07.01.2020

Statistical Methods in AI (CSE/ECE 471)

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

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

Vineet Gandhi (v.gandhi@iiit.ac.in)

Center for Visual Information Technology (CVIT)

IIIT Hyderabad

Announcements

- IMPORTANT: All assignments/projects will need to be in Python.
- Tutorial on Python, Pandas, Jupyter notebook, Plotting tools. **Bring your laptops**.
- Ask questions.

Lecture Outline

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

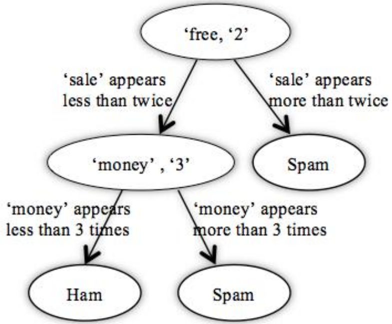
Machine Learning

Algorithmic methods that use data to improve their knowledge of a task

Task: Detect spam email

Data: Labelled emails (in inboxes of other users as well !)

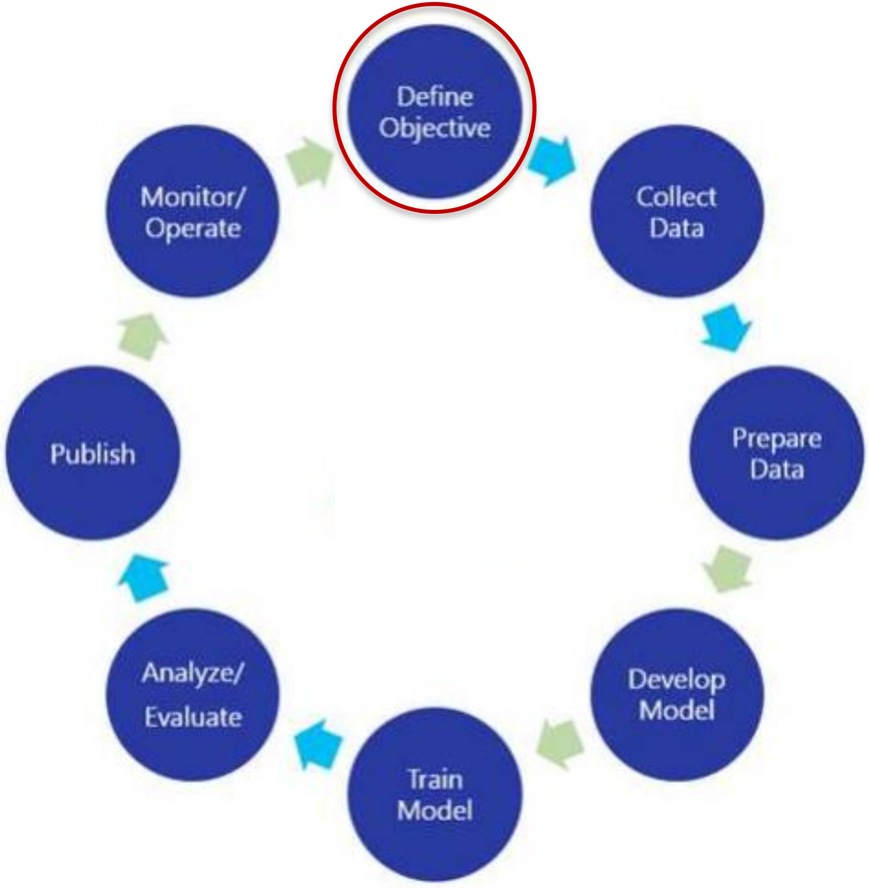
Knowledge:



Improve → 85% reduction of spam emails in Inbox over 3 months

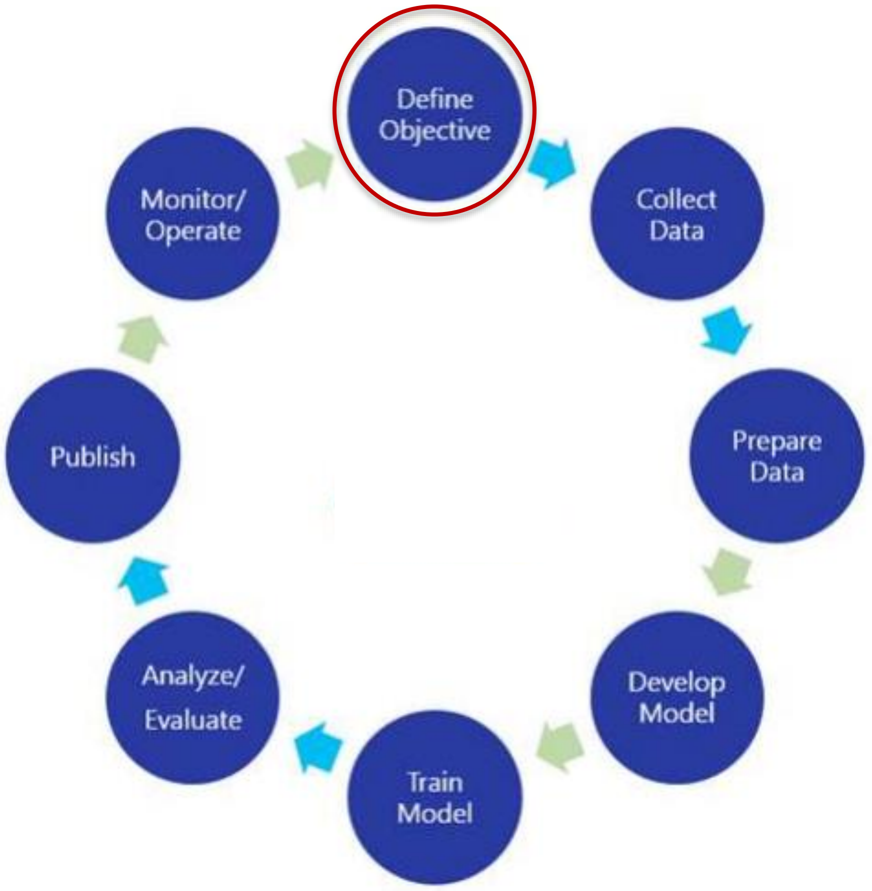
Algorithmic method: Decision Tree

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



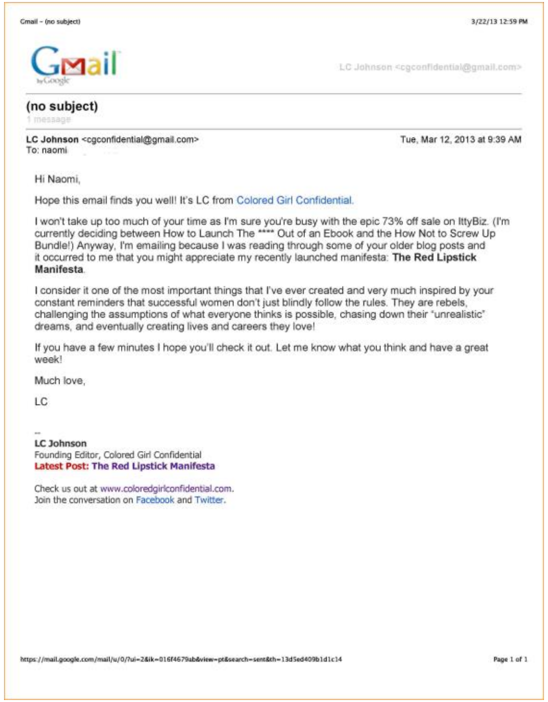
No Data, no ML !


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


Gmail - (no subject) 3/22/13 12:59 PM

Gmail by Google LC Johnson <cgconfidential@gmail.com>

(no subject)
1 message

LC Johnson <cgconfidential@gmail.com> Tue, Mar 12, 2013 at 9:39 AM
To: naomi

Hi Naomi,

Hope this email finds you well! It's LC from Colored Girl Confidential.

I won't take up too much of your time as I'm sure you're busy with the epic 73% off sale on IttyBiz. (I'm currently deciding between How to Launch The **** Out of an Ebook and the How Not to Screw Up Bundle!) Anyway, I'm emailing because I was reading through some of your older blog posts and it occurred to me that you might appreciate my recently launched manifesta: **The Red Lipstick Manifesta**.

I consider it one of the most important things that I've ever created and very much inspired by your constant reminders that successful women don't just blindly follow the rules. They are rebels, challenging the assumptions of what everyone thinks is possible, chasing down their "unrealistic" dreams, and eventually creating lives and careers they love!

If you have a few minutes I hope you'll check it out. Let me know what you think and have a great week!

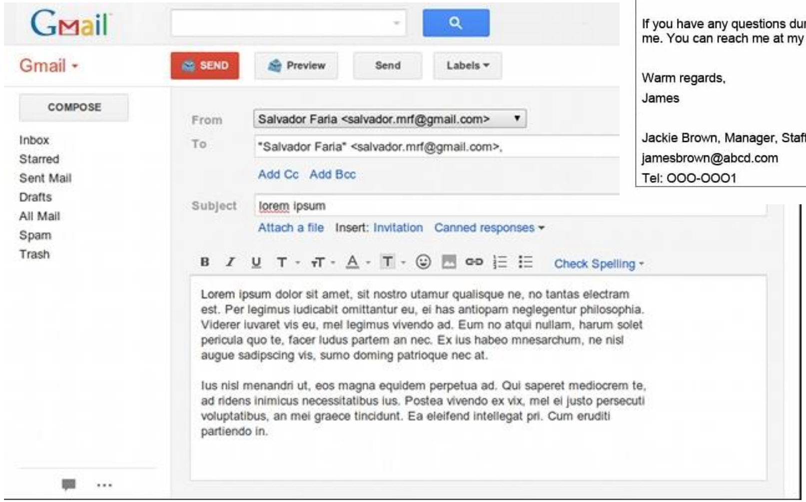
Much love,

LC

--
LC Johnson
Founding Editor, Colored Girl Confidential
Latest Post: The Red Lipstick Manifesta

Check us out at www.coloredgirlconfidential.com.
Join the conversation on Facebook and Twitter.

https://mail.google.com/mail/u/0/?ui=2&ik=016f4679ab&view=pt&search=sent&th=13d5ed409b1d1c14 Page 1 of 1




Business Email Sample

To: "Anna Jones" <annajones@buzzle.com.>
CC: All Staff
From: "James Brown"
Subject: Welcome to our Hive!

Dear Anna,

Welcome to our Hive!
It is a pleasure to welcome you to the team of________________. We are excited to have you join our team, and we hope that you will enjoy Working with our Company.

On the last Saturday of each month we hold a special staff party to welcome any new employees. Please be sure to come next Week to meet all of our senior staff and any other new staff members who have joined________________ this month. You will receive an e-mail regarding the same with further details.

If you have any questions during your training period, please do not hesitate to contact me. You can reach me at my email address or on my office line at OOO-OOO1.

Warm regards,
James

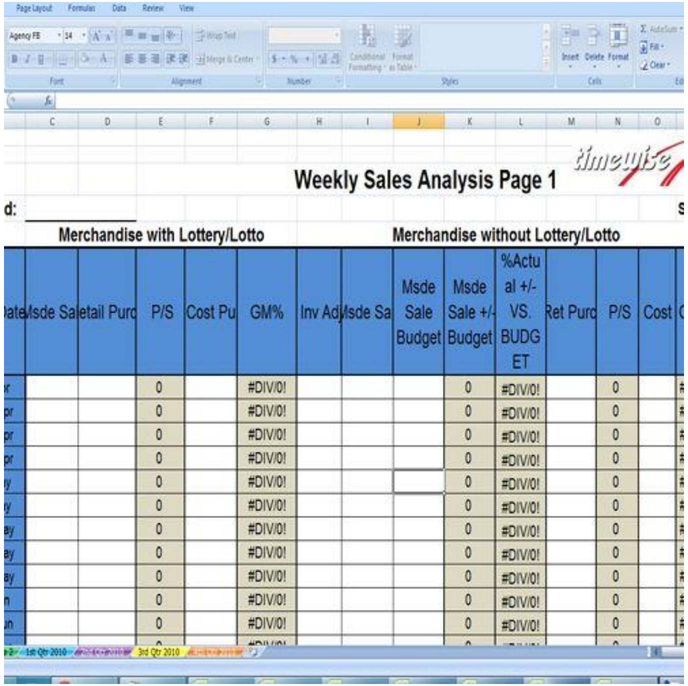
Jackie Brown, Manager, Staff
jamesbrown@abcd.com
Tel: OOO-OOO1

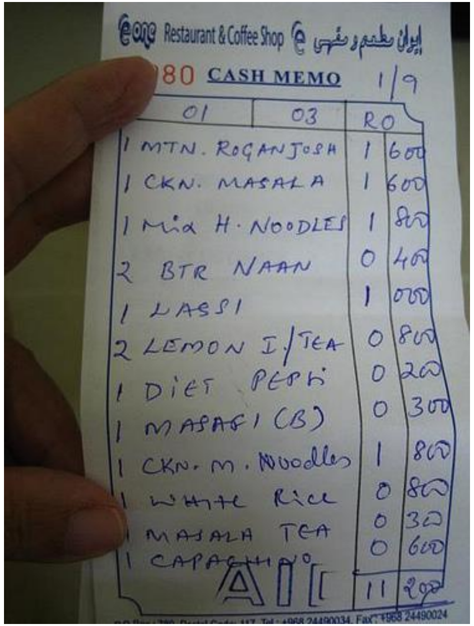
Sources of data

- Predict value of a stock

Sources of data

- Predict effect of advertising on sales

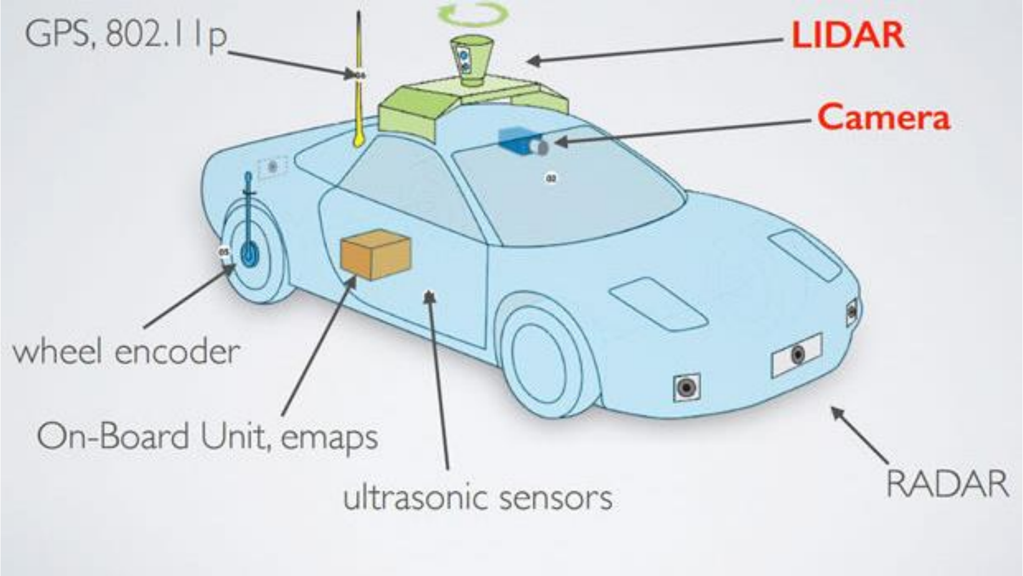




Raw Data may not always be digital in nature !

Sources of data

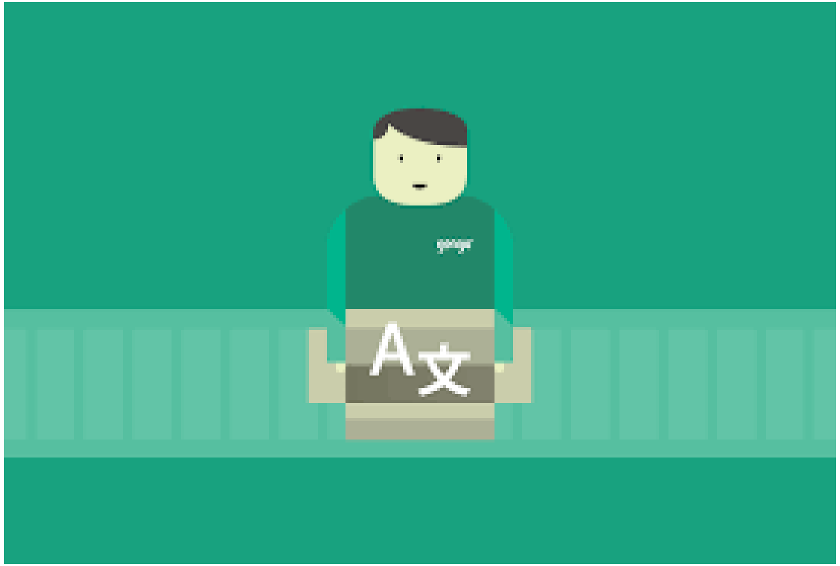
- Drive car safely without human intervention



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

Two fundamental questions

- What data to collect ?
- How 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

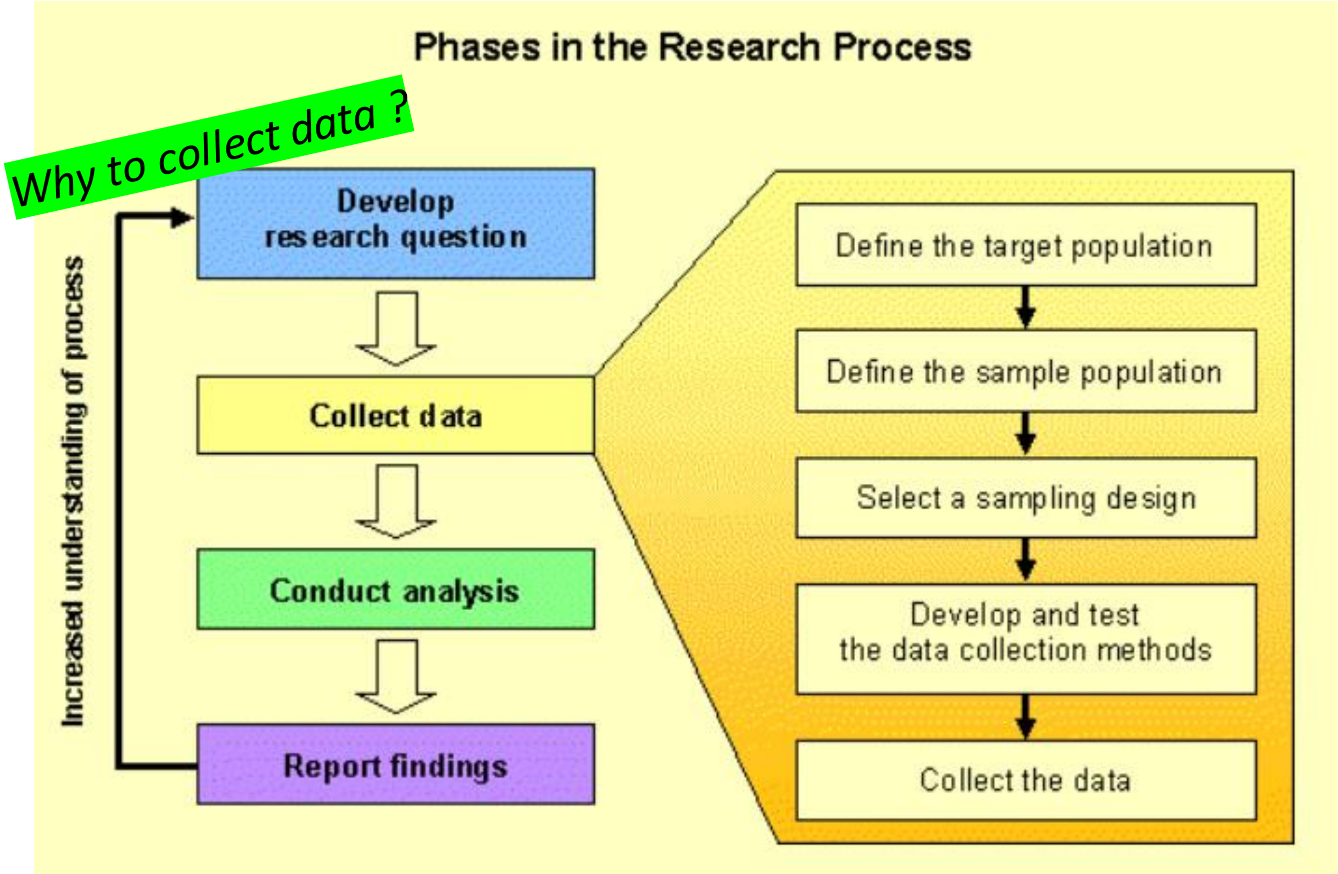
- Often not directly usable
  - Filter (needed data)
  - **Transform (to numerical data)**


```
{
    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"
}
```

Two fundamental questions

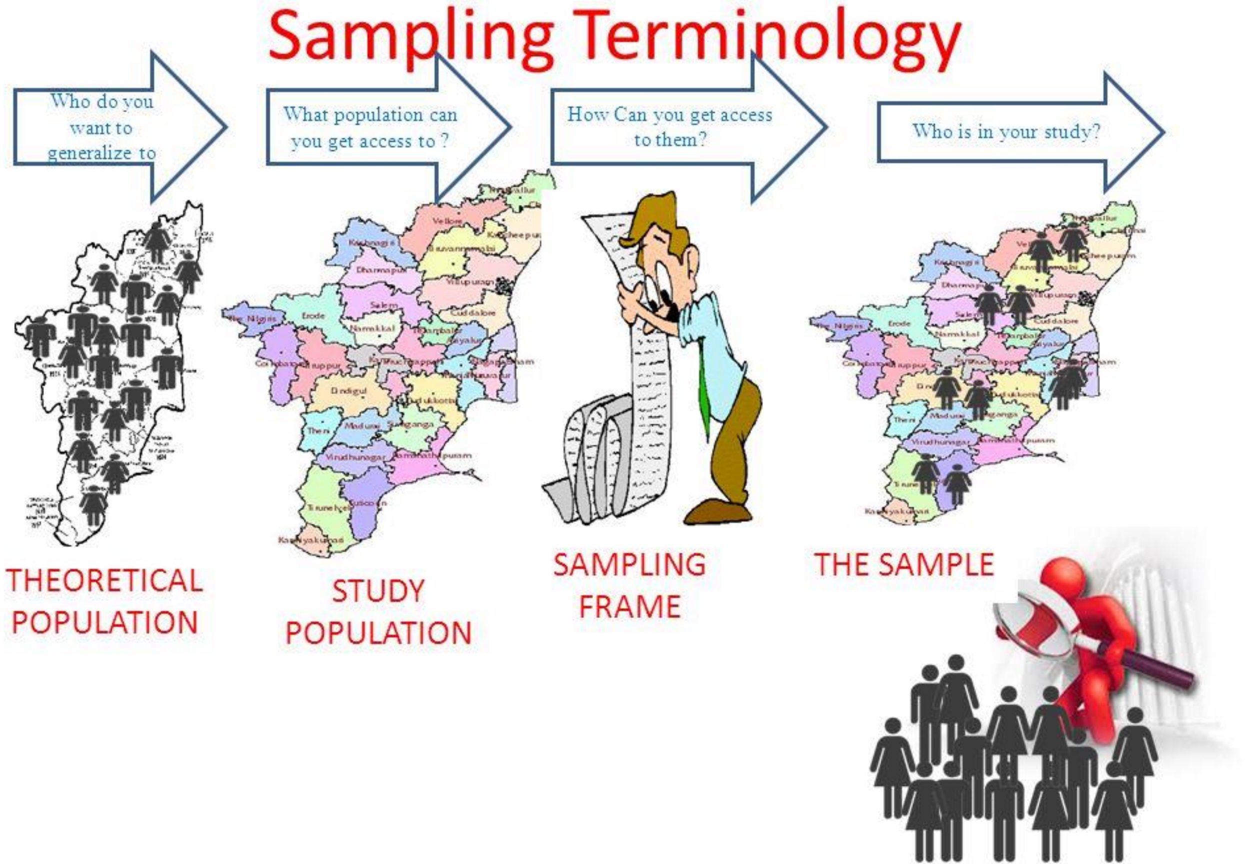
- What data to collect ?
- How (much) to collect ?

The Research Method

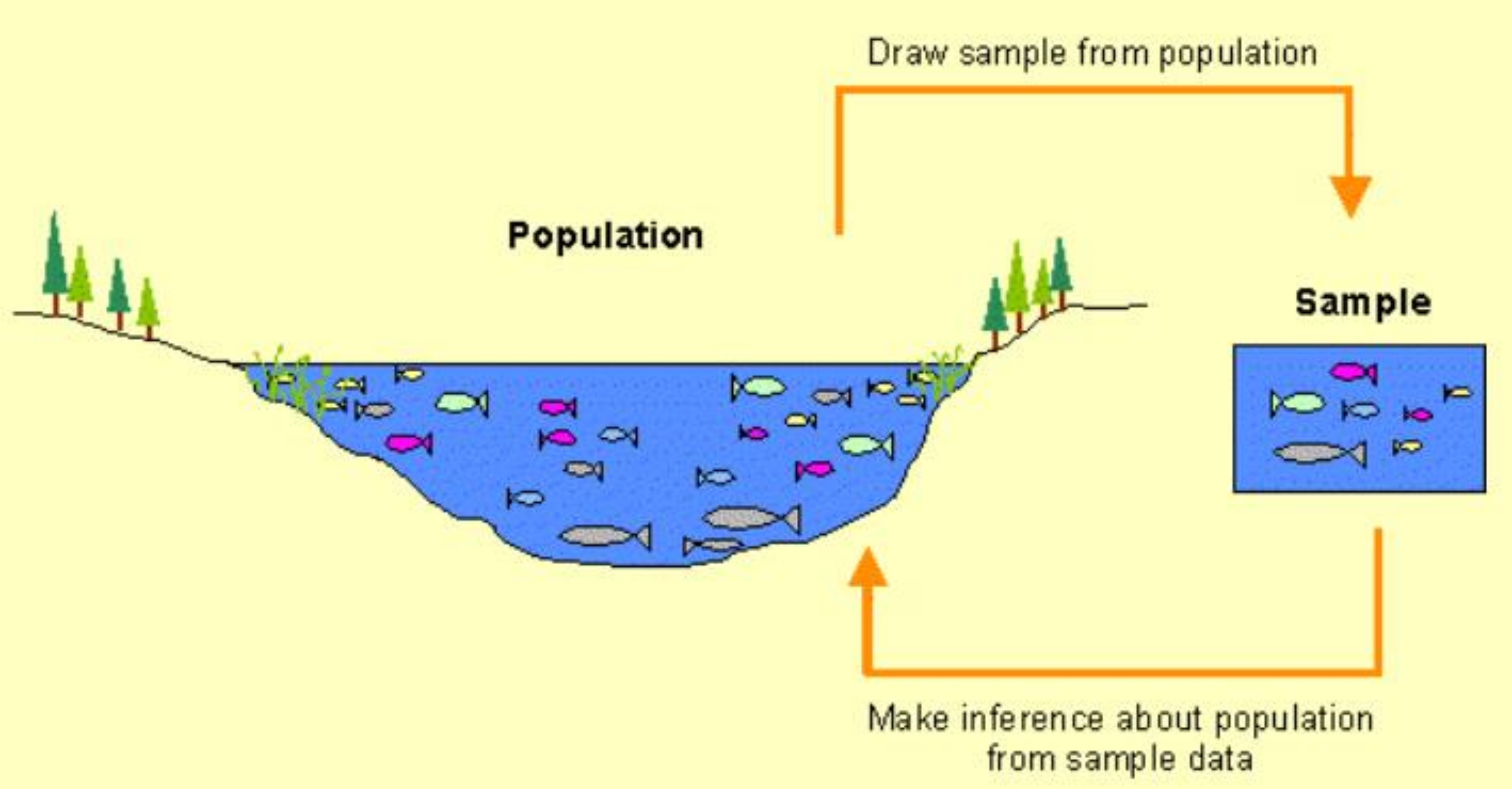


http://labs.geog.uvic.ca/geog226/frLab3.html

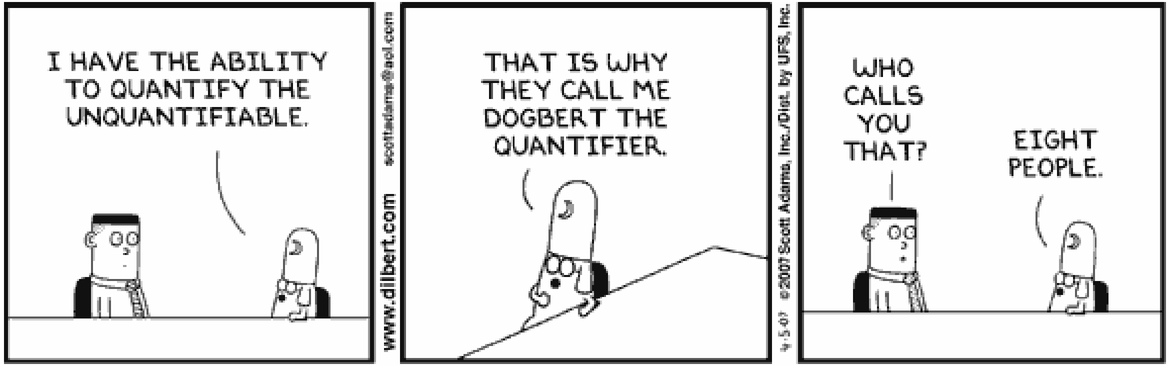
The Research Method



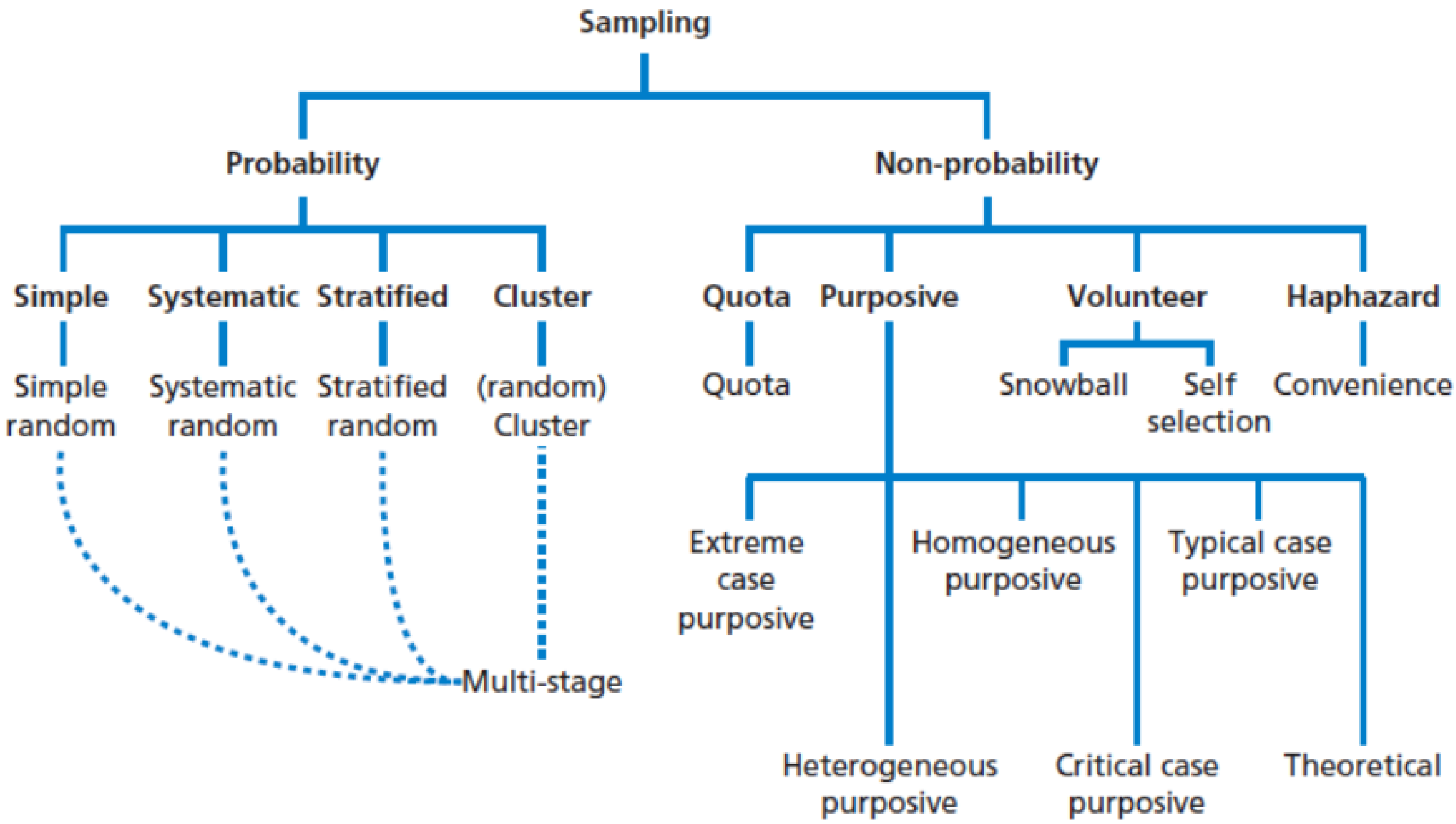
The Research Method



Are our samples ‘nice’ ?

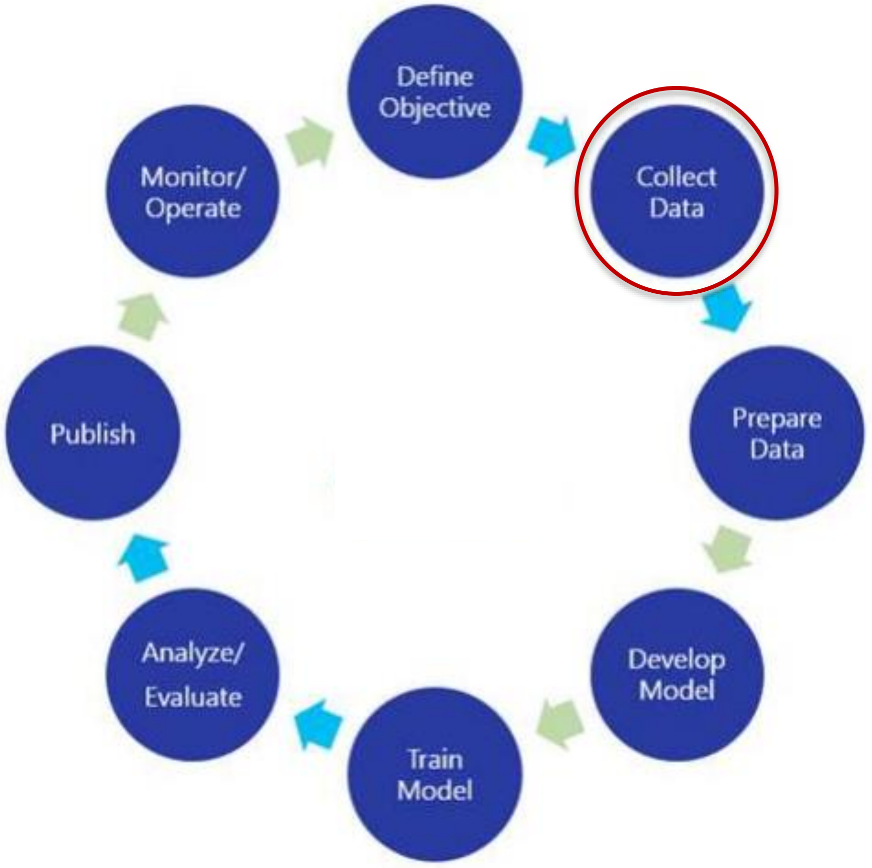


Sampling Techniques



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

Workflow of a Machine Learning Problem



Taxonomy of data variables

Data Type

Counts, Measurements

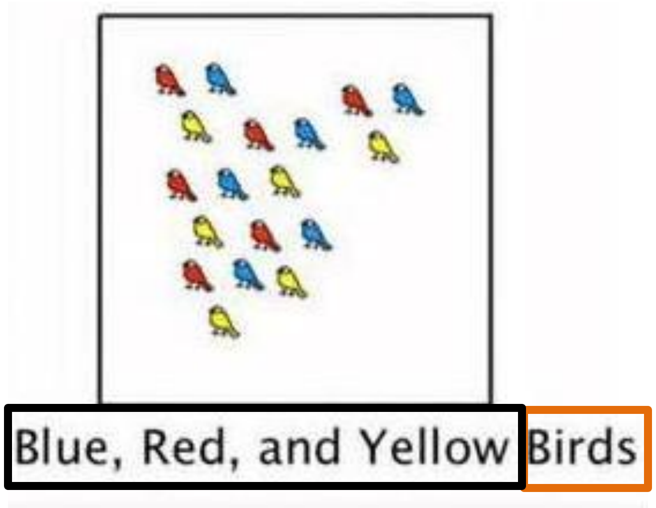
Quantitative

Attributes, Labels

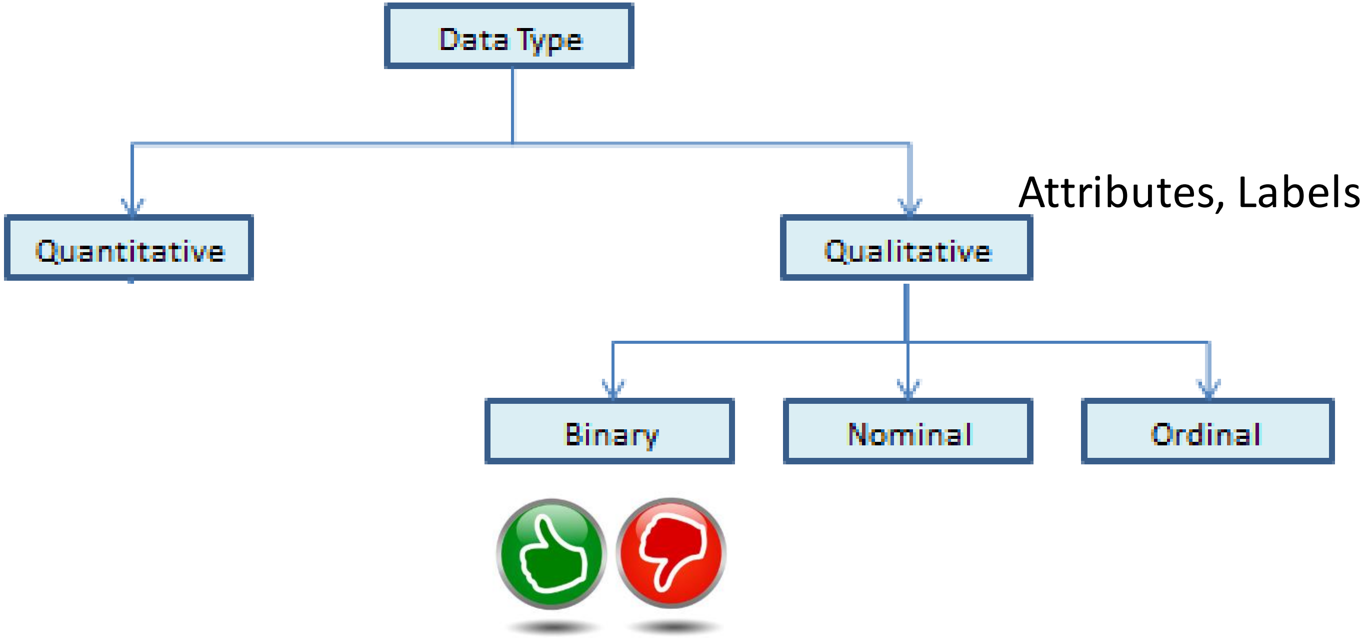
Qualitative

Data → VALUE

Information → VARIABLE



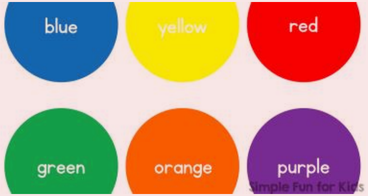
Taxonomy of data



Preference

Taxonomy of data

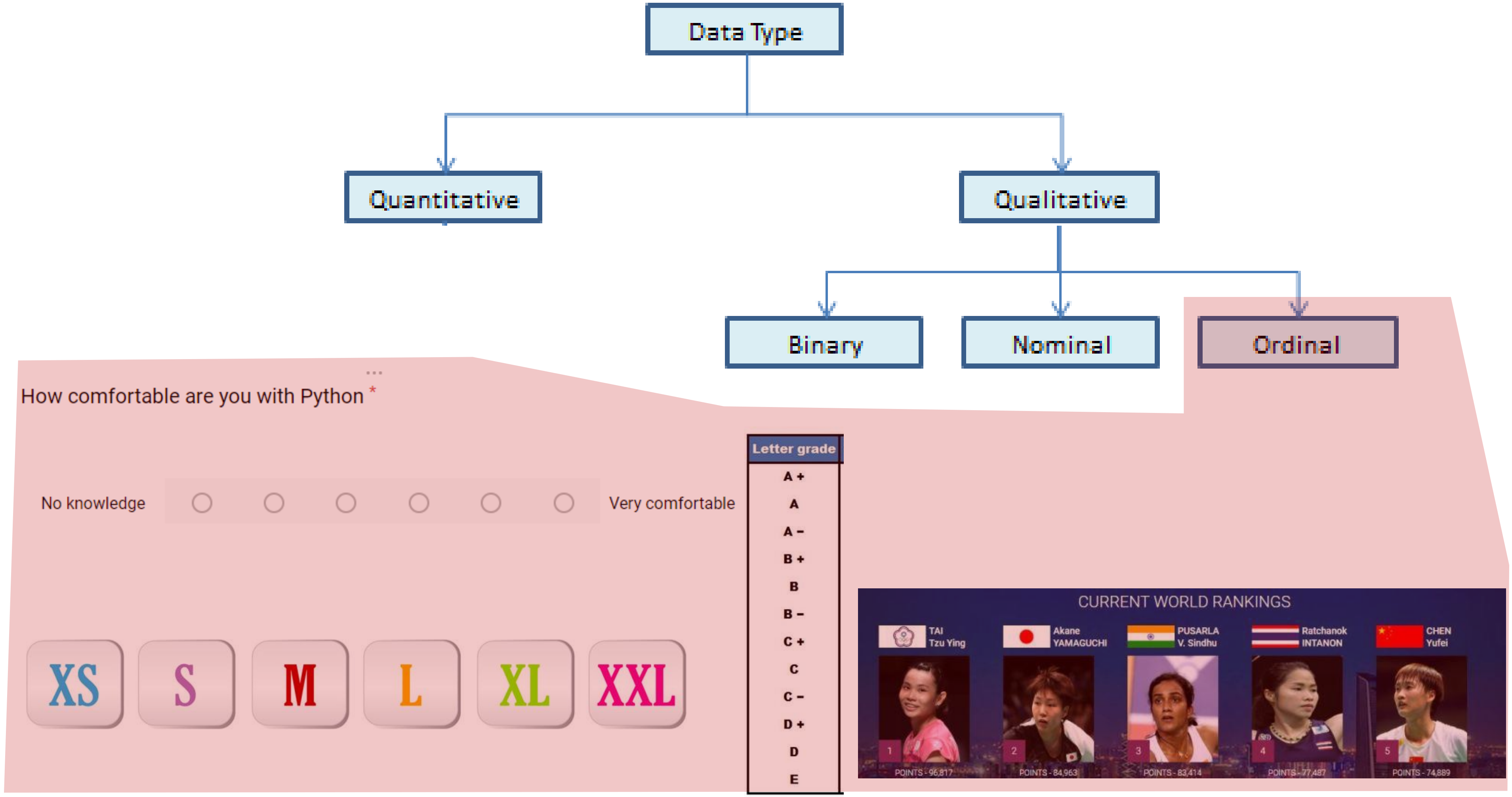
- Data Type
  - Quantitative
  - Qualitative — Attributes, Labels
    - Binary
    - Nominal
    - Ordinal



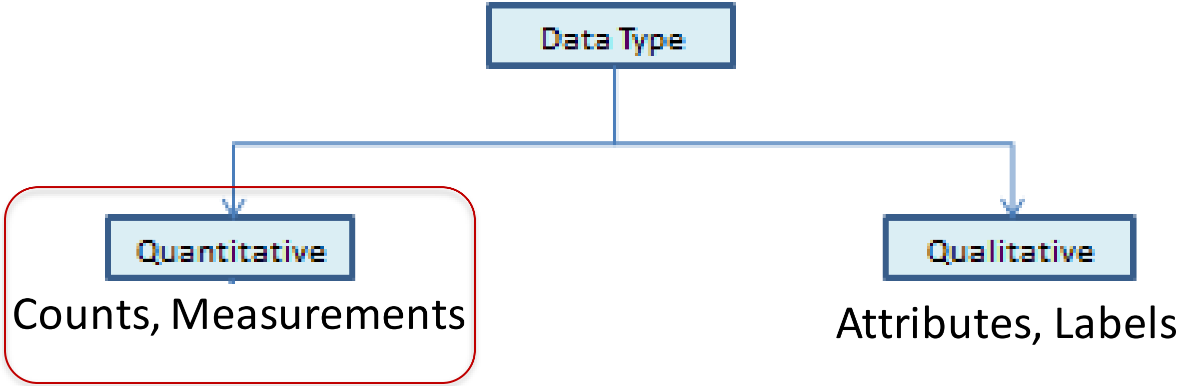
Color

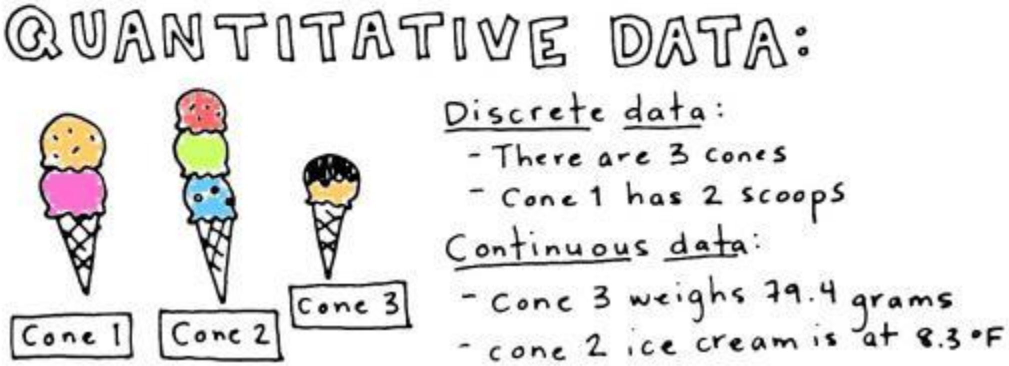
Make

Pin Code

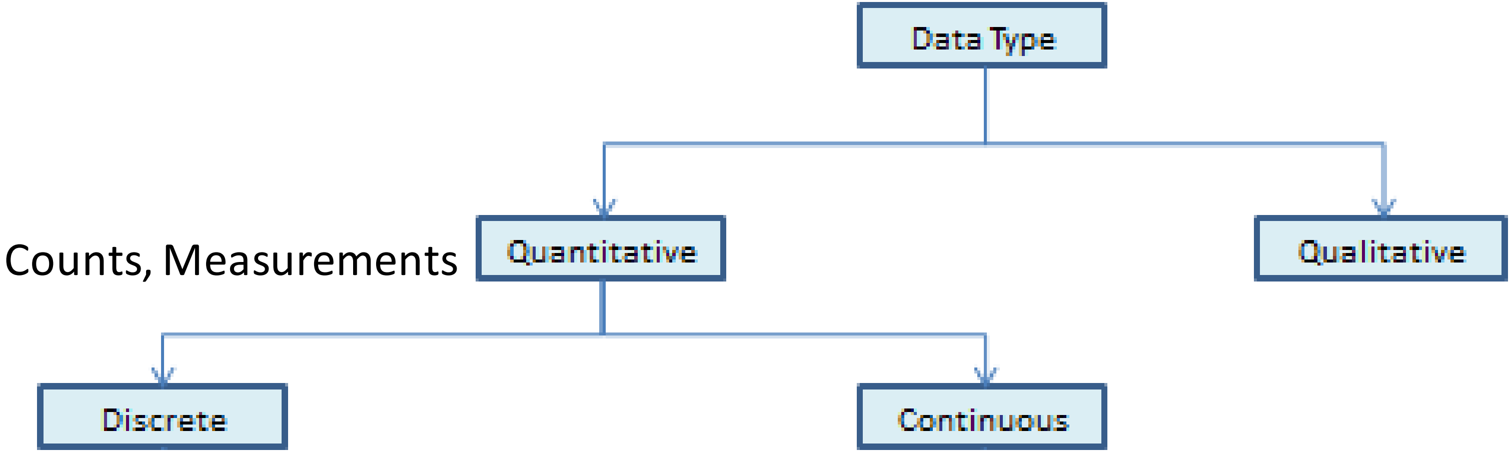

Data Type
Quantitative
Qualitative
Binary
Nominal
Ordinal
How comfortable are you with Python *
No knowledge
Very comfortable
XS
S
M
L
XL
XXL
Letter grade
A +
A
A –
B +
B
B –
C +
C
C –
D +
D
E
CURRENT WORLD RANKINGS
TAI Tzu Ying
Akane YAMAGUCHI
PUSARLA V. Sindhu
Ratchanok INTANON
CHEN Yufei
1
2
3
4
5
POINTS - 96,817
POINTS - 84,963
POINTS - 83,414
POINTS - 77,487
POINTS - 74,889

Taxonomy of data





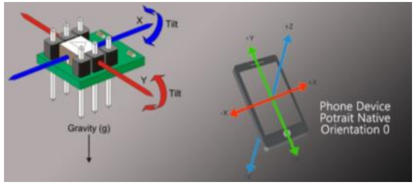
Taxonomy of data



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

Taxonomy of data

- Data Type
  - Quantitative — Counts, Measurements
    - Discrete
    - Continuous
  - Qualitative



Samples and Features

Feature / Attribute

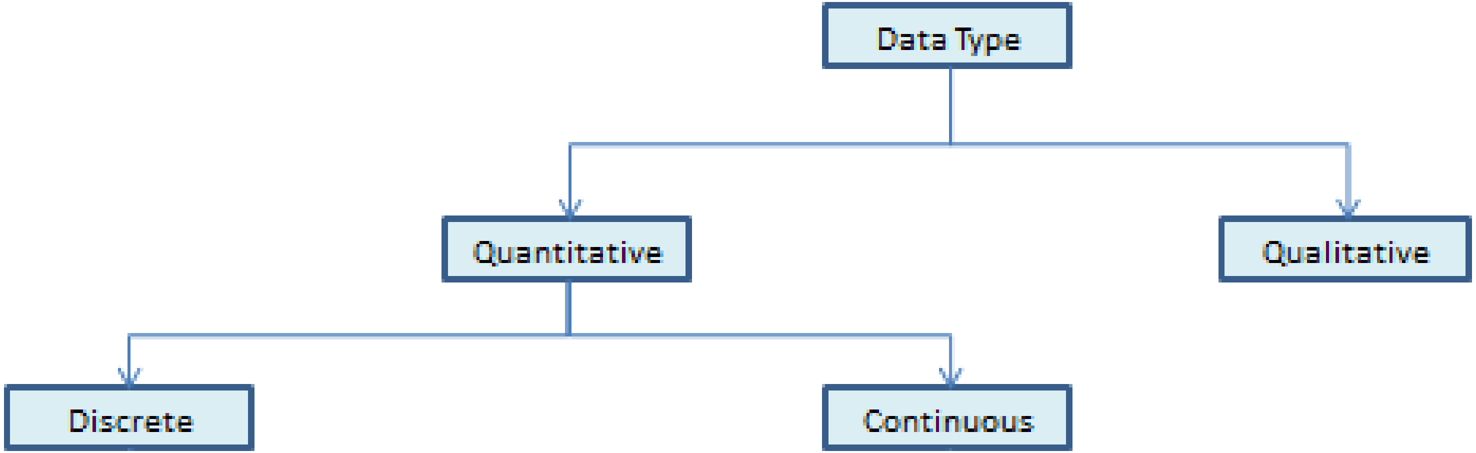
Dataset

Sample

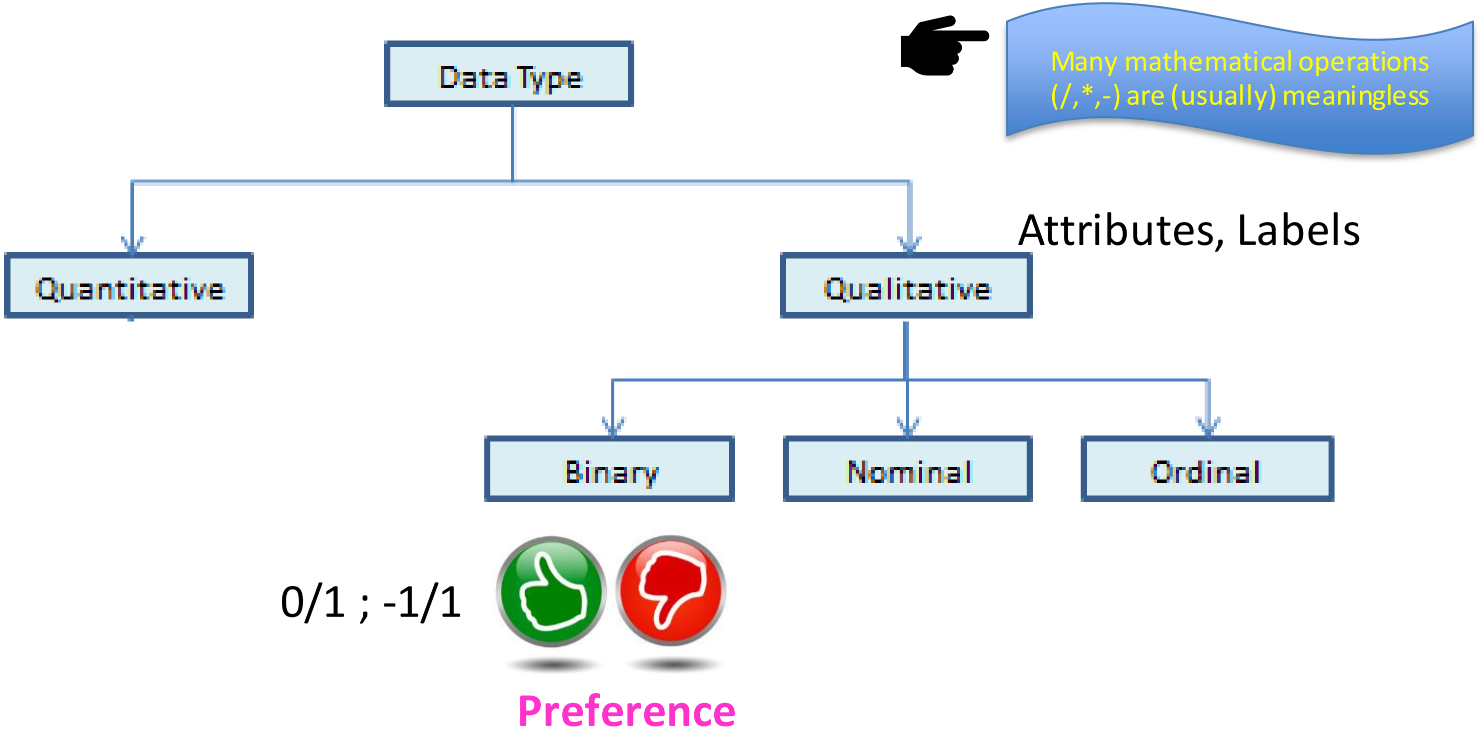
B	C	D	E	F	G	H	I
Quality	Usage	Dioxane [mol%]	Toluene [mol%]	Cyclohexane [mol%]	Temperature [°C]	{Instrument}	Timestamp
Good	train	18.238	59.40672	22.3555	22.1	RXN1	2019-11-14
Good	train	23.315	37.88732	38.7977	22.2	RXN1	2019-11-14
Good	train	16.405	56.02367	27.5714	22.0	RXN1	2019-11-14
Good	train	41.196	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

Ultimately, all data needs to be quantitative

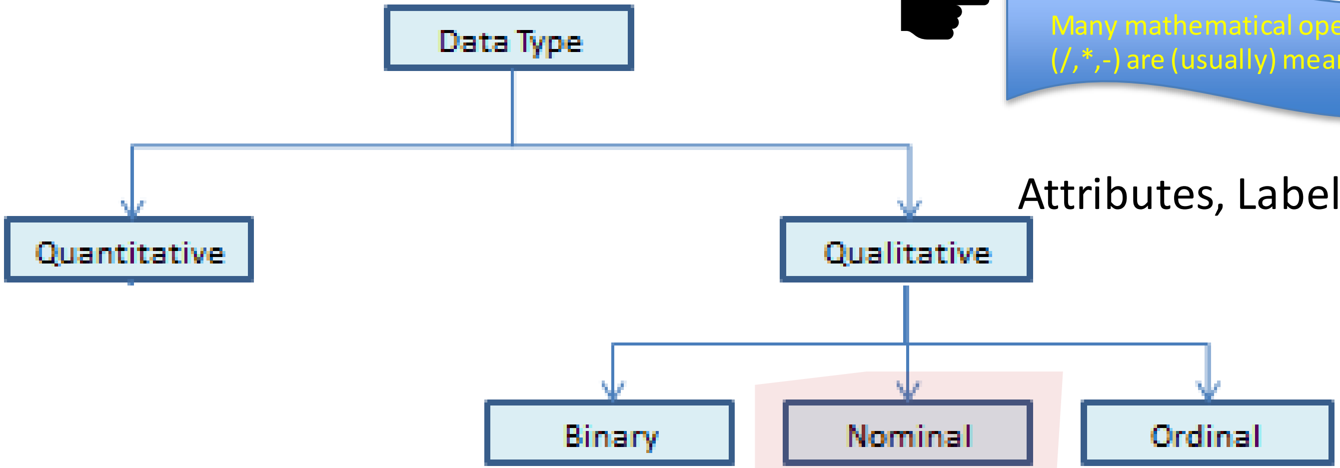


Taxonomy of data: Qualitative → Quantitative



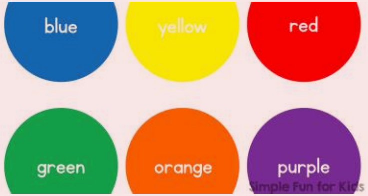
Taxonomy of data: Qualitative → Quantitative

Many mathematical operations (/,*,-) are (usually) meaningless



Attributes, Labels

0 1 2



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

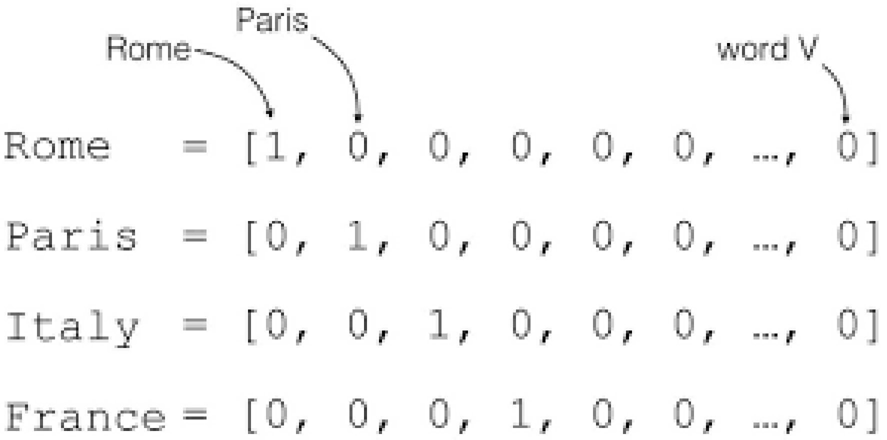
Original data:

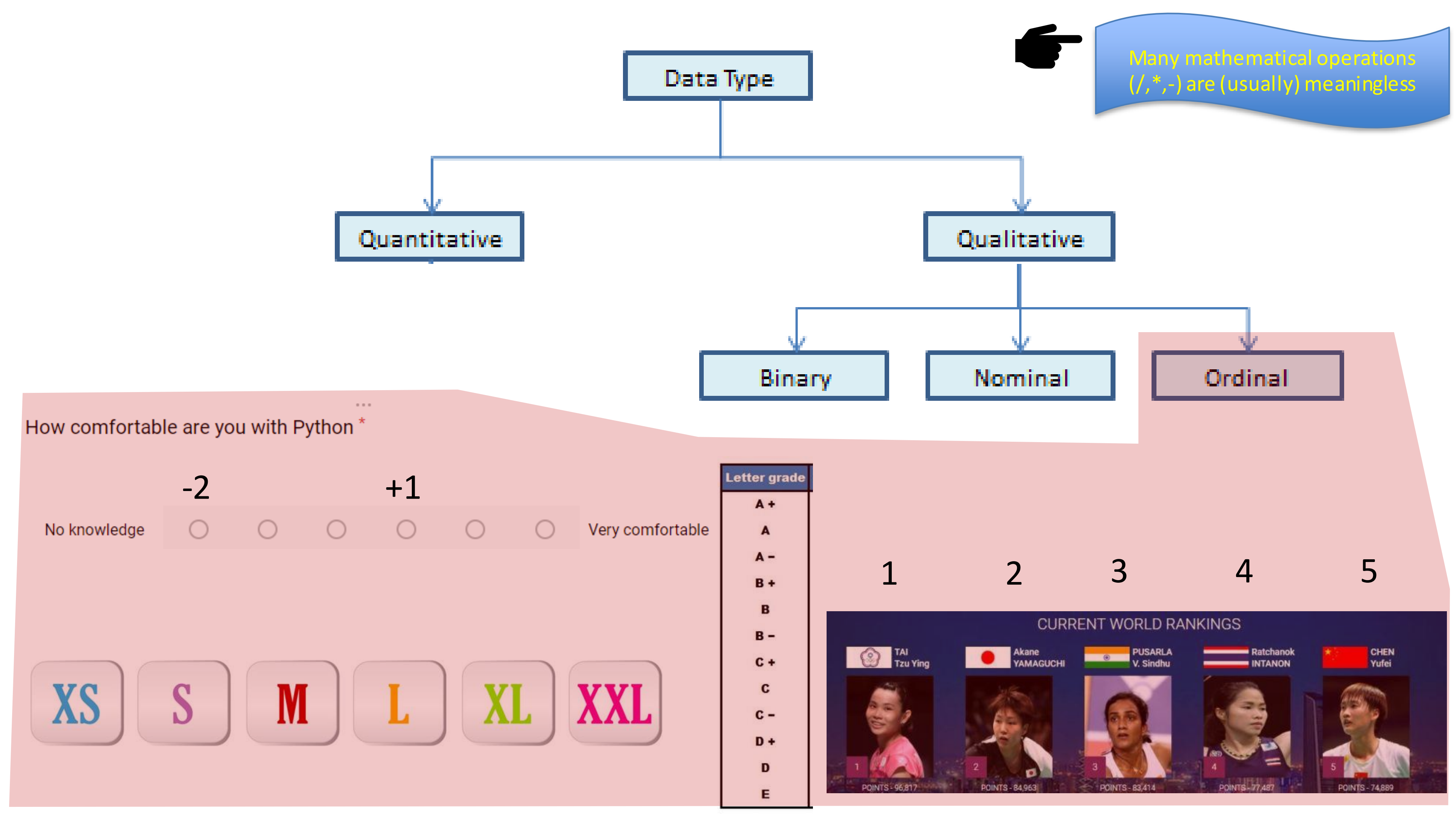
id	Color
1	White
2	Red
3	Black
4	Purple
5	Gold

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




Data Type
Quantitative
Qualitative
Binary
Nominal
Ordinal
Many mathematical operations (/,*,-) are (usually) meaningless
How comfortable are you with Python *
-2
+1
No knowledge
Very comfortable
Letter grade
A +
A
A –
B +
B
B –
C +
C
C –
D +
D
E
XS
S
M
L
XL
XXL
1
2
3
4
5
CURRENT WORLD RANKINGS
TAI Tzu Ying
Akane YAMAGUCHI
PUSARLA V. Sindhu
Ratchanok INTANON
CHEN Yufei
POINTS - 96,817
POINTS - 84,963
POINTS - 83,414
POINTS - 77,487
POINTS - 74,889

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

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
...	...	...	...	...

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

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

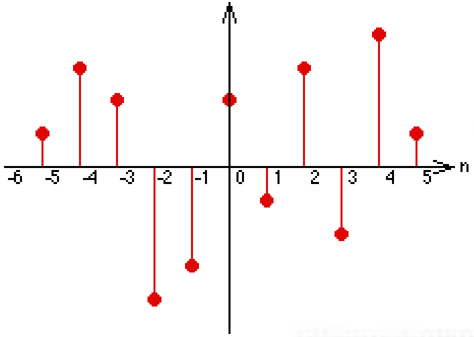
Feature / Attribute

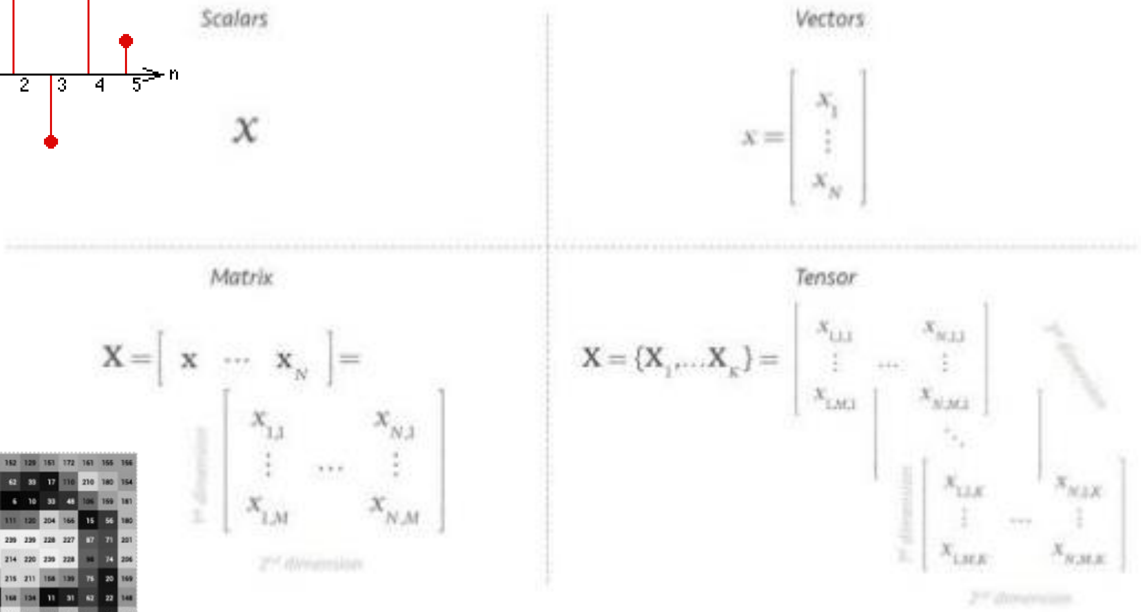
Sample

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

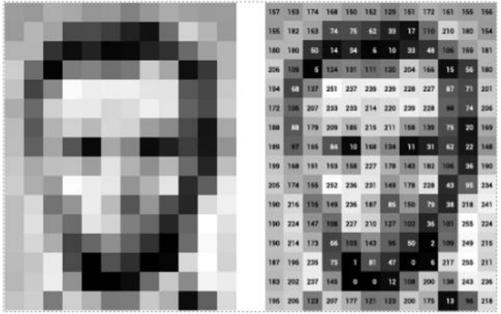
able1

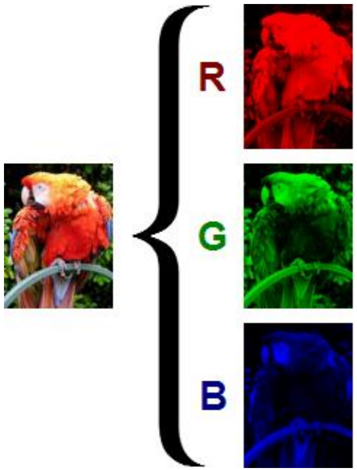
Data Sample Representations



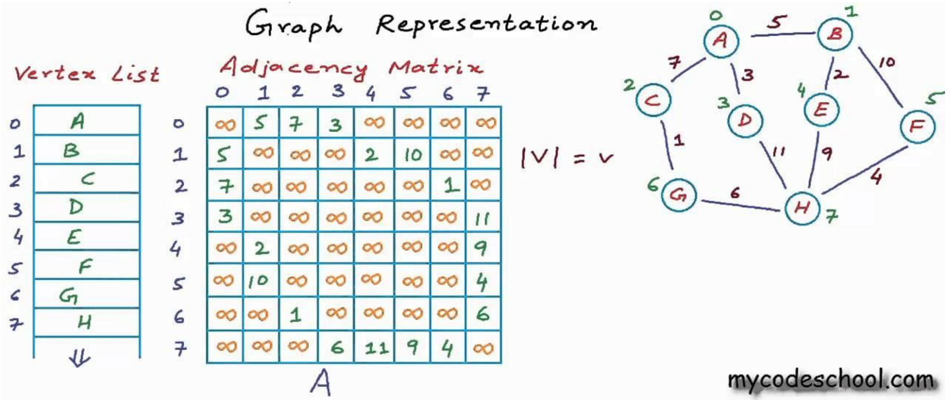


2-d image





Data Representations



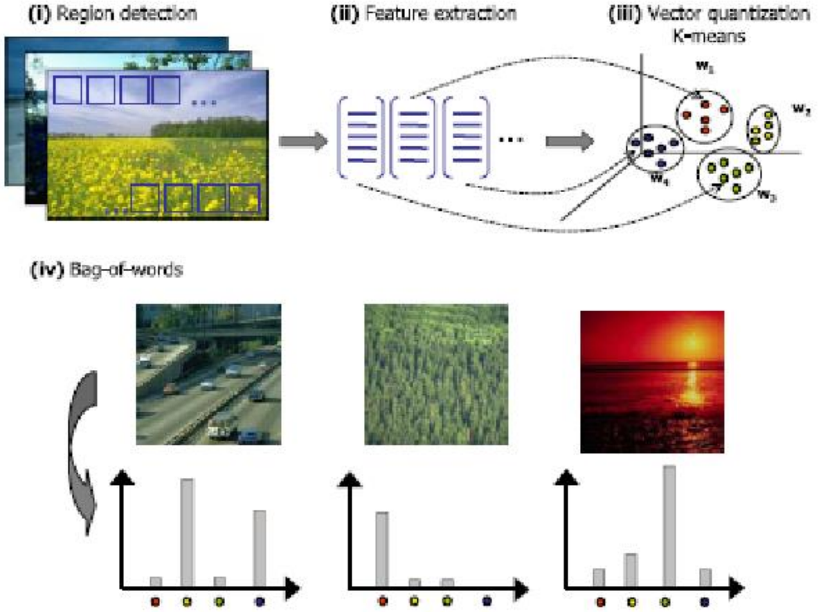
Feature Extraction (FE)

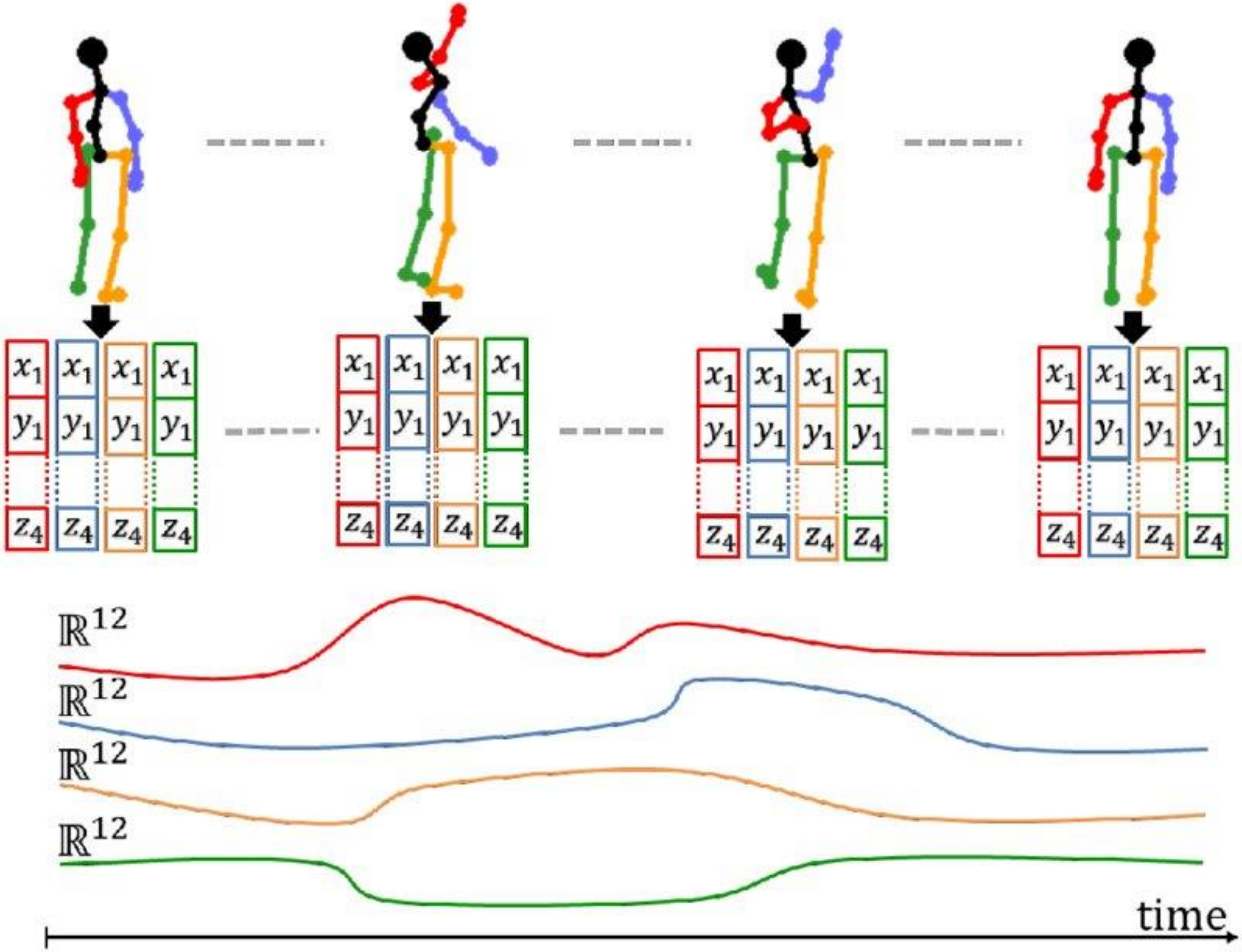
- Def: Feature Extraction (FE) is any algorithm that transformation raw data into features that can be used as an input for a learning algorithm.

The Bag of Words Representation

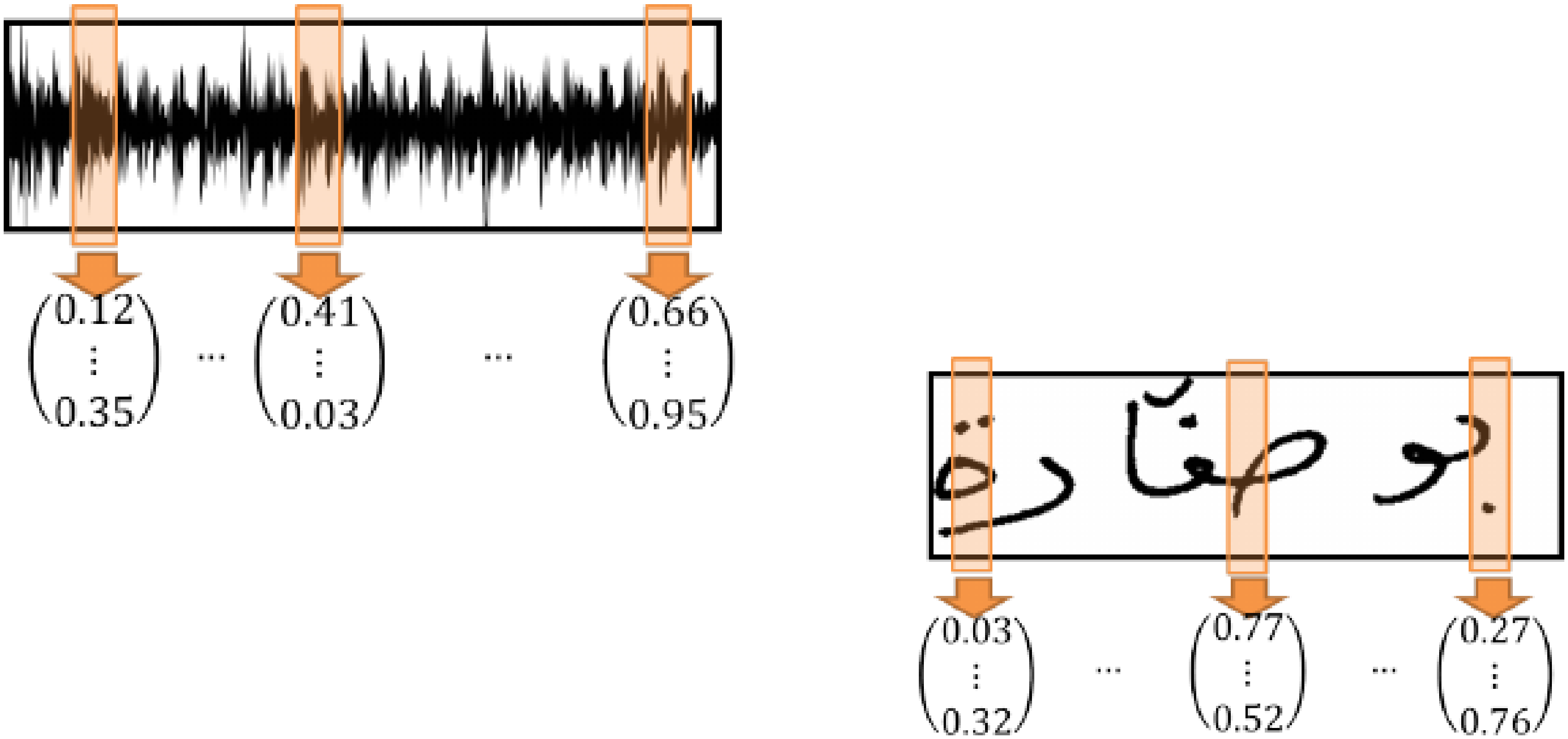
I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy 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
...	...

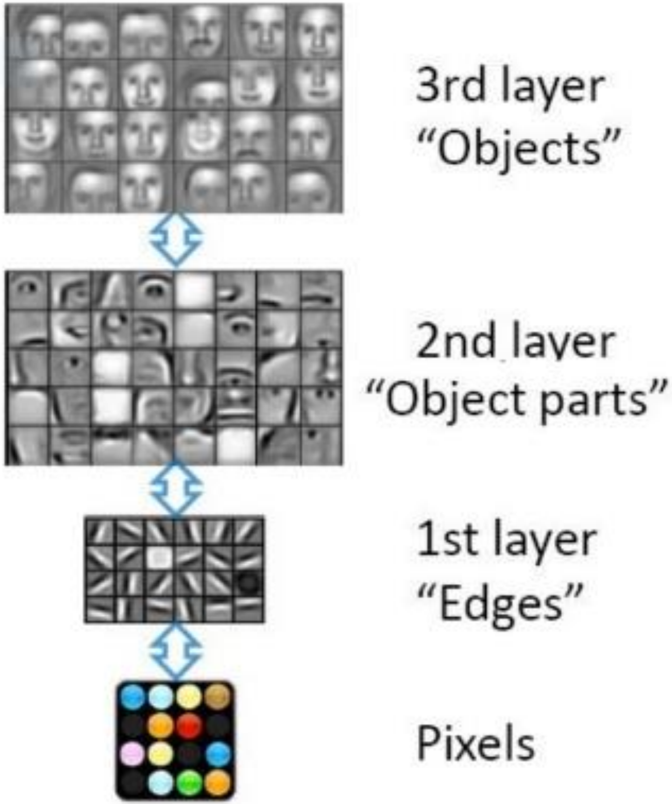


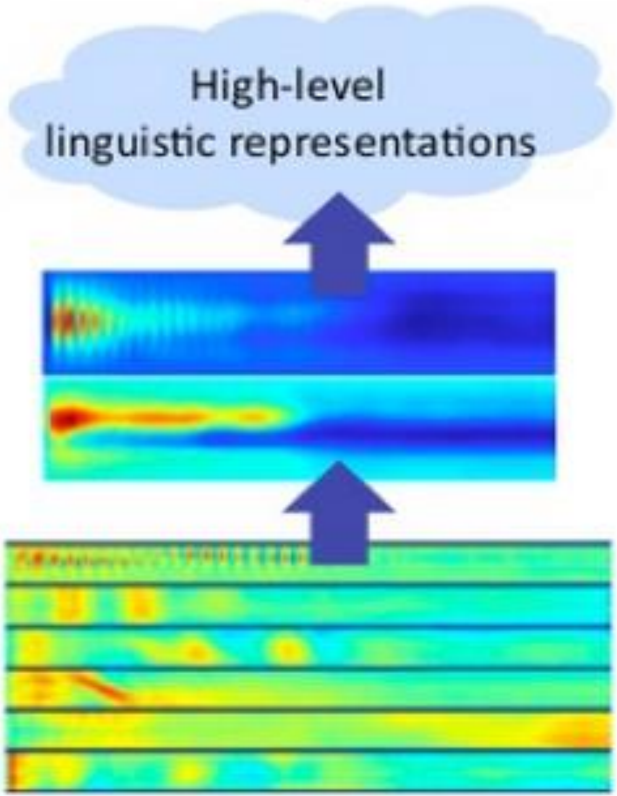


x_1 x_1 x_1 x_1
y_1 y_1 y_1 y_1
z_4 z_4 z_4 z_4
$\mathbb{R}^{12}$
$\mathbb{R}^{12}$
$\mathbb{R}^{12}$
$\mathbb{R}^{12}$
time


0.12
0.35
0.41
0.03
0.66
0.95
0.03
0.32
0.77
0.52
0.27
0.76

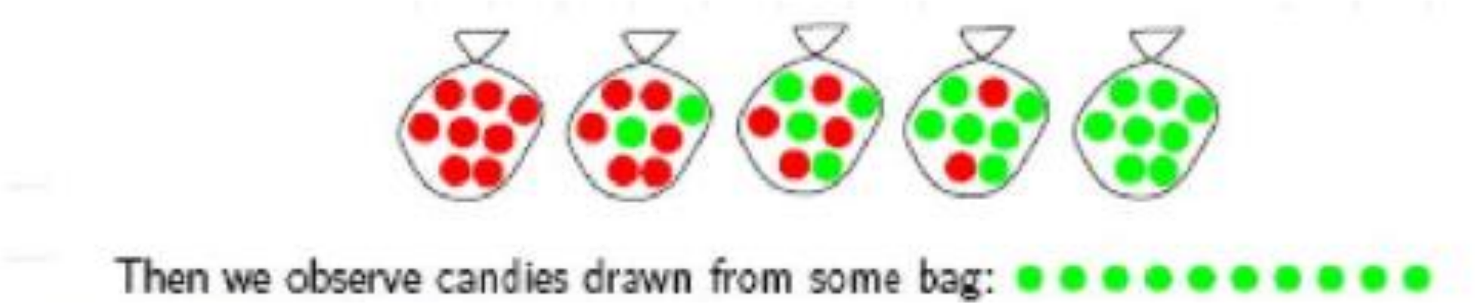
Feature-based, Hierarchical Data Representations





Data – a probability-based perspective

- The basis for Statistical Learning Theory



  - 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’)

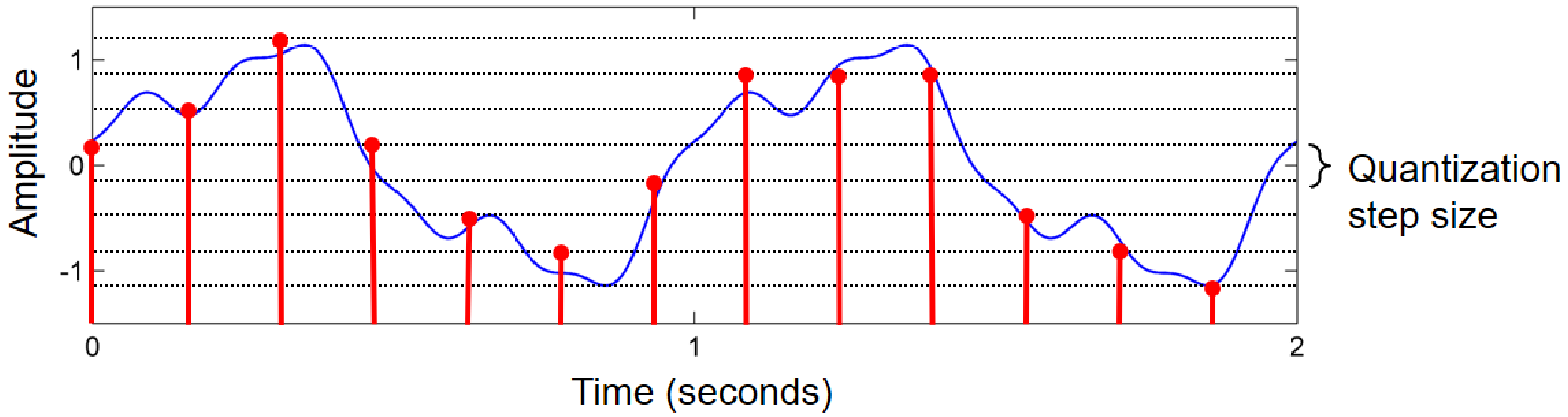
- Statistical assumptions on data
  - i.i.d (independent and identically distributed)
  - Online (e.g. time-series data)

Lecture Outline

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

Quantization

1. Continuous → Discrete ('Rounding off')

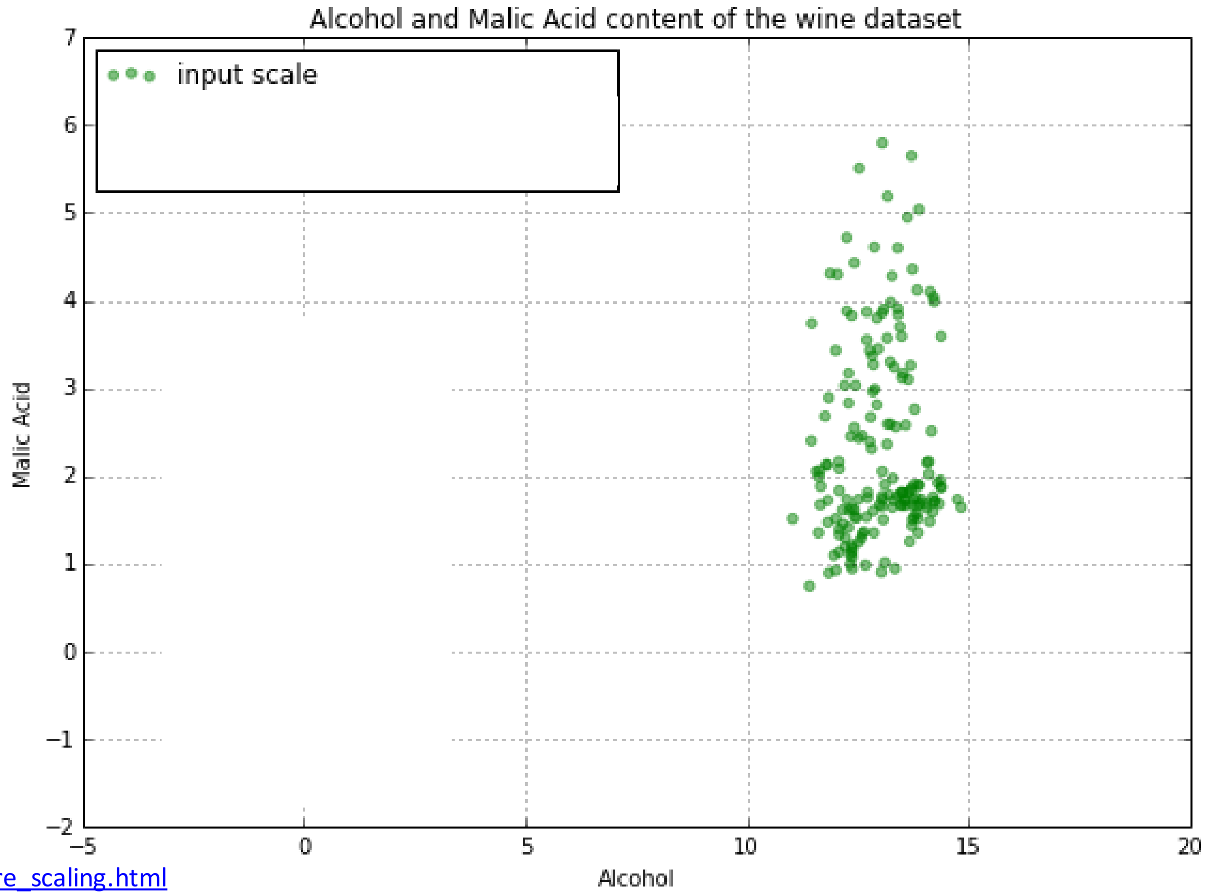


2. Binary Quantization ('Thresholding')

Data Normalization

	Class label	Alcohol	Malic acid
0	1	14.23	1.71
1	1	13.20	1.78
2	1	13.16	2.36
3	1	14.37	1.95
4	1	13.24	2.59

- ⋮



https://sebastianraschka.com/Articles/2014_about_feature_scaling.html

Popular normalization approaches

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

MinMax Scaling

$$z = \frac{x - \mu}{\sigma}$$

$$\mu = \frac{1}{N}\sum_{i=1}^{N}(x_i)$$

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

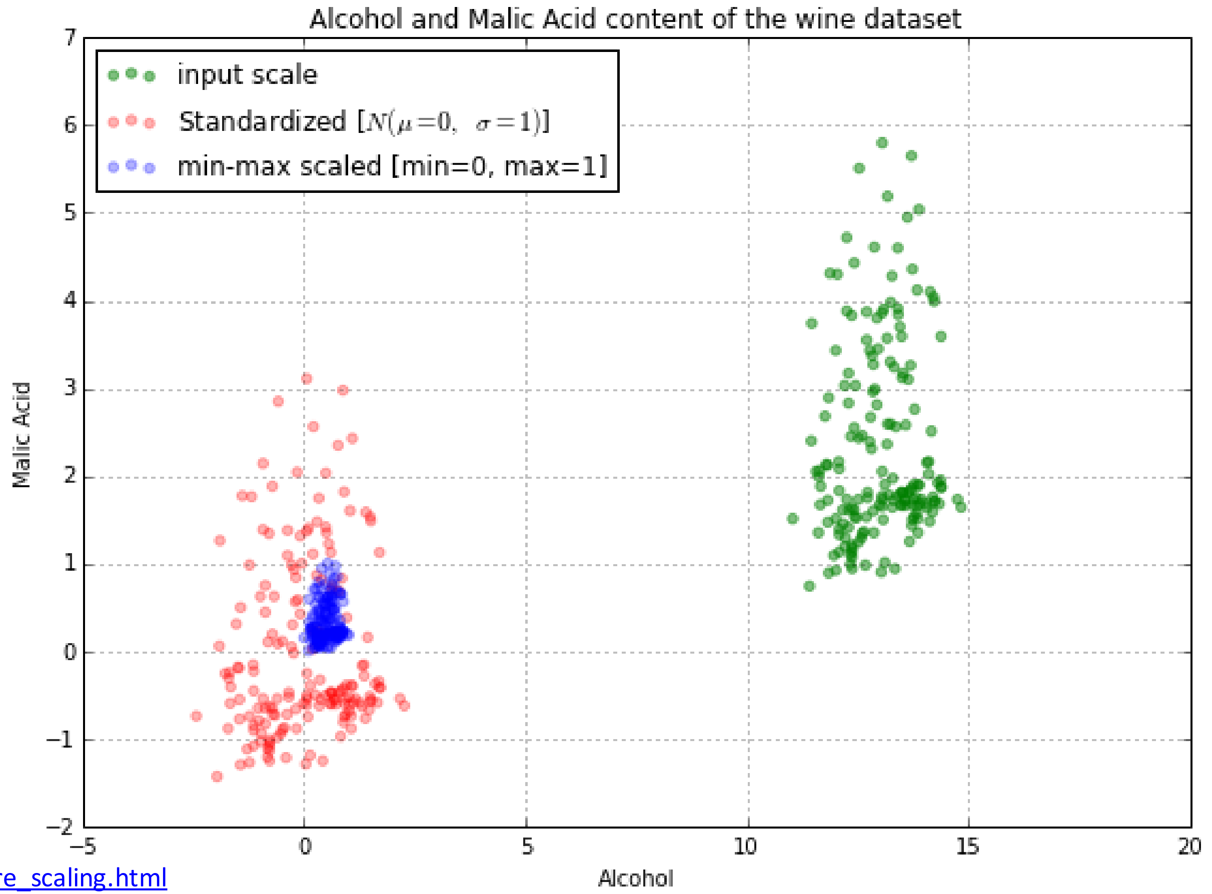
Standardization
(Unit Normal Scaling)

	Class label	Alcohol	Malic acid
0	1	14.23	1.71
1	1	13.20	1.78
2	1	13.16	2.36
3	1	14.37	1.95
4	1	13.24	2.59

Data Normalization (applied to each feature)

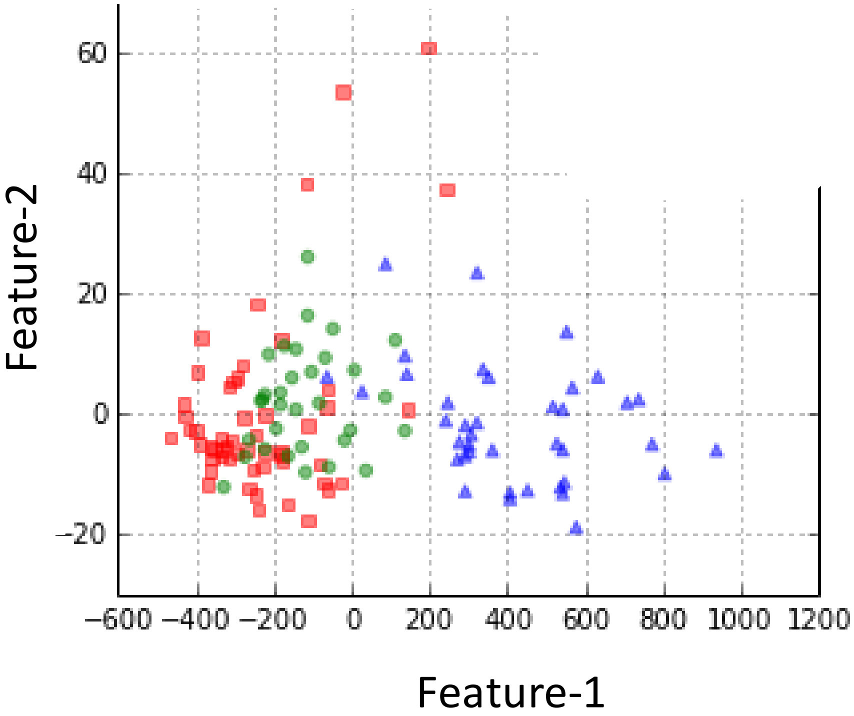
	Class label	Alcohol	Malic acid
0	1	14.23	1.71
1	1	13.20	1.78
2	1	13.16	2.36
3	1	14.37	1.95
4	1	13.24	2.59

- ⋮

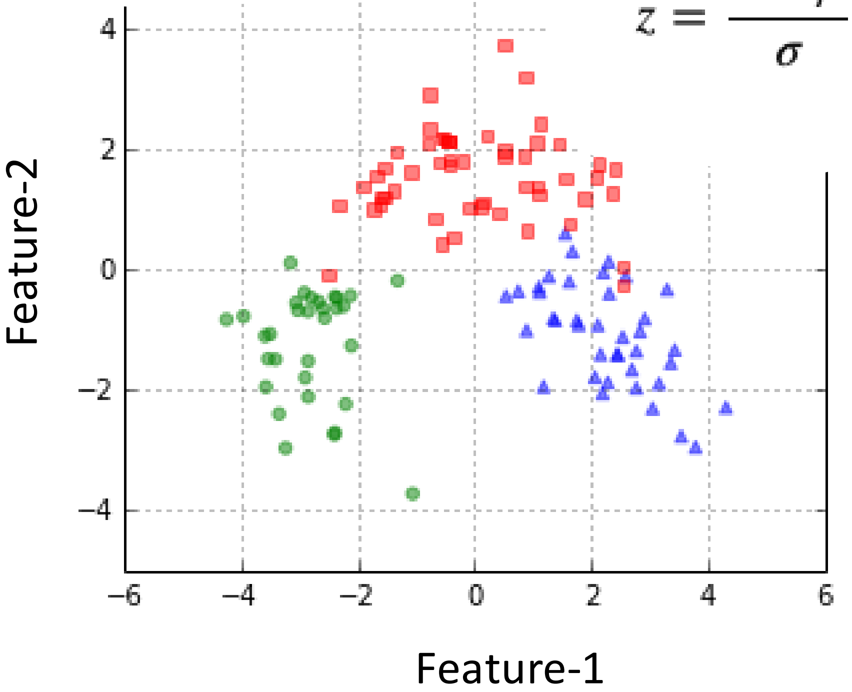


https://sebastianraschka.com/Articles/2014_about_feature_scaling.html

Before standardization



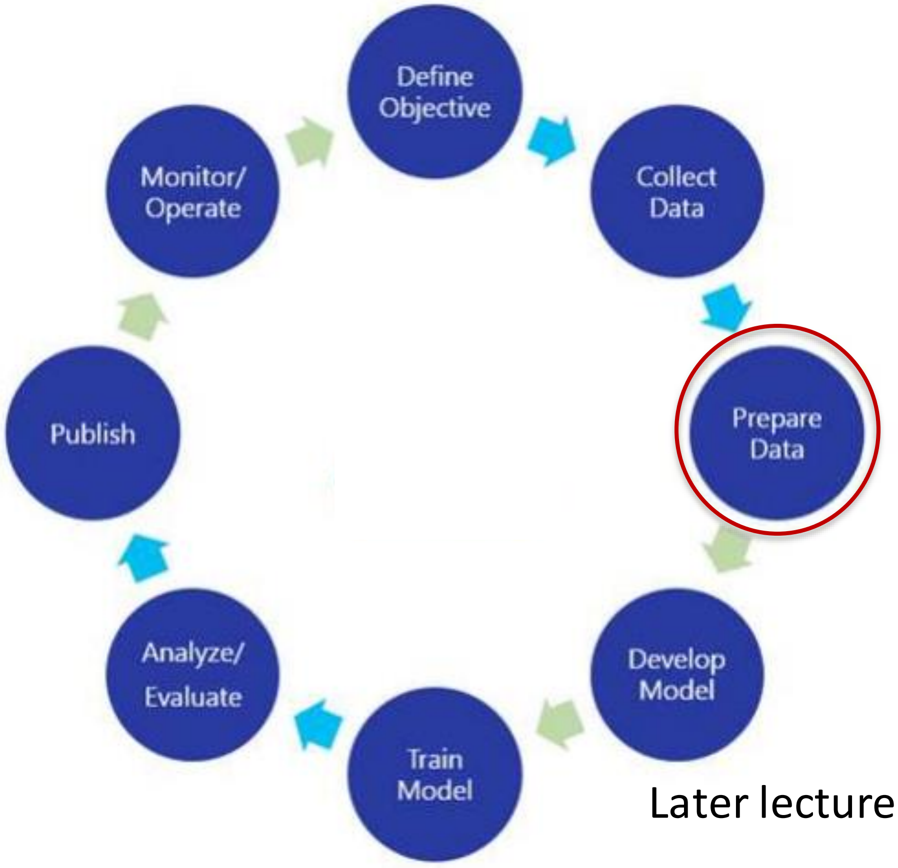
After standardization



$$z = \frac{x - \mu}{\sigma}$$

https://sebastianraschka.com/Articles/2014_about_feature_scaling.html

Workflow of a Machine Learning Problem



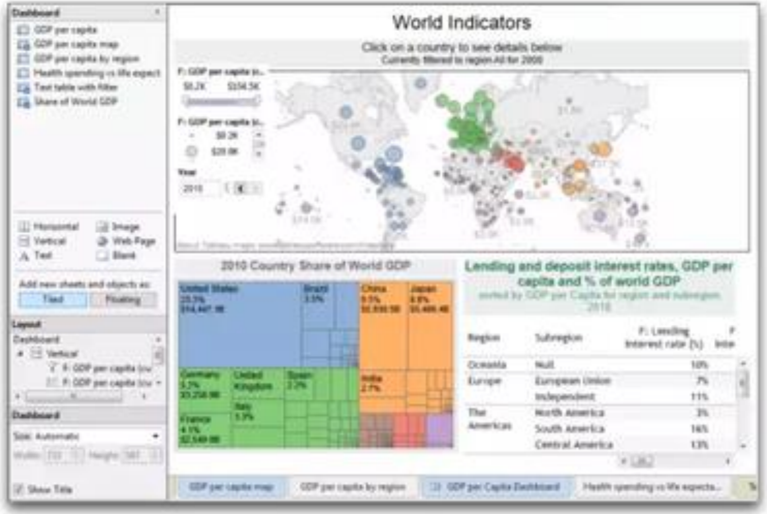
Later lectures …

Lecture Outline

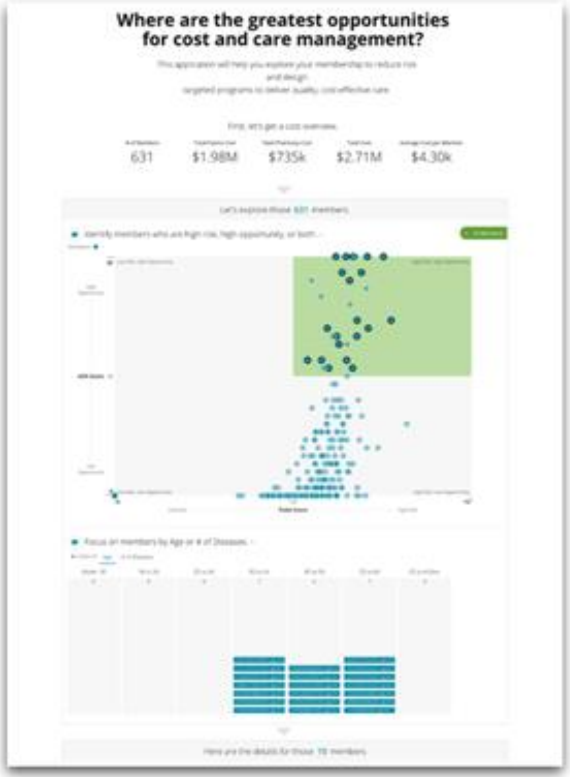
- *ML Workflow*
- *Data Representations*
- *Basic Data Transformations*
- Data Visualization

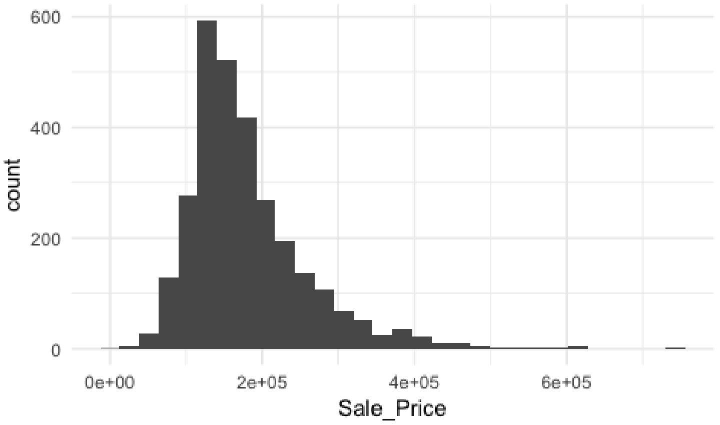
Gazing at Data: Data visualization

data exploration

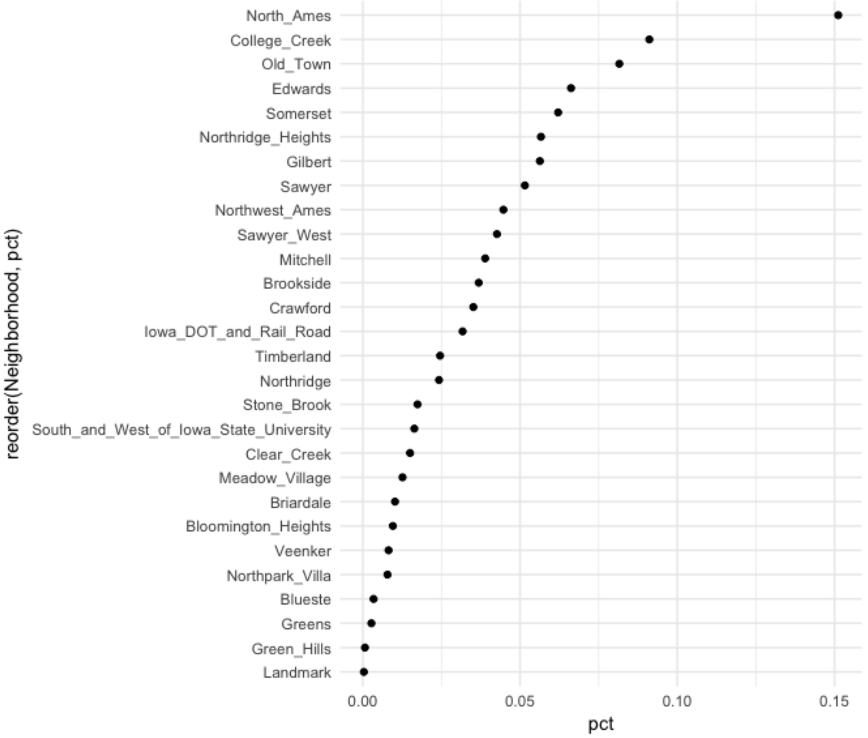


data presentation

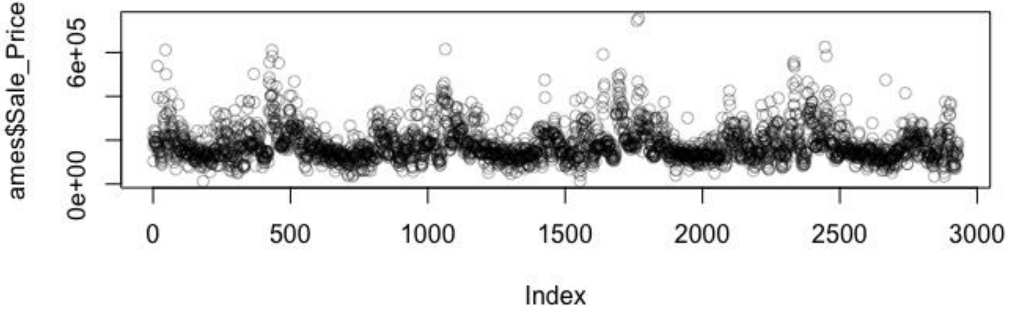




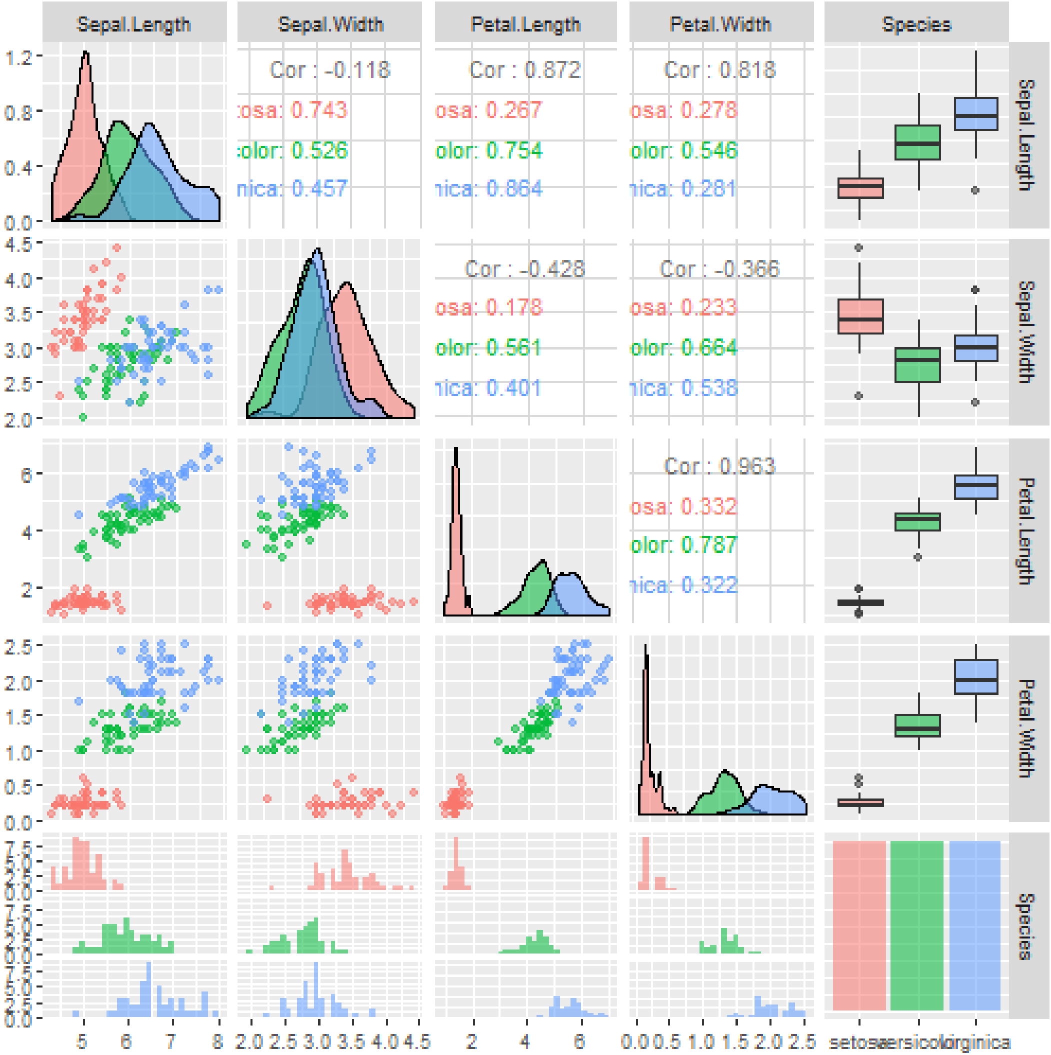
600
400
200
0
count
0e+00
2e+05
4e+05
6e+05
Sale_Price



North_Ames
College_Creek
Old_Town
Edwards
Somerset
Northridge_Heights
Gilbert
Sawyer
Northwest_Ames
Sawyer_West
Mitchell
Brookside
Crawford
Iowa_DOT_and_Rail_Road
Timberland
Northridge
Stone_Brook
South_and_West_of_Iowa_State_University
Clear_Creek
Meadow_Village
Briardale
Bloomington_Heights
Veenker
Northpark_Villa
Blueste
Greens
Green_Hills
Landmark
reorder(Neighborhood, pct)
0.00
0.05
0.10
0.15
pct



ames$Sale_Price
0e+00
6e+05
0
500
1000
1500
2000
2500
3000
Index



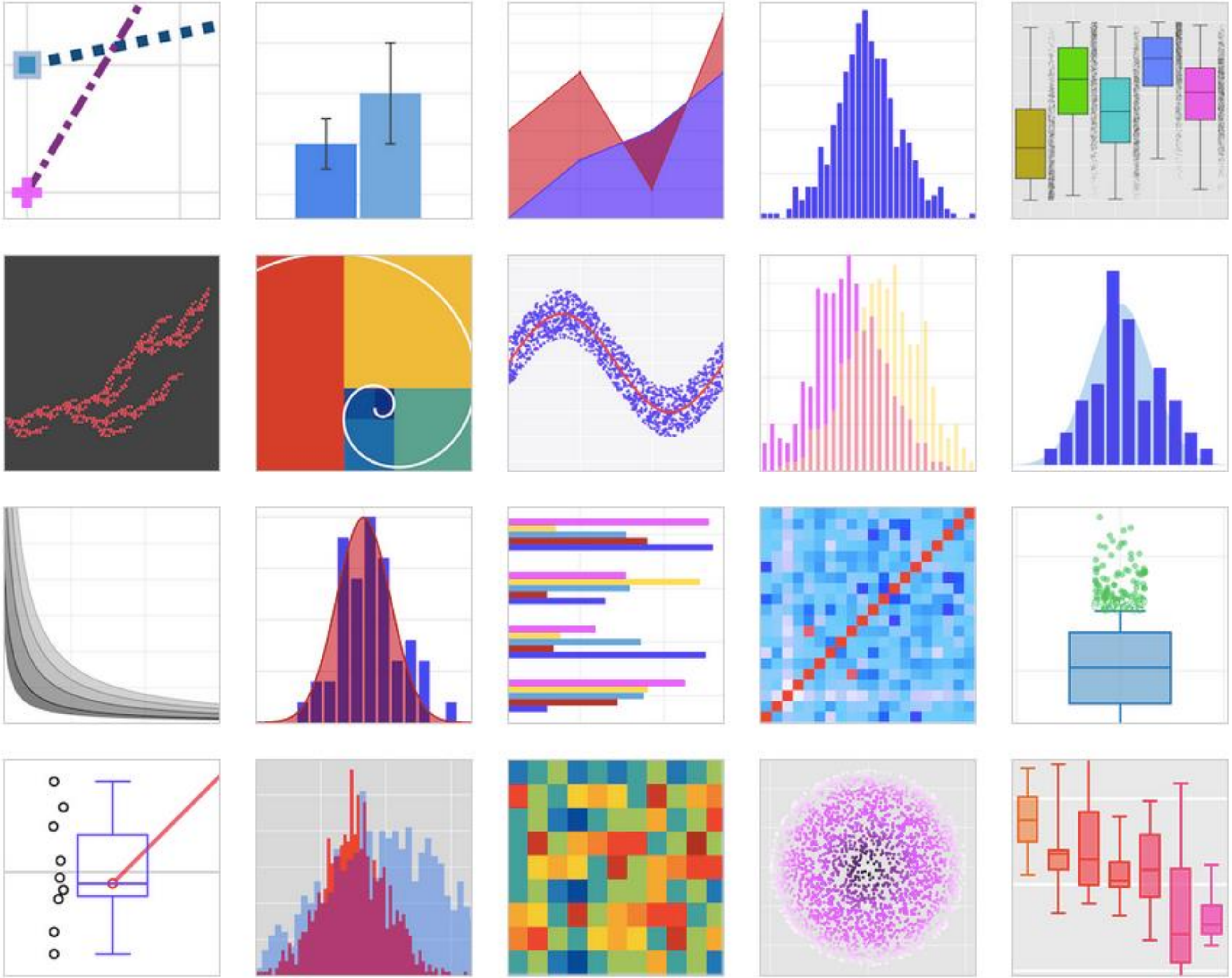
Sepal.Length
Sepal.Width
Petal.Length
Petal.Width
Species
Cor : -0.118
Cor : 0.872
Cor : 0.818
Cor : -0.428
Cor : -0.366
Cor : 0.963

Data visualization

treemap

leaderboard

Shuttle		40 Yard		Bench Press		Vert Leap (in)		Broad Jump (in)	
Jordan Jeffer...	4.06	Robert Griffin	4.41	Jordan Jeffer...	14	Robert Griffin	39	Andrew Luck	124
Russell Wilson	4.09	Russell Wilson	4.55	Darron Thomas	14	Jacory Harris	37	Darron Thomas	121
Austin Davis	4.11	Jordan Jeffer...	4.65	Robert Griffin	---	Jordan Jeffer...	37	Robert Griffin	120
Chandler Harn...	4.15	Andrew Luck	4.67	Russell Wilson	---	Darron Thomas	36	Russell Wilson	118
Andrew Luck	4.28	Aaron Corp	4.72	Andrew Luck	---	Andrew Luck	36	Jordan Jeffer...	116
Darron Thomas	4.28	Jacory Harris	4.72	Aaron Corp	---	Russell Wilson	34	Jacory Harris	113
Aaron Corp	4.30	Chandler Harn...	4.76	Jacory Harris	---	Chandler Harn...	33	Tyler Hansen	113
Patrick Witt	4.37	Tyler Hansen	4.78	Chandler Harn...	---	Case Keenum	33	Chandler Harn...	112
B.J. Coleman	4.38	Darron Thomas	4.80	Tyler Hansen	---	Aaron Corp	32	Nick Foles	112
Jacory Harris	4.40	Case Keenum	4.82	Case Keenum	---	Patrick Witt	32	Austin Davis	109



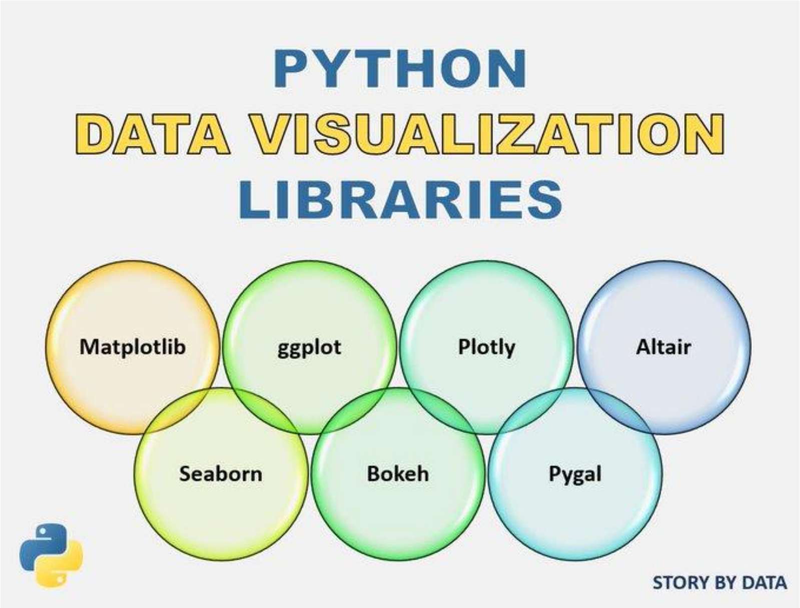
"In good information visualization, there are no rules, no guidelines, no templates, no standard technologies, no stylebooks … You must simply do whatever it takes."

—*Edward Tufte*

VISAGE.CO

Resources

- https://towardsdatascience.com/5-quick-and-easy-data-visualizations-in-python-with-code-a2284bae952f



https://twitter.com/storybydata/status/1166337648341991424