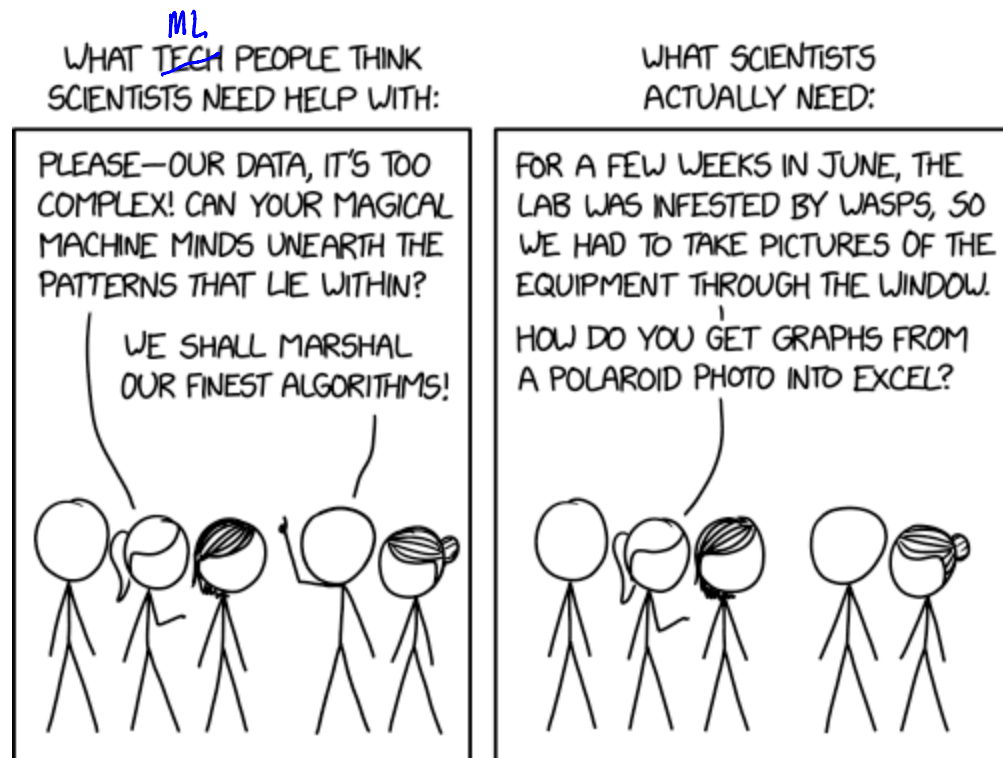


# Chapter 1: Introduction

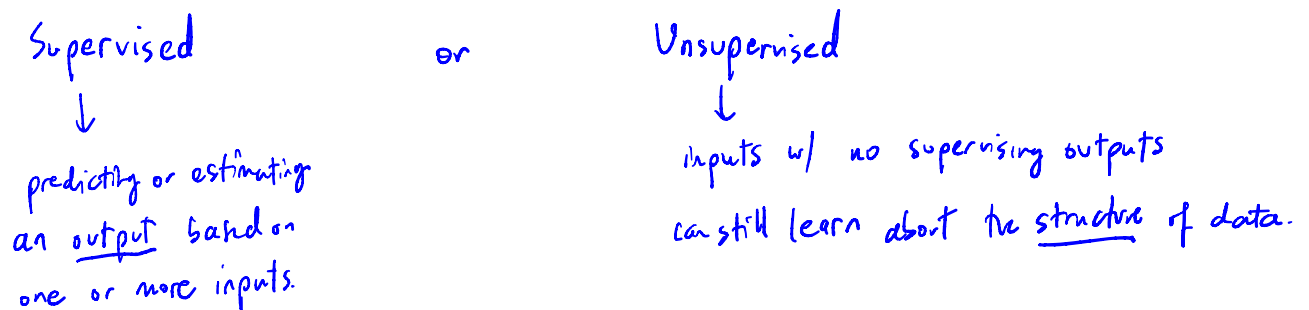
*Statistical learning* refers to a vast set of tools for understanding data.



<https://xkcd.com/2341/>

**Alternative text:** I vaguely and irrationally resent how useful WebPlotDigitizer is.

These tools can broadly be thought of as

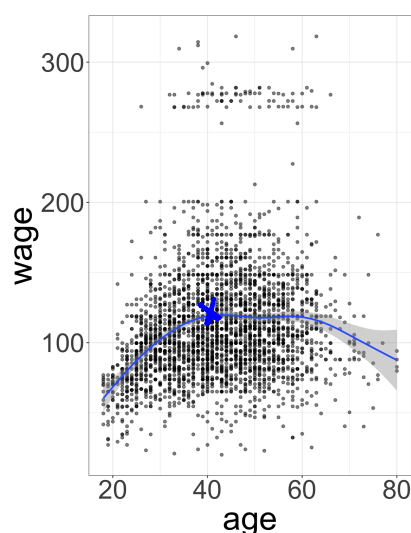


Examples:

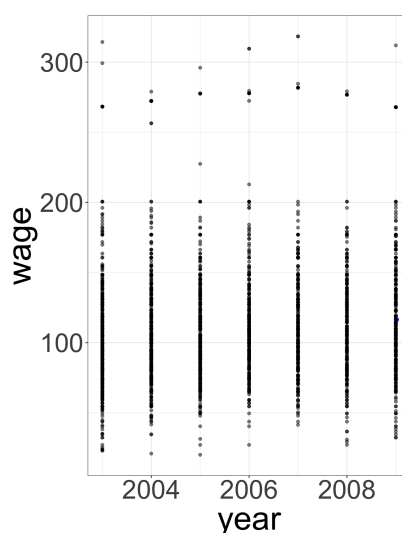
## Wage data

year	age	maritl	race	edu- cation	region	job- class	health	health_ins	logwage	wage
2006	18	1. Never Mar- ried	1. White	1. < HS Grad	2. Mid- dle At- lantic	1. Indus- trial	1. <=Good	2. No	4.318063	75.04315
2004	24	1. Never Mar- ried	1. White	4. Col- lege Grad	2. Mid- dle At- lantic	2. Infor- ma- tion	2. >=Very Good	2. No	4.255273	70.47602
2003	45	2. Mar- ried	1. White	3. Some Col- lege	2. Mid- dle At- lantic	1. Indus- trial	1. <=Good	1. Yes	4.875061	130.98218

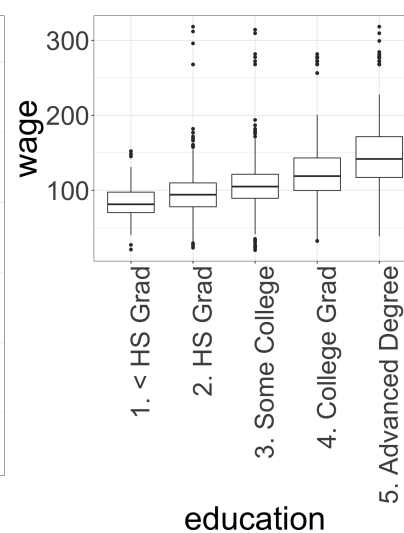
Factors related to wages for a group of males from the Atlantic region of the United States. We might be interested in the association between an employee's age, education, and the calendar year on his wage. *relationship*



wage looks to increase w/ age but then decreases after age 60.



slow but slight increase in wage over time.  
lot of variability.



wage typically higher for individuals w/ greater education levels.

could use 1 factor to predict wage, but lots of variability.

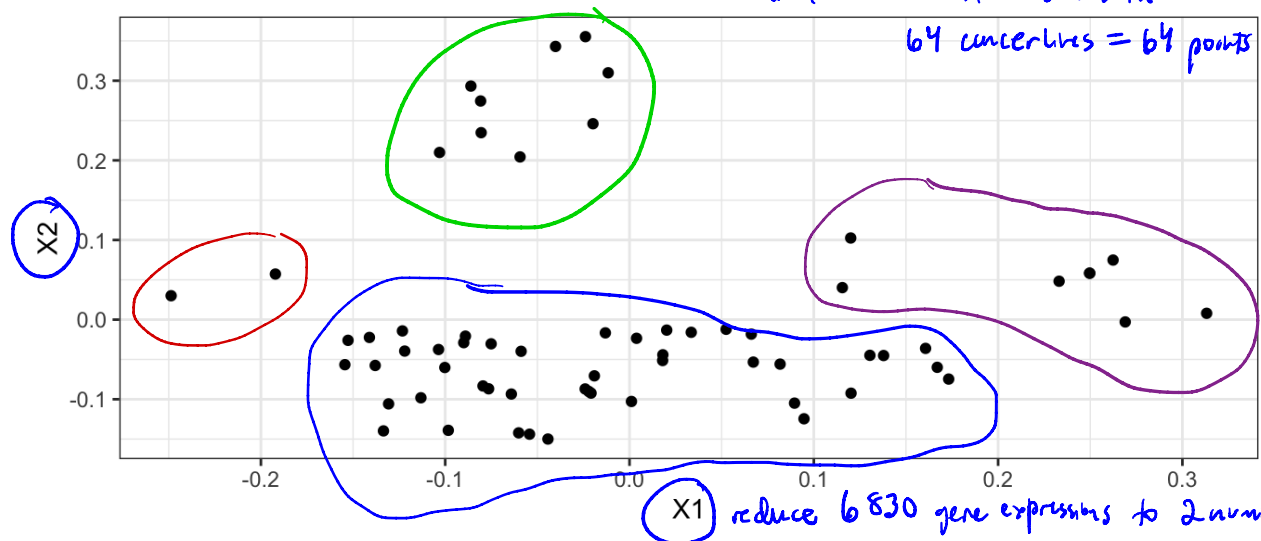
would be better (more accurate) to combine age, education, & year and also account for nonlinear relationship w/ age and wage.

## Gene Expression Data

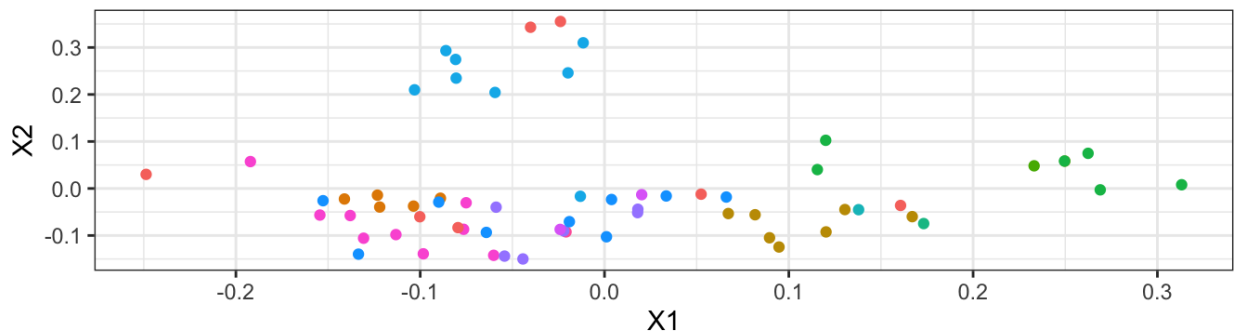
Consider the NCI60 data, which consists of 6,830 gene expression measurements for 64 cancer lines. We are interested in determining whether there are groups among the cell lines based on their gene expression measurements.

*we have no known output (cancer type), instead we want to look for structure in data*

*64 cancer lines = 64 points*



*reduce 6,830 gene expressions to 2 numbers "principal components" to describe the structure ("dimension reduction")*



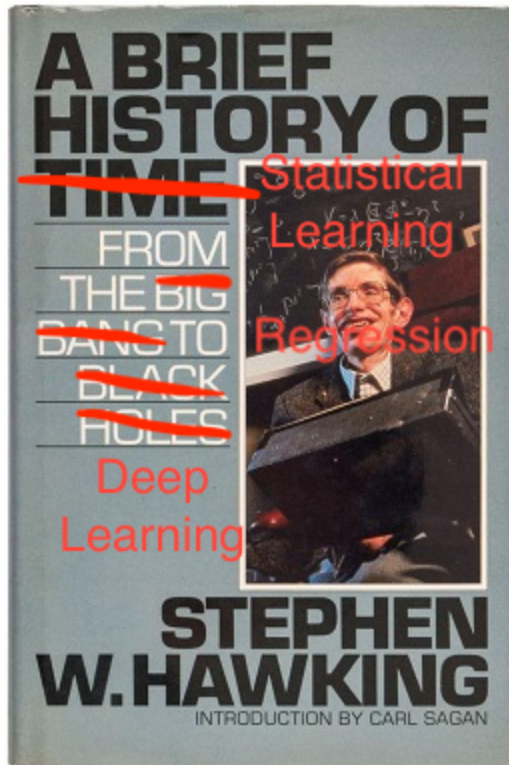
NCI60\$labs

BREAST	K562A-repro	MCF7A-repro	NSCLC	RENAL
CNS	K562B-repro	MCF7D-repro	OVARIAN	UNKNOWN
COLON	LEUKEMIA	MELANOMA	PROSTATE	

*"true" cancer types*

*cell lines w/ same cancer type are close in 2D representation  
and our clustering (top) was able to find some of these types*

# 1 A Brief History



Although the term “statistical machine learning” is fairly new, many of the concepts are not. Here are some highlights:

early 19th century - Legendre and Gauss publish method of least squares  $\Rightarrow$  linear regression  
1936 - Linear discriminant analysis  
1940s - Logistic regression  
1960s - Bayesian Methods (1980s popularized)

## 2 Notation and Simple Matrix Algebra

I'll try to keep things consistent notationally throughout this course. Please call me out if I don't!

$n$

$p$

$x_{ij}$

$\mathbf{X}$

$\mathbf{y}$

$a, \mathbf{A}, A$

$a \in \mathbb{R}$

Matrix multiplication