

# 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](mailto:ravi.kiran@iiit.ac.in))

Vineet Gandhi ([v.gandhi@iiit.ac.in](mailto:v.gandhi@iiit.ac.in))



Center for Visual Information Technology (CVIT)  
IIIT Hyderabad

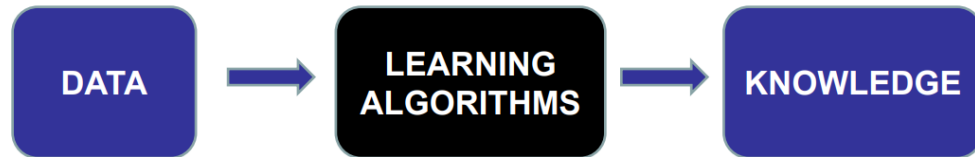
# Announcements

- 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



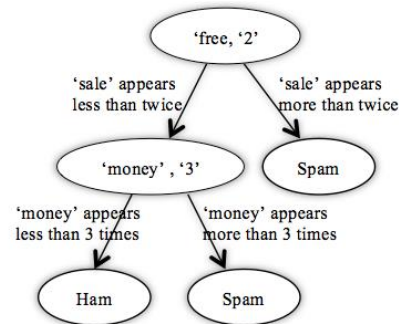
Algorithmic methods that use data to improve their knowledge of a task

Task: Detect spam email



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

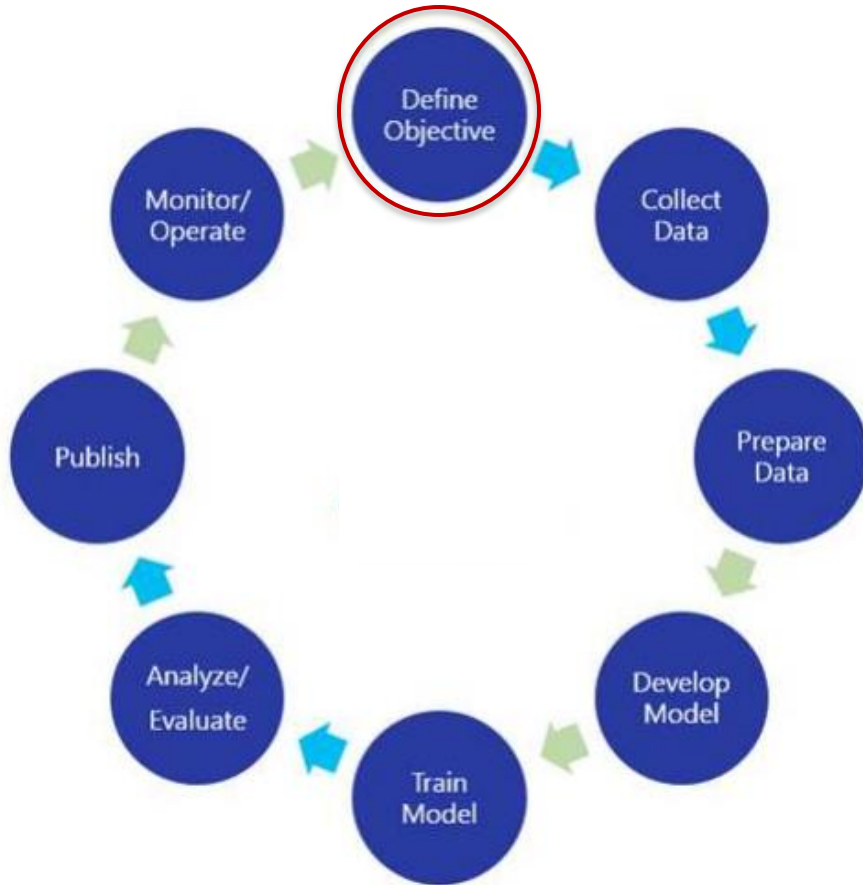
Knowledge:



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

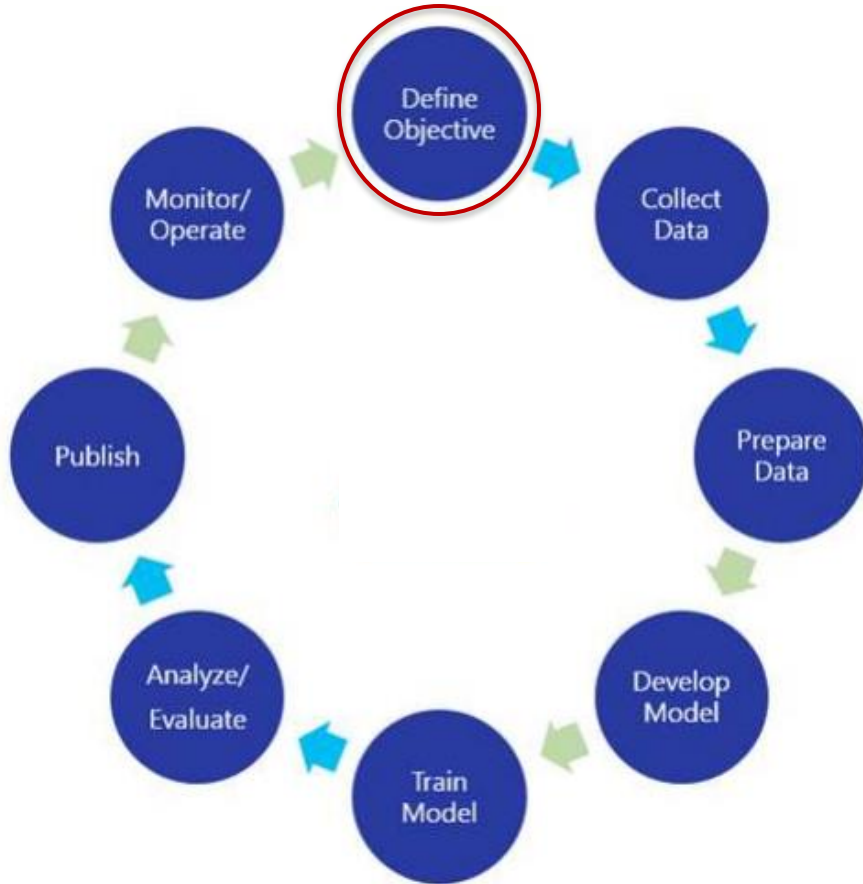
Algorithmic method: Decision Tree

# Workflow of a Machine Learning Problem

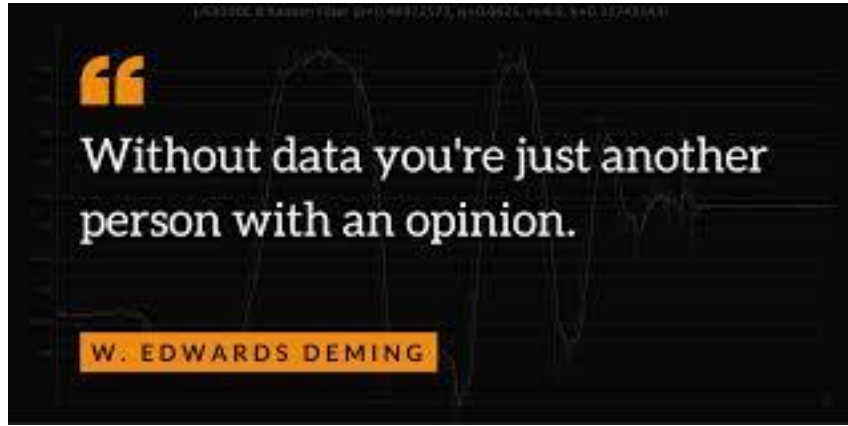


- Detect spam email
- Predict value of a stock
- Predict effect of advertising on sales
- Drive car 'safely' without human intervention
- Translate text from one language to another
- Sentiment Analysis
- ...

# Workflow of a Machine Learning Problem

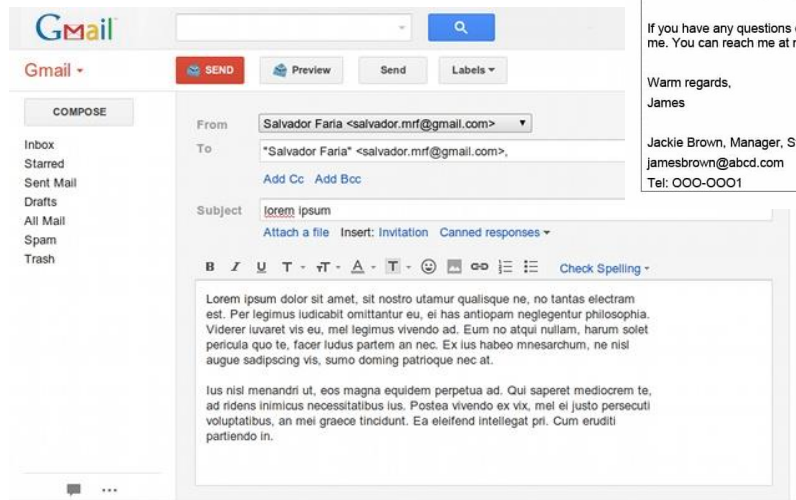
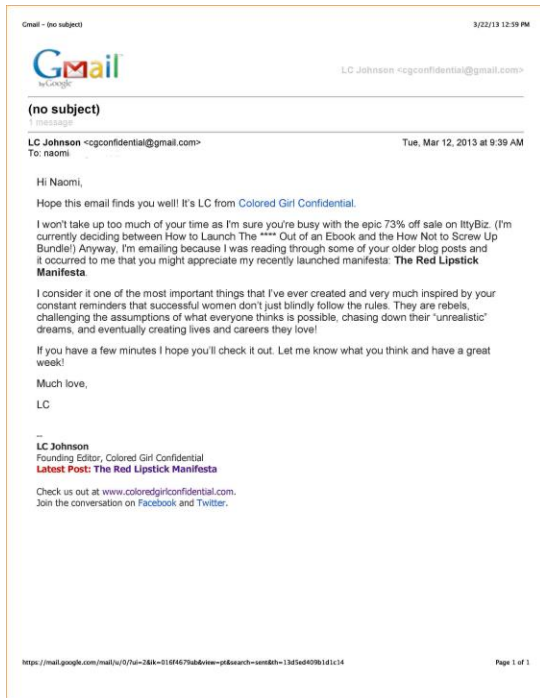


# No Data, no ML !



# Sources of data

## - Detect spam email





# Sources of data

- Predict value of a stock



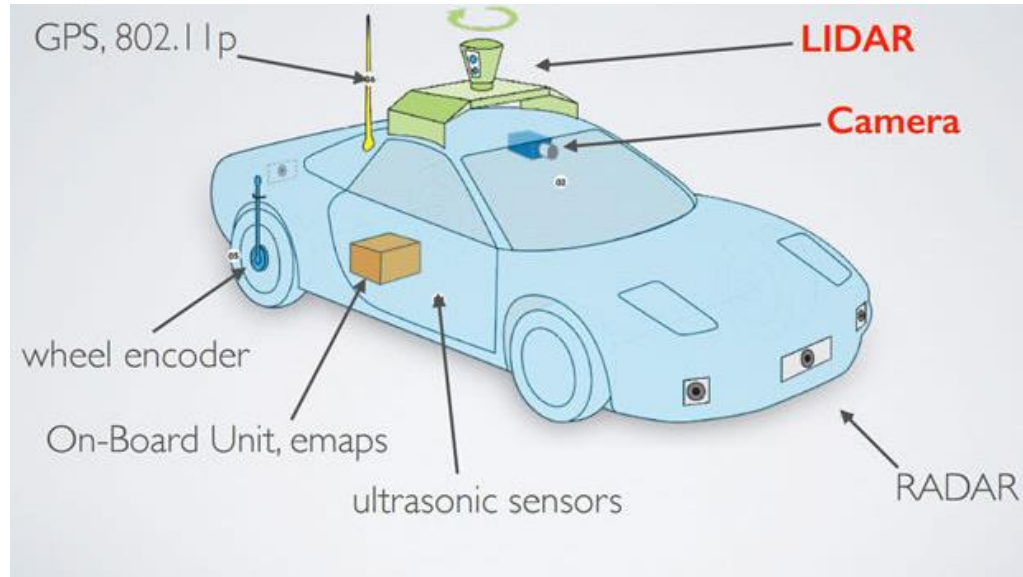
- Predict effect of advertising on sales

80 CASH MEMO		1/9
01	03	RO
1 MTN. ROGAN JOSH	1	600
1 CKN. MASALA	1	600
1 MID H. NOODLES	1	800
2 BTR NAAN	0	400
1 LASSI	1	000
2 LEMON I/TEA	0	800
1 DIET PEPSI	0	200
1 MASALI (B)	0	300
1 CKN. M. Noodles	1	800
1 WHITE RICE	0	800
MASALA TEA	0	300
1 CAPA CHAI	0	600
TOTAL		11 200

Raw Data may not always be digital in nature !

# Sources of data

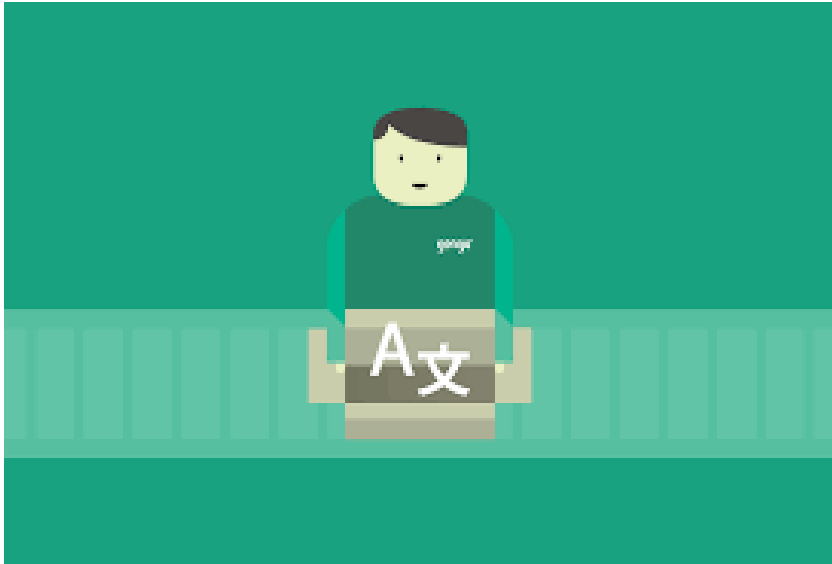
- Drive car safely without human intervention



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



A screenshot of a web browser window displaying a JSON response from the Mailgun API. The address bar shows the URL `https://api.mailgun.net/v2/domains/mailgun.com/messages/WyJlMTFiZ'`. The JSON data is as follows:

```
{
  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)**



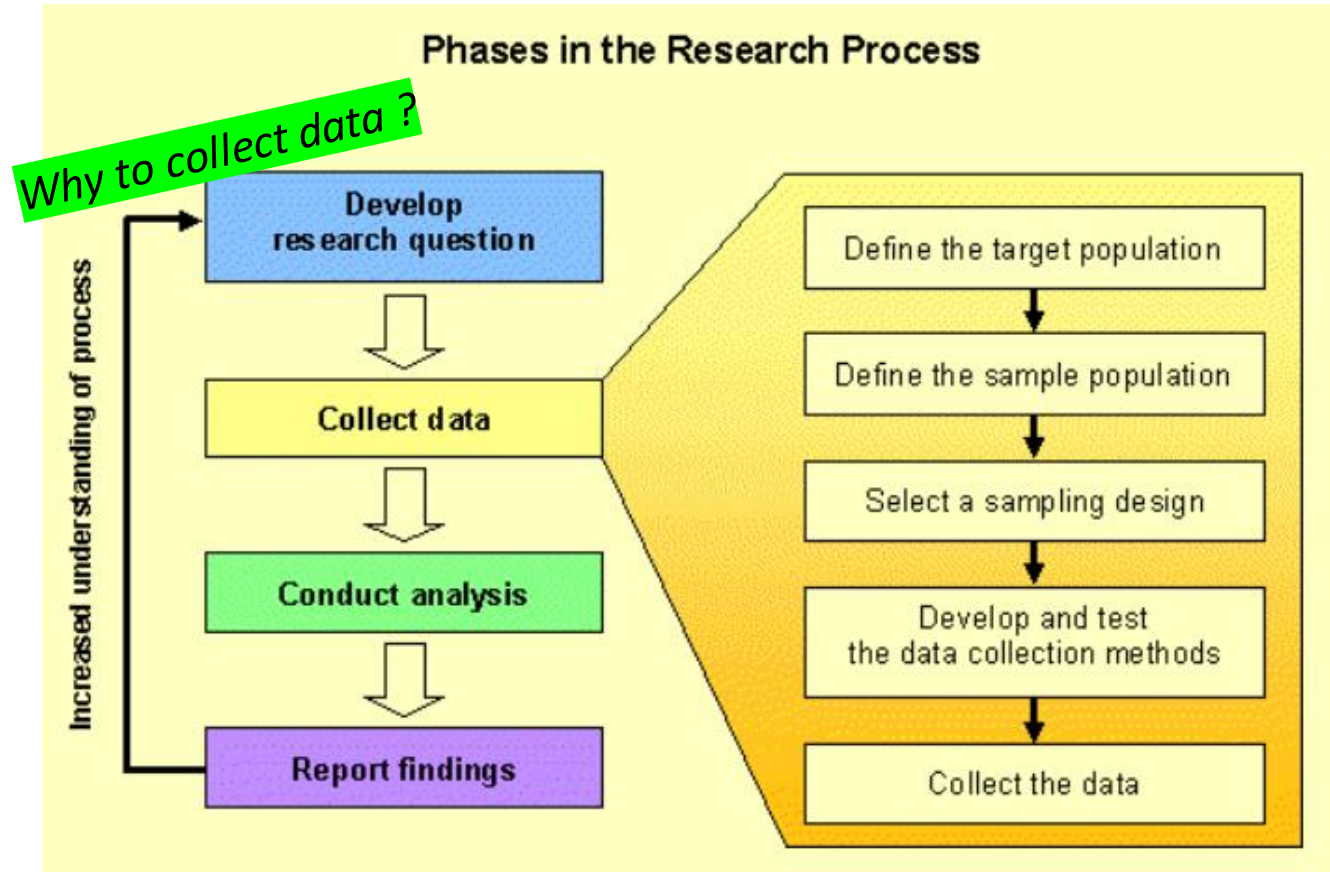
A screenshot of a web browser window displaying a JSON response from the Mailgun API. The address bar shows the URL `https://api.mailgun.net/v2/domains/mailgun.com/messages/WyJlMTFiZ`. The JSON data is as follows:

```
{
  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

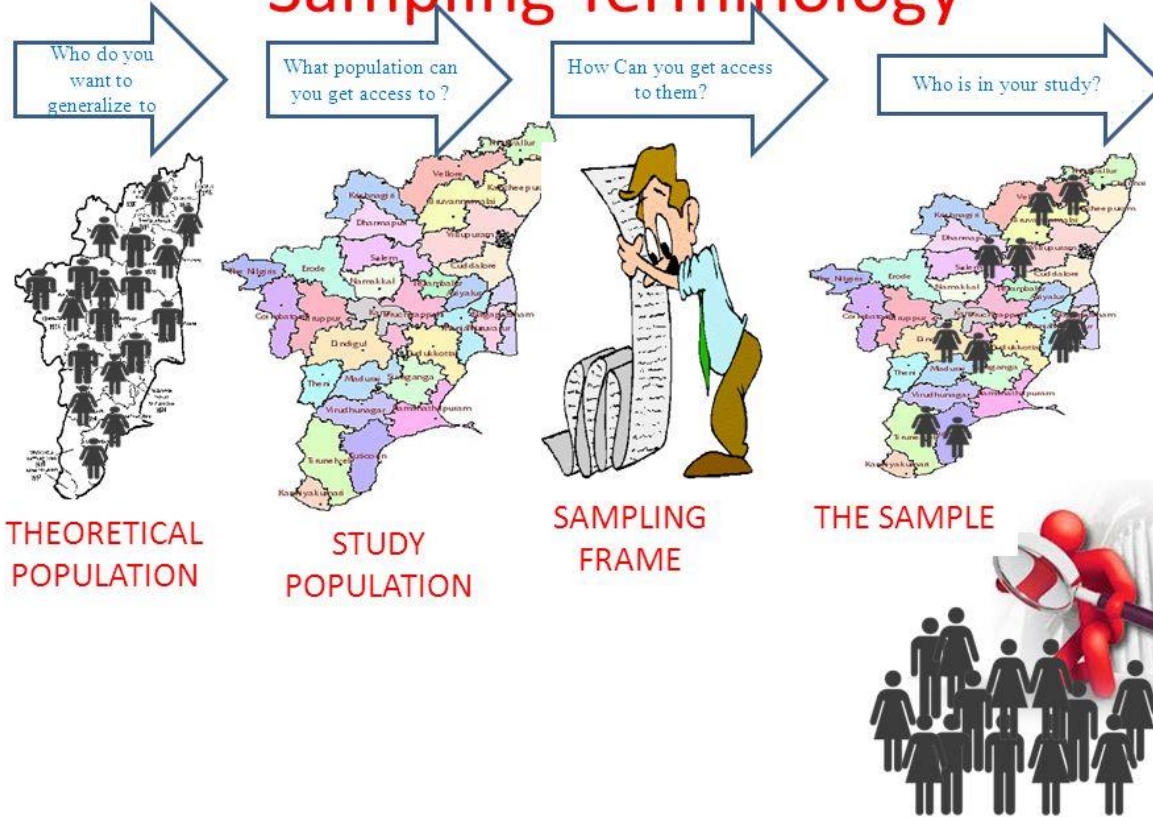
- What data to collect ?
- How (much) to collect ?

# The Research Method

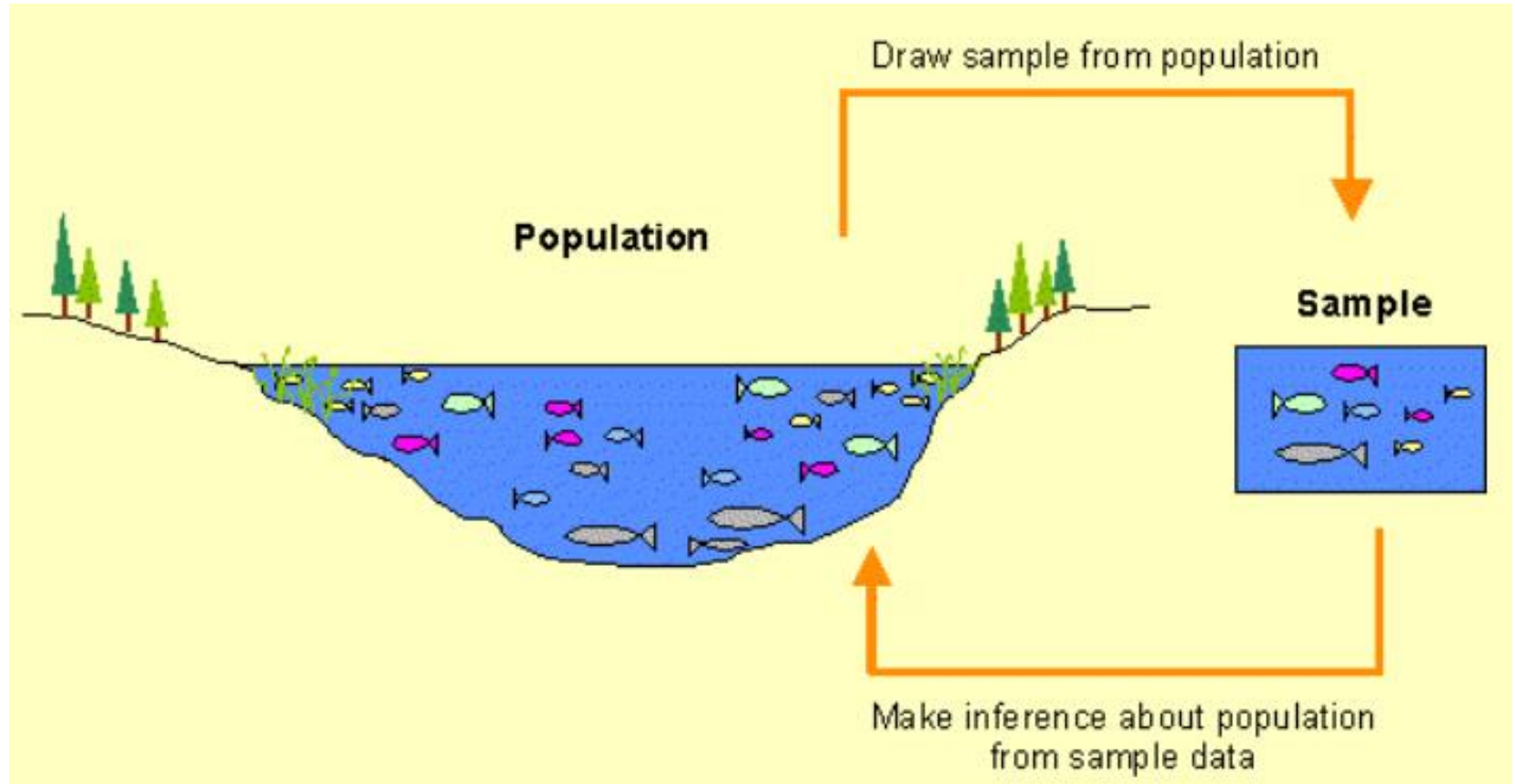


# The Research Method

## Sampling Terminology



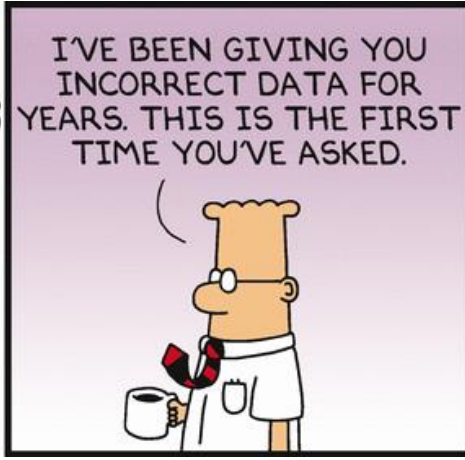
# The Research Method



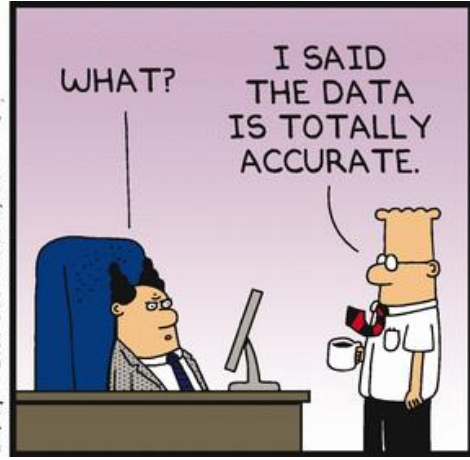
# Are our samples 'nice' ?



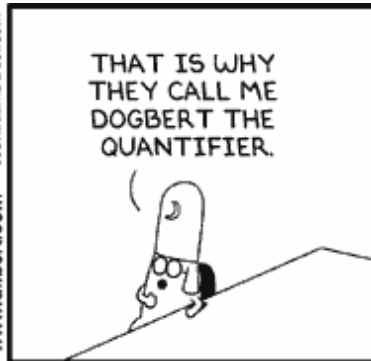
Dilbert.com DilbertCartoonist@gmail.com



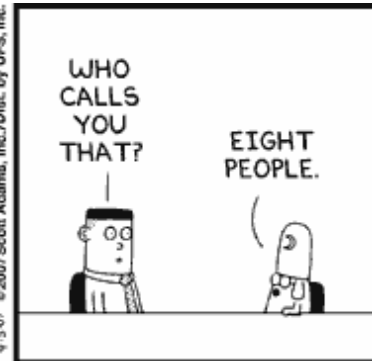
5-7-14 ©2014 Scott Adams, Inc./Dist. by Universal Uclick



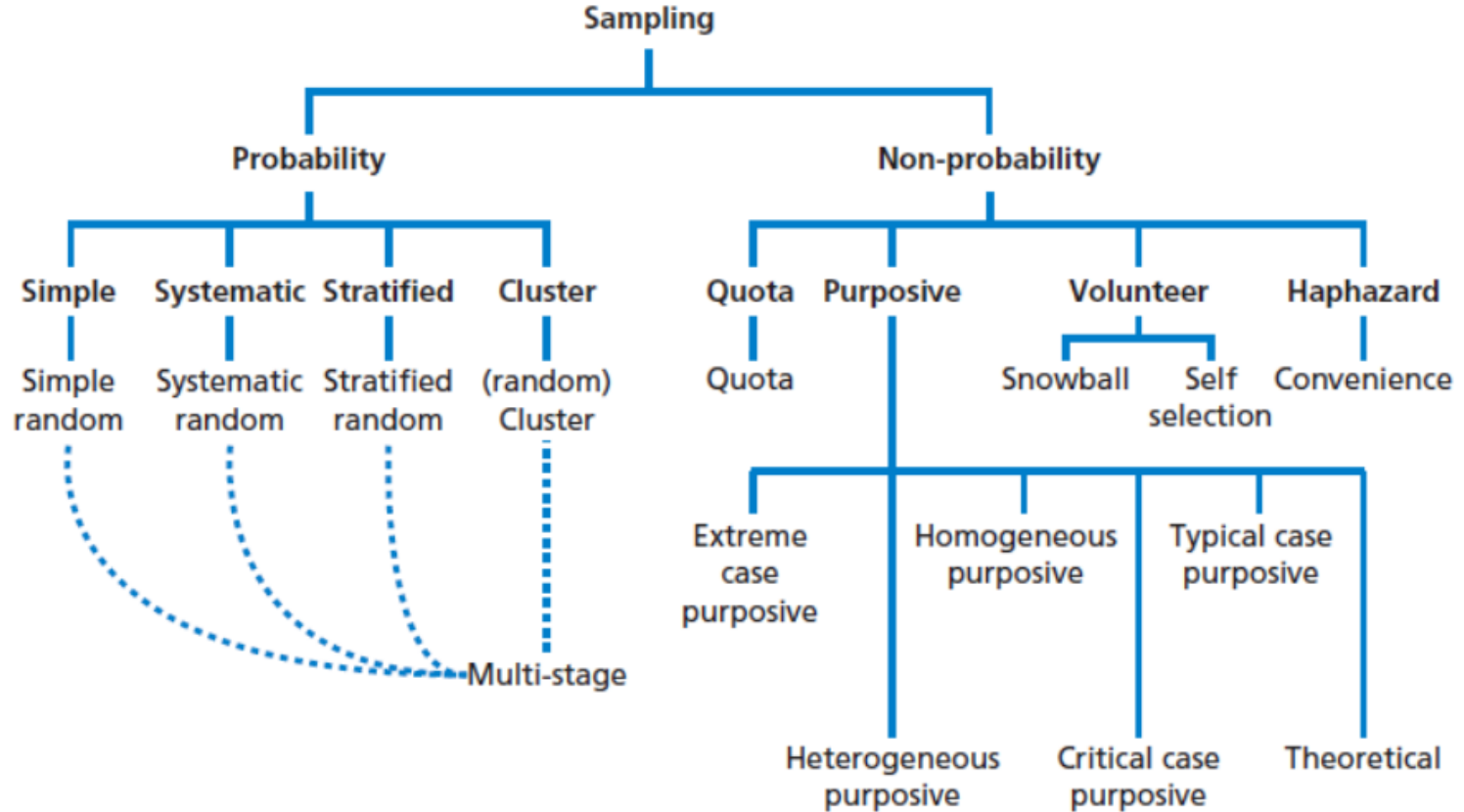
www.dilbert.com scottadams@aol.com



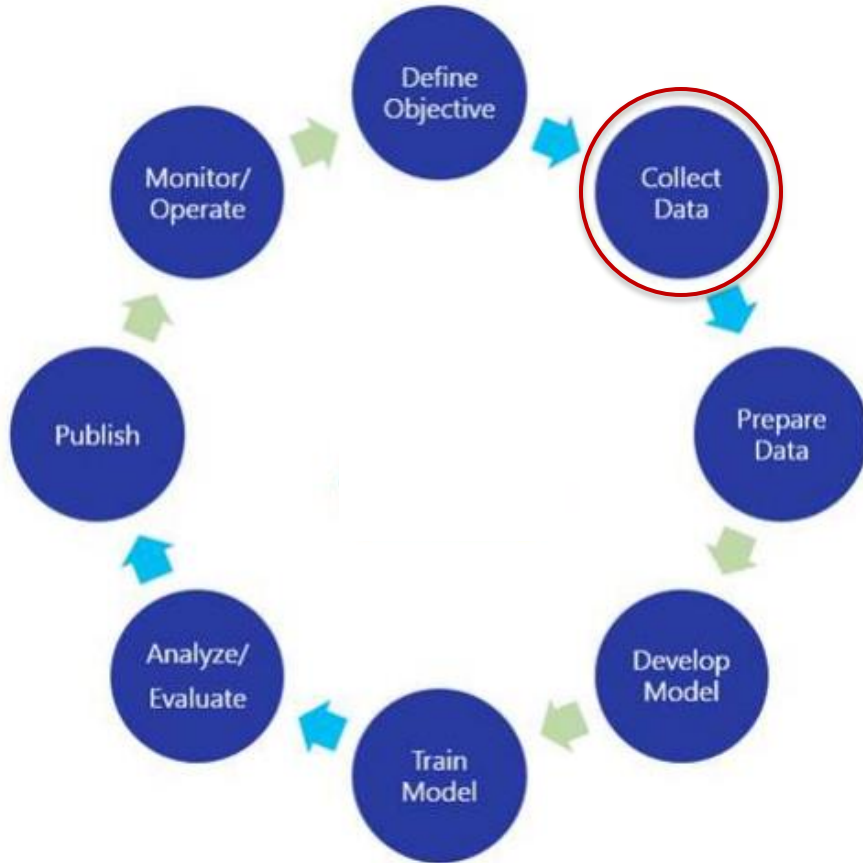
4-5-07 ©2007 Scott Adams, Inc./Dist. by UFS, Inc.



# Sampling Techniques

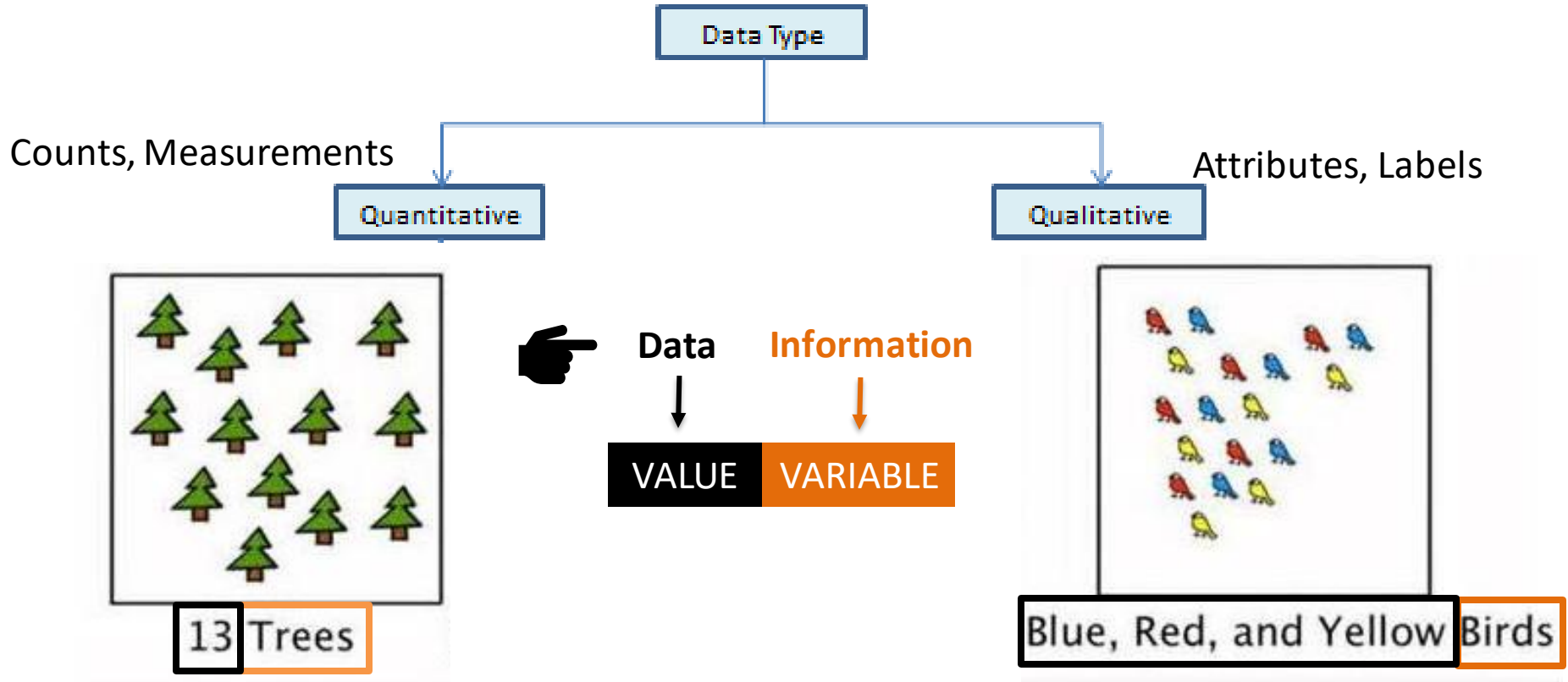


# Workflow of a Machine Learning Problem

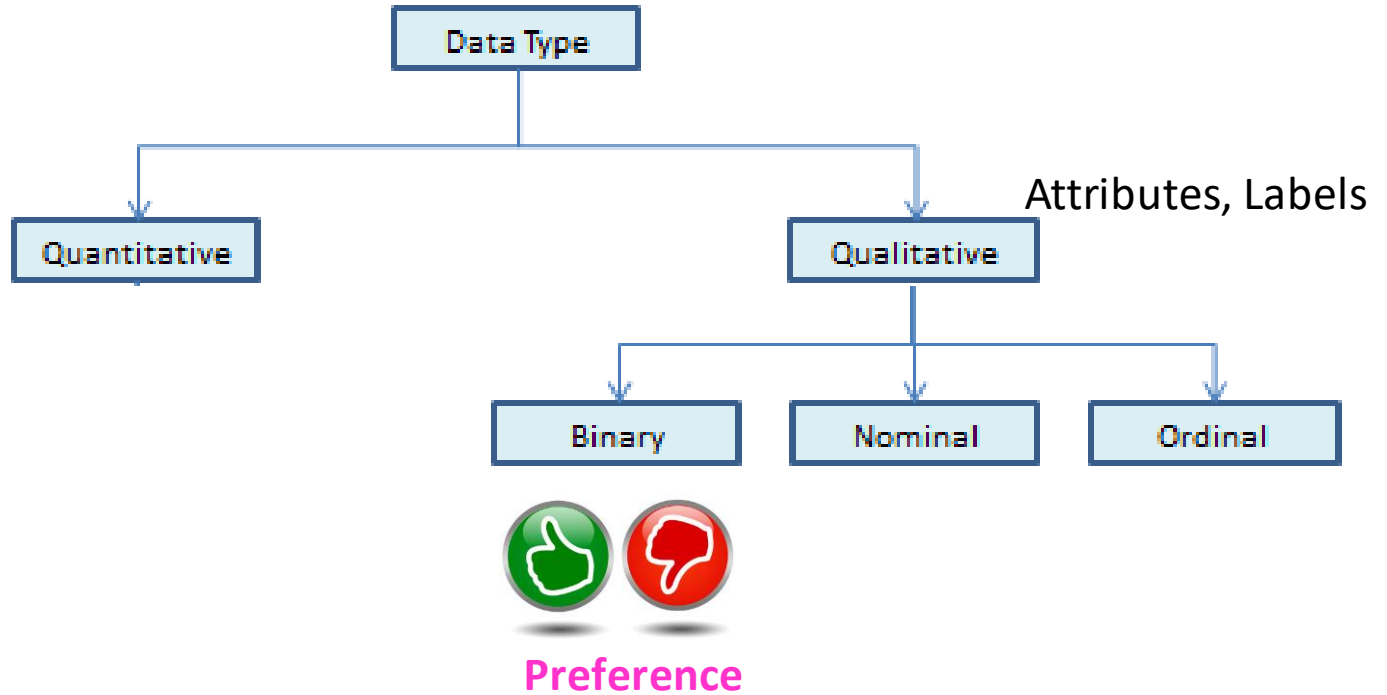




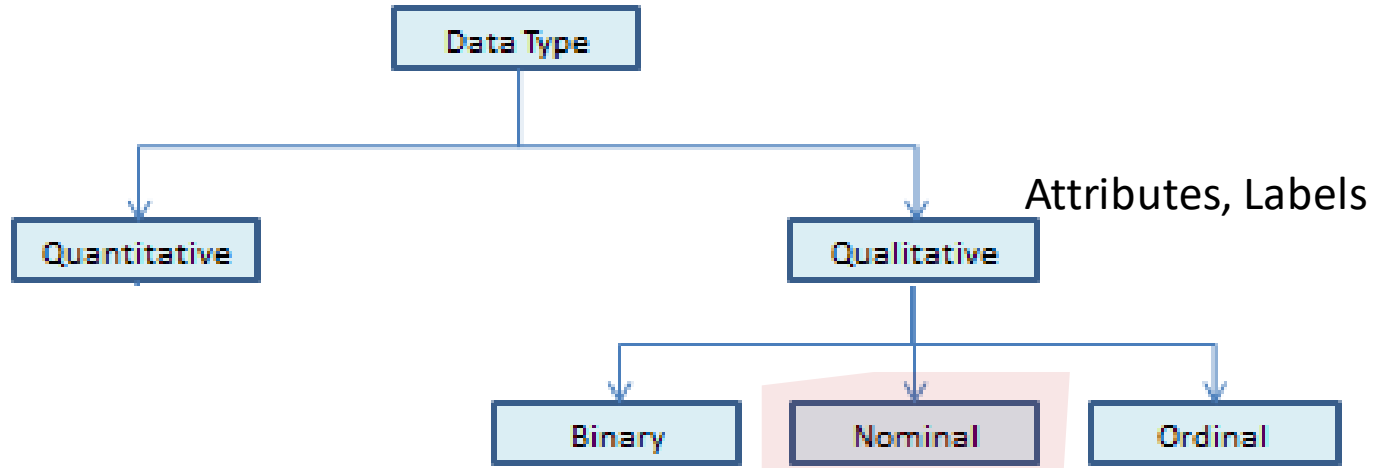
# Taxonomy of data variables



# Taxonomy of data



# Taxonomy of data



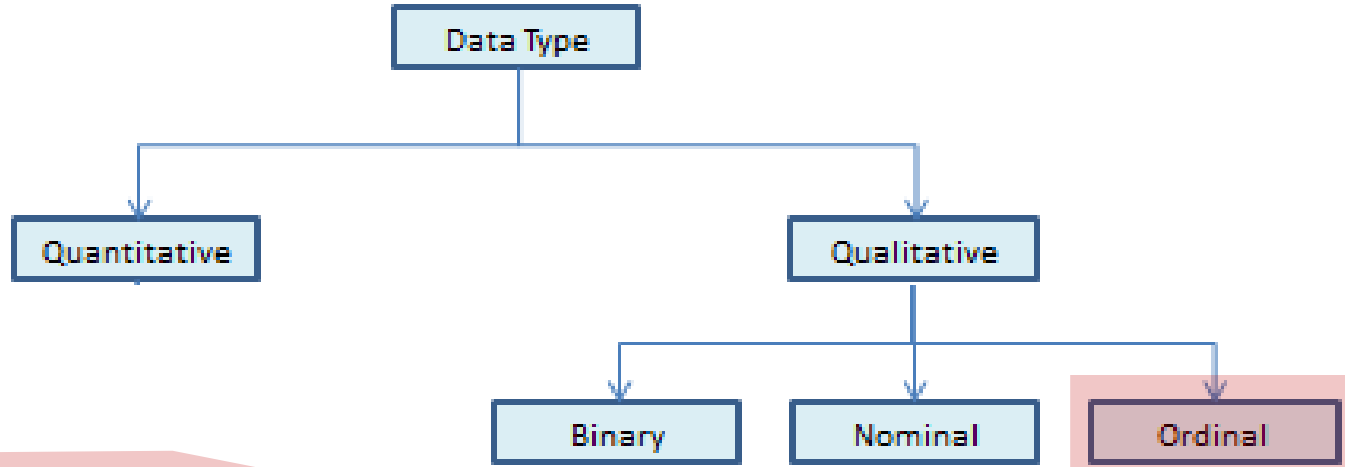
Color



Make



Pin Code



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



1

POINTS - 96,817



Akane  
YAMAGUCHI



2

POINTS - 84,963



PUSARLA  
V. Sindhu



3

POINTS - 83,414



Ratchanok  
INTANON



4

POINTS - 77,487



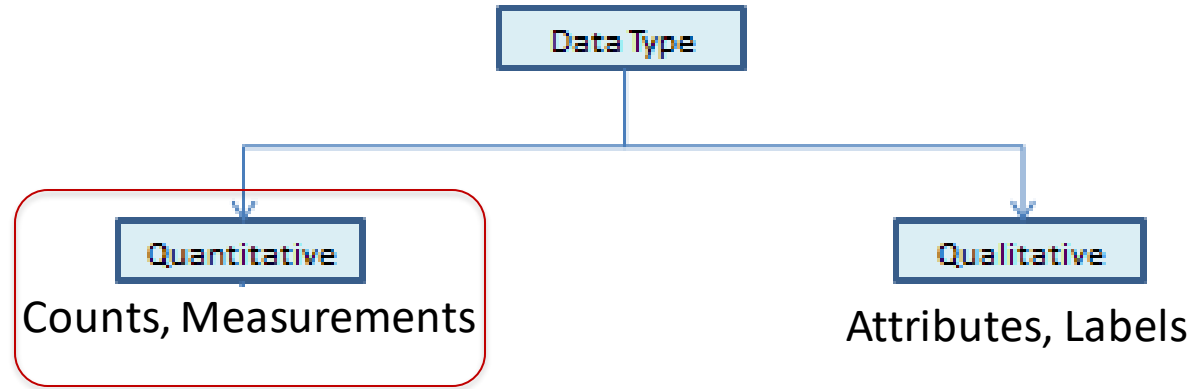
CHEN  
Yufei



5

POINTS - 74,889

# Taxonomy of data



## QUANTITATIVE DATA:



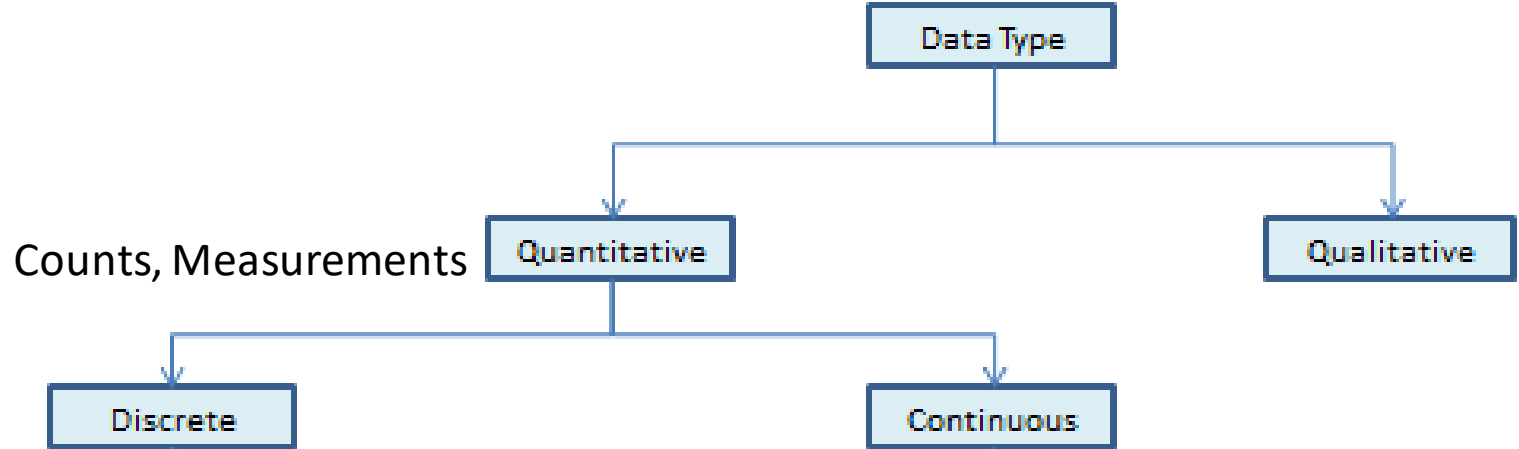
### Discrete data:

- There are 3 cones
- Cone 1 has 2 scoops

### Continuous data:

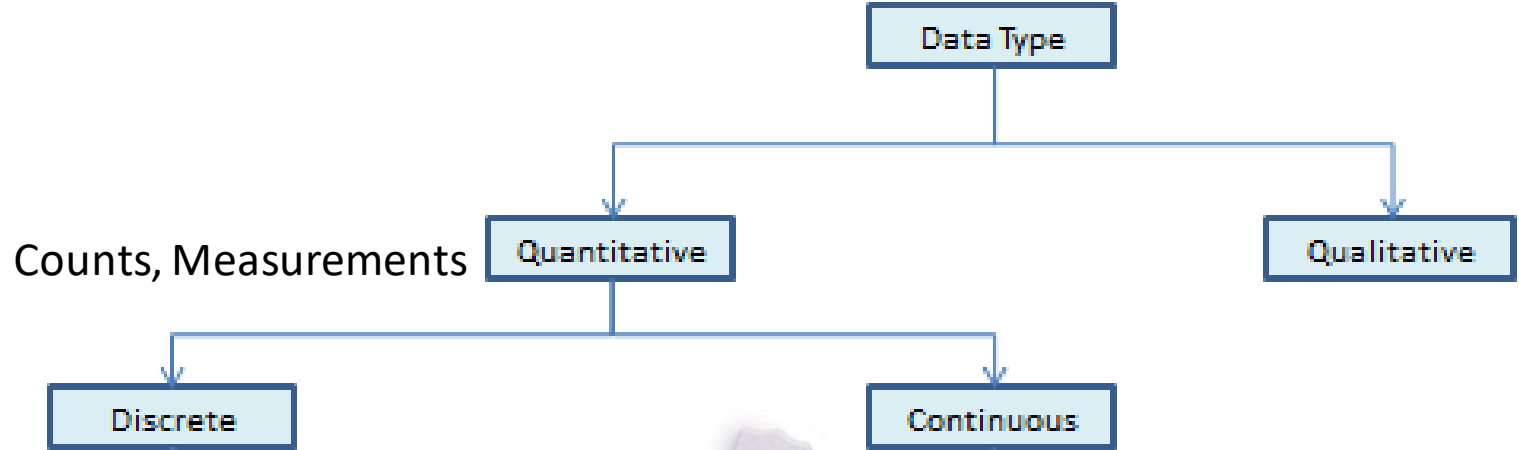
- Cone 3 weighs 79.4 grams
- cone 2 ice cream is at 8.3°F

# Taxonomy of data



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

# Taxonomy of data



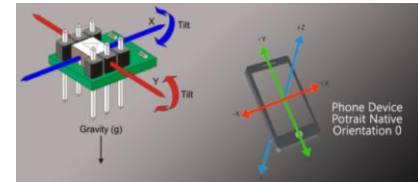
Counts, Measurements

Quantitative

Qualitative

Discrete

Continuous



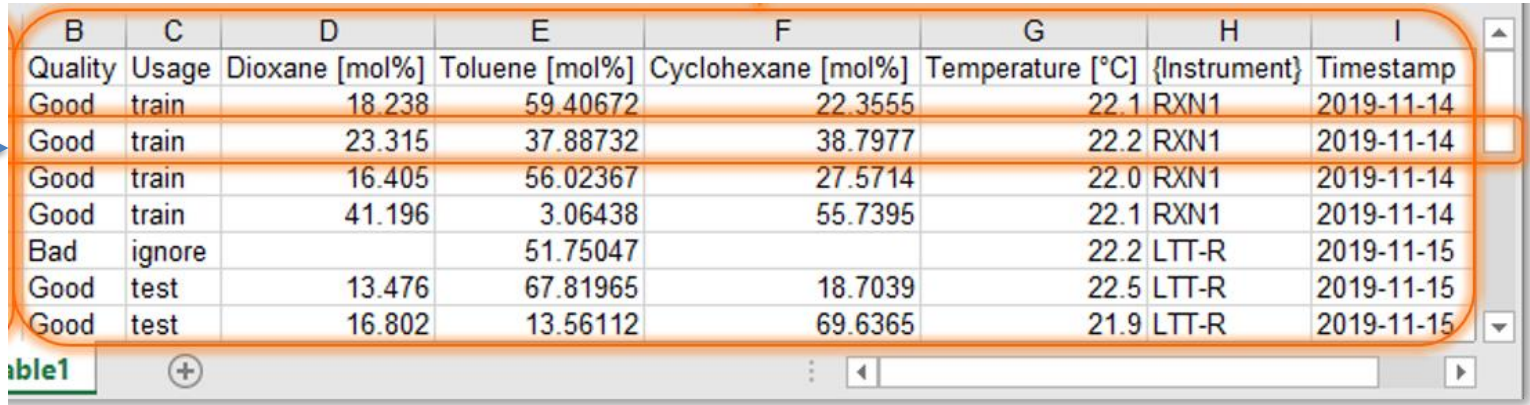
# Samples and Features

Feature / Attribute



Dataset

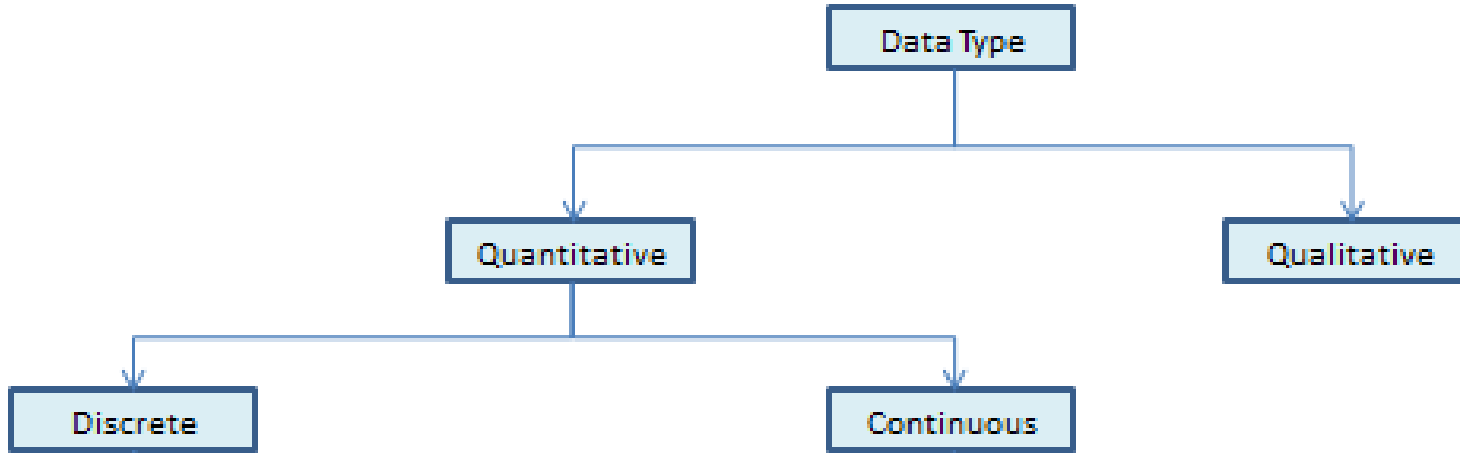
Sample



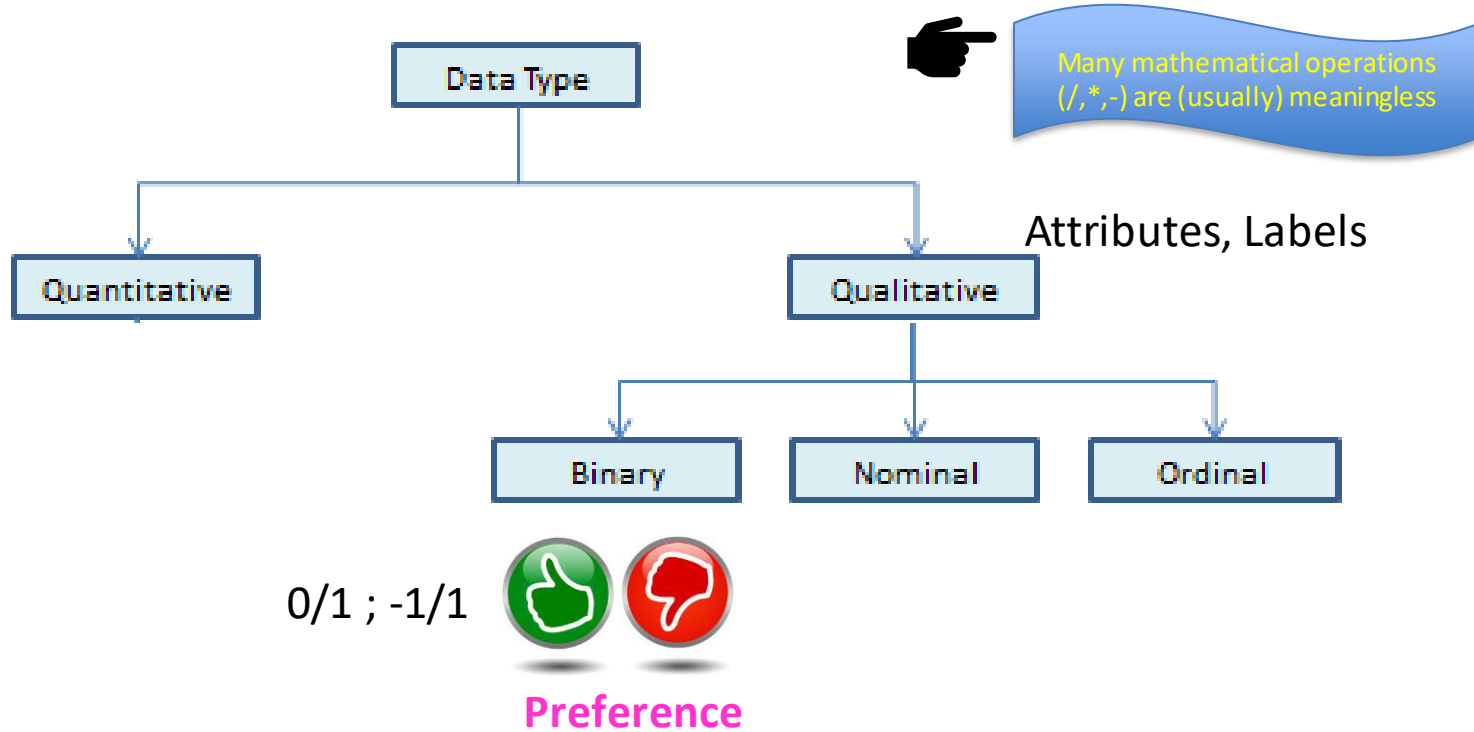
B	C	D	E	F	G	H	I
Quality	Usage	Dioxane [mol%]	Toluene [mol%]	Cyclohexane [mol%]	Temperature [°C]	{Instrument}	Timestamp
Good	train	18.238	59.40672	22.3555	22.1	RXN1	2019-11-14
Good	train	23.315	37.88732	38.7977	22.2	RXN1	2019-11-14
Good	train	16.405	56.02367	27.5714	22.0	RXN1	2019-11-14
Good	train	41.196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



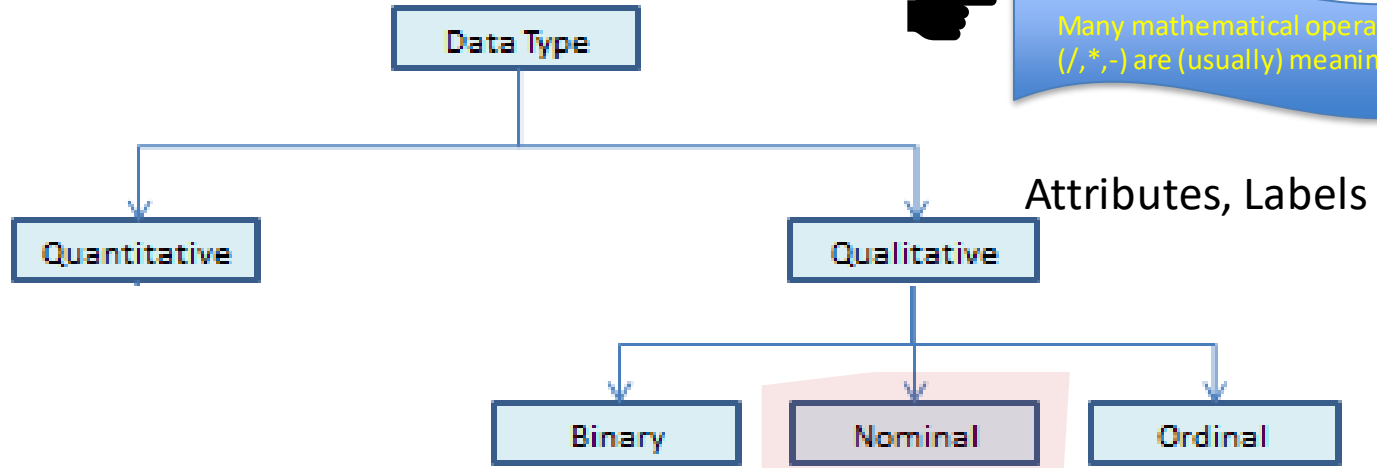
# Ultimately, all data needs to be quantitative



# Taxonomy of data: Qualitative → Quantitative



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

# Numerical encoding of categorical variables

Rome = [1, 0, 0, 0, 0, 0, ..., 0]

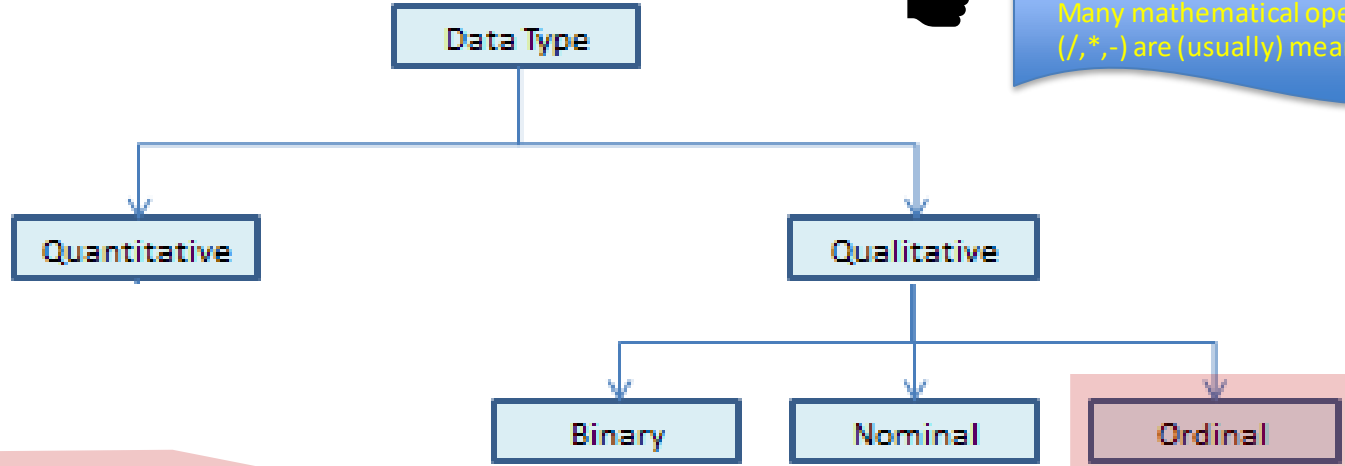
Paris = [0, 1, 0, 0, 0, 0, ..., 0]

Italy = [0, 0, 1, 0, 0, 0, ..., 0]

France = [0, 0, 0, 1, 0, 0, ..., 0]



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



How comfortable are you with Python \*

-2                      +1

No knowledge      ○   ○   ○   ○   ○   ○   Very comfortable

XS   S   M   L   XL   XXL

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

1                      2                      3                      4                      5

CURRENT WORLD RANKINGS

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	<b>Chicago</b>	IL	<b>60608</b>	 Conflicts
t2	John Veliotis Sr.	Johnnyo's	3465 S Morgan ST	Chicago	IL	<b>60609</b>	
t3	John Veliotis Sr.	Johnnyo's	3465 S Morgan ST	Chicago	IL	<b>60609</b>	
t4	<b>Johnnyo's</b>	Johnnyo's	3465 S Morgan ST	<b>Cicago</b>	IL	60608	

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

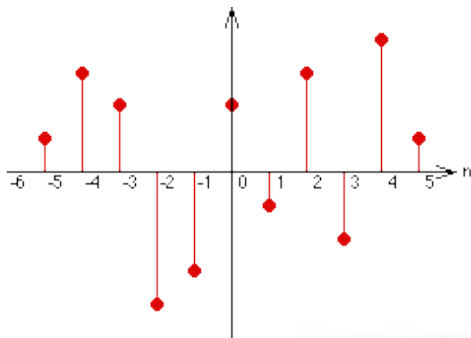
Label

Feature / Attribute

The diagram illustrates a data table with annotations. A red box labeled 'Label' has a red arrow pointing to the 'Quality' column (column B). A blue box labeled 'Feature / Attribute' has a blue arrow pointing to the 'Usage' column (column C). A blue arrow labeled 'Sample' points to the first row of the table. The table itself is outlined in orange and contains the following data:

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

# Data Sample Representations



Scalars

$X$

Vectors

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

Matrix

$$X = \begin{bmatrix} x & \dots & x_N \end{bmatrix} = \begin{bmatrix} x_{1,1} & \dots & x_{N,1} \\ \vdots & \dots & \vdots \\ x_{1,M} & \dots & x_{N,M} \end{bmatrix}$$

$2^{nd} \text{ dimension}$

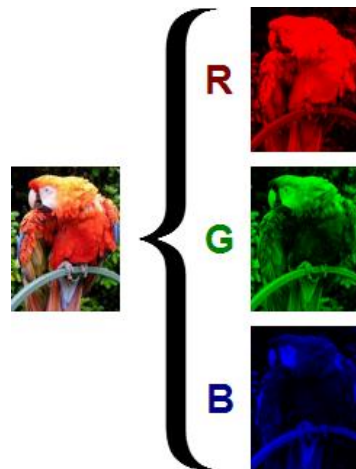
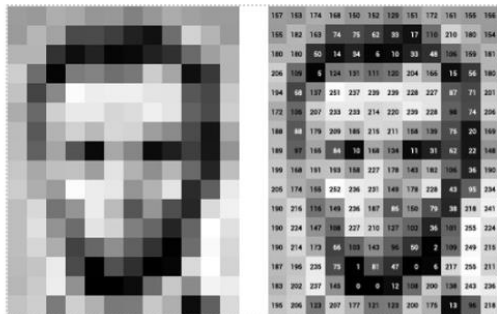
Tensor

$$X = \{X_1, \dots, X_k\} = \begin{bmatrix} x_{1,1,1} & \dots & x_{N,1,1} \\ \vdots & \dots & \vdots \\ x_{1,M,1} & \dots & x_{N,M,1} \end{bmatrix} \dots \begin{bmatrix} x_{1,1,k} & \dots & x_{N,1,k} \\ \vdots & \dots & \vdots \\ x_{1,M,k} & \dots & x_{N,M,k} \end{bmatrix}$$

$2^{nd} \text{ dimension}$



2-d image



# Data Representations



Graph Representation

Vertex List

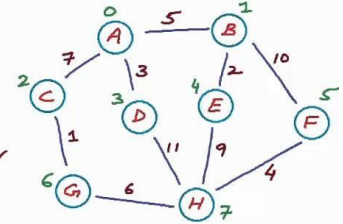
	0	1	2	3	4	5	6	7
0	A							
1	B							
2	C							
3	D							
4	E							
5	F							
6	G							
7	H							
	↓							

Adjacency Matrix

	0	1	2	3	4	5	6	7
0	∞	5	7	3	∞	∞	∞	∞
1	5	∞	∞	∞	2	10	∞	∞
2	7	∞	∞	∞	∞	∞	1	∞
3	3	∞	∞	∞	∞	∞	∞	11
4	∞	2	∞	∞	∞	∞	∞	9
5	∞	10	∞	∞	∞	∞	∞	4
6	∞	∞	1	∞	∞	∞	∞	6
7	∞	∞	∞	6	11	9	4	∞

A

$$|V| = v$$





# 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.

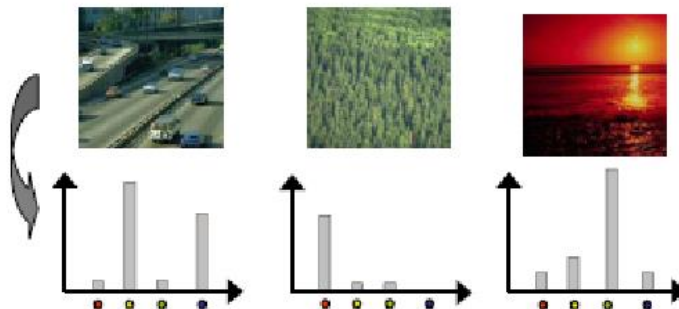
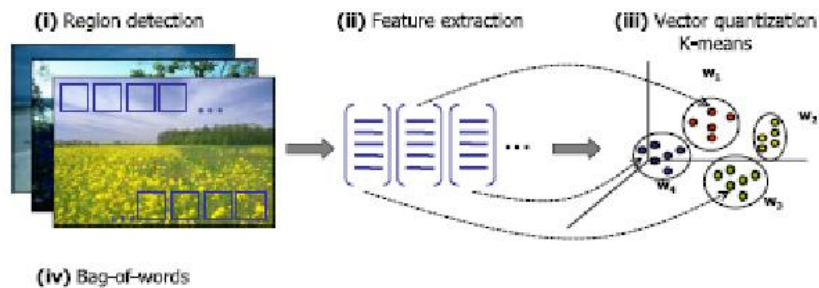
# The Bag of Words Representation

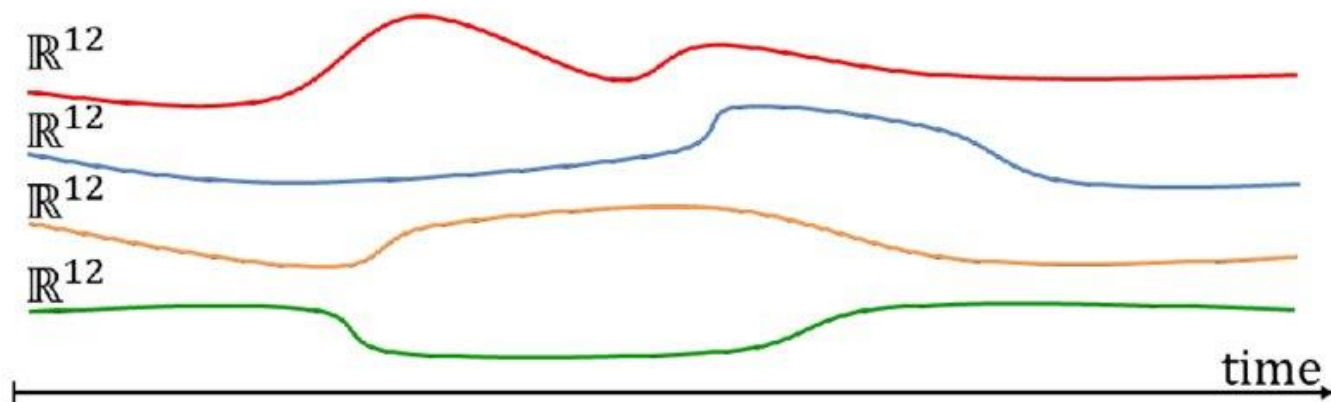
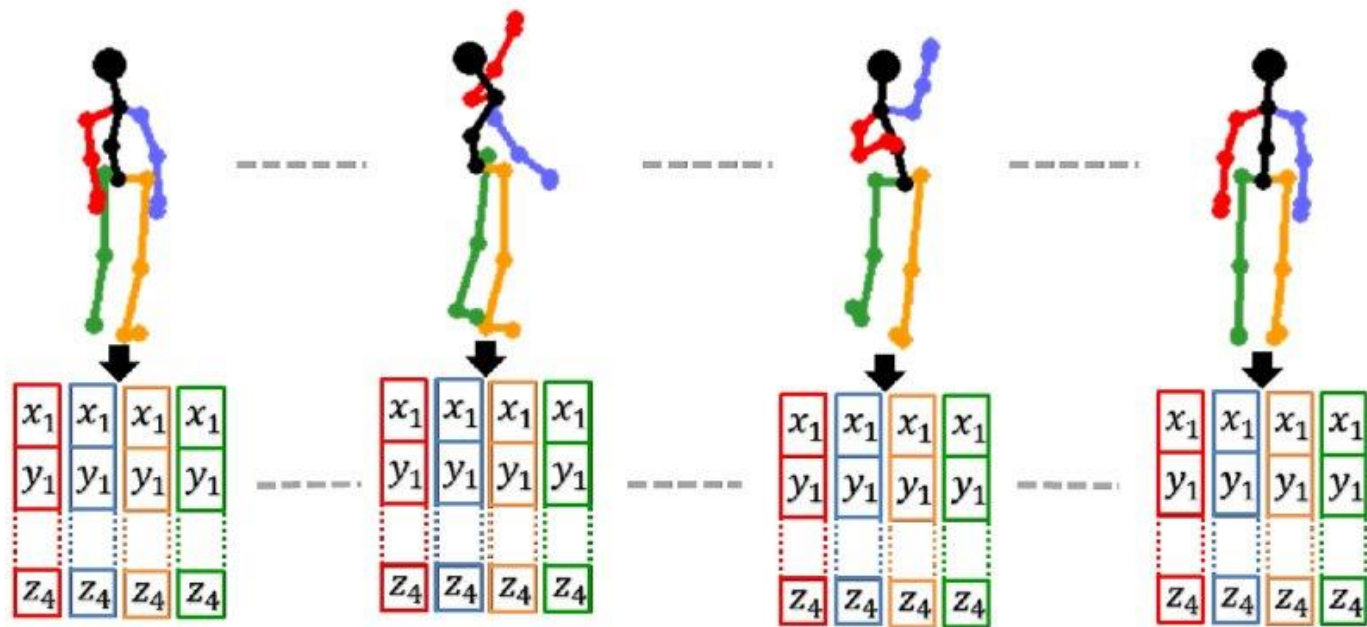
I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy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

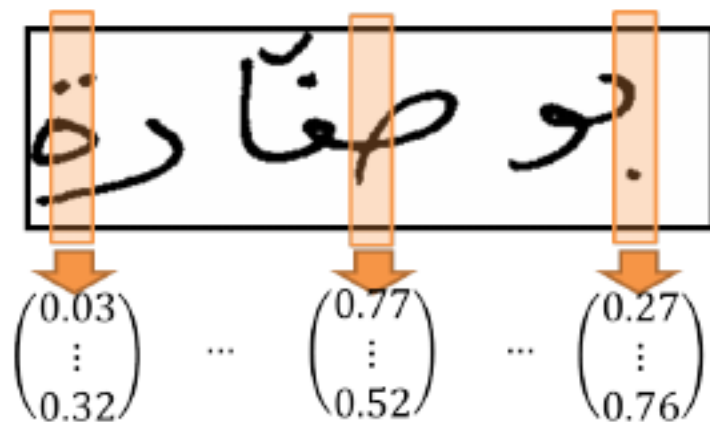
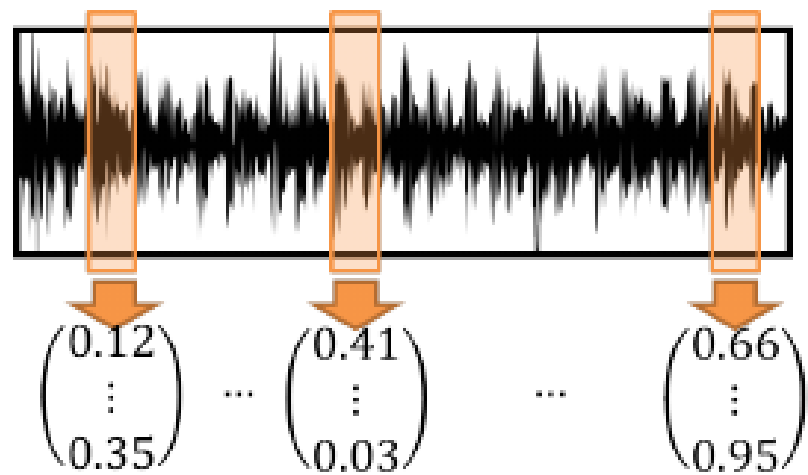


it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
...	...

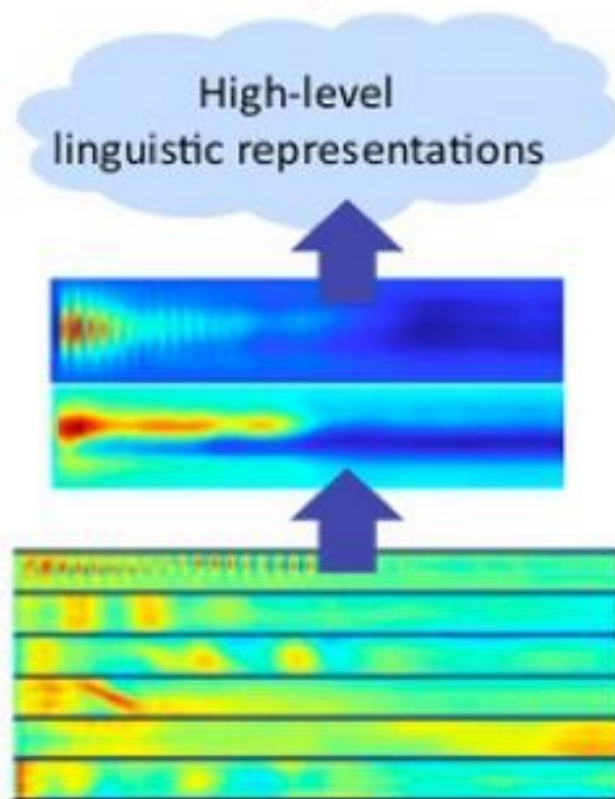
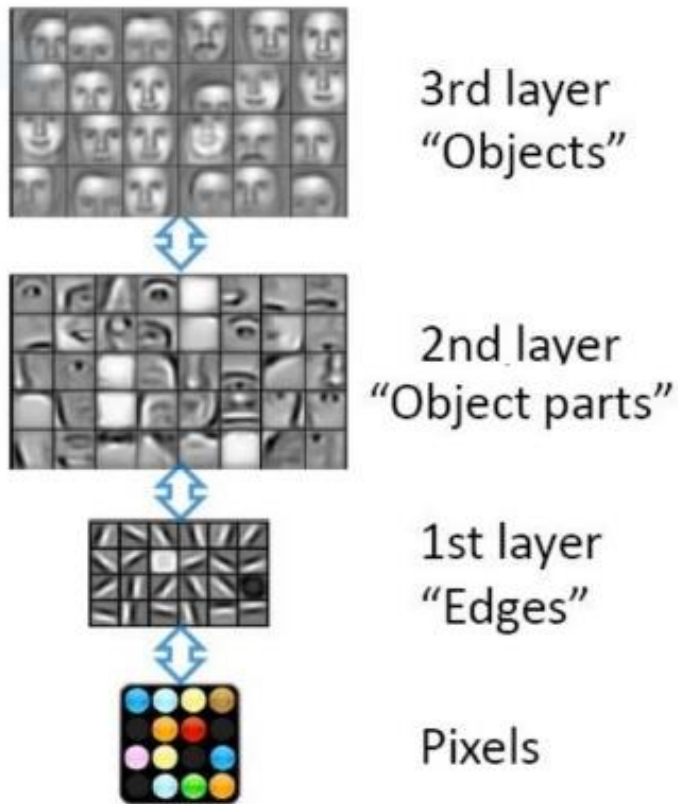
15







# 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.)
  - $X = \{\text{apple, grape}\}$
  - $b_i \in [1,5]$
- Data = Instantiation of some or all r.v.'s in the domain

# Data: a probabilistic perspective

## Output

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

Conflicts

Does not obey data distribution

Conflict



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	<b>60608</b>
t3	John Veliotis Sr.	3465 S Morgan ST	Chicago	IL	<b>60608</b>
t4	<b>John Veliotis Sr.</b>	3465 S Morgan ST	<b>Chicago</b>	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)

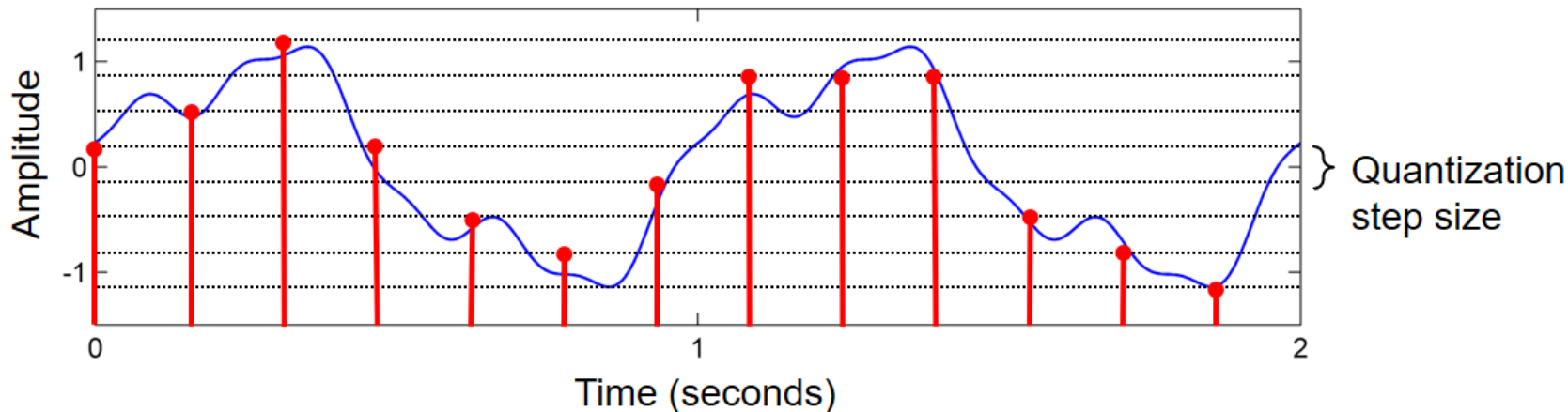


# Lecture Outline

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

# Quantization

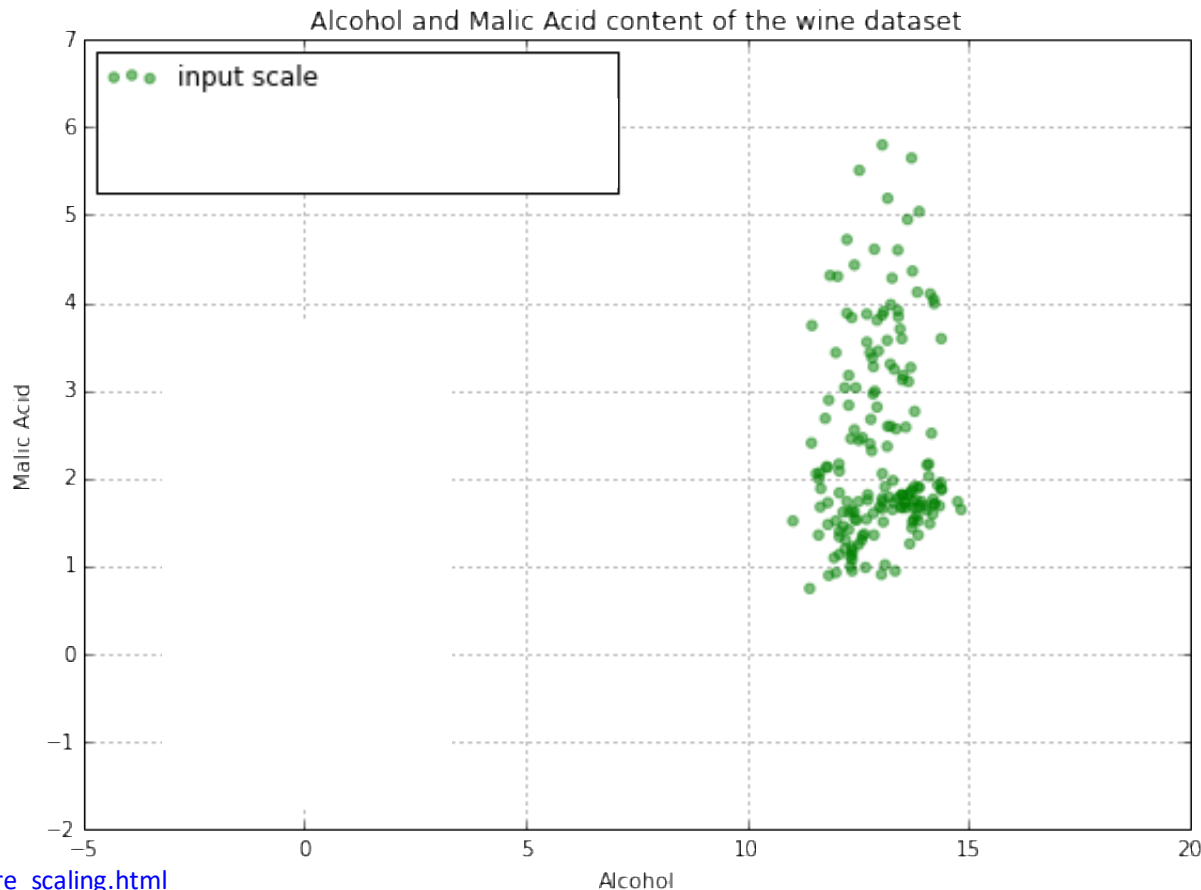
1. Continuous  $\rightarrow$  Discrete ('Rounding off')



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



# Popular normalization approaches

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

MinMax Scaling

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

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

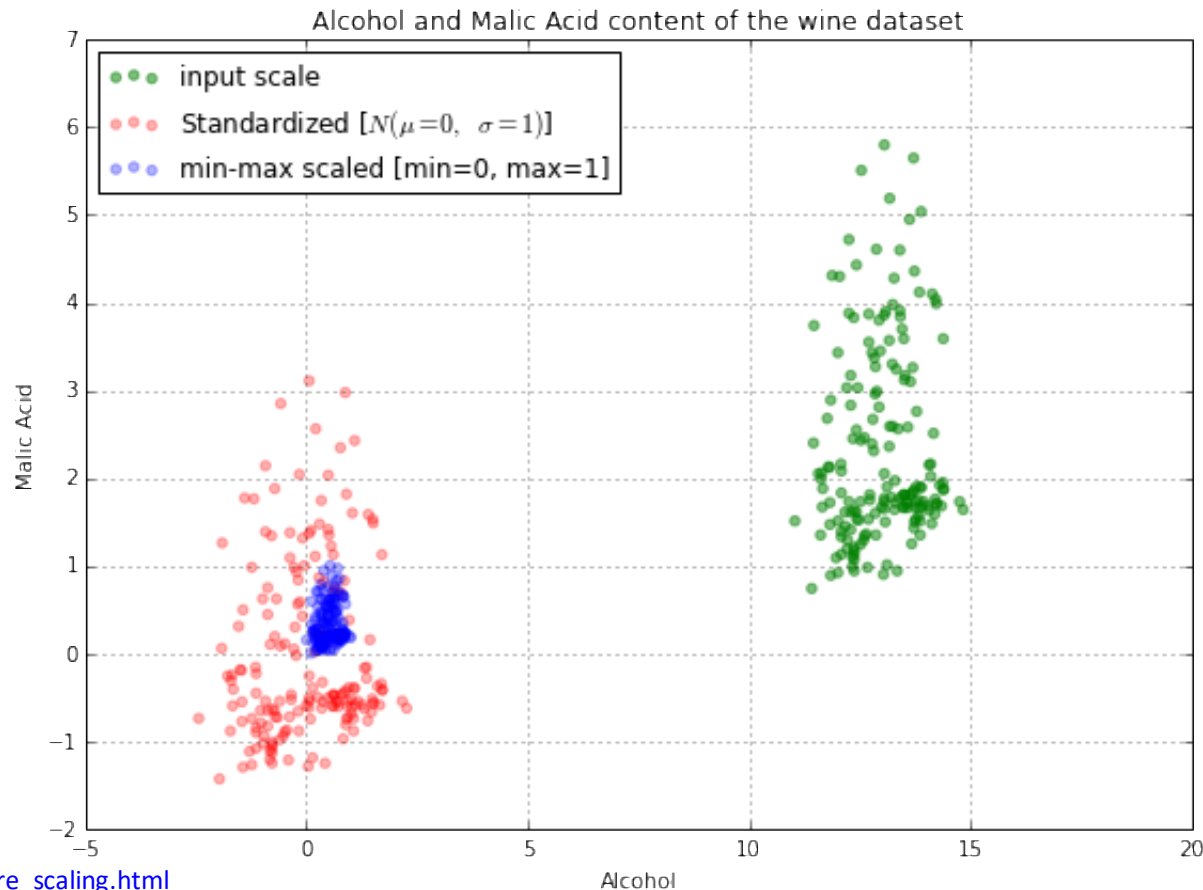
$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2}$$

Standardization  
(Unit Normal Scaling)

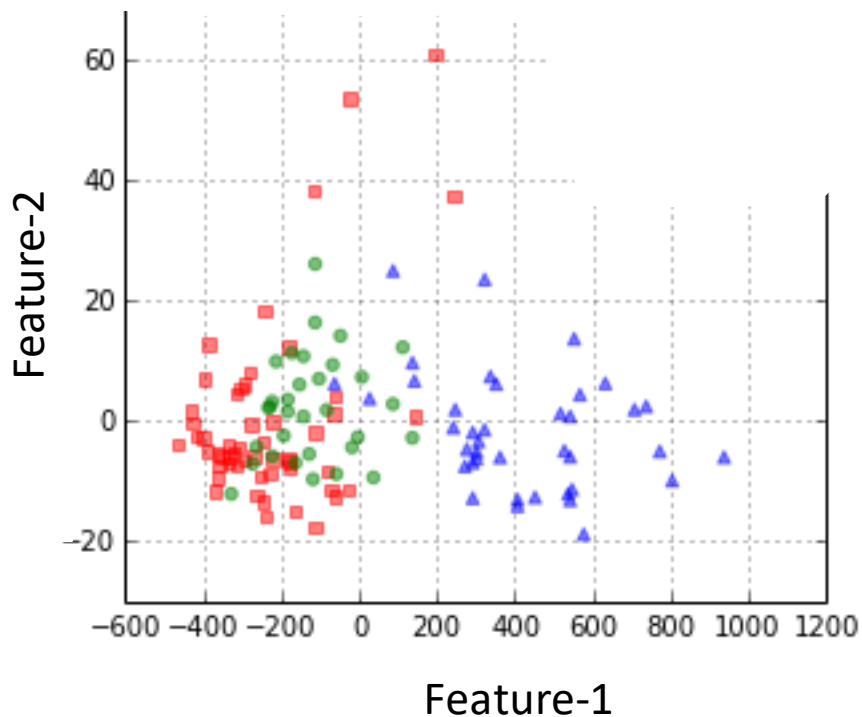
	Class label	Alcohol	Malic acid
0	1	14.23	1.71
1	1	13.20	1.78
2	1	13.16	2.36
3	1	14.37	1.95
4	1	13.24	2.59

# Data Normalization (applied to each feature)

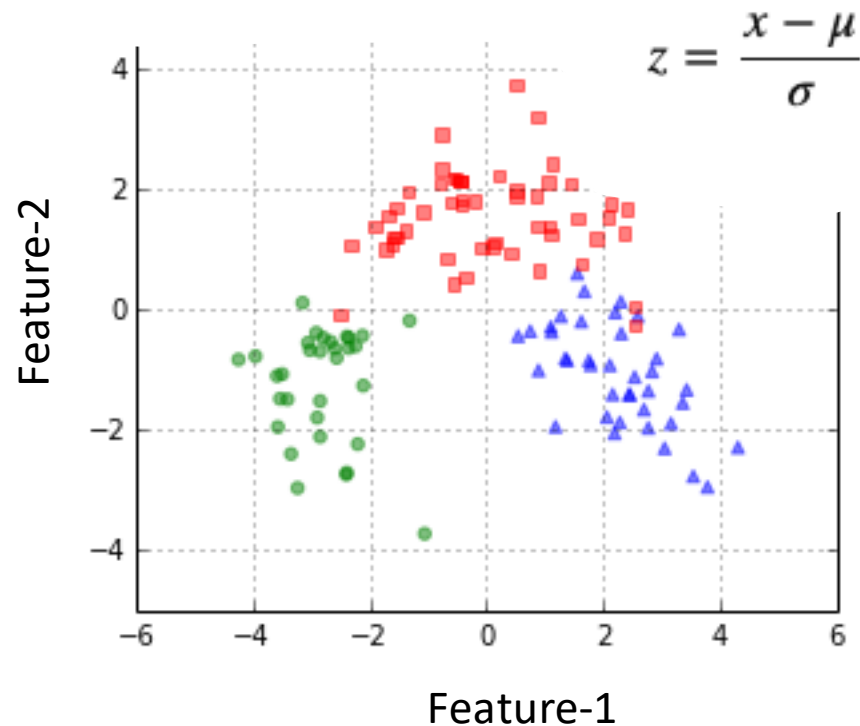
	Class label	Alcohol	Malic acid
0	1	14.23	1.71
1	1	13.20	1.78
2	1	13.16	2.36
3	1	14.37	1.95
4	1	13.24	2.59



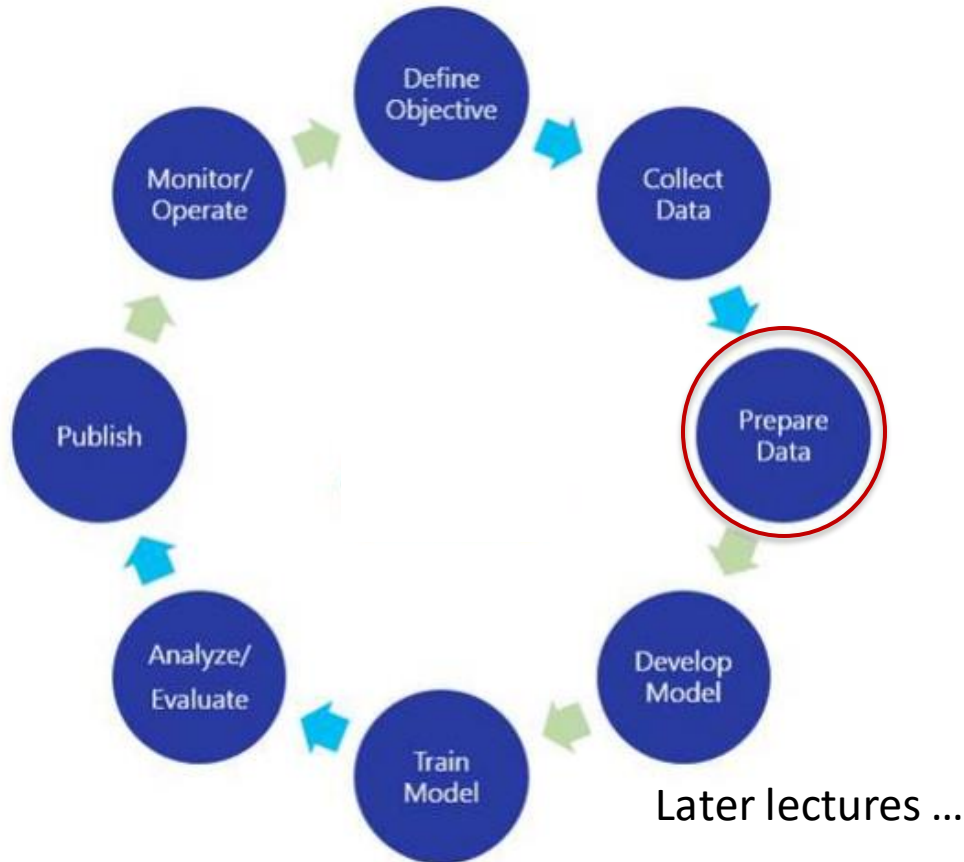
Before standardization



After standardization



# Workflow of a Machine Learning Problem



# Lecture Outline

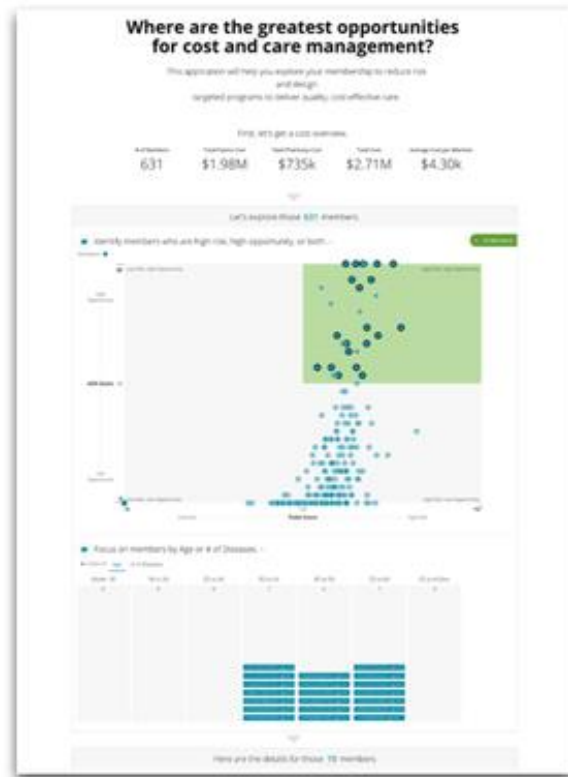
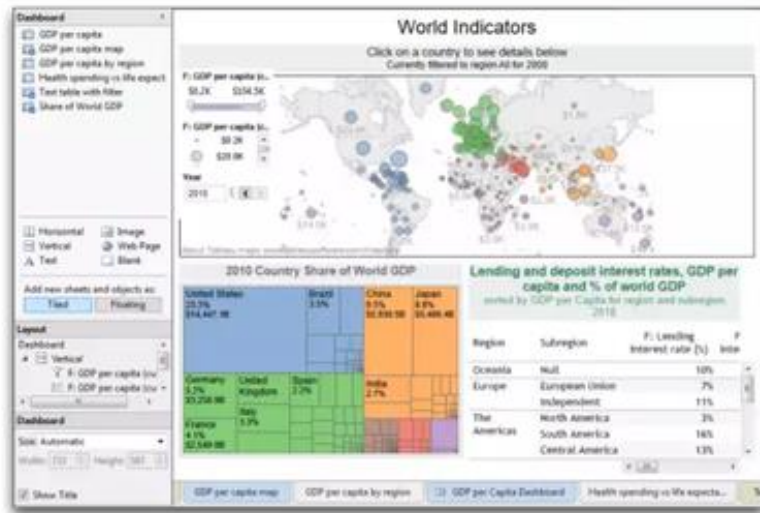
- *ML Workflow*
- *Data Representations*
- *Basic Data Transformations*
- **Data Visualization**

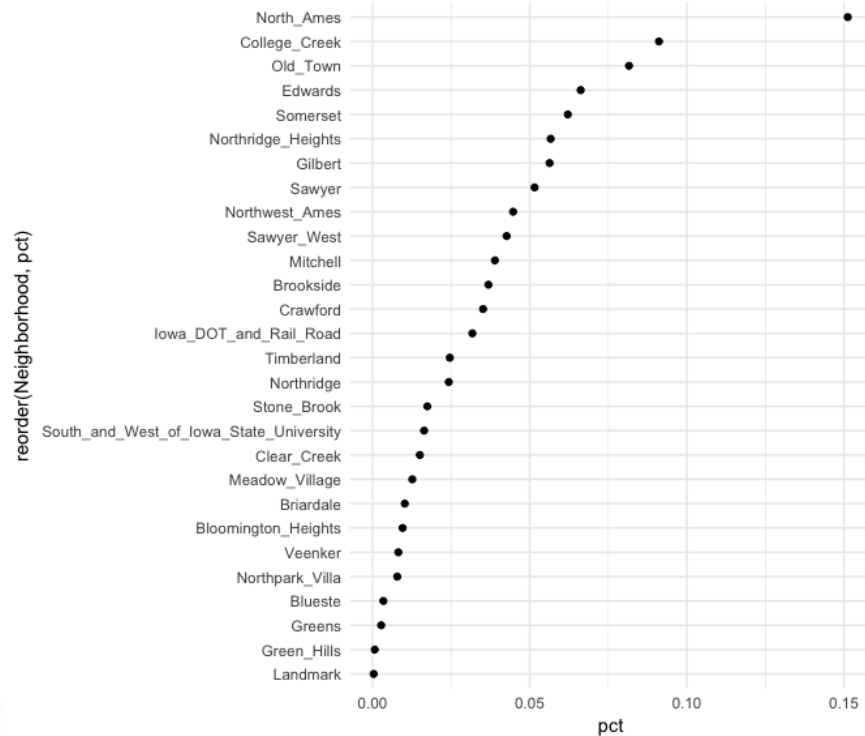
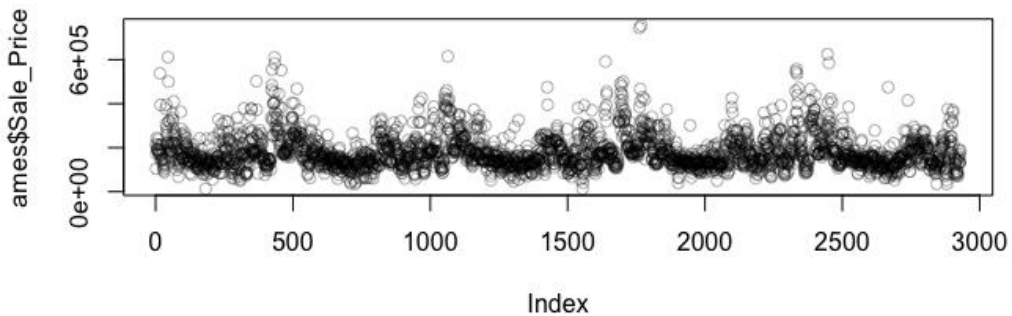
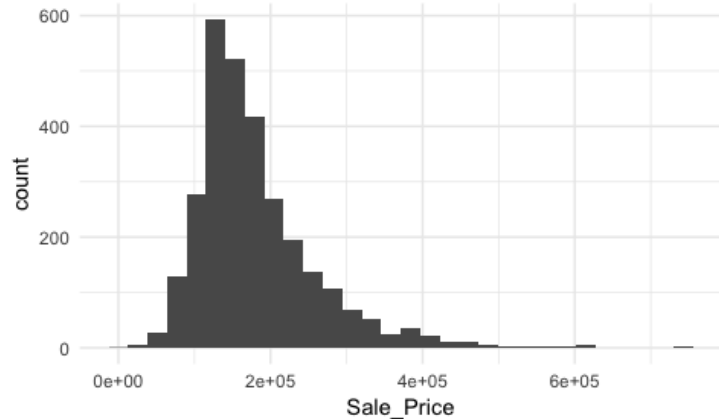


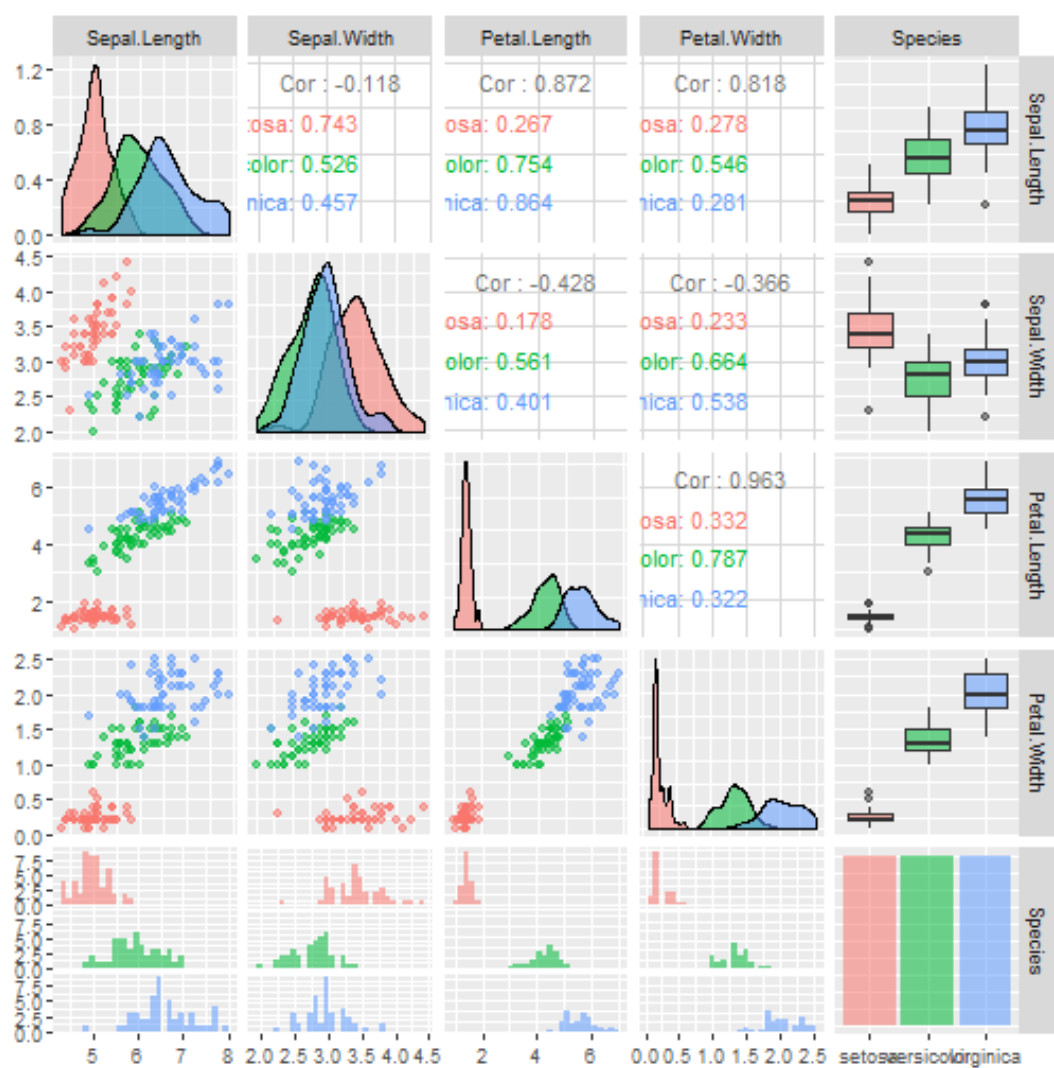
# Gazing at Data: Data visualization

## data exploration

## data presentation

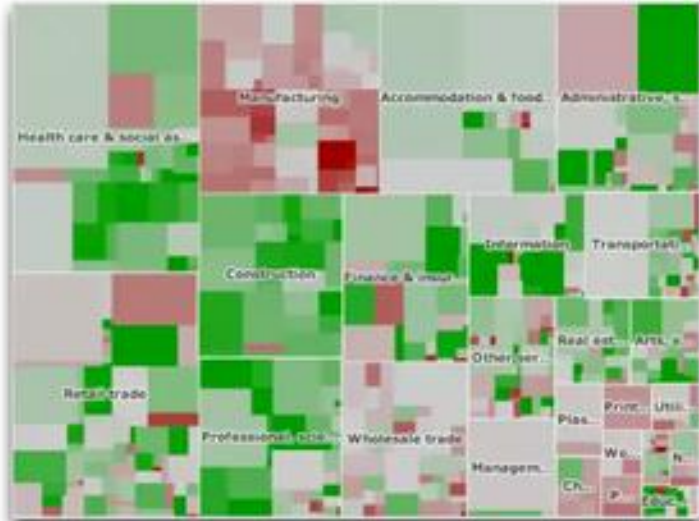






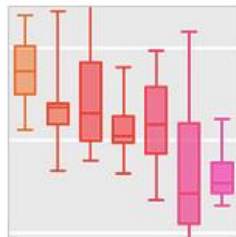
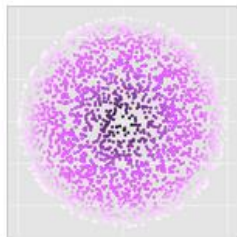
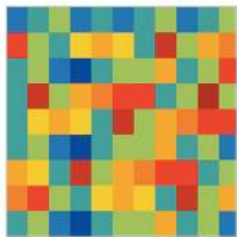
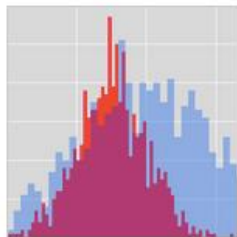
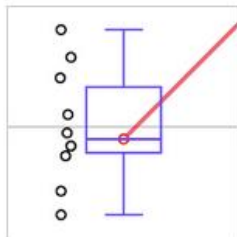
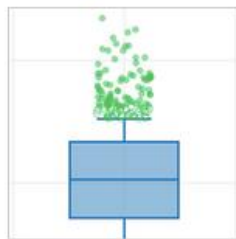
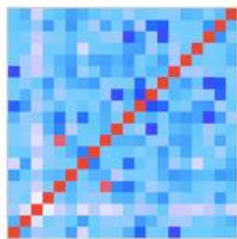
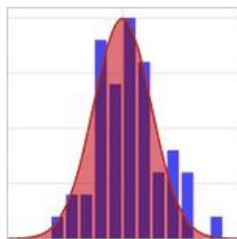
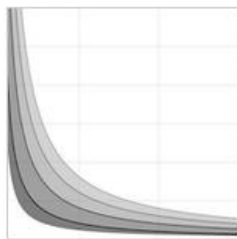
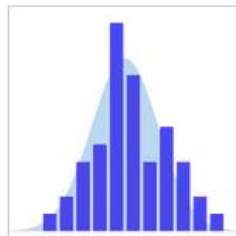
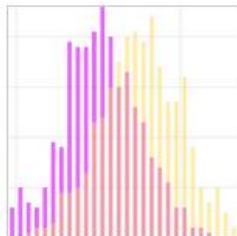
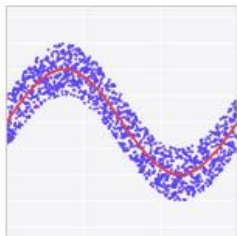
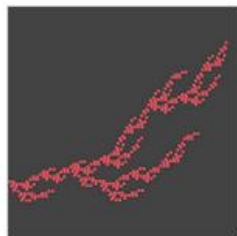
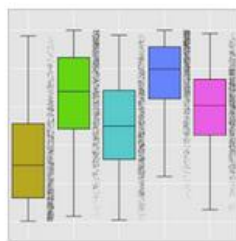
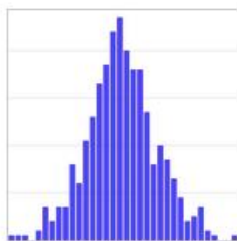
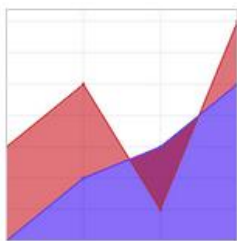
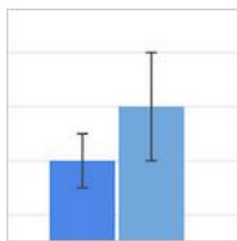
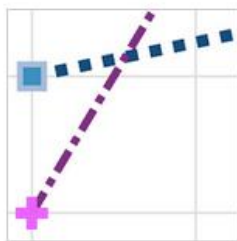
# Data visualization

## treemap



## leaderboard

SHUTTLE		40 YARD		BENCH PRESS		VERT LEAP (in)		BROAD JUMP (in)	
Jordan Jaffer...	4.06	1st Robert Griffin	4.41	Jordan Jaffer...	14	1st 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 Jaffer...	4.65	Robert Griffin	---	Jordan Jaffer...	37	1st Robert Griffin	120
Chandler Han...	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 Jaffer...	116
Darron Thomas	4.28	Jacory Harris	4.72	Aaron Corp	---	Russell Wilson	34	Jacory Harris	113
Aaron Corp	4.30	Chandler Han...	4.76	Jacory Harris	---	Chandler Han...	33	Tyler Hansen	113
Patrick Witt	4.37	Tyler Hansen	4.78	Chandler Han...	---	Capt Keshum	33	Chandler Han...	112
B.J. Coleman	4.38	Darron Thomas	4.80	Tyler Hansen	---	Aaron Corp	32	Nick Foles	112
Jacory Harris	4.40	Capt Keshum	4.82	Capt Keshum	---	Patrick Witt	32	Austin Davis	109

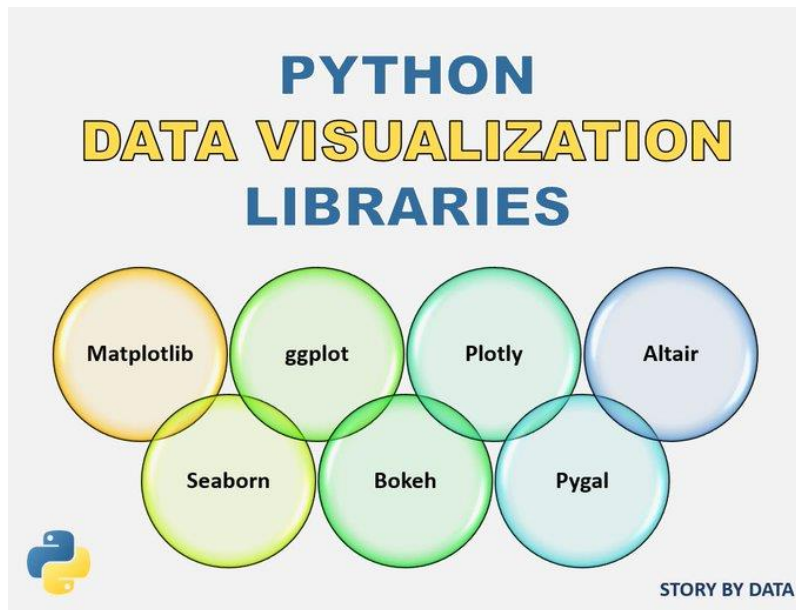


// In good information  
visualization, there are  
no rules, no guidelines,  
no templates, no  
standard technologies,  
no stylebooks ... You  
must simply do  
whatever it takes. //

—Edward Tufte

# Resources

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



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