Occupant-Centric Grid-Interactive Buildings

2. Machine Learning I

CE397

Spring 2024

Prof. Dr. Zoltan Nagy

Tentative Course Outline / Schedule

Week	Class	Topic	Guest Lecture
1	01/17	Introduction / Overview / Python	
2	01/24	Machine Learning I	
3	01/31	Machine Learning II	
4	02/07	Machine Learning III	Justin Hill (Southern)
5	02/14	Occupant Behavior Modeling	
6	02/21	Occupant Behavior Modeling	Tanya Barham (CEL)
7	02/28	Occupant Behavior Modeling	Jessica Granderson (LBNL)
8	03/06	Occupant Behavior Modeling	Hussain Kazmi (KU Leuven)
9	03/13	Spring Break	
10	03/20	Advanced Control & Calibration	Ankush Chakrabarty (MERL)
11	04/27	Calibration	Donghun Kim (LBNL)
12	04/03	Introduction to CityLearn	
13	04/10	Project Work	Siva Sankaranarayanan (EPRI)
14	04/17	Project work	
15	04/24	Project work	

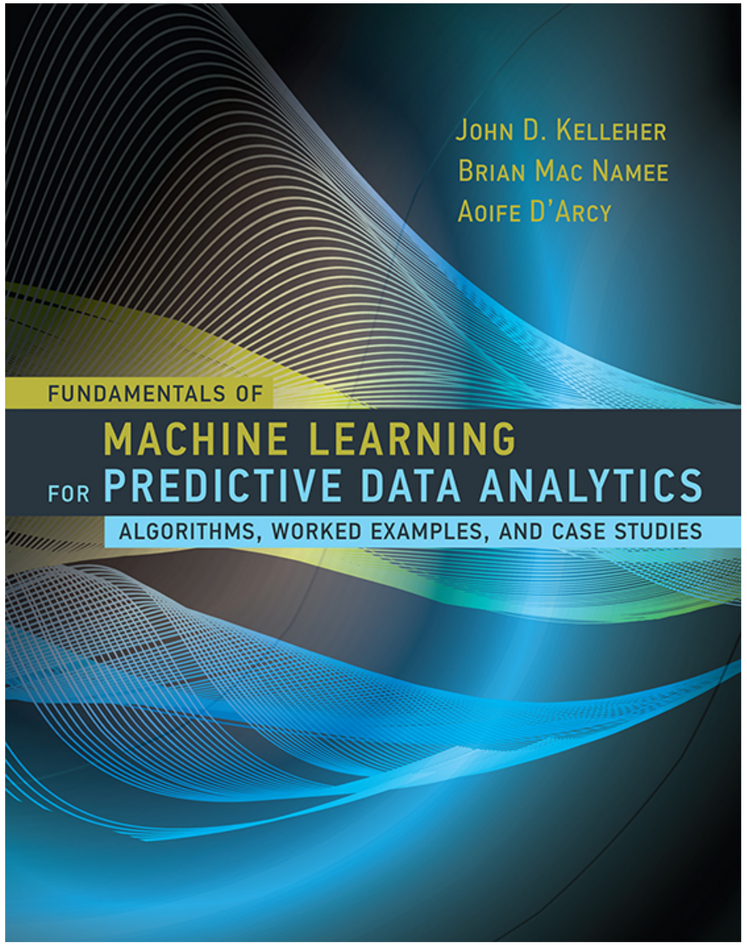
The Plan for Today

- Data & Features
- What can go wrong?
- Machine Learning Procedure
- Machine Learning Algorithms

Reference

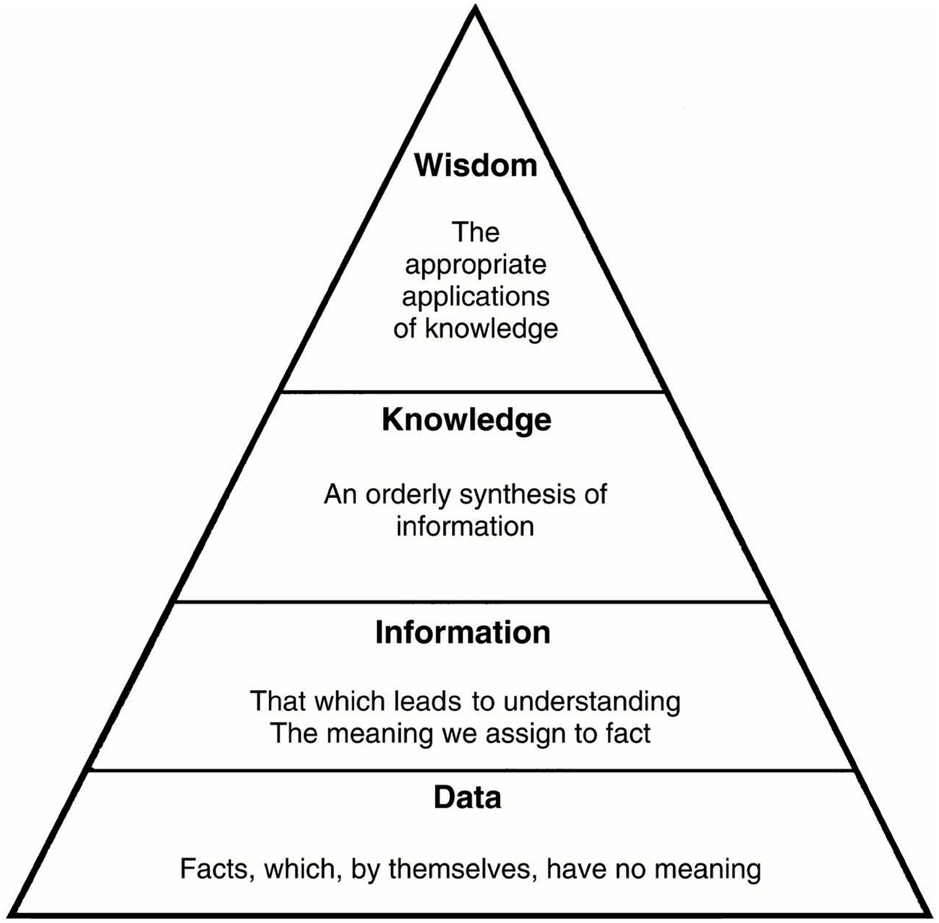
- Most of today's lecture is from

Kelleher, J. D., Mac Namee, B., & D'Arcy, A. (2015). *Fundamentals of Machine Learning for Predictive Data Analytics.* MIT Press



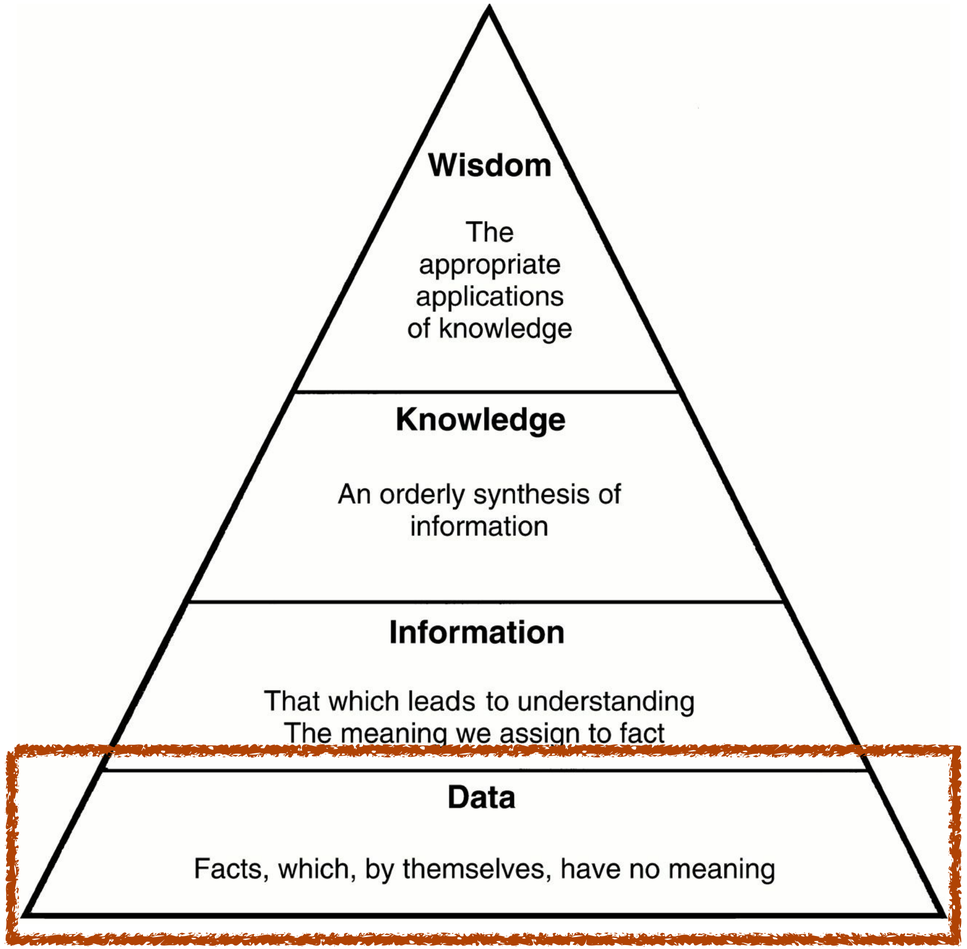
- Chapters 1.3, 1.4, 1.5, 2.4

The DIKW Pyramid



http://jbjs.org/content/82/6/888

The DIKW Pyramid

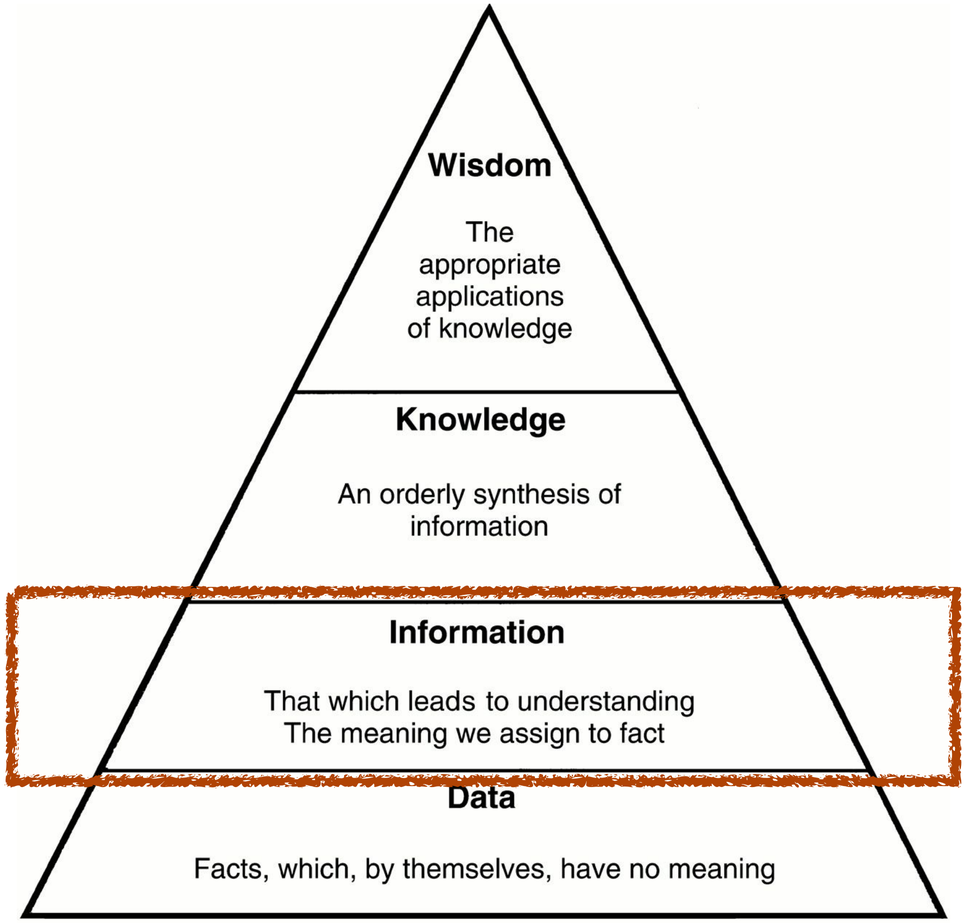


http://jbjs.org/content/82/6/888

Data

Facts, which, by themselves have no meaning

The DIKW Pyramid



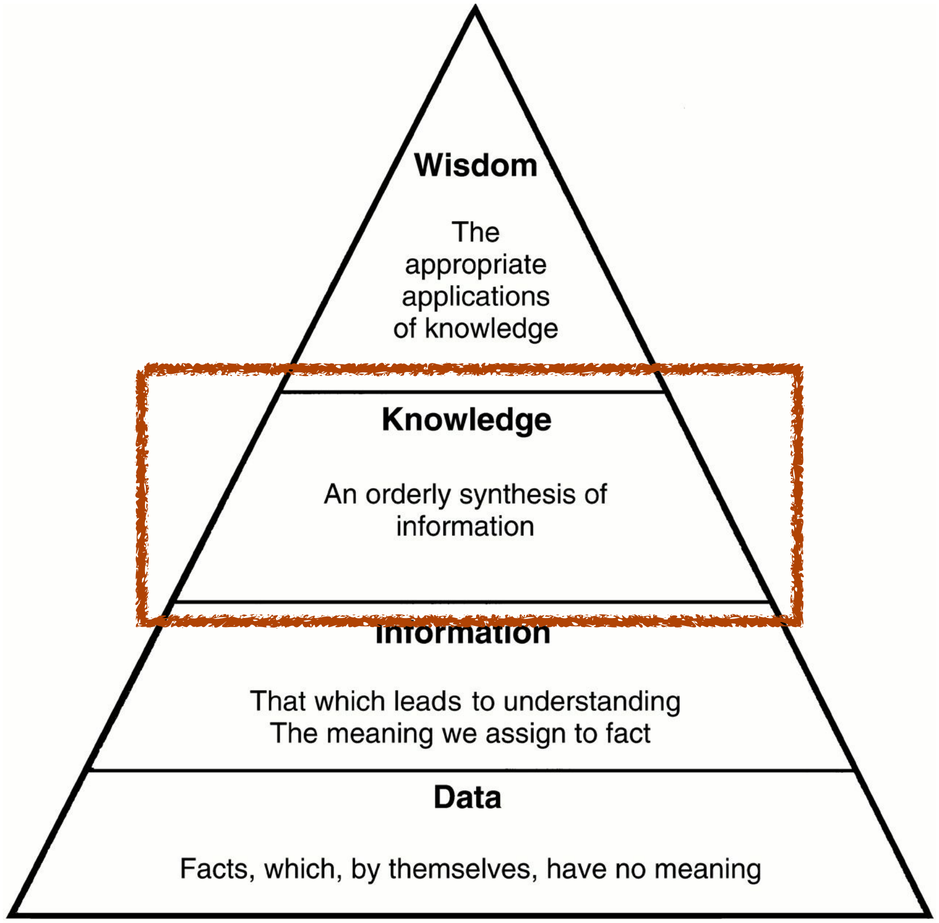
http://jbjs.org/content/82/6/888

Information

That which leads to understanding

The meaning we assign to facts

The DIKW Pyramid

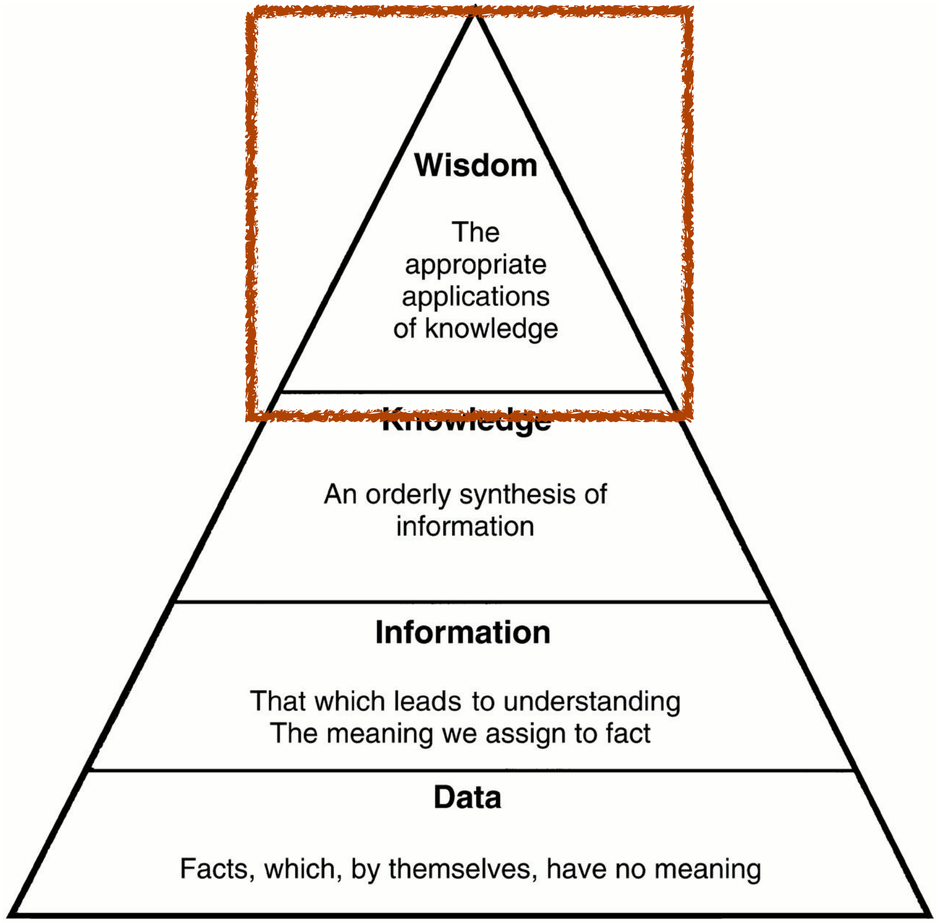


http://jbjs.org/content/82/6/888

Knowledge

An orderly synthesis of information

The DIKW Pyramid

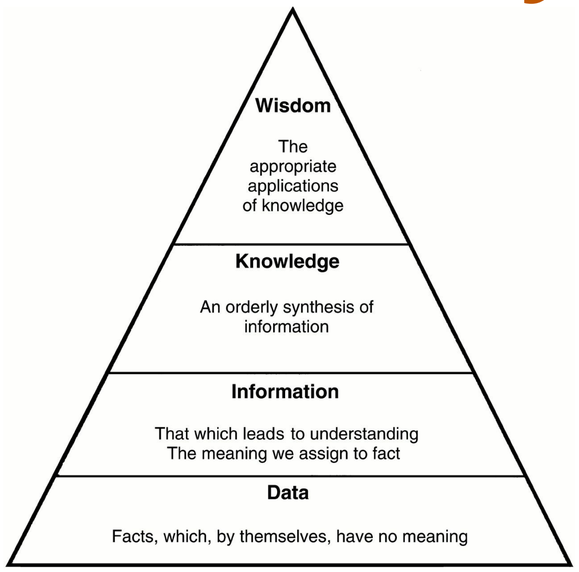


http://jbjs.org/content/82/6/888

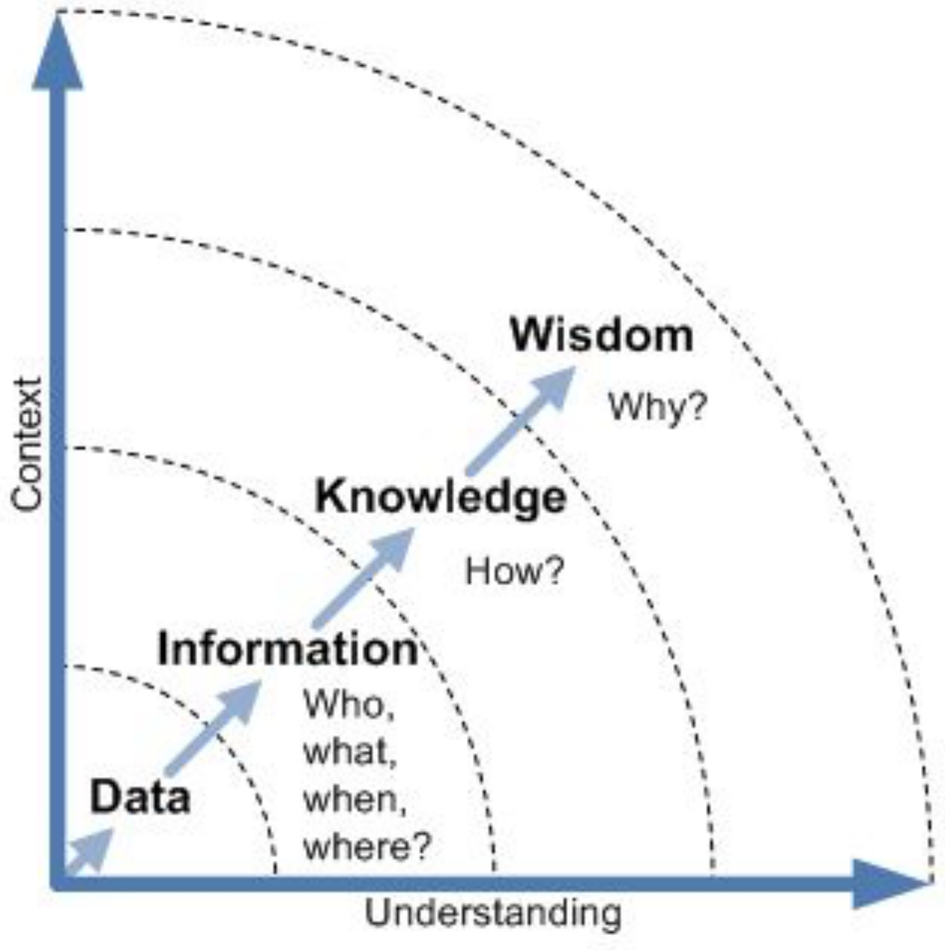
Wisdom (Insight)

The appropriate applications of knowledge

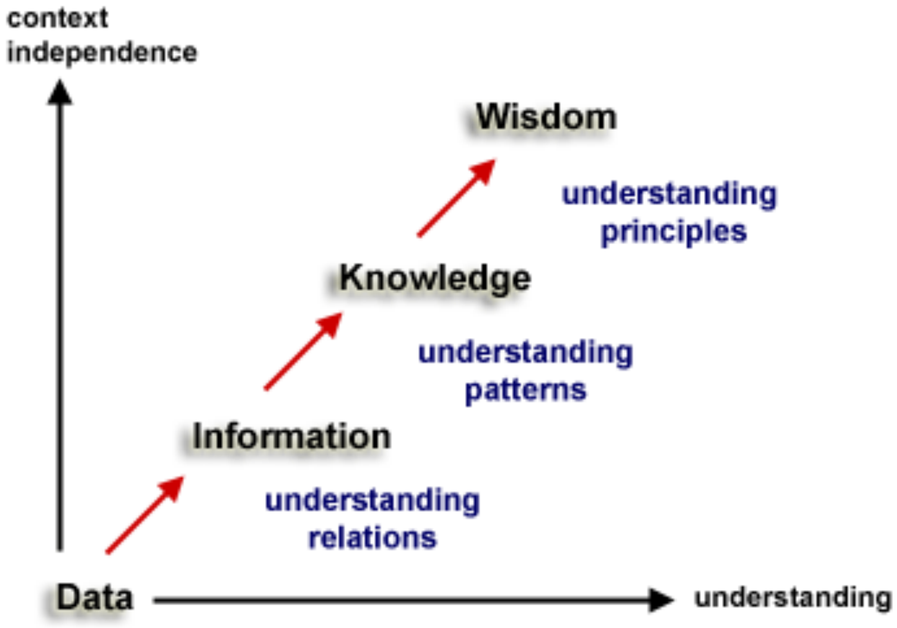
The DIKW Pyramid



http://jbjs.org/content/82/6/888

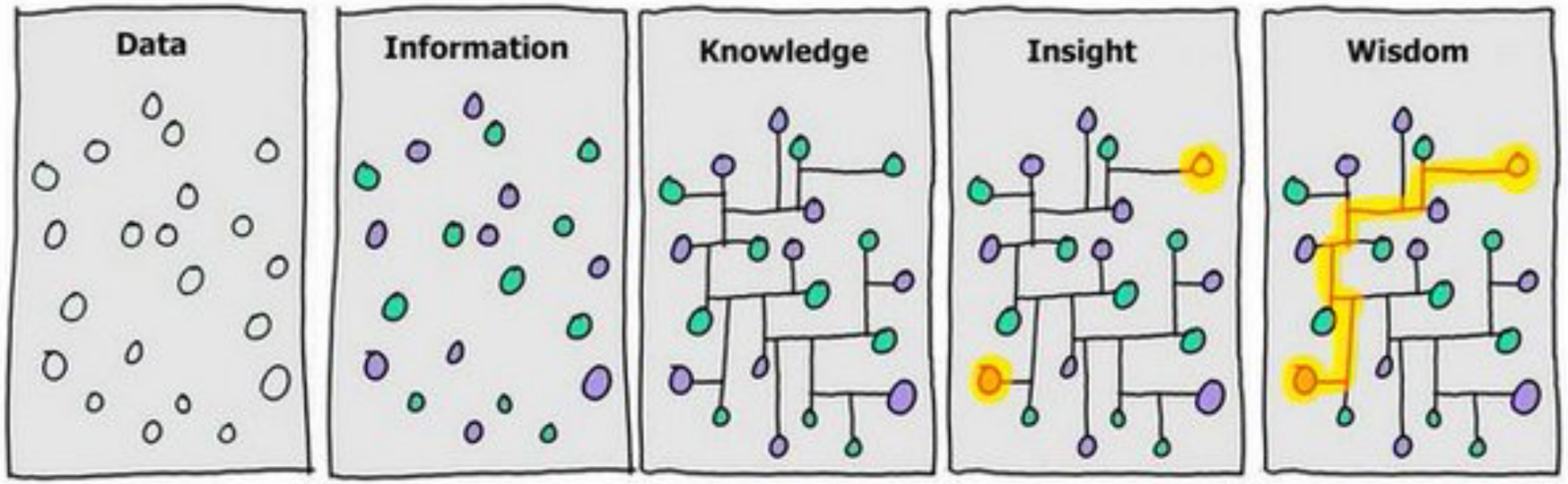


http://www.theitsmreview.com



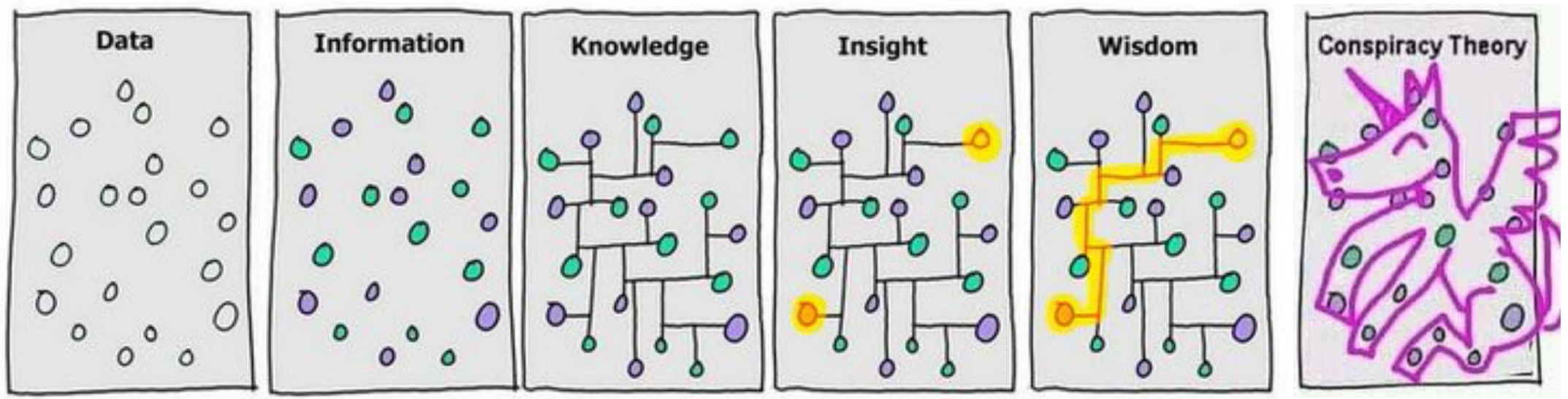
https://qph.ec.quoracdn.net

The DIKW Pyramid



https://www.kaushik.net

The DIKW Pyramid - don’t make stuff up :)



https://www.kaushik.net

Data

data [dey-*tuh*, dat-*uh*, dah-*tuh*] SHOW IPA

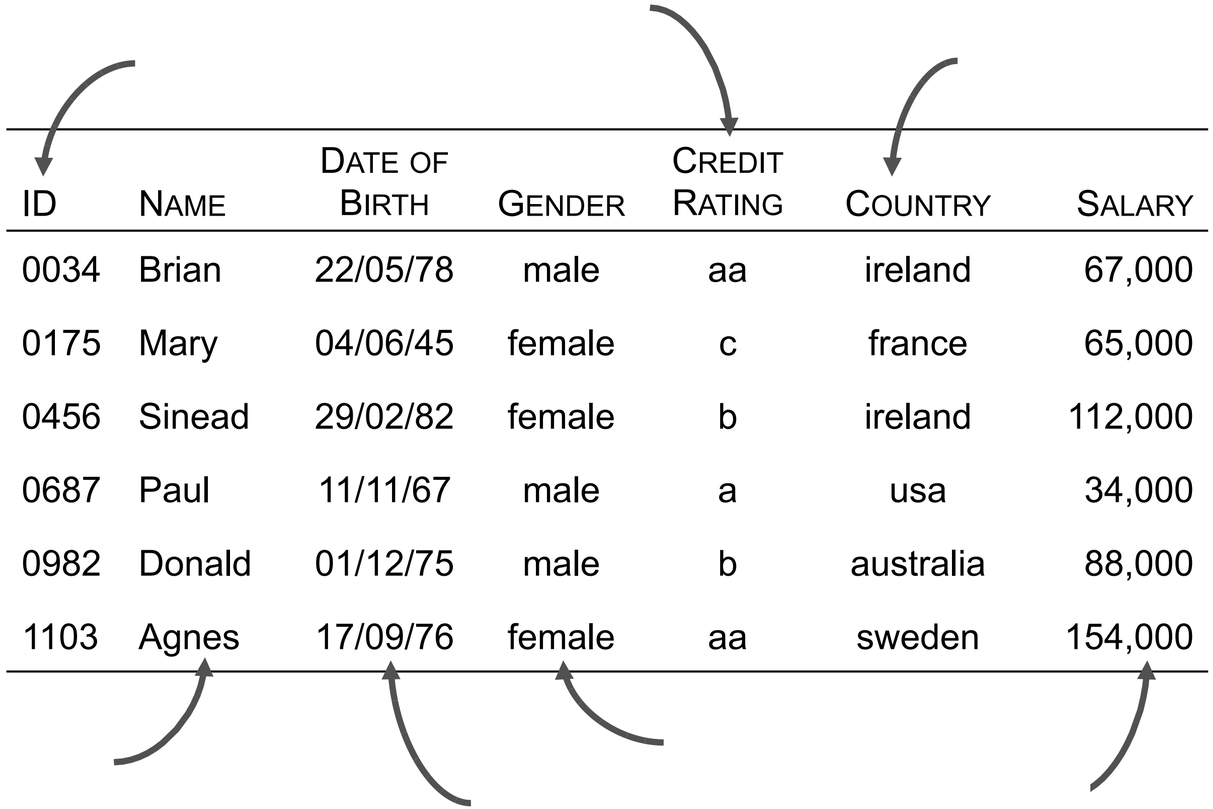
See synonyms for *data* on Thesaurus.com

noun

1. a plural of datum.
2. (*used with a plural verb*) individual facts, statistics, or items of information:
   These data represent the results of our analyses.
3. (*usually used with a singular verb*)*Digital Technology*. information in digital format, as encoded text or numbers, or multimedia images, audio, or video:
   The data was corrupted and can't be retrieved.
   Data are entered by terminal for immediate processing by the computer.
4. (*used with a singular verb*) a body of facts; information:
   Additional data is available from the president of the firm.

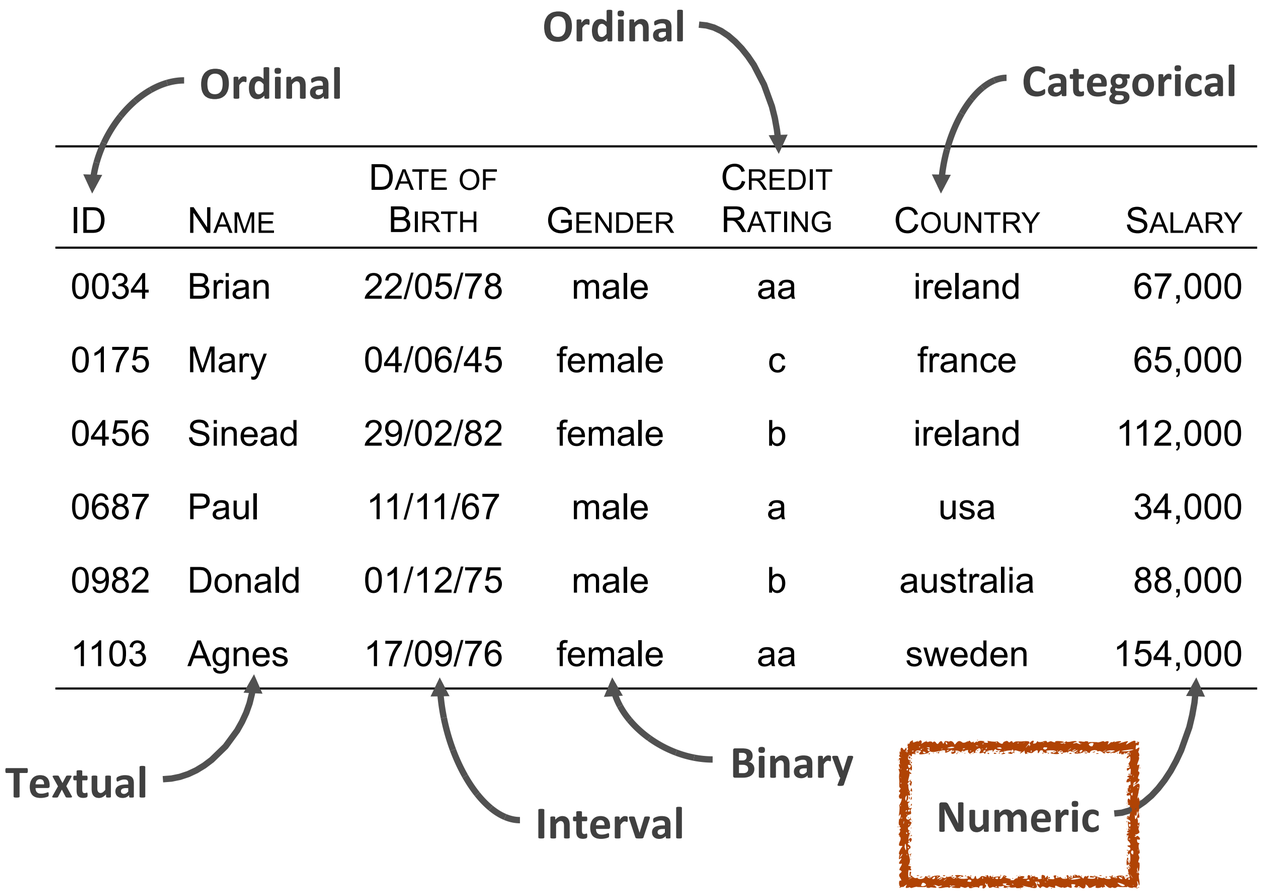
https://www.dictionary.com

Different types of Data



ID	NAME	DATE OF BIRTH	GENDER	CREDIT RATING	COUNTRY	SALARY
0034	Brian	22/05/78	male	aa	ireland	67,000
0175	Mary	04/06/45	female	c	france	65,000
0456	Sinead	29/02/82	female	b	ireland	112,000
0687	Paul	11/11/67	male	a	usa	34,000
0982	Donald	01/12/75	male	b	australia	88,000
1103	Agnes	17/09/76	female	aa	sweden	154,000

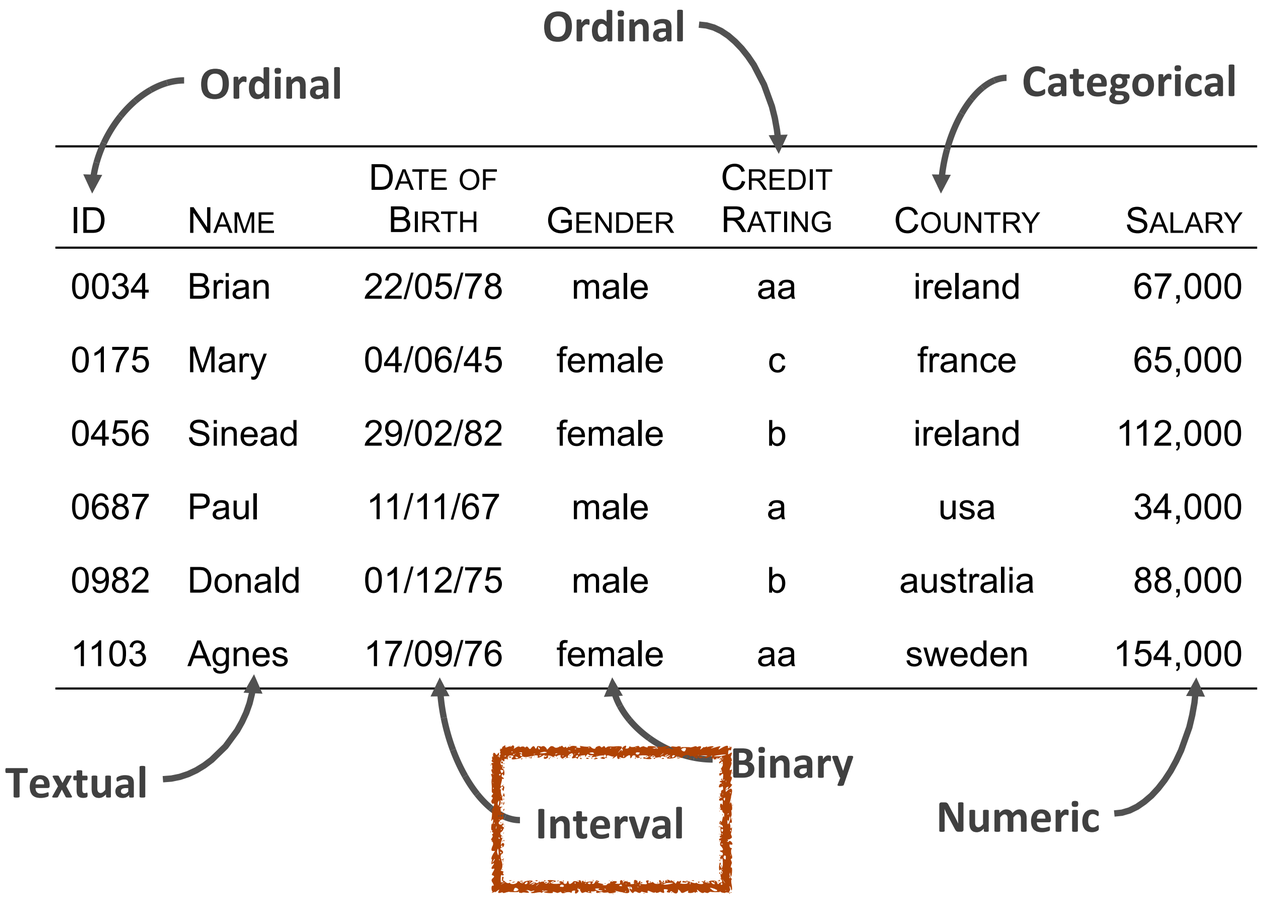
Different types of Data



Numeric

Values that allow arithmetic operations (price, age, …)

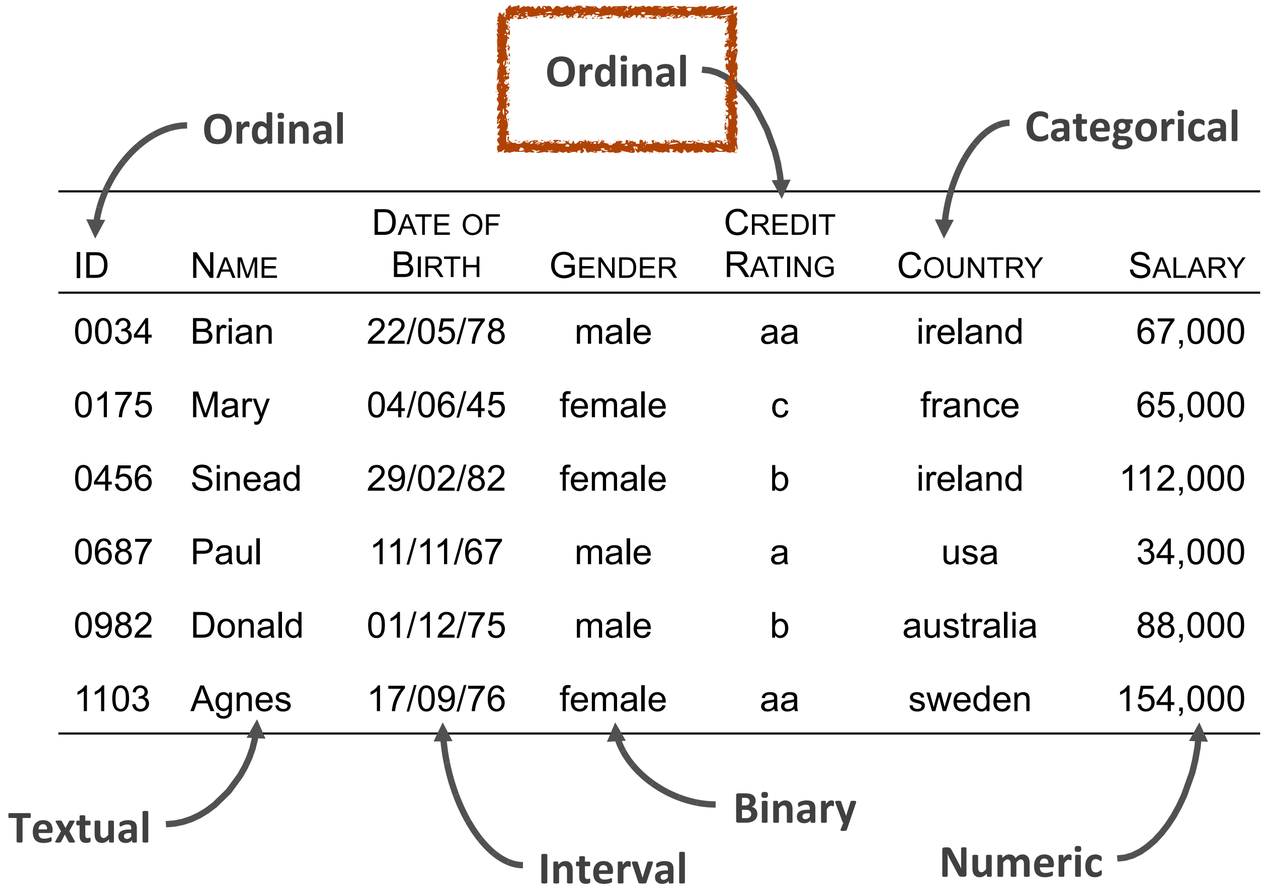
Different types of Data



Interval

Values that allow ordering and subtraction, but no other arithmetic operation (date, time)

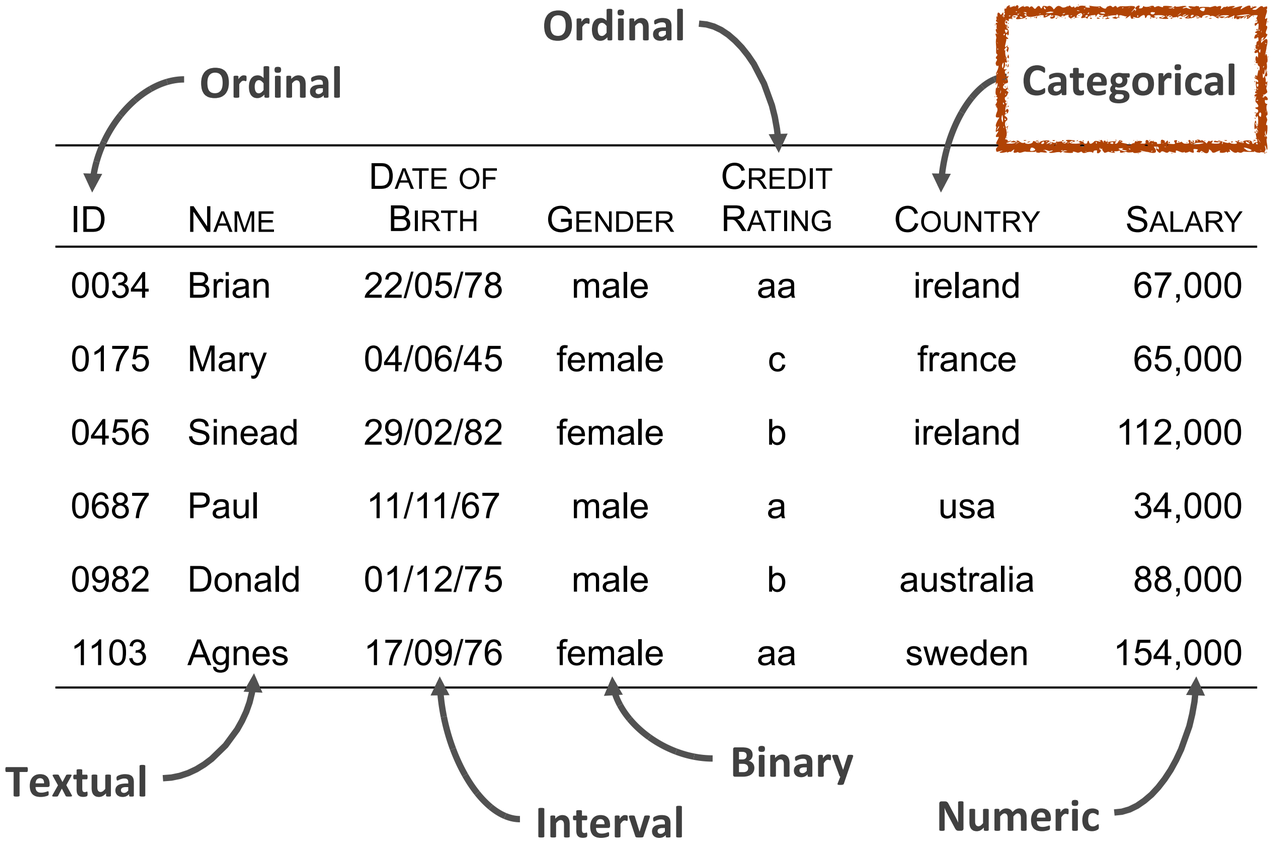
Different types of Data



Ordinal

Allow ordering but no arithmetic operation (e.g. size as small/medium/large)

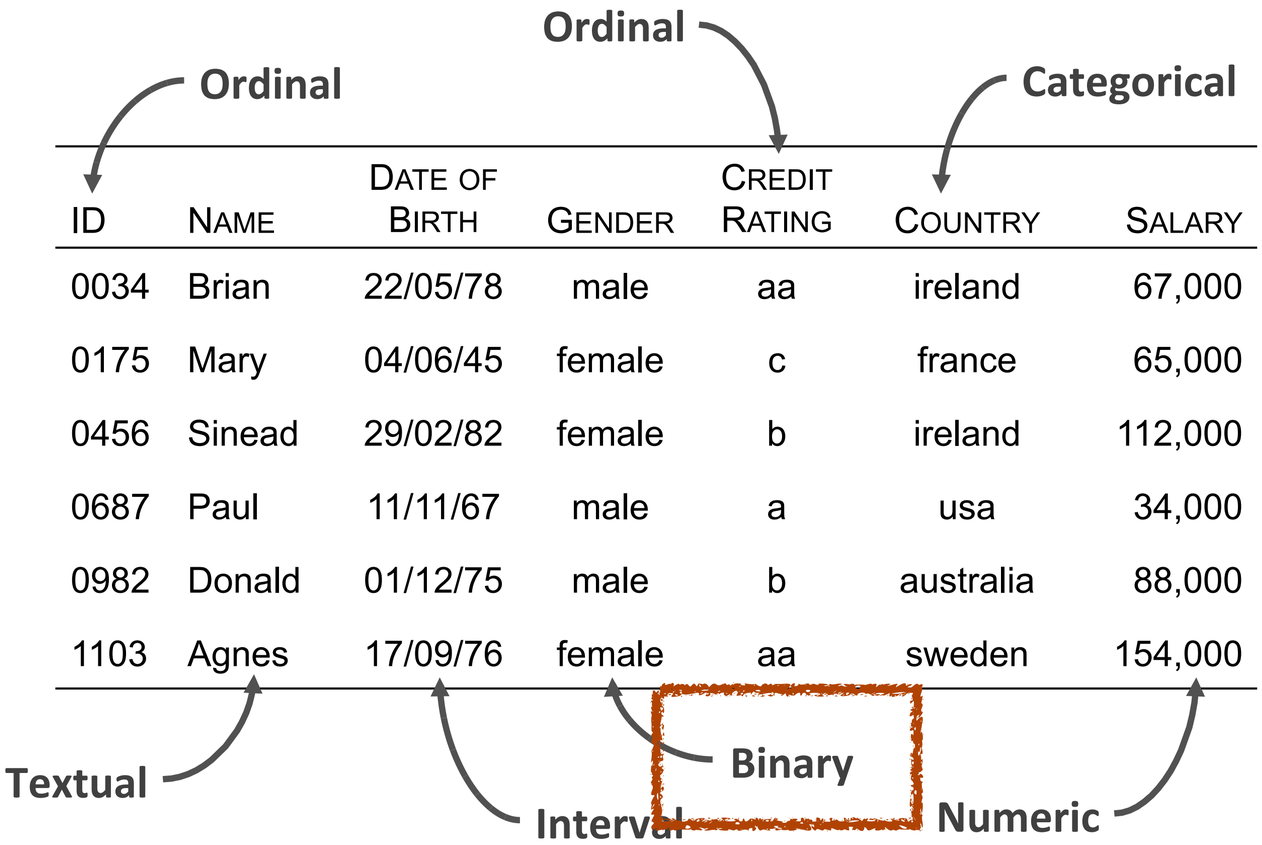
Different types of Data



Categorical

A finite set that cannot be ordered and allow no arithmetic

Different types of Data

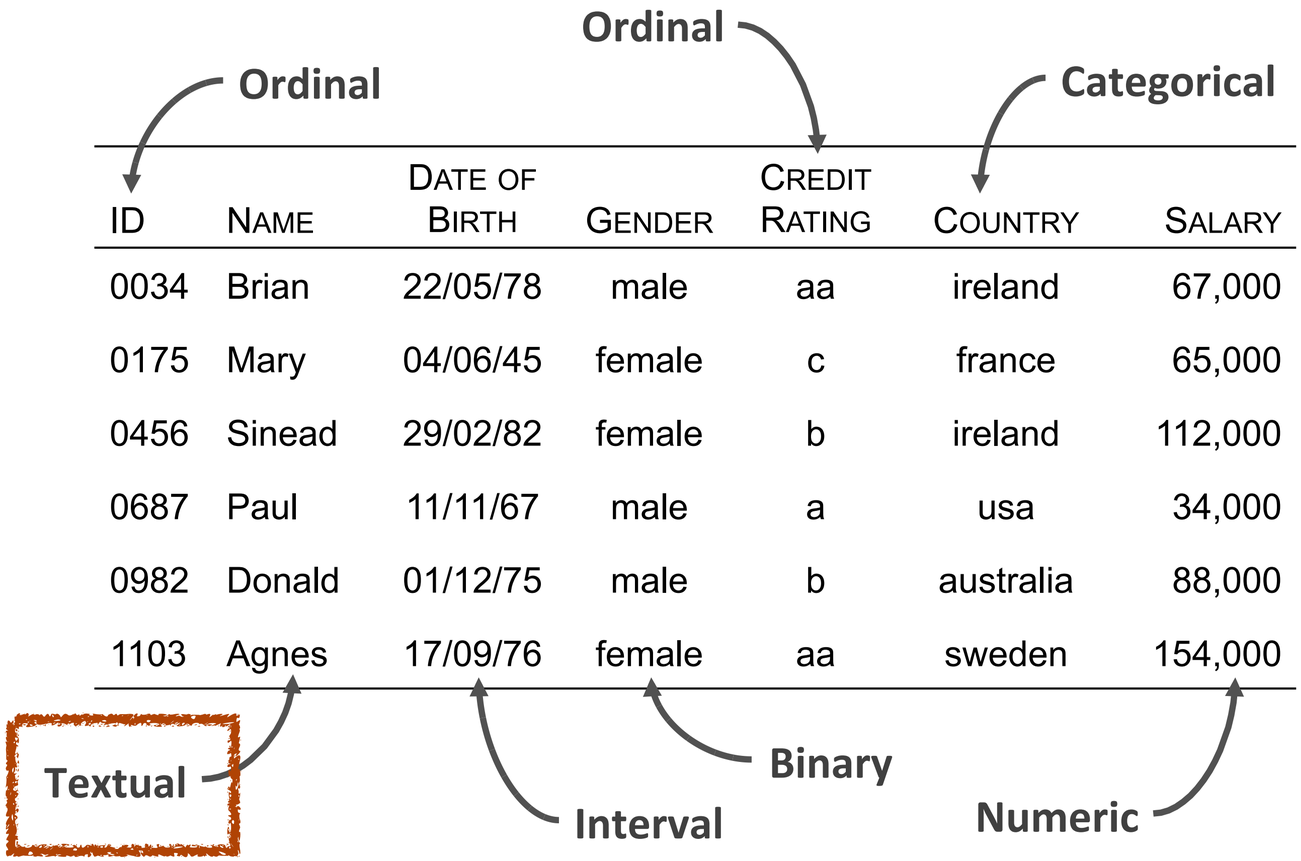


Binary

A set of only two values

(special case of categorical)

Different types of Data



Textual

Free form, usually short, text data (name, address)

Exercise

(a) State whether each descriptive feature contains numeric, interval, ordinal, categorical, binary, or textual data.

(b) How many levels does each categorical and ordinal feature have?

ID	Occupation	Gender	Age	Motor Value	Policy Type	Pref Channel
1	lab tech	female	43	42,632	planC	sms
2	farmhand	female	57	22,096	planA	phone
3	biophysicist	male	21	27,221	planA	phone
4	sheriff	female	47	21,460	planB	phone
5	painter	male	55	13,976	planC	phone
6	manager	male	19	4,866	planA	email
7	geologist	male	51	12,759	planC	phone
8	messenger	male	49	15,672	planB	phone
9	nurse	female	18	16,399	planC	sms
10	fire inspector	male	47	14,767	planC	email

Exercise

(a) State whether each descriptive feature contains numeric, interval, ordinal, categorical, binary, or textual data.

(b) How many levels does each categorical and ordinal feature have?

Exercise

(a) State whether each descriptive feature contains numeric, interval, ordinal, categorical, binary, or textual data.

ID	Ordinal	MOTORVALUE	Numeric
OCCUPATION	Textual	POLICYTYPE	Ordinal
GENDER	Categorical	AGE	Numeric
PREFCHANNEL	Categorical		

(b) How many levels does each categorical and ordinal feature have?

Exercise

(a) State whether each descriptive feature contains numeric, interval, ordinal, categorical, binary, or textual data.

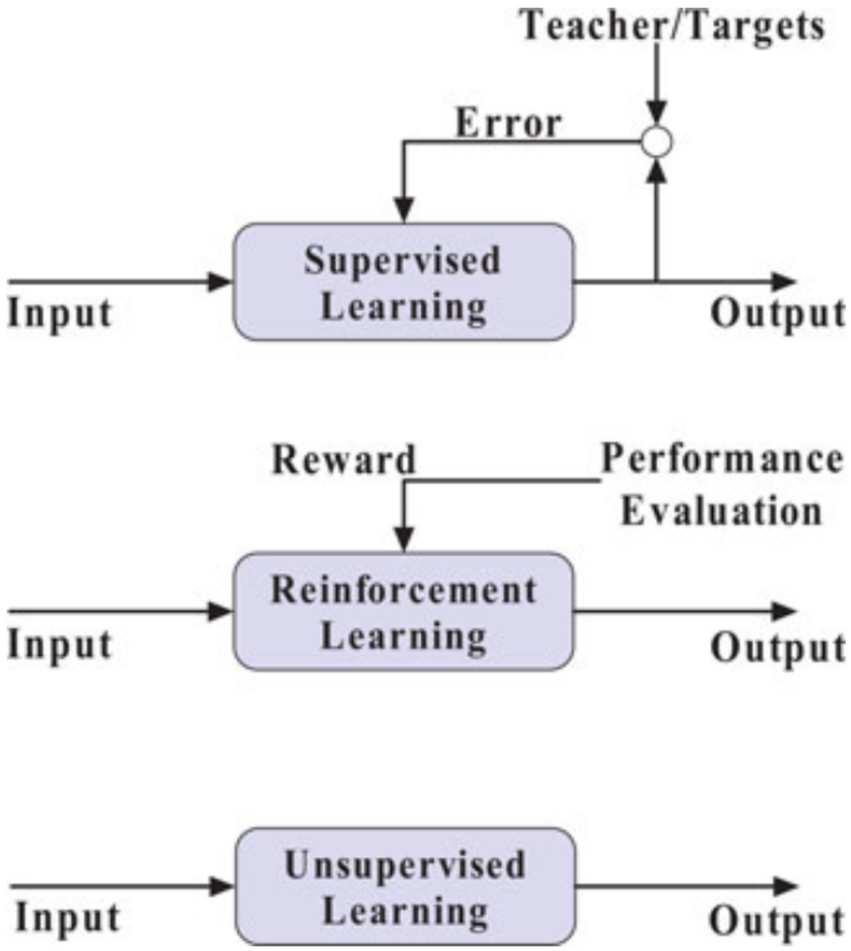
ID	Ordinal	MOTORVALUE	Numeric
OCCUPATION	Textual	POLICYTYPE	Ordinal
GENDER	Categorical	AGE	Numeric
PREFCHANNEL	Categorical		

(b) How many levels does each categorical and ordinal feature have?

ID	10 are present in the sample, but there is likely to be 1 per customer
GENDER	2 (*male*, *female*)
POLICYTYPE	3 (*planA*, *planB*, *planC*)
PREFCHANNEL	3 (*sms*, *phone*, *email*)

Machine Learning

Machine Learning



From Data to Features

- A **feature** is any measure derived from data that can be used by the machine learning algorithm.

From Data to Features

- A **feature** is any measure derived from data that can be used by the machine learning algorithm.

1. raw features = raw data source (ex. age, etc)

From Data to Features

- A **feature** is any measure derived from data that can be used by the machine learning algorithm.

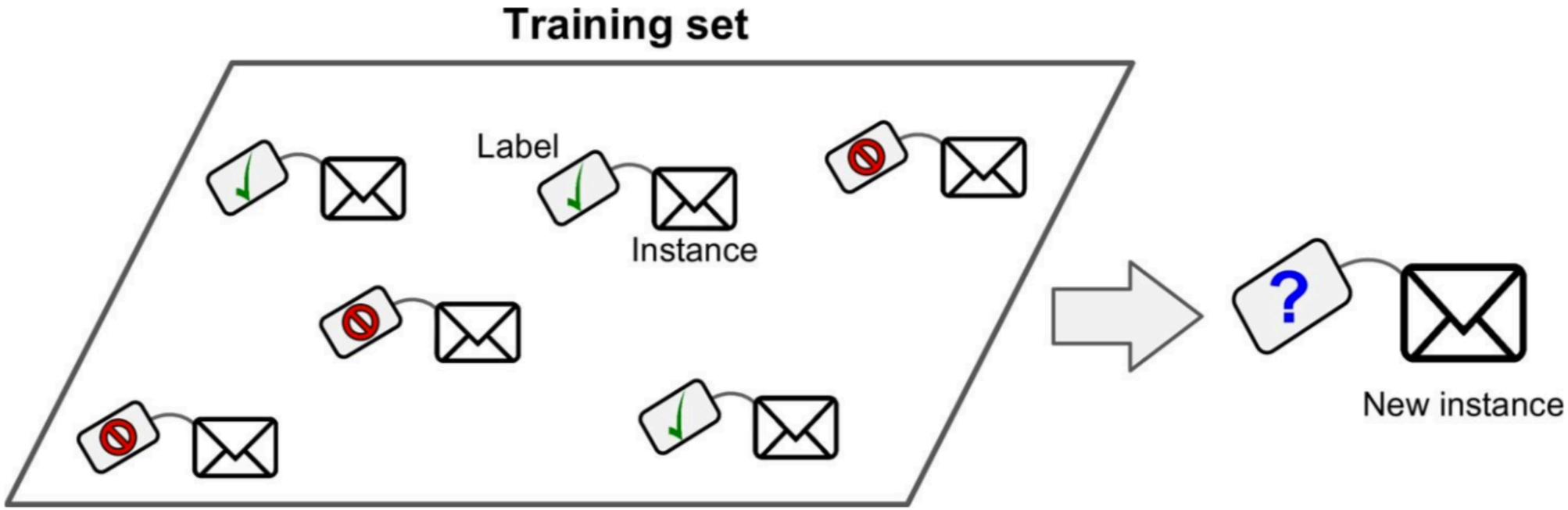
1. raw features = raw data source (ex. age, etc)
2. derived features
    - Aggregates (count, sum, average, min, max, …)
    - Flags - presence/absence of characteristic
    - Ratios - relationships between raw values
    - Mappings - convert continuous to categorical
    - Others - the sky is the limit

From Data to Features

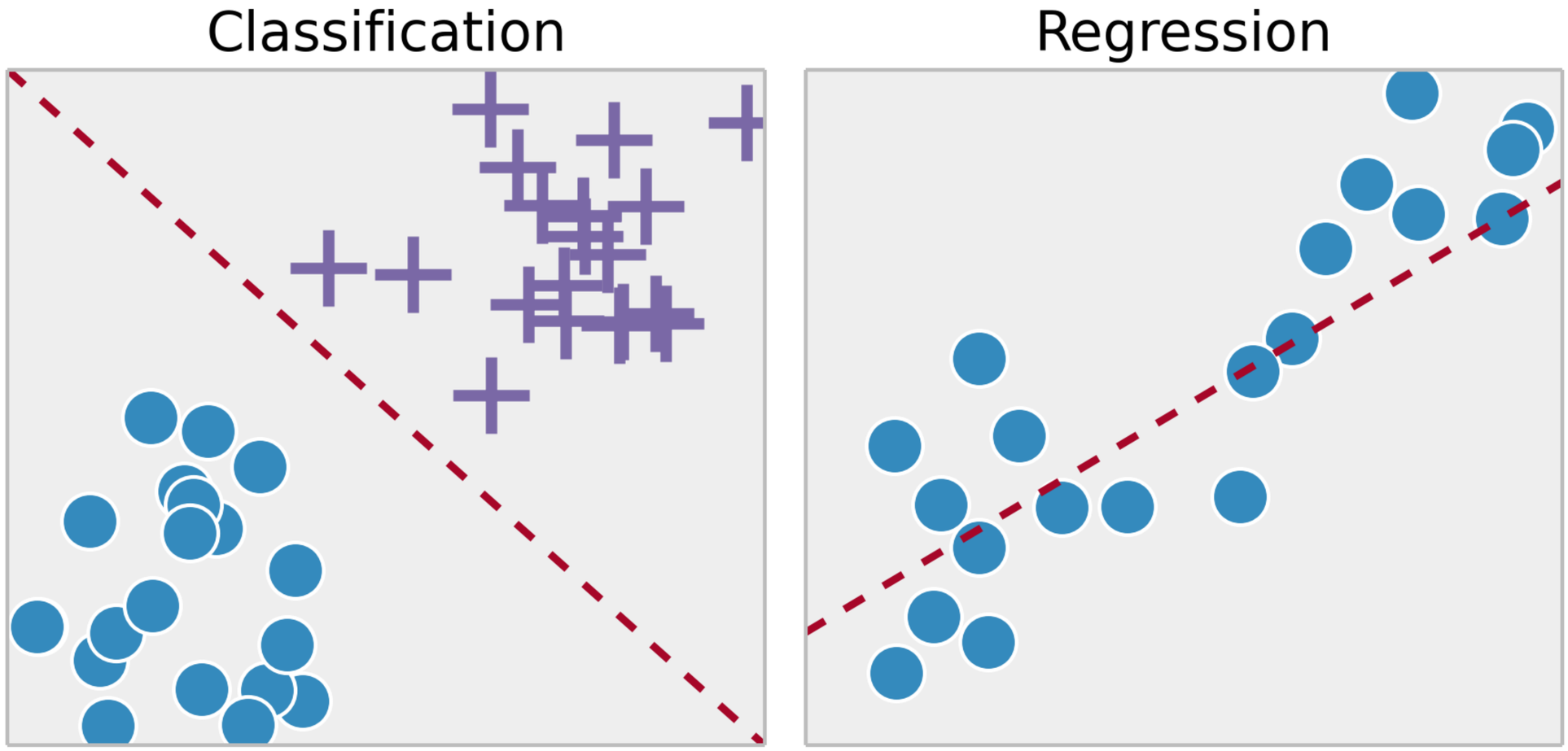
- Key consideration when designing features
  1. Data availability
  2. Timing
  3. Longevity

Supervised Learning

Modeling the relationship between measured features of data and some label associated with the data; once this model is determined, it can be used to apply labels to new, unknown data.



Supervised Learning



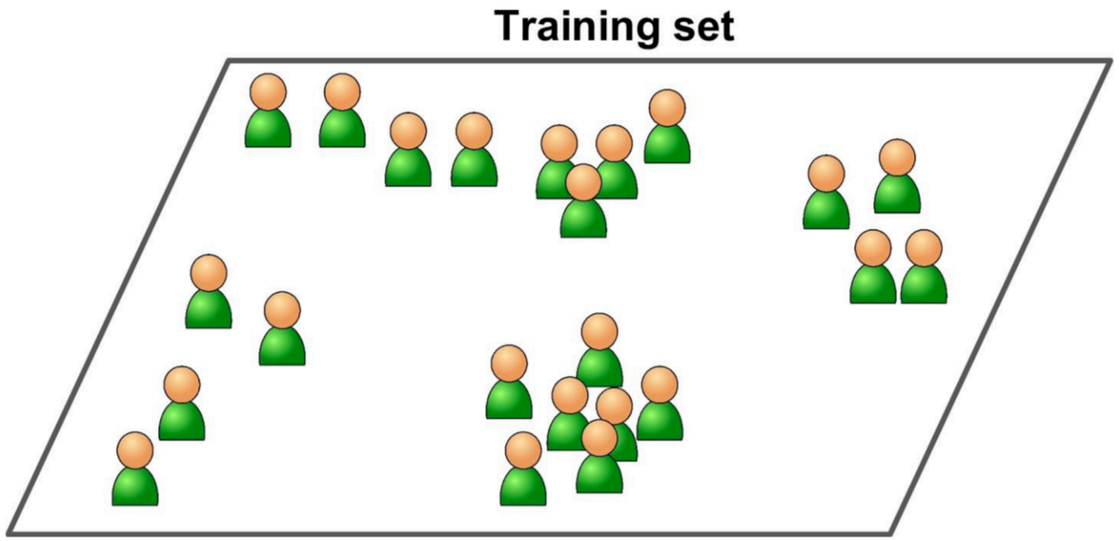
http://ipython-books.github.io/images/ml.png

Supervised Learning - Algorithms

- Linear Regression
- Logistic Regression
- KNN (*k* nearest neighbor)
- Naïve Bayes Classifier
- Decision Trees
- Support Vector Machines
- Ensemble Methods

Unsupervised Learning

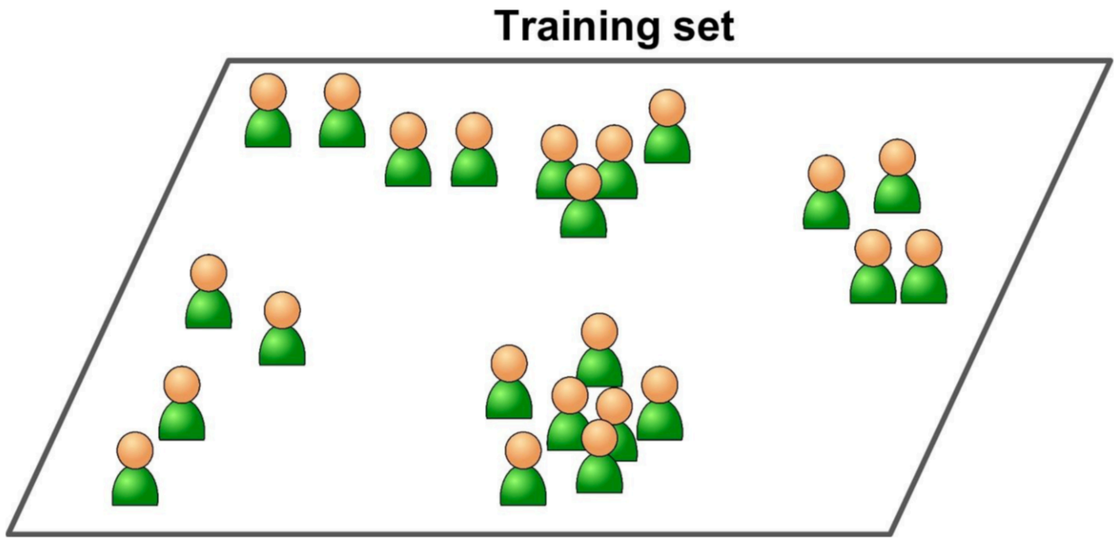
Modeling the features of a dataset *without* reference to any label, and is often described as “letting the dataset speak for itself.”

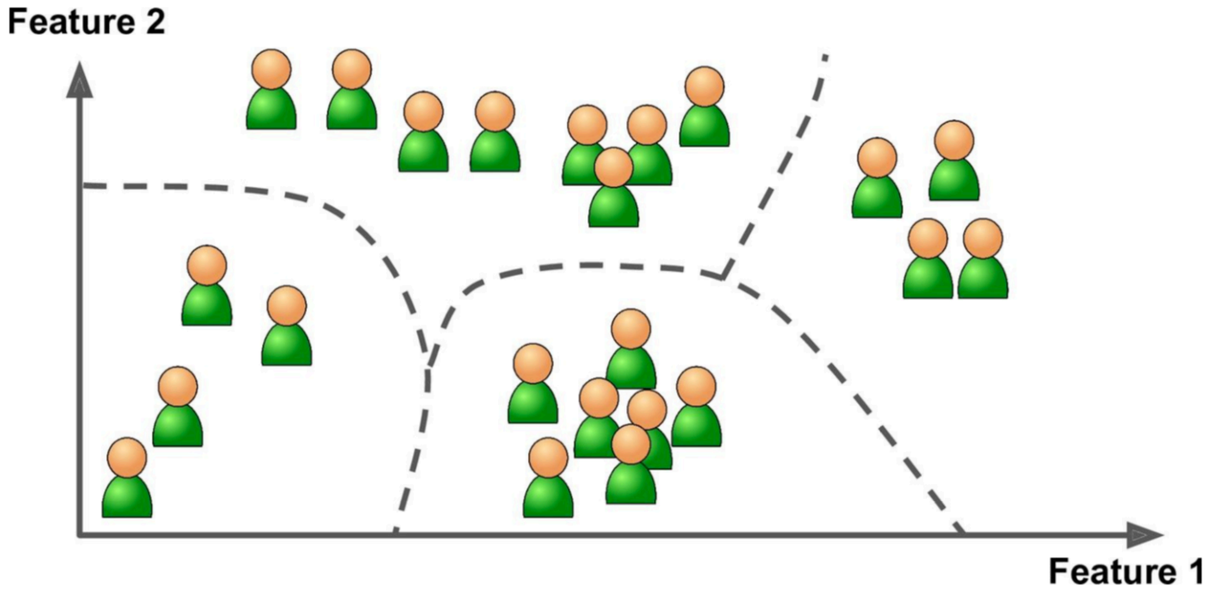


http://www.frankichamaki.com/data-driven-market-segmentation-more-effective-marketing-to-segments-using-ai/

Unsupervised Learning

Modeling the features of a dataset *without* reference to any label, and is often described as “letting the dataset speak for itself.”





http://www.frankichamaki.com/data-driven-market-segmentation-more-effective-marketing-to-segments-using-ai/

Unsupervised Learning - Algorithms

- k-means clustering
- (self organizing map)

- Similar to Human & Animal learning
- No need for data labeling
- Little information

When we're learning to see, nobody's telling us what there might answers are - we just look. *(Geoffrey Hinton, 1996)*

Break

Goal of Machine Learning

- A good ML algorithm / model should provide a good **generalization**

Goal of Machine Learning

- A good ML algorithm / model should provide a good **generalization**
- Consistency with data is a necessary but not a sufficient condition

Goal of Machine Learning

- A good ML algorithm / model should provide a good **generalization**
- Consistency with data is a necessary but not a sufficient condition
- Important! All models/algorithms have their own set of assumptions: **Inductive bias**

Goal of Machine Learning

- A good ML algorithm / model should provide a good **generalization**
- Consistency with data is a necessary but not a sufficient condition
- Important! All models/algorithms have their own set of assumptions: **Inductive bias**
  1. Restriction bias: constrain learning to certain models
  2. Preference bias: guide learning to prefer certain models over others

What can go wrong?

- **Note**: Inductive bias is a necessary prerequisite for learning to occur.

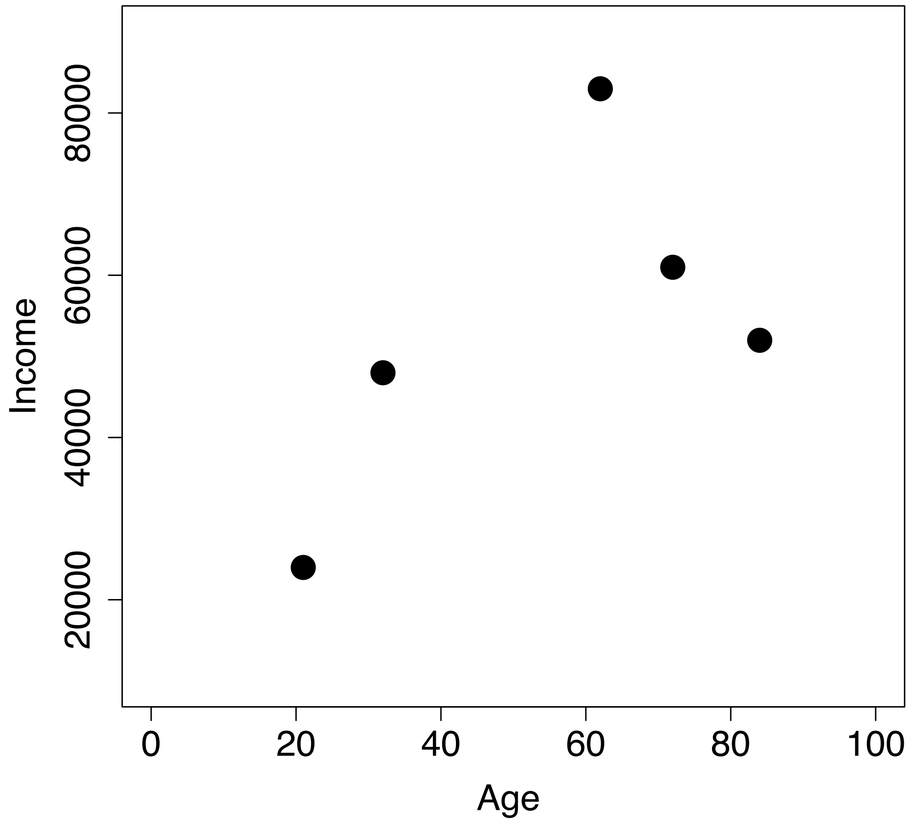
What can go wrong?

- **Note**: Inductive bias is a necessary prerequisite for learning to occur.
- There is no particular inductive bias that on average is the best one to use
  No Free Lunch Theorem

What can go wrong?

- **Note**: Inductive bias is a necessary prerequisite for learning to occur.
- There is no particular inductive bias that on average is the best one to use
  No Free Lunch Theorem
- Inappropriate inductive bias can lead to
  1. **Underfitting:** model too simplistic (relationships not captured well)
  2. **Overfitting:** model too complex (model becomes very sensitive to noise)

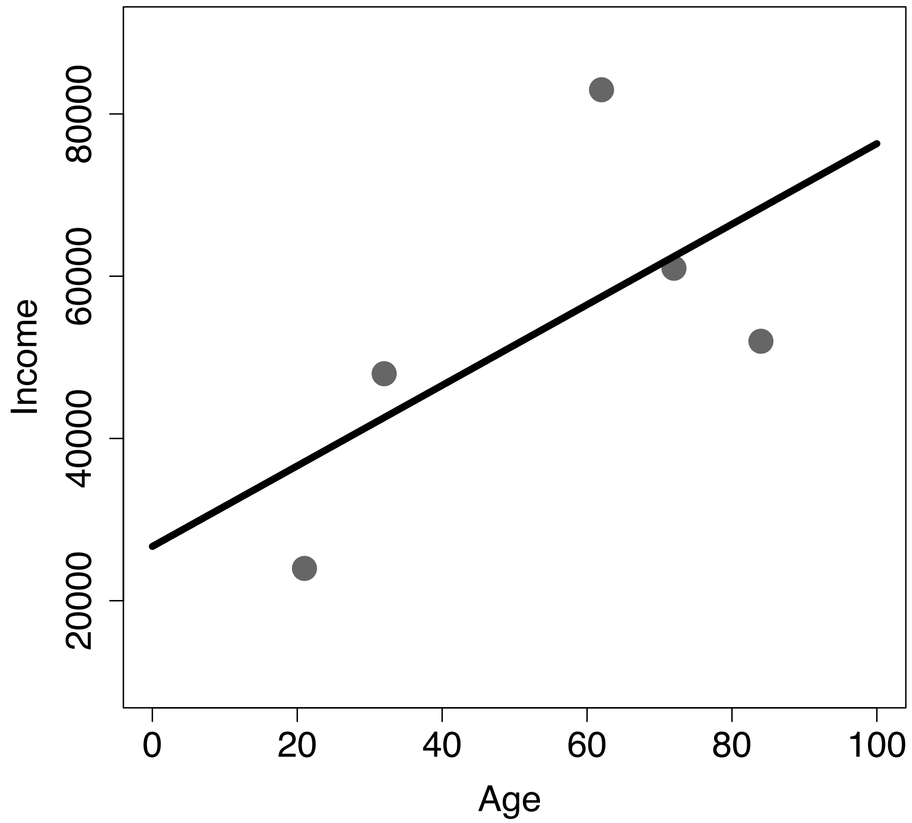
Example: Age-Income Dataset



ID	AGE	INCOME
1	21	24,000
2	32	48,000
3	62	83,000
4	72	61,000
5	84	52,000

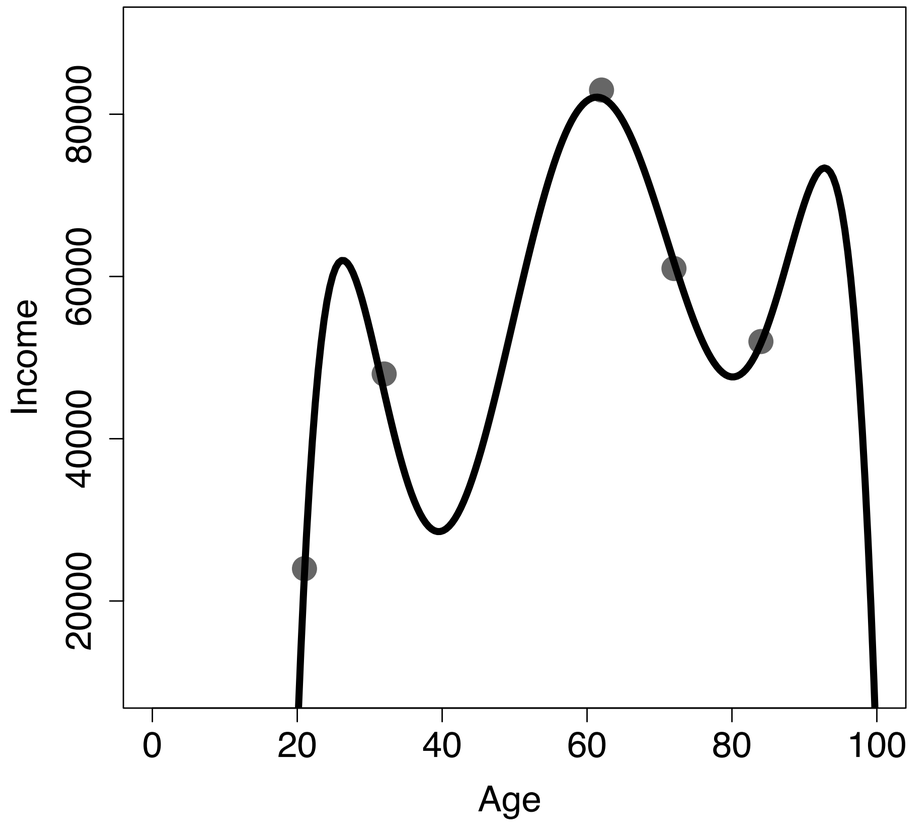
Example: Age-Income Dataset

Underfitting



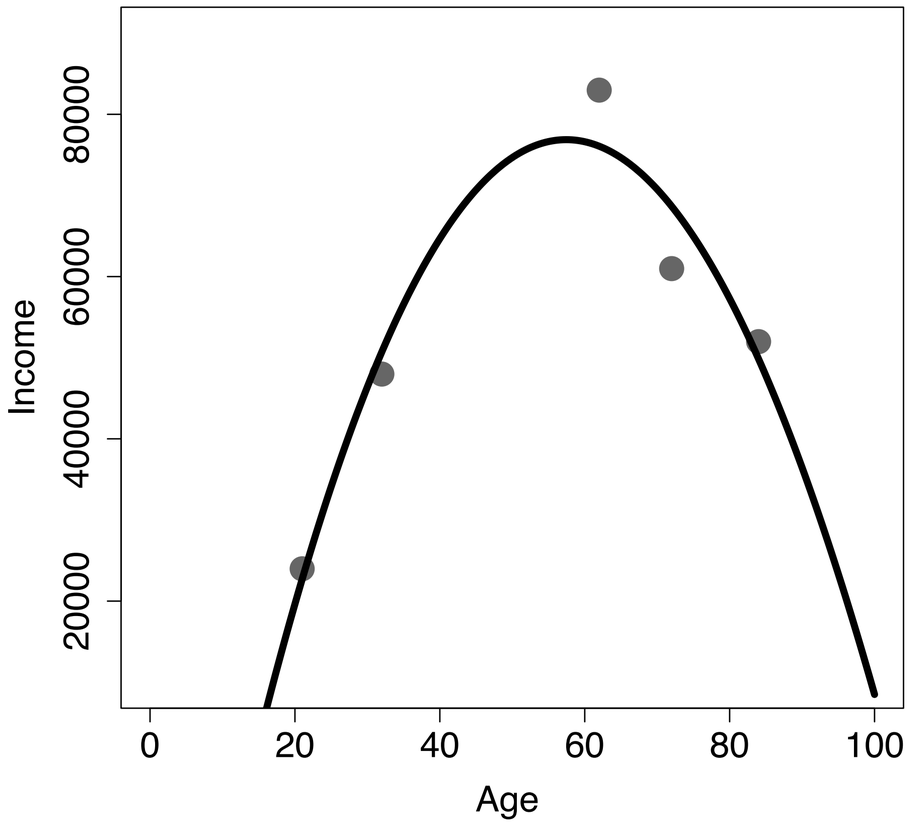
Example: Age-Income Dataset

Overfitting



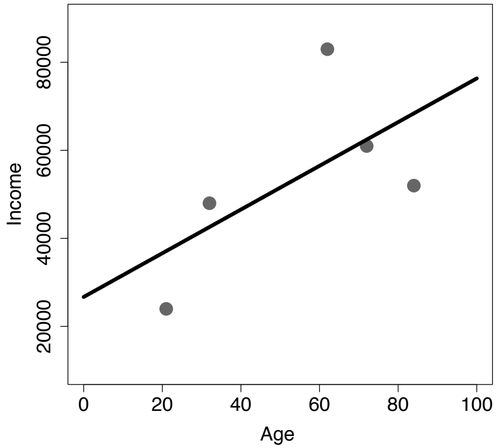
Example: Age-Income Dataset

Just right



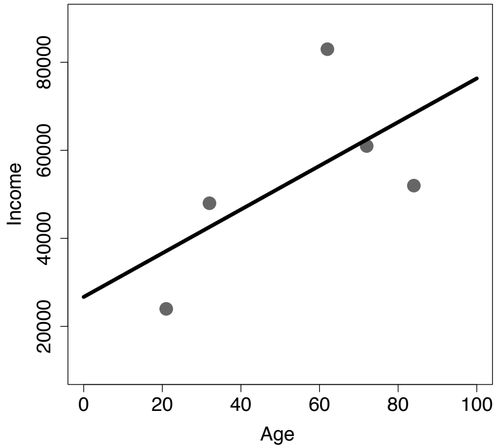
Bias-variance tradeoff in ML

- *Bias*: error due to assumptions
  - High bias = underfitting
  - model does not capture sufficient characteristics of the data

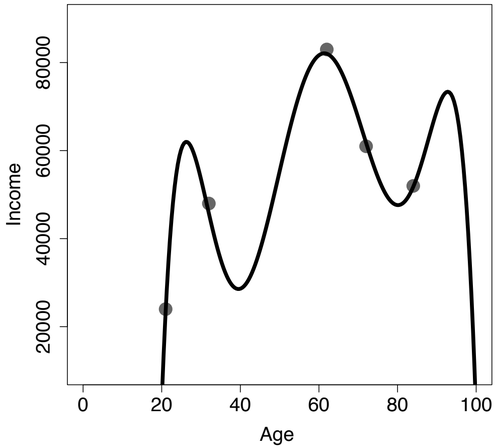


Bias-variance tradeoff in ML

- *Bias*: error due to assumptions
  - High bias = underfitting
  - model does not capture sufficient characteristics of the data



- Variance: *error* due to sensitivity of model
  - High variance = overfitting
  - model captures noise in the data



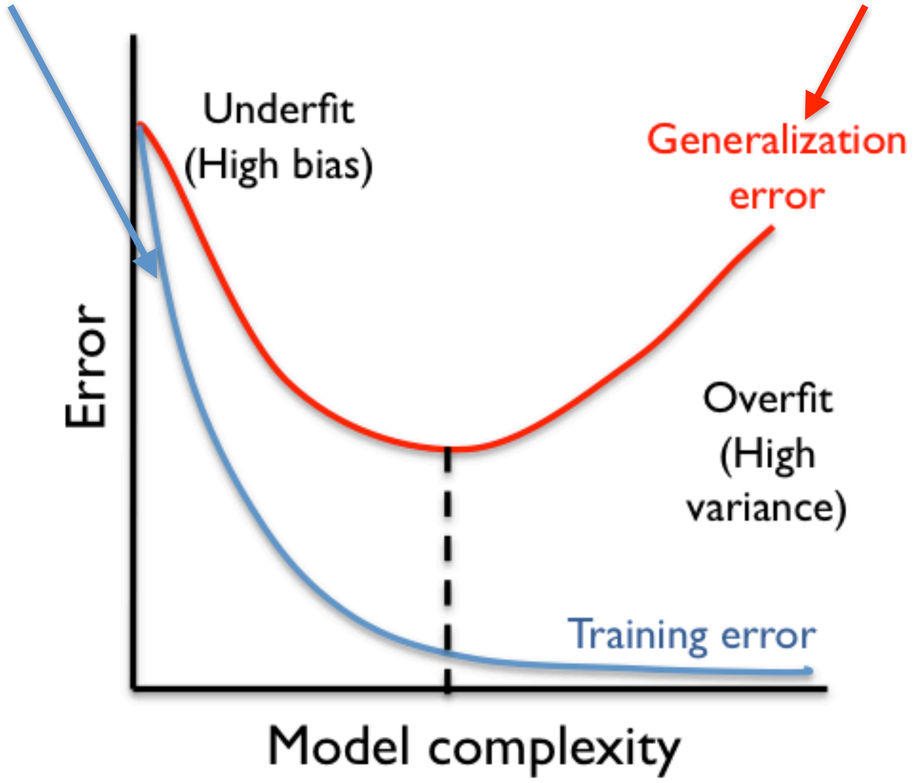
How can we then evaluate (and improve) the performance of a given model?

Approach: Split input data

Training set (70 %)	Validation set (30 %)

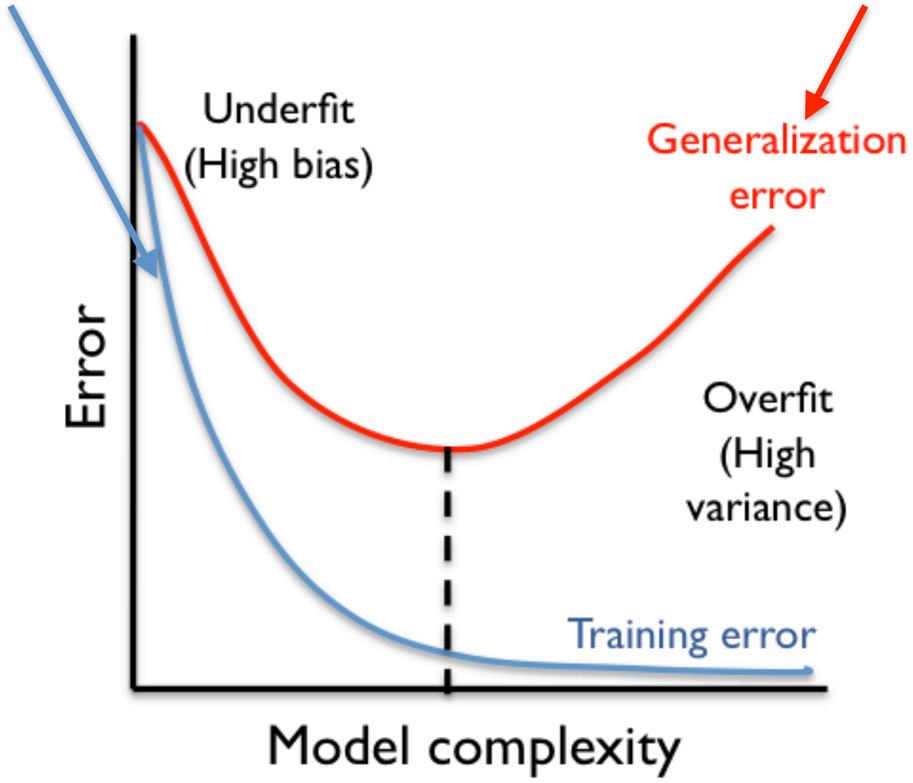
Approach: Split input data

Training set (70 %)	Validation set (30 %)



Approach: Split input data

Training set (70 %)	Validation set (30 %)

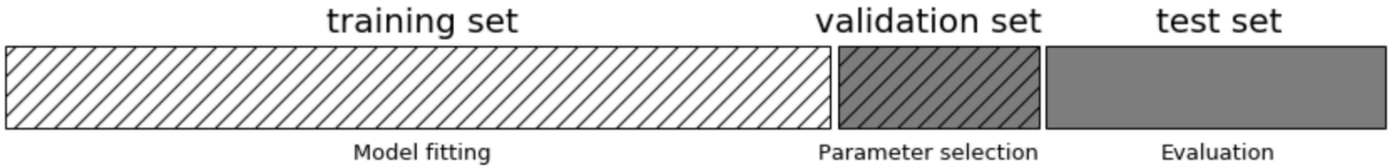


Note:
If choosing between rival models, one often uses 3-fold split

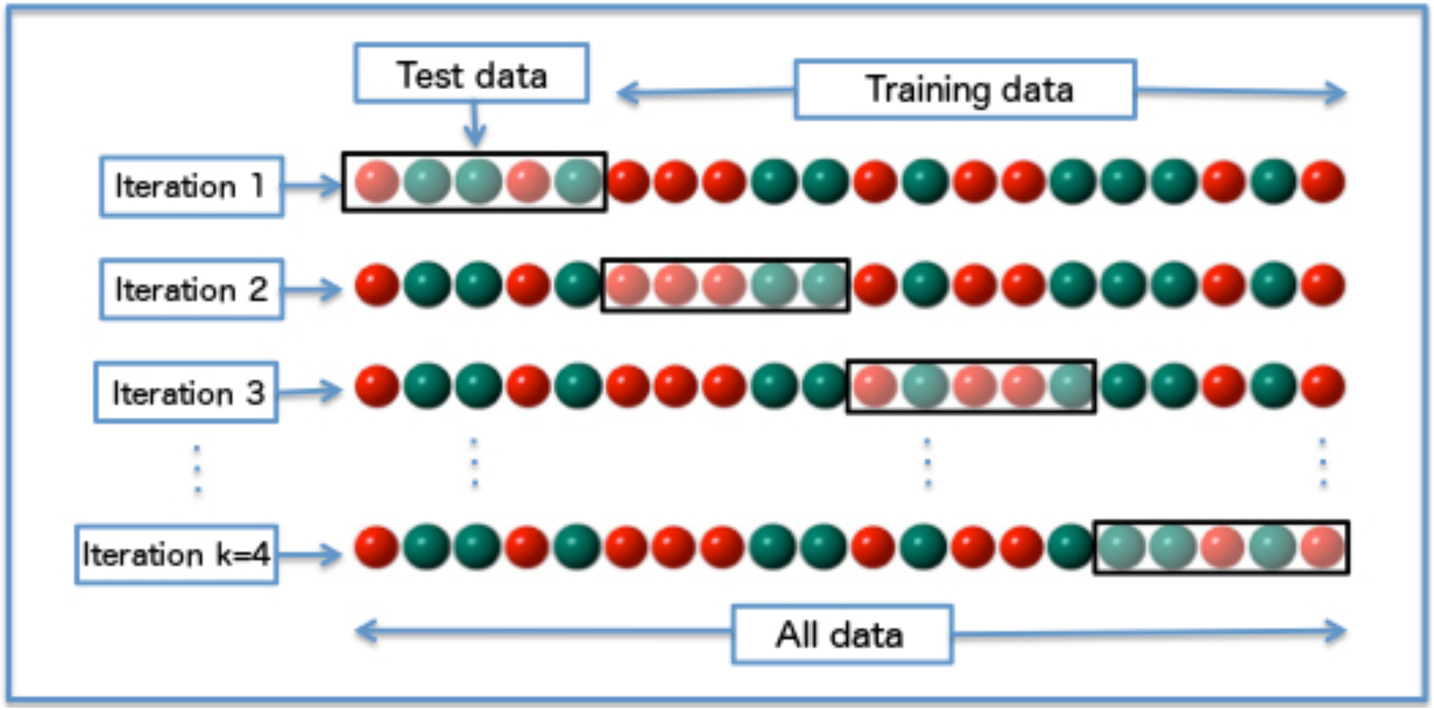
Training: Train

Validation: Compare models

Test: Estimate accuracy (no model tuning anymore here)



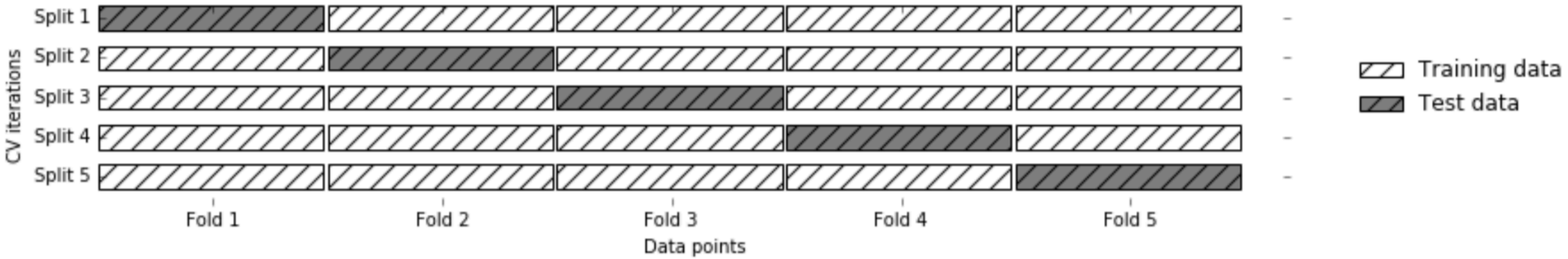
Cross-Validation: Split & Repeat



Cross-Validation: Split & Repeat

- Statistical method of evaluating generalization performance
- More stable and thorough than using a split into a training and a test set.
- Data is instead split repeatedly and multiple models are trained.
- Most commonly version: *k*-fold cross-validation, where k is a user-specified number, usually 5 or 10.

k-fold Cross Validation



Characteristics of Cross Validation

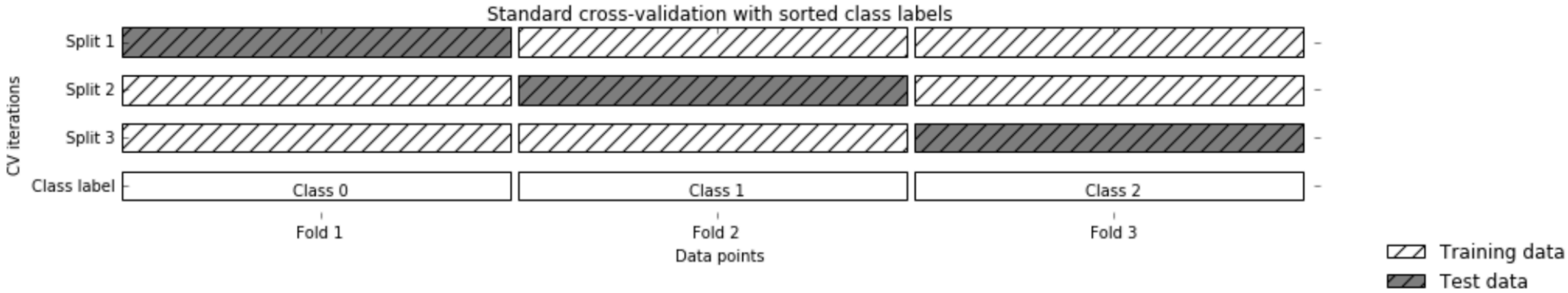
+ Avoids bad/good choice of a single training & test set
+ provides sensitivity of model to data
+ use data more efficiently
- Computational cost

Keep in mind:

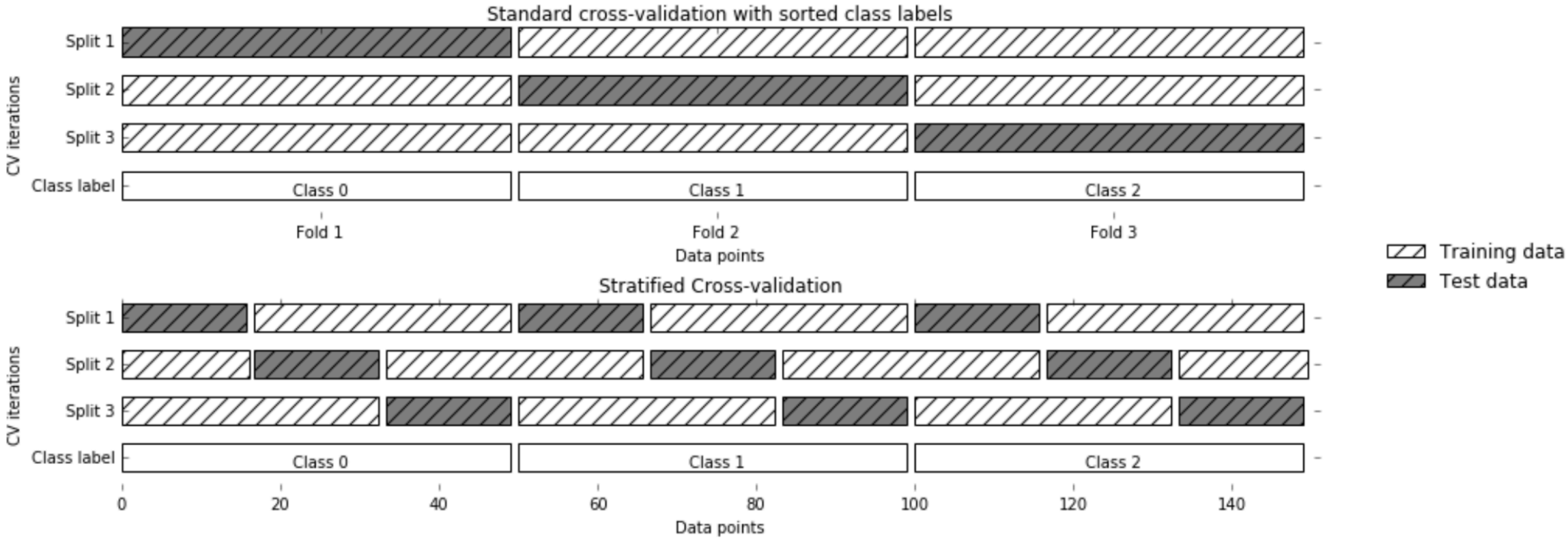
Cross-validation is not mode building! The purpose of cross-validation is only to evaluate how well a given algorithm will generalize when trained on a specific dataset.

Stratified Cross Validation

Stratified Cross Validation



Stratified Cross Validation

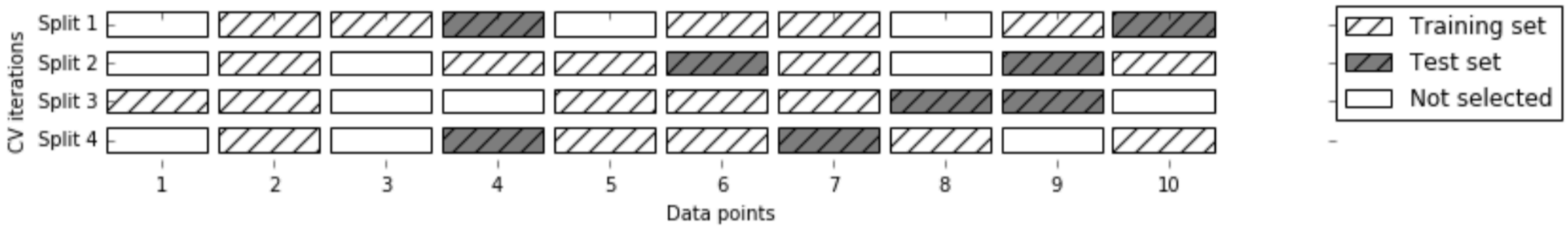


Split the data such that the proportions between classes are the same in each fold as they are in the whole dataset

Other Cross Validation Techniques

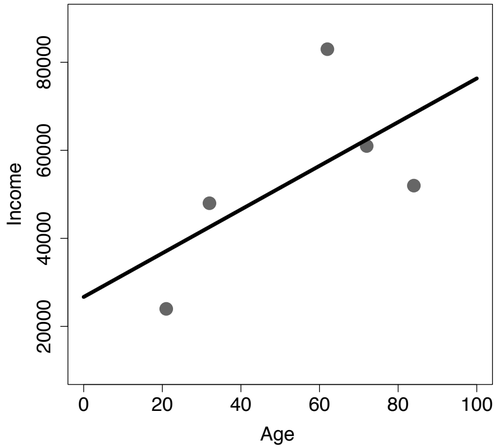
- *Leave-One-Out* Cross Validation
  - in each split, one datapoint is the test set
  - time consuming for large datasets
  - better estimation for smaller datasets

- *Shuffle-Split* Cross Validation

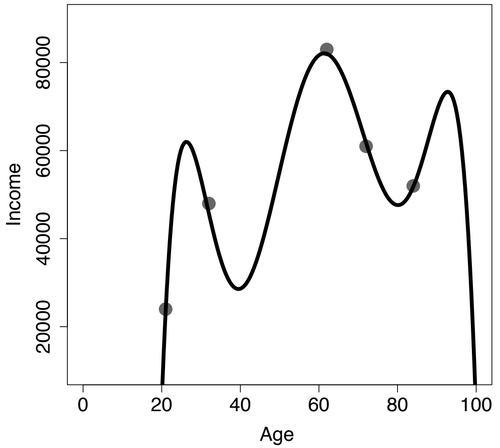


Bias-variance tradeoff in ML

- *Bias*: error due to assumptions
  - High bias = underfitting
  - model does not capture sufficient characteristics of the data



- Variance: *error* due to sensitivity of model
  - High variance = overfitting
  - model captures noise in the data



Exercise

The following is a table for credit scoring. Given are two consistent models **Model 1** and **Model 2**

ID	Occupation	Age	Loan-Salary Ratio	Outcome
1	industrial	39	3.40	default
2	industrial	22	4.02	default
3	professional	30	2.7 0	repay
4	professional	27	3.32	default
5	professional	40	2.04	repay
6	professional	50	6.95	default
7	industrial	27	3.00	repay
8	industrial	33	2.60	repay
9	industrial	30	4.5 0	default
10	professional	45	2.78	repay

Model 1

if Loan-Salary Ratio $>$ 3.00 **then**
  Outcome = *default*
else
  Outcome = *repay*

Model 2

if Age= 50 **then**
  Outcome = *default*
else if Age= 39 **then**
  Outcome = *default*
else if Age= 30 **and** Occupation = *industrial* **then**
  Outcome = *default*
else if Age= 27 **and** Occupation = *professional* **then**
  Outcome = *default*
else
  Outcome = *repay*

ID	Occupation	Age	Loan-Salary Ratio	Outcome
1	industrial	39	3.40	default
2	industrial	22	4.02	default
3	professional	30	2.7 0	repay
4	professional	27	3.32	default
5	professional	40	2.04	repay
6	professional	50	6.95	default
7	industrial	27	3.00	repay
8	industrial	33	2.60	repay
9	industrial	30	4.5 0	default
10	professional	45	2.78	repay

Model 1

if Loan-Salary Ratio $>$ 3.00 **then**
    Outcome = *default*
else
    Outcome = *repay*

Model 2

if Age= 50 **then**
    Outcome = *default*
else if Age= 39 **then**
    Outcome = *default*
else if Age= 30 **and** Occupation = *industrial* **then**
    Outcome = *default*
else if Age= 27 **and** Occupation = *professional* **then**
    Outcome = *default*
else
    Outcome = *repay*

(a) Which of these two models will generalize better to the instances not contained in the dataset?

(b) Propose an inductive bias that would enable a machine learning algorithm to make the same preference choice as you did in part (a)

(c) Do you think the model you rejected in part (a) is overfitting or underfitting the data?

Example

(a) Which of these two models will generalize better to the instances not contained in the dataset?

(b) Propose an inductive bias that would enable a machine learning algorithm to make the same preference choice as you did in part (a)

(c) Do you think the model you rejected in part (a) is overfitting or underfitting the data?

Example

(a) Which of these two models will generalize better to the instances not contained in the dataset?

Model 1 is more likely to generalise beyond the training dataset because it is simpler and appears to be capturing a real pattern in the data.

(b) Propose an inductive bias that would enable a machine learning algorithm to make the same preference choice as you did in part (a)

Example

(a) Which of these two models will generalize better to the instances not contained in the dataset?

Model 1 is more likely to generalise beyond the training dataset because it is simpler and appears to be capturing a real pattern in the data.

(b) Propose an inductive bias that would enable a machine learning algorithm to make the same preference choice as you did in part (a)

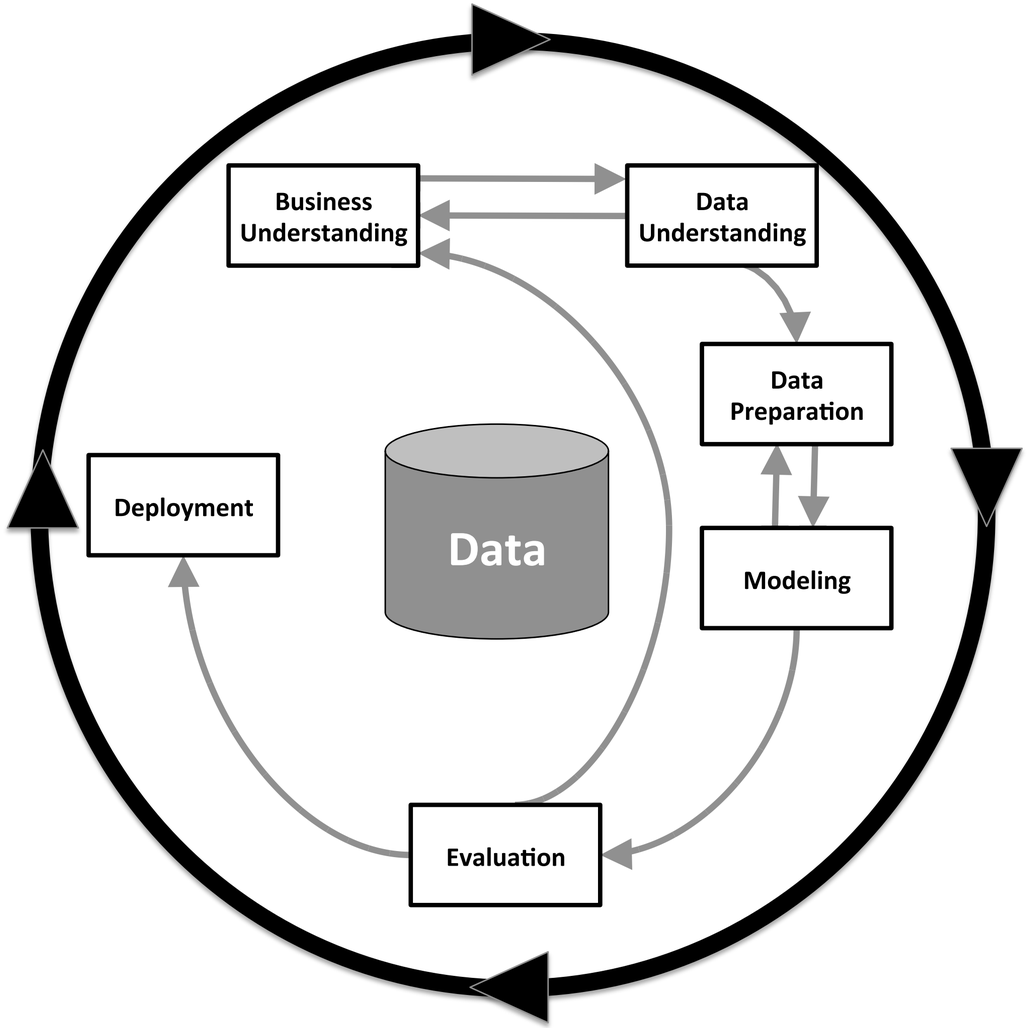
If you are choosing between a number of models that perform equally well then prefer the simpler model over the more complex models.

Example

(c) Do you think the model you rejected in part (a) is overfitting or underfitting the data?

Model 2 is overfitting the data. All of the decision rules in this model that predict OUTCOME = *default* are specific to single instances in the dataset. Basing predictions on single instances is indicative of a model that is overfitting.

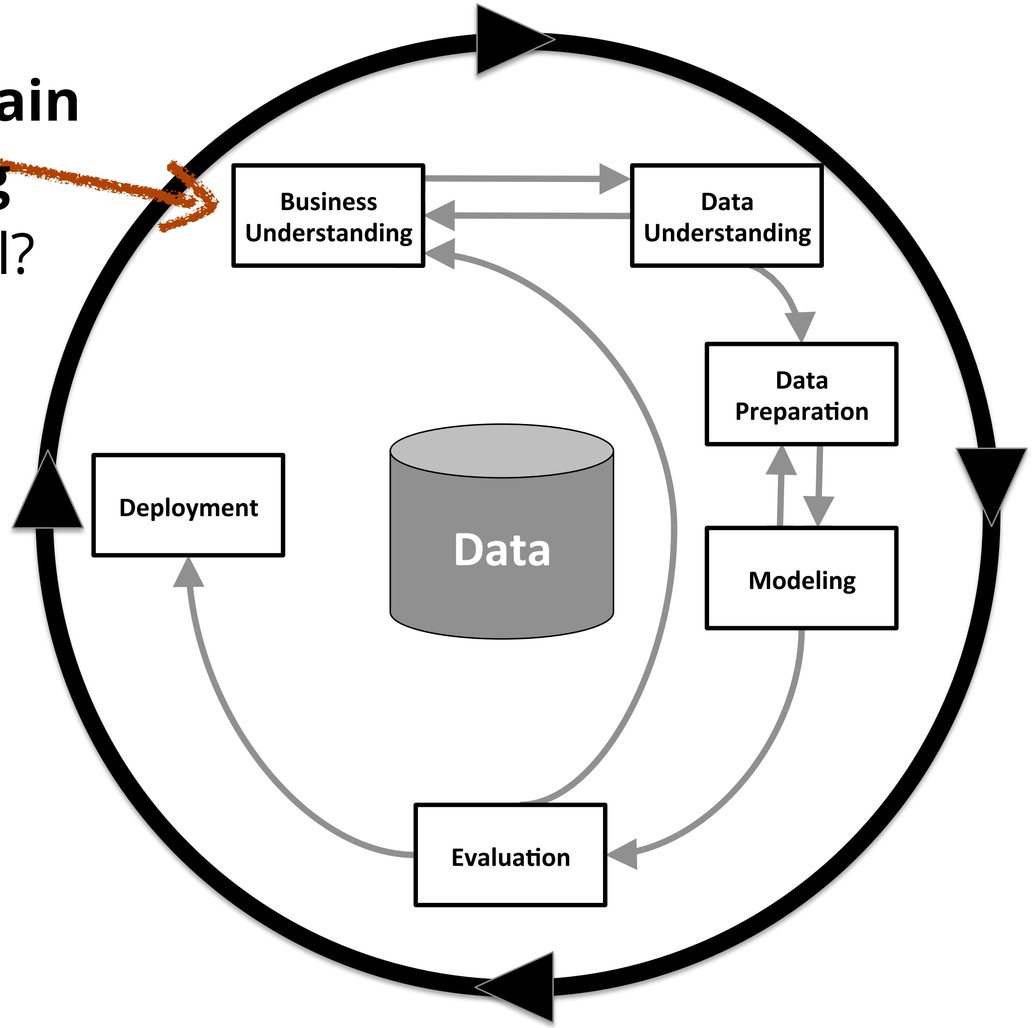
Machine Learning in Practice



Machine Learning in Practice

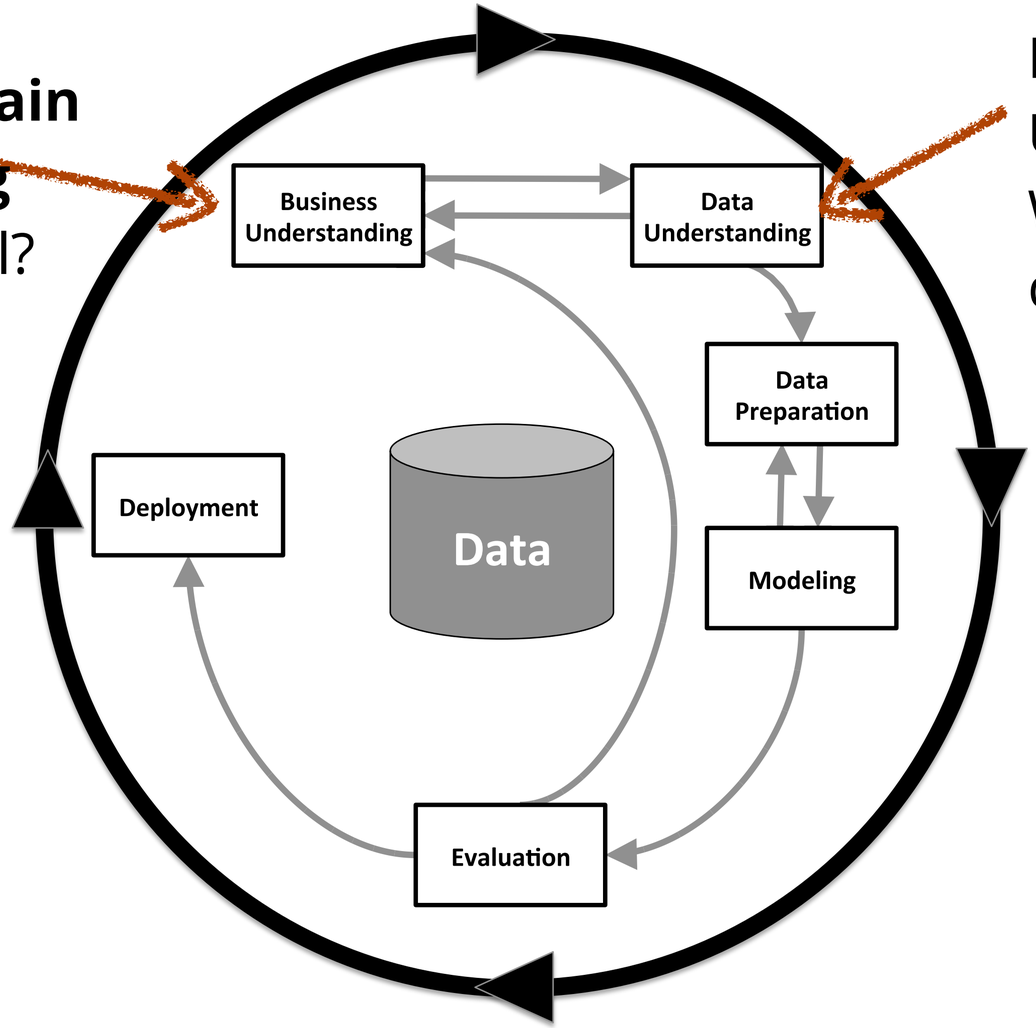
Business/Domain Understanding

What is the goal?



Machine Learning in Practice

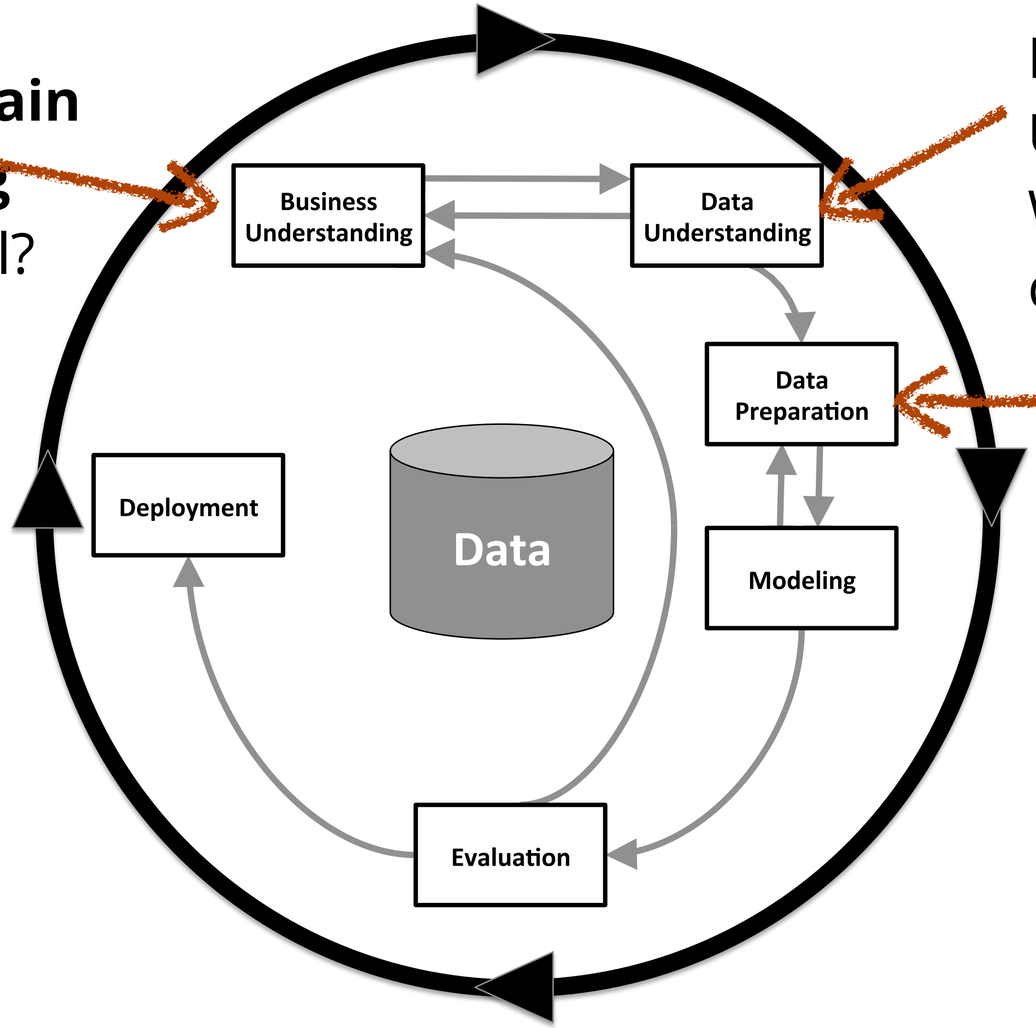
Business/Domain Understanding
What is the goal?



Data Understanding
What are the data sources?

Machine Learning in Practice

Business/Domain Understanding
What is the goal?

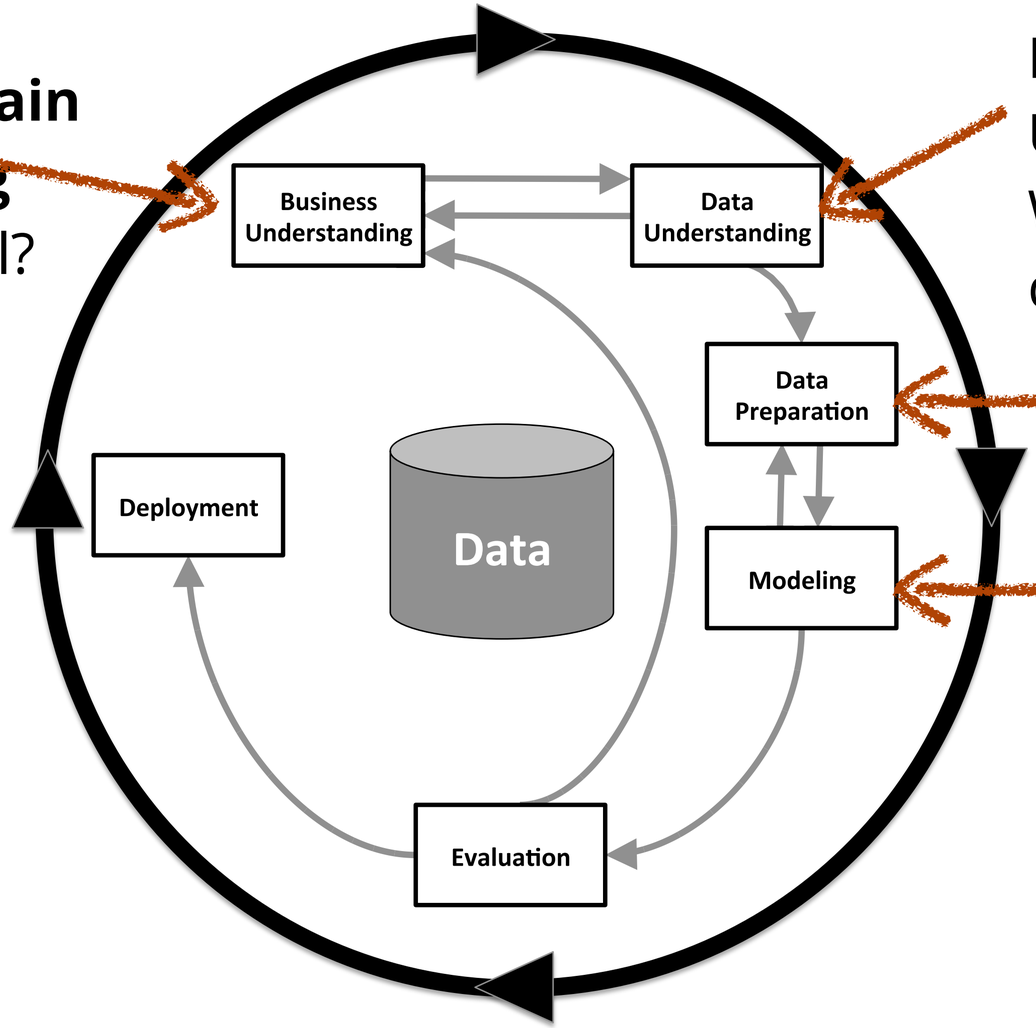


Data Understanding
What are the data sources?

Data Preparation

Machine Learning in Practice

Business/Domain Understanding
What is the goal?



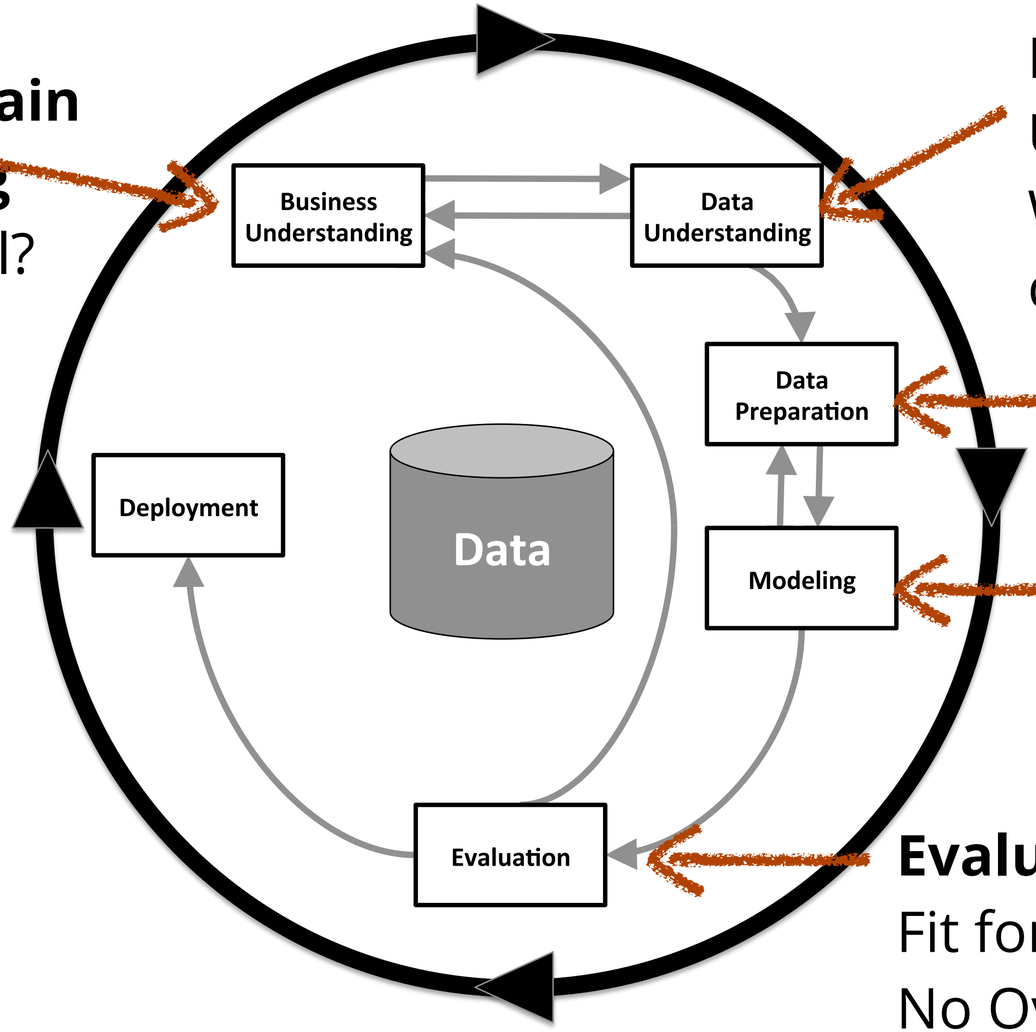
Data Understanding
What are the data sources?

Data Preparation

Modeling
Machine learning

Machine Learning in Practice

Business/Domain Understanding
What is the goal?



Data Understanding
What are the data sources?

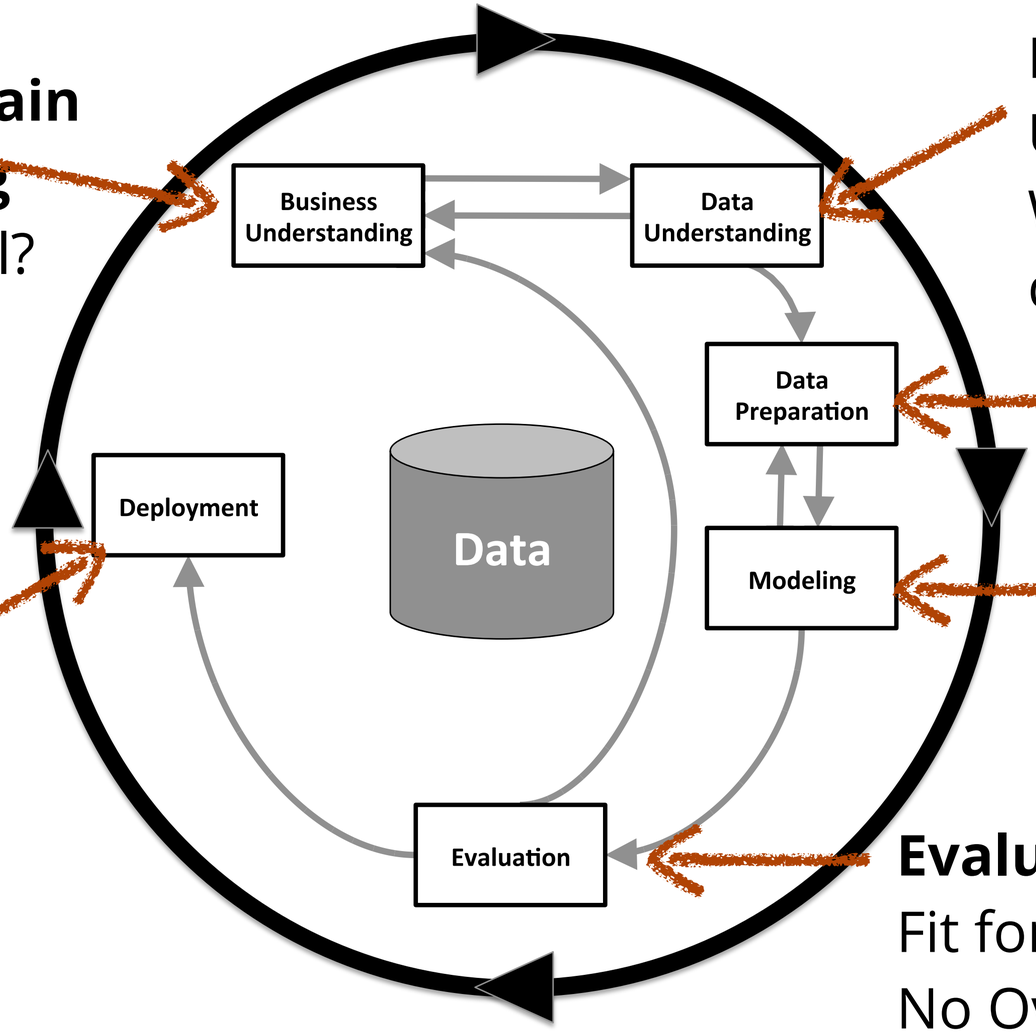
Data Preparation

Modeling
Machine learning

Evaluation
Fit for purpose
No Over/Underfitting

Machine Learning in Practice

Business/Domain Understanding
What is the goal?



Data Understanding
What are the data sources?

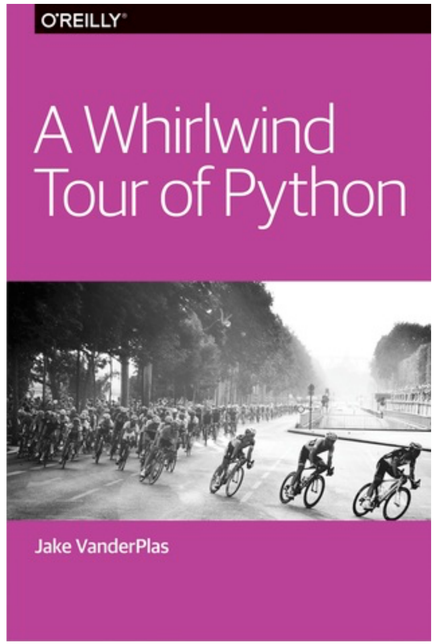
Data Preparation

Modeling
Machine learning

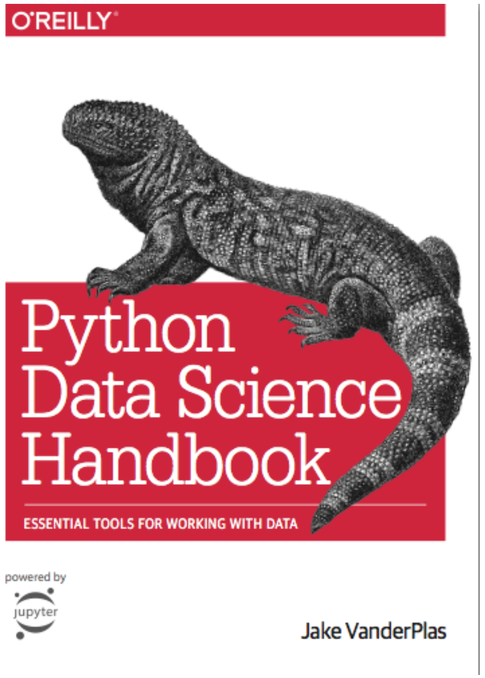
Deployment
Integrate into real life environment

Evaluation
Fit for purpose
No Over/Underfitting

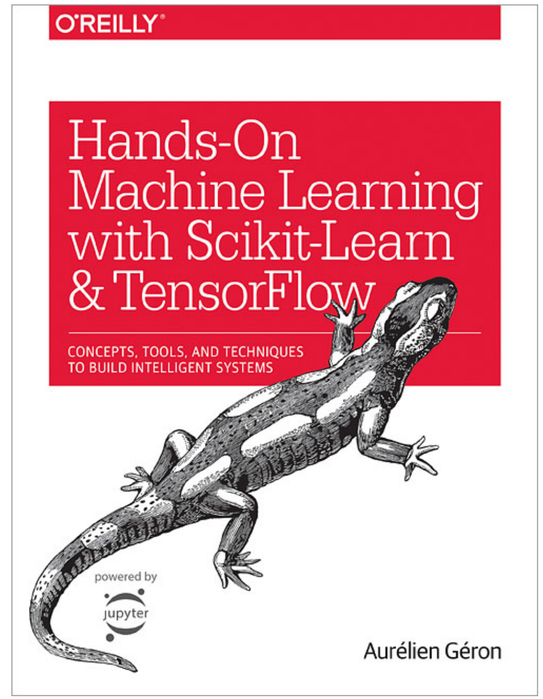
Python Resources



https://github.com/jakevdp/WhirlwindTourOfPython



https://github.com/jakevdp/PythonDataScienceHandbook



https://github.com/ageron/handson-ml

Python activity

- A2-scikit-learn.ipynb
- Neural network based classification and regression
- use of scikit-learn library & built-in datasets

Questions?