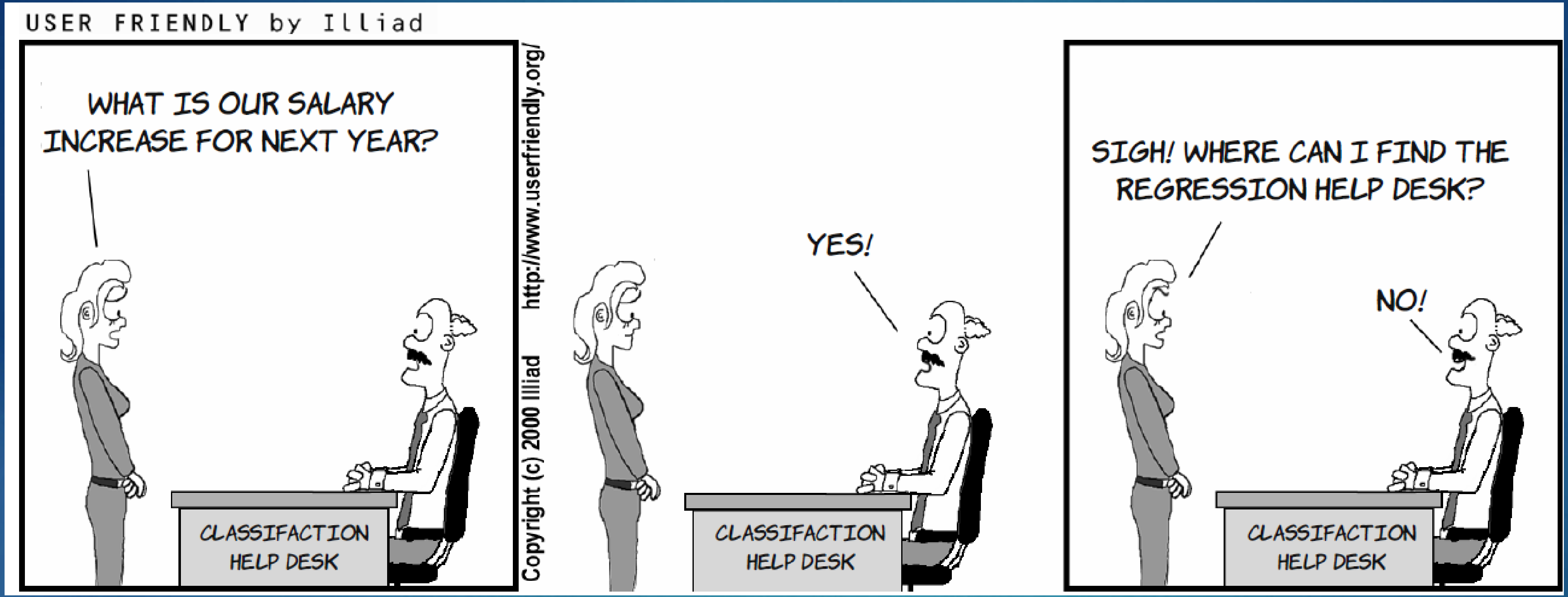
COMP-2704: Supervised Machine Learning

WEEK 2


USER FRIENDLY by Illiad
WHAT IS OUR SALARY INCREASE FOR NEXT YEAR?
CLASSIFACTION HELP DESK
http://www.userfriendly.org/
Copyright (c) 2000 Illiad
YES!
CLASSIFACTION HELP DESK
SIGH! WHERE CAN I FIND THE REGRESSION HELP DESK?
NO!
CLASSIFACTION HELP DESK

Chapter 2: Types of Machine Learning

Three main families of machine learning

- Different types of data need to be treated differently in ML.
- ML models can be grouped into three main families according to the type of data they are built for:
  - Supervised learning
  - Unsupervised learning
  - Reinforcement learning
- Each family contains many different kinds of models.

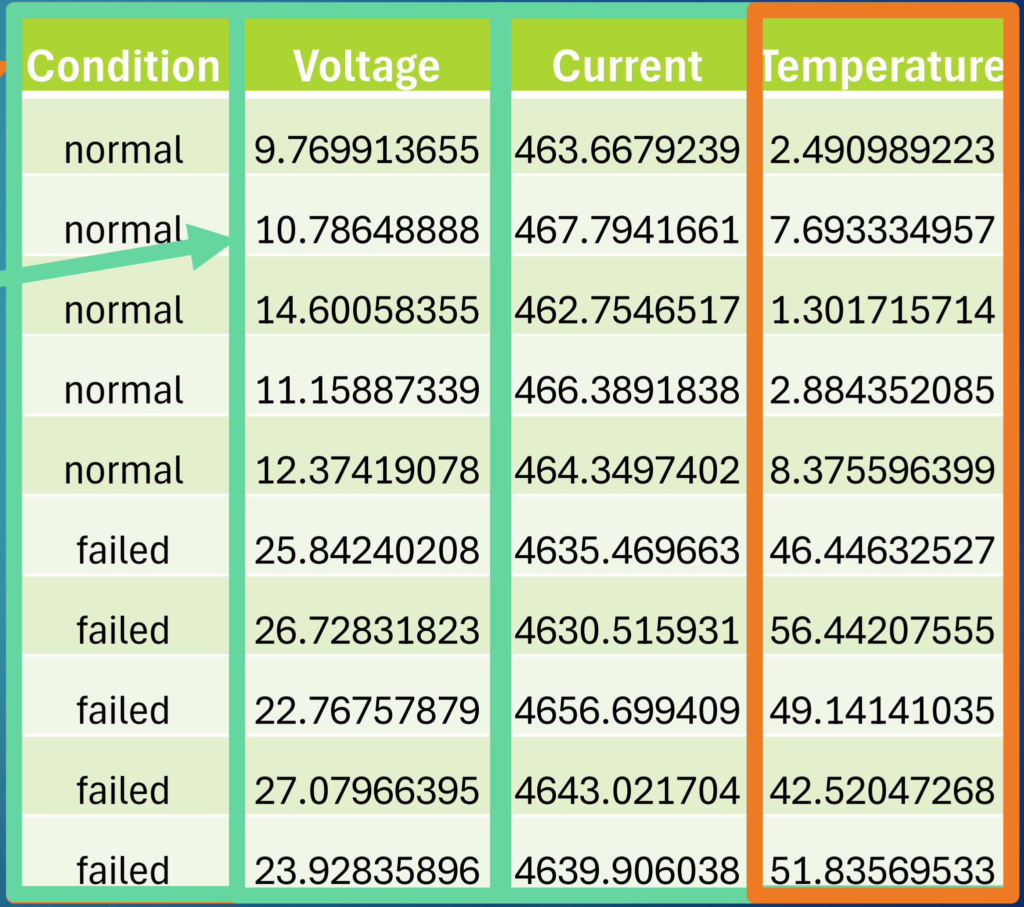
Data

- **Q: What is data?**
  - Data is information.
- **Q: What forms does data come in?**
  - Tables of numbers, strings, GPS coordinates, timestamps, …
  - Images
  - Written text
  - Audio files
  - Sensor readings
  - …

Labeled data is for supervised learning

- This table shows 10 ***samples*** of data.
- The condition column gives a ***label*** for each sample.
- The other columns are called ***features***.
- ***Supervised*** models can learn to ***predict*** the label from the features.
- Which column is treated as the label depends on the use case.

Q: What is the prediction made in this case?



Condition	Voltage	Current	Temperature
normal	9.769913655	463.6679239	2.490989223
normal	10.78648888	467.7941661	7.693334957
normal	14.60058355	462.7546517	1.301715714
normal	11.15887339	466.3891838	2.884352085
normal	12.37419078	464.3497402	8.375596399
failed	25.84240208	4635.469663	46.44632527
failed	26.72831823	4630.515931	56.44207555
failed	22.76757879	4656.699409	49.14141035
failed	27.07966395	4643.021704	42.52047268
failed	23.92835896	4639.906038	51.83569533

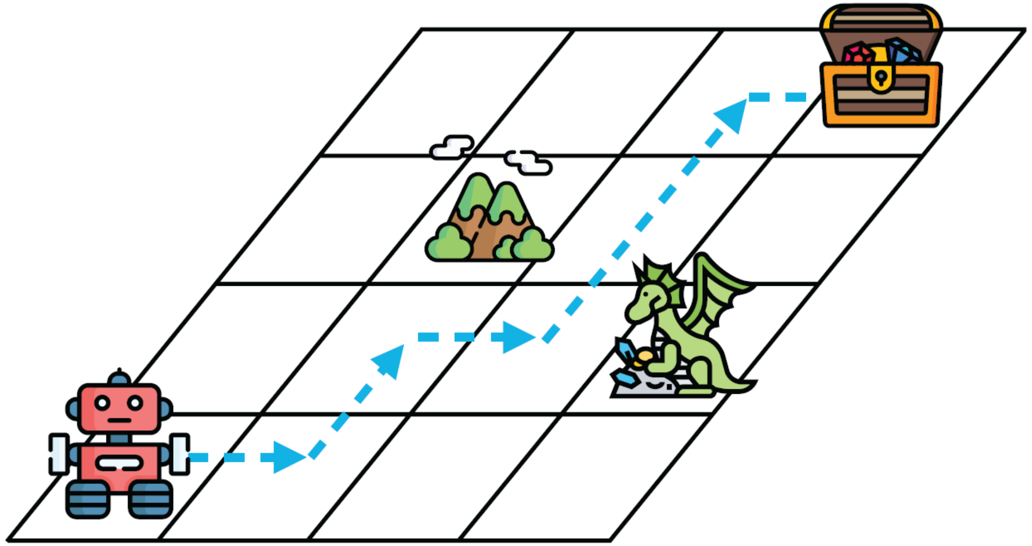
Unlabeled data is for unsupervised learning

- What if there is no label?
- This type of data requires a different approach.
- ***Unsupervised*** models can:
  - Find similarities between samples and group them into ***clusters***.
  - Reduce the number of features used to represent each sample (***dimensionality reduction***).
  - ***Generate*** new data that is similar.

Voltage	Current	Temperature
9.769913655	463.6679239	2.490989223
10.78648888	467.7941661	7.693334957
14.60058355	462.7546517	1.301715714
11.15887339	466.3891838	2.884352085
12.37419078	464.3497402	8.375596399
25.84240208	4635.469663	46.44632527
26.72831823	4630.515931	56.44207555
22.76757879	4656.699409	49.14141035
27.07966395	4643.021704	42.52047268
23.92835896	4639.906038	51.83569533

Environmental data is for reinforcement learning

- In ***Reinforcement*** learning, an ***agent*** learns to achieve goals based on environment data.
- A simple example is learning to traverse a board game to reach a reward.
- More complex examples are a self-driving car or chess engines.
- Realtime Data is input to the model, often using cameras and other sensors.
- The agent learns from trial and error which actions lead to the most favourable results.



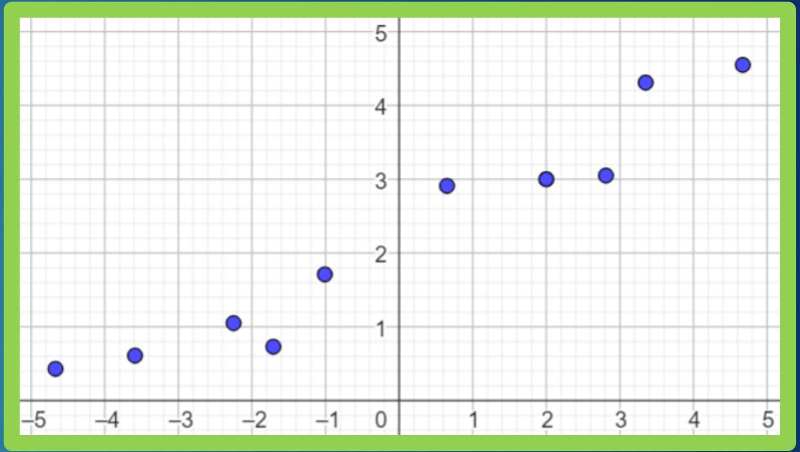
Data quality

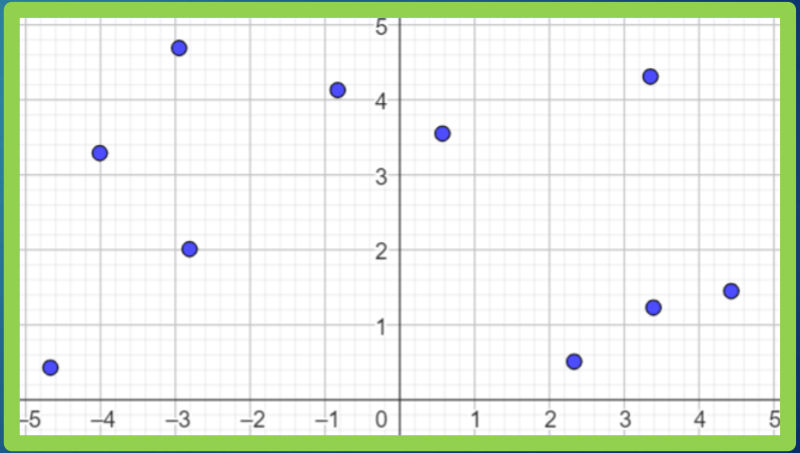
- All kinds of machine learning models need to be ***trained*** on data.
- A model can only be as good as the data it is trained on.

Q: Which dataset would produce a linear regression model with lower error?

Q: Is there anything that could be done to lower the error for the data on the bottom?

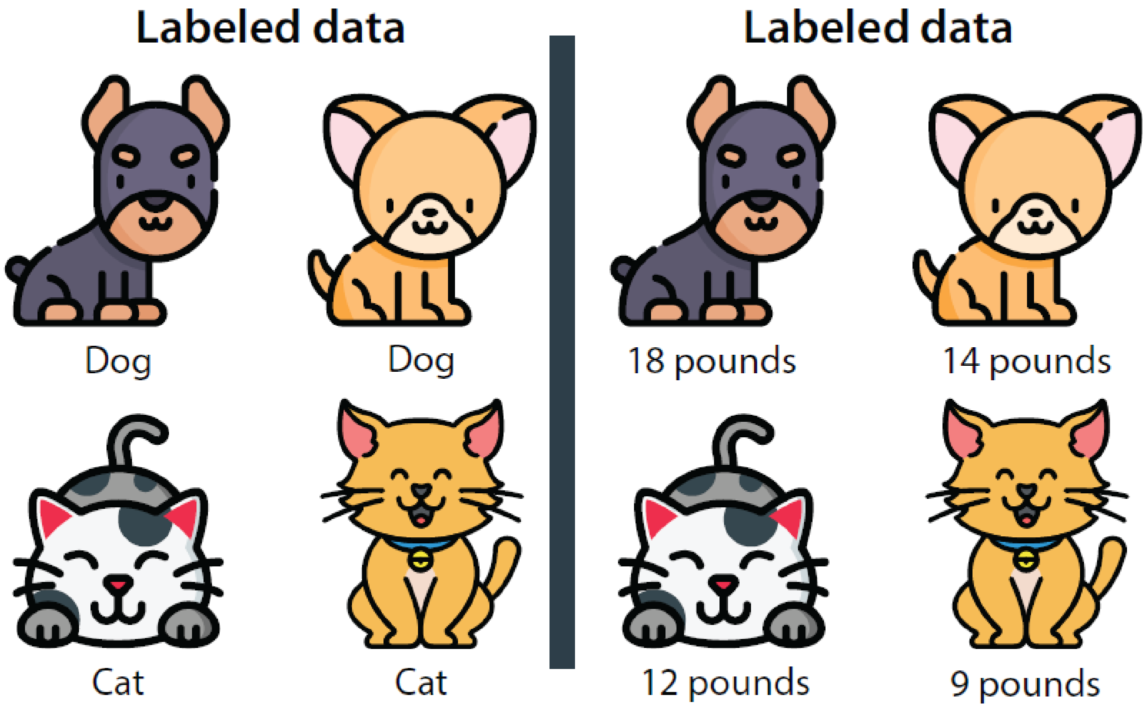
- Data quality is important for all kinds of ML, although it is often harder to visualize.



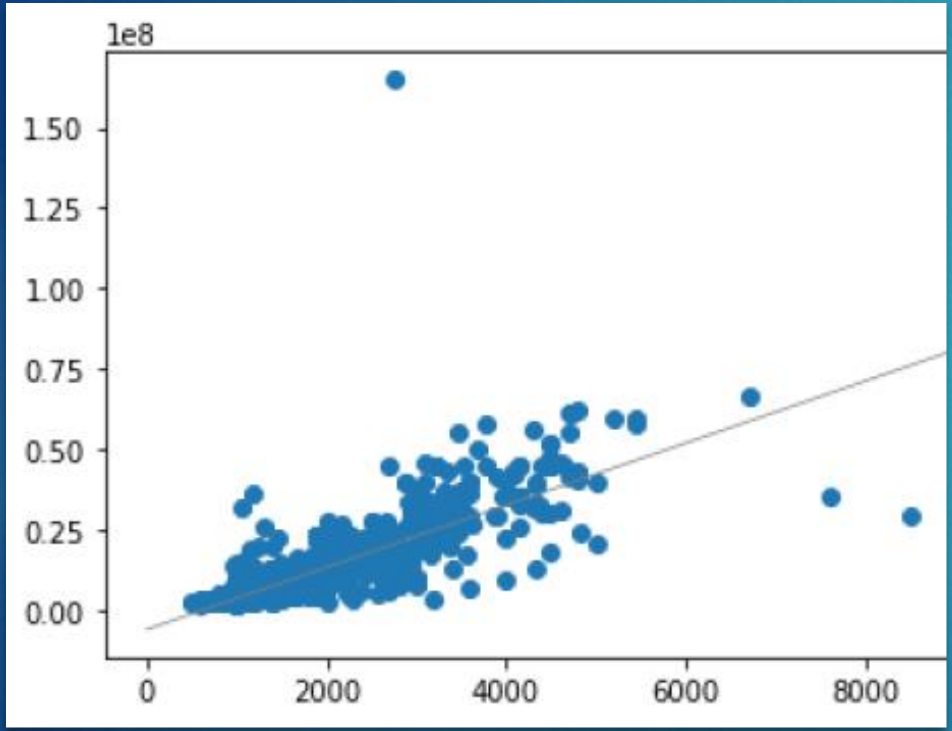


Back to supervised learning

- Labels can be categorical or numerical.
- There are two kinds of supervised learning:
  - ***Regression*** models predict numerical labels.
  - ***Classification*** models predict categorical labels.



Regression



- A regression model can predict the numerical label for a sample from its features.
- The features need to supply useful information.
- Let’s discuss a use case for linear regression: house price prediction.

Q: What data (features and label) would you use to train a model for house price predictions?

Q: What data type is each feature? What data type is the label?

Q: How would you obtain this data?

Q: How would the model be used in practice?

Classification

- A classification model can predict the categorical label of a sample from its features.
- ***Binary*** classification is when there are only two choices of label in the dataset.
  - Example: given a set of images of animals (including many dogs), predict whether or not a dog is in an image.
- ***Multiclass*** classification is when there are more than two choices of label in the dataset.
  - Example: given a set of images of only dogs, cats and frogs, predict which of these three is in an image.

Classification use cases

Q: How would you create the following models if you were an employee working in the related area? Discuss a) what features and label you would use, b) the datatype of each, c) where you would get the data, and d) how you would implement the model.

1. A model which recommends videos to a user.
2. A model that predicts whether a user will click a link.
3. A model that classifies a movie review as positive or negative.
4. A model that predicts whether a social media user will make friends with another user.