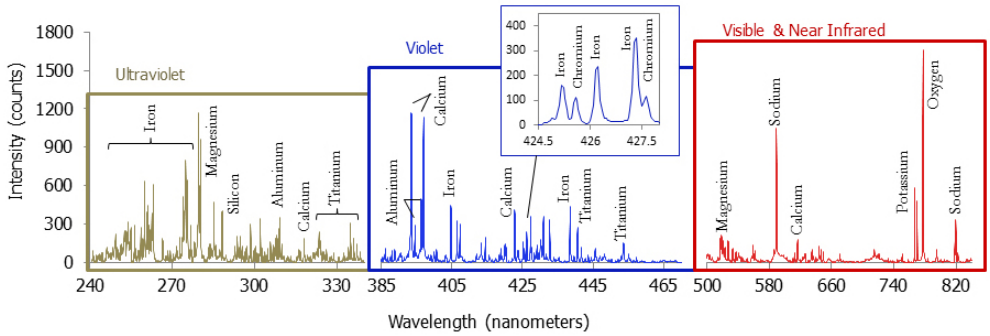
Proximity in data spaces

A: Finding similar items from the past

Example: an instrument on the Mars Rover

Mars Curiosity Rover: the ChemCam instrument

- Laser-induced breakdown spectroscopy (LIBS)
- Gives detailed information about chemical composition of rock



Given an observation: Have we seen something like this before?

Novelty detection

- Past observations: $x_1, x_2, \ldots, x_n$ from some space $\mathcal{X}$
- Now you see x
- Is it something familiar? Or something new that warrants attention?

Nearest neighbor approach:

- Fix a distance function d on $\mathcal{X}$
- Find $\min_i d(x_i, x)$
- If this distance is large: x is something new

ChemCam example: What is $\mathcal{X}$, and what is the distance function?

A ChemCam observation

# wave	shot1	shot2	shot3	shot4	shot5	shot6
240.811	2.97E+11	2.61E+11	3.45E+11	2.99E+11	2.93E+11	3.07E+11
240.86501	1.50E+11	1.32E+11	1.22E+11	1.17E+11	6.16E+10	9.10E+10
240.918	1.06E+11	1.31E+11	8.70E+10	7.35E+10	1.04E+11	7.50E+10
240.972	1.09E+11	1.09E+11	1.67E+11	1.92E+11	1.43E+11	1.75E+11
241.02699	3.59E+11	4.78E+11	5.33E+11	4.23E+11	4.35E+11	5.27E+11
241.07899	8.83E+11	9.92E+11	1.13E+12	1.01E+12	1.04E+12	1.08E+12
241.133	1.06E+12	1.18E+12	1.42E+12	1.26E+12	1.28E+12	1.38E+12
241.188	7.63E+11	8.49E+11	1.06E+12	9.59E+11	9.22E+11	1.02E+12
241.24001	2.88E+11	3.21E+11	4.30E+11	4.09E+11	3.71E+11	4.04E+11
241.29401	1.88E+11	1.79E+11	2.78E+11	2.30E+11	1.85E+11	2.15E+11
241.34801	3.14E+11	4.13E+11	4.12E+11	4.25E+11	4.04E+11	3.66E+11
241.401	4.71E+11	5.03E+11	5.99E+11	4.97E+11	5.12E+11	5.65E+11
241.45599	3.62E+11	3.52E+11	3.61E+11	3.50E+11	3.69E+11	4.13E+11
241.508	1.10E+11	1.65E+11	1.89E+11	1.72E+11	1.27E+11	1.92E+11
241.562	5.23E+10	6.94E+10	1.30E+11	3.23E+10	6.19E+10	9.61E+10

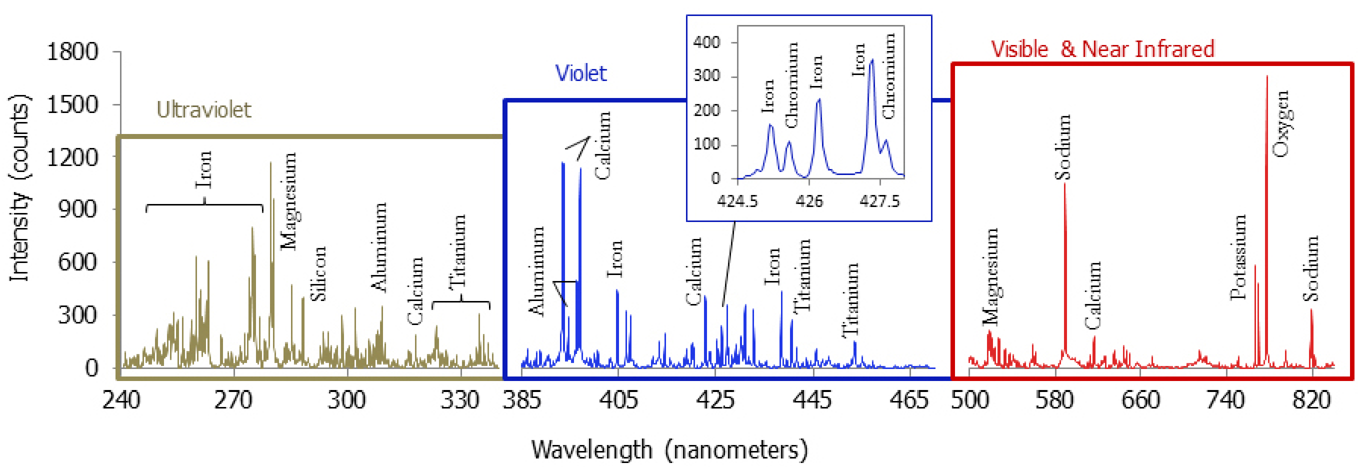
A single observation can be represented by a 6144-dimensional vector.

Picking a distance function

The most familiar option: **Euclidean, or ℓ_2, distance**.

B: Representations and distances

A ChemCam observation



A single observation can be represented by a 6144-dimensional vector.

Problem: Scaling

Consider these two observations:

- One solution: **normalize** each vector to sum to 1:

$$x'_j = \frac{x_j}{\sum_i x_i}.$$

The normalized vectors can be thought of **probability distributions**.

- Modified input space: the space of probability distributions over $m = 6144$ outcomes, also known as the **probability simplex**:

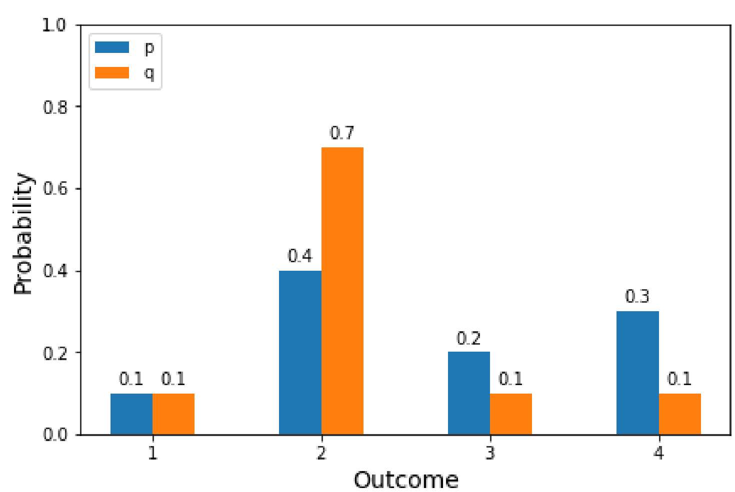
$$\Delta_m = \left\{(p_1, \ldots, p_m) : p_i \geq 0, \sum_i p_i = 1\right\}.$$

L_1 distance

The ℓ_1 distance between vectors $x, z \in \mathbb{R}^m$ is

$$\|x - z\|_1 = \sum_{i=1}^{m} |x_i - z_i|.$$

Example: distributions p, q



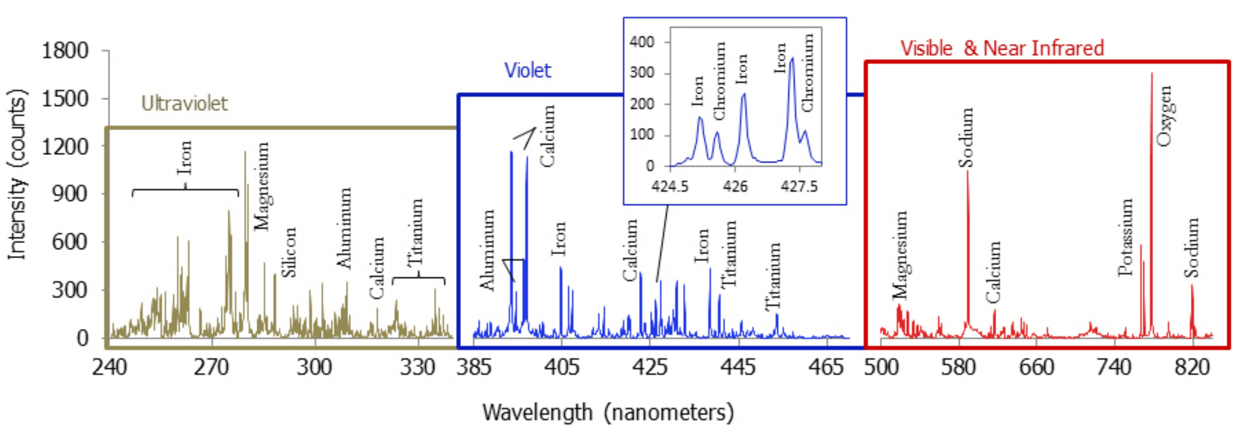
A popular distance function between distributions

Let p, q be probability distributions over a set of m outcomes.

The **Kullback-Leibler divergence** or **relative entropy** between p, q is:

$$KL(p, q) = \sum_{i=1}^{m} p_i \log \frac{p_i}{q_i}.$$

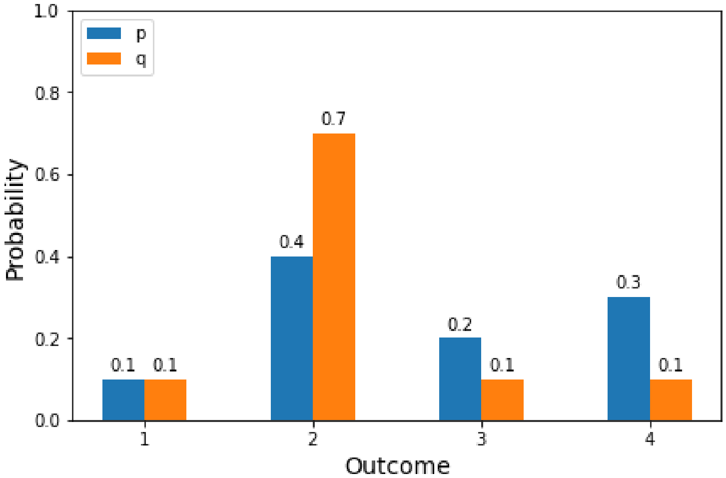
Problem: Noise



What if the wavelength measurements are noisy and bleed into neighboring values?

Some ways of handling noise

- Alternative distance function
  E.g. Earthmover or Wasserstein distance.



- Alternative representation, e.g. using binning or blurring.

C: Picking a good representation

From Herbert Simon, *Sciences of the Artificial*:

> Solving a problem simply means representing it so as to make the solution transparent.

Representations for text

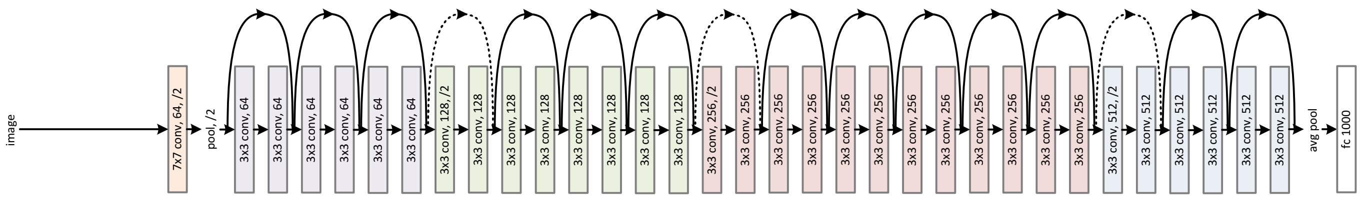
- Bag of words
- Latent semantic indexing
- Brown clustering
- Topic models
- Word2Vec and Glove
- BERT and beyond

Example: Word2Vec

Representations for images

- Principal component analysis, for images or image-patches
- Wavelets
- Sparse coding
- SIFT
- HOG
- Deep belief nets
- Self- or fully-supervised deep representations

Example: Residual Network (ResNet)



Representations for other domains

- Audio: speech, music, animal sounds, etc.
- Biological sequences: DNA, proteins, etc.
- And many others.