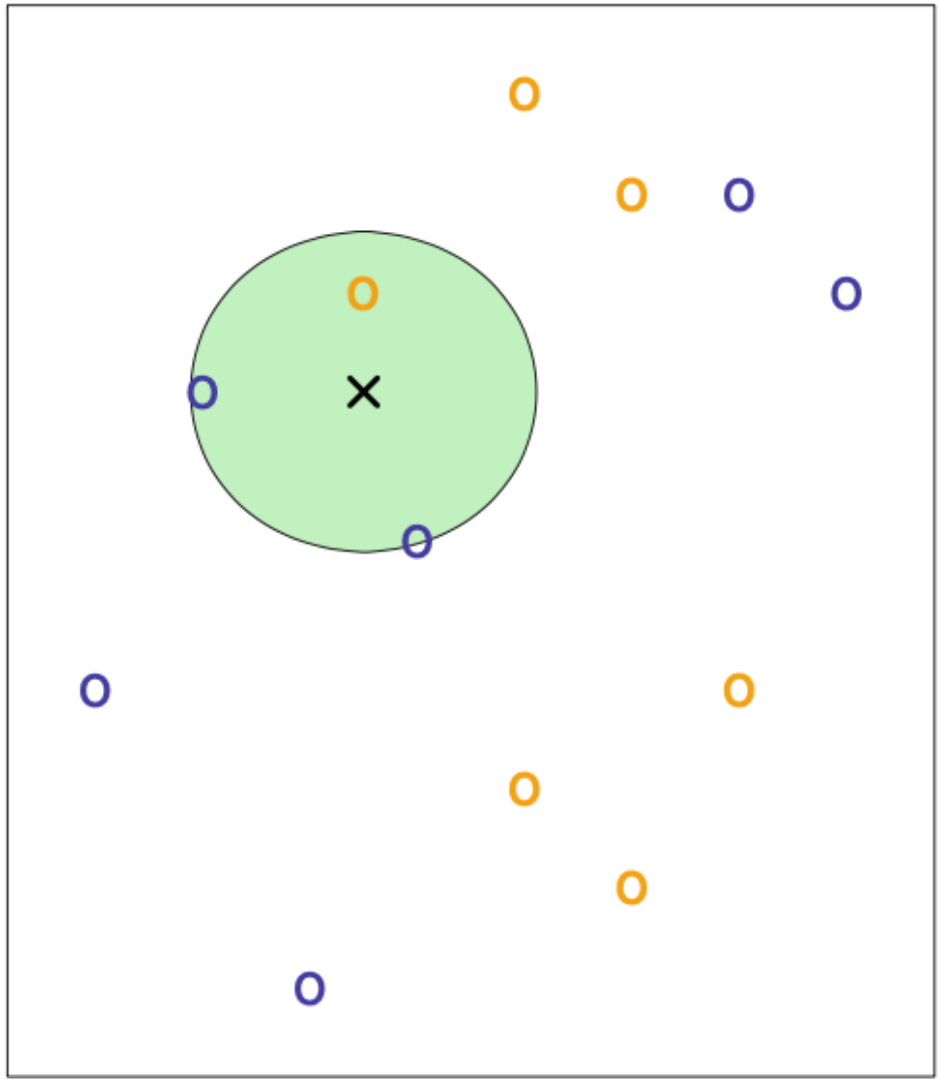
DATA221 Intro Machine Learning
09 K-nearest-neighbor classifier
The curse of dimensionality

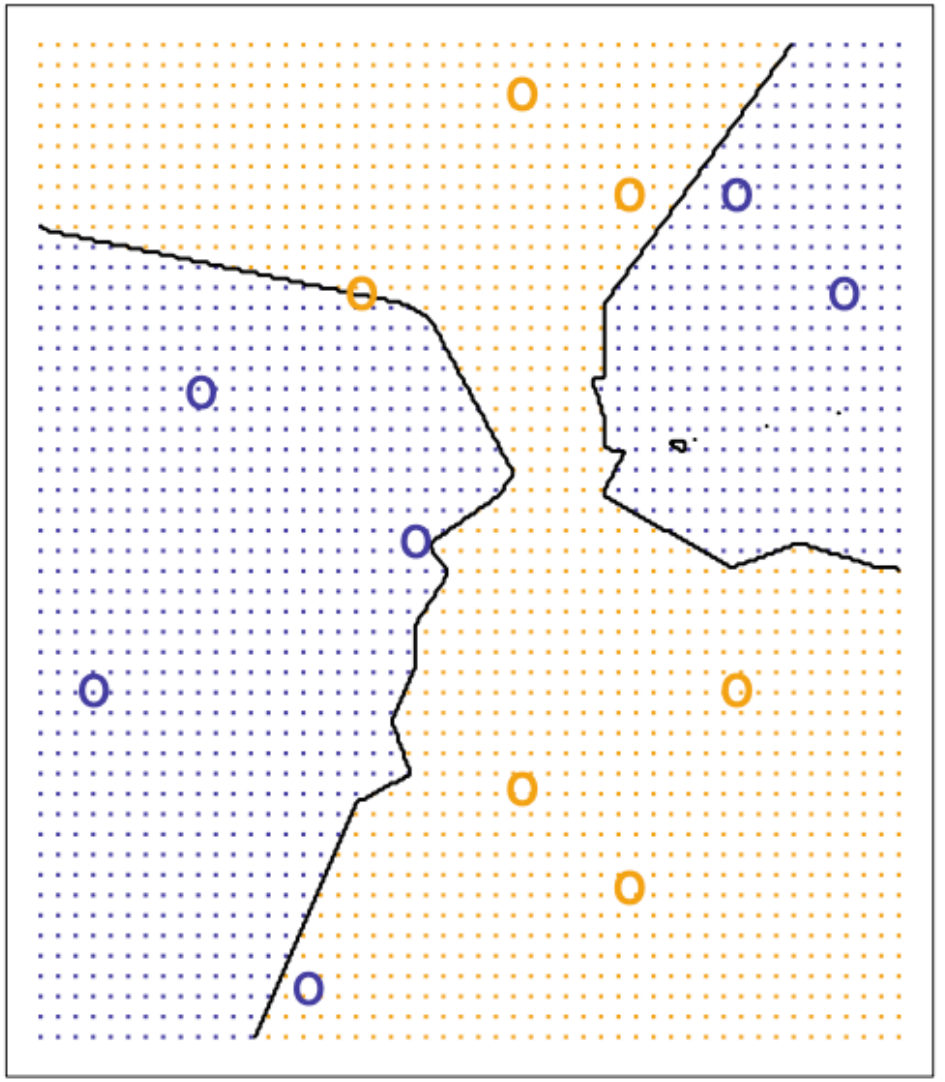
William Trimble
Winter 2023

THE UNIVERSITY OF CHICAGO

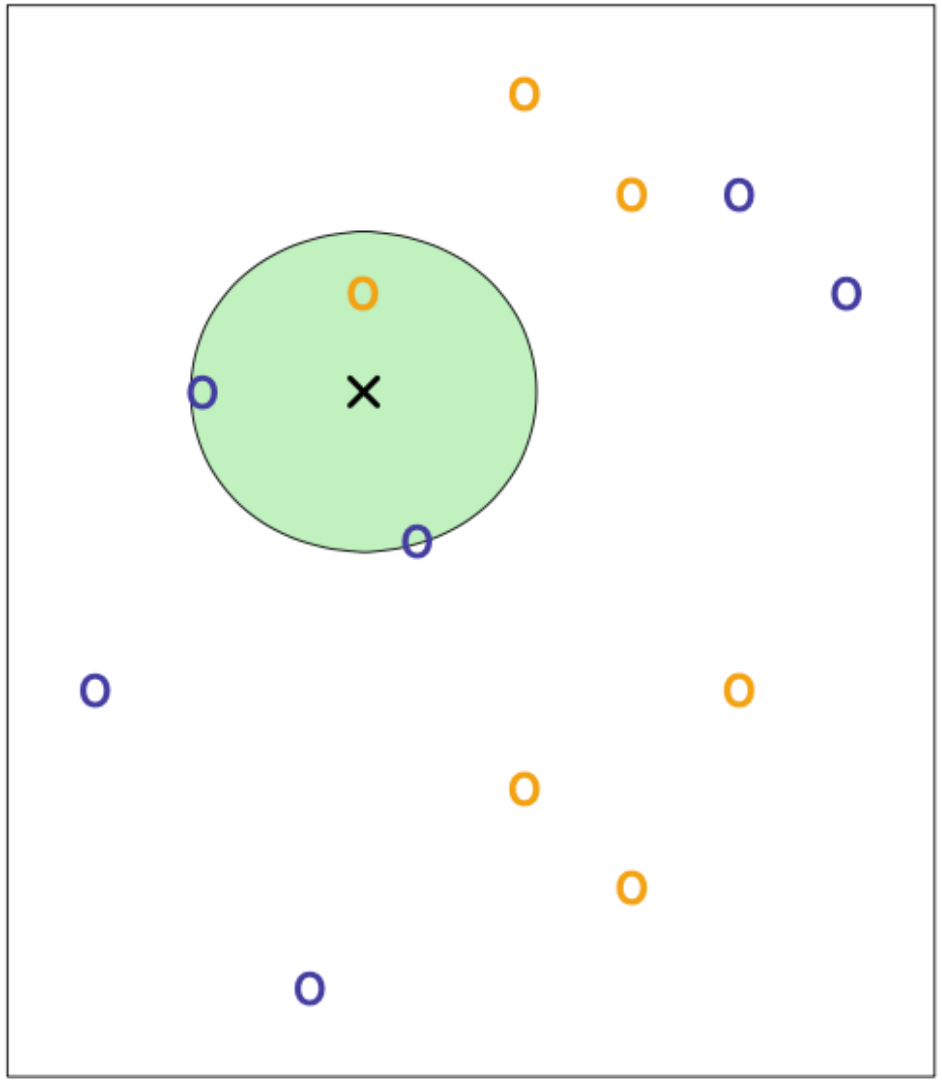
3 similar techniques…

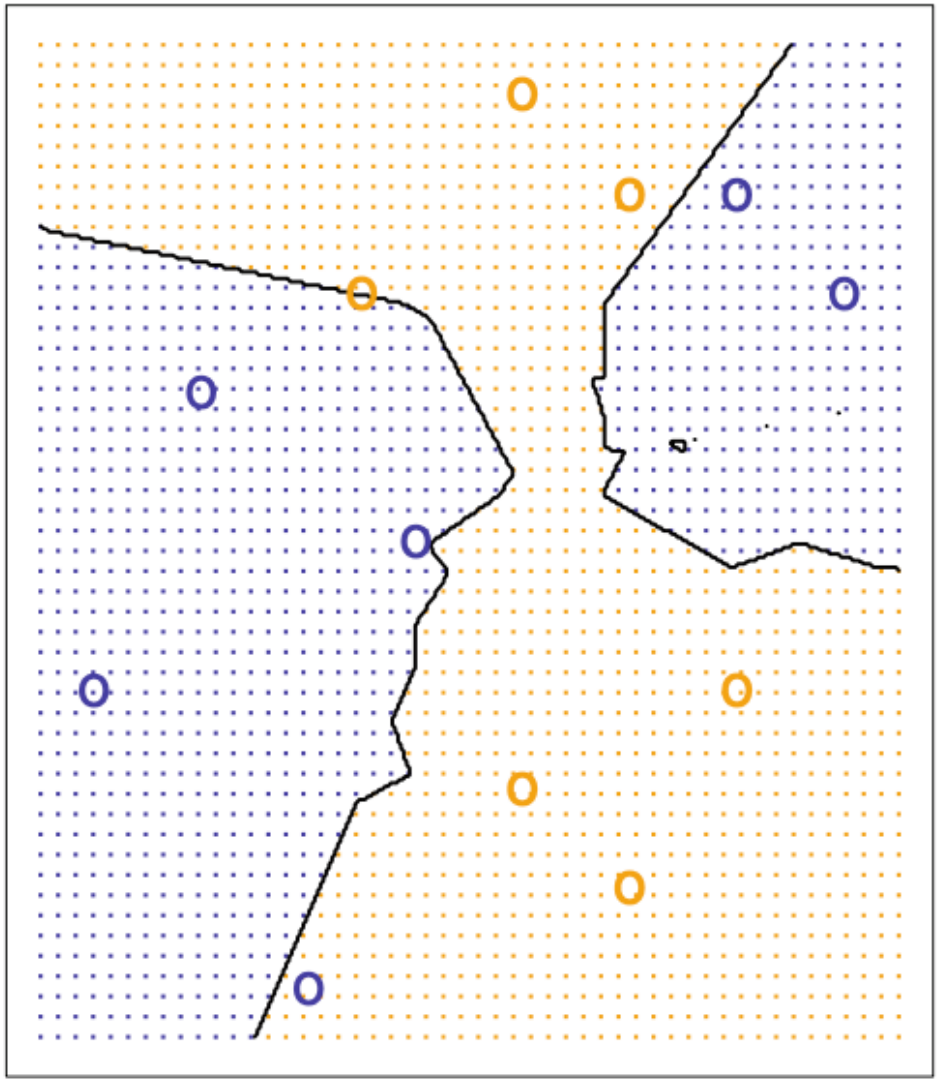
- knn classification
- knn regression
- knn clustering (of which k-means is one algorithm)





KNN classification : what do you need to be able to do to pull this off?





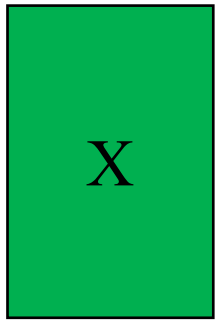
KNN classification : What does the green circle mean?

Features and labels as matrices…

r features

n labels

n examples



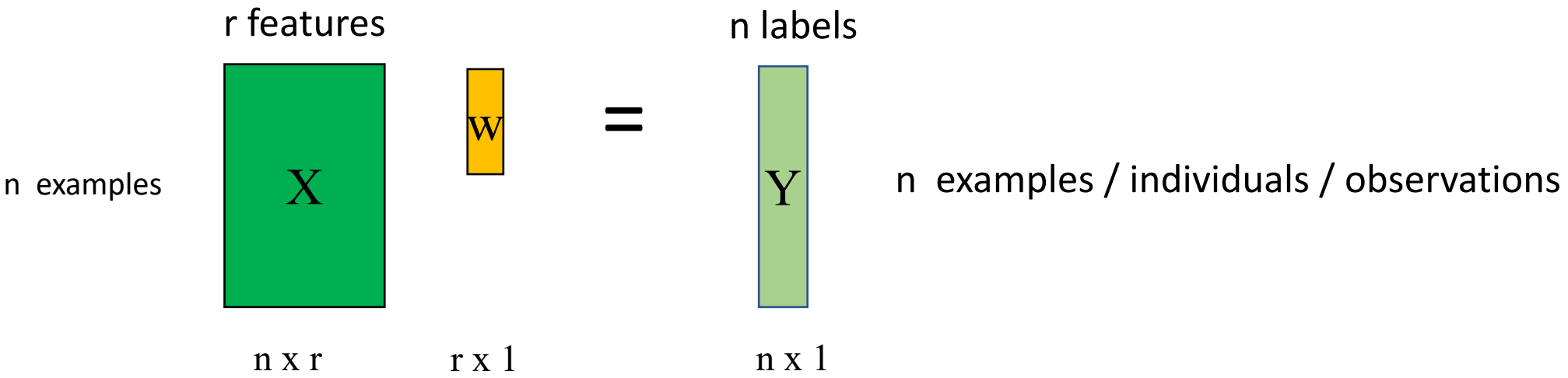
n examples / individuals / observations

n x r

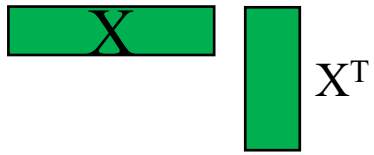
n x 1

This is the form a lot of our library functions will expect.

Features and labels as matrices…



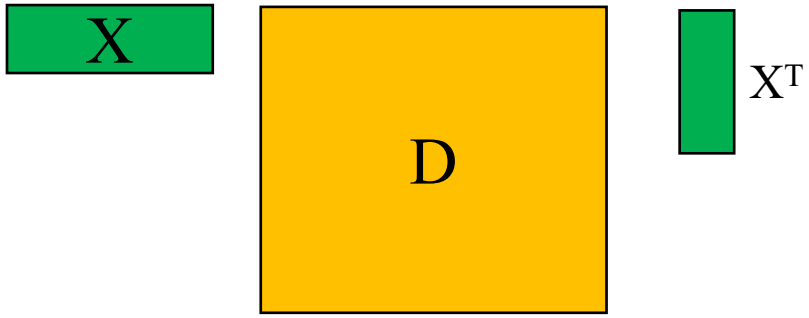
Distance matrices



$= |X|^2$

Generalization of L2-norm
($X^TX = \Sigma x_i^2$)
that allows different components of
x different weights.

r features



1 example

$= |X|_D^2$

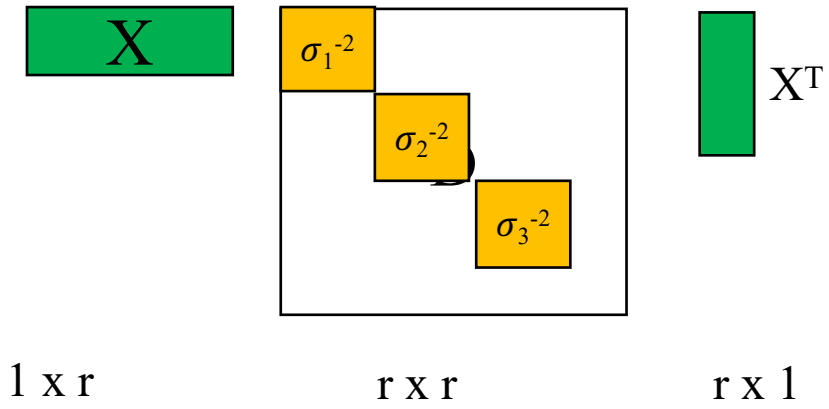
1 x r r x r r x 1

1 x 1

Distance matrices

r features

1 example



If the distance metric matrix is diagonal, this is just a weighted sum:

$$X\,D\,X^T = \Sigma\, x_i^2 \sigma_i^{-2}$$

KNN

- Always put attention on the distance metric.
- Gets more accurate with increasing n (at constant n/N) but becomes very slow.
- Tends not to work well in high numbers of dimensions. Why? Density of training points in high-dimensional space is low; will tend not to have examples that are alike enough.
- Prototypical high-dimensional space is bag-of-words. 2,000-5,000 parameters for starters. (Spam classifier)
- We didn't have to fit those, though! We used empirical means (easy to calculate, didn't have to run optimizer) that were maximum-likelihood estimates of the parameters. Not all problems are so easy.

“The flaw of averages”

- In the late 1940s, pilots were crashing planes at rates that were concerning to the US Air Force.
- Aviation cockpits were based on measurements from 1926.. was it possible that pilots 20 years later were bigger?
- Gilbert Daniels measures 4063 pilots in 10 aviation-relevant dimensions. Roughly normal distribution on each one, right?

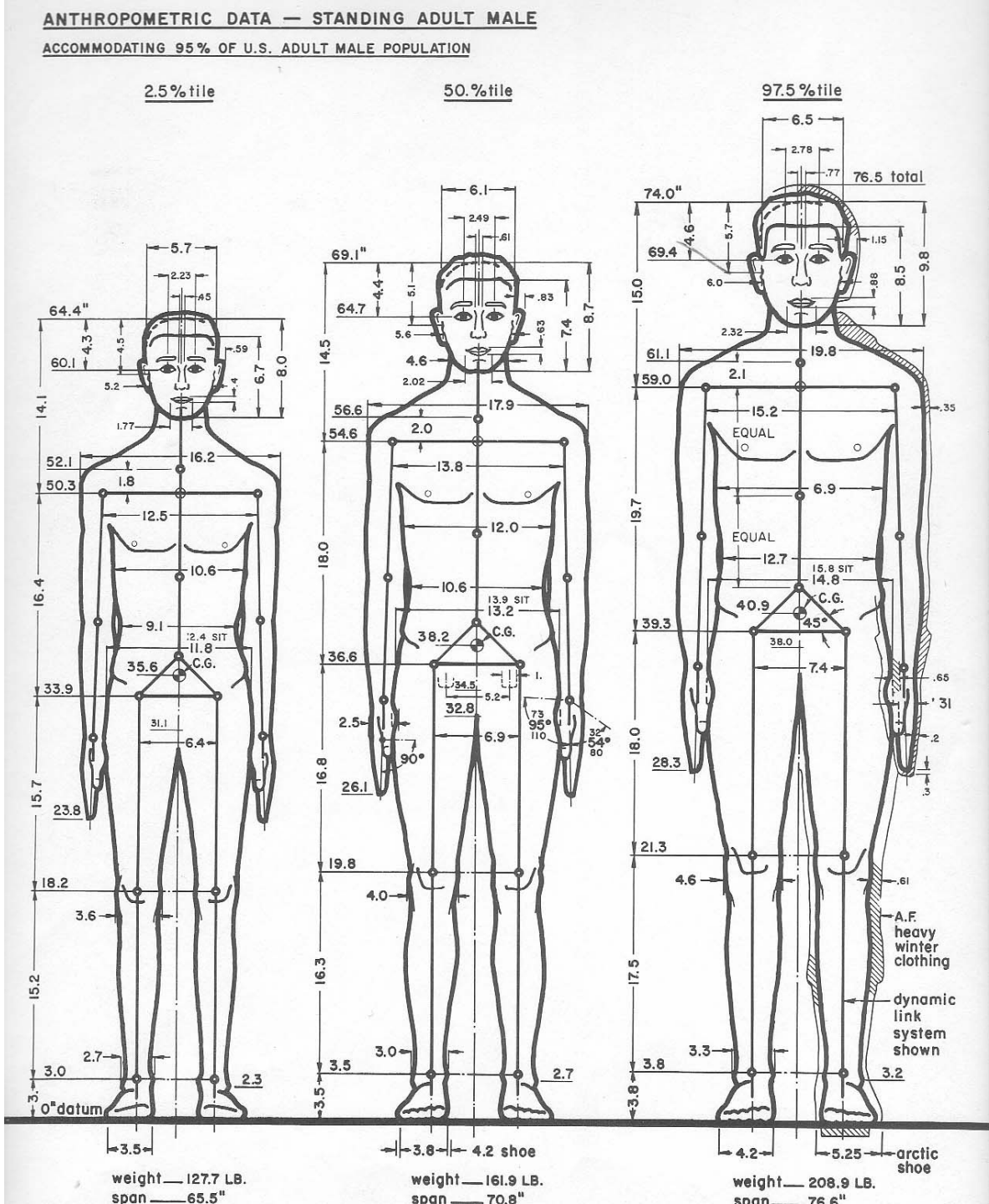
https://www.thestar.com/news/insight/2016/01/16/when-us-air-force-discovered-the-flaw-of-averages.html

How many pilots really were average?

Given a set of measurements (relevant for clothing design or cockpit design), there is a distribution of measurements in each dimension.

How many pilots are close to the average in all dimensions?

Gilbert Daniels, ‘The “Average Man”?’ Air force technical report
https://apps.dtic.mil/sti/citations/AD0010203



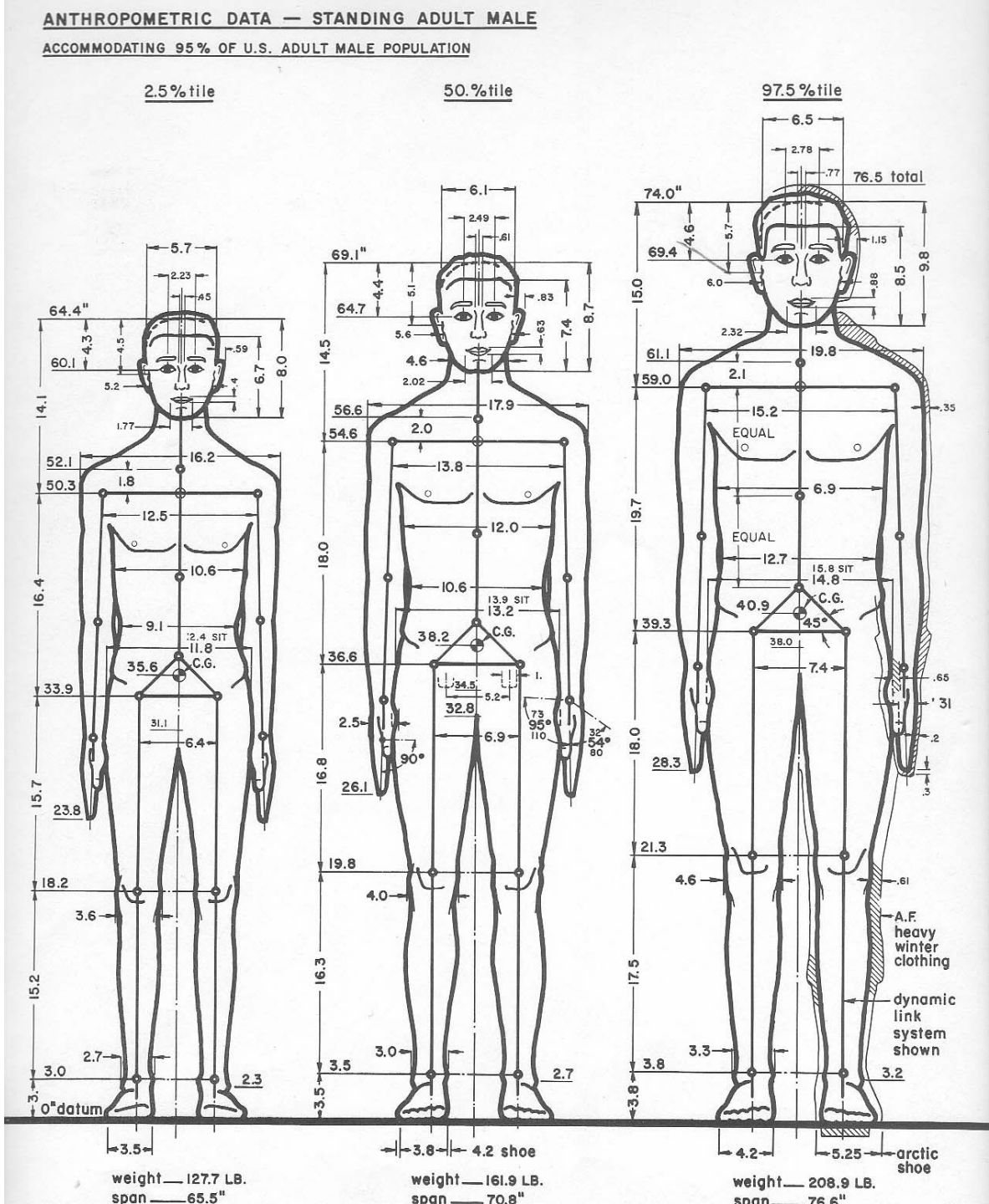
How many pilots really were average?

Take plus or minus 0.3 standard deviations in each dimension to be “approximately average” – this is about the middle third.

Of the original 4063 men, 1055 were of approximately average stature.
of these 1055 men, 302 were also of approximately average chest circumference.
of these 302 men, 143 were also of approximately average sleeve length.
of thsee 143 men, 73 were also of approximately average crotch height.
of these 73 men, 28 were also of approximately averages torso circumference
of these 28 men, 12 were also of approximately average hip circumferences.
of these 12 men, 6 were also of approximately average neck circumference.
of thsee 6 men, 3 were also of approximately average waist circumference.
of these 3 men, 2 were also of approximately average thigh circumference.
of these 2 men, 0 were approximately average in crotch length.

Huh. Requiring close-to-average on many dimensions becomes an impossible selection problem.

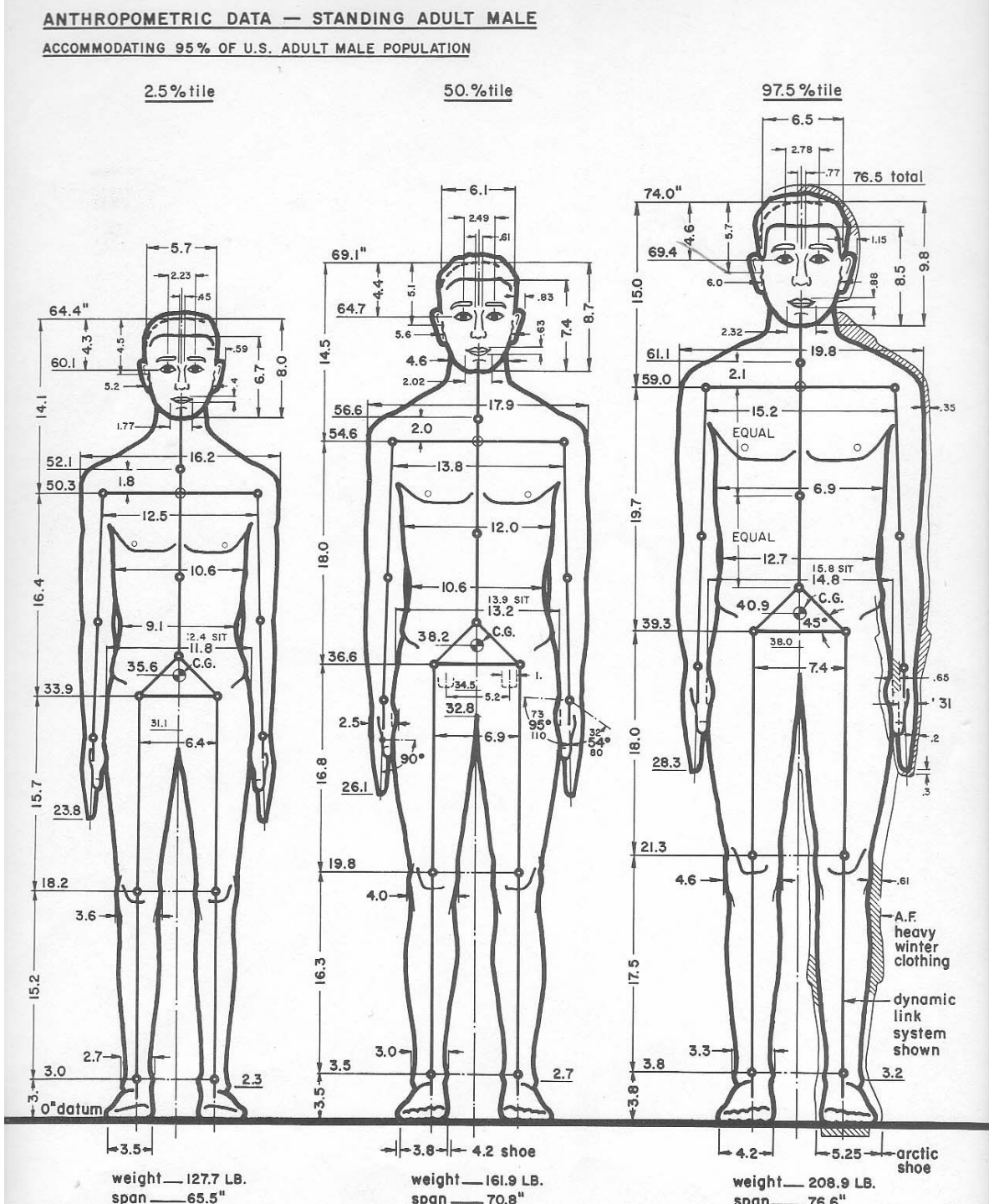
Gilbert Daniels, ‘The “Average Man”?’ Air force technical report
https://apps.dtic.mil/sti/citations/AD0010203



How many pilots really were average?

"Out of 4,063 pilots, not a single airman fit within the average range on all 10 dimensions. One pilot might have a longer-than-average arm length, but a shorter-than-average leg length. Another pilot might have a big chest but small hips. Even more astonishing, Daniels discovered that if you picked out just three of the ten dimensions of size — say, neck circumference, thigh circumference and wrist circumference — less than 3.5 per cent of pilots would be average sized on all three dimensions. Daniels's findings were clear and incontrovertible. *There was no such thing as an average pilot.* If you've designed a cockpit to fit the average pilot, you've actually designed it to fit no one."

Todd Rose The End of Average (2016)
https://www.thestar.com/news/insight/2016/01/16/when-us-air-force-discovered-the-flaw-of-averages.html

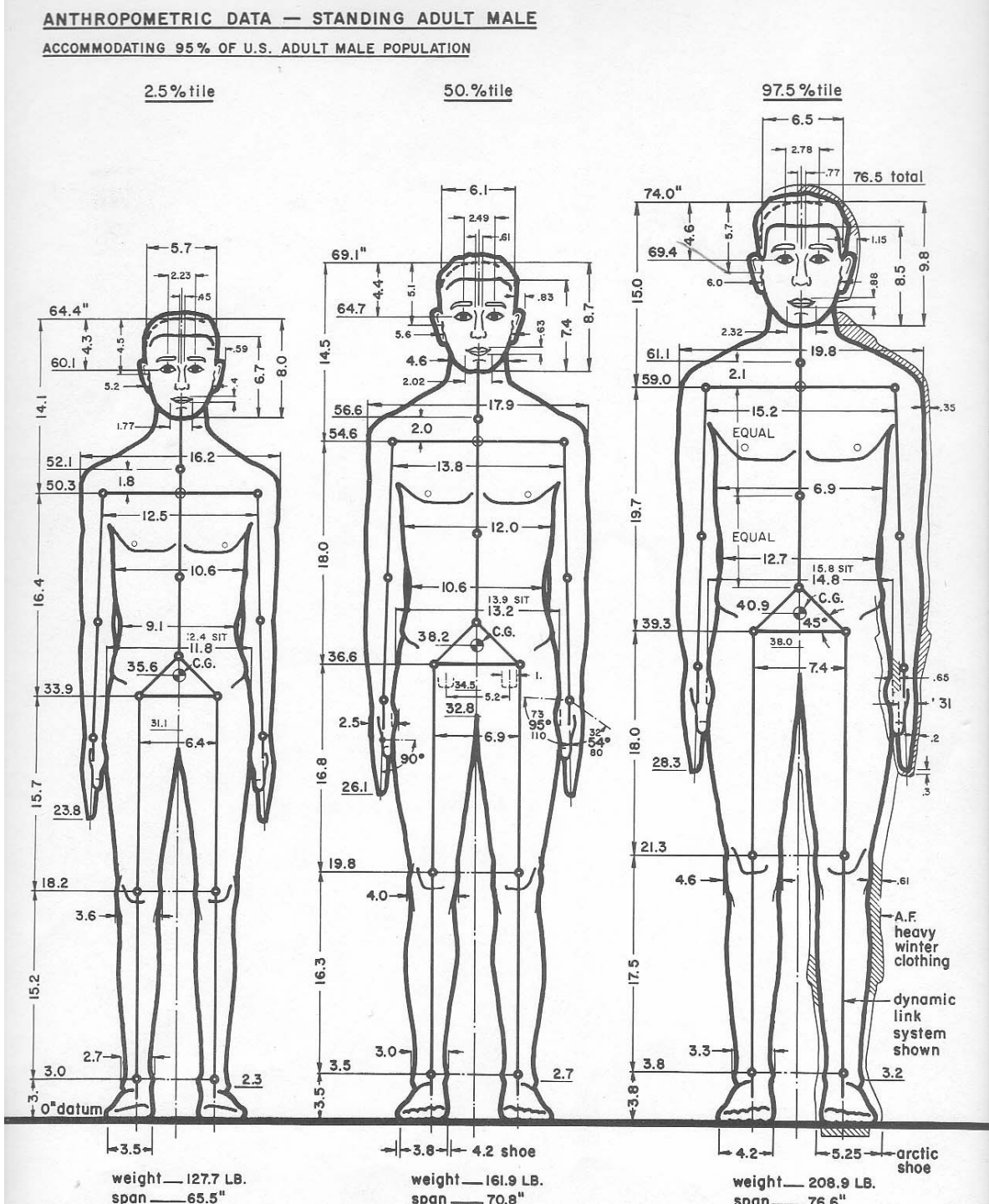


How many pilots really were average?

The historians tell us that after 1952, the military demanded that aviation contractors make the aircraft fit the pilots rather than selecting the pilots that fit well in the aircraft.

…but still only 9% of women in the Air Force in 2020 were of the right size to use the F-15 cockpit.

https://www.airforcetimes.com/news/your-air-force/2020/08/19/to-get-more-female-pilots-the-air-force-is-changing-the-way-it-designs-weapons/



Curse of dimensionality

- Things don't work in high numbers of dimensions the way we expect them to.
- Computing n 1st derivative components and keeping track of n(n+1) 2nd derivative components is not the worst of our troubles.

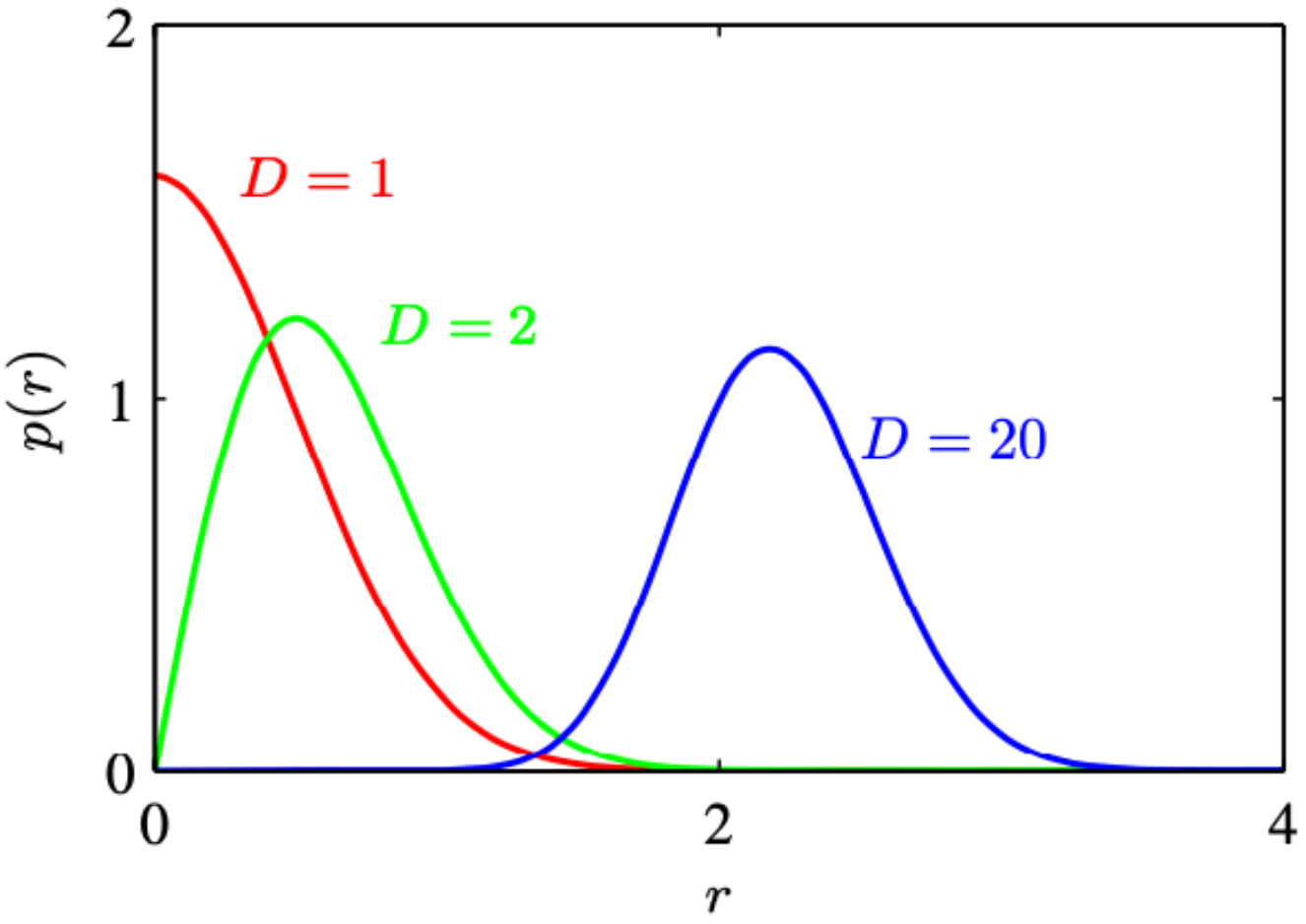
- Data become sparse in vast high-dimensional spaces (where machine learning parameters live)
- Limited data .. all linear combinations of N data points in $\mathbb{R}^D$ allow only vectors in an $\mathbb{R}^N$ subspace
- In high dimensions it's clear you don't have the data to do combinatorical tests. When $2^D >> N$ situation is hopeless and there are always dimensions you never sample from.

Curse of dimensionality…

- Picture a unit cube in D dimensions with N datapoints uniformly distributed in it.
- Average volume per point is $\frac{1^D}{N}$
- Distance between points scales like d = $\frac{1}{N^{1/D}}$
- Penguins? total number of dimensions ~42
- d = 1/350**(1/40) = 0.864

- To get just 1/350 of the data, I need a cube 0.864 on a side? This is a huge fraction of the span of each dimension.
- It will take a very large radius just to capture on average a handful of datapoints, and my nearest datapoints may not be very similar.

Multivariate normal, how bad can it be?



- The volume of a sphere in N dimensions is $\frac{\pi^{n/2} R^n}{\Gamma\left(\frac{n}{2}+1\right)}$
- Surface area is proportional to R^{n-1}
- MVN density centered at 0 in D dimensions, but the vast majority of the probability density is in a shell at the surface where the terms in $x^D \exp(-x^2)$ balance.