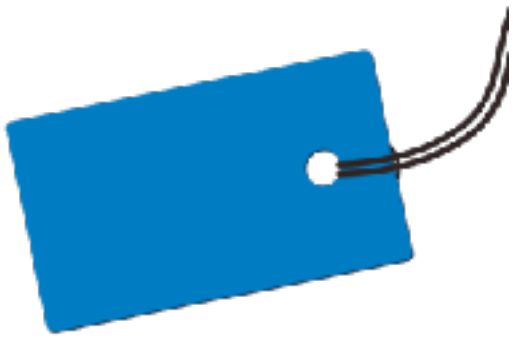
Defining The Right Label

Smart setting of population tag

Adi Nesher I Sep 2016

Table of contents



Why do we need a Label?

- Supervised and Unsupervised learning
- Semi supervised learning

Manual labeling

Situations where a tag is not obvious – PayPal use case

- Different approached for affective labeling
- Best practices

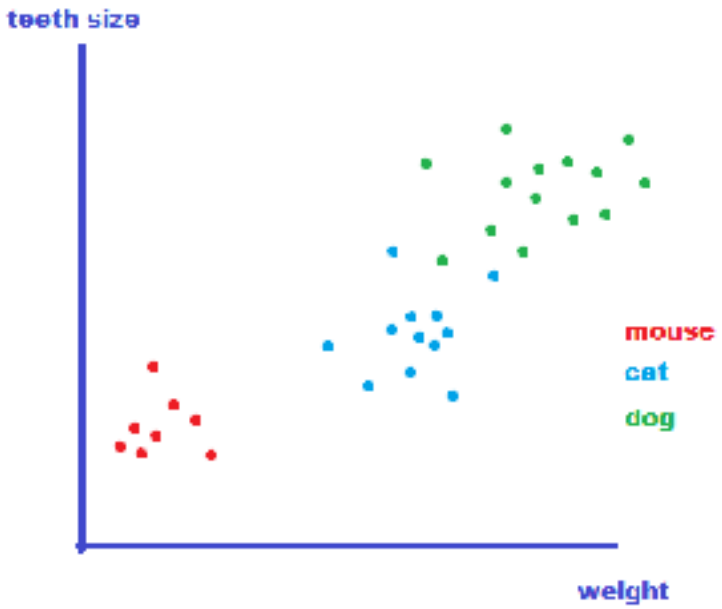
Why do we need a label?

- Label is the predicted value
- Sometimes it's obvious definition your data will generate:
  - Job failed/succeeded
  - Weather temp
  - Conversion Rate
  - Patient dead or alive

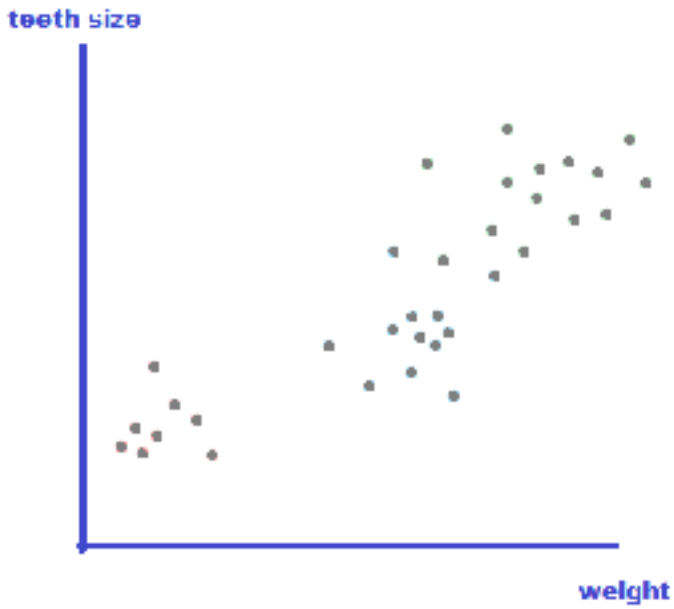
- Sometime it's not obvious at all:
  - Good/bad client
  - Movie category
  - Picture meaning

Supervised and Unsupervised learning

Supervised learning



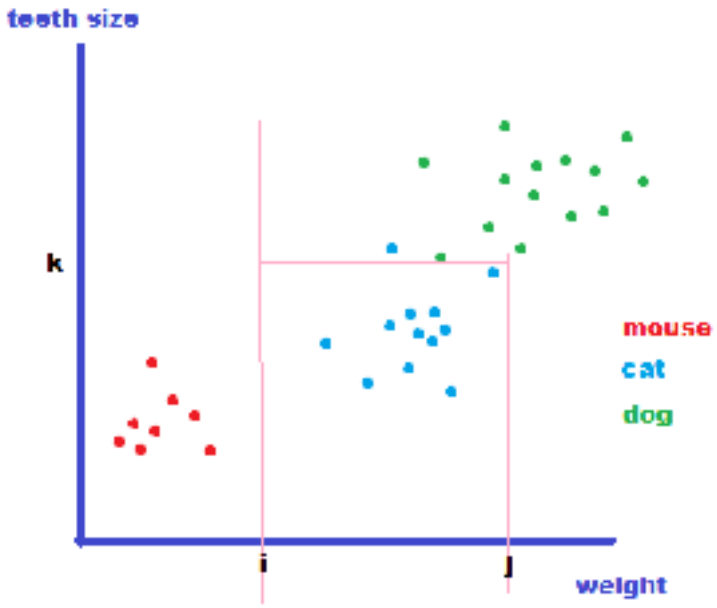
Unsupervised learning



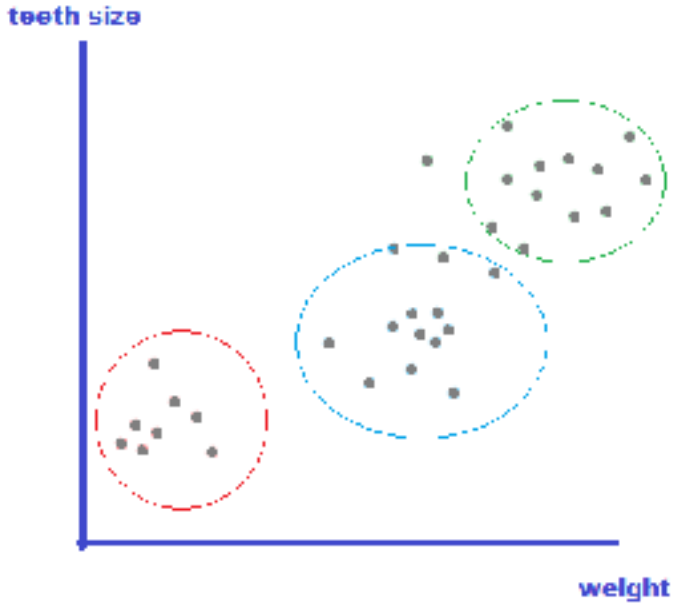
Different problems require different algorithms

Supervised and Unsupervised learning

Supervised learning



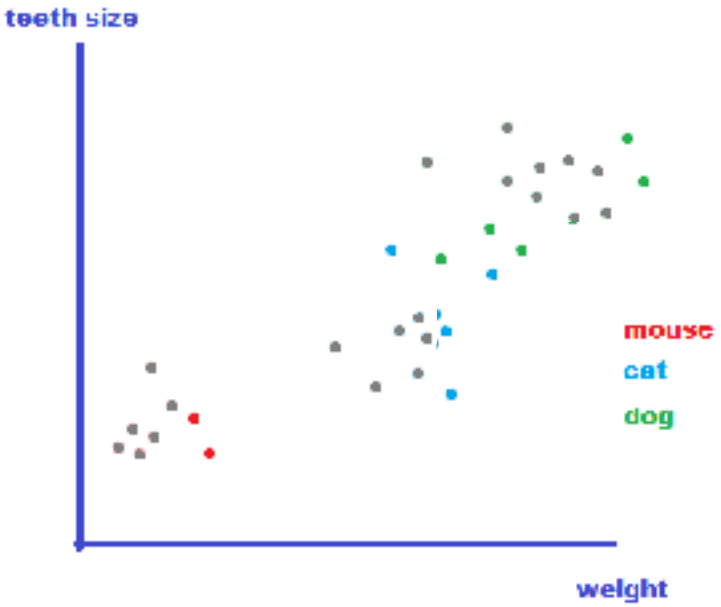
Unsupervised learning



Different problems require different algorithms

Semi Supervised learning

Semi Supervised learning



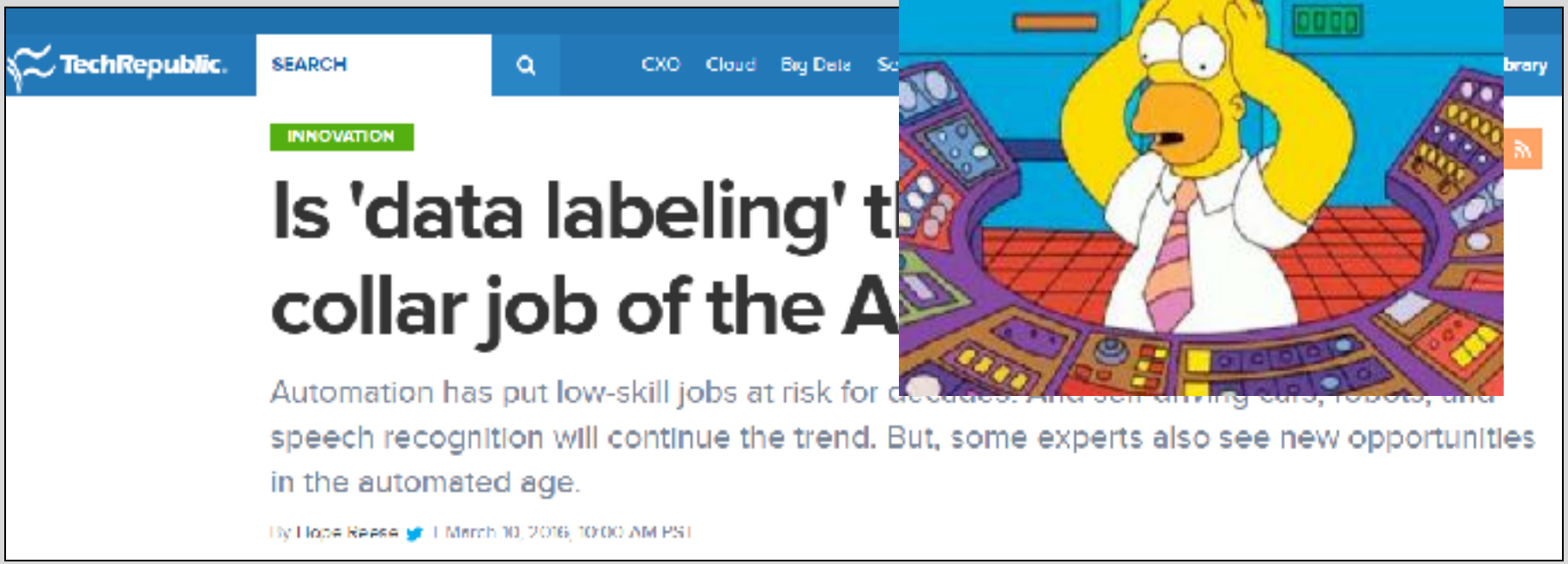
Many methods of leveraging the label data to help classify the whole population:

- Self training
- Generative models
- S3VMs
- Many More…

Manual Labeling

Google image labeler

Manual Labeling



Creating a Label in situations of uncertainty

- PayPal's use case:

Who is a bad client?

What is fraud?

PayPal Risk Product in a Nutshell

Creating a Label in situations of uncertainty

- PayPal's use case: **who is a bad client?** **What is fraud?**

5 Claims
$100 net loss

Lifetime transactions: 7
Selling iPhone
6 didn't reach destination

1 Claims
$500 net loss

Lifetime transactions: 39
Selling furniture
1 order had a broken tip

Creating a Label in situations of uncertainty

- different approaches:

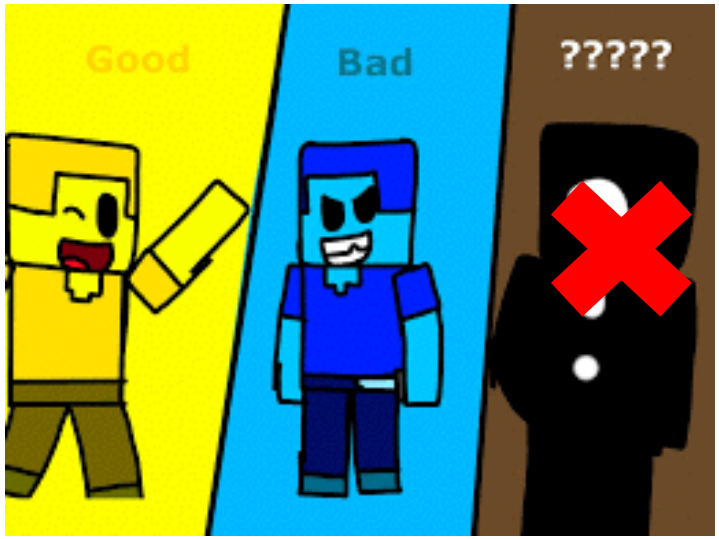
- Logically define thresholds

- Business definition that would keep us margin positive (if exist)
- Create a learning model on absolute ‘bads’

  - Regression or classification model to apply the right thresholds for bads

Creating a Label in situations of uncertainty

- Best practice:
  - Explain the factor we try to optimize (loss in our case)
  - Avoid gray population in future training
  - Occam's razor : Keep it simple, even if not with perfect accuracy



PayPal