Mining Data for Patterns

Fall 2013

Carola Wenk

Analyzing Data

- Modern technology allows us to gather many forms of information at an unprecedented rate. What do we do with all of it?

- We'd like to learn from it. What does this mean?

  - 1. Find important "stuff" in the data.

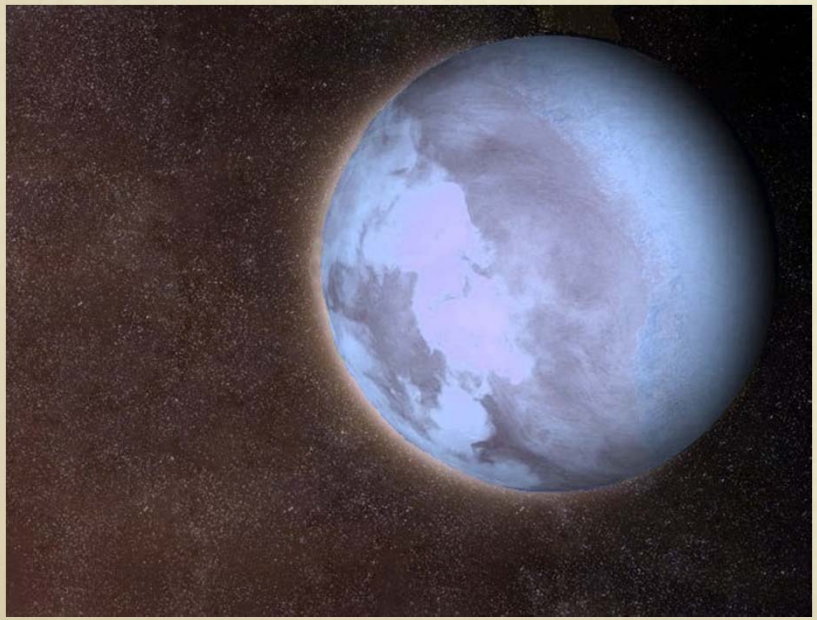
  - 2. Learn the "rules" of the data.

  - 3. Make automatic decisions.

Is There Extraterrestrial Life?

- In 1961, Frank Drake estimated that roughly 10,000 star systems in the Milky Way could support life (i.e., advanced civilization).

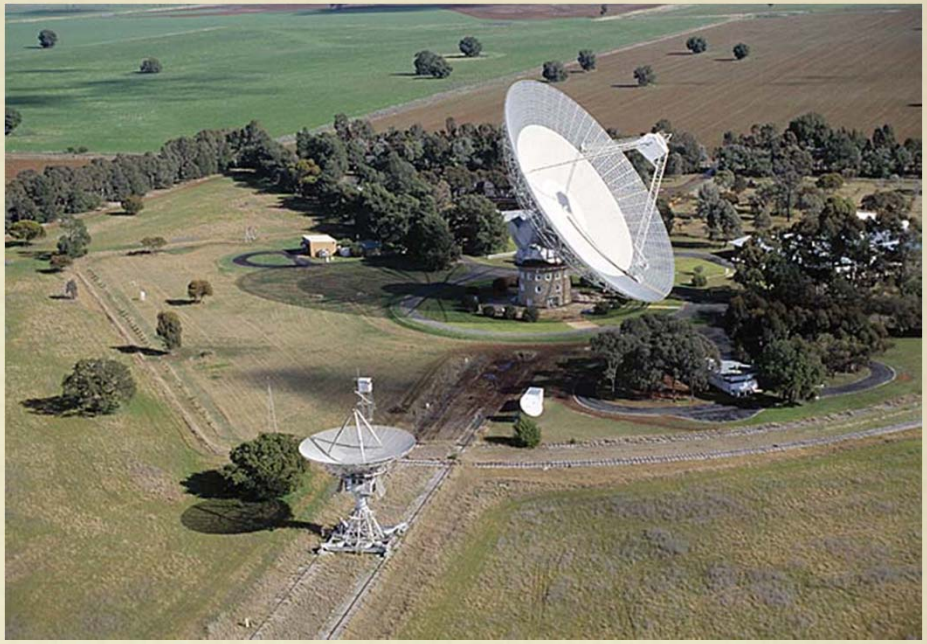
- Using the same criteria, we estimate there are about 6 billion such systems in the observable Universe.



Gliese 581C

Martian Meteorite ALH84001

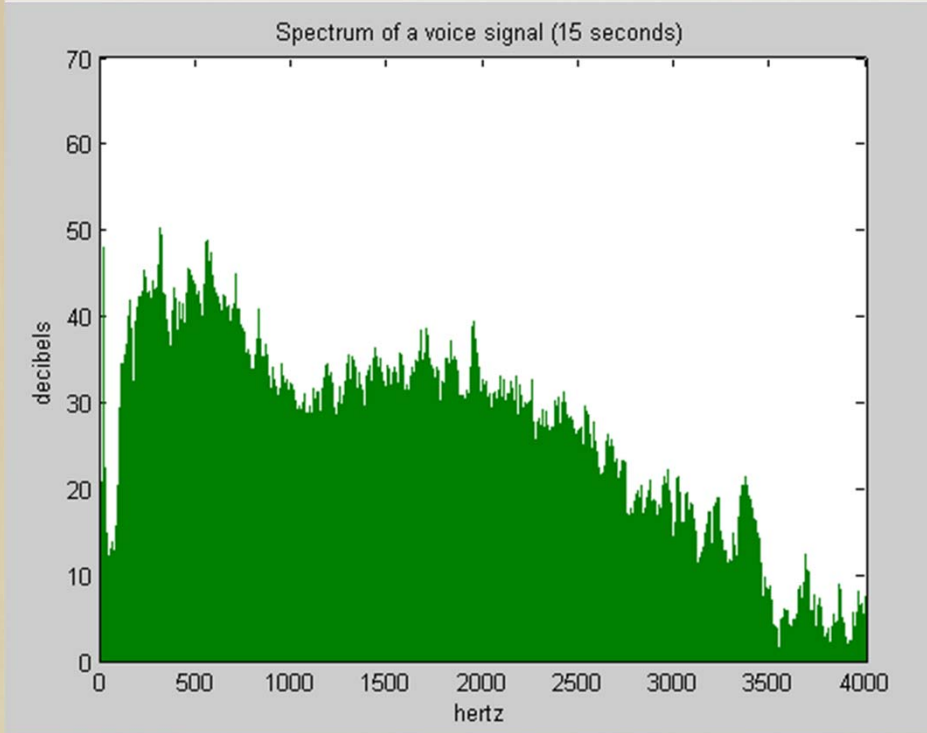
Radio Astronomy



As early as the 1930s, it was observed that radio antennas could pick up non-terrestrial signals (e.g. nearby stars).

Radio “telescopes” can image most astronomical objects (stars, galaxies, planets), and have been used since radio technology matured after World War 2.

Searching for Extraterrestrial Life



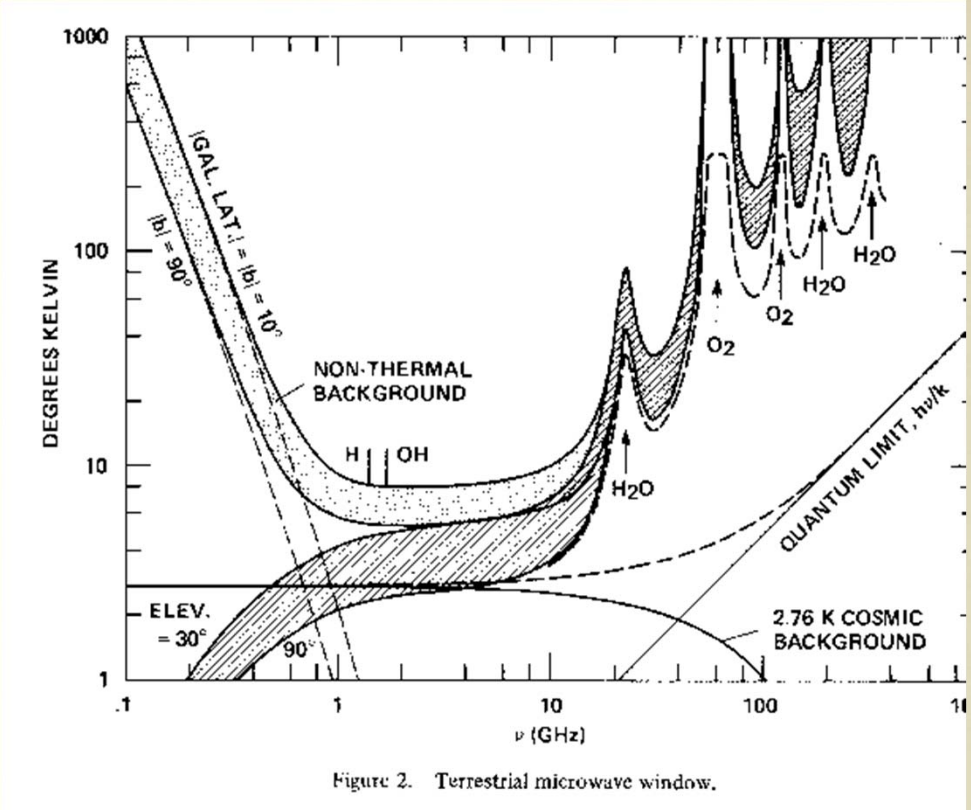


Figure 2. Terrestrial microwave window.

Radio waves are the primary means (that we know of) to communicate over long distances.

We postulate that sufficiently advanced extraterrestrial life would use a particular part of the radio spectrum. Can we observe their communication?

The SETI Project

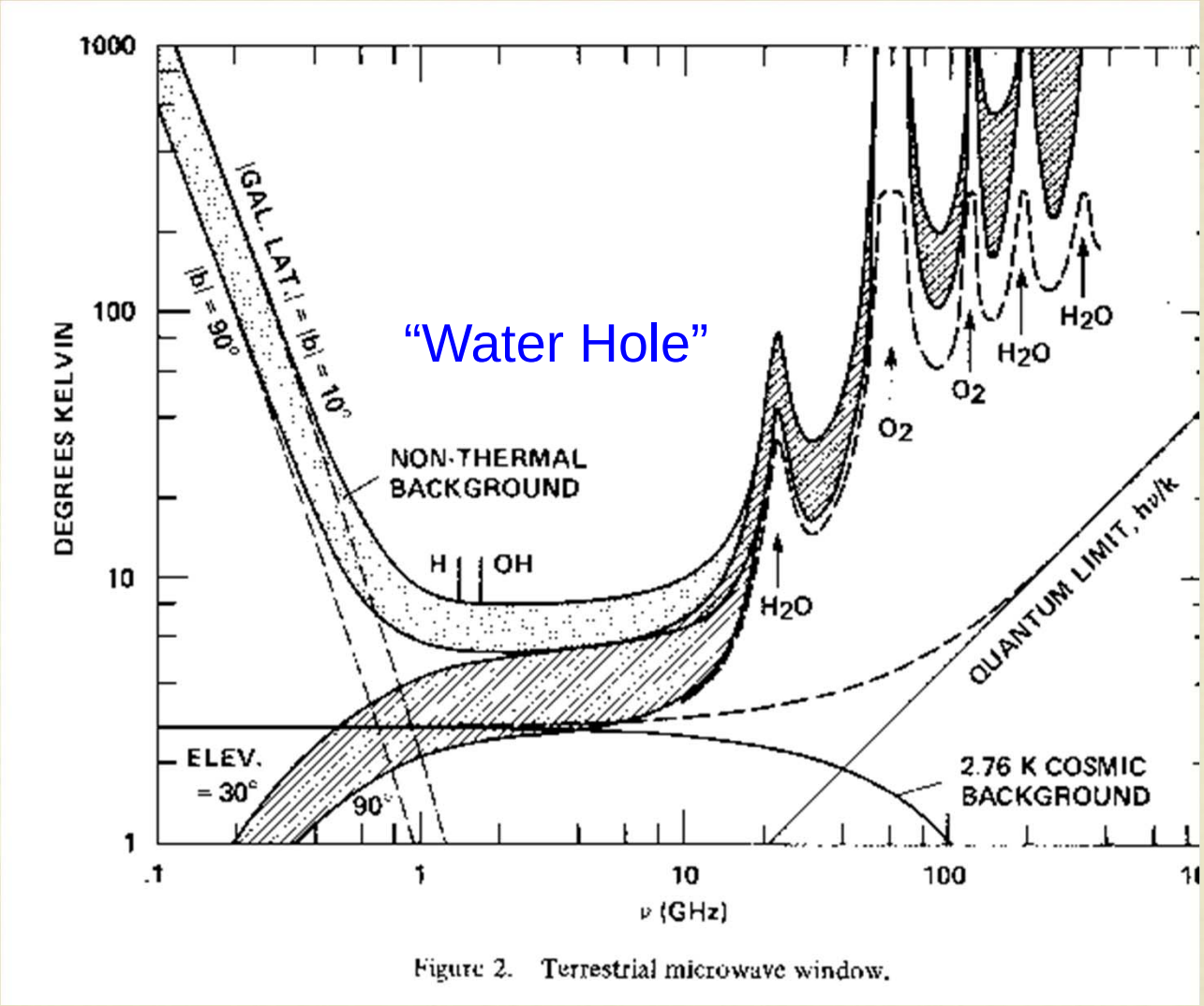


Figure 2. Terrestrial microwave window.

The SETI project began in the 1990s as a way to use idle radio telescopes to probe the galaxy for signs of intelligent life.

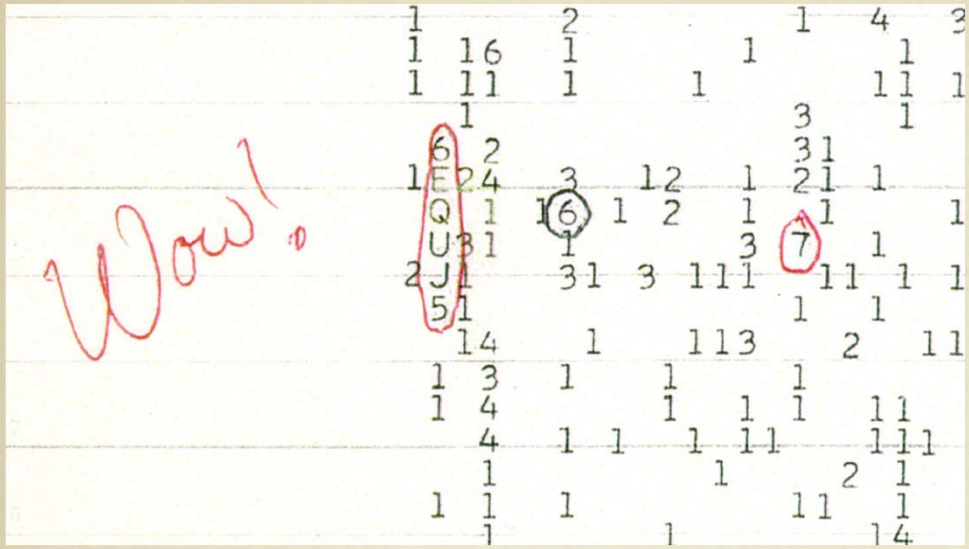
The SETI Project

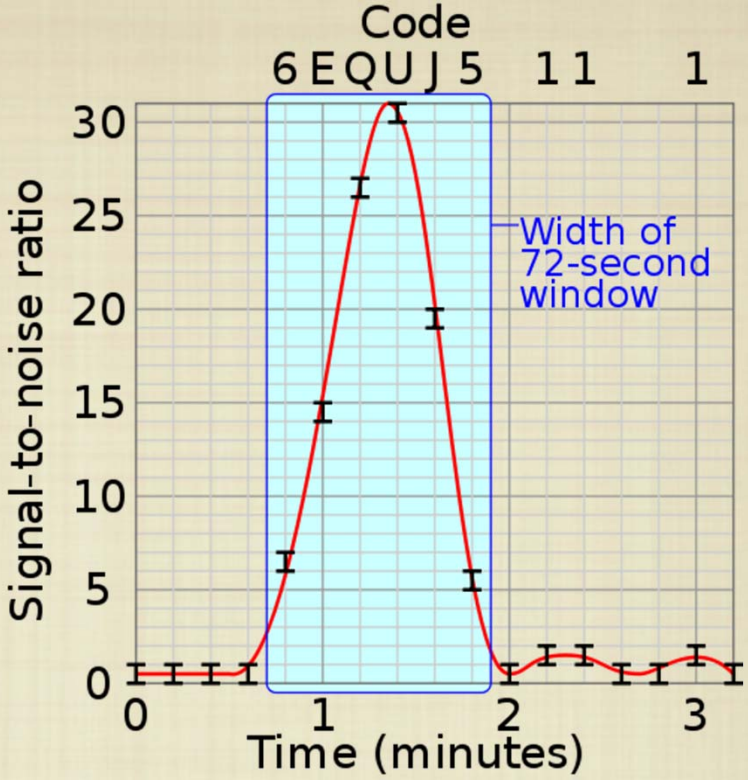
Algorithm:

1. For each time period, record the frequency spectrum for a specific location in the skies (e.g. near a sun-like star).

2. Identify all signals in the frequency spectrum that appear at regular time intervals at sufficiently large intensity, and correspond to possible the frequencies used (by us) for communication.

The SETI Project

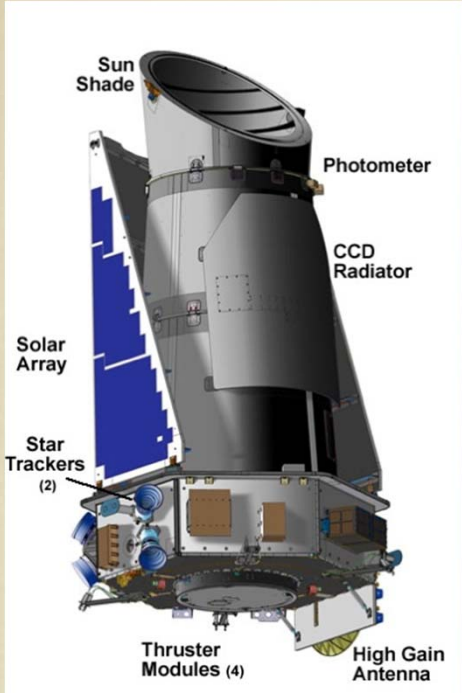


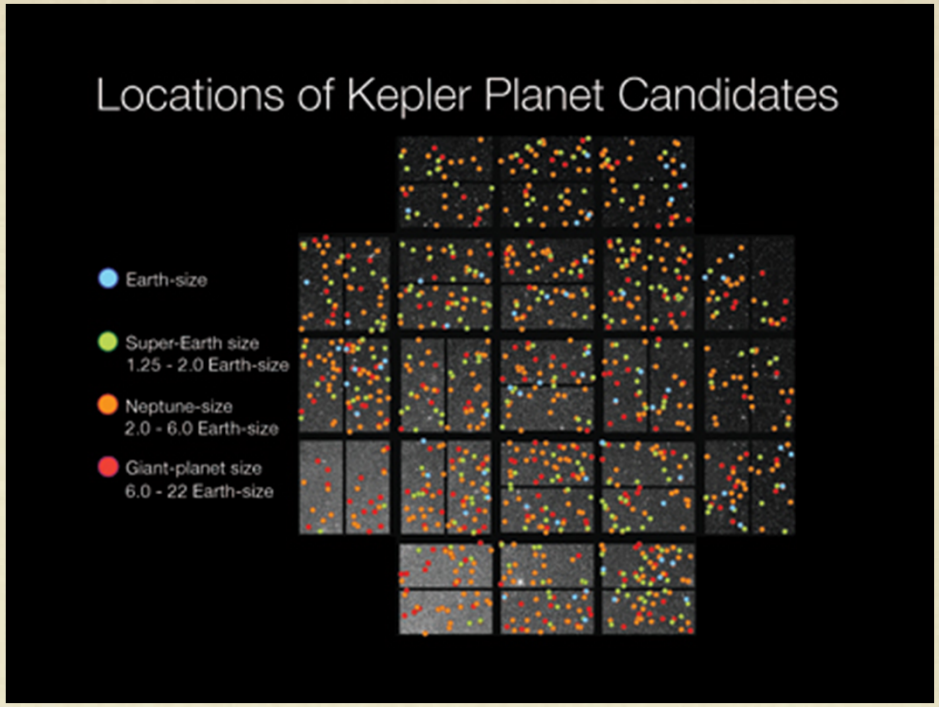


In 1977, researchers observed the “WOW” signal, which was significantly louder than background noise.

The signal appeared at 1420 GHz, the same resonant frequency as Hydrogen, a possible indication of a “message”. It was also seen in 2003, leading to more speculation.

The SETI Project

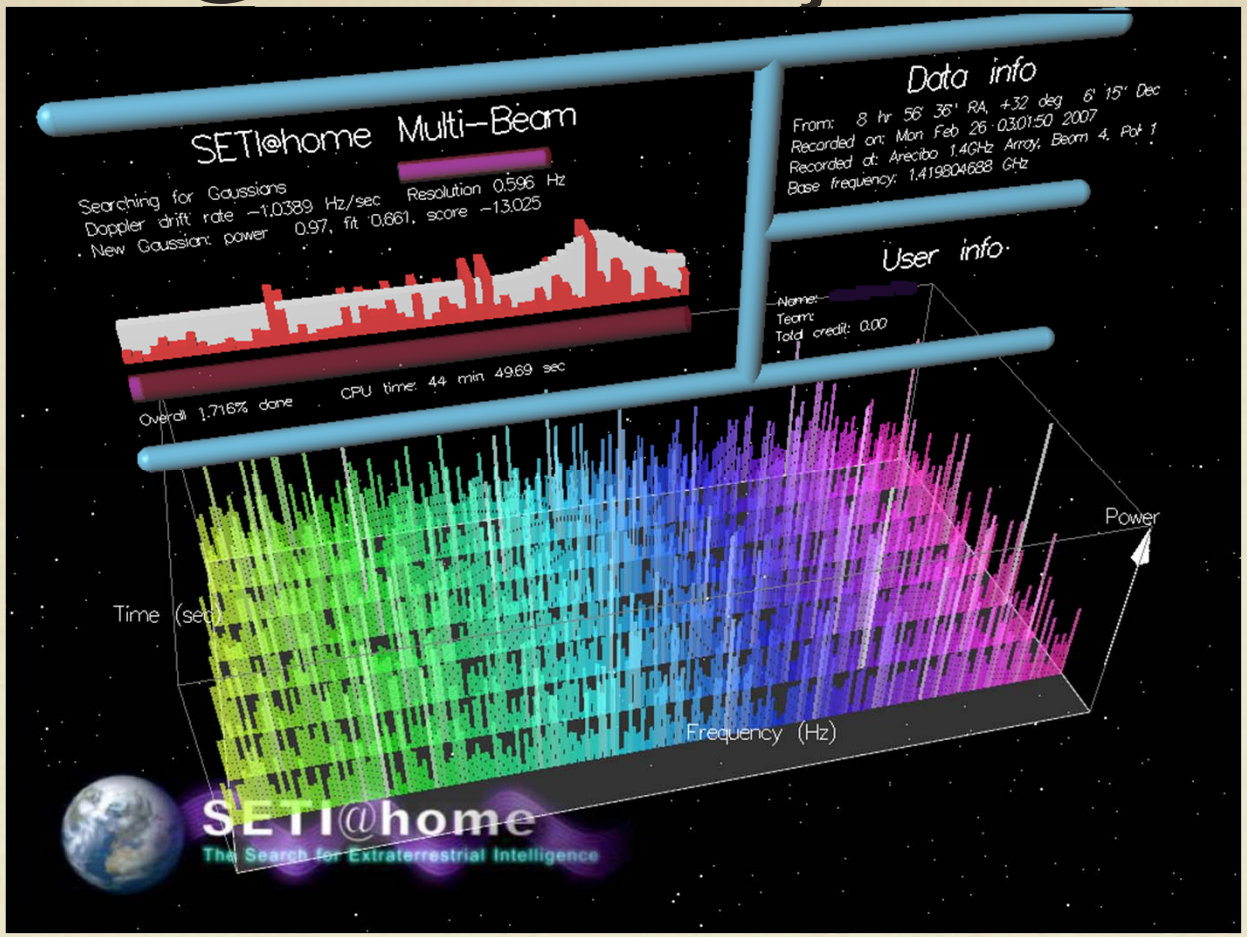




More recently, the Kepler telescope has been scanning for planets that are in the “Goldilocks” zone near sun-like stars.

So far, 1235 planets have been found, and 86 have been chosen for further examination by the SETI project.

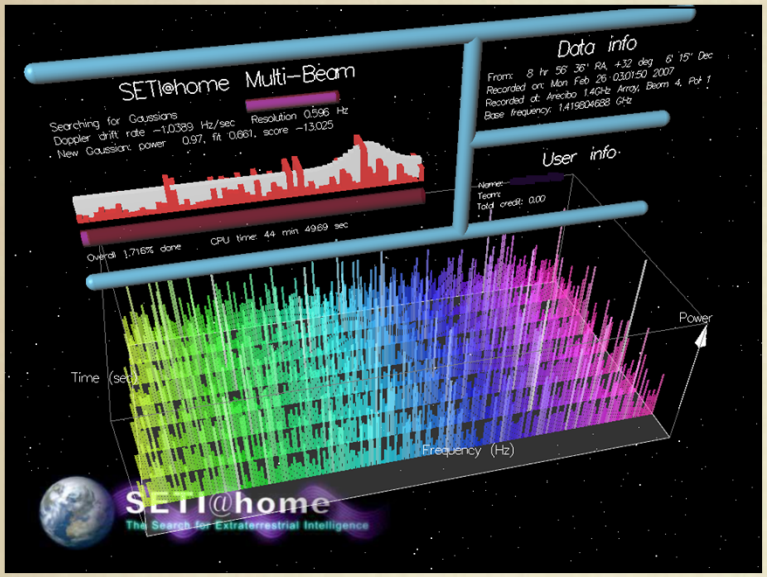
The SETI@Home Project



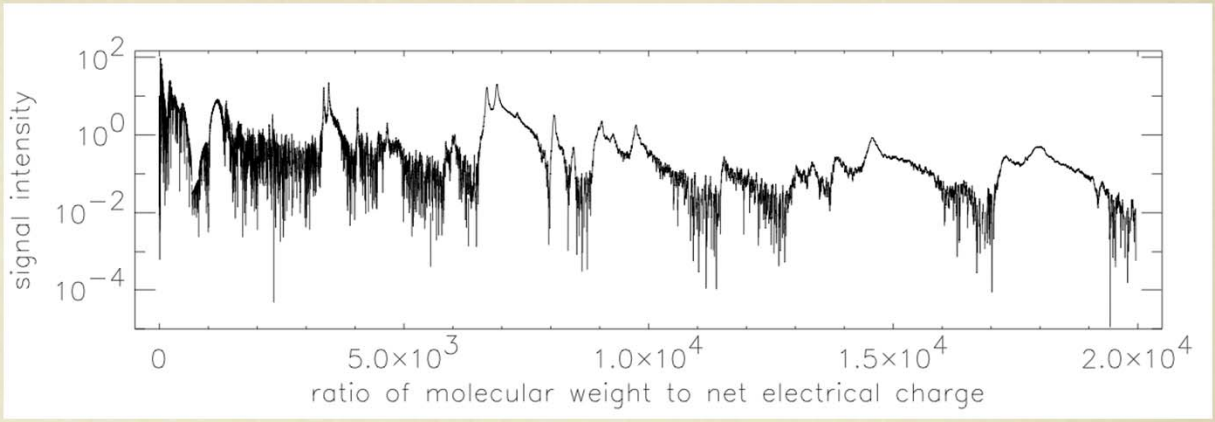
The SETI@Home project is an attempt to utilize idle time of computers on the Internet to scan radio telescope observations.

The advantage of distributed computing is two-fold: frequency spectra can be analyzed in greater quantities and in higher detail.

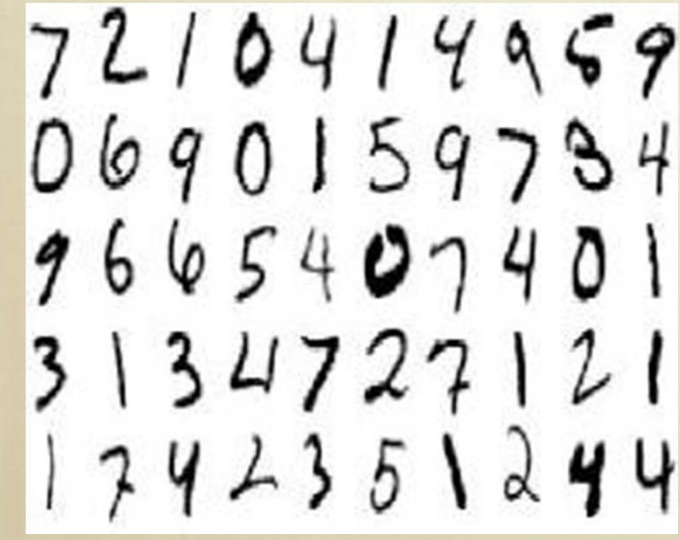
Analyzing and Classifying Data



Radio Astronomy Dataset



Mass Spectrogram of Blood Serum

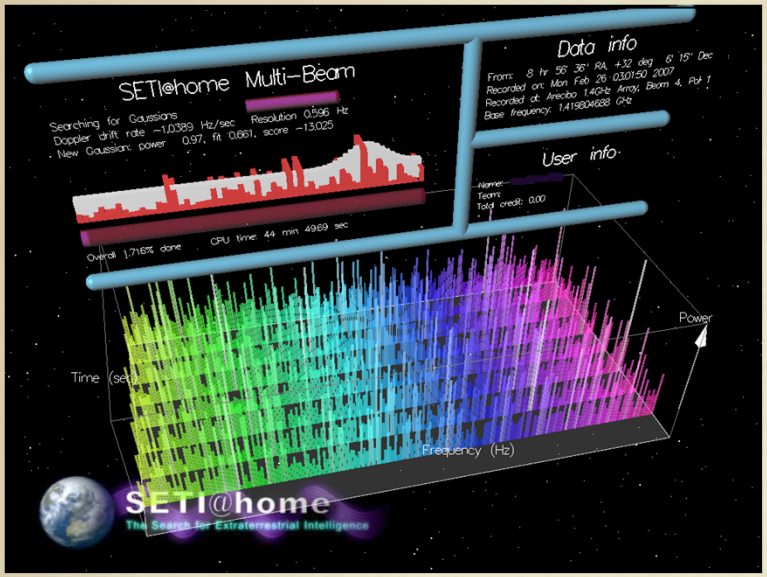


Images of Handwritten Digits

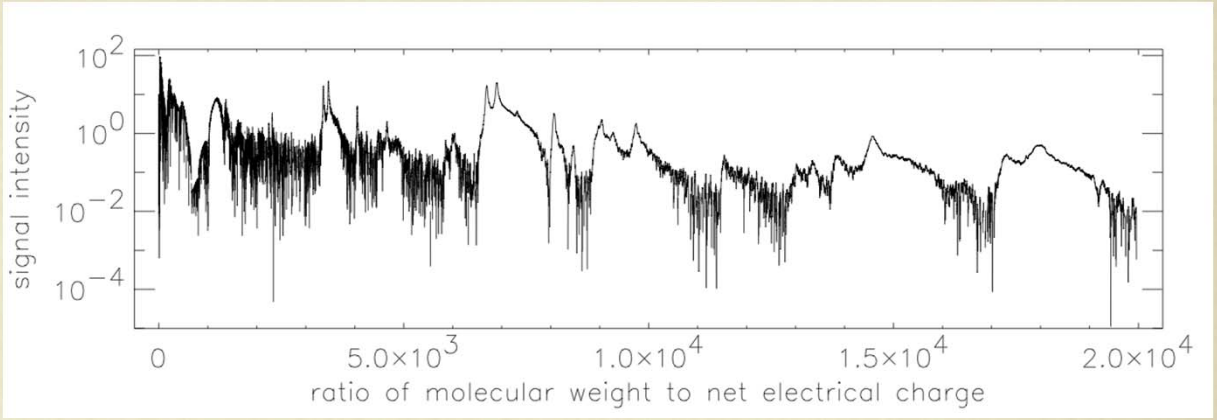
At a high-level, we want to see if a data set contains a pattern that we know about, or even a new pattern that is not just due to chance.

While data can come from a variety of application areas, we will see general techniques that are broadly applicable.

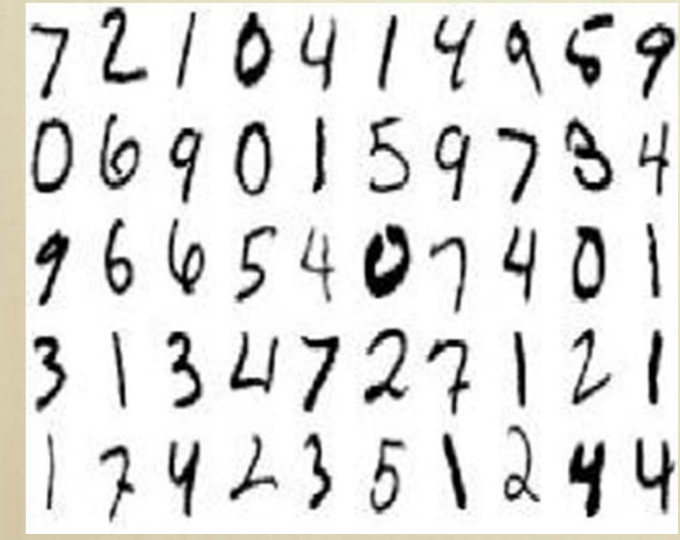
Analyzing and Classifying Data



E.T. Communication?



Healthy or Diseased Sample?

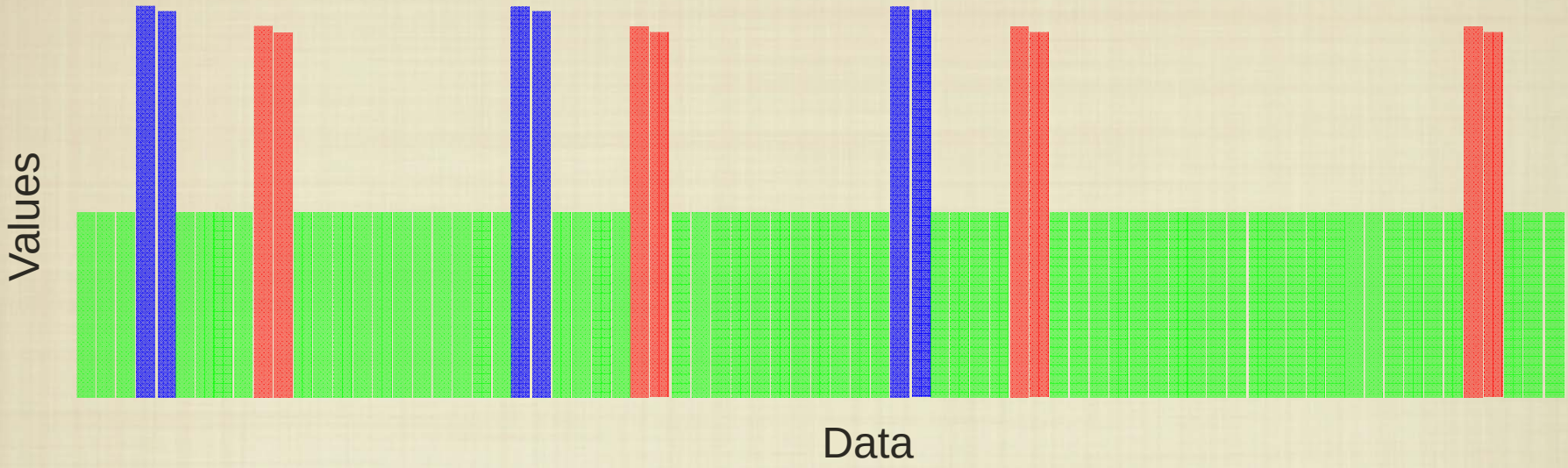


Which digit was written?

At a high-level, we want to see if a data set contains a pattern that we know about, or even a new pattern that is not just due to chance.

While data can come from a variety of application areas, we will see general techniques that are broadly applicable.

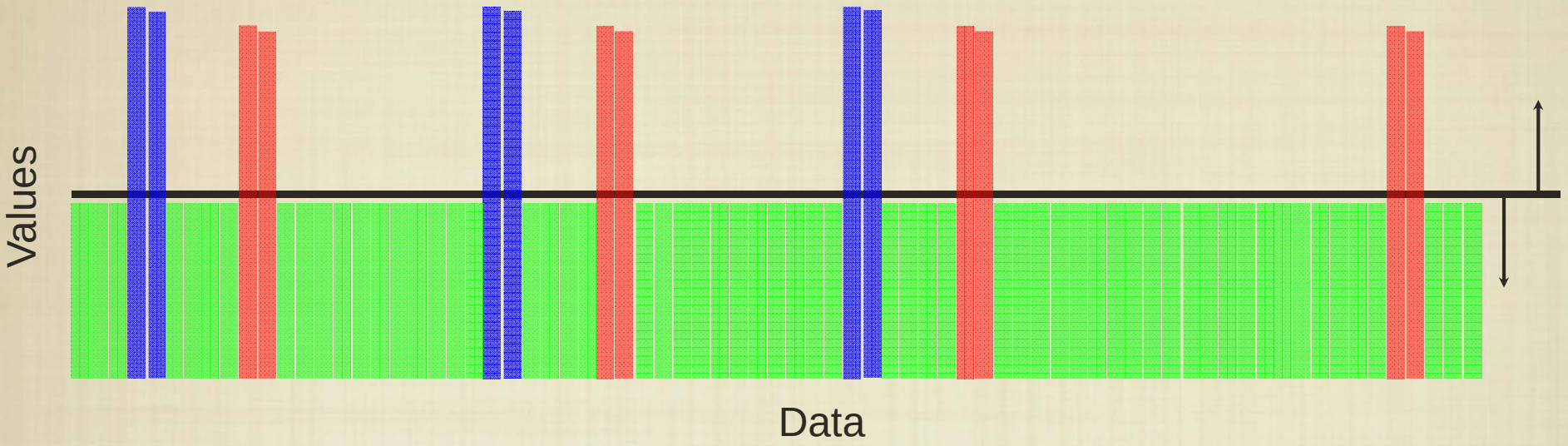
Finding Patterns



If we were given a set of scalar data, how could we identify “significant” peaks?

When does the intensity at a particular frequency exceed the background “noise”?

Finding Patterns

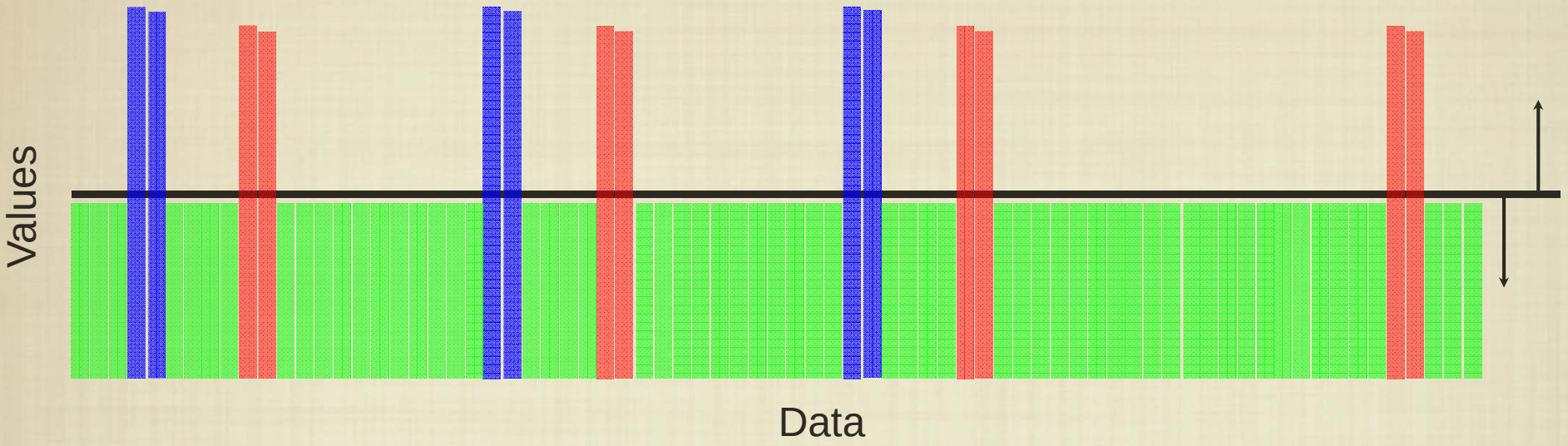


To determine the background noise, we could use the average over all data, and consider values above the average as outliers.

For SETI, these are communication frequencies at which there are “spikes”, and so we can flag them for further consideration.

How long does it take to do this?

Finding Patterns

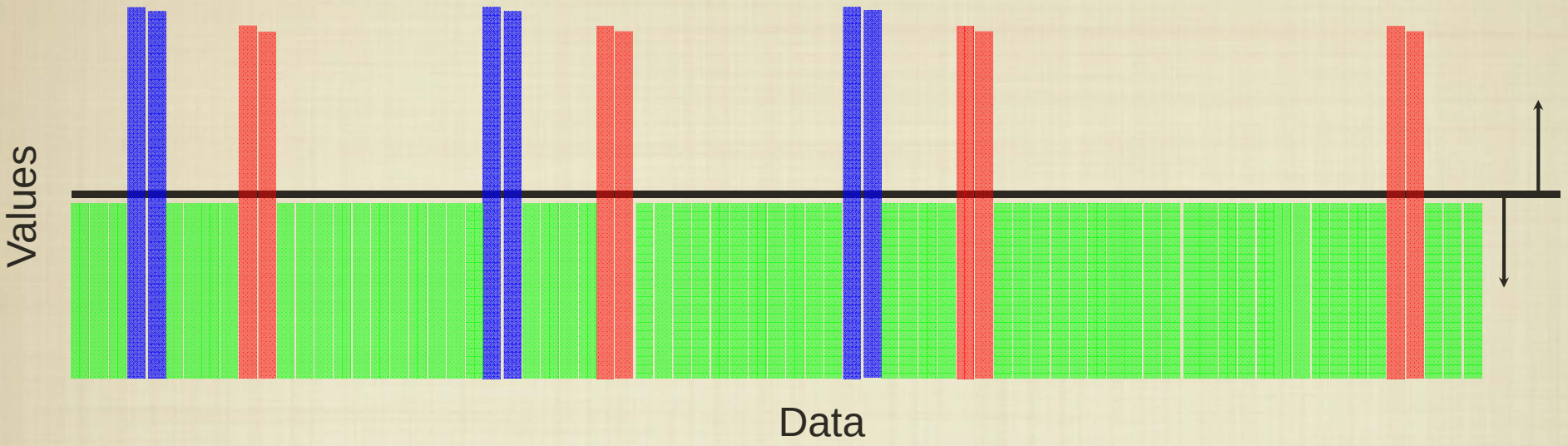


Algorithm:

1. Compute average over all data points.
2. Flag any position at which we have an intensity that is twice the average.

Running time?

Finding Patterns

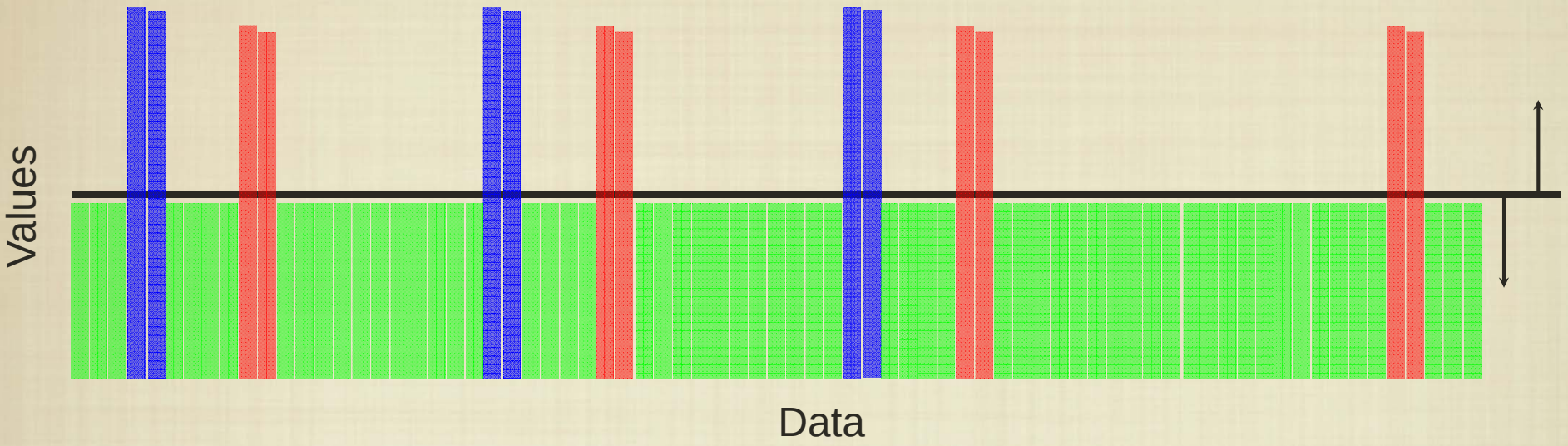


Algorithm:

1. Compute average over all data points.
2. Flag any position at which we have an intensity that is twice the average.

We can scan the data once to compute the average, so this algorithm takes time that is linear in the data size.

Finding Patterns

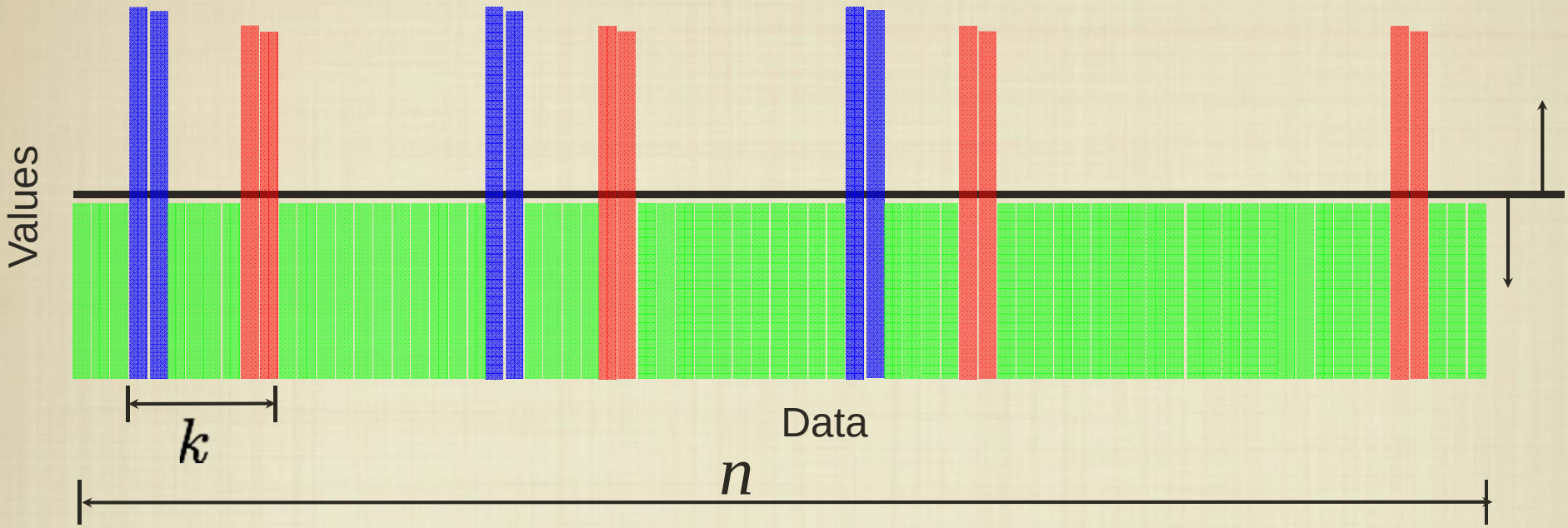


In many cases, we only care about a peak in the signal when it is part of a pattern.

What about more complicated data sets, in which we are interested in a particular pattern between frequencies?

If we are given a pattern, can we efficiently look for it? Can we identify whether a pattern exists in the data? For example, in SETI we look for signal “triplets”.

Sequential Patterns

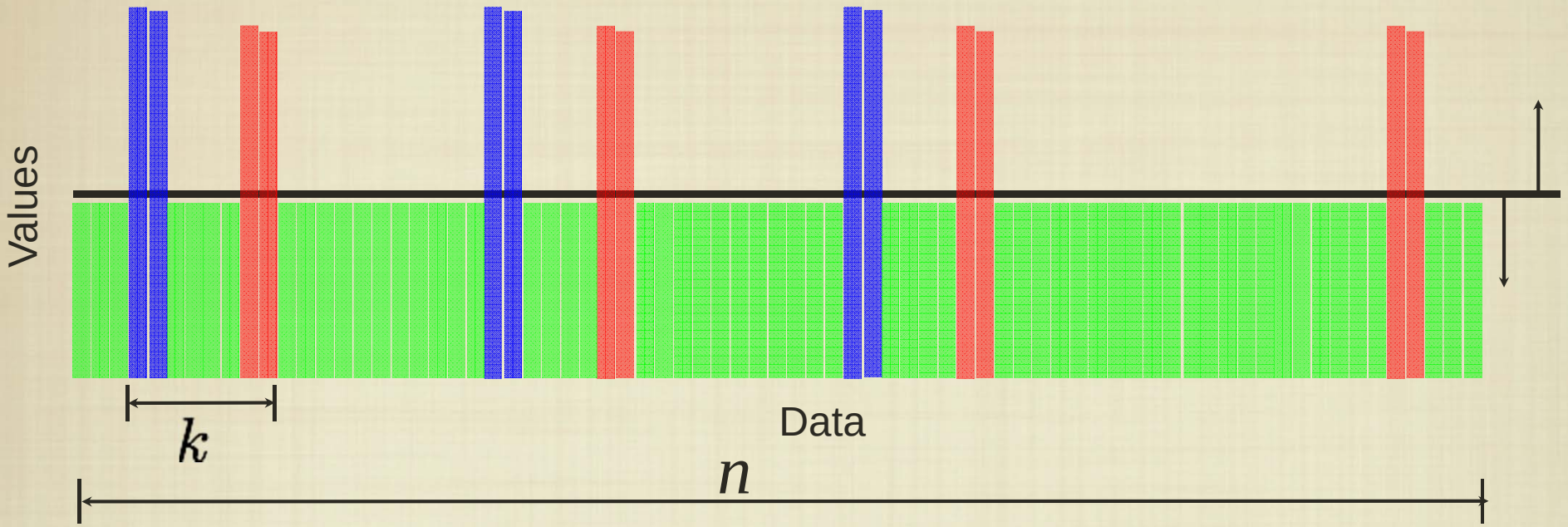


Suppose that we knew that we had a pattern that consisted of multiple peaks, where one outlier follows another within k “steps”.

How do we detect this pattern?

How long does it take?

Sequential Patterns

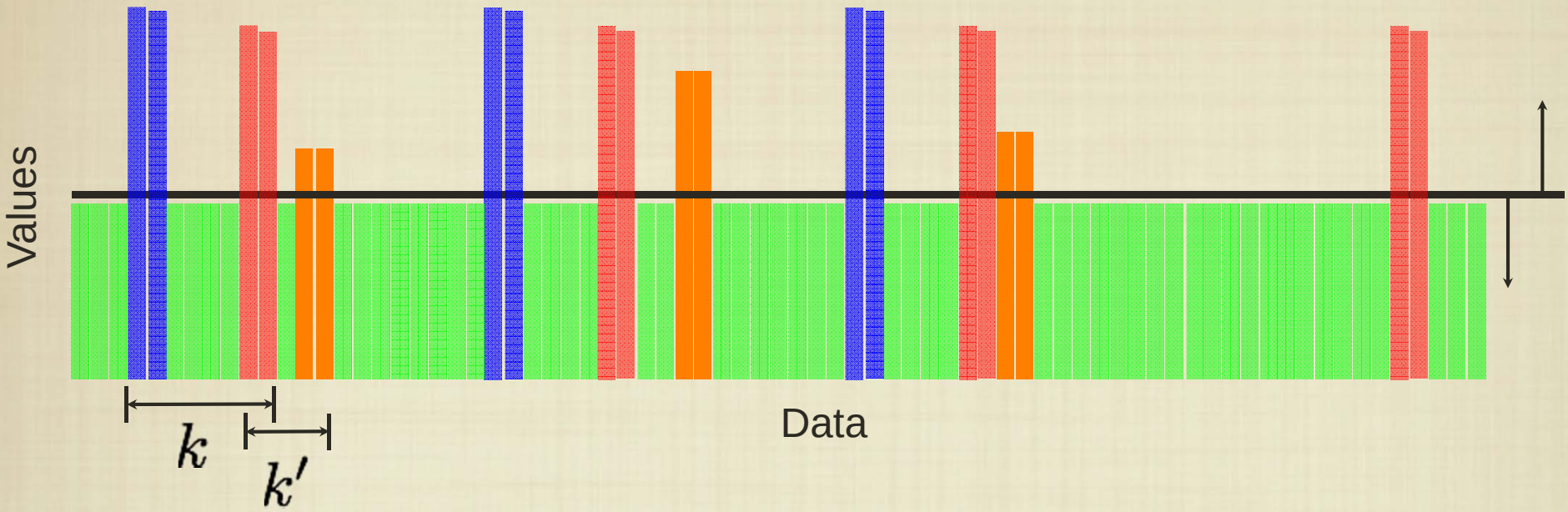


Suppose that we knew that we had a pattern that consisted of multiple peaks, where one outlier follows another within k “steps”.

For each outlier we find, we must look within k entries to see if there is another outlier nearby. In the worst case, this can take $O(nk)$ time.

What if we had multiple sequential outliers?

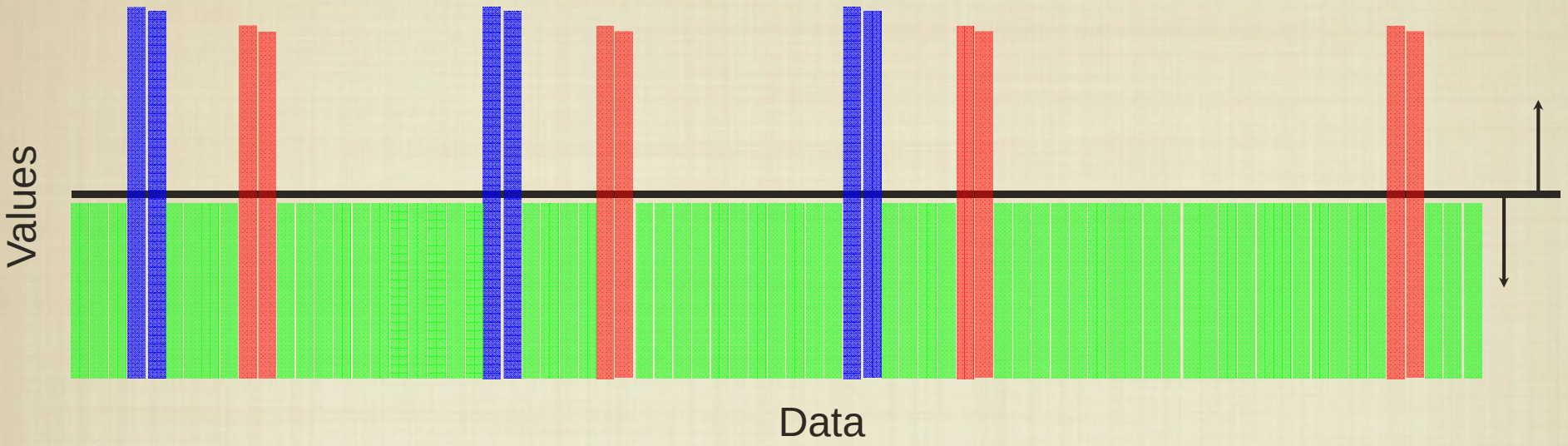
Sequential Patterns



If we know the pattern to look for, then the task of finding outliers to match the pattern just depends on the “width” of the pattern.

What if we do not know what to look for? In many cases, we do not know the exact pattern. And it may be the case that simply thresholding the data is not sufficient.

More Subtle Data



We’ve been discussing using the average to determine an appropriate threshold for outliers: anything larger than twice the average.

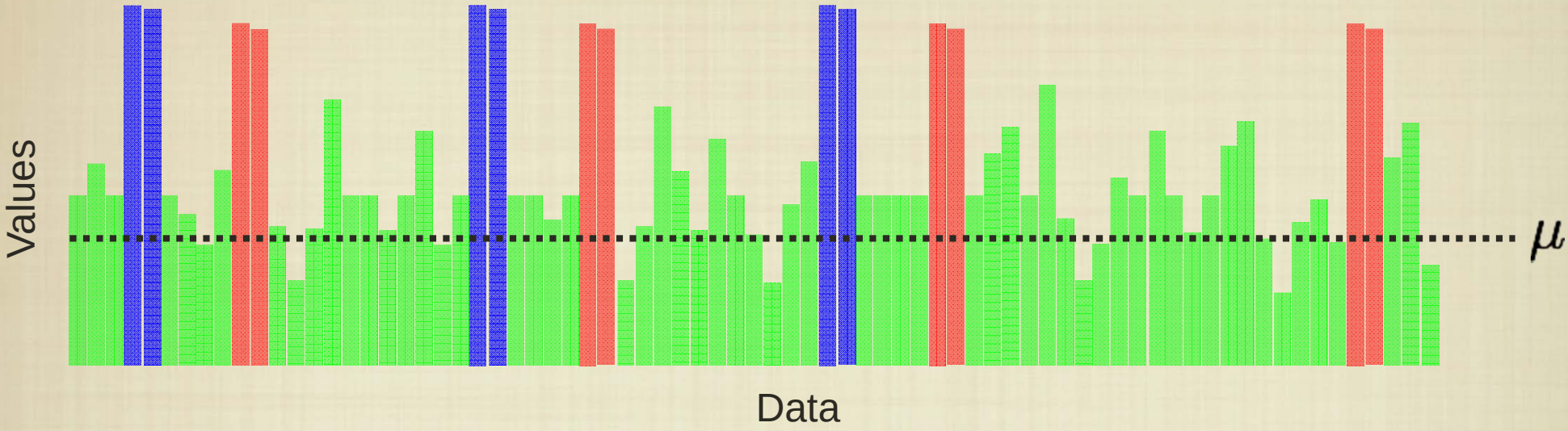
But how realistic is our picture of the problem?

More Subtle Data

We’ve been discussing using the average to determine an appropriate threshold for outliers: anything larger than twice the average.

In certain settings, the data most likely have a fair degree of variability...

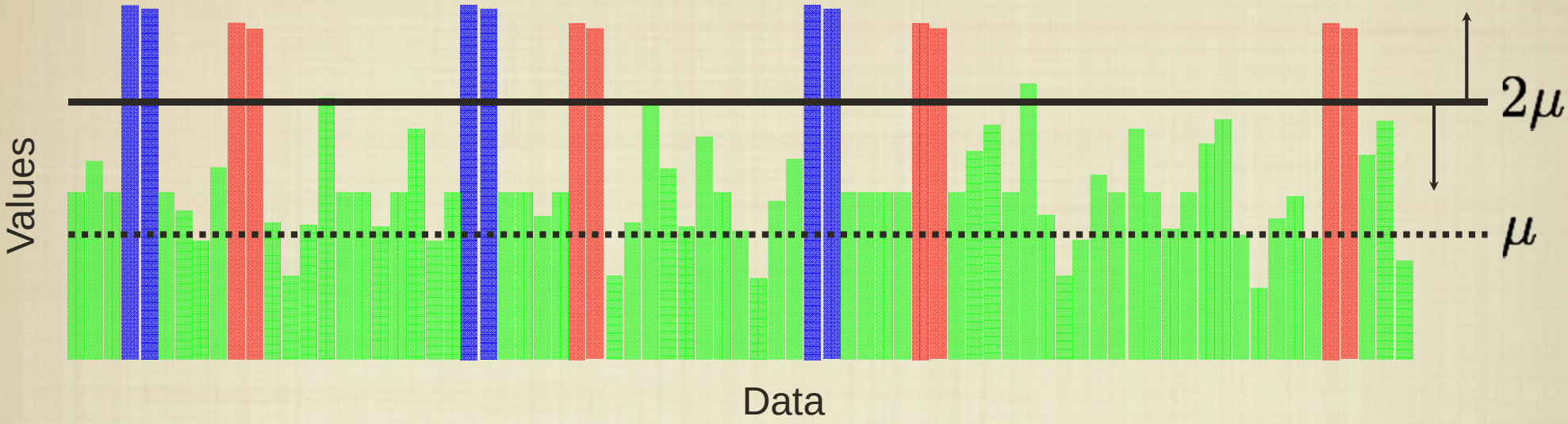
More Subtle Data



We've been discussing using the average to determine an appropriate threshold for outliers: anything larger than twice the average.

How well does our "twice the mean" threshold work?

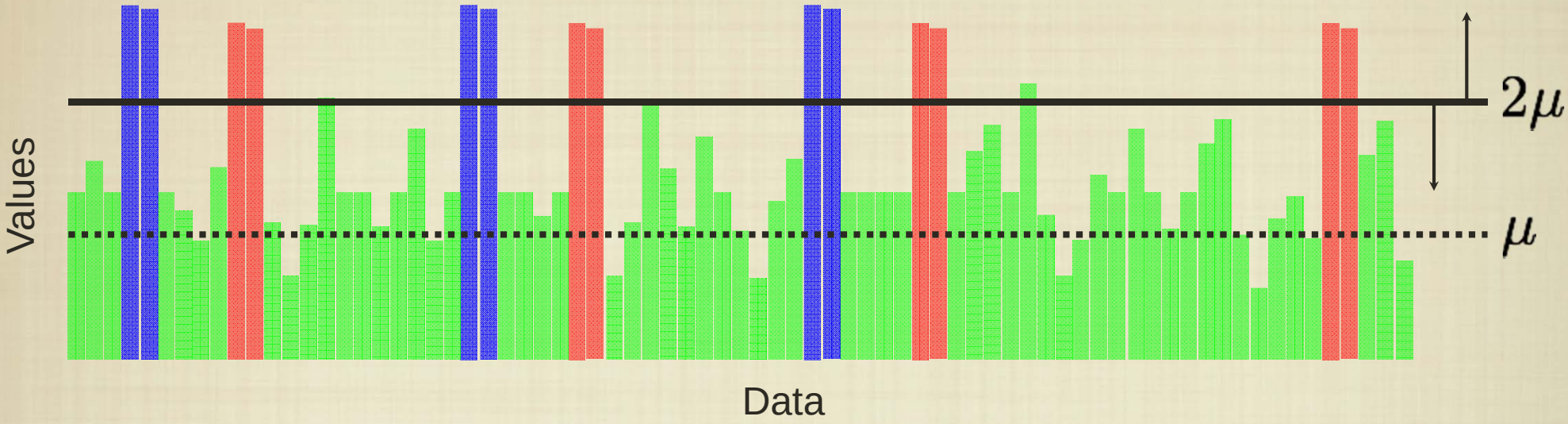
More Subtle Data



We've been discussing using the average to determine an appropriate threshold for outliers: anything larger than twice the average.

How well does our "twice the mean" threshold work?

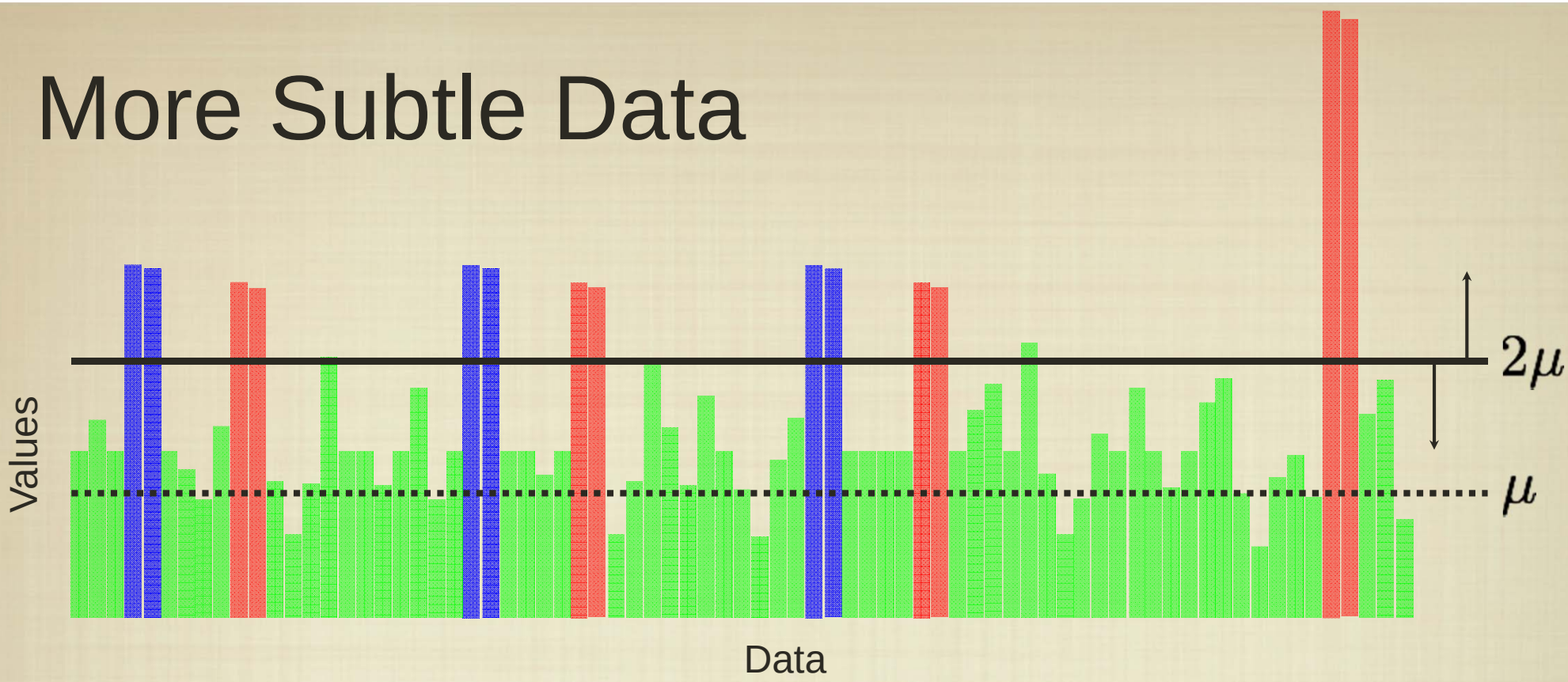
More Subtle Data



We've been discussing using the average to determine an appropriate threshold for outliers: anything larger than twice the average.

Not bad, but are there cases where it can fail?

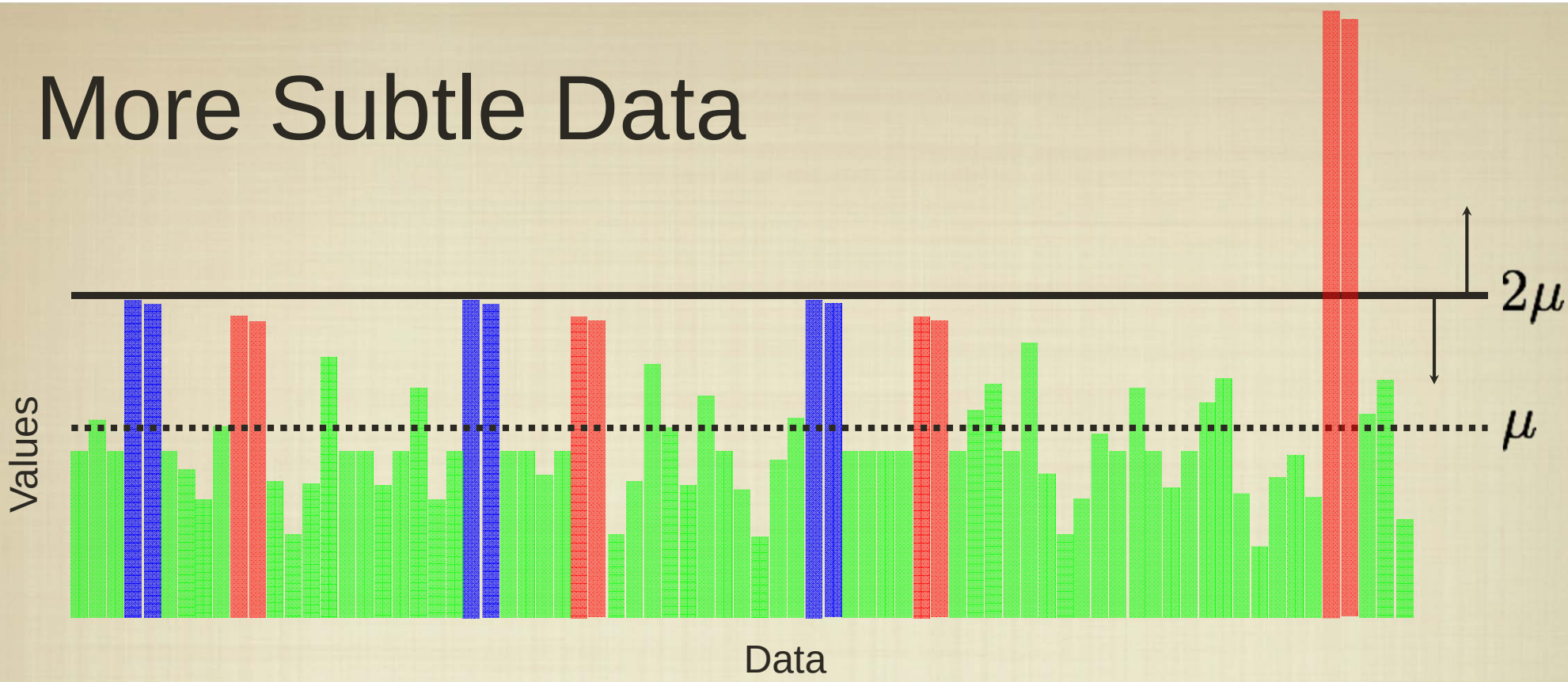
More Subtle Data



We've been discussing using the average to determine an appropriate threshold for outliers: anything larger than twice the average.

What if there is one particular data point that is really large?

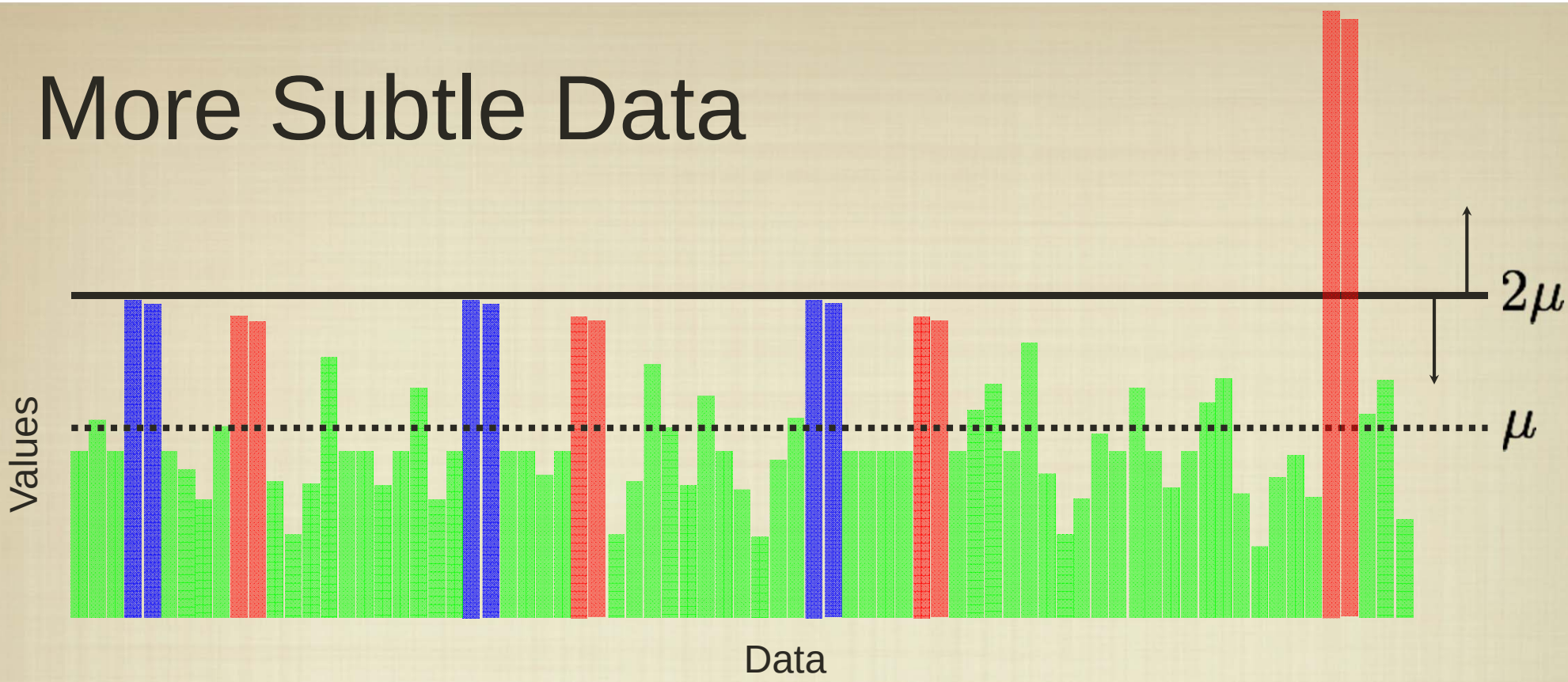
More Subtle Data



We've been discussing using the average to determine an appropriate threshold for outliers: anything larger than twice the average.

What if there is one particular data point that is really large?

More Subtle Data

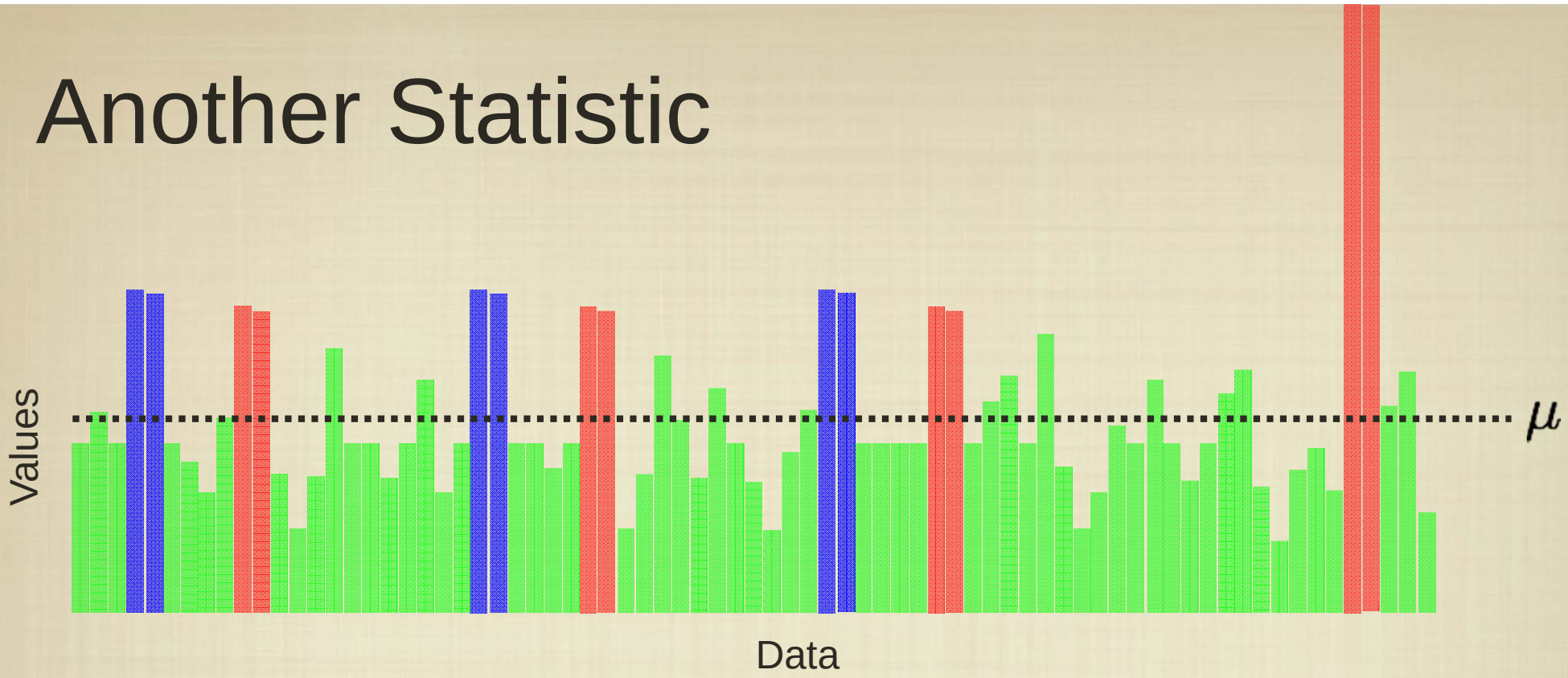


We've been discussing using the average to determine an appropriate threshold for outliers: anything larger than twice the average.

It can also be the case that some data doesn't have that much variation: for example, height and age don't match this profile.

Can we construct a measurement that incorporates the mean?

Another Statistic

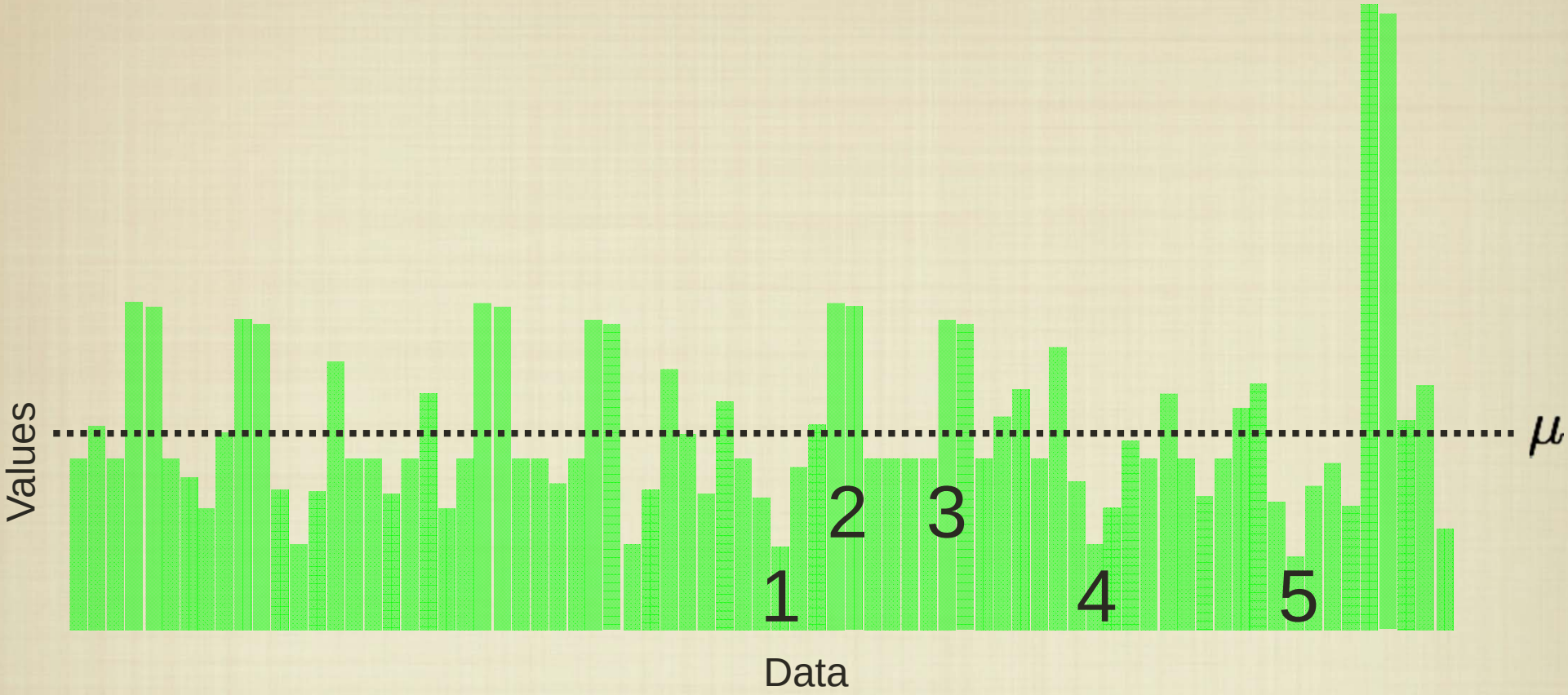


The standard deviation is a sort of average of how much values deviate from the mean:

$$\sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu)^2}$$

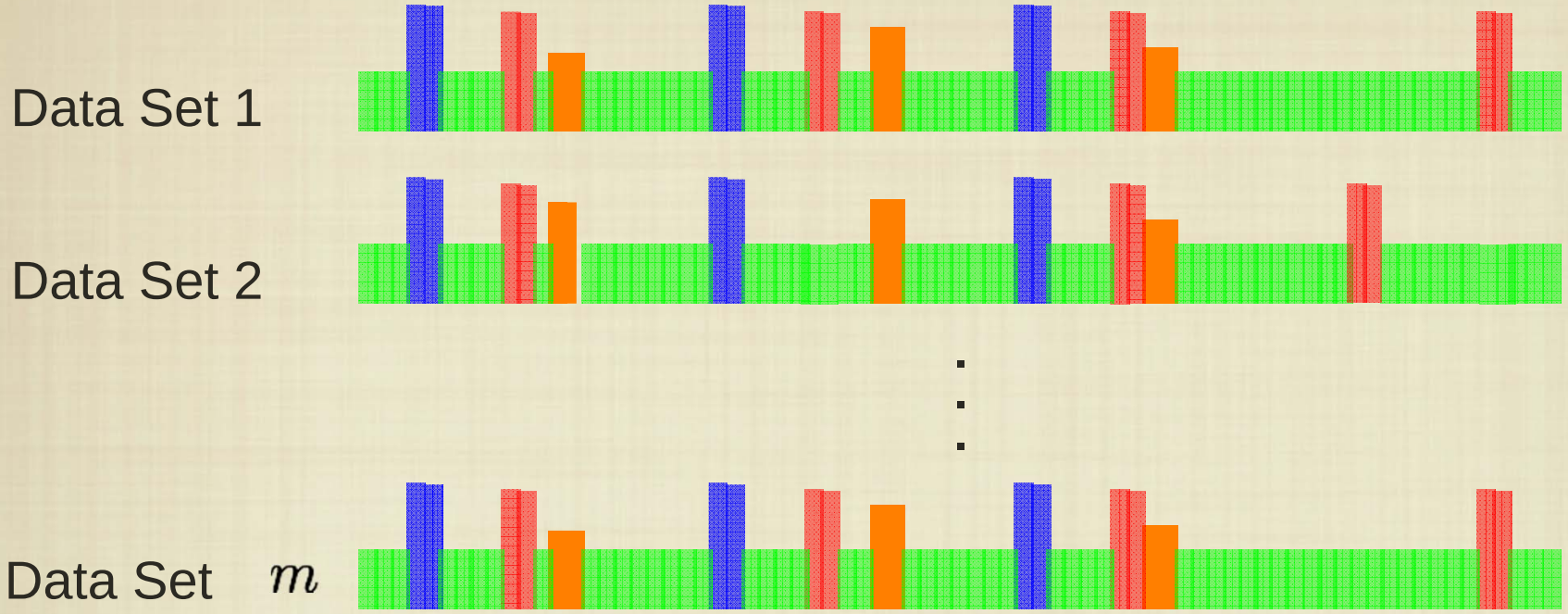
Notice that the standard deviation takes the average into account, and that all data points contribute equally.

Will Statistics Always Work?



Thresholding by any value is not necessarily what we may want; we cannot capture patterns between elements. This may not at all be intuitive if we are exploring the data for a new phenomenon.

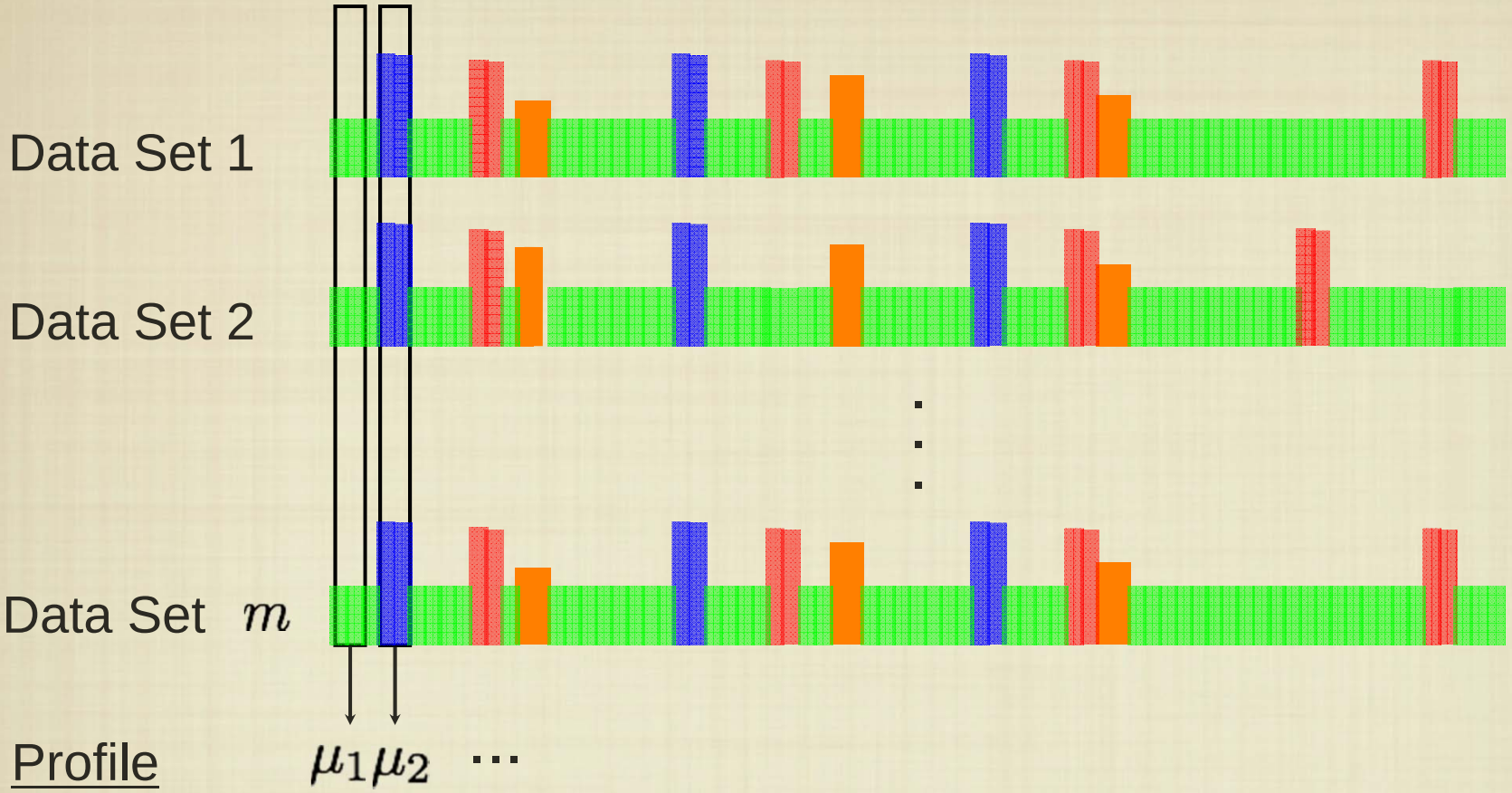
Learning From Data



If we do not know what the data might contain, can we learn from examples of what we are trying to identify?

Suppose we have a number of examples of what we are trying to identify, how can we build a profile?

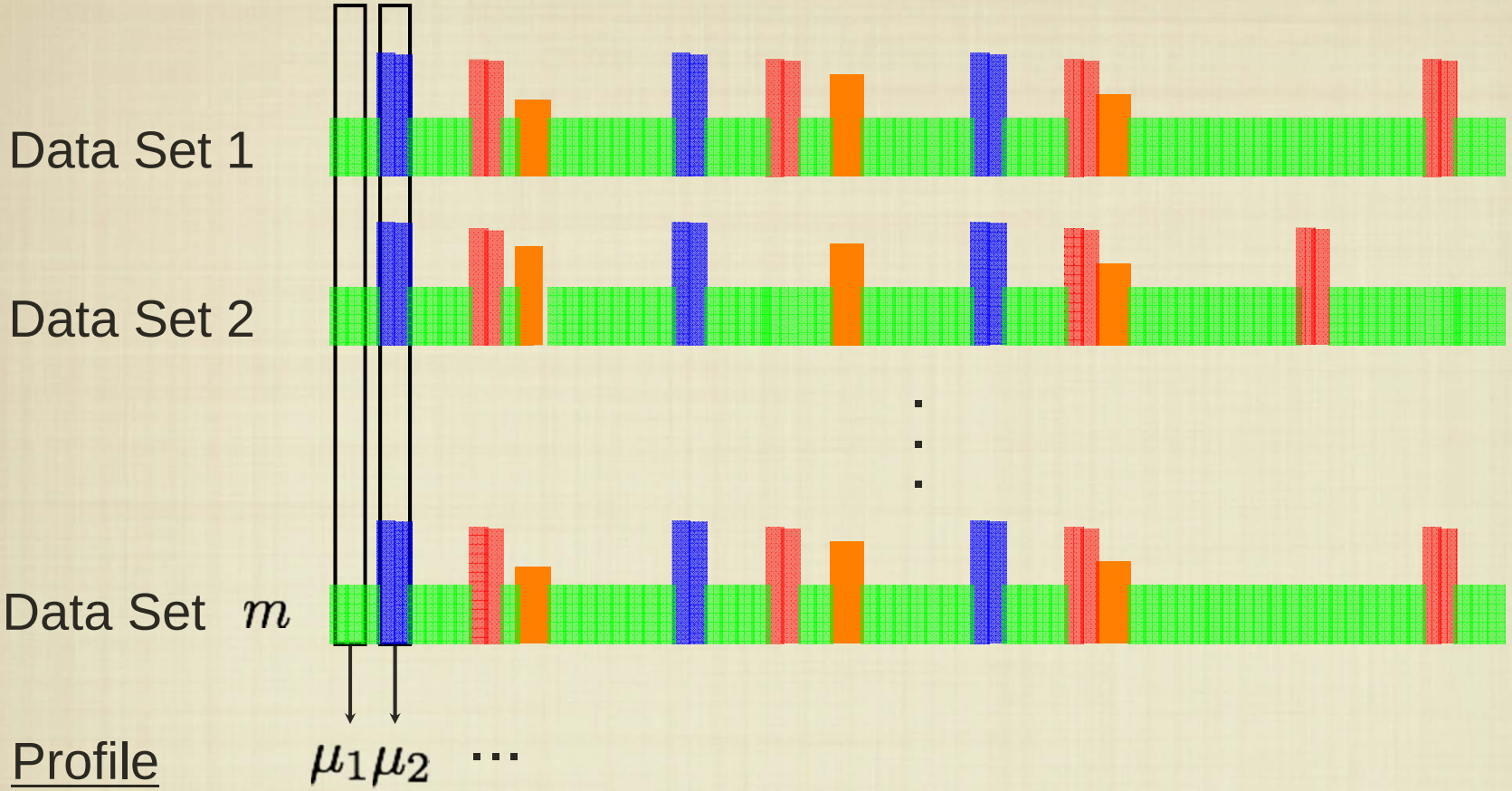
Learning From Data



We can take an average of each position, and use the resulting sequence of averages as a “profile”.

How long does it take to compute a profile?

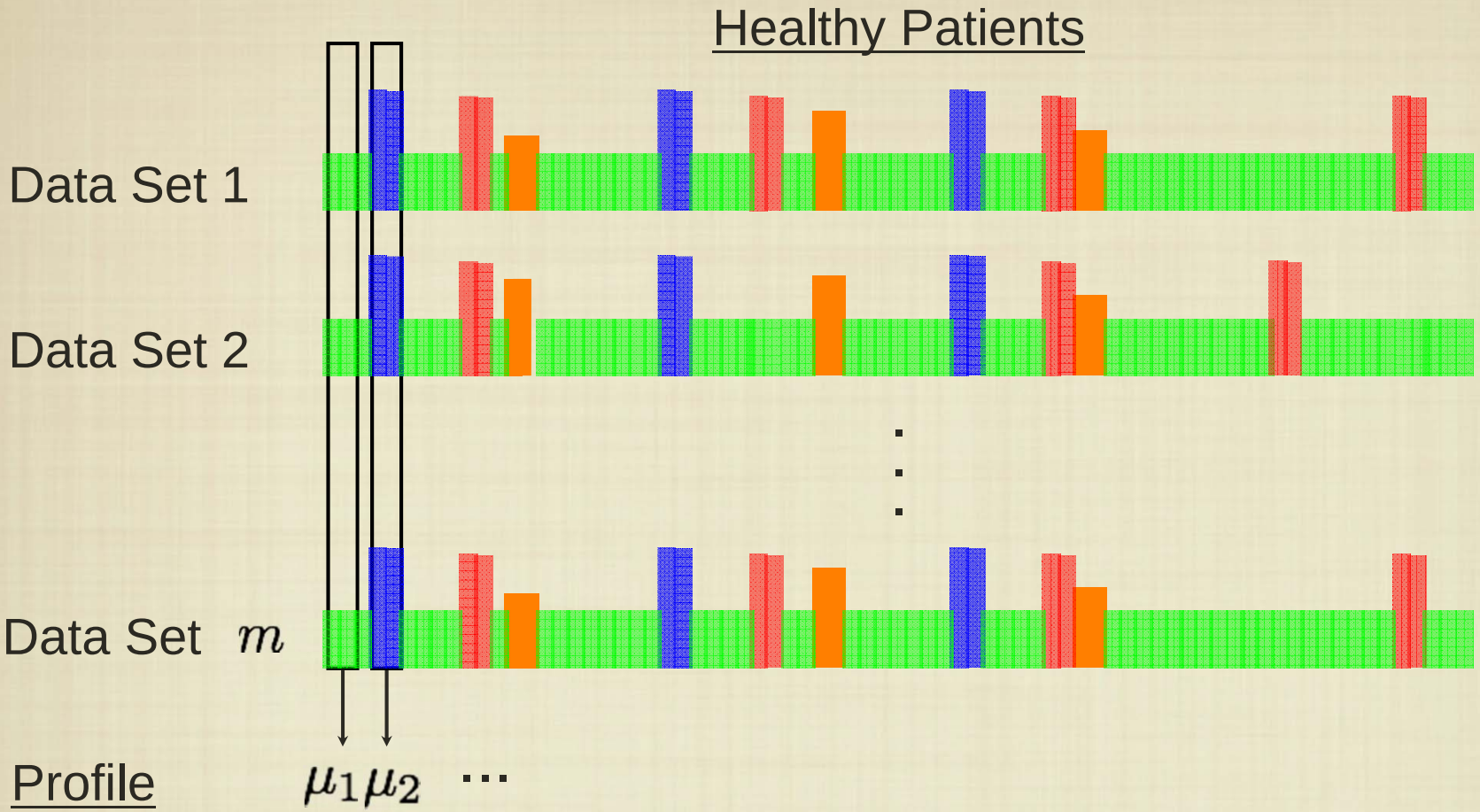
Learning From Data



A profile can be computed in $O(mn)$ time, just by considering each position in the data in turn.

Note that the profile is “position specific”; does this matter?

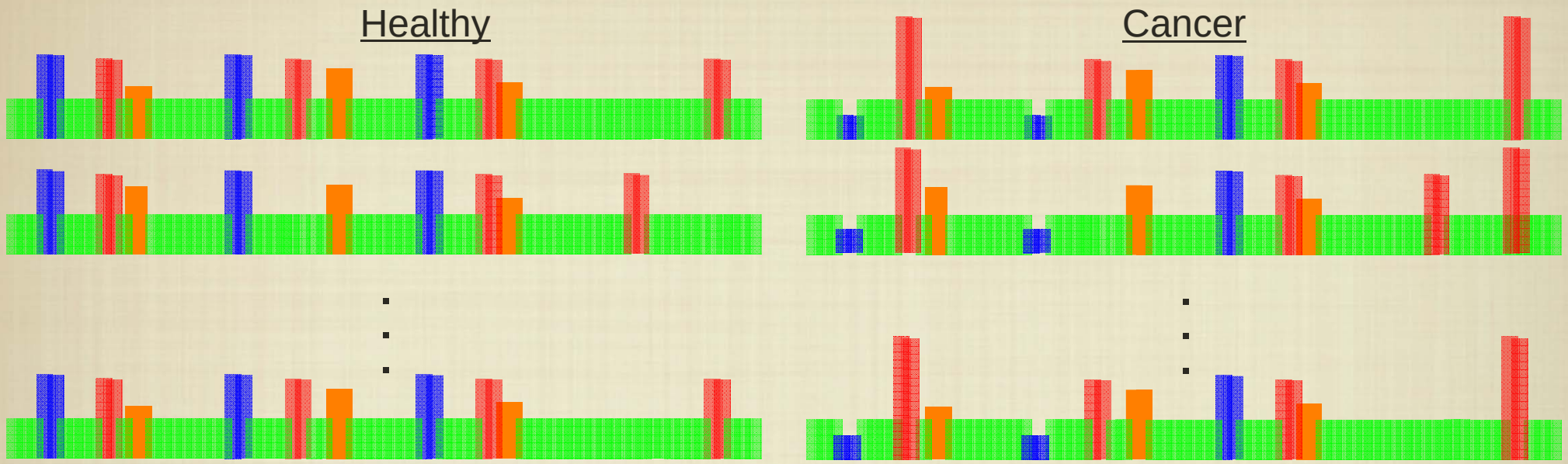
Learning From Data



The computed profile is position specific, but we can break out patterns and apply them to new data.

Can we tell the difference between diseased and healthy patients?

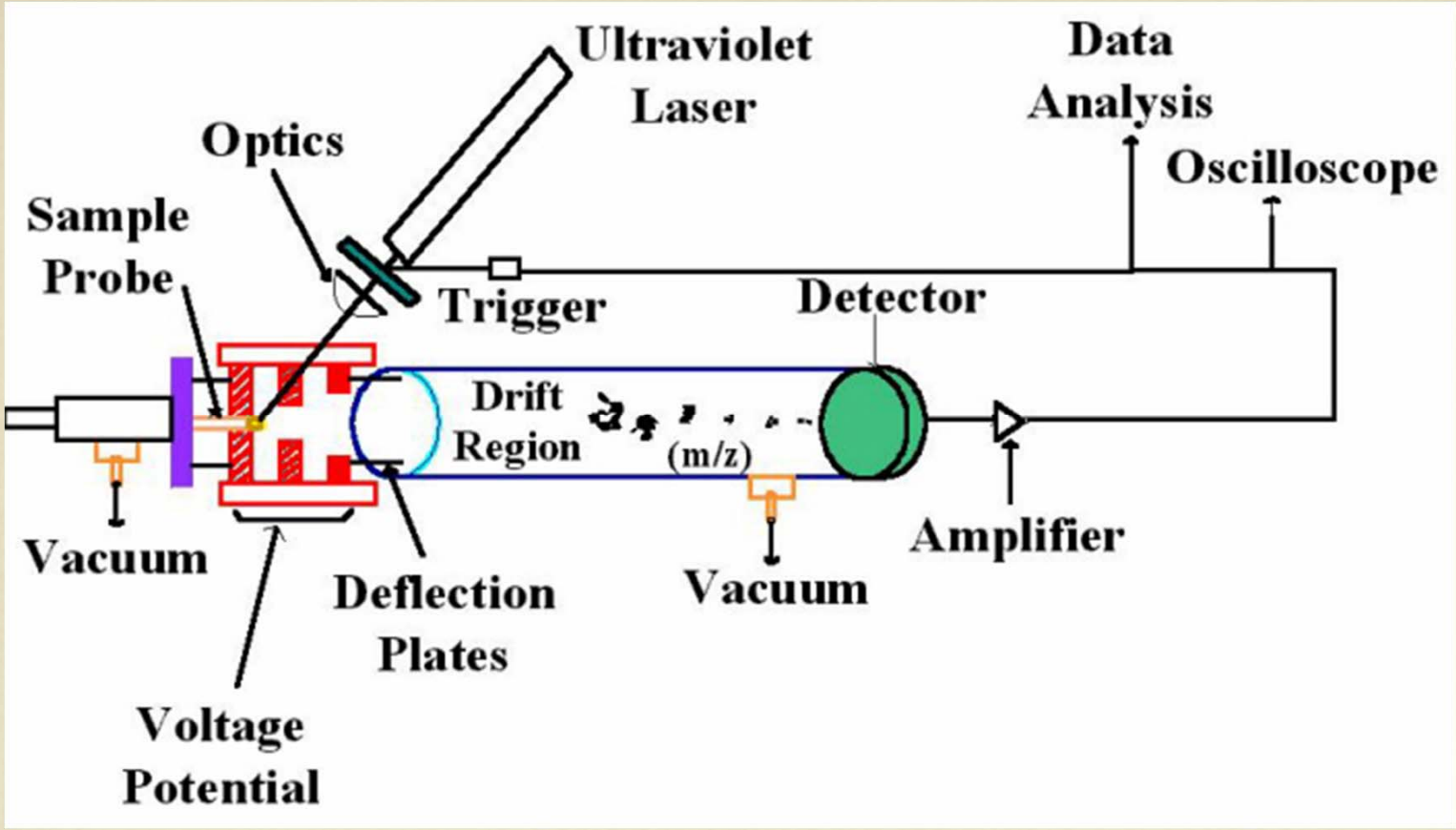
Supervised Learning



Dependencies between data points is still not captured in the position-specific approach, and it may be important for feature-rich data.

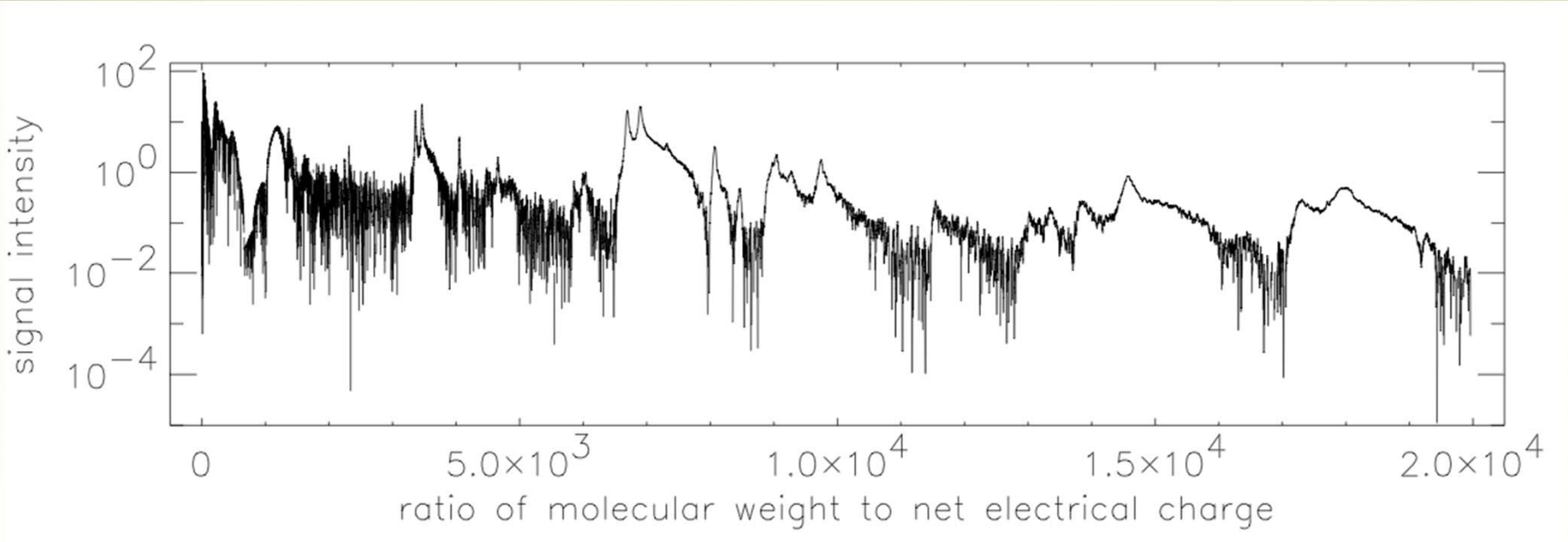
Can we learn these dependencies if we are given labelled examples of each condition?

Mass Spectrometry



Mass spectrometry has become a widely used technique because even very slightly different samples will have different chemical characteristics that give rise to different spectra.

Mass Spectrometry



Mass spectrometry in fact just produces a list of numbers for each sample, indicating the abundance of various masses (i.e., chemical components).

How do we construct a profile that tells us about possible correlations?

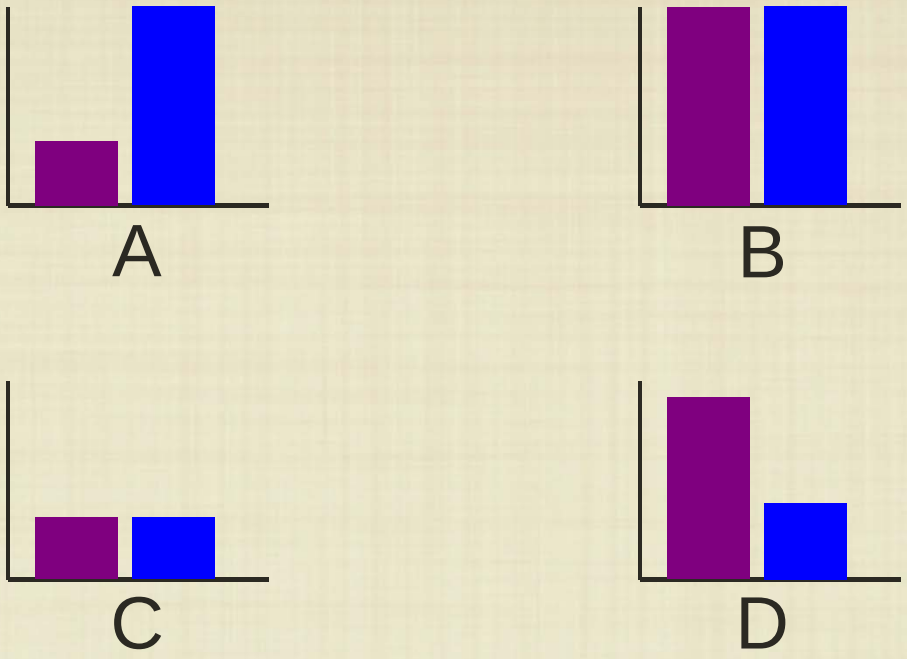
Mass Spectrometry

[Baggerly Et. Al. ’04]

Mass spectrometry in fact just produces a list of numbers for each sample, indicating the abundance of various masses (i.e., chemical components).

How do we construct a profile that tells us about possible correlations?

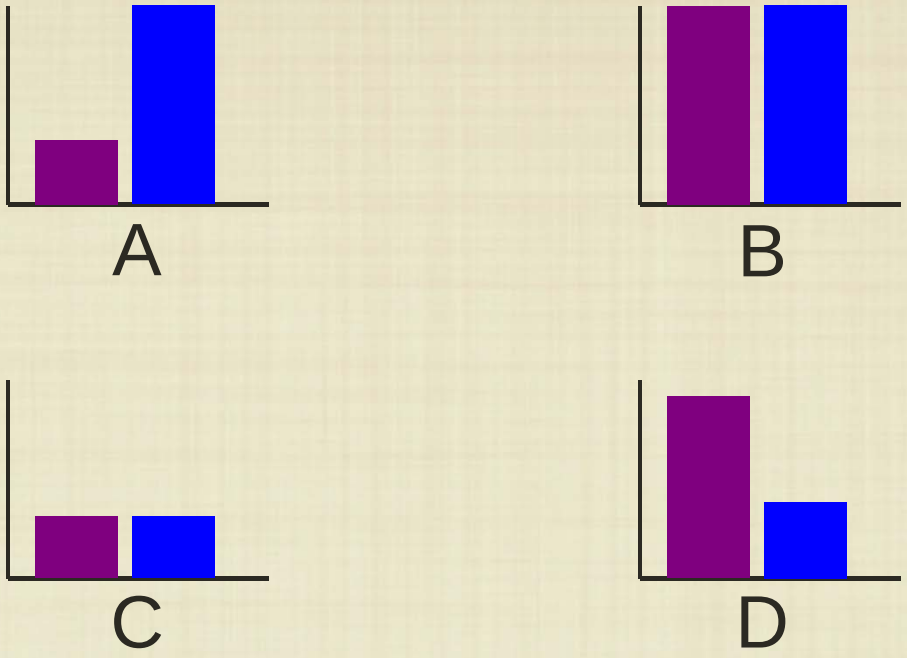
Two Dimensional Data



As a simple example, consider the above types of data, consisting of just two masses each.

How can we distinguish between these four kinds of spectra?

Two Dimensional Data

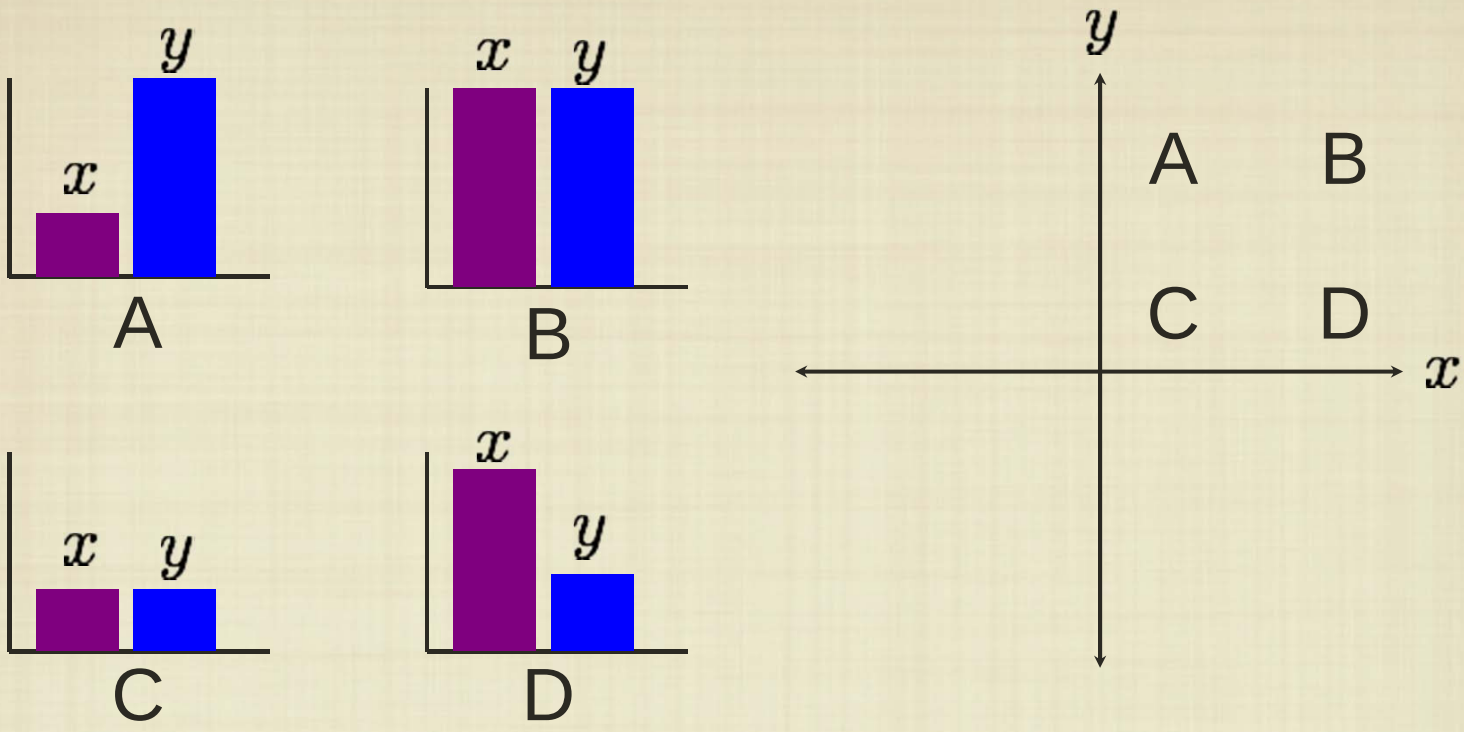


As a simple example, consider the above types of data, consisting of just two masses each.

How can we distinguish between these four kinds of spectra?

What if we represent this data in two dimensions?

Two Dimensional Data

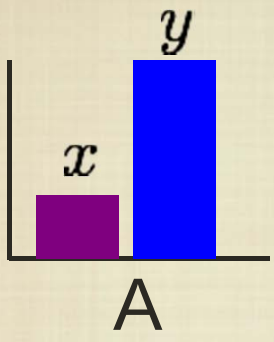


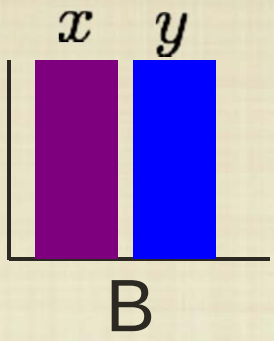
These four pieces of data all have a spatial region that they inhabit, as defined according to the mass spectrum.

How do we use this to classify patients?

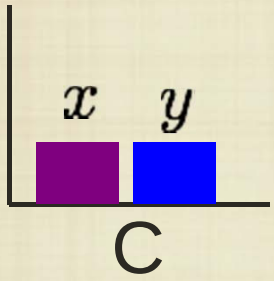
Training and Testing

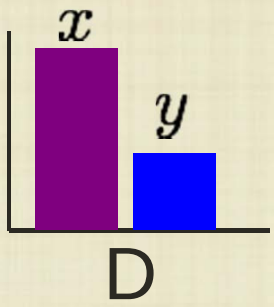
Healthy

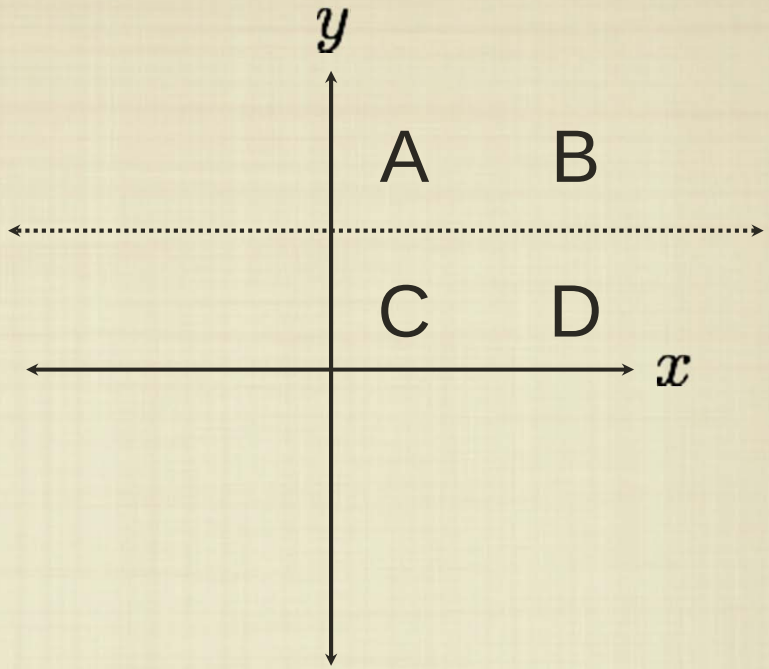




Cancer







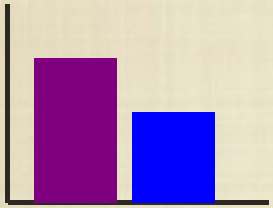
These four pieces of data all have a partial spatial region that they inhabit, as defined according to the mass spectrum.

How do we use this to classify patients?

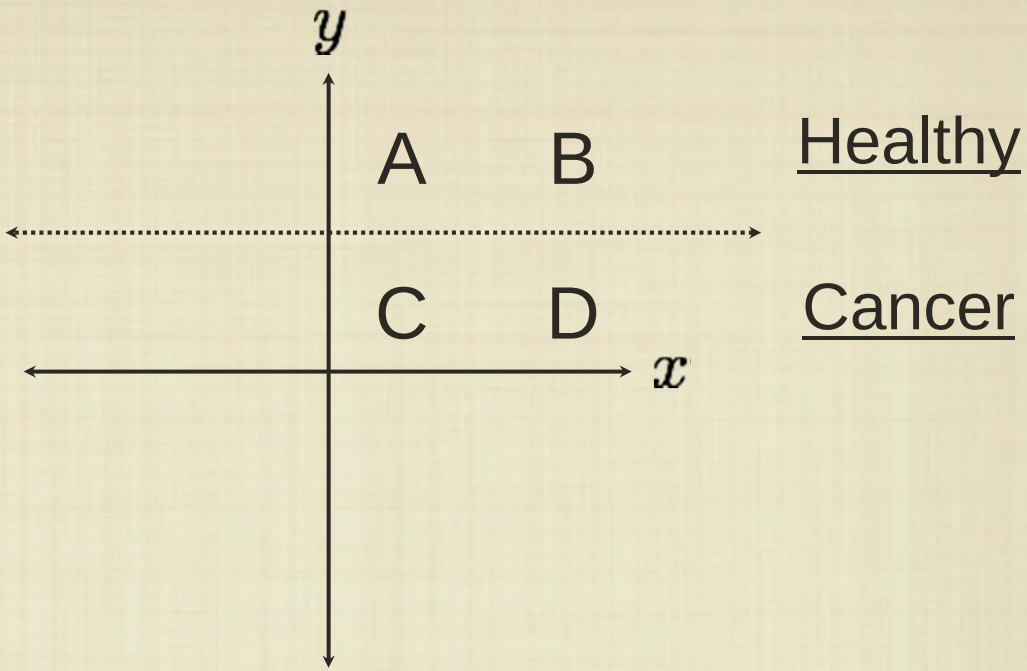
We can draw a threshold between groups of data, viewing each data set as a point in space.

Training and Testing

Healthy?



E

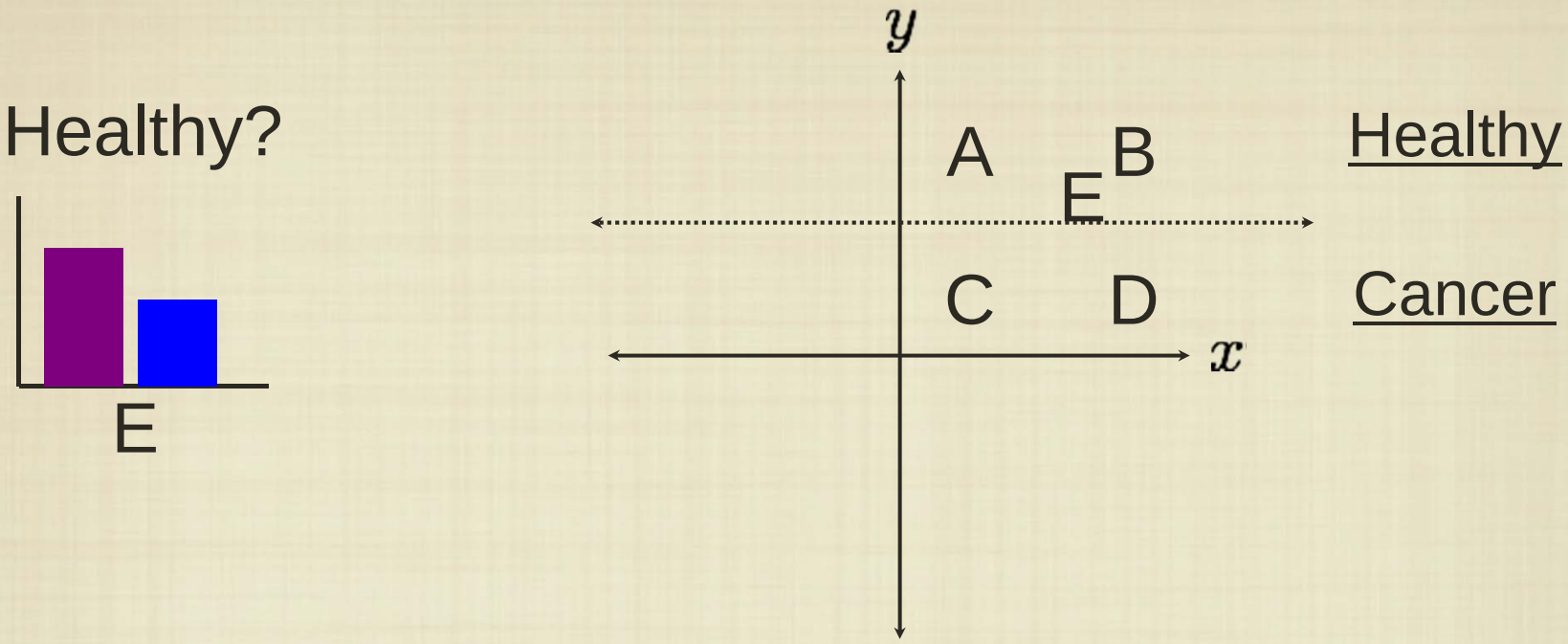


These four pieces of data all have a partial spatial region that they inhabit, as defined according to the mass spectrum.

How do we use this to classify patients?

When we get a new patient sample, we can see which category it belongs to, in a geometric sense.

Training and Testing

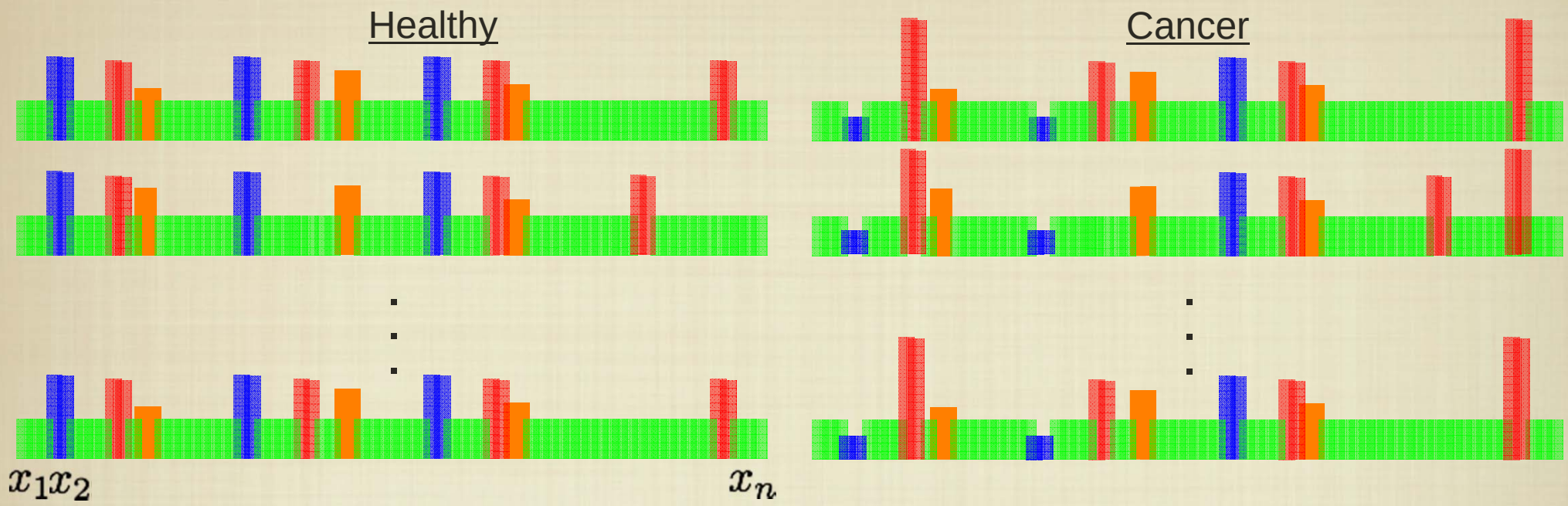


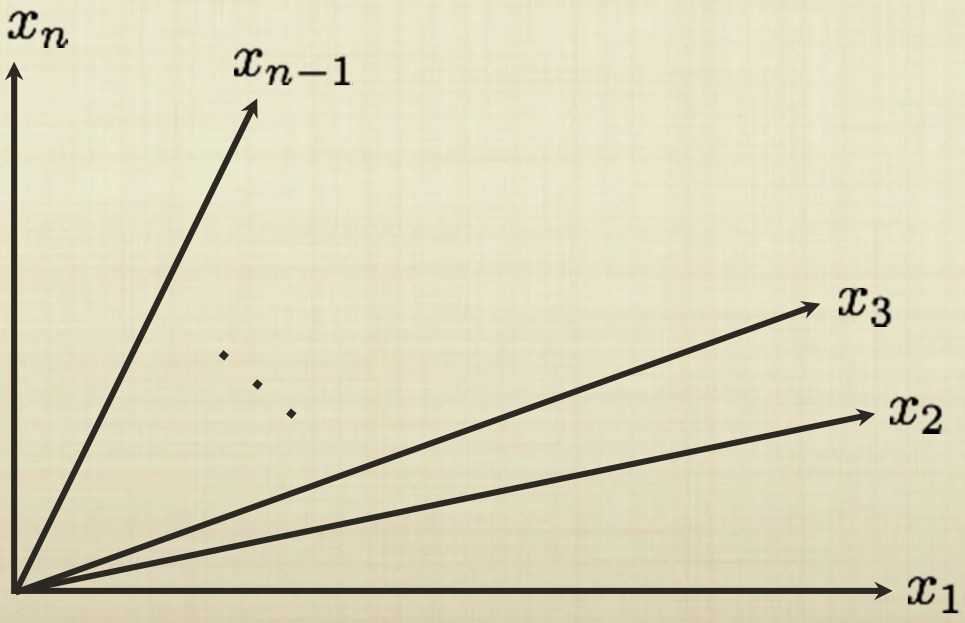
These four pieces of data all have a partial spatial region that they inhabit, as defined according to the mass spectrum.

How do we use this to classify patients?

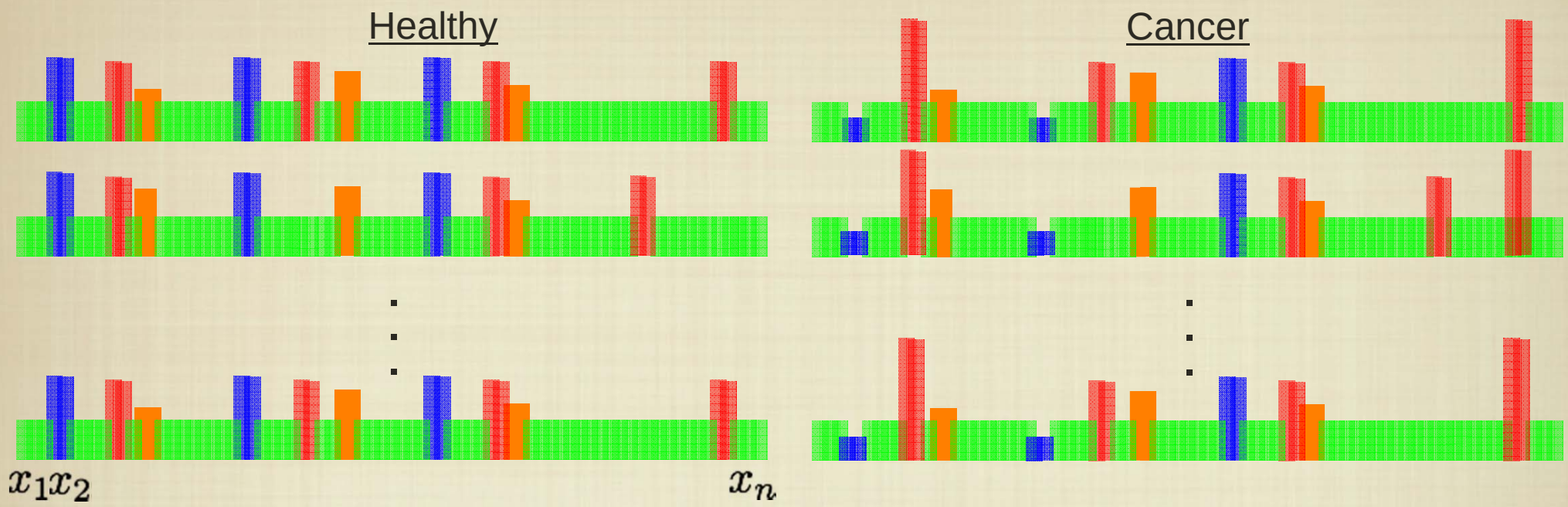
When we get a new patient sample, we can see which category it belongs to, in a geometric sense.

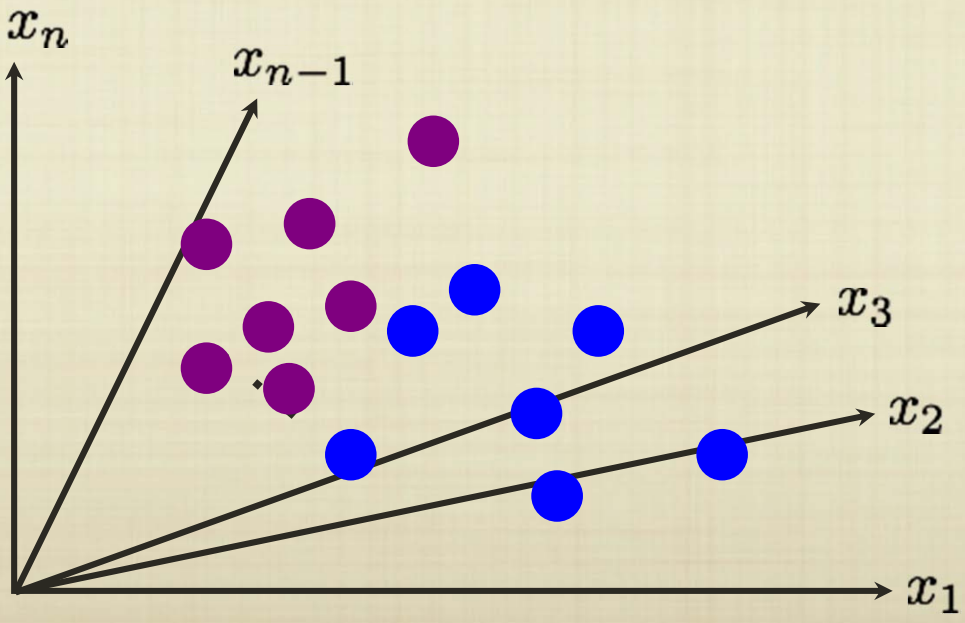
In Higher Dimensions



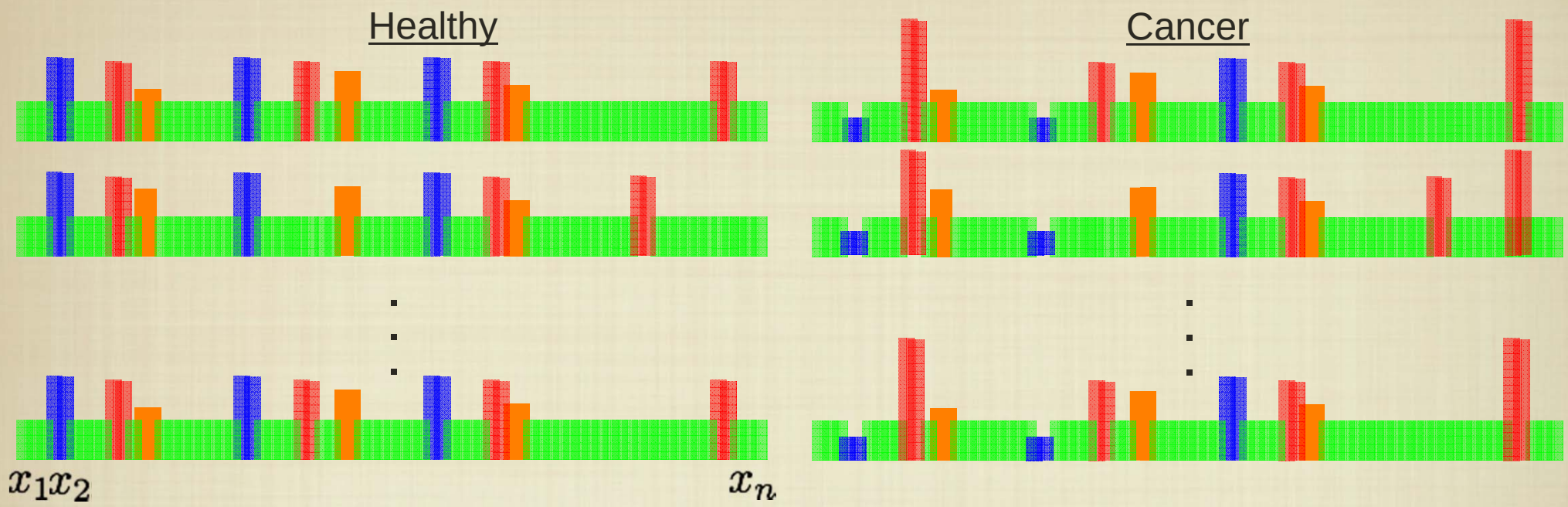


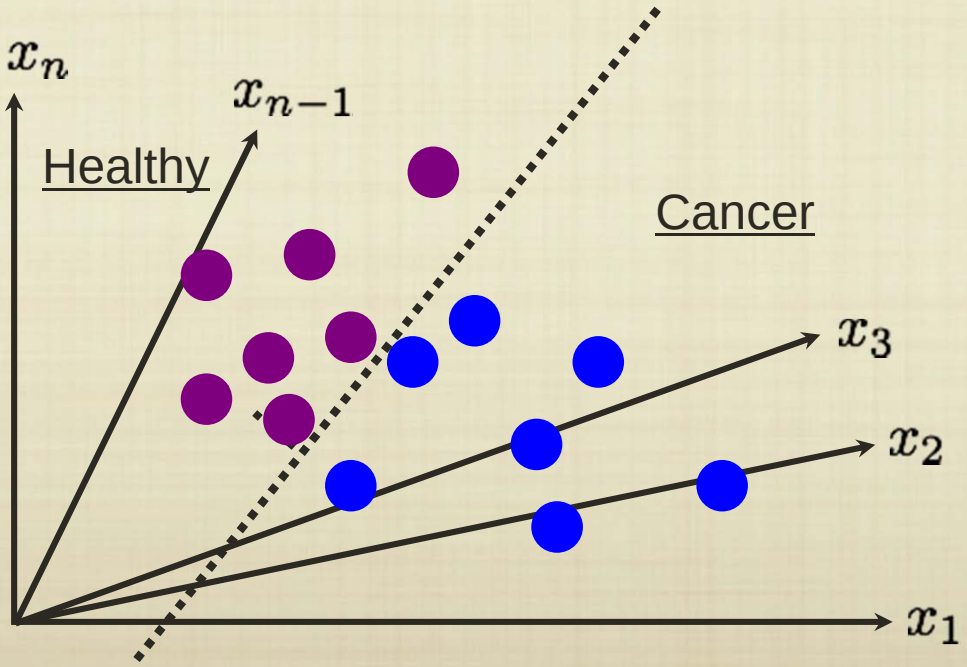
In Higher Dimensions



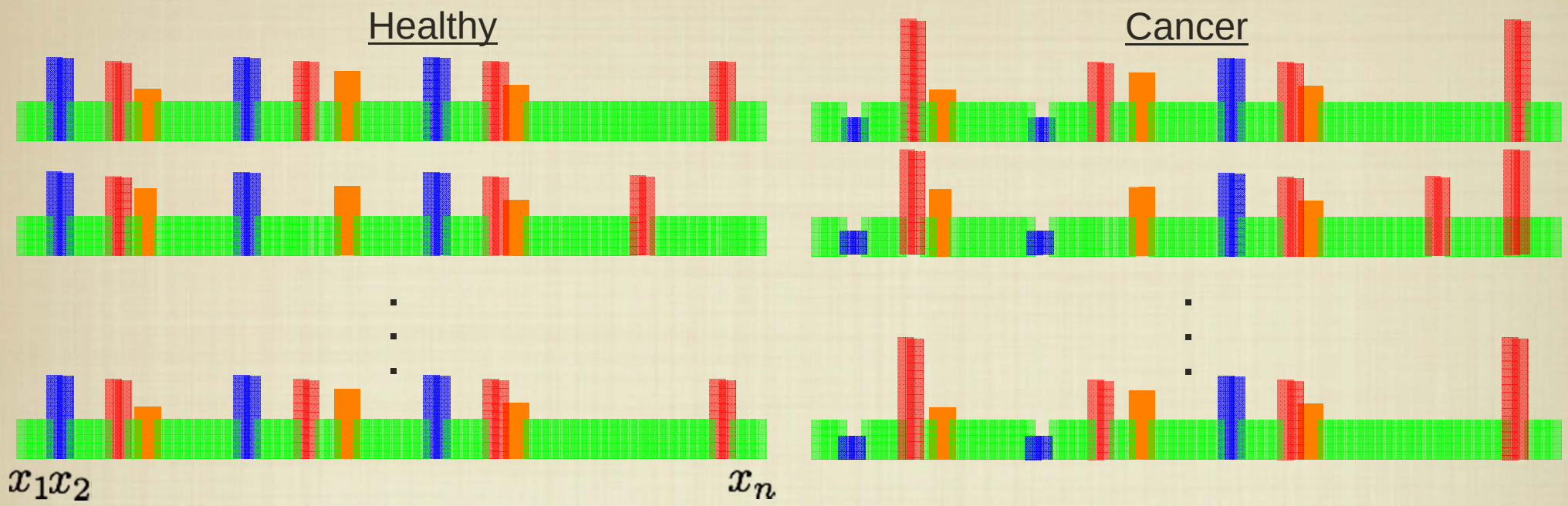


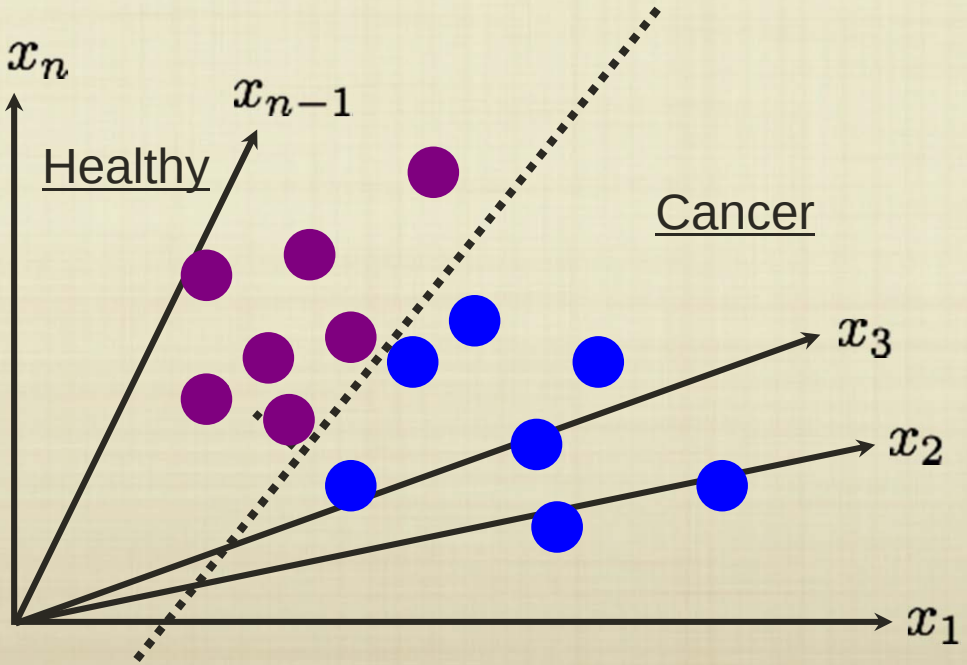
In Higher Dimensions





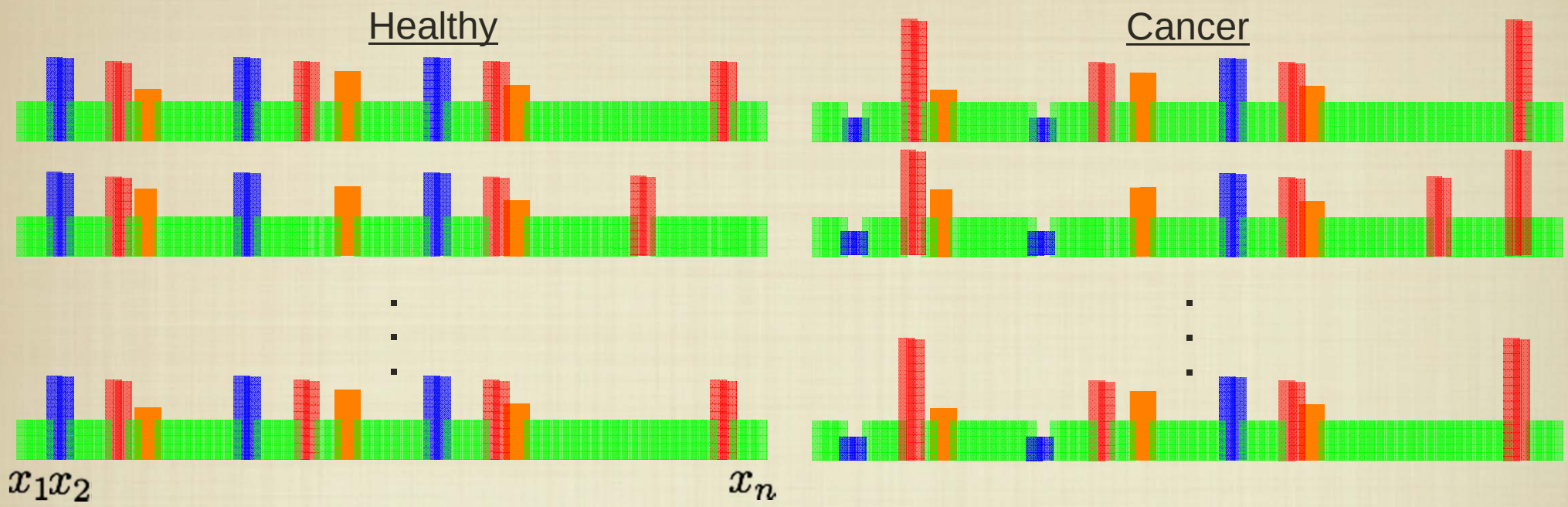
In Higher Dimensions

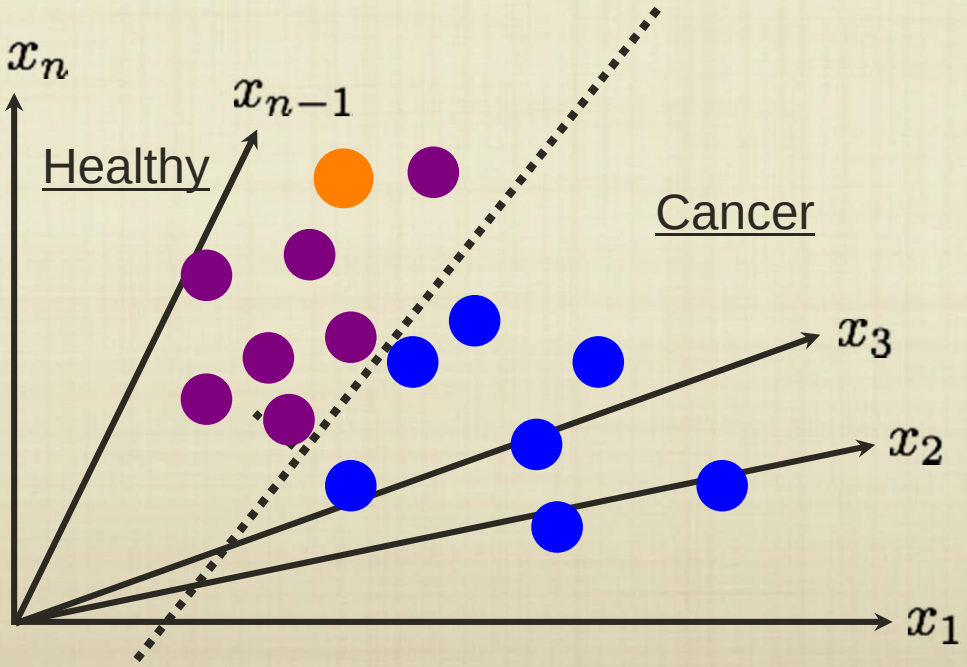




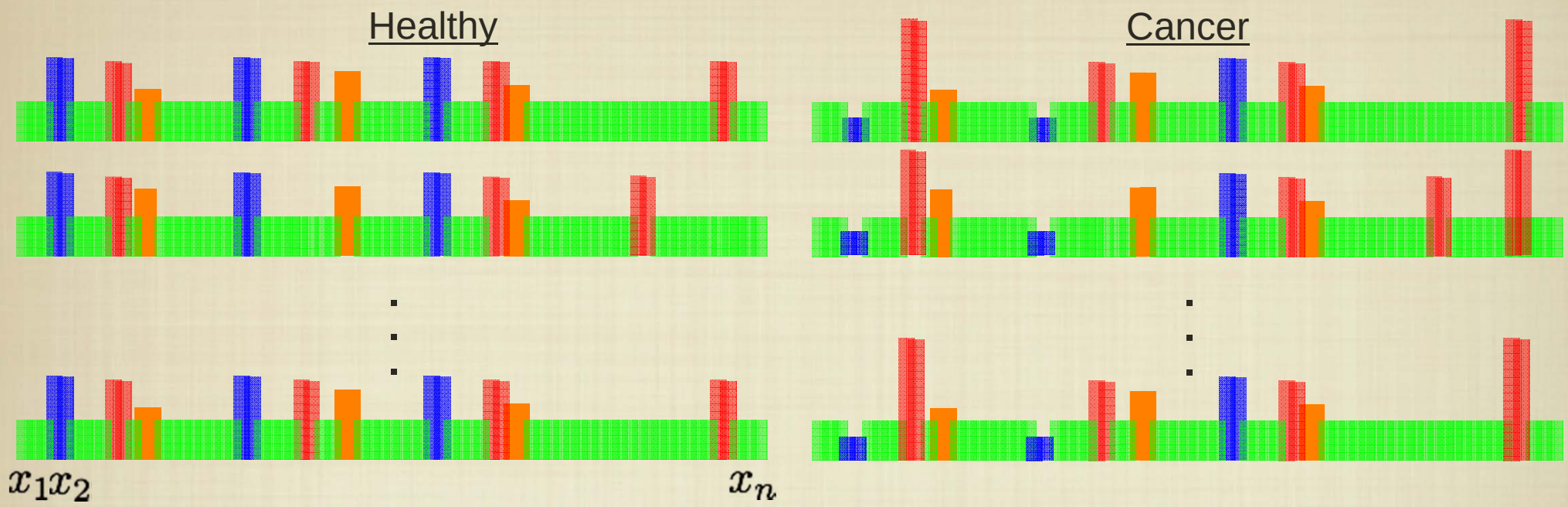
?

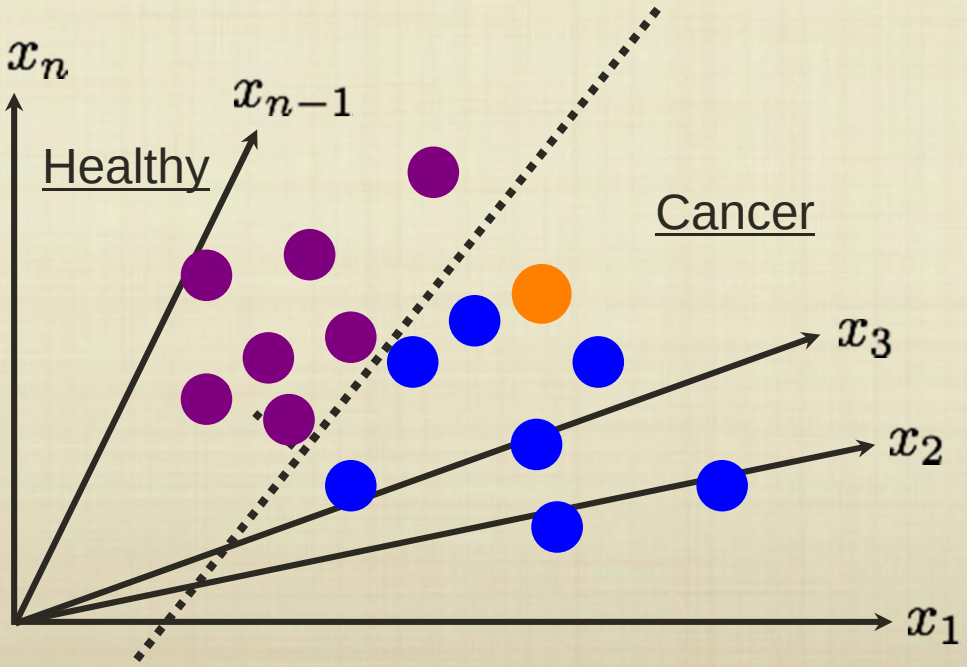
In Higher Dimensions



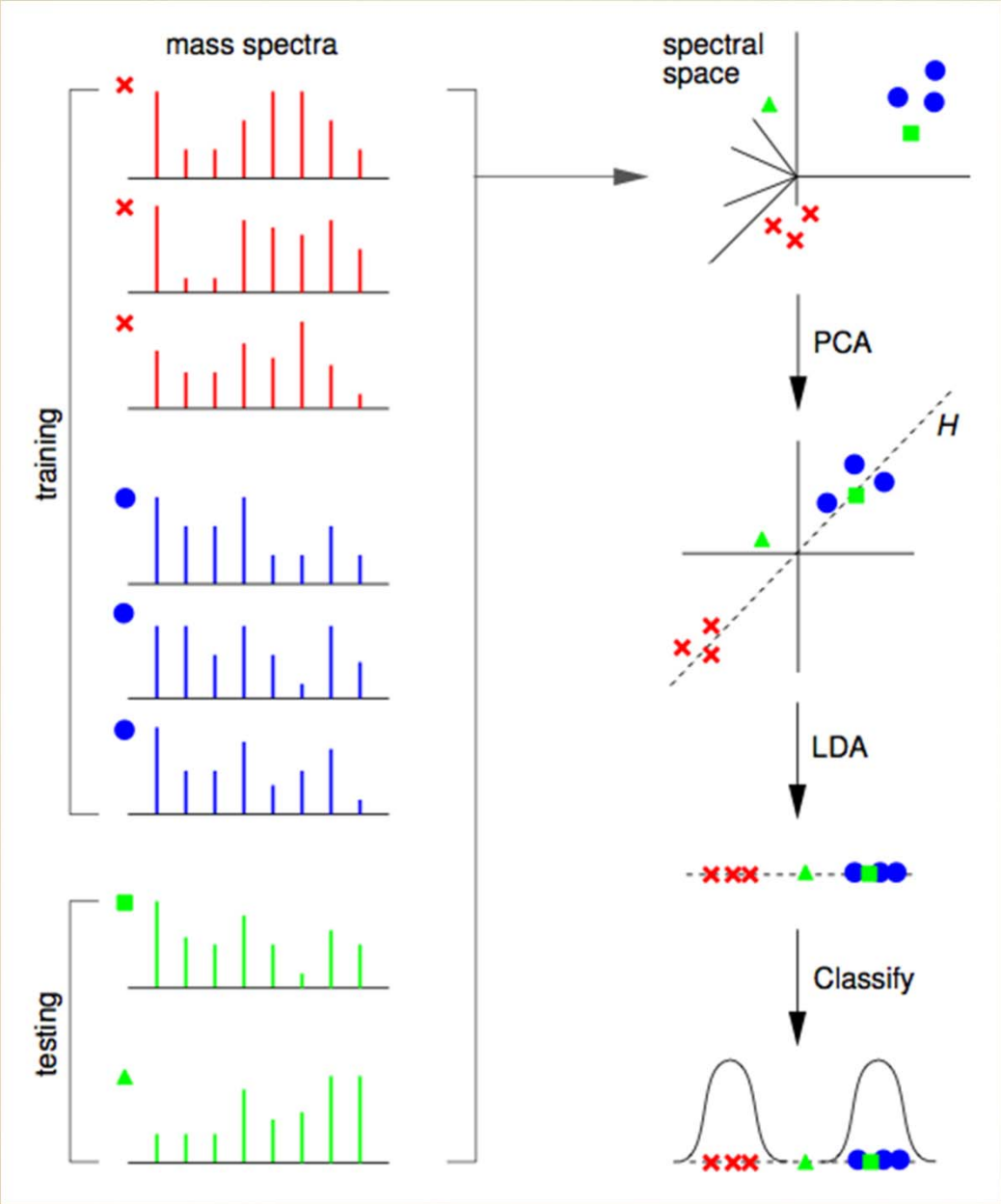


In Higher Dimensions





Curse of Dimensionality



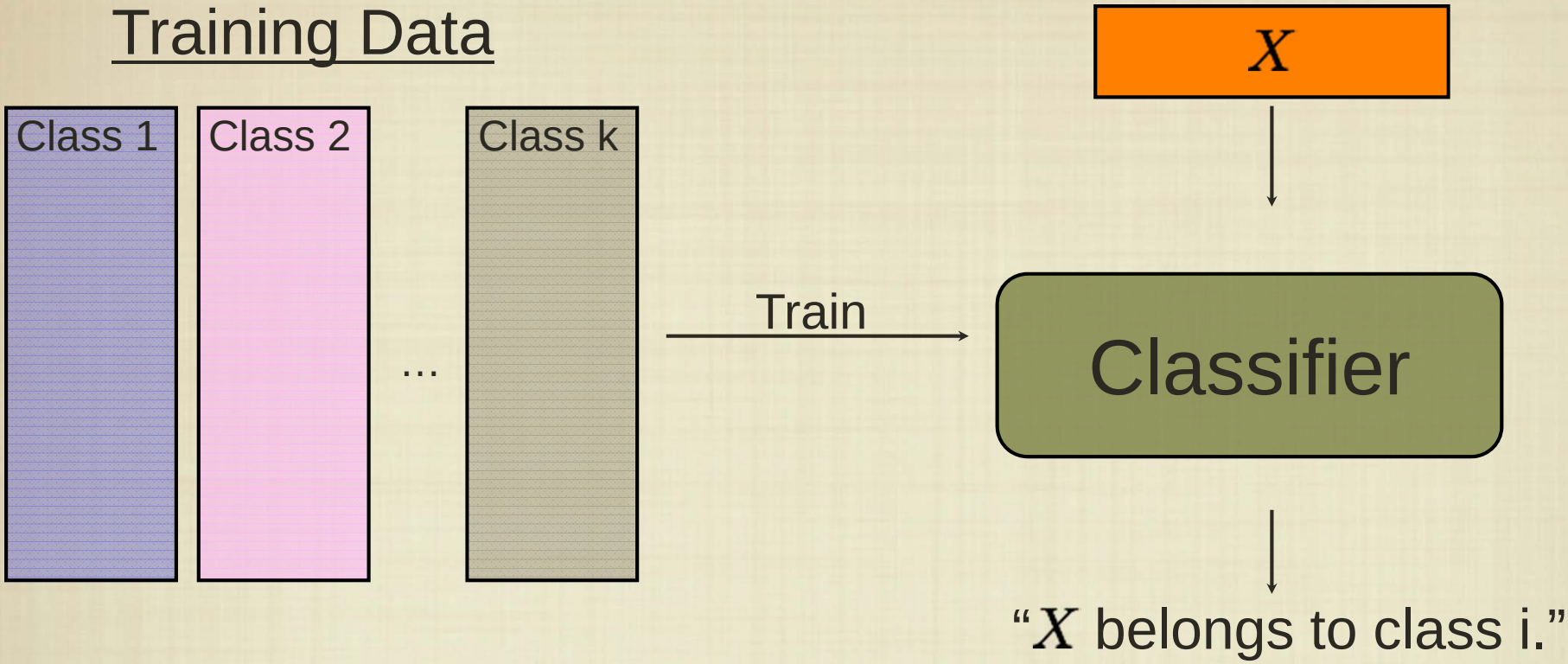
[Lilien et al. ’03]

In very high dimensions, even massive differences along one dimension may not be noticeable.

An active area of research is to design algorithms for “feature selection” to improve classification accuracy.

A by-product of feature selection is the identification of biomarkers, which could also lead to therapies.

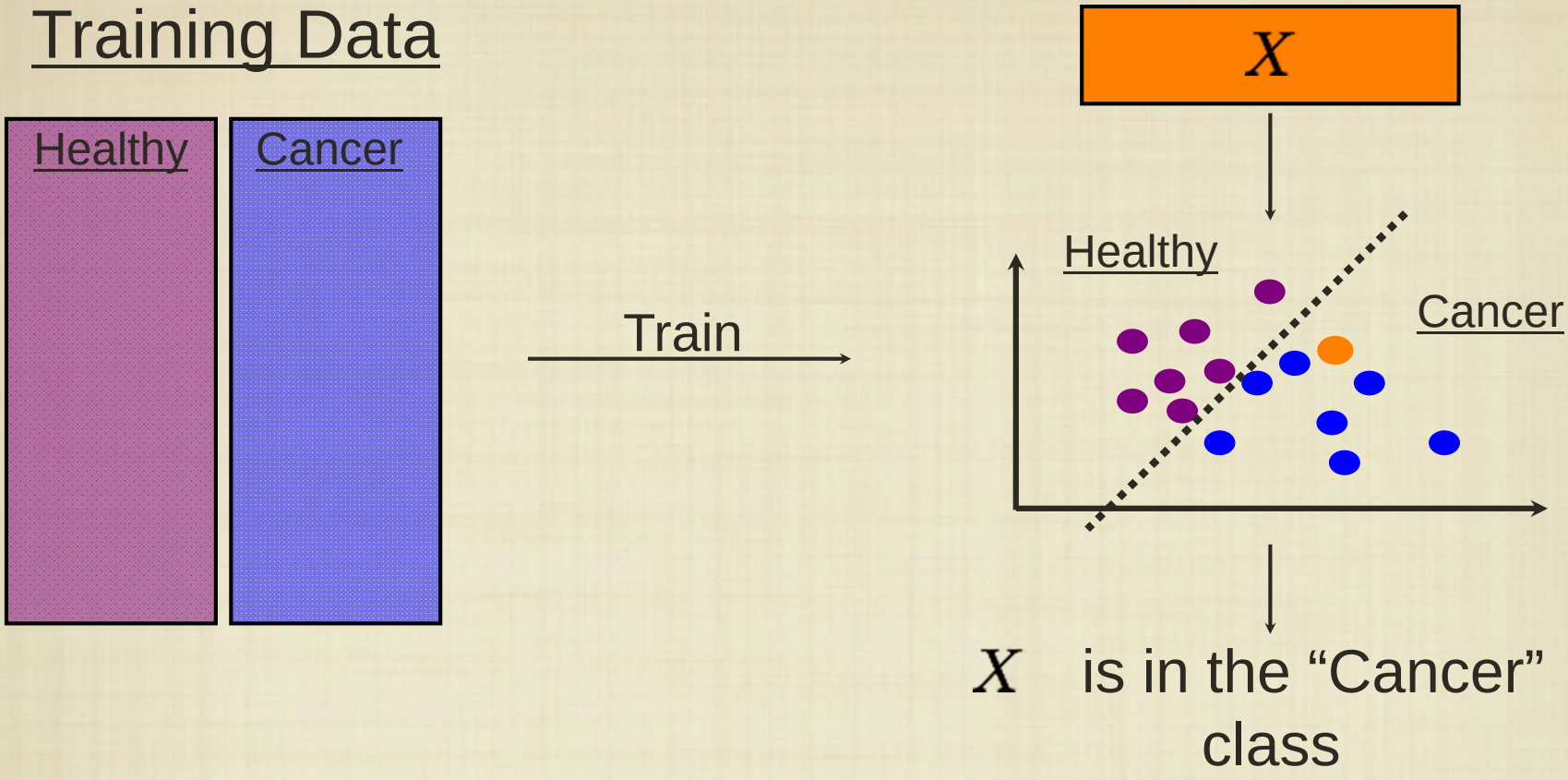
“Supervised” Learning



In supervised machine learning, we take data from known categories to develop a classifier. The classifier essentially summarizes everything we have learned.

Given new data, the classifier is used to determine which category it belongs to.

“Supervised” Learning



We had two classes for which we had mass spectra of blood serum, and “trained” by mapping each spectrum to a high-dimensional space.

After computing a surface to geometrically separate the two classes, we have a way to classify any new data.