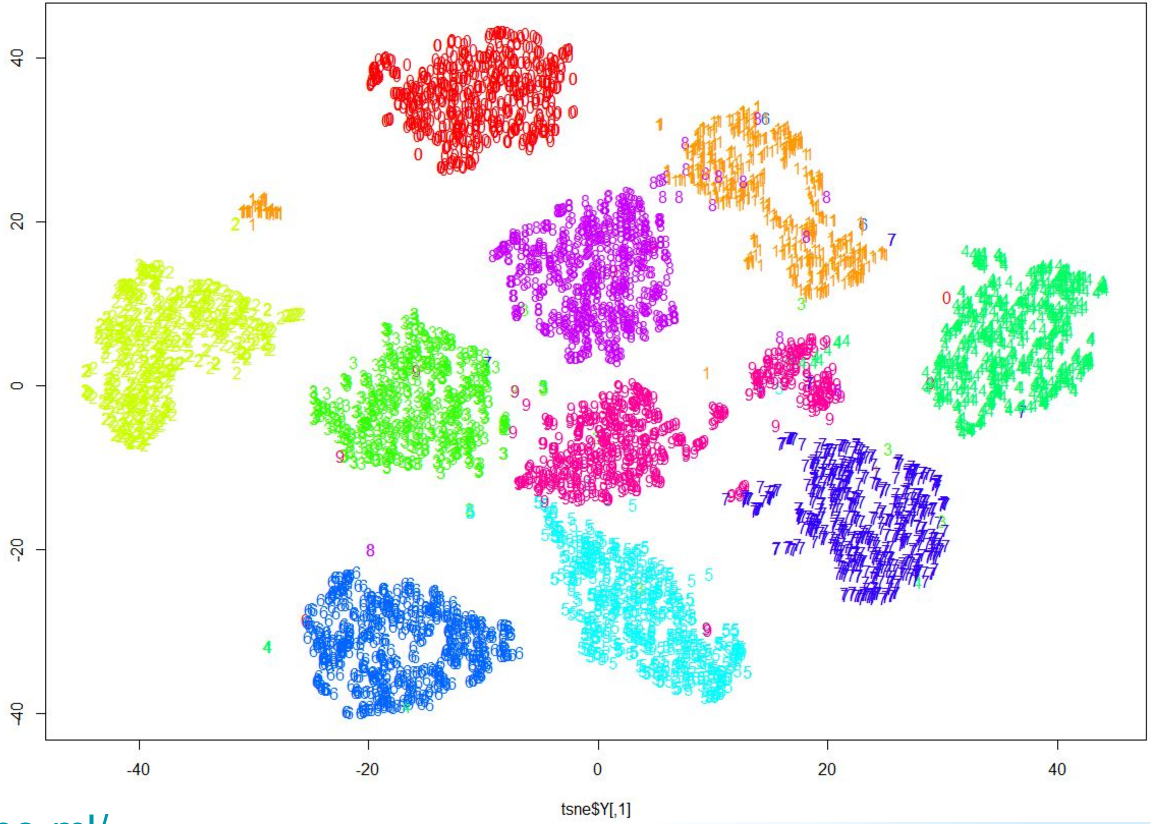


Unsupervised

- Given $x_1, x_2, \ldots, x_n$ (without labels)
- Output hidden structure behind the x's
  - E.g., clustering



Unsupervised Example

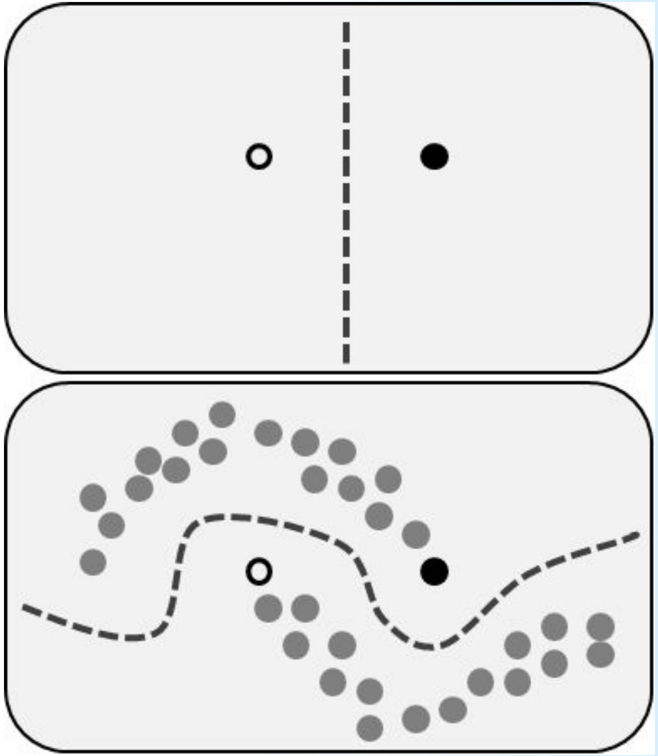


https://mark-borg.github.io/blog/2016/tsne-ml/

Semi-supervised Learning Example

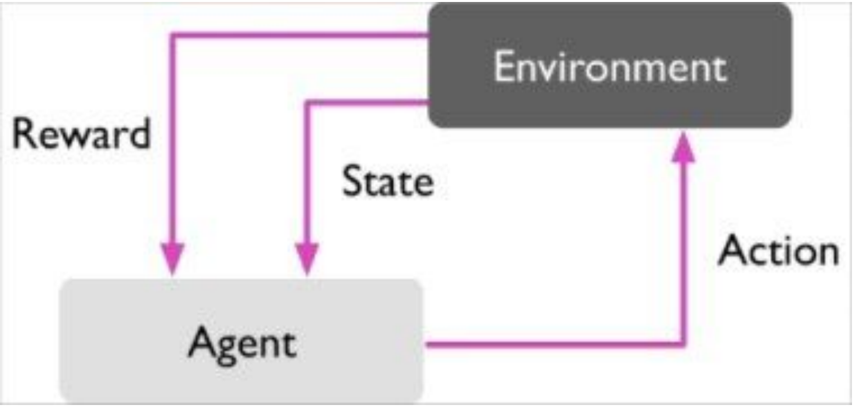
Antivirus:

- Some files are known to be safe
- Some are known to be malicious
- Most are unknown

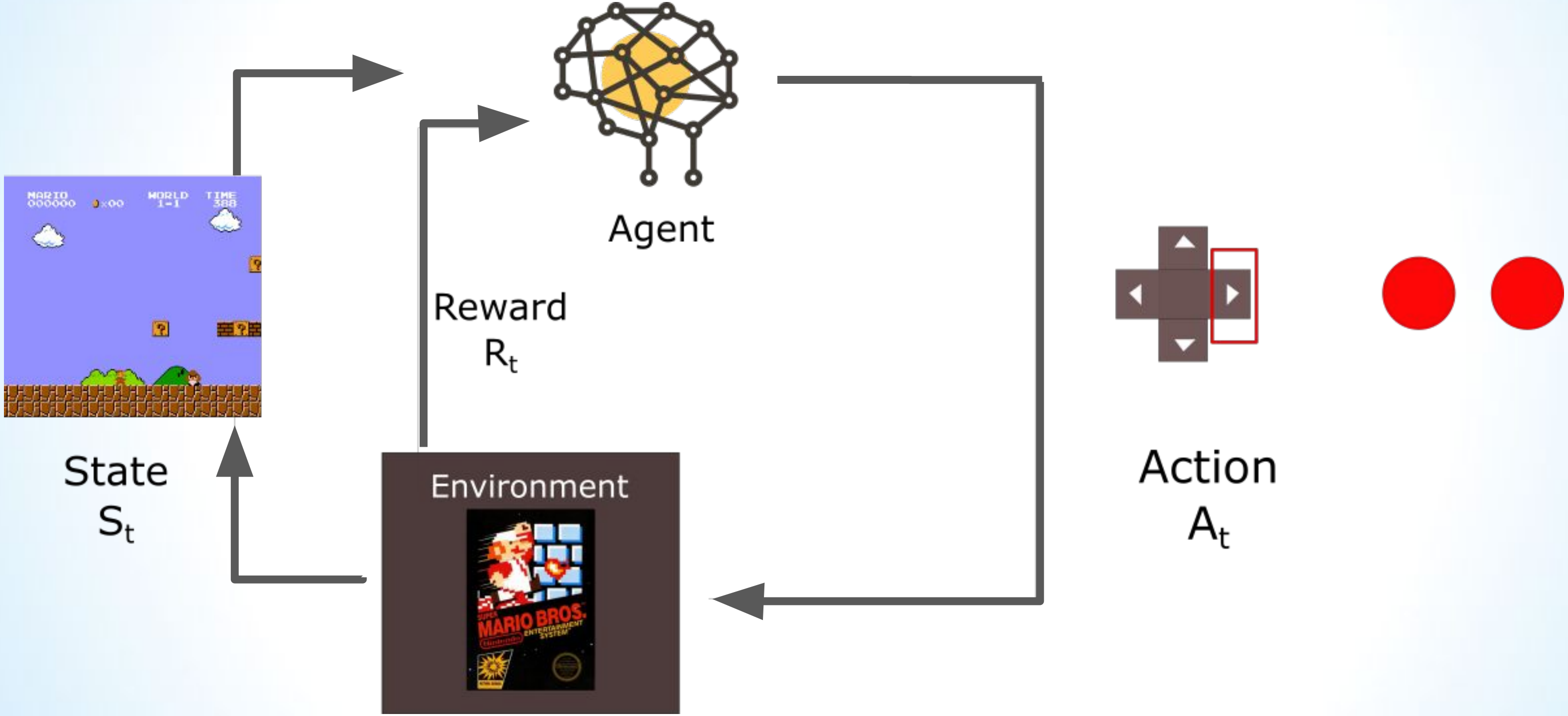


https://en.wikipedia.org/wiki/Semi-supervised_learning

Reinforcement Learning

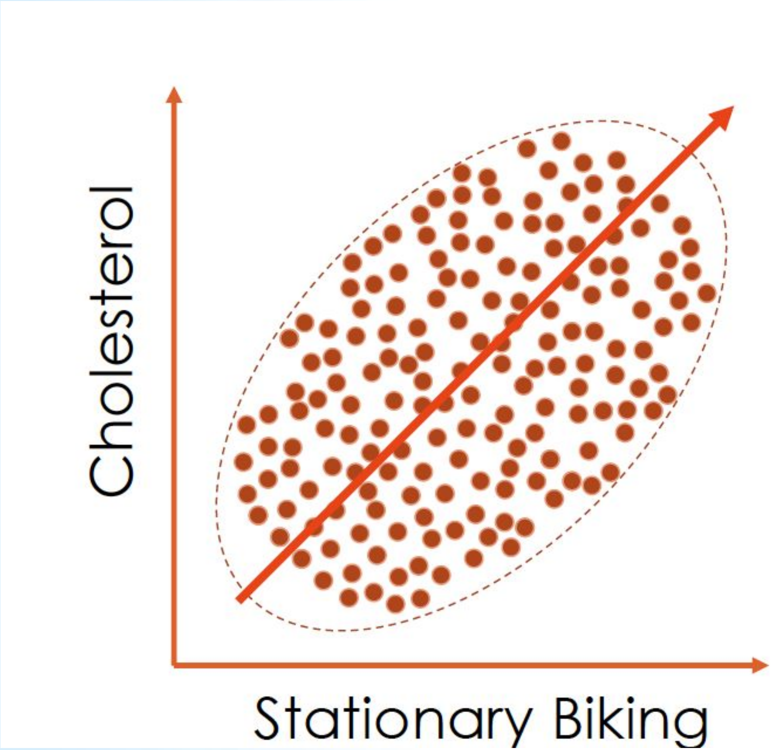


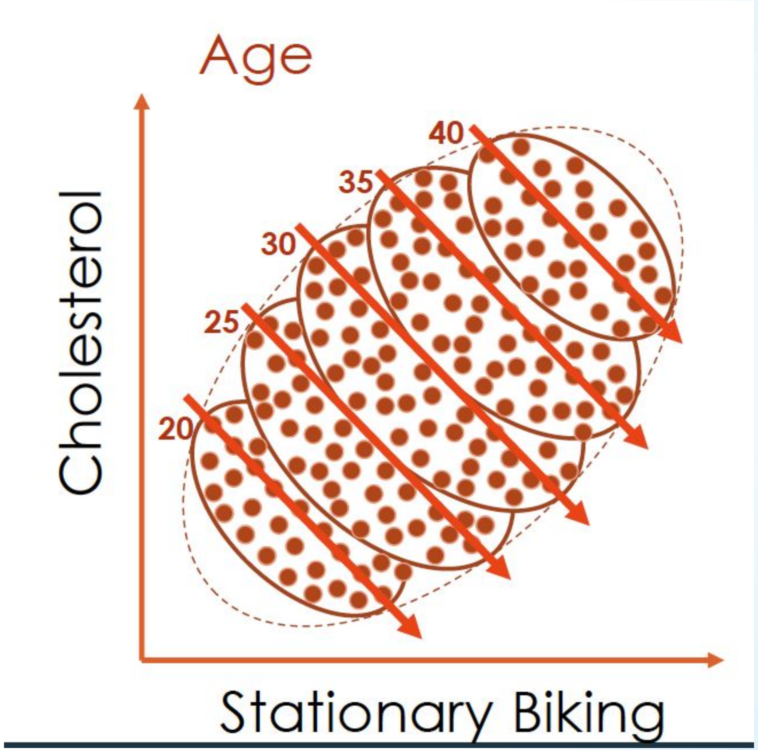
RL Example



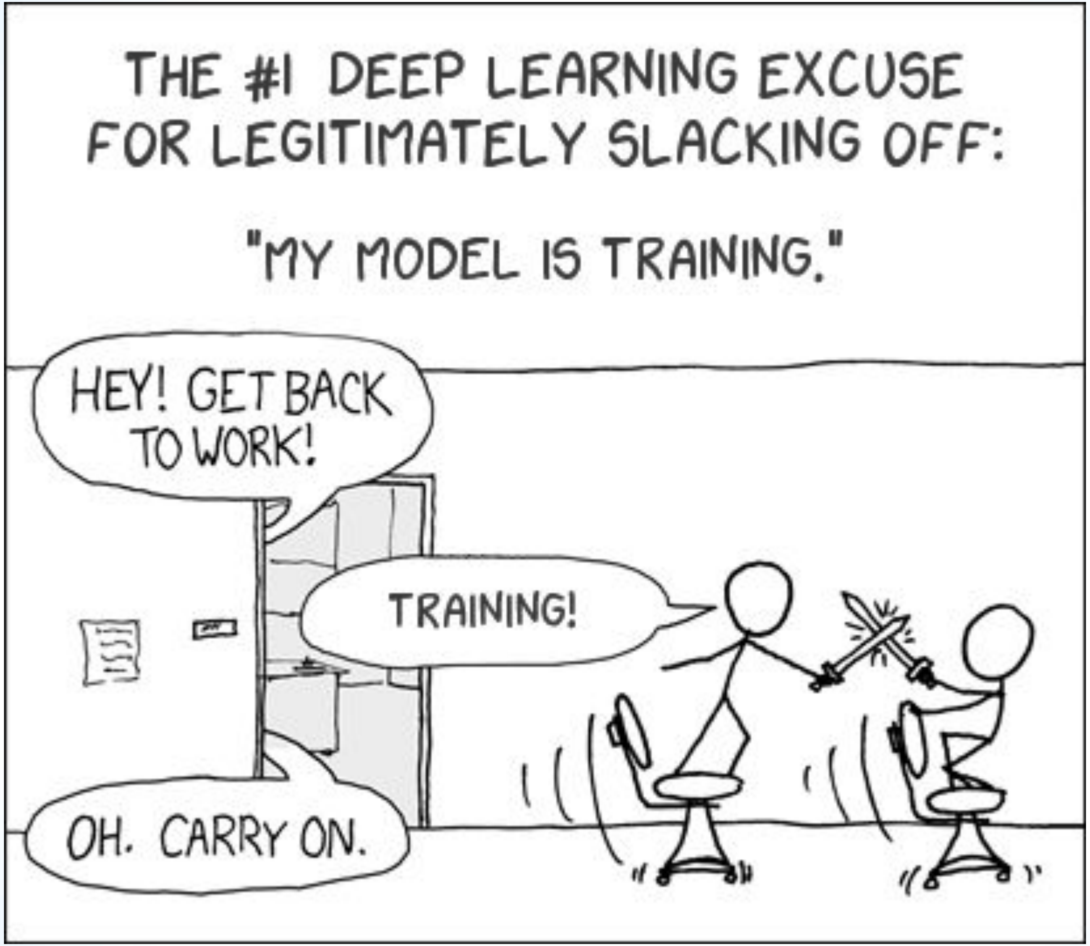
https://medium.freecodecamp.org/an-introduction-to-reinforcement-learning-4339519de419

Causal Inference

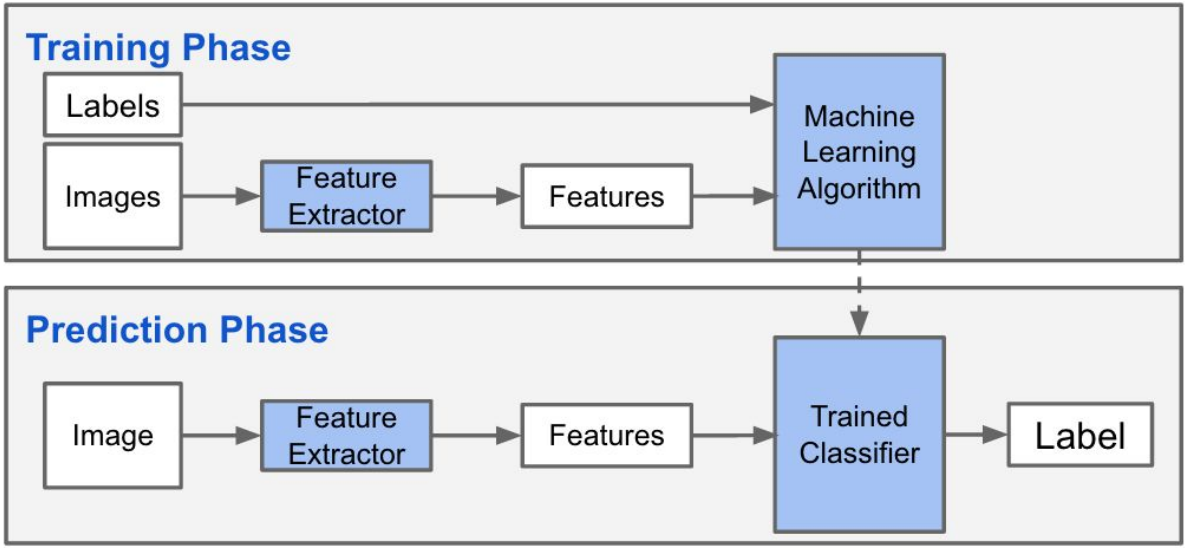




3. How is ML in Practice?


THE #1 DEEP LEARNING EXCUSE
FOR LEGITIMATELY SLACKING OFF:
"MY MODEL IS TRAINING."
HEY! GET BACK TO WORK!
TRAINING!
OH. CARRY ON.

Train vs Test Pipelines



Machine Learning Phases

ML $\in$ DS

1. Understand the domain and explore the data
2. Data integration, selection, cleaning and pre-processing.
3. Learning models ← ML part
4. Interpreting results.
5. Consolidating and deploying discovered knowledge.

It is not a one-shot process, it is a cycle. You need to run the loop until you get a result that you can use in practice. Also, the data can change, requiring a new loop.

~~Hidden~~ Technical Debt in Machine Learning Systems

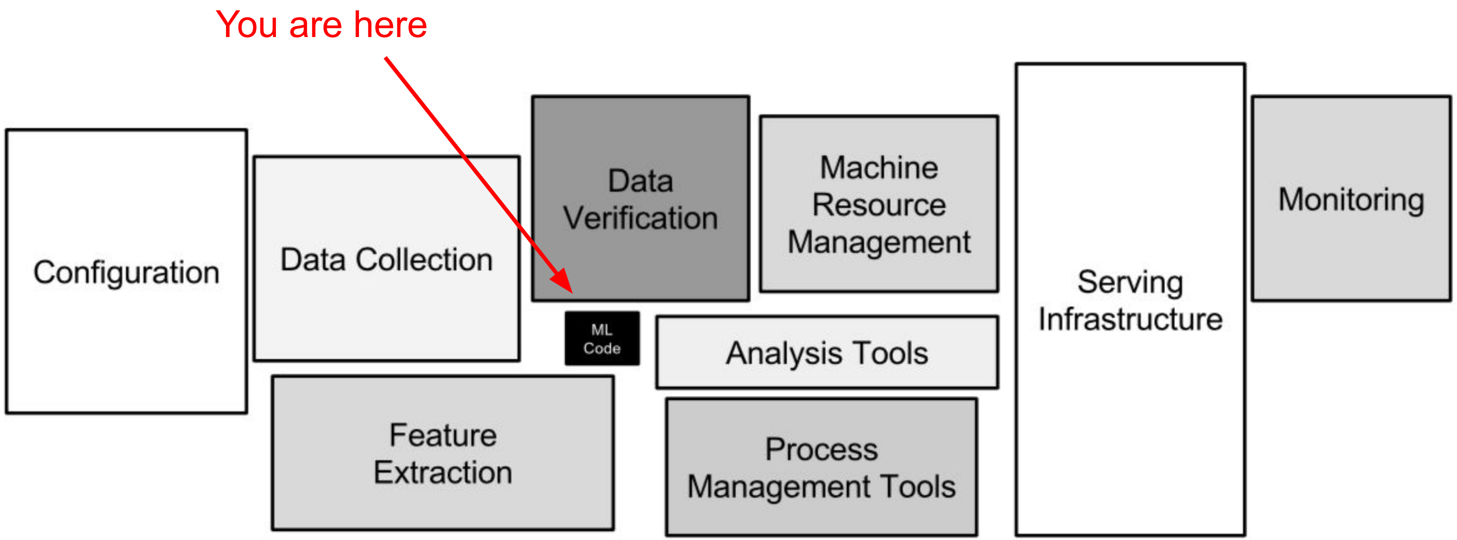
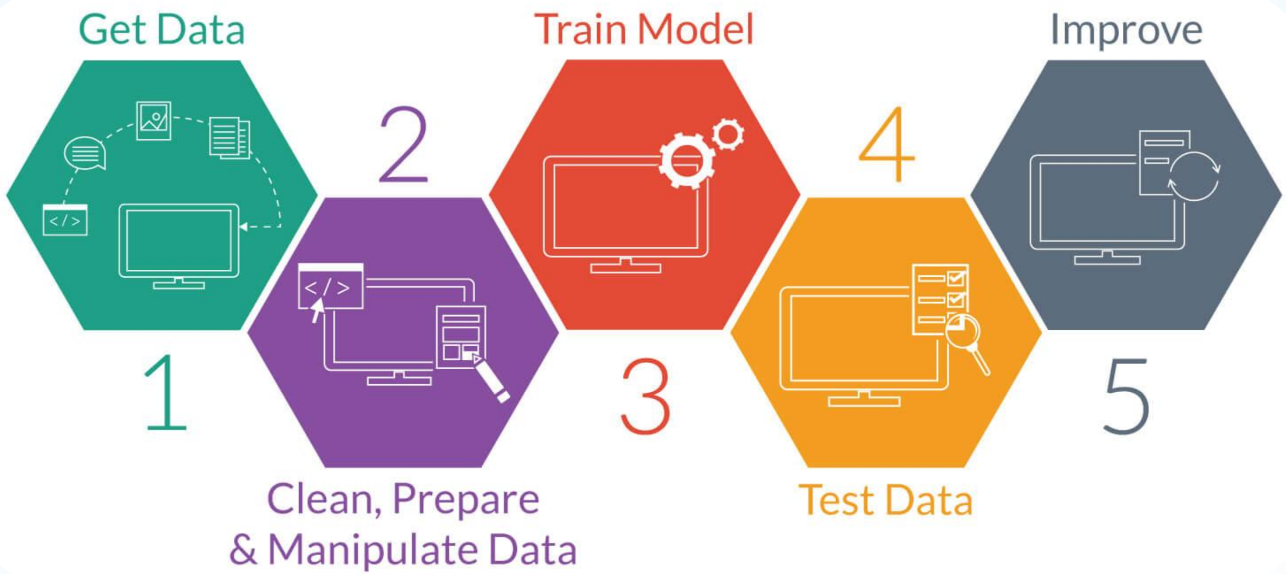


Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

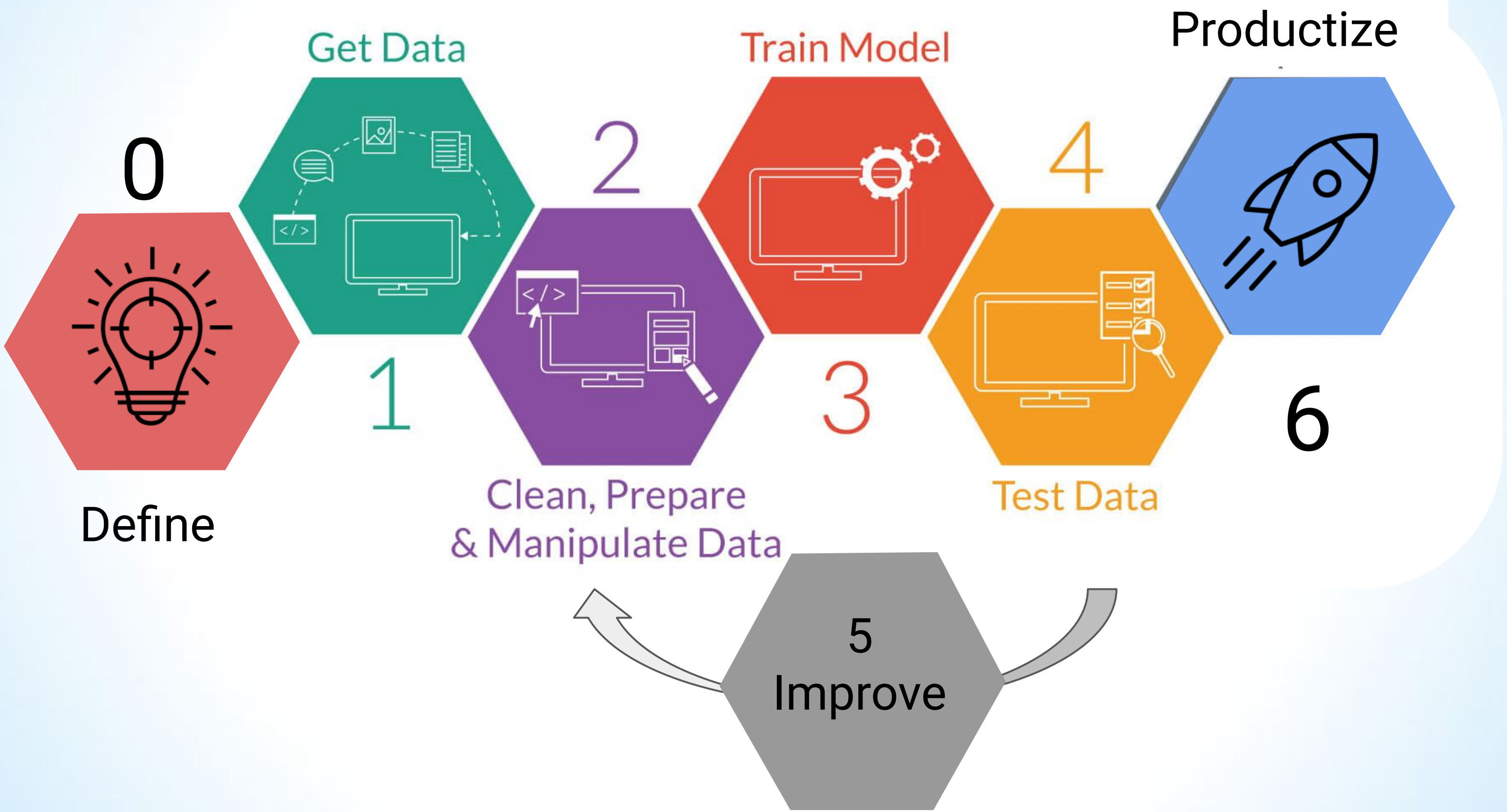
https://papers.nips.cc/paper/5656-hidden-technical-debt-in-machine-learning-systems.pdf

IV. DS project management

Steps to Predictive Modeling



Steps to Predictive Modeling



Y-DATA SCHOOL OF DATA SCIENCE

Yandex

“Tachles” - In Reality

1. Build a Baseline model as soon as possible!
   a. Understand the problem.
   b. Have a SIMPLE data validation pipeline.
   c. Train the SIMPLEST model.
2. Iteratively improve the model
   a. Incremental - add features, tune models, clean data etc.
   b. Go Wild - re-model the data, add sophisticated features, use SOTA approaches.

Task

Given a **KPI** define a feature in the product.

Transaction made weekly by user

1. Stock recommendation - Which Stock should I buy today?
2. Decision support - Should I buy Apple stock?

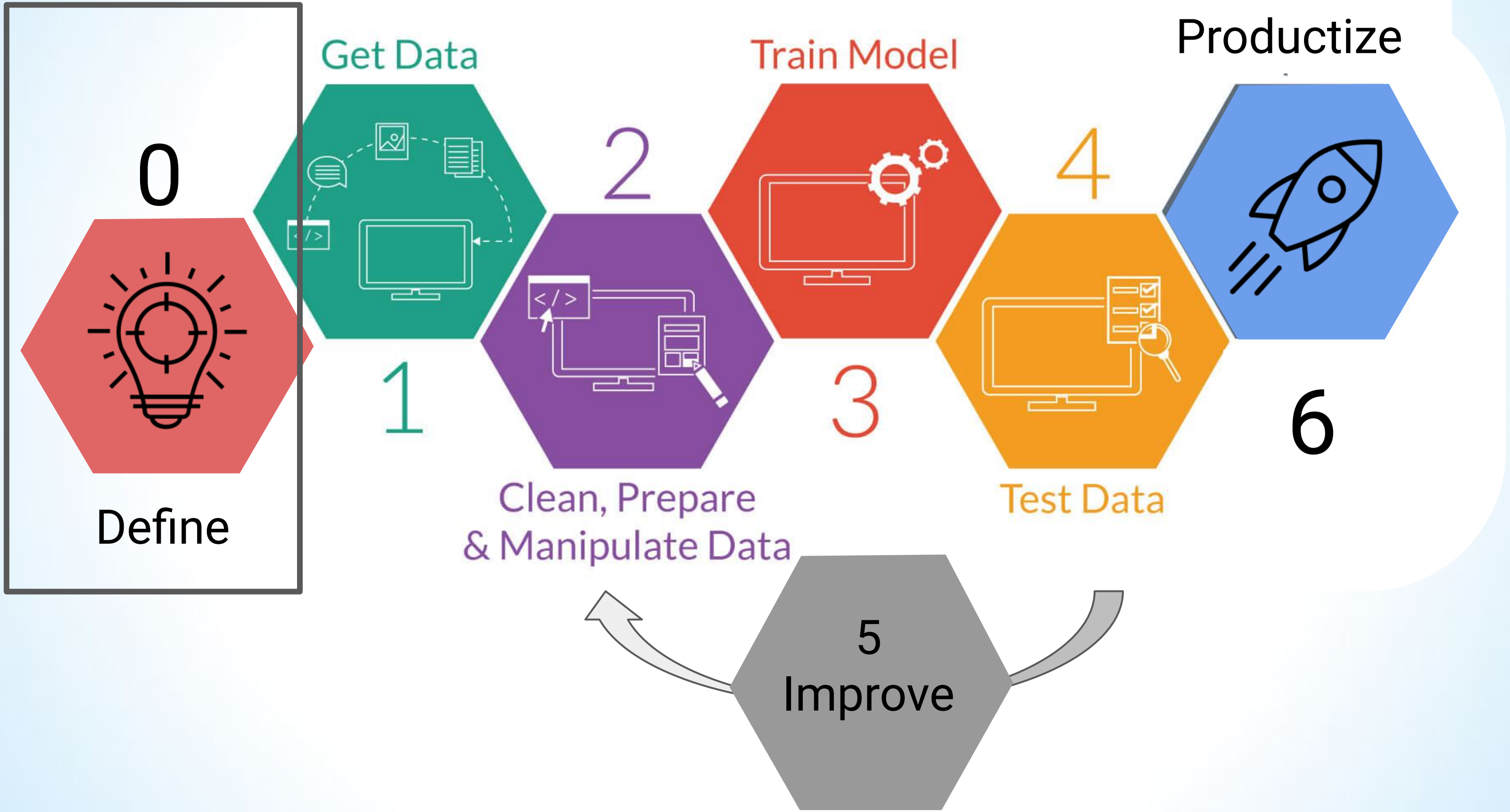
Hours spend daily by customer

1. Decision support - What will be the Microsoft stock price in the next 20 days?
2. Exploration - Which stock is similar to Amazon stock?

Y-DATA SCHOOL OF DATA SCIENCE

Yandex

Steps to Predictive Modeling



Define Task

Which Stock should I buy today?

- Story: Beginning of day a customer will get a list of stocks

Steps

1. Literature overview
2. Data source
3. Data modeling: Entities
4. Labeling function: Rules, manual, data source

Define Data

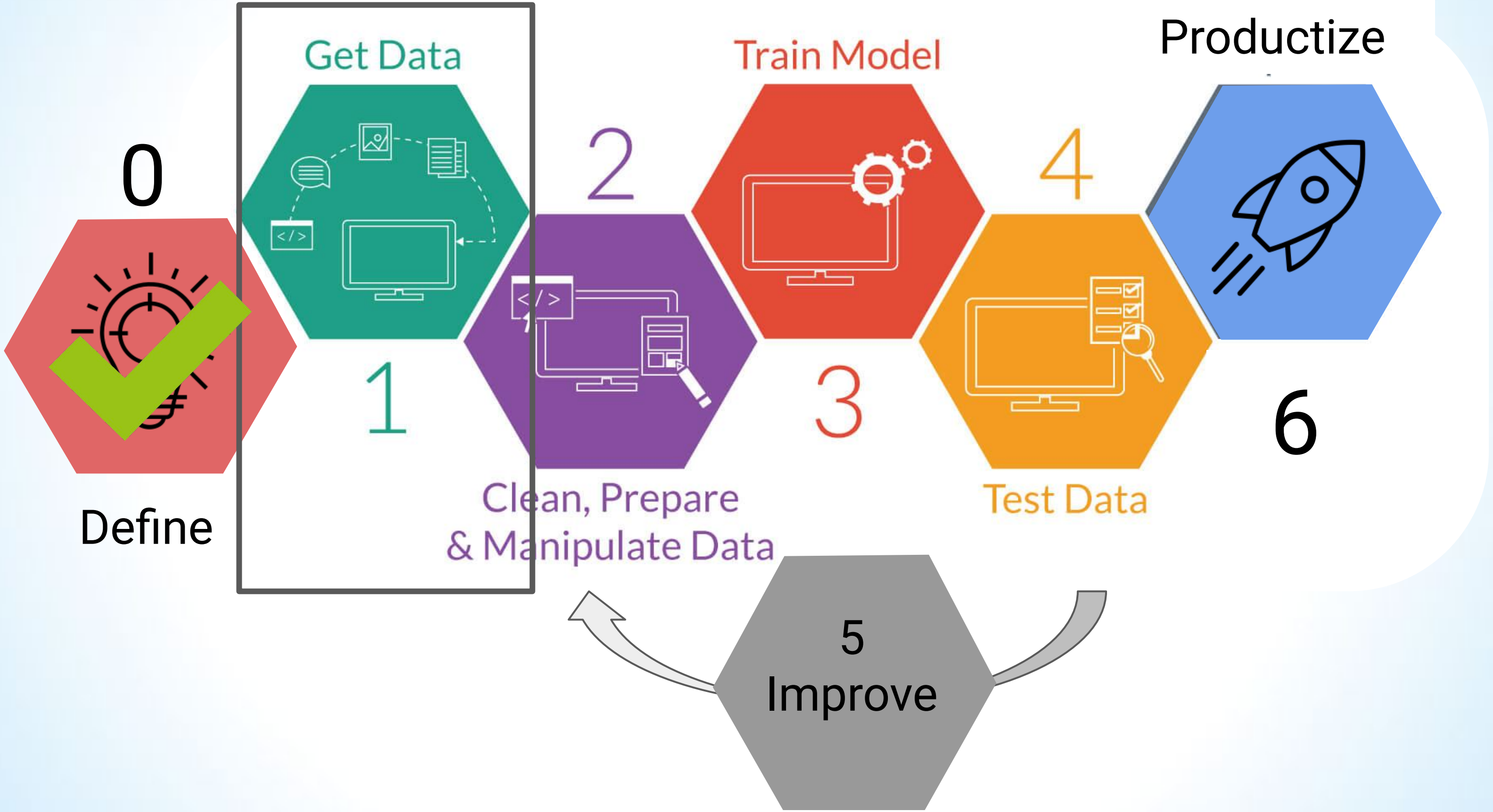
Which Stock should I buy today?

- Story: Beginning of day a customer will get a list of stocks

Data sources - based on “what” we make a decision?

- **Data** - All stock of S&P 500 - after 2009 and before 2020.
- **Entity** - stock
- **Horizon** - next Day

Steps to Predictive Modeling



Ground Truth

Which Stock should I buy today?

- Story: Beginning of day a customer will get a list of stocks

Data sources - based on “what” we make a decision?

- **Data** - All stock of S&P 500 - after 2009 and before 2020.
- **Entity** - stock
- **Horizon** - next Day

Labeling - What is a good stock?

- A stock with low volatility in the next day -> low Sharp ratio
- A stock with high revenue potential -> 4% price increment

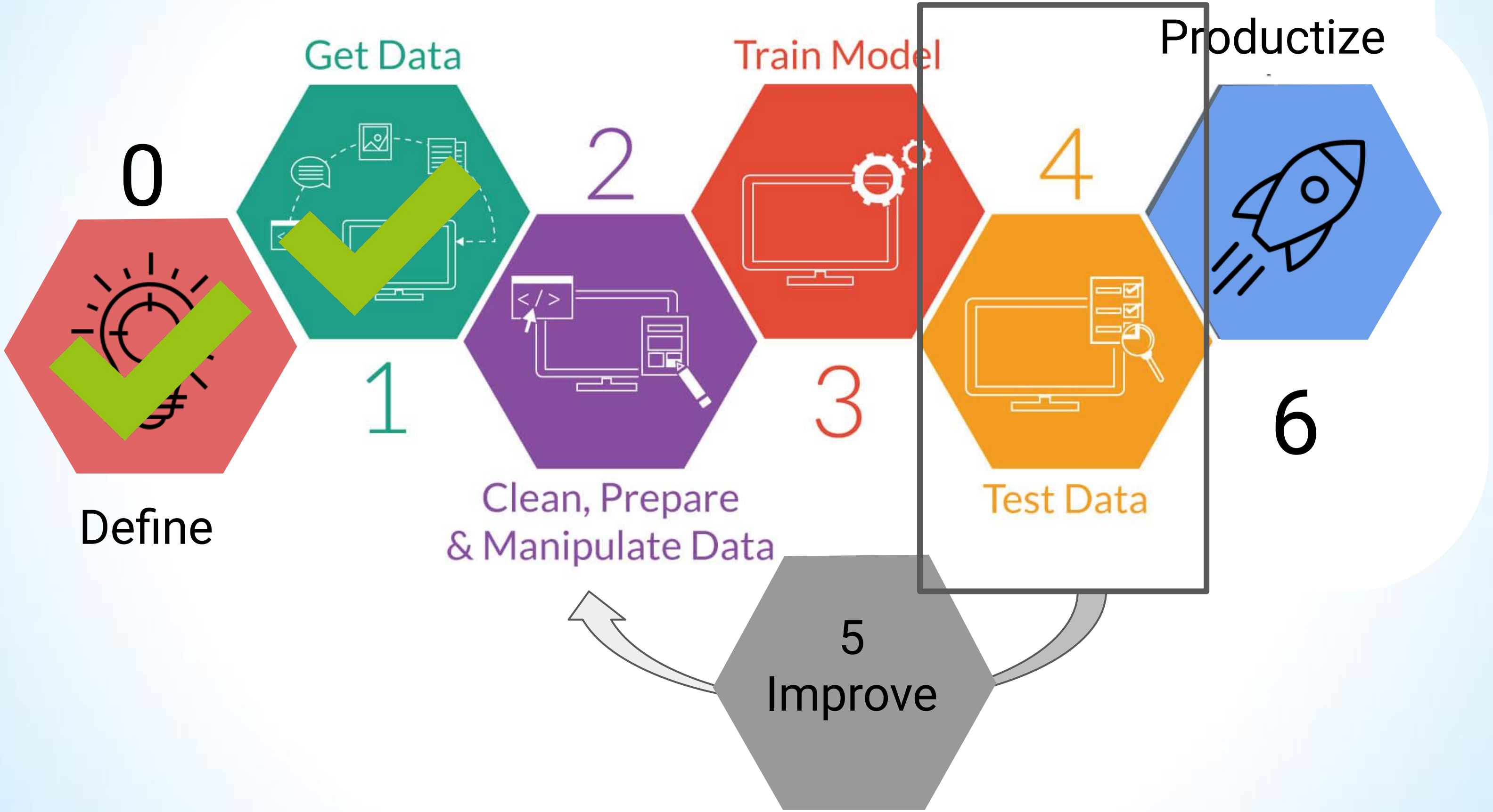
Labeling function

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

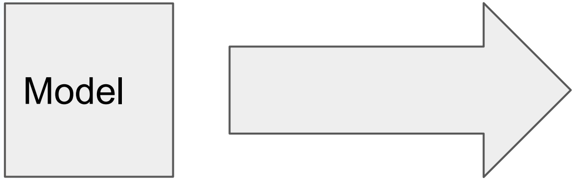
Date	Stock	Price	Volatile	Profit	Label
1/1/2020	TSLA	50	Low	High	1
1/1/2020	AMZ	63	Low	Low	0
1/1/2020	APPL	42	High	Low	0
2/1/2020	TSLA	55	High	High	0
2/1/2020	AMZ	60	High	Low	0
2/1/2020	APPL	39	Low	Low	0
3/1/2020	TSLA	60	Low	High	1

4. How to Estimate Model’s Performance

Steps to Predictive Modeling



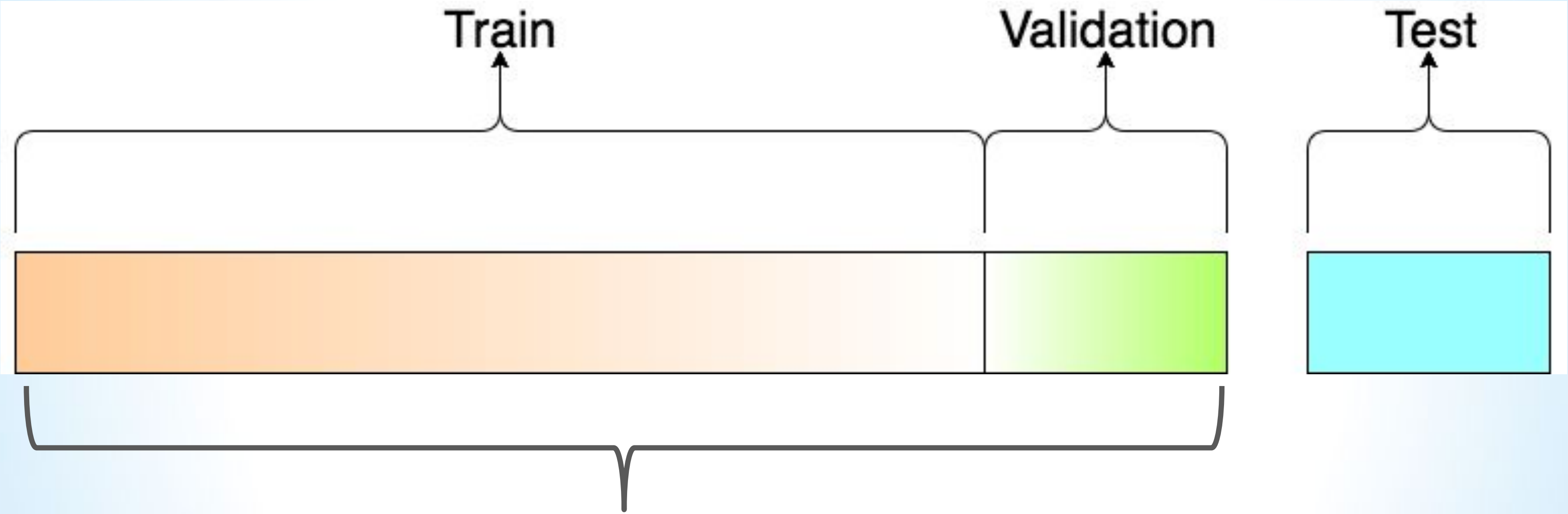
How To Evaluate?



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

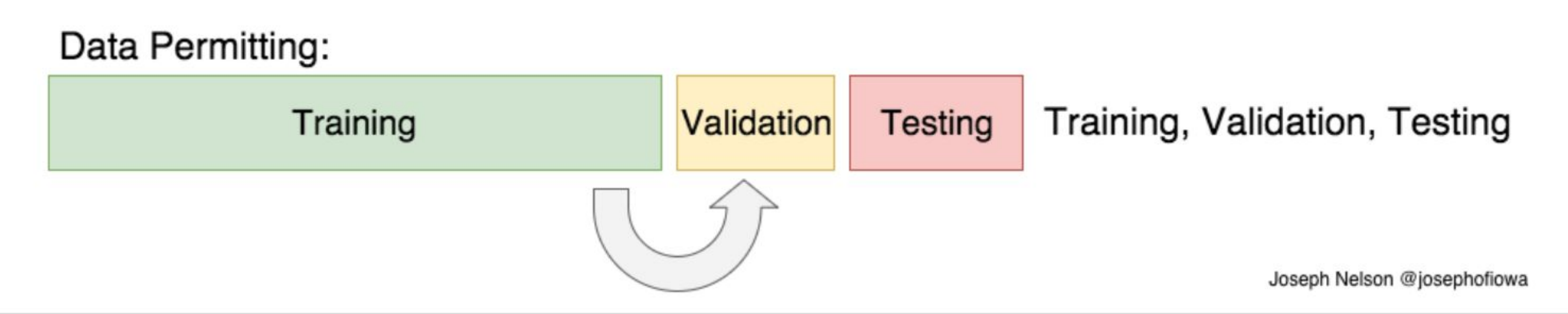
Amazon
- Buy Confidence 49%
- Not Buy Confidence 51%

Estimating Performance



E.g. Choose best model
Hyper parameter optimization

Estimating Performance - Data is Abundant

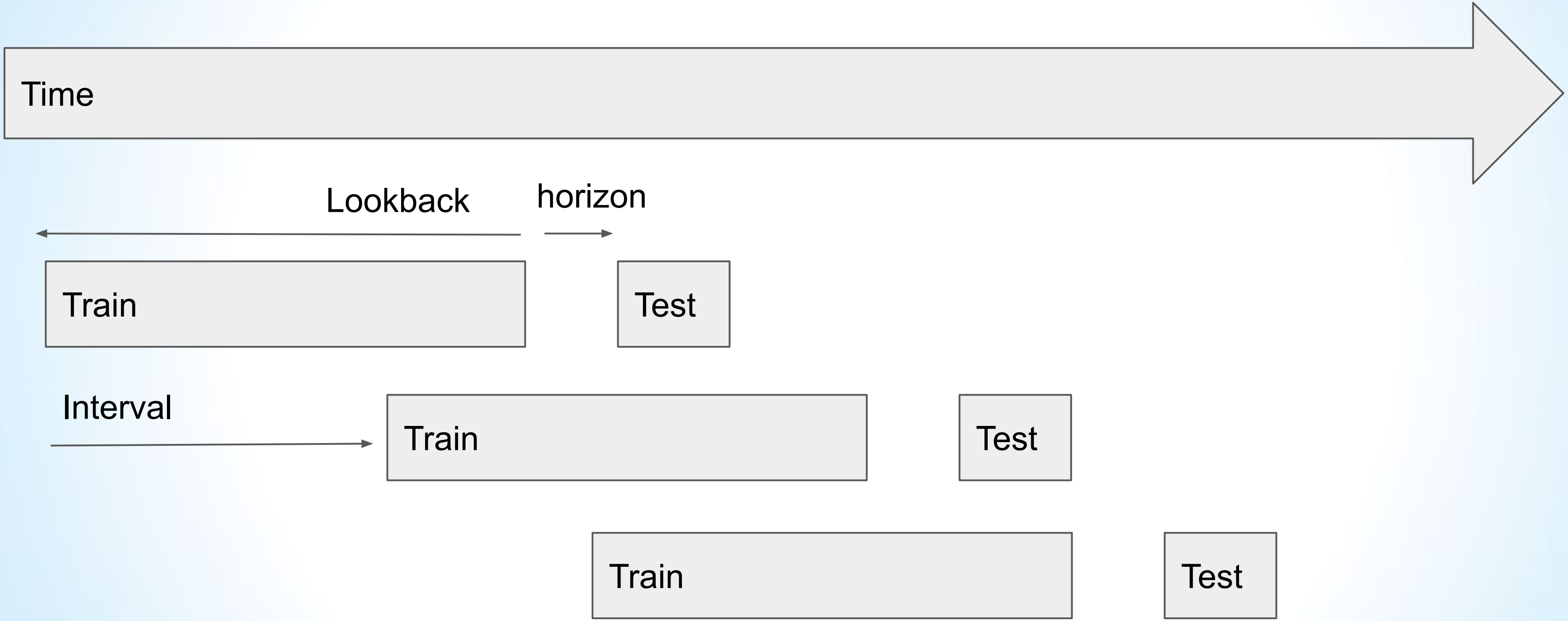


Datasets distribution: Training <> Validation == Test ~ Real world = Random

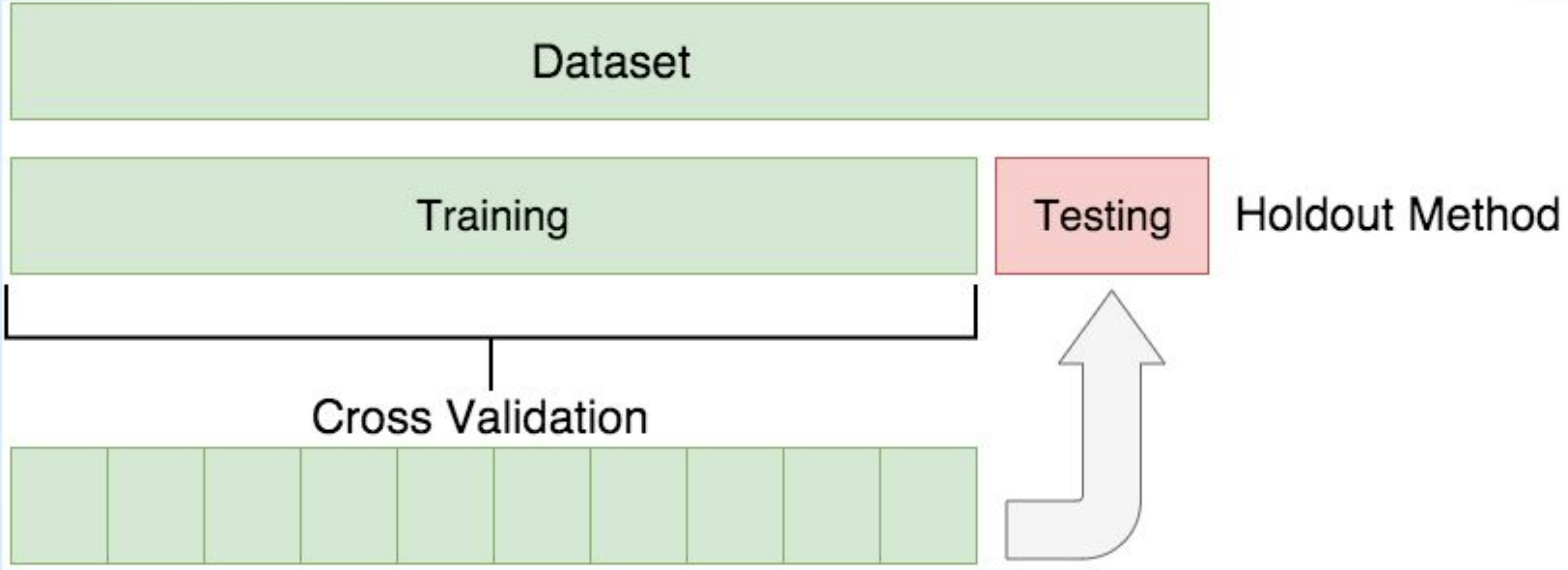
Validation used for Hypertuning and model Calibration

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)*			

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

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	**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

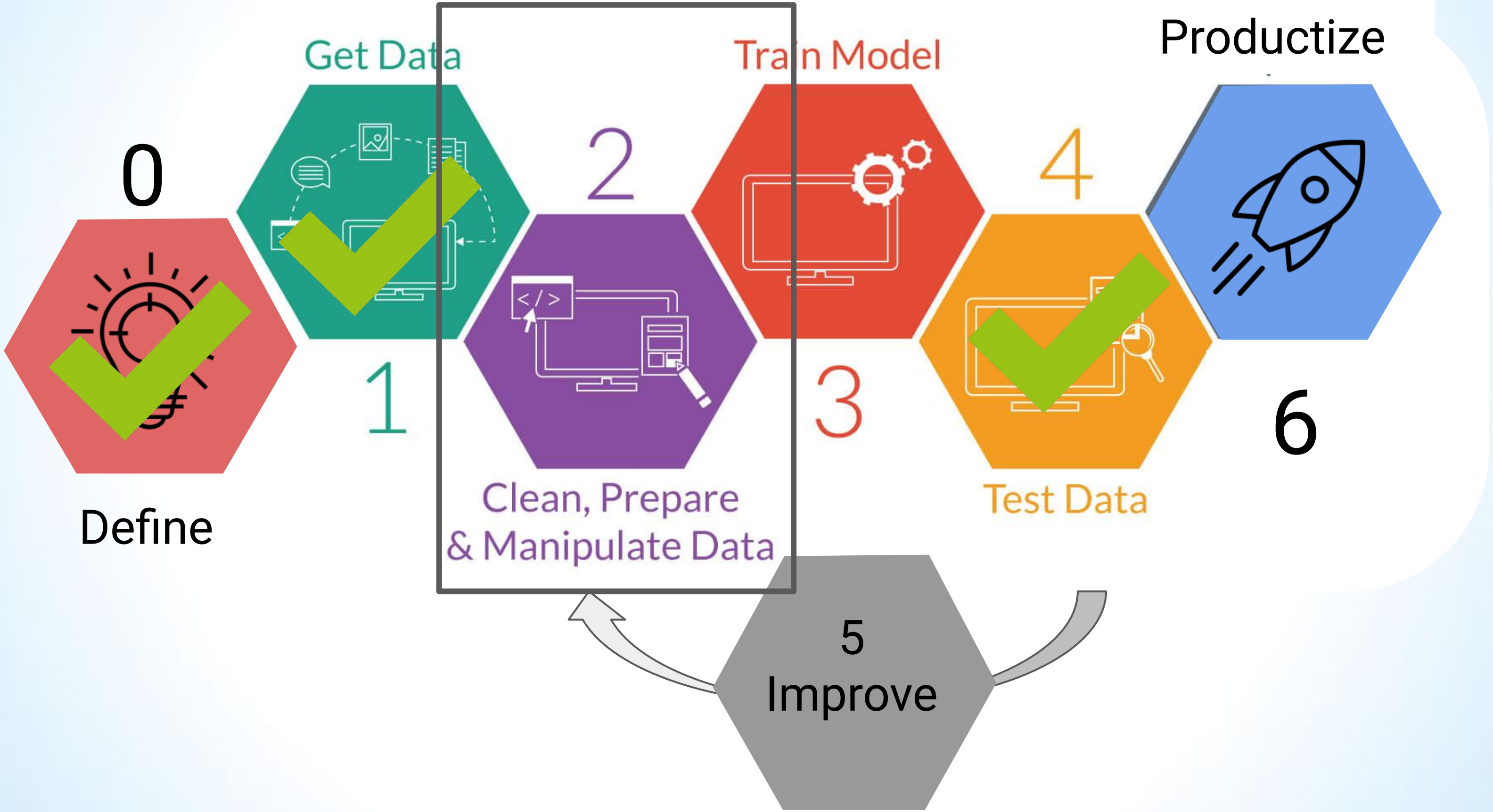
		Label		
		Condition Positive *(Buy)*	**Condition Negative** *(Don't Buy)*	
Classifier	**Predict Positive** *(should buy)*	**True Positive** (TP) = 20	**False Positive** (FP) = 180	**Positive predictive value | Precision** TP / (TP + FP) = 20 / (20 + 180) = **10%**
	Predict Negative *(shouldn't buy)*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

5. How to prepare data for ML

Steps to Predictive Modeling



Classical vs Deep Learning Framework

Traditional Machine Learning Flow

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

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 Change	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

Feature Selection

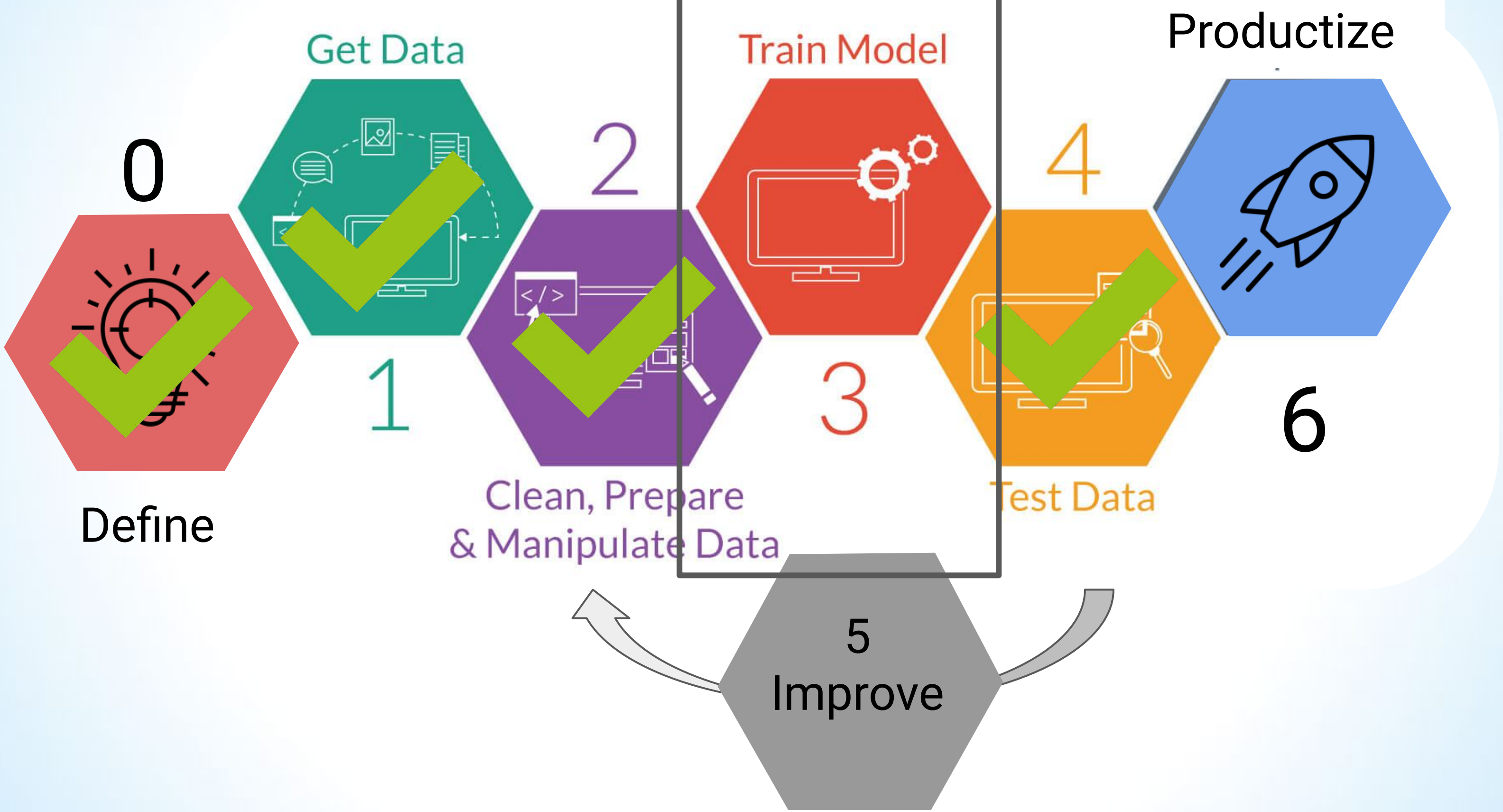
Days price increase	Price Change	RSI	Sector	Label
4	0.12	0.4	Auto	1
2	0.35	0.7	Software	0
4	0.8	0.5	Energy	0
3	0.22	0.3	Materials	1
5	0.3	0.6	Health	0
1	0.1	0.3	Telco	1

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

6. What types of ML Algorithms are there?

Steps to Predictive Modeling



Parametric vs. Non-parametric Models

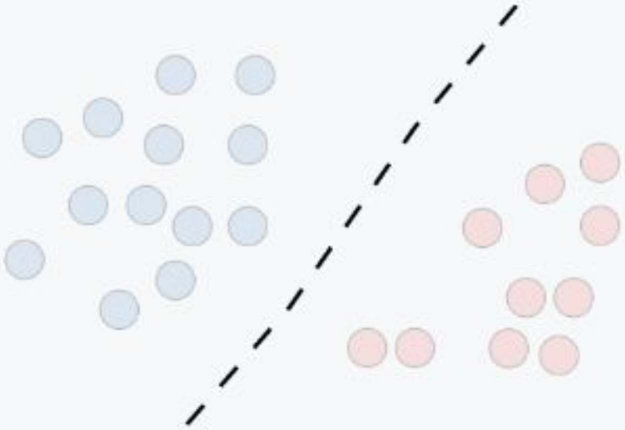
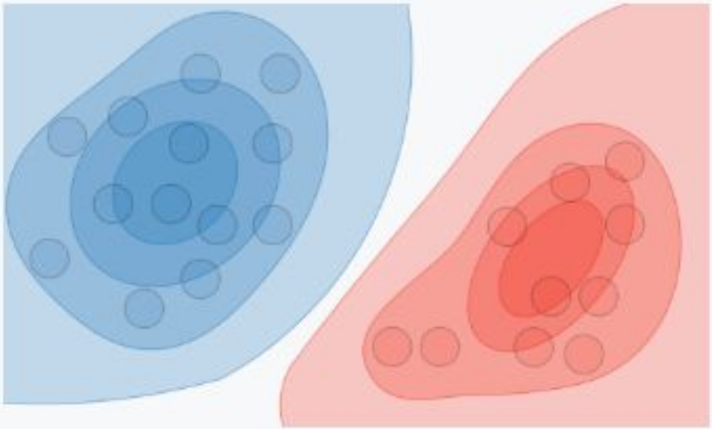
Almost all models for machine learning have "parameters" or "weights" that need to be learned.

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?

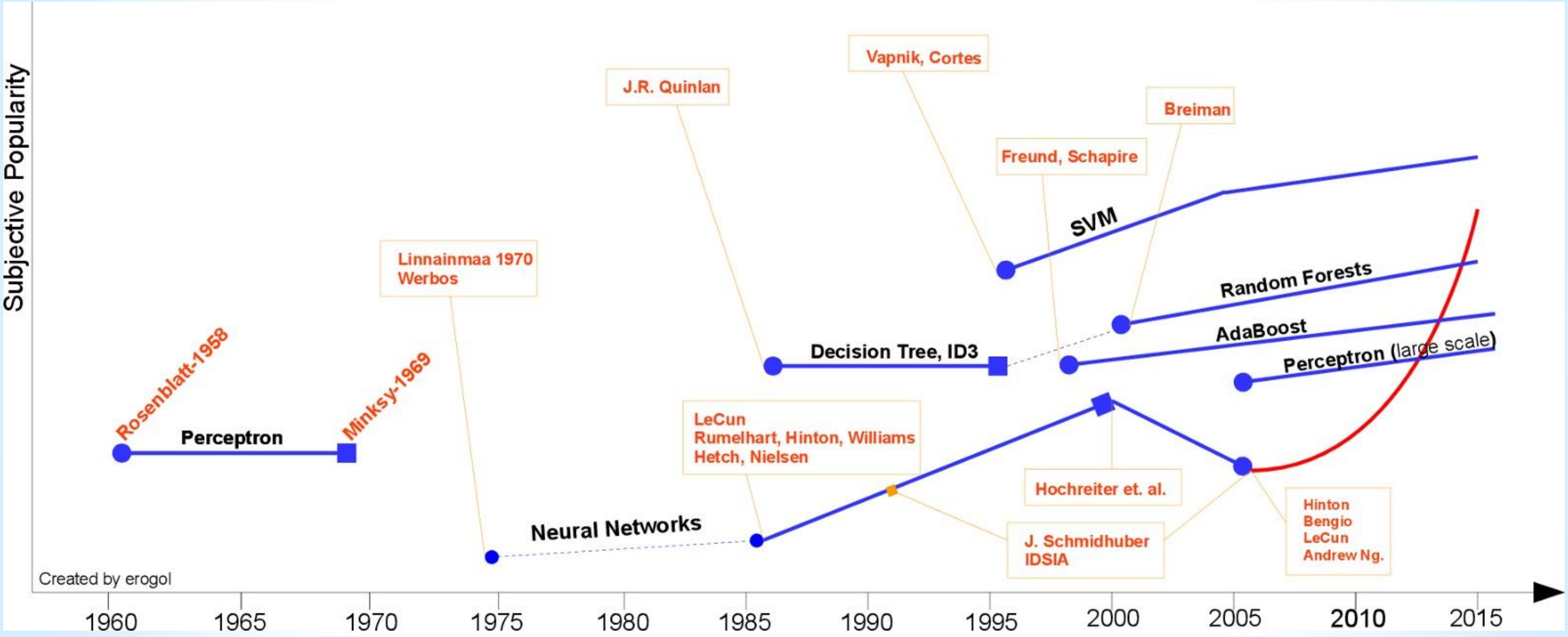
Can you think of an example for parametric and non-parametric method?

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

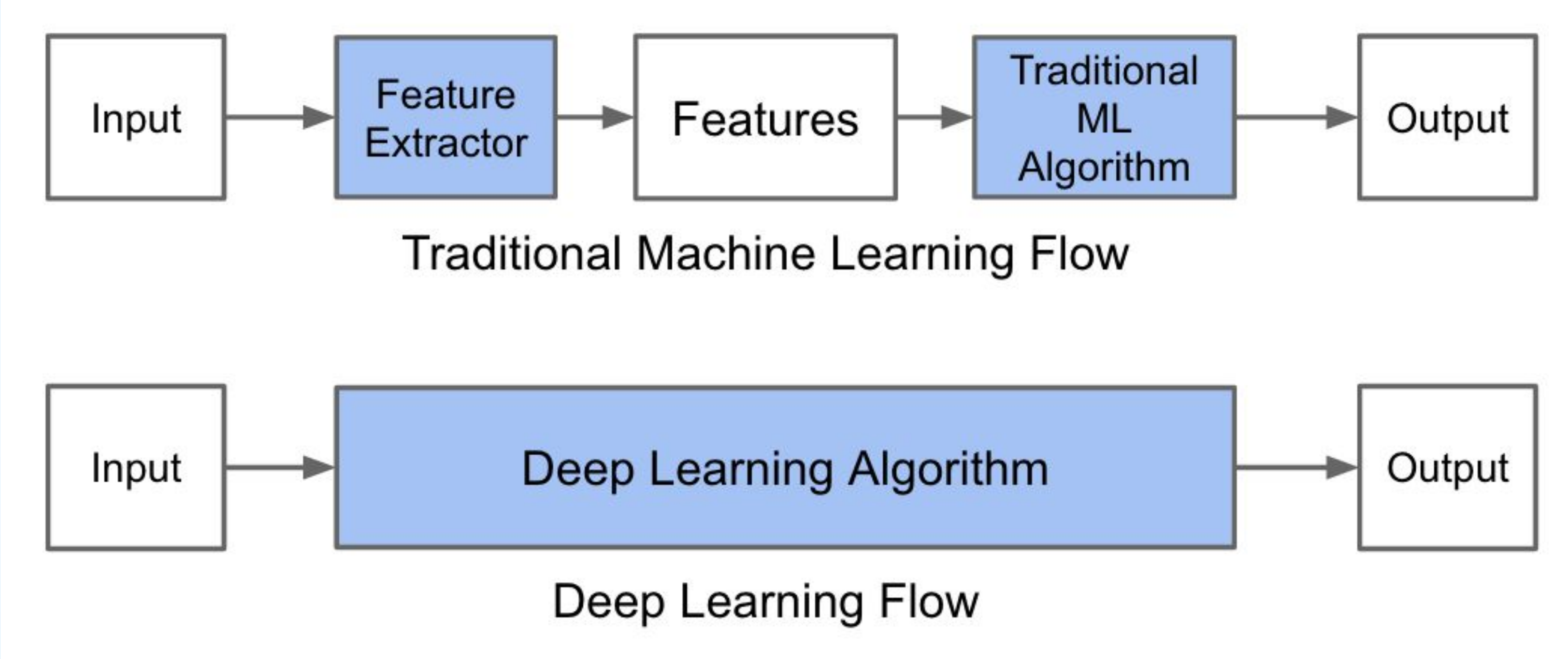
https://stackoverflow.com/questions/879432/what-is-the-difference-between-a-generative-and-discriminative-algorithm

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

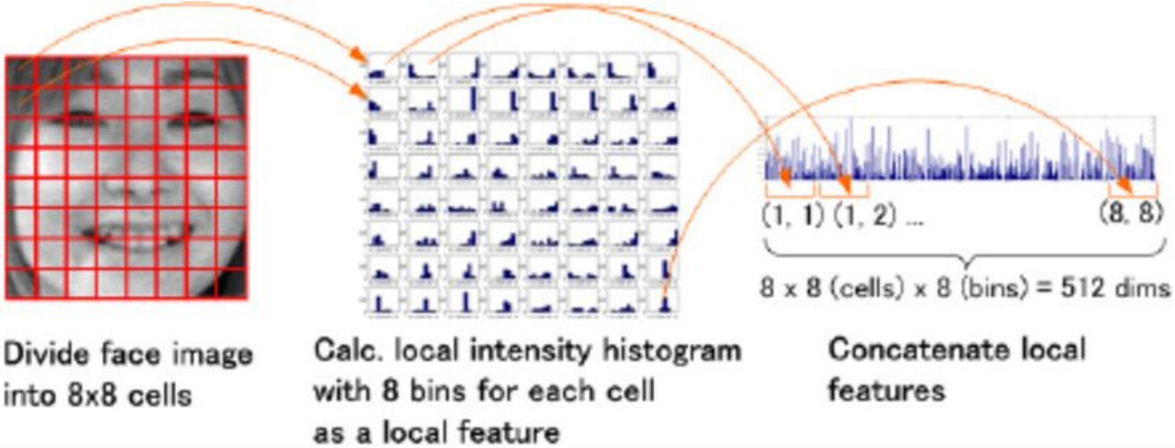


http://adilmoujahid.com/posts/2016/06/introduction-deep-learning-python-caffe/

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…

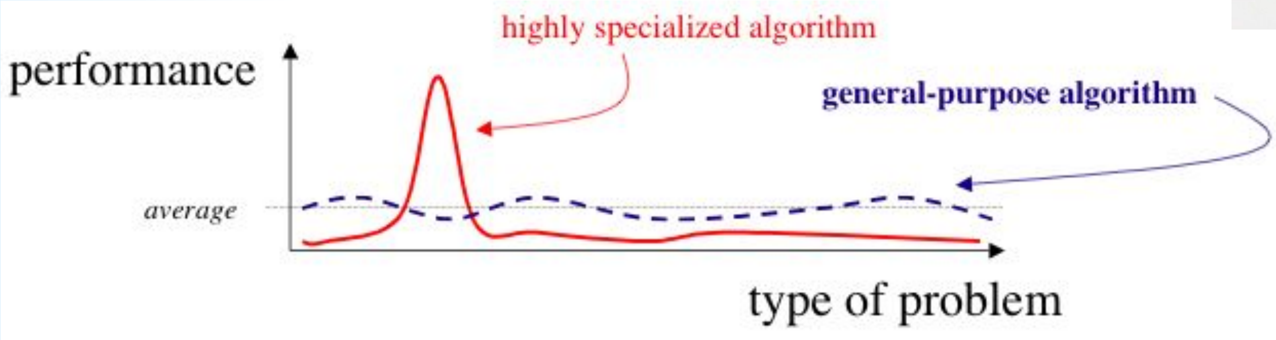
Y-DATA SCHOOL OF DATA SCIENCE

Yandex

No Free Lunch Theorem - Best Model Does Not Exists

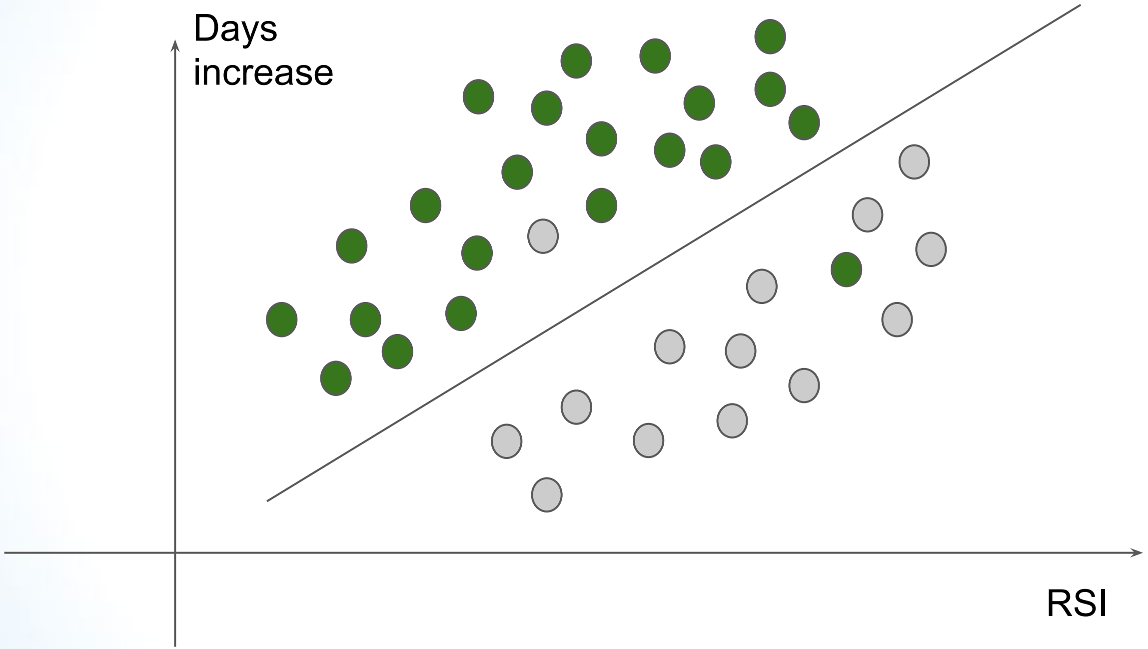
A superior black-box optimisation strategy, which is better than anything else for any kind of problem, is impossible.

Deep cannot be always better.

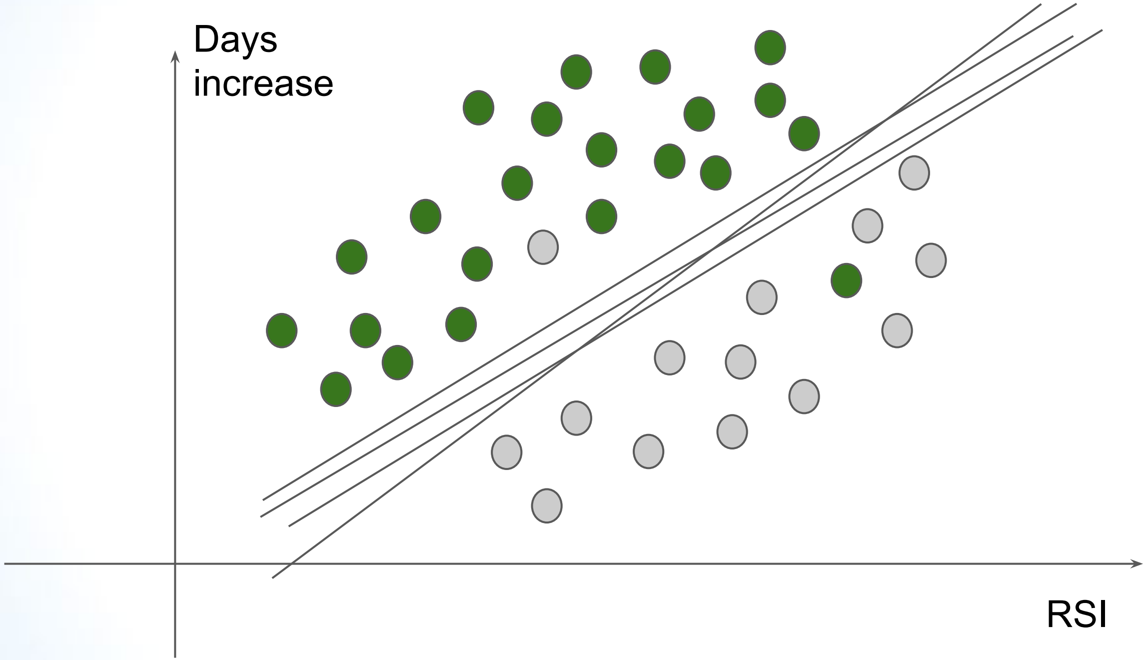


https://medium.com/@LeonFedden/the-no-free-lunch-theorem-62ae2c3ed10c

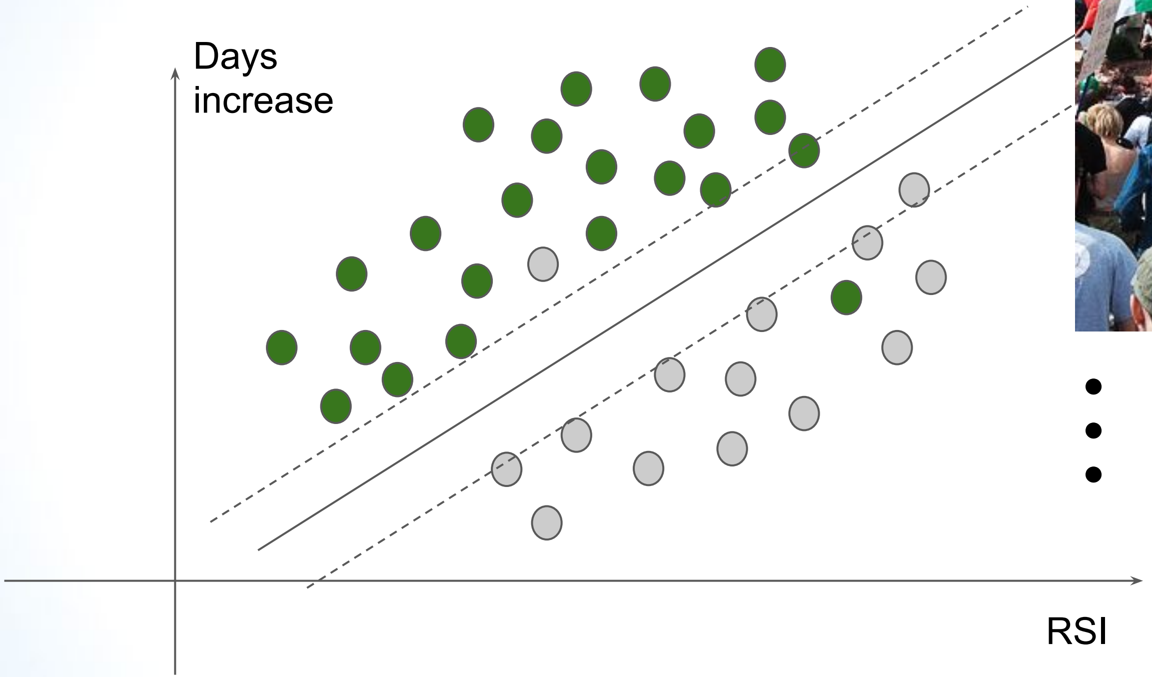
Modeling - Linear



Modeling - Linear

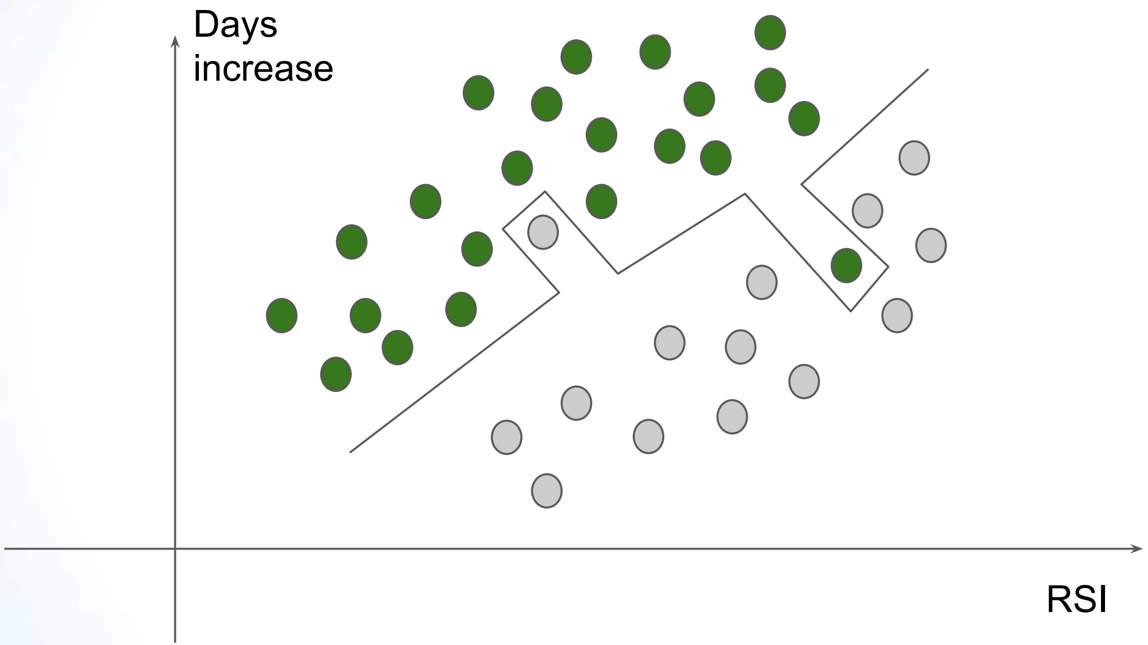


Modeling - Maximum Margin

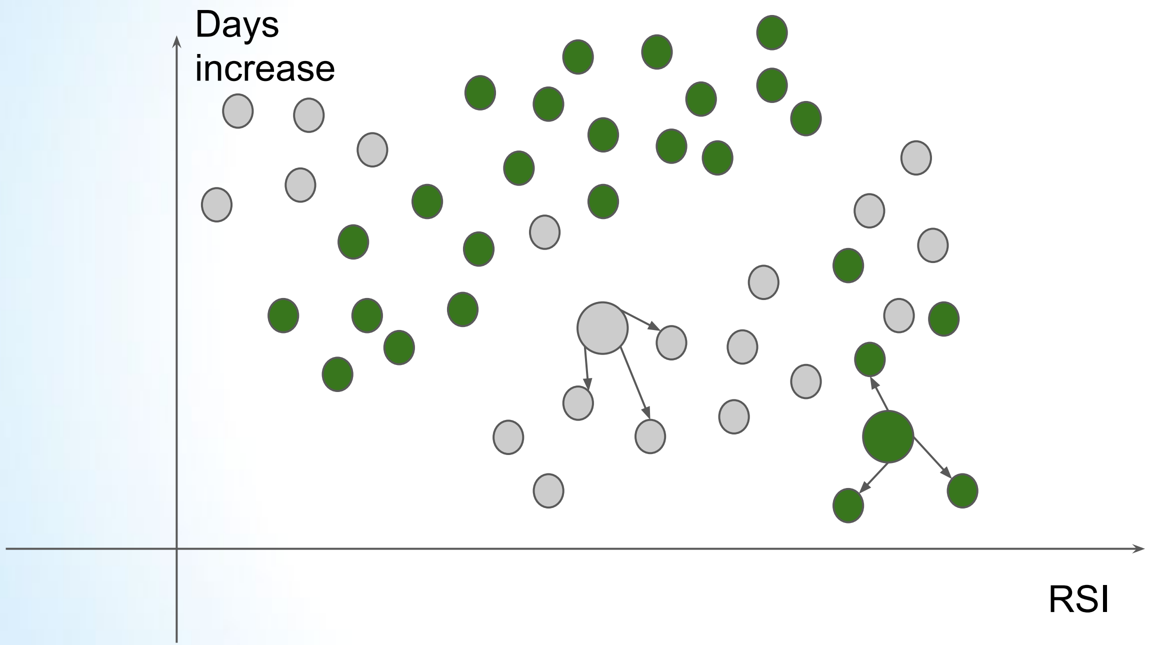


- A Linear classifier
- Maximize the margin
- Simple Linear SVM - LSVM

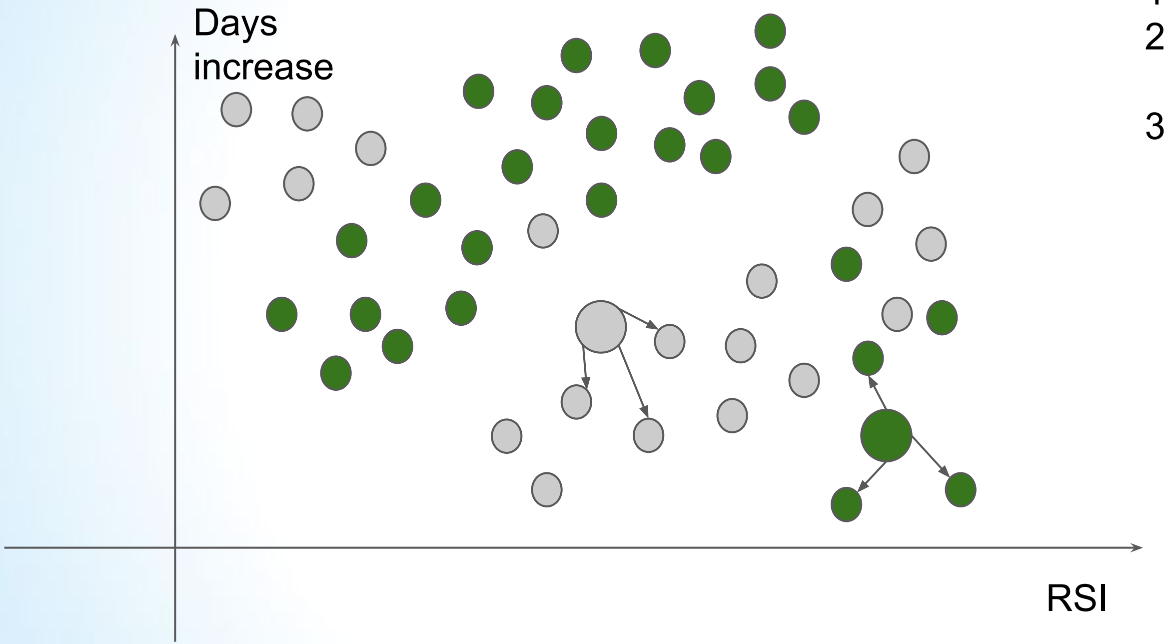
Modeling



Modeling - K nearest neighbors

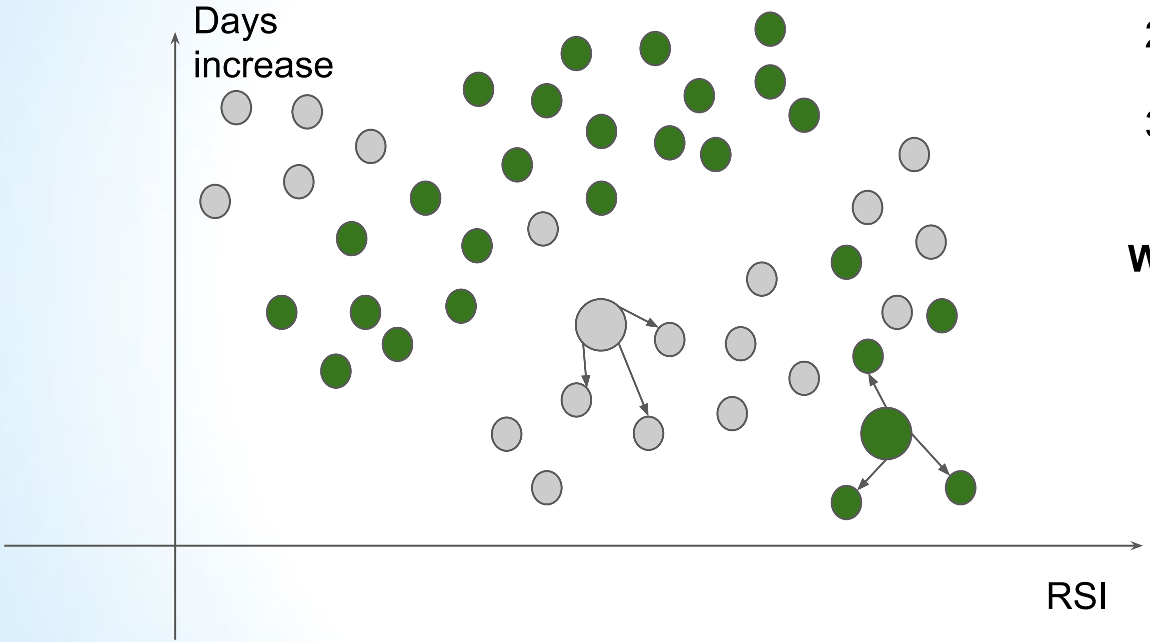


Modeling - K nearest neighbors



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?

KNN Best Practices

When to Consider

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

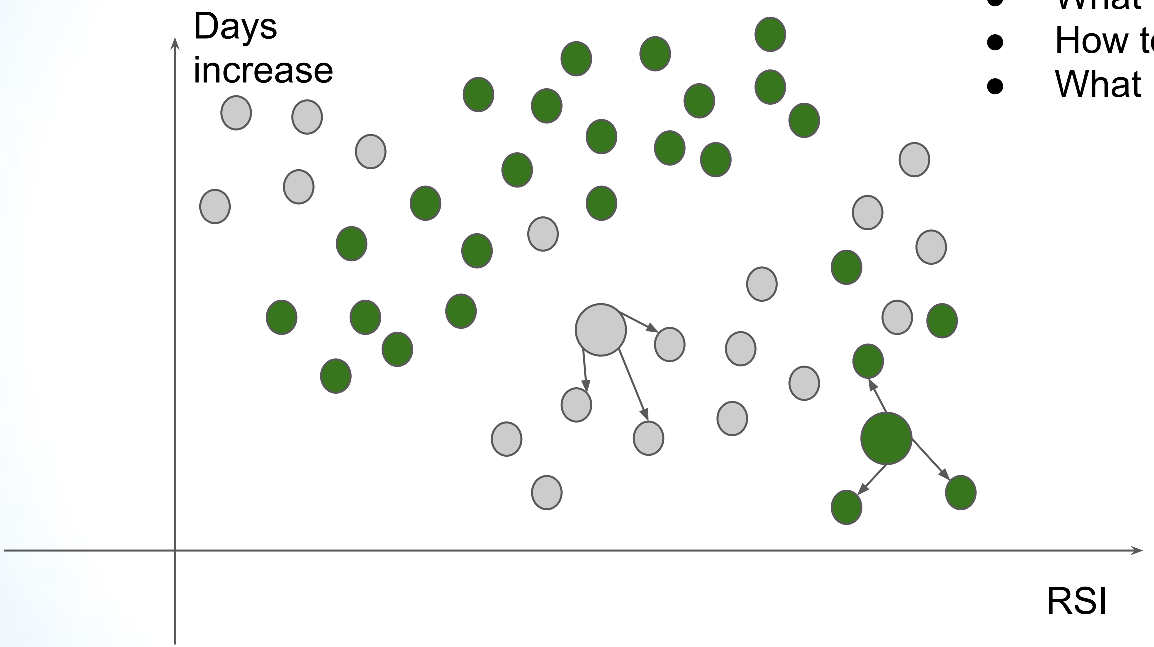
Advantages

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

Disadvantages

- Slow at query time
- Easily fooled by irrelevant attributes

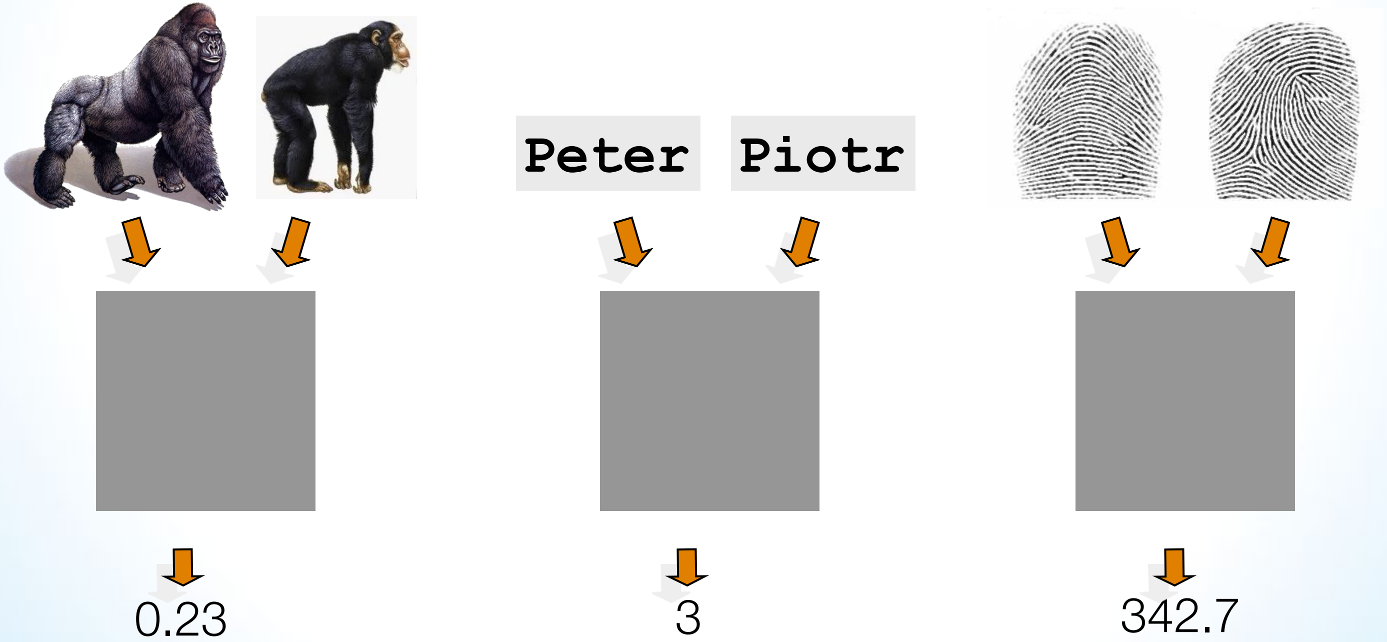
Modeling - K nearest neighbors



- 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 $D(O_1,O_2)$

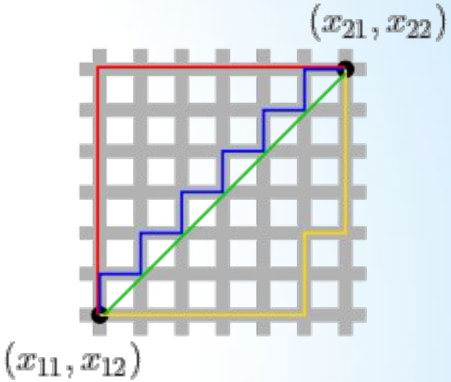


Distance function behavior

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

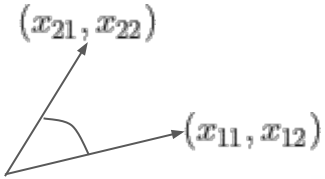
Distance Function

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



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

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



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

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

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

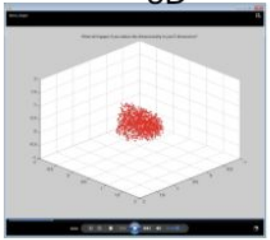
Any ideas about issues with using absolute ?

Curse of Dimensionality


ADHESIVE

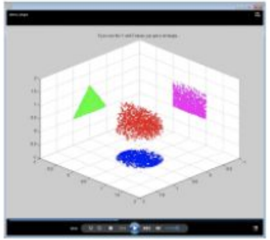
www.cs.gmu.edu/~jessica/DimReducDanger.html eamonn keogh

A cloud of points in 3D

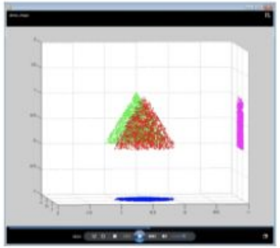


Can be projected into 2D

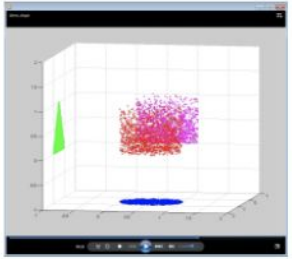
XY or XZ or YZ



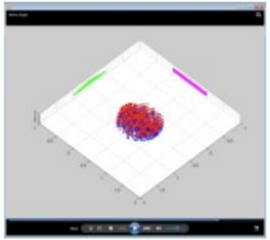
In 2D XZ we see a triangle



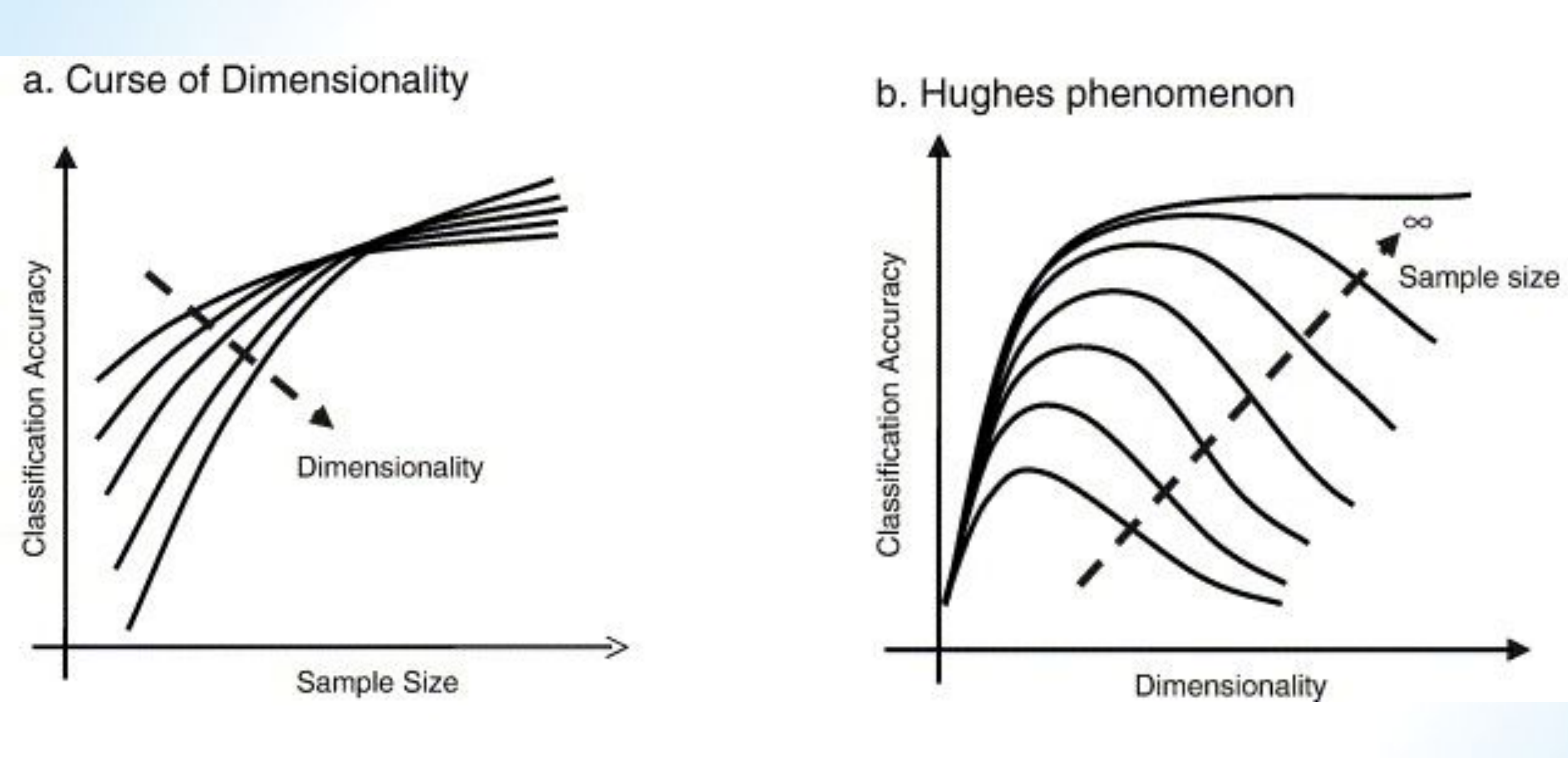
In 2D YZ we see a square



In 2D XY we see a circle



Hughes phenomenon (1968) (Peaking Paradox)

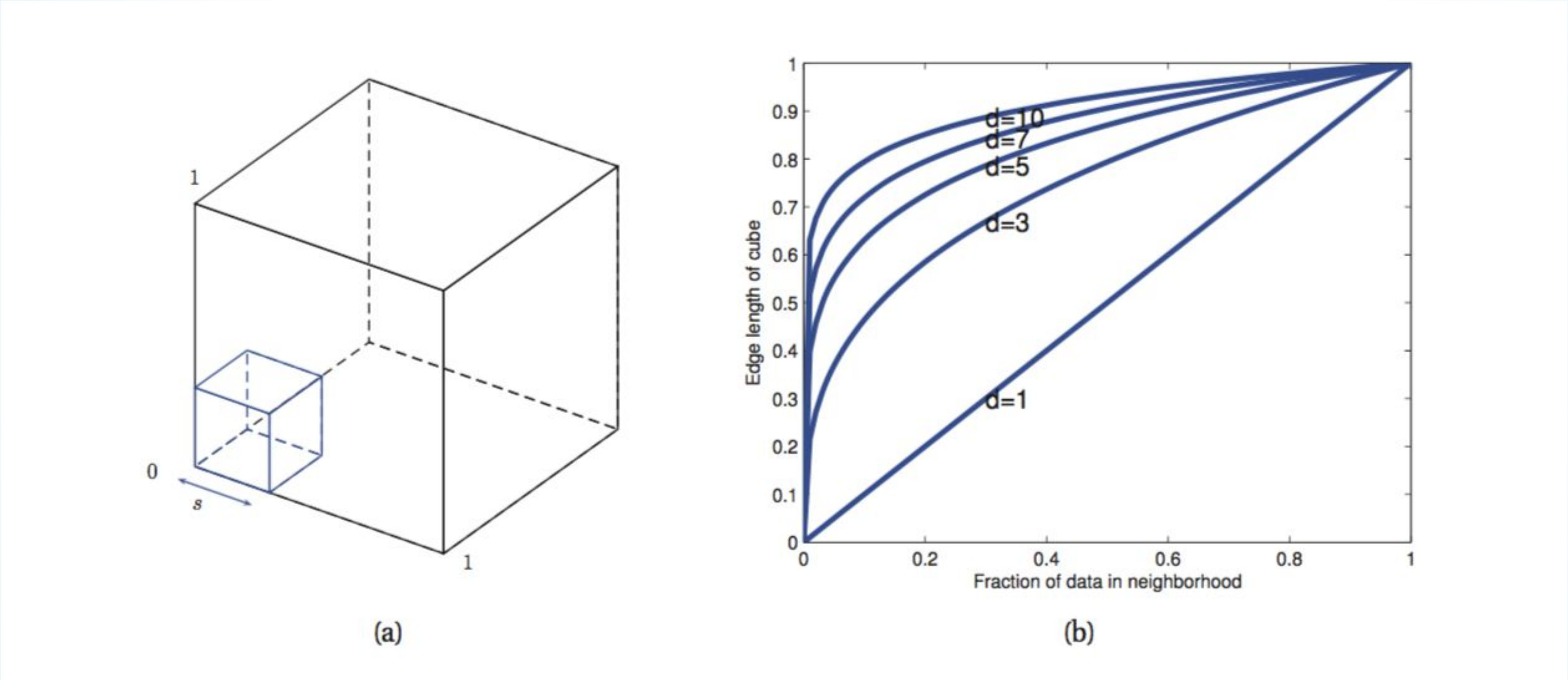


http://37steps.com/2322/hughes-phenomenon/

Y-DATA SCHOOL OF DATA SCIENCE

Yandex

Why Nearest Neighbours Fails in High Dimensions?

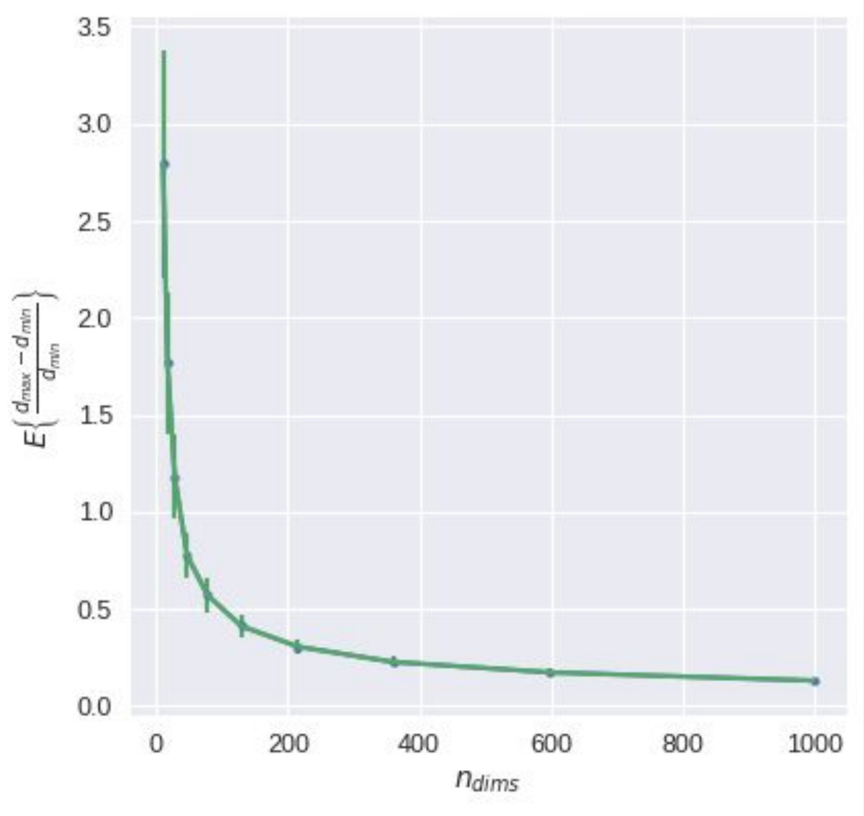


Kevin Murphy's book: Machine Learning - A probabilistic Perspective

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.



Example: Detecting Suspicious URL Names

- DDOS
- Botnets
- Derive by download
- Phishing - How phishing is different from other attacks?

Task:

- Build an analytics tools to explore and detect NEW types of malicious URLs.
- What we wish to gain? Precision / Recall
- How you can represent domain Names?

Appendix - helix

Helix: DGA Domain Embeddings for Tracking and Exploring Botnets
L Sidi, Y Mirsky, A Nadler, Y Elovici, A Shabtai - Proceedings of the 29th ACM International Conference …, 2020

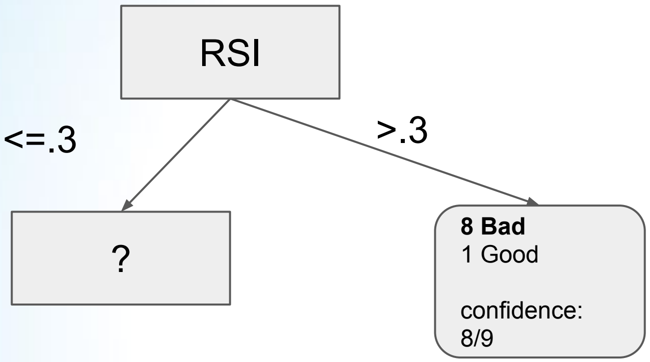
Review Homework

Part 1 - Implement k-Nearest Neighbours (KNN)

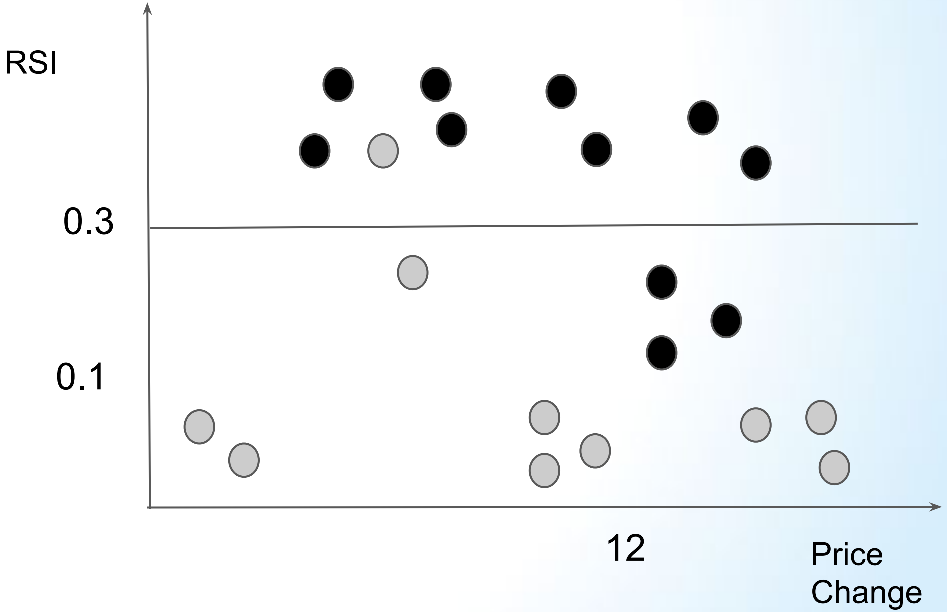
Part 2.1 - Learn and evaluate kNN algorithm on artificial data + Analyse the properties of KNN

Part 3 - bonus

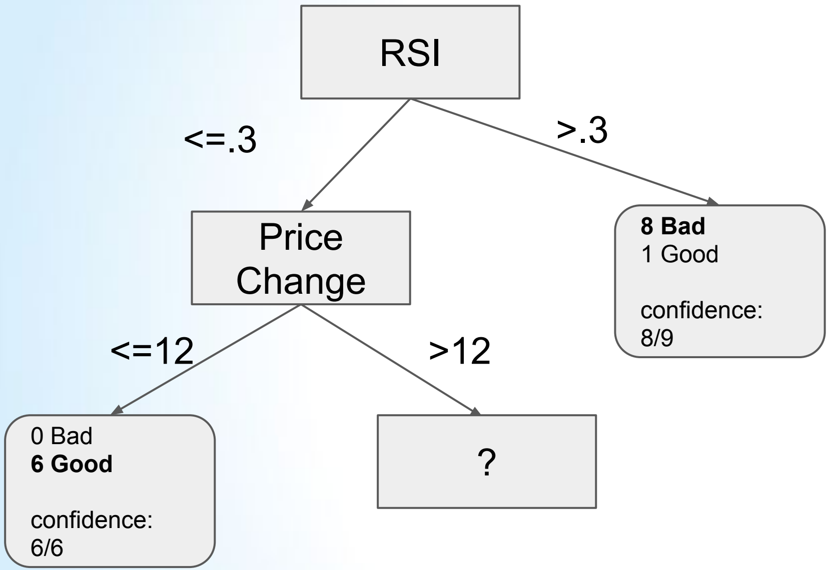
Decision Tree - Adding splits



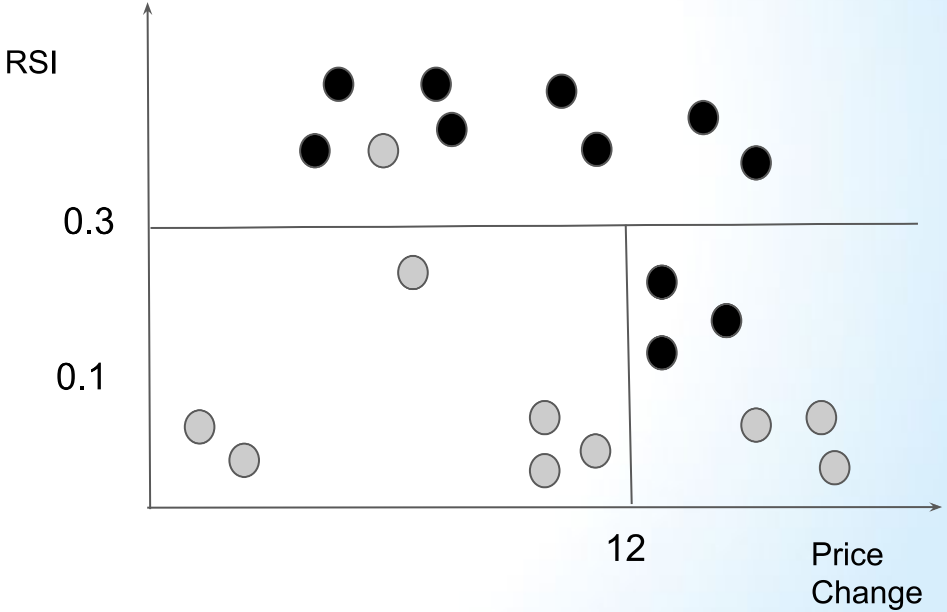
- Nodes = attribute decision
- Leaf nodes = classification decision



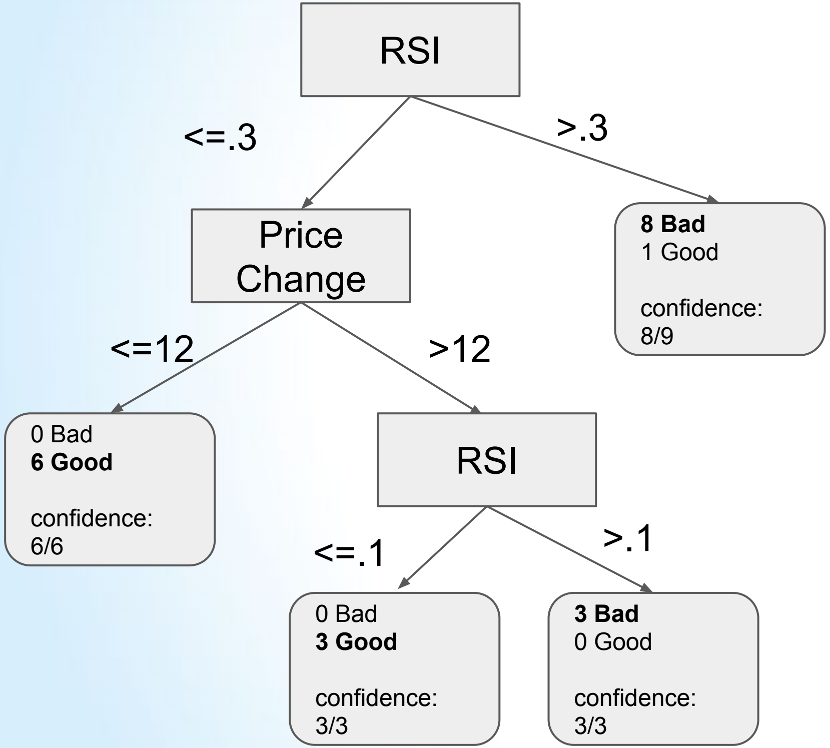
Decision Tree - Adding splits



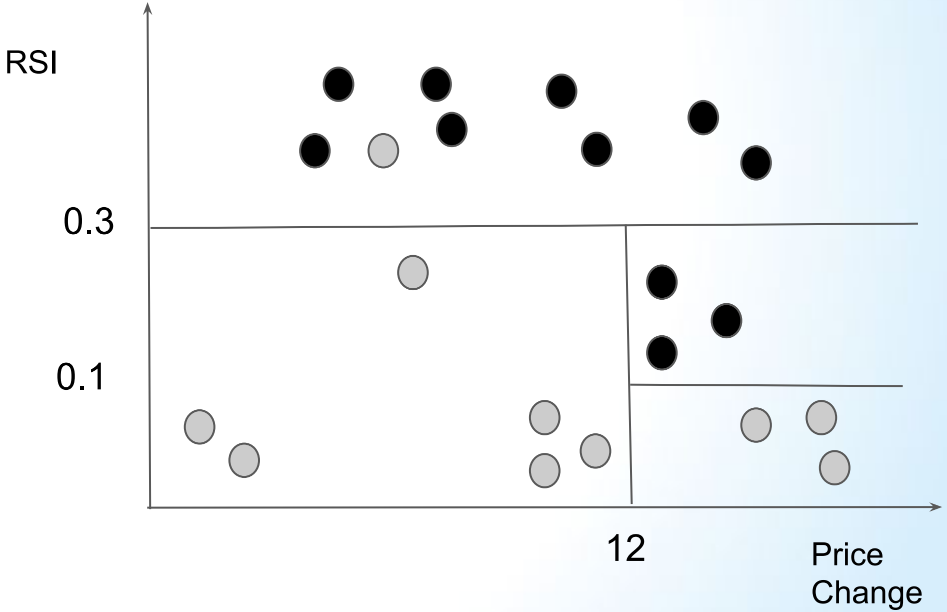
- Nodes = attribute decision
- Leaf nodes = classification decision



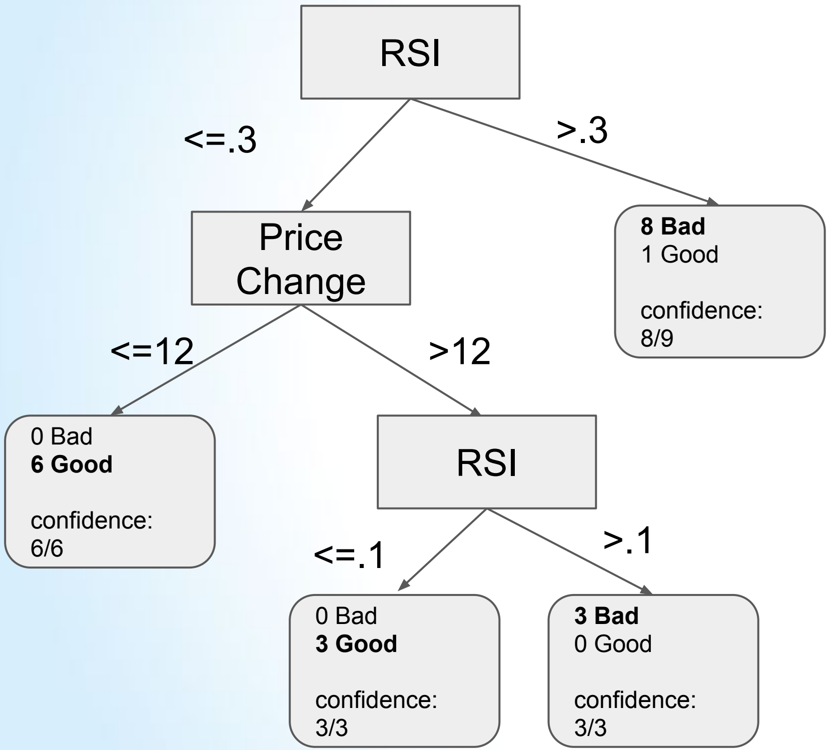
Decision Tree - Adding splits



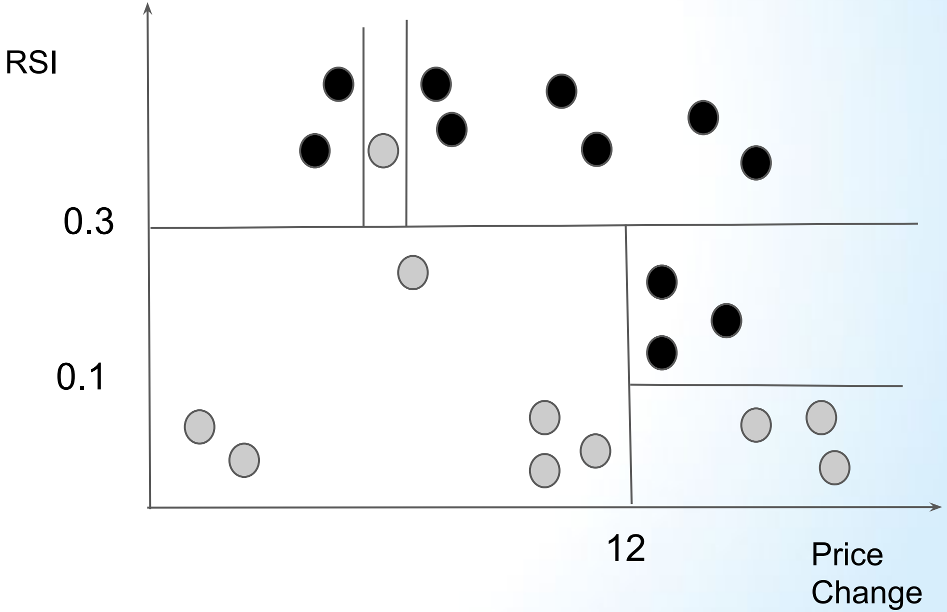
- Nodes = attribute decision
- Leaf nodes = classification decision



Decision Tree - Adding splits



- Nodes = attribute decision
- Leaf nodes = classification decision

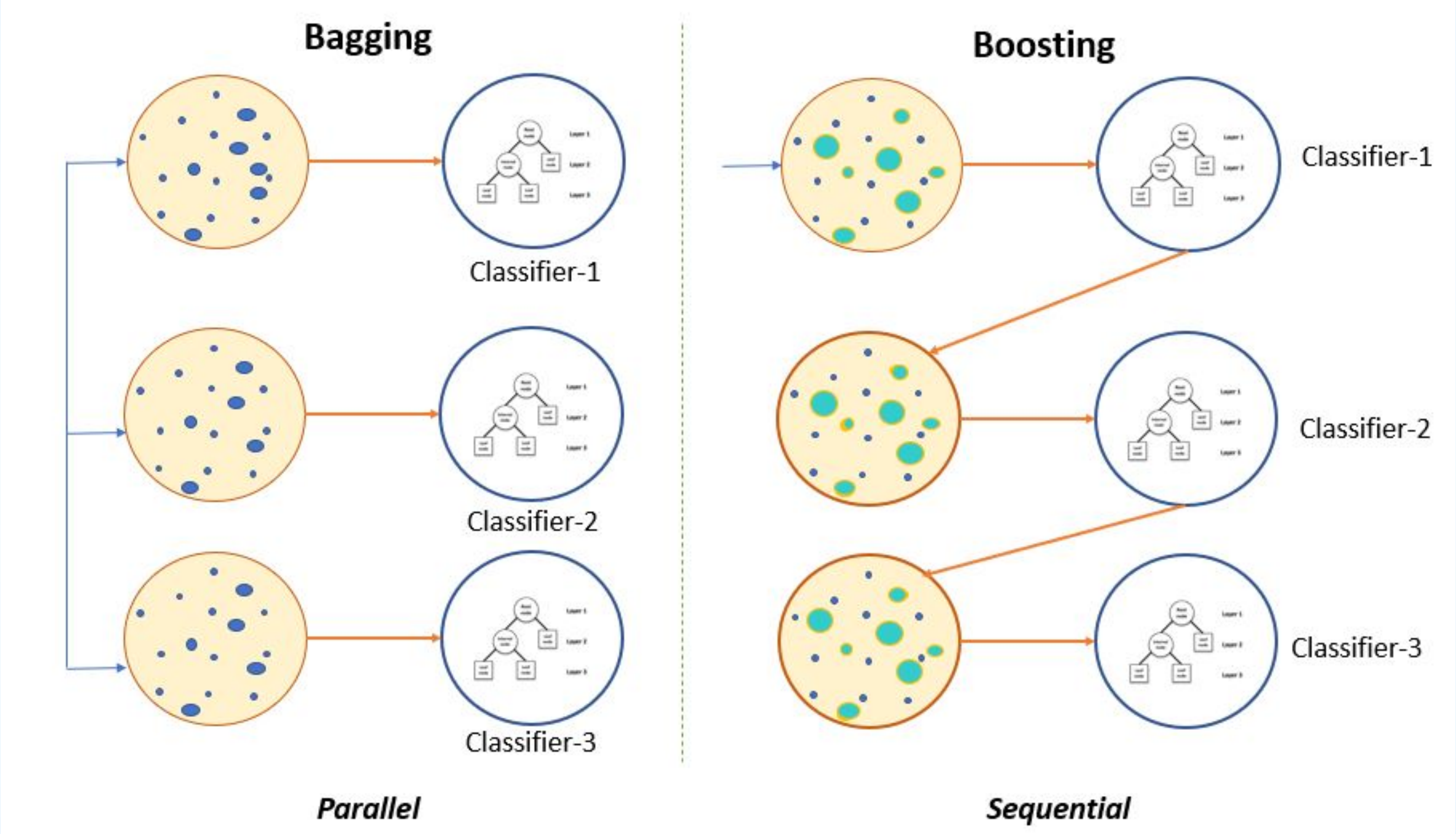


Ensemble Learning

Use multiple learning algorithms to obtain better predictive performance than could be obtained from any of the constituent learning algorithms alone.

Empirically, ensembles tend to yield better results when there is a significant diversity among the models.

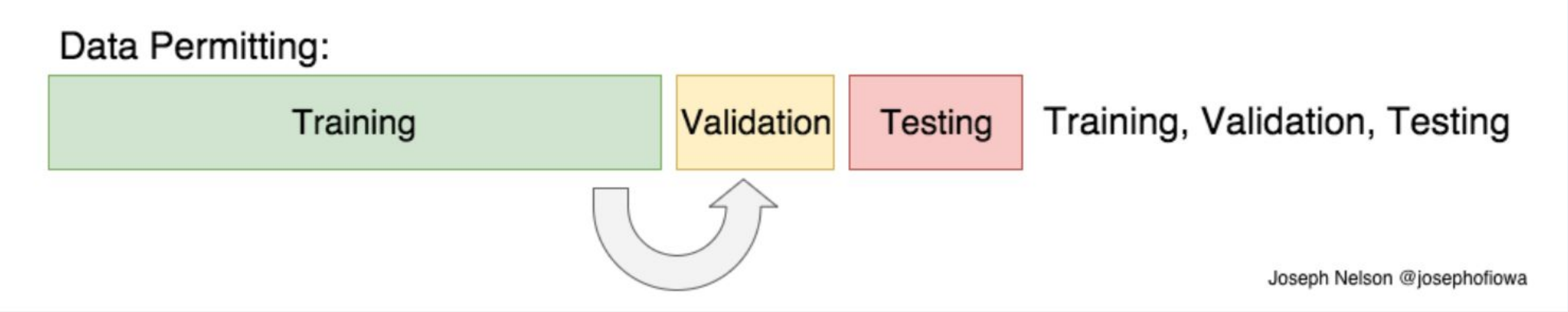
https://en.wikipedia.org/wiki/Ensemble_learning


Bagging
Classifier-1
Classifier-2
Classifier-3
Parallel
Boosting
Classifier-1
Classifier-2
Classifier-3
Sequential

7. How to improve ML models?

Steps to Predictive Modeling

Estimating Performance - Data is Abundant

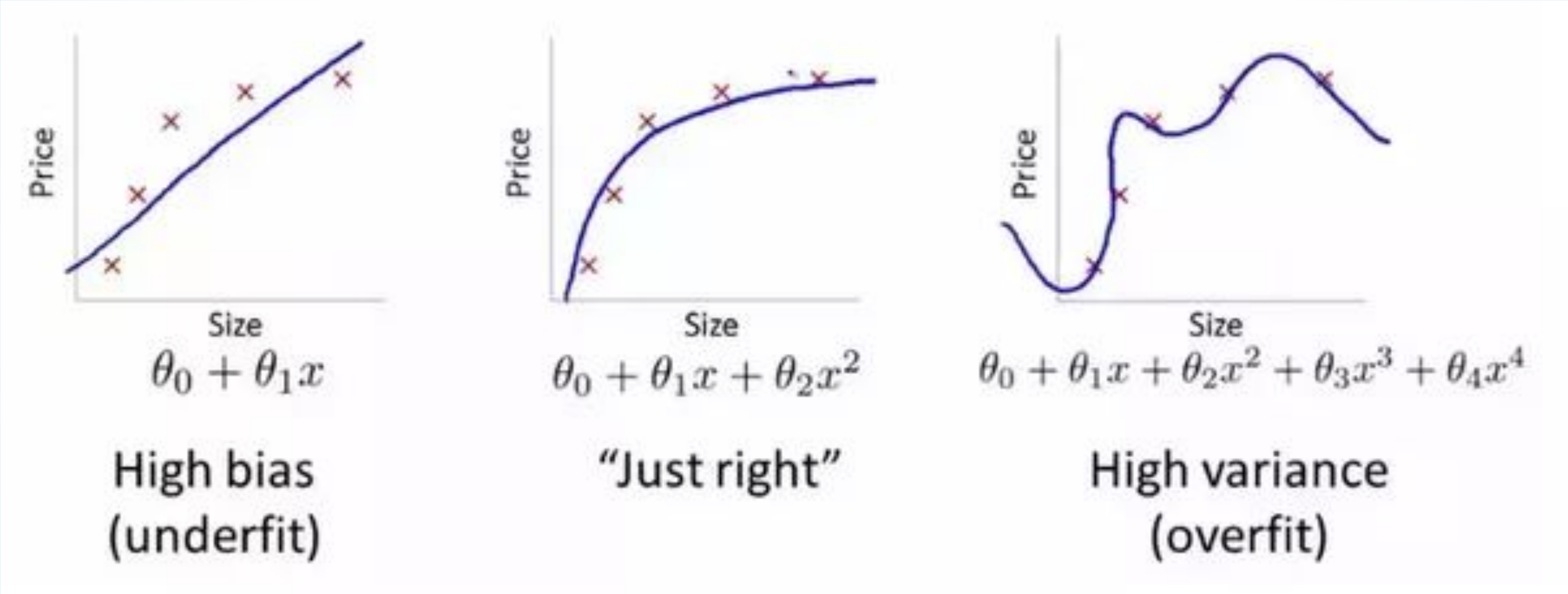


Datasets distribution: Training <> Validation == Test ~ Real world = Random

Hypertuning, Calibration: Validation

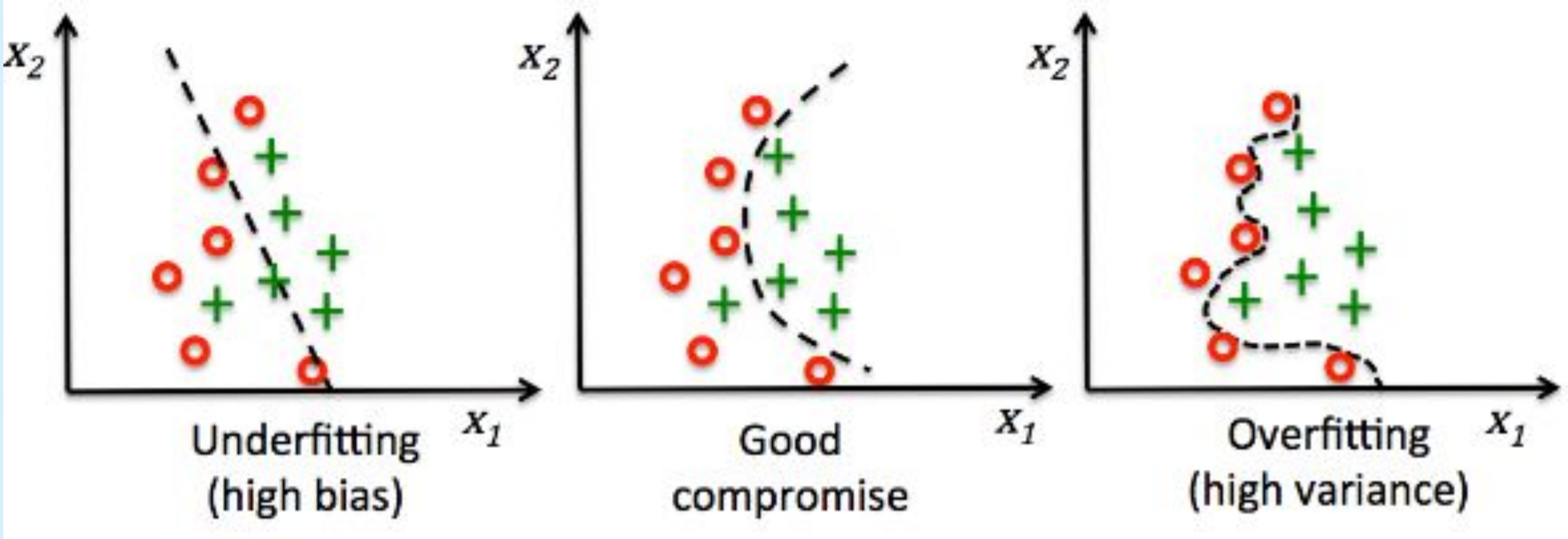
Evaluation: Testing

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

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

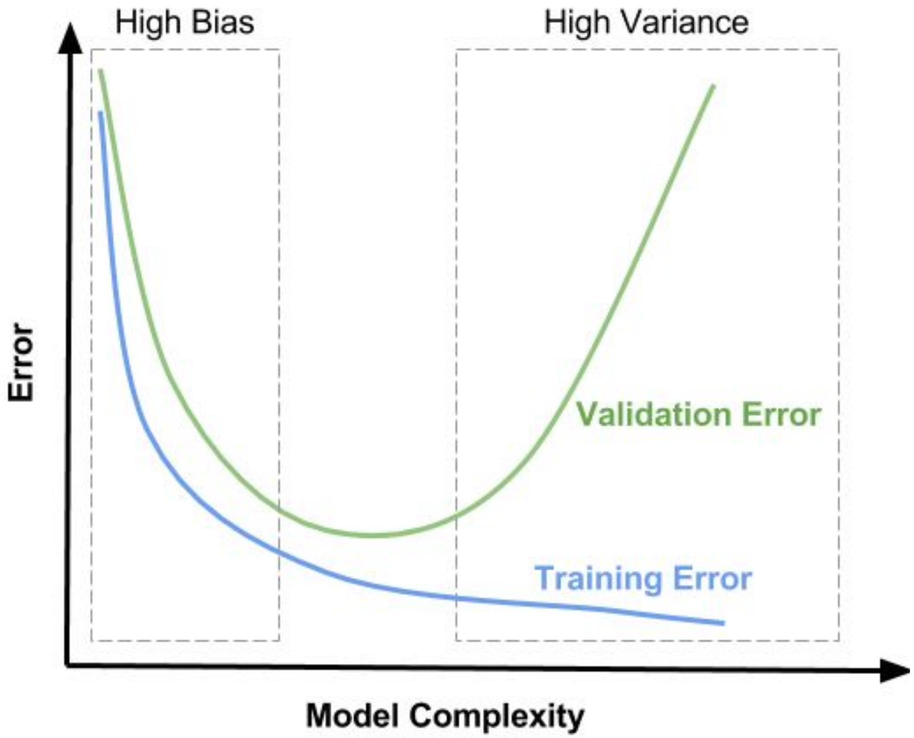
Optional pencil and paper exercise: prove it in details

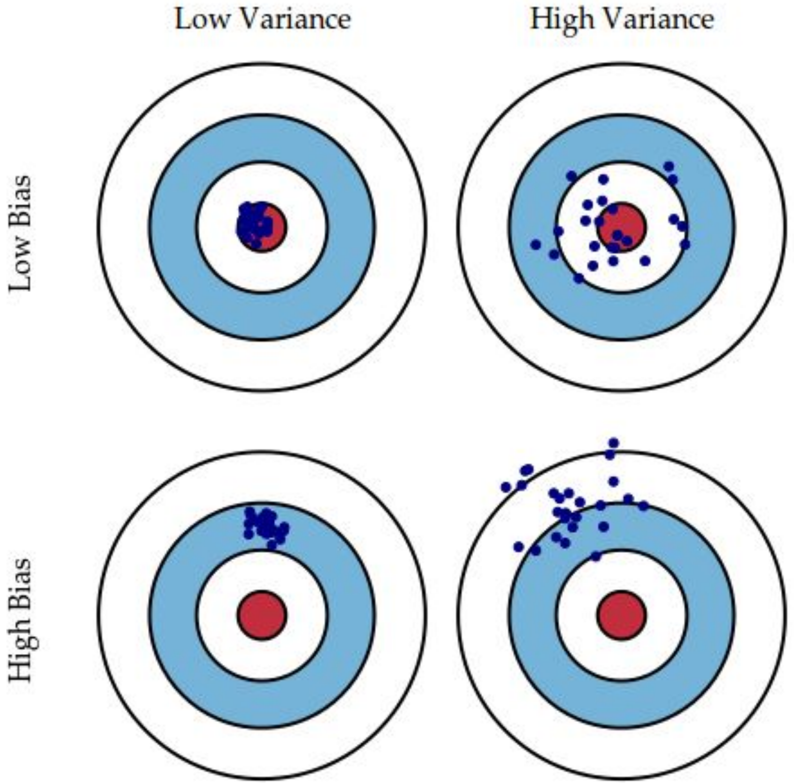
Y-DATA SCHOOL OF DATA SCIENCE

Yandex

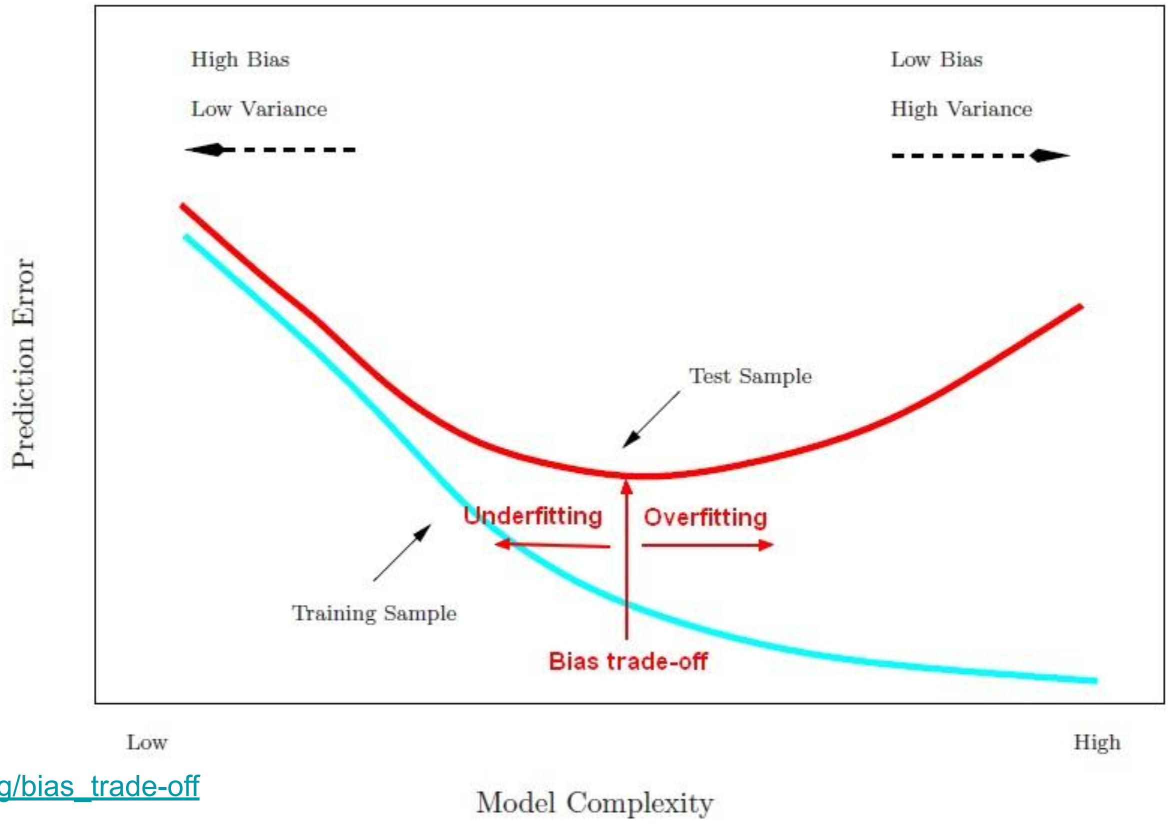
Bias Variance Tradeoff

IMPORTANT NOTICE



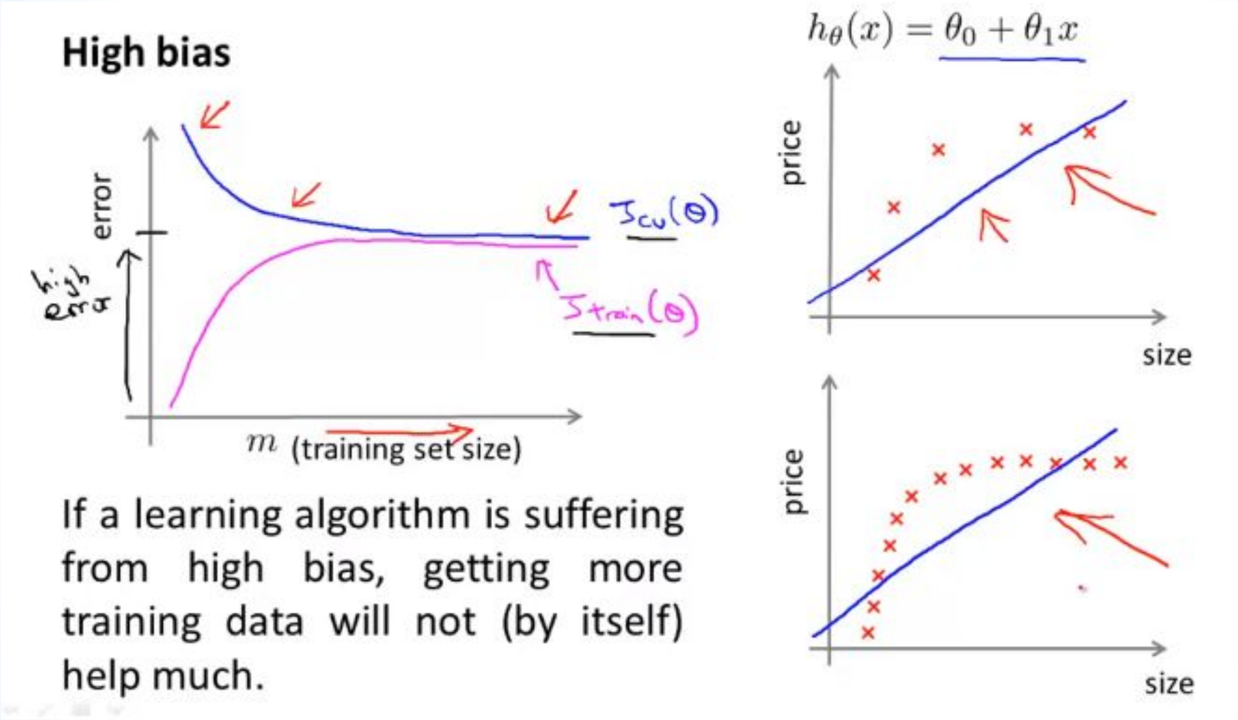


Over and Underfitting



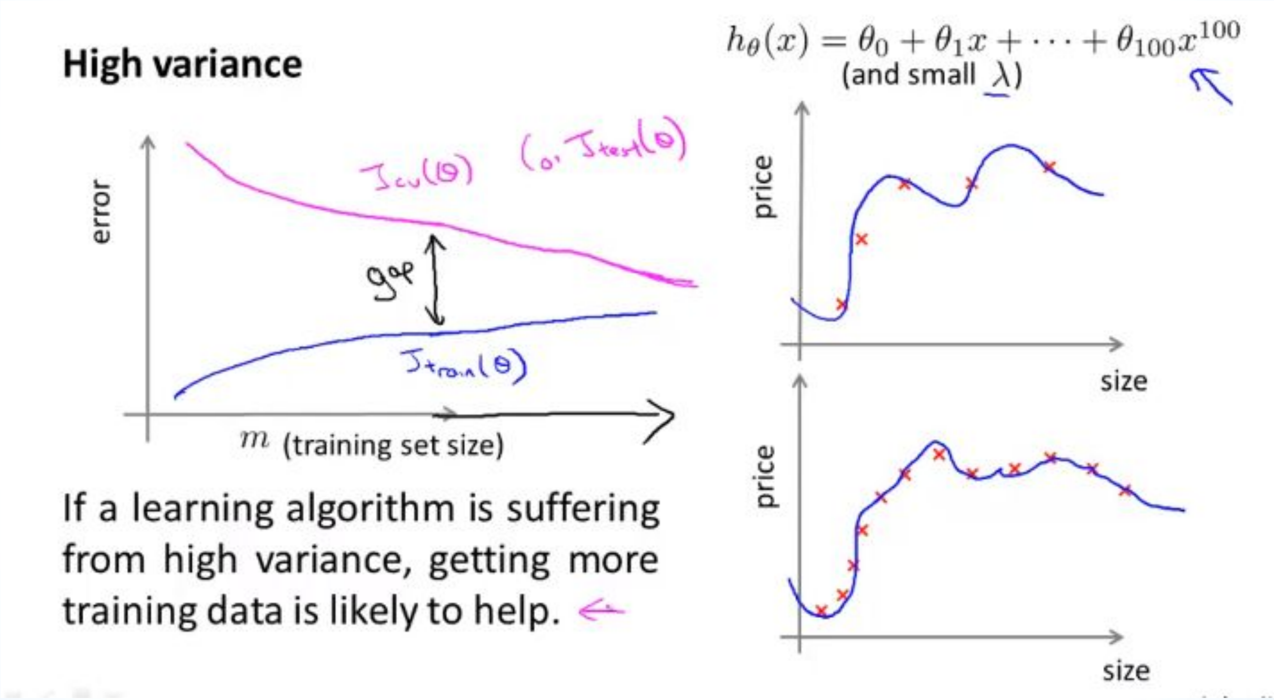
https://gerardnico.com/wiki/data_mining/bias_trade-off

Bias Variance Analysis - Learning Curve



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

Bias Variance Analysis - Learning Curve



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

Hypertunning

Choosing the set of “optimal” hyperparameters for a training algorithm

KNN:

Decision Tree:

Hypertunning

Choosing the set of “optimal” hyperparameters for a training algorithm

KNN:

- Number of K
- Distance function
- Distance function params
- Post processing

Decision Tree:

- Split criteria
- Tree depth
- Minimum instances per split
- Data discretization

Review Homework

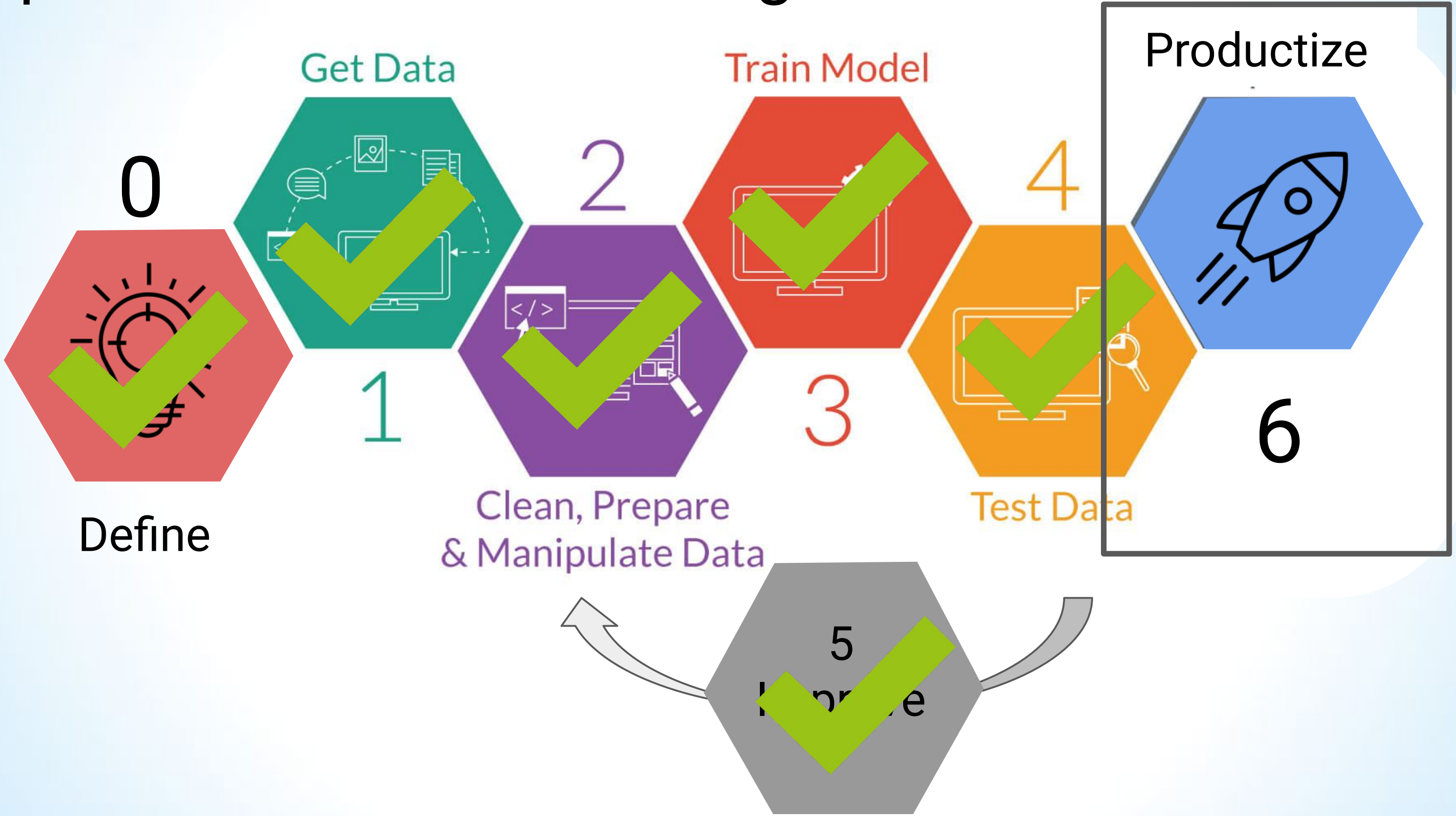
Part 2.2 - Finding the optimal k

Part 2.3 - Using cross validation

Y-DATA SCHOOL OF DATA SCIENCE

Yandex

Steps to Predictive Modeling

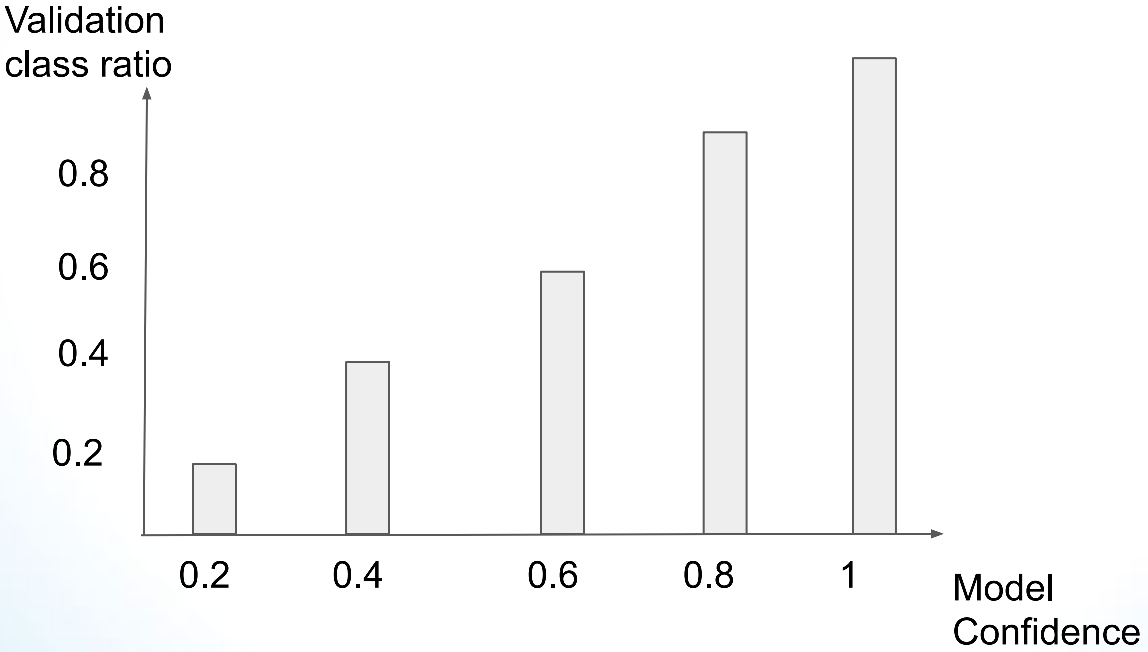


Productization

- Make the Model (and the Data) accessible for prediction
- Monitor model’s performance - what type of performance?
- Concept drift - changes in data, labels and their relation
- Feedback loop - label more data and retrain the model

Calibration

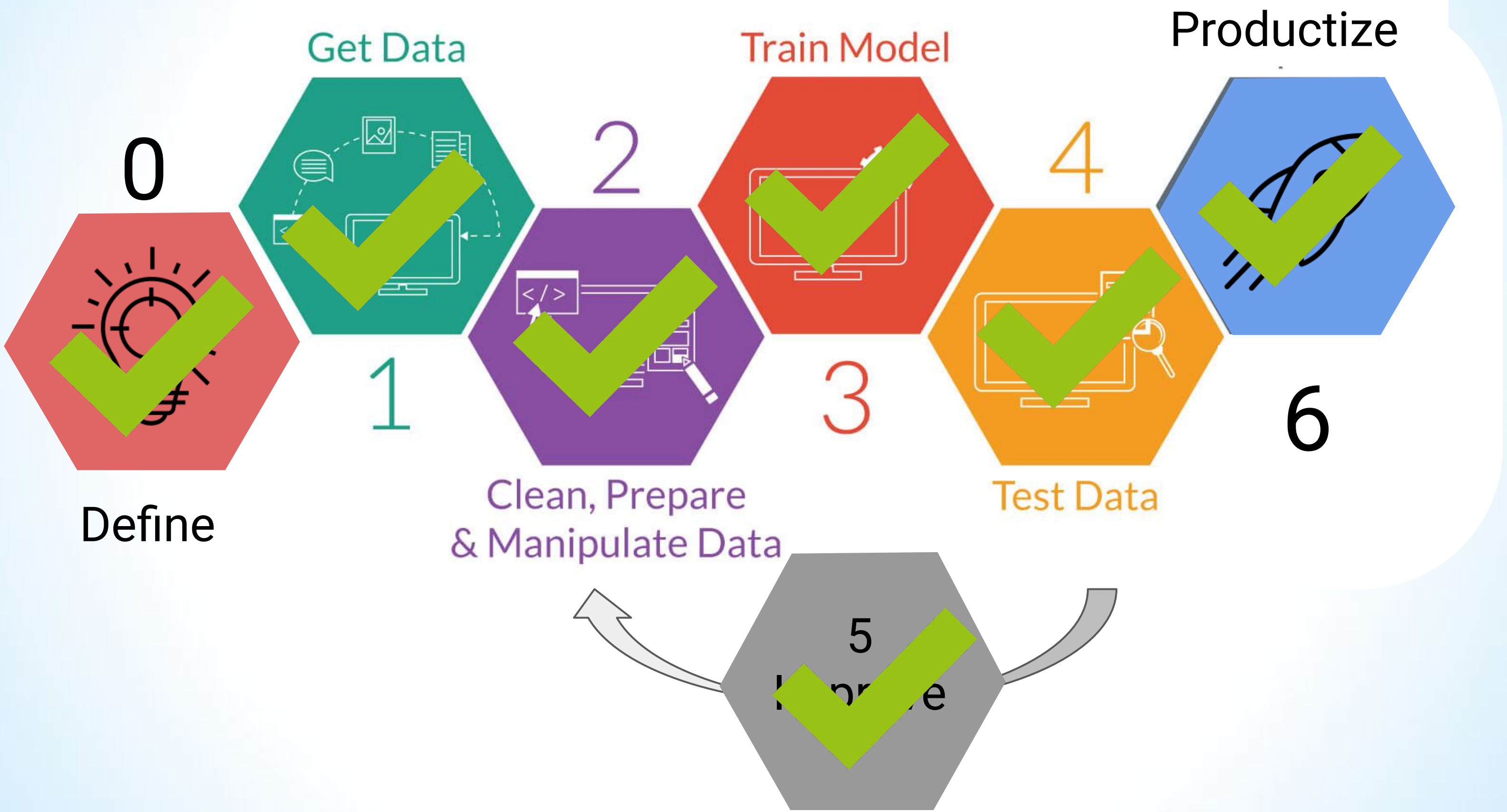
Calibration in classification means turning transform classifier scores into class membership probabilities.



Y-DATA SCHOOL OF DATA SCIENCE

Yandex

Steps to Predictive Modeling



TL,DR

- Data Science is a practical profession
- There are many topics -> understand over memorize
- Understand the product task + Have a clear end2end intuition of the project
- Build a Baseline model as soon as possible!
- Keep Asking questions
- There is no silver bullet / free lunch theory

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. Introduction to Statistical Learning Theory.pdf
5. On the Surprising Behavior of Distance Metrics in High Dimensional Space.pdf
6. Statistical learning theory - a primer.pdf
7. Statistical Machine Learning- Introduction.pdf
8. 2012b A Geometrical Explanation of Stein Shrinkage.pdf
9. INADMISSIBILITY OF THE USUAL ESTIMATOR FOR THE MEAN OF MULTIVARIATE NORMAL DISTRIBUTION - STEIN.pdf
10. THE USE OF MULTIPLE MEASUREMENTS IN TAXONOMIC PROBLEMS - FISHER - 1936 - Annals of Eugenics - Wiley Online Library.pdf

QA