JDBI Talk

Lokesh Nagalapatti

Talk Outline

- Data is an important asset
- Overview of Data Quality
- Some popular Data Quality issues
  - Data Cleaning
  - Class Imbalance
  - Label Noise
  - Data Valuation
  - Data Homogeneity
  - Data Transformations
- An intuitive ML algorithm: Decision Tree

Data is the new oil

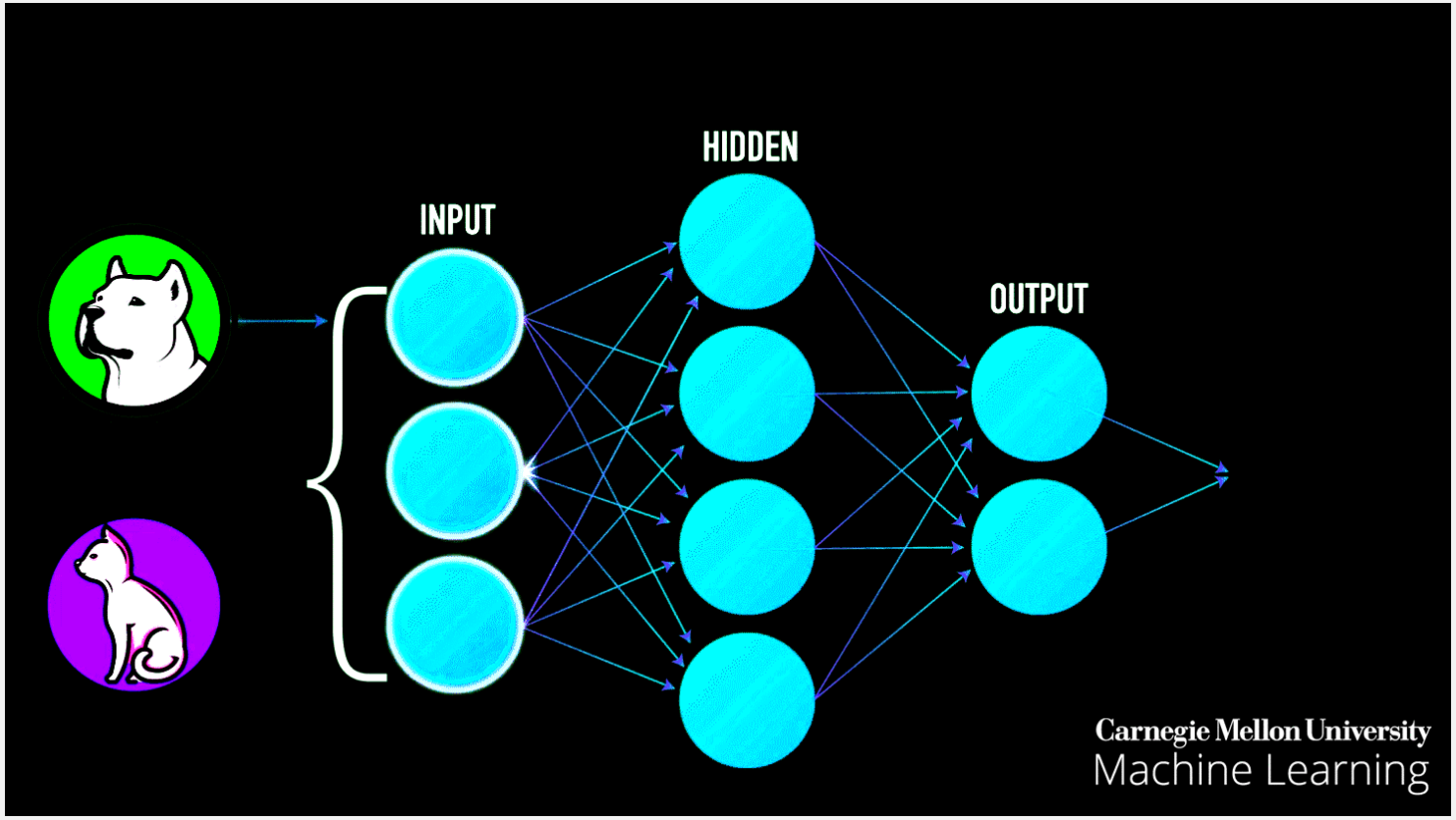
- William Edwards Deming, "**In God we trust; all others must bring data**."

https://medium.com/@adeolaadesina/data-is-the-new-oil-2947ed8804f6

Why do we say that?

- Data is an essential resource that powers the information economy in much the way that oil has fueled the industrial economy
- Information can be extracted from data just as energy can be extracted from oil
- Data flows like oil but we must “drill down” into data to extract value from it
- Oil is a scarce resource. Data isn’t just abundant, it is a cumulative resource

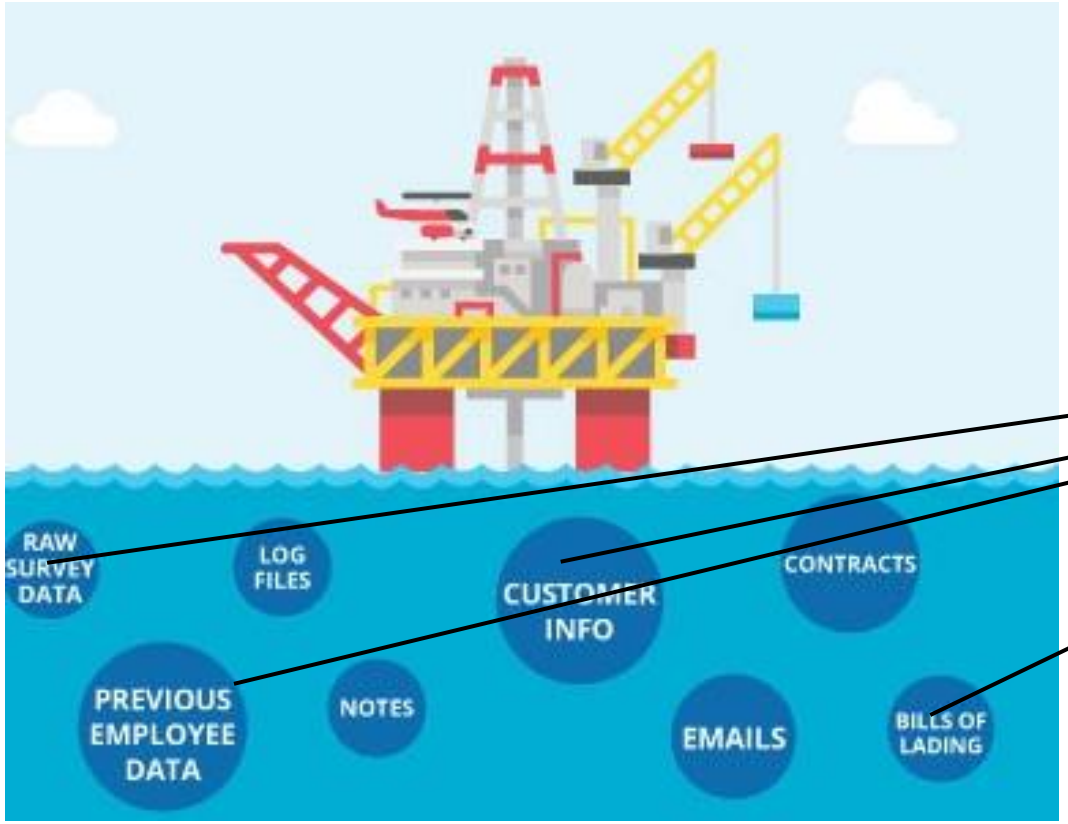
What is Machine Learning?



- **Machine learning is the study of computer algorithms that allow computer programs to automatically improve through data.**

- **Here is the catch: Good the data, better the algorithm.**

Real world scenario



Real world data is often noisy and is not readily consumable by ML algorithms

Hence, we would discuss first on “Overview and importance of **Data Quality** for AI”

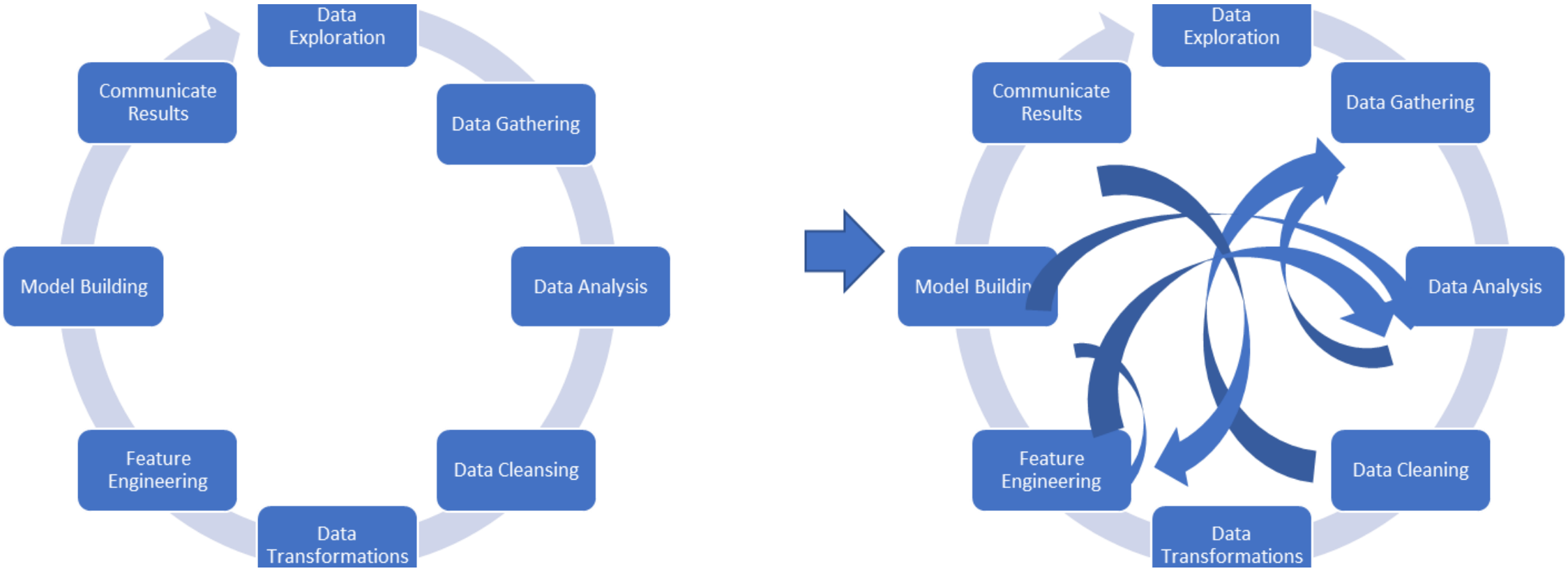
Data Preparation in Machine Learning

- "**Data collection and preparation** are typically
- the **most time-consuming activities** in developing
- an AI-based application, much more so than
- selecting and tuning a model." – MIT Sloan Survey

- https://sloanreview.mit.edu/projects/reshaping-business-with-artificial-intelligence/

Data preparation accounts for about **80%** of the work of data scientists" - Forbes

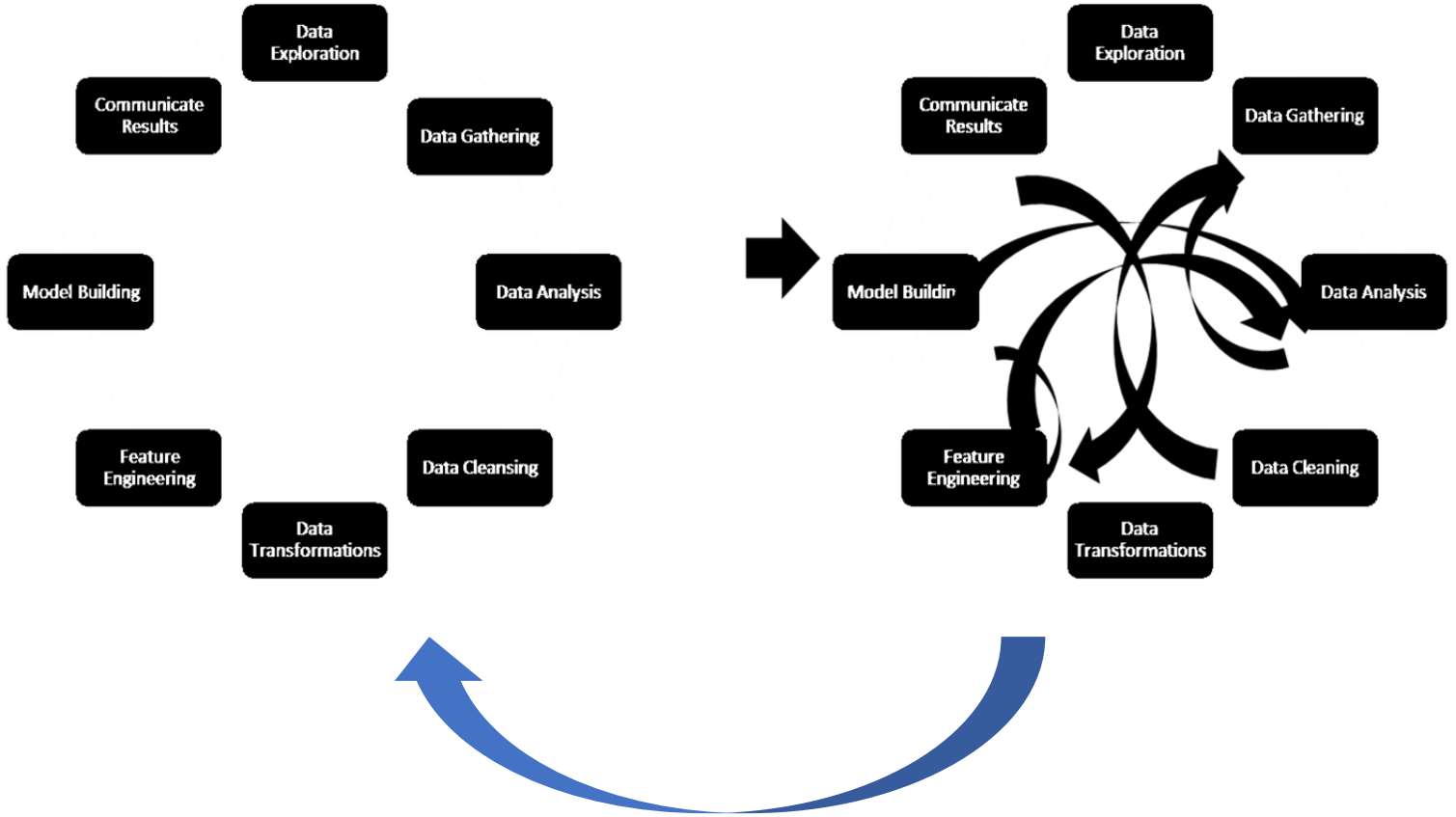
https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#70d9599b6f63



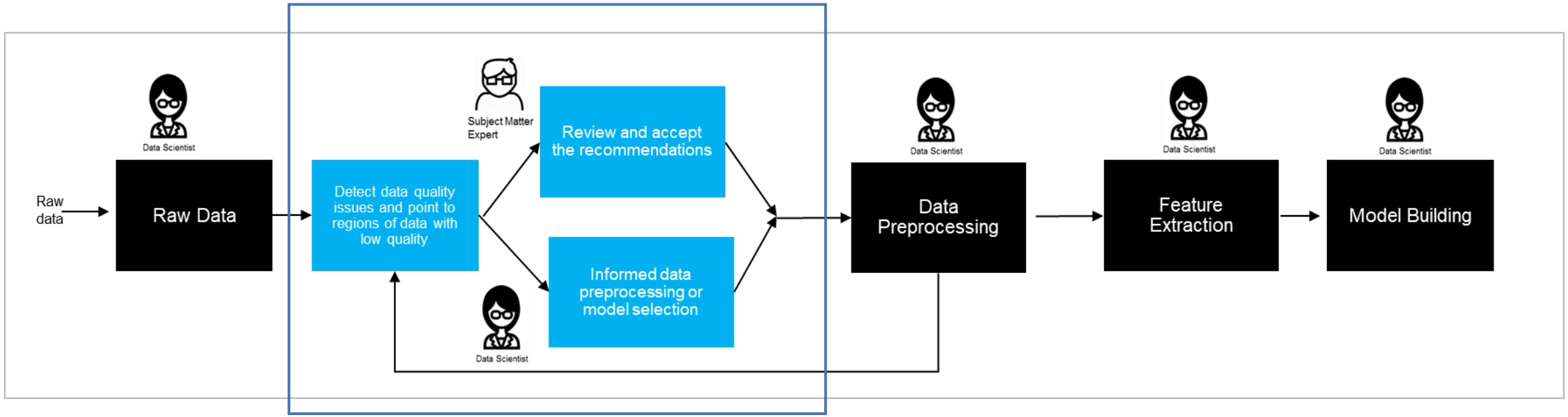
Challenges with Data Preparation

Data Quality Analysis can help..

Know	Know the issues in the data beforehand, example noise in labels, overlap between classes.
Make	Make informed choices for data preprocessing and model selection
Reduce	Reduce turn around time for data science projects.



To put it all together



- Need for algorithms that can assess training datasets
- Across modalities.. Structured/unstructured/timeseries etc
- Allow for complex interaction between the different personas and human in loop techniques

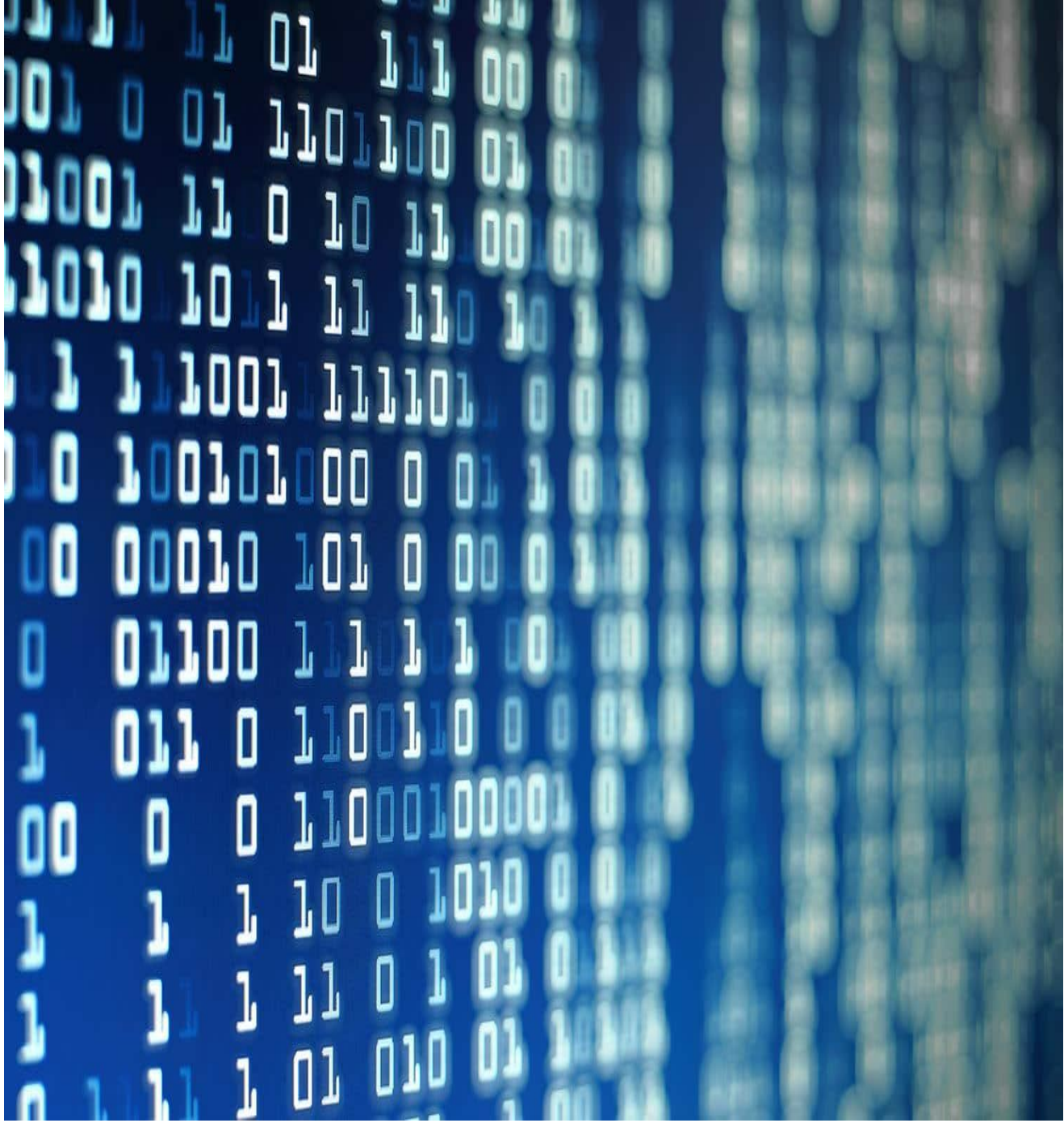
- **Need for automation**

To summarize:

- Lot of progress in last several years on improving ML algorithms including building automated machine learning toolkits (AutoML)
- However, Quality of a ML model is directly proportional to Quality of Data
- Hence, there is a need for systematic study of measuring quality of data with respect to machine learning tasks.

Some popular Data Quality Issues

- **Data Cleaning**
- Class Imbalance
- Label Noise
- Data Valuation
- Data Homogeneity
- Data Transformations

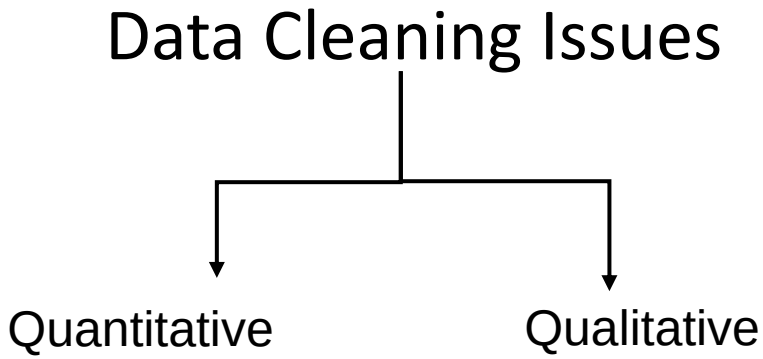


Source: https://www.analyticsinsight.net/data-literacy-helping-enterprises-lead-with-data-through-challenging-times/

Data Cleaning

- What are some common data cleaning techniques used for machine learning?
- Do data cleaning techniques always help in building a machine learning pipeline
- Joint cleaning and model building techniques

Common Data Cleaning Techniques



- Missing Values
- Outliers
- Noise in the data
- Noise in labels
- Data Linters
- …

Integrity Constraints based errors (not covered)
- Functional Dependency Constraints
- Denial Constraints
- Others ..

Some insights on Data Cleaning

- [LXJ19] [LAU19]

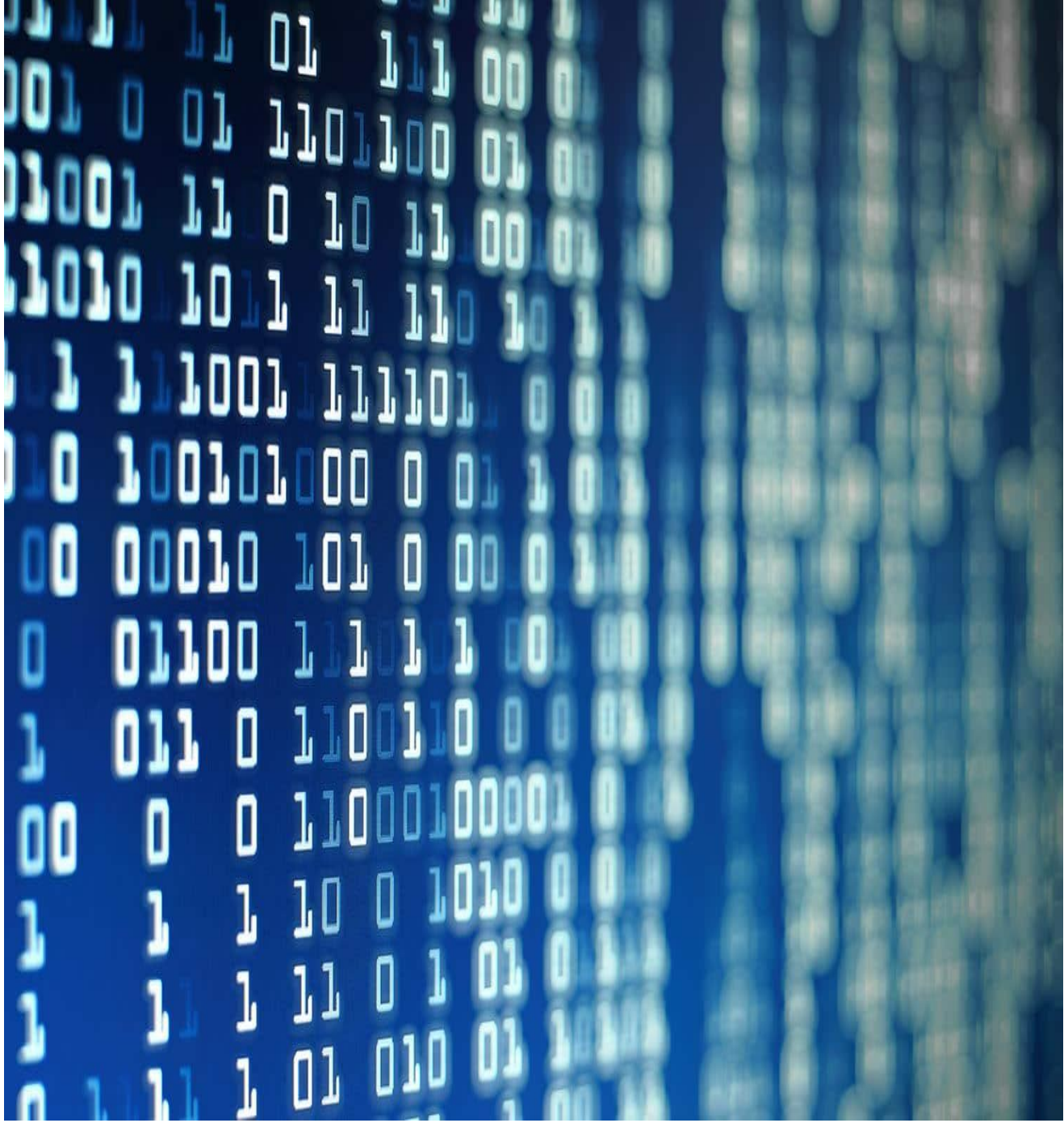
- Data cleaning does not necessarily improve the quality of downstream ML models

- Impact depends on different factors:
  - Cleaning algorithm and its set of parameters
  - ML model dependent
  - Order of cleaning operators

- Model selection and cleaning algorithm can increase robustness of impacts – no one solution!

Data Quality Metrics

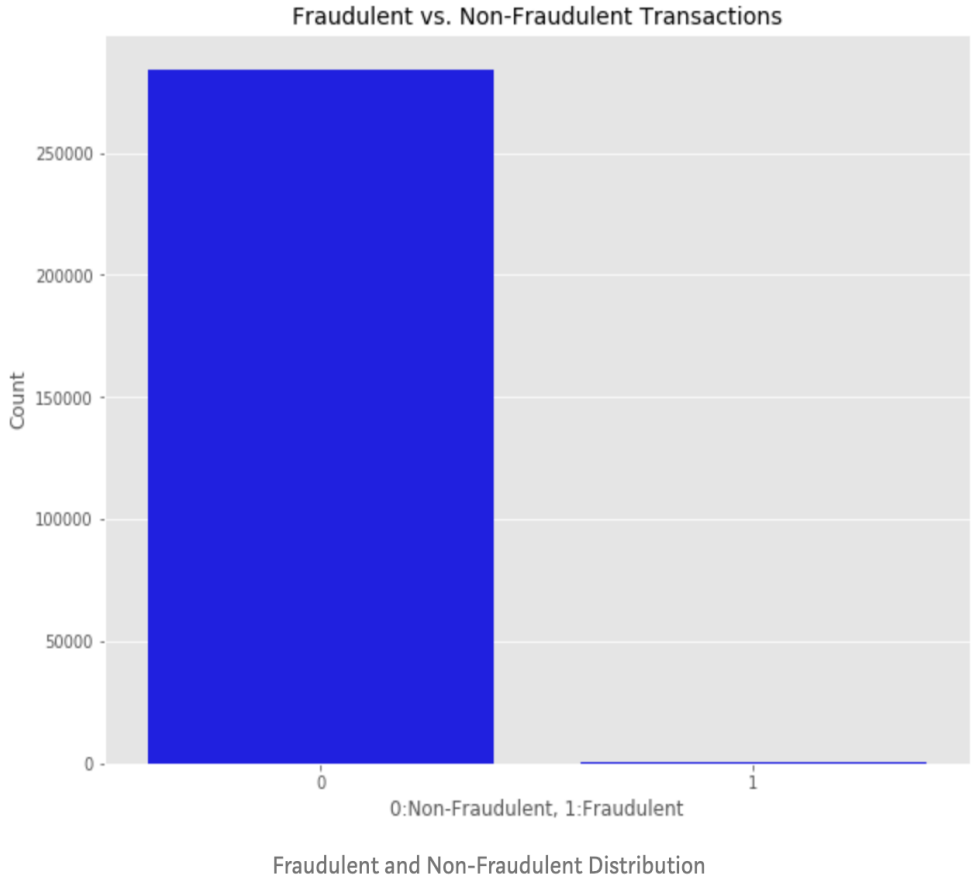
- We will cover the following topics:
  - Data Cleaning
  - **Class Imbalance**
  - Label Noise
  - Data Valuation
  - Data Homogeneity
  - Data Transformations



Source: https://www.analyticsinsight.net/data-literacy-helping-enterprises-lead-with-data-through-challenging-times/

Class Imbalance

- Unequal distribution of classes within a dataset



Source: https://towardsdatascience.com/credit-card-fraud-detection-a1c7e1b75f59

Why it happens?

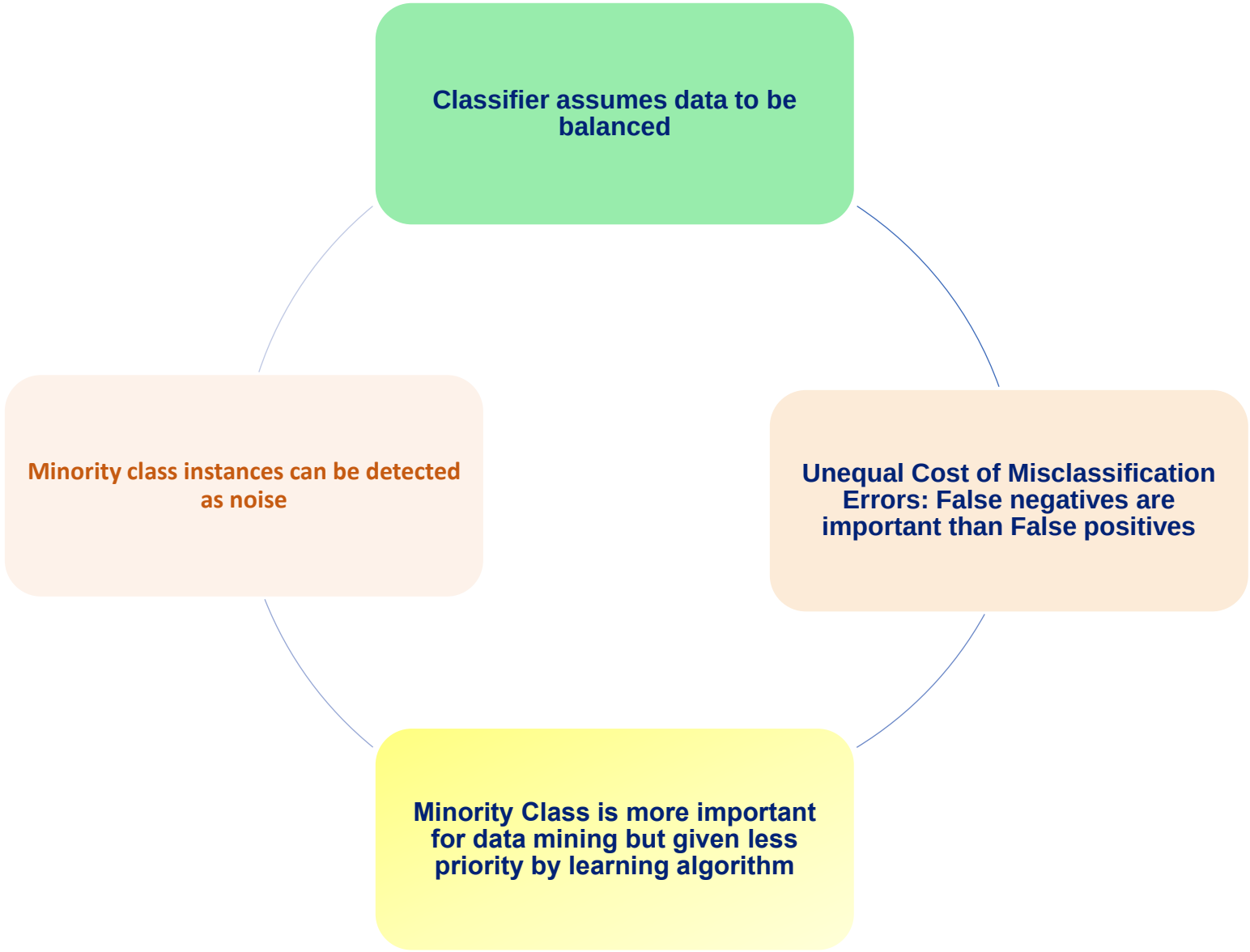
[Kra16]

- Expected in domains where data with one pattern is more common than other.

Application area	Problem description
Activity recognition [19]	Detection of rare or less-frequent activities (multi-class problem)
Behavior analysis [3]	Recognition of dangerous behavior (binary problem)
Cancer malignancy grading [30]	Analyzing the cancer severity (binary and multi-class problem)
Hyperspectral data analysis [50]	Classification of varying areas in multi-dimensional images (multi-class problem)
Industrial systems monitoring [44]	Fault detection in industrial machinery (binary problem)
Sentiment analysis [65]	Emotion and temper recognition in text (binary and multi-class problem)
Software defect prediction [48]	Recognition of errors in code blocks (binary problem)
Target detection [45]	Classification of specified targets appearing with varied frequency (multi-class problem)
Text mining [39]	Detecting relations in literature (binary problem)
Video mining [20]	Recognizing objects and actions in video sequences (binary and multi-class problem)

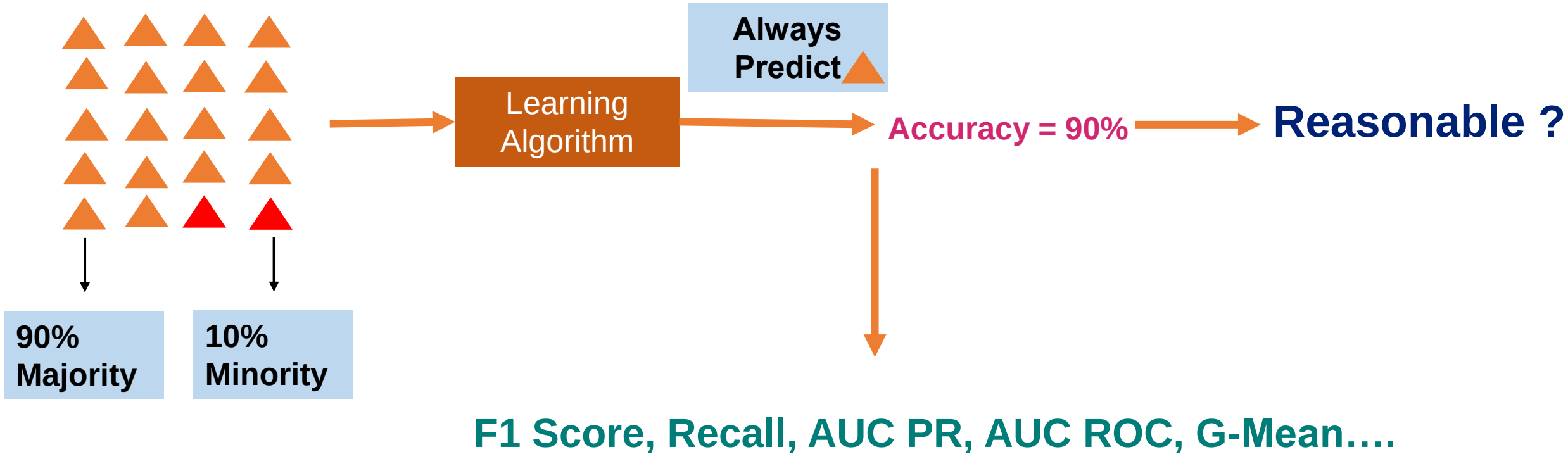
Source: Krawczyk et al, 2016. Learning from imbalanced data: open challenges and future direction

Why Imbalanced Classification is Hard?



Evaluation Metrics for Imbalanced Datasets

- **Accuracy Paradox**



Source: https://en.wikipedia.org/wiki/Accuracy_paradox

Factors affecting class imbalance

- Imbalance Ratio
- Overlap between classes
- Smaller sub-concepts
- Dataset Size
- Label Noise
- Combination
- …

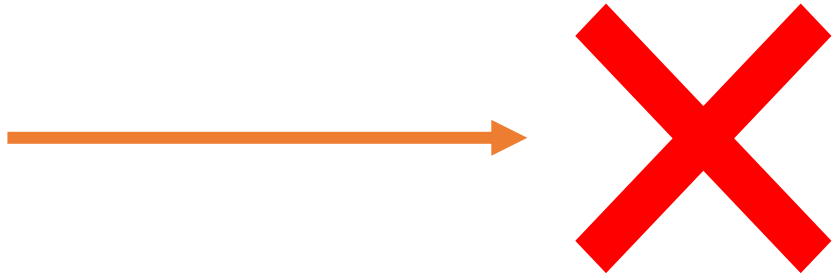
To overcome Imbalance: Sampling datasets

- Oversampling
- Undersampling
- Ensemble Based Techniques

Does imbalance recovery method always help?

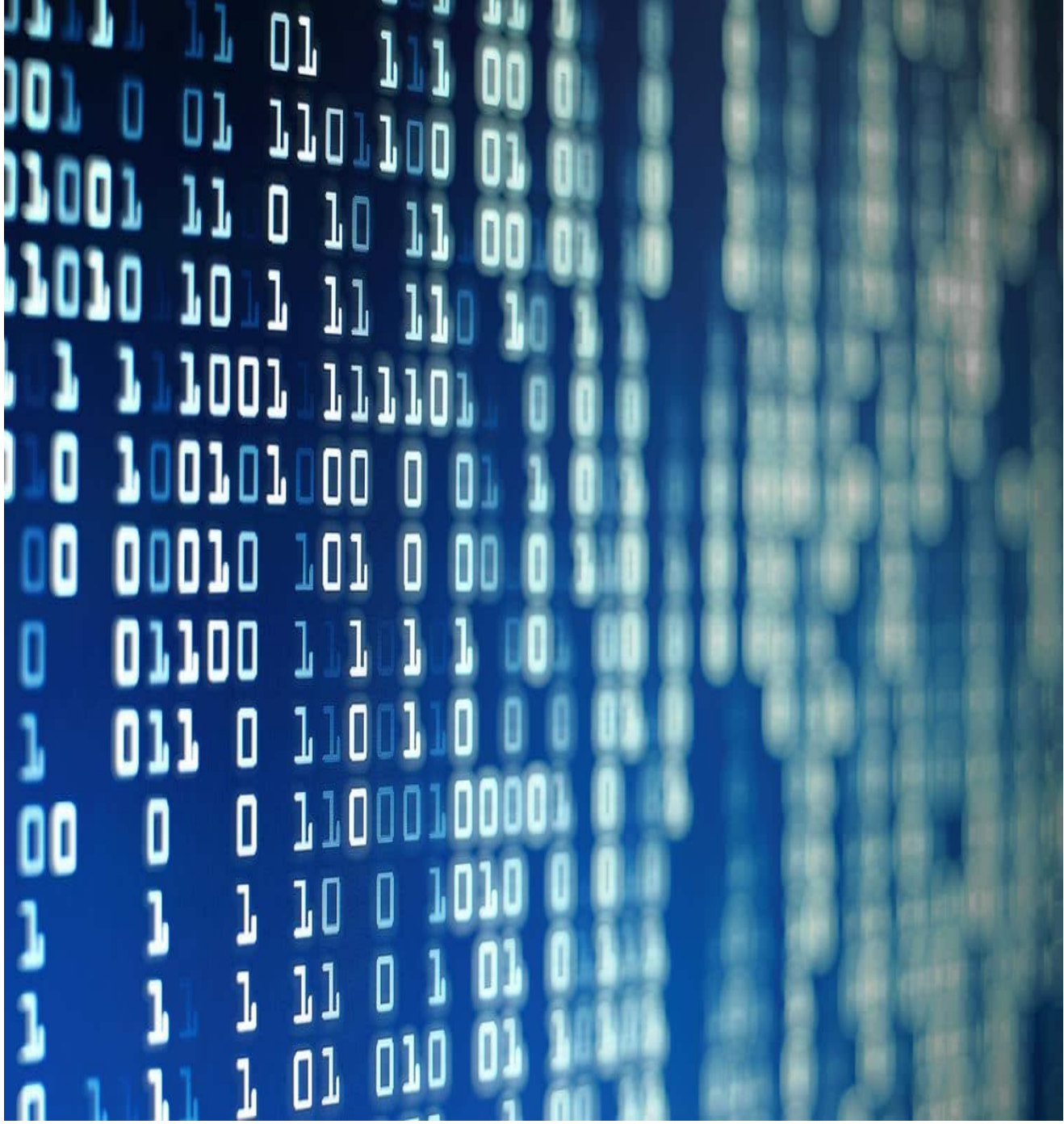
OR

Does the Impact of imbalance recovery method same on all datasets.



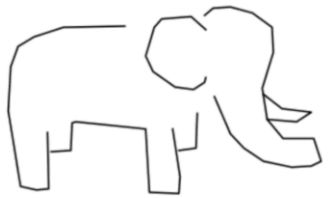
Data Quality Metrics

- We will cover the following topics:
  - Data Cleaning
  - Class Imbalance
  - **Label Noise**
  - Data Valuation
  - Data Homogeneity
  - Data Transformations



Source: https://www.analyticsinsight.net/data-literacy-helping-enterprises-lead-with-data-through-challenging-times/

Label Noise



Dataset: **Google Quickdraw!**
Given Label: **Mosquito**

Dataset: **MNIST**
Given Label: **5**

"I've had this for over a year and it works very well. I am very happy with this purchase."

Dataset: **Amazon Reviews**
Given Label: **1 star review**

Source: https://l7.curtisnorthcutt.com/confident-learning

- Most of the large data generated or annotated have some noisy labels.
- In this metric we discuss "How one can identify these label errors and correct them to model data better?".

SepalLength2	SepalWidth	PetalLength	PetalWidth	Species
5.1	3.5	1.4	0.2	Iris-setosa
4.9	3.0	1.4	0.2	Iris-setosa
4.7	2.1	1.3	0.1	Iris-setosa
6.2	3.1	4.1	1.2	Iris-virginica
6.7	3.0	4.4	1.2	Iris-virginica
6.3	3.1	4.2	1.3	Iris-virginica
6.5	3.2	4.4	1.2	Iris-setosa

Given Label – Iris-setosa
Correct Label –Iris-virginica (based on attributes analysis)

There are atleast 100,000 label issues is ImageNet!

Effects of Label Noise

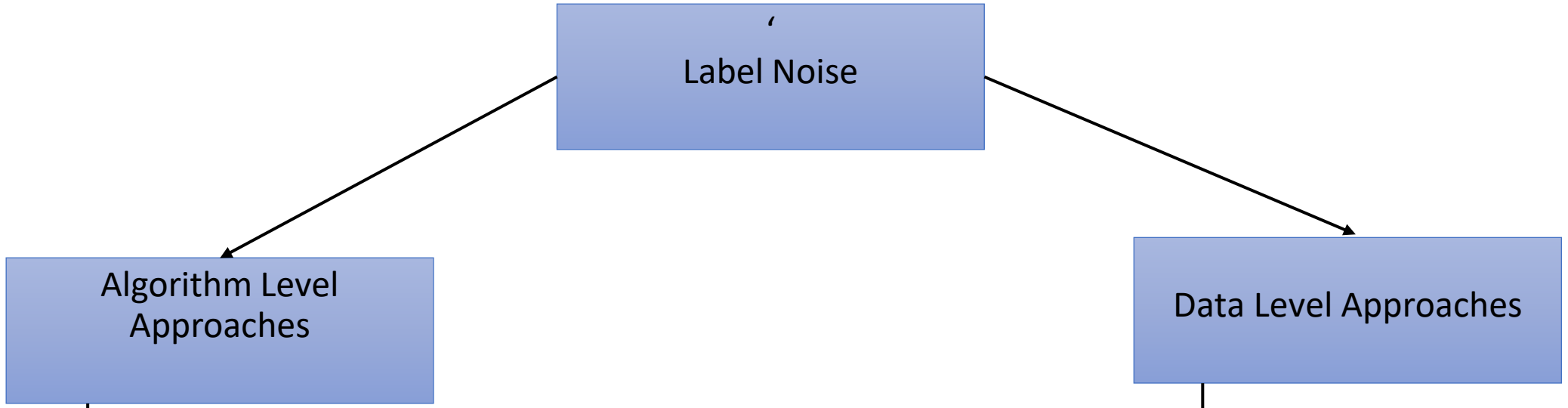
Possible Sources of Label Noise:

- Insufficient information provided to the labeler
- Errors in the labelling itself
- Subjectivity of the labelling task
- Communication/encoding problems

Label noise can have several effects:

- Decrease in classification performance
- Pose a threat to tasks like feature selection
- In online settings, new labelled data may contradict the original labelled data

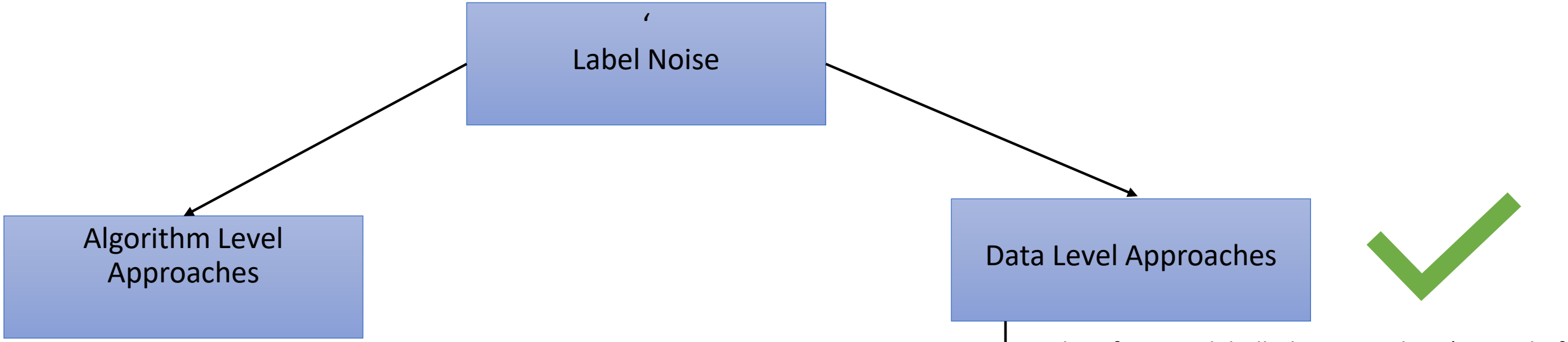
Label Noise Techniques



- Designing robust algorithms that are insensitive to noise
- Not directly extensible to other learning algorithms
- Requires to change an existing method, which neither is always possible nor easy to develop

- Filtering out noise before passing to underlying ML task
- Independent of the classification algorithm
- Helps in improving classification accuracy and reduced model complexity.

Label Noise Techniques

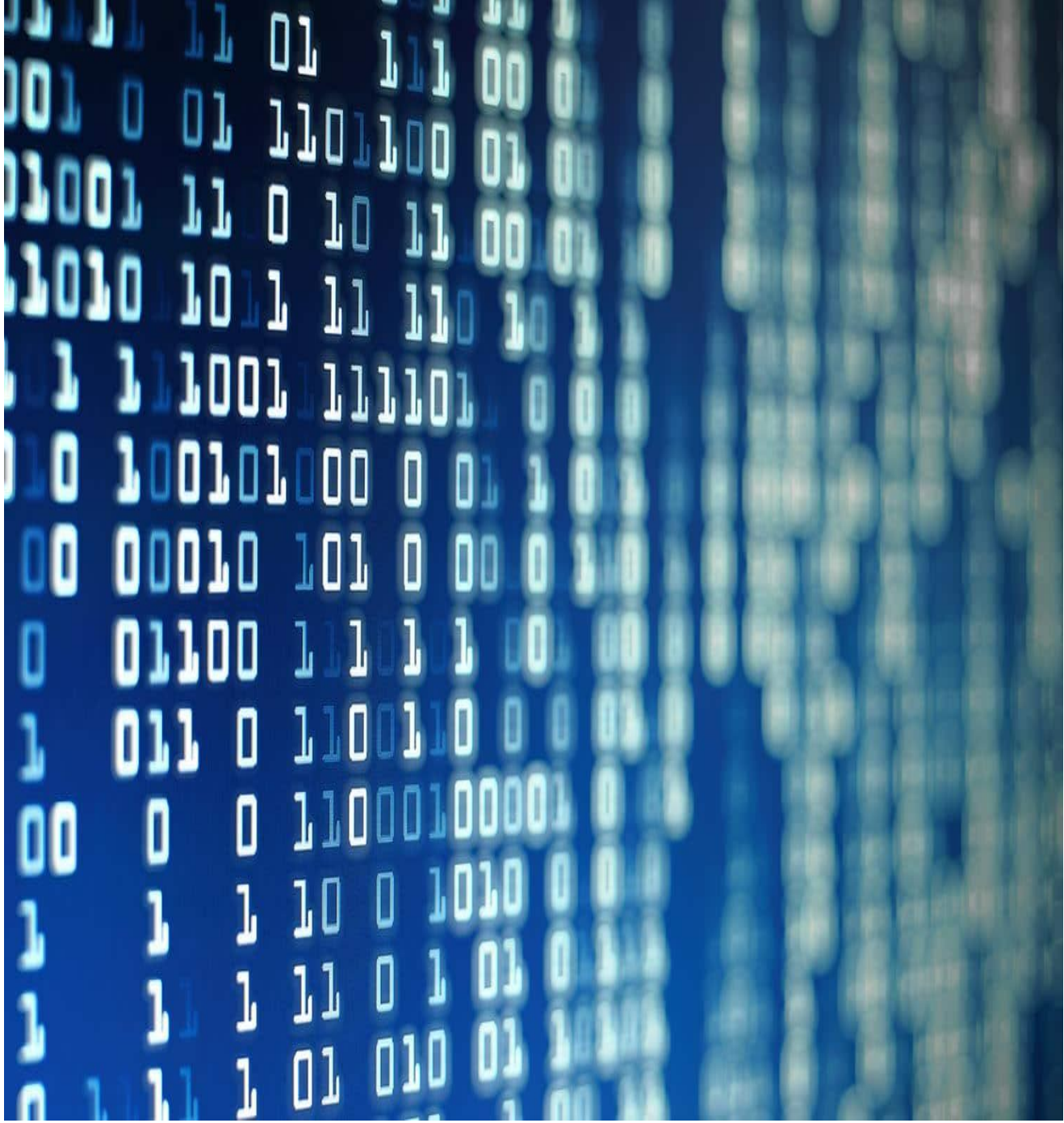


- Learning with Noisy Labels (NIPS-2014)
- Robust Loss Functions under Label Noise for Deep Neural Networks (AAAI-2017)
- Probabilistic End-To-End Noise Correction for Learning With Noisy Labels (CVPR-2019)
- Can Gradient Clipping Mitigate Label Noise? (ICLR-2020)

- Identifying mislabelled training data (Journal of artificial intelligence research 1999)
- On the labeling correctness in computer vision datasets (IAL 2018)
- Finding label noise examples in large scale datasets (SMC 2017)
- Confident Learning: Estimating Uncertainty in Dataset Label (Arxiv -2019)

Data Quality Metrics

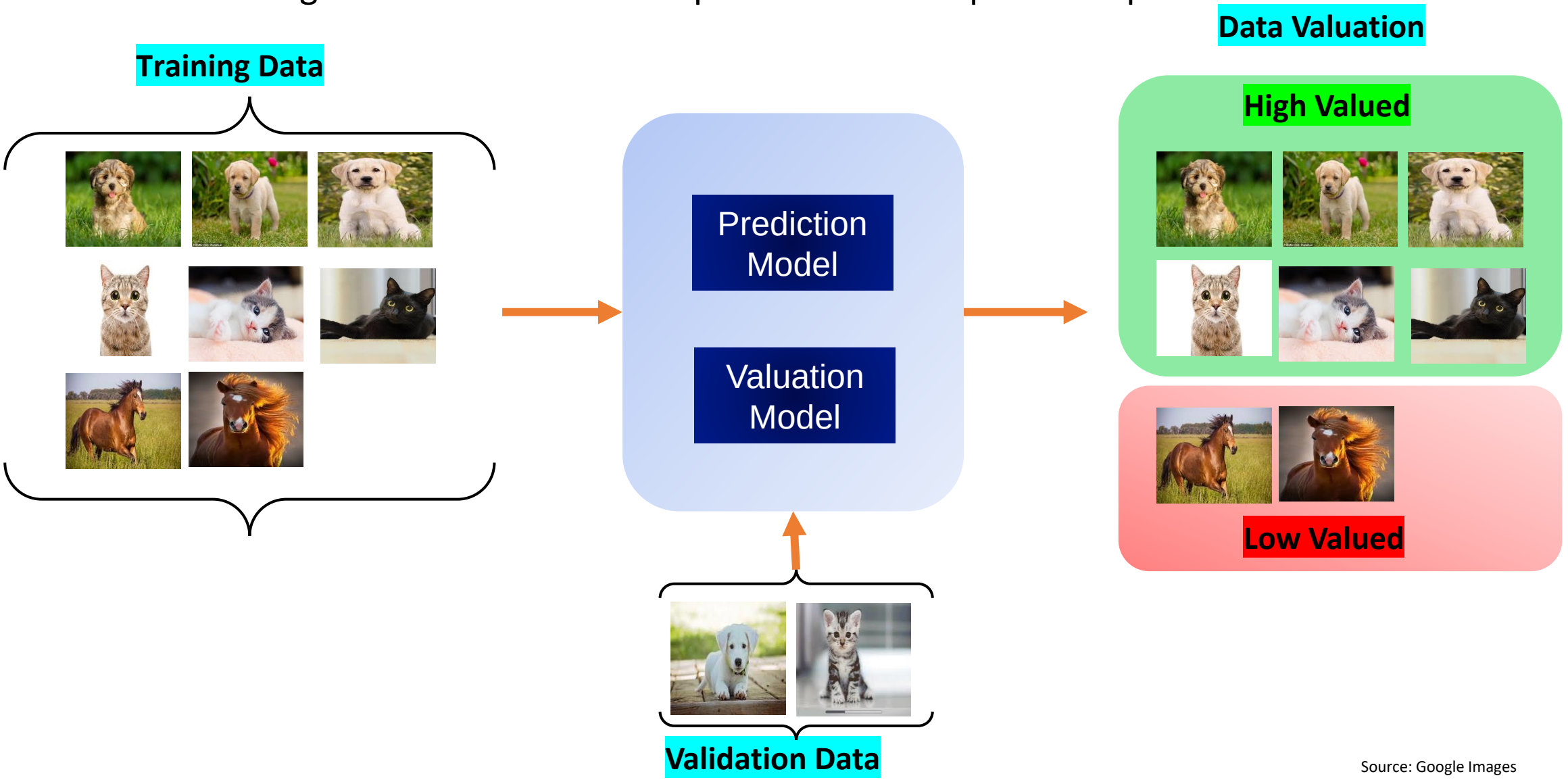
- We will cover the following topics:
  - Data Cleaning
  - Class Imbalance
  - Label Noise
  - **Data Valuation**
  - Data Homogeneity
  - Data Transformations



Source: https://www.analyticsinsight.net/data-literacy-helping-enterprises-lead-with-data-through-challenging-times/

Data Valuation: This is a new concept

- Value of a training datum is how much impact it has in the predictor performance.

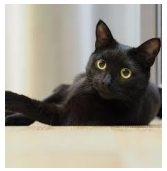


Source: Google Images

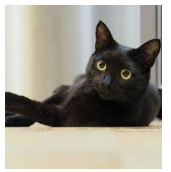


A: *Learning Algorithm*
V: *Validation Set*
x: *Training Datum*
$P(x, A, V)$: *Performance of A on V using x*
$Val(x, A, V)$: *Impact of x in performance of A on V*

Is the impact of all the cat images from training set same?

Is $P($  $, A, V) == P($  $, A, V)$ **?** → ✗

Then

$Val($  $, A, V) == Val($  $, A, V)$ → ✗

Source: Google Images

Scenarios in which datum has low value:

- Incorrect label


Label: Dog Label: Dog


Label: Dog


Label: Dog

- Input is noisy or low quality


- Usefulness for target task




Validation Set

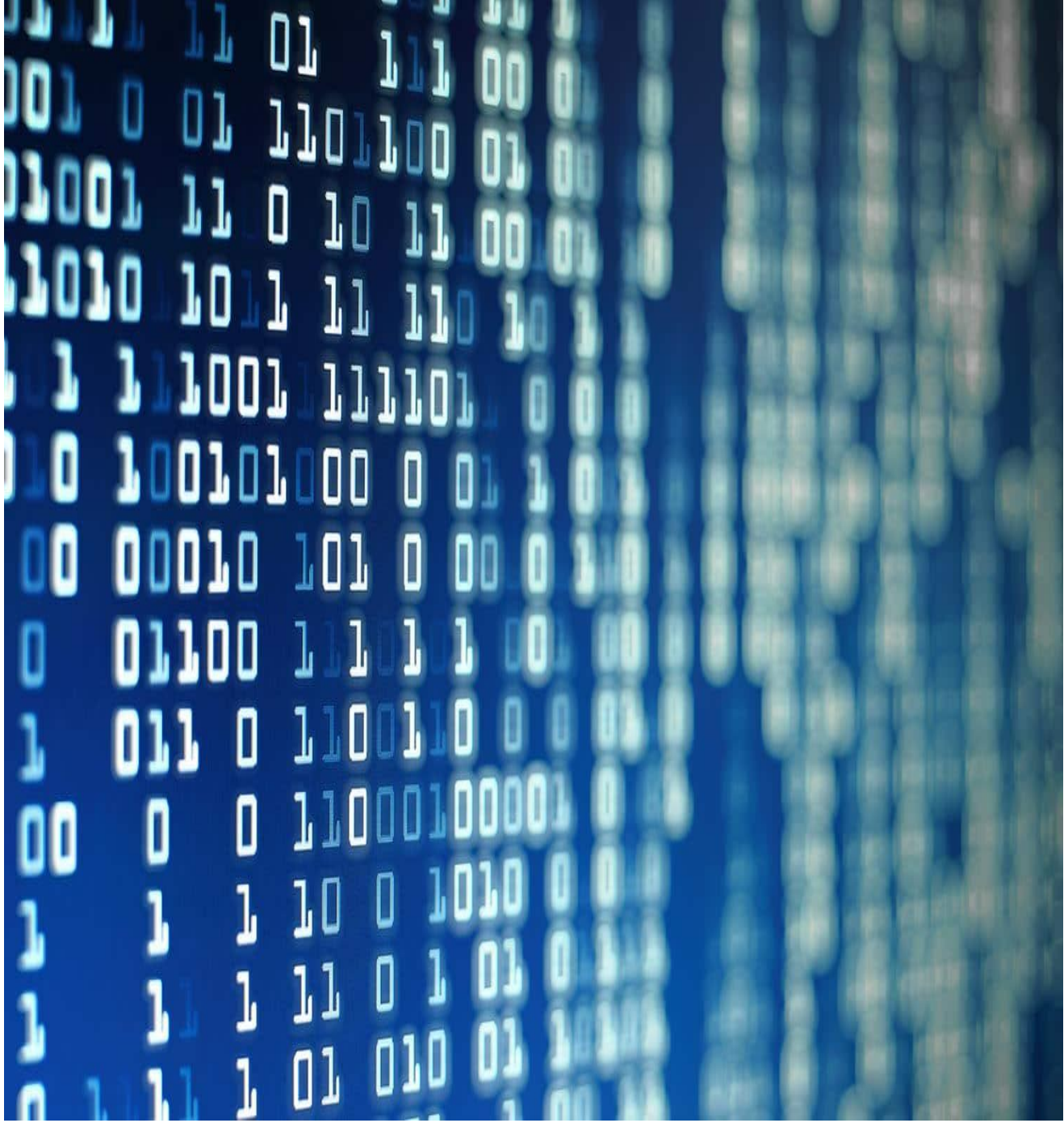

Source: Google Images

Application

- Identifying Data Quality
  - High value data → **Significant contribution**
  - Low value data → **Noise or outliers or not useful for task**
- Domain Adaptation
- Data Gathering

Data Quality Metrics

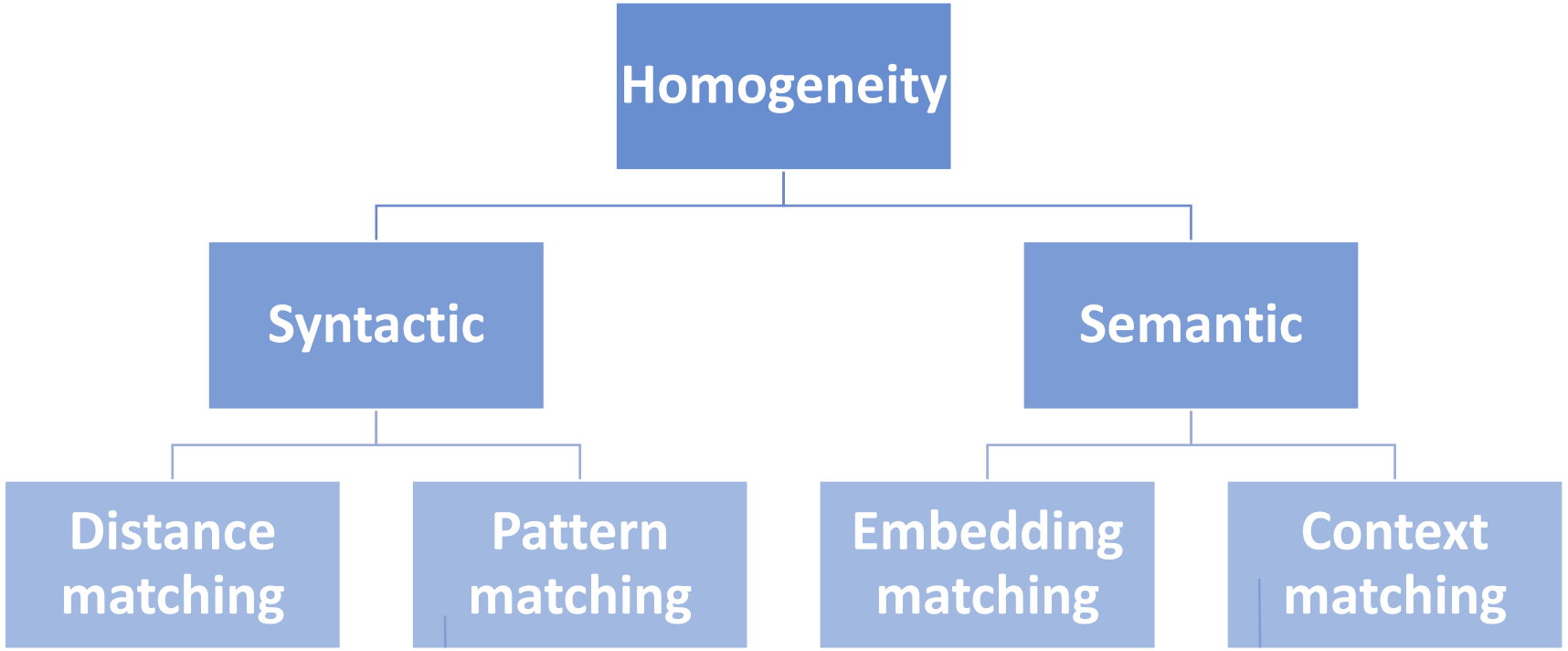
- We will cover the following topics:
  - Data Cleaning
  - Class Imbalance
  - Label Noise
  - Data Valuation
  - **Data Homogeneity**
  - Data Transformations



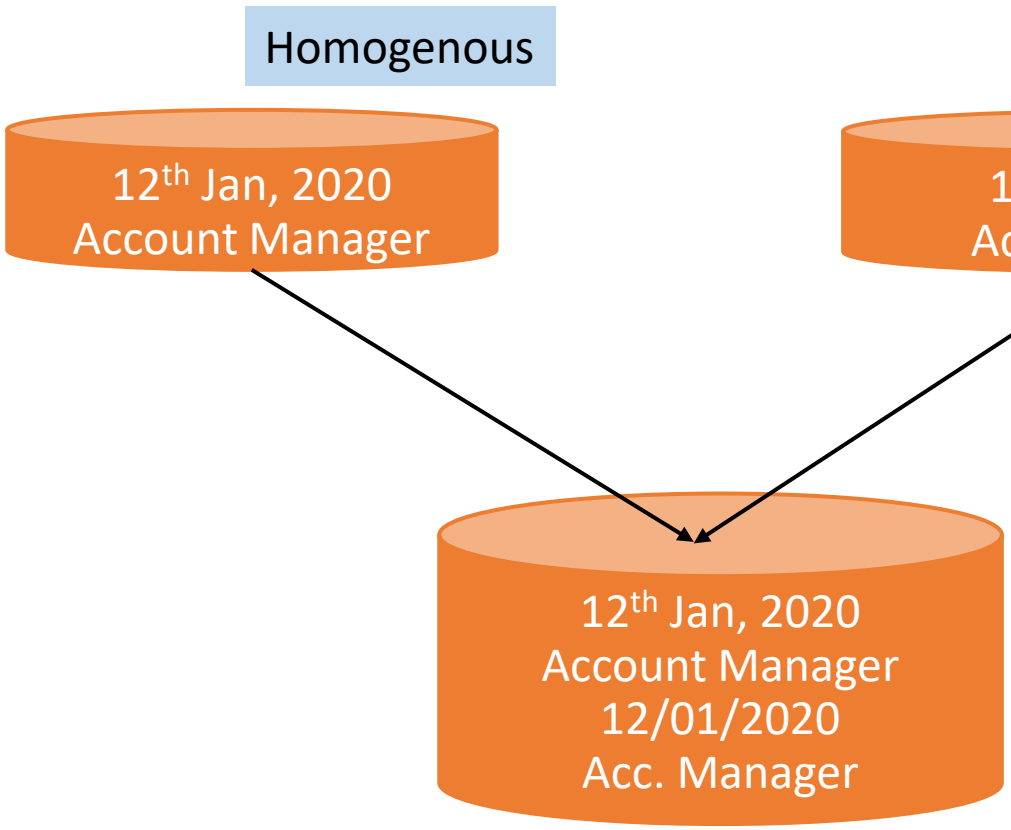
Source: https://www.analyticsinsight.net/data-literacy-helping-enterprises-lead-with-data-through-challenging-times/

Data Homogeneity

We call a data homogenous, if all the entries follow a unique pattern



Data Homogeneity



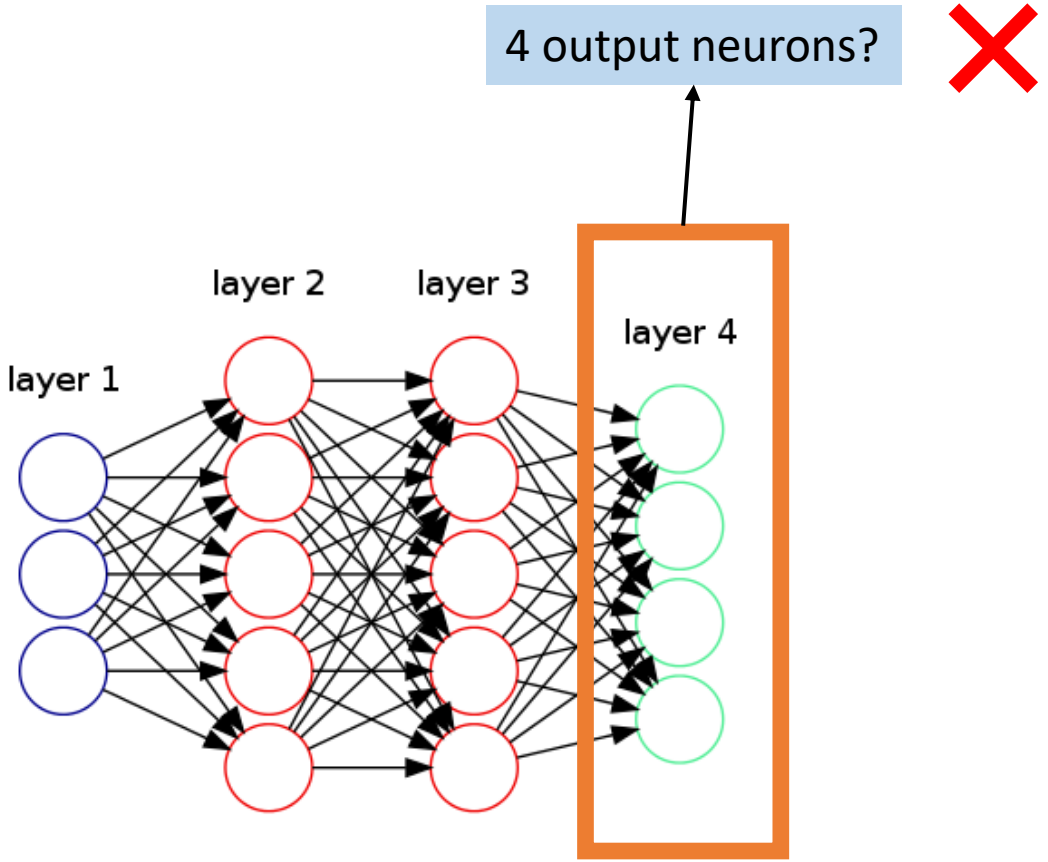
In-homogenous

- Two different date formats
- Two different values for the same category

In-homogeneity affects ML pipeline

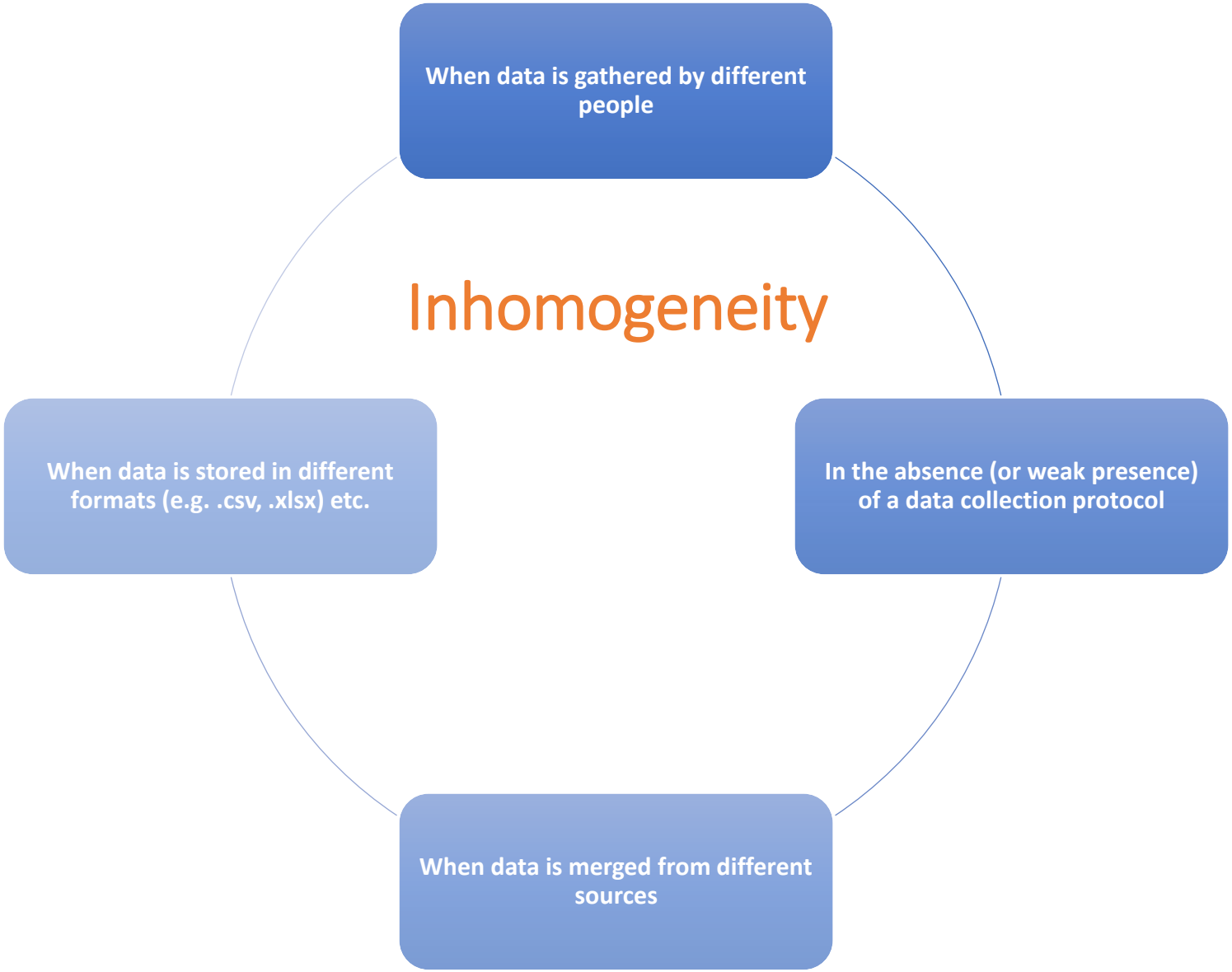
- Adult Census dataset downloaded from Kaggle
- Task is to predict Income level (>50k/<=50k) given several attributes of a person

Income level (Train)	Income level (Test)
<=50K	<=50K.
<=50K	<=50K.
>50K	>50K.
<=50K	>50K.
<=50K	<=50K.
<=50K	<=50K.
>50K	<=50K.
>50K	>50K.



Source: Image from Google Images

Inhomogeneity causes



Syntactic Homogeneity

Distance based similarity

Edit distance

		y0	y1	y2	y3	y4
			d	**a**	**v**	**e**
	0	1	2	3	4
d	1	0	1	2	3
v	2	1	1	1	2
a	3	2	1	2	2

x = d – v a
| | | |
y = d a v e

substitute a with e
insert a (after d)

Sequence based distance measure

Jaccard Similarity

$$J(x,y) = \frac{|B_x \cap B_y|}{|B_x \cup B_y|}$$

- Eg. x = dave, y = dav
- $B\_x = \{\#d, da, av, ve, e\#\}$,
  $B\_y = \{\#d, da, av, v\#\}$
- $J(x,y) = 3/6$

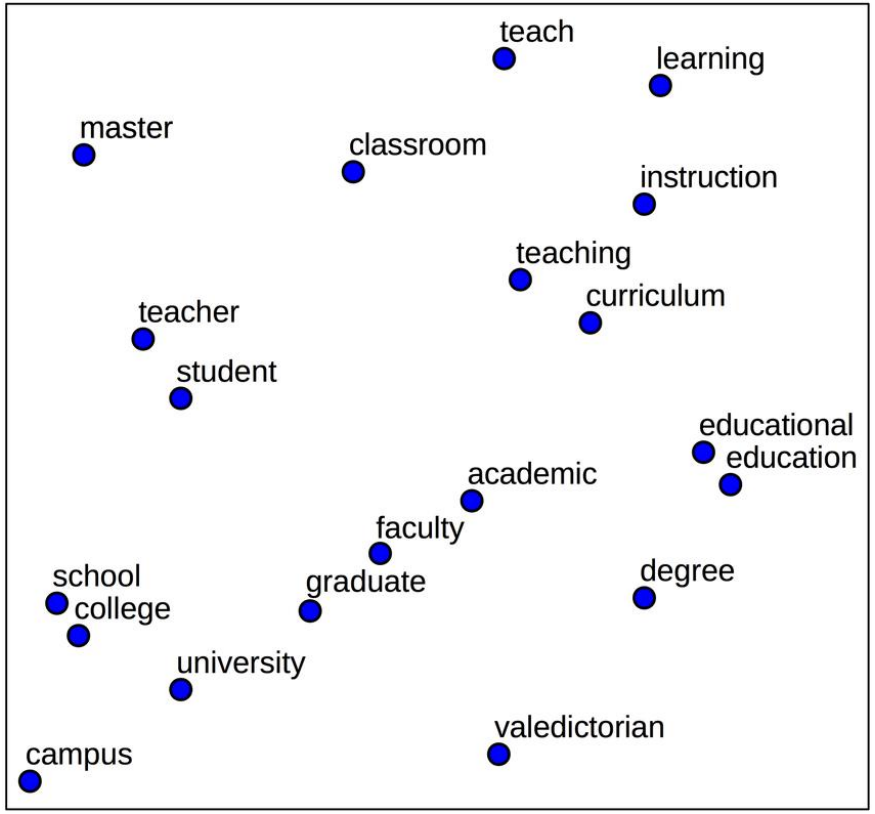
Set based distance measure

Source: Doan et al, 2012. Principles of Data Integration

[MCCD13]

Semantic Homogeneity

Embeddings based similarity

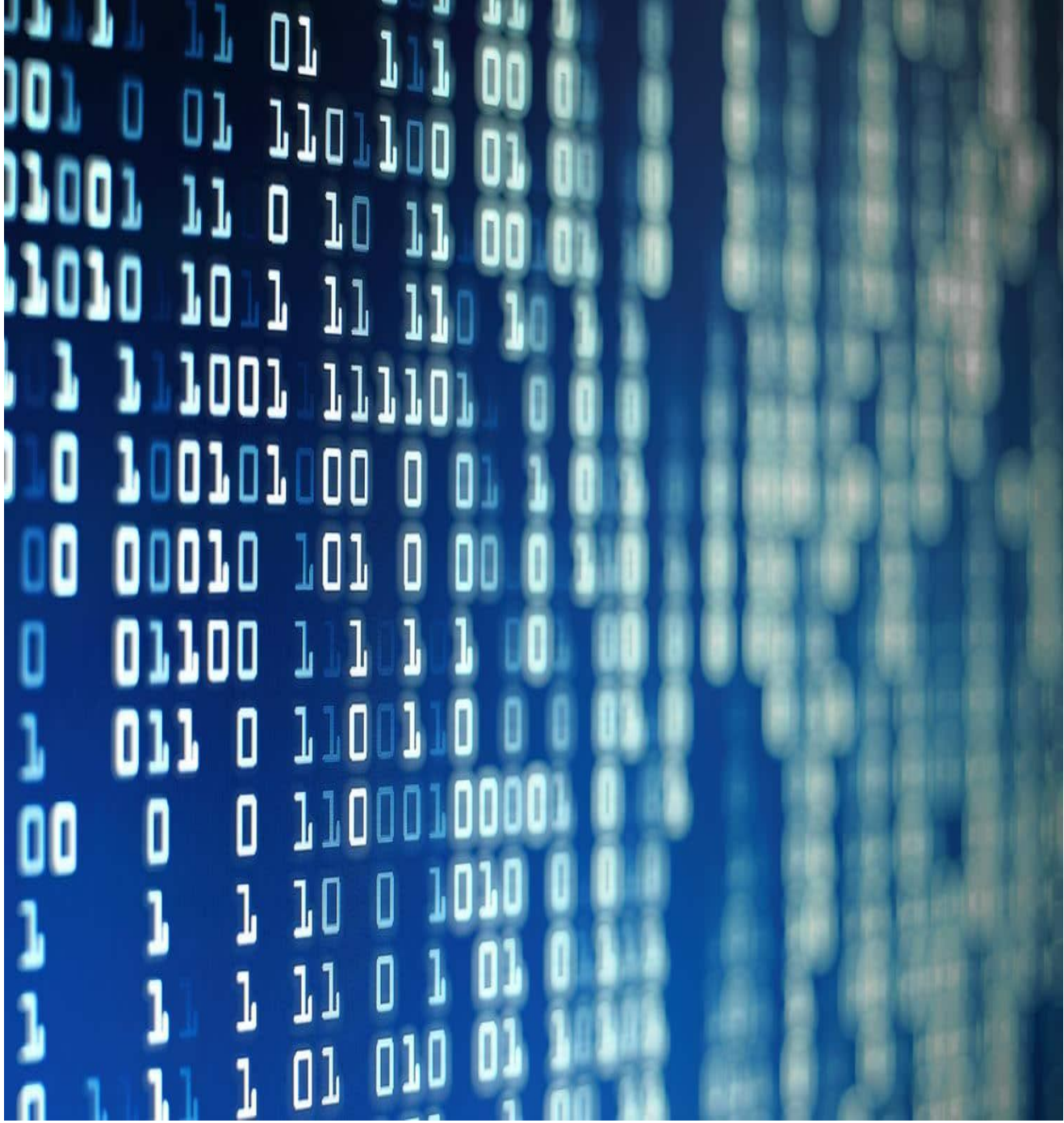


- Applicable when the entries in the data are meaningful English words
- Can use off the shelf embeddings like word2vec, Glove etc.
- $similarity(x, y) = cosine\big(\phi(x), \phi(y)\big);$
- $\phi(.) = word\ embedding$
- Embeddings capture semantic information as well

Source: Mikolov et al, 2012. Efficient estimation of word representations in vector space

Data Quality Metrics

- We will cover the following topics:
  - Data Cleaning
  - Class Imbalance
  - Label Noise
  - Data Valuation
  - Data Homogeneity
  - **Data Transformations**



Source: https://www.analyticsinsight.net/data-literacy-helping-enterprises-lead-with-data-through-challenging-times/

Data Transformation

- The goal is to allow the business user to transform provided data (heterogeneous) into user intended format (homogeneous), by showing a few examples of expected output for the given input samples.

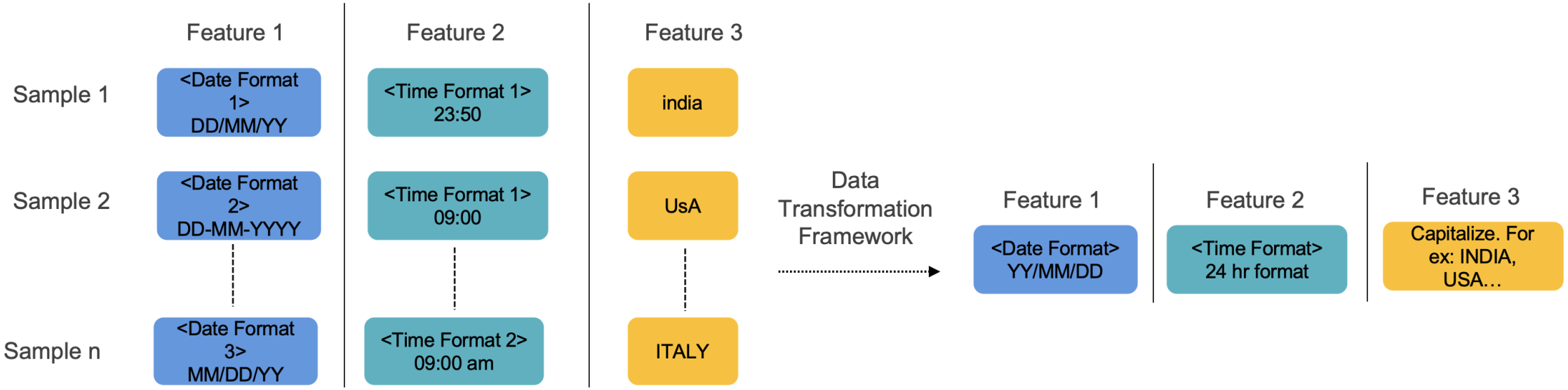


Fig 1. Each column represents a feature of the tabular data which is represented in multiple formats (left). The Data Transformation framework must convert them into user-intended format (right).

Data Transformation Examples

input	output
Male	M
M	M
Female	F
F	F

Gender

input	output
r. ponmajalcv	ponmajalcv
j. ndsb	ndsb
n. gupta	gupta
p. sharma	sharma

Surnames

input	output
Rs. 9723409119	9723409119
Rs. 2,235,313,002	2235313002
870,416,581	870416581
$ 3409245079.0	3409245079

Currency

input	Intended Output
Was 10%	10%
Only 10%	10%
10%	10%
10	10%

Product Discount Details

input	Intended Output
2/2/1988	2-2-88
20/9/11	20-9-11
3-1-1909	3-1-09
20-10-03	20-10-03

Dates

input	Intended Output
2,812,713,043	2812713043
427,373,287	427373287
4234567	4234567
13456789	13456789

Comma Numbers

Data Transformation Using PbE Systems

Programming-by-Example systems capture user intent using sample input-output example pairs and learn transformation programs that convert inputs into their corresponding outputs.

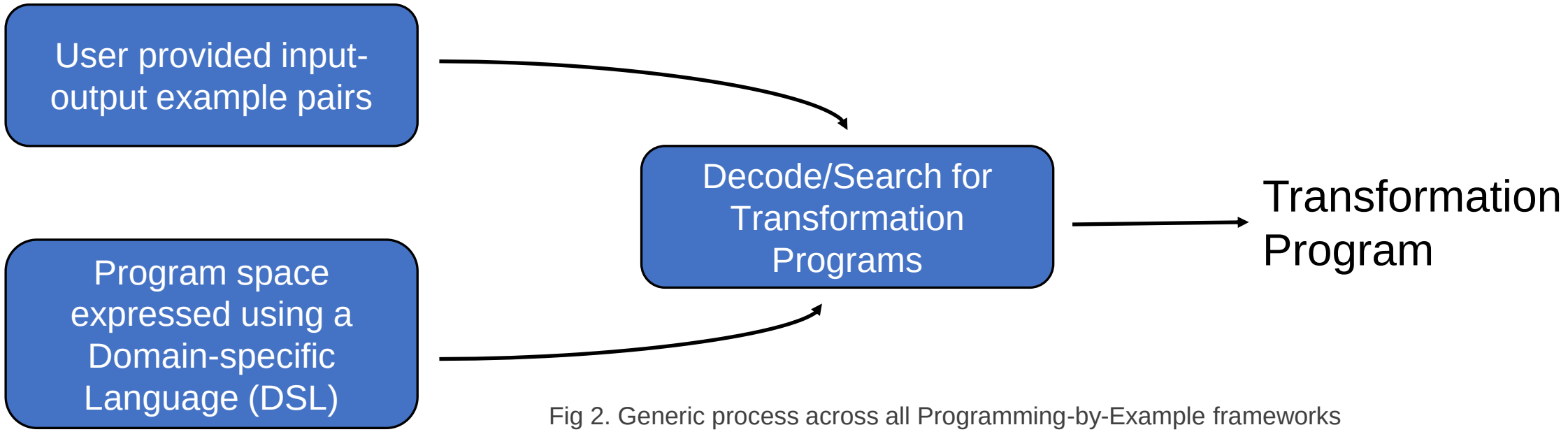


Fig 2. Generic process across all Programming-by-Example frameworks

We have learnt a lot of Quality issues
Be rest assured that it is only the tip of the iceberg

Now, let us see how to use data in building a Classifier

Assuming we have the correct data, let us look at a very intuitive ML Algorithm: Decision Tree

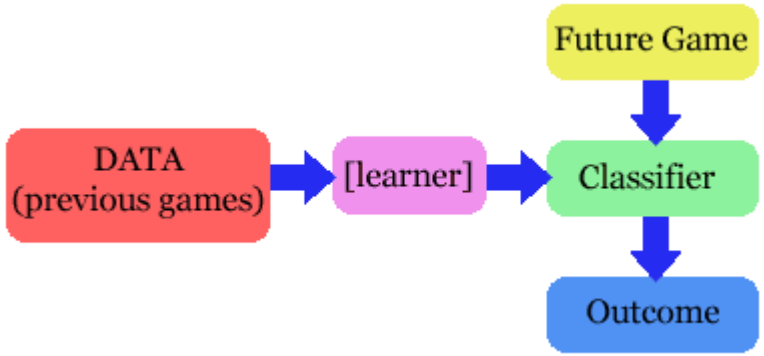
Decision Tree Algorithm

Comp328 tutorial 1

Thanks to: Kai Zhang (https://cszn.github.io/)

The problem

- Given a set of training cases/objects and their attribute values, try to determine the target attribute value of new examples.
  - Classification
  - Prediction



Why decision tree?

- Decision trees are powerful and popular tools for classification and prediction.
- Decision trees represent *rules*, which can be understood by humans and used in knowledge system such as database.

A simple Example: Basketball data

Where	When	Fred Starts	Joe offense	Joe defense	Opp C	*OutCome*
Home	7pm	Yes	Center	Forward	Tall	*Won*
Home	7pm	Yes	Forward	Center	Short	*Won*
Away	7pm	Yes	Forward	Forward	Tall	*Won*
Home	5pm	No	Forward	Center	Tall	*Lost*
Away	9pm	Yes	Forward	Forward	Short	*Lost*
Away	7pm	No	Center	Forward	Tall	*Won*
Home	7pm	No	Forward	Center	Tall	*Lost*
Home	7pm	Yes	Center	Center	Talls	*Won*
Away	7pm	Yes	Center	Center	Short	*Won*
Home	9pm	No	Forward	Center	Short	*Lost*

●
●
●

What we know

- The game will be away, at 9pm, and that Joe will play center on offense…

Where	When	Fred Starts	Joe offense	Joe defense	Opp C	*Outcome*
Away	9pm	No	Center	Forward	Tall	??

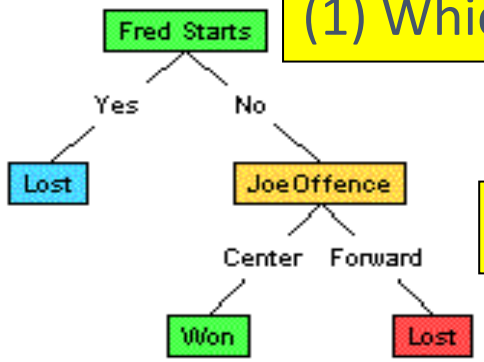
- A classification problem
- Generalizing the learned rule to new examples

Definition

- Decision tree is a classifier in the form of a tree structure
  - Decision node: specifies a test on a single attribute
  - Leaf node: indicates the value of the target attribute
  - Arc/edge: split of one attribute

- Decision trees classify instances or examples by starting at the root of the tree and moving through it until a leaf node.

Illustration



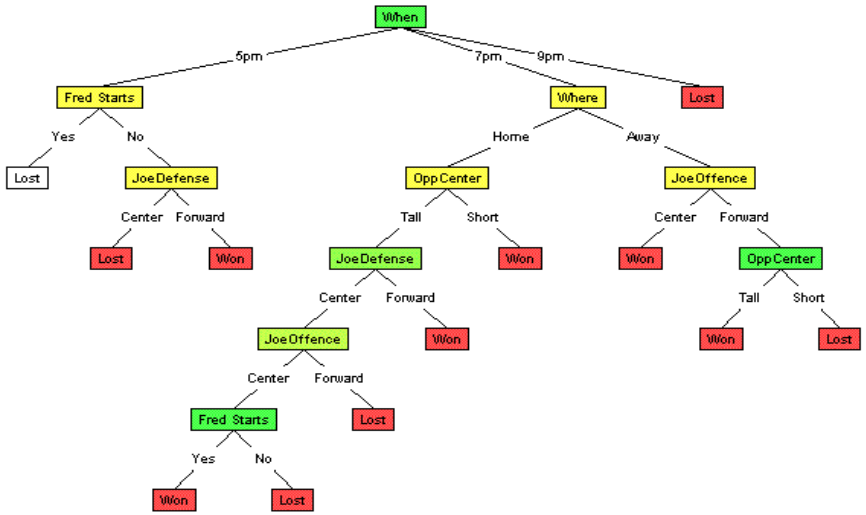
(1) Which to start? (root)

(2) Which node to proceed?

(3) When to stop/ come to conclusion?

Random split

- The tree can grow huge
- These trees are hard to understand.
- Larger trees are typically less accurate than smaller trees (Occams Razor).
- Finding the simplest Tree is NP-Hard (as all good problems are!)
- We employ heuristics to construct trees in practice



Principled Criterion

- Selection of an attribute to test at each node - choosing the most useful attribute for classifying examples.
- information gain
  - measures how well a given attribute separates the training examples according to their target classification
  - This measure is used to select among the candidate attributes at each step while growing the tree

Entropy

- A measure of homogeneity of the set of examples.

- Given a set S of positive and negative examples of some target concept (a 2-class problem), the entropy of set S relative to this binary classification is

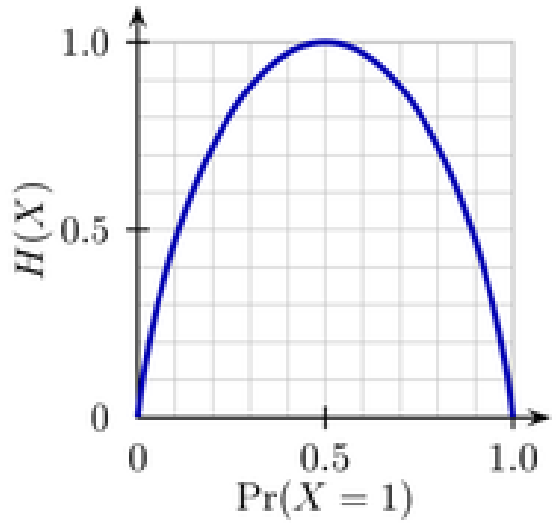
$$E(S) = - p(P)\log_2 p(P) – p(N)\log_2 p(N)$$

- Suppose S has 25 examples, 15 positive and 10 negatives [15+, 10-]. Then the entropy of S relative to this classification is

$$E(S)=-(15/25)\log_2(15/25) - (10/25)\log_2(10/25)$$

Some Intuitions

- The entropy is 0 if the outcome is ``certain''.
- The entropy is maximum if we have no knowledge of the system (or any outcome is equally possible).



Entropy of a 2-class problem with regard to the portion of one of the two groups

Information Gain

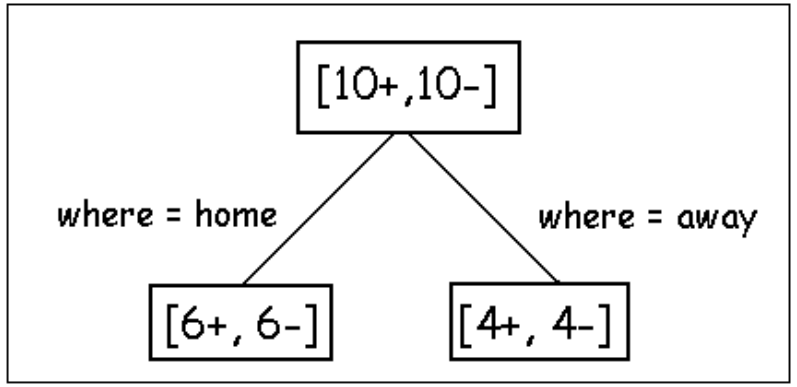
- Information gain measures the expected reduction in entropy, or uncertainty.

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

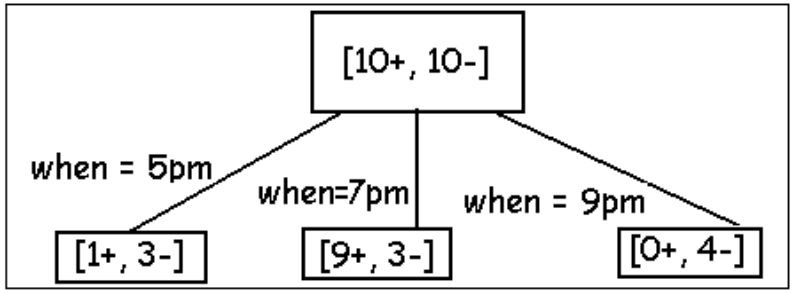
  - Values(A) is the set of all possible values for attribute A, and Sv the subset of S for which attribute A has value v Sv = {s in S | A(s) = v}.
  - the first term in the equation for *Gain* is just the entropy of the original collection *S*
  - the second term is the expected value of the entropy after S is partitioned using attribute A

- It is simply the expected reduction in entropy caused by partitioning the examples according to this attribute.
- It is the number of bits saved when encoding the target value of an arbitrary member of *S*, by knowing the value of attribute *A*.

Examples



- Before partitioning, the entropy is
  - *H(10/20, 10/20) = - 10/20 log(10/20) - 10/20 log(10/20) = 1*
- Using the ``where'' attribute, divide into 2 subsets
  - Entropy of the first set H(home) = - 6/12 log(6/12) - 6/12 log(6/12) = 1
  - Entropy of the second set H(away) = - 4/8 log(6/8) - 4/8 log(4/8) = 1
- Expected entropy after partitioning
  - 12/20 * H(home) + 8/20 * H(away) = 1



- Using the ``when'' attribute, divide into 3 subsets
  - Entropy of the first set H(5pm) = - 1/4 log(1/4) - 3/4 log(3/4);
  - Entropy of the second set H(7pm) = - 9/12 log(9/12) - 3/12 log(3/12);
  - Entropy of the second set H(9pm) = - 0/4 log(0/4) - 4/4 log(4/4) = 0
- Expected entropy after partitioning
  - 4/20 * H(1/4, 3/4) + 12/20 * H(9/12, 3/12) + 4/20 * H(0/4, 4/4) = 0.65
- Information gain 1-0.65 = 0.35

Decision

- Knowing the ``when'' attribute values provides larger information gain than ``where''.
- Therefore the ``when'' attribute should be chosen for testing prior to the ``where'' attribute.
- Similarly, we can compute the information gain for other attributes.
- At each node, choose the attribute with the largest information gain.

- Stopping rule
  - Every attribute has already been included along this path through the tree, or
  - The training examples associated with this leaf node all have the same target attribute value (i.e., their entropy is zero).

Evaluation

- Training accuracy
  - How many training instances can be correctly classify based on the available data?
  - Is high when the tree is deep/large, or when there is less confliction in the training instances.
  - however, higher training accuracy does not mean good generalization
- Testing accuracy
  - Given a number of new instances, how many of them can we correctly classify?
  - Cross validation

References

[BF99] Carla E Brodley and Mark A Friedl. Identifying mislabeled training data. Journal of artificial intelligence research, 11:131–167, 1999.

[ZWC03] Xingquan Zhu, Xindong Wu, and Qijun Chen. Eliminating class noise in large datasets. In Proceedings of the 20th International Conference on Machine Learning(ICML-03), pages 920–927, 2003.

[DLG17]J unhua Ding, XinChuan Li, and Venkat N Gudivada. Augmentation and evaluation of training data for deep learning. In2017 IEEE International Conference on Big Data(Big Data), pages 2603–2611. IEEE, 2017.

[ARK18] Mohammed Al-Rawi and Dimosthenis Karatzas. On the labeling correctness in computer vision datasets. In IAL@PKDD/ECML, 2018.

[NJC19] Curtis G Northcutt, Lu Jiang, and Isaac L Chuang. Confident learning: Estimating uncertainty in dataset labels. arXiv preprint arXiv:1911.00068, 2019.

[Coo77] R Dennis Cook. Detection of influential observation in linear regression.T echnometrics

[EGH17] Rajmadhan Ekambaram, Dmitry B Goldgof, and Lawrence O Hall. Finding label noise examples in large scale datasets. In2017 IEEE International Conference on Systems, Man, and Cybernetics pages 2420–2424., 2017.

References

[GZ19] Amirata Ghorbani and James Zou. Data shapley: Equitable valuation of data for machine learning. arXiv preprint arXiv:1904.02868, 2019

[YAP19] Jinsung Yoon, Sercan O Arik, and Tomas Pfister. Data valuation using reinforcement learning. arXiv preprint arXiv:1909.11671, 2019.

[DKO+07] Bing Tian Dai, Nick Koudas, Beng Chin Ooi, Divesh Srivastava, and Suresh Venkatasubramanian. Column heterogeneity as a measure of data quality. 2007

[sim] http://www.countrysideinfo.co.uk/simpsons.htm.

[DHI12] AnHai Doan, Alon Halevy, and Zachary Ives. Principles of data integration. Elsevier, 2012.

[MCCD13] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 2013

[ALI15] Aida Ali and Siti Mariyam Hj. Shamsuddin and Anca L. Ralescu. Classification with class imbalance problem: A review. SOCO 2015

References

[PG15] Oleksandr Polozov and Sumit Gulwani. Flashmeta: a framework for inductive program synthesis. In Proceedings of the 2015 ACM SIGPLAN International Conference on Object-Oriented Programming, Systems, Languages, and Applications, pages 107–126, 2015.

[DUB+17] Jacob Devlin, Jonathan Uesato, Surya Bhupatiraju, Rishabh Singh, Abdel-rahman Mohamed, and Pushmeet Kohli. Robustfill: Neural program learning under noisy i/o. In Proceedings of the 34th International Conference on Machine Learning-Volume 70, pages 990–998, 2017.

[KMP+18] Ashwin Kalyan, Abhishek Mohta, Oleksandr Polozov, Dhruv Batra, Prateek Jain, and Sumit Gulwani. Neural-guided deductive search for real-time program synthesis from examples. arXiv preprint arXiv:1804.01186, 2018.

[AYR16] Dimitrios Alikaniotis, Helen Yannakoudakis, and Marek Rei. Automatic text scoring using neural networks. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 715–725, Berlin, Germany, August 2016. Association for Computational Linguistics.

[CPR19] Richard Csaky, Patrik Purgai, and Gabor Recski. Improving neural conversational models with entropy-based data filtering. arXiv preprint arXiv:1905.05471, 2019.

References

[CRZ18] Edward Collins, Nikolai Rozanov, and Bingbing Zhang. Evolutionary data measures: Understanding the difficulty of text classification tasks. arXiv preprint arXiv:1811.01910, 2018.

[JMAS19] Parag Jain, Abhijit Mishra, Amar Prakash Azad, and Karthik Sankaranarayanan. Unsupervised controllable text formalization. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, pages 6554–6561, 2019.

[KWAP17] Ramakrishnan Kannan, Hyenkyun Woo, Charu C Aggarwal, and Haesun Park. Outlier detection for text data. In Proceedings of the 2017 siam international conference on data mining, pages 489–497. SIAM, 2017.

[MS18] Mohsen Mesgar and Michael Strube. A neural local coherence model for text quality assessment. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 4328–4339, 2018.

[ÖG17] Robert Östling and Gintar˙e Grigonyt˙e. Transparent text quality assessment with convolutional neural networks. In Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications, pages 282–286, 2017.

References

[PLN19] Nicole Peinelt, Maria Liakata, and Dong Nguyen. Aiming beyond the obvious: Identifying non-obvious cases in semantic similarity datasets. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 2792–2798, 2019.

[RDNMJ13] Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. Linguistic models for analyzing and detecting biased language. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1650–1659, 2013.

[RWGS20] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of nlp models with checklist. arXiv preprint arXiv:2005.04118, 2020.

[RZV+19] Lukas Ruff, Yury Zemlyanskiy, Robert Vandermeulen, Thomas Schnake, and Marius Kloft. Self-attentive, multi-context one-class classification for unsupervised anomaly detection on text. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4061–4071, 2019.

[RH01] Vijayshankar Raman and Joseph M Hellerstein. Potter's wheel: An interactive data cleaning system. In VLDB, volume 1, pages 381–390, 2001.

[ROE+19] El Kindi Rezig, Mourad Ouzzani, Ahmed K Elmagarmid, Walid G Aref, and Michael Stonebraker. Towards an end-to-end human-centric data cleaning framework. In Proceedings of the Workshop on Human-In-the-Loop Data Analytics, pages 1–7, 2019.

References

[Ham13] Kelli Ham. Openrefine (version 2.5). http://openrefine. org. free, open-source tool for cleaning and transforming data. Journal of the Medical Library Association: JMLA, 101(3):233, 2013.

[KHFW16] Sanjay Krishnan, Daniel Haas, Michael J Franklin, and Eugene Wu. Towards reliable interactive data cleaning: A user survey and recommendations. In Proceedings of the Workshop on Human-In-the-Loop Data Analytics, pages 1–5, 2016.

[KPHH11] Sean Kandel, Andreas Paepcke, Joseph Hellerstein, and Jeffrey Heer. Wrangler: Interactive visual specification of data transformation scripts. In ACM Human Factors in Computing Systems (CHI), 2011.

[MLW+19] Michael Muller, Ingrid Lange, Dakuo Wang, David Piorkowski, Jason Tsay, Q Vera Liao, Casey Dugan, and Thomas Erickson. How data science workers work with data: Discovery, capture, curation, design, creation. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pages 1–15, 2019.

[BTR15] Paula Branco, Luis Torgo, and Rita Ribeiro. A survey of predictive modelling under imbalanced distributions.arXiv preprint arXiv:1505.01658, 2015

References

[DT10] Misha Denil and Thomas Trappenberg. Overlap versus imbalance. In Canadian conference on artificial intelligence, pages 220–231. Springer, 2010.

[GBSW04] Jerzy W Grzymala-Busse, Jerzy Stefanowski, and Szymon Wilk. A comparison of two approaches to data mining from imbalanced data. In International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, pages 757–763. Springer, 2004

[Jap03] Nathalie Japkowicz. Class imbalances: are we focusing on the right issue. 2003.

[JJ04] Taeho Jo and Nathalie Japkowicz. Class imbalances versus small disjuncts. ACM Sigkdd Explorations Newsletter, 6(1):40–49, 2004.

[Kra16] Bartosz Krawczyk. Learning from imbalanced data: open challenges and future directions. InProgress in Artificial Intelligence volume, 2016.

[LCT19] Yang Lu, Yiu-ming Cheung, and Yuan Yan Tang. Bayes imbalance impact index: A measure of class imbalanced data set for classification problem. IEEE Transactions on Neural Networks and Learning Systems, 2019

[Jap03] Nathalie Japkowicz. Class imbalances: are we focusing on the right issue. 2003.

References

[DT10] Misha Denil and Thomas Trappenberg. Overlap versus imbalance. In Canadian conference on artificial intelligence, pages 220–231. Springer, 2010.

[GBSW04] Jerzy W Grzymala-Busse, Jerzy Stefanowski, and Szymon Wilk. A comparison of two approaches to data mining from imbalanced data. In International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, pages 757–763. Springer, 2004

[JJ04] Taeho Jo and Nathalie Japkowicz. Class imbalances versus small disjuncts. ACM Sigkdd Explorations Newsletter, 6(1):40–49, 2004.

[Kra16] Bartosz Krawczyk. Learning from imbalanced data: open challenges and future directions. InProgress in Artificial Intelligence volume, 2016.

[LCT19] Yang Lu, Yiu-ming Cheung, and Yuan Yan Tang. Bayes imbalance impact index: A measure of class imbalanced data set for classification problem. IEEE Transactions on Neural Networks and Learning Systems, 2019

WP03] Gary M Weiss and Foster Provost. Learning when training data are costly: The effect of class distribution on tree induction. Journal of artificial intelligence research,19:315–354, 2003