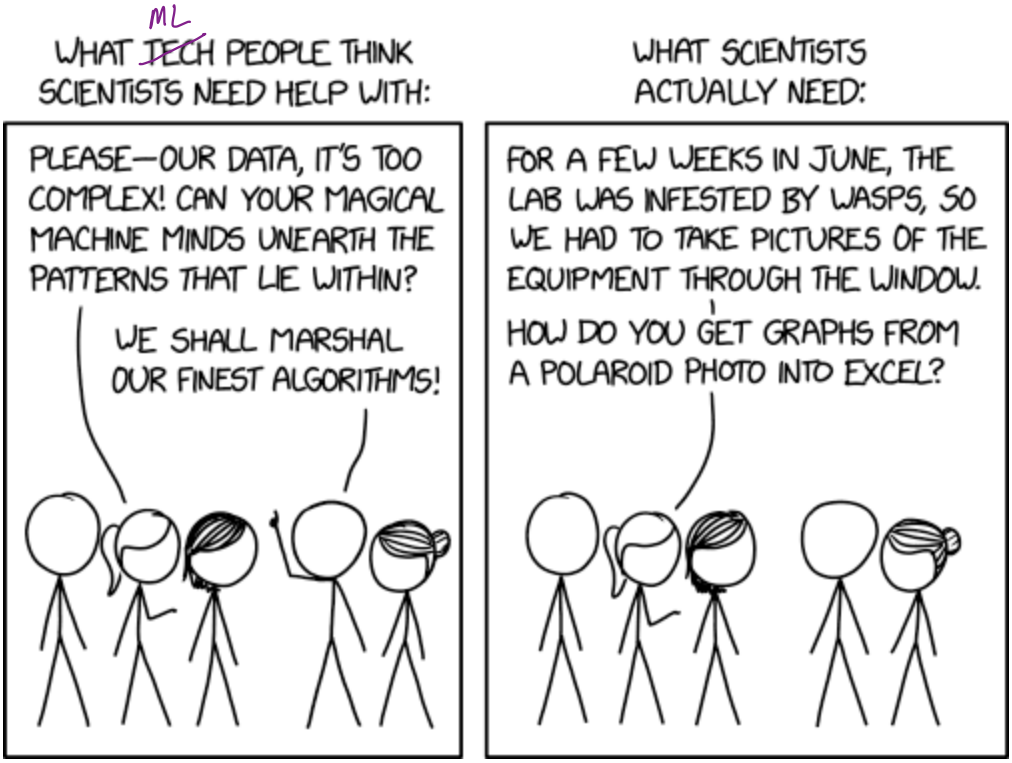
Chapter 1: Introduction

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
↓
predict or estimate an output based on one or more inputs.

or

Unsupervised
inputs w/ no supervising outputs
can still learn about the structure of our data.

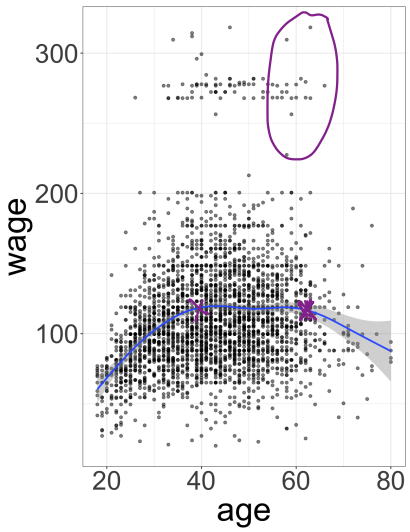
Examples:

Wage data

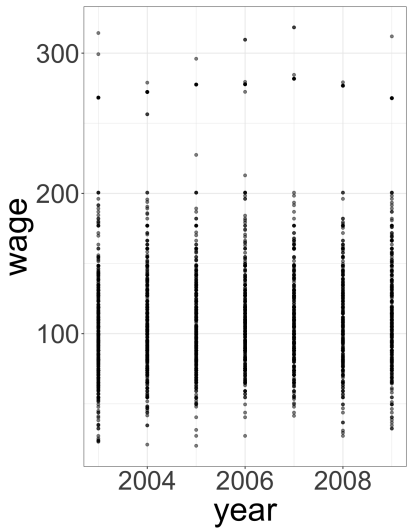
year	age	maritl	race	education	region	jobclass	health	health_ins	logwage	wage
2006	18	1. Never Married	1. White	1. < HS Grad	2. Middle Atlantic	1. Industrial	1. <=Good	2. No	4.318063	75.04315
2004	24	1. Never Married	1. White	4. College Grad	2. Middle Atlantic	2. Information	2. >=Very Good	2. No	4.255273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

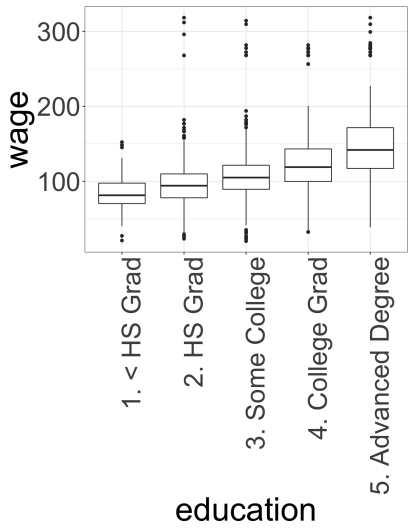
inputs — output — relationship — inputs



Wage looks to increase w/ age then decreases after 60



slight increase in wage over time but lots of variability.



Wage typically higher for individuals w/ greater education levels.

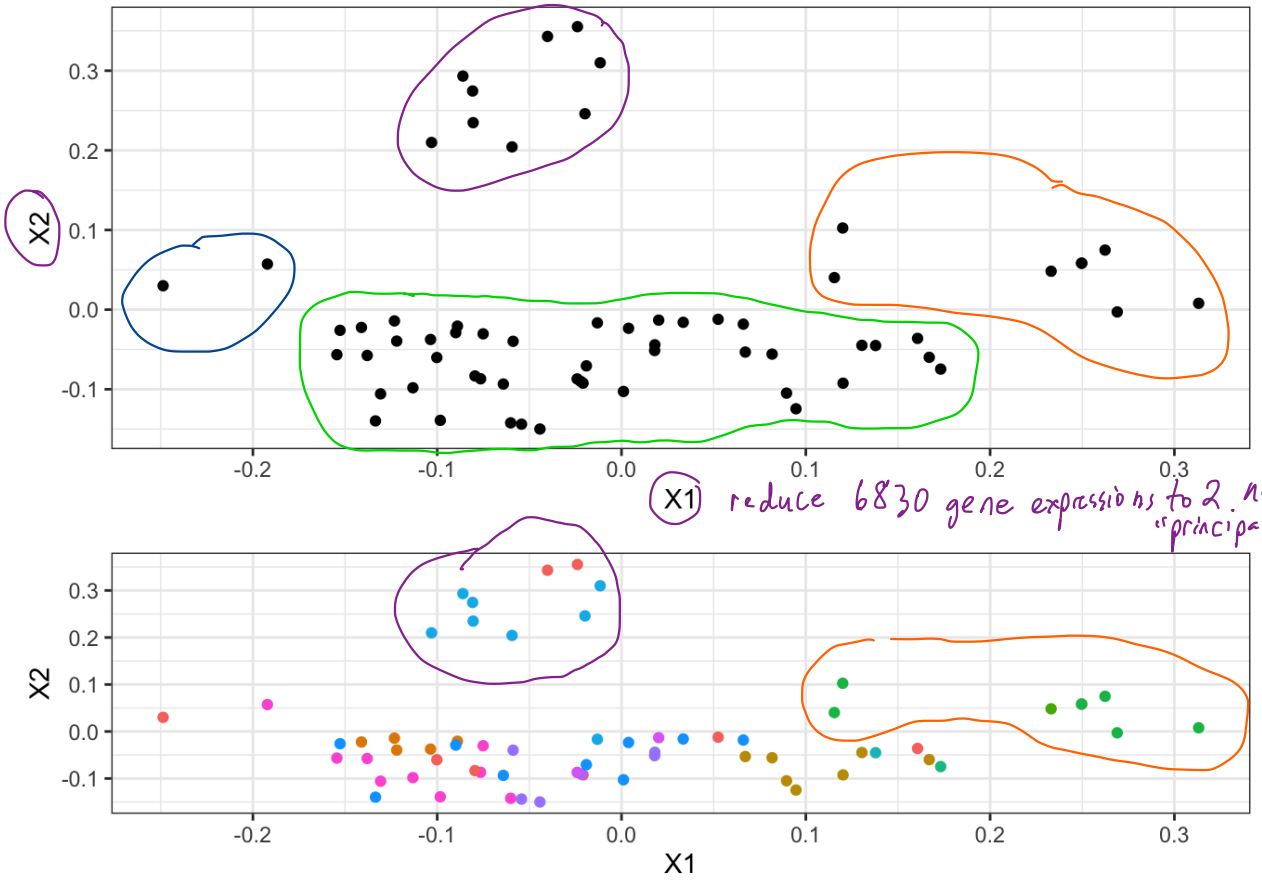
We could use 1 factor to predict wage, but lots of variability.

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

Gene Expression Data

unsupervised problem.

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 don't have known output (cancer type) instead we can look for structure in data

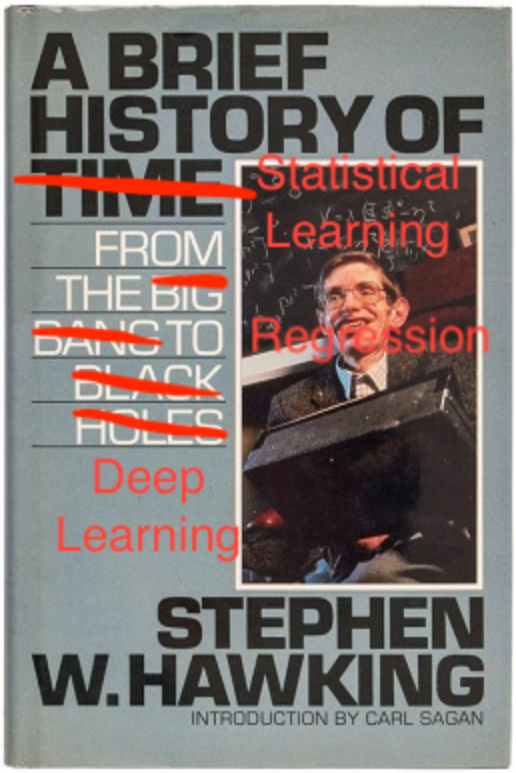


reduce 6830 gene expressions to 2 numbers "principal components" to describe the structure (dimension reduction).

"true" cancer types

cell lines w/ same cancer type are "close" in 2D representation and 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 ⇒ linear regression.

1936 – Linear discriminant analysis

1940 – Logistic regression.

1960s – Bayesian methods (1980s popularized)

1970s – generalized linear regression (includes linear & logistic)

non-linear methods too computationally complex at this point →

1980s – Breiman & Friedman introduced classification & regression trees (random forest) & cross-validation

1990s – ML Boom! Shift to data-driven approach

Support vector Machines

recurrent neural nets.

2000s – kernel methods, unsupervised learning becomes more popular

2010s – "deep learning"

more data

more computational complexity.

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 $p = 12$ variables & $n = 3{,}000$ people.

lowercase, not bolded

x_{ij} - value of the jth variable for ith individual.

$i = 1, \ldots, n$

$j = 1, \ldots, p$.

capital, bolded

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

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

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

$\underline{x}_i^T = \underline{x}_i' = (x_{i1}, \ldots, x_{ip})$ "transpose"

lowercase bolded

$\boldsymbol{y}$ - variable on which we wish to make a prediction

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

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

scalar, matrix

$\underline{a}$ = vector

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

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

$\underline{y} \in \mathbb{R}^n$

Matrix multiplication

must be equal.

Let $A \in \mathbb{R}^{r \times d}$ and $B \in \mathbb{R}^{d \times s}$ the product of A and B is "AB" → multiply rows of A by columns of B elementwise & sum.

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

e.g. $A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$, $B = \begin{pmatrix} 5 & 6 \\ 7 & 8 \end{pmatrix}$ $\Rightarrow AB = \begin{pmatrix} 1\times5+2\times7 & 1\times6+2\times8 \\ 3\times5+4\times7 & 3\times6+4\times8 \end{pmatrix} = \begin{pmatrix} 19 & 22 \\ 43 & 50 \end{pmatrix}$

result is $r \times s$ matrix.