Doc. RNDr. Irena Holubová, Ph.D. & PROFINIT

DATA SCIENCE

NDBI048

Data Preparation

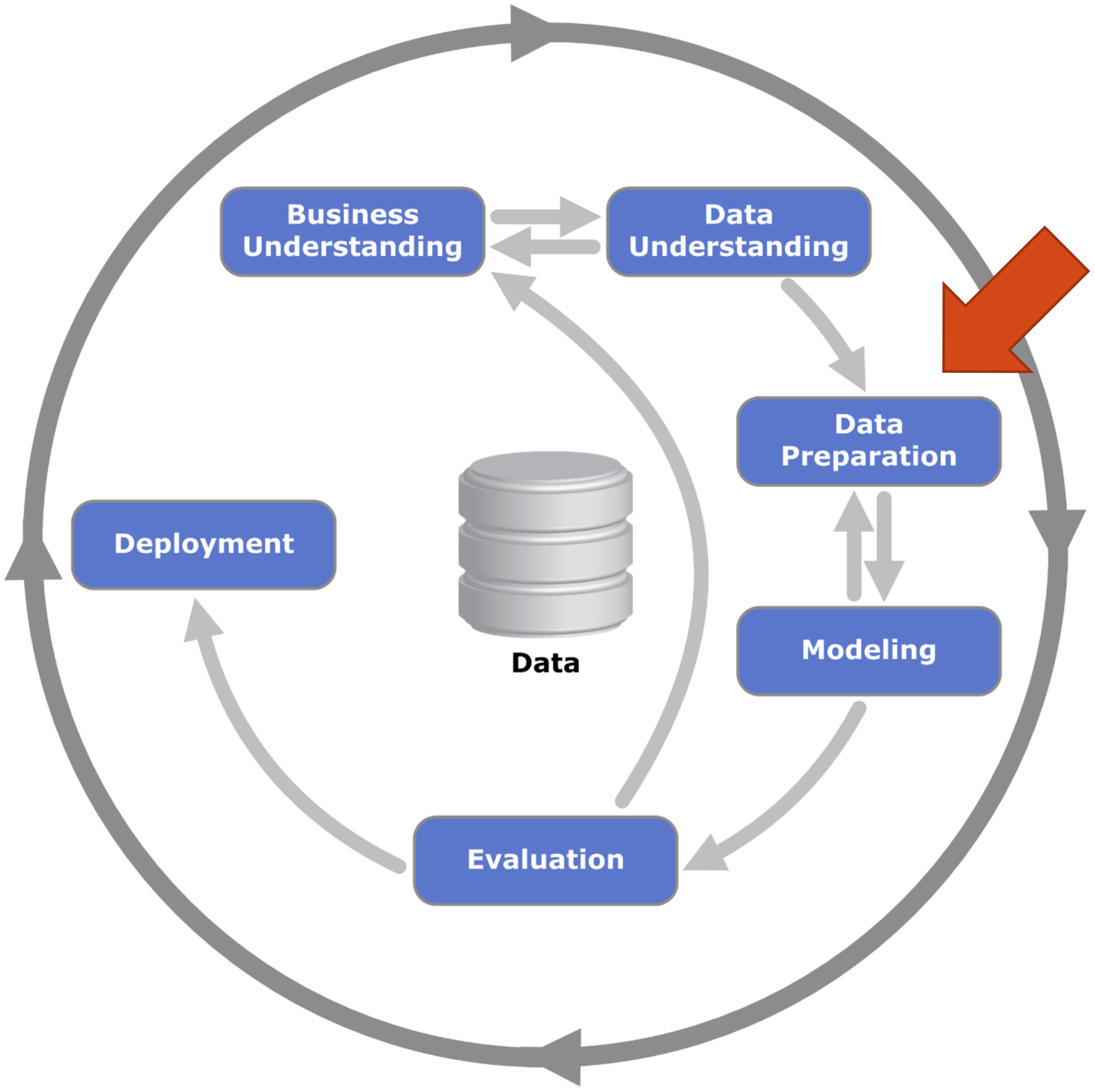
https://www.ksi.mff.cuni.cz/~holubova/NDBI048/

OUTLINE

- Data cleaning
- Data transformation

CRISP-DM PHASES

I. Business Understanding
II. Data Understanding
III. Data Preparation
IV. Modeling
V. Evaluation
VI. Deployment



https://www.datascience-pm.com/crisp-dm-2/

WHAT IS DATA PREPARATION?

Rule 80/20:
20% of time: data analysis
80% of time: finding, cleaning, and reorganizing amounts of data

- The process of cleaning and transforming raw data prior to processing and analysis
  - Reformatting data
  - Making corrections
    - Incomplete, noisy, inconsistent, …
  - Combining of data sets
    - To enrich data
  - …
- The purpose is to transform data sets so that their information content is best exposed to processing tools
- Error prediction rate should be lower (or the same) after the preparation as before it

age = “”
age = -17
age = 41, birth = “2010-07-17”

Business Understanding

Determine Business Objectives
Background
Business Objectives
Business Success Criteria

Assess Situation
Inventory of Resources
Requirements, Assumptions, and Constraints
Risks and Contingencies
Terminology
Costs and Benefits

Determine Data Mining Goals
Data Mining Goals
Data Mining Success Criteria

Produce Project Plan
Project Plan
Initial Assessment of Tools and Techniques

Data Understanding

Collect Initial Data
Initial Data Collection Report

Describe Data
Data Description Report

Explore Data
Data Exploration Report

Verify Data Quality
Data Quality Report

Data Preparation

Select Data
Rationale for Inclusion/ Exclusion

Clean Data
Data Cleaning Report

Construct Data
Derived Attributes
Generated Records

Integrate Data
Merged Data

Format Data
Reformatted Data

Dataset
Dataset Description

Modeling

Select Modeling Techniques
Modeling Technique
Modeling Assumptions

Generate Test Design
Test Design

Build Model
Parameter Settings
Models
Model Descriptions

Assess Model
Model Assessment
Revised Parameter Settings

Evaluation

Evaluate Results
Assessment of Data Mining Results w.r.t. Business Success Criteria
Approved Models

Review Process
Review of Process

Determine Next Steps
List of Possible Actions
Decision

Deployment

Plan Deployment
Deployment Plan

Plan Monitoring and Maintenance
Monitoring and Maintenance Plan

Produce Final Report
Final Report
Final Presentation

Review Project
Experience Documentation

https://www.mygreatlearning.com/blog/why-using-crisp-dm-will-make-you-a-better-data-scientist/

CRISP-DM DATA PREPARATION

- **Select data**: Which (portions of) data sets will (not) be used and why?
  - Collect additional data (internal, external)
- **Clean data**: The data is unlikely to be perfectly clean (error-free)
  - Correct, replace, remove, ignore noise
    - Track down sources to make specific data corrections
  - Decide how to deal with special values and their meaning
  - Aggregation level, missing values
  - Outliers
- **Construct data**: Extract new attributes (or re-construct missing)
  - E.g., body mass index
- **Integrate data**: Create new data sets by combining data from multiple sources
- **Format data**: Re-arrange, re-order, re-format
  - E.g., convert string values that store numbers to numeric values

The most time consuming

TYPES OF DATA

Information content

- **Nominal** – cannot be compared (ordered)
  - ID numbers, zip codes
  - Categorical - blood types, marital status
  - Dichotomous = yes/no
- **Ordinal** – can be compared (ordered)
  - Rankings (e.g., a scale from 1-10), grades, height (tall, medium, short)
- **Interval** – distance between data makes sense
  - Calendar dates, temperatures, IQ scores
- **Ratio-like** – the ratio between the values makes sense
  - Length, time, counts

Qualitative

Quantitative

Discrete

Continuous

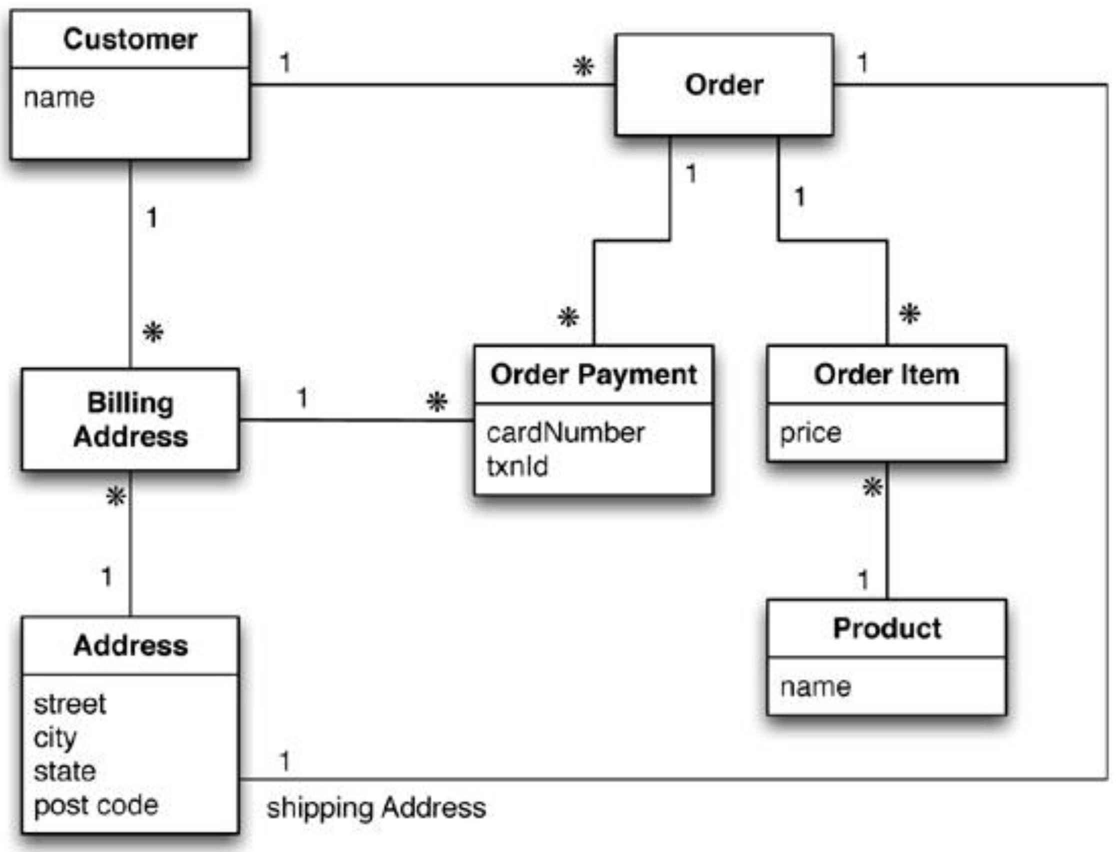
TYPES OF DATA – OTHER CLASSIFICATIONS

- Structured vs. semi-structured vs. unstructured
  - Aggregate-oriented vs. aggregate-ignorant
- Single-model vs. multi-model
- Schema-less vs. schema-full vs. schema-mixed
- Small or big … Big Data
- Stable, long term changing, frequently changing
- Federated data (come form different heterogeneous sources), massive high dimensional data, time series, Web data, …
- …

AGGREGATES

- Data model = the model by which the database organizes data
- Various types of databases depending on their model
  - Relational, object, array, key-value, document, column-family, graph…
- Aggregate
  - A data unit with a complex structure
  - Domain-Driven Design: "an aggregate is a collection of related objects that we wish to treat as a unit"
    - A unit for data manipulation and management of consistency

EXAMPLE – AGGREGATE-IGNORANT



Customer	
Id	Name
1	Martin

Orders		
Id	CustomerId	ShippingAddressId
99	1	77

Product	
Id	Name
27	NoSQL Distilled

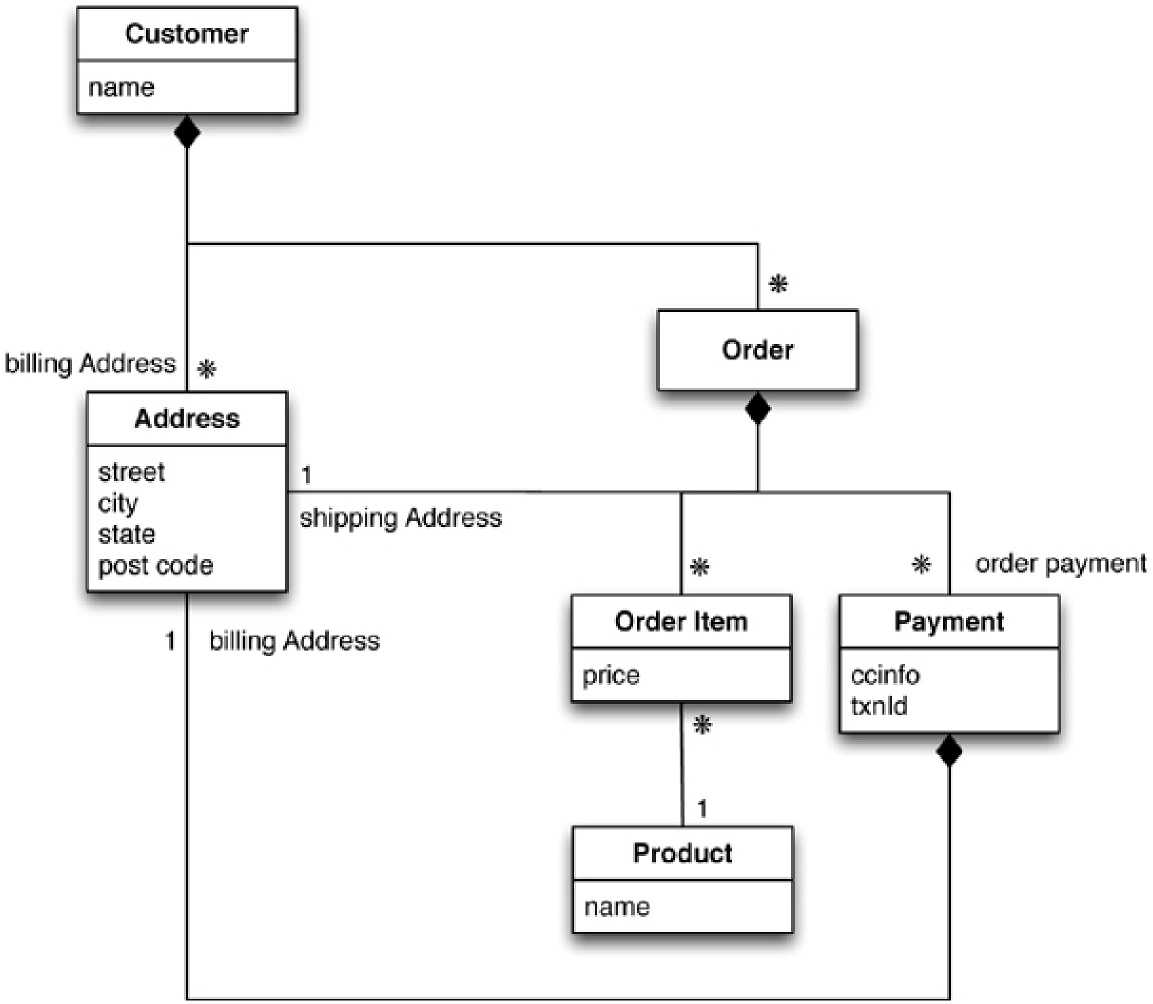
BillingAddress		
Id	CustomerId	AddressId
55	1	77

OrderItem			
Id	OrderId	ProductId	Price
100	99	27	32.45

Address	
Id	City
77	Chicago

OrderPayment				
Id	OrderId	CardNumber	BillingAddressId	txnId
33	99	1000-1000	55	abelif879rft

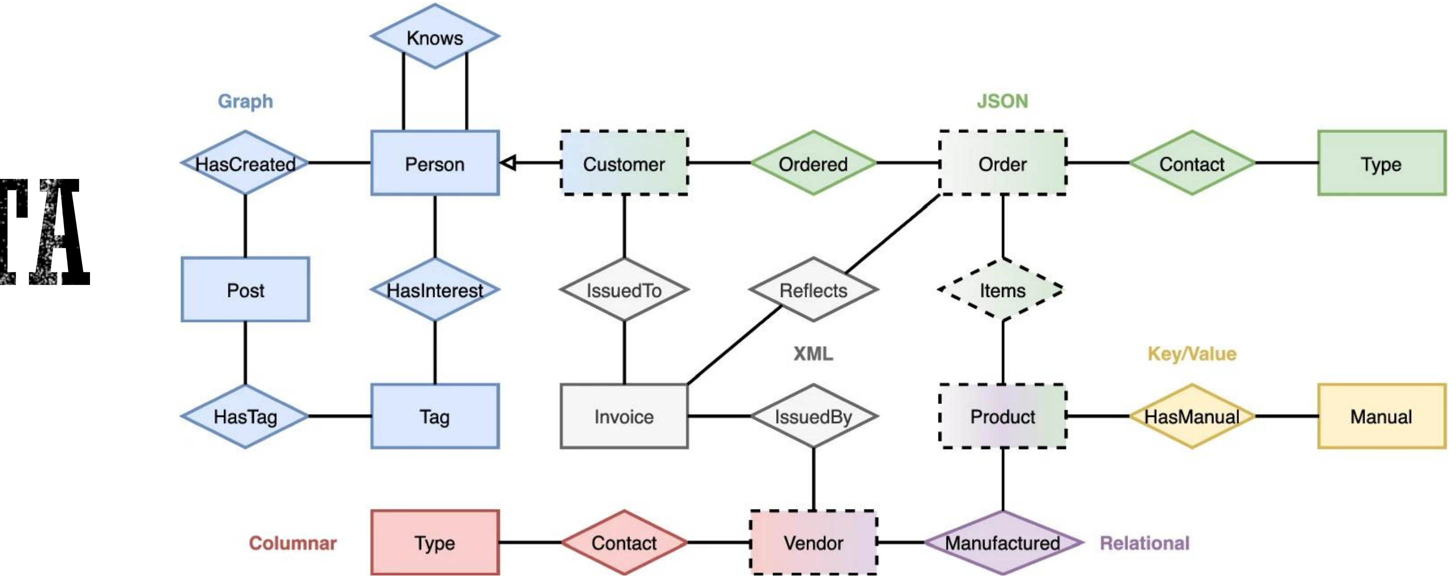
EXAMPLE – AGGREGATE-ORIENTED

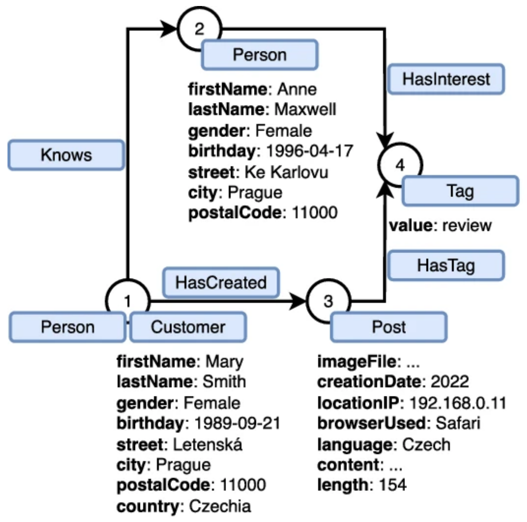



```
// in customers
{
"customer": {
"id": 1,
"name": "Martin",
"billingAddress": [{"city": "Chicago"}],
"orders": [
  {
    "id":99,
    "customerId":1,
    "orderItems":[
    {
    "productId":27,
    "price": 32.45,
    "productName": "NoSQL Distilled"
    }
  ],
  "shippingAddress":[{"city":"Chicago"}]
  "orderPayment":[
    {
    "ccinfo":"1000-1000-1000-1000",
    "txnId":"abelif879rft",
    "billingAddress": {"city": "Chicago"}
    }],
  }]
}
}
```

MULTI-MODEL DATA

- A set of interlinked data, each having its own model
- Types of combination:
  - Inter-model references
  - Embedding
  - Cross-model redundancy






```xml
<invoice invoiceNo="2022001">
  <customerId>1</customerId>
  <creationDate>2022-02-22</creationDate>
  <items>
    <product productid="107" title="RacingCar" price="1500" quantity="2" />
    <product productid="105" title="Death" price="300" quantity="1" />
  </items>
  <totalPrice currency="CZK">1800</totalPrice>
  <paid/>
</invoice>
```

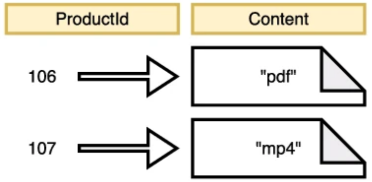
```json
{
  _id : 2022001
  customer: {
    customerId: 1,
    firstName: Mary,
    lastName: Smith,
    street: Letenská,
    city: Prague,
    postalCode: 11000,
    country: Czechia
  },
  contact : {
    cellphone: +420123456789,
    email: mary@smith.cz
  },
  items: [
    {
      productid: 107,
      title: RacingCar,
      brand: Toy,
      price: 1500,
      quantity: 3
    }, {
      productid: 105,
      title: Death,
      brand: Book,
      price: 300,
      quantity: 1
    }
  ]
}
```

id	name	country	cdf	Industry	Contact	
10001	Books Inc.	Czechia	...	Printhouse	address	...
				Audiomedia	phone	...

id	name	country	cdf	Contact	
10002	Toys Ins.	Slovakia	...	address	...
				phone	...
				website	...

productId	asin	title	price	brand	imgUrl
103	...	Pyramids	300	Book	...
104	...	Pyramids	450	Audiobook	null
105	...	Death	300	Book	...
106	...	Baby Doll	2500	Toy	...
107	...	RacingCar	1500	Toy	...

id	productId
10001	103
10001	104
10001	105
10002	106
10002	107



http://www.ibmbigdatahub.com/

WHAT IS BIG DATA?

Mobile devices
(tracking all objects all the time)

Social media and networks
(all of us are generating data)

Scientific instruments
(collecting all sorts of data)

Sensor technology and networks
(measuring all kinds of data)

Gartner: *"**Big Data**" is high **volume**, high **velocity**, and/or high **variety** information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization.*

IBM: *Depending on the industry and organization, **Big Data** encompasses information from internal and external sources such as transactions, social media, enterprise content, sensors, and mobile devices.*
Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

FACEBOOK BY THE NUMBERS:
STATS, DEMOGRAPHICS & FUN FACTS
(LAST UPDATE: APRIL 2020)

- 2.5 billion monthly active users
- 5 billion comments are left on Facebook pages monthly
- 55 million status updates are made every day
- Every 60 seconds
  - 317,000 status updates
  - 147,000 photos uploaded
  - 54,000 shared links

https://www.omnicoreagency.com/facebook-statistics/

DATA CLEANING

DATA CLEANING

Dimensions:
- Completeness
- Accuracy
- Consistency

- Data in the real world is dirty
- Incomplete:
  - Lacking attribute values, lacking certain attributes of interest, or containing only aggregate values
  - e.g., Occupation=“”
- Noisy:
  - Containing errors or outliers (spelling, phonetic and typing errors, word transpositions, multiple values in a single free-form field)
  - e.g., Salary=“-10”
- Inconsistent:
  - Containing discrepancies in codes or names (synonyms and nicknames, prefix and suffix variations, abbreviations, truncation and initials)
  - e.g., Age=“42” Birthday=“03/07/1997”
  - e.g., Was rating “1,2,3”, now rating “A, B, C”

https://slideplayer.com/slide/5116671/

WHY IS DATA DIRTY?

- **Incomplete data** comes from
  - Non available data value when collected
  - Different criteria between the time when the data was collected and when it is analyzed
  - Human/hardware/software problems
- **Noisy data** comes from
  - Data collection: faulty instruments
  - Data entry: human or computer errors
  - Data transmission
- **Inconsistent** (and redundant) data comes from:
  - Different data sources, so non uniform naming conventions/data codes
  - Functional dependency and/or referential integrity violation

Low quality data = low quality decisions!!

MISSING VALUES

- Missing can be a column, a value, a label
- Patterns of missing data
  - Missing completely at random (MCAR)
    - **No difference** between our primary variable of interest and the missing and non-missing values
  - Missing at random (MAR)
    - **No significant difference** between our primary variable of interest and the missing and non-missing values
    - Not a realistic assumption for many real-time data
  - Missing not at random (MNAR)
    - Depending on other values
    - Non-ignorable

DEALING WITH MISSING VALUES

- Delete the missing data
- Ignore the missing data
  - Applying methods unaffected by the missing values
- Fill in missing values
  - Manually
  - Use global constant such as "N/A" or "Unknown"
  - Use an imputation method
    - Expectation–Maximization algorithm

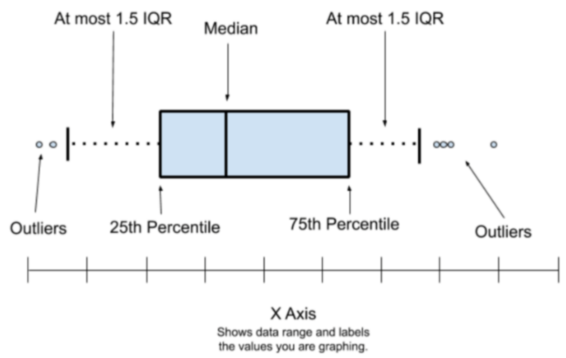
MISSING VALUES – IMPUTATION METHODS

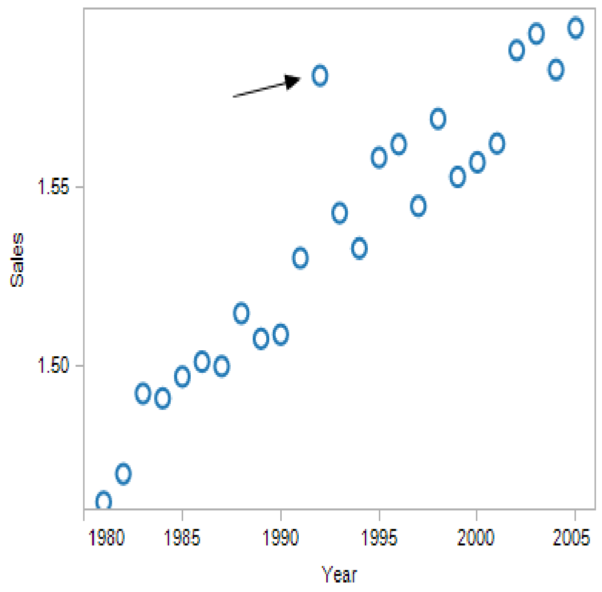
Motivation: Data is expensive to collect => replace the missing values with some possible values (minimize the bias)

Imputation methods = process of estimating missing data of an observation, based on valid values of other variables

- Hot deck imputation
  - Random observed value
- Mean/majority/median/mode-based imputation
- Imputation using regression or a decision tree to predict the missing values
- …
- K-Nearest Neighbors
  - E.g. estimate from the same class of data (not all) to narrow the down the scope

NOISY DATA





- Noise is a random error or variance in a measured variable
- Box plot / scatter plot / … can help to find outliers
- Outliers:
  - Remove
  - Have an extra group / statistical methods
    - May need other strategy for processing
- Data smoothing techniques:
  - Binning - the sorted values are divided into 'bins' and values are replaced by using the values around them
    - e.g., with mean/median of the given bin
  - Clustering - group values in clusters and then detect and remove outliers (automatic or manual)
  - Regression - fitting the data into regression functions, i.e. linear regression

INCONSISTENT / INVALID DATA

- Inconsistent representation of the same real world object in the database
  - E.g. "Raspberry", "raspberry", "RASPBERRY"; "Raspberry pi", "Raspberry pie";
- Solutions:
  - Domain/business knowledge
    - Sometimes only the domain expert can fix it
      - E.g. pi vs. pie
  - Levenshtein distance
  - Association Rule
  - Clustering
  - …
  - Remove

DATA TYPE ISSUES

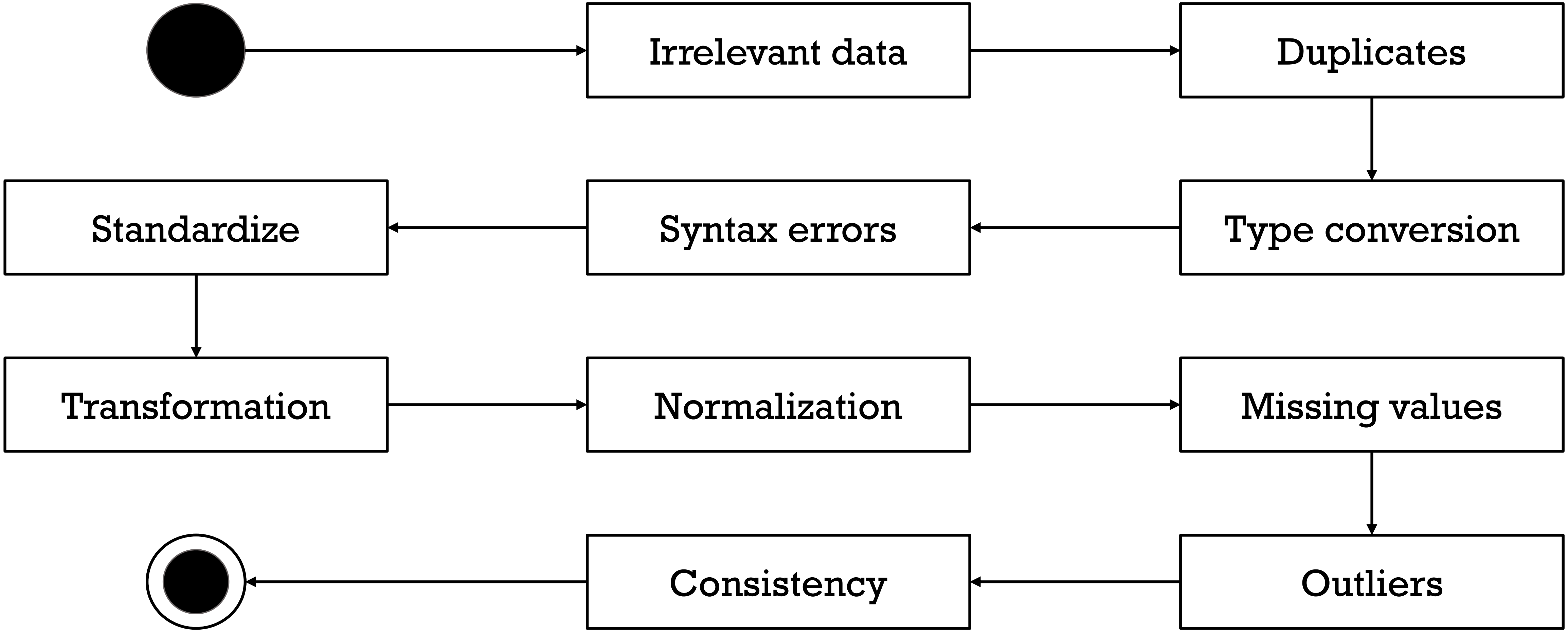
String

- Standardize casing
- Remove whitespaces, new lines
- Correcting typos
- Standardize encoding
- Map to type / categorical variables
- Remove stop words

Date and time

- Format
- Time zones

DATA CLEANING



DATA CLEANING

- Irrelevant data: Not actually needed
- Duplicates: Points that are repeated in your dataset.
- Type conversion: Numbers are stored as numerical data types, …
- Syntax errors: Remove white spaces, pad strings, fix typos, …
- Standardize: For strings, make sure all values are either in lower or upper case. For numerical values, make sure all values have a certain measurement unit.
- Scaling / Transformation
- Normalization: If we're going to be using statistical methods that rely on normally distributed data (e.g., log function)
- Missing values
- Outliers : Outliers are innocent until proven guilty
- In-record & cross-datasets errors : These errors result from having two or more values in the same row or across datasets that contradict with each other.

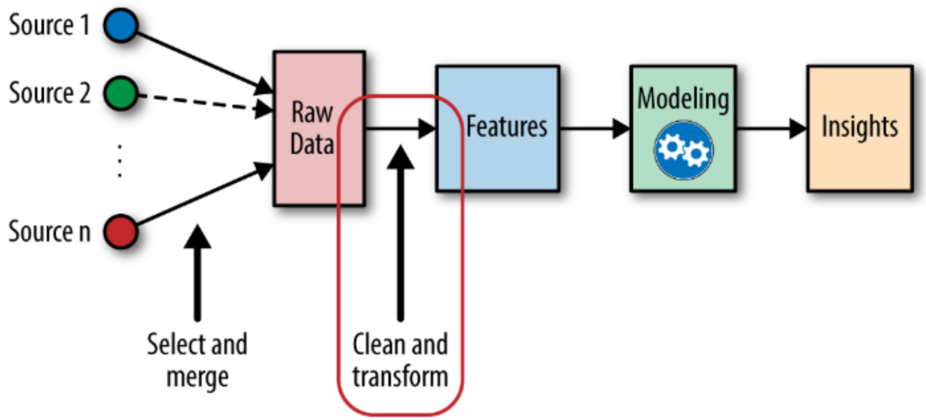
DATA QUALITY

- Assesses whether information can serve its purpose in a particular context

Characteristic	How it’s measured
Accuracy	Is the information correct in every detail?
Completeness	How comprehensive is the information?
Reliability	Does the information contradict other trusted resources?
Relevance	Do you really need this information?
Timeliness	How up- to-date is information? Can it be used for real-time reporting?

DATA TRANSFORMATION

FEATURE ENGINEERING



- Process of using domain knowledge to select and transform the most relevant variables from raw data into features (variables) that can be used in ML (DS) tasks
- **Feature Creation:** Identifying the variables that will be most useful in the predictive model
  - Subjective process - requires human intervention and creativity
  - Features can be combined to create new derived features that have greater predictive power
    - Mixed via addition, subtraction, multiplication, ratio…
- **Transformations:** Manipulating the predictor variables to improve model performance
  - E.g. variables are on the same scale, within an acceptable range, have suitable formats and variety, …
- **Feature Extraction:** Feature extraction is the automatic creation of new variables by extracting them from raw data
  - Automatically reduce the volume of data into a more manageable set
  - E.g., cluster analysis, text analytics, edge detection algorithms, principal components analysis, …
- **Feature Selection:** Analyse, judge, and rank various features to determine which features are irrelevant, redundant, most useful, …

Feature selection = keeps a subset of the original features
Feature extraction = creates brand new ones

https://www.omnisci.com/technical-glossary/feature-engineering

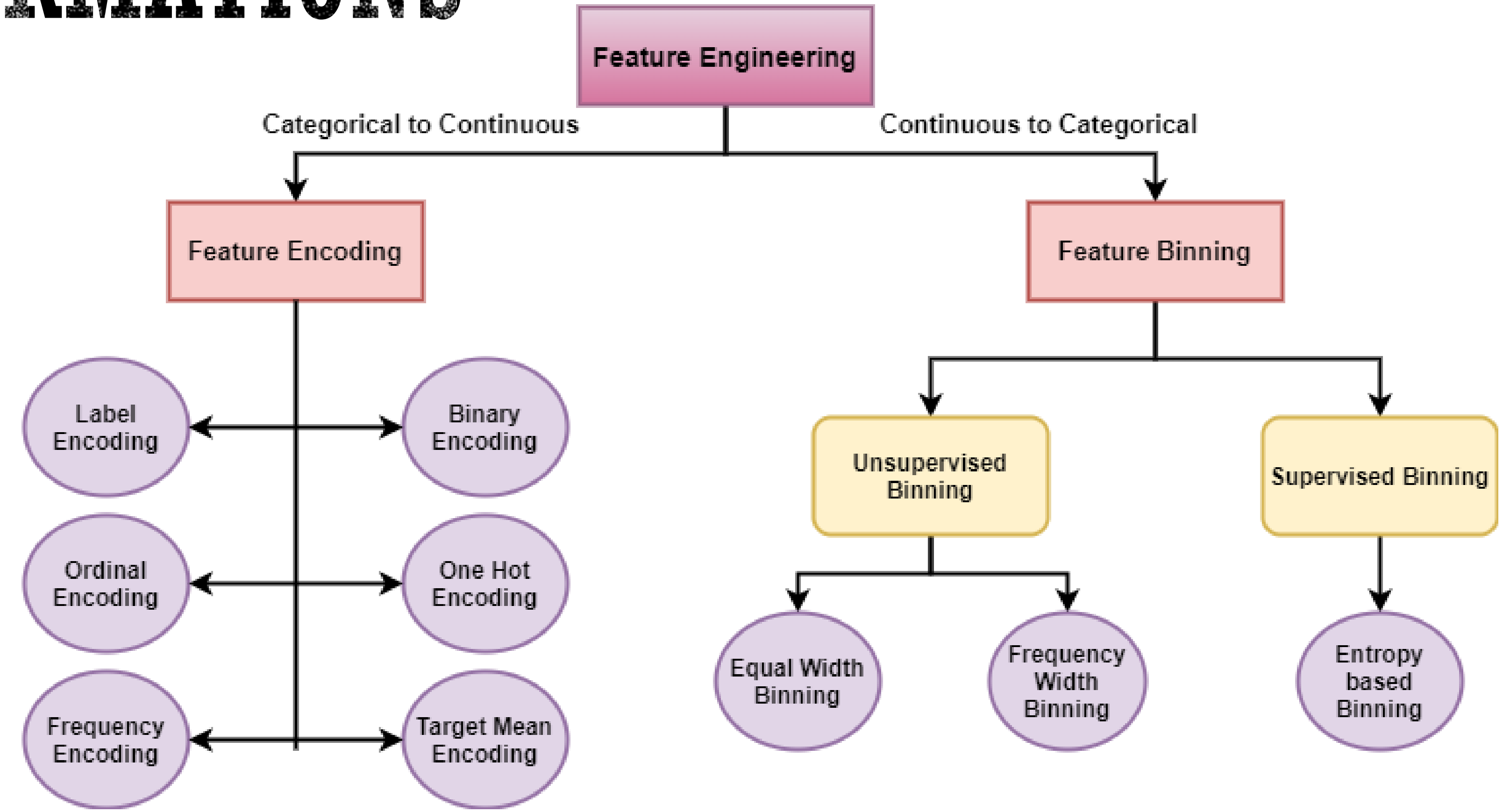
BASIC DATA TRANSFORMATIONS

- Some tools/approaches need data to be only numeric…
- Assign numbers
  - 50 US states => 50 numbers
- Use numbers so that you can compare
  - A, B+, B, B-, … => 5.0, 4.8, 4.2, …
- Multi-valued with small domain – (vector of) binary flags
  - Colors of a species
- Many values
  - Natural grouping: 50 US states => 4-5 regions
  - Groups for most frequent, one group for the rest

BASIC DATA TRANSFORMATIONS

- Feature binarization
  - Thresholding numerical features to get Boolean values
- Dataset standardization
  - Distributions of values of different features can be radically different
  - Move the center (toward zero mean) and scale (towards unit variance)
- Vector normalization
  - Scaling individual samples to have unit norm
- Simple feature selection
  - Remove all but the **k** highest scoring features, a user-specified highest scoring percentile of features, …

CATEGORICAL TO/FROM CONTINUOUS TRANSFORMATIONS



https://towardsdatascience.com/feature-engineering-deep-dive-into-encoding-and-binning-techniques-5618d55a6b38

FEATURE ENCODING

- Transformation of a categorical feature into a numerical variable
- Most of the ML algorithms cannot handle categorical variables

Label Encoding

- Assigning a numerical value to each of the categories
- Can be used for ordinal variables

[male, female]	[0, 1]
[blue, green, red, black]	[0, 1, 2, 3]
[10, 21], [22, 33], [34, 45], [46, 55]	[0, 1, 2, 3]

Ordinal encoding

- Transform an original categorical variable to a numerical variable by ensuring the ordinal nature of the variables is sustained

[male, female]	[0, 1]
[10, 21], [22, 33], [34, 45], [46, 55]	[0, 1, 2, 3]
[cold, warm, hot]	[0, 1, 2]
[poor, fair, good, very good, excellent]	[0, 1, 2, 3, 4]

FEATURE ENCODING

Frequency encoding

- Considering the frequency distribution of the data
- Useful for nominal features

Column	Freq_Encoding
red	5
green	3
red	5
green	3
blue	4
red	5
red	5
blue	4
red	5
blue	4
blue	4
green	3

Binary encoding

- Encoding the categories as integer and converted into binary code
- For variables having a large number of categories
- Example:100 category variable
  - Label Encoding: 100 categories
  - Binary encoding: 7 categories
  - One-hot encoding: 100 columns
    - See next slide

Column	Label Enc	Binary enc1	Binary enc2	Binary enc3
red	1	0	0	1
green	2	0	1	0
red	1	0	0	1
green	2	0	1	0
blue	3	0	1	1
red	1	0	0	1
grey	4	1	0	0
blue	3	0	1	1
red	1	0	0	1
blue	3	0	1	1
blue	3	0	1	1
green	2	0	1	0
grey	4	1	0	0

FEATURE ENCODING

One hot encoding

- Creates ***k*** different columns each for a category and replaces one column with 1 rest is 0

(Target) Mean encoding

- Takes the target class label into account
- Idea: to replace the categorical variable with the mean of its corresponding target variable
  - Ratio of the occurrence of the positive class in the target variable
  - Probability of the target variable, conditional on each value of the feature
- Encoding task + feature that is more representative of the target variable

Column	red	green	blue
red	1	0	0
green	0	1	0
red	1	0	0
green	0	1	0
blue	0	0	1
red	1	0	0
red	1	0	0
blue	0	0	1
red	1	0	0
blue	0	0	1
blue	0	0	1
green	0	1	0

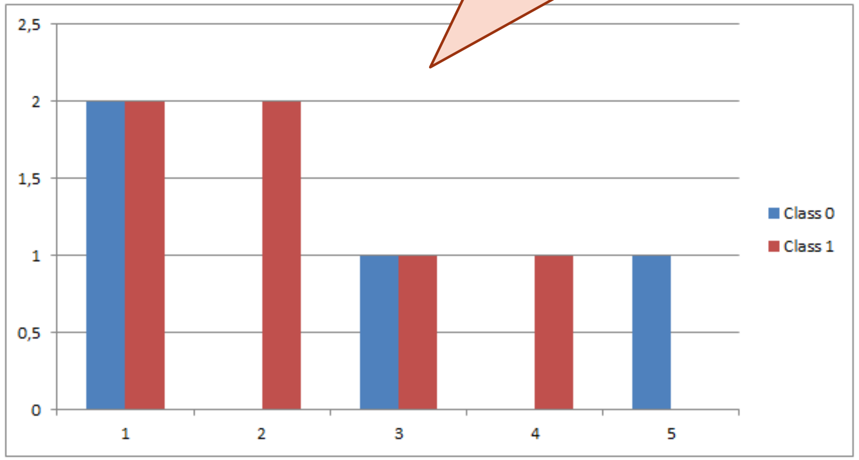
Column	Target	Target Mean	Target Mean (numerical value)
red	1	3/5	0.6
green	1	2/3	0.67
red	0	3/5	0.6
green	0	2/3	0.67
blue	1	2/4	0.5
red	0	3/5	0.6
red	1	3/5	0.6
blue	0	2/4	0.5
red	1	3/5	0.6
blue	0	2/4	0.5
blue	1	2/4	0.5
green	1	2/3	0.67

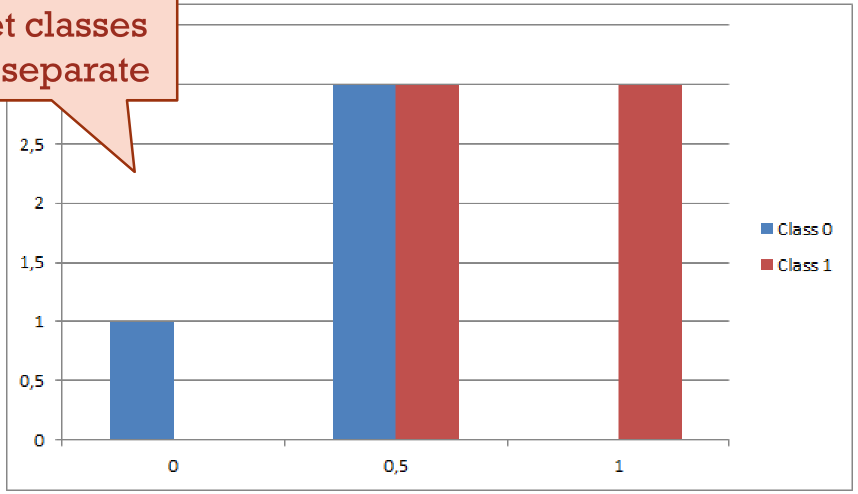
MEAN ENCODING – EXAMPLE

id	job	age	target
1	Doctor	54	1
2	Doctor	35	0
3	Doctor	28	1
4	Doctor	75	0
5	Teacher	29	1
6	Teacher	37	1
7	Engineer	60	0
8	Engineer	38	1
9	Waiter	31	1
10	Driver	22	0

id	job	job_label	target
1	Doctor	1	1
2	Doctor	1	0
3	Doctor	1	1
4	Doctor	1	0
5	Teacher	2	1
6	Teacher	2	1
7	Engineer	3	0
8	Engineer	3	1
9	Waiter	4	1
10	Driver	5	0

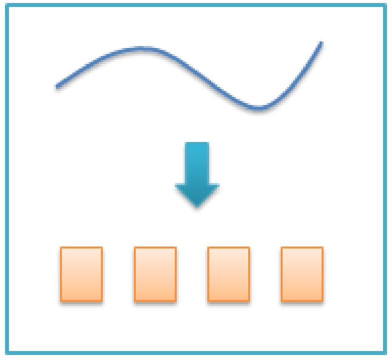
id	job	job_mean	target
1	Doctor	0,50	1
2	Doctor	0,50	0
3	Doctor	0,50	1
4	Doctor	0,50	0
5	Teacher	1	1
6	Teacher	1	1
7	Engineer	0,50	0
8	Engineer	0,50	1
9	Waiter	1	1
10	Driver	0	0





https://towardsdatascience.com/why-you-should-try-mean-encoding-17057262cd0

DISCRETIZATION OF CONTINUOUS VALUES (BINNING)



- Divide the range of a continuous attribute into intervals
  - Some methods require discrete values (e.g. most versions of Naïve Bayes)
- Reduction of data size
- Supervised vs. unsupervised

UNSUPERVISED BINNING

Equal width binning

- Divides the range into ***N*** intervals of equal size (range)
- Uniform grid
- Width of intervals:
  W = (max_value – min_value) / N
- Pros:
  - Simple, easy to implement
  - Reasonable abstraction of data
    - E.g., people 30+, 40+, 50+
- Cons:
  - How to set ***N***?
  - Sensitive to outliers
  - Unsupervised

Equal depth (height, frequency) binning

- ***N*** intervals, each containing approximately the same number of samples
- Pros:
  - Avoids clumping
  - More intuitive breakpoints

EXAMPLE

$$X = [10, 15, 16, 18, 20, 30, 35, 42, 48, 50, 52, 55]$$

N = 4
W = (55 - 10)/4 = 12

[10,21]
[22.33]
[34,45]
[46,55]

AGE	AGE_bins
10	[10, 21]
15	[10, 21]
16	[10, 21]
18	[10, 21]
20	[10, 21]
30	[22, 33]
35	[34, 45]
42	[34, 45]
48	[46, 55]
50	[46, 55]
52	[46, 55]
55	[46, 55]

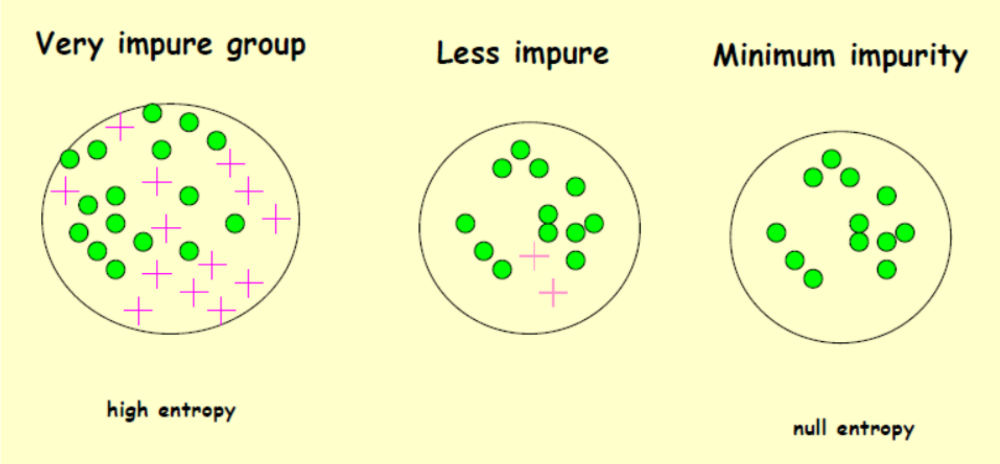
AGE	AGE_bins
10	[10, 16]
15	[10, 16]
16	[10, 16]
18	[17, 30]
20	[17, 30]
30	[17, 30]
35	[31, 48]
42	[31, 48]
48	[31, 48]
50	[49, 55]
52	[49, 55]
55	[49, 55]

https://towardsdatascience.com/feature-engineering-deep-dive-into-encoding-and-binning-techniques-5618d55a6b38

UNSUPERVISED BINNING

- **Rank**
  - Rank of a number = its size relative to other values
  - Sort the list of values, then assign the position of a value as its rank
    - Same values receive the same rank
    - The presence of duplicate values affects the ranks of subsequent values
      - e.g., 1,2,3,3,5
  - Drawback: values can have different ranks in different lists
- **Quantiles** (median, quartiles, percentiles, ...)
  - Useful
  - Same problem as Rank
- **Math functions**:
  - E.g., FLOOR(LOG(X)) - effective binning method for the numerical variables with highly skewed distribution (e.g., income)

SUPERVISED BINNING



- Considers target (class, label) into account

Entropy-based binning

- Uses a split approach
- Entropy (information content) is calculated based on the class label
- Sort the input and iteratively (until a reasonable solution is found) find the best split so that the bins are as pure as possible
  - The majority of the values in a bin have the same class label
  - The split with maximal information gain
- S – set of data
- $C_1, \ldots C_n$ – classes
- p_c – proportion of C_c in S
- Measure of the impurity (entropy):

$$\mathrm{Impurity}(S) = -\sum_{c=1}^{N} p_c \cdot \log_2 p_c$$

SUPERVISED BINNING – EXAMPLE

O-Ring Failure	Temperature
Y	53
Y	56
Y	57
N	63
N	66
N	67
N	67
N	67
N	68
N	69
N	70
Y	70
Y	70
Y	70
N	72
N	73
N	75
Y	75
N	76
N	76
N	78
N	79
N	80
N	81

1. Calculate "Entropy" for the target

$$E(S) = \sum_{i=1}^{c} -p_i \log_2 p_i$$

O-Ring Failure	
Y	N
7	17

p(Y) = 7/24 = 0.29
p(N) = 17/24 = 0.71
E (Failure) = E(7,17) = -0.29 x $\log_2(0.29)$ - 0.71 x $\log_2(0.71)$
= **0.871**

2. Calculate "Entropy" for the target given a bin

$$E(S,A) = \sum_{v \in A} \frac{|S_v|}{|S|} E(S_v)$$

		O-Ring Failure	
		Y	N
Temp.	<= 60	3	0
	> 60	4	17

E (Failure,Temperature) = p(<=60) x E(3,0) + p(>60) x E(4,17) =
3/24 x 0 + 21/24 x 0.7= **0.615**

Gain = 0.256

		O-Ring Failure	
		Y	N
Temperature	<= 60	3	0
	> 60	4	17

Gain = 0.101

		O-Ring Failure	
		Y	N
Temperature	<= 70	6	8
	> 70	1	9

Gain = 0.148

		O-Ring Failure	
		Y	N
Temperature	<= 75	7	11
	> 75	0	6

3. Calculate "Information Gain" given a bin

$$\textit{Information Gain} = E(S) - E(S,A)$$

Information Gain (Failure, Temperature) = **0.256**

https://www.saedsayad.com/supervised_binning.htm

REFERENCES

- Dorian Pyle: Data Preparation for Data Mining
- https://www.cs.princeton.edu/courses/archive/spring10/cos424/slides/18-feat.pdf
- https://paginas.fe.up.pt/~ec/
- https://towardsdatascience.com/all-about-categorical-variable-encoding-305f3361fd02
- Courses:
  - NDBI046 – Data Management http://skoda.projekty.ms.mff.cuni.cz/NDBI046/
  - NPFL104 – Machine Learning Methods https://ufal.mff.cuni.cz/courses/npfl104#classes