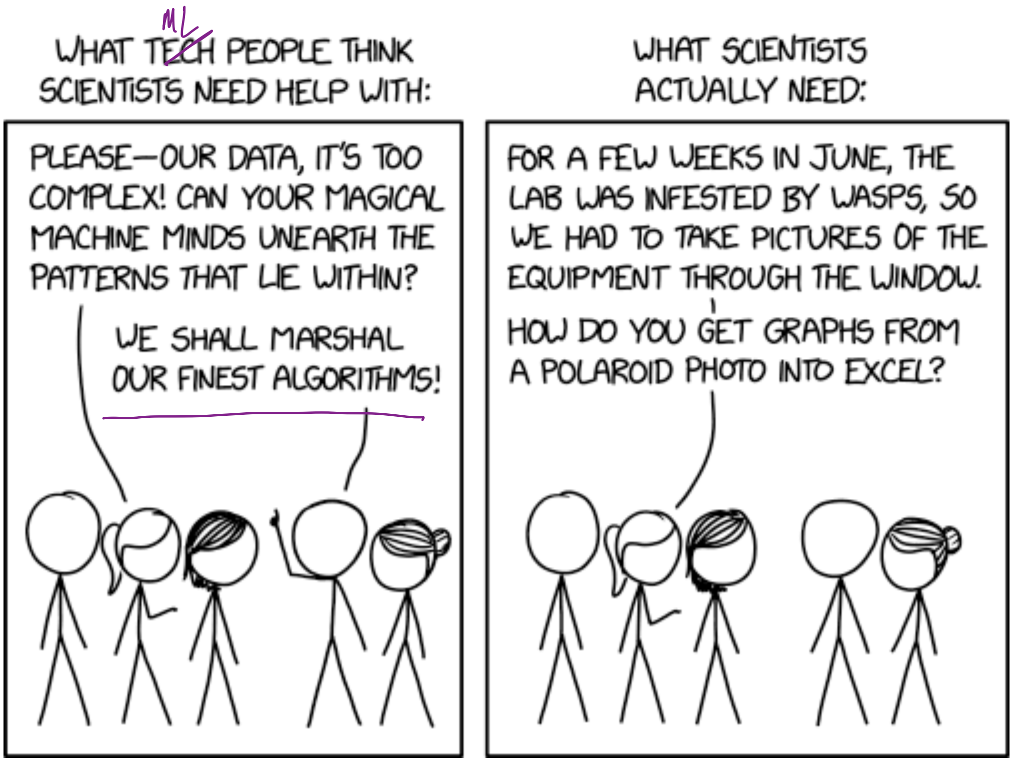
No office hours today.

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

Supervised or Unsupervised

Supervised ↓ predicting or estimating an output based on one or more inputs.

Unsupervised: inputs w/ no supervising outputs
can still learn about structure of data

Supervised

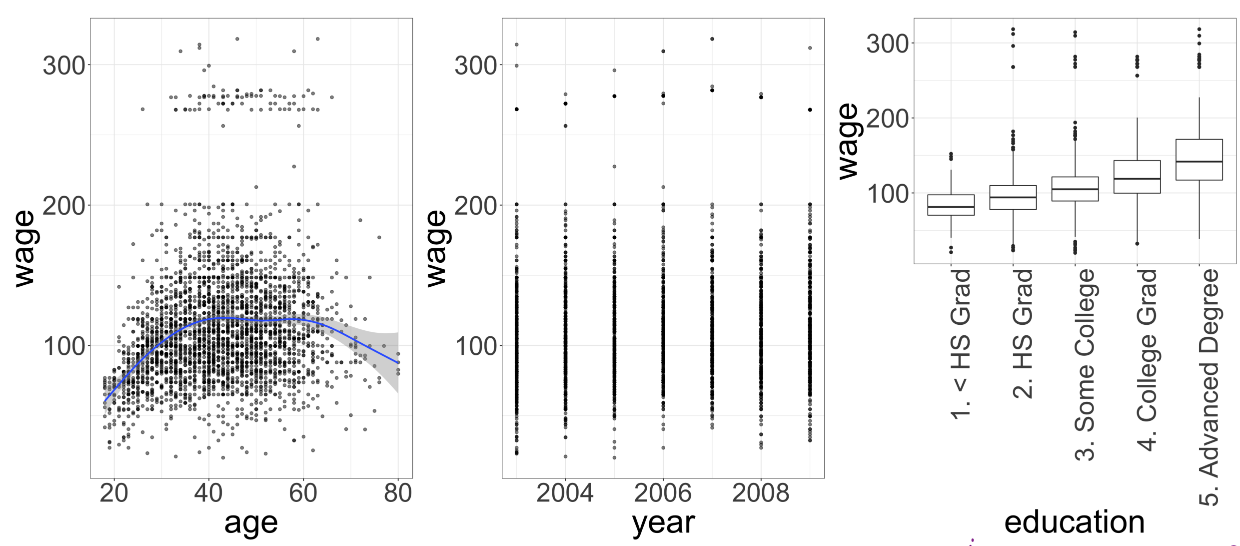
Examples:

Wage data

year	age	maritl	race	education	region	jobclass	health	health_ins	logwage	wage
2006	18	1. Never Married	1. White	1. < HS Grad	2. Middle Atlantic	1. Industrial	1. <=Good	2. No	4.318063	75.04315
2004	24	1. Never Married	1. White	4. College Grad	2. Middle Atlantic	2. Information	2. >=Very Good	2. No	4.255273	70.47602
2003	45	2. Married	1. White	3. Some College	2. Middle Atlantic	1. Industrial	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.

Very slight increase in wage over time
lots of variability.

Wages typically higher for individuals w/ greater education levels.

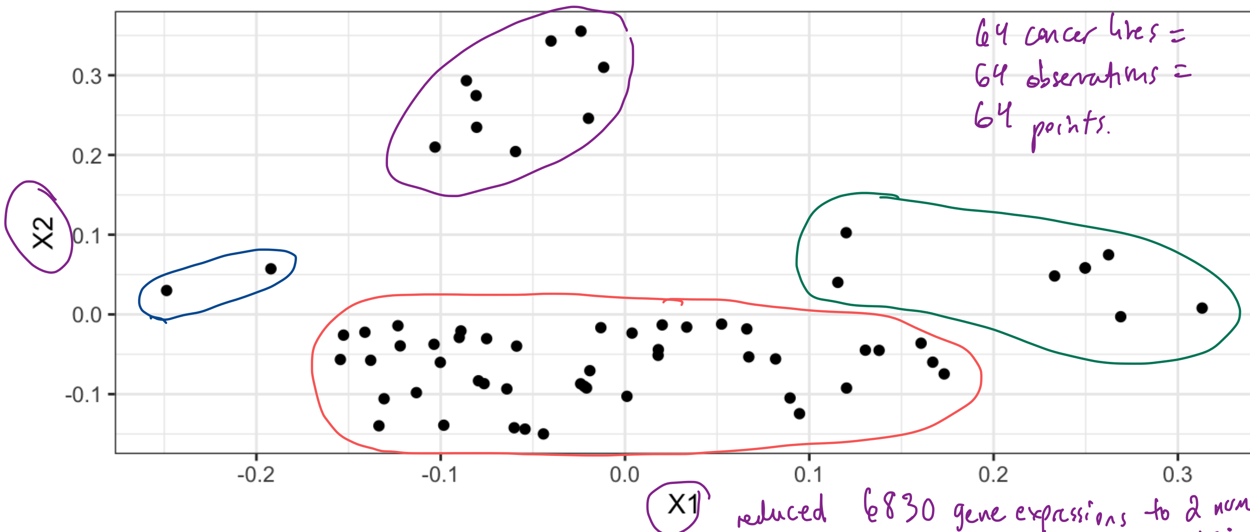
could use 1 factor to predict wage, but lots of variability.

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

Unsupervised.

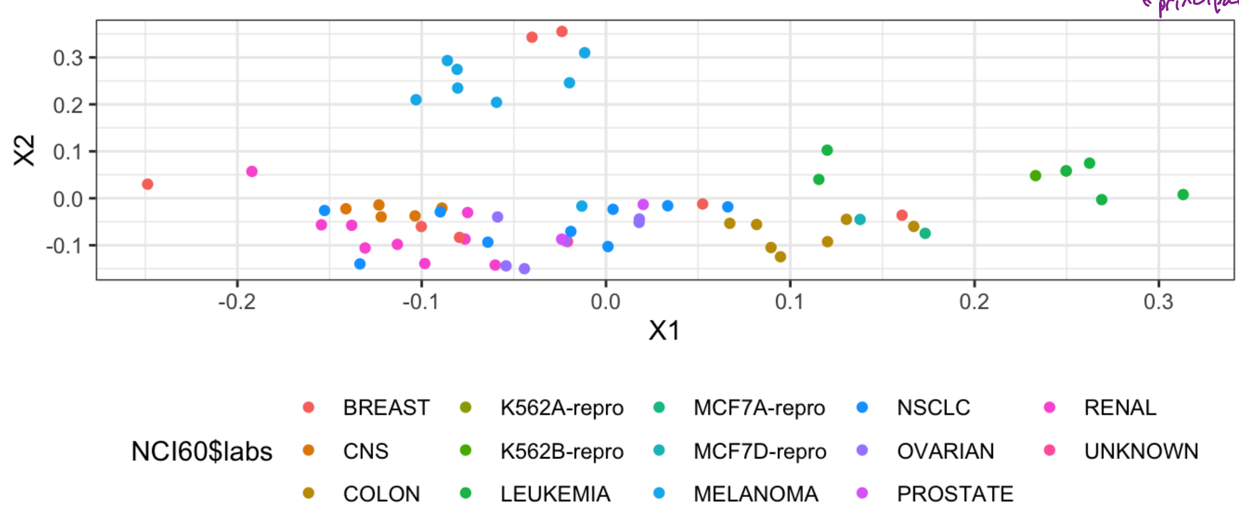
Gene Expression Data

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.



visual inspection maybe 4 groups?

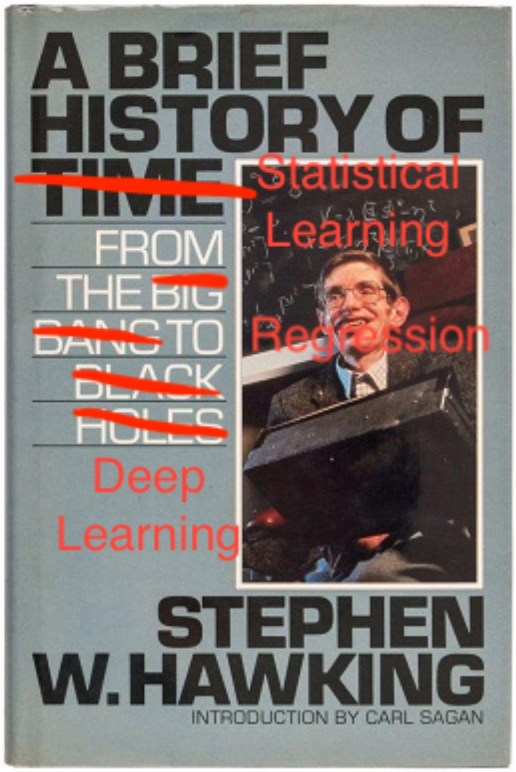
reduced 6830 gene expressions to 2 numbers "principal components" to describe the structure (dimension reduction).



"true" cancer types.

cell lines w/ same cancer type are close in 2D representation. and visual 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 & Gauss publish method of least squares => linear regression.

1936 – Fisher proposes Linear discriminant analysis.

1940s – logistic regression.

1960s – Bayesian methods.

1970s – generalized linear regression (includes linear + logistic).

non-linear models too computationally complex at this point. →

1980s – Brieman & Friedman introduce classification and regression trees (random forest, cross-validation).

1990s – ML boom! Shift to data-driven approach.

- support vector machines
- recurrent neural networks.

2000s – kernel methods, unsupervised learning becomes more popular.

more data
more computational complexity.

2010s – "deep" learning.

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't!

n - number of distinct data points or observations in our sample

p - # of variables available to us for making predictions.

e.g. Wage data has 12 variables collected for 3,000 people. $n = 3000$, $p = 12$.

x_{ij} ← value of j^{th} variable for i^{th} observation.

$i = 1, \dots, n$
$j = 1, \dots, p$.

$\boldsymbol{X}$ – $n \times p$ matrix whose $(i,j)^{th}$ element is x_{ij}

$$\underline{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

$\underline{x}_i = i^{th}$ row of $\underline{X}$ (vector of length p) $= \begin{pmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{pmatrix}$.

$\underline{x}_i^T = (x_{i1} \cdots x_{ip})$ "transpose"

$\boldsymbol{y}$ – variable on which we wish to make a prediction "response"

$y_i = i^{th}$ observation of $\underline{y}$.

$a, \boldsymbol{A}, A$ ← scalar, matrix, random variable

$\underline{a}$ – vector

$a \in \mathbb{R}$ ← indicates dimension

$\underline{A} \in \mathbb{R}^{r \times s}$ = $r \times s$ matrix.

Matrix multiplication

Let $\underline{A} \in \mathbb{R}^{r \times d}$ and $\underline{B} \in \mathbb{R}^{d \times s}$ (must be equal.) then product of A and B is "AB" → multiplying rows of A elementwise w/ columns of B, adding.

$$(AB)_{ij} = \sum_{k=1}^{d} a_{ik} b_{kj}$$

e.g. $A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$, $B = \begin{pmatrix} 5 & 6 \\ 7 & 8 \end{pmatrix}$.

$$AB = \begin{pmatrix} 1\times5 + 2\times7 & 1\times6 + 2\times8 \\ 3\times5 + 4\times7 & 3\times6 + 4\times8 \end{pmatrix} = \begin{pmatrix} 19 & 22 \\ 43 & 50 \end{pmatrix}$$ ← result is $r \times s$ matrix