

DATA221 Intro Machine Learning

09 K-nearest-neighbor classifier

The curse of dimensionality

William Trimble
Spring 2023



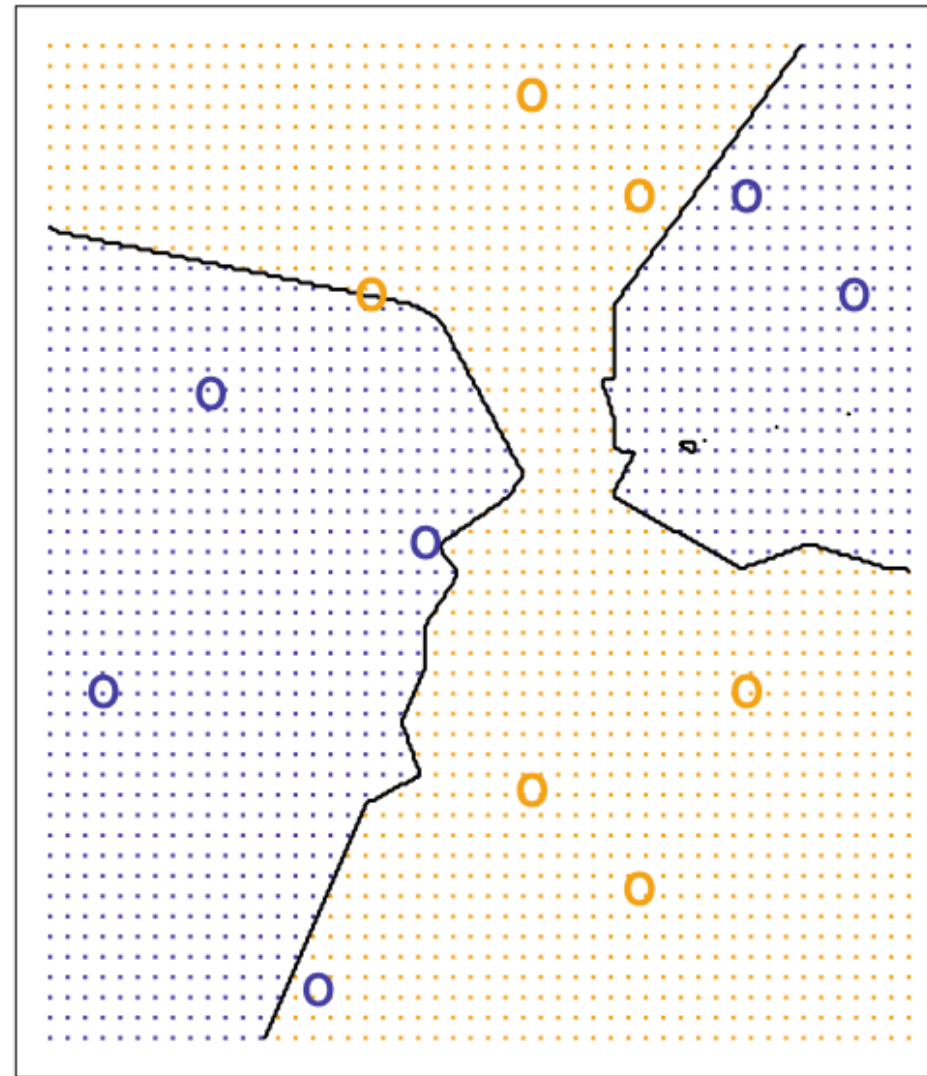
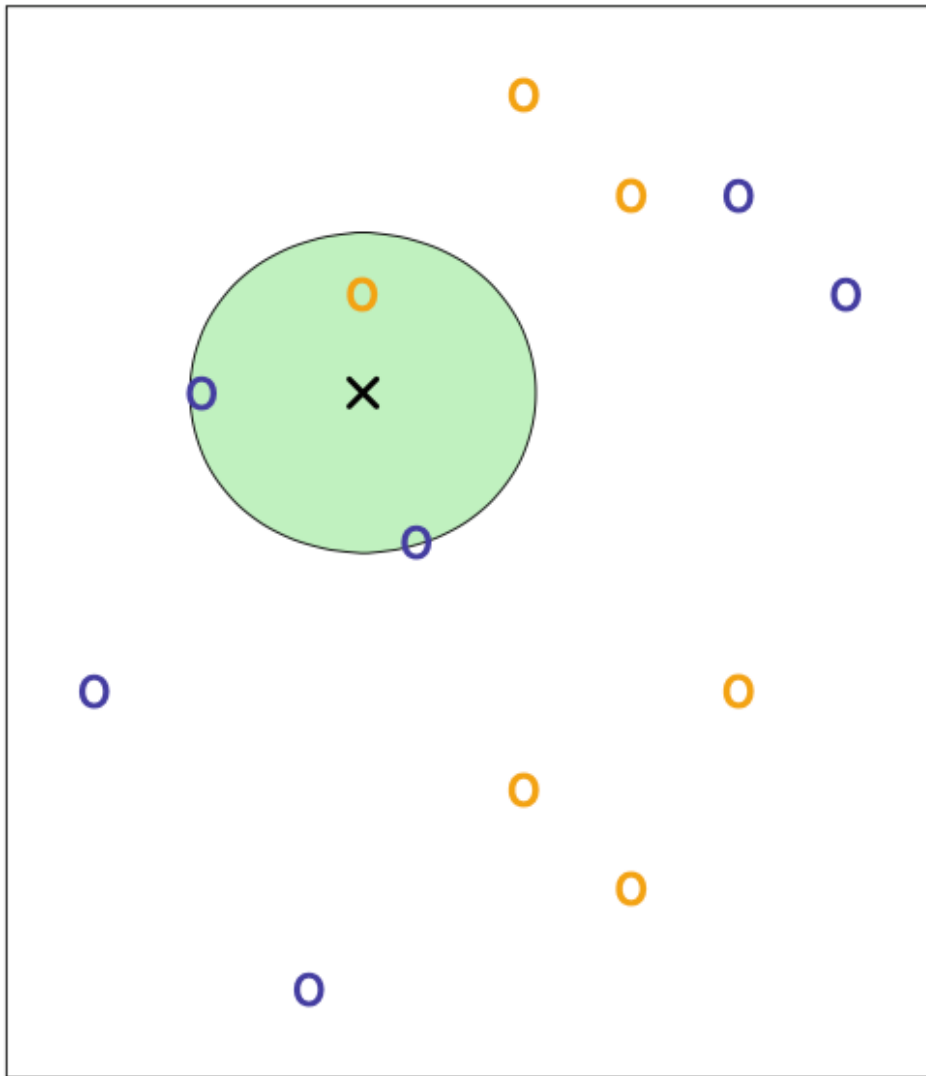
THE UNIVERSITY OF
CHICAGO

Testing / validation / training

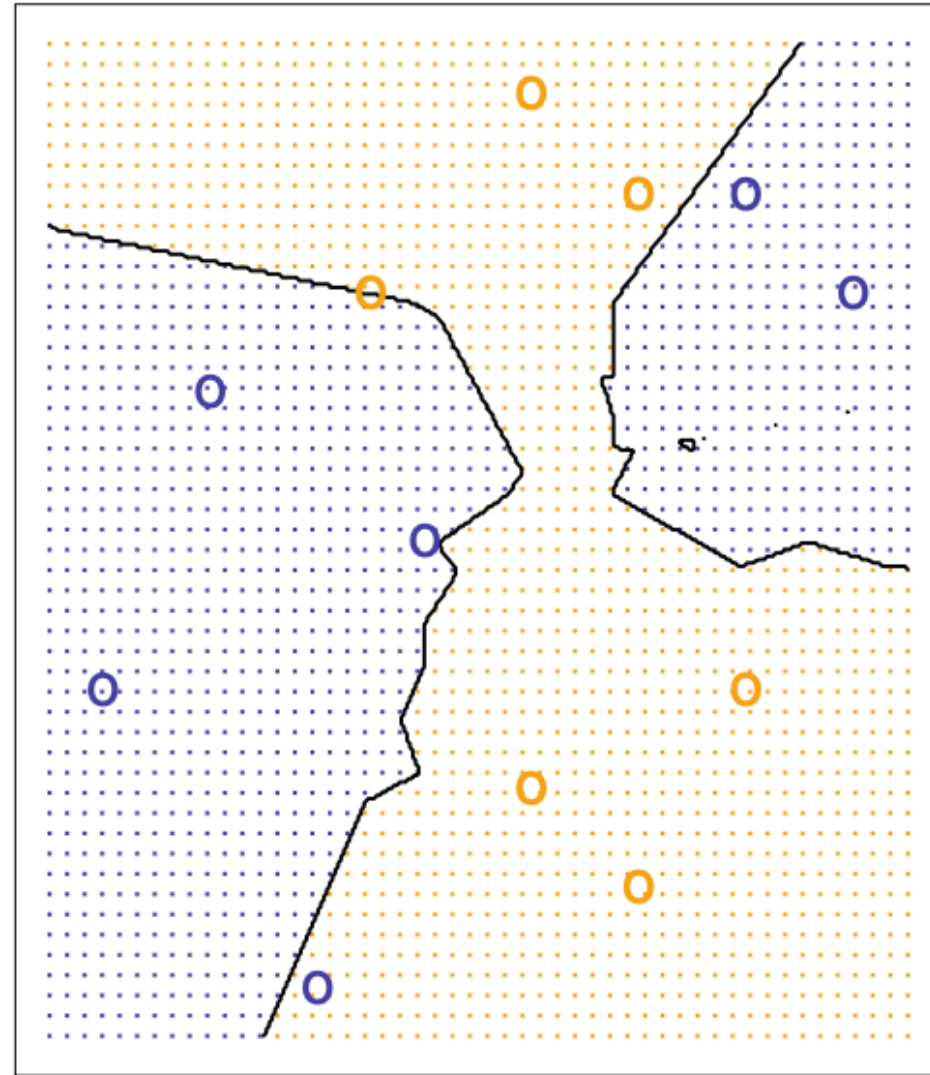
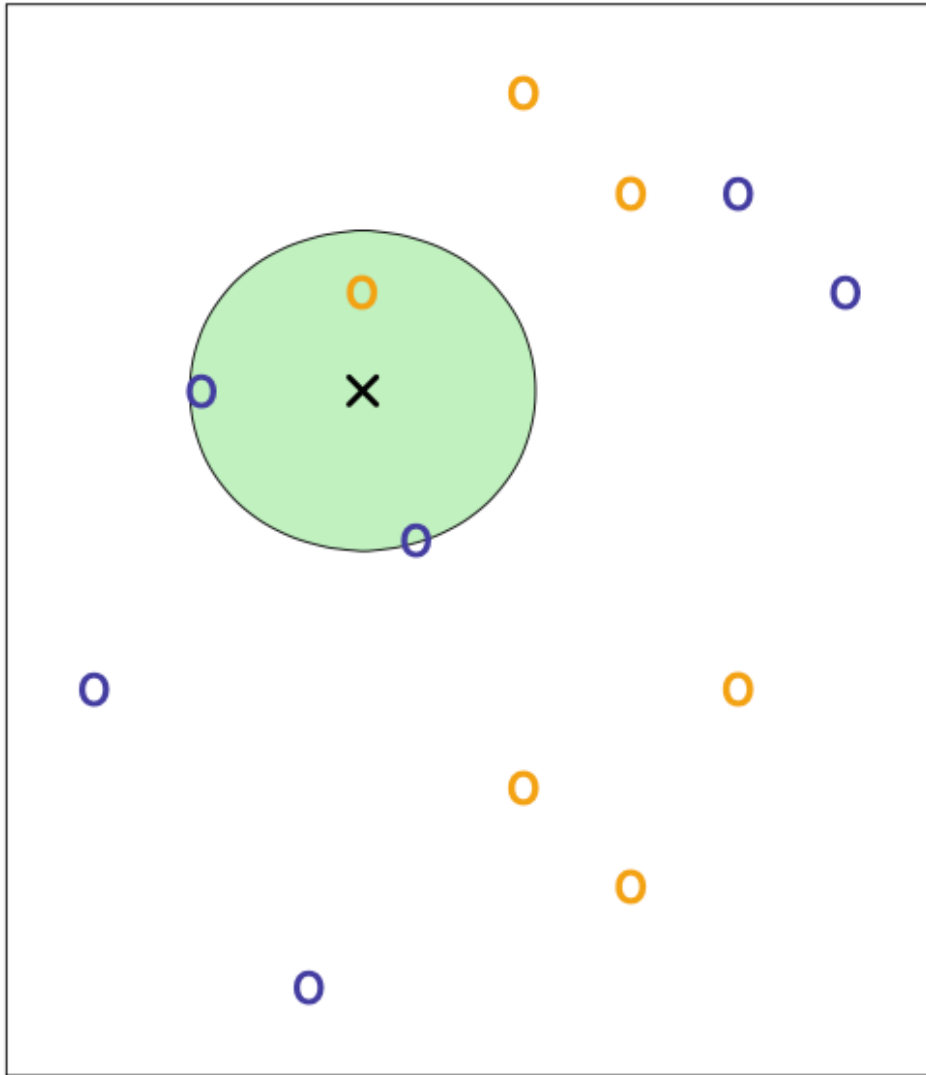
- More parameters always gives better fit.
- "Fits always look good"!!! (Fits are biased toward the data and away from the truth !!)
- Semantically:
 - Model selection: find the best model among a handful of alternatives
 - Model evaluation: estimate the error rate that the model will have on out-of-training-sample data
- The need for avoiding bias in model evaluation here gives rise to the
 - Training / Validation / Test set convention:
 - The validation set you can peek at the answers to tune hyperparameters (regularization strength, k)
 - **Can't use cross-validation both for hyperparameter tuning and evaluation.**

3 similar techniques...

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

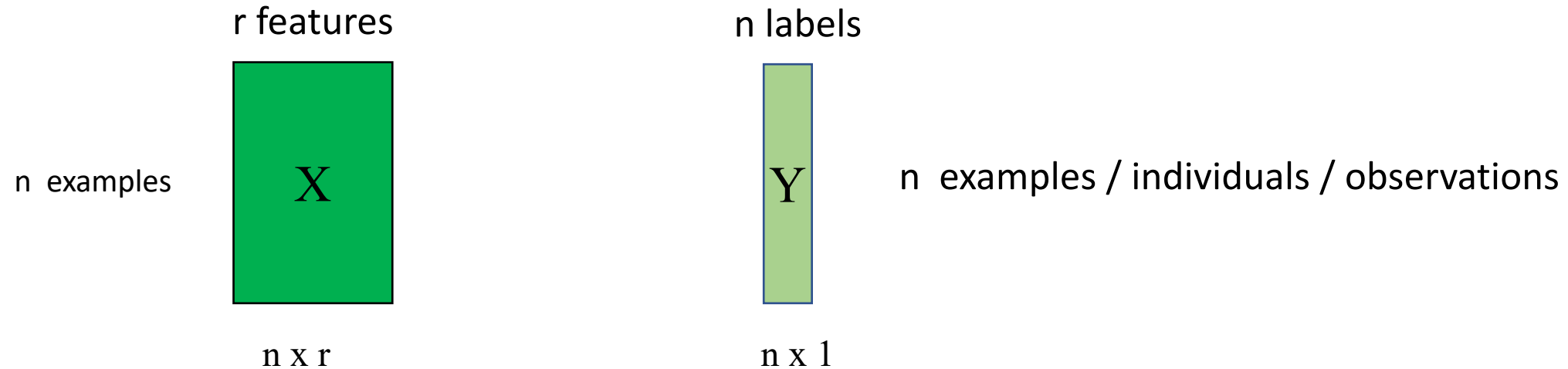


KNN classification : what do you need to be able to do to pull this off?



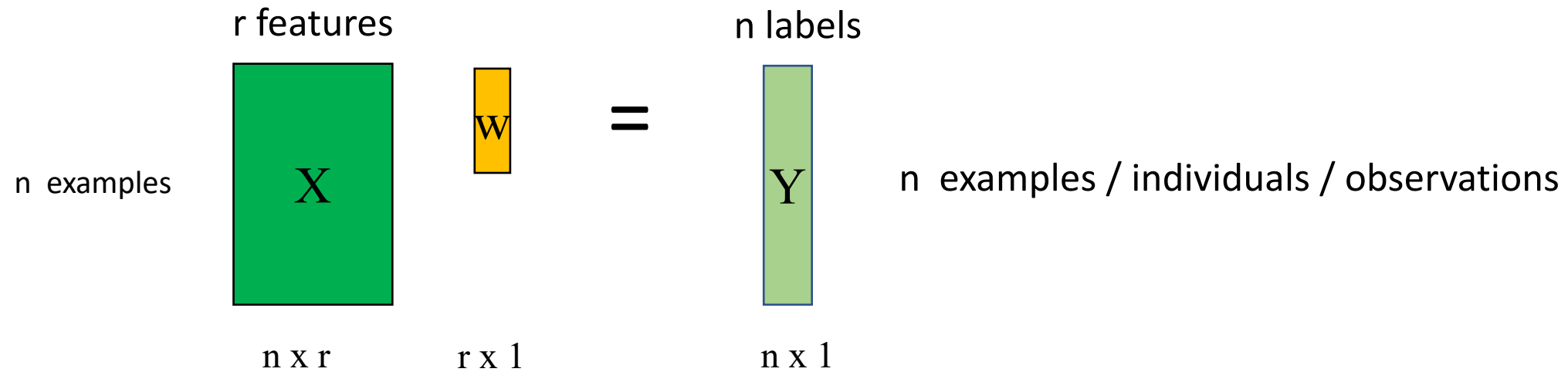
KNN classification : What does the green circle mean?

Features and labels as matrices...

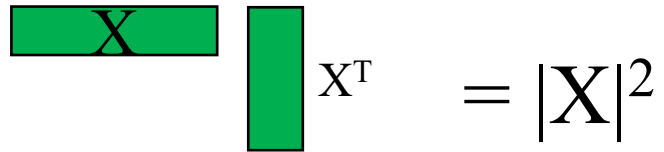


This is the form a lot of our library functions will expect.

Features and labels as matrices...

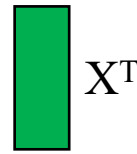
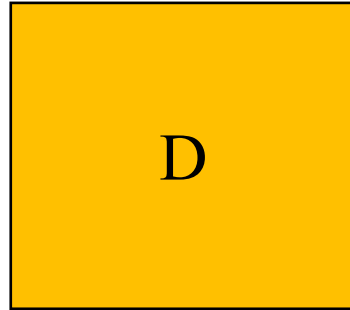


Distance matrices



$$\boxed{X} \boxed{X^T} = |X|^2$$

r features



1 example

1 x r

r x r

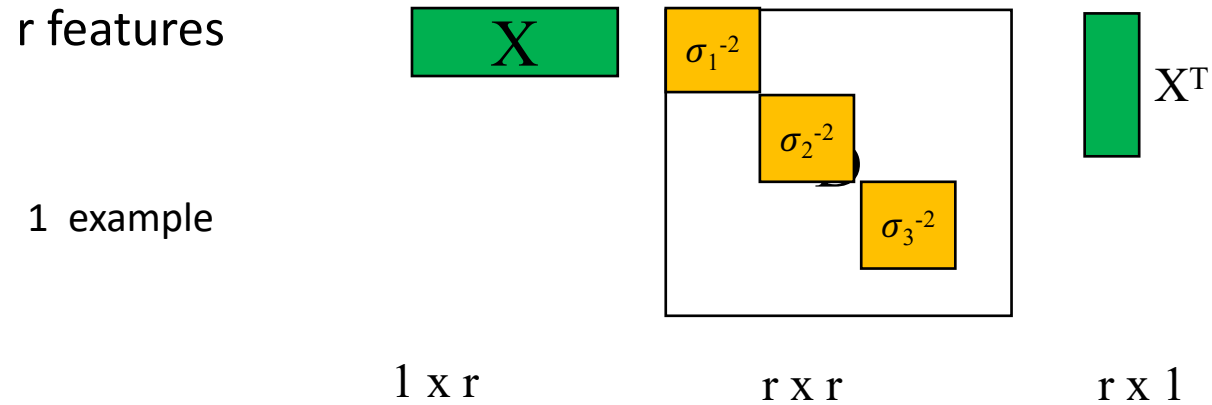
r x 1

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

$$= |X|_D^2$$

1 x 1

Distance matrices



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

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

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”

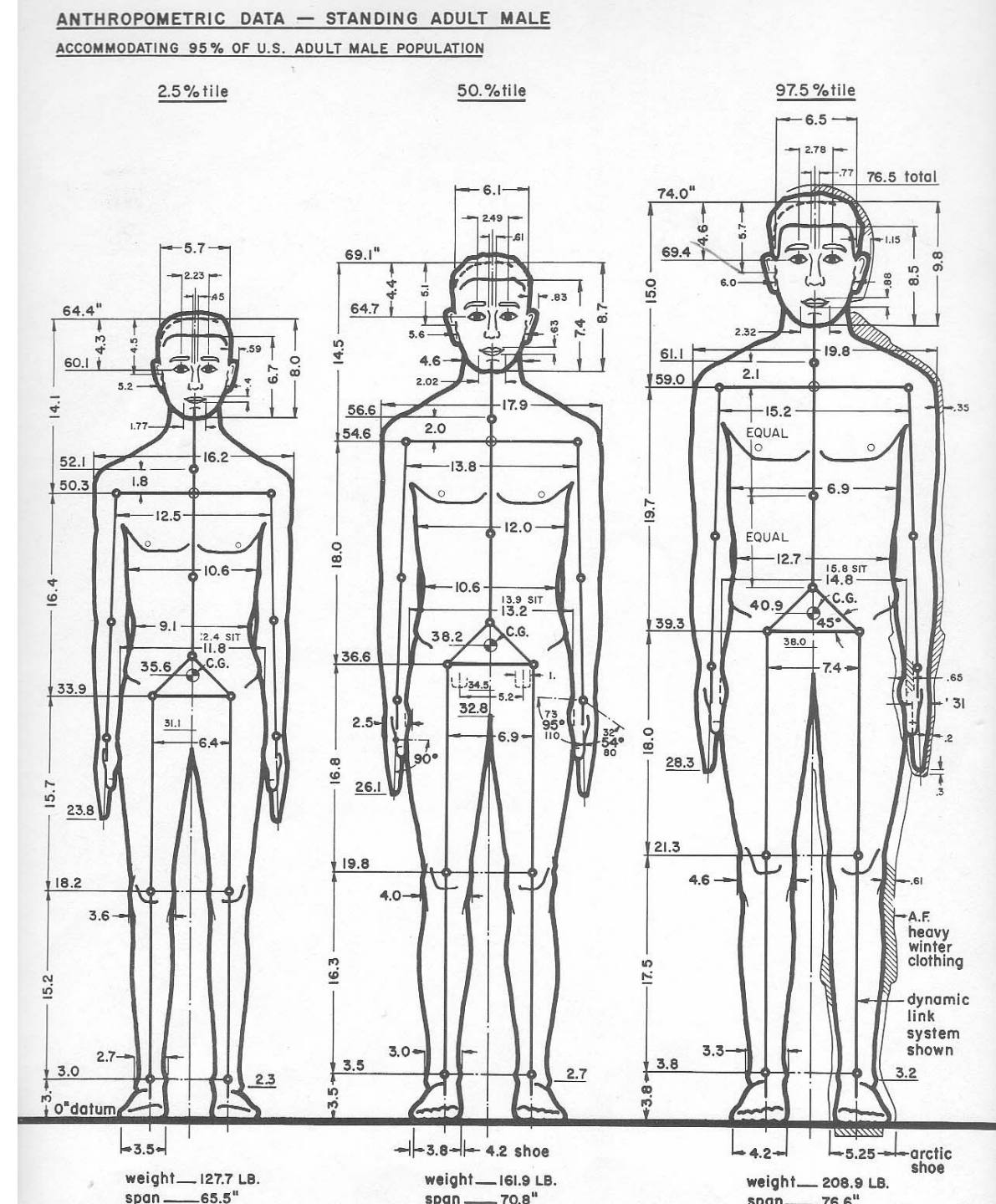
- In the late 1940s, pilots were crashing planes at rates that were concerning to the US Air Force.
- Aviation cockpits were based on measurements from 1926.. was it possible that pilots 20 years later were bigger?
- Gilbert Daniels measures 4063 pilots in 10 aviation-relevant sartorial dimensions. Roughly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

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>



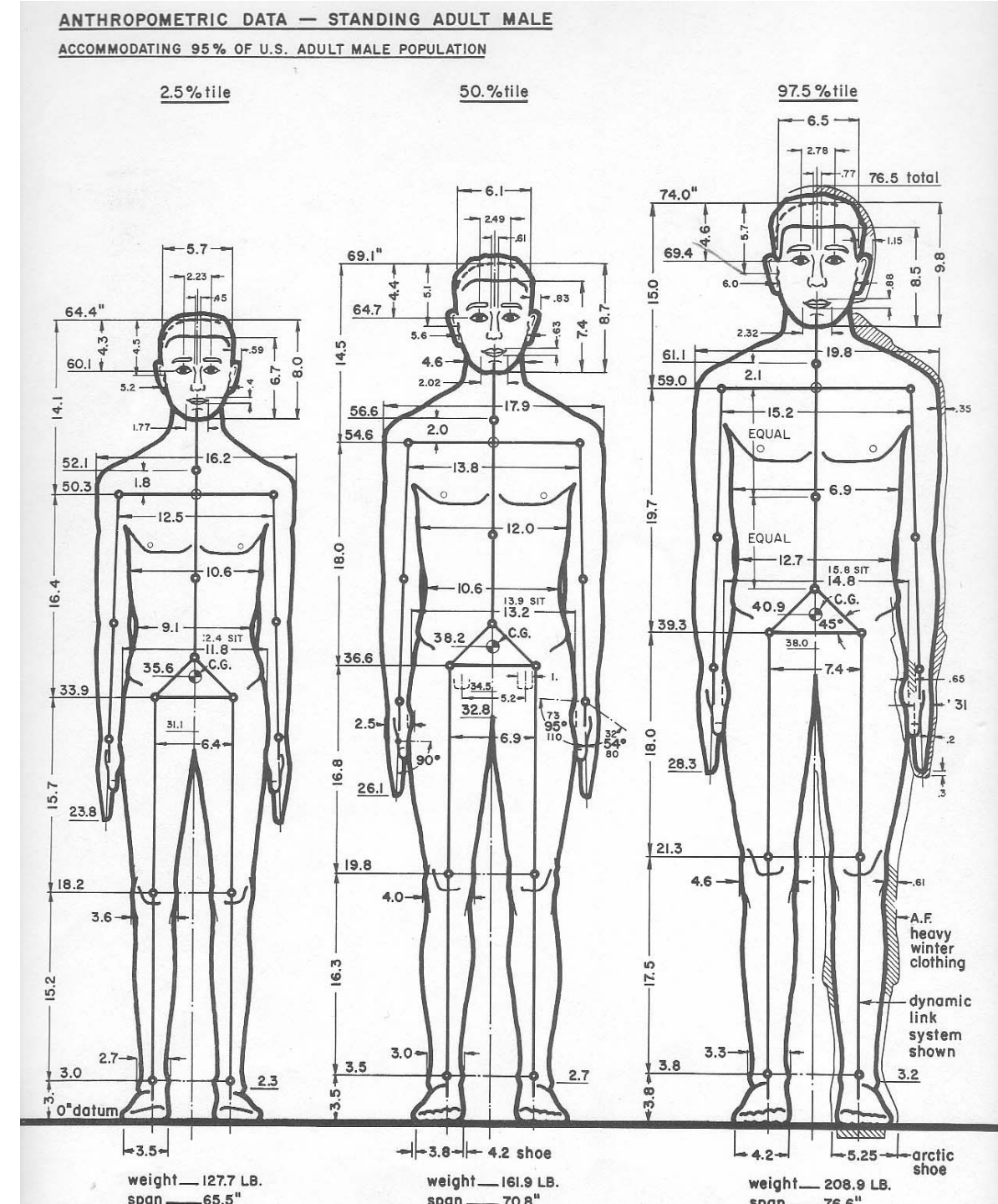
How many pilots really were average?

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 these 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 these 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.

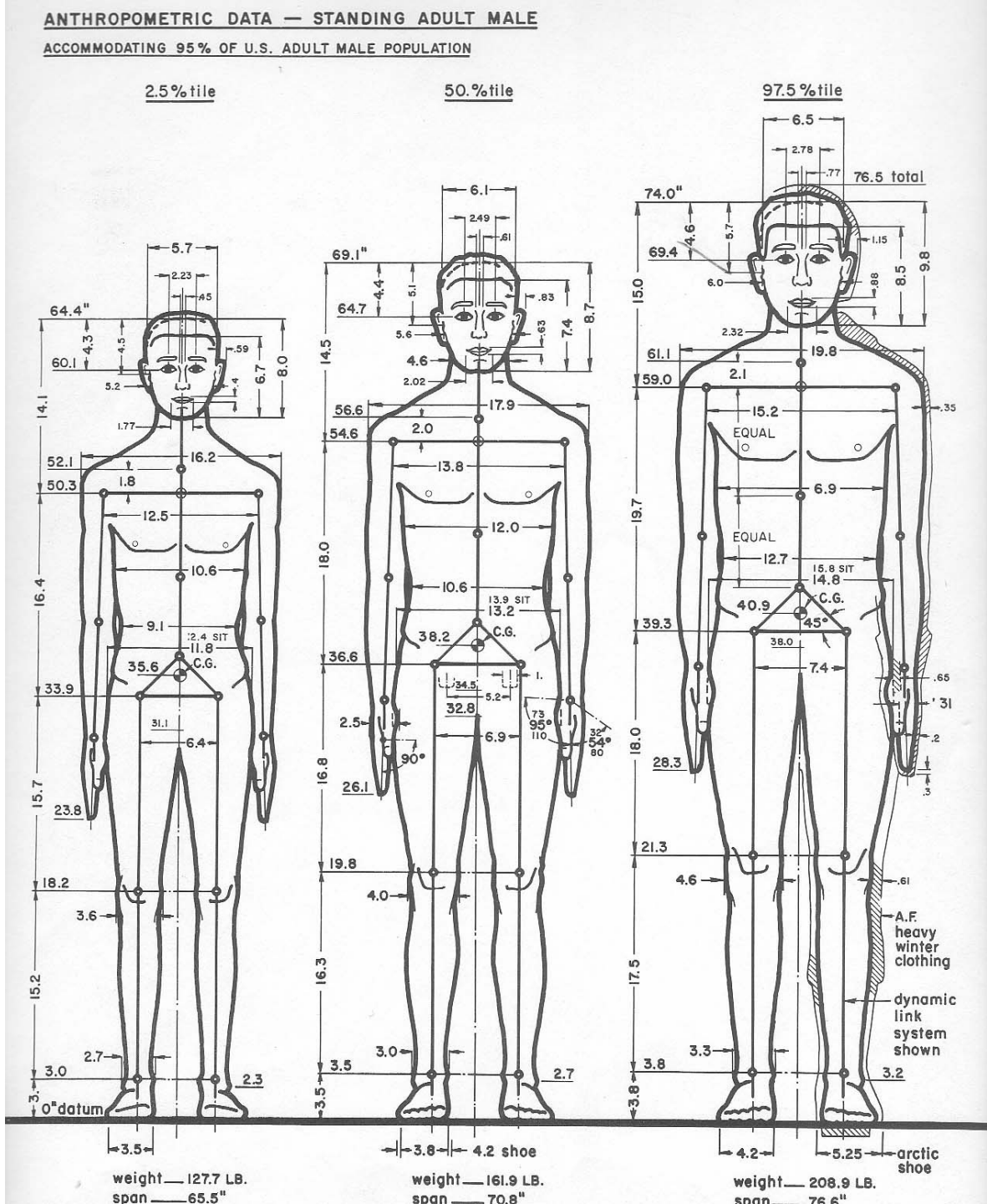
Gilbert Daniels, ‘The “Average Man”?’ Air force technical report
<https://apps.dtic.mil/sti/citations/AD0010203>



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>

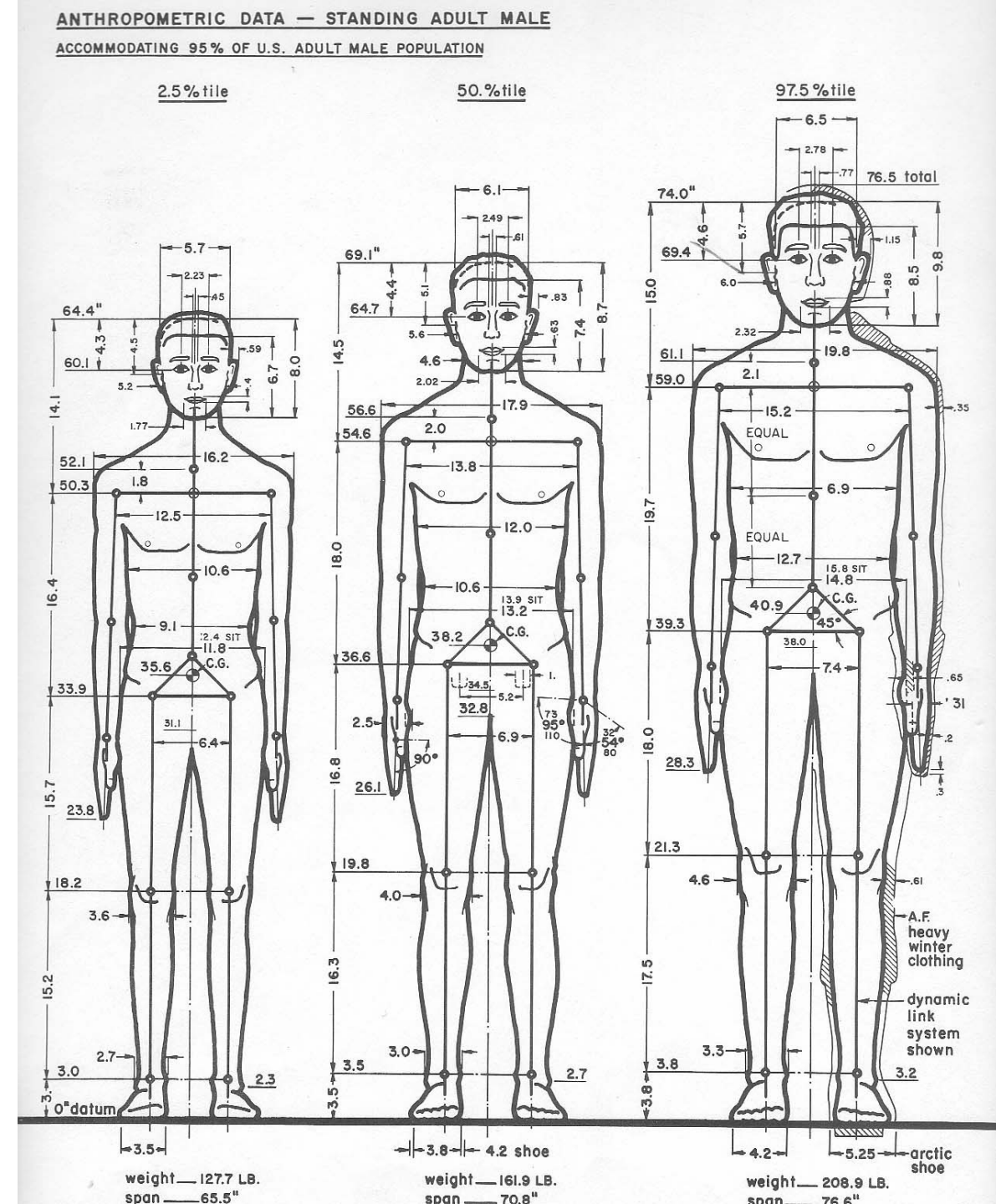


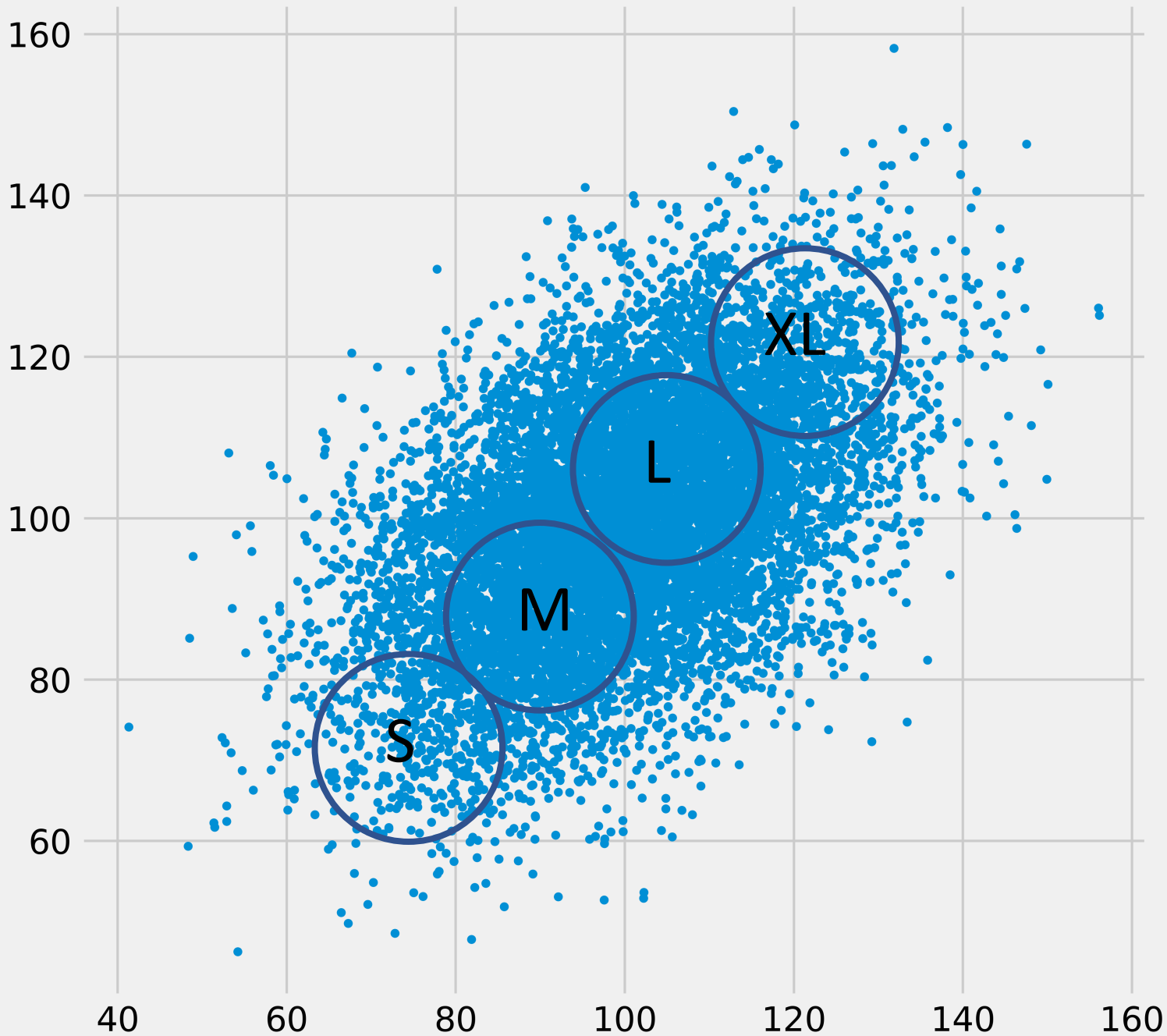
How many pilots really were average?

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/>

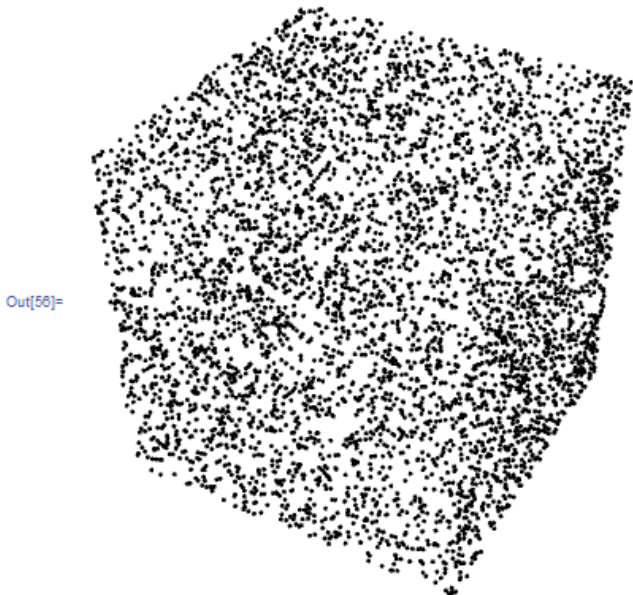




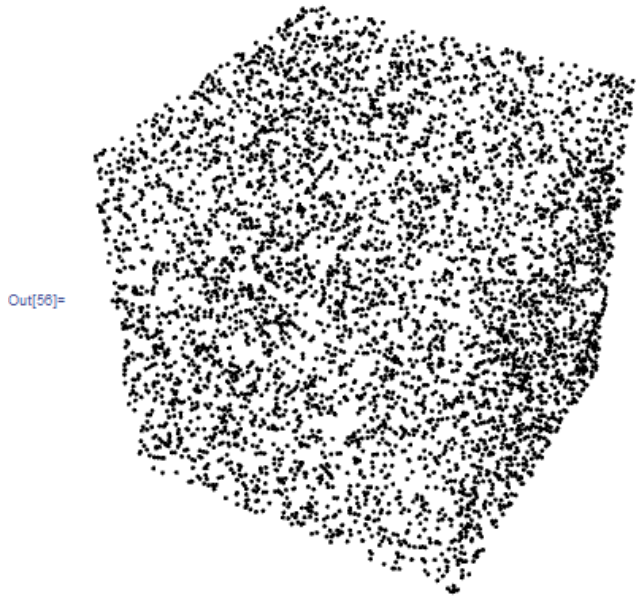
If you ever wonder why your clothes don't fit, blame the curse of dimensionality.

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 combinatorial tests. When $2^D \gg N$ situation is hopeless and there are always dimensions you never sample from.

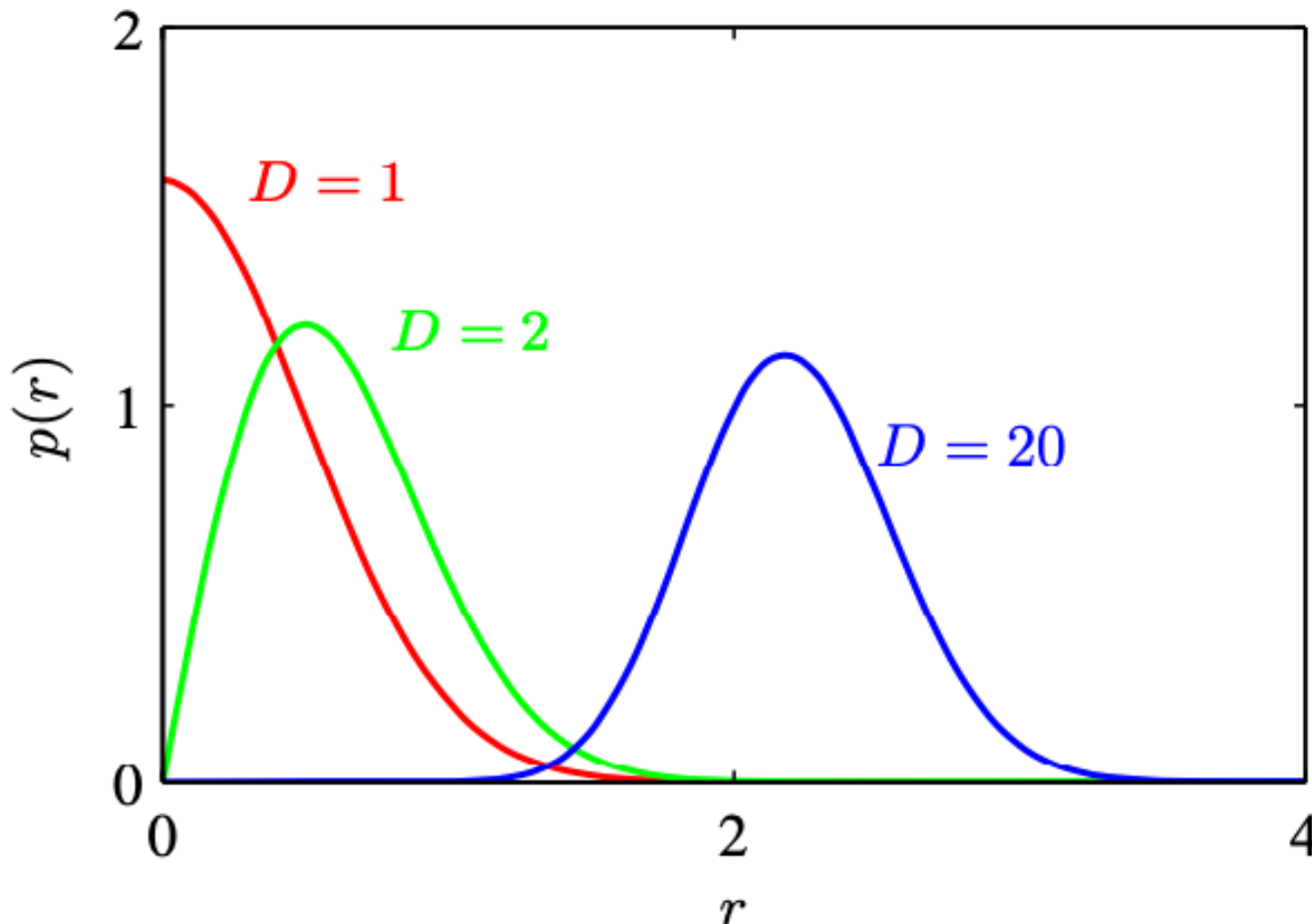


Curse of dimensionality...



- 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
$$\frac{\pi^{n/2} R^n}{\Gamma(\frac{n}{2} + 1)}$$
- 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.

Large-dimensional optimization

- The Linear discriminant and naïve Bayesian attacks didn't require optimization; we could use summary statistics to get the parameters of models with 20,000 words without fitting. When we are lucky...
- Optimizing in vast numbers of dimensions... we don't find the same answer every time.
- ChatGPT didn't use summary statistics to fit 1.3 billion parameters.. it used stochastic gradient descent optimization.

$$\text{posterior} = \frac{\text{likelihood} * \text{prior}}{\text{evidence}}$$

$$\frac{p(\text{spam}|\text{sms})}{p(\text{ham}|\text{sms})} = \frac{p(\text{sms}|\text{spam})}{p(\text{sms}|\text{ham})} \frac{p(\text{spam})}{p(\text{ham})}$$

And the evidence, which is $p(\text{sms})$ which is $p(\text{sms}|\text{spam})p(\text{spam}) + p(\text{sms}|\text{ham})p(\text{ham})$ appears in both the numerator and the denominator so we don't need to calculate it.

$$\text{posterior} = \frac{\text{likelihood} * \text{prior}}{\text{evidence}}$$

$$\frac{p(\textit{spam}|\textit{sms})}{p(\textit{ham}|\textit{sms})} = \frac{p(\textit{sms}|\textit{spam})}{p(\textit{sms}|\textit{ham})} \frac{p(\textit{spam})}{p(\textit{ham})}$$

$$\frac{p(\textit{spam}|\textit{sms})}{p(\textit{ham}|\textit{sms})} = \frac{p(\textit{spam})}{p(\textit{ham})} \Pi_{\textit{word}} \frac{p(\textit{word}|\textit{spam})}{p(\textit{word}|\textit{ham})}$$

$$\text{posterior} = \frac{\text{likelihood} * \text{prior}}{\text{evidence}}$$

$$\frac{p(\text{spam}|\text{sms})}{p(\text{ham}|\text{sms})} = \frac{p(\text{sms}|\text{spam})}{p(\text{sms}|\text{ham})} \frac{p(\text{spam})}{p(\text{ham})}$$

$$\frac{p(\text{spam}|\text{sms})}{p(\text{ham}|\text{sms})} = \frac{p(\text{spam})}{p(\text{ham})} \Pi_{\text{word}} \frac{p(\text{word}|\text{spam})}{p(\text{word}|\text{ham})}$$

$$\text{logodds spam:ham} = \log\left(\frac{p(\text{spam})}{p(\text{ham})}\right) + \sum_{\text{word}} \log\left(\frac{p(\text{word}|\text{spam})}{p(\text{word}|\text{ham})}\right)$$

$$\text{logodds spam:ham} = \log\left(\frac{p(\text{spam})}{p(\text{ham})}\right) + \sum_{word} n(\text{word}|\text{sms}) \text{score}(\text{word})$$

This is remarkably like a linear model for logodds with a per-word weight and overall bias !!

$$\text{score}(\text{word}) = \log(n_{\text{spam}}) - \log(N_{\text{spam}}) - \log(n_{\text{ham}}) + \log(N_{\text{ham}})$$

$$\text{score}(\text{word}) = \log \frac{n_{\text{spam}}}{n_{\text{ham}}} \frac{N_{\text{ham}}}{N_{\text{spam}}}$$

$$\text{logodds spam:ham} = \log\left(\frac{p(\text{spam})}{p(\text{ham})}\right) + \sum_{word} n(\text{word}|\text{sms}) \text{score}(\text{word})$$

This is remarkably like a linear model for logodds with a per-word weight and overall bias !!

$$\text{score}(\text{word}) = \log(n_{\text{spam}}) - \log(N_{\text{spam}}) - \log(n_{\text{ham}}) + \log(N_{\text{ham}})$$

$$\text{score}(\text{word}) = \log \frac{n_{\text{spam}}}{n_{\text{ham}}} \frac{N_{\text{ham}}}{N_{\text{spam}}}$$

$$\text{score}(\text{word}) = \log \frac{n_{\text{spam}} + \alpha}{n_{\text{ham}} + \alpha} \cdot \frac{N_{\text{ham}} + \beta}{N_{\text{spam}} + \beta}$$

That “prior”.... in this problem the prior is a single number (almost like an intercept term in logistic regression)

We can estimate the fraction of spam to ham, (from judgement)..

We can also adjust the prior to improve the behavior of the model. If the balance between FP and FN is not to our liking, turn the knob.