Understanding Feature Space in Machine Learning

Alice Zheng

Director of Data Science, Dato

RainyData Datoinc

#datapopupseattle

UNSTRUCTURED

Data Science POP-UP in Seattle

Produced by **Domino Data Lab**

Domino's enterprise data science platform is used by leading analytical organizations to increase productivity, enable collaboration, and publish models into production faster.

www.dominodatalab.com

Understanding Feature Space in Machine Learning

Alice Zheng, Dato

October, 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

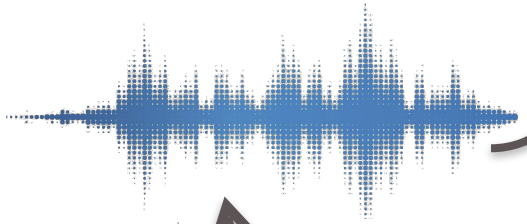
Why machine learning?

Model data.
Make predictions.
Build intelligent applications.

The machine learning pipeline

Raw data

I fell in love the instant I laid my eyes on that puppy. His big eyes and playful tail, his soft furry paws, …



Features

Models

Deploy in production

Predictions

Three things to know about ML

- Feature = numeric representation of raw data
- Model = mathematical “summary” of features
- Making something that works = choose the right model and features, given data and task

Feature = Numeric representation of raw data

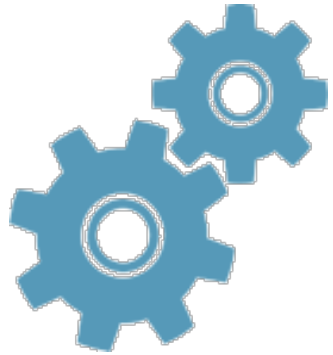
Representing natural text

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 }

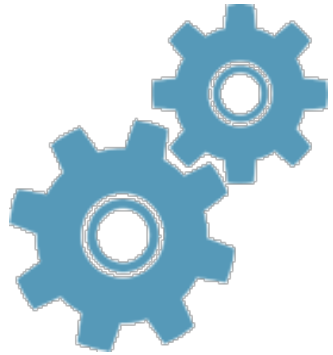


Representing natural text

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

It is a puppy and it is extremely cute.

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

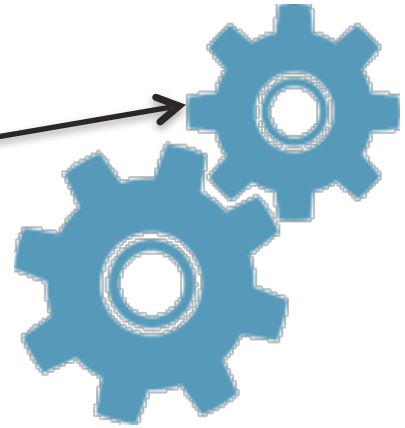


Sparse vector representation

Representing images

Raw Image → Bag of Visual Words

Classify:
person or animal?



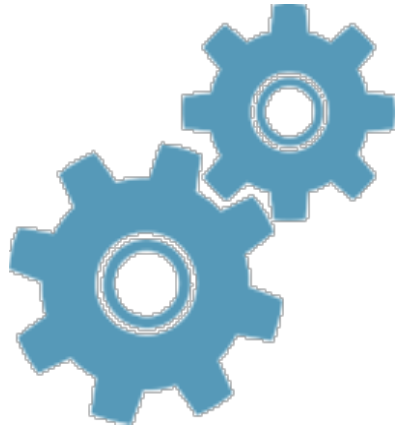
Raw image:
millions of RGB triplets,
one for each pixel

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
- Feature engineering = creating features of the appropriate granularity for the task

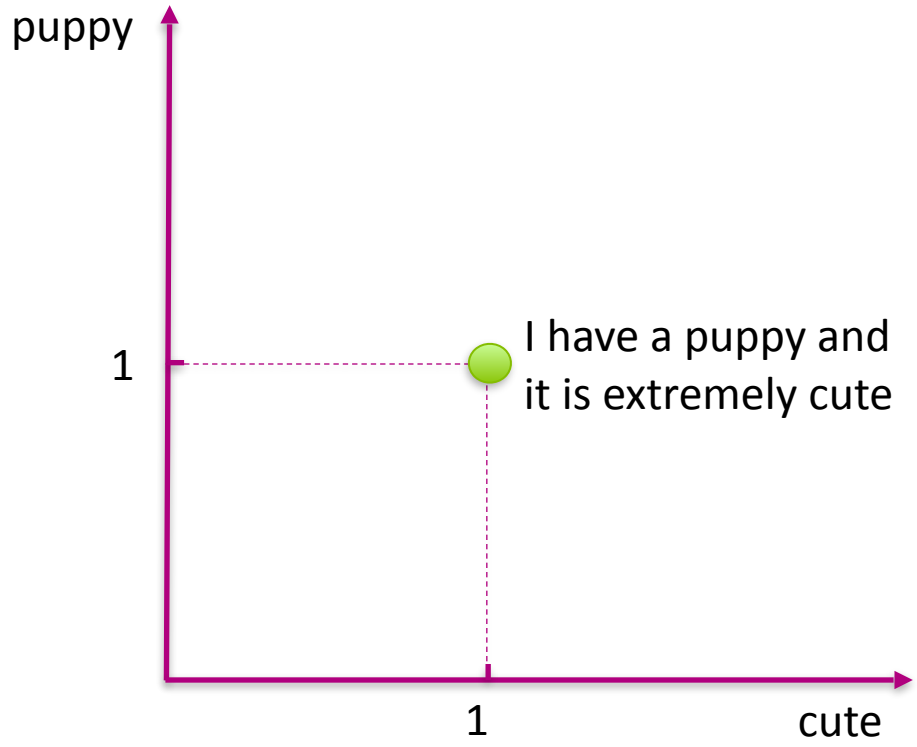
Visualizing Feature Space

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.

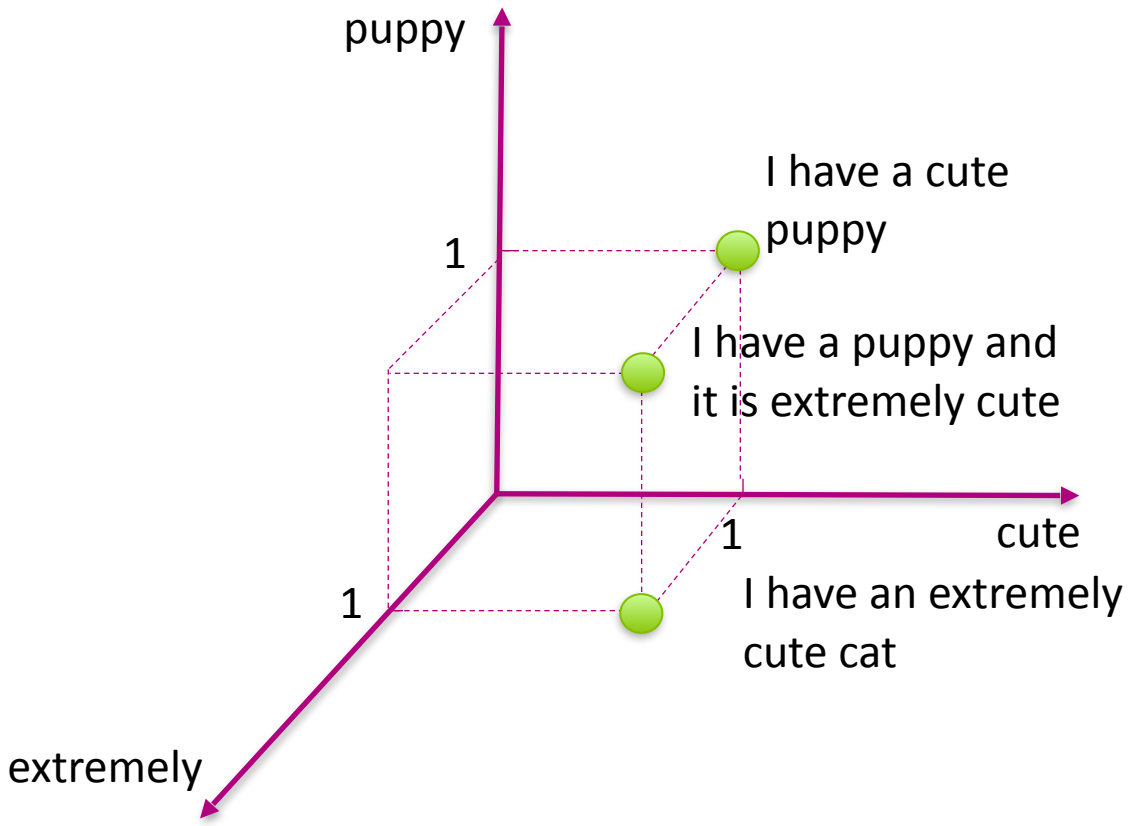
-- Masha Gessen, “Perfect Rigor”

Visualizing bag-of-words

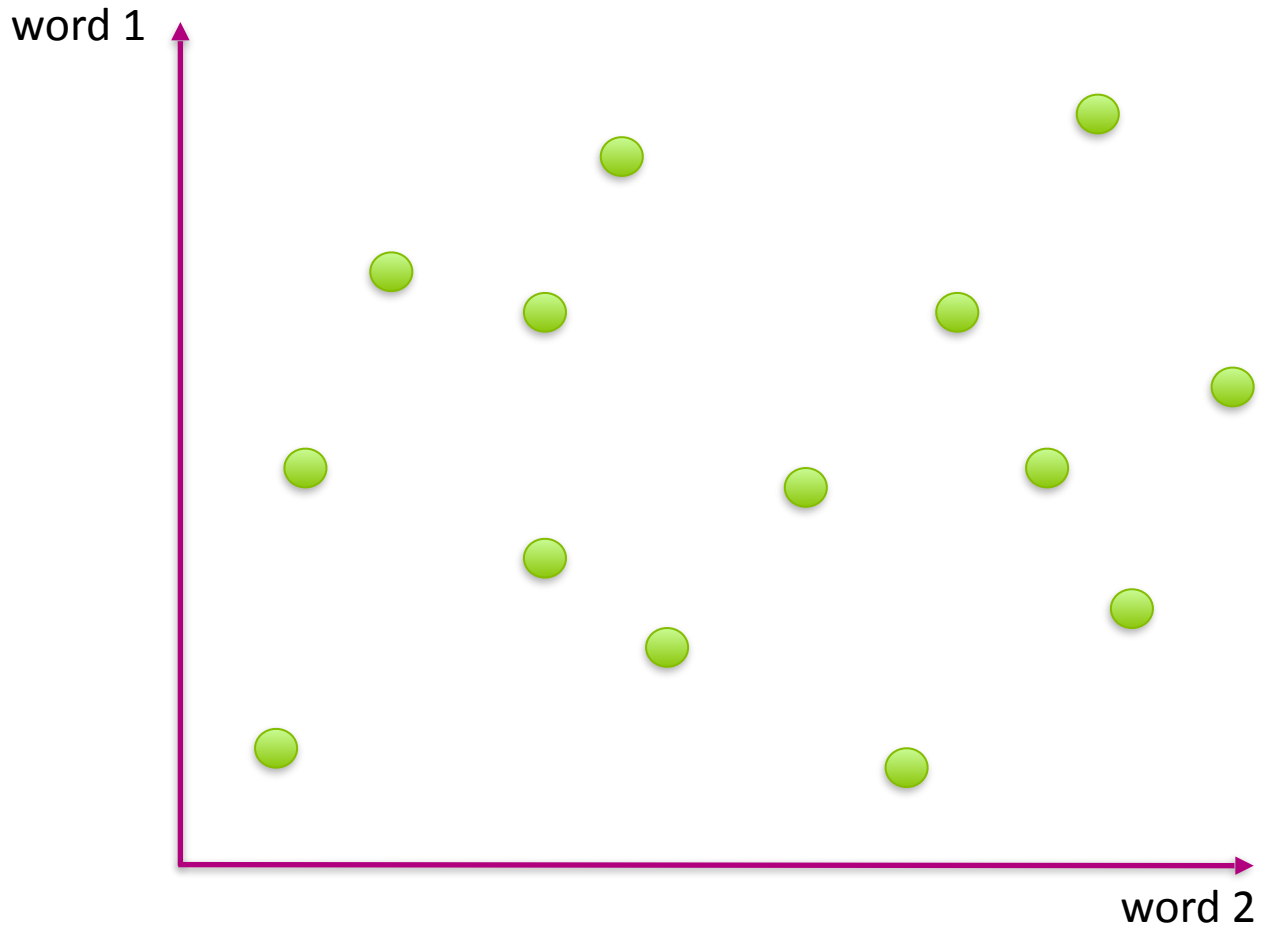
I have a puppy and it is extremely cute



Visualizing bag-of-words



Document point cloud



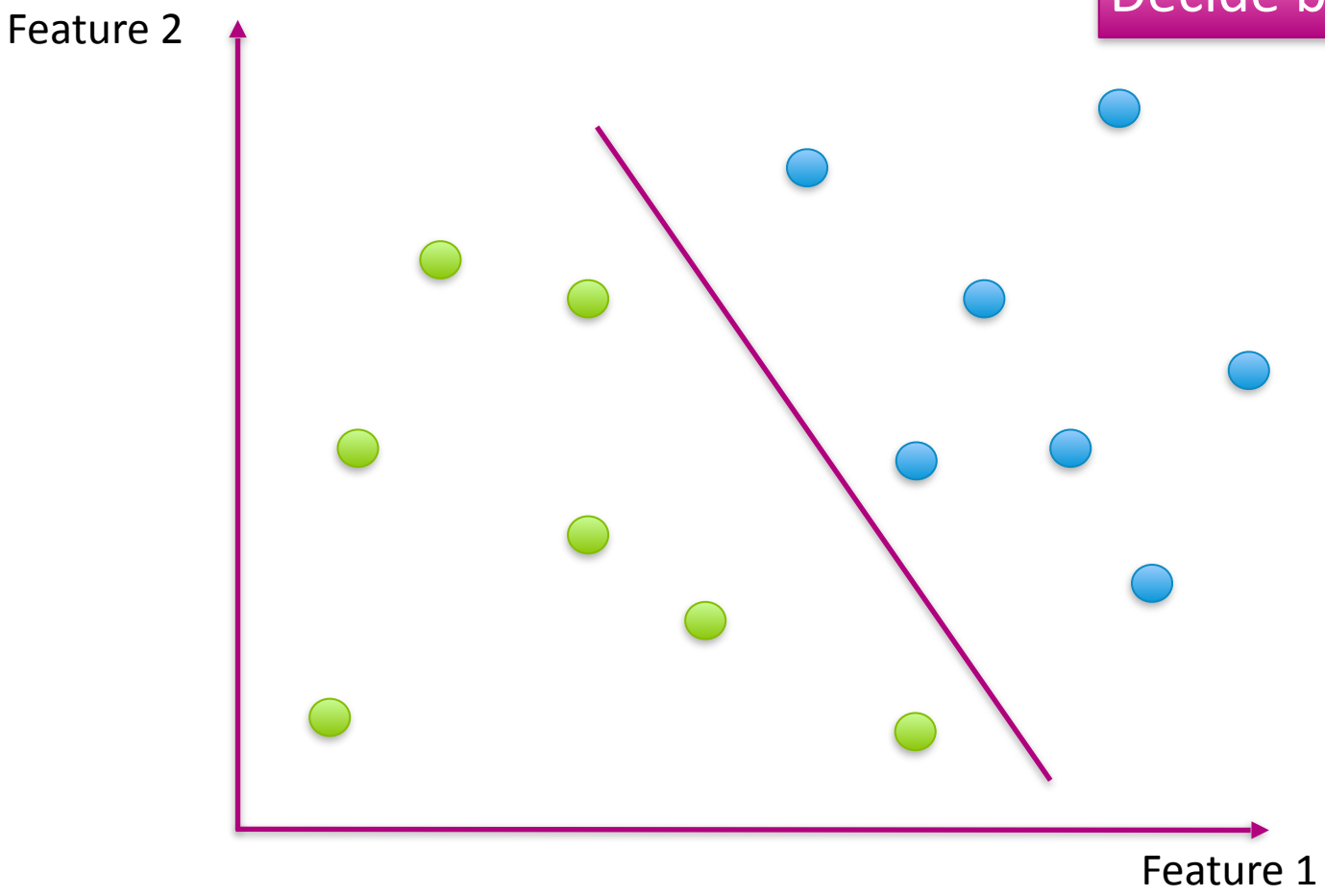
Model = Mathematical “summary” of features

What is a summary?

- Data → point cloud in feature space
- Model = a geometric shape that best “fits” the point cloud

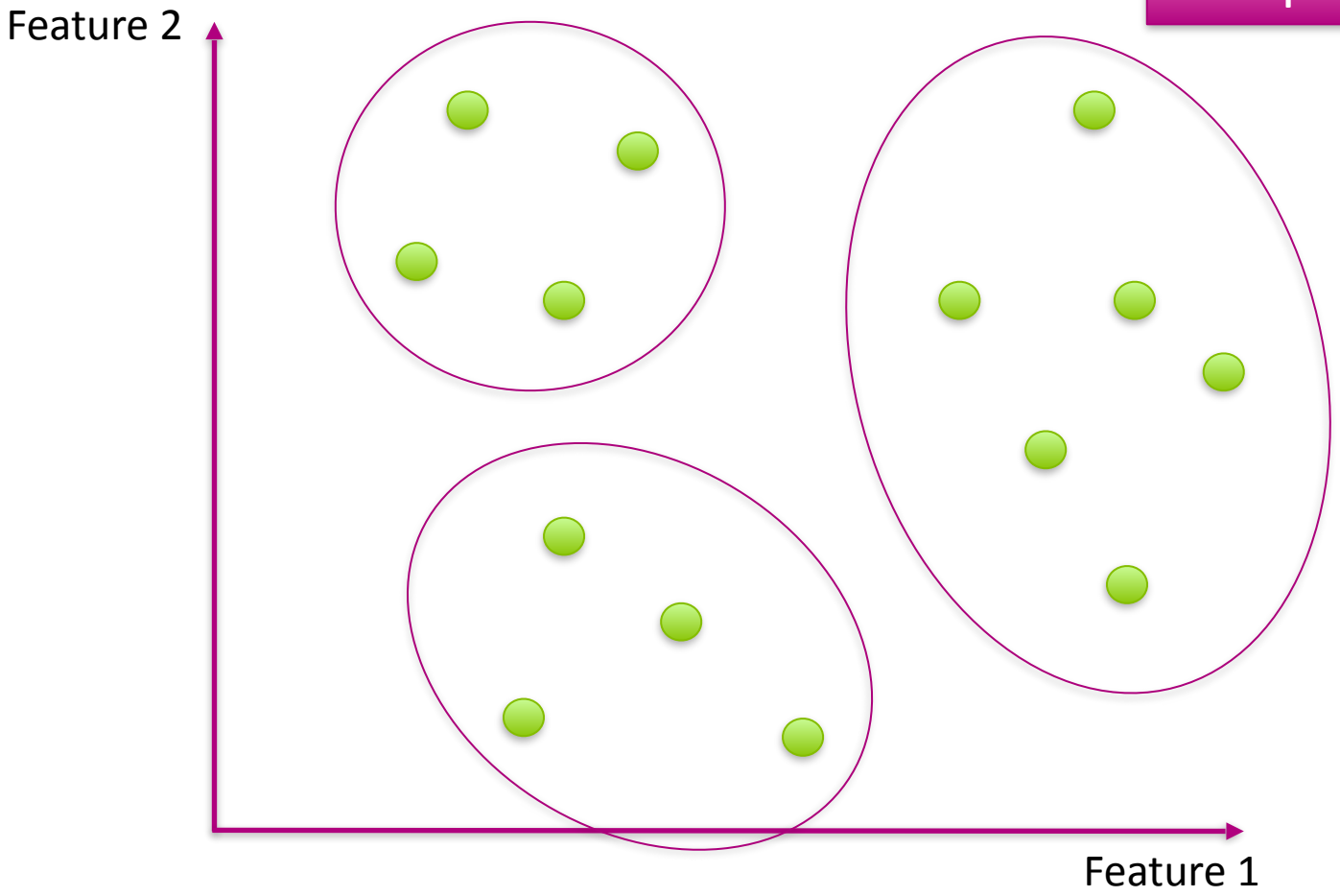
Classification model

Decide between two classes

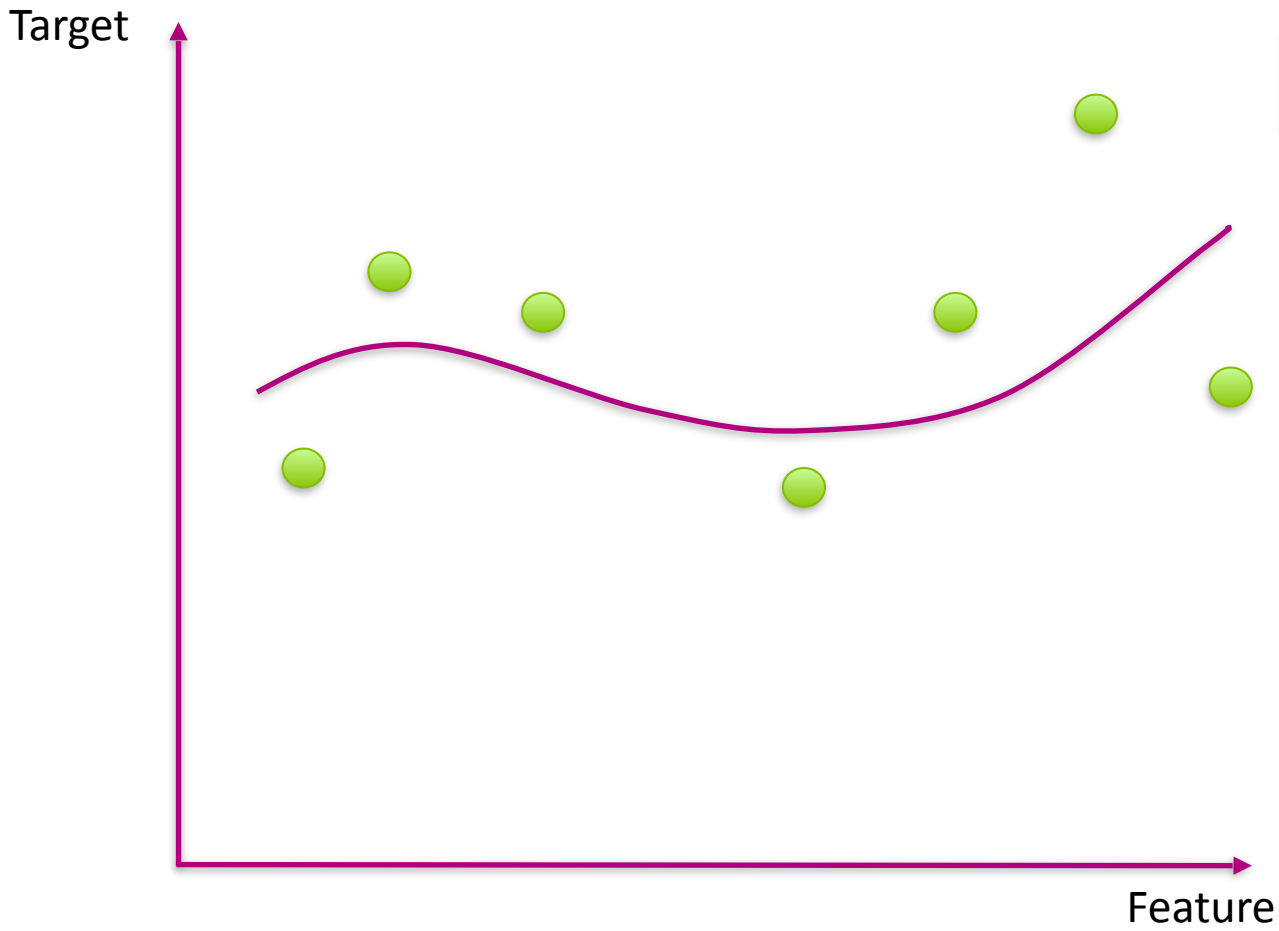


Clustering model

Group data points tightly



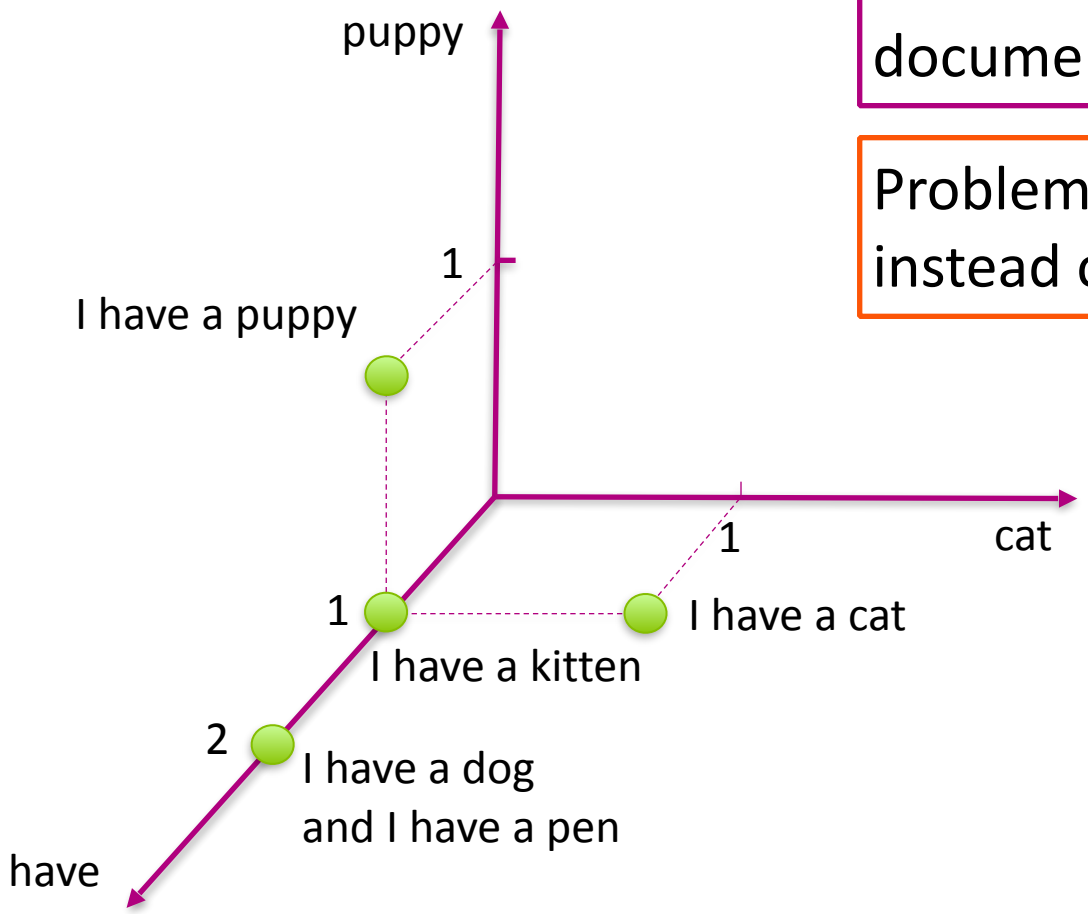
Regression model



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

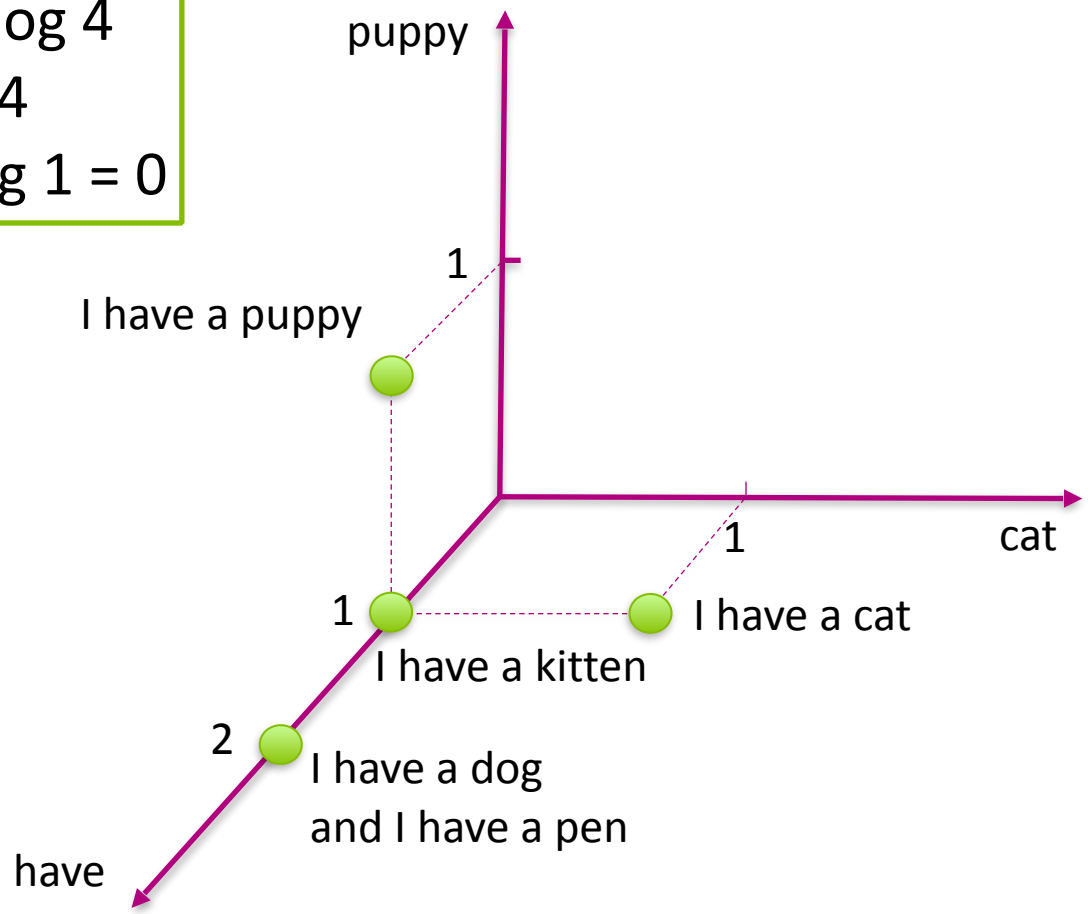
- Idea: “normalize” word counts so that popular words are discounted
- Term frequency (tf) = Number of times a terms appears in a document
- Inverse document frequency of word (idf) =

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

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

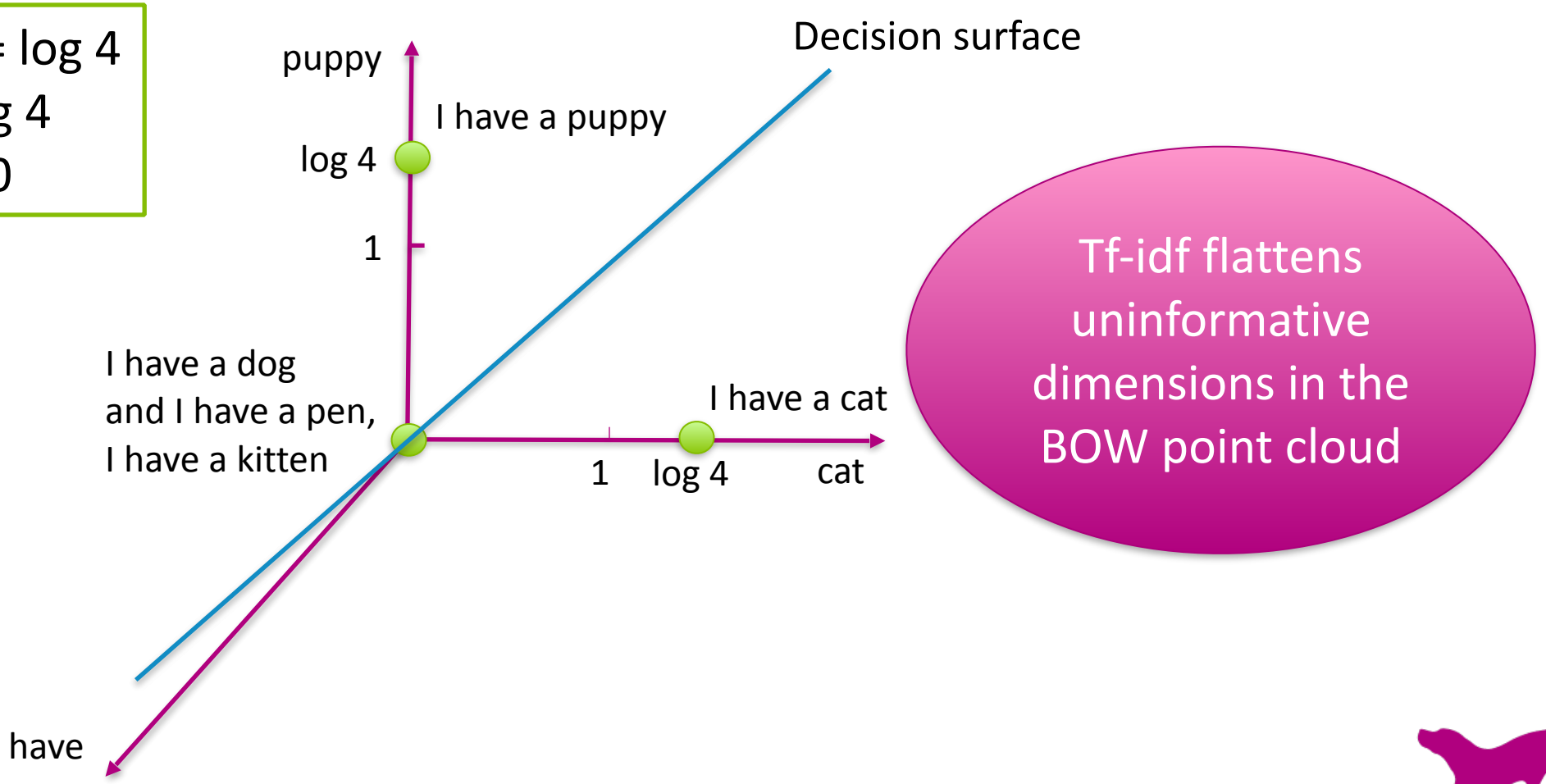
From BOW to tf-idf

idf(puppy) = log 4
idf(cat) = log 4
idf(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

That's not all, folks!

- Geometry is the key to understanding feature space and machine learning
- Many other fun topics:
  - Feature normalization
  - Feature transformations
  - Model regularization

- Dato is hiring! jobs@dato.com

@RainyData, @DatoInc

@datapopup
#datapopupseattle

Thank You To Our Sponsors

galvanize

H2O.ai

O'REILLY

DOMINO

datasciencedojo
unleash the data scientist in you

Dato

Greythorn
Specialist Technology Recruitment

ebay

CONCUR

KOVERSE

#datapopupseattle