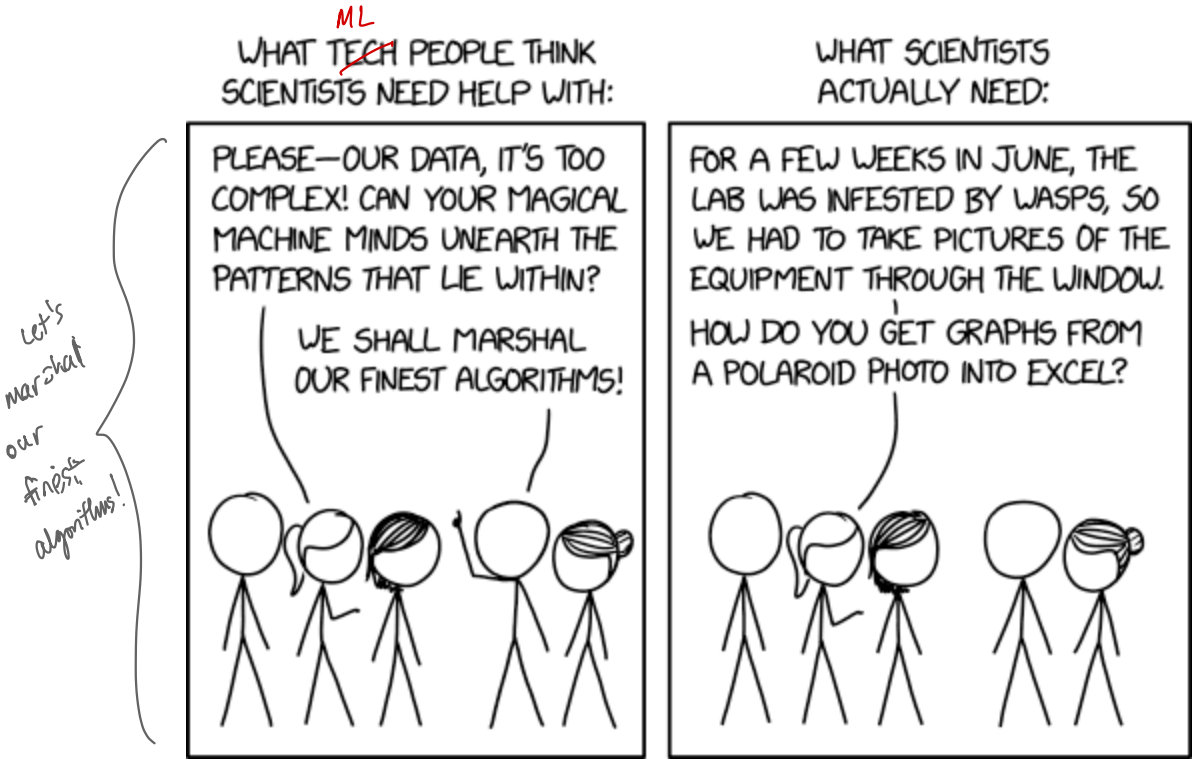
Chapter 1: Introduction

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 → predicting or estimating an output based on one or more inputs.

or

Unsupervised. → inputs w/ no supervising outputs

can still learn about the structure of data.

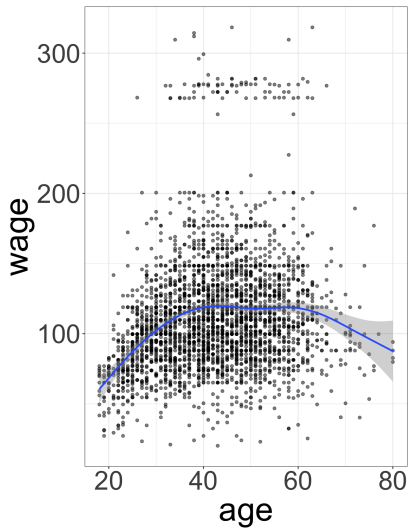
Examples:

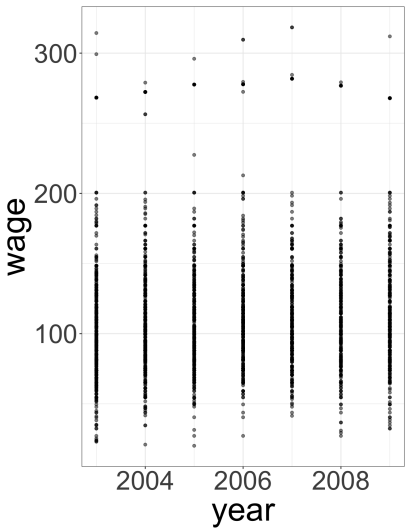
Wage data

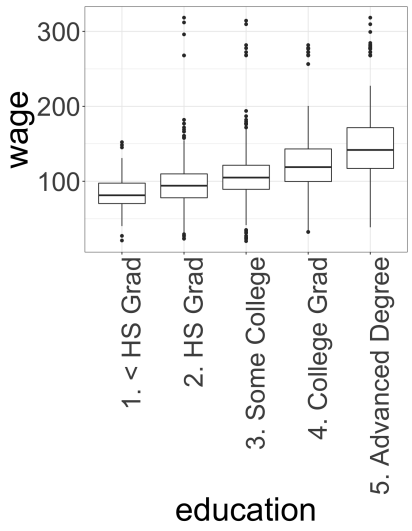
year	age	maritl	race	education	region	jobclass	health	health_ins	logwage	wage
2006	18	1. Never Married	1. White	1. < HS Grad	2. Middle Atlantic	1. Industrial	1. <=Good	2. No	4.318063	75.04315
2004	24	1. Never Married	1. White	4. College Grad	2. Middle Atlantic	2. Information	2. >=Very Good	2. No	4.255273	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 increases until 40, plateaus until 60 then decreases.

slight increase

lots of variability.

Wages typically higher for greater education levels.

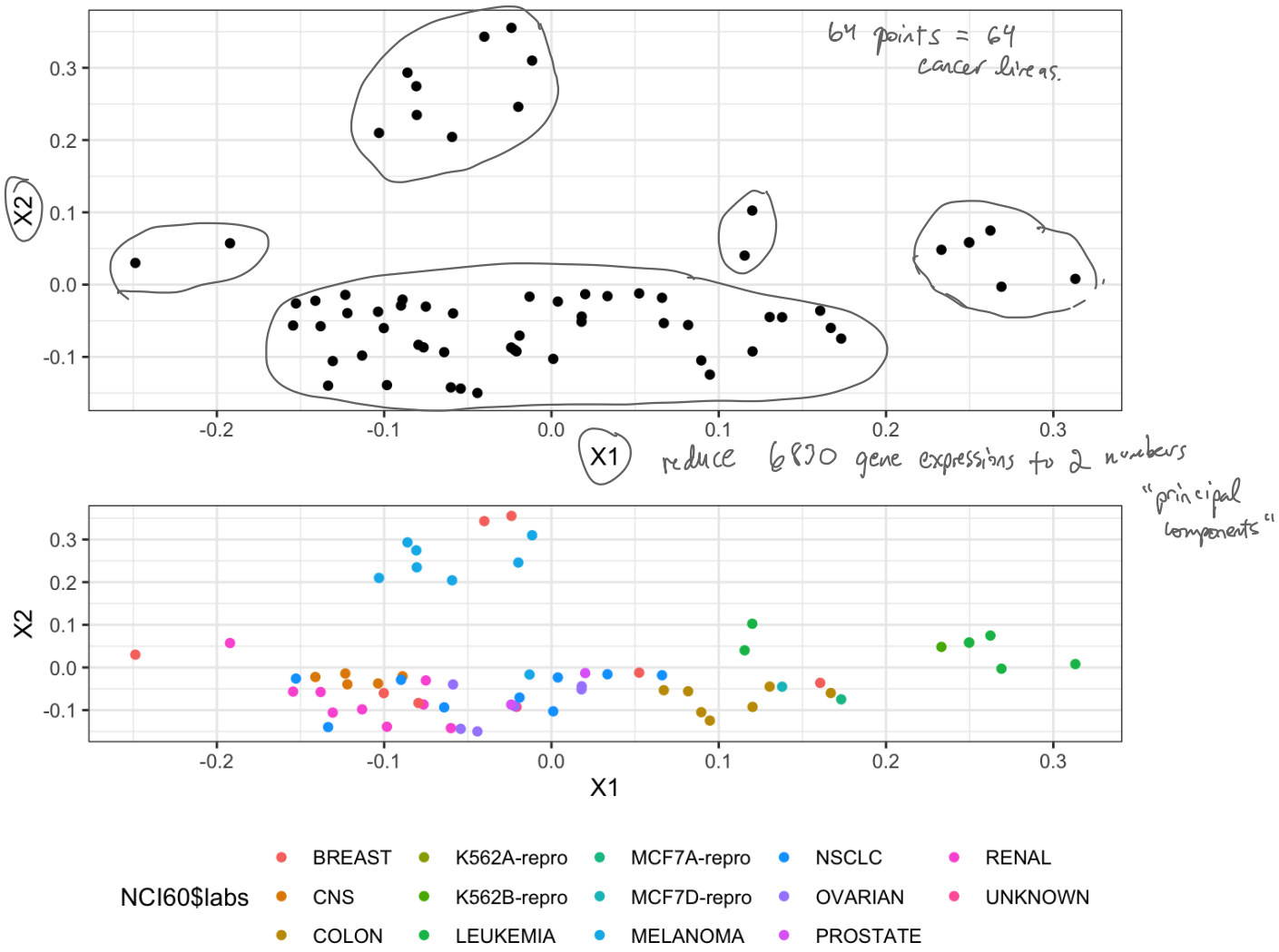
could use 1 factor to predict wage, but lots of variability.

Would be more accurate to combine age, education, & year and also account for nonlinear relationship b/ age & wage.

Gene Expression Data

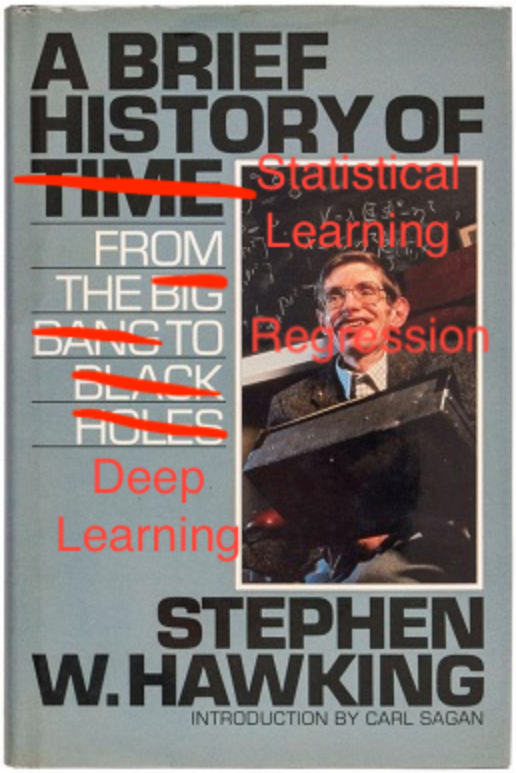
Consider the `NCI60` data, which consists of 6,830 gene expression measurements for 64 cancer lines. We are interested ind determining whether there are **groups** among the cell lines based on their gene expression measurements.

we have no known output (cancer cluster)



cell lines clustered based on proximity in 2D representation.

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 square ⇒ linear regression

1936 - Fisher proposes Linear discriminant analysis

1940s - logistic regression

1960s - Bayesian methods.

1970 - generalized linear regression (includes linear & logistic)

non-linear models too computationally complex.

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

more data

more computational complexity

1990s - ML boom! shift to data driven approach
- support vector machines
- recurrent neural networks

2000s - kernel methods, unsupervised methods become more popular.

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!

n – number of distinct data points (observations) in a sample.

p – # of variables available to us.

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

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}

$$\boldsymbol{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

$\underline{x}_i$ = i^{th} row of $\boldsymbol{X}$ (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

y_i = i^{th} observation of y.

$a, \boldsymbol{A}, A$ ← random variable.

↑ scalar ↖ matrix

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

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

Matrix multiplication

Let $A \in \mathbb{R}^{r \times d}$ and $B \in \mathbb{R}^{d \times s}$ (must match!) then product of A and B is "AB"

→ multiply rows of A (element wise) by columns of B

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

e.g. $A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$ $B = \begin{pmatrix} 5 & 6 \\ 7 & 8 \end{pmatrix}$, $AB = \begin{pmatrix} 1 \cdot 5 + 2 \cdot 7 & 1 \cdot 6 + 2 \cdot 8 \\ 3 \cdot 5 + 4 \cdot 7 & 3 \cdot 6 + 4 \cdot 8 \end{pmatrix} = \begin{pmatrix} 19 & 22 \\ 43 & 50 \end{pmatrix}$ (2×2).

result is $r \times s$ matrix