

04.01.2019

Statistical Methods in AI (CSE/ECE 471)

Lecture-2: ML Workflow, Data Representations, Supervised Learning, Intro to Classification

Ravi Kiran

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 this Saturday. **Bring your laptops.**
- Ask questions.

Announcements

- Assignments – Questions involving equations/mathematical derivation
 - Write up in latex [overleaf.com] → submit
 - Write neatly on paper → scan as photo/pdf [camscanner] → submit
- TA office hours, locations will be announced shortly.

Lecture Outline

- ML Workflow
- Intro to Supervised Learning
 - Taxonomy
 - Data Representations
 - Models

Machine Learning



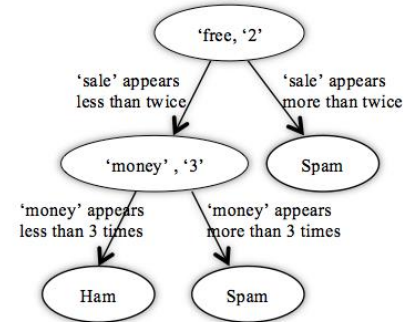
Algorithmic methods that use data to improve their knowledge of a task

Task: Detect spam email



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

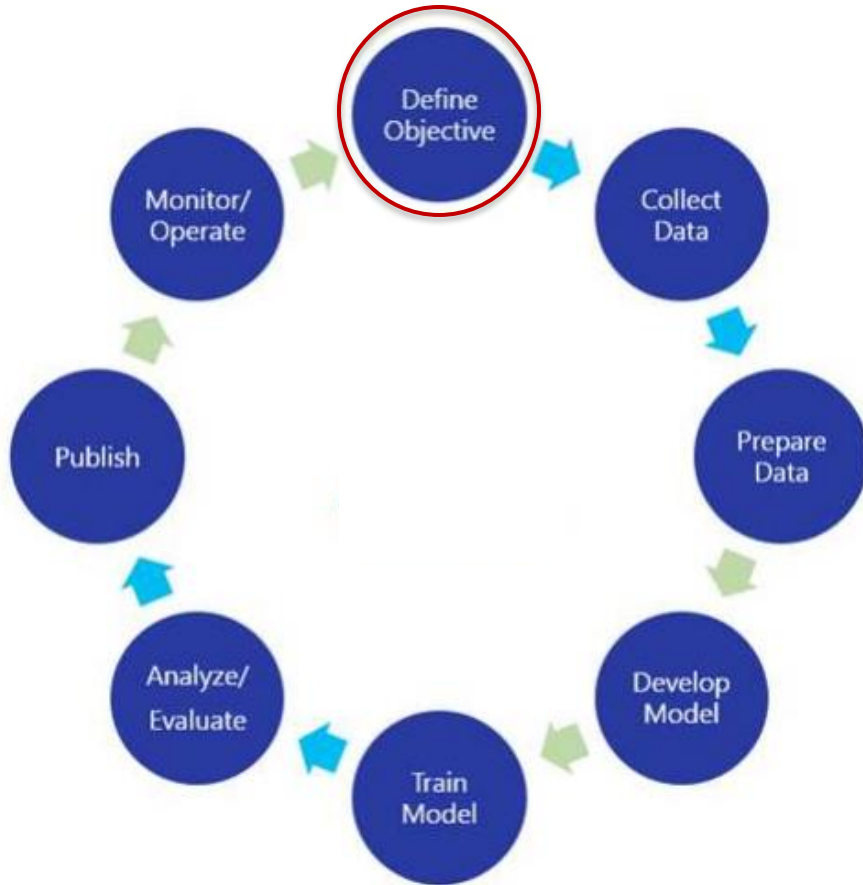
Knowledge:



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

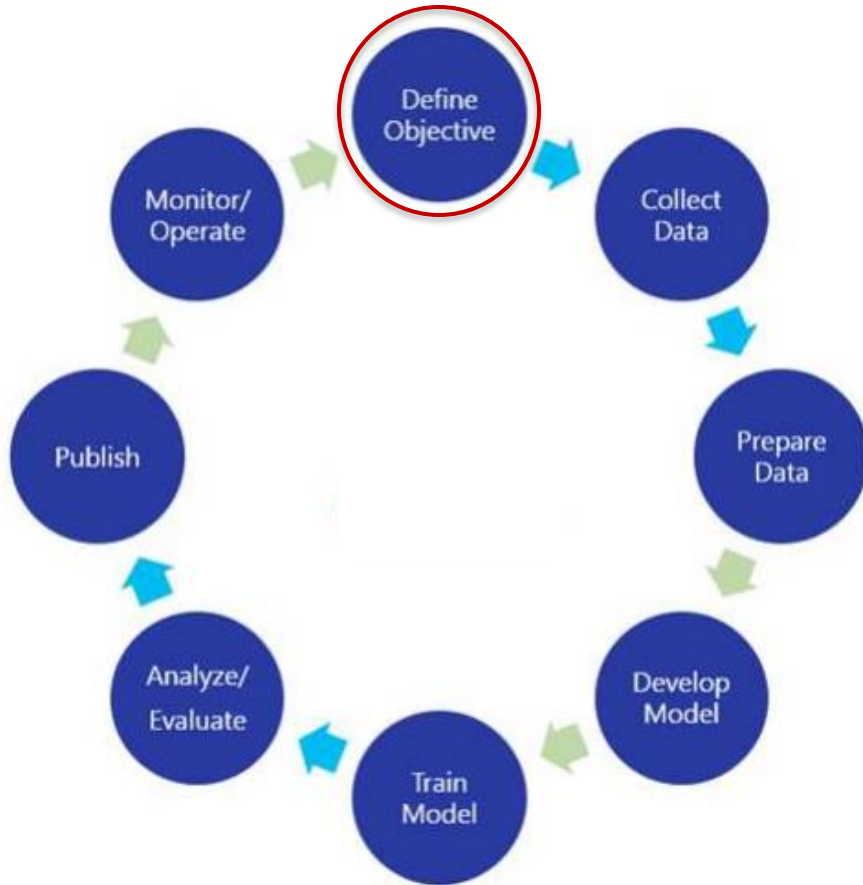
Algorithmic method: Decision Tree

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.

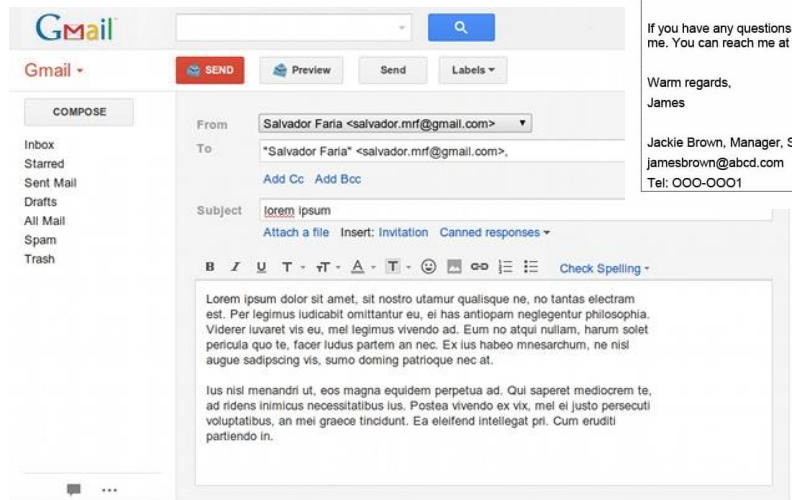
W. EDWARDS DEMING

"In God we trust, all others must bring data."

W. Edwards Deming

© iStockphoto / Thinkstock

- Detect spam email



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 000-0001.

Warm regards,
James

Jackie Brown, Manager, Staff
jamesbrown@abcd.com
Tel: 000-0001

Sources of data

- Predict value of a stock



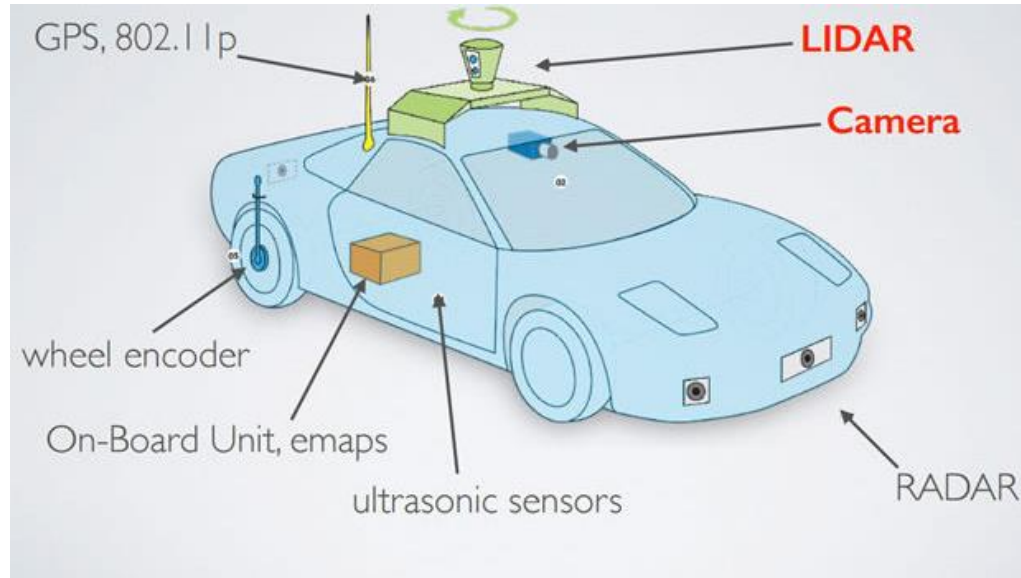
- Predict effect of advertising on sales

Restaurant & Coffee Shop		ایران کاظم و صفی
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

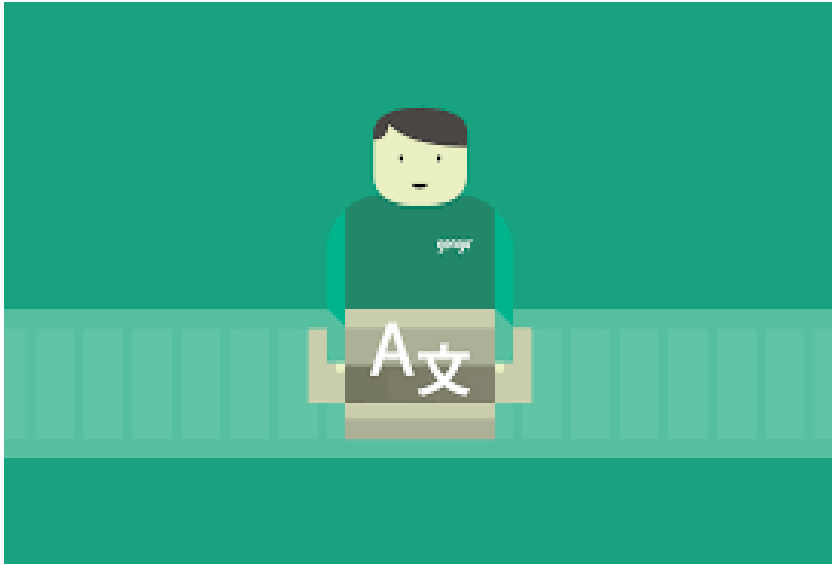
- Drive car safely without human intervention



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

Raw data

- Not all of it relevant



```
{
  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 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:

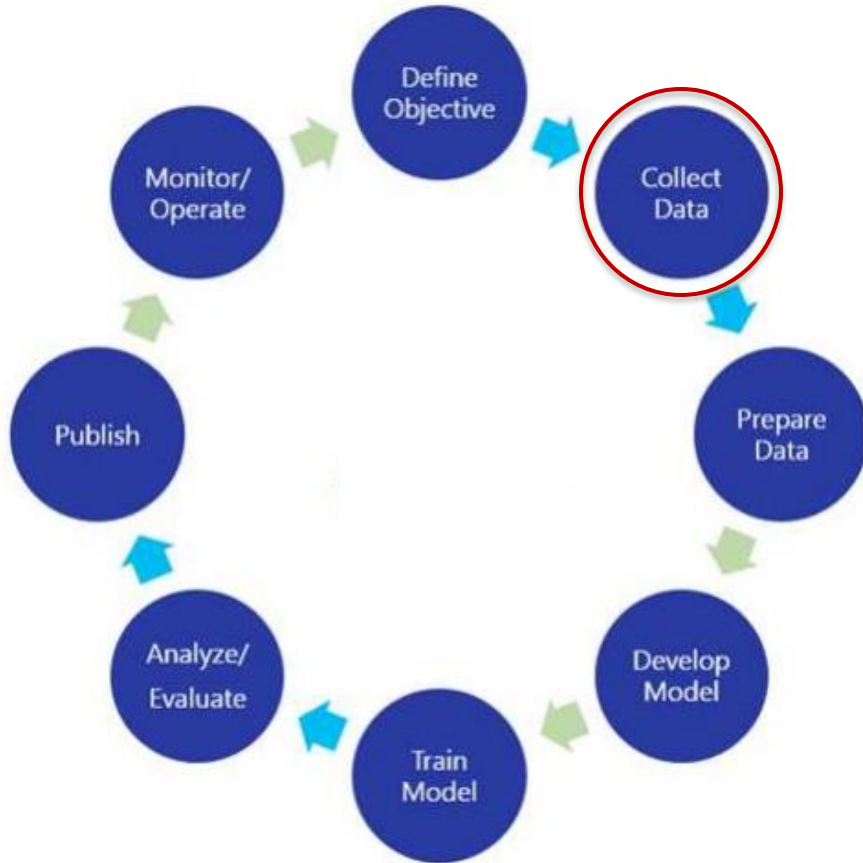
```
{
  Received: "by luna.mailgun.net with HTTP; Fri, 26 Feb 2016 20:12:03 +0000",
  stripped-signature: "",
  Message-Id: "<20160226201203.54979.26875@mailgun.com>",
  from: "Sample Email <me@mailgun.com>",
  sender: "me@mailgun.com",
  recipients: "anton@mailgunhq.com",
  Subject: "Test Message",
  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

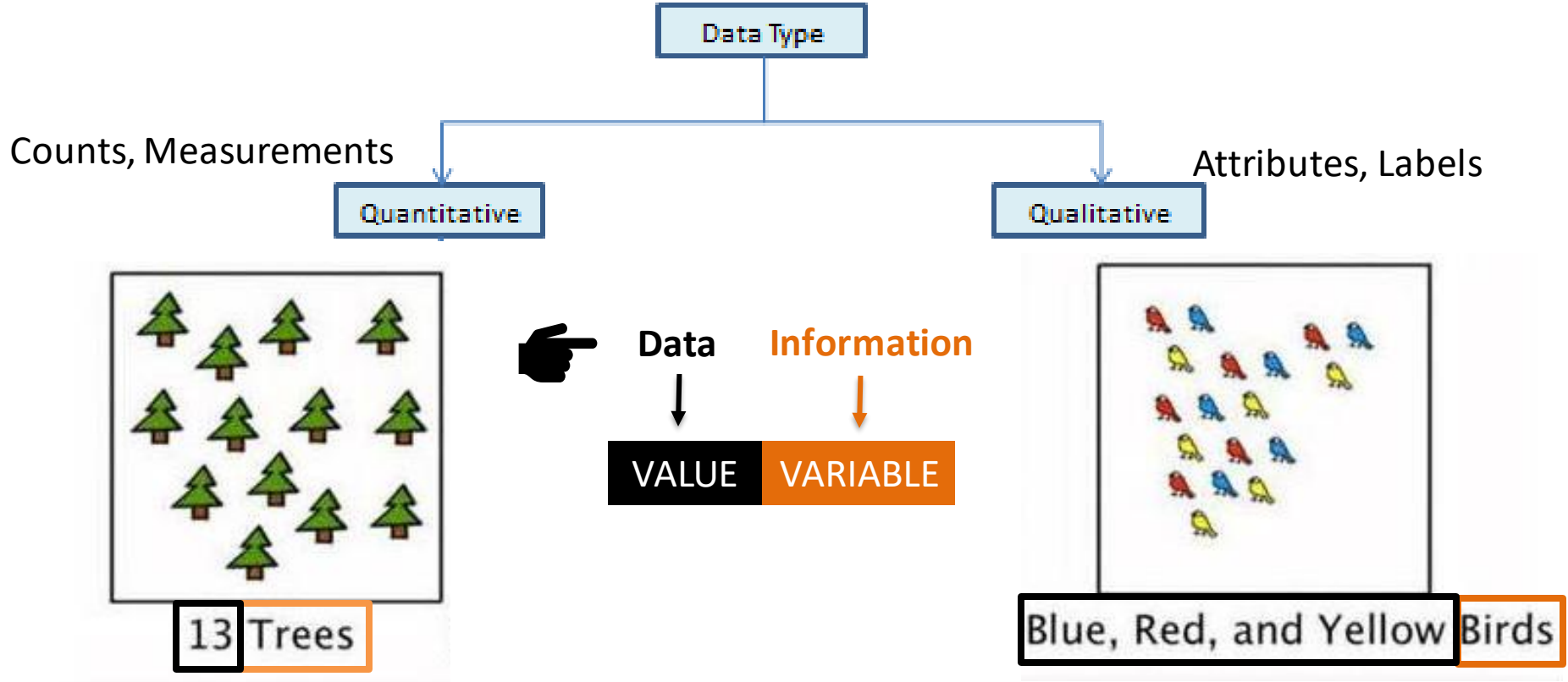
- May be too much in quantity
 - Limitations on system end (compute, storage)



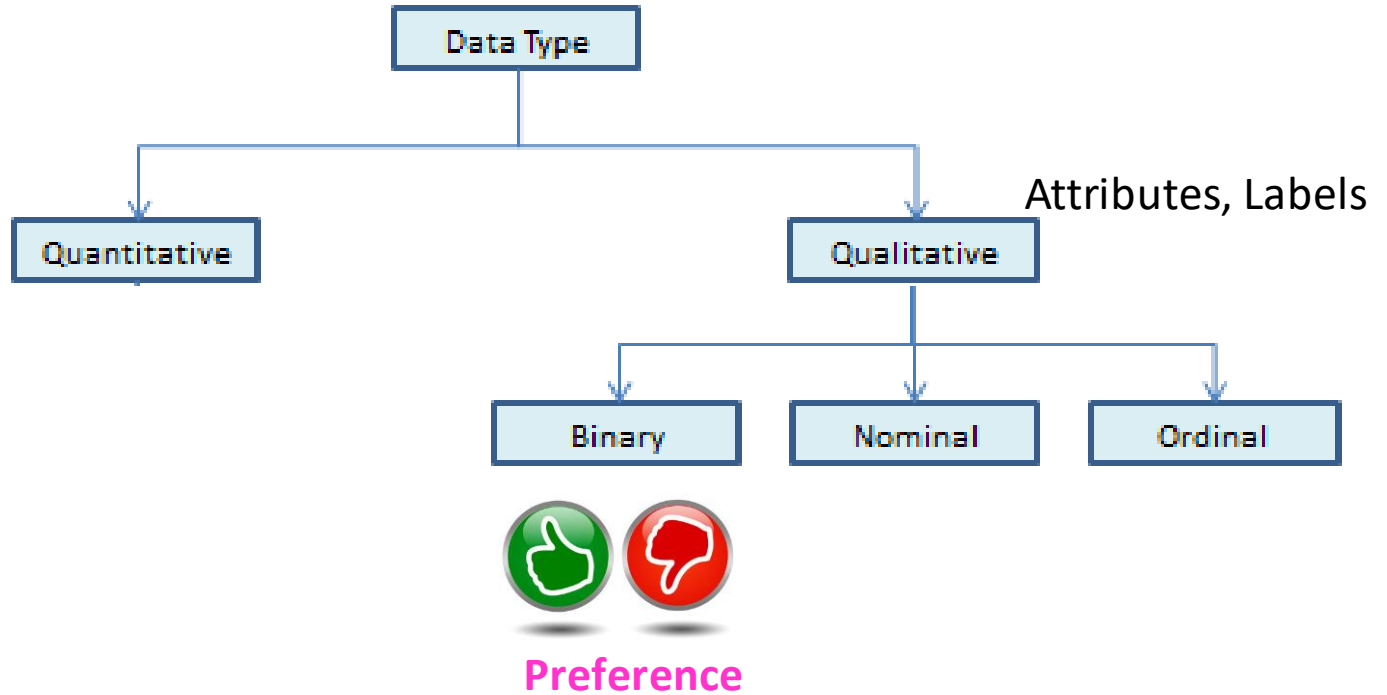
Workflow of a Machine Learning Problem



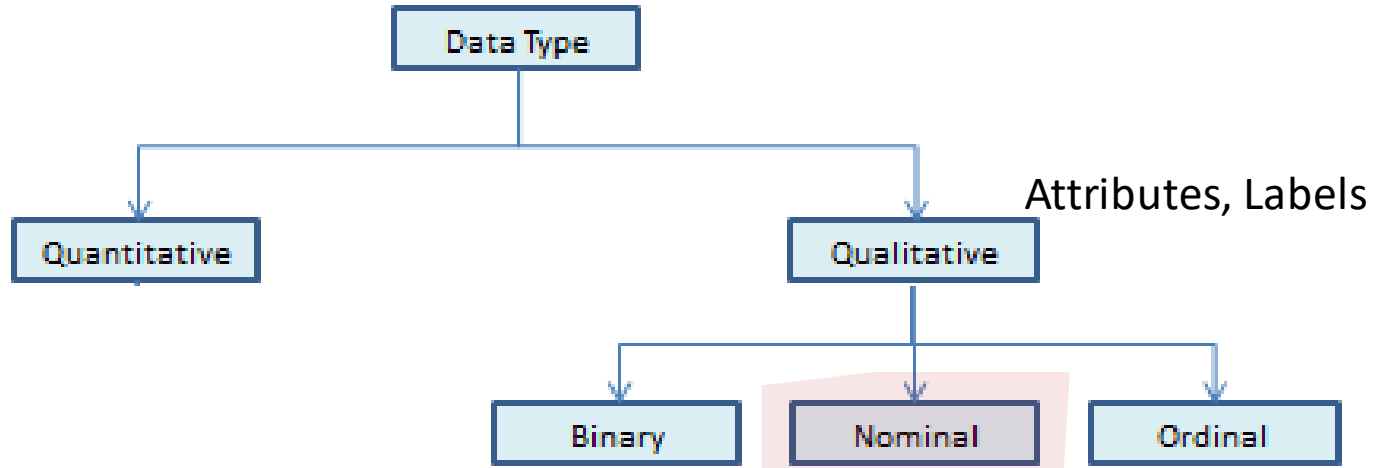
Taxonomy of data variables



Taxonomy of data



Taxonomy of data



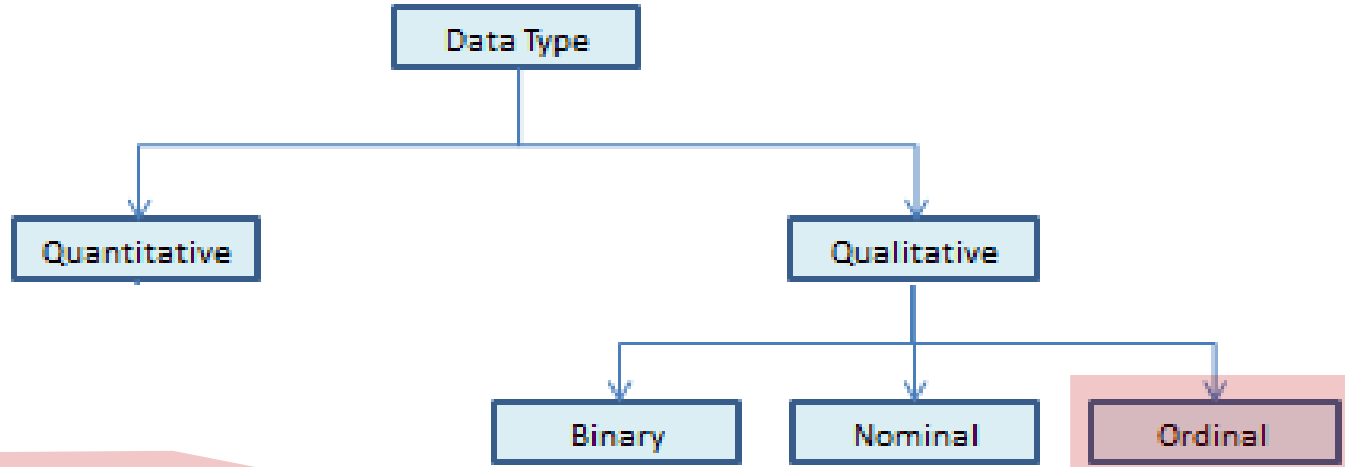
Color



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



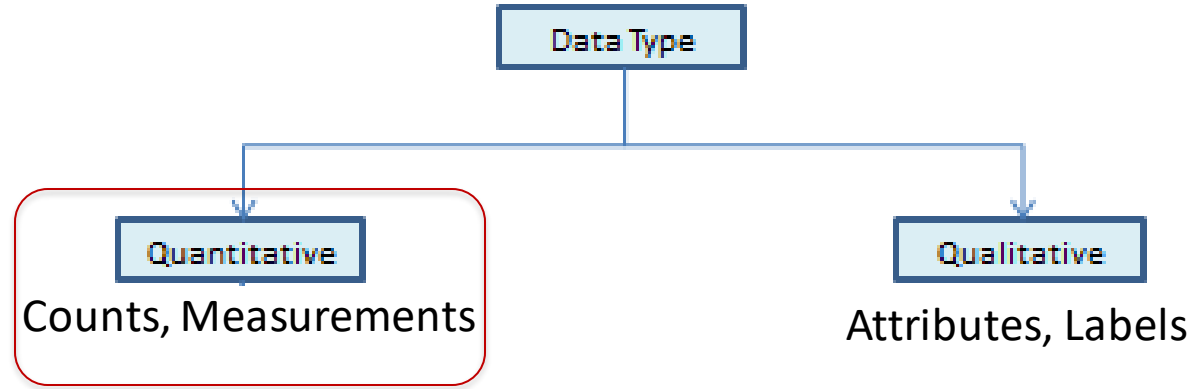
CHEN
Yufei



5

POINTS - 74,889

Taxonomy of data



QUANTITATIVE DATA:



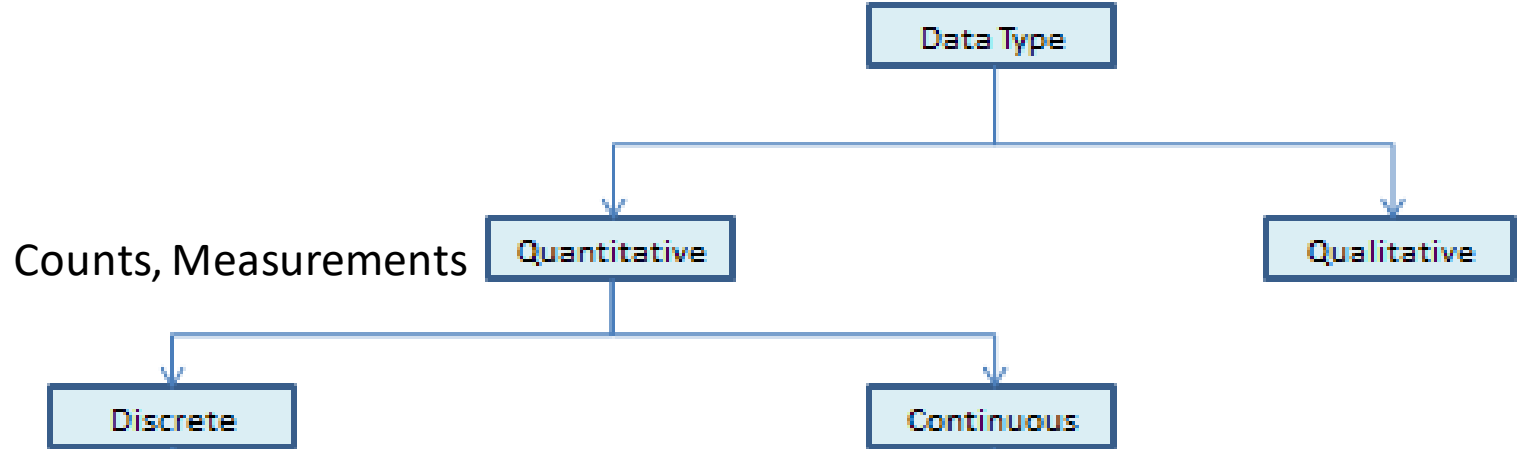
Discrete data:

- There are 3 cones
- Cone 1 has 2 scoops

Continuous data:

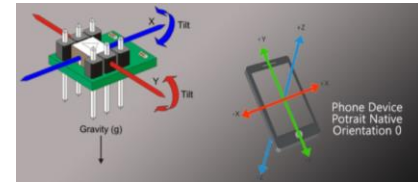
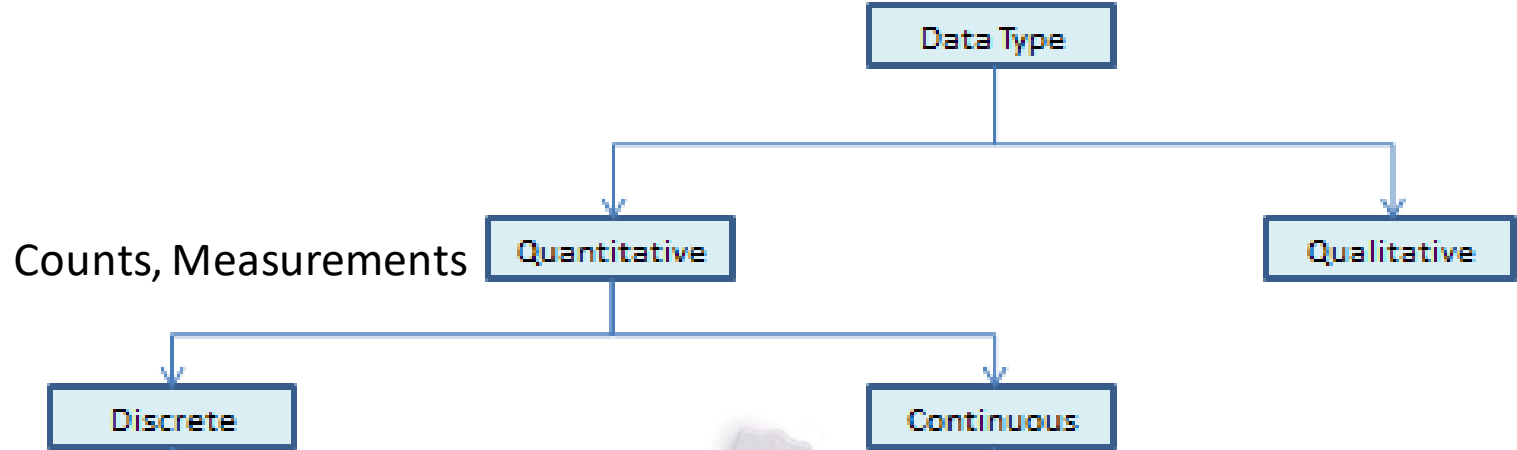
- Cone 3 weighs 79.4 grams
- cone 2 ice cream is at 8.3°F

Taxonomy of data

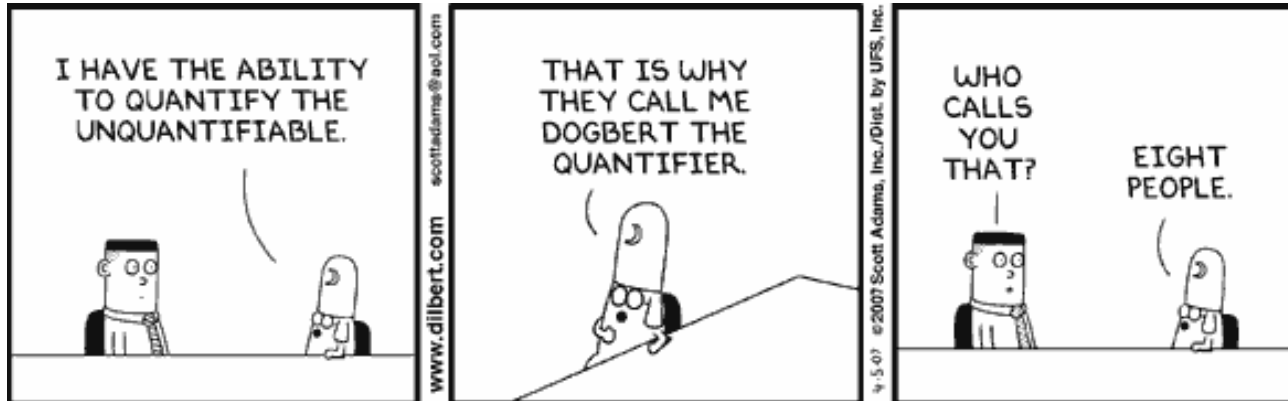
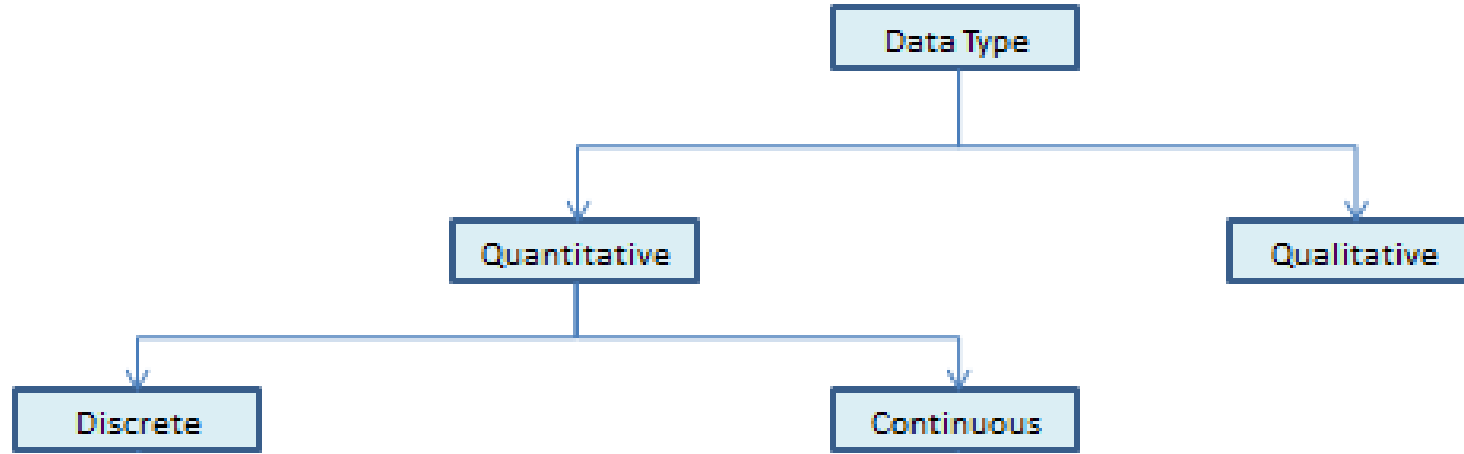


- # of CPU cores
- # of courses taken in a semester
- # of times word 'sale' appears in a doc

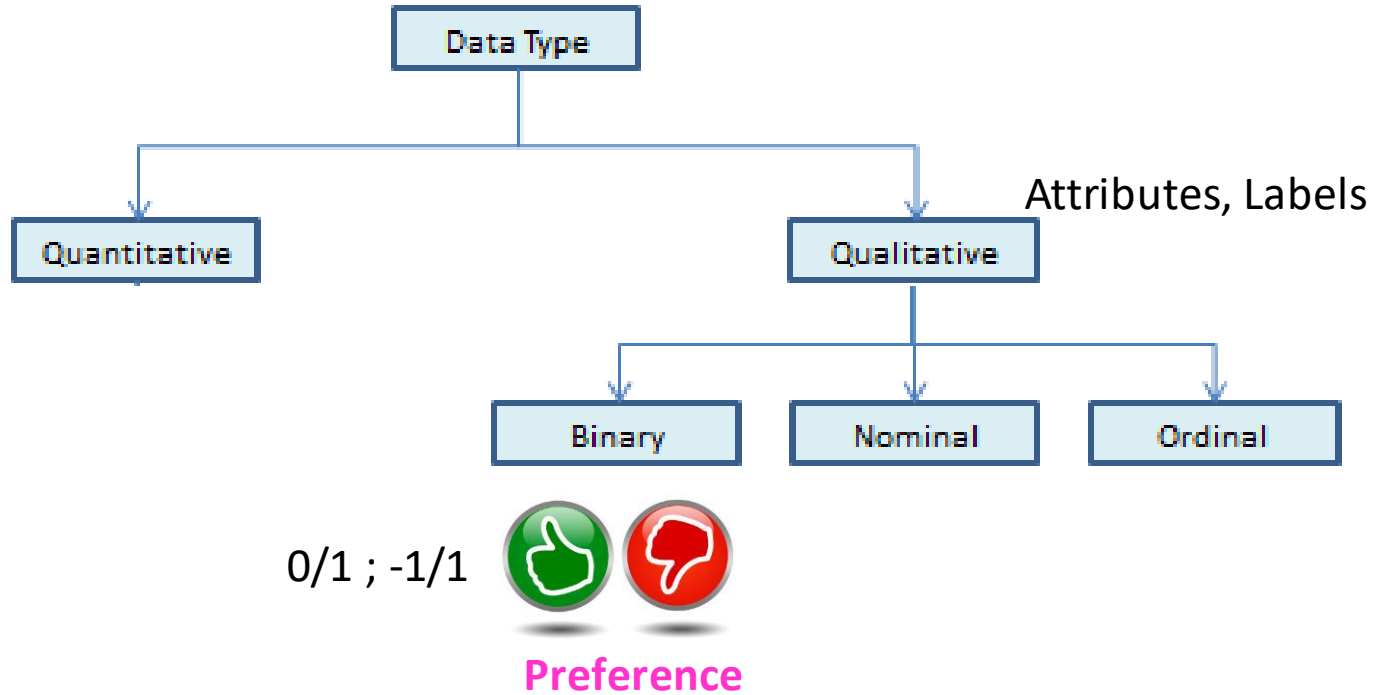
Taxonomy of data



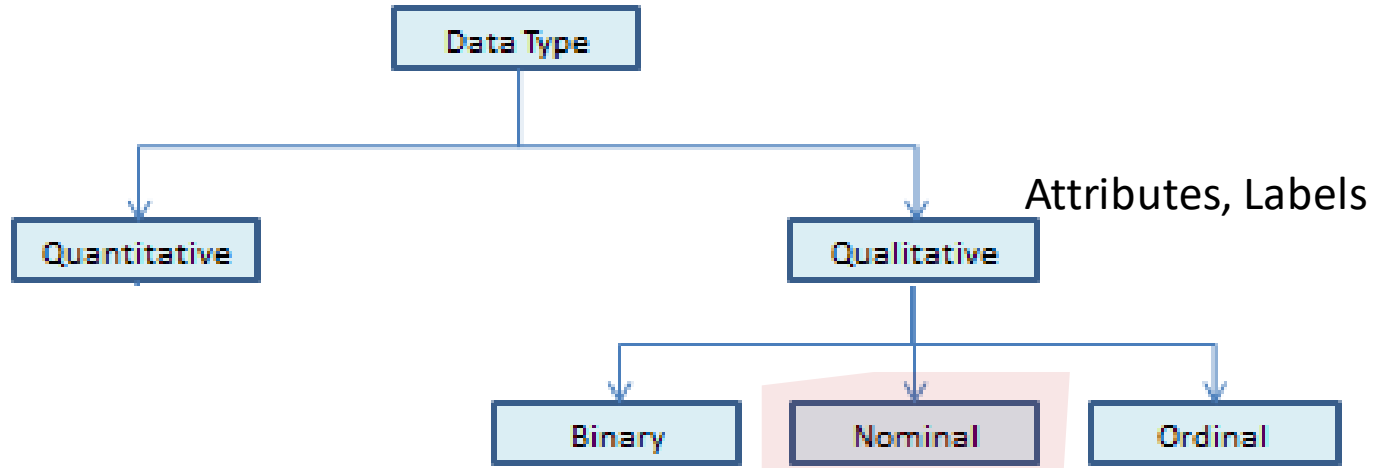
Ultimately, all data needs to be quantitative



Taxonomy of data: Qualitative → Quantitative



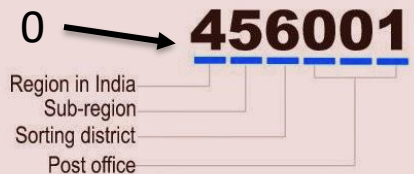
Taxonomy of data: Qualitative → Quantitative



Color



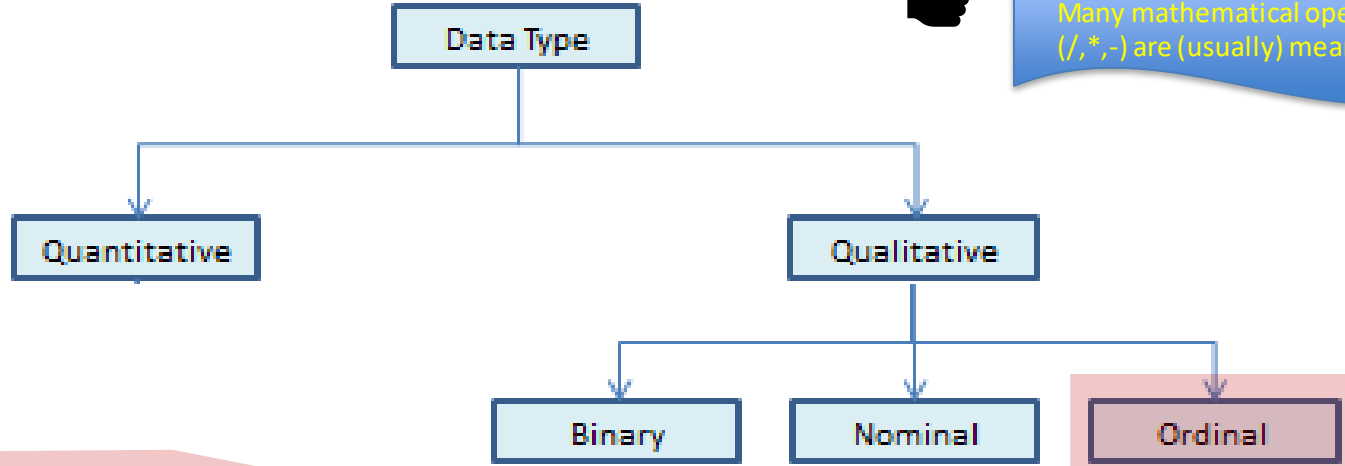
Make



Pin Code



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



How comfortable are you with Python *

No knowledge -2 +1 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

 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	 Conflicts
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	

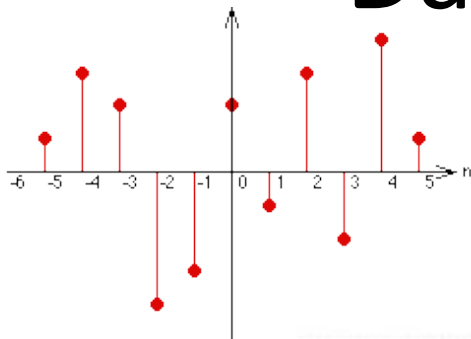
Does not obey data distribution

Conflict

Data imputation

- Approaches that aim to estimate missing data

Data Representations



Scalars

X

Vectors

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

Matrix

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

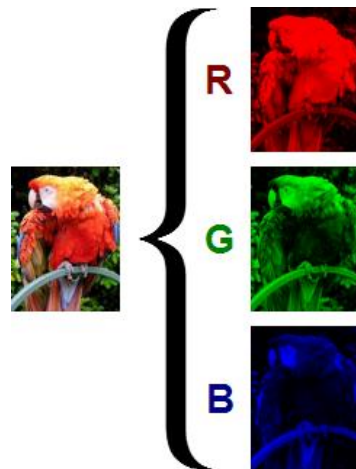
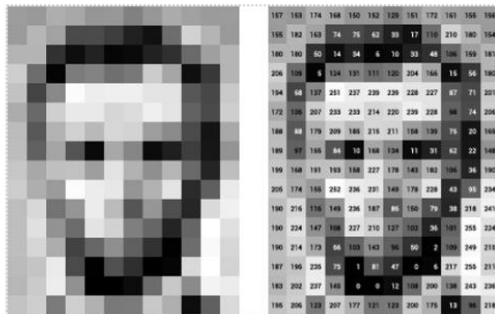
$M^{\text{th}} \text{ dimension}$

Tensor

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

$M^{\text{th}} \text{ dimension}$

2-d image



Data Representations



Graph Representation

Vertex List

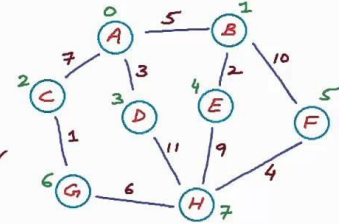
0	A
1	B
2	C
3	D
4	E
5	F
6	G
7	H
	↓

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.



■ Examples

- Construct bag-of-words vector from an email
- Remove stopwords in a sentence
- Apply PCA projection to high-dimensional data

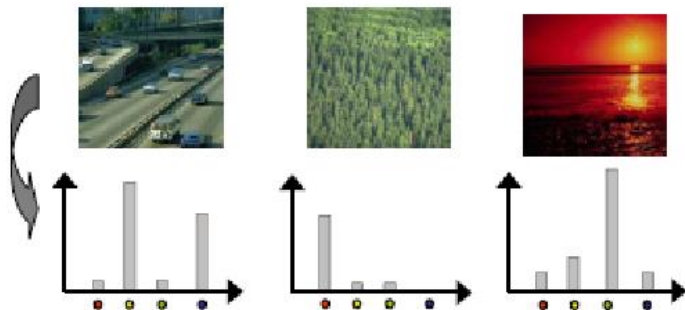
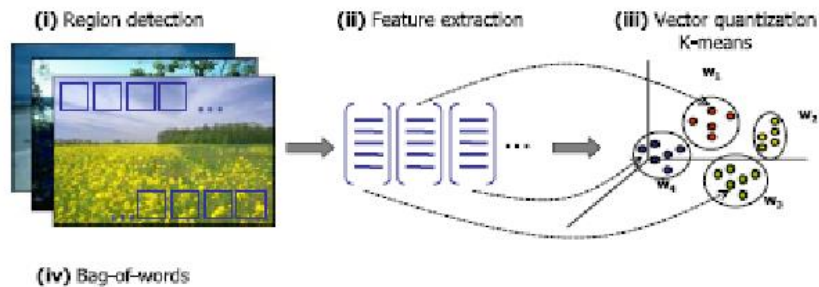
The Bag of Words Representation

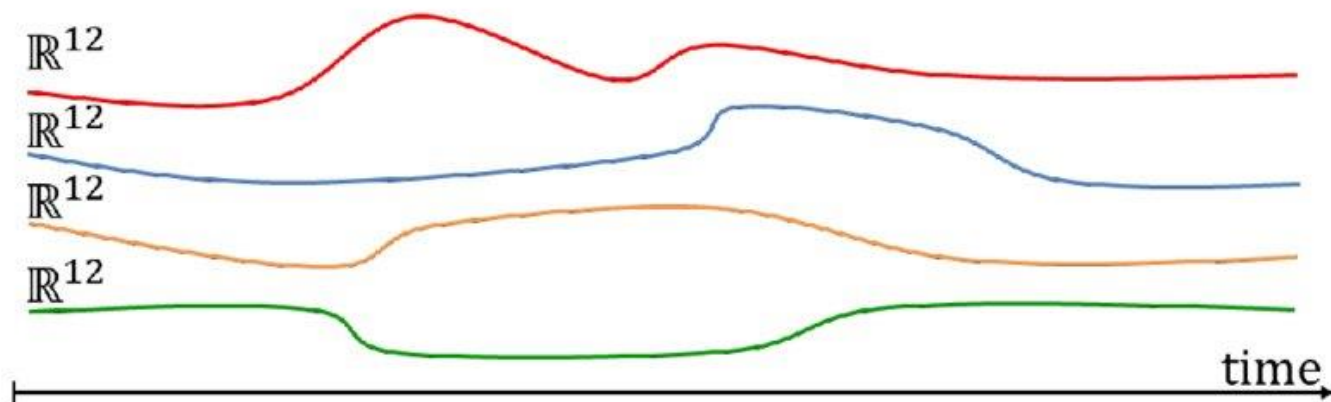
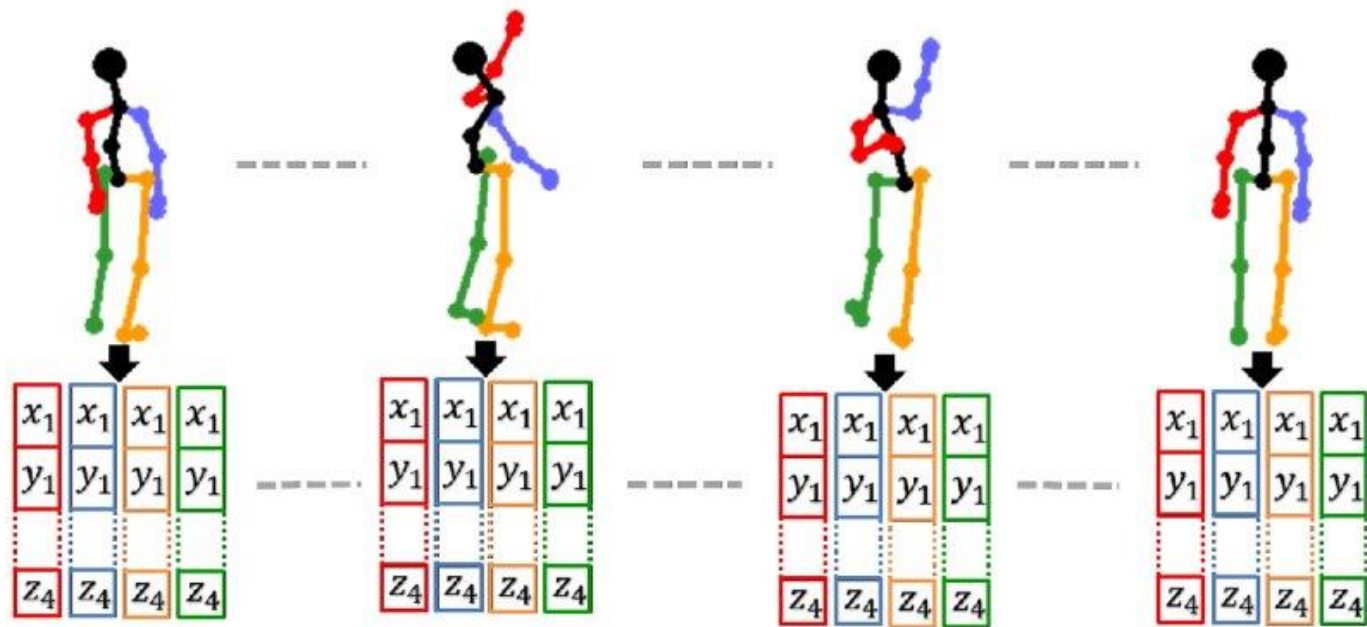
I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy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

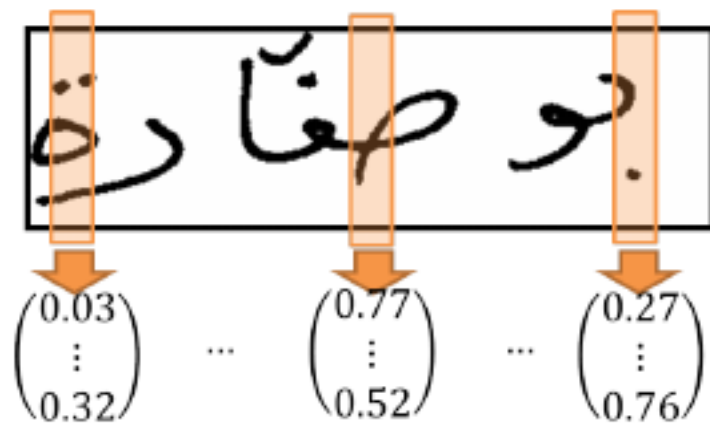
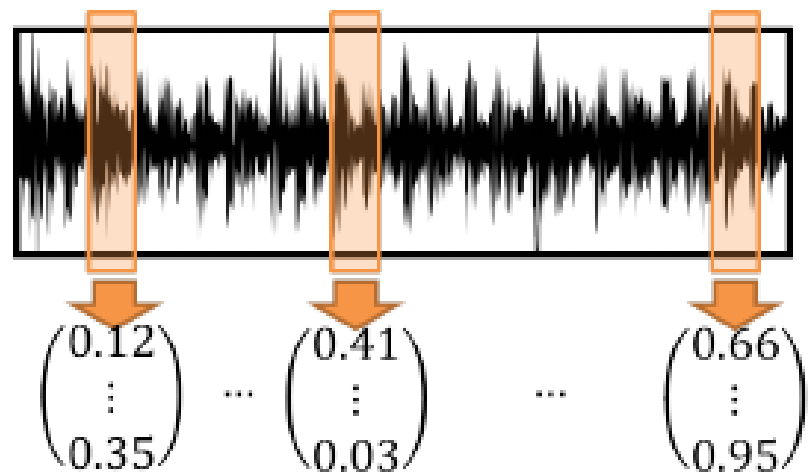


it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
...	...

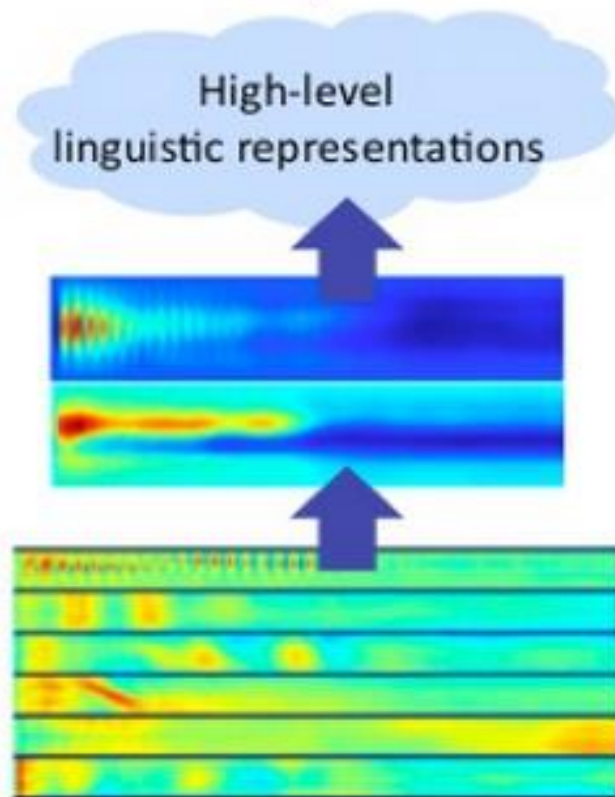
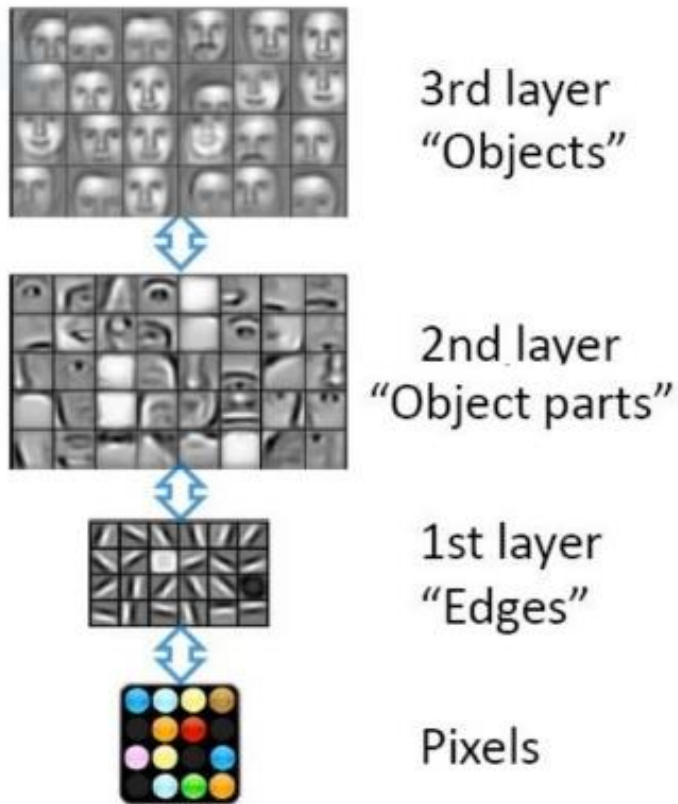
15







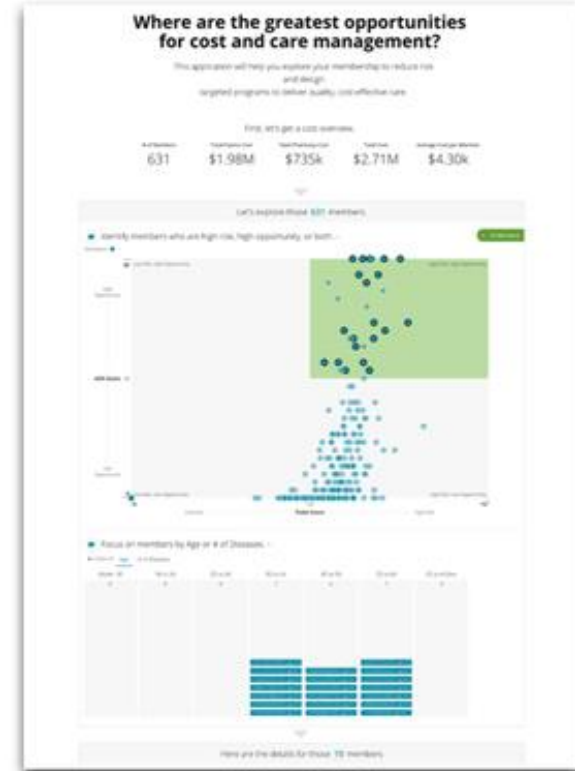
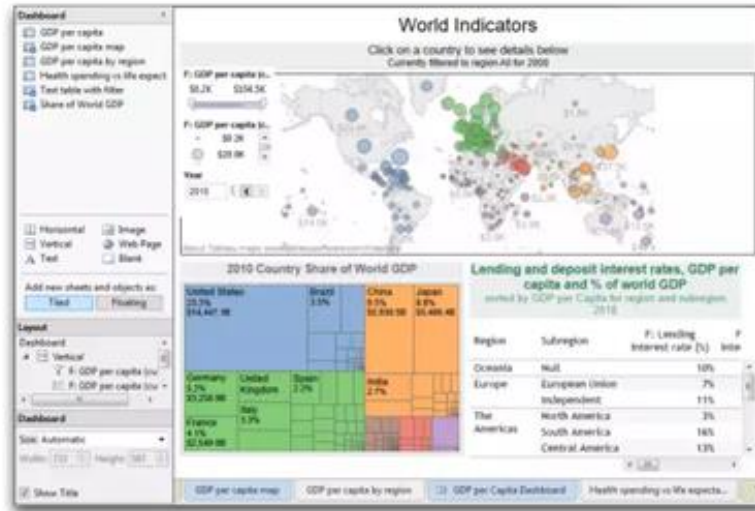
Feature-based, Hierarchical Data Representations

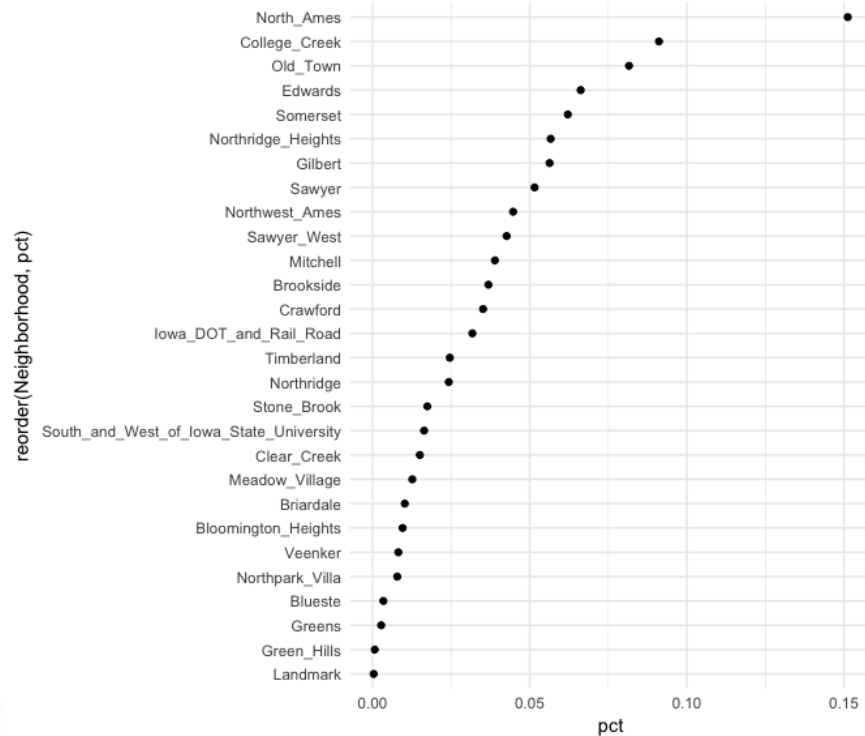
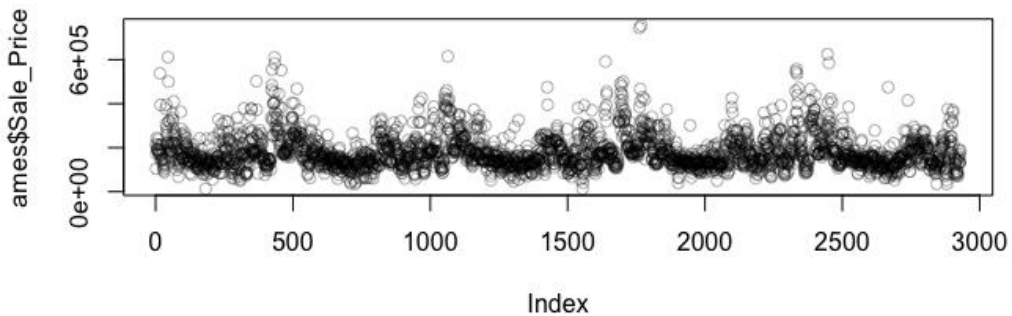
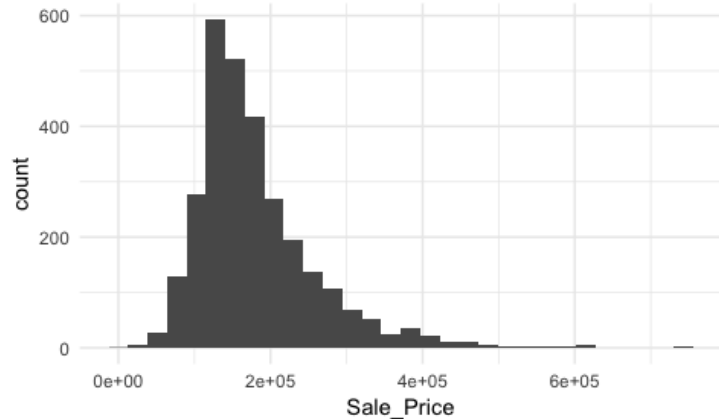


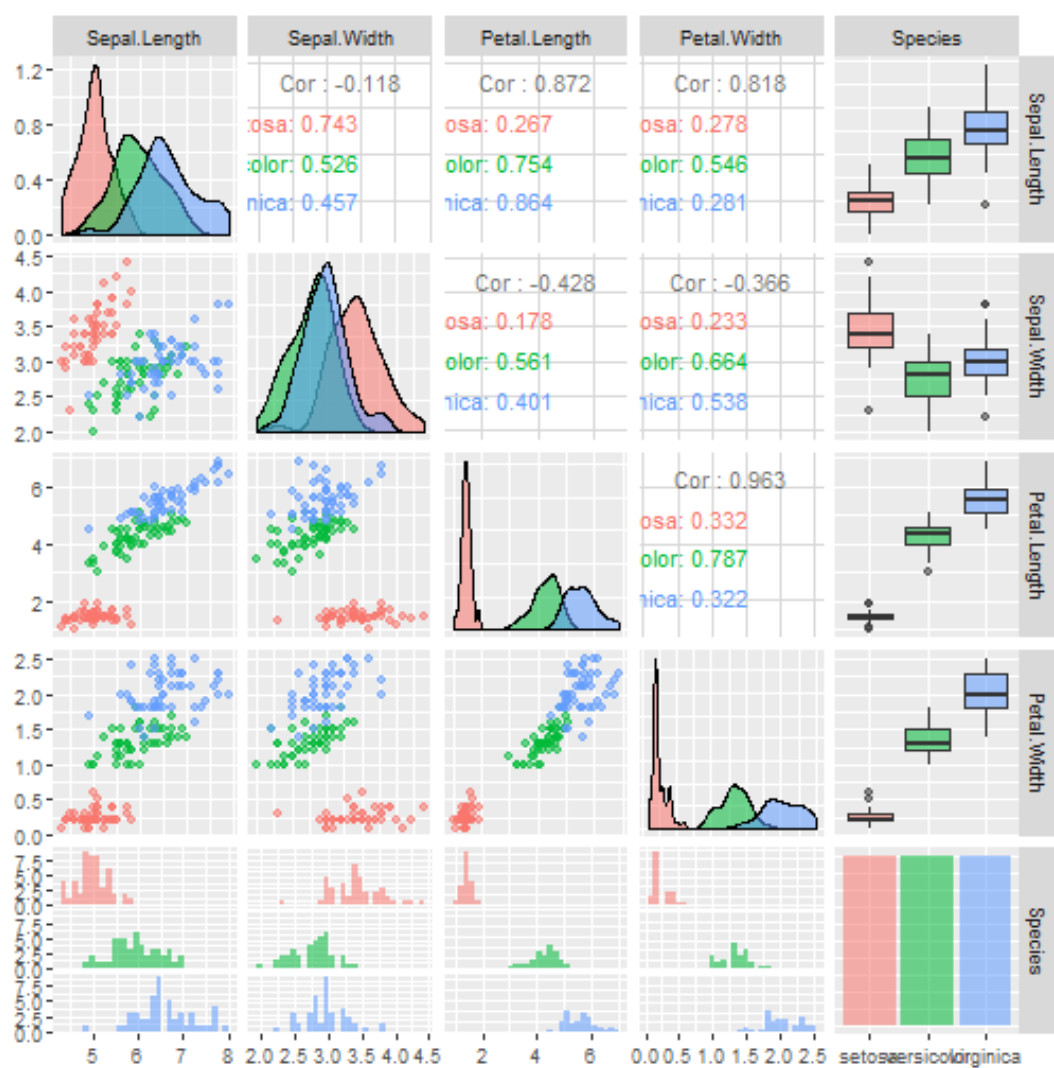
Gazing at Data: Data visualization

data exploration

data presentation

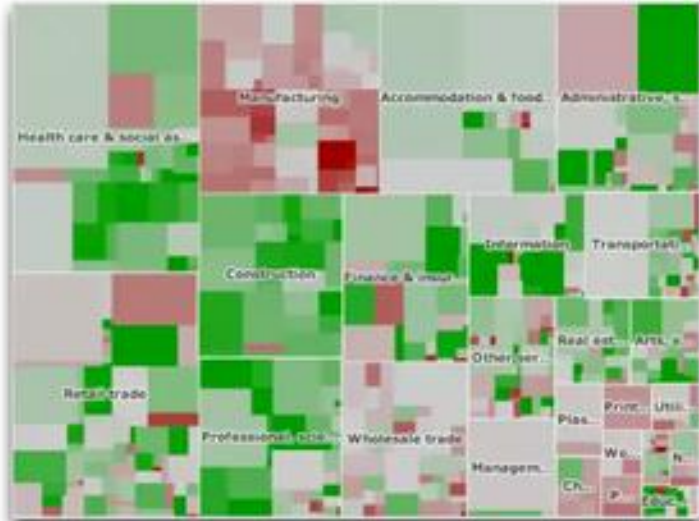






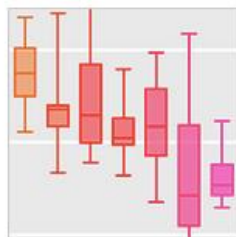
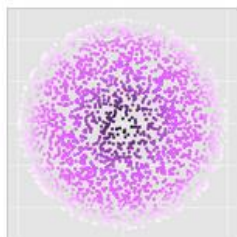
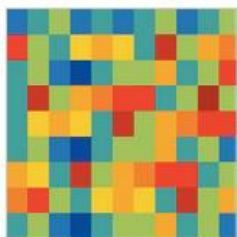
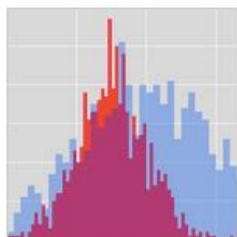
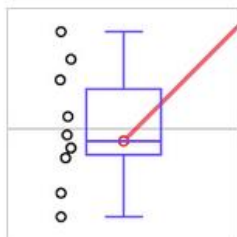
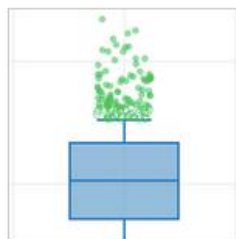
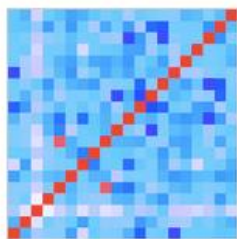
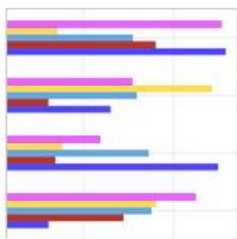
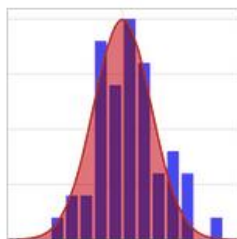
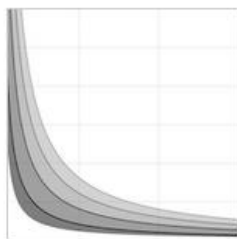
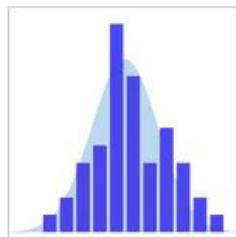
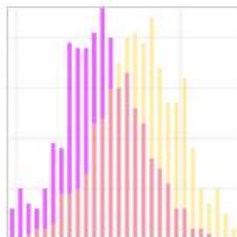
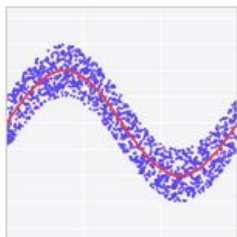
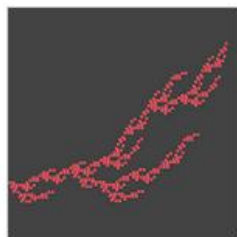
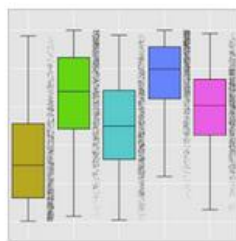
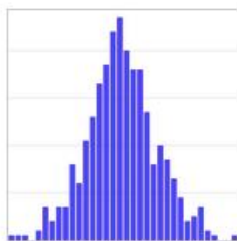
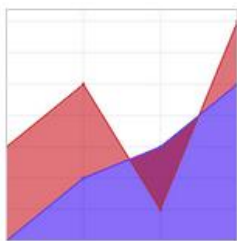
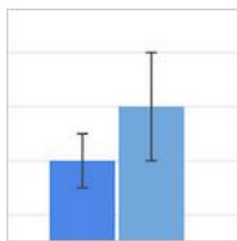
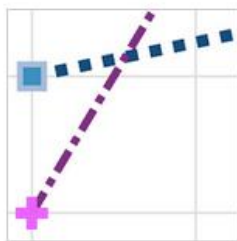
Data visualization

treemap

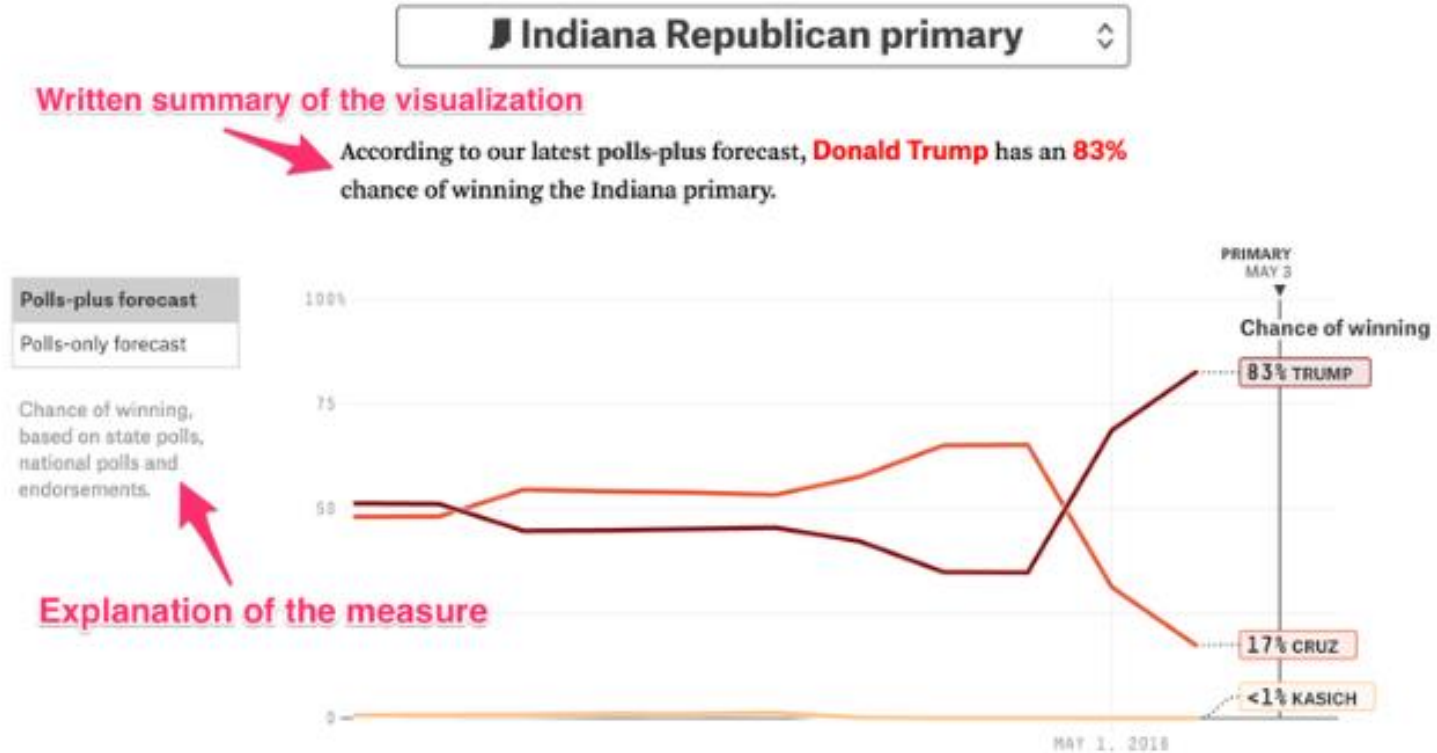


leaderboard

SHUTTLE		40 YARD		BENCH PRESS		VERT LEAP (in)		BENCH 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	Darion Thomas	14	Jacory Harris	37	Darion 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	---	Darion Thomas	36	Russell Wilson	118
Andrew Luck	4.28	Aaron Corp	4.72	Andrew Luck	---	Andrew Luck	36	Jordan Jaffer...	116
Darion 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	Darion 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



Data Exploration and Visualization



Fivethirtyeight provides explanation surrounding their visualization to ensure readers understand what they are looking at.

Data Exploration and Visualization

.. To be covered in next week's tutorial

// 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

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	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



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

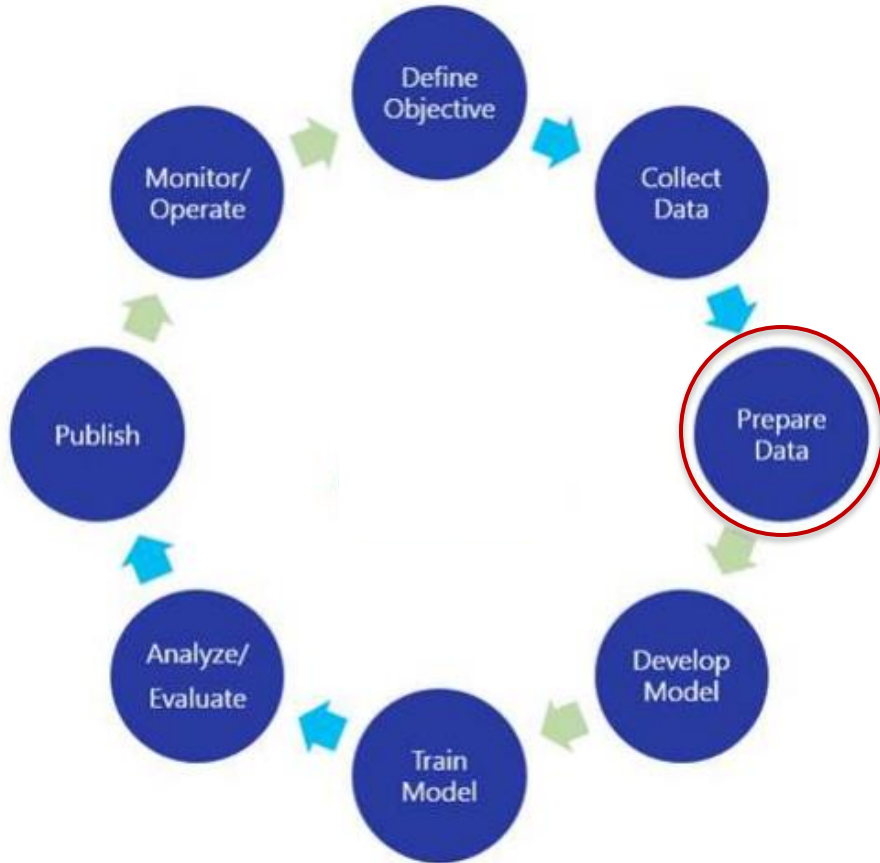
Marginal Distribution of Cell Assignments

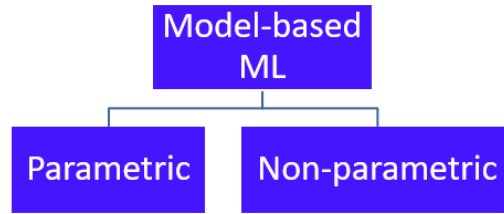
Cell	Possible Values	Probability
t2.Zip	60608	0.84
	60609	0.16
t4.City	Chicago	0.95
	Cicago	0.05
t4.DBAName	John Veliotis Sr.	0.99
	Johnnyo's	0.01

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)

Workflow of a Machine Learning Problem

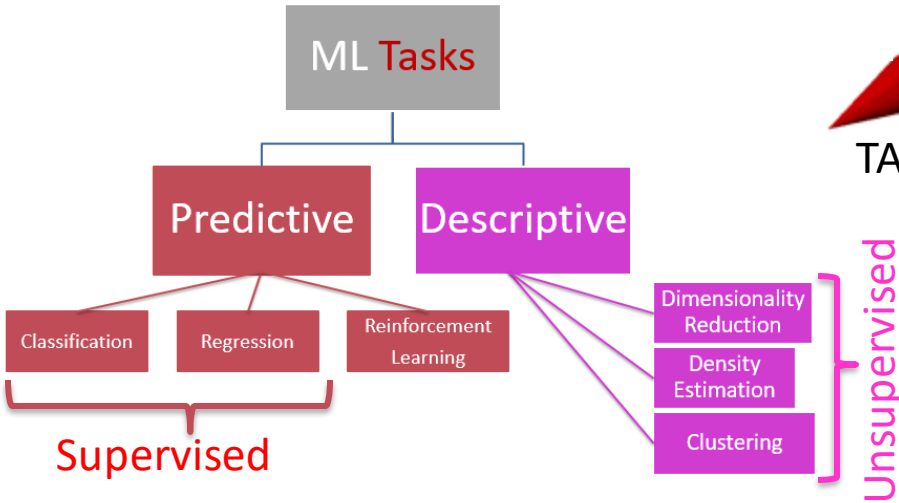




ALGORITHMS

DATA

TASKS



- 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)
- Actively collect / sense data (e.g. exploration robots)

ML Tasks

```
graph TD; A[ML Tasks] --> B[Supervised Learning]; A --> C[Unsupervised Learning];
```

Supervised
Learning

Given an input,
estimate output

Unsupervised
Learning

ML::Tasks → Predictive

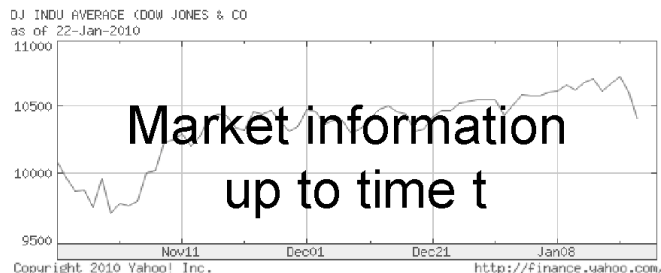
Feature Space \mathcal{X}



Words in a document

Label Space \mathcal{Y}

“Sports”
“News”
“Science”
...



Share Price
“\$ 24.50”



Task: Given $X \in \mathcal{X}$, predict $Y \in \mathcal{Y}$.

ML::Tasks → Predictive → Classification

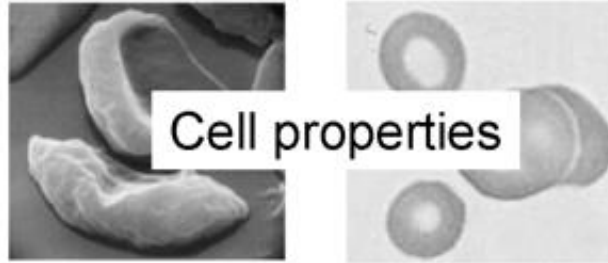
Feature Space \mathcal{X}



Label Space \mathcal{Y}



"Sports"
"News"
"Science"
...



"Anemic cell"
"Healthy cell"

Discrete Labels

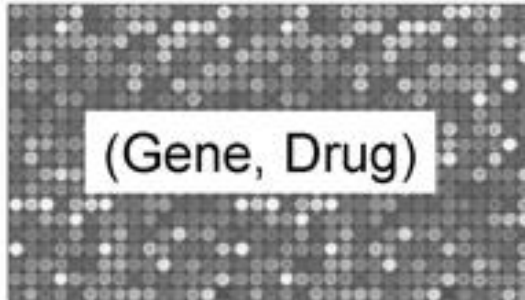
ML::Tasks \rightarrow Predictive \rightarrow Regression

Feature Space \mathcal{X}

Label Space \mathcal{Y}



Share Price
"\$ 24.577"



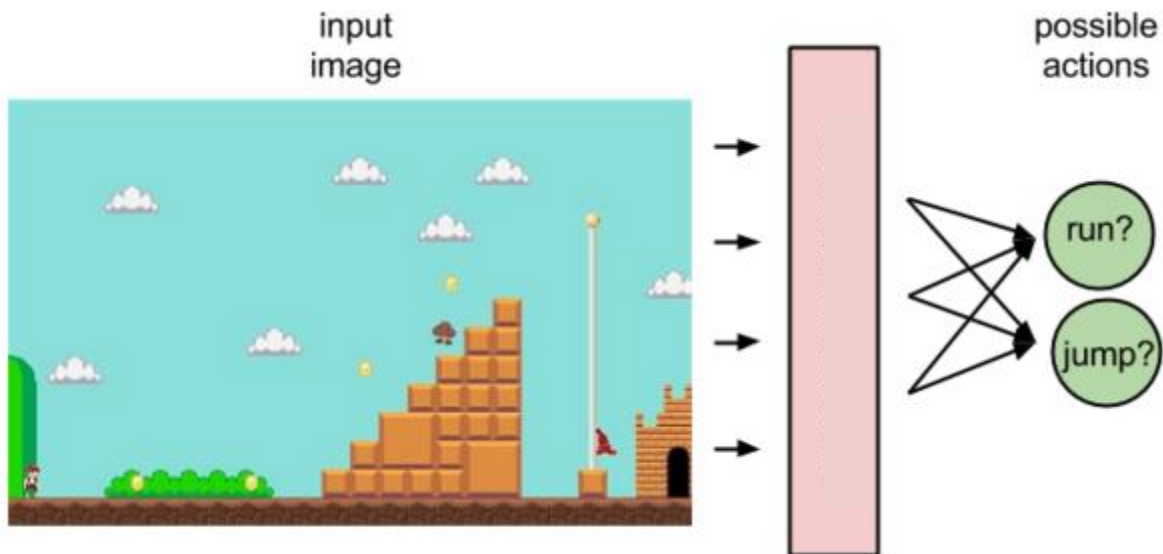
Expression level
"6.88"

Continuous Labels

ML::Tasks \rightarrow Predictive \rightarrow Reinforcement Learning

Feature Space \mathcal{X}

Label Space \mathcal{Y}



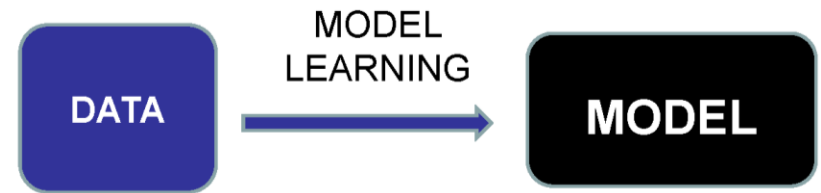
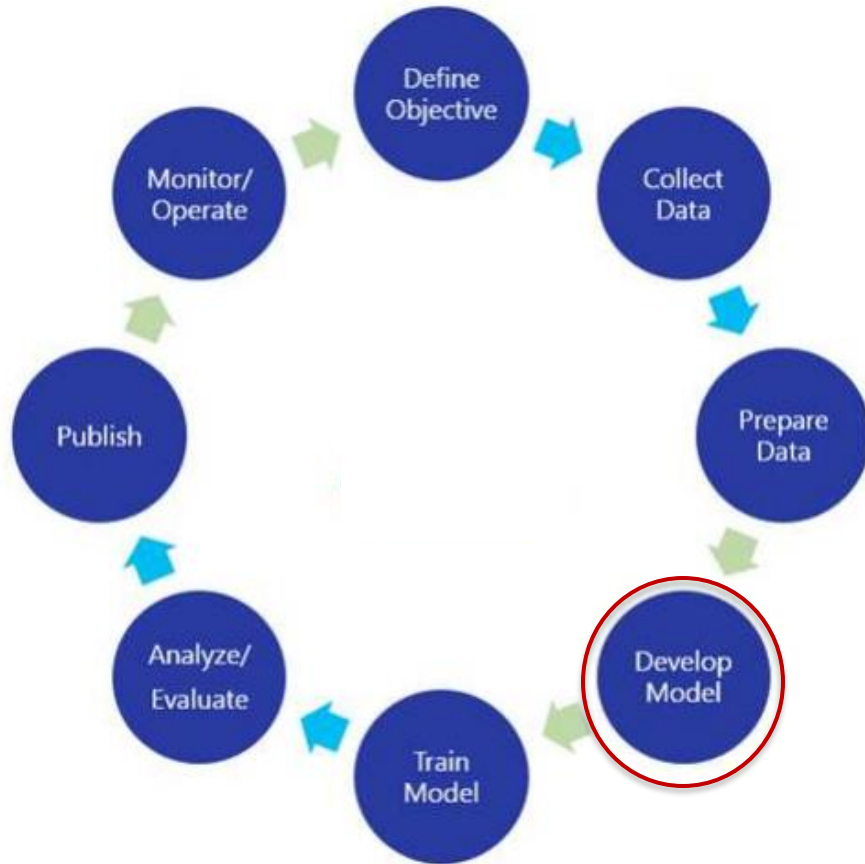
ML::Tasks \rightarrow Predictive \rightarrow Reinforcement Learning

Feature Space \mathcal{X}

Label Space \mathcal{Y}



Workflow of a Machine Learning Problem



Strategy for fulfilling preferences

Optimization

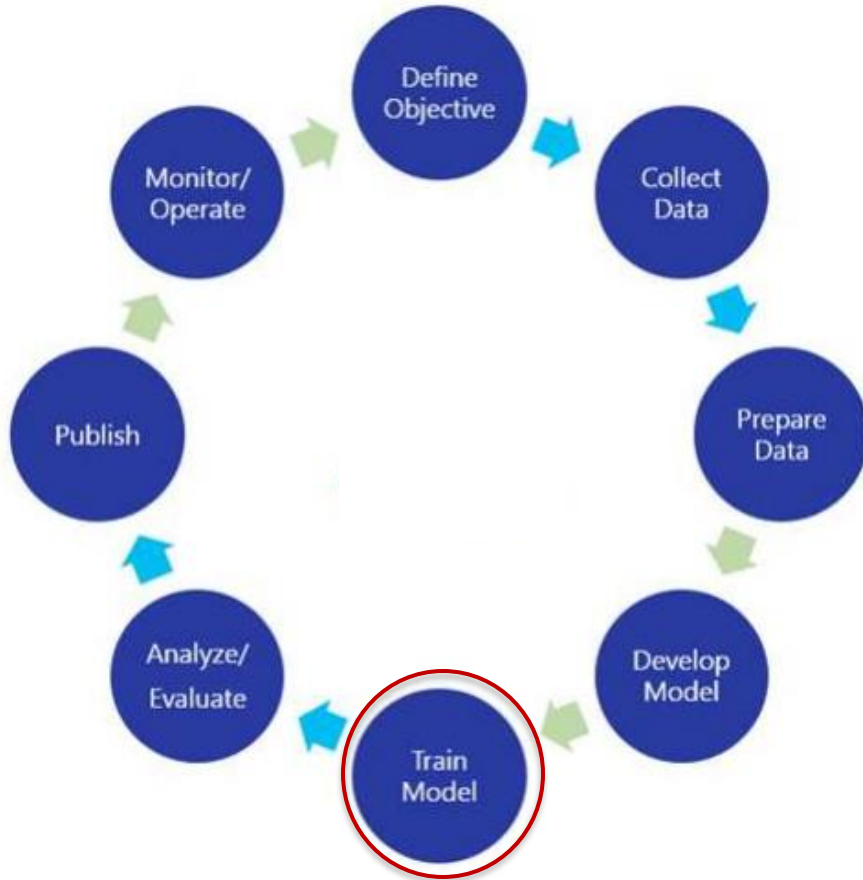
Evaluation

Representation

The landscape of allowed models

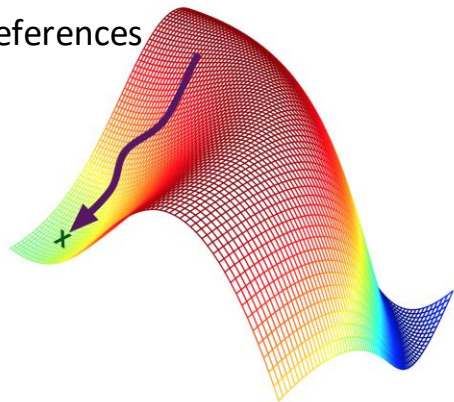
Preferences over the landscape

Workflow of a Machine Learning Problem



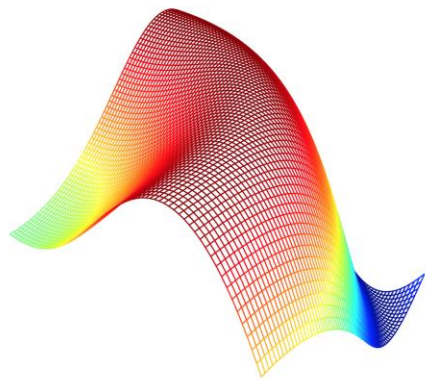
Strategy for fulfilling preferences

Optimization

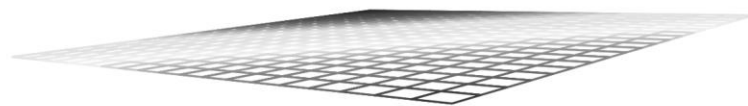


Evaluation

Representation

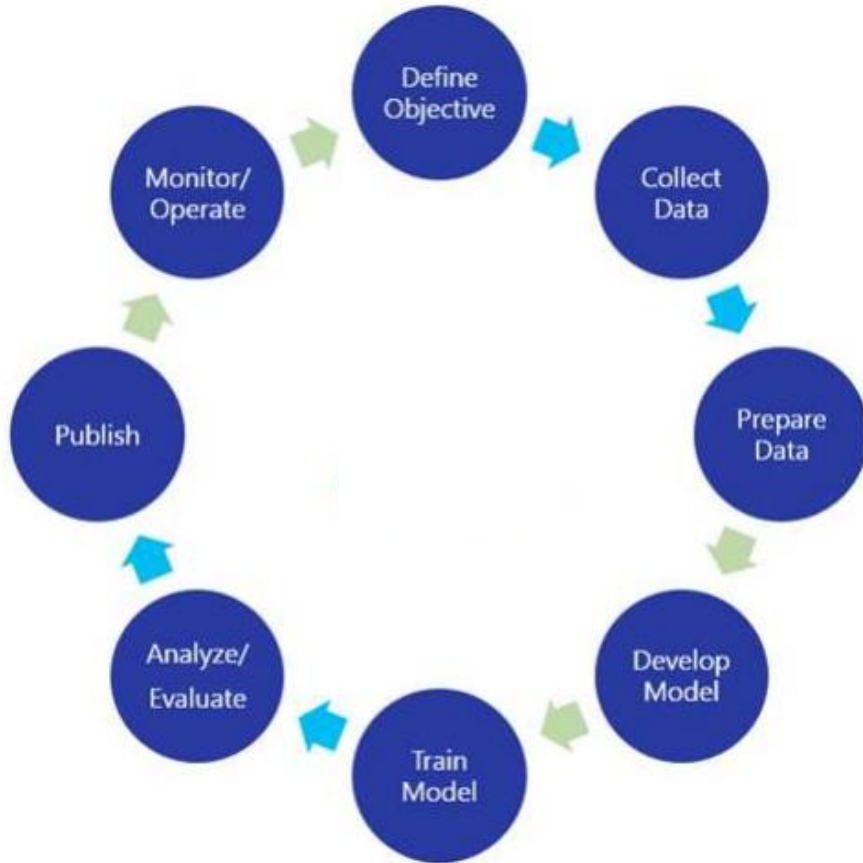


Preferences over the landscape



The landscape of allowed models

Workflow of a Machine Learning Problem



ML Tasks

```
graph TD; A[ML Tasks] --> B[Supervised Learning]; A --> C[Unsupervised Learning];
```

Supervised
Learning

Given an input,
estimate output

Unsupervised
Learning

Supervised Learning

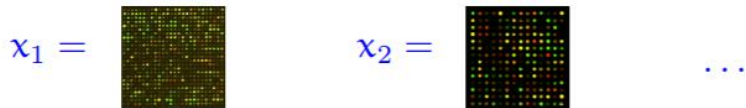
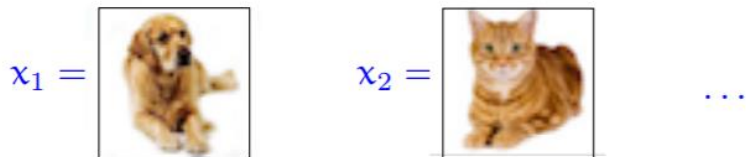


Data

Space of inputs (or, predictors): \mathcal{X}

▷ e.g. $\mathbf{x} \in \mathcal{X} \subset \{0, 1, \dots, 2^{16}\}^{64}$ is a string of pixel intensities in an 8×8 image.

▷ e.g. $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^{33,000}$ is a set of gene expression levels.



$\mathbf{x}_1 = \begin{bmatrix} 5 \\ 1 \\ 22 \\ \vdots \end{bmatrix}$ $\mathbf{x}_2 = \begin{bmatrix} 1 \\ 0 \\ 17 \\ \vdots \end{bmatrix}$ # cigarettes/day
drinks/day
BMI

Space of outputs (or, responses): \mathcal{Y}

▷ e.g. $y \in \mathcal{Y} = \{0, 1\}$ is a binary label (1 = “cat”)

▷ e.g. $y \in \mathcal{Y} = [0, 200]$ is life expectancy

Space of outputs (or, responses): \mathcal{Y}

▷ e.g. $y \in \mathcal{Y} = \{0, 1\}$ is a binary label ($1 = \text{"cat"}$)

▷ e.g. $y \in \mathcal{Y} = [0, 200]$ is life expectancy

A pair (x, y) is a *labeled* example.

▷ e.g. (x, y) is an example of an image with a label $y = 1$, which stands for the presence of a face in the image x

Space of outputs (or, responses): \mathcal{Y}

- ▷ e.g. $y \in \mathcal{Y} = \{0, 1\}$ is a binary label (1 = “cat”)
- ▷ e.g. $y \in \mathcal{Y} = [0, 200]$ is life expectancy

A pair (x, y) is a labeled example.

- ▷ e.g. (x, y) is an example of an image with a label $y = 1$, which stands for the presence of a face in the image x

Dataset (or *training data*): examples $\{(x_1, y_1), \dots, (x_n, y_n)\}$

- ▷ e.g. a collection of images labeled according to the presence or absence of a face

Supervised Learning

```
graph TD; A[Supervised Learning] --> B[Classification]; A --> C[Regression]; A --> D[Reinforcement Learning]; B --- E[ ]; C --- E; E --- F[We'll focus on these two];
```

Classification

Regression

Reinforcement
Learning

We'll focus on these two

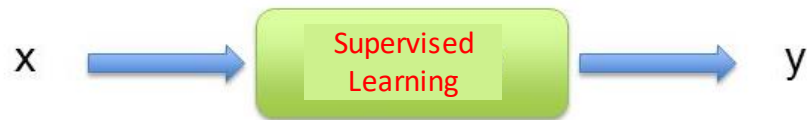
Supervised Learning

```
graph TD; A[Supervised Learning] --> B[Classification]; A --> C[Regression]; A --> D[Reinforcement Learning];
```

Classification

Regression

Reinforcement
Learning



Classification

Binary

$\{0,1\}$

Multi-class

1-of-K

Multi-label

n-of-K

Structure

E.g. graph/sequence

