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

DATA → LEARNING ALGORITHMS → KNOWLEDGE

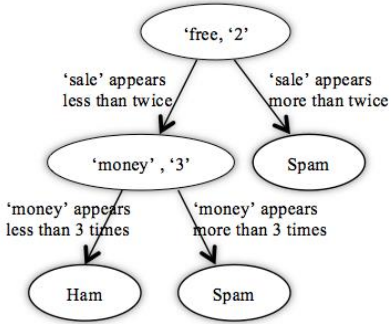
Algorithmic methods that use data to improve their knowledge of a task

Task: Detect spam email



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

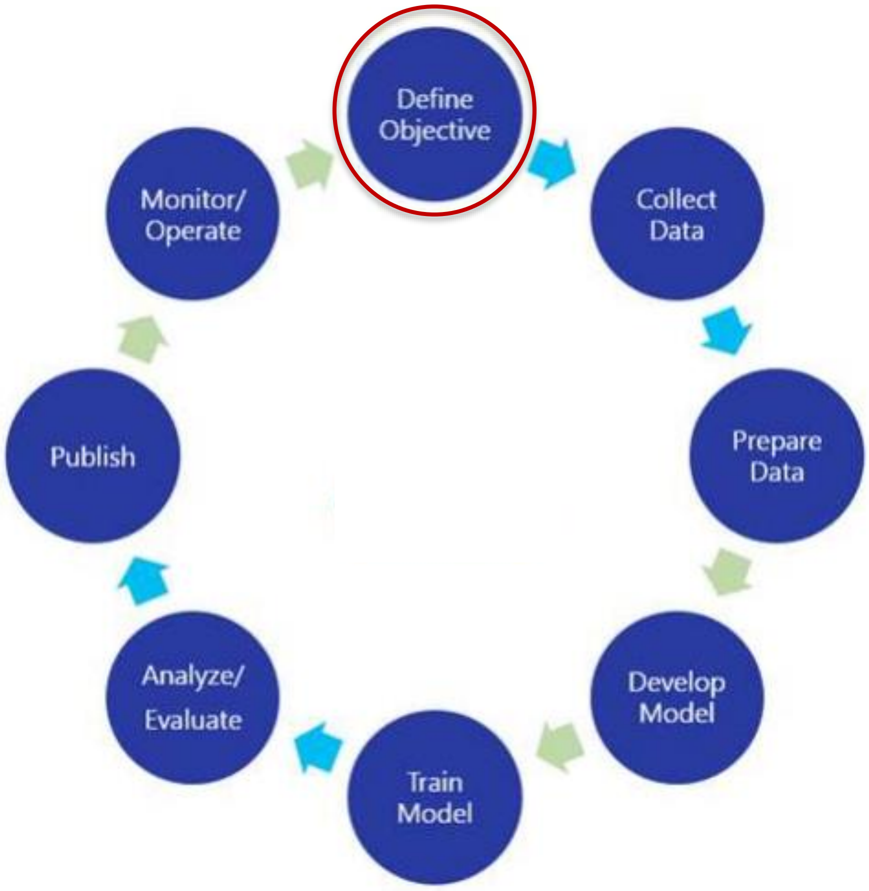
Knowledge:




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

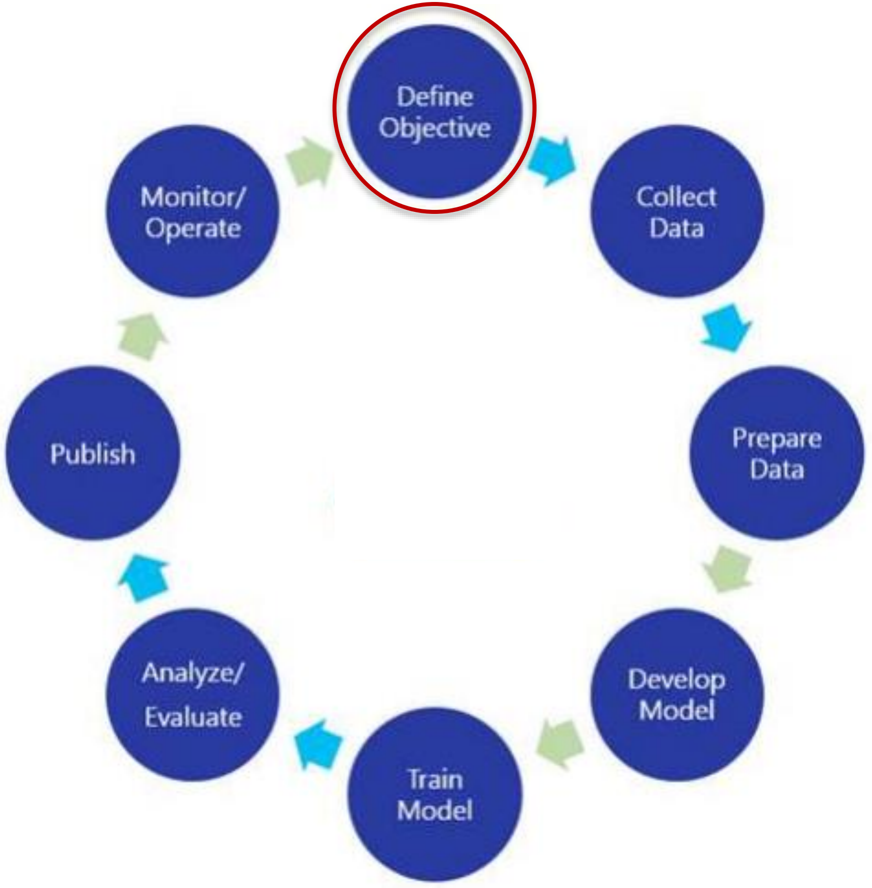
Algorithmic method: Decision Tree

# 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



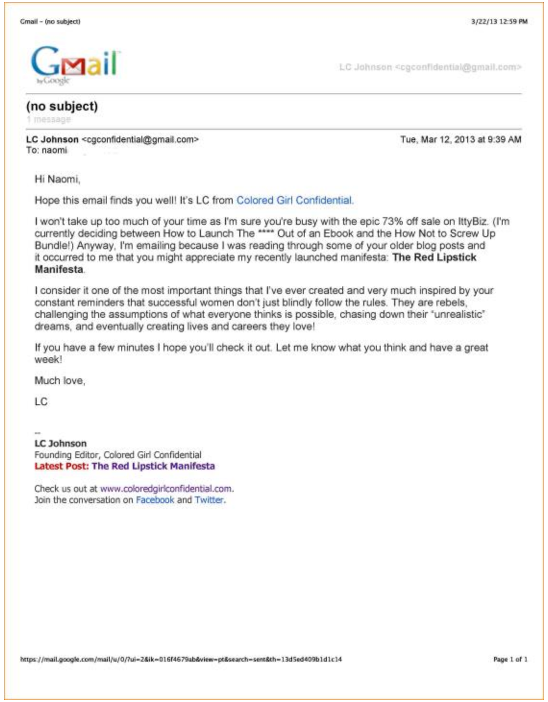
# No Data, no ML !

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

# Sources of data

- Detect spam email



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

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!

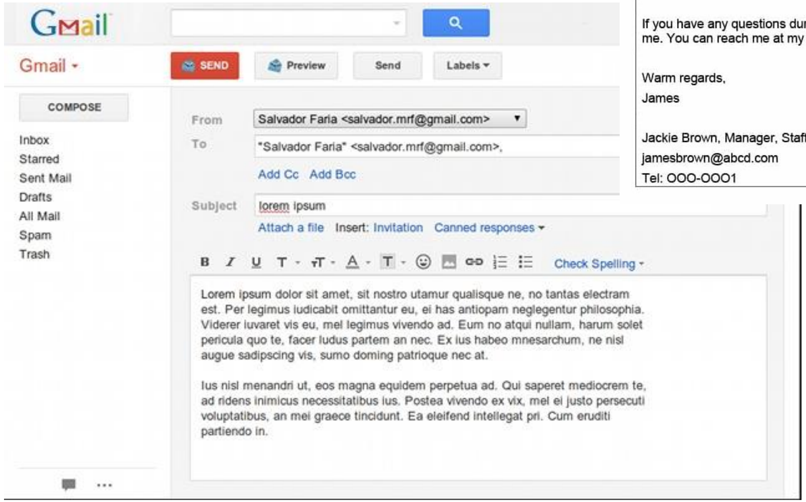
Much love,

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

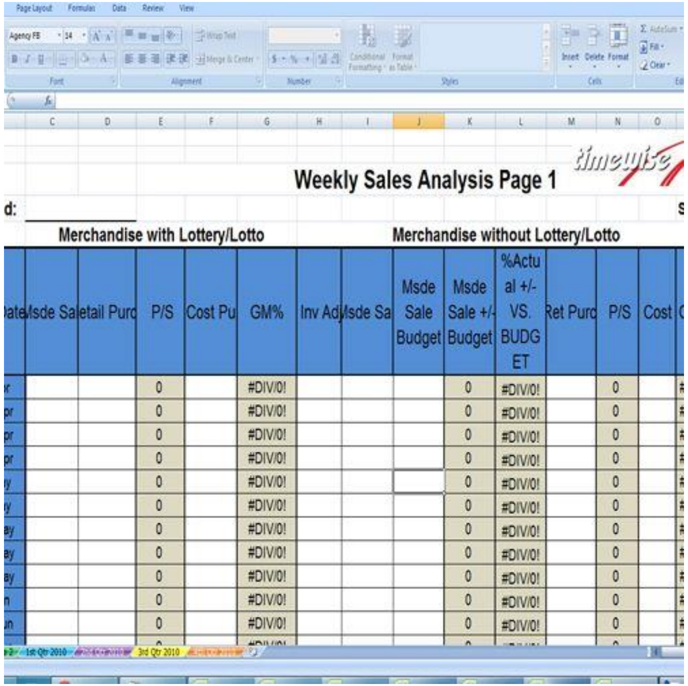
Jackie Brown, Manager, Staff
jamesbrown@abcd.com
Tel: OOO-OOO1

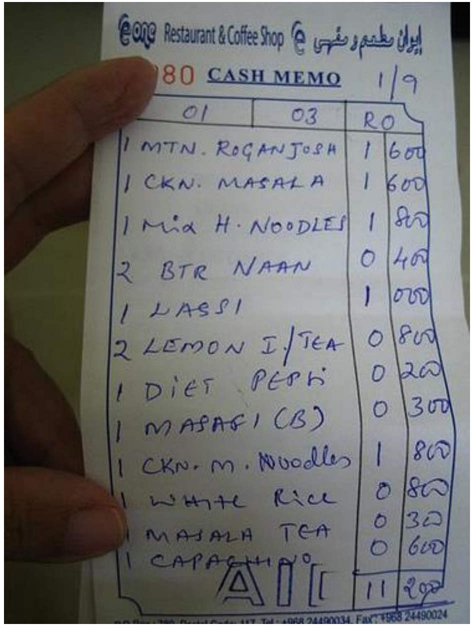
# Sources of data

- Predict value of a stock

# Sources of data

- Predict effect of advertising on sales

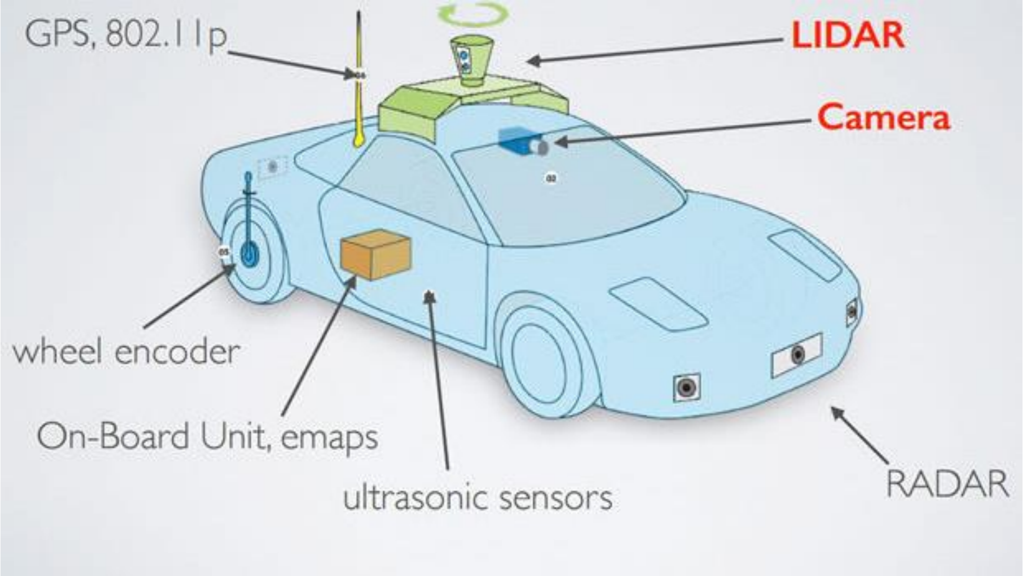



Raw Data may not always be digital in nature !

# Sources of data

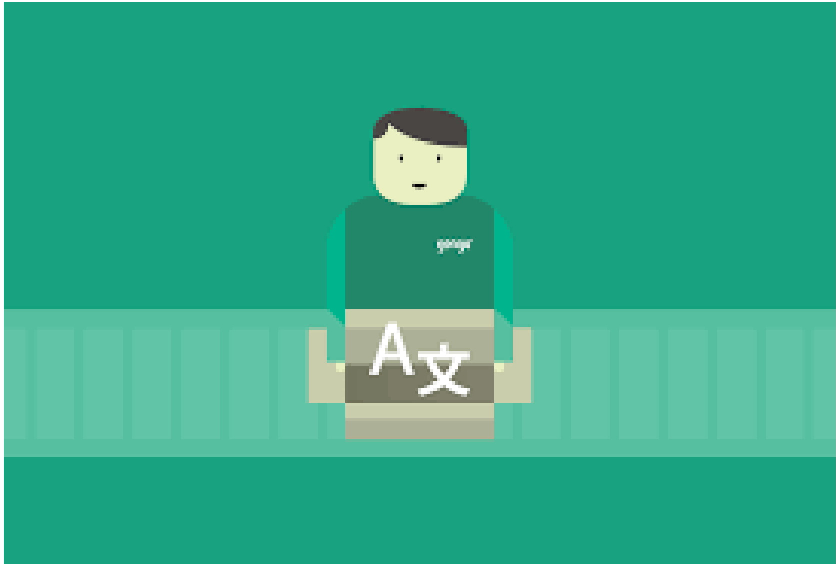
- Drive car safely without human intervention




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

# 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

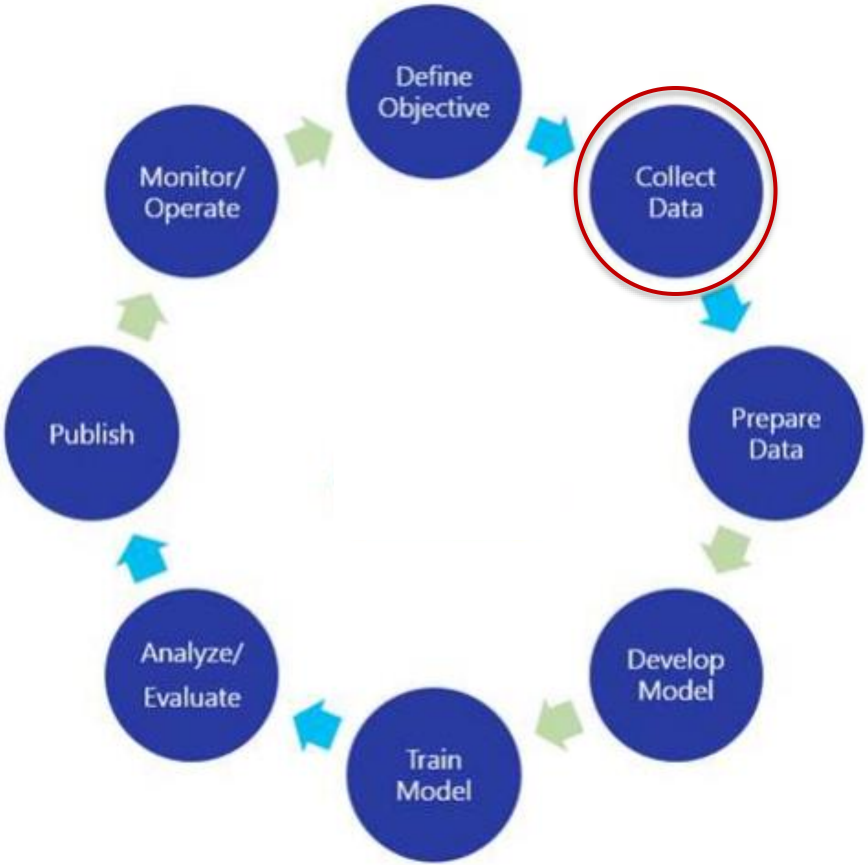
- Often not directly usable
  - Filter (needed data)
  - **Transform (to numerical data)**


```
{
  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"
}
```

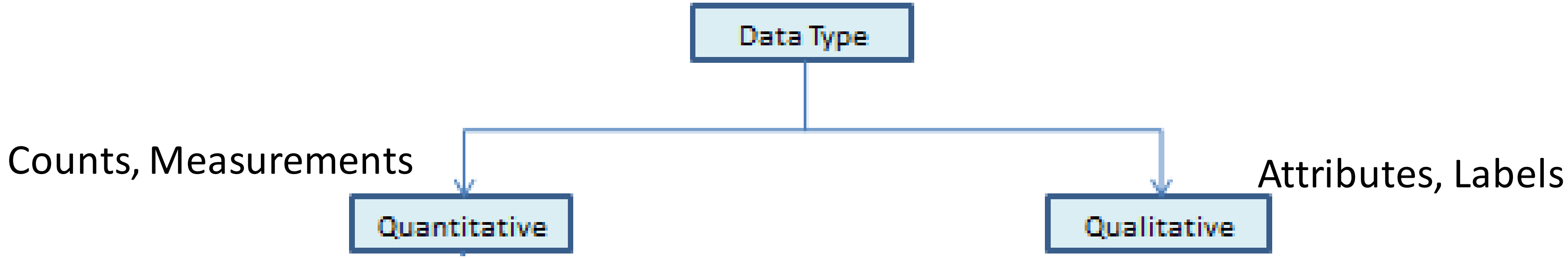
# Raw data

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

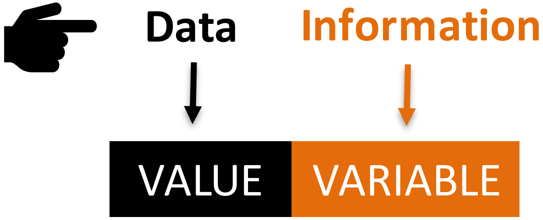
# Workflow of a Machine Learning Problem

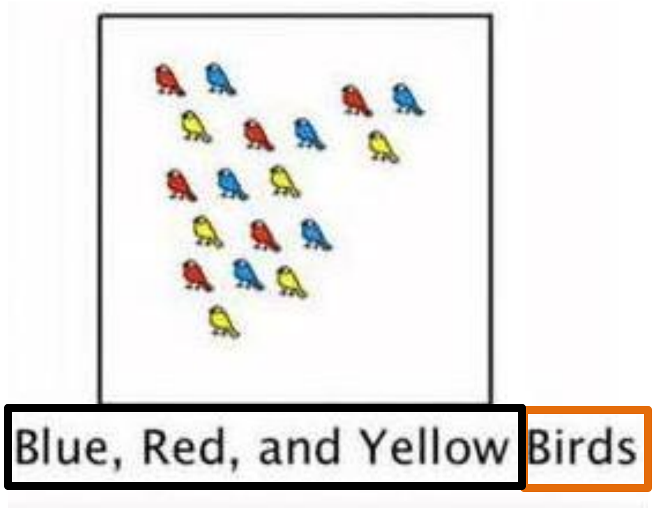


# Taxonomy of data variables

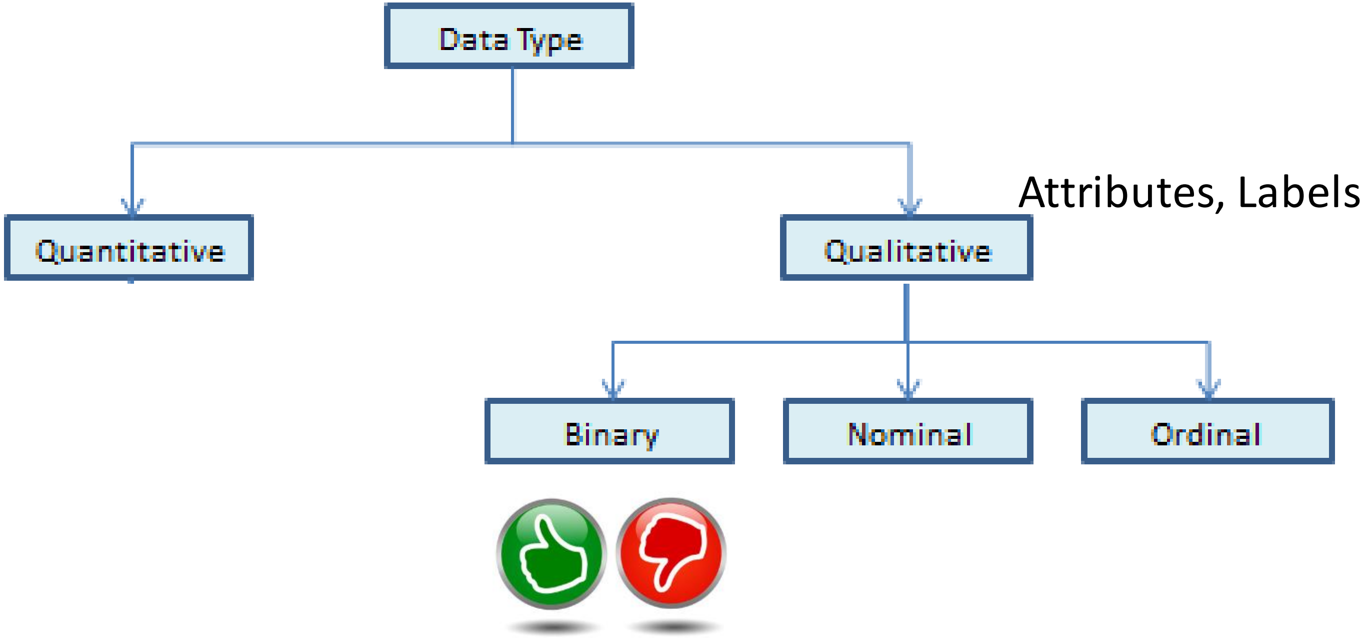






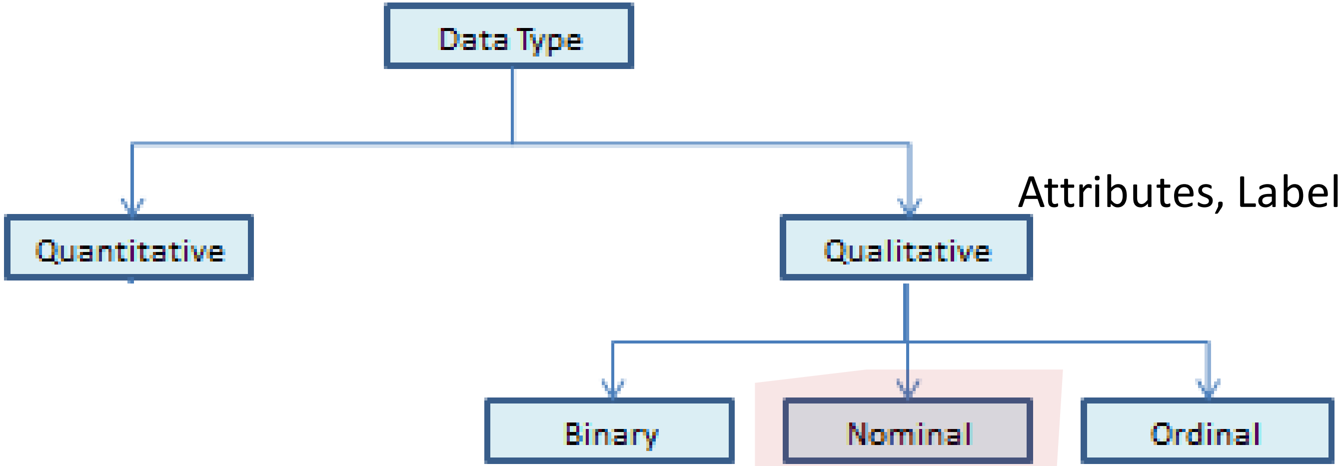



# Taxonomy of data

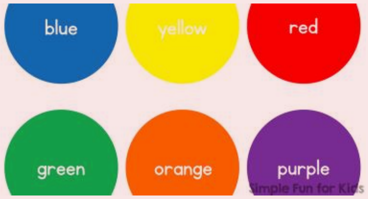



Preference

# Taxonomy of data



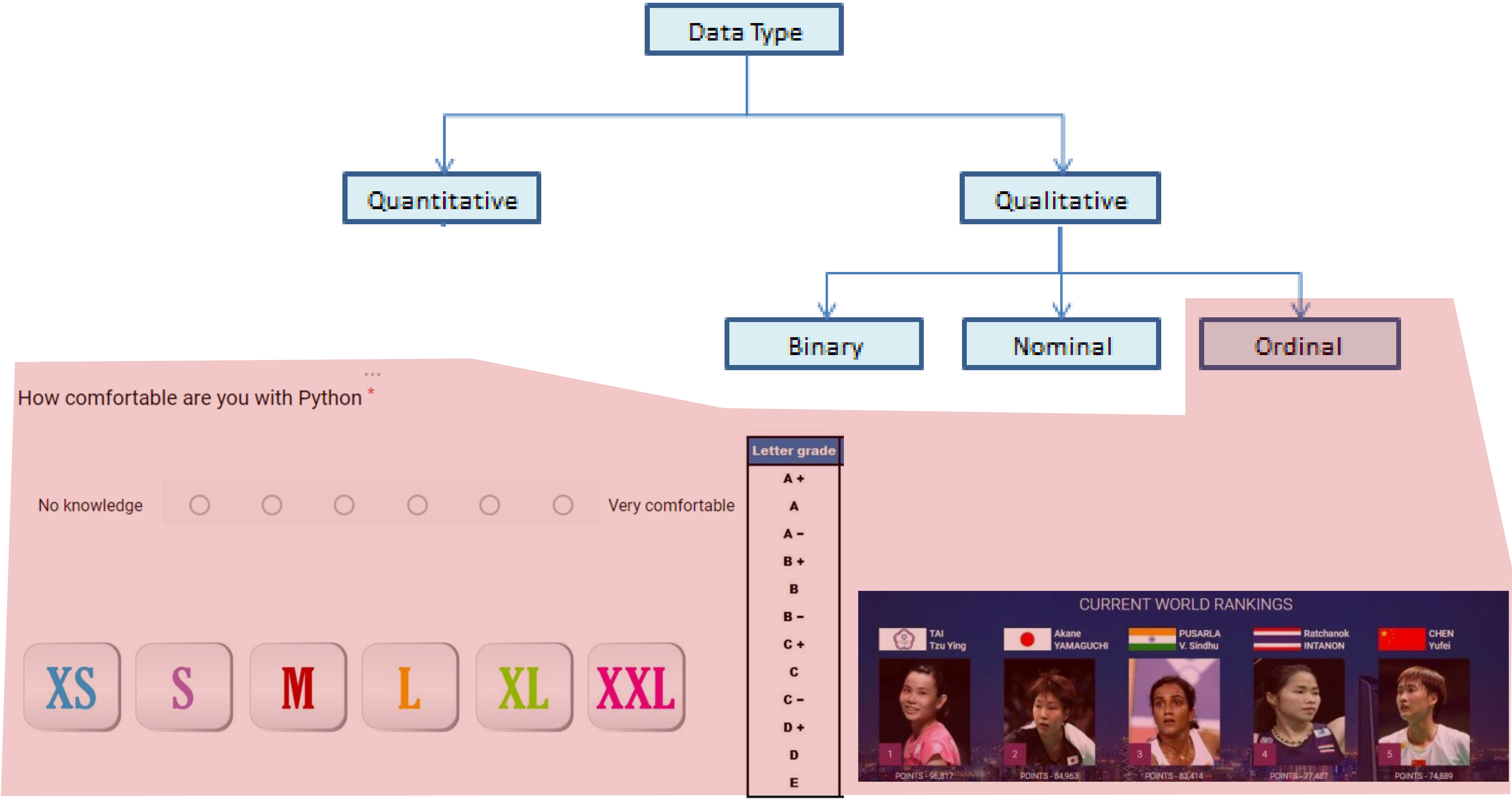

Attributes, Labels




Color

Make

Pin Code


Data Type
Quantitative
Qualitative
Binary
Nominal
Ordinal
How comfortable are you with Python *
No knowledge
Very comfortable
Letter grade
A +
A
A –
B +
B
B –
C +
C
C –
D +
D
E
XS
S
M
L
XL
XXL
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

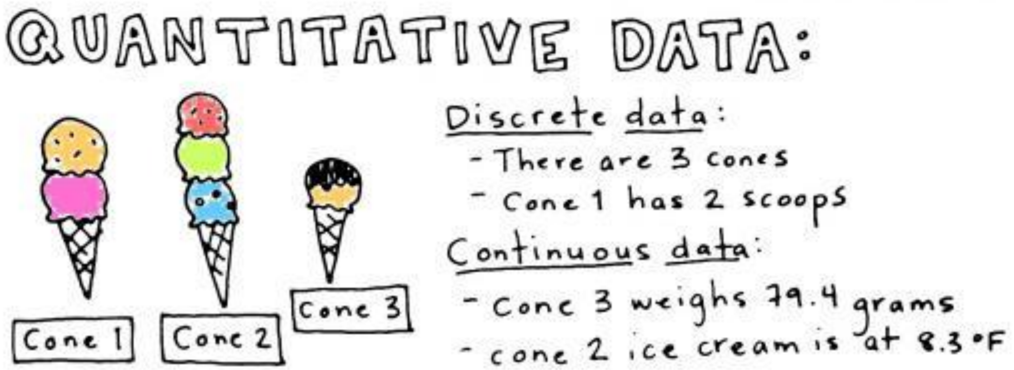
Data Type

Quantitative

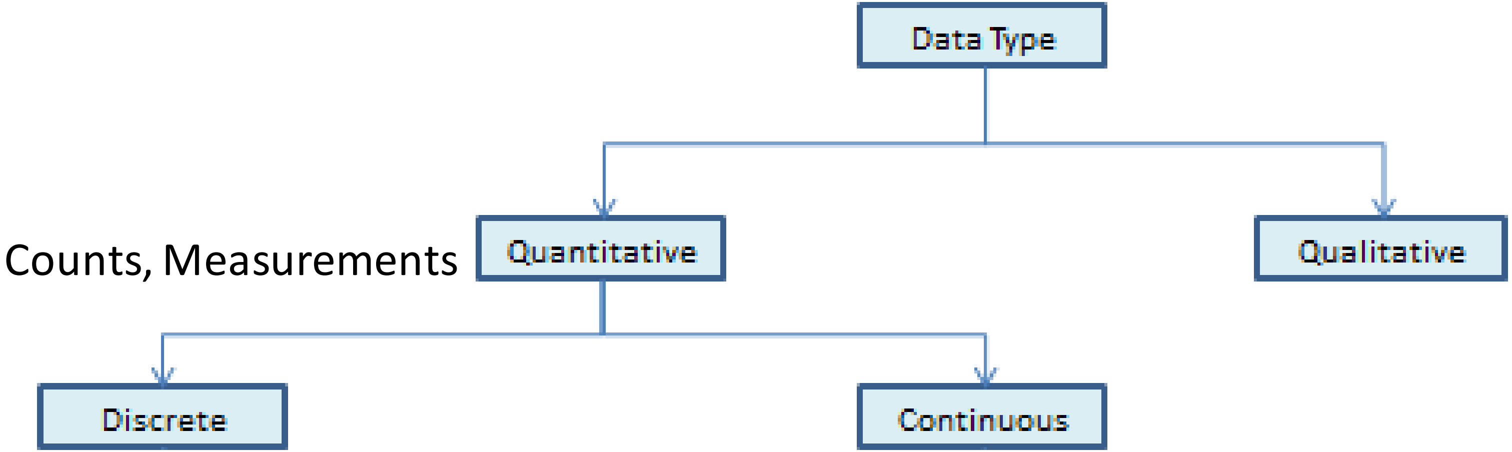
Counts, Measurements

Qualitative

Attributes, Labels



# Taxonomy of data




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

# Taxonomy of data

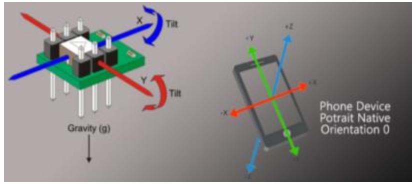
Data Type

Counts, Measurements — Quantitative

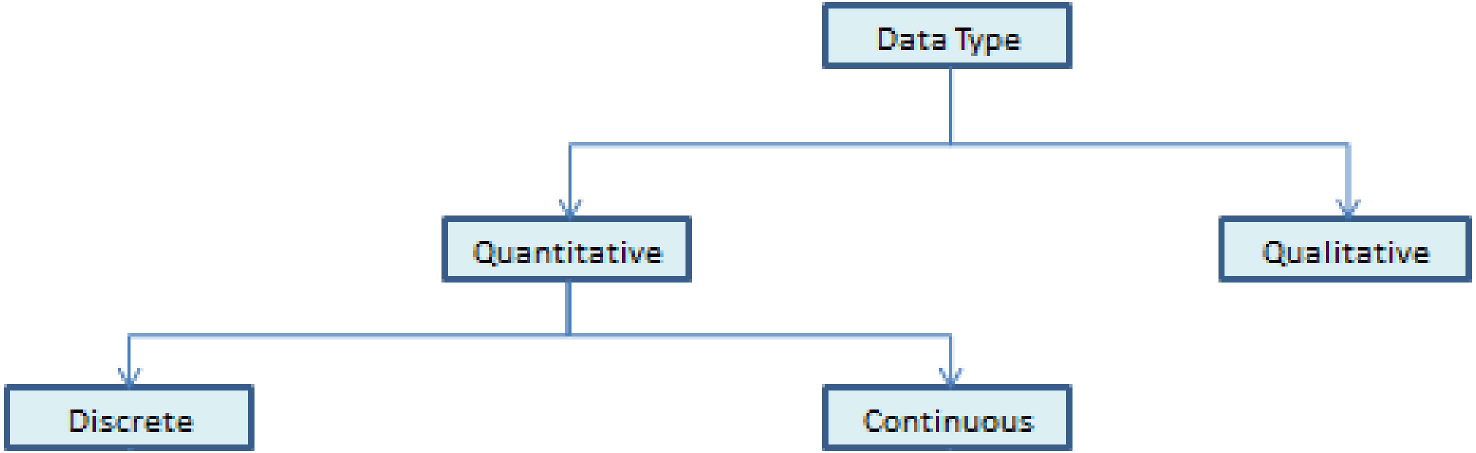
Qualitative

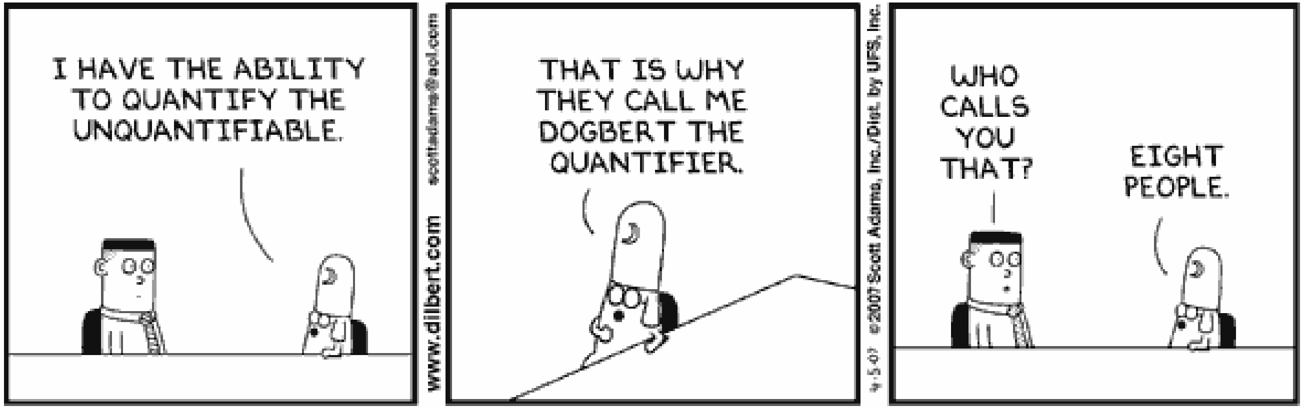
Discrete

Continuous

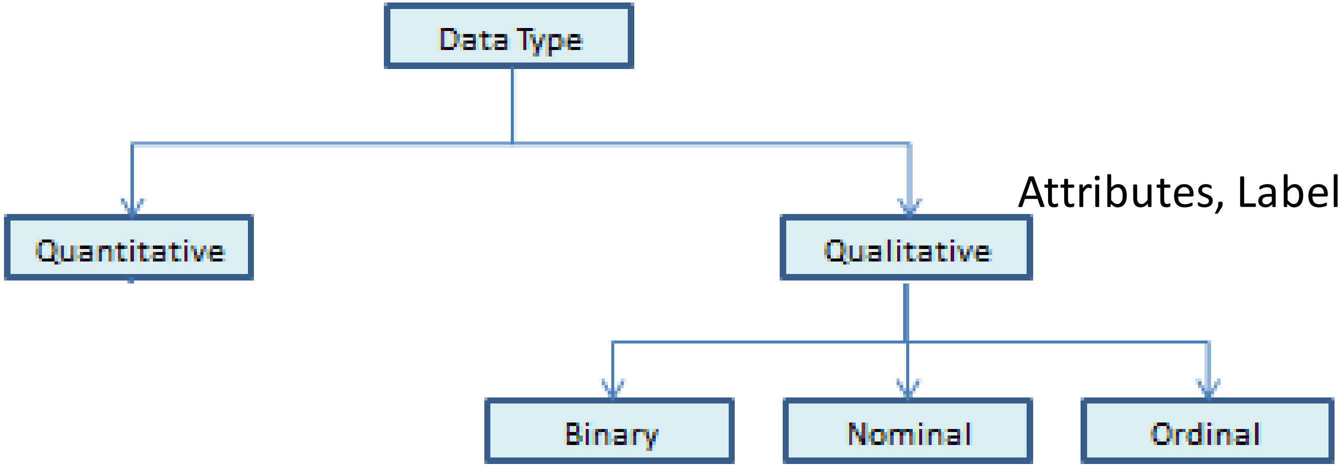


# Ultimately, all data needs to be quantitative



# Taxonomy of data: Qualitative → Quantitative

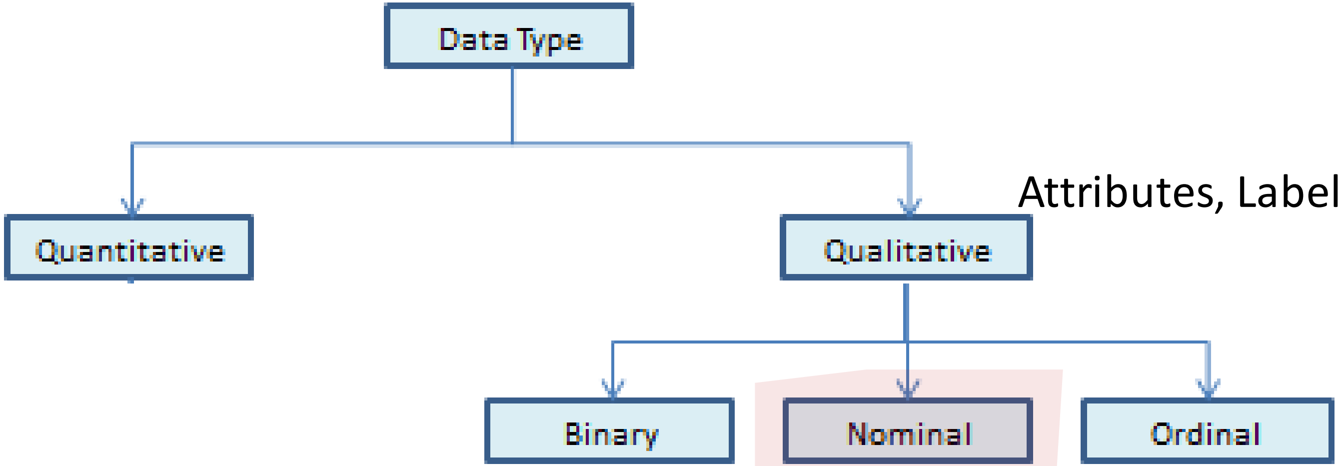



Attributes, Labels

0/1 ; -1/1

Preference

# Taxonomy of data: Qualitative → Quantitative



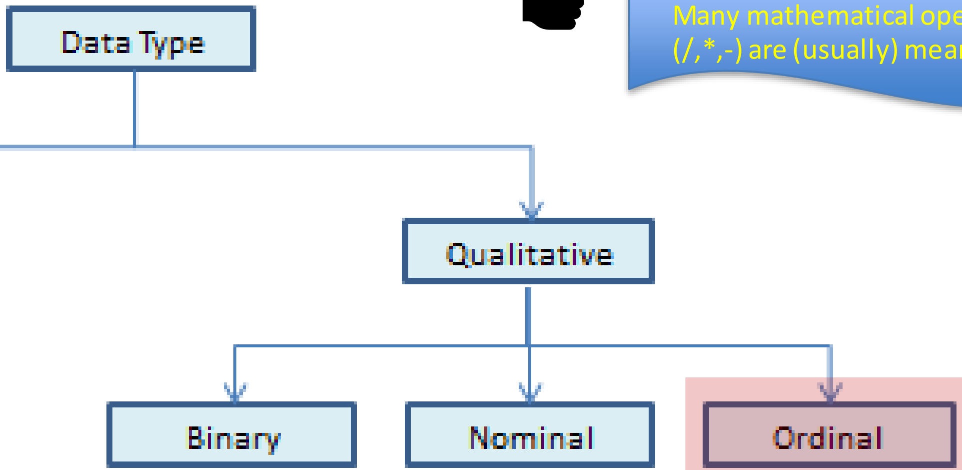

Attributes, Labels


Color

Make

Pin Code

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 |

1 2 3 4 5

# 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

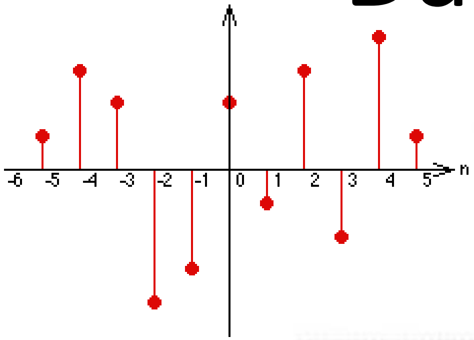
Does not obey data distribution

Conflict

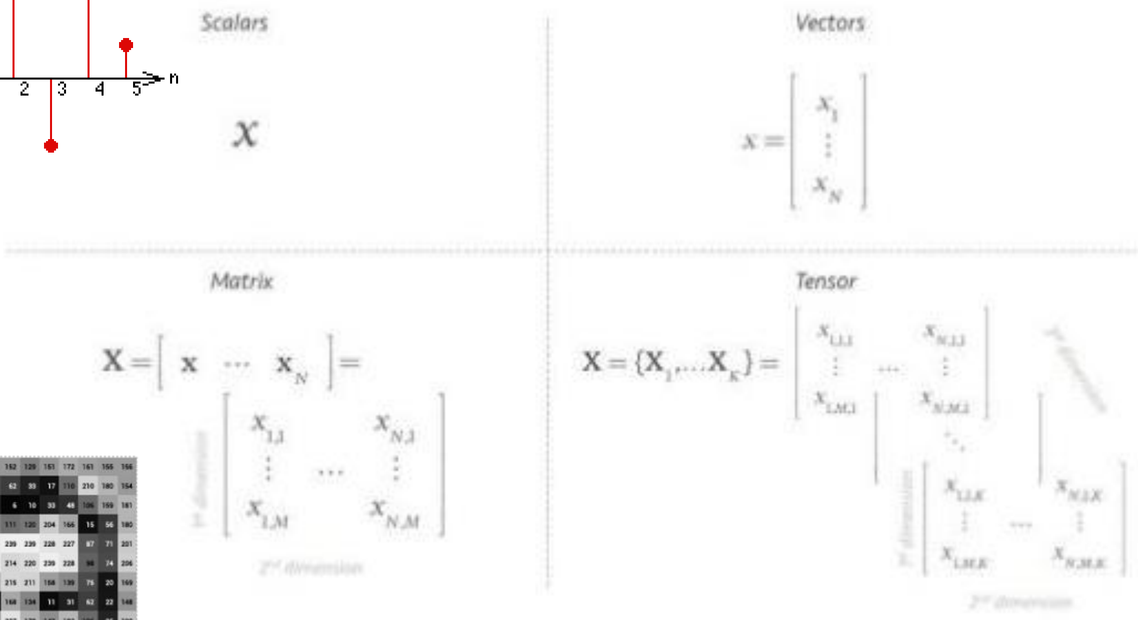
# Data imputation

- Approaches that aim to estimate missing data

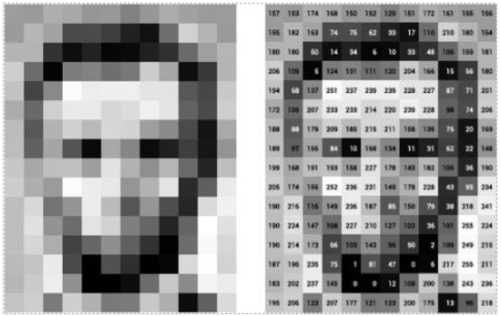
# Data Representations

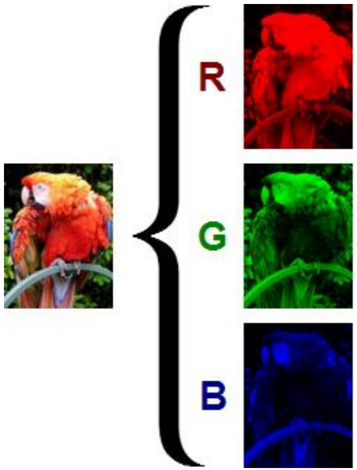




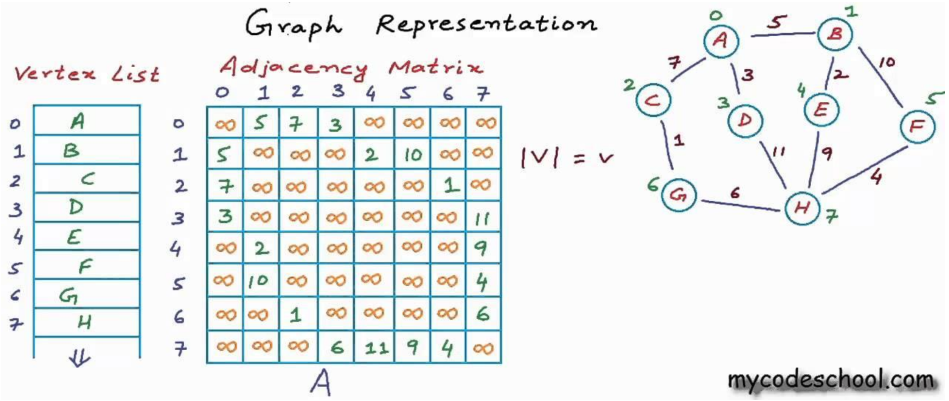



2-d image



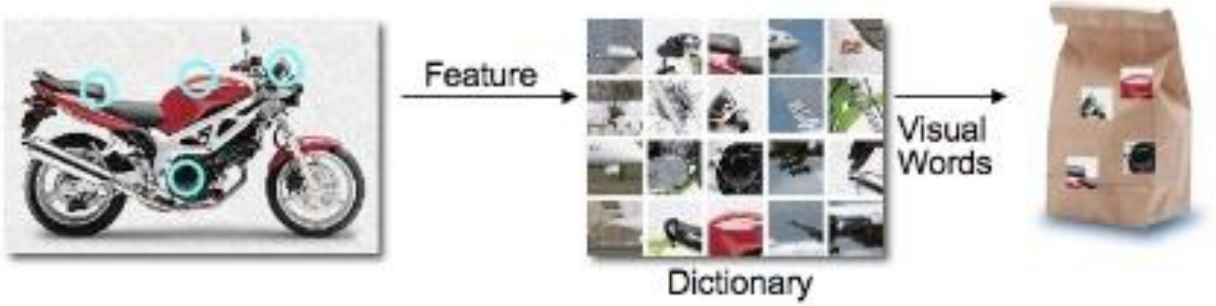


# Data Representations



# 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.

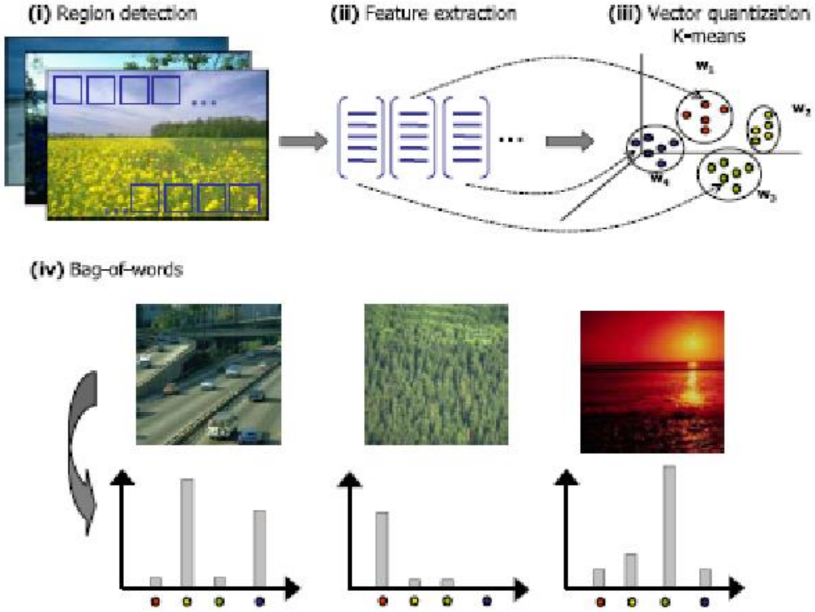


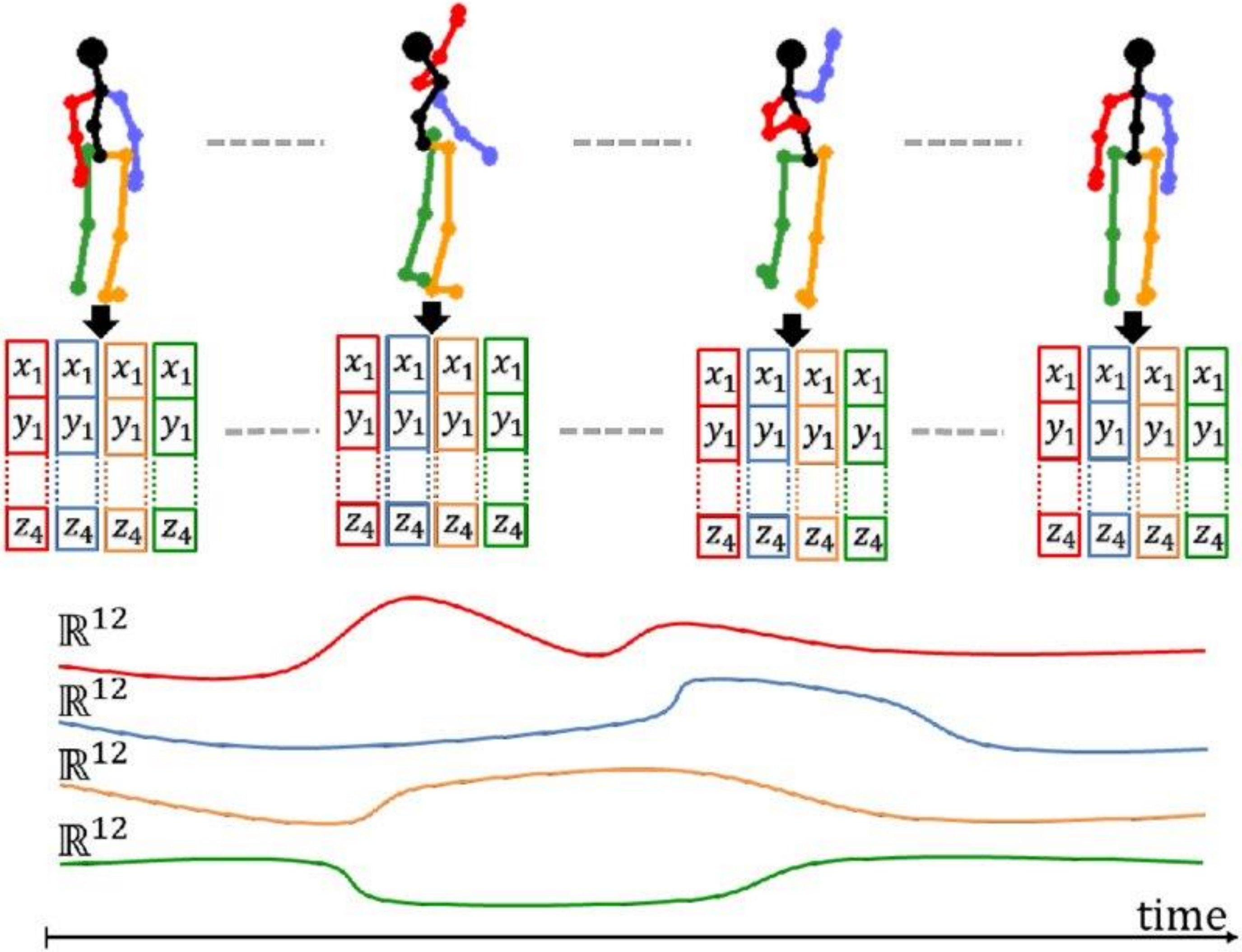

- Examples
  - Construct bag-of-words vector from an email
  - Remove stopwords in a sentence
  - Apply PCA projection to high-dimensional data

# The Bag of Words Representation

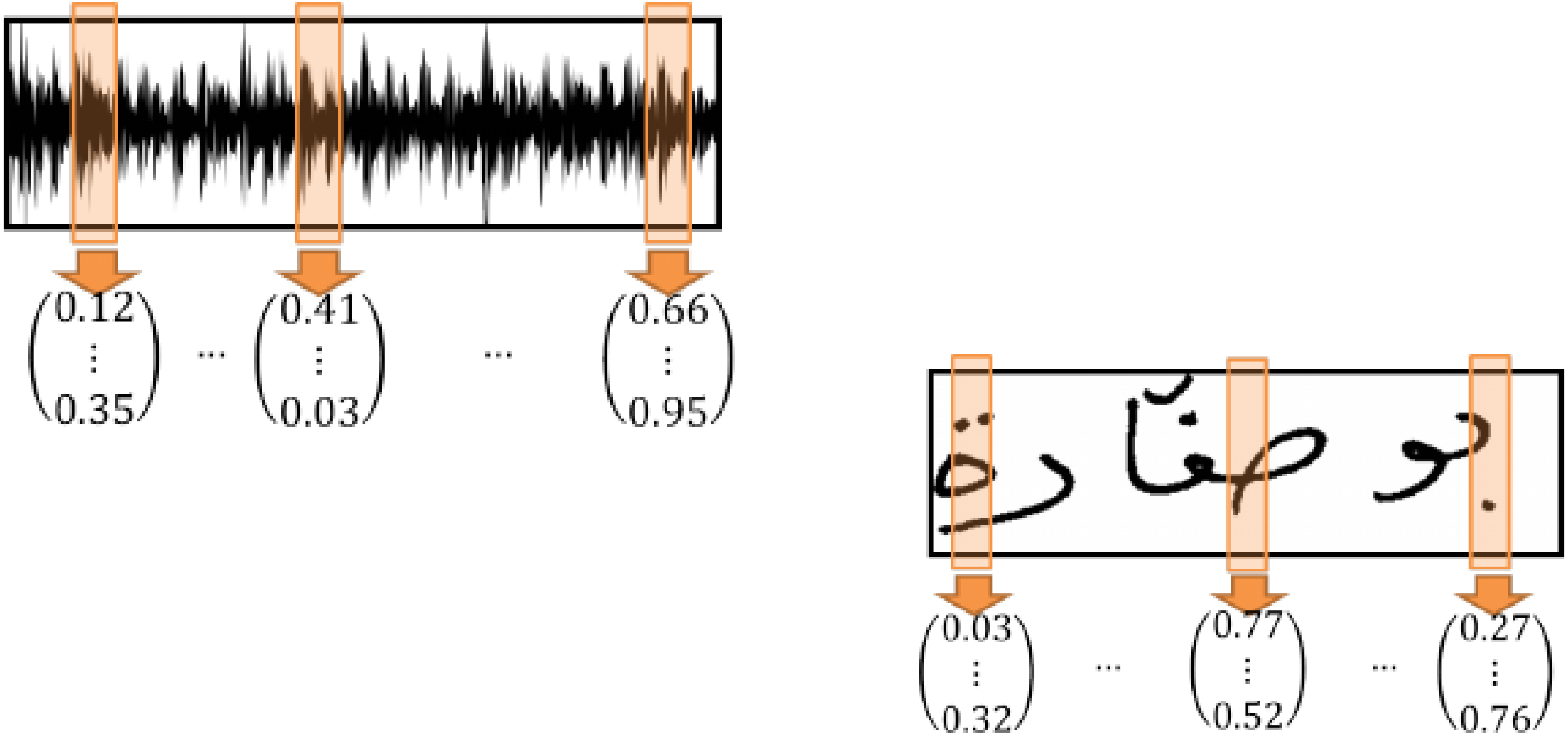
I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy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

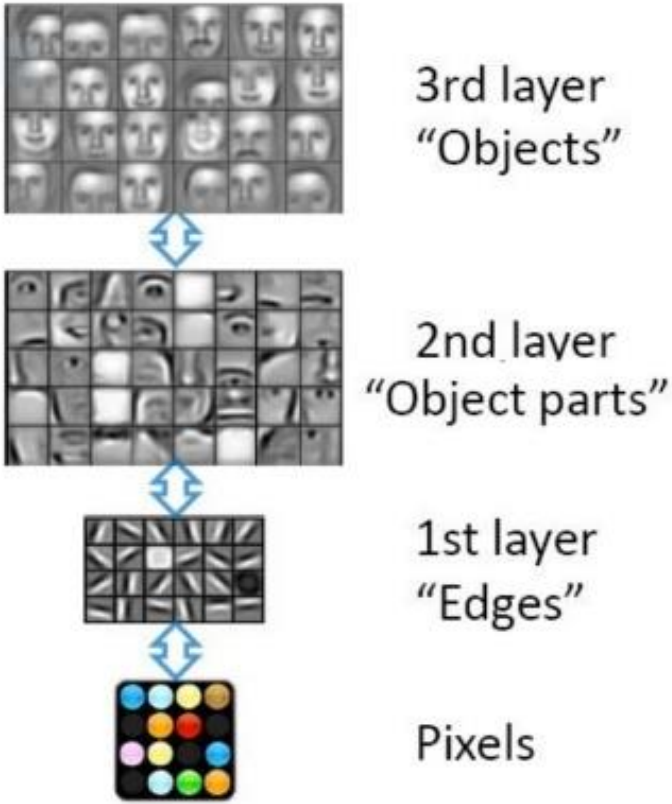


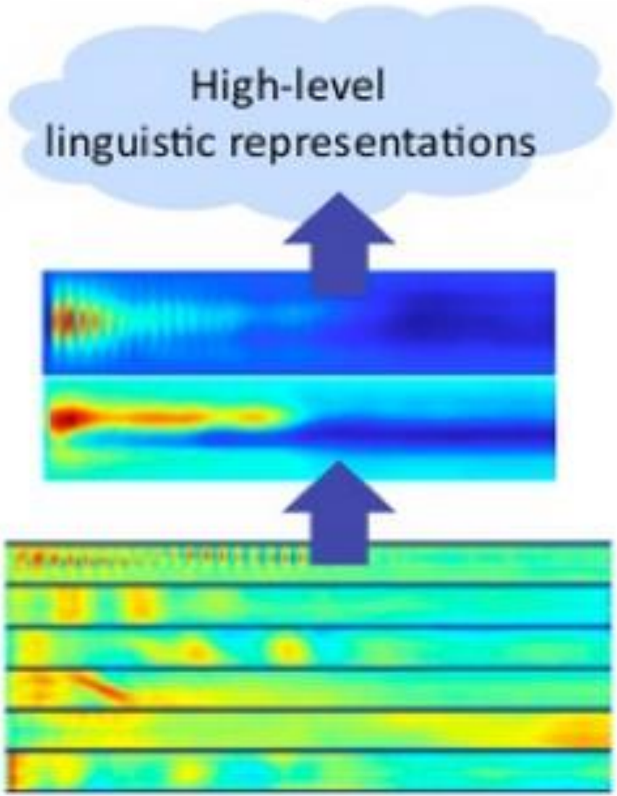


$x_1$
$y_1$
$z_4$
$\mathbb{R}^{12}$
time


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

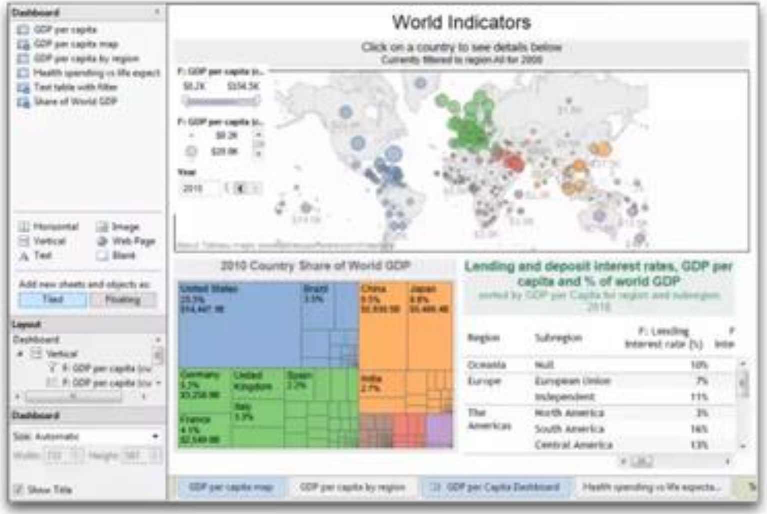
# Feature-based, Hierarchical Data Representations



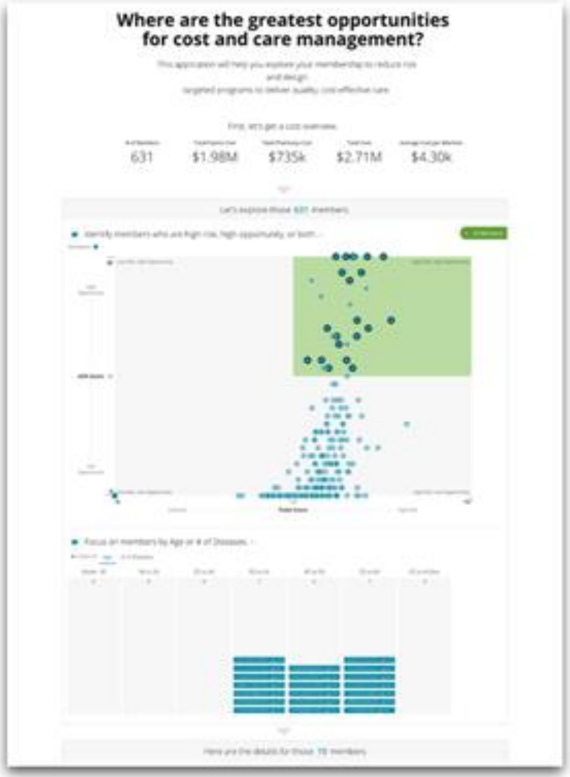



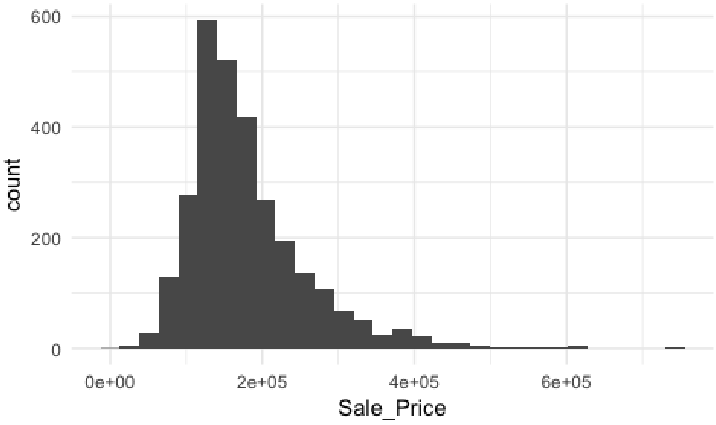
# Gazing at Data: Data visualization

## data exploration

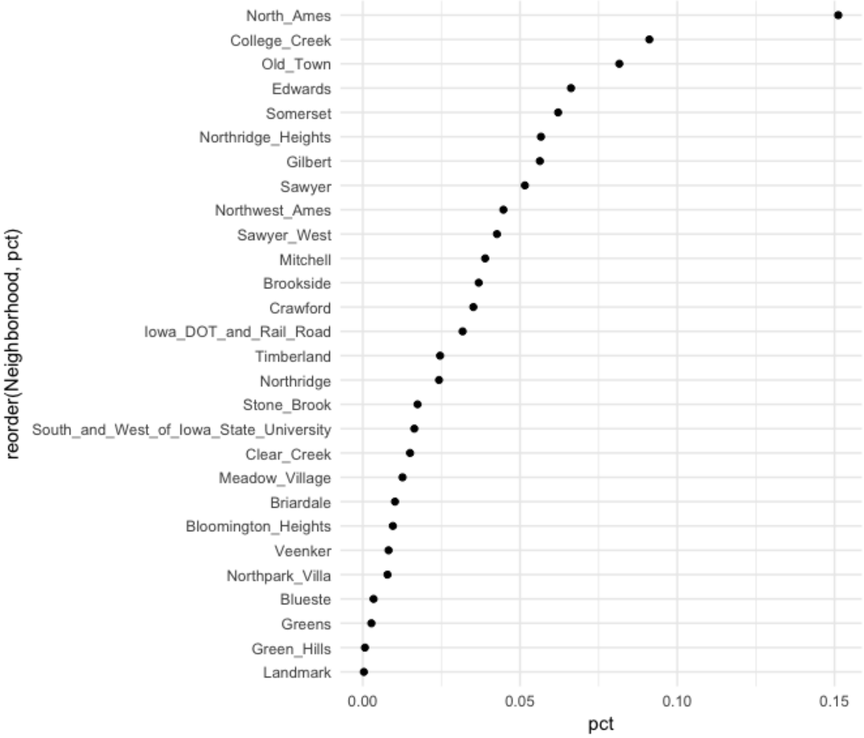



## data presentation

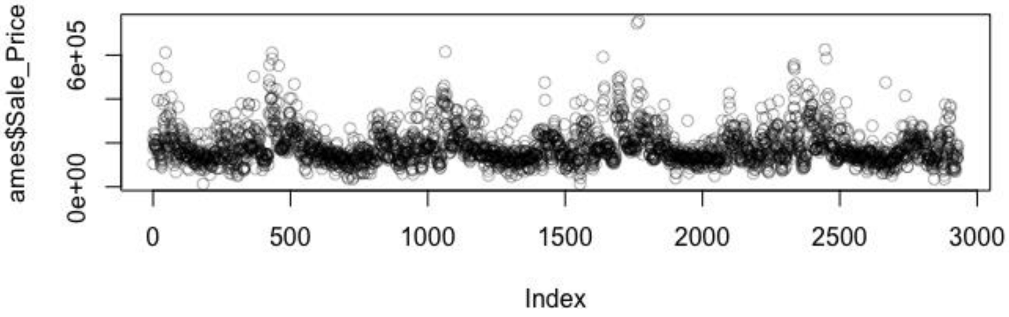




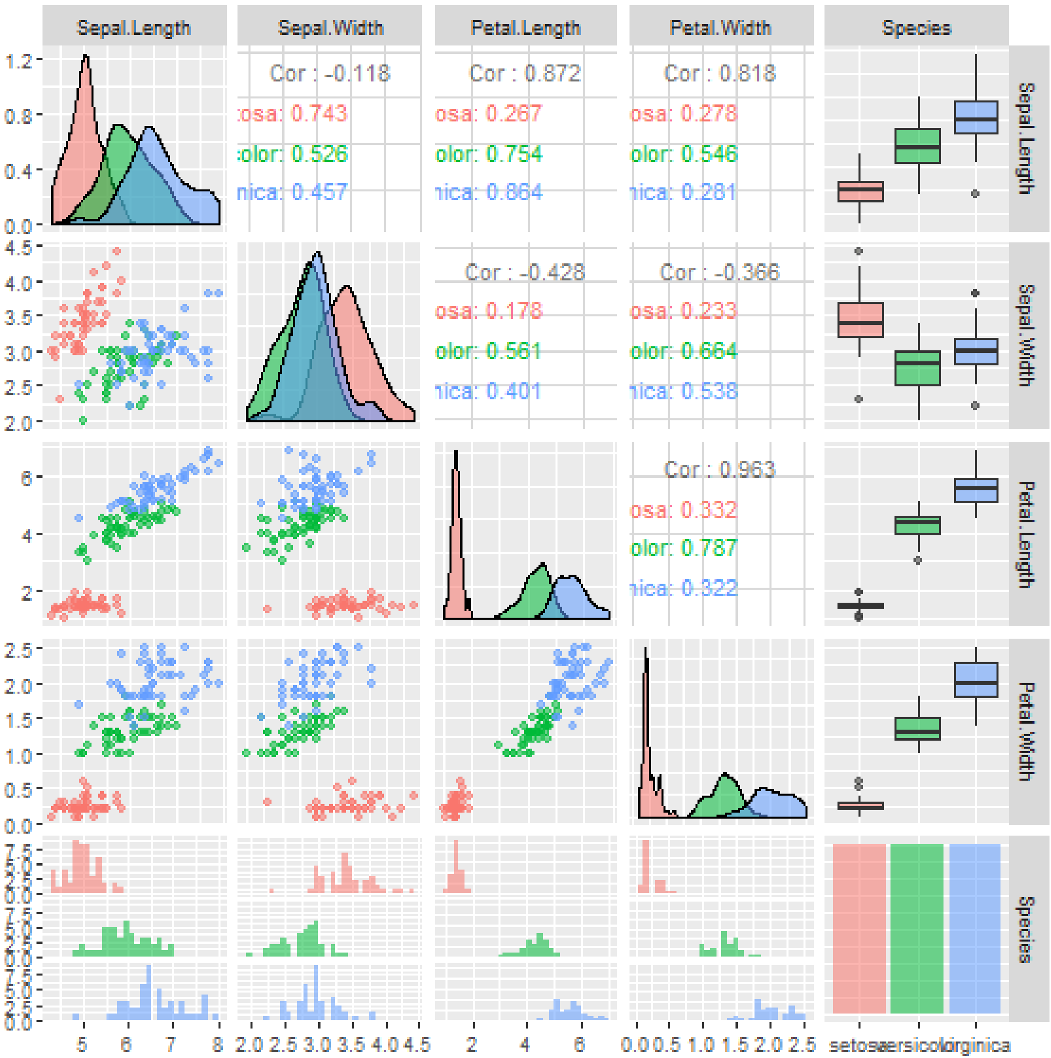
600
400
200
0
count
0e+00
2e+05
4e+05
6e+05
Sale_Price

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

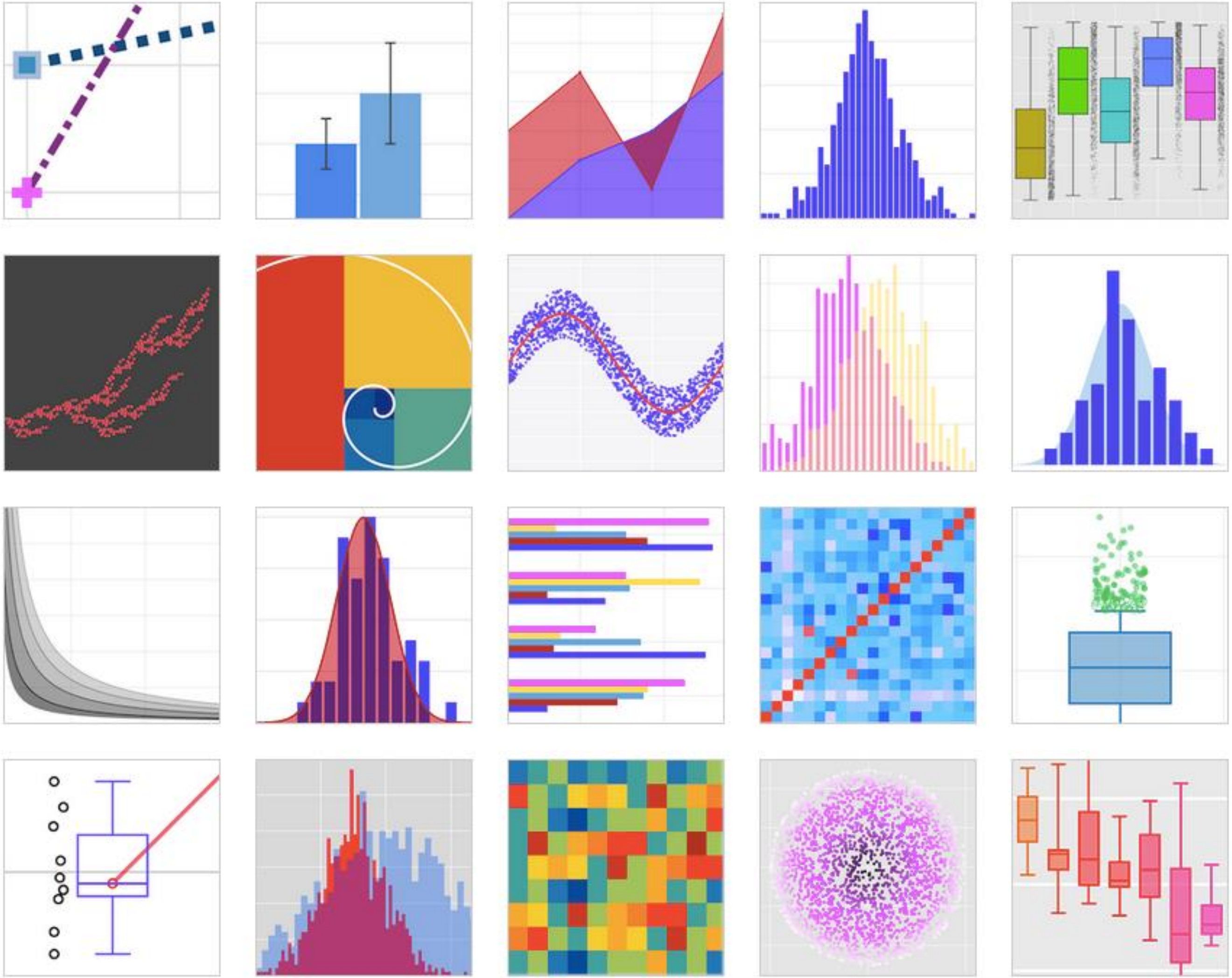


Sepal.Length
Sepal.Width
Petal.Length
Petal.Width
Species
Cor : -0.118
osa: 0.743
olor: 0.526
nica: 0.457
Cor : 0.872
osa: 0.267
olor: 0.754
nica: 0.864
Cor : 0.818
osa: 0.278
olor: 0.546
nica: 0.281
Cor : -0.428
osa: 0.178
olor: 0.561
nica: 0.401
Cor : -0.366
osa: 0.233
olor: 0.664
nica: 0.538
Cor : 0.963
osa: 0.332
olor: 0.787
nica: 0.322
setosa
versicolor
virginica

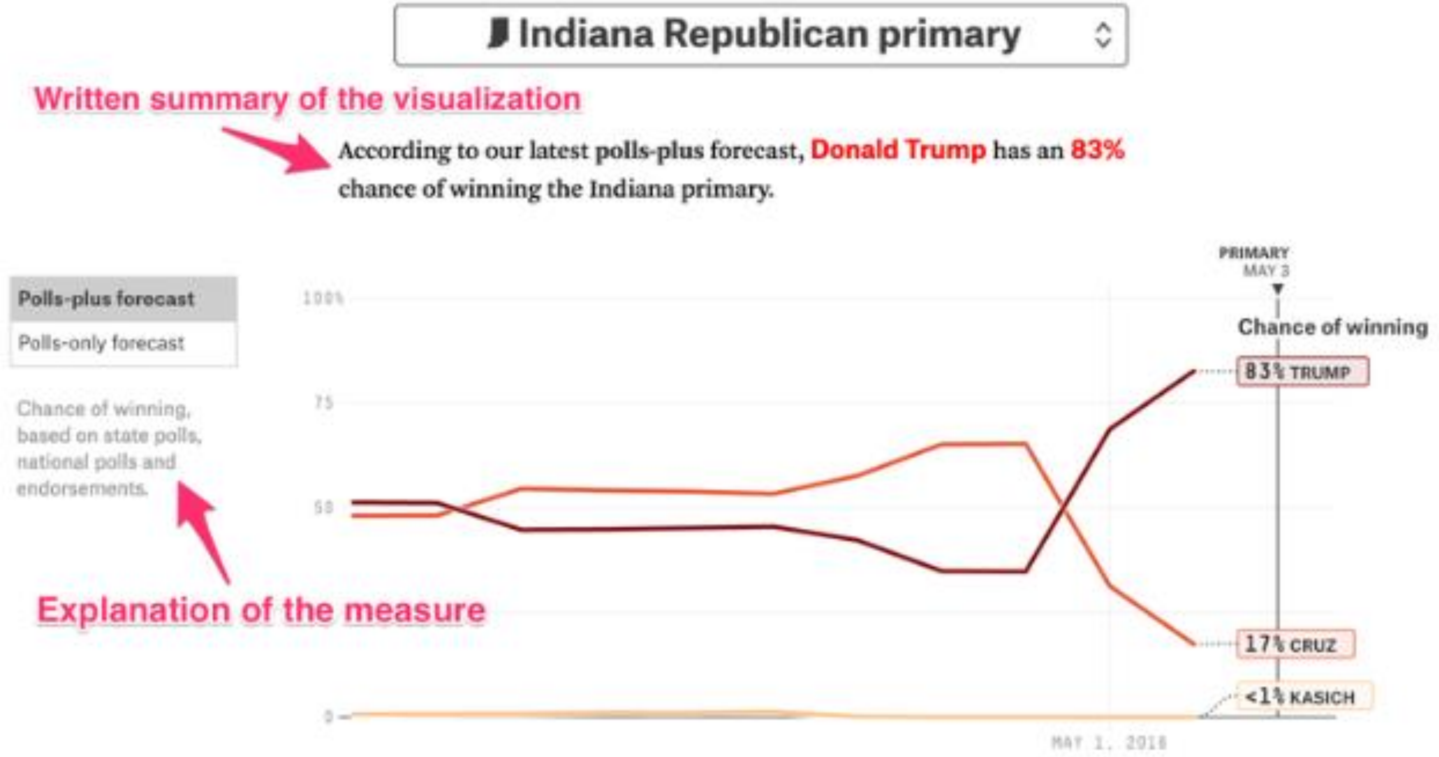
# Data visualization

## treemap


## leaderboard



# Data Exploration and Visualization




Fivethiryeight 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*

VISAGE.CO

# 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

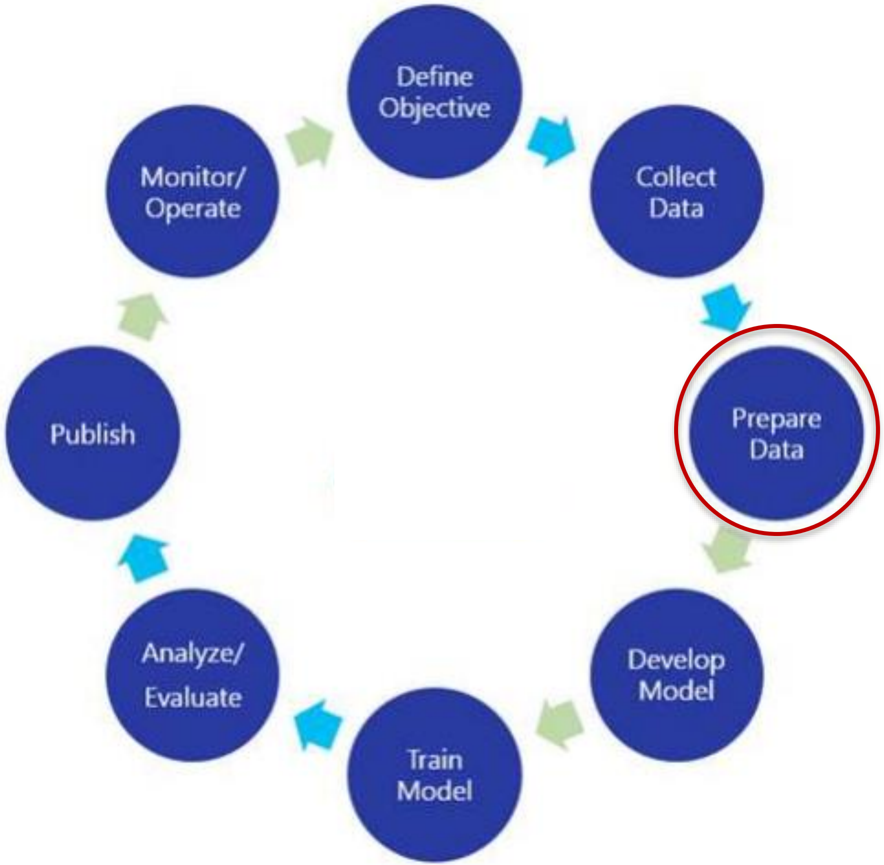
| Cell | Possible Values | Probability |
|---|---|---|
| t2.Zip | 60608 | 0.84 |
| | 60609 | 0.16 |
| t4.City | Chicago | 0.95 |
| | Cicago | 0.05 |
| t4.DBAName | John Veliotis Sr. | 0.99 |
| | Johnnyo's | 0.01 |

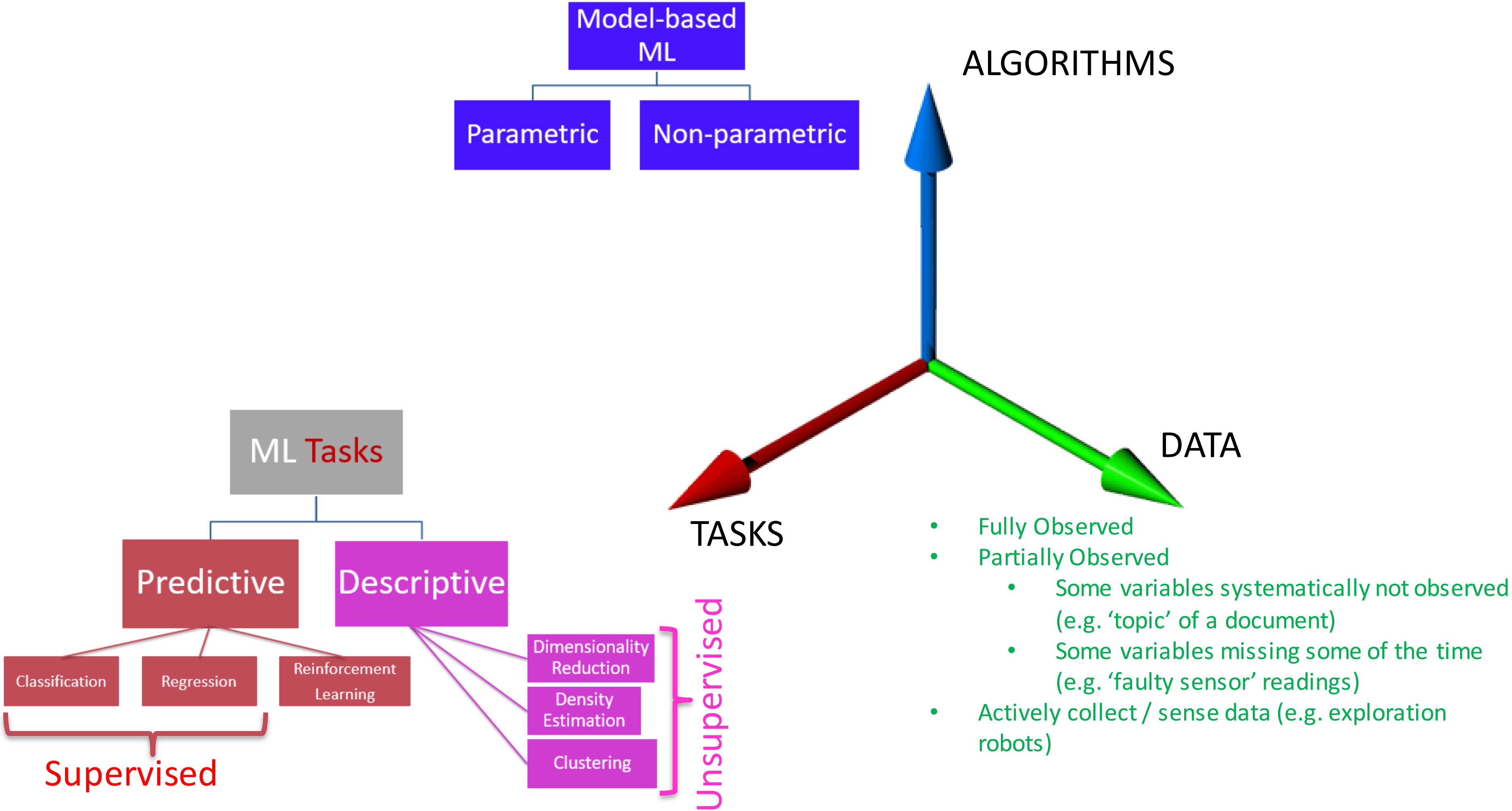
# Other important aspects of data

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

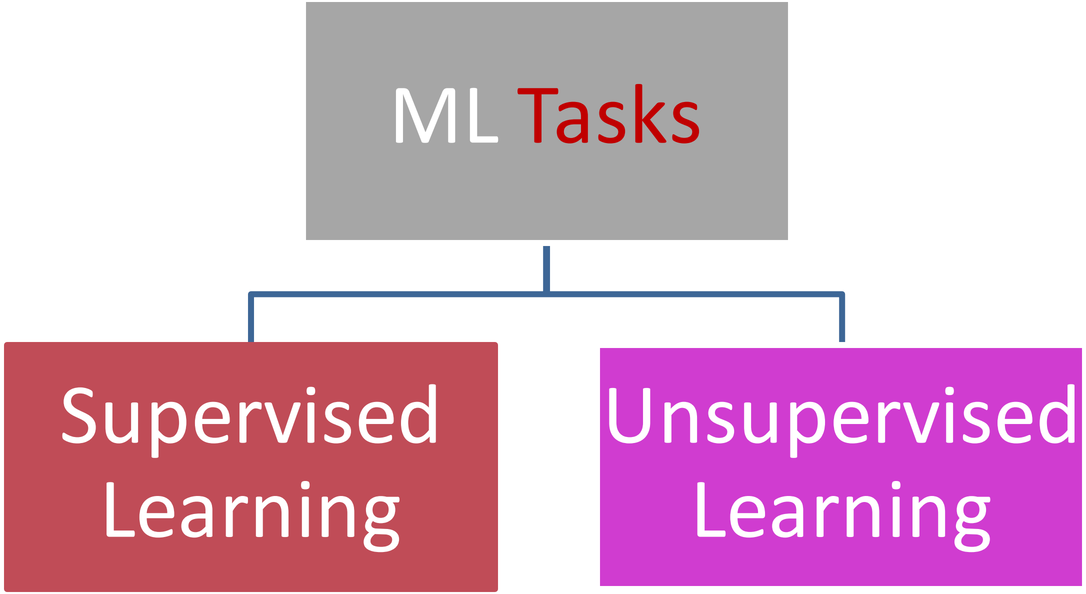
- Statistical assumptions on data
  - i.i.d (independent and identically distributed)
  - Online (e.g. time-series data)

# Workflow of a Machine Learning Problem





Model-based ML
Parametric
Non-parametric
ALGORITHMS
ML Tasks
Predictive
Descriptive
Classification
Regression
Reinforcement Learning
Dimensionality Reduction
Density Estimation
Clustering
Supervised
Unsupervised
TASKS
DATA
• 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)




Given an input,
estimate output

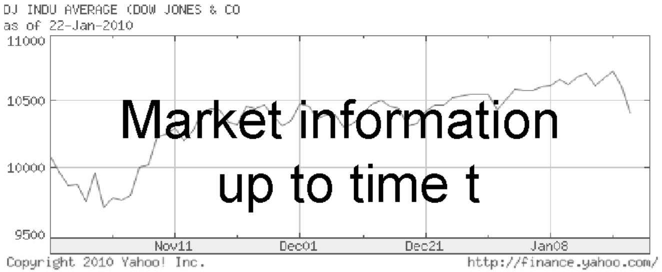
# ML::Tasks → Predictive

**Feature Space** $\mathcal{X}$

**Label Space** $\mathcal{Y}$

Words in a document

→ "Sports" "News" "Science" …

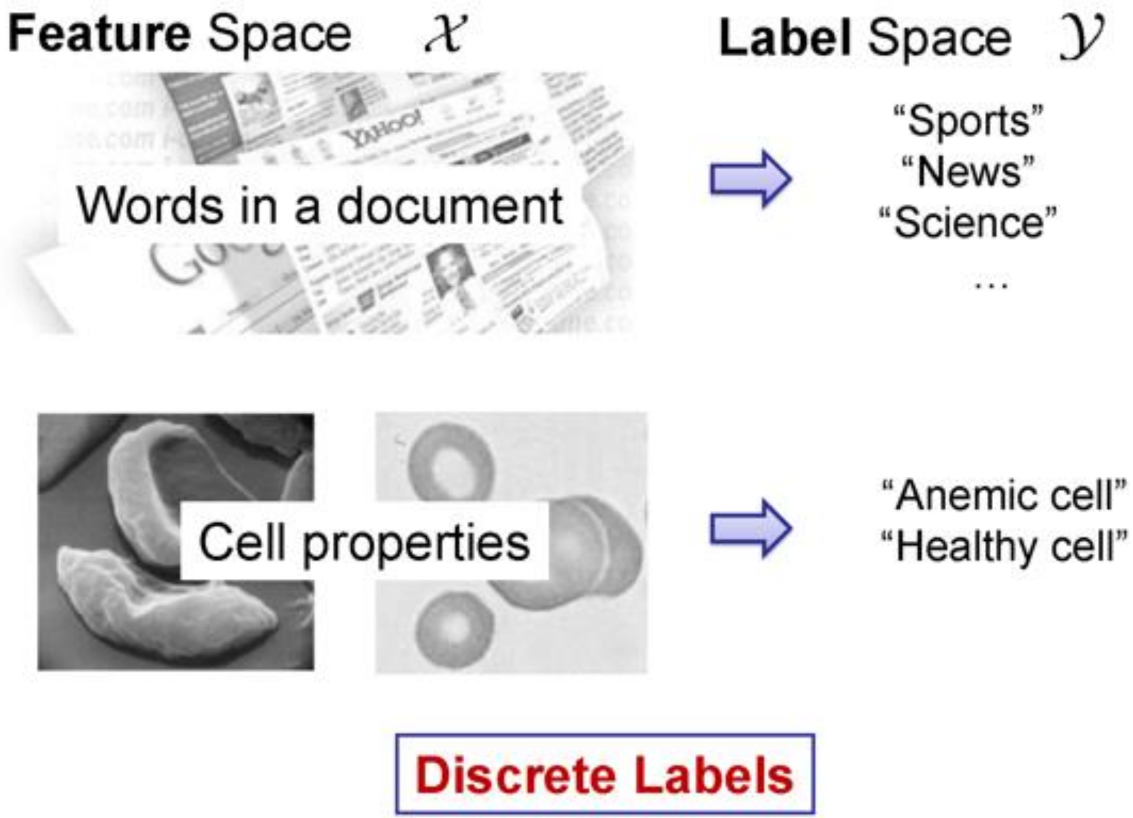



Market information up to time t

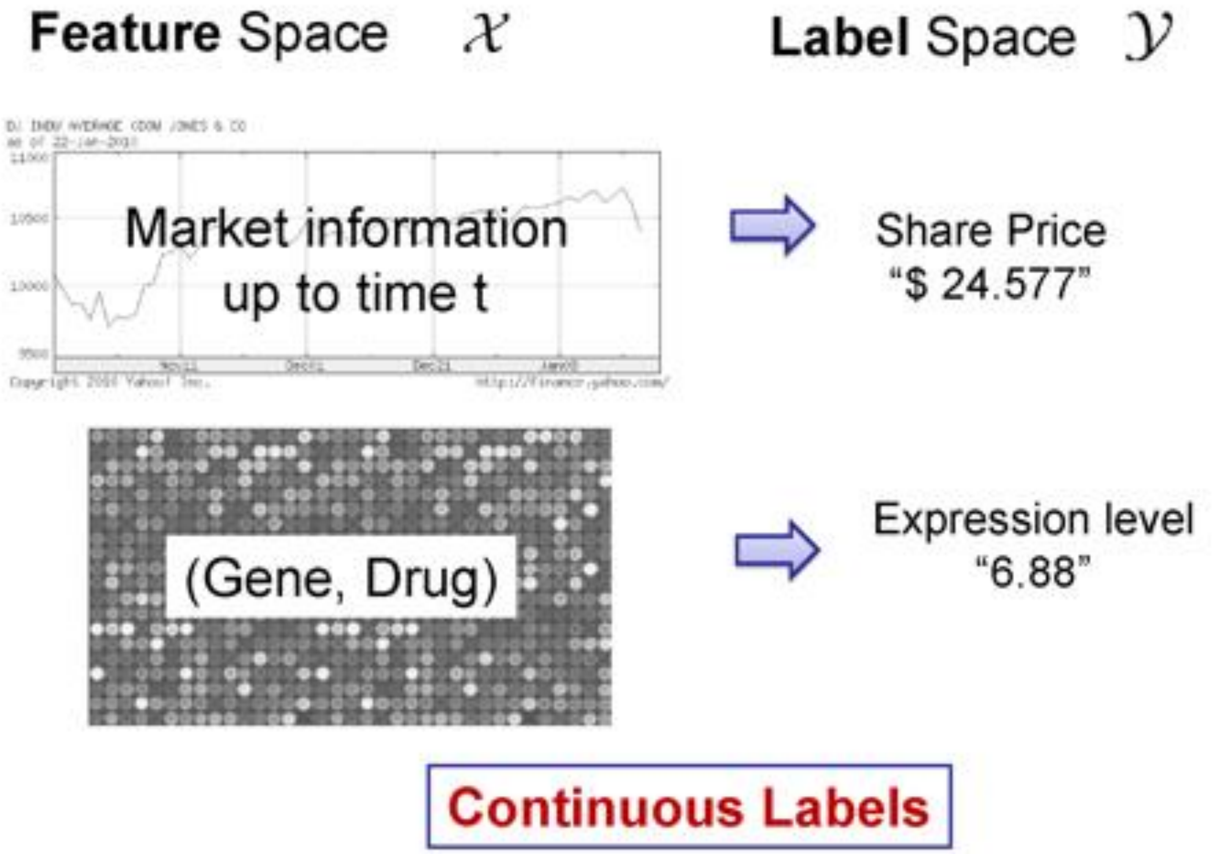
→ Share Price "$ 24.50"

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

# ML::Tasks → Predictive → Classification

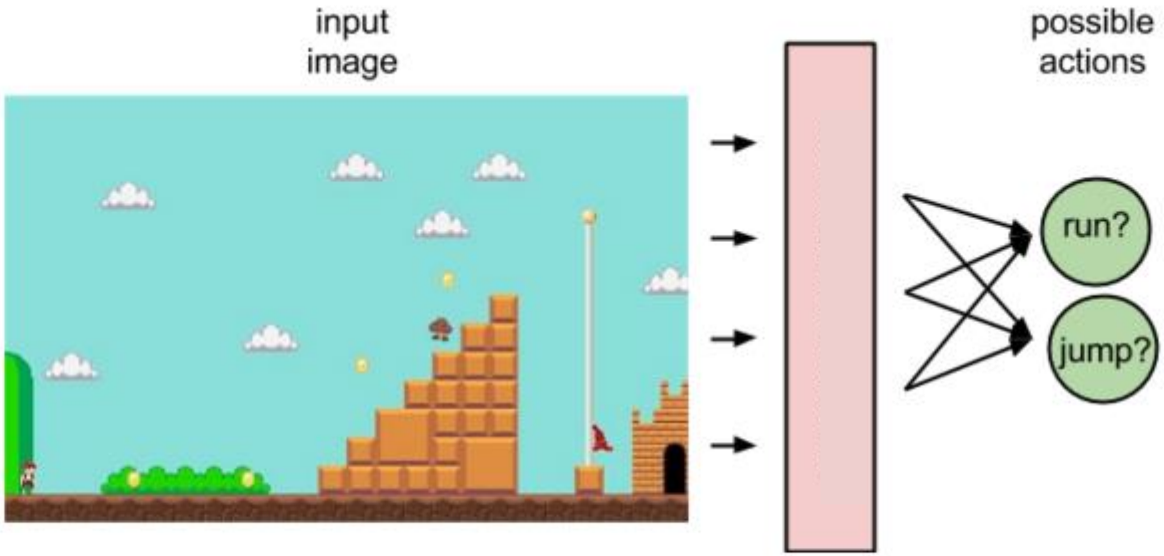


# ML::Tasks → Predictive → Regression



# ML::Tasks → Predictive → Reinforcement Learning

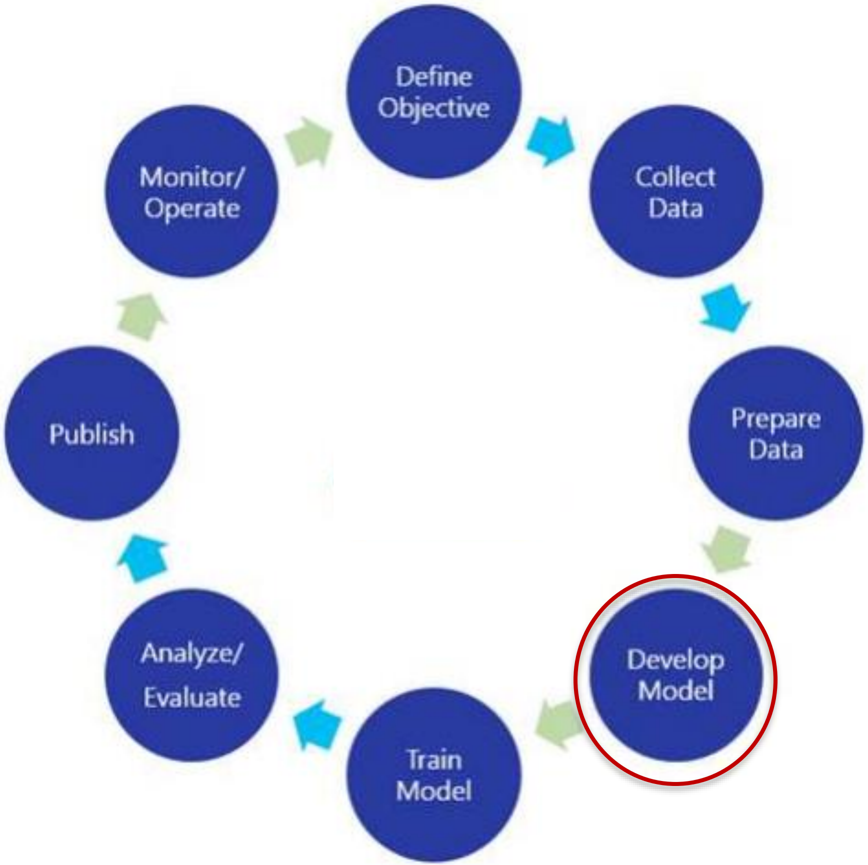
**Feature** Space $\mathcal{X}$ **Label** Space $\mathcal{Y}$

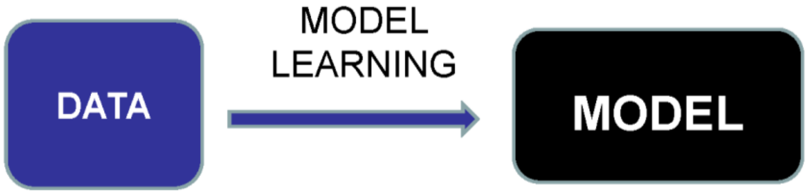


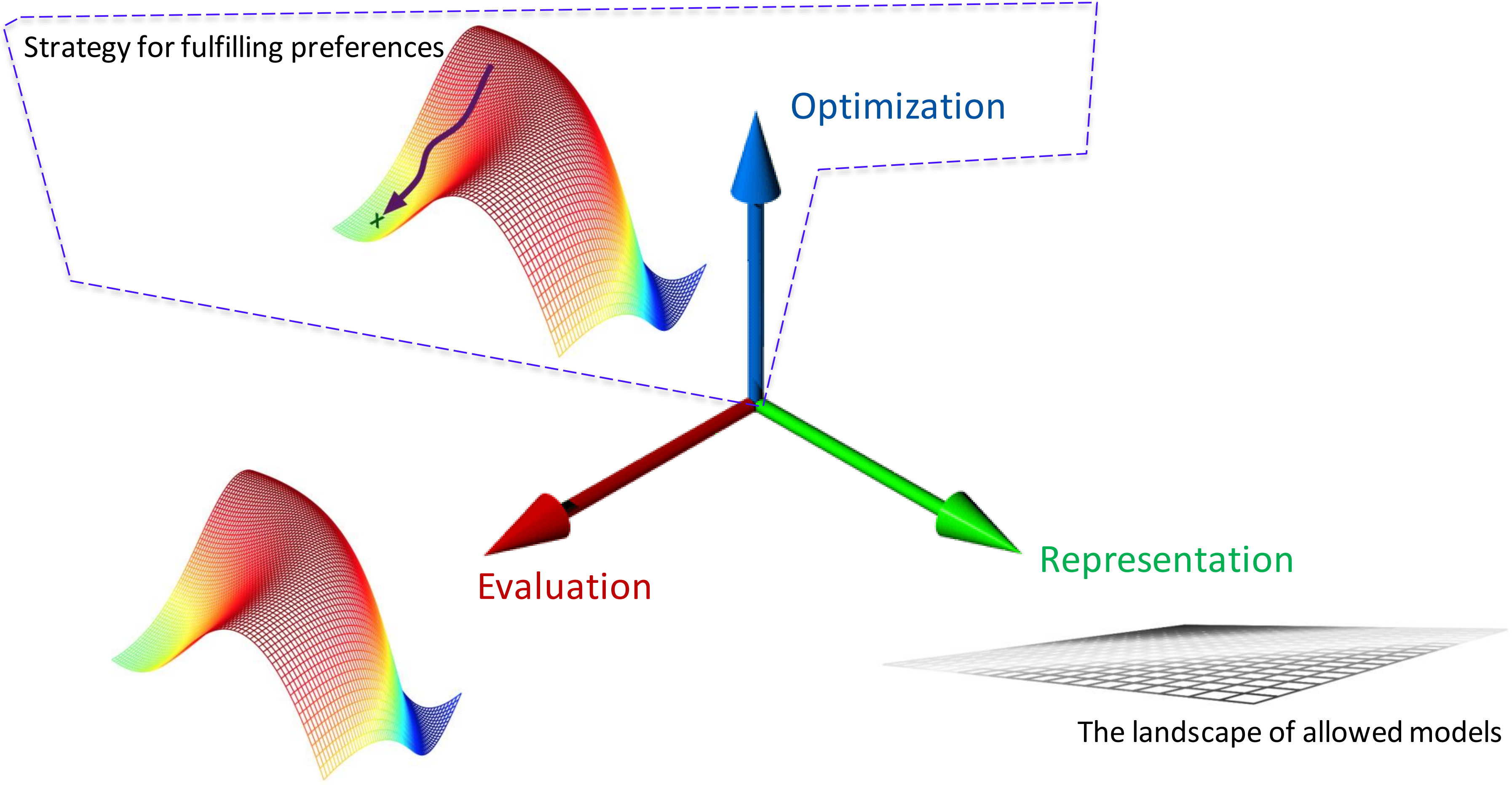
# ML::Tasks → Predictive → Reinforcement Learning

**Feature** Space $\mathcal{X}$ **Label** Space $\mathcal{Y}$

# Workflow of a Machine Learning Problem

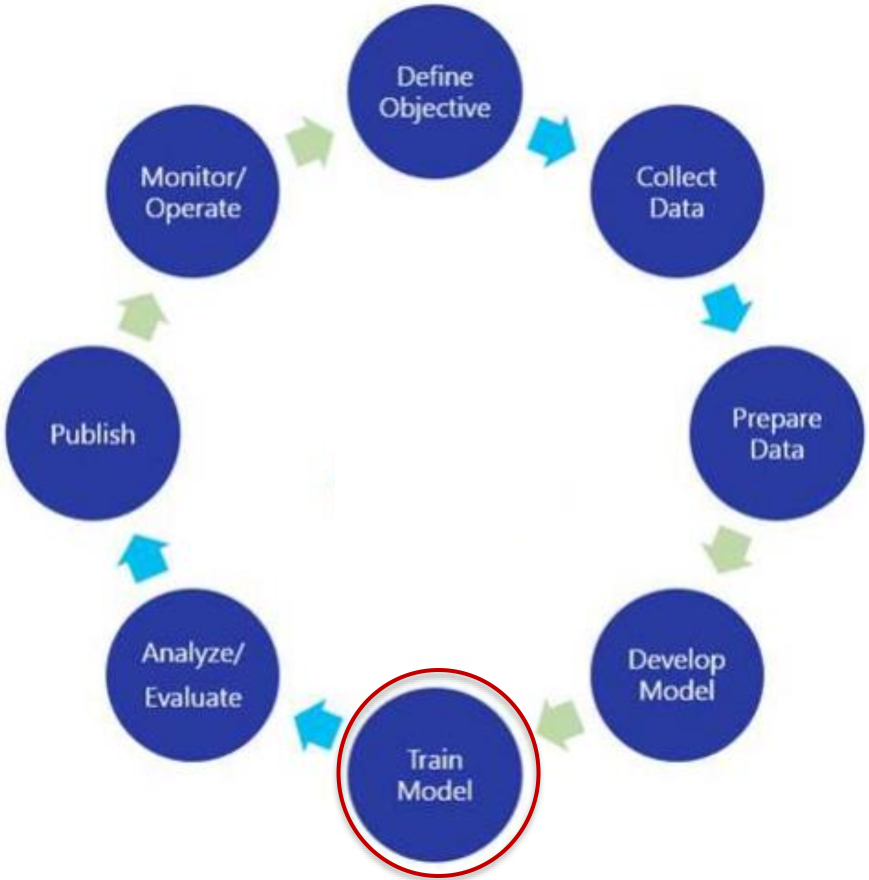


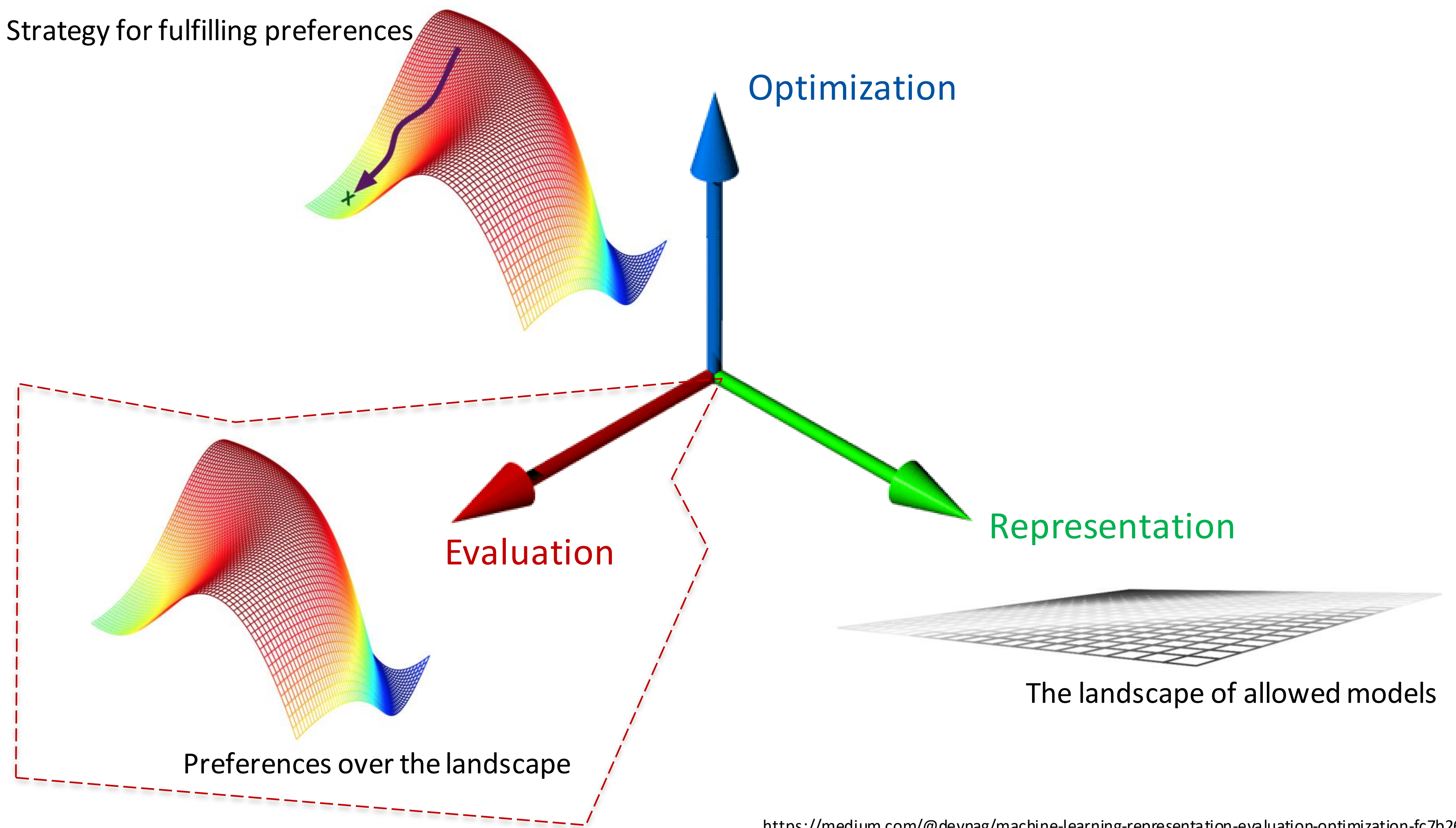






https://medium.com/@devnag/machine-learning-representation-evaluation-optimization-fc7b26b38fdb

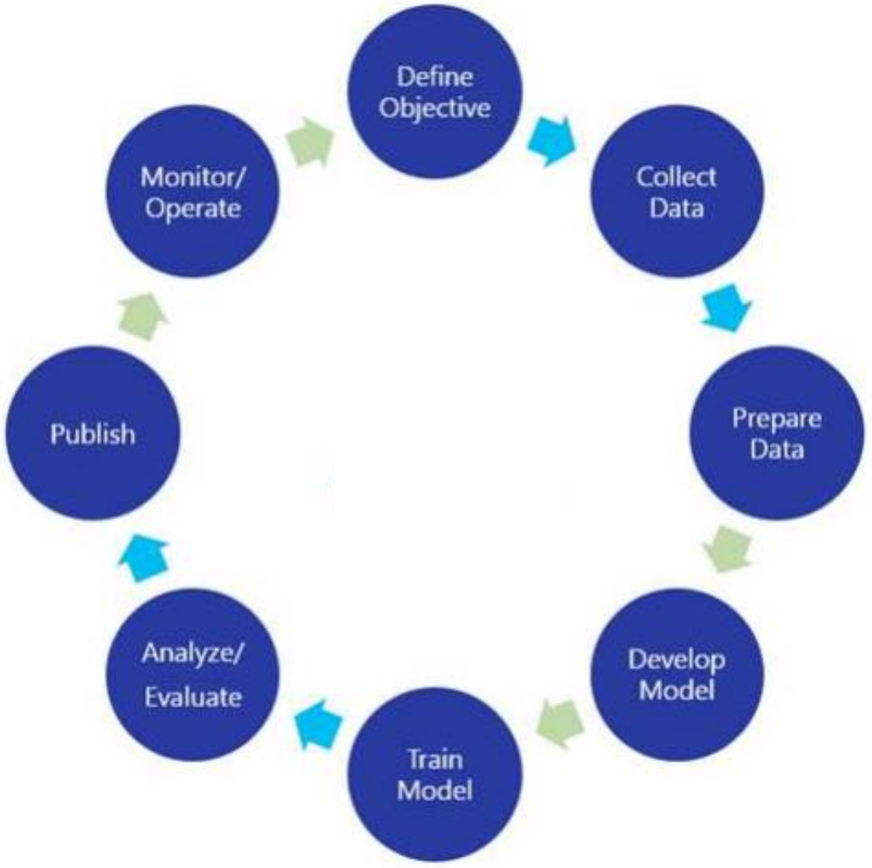
# Workflow of a Machine Learning Problem




https://medium.com/@devnag/machine-learning-representation-evaluation-optimization-fc7b26b38fdb

# Workflow of a Machine Learning Problem




ML Tasks

Supervised Learning

Given an input,
estimate output

Unsupervised Learning

# Supervised Learning

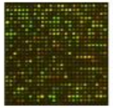
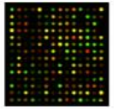
# Data

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

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

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

$x_1 =$  $x_2 =$  $\ldots$

$x_1 =$  $x_2 =$  $\ldots$

$$x_1 = \begin{bmatrix} 5 \\ 1 \\ 22 \\ \vdots \end{bmatrix} \qquad x_2 = \begin{bmatrix} 1 \\ 0 \\ 17 \\ \vdots \end{bmatrix} \quad \begin{array}{l} \text{\# cigarettes/day} \\ \text{\# drinks/day} \\ \text{BMI} \\ \phantom{\vdots} \end{array}$$

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

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

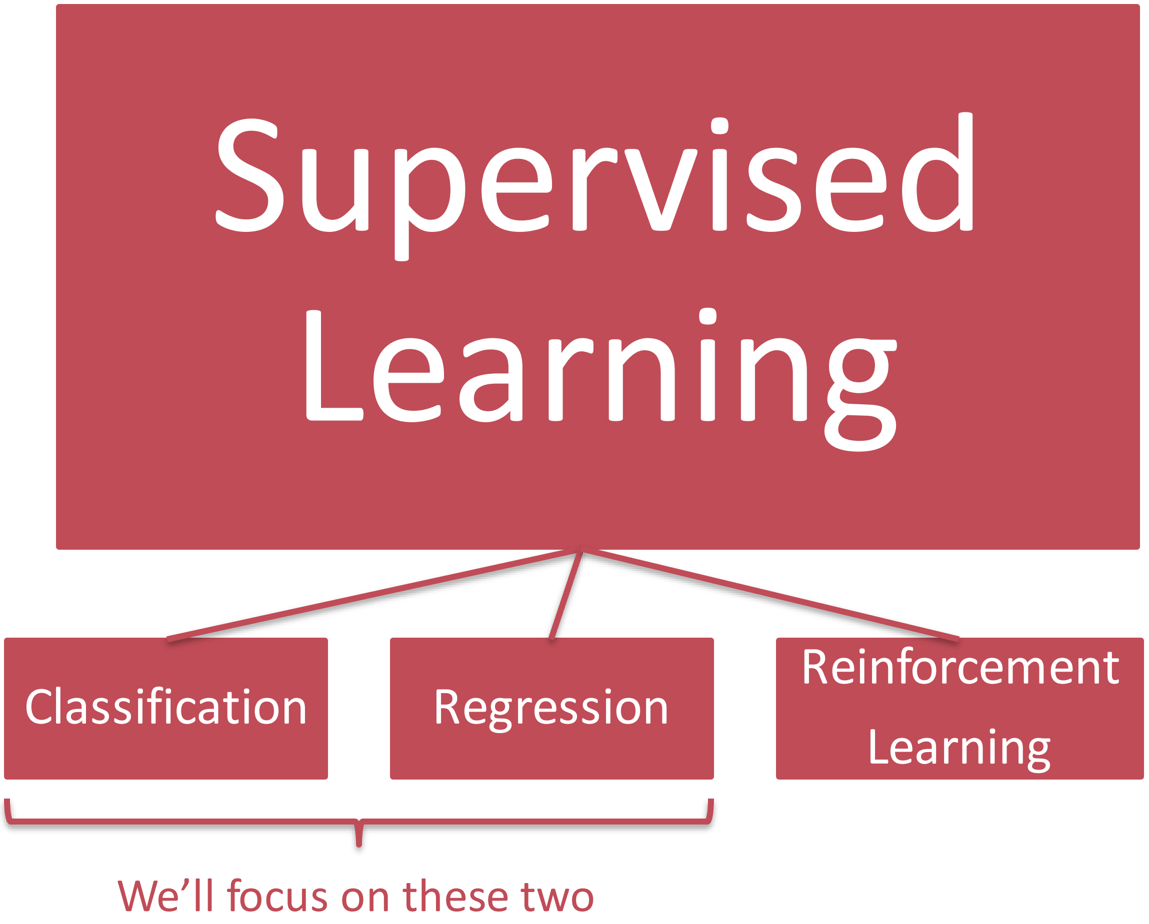
▷ e.g. $y \in \mathcal{Y} = [0, 200]$ is life expectancy

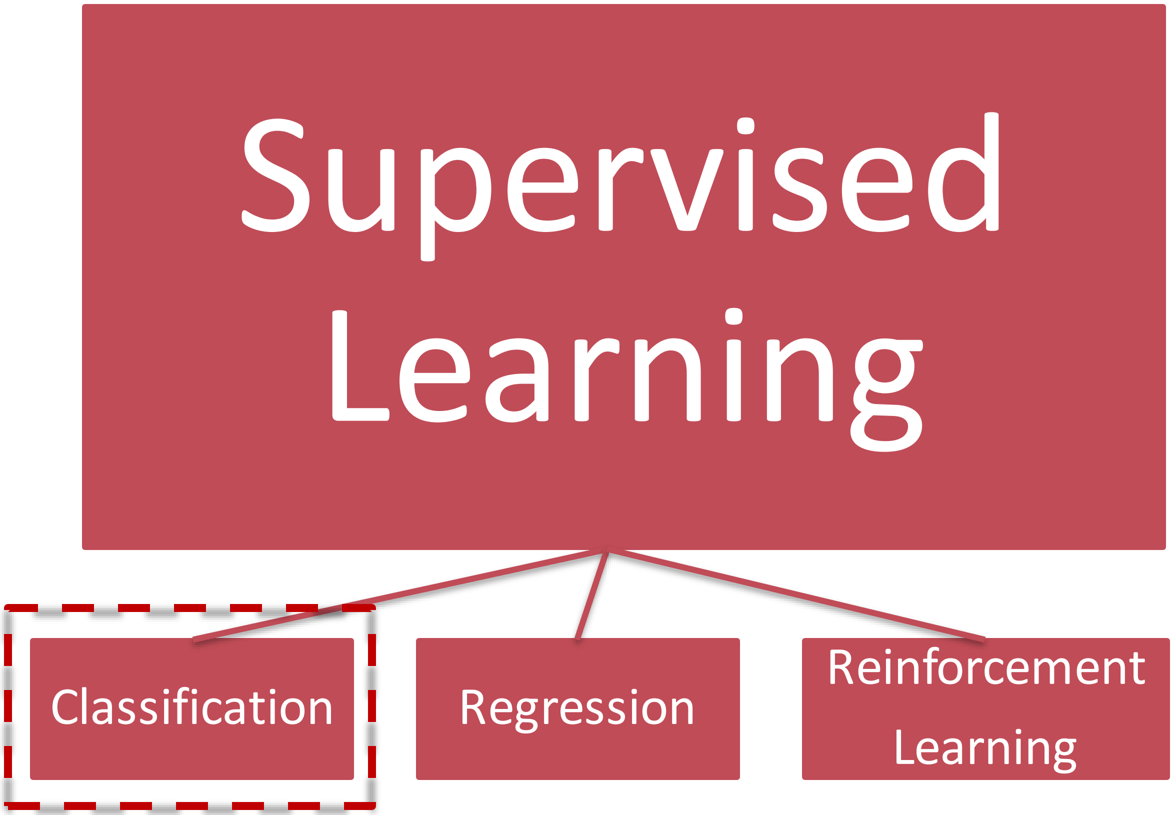
A pair $(x, y)$ is a *labeled* example.

▷ 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

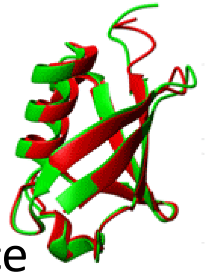

Supervised Learning
Classification
Regression
Reinforcement Learning
We'll focus on these two


Supervised Learning
Classification
Regression
Reinforcement Learning

x → Supervised Learning → y

Classification

- Binary: {0,1}
- Multi-class: 1-of-K
- Multi-label: n-of-K
- Structure:



E.g. graph/sequence