02.08.2024

Statistical Methods in AI (CS7.403)

Lecture-2: ML Workflow, Data Representations, Basic Data Transformations, Data Visualization

Ravi Kiran (ravi.kiran@iiit.ac.in)

https://ravika.github.io

@vikataravi

Center for Visual Information Technology (CVIT)

IIIT Hyderabad

Announcements

- **Update** on one-time bonus rule for assignments
  - You are allowed to use a maximum of 3 days for assignment extensions
  - NEW: You can split them across assignments, BUT each usage needs to be at least 1 day
    - ➔ you can use the assignment extension bonus for at most 3 of your assignments

Announcements

- Tutorial (11.40a – 1.05p Saturday)
  - SH2 (ONLY FOR THIS WEEK)
  - H-205 (NEXT WEEK ONWARDS)
- TOPICS: Python, Pandas, Jupyter notebook, Colab, Plotting tools.
- **Bring your (fully charged) laptops**.
- Ask questions.

Announcements

- IMPORTANT: All assignments/projects will need to submitted via Github Classroom
- Tutorial
  - Git
  - Github
- Ask questions.

Announcements

- TAs will share SMAI Course Calendar on Moodle
- You can add it to your MS Teams Calendar
- Will contain assignment release/due/eval dates
- Will contain exam paper showing dates

Queries

- Post queries on Moodle
- Helps all (many may have same question)
- Do not DM TAs !

Announcements

- Do <u>not</u> use Python libraries for assignments unless explicitly allowed/specified.

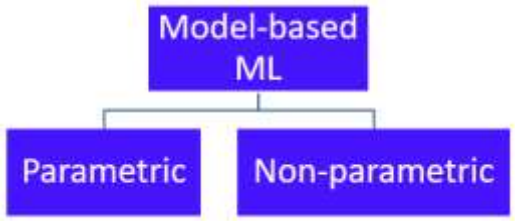
Additionally …

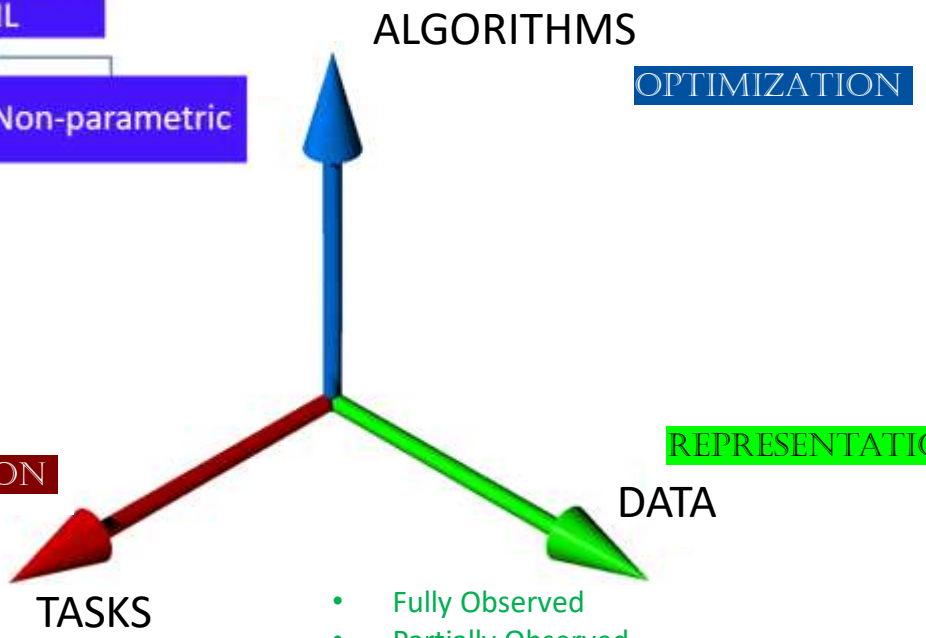
- Spending time <u>everyday</u> on material covered in class helps
  - Take notes
  - Revise
  - Reflect
- **Ask if you wish to take something down, but slide is no longer on screen**

Course TAs

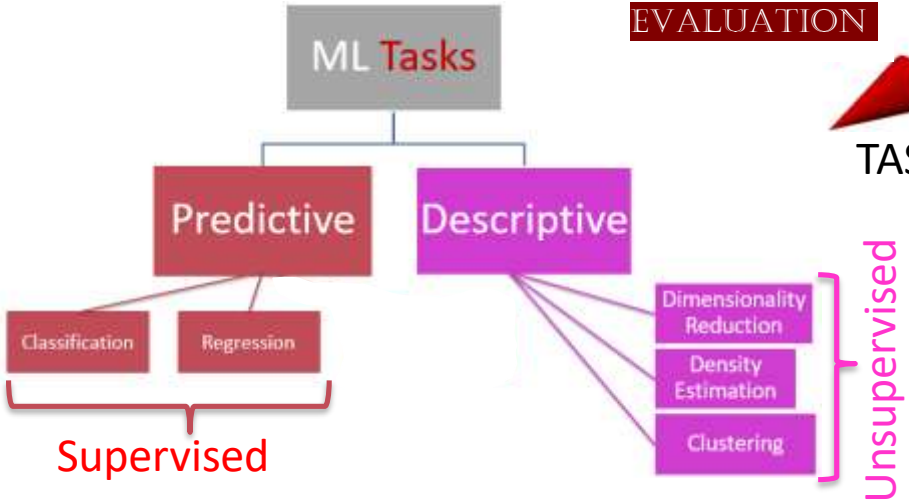
Meghana (SERC, CSE4)
Rudransh (CVIT, CSE4)
Sriram (C2S2, CSE4)
Veda Nivas (CSE4)
Vanshita (CSE4)
Harshavardhan (CVIT, CSE4-DD)
Vedansh (CogSci, CSE4-DD)
Arghya (CogSci, CSE4-DD)
Sanika (HSRC, CSE4-DD)

Recap





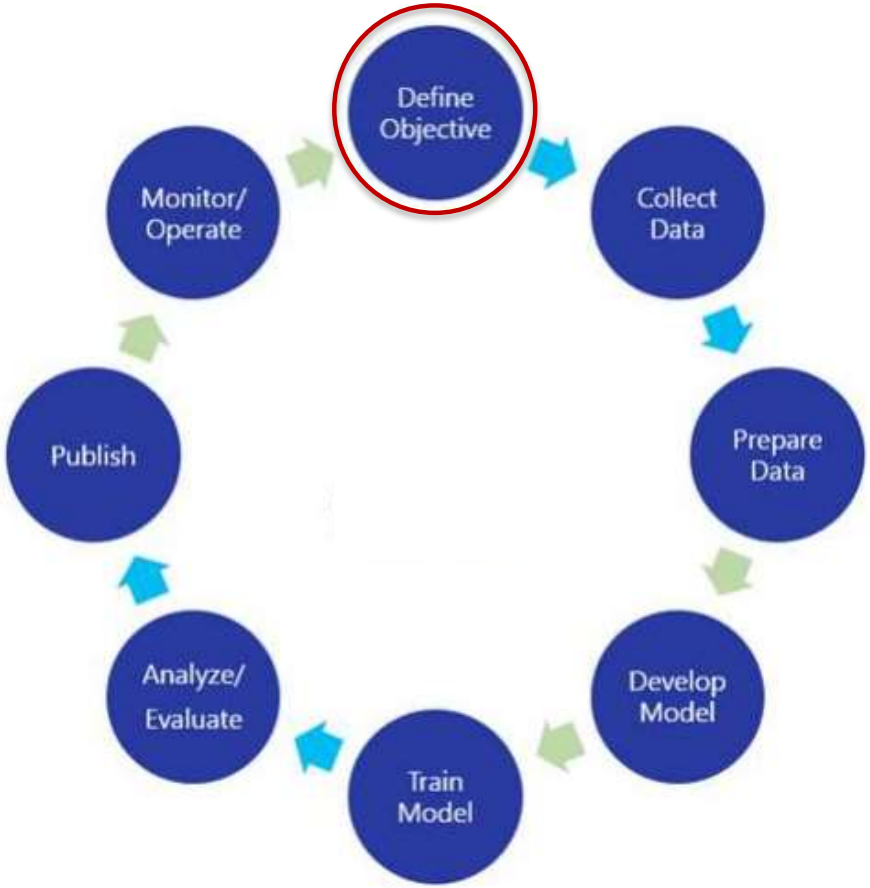
- Fully Observed
- Partially Observed
  - Some variables systematically not observed (e.g. 'topic' of a document)
  - Some variables missing some of the time (e.g. 'faulty sensor' readings)



Lecture Outline

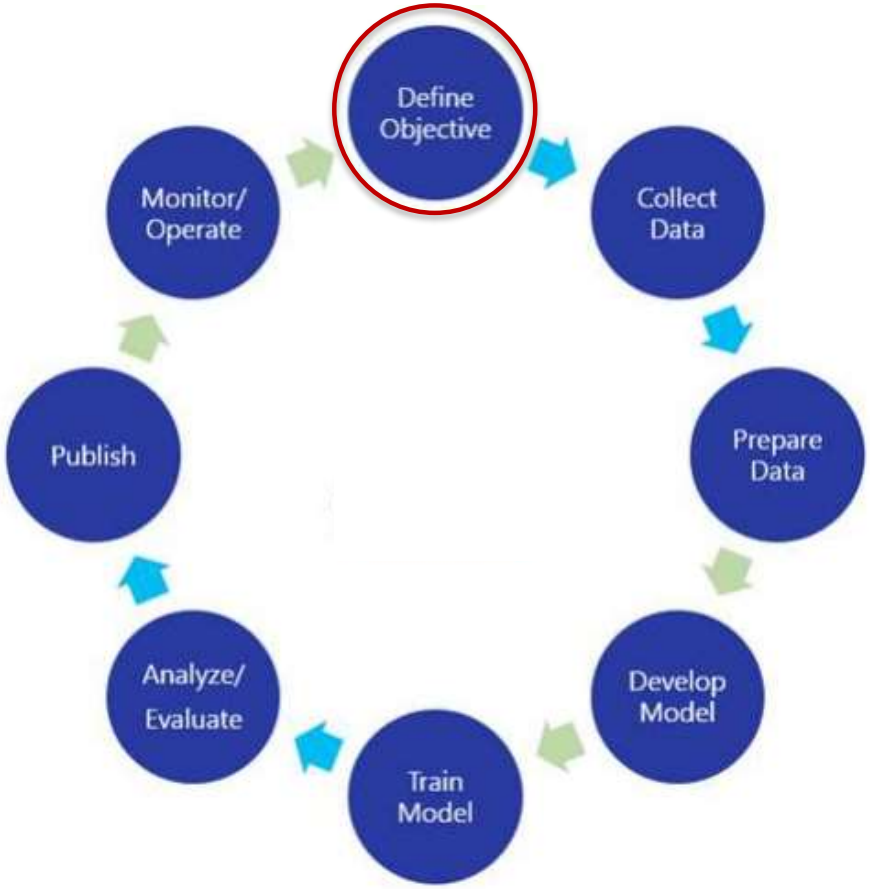
- ML Workflow
- Data Representations
- Basic Data Transformations
- Data Visualization

Workflow of a Machine Learning Problem



- Detect spam email
- Predict value of a stock
- Predict effect of advertising on sales
- Drive car ‘safely’ without human intervention
- Translate text from one language to another
- Sentiment Analysis
- …

Workflow of a Machine Learning Problem



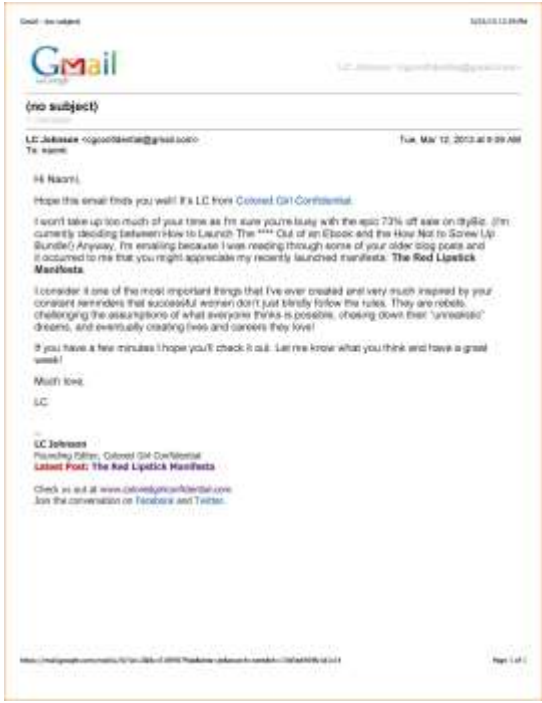
No Data, no ML !

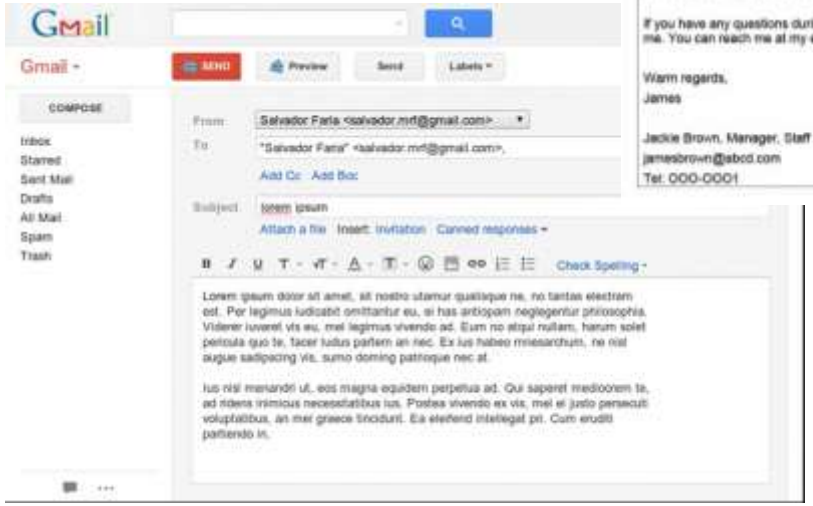

Without data you're just another person with an opinion.
W. EDWARDS DEMING


"In God we trust, all others must bring data."
W. Edwards Deming
(c) iStockphoto / Thinkstock

Sources of data

- Detect spam email



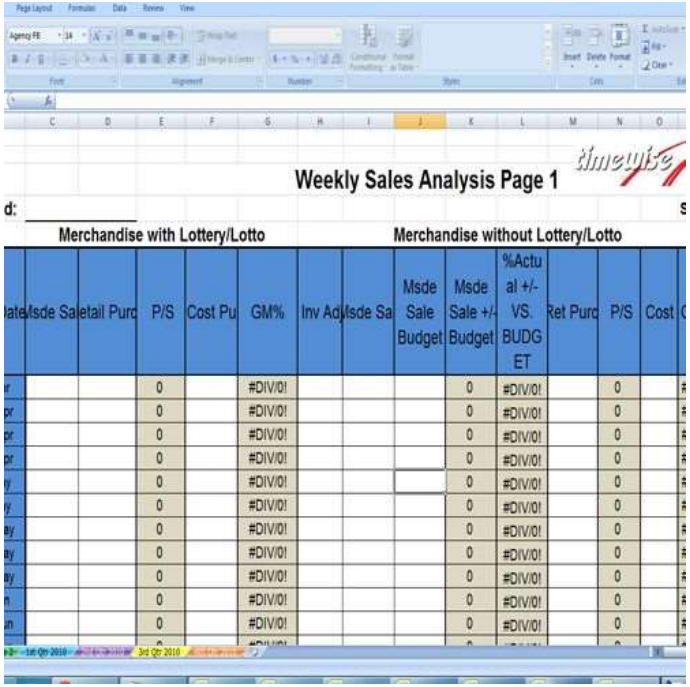


Sources of data

- Predict value of a stock

Sources of data

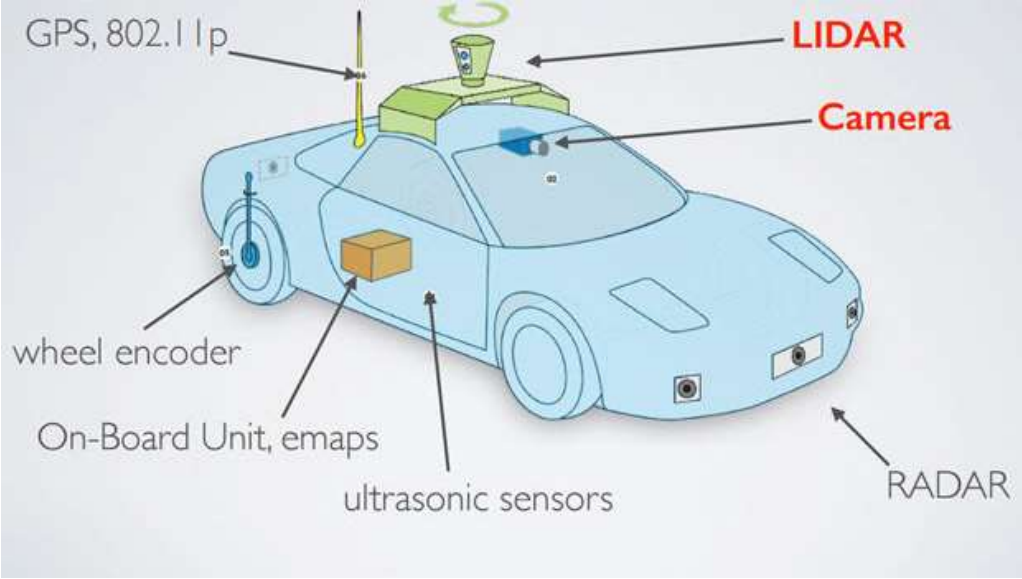
- Predict effect of advertising on sales



Raw Data may not always be digital in nature !

Sources of data

- Drive car safely without human intervention

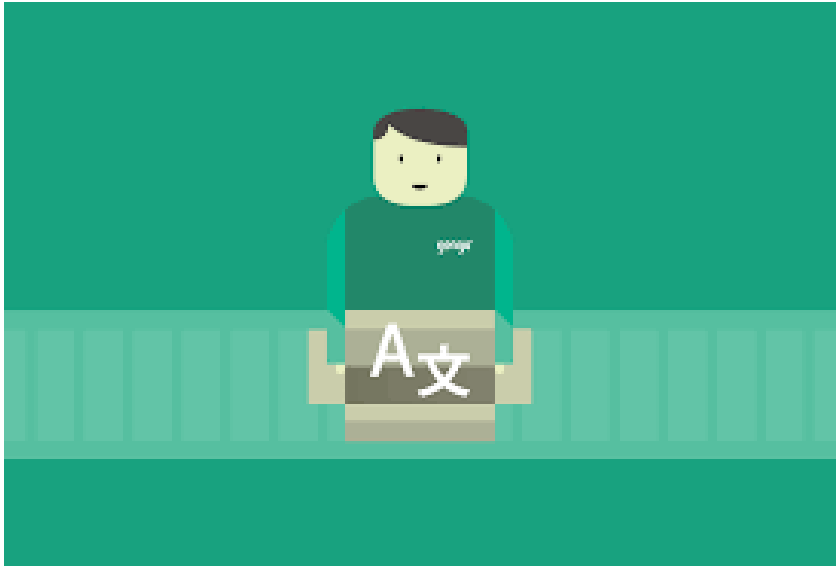


https://inai.iiit.ac.in/bodhyaan.html

Data can be multi-modal and may need to be 'synchronized'

Sources of data

- Translate text from one language to another



A human domain expert may be required to obtain raw data

Sources of data

kaggle

Google Dataset Search Beta

VisualData

UC Irvine Machine Learning Repository

OpenML

DATA HUB

Datasets

DATA

HealthData.gov

The NLP Index

zenodo

सत्यमेव जयते

data.gov.in

Open Government Data (OGD) Platform India

Three fundamental questions

- What data to collect ?
- How to collect ?
- How much to collect ?

Raw data

- May be too little in quantity

Raw data

- May be too much in quantity
  - Limitations on system end (compute, storage)

Raw data

- Not all of it relevant

https://api.mailgun.net/v2/domains/mailgun.com/messages/WyJlMTFiZ


```
{
    Received: "by luna.mailgun.net with HTTP; Fri, 26 Feb 2016 20:12:03 +0000",
    stripped-signature: "",
    Message-Id: "<20160226201203.54979.26875@mailgun.com>",
    from: "Sample Email <me@mailgun.com>",
    sender: "me@mailgun.com",
    recipients: "anton@mailgunhq.com",
    Subject: "Test Message",
    Content-Transfer-Encoding: "7bit",
    attachments: [ ],
    To: "anton@mailgunhq.com",
    stripped-html: "<p>Testing some Mailgun awesomness!</p>",
    content-id-map: { },
    stripped-text: "Testing some Mailgun awesomness!",
    From: "Sample Email <me@mailgun.com>",
  + message-headers: […],
    Mime-Version: "1.0",
    Content-Type: "text/plain; charset="ascii"",
    body-plain: "Testing some Mailgun awesomness!",
    subject: "Test Message"
}
```

Raw data

- Often not directly usable
  - Filter (needed data)
  - **Transform (to numerical data)**

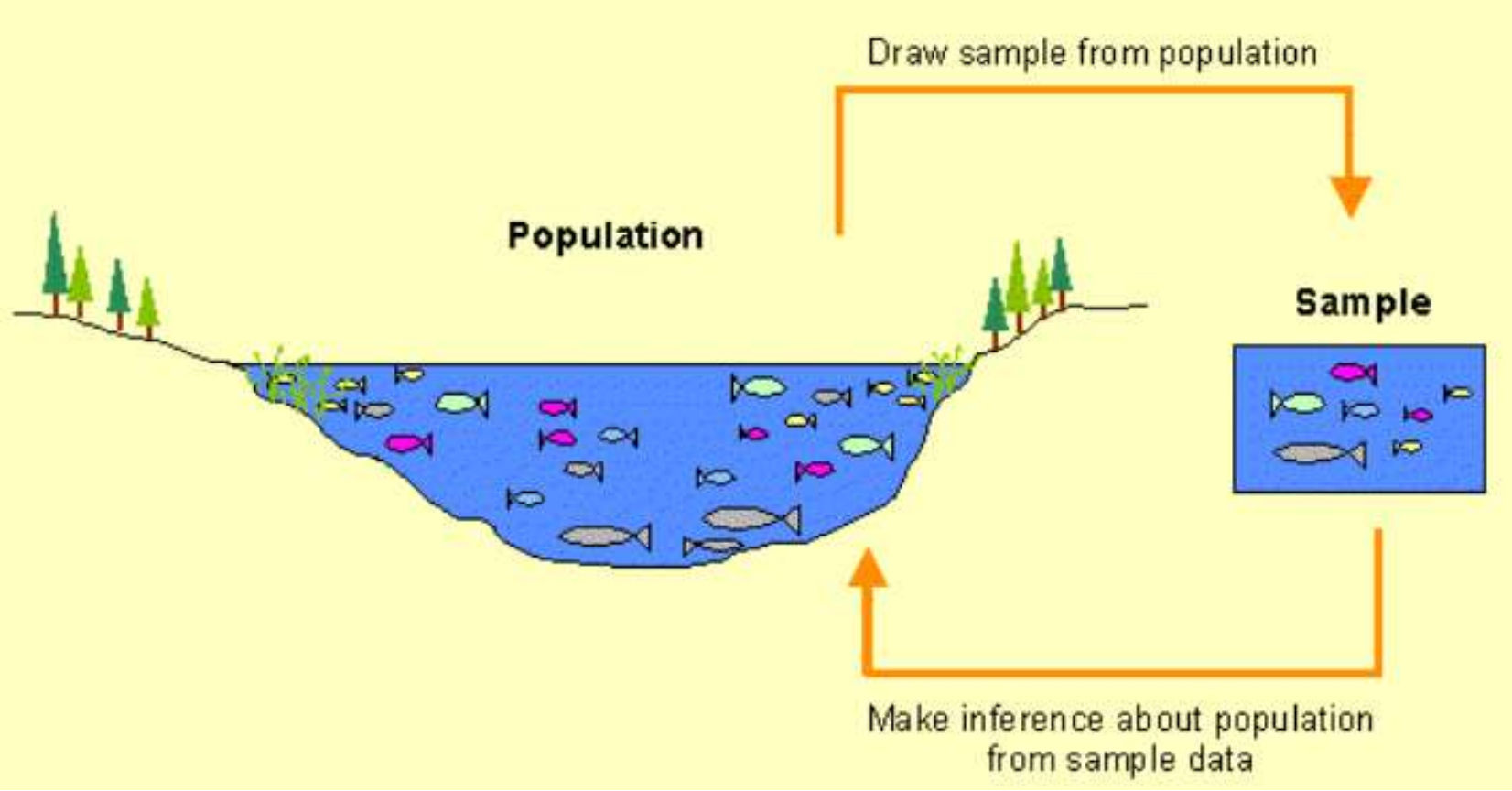
https://api.mailgun.net/v2/domains/mailgun.com/messages/WyJlMTFiZ


```
{
    Received: "by luna.mailgun.net with HTTP; Fri, 26 Feb 2016 20:12:03 +0000",
    stripped-signature: "",
    Message-Id: "<20160226201203.54979.26875@mailgun.com>",
    from: "Sample Email <me@mailgun.com>",
    sender: "me@mailgun.com",
    recipients: "anton@mailgunhq.com",
    Subject: "Test Message",
    Content-Transfer-Encoding: "7bit",
    attachments: [ ],
    To: "anton@mailgunhq.com",
    stripped-html: "<p>Testing some Mailgun awesomness!</p>",
    content-id-map: { },
    stripped-text: "Testing some Mailgun awesomness!",
    From: "Sample Email <me@mailgun.com>",
  + message-headers: […],
    Mime-Version: "1.0",
    Content-Type: "text/plain; charset="ascii"",
    body-plain: "Testing some Mailgun awesomness!",
    subject: "Test Message"
}
```

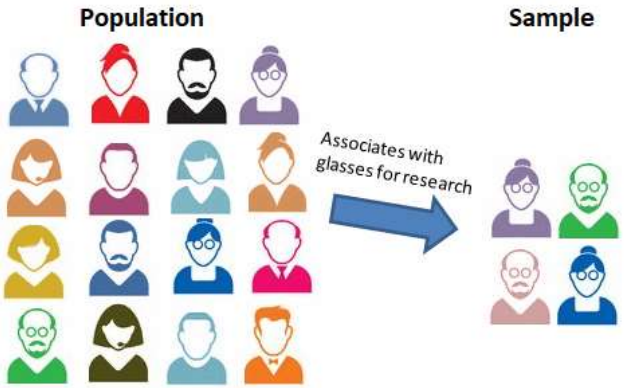
Three fundamental questions

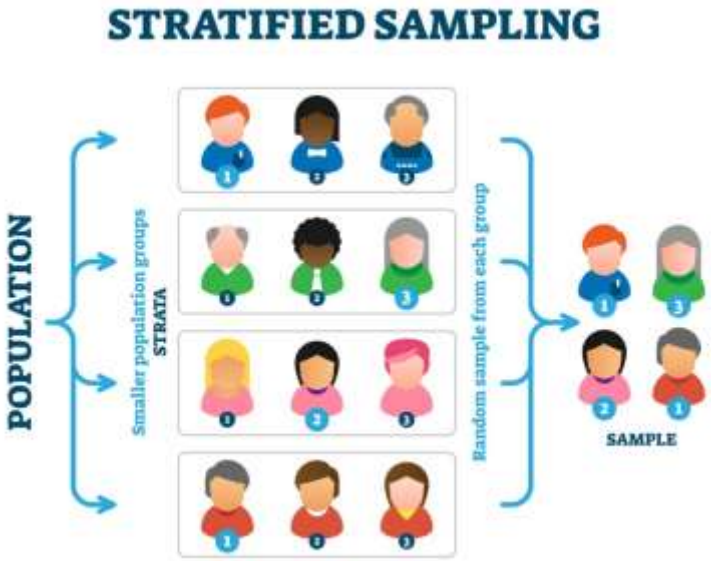
- What data to collect ?
- How to collect ?
- How much to collect ?

Sample v/s Population



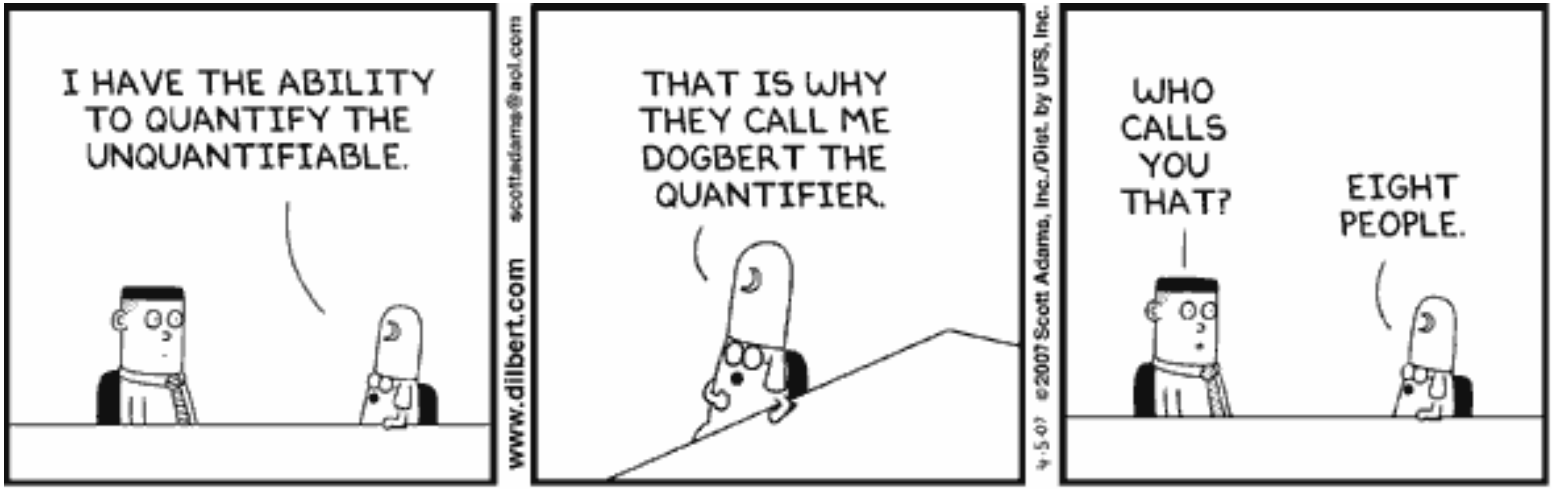
Data Sampling



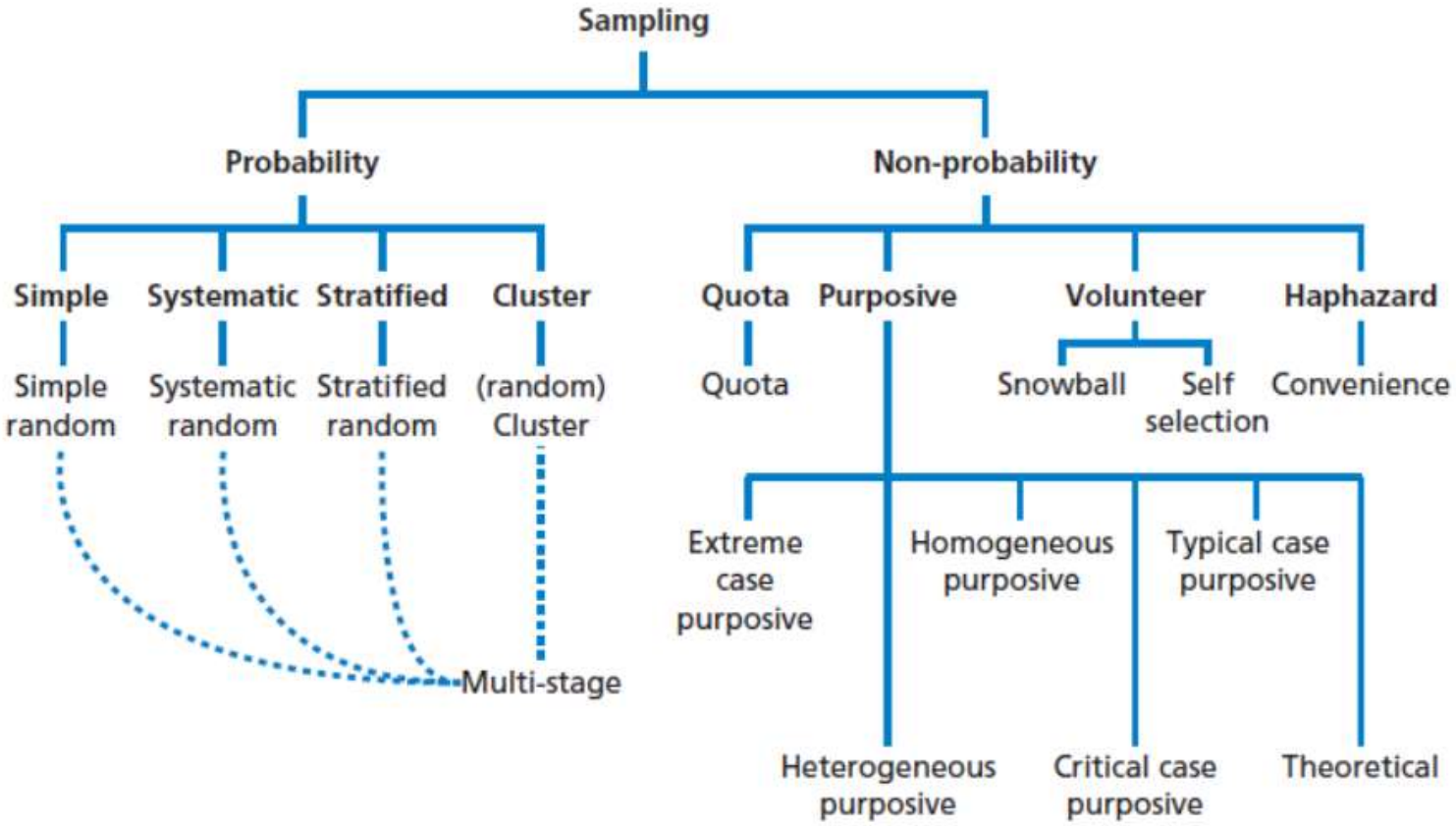


Important to understand how the dataset has been created

Are our samples ‘representative’ of the population?

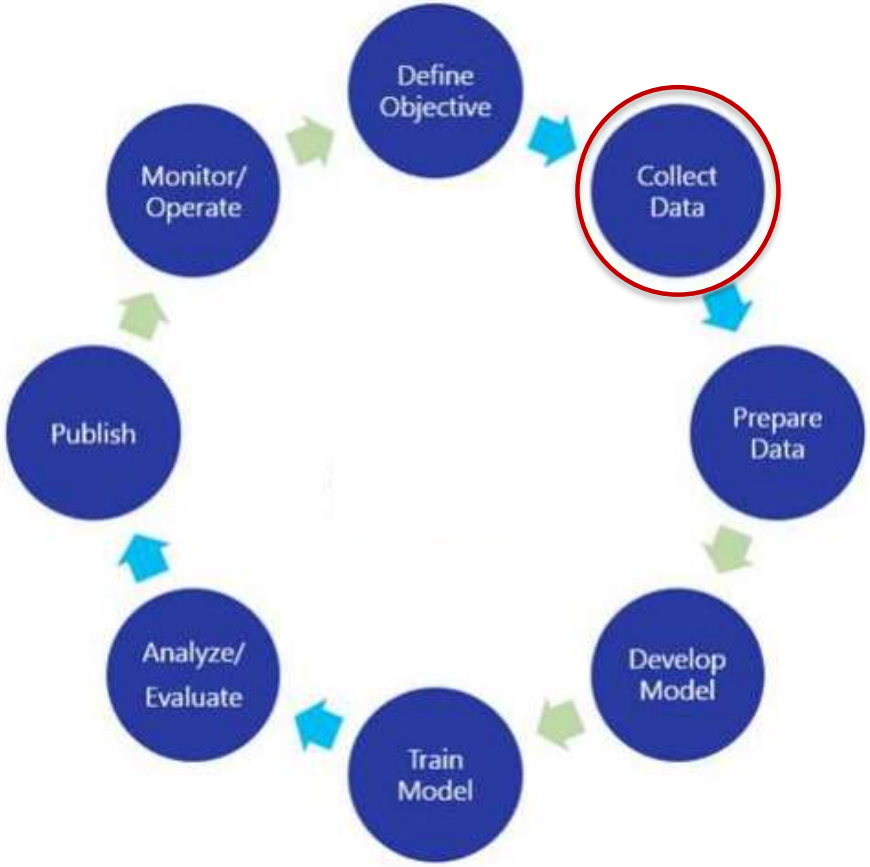


Sampling Techniques



https://research-methodology.net/sampling-in-primary-data-collection/

Workflow of a Machine Learning Problem



Taxonomy of data variables

Data Type

Counts, Measurements

Quantitative

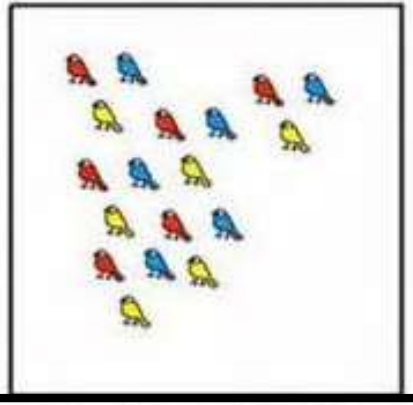
Attributes, Labels

Qualitative

13 Trees

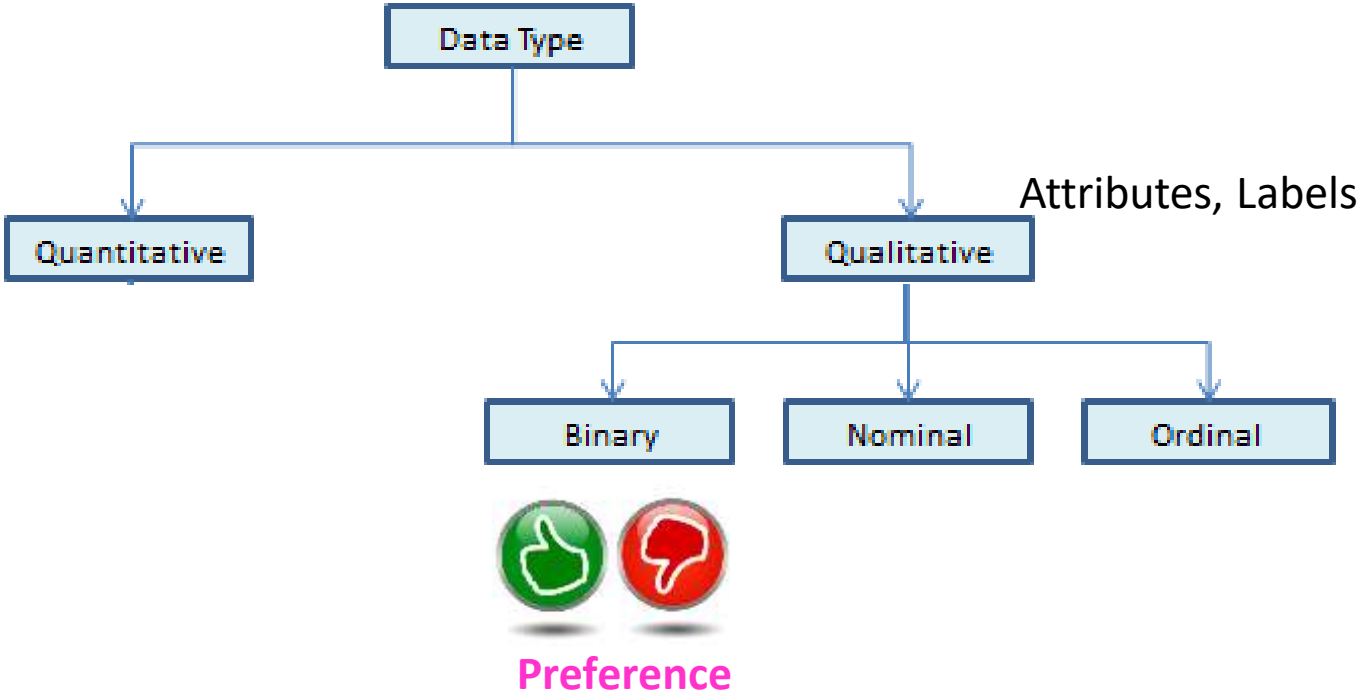
Data → VALUE

Information → VARIABLE

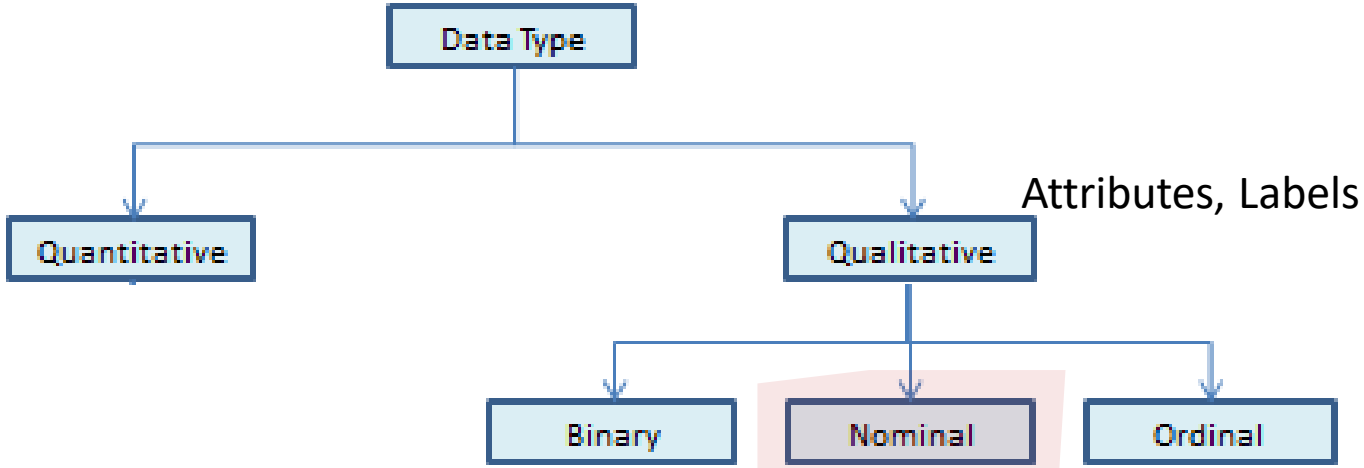


Blue, Red, and Yellow Birds

Taxonomy of data



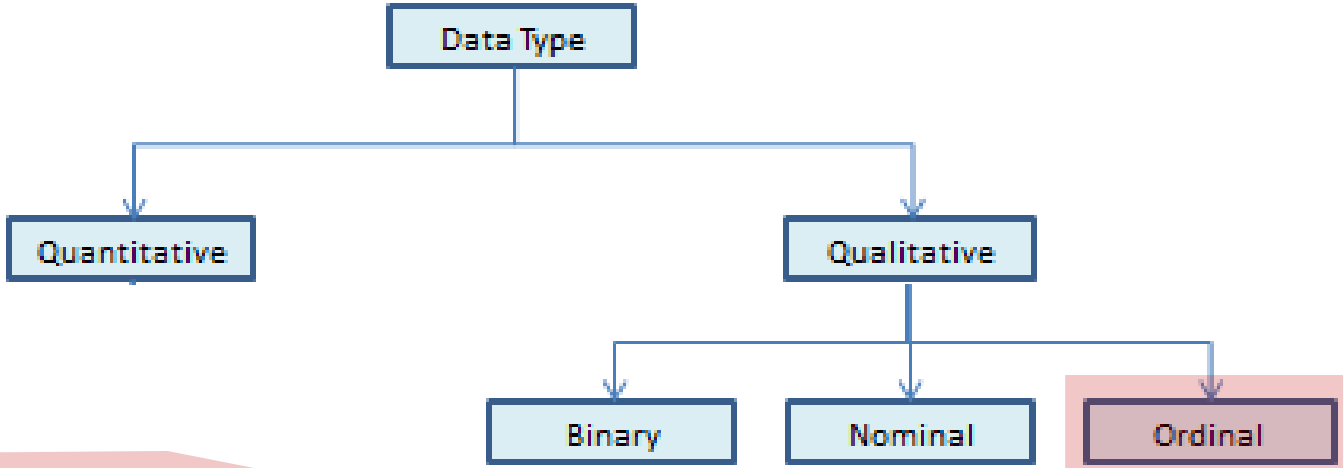
Taxonomy of data



Color

Make

Pin Code

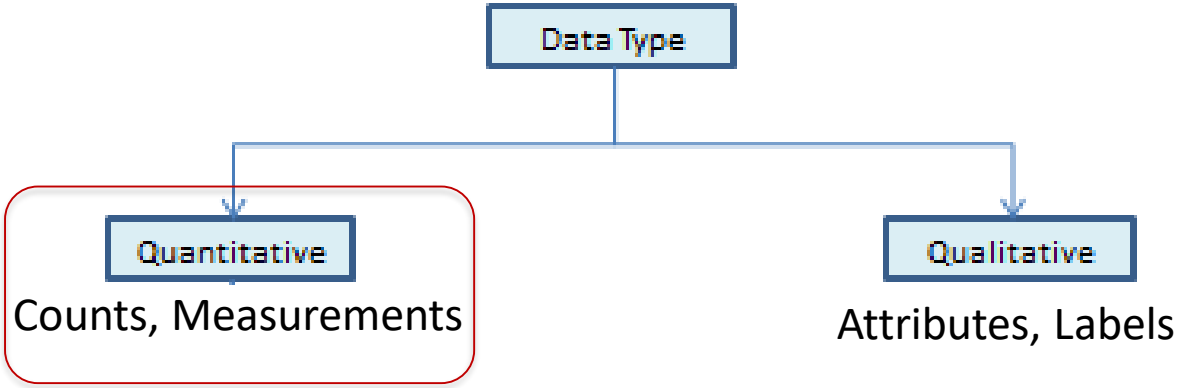


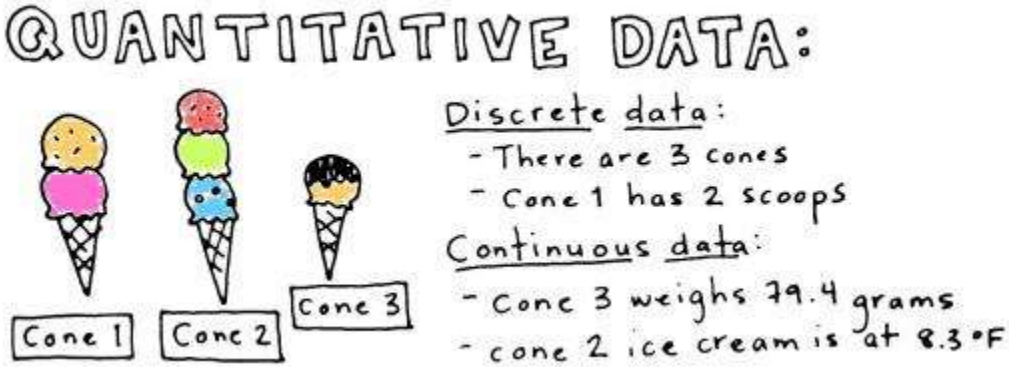
How comfortable are you with Python *

No knowledge ○ ○ ○ ○ ○ ○ Very comfortable

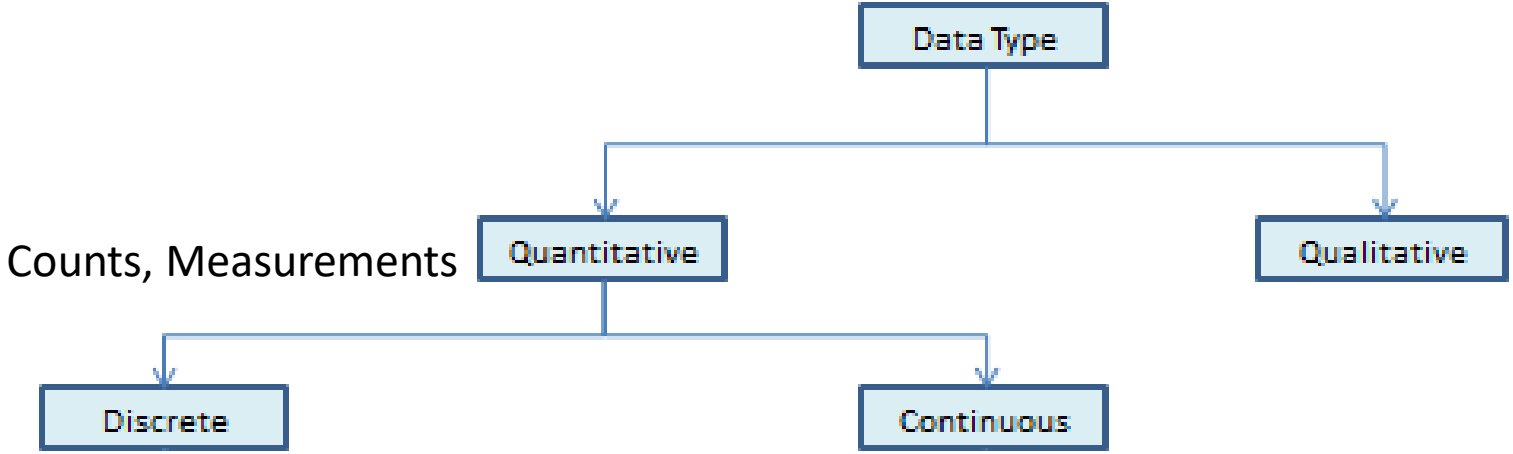
Letter grade
A+
A
A-
B+
B
B-
C+
C
C-
D+
D
E

Taxonomy of data



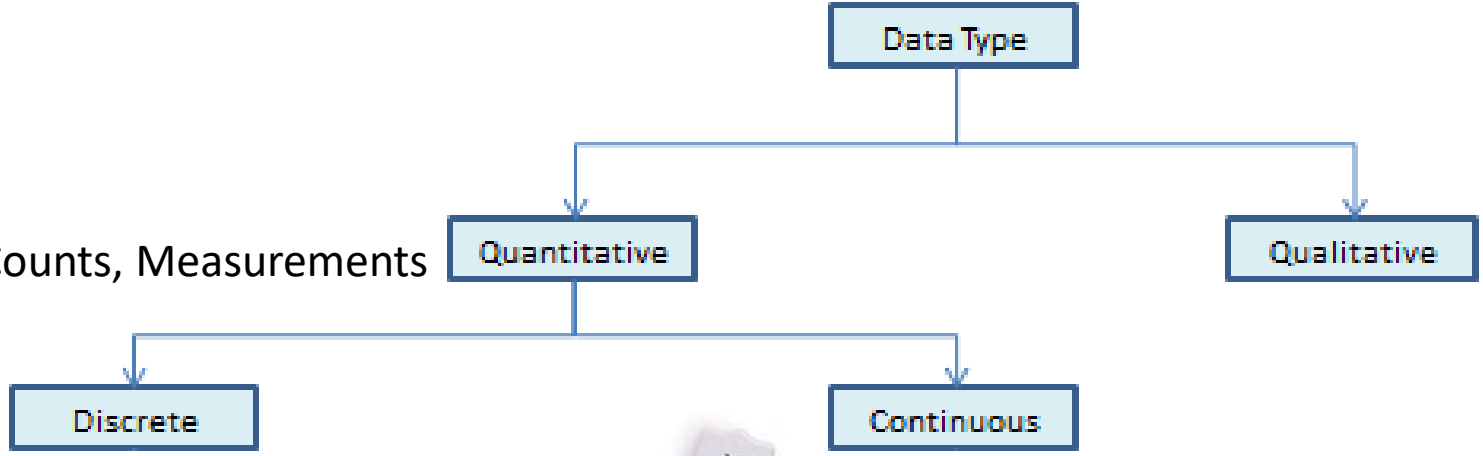


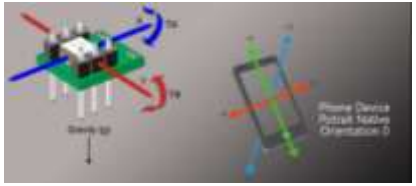
Taxonomy of data



- # of CPU cores
- # of courses taken in a semester
- # of times word ‘sale’ appears in a document

Taxonomy of data





Samples and Features

Feature / Attribute

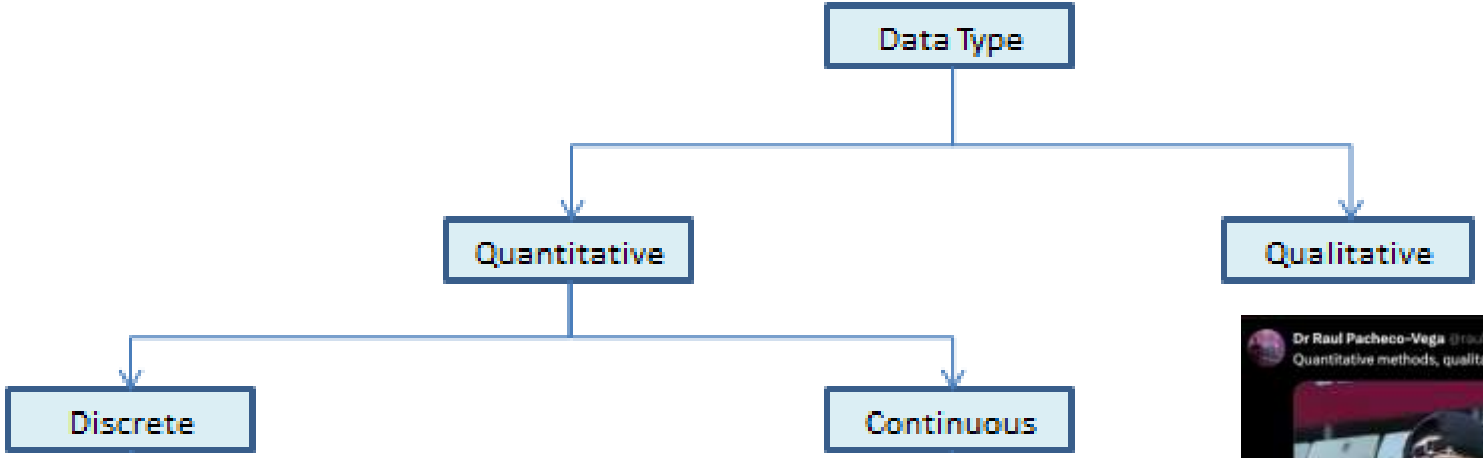
Dataset

Data Sample /Data Point

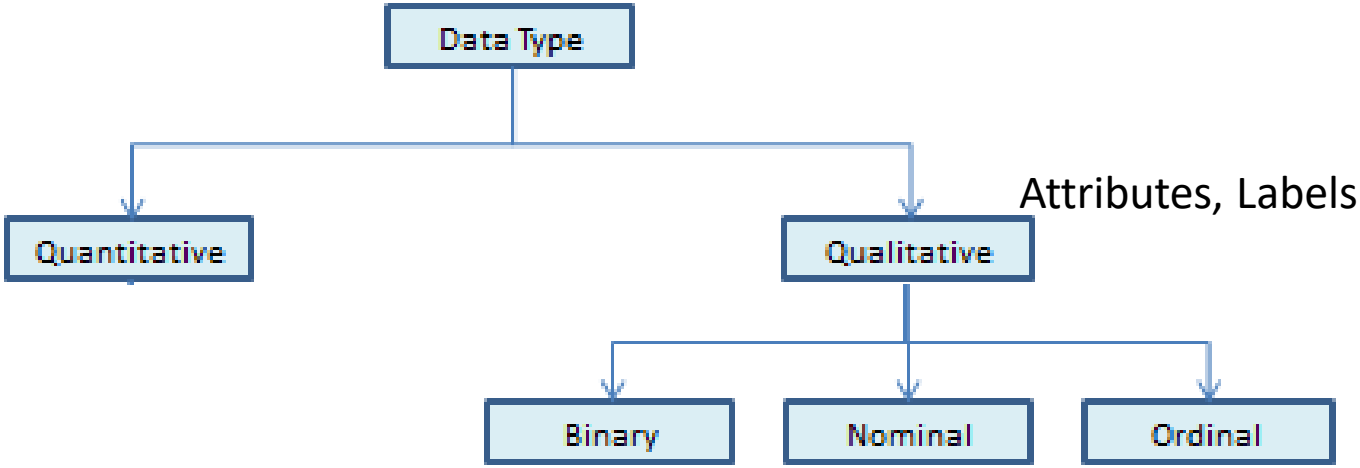
B	C	D	E	F	G	H	I
Quality	Usage	Dioxane [mol%]	Toluene [mol%]	Cyclohexane [mol%]	Temperature [°C]	{Instrument}	Timestamp
Good	train	18.238	59.40672	22.3555	22.1	RXN1	2019-11-14
Good	train	23.315	37.88732	38.7977	22.2	RXN1	2019-11-14
Good	train	16.405	56.02367	27.5714	22.0	RXN1	2019-11-14
Good	train	41.196	3.06438	55.7395	22.1	RXN1	2019-11-14
Bad	ignore		51.75047		22.2	LTT-R	2019-11-15
Good	test	13.476	67.81965	18.7039	22.5	LTT-R	2019-11-15
Good	test	16.802	13.56112	69.6365	21.9	LTT-R	2019-11-15

able1

Ultimately, all data needs to be quantitative



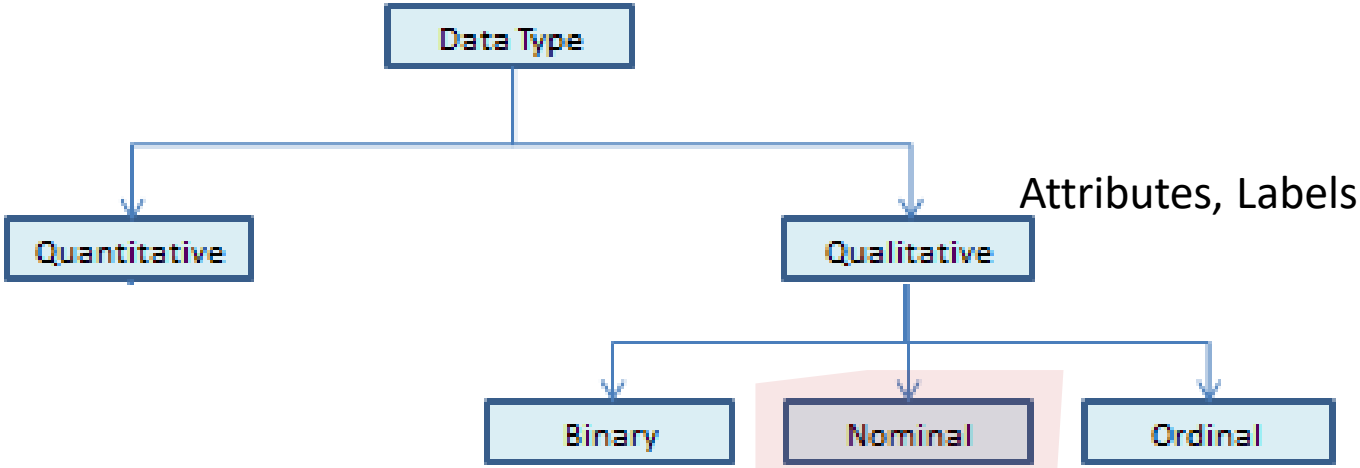
Taxonomy of data: Qualitative → Quantitative



0/1 ; -1/1

Preference

Taxonomy of data: Qualitative → Quantitative



Color

Make

Pin Code

Numerical encoding of categorical variables

Original data:	
id	Color
1	White
2	Red
3	Black
4	Purple
5	Gold

Numerical encoding of categorical variables

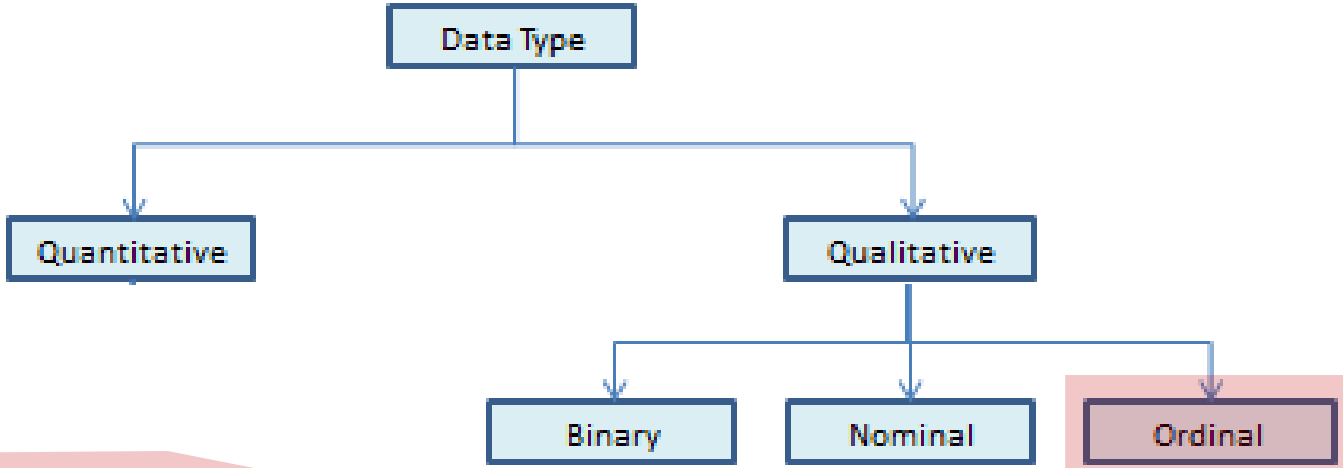
Original data:	
id	Color
1	White
2	Red
3	Black
4	Purple
5	Gold

One-hot encoding format:					
id	White	Red	Black	Purple	Gold
1	1	0	0	0	0
2	0	1	0	0	0
3	0	0	1	0	0
4	0	0	0	1	0
5	0	0	0	0	1

Numerical encoding of categorical variables

Rome Paris word V

```
Rome   = [1, 0, 0, 0, 0, 0, …, 0]
Paris  = [0, 1, 0, 0, 0, 0, …, 0]
Italy  = [0, 0, 1, 0, 0, 0, …, 0]
France = [0, 0, 0, 1, 0, 0, …, 0]
```



How comfortable are you with Python *

-2 +1

No knowledge ○ ○ ○ ○ ○ ○ Very comfortable

1 2 3 4 5

Encoding Choices

- One-hot
- Numerical
  - Equal
  - Unequal/Weighted
- Binary
- Target*

Q

- How should a binary attribute be encoded ?

Example: PlayTennis dataset

Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	Normal	False	Yes
...	...	...	...	...

Outlook	Temperature	Humidity	Windy	Play
Sunny	85	85	False	No
Sunny	80	90	True	No
Overcast	83	86	False	Yes
Rainy	75	80	False	Yes
...	...	...	...	...

Example: Contact Lenses dataset

☛ No patient id

☛ Age is not a number !

Age	Spectacle prescription	Astigmatism	Tear production rate	Recommended lenses
Young	Myope	No	Reduced	None
Young	Myope	No	Normal	Soft
Young	Myope	Yes	Reduced	None
Young	Myope	Yes	Normal	Hard
Young	Hypermetrope	No	Reduced	None
Young	Hypermetrope	No	Normal	Soft
Young	Hypermetrope	Yes	Reduced	None
Young	Hypermetrope	Yes	Normal	hard
Pre-presbyopic	Myope	No	Reduced	None
Pre-presbyopic	Myope	No	Normal	Soft
Pre-presbyopic	Myope	Yes	Reduced	None
Pre-presbyopic	Myope	Yes	Normal	Hard
Pre-presbyopic	Hypermetrope	No	Reduced	None
Pre-presbyopic	Hypermetrope	No	Normal	Soft
Pre-presbyopic	Hypermetrope	Yes	Reduced	None
Pre-presbyopic	Hypermetrope	Yes	Normal	None
Presbyopic	Myope	No	Reduced	None
Presbyopic	Myope	No	Normal	None
Presbyopic	Myope	Yes	Reduced	None
Presbyopic	Myope	Yes	Normal	Hard
Presbyopic	Hypermetrope	No	Reduced	None
Presbyopic	Hypermetrope	No	Normal	Soft
Presbyopic	Hypermetrope	Yes	Reduced	None
Presbyopic	Hypermetrope	Yes	Normal	None

Sometimes data can be missing

Outlook	Temperature	Humidity	Windy	Play
Sunny	85	85	False	No
Sunny	80		True	No
Overcast	83	86	False	Yes
Rainy	75	80	False	Yes
...	...	...	...	...

→ Unknown or unrecorded

… or incorrect

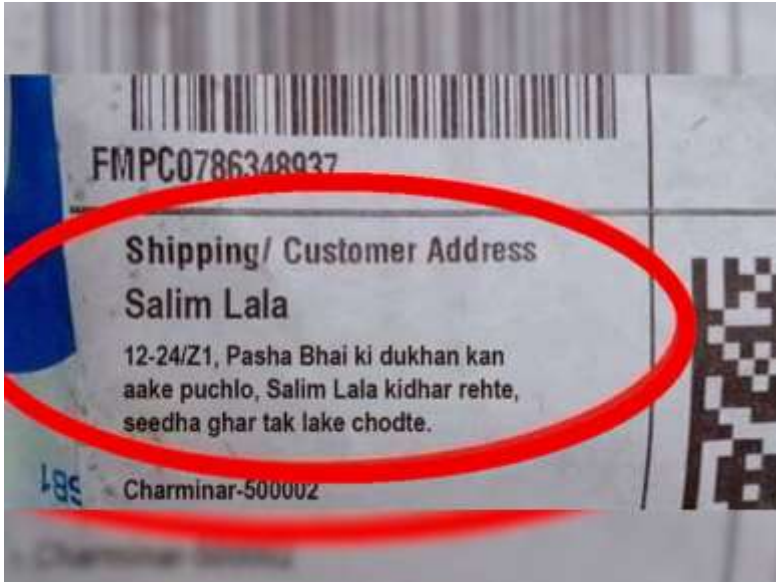
	DBAName	AKAName	Address	City	State	Zip
t1	John Veliotis Sr.	Johnnyo's	3465 S Morgan ST	***Chicago***	IL	***60608***
t2	John Veliotis Sr.	Johnnyo's	3465 S Morgan ST	Chicago	IL	***60609***
t3	John Veliotis Sr.	Johnnyo's	3465 S Morgan ST	Chicago	IL	***60609***
t4	***Johnnyo's***	Johnnyo's	3465 S Morgan ST	***Cicago***	IL	60608

Conflicts

Does not obey data distribution

Conflict

… deal with reality out there !



… deal with reality out there !

Data imputation

- Approaches that aim to estimate missing data
- Options
  - Remove sample
  - Fill with 0
  - Fill with constant
  - Fill with a statistical measure (mean, median, mode)
  - Do nothing. Use a learning method which can handle missing data.

Lecture Outline

- *ML Workflow*
- Data sample Representations
- Basic Data Transformations
- Data Visualization

Samples, Features, Labels

Label

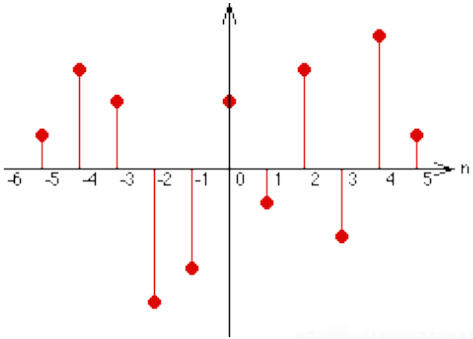
Feature / Attribute

Sample

B	C	D	E	F	G	H	I
Quality	Usage	Dioxane [mol%]	Toluene [mol%]	Cyclohexane [mol%]	Temperature [°C]	{Instrument}	Timestamp
Good	train	18.238	59.40672	22.3555	22.1	RXN1	2019-11-14
Good	train	23.315	37.88732	38.7977	22.2	RXN1	2019-11-14
Good	train	16.405	56.02367	27.5714	22.0	RXN1	2019-11-14
Good	train	41.196	3.06438	55.7395	22.1	RXN1	2019-11-14
Bad	ignore		51.75047		22.2	LTT-R	2019-11-15
Good	test	13.476	67.81965	18.7039	22.5	LTT-R	2019-11-15
Good	test	16.802	13.56112	69.6365	21.9	LTT-R	2019-11-15

able1

Data Sample Representations



Scalars

$$x$$

Vectors

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix}$$

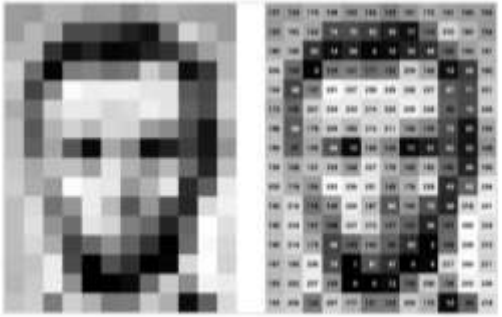
Matrix

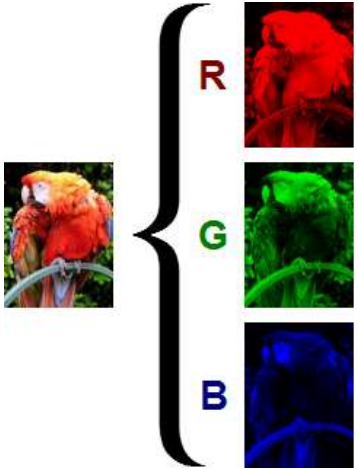
$$\mathbf{X} = \begin{bmatrix} \mathbf{x} & \cdots & \mathbf{x}_N \end{bmatrix} = \begin{bmatrix} x_{1,1} & & x_{N,1} \\ \vdots & \cdots & \vdots \\ x_{1,M} & & x_{N,M} \end{bmatrix}$$

Tensor

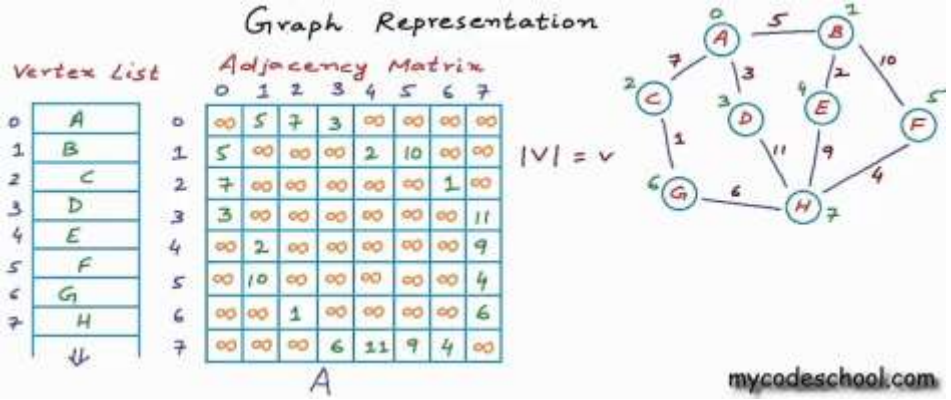
$$\mathbf{X} = \{\mathbf{X}_1, \ldots \mathbf{X}_K\} = \begin{bmatrix} x_{1,1,1} & \cdots & x_{N,1,1} \\ \vdots & & \vdots \\ x_{1,M,1} & & x_{N,M,1} \end{bmatrix} \begin{bmatrix} x_{1,1,K} & & x_{N,1,K} \\ \vdots & \cdots & \vdots \\ x_{1,M,K} & & x_{N,M,K} \end{bmatrix}$$

2-d image





Data Representations



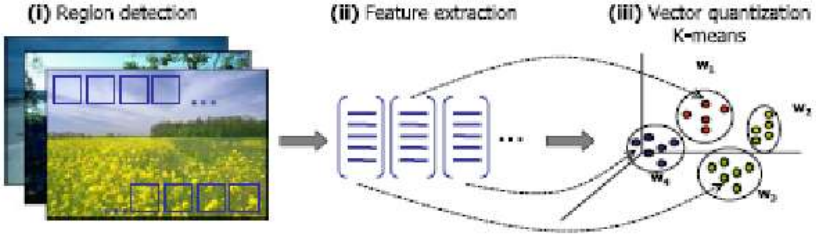
Feature Extraction (FE)

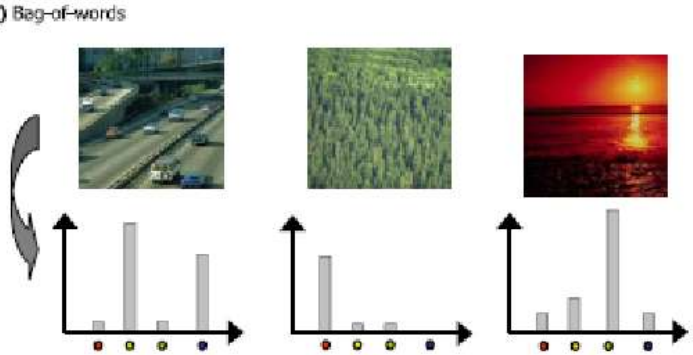
- Def: Feature Extraction (FE) is any algorithm that transformation raw data into features that can be used as an input for a learning algorithm.

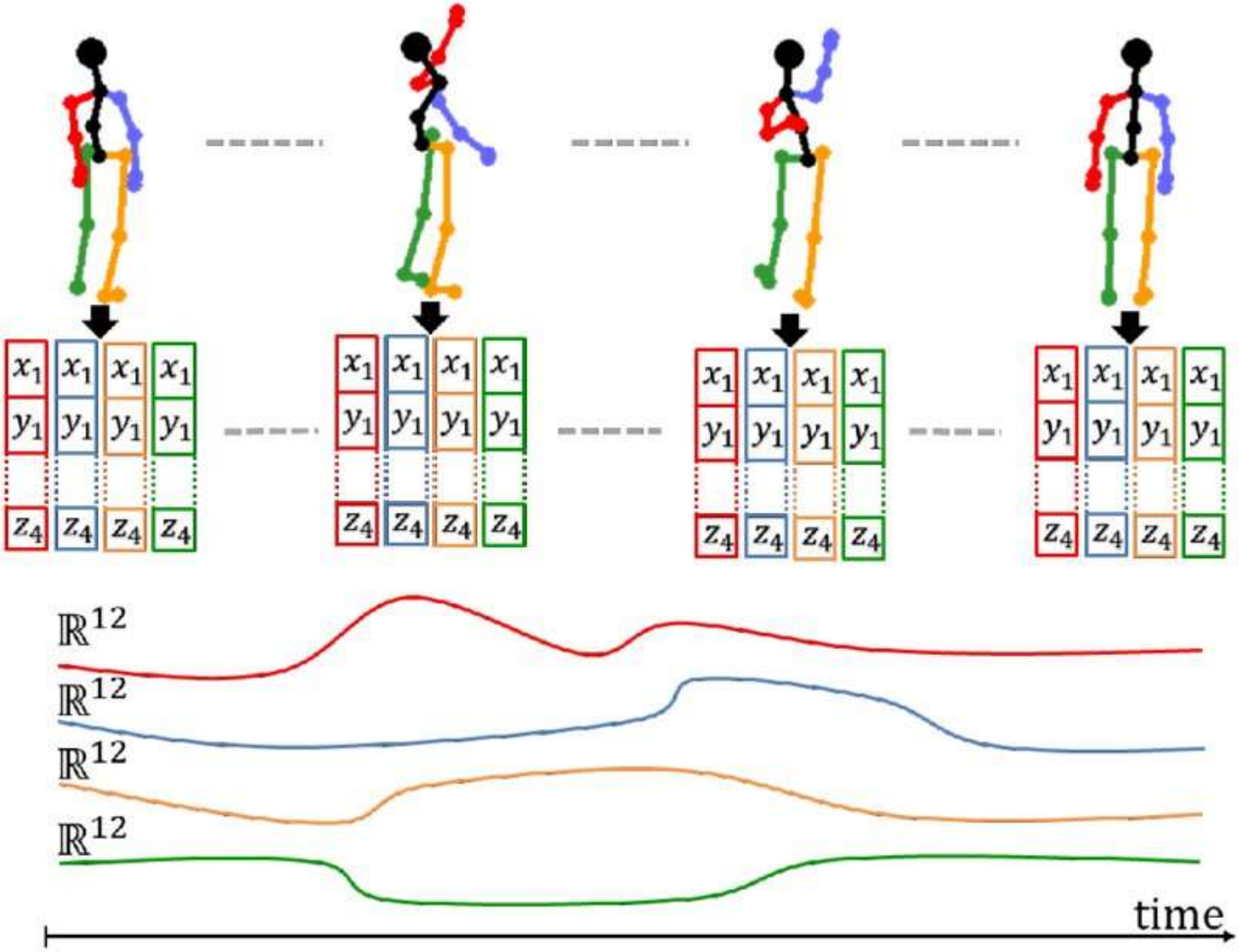
The Bag of Words Representation

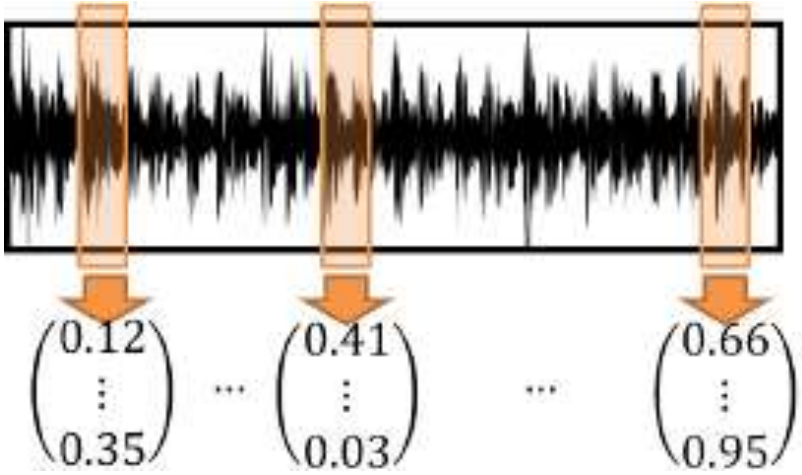
I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!

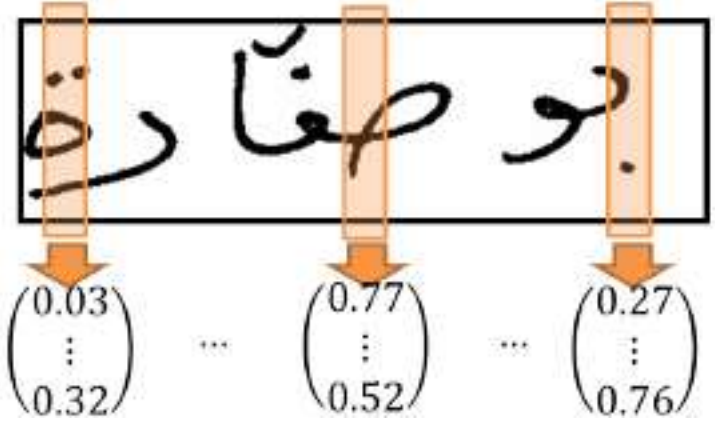
it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
…	…



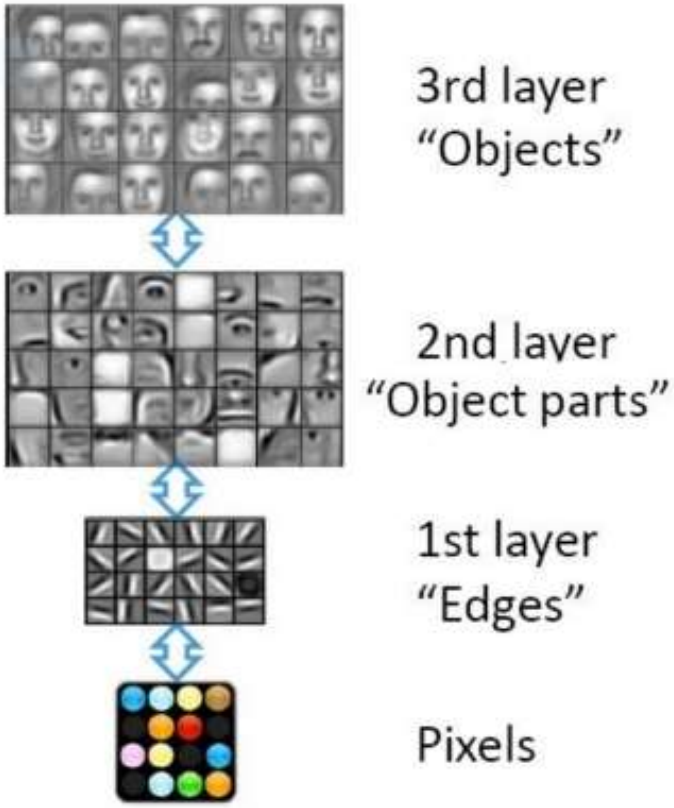


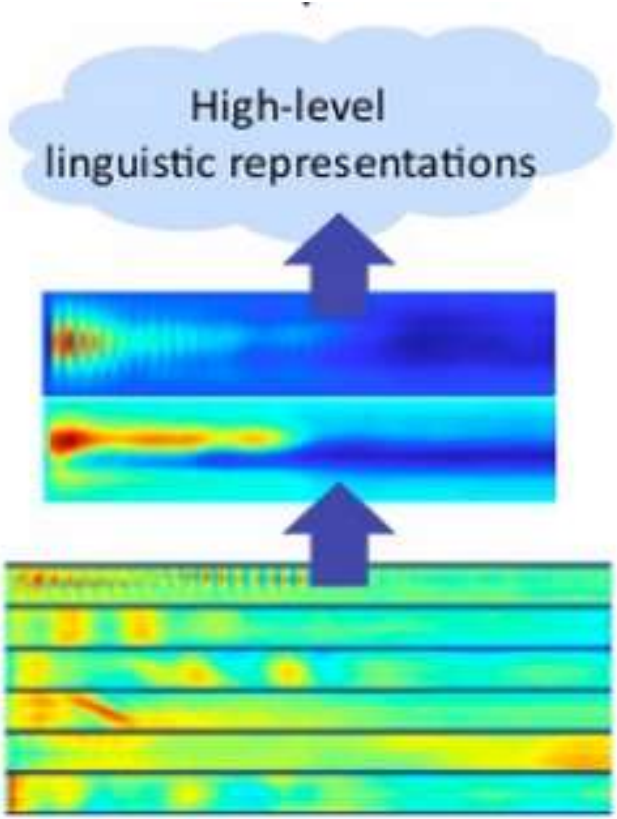

x_1 y_1 z_4
$\mathbb{R}^{12}$
$\mathbb{R}^{12}$
$\mathbb{R}^{12}$
$\mathbb{R}^{12}$
time


0.12
0.35
0.41
0.03
0.66
0.95


0.03
0.32
0.77
0.52
0.27
0.76

Feature-based, Hierarchical Data Representations





Data – a probability-based perspective

- The basis for Statistical Learning Theory

Then we observe candies drawn from some bag: 

  - Domain described by random variables (r.v.)
    - X = {apple, grape}
    - $b_i \in [1,5]$

  - Data = Instantiation of some or all r.v.'s in the domain

Data: a probabilistic perspective

	DBAName	AKAName	Address	City	State	Zip
t1	John Veliotis Sr.	Johnnyo's	3465 S Morgan ST	***Chicago***	IL	***60608***
t2	John Veliotis Sr.	Johnnyo's	3465 S Morgan ST	Chicago	IL	***60609***
t3	John Veliotis Sr.	Johnnyo's	3465 S Morgan ST	Chicago	IL	***60609***
t4	***Johnnyo's***	Johnnyo's	3465 S Morgan ST	***Cicago***	IL	60608

Conflicts

Does not obey data distribution

Conflict

Output

Proposed Cleaned Dataset

	DBAName	Address	City	State	Zip
t1	John Veliotis Sr.	3465 S Morgan ST	Chicago	IL	60608
t2	John Veliotis Sr.	3465 S Morgan ST	Chicago	IL	**60608**
t3	John Veliotis Sr.	3465 S Morgan ST	Chicago	IL	**60608**
t4	**John Veliotis Sr.**	3465 S Morgan ST	**Chicago**	IL	60608

Marginal Distribution of Cell Assignments

Cell	Possible Values	Probability
t2.Zip	60608	0.84
	60609	0.16
t4.City	Chicago	0.95
	Cicago	0.05
t4.DBAName	John Veliotis Sr.	0.99
	Johnnyo's	0.01

Other important aspects of data

- Mode of collection
  - Passive (‘sense’)
  - Active (‘explore, sense, repeat’)

- Statistical assumptions on data
  - i.i.d (independent and identically distributed)
  - Online (e.g. time-series data)