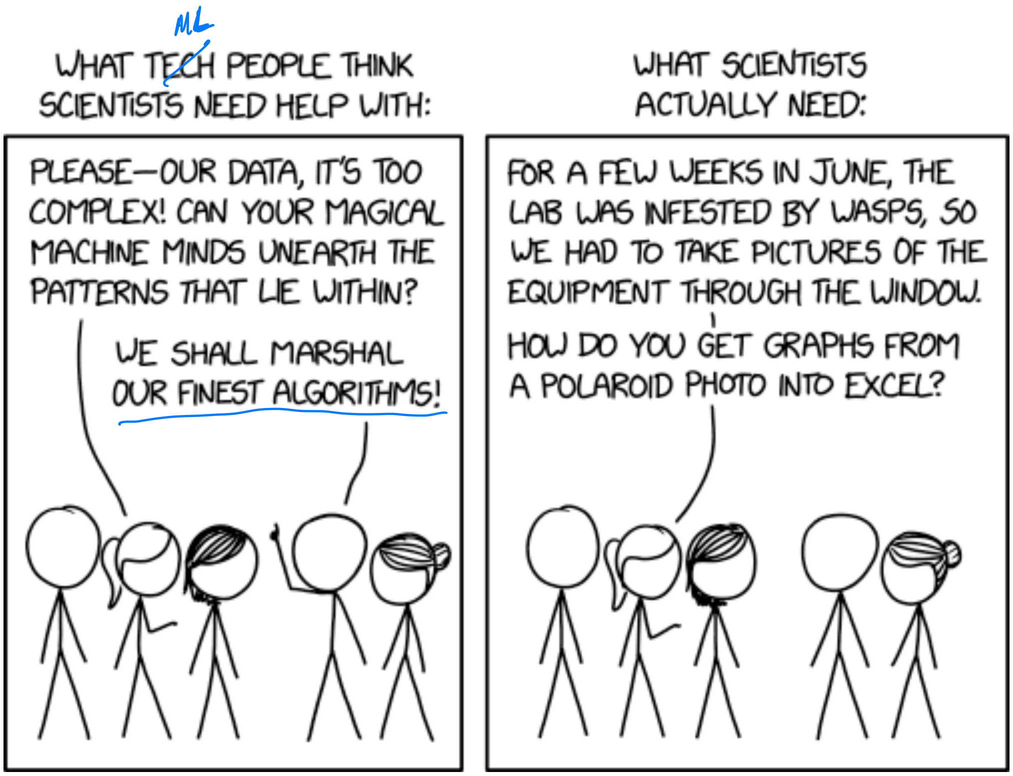
Chapter 1: Introduction

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

predicting or estimating an output based on one or more inputs

inputs w/ no supervising outputs

can still learn about the structure of data.

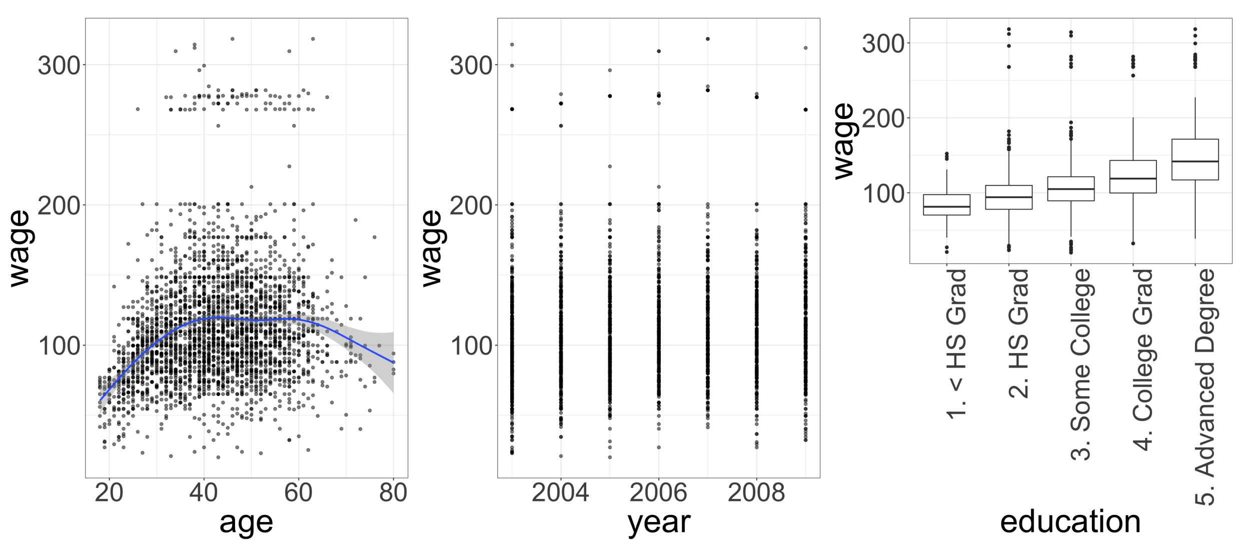
Examples:

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 w/ age then flattens and decreases after 60.

slight increase over time.

wages typically higher for greater education levels.

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

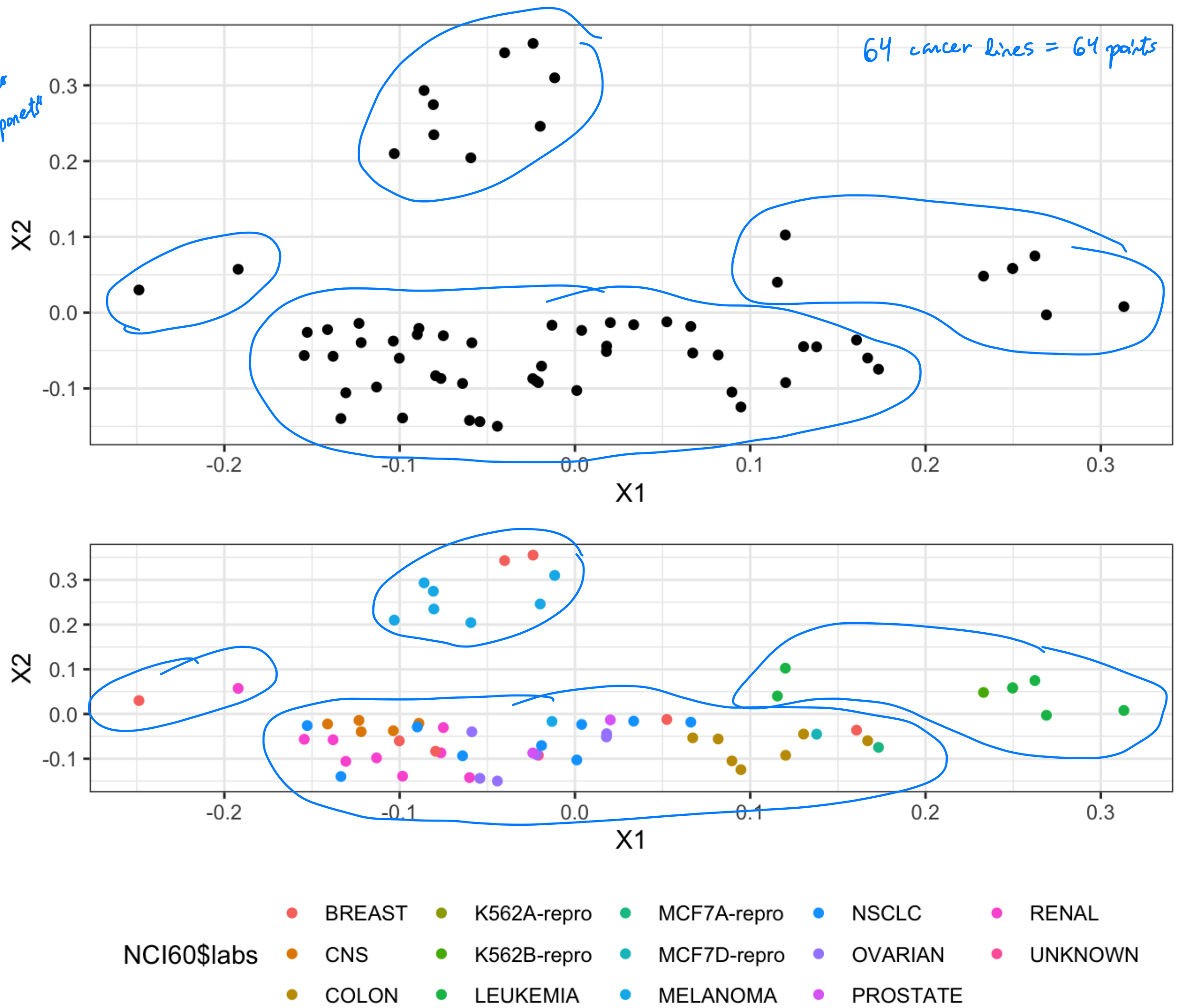
> would be better to combine age, education, & year + nonlinear behavior w/ age and wage

Gene Expression Data

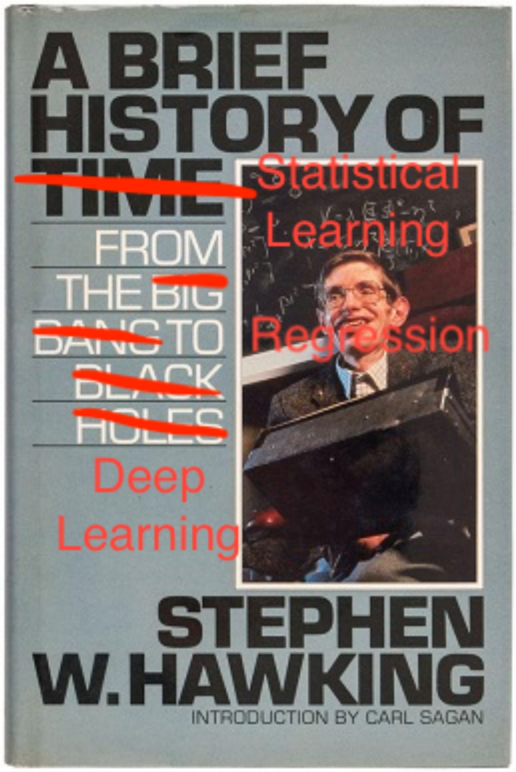
Consider the `NCI60` data, which consists of 6,830 gene expression measurements for 64 cancer lines. We are interested ind determining whether there are **groups** among the cell lines based on their gene expression measurements.

reduce 6830 gene expressions to 2 " principal components (see later).

look for structure.



1 A Brief History



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

early 19^{th} 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

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

1990s – ML boom! shift to data-driven approach

- support vector machines
- recurrent neural networks.

more data

more computational complexity.

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

2010s – "deep learning"

2020s – generative AI?

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 or observations in our sample.

p – # of variables available for making predictions.

e.g. Wage data has 12 variables collected for 3,000 people $\Rightarrow$ $n = 3{,}000$ $p = 12$.

x_{ij} – value of j^{th} variable for ith observation.

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

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

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

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

"transpose"

$\underline{x}_i^T = (x_{i1}, \ldots, x_{ip})$

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

y_i = i^{th} observation of y.

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

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

$\underline{a} \in \mathbb{R}^d$

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

Matrix multiplication

must equal!

let $\underline{A} \in \mathbb{R}^{r \times d}$ and $\underline{B} \in \mathbb{R}^{d \times s}$ then product of A and B is "AB"

→ multiply rows of A (elementwise) by columns of B and sum.

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