Understanding Feature Space in Machine Learning

Alice Zheng, Dato

September 9, 2015

My journey so far

Carnegie Mellon University

Microsoft Research

GraphLab

Dato

Applied machine learning
(Data science)

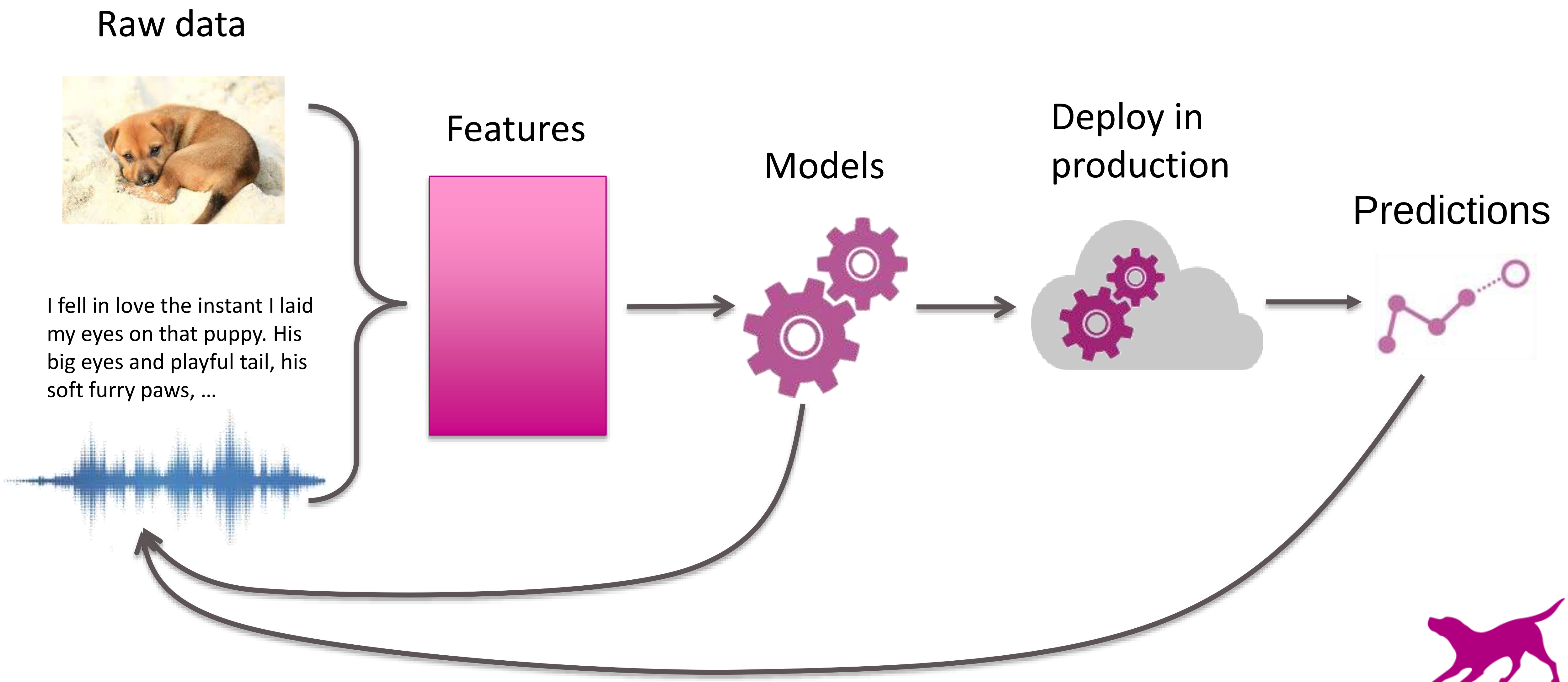
Shortage of experts
and good tools.

Build ML tools

Why machine learning?

Model data.
Make predictions.
Build intelligent applications.

The machine learning pipeline



Feature = numeric representation of raw data

Representing natural text

Raw Text → Bag of Words → Classify: puppy or not?

It is a puppy and it is extremely cute.

What’s important? Phrases? Specific words? Ordering? Subject, object, verb?

{“it”:2,
“is”:2,
“a”:1,
“puppy”:1,
“and”:1,
“extremely”:1,
“cute”:1 }

Representing natural text

Raw Text → Bag of Words → Classify: puppy or not?

It is a puppy and it is extremely cute.

it	2
they	0
I	1
am	0
how	0
puppy	1
and	1
cat	0
aardvark	0
cute	1
extremely	1
…	…

Sparse vector representation

Representing images

Raw Image → Bag of Visual Words → Classify: person or animal?

Raw image:
millions of RGB triplets,
one for each pixel

Image source: "Recognizing and learning object categories,"
Li Fei-Fei, Rob Fergus, Anthony Torralba, ICCV 2005—2009.

Representing images

Raw Image → Deep learning features → Classify: person or animal?

3.29
-15
-5.24
48.3
1.36
47.1
-
1.92
36.5
2.83
95.4
-19
-89
5.09
37.8

Dense vector representation

Feature space in machine learning

- Raw data → high dimensional vectors
- Collection of data points → point cloud in feature space
- Model = geometric summary of point cloud
- Feature engineering = creating features of the appropriate granularity for the task

Crudely speaking, mathematicians fall into two categories: the algebraists, who find it easiest to reduce all problems to sets of numbers and variables, and the geometers, who understand the world through shapes.

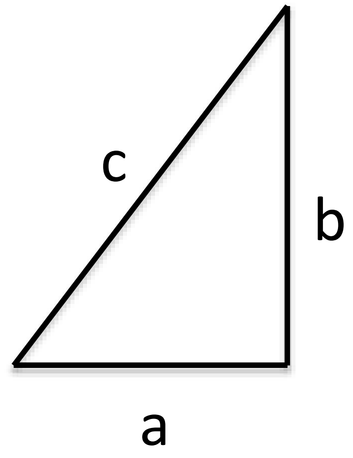
-- Masha Gessen, “Perfect Rigor”

Algebra vs. Geometry

Algebra — Pythagorean Theorem — Geometry

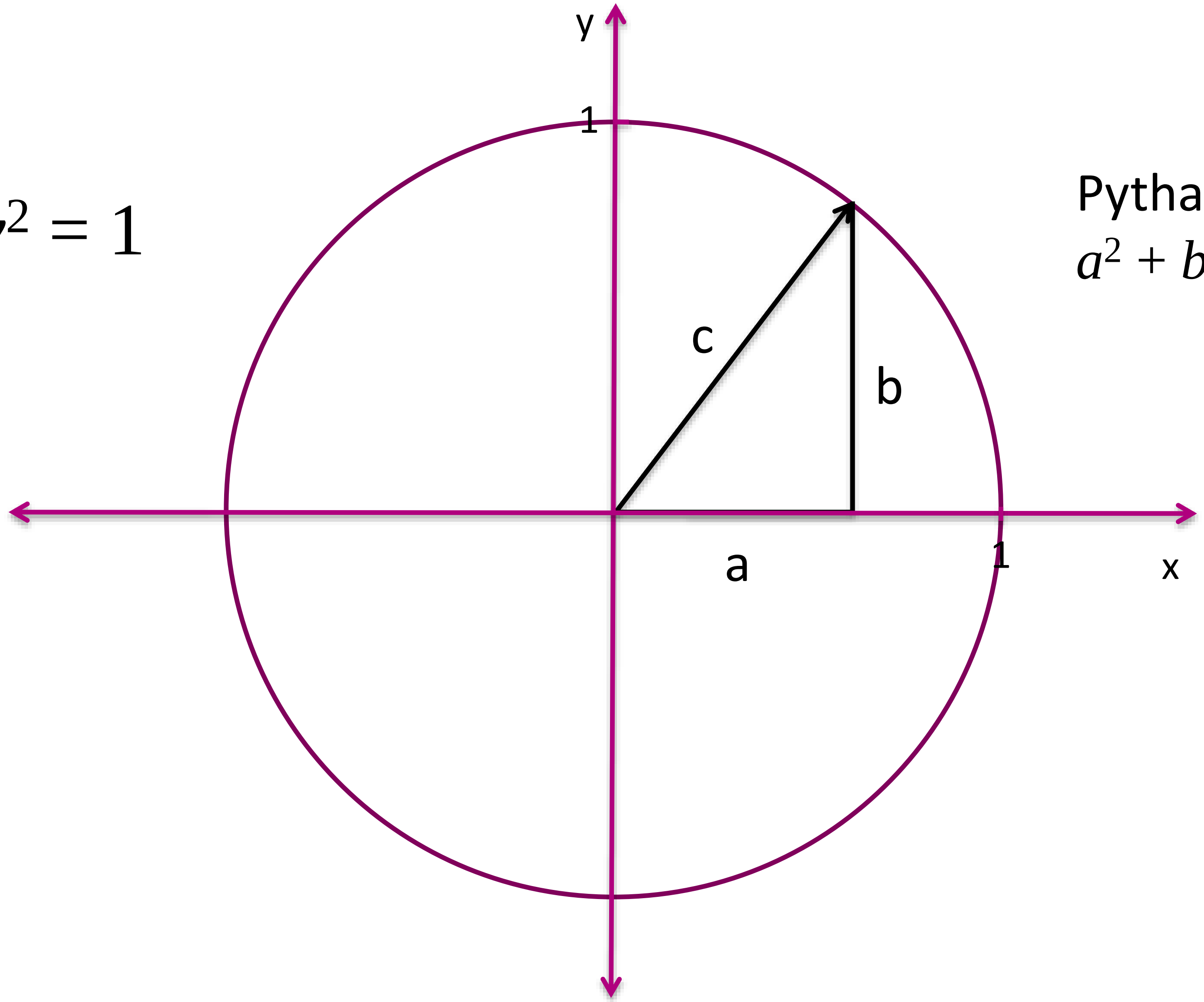
$a^2 + b^2 = c^2$

(Euclidean space)



Visualizing a sphere in 2D

$$x^2 + y^2 = 1$$

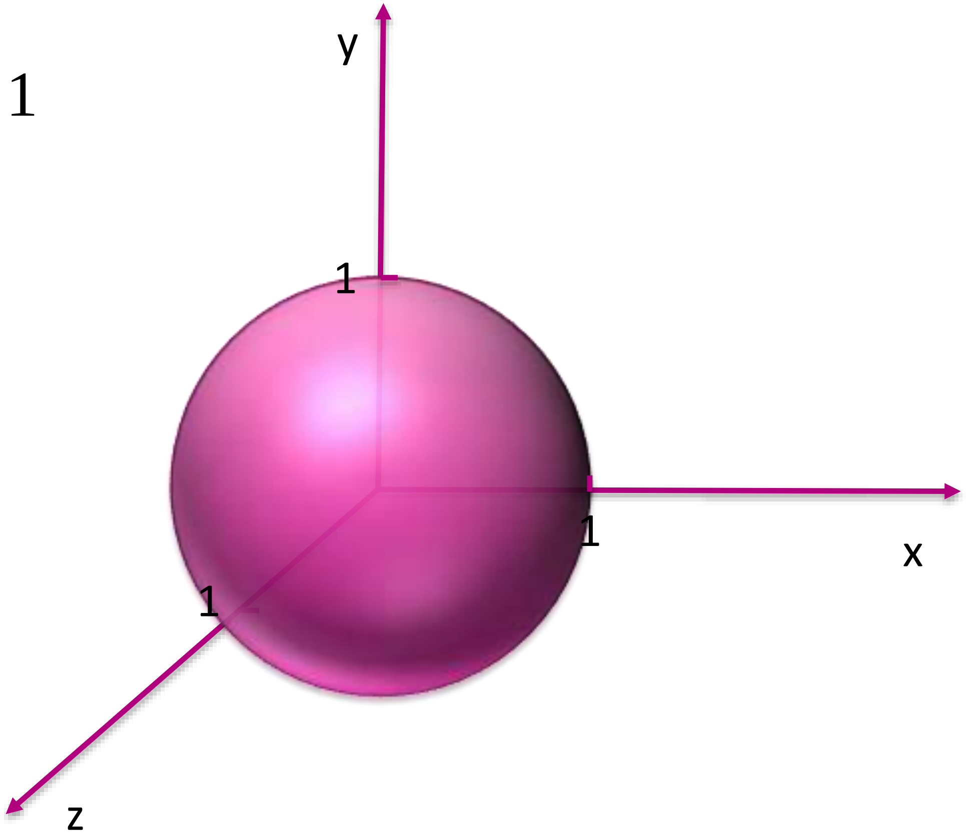


Pythagorean theorem:

$a^2 + b^2 = c^2$

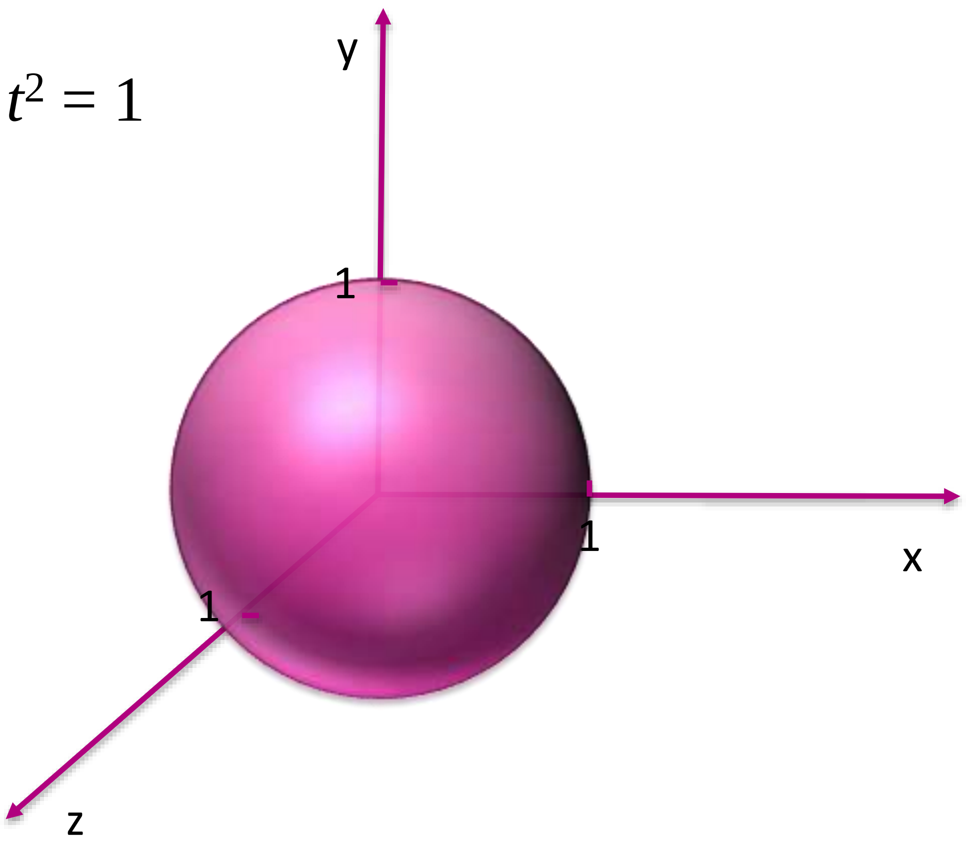
Visualizing a sphere in 3D

$$x^2 + y^2 + z^2 = 1$$

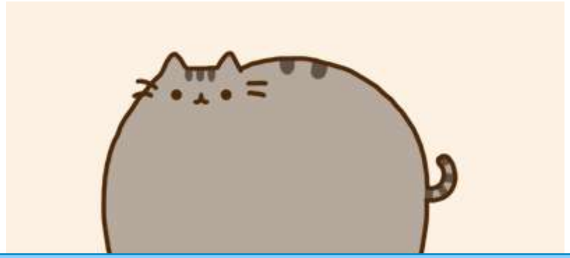


Visualizing a sphere in 4D

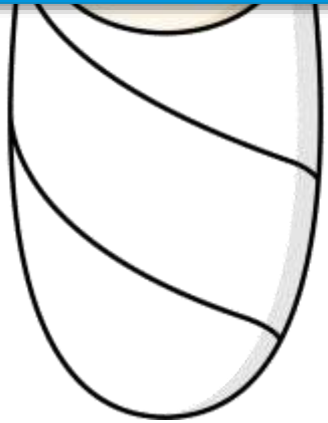
$$x^2 + y^2 + z^2 + t^2 = 1$$



Why are we looking at spheres?



Poincaré Conjecture:
All physical objects without holes is “equivalent” to a sphere.



The power of higher dimensions

- A sphere in 4D can model the birth and death process of physical objects
- Point clouds = approximate geometric shapes
- High dimensional features can model many things

Visualizing Feature Space

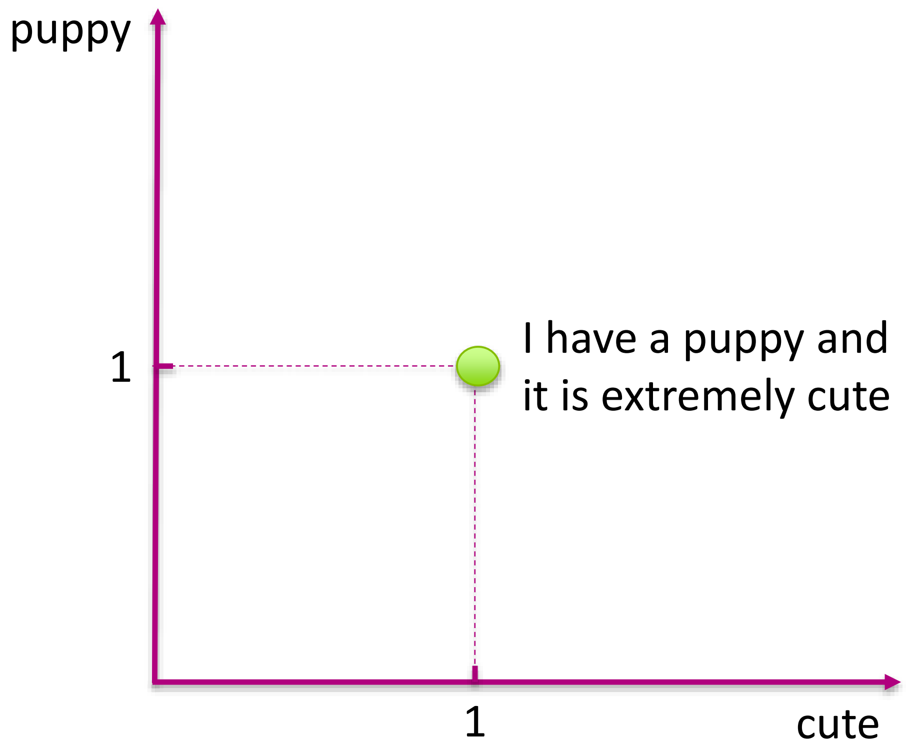
The challenge of high dimension geometry

- Feature space can have hundreds to millions of dimensions
- In high dimensions, our geometric imagination is limited
  - Algebra comes to our aid

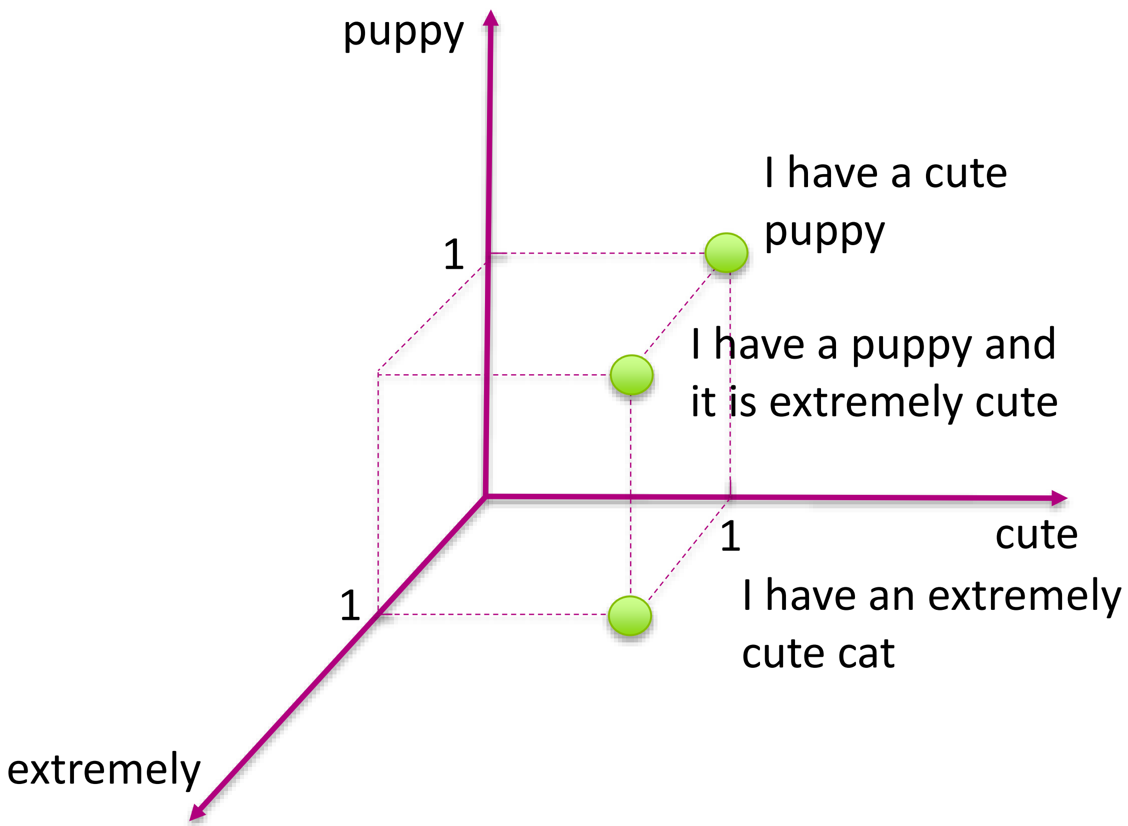
Visualizing bag-of-words

I have a puppy and it is extremely cute

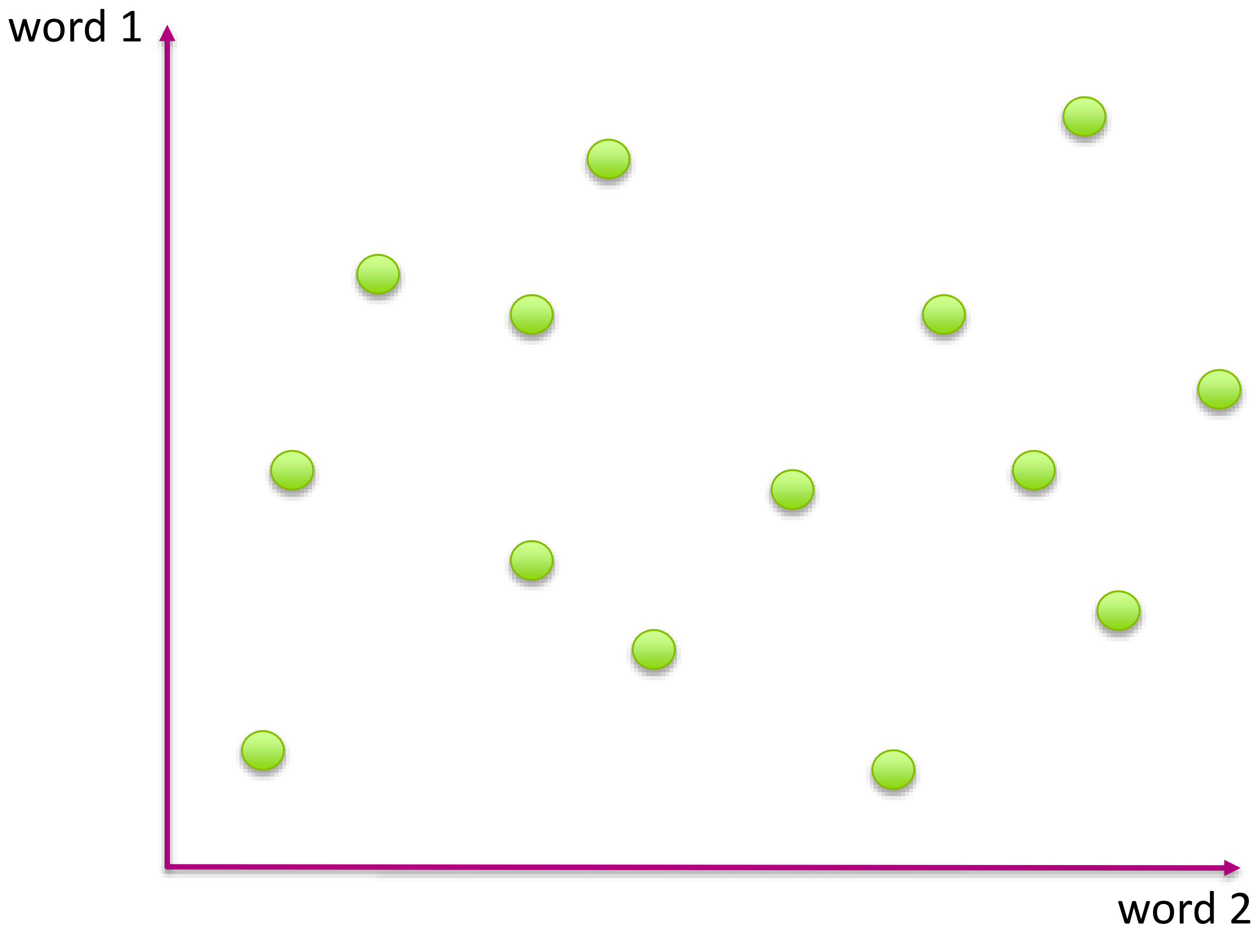
it	1
they	0
I	1
am	0
how	0
puppy	1
and	1
cat	0
aardvark	0
zebra	0
cute	1
extremely	1
…	…



Visualizing bag-of-words



Document point cloud

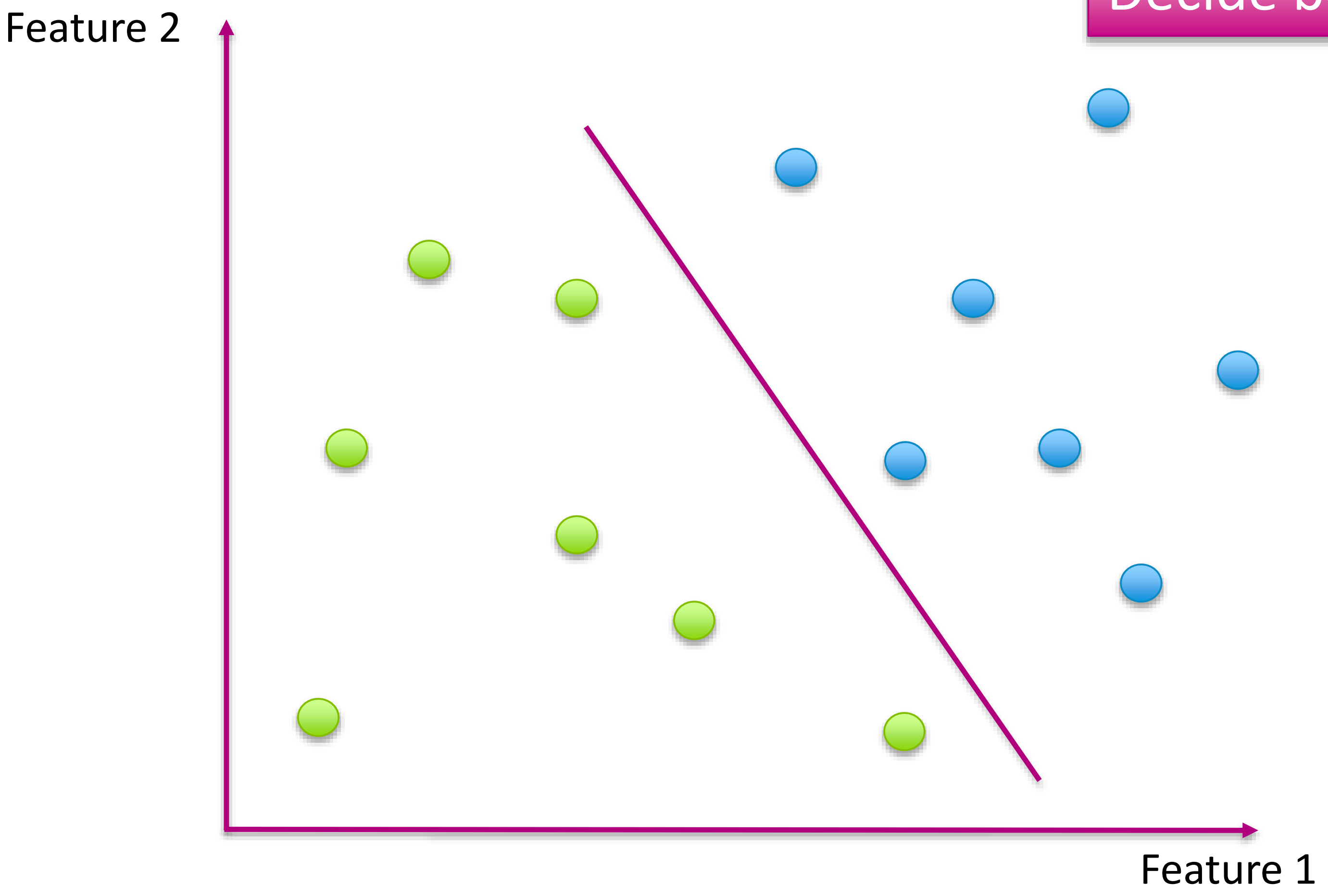


What is a model?

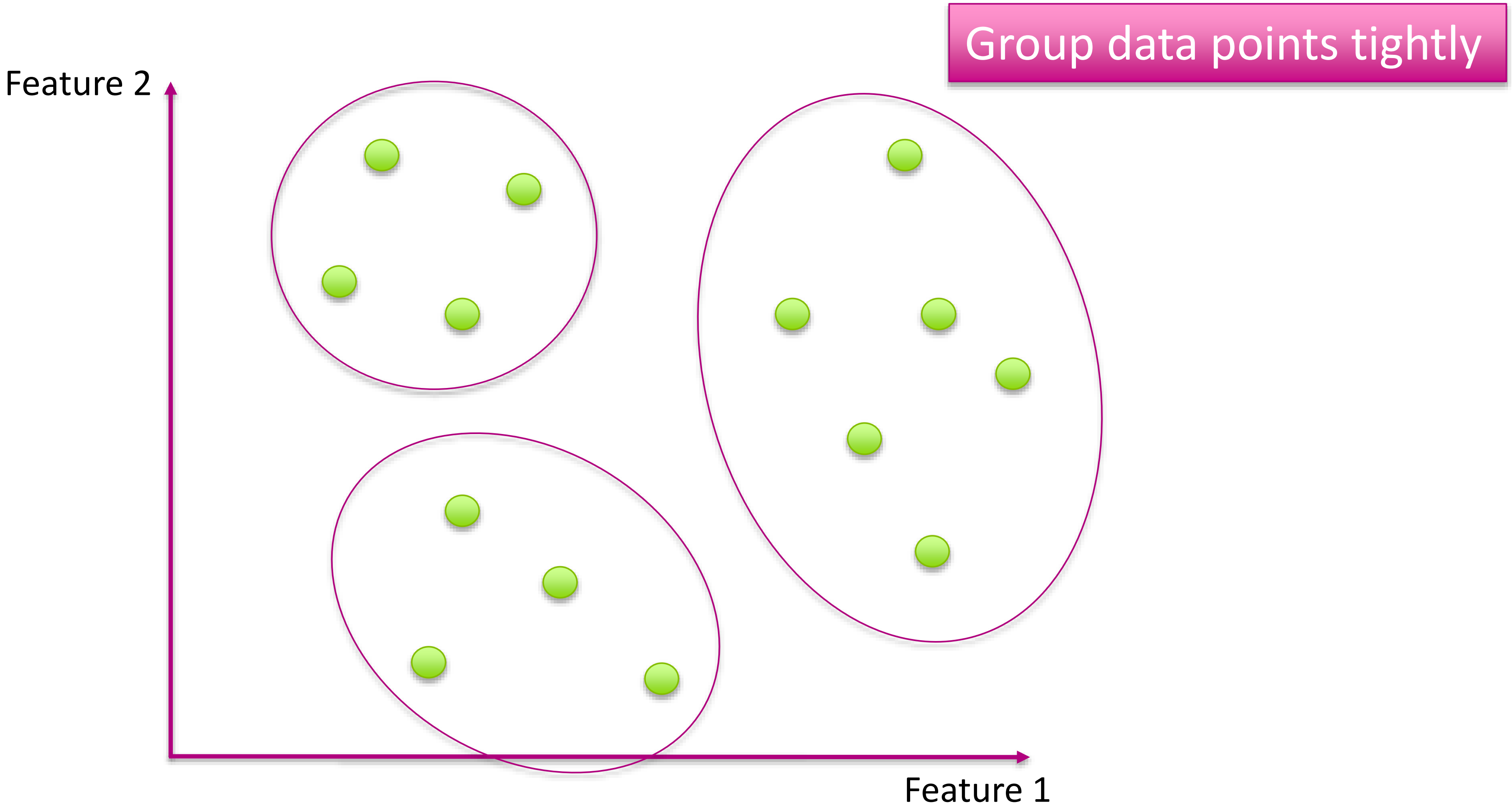
- Model = mathematical “summary” of data
- What’s a summary?
  - A geometric shape

Classification model

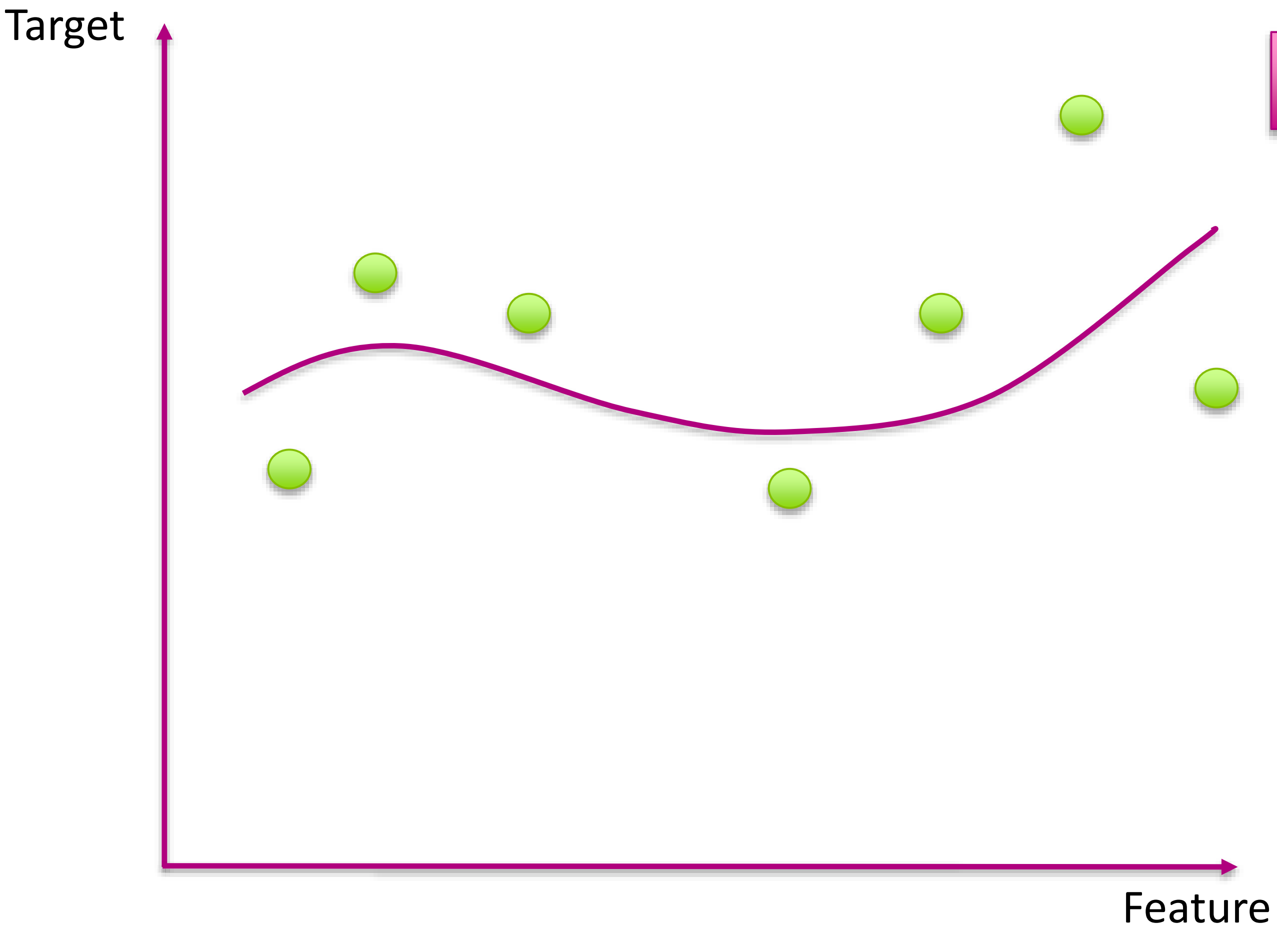
Decide between two classes



Clustering model



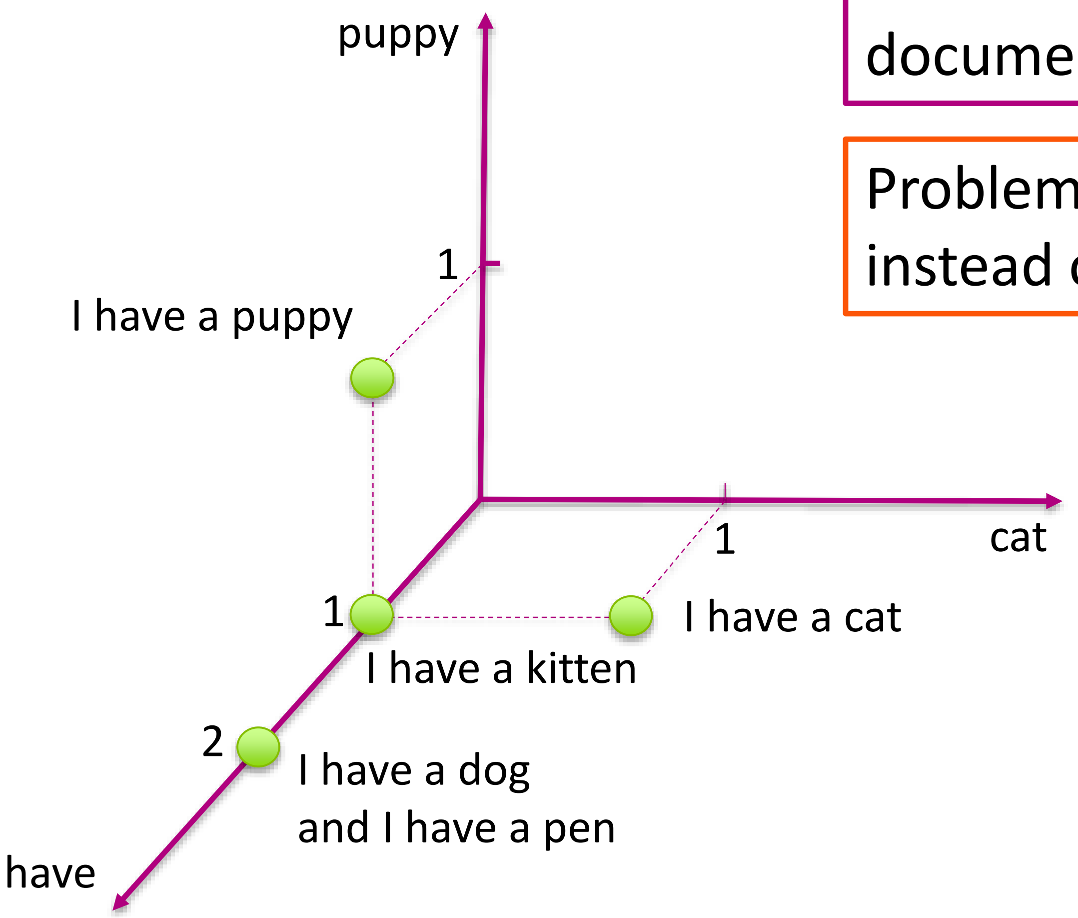
Regression model



Fit the target values

Visualizing Feature Engineering

When does bag-of-words fail?



Task: find a surface that separates documents about dogs vs. cats

Problem: the word “have” adds fluff instead of information

Improving on bag-of-words

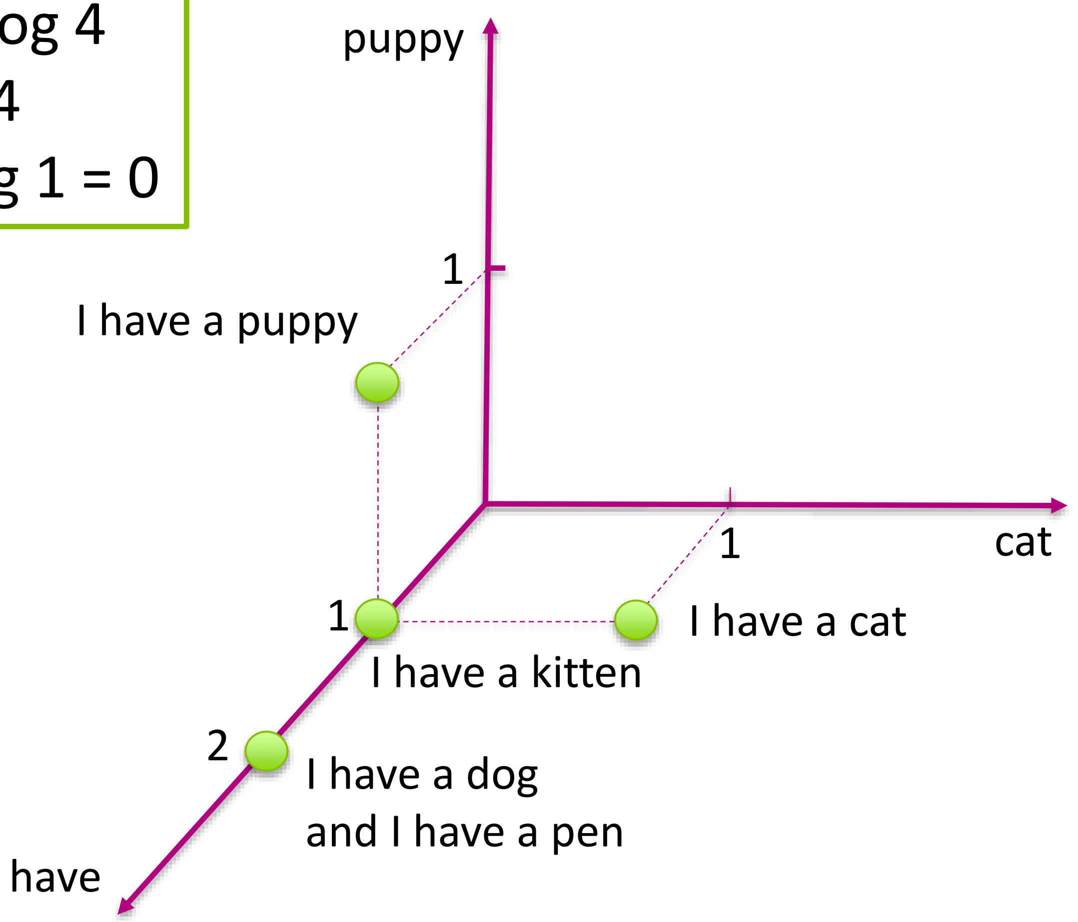
- Idea: “normalize” word counts so that popular words are discounted
- Term frequency (tf) = Number of times a terms appears in a document
- Inverse document frequency of word (idf) =

$$\log\left(\frac{N}{\text{\# docs containing word } w}\right)$$

  - N = total number of documents
- Tf-idf count = tf x idf

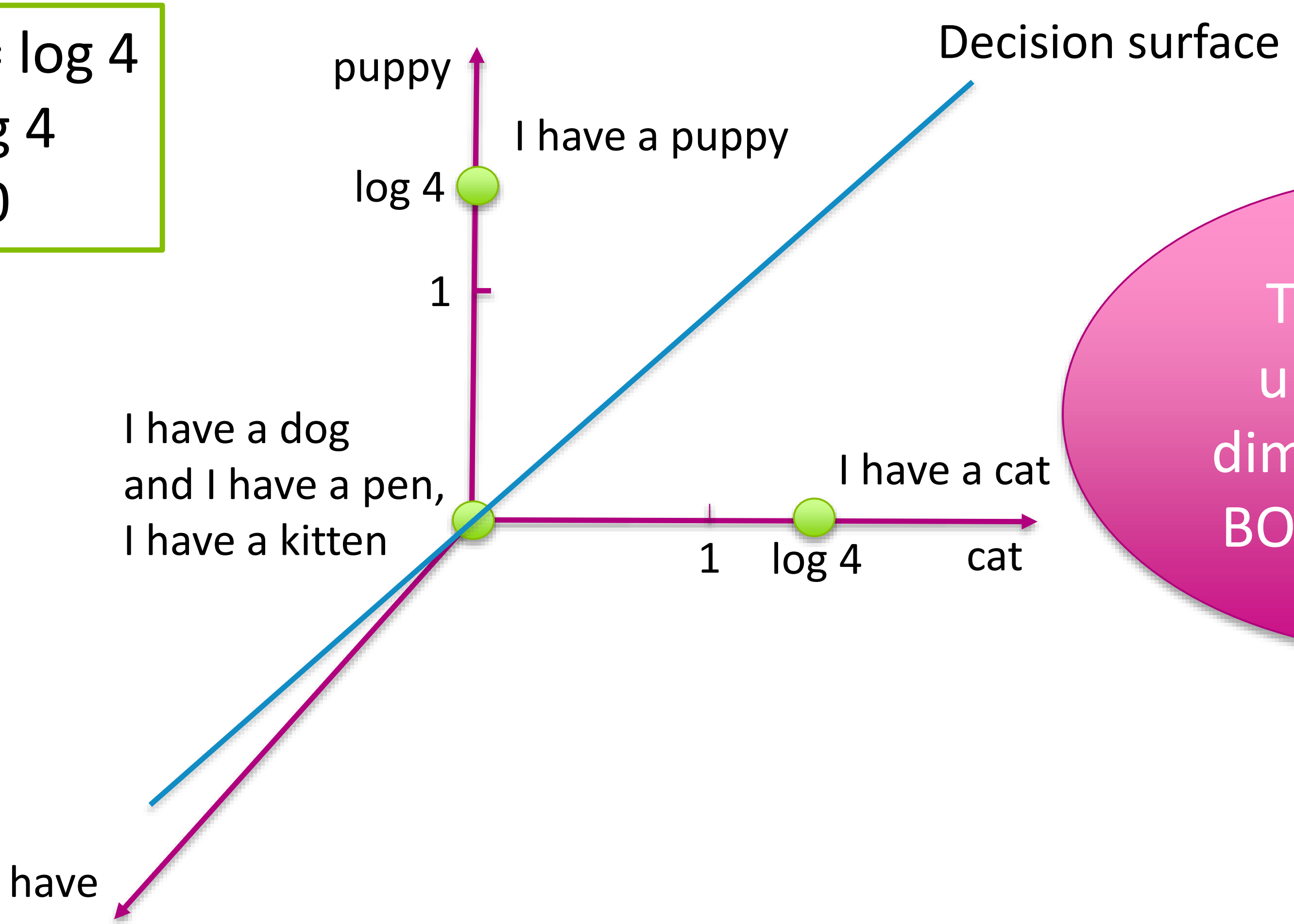
From BOW to tf-idf

$\mathrm{idf}(\mathrm{puppy}) = \log 4$
$\mathrm{idf}(\mathrm{cat}) = \log 4$
$\mathrm{idf}(\mathrm{have}) = \log 1 = 0$



From BOW to tf-idf

tfidf(puppy) = log 4
tfidf(cat) = log 4
tfidf(have) = 0



Tf-idf flattens uninformative dimensions in the BOW point cloud

Entry points of feature engineering

- Start from data and task
  - What’s the best text representation for classification?
- Start from modeling method
  - What kind of features does k-means assume?
  - What does linear regression assume about the data?

That’s not all, folks!

- There’s a lot more to feature engineering:
  - Feature normalization
  - Feature transformations
  - “Regularizing” models
  - Learning the right features

- Dato is hiring! jobs@dato.com

alicez@dato.com @RainyData