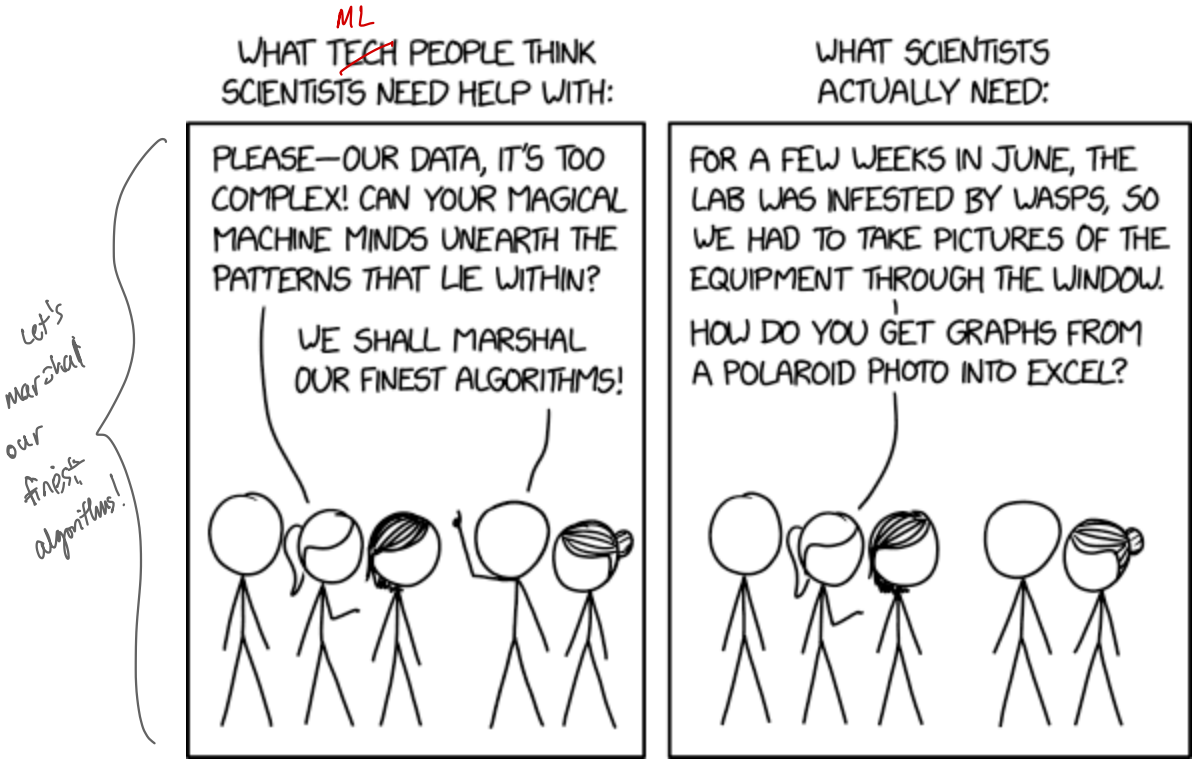
Chapter 1: Introduction

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.

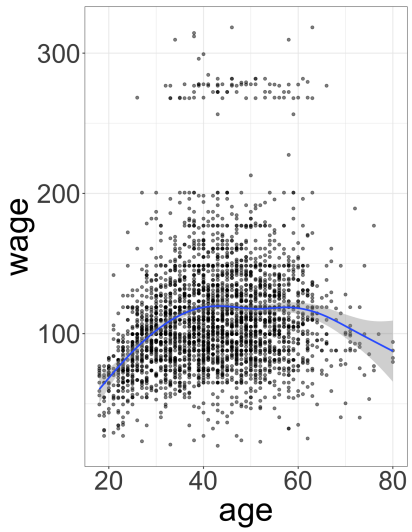
Examples:

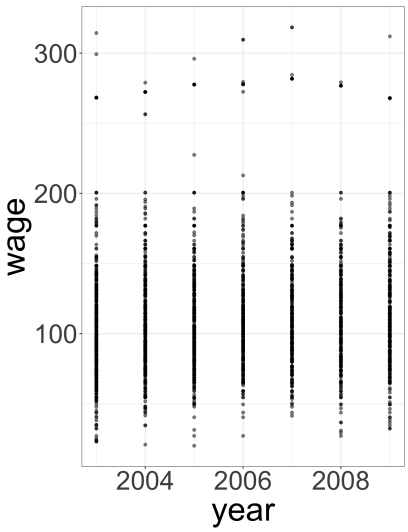
Wage data

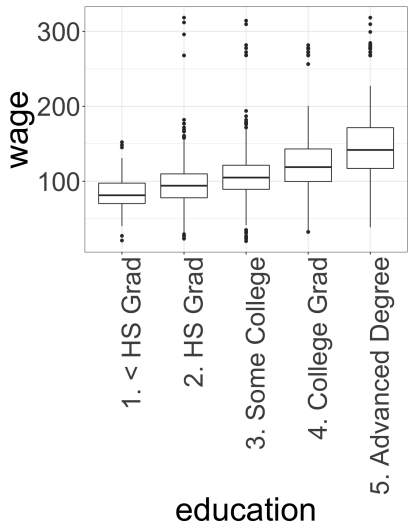
year	age	maritl	race	education	region	jobclass	health	health_ins	logwage	wage
2006	18	1. Never Married	1. White	1. < HS Grad	2. Middle Atlantic	1. Industrial	1. <=Good	2. No	4.318063	75.04315
2004	24	1. Never Married	1. White	4. College Grad	2. Middle Atlantic	2. Information	2. >=Very Good	2. No	4.255273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

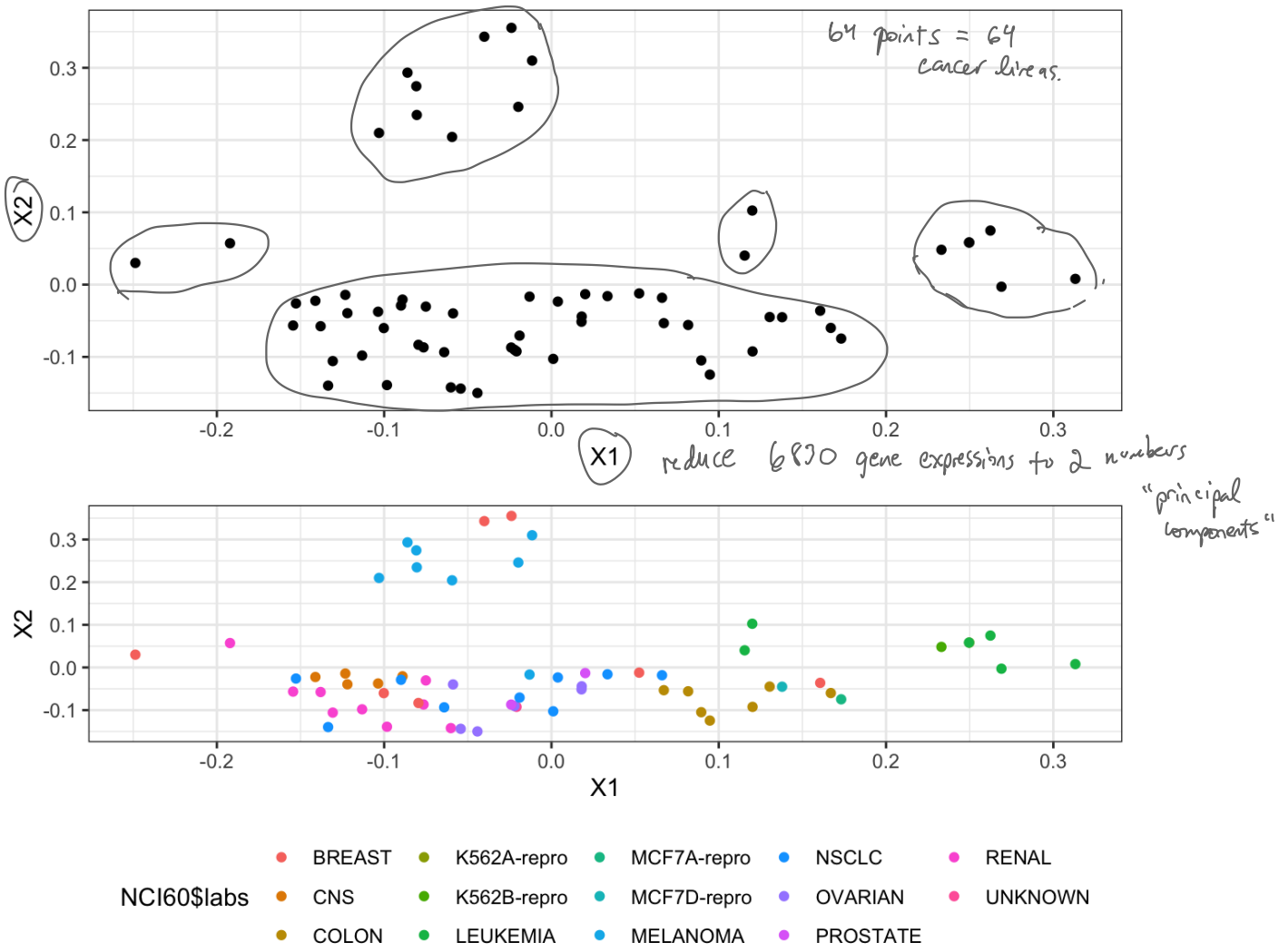
could use 1 factor to predict wage, but lots of variability.

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

Gene Expression Data

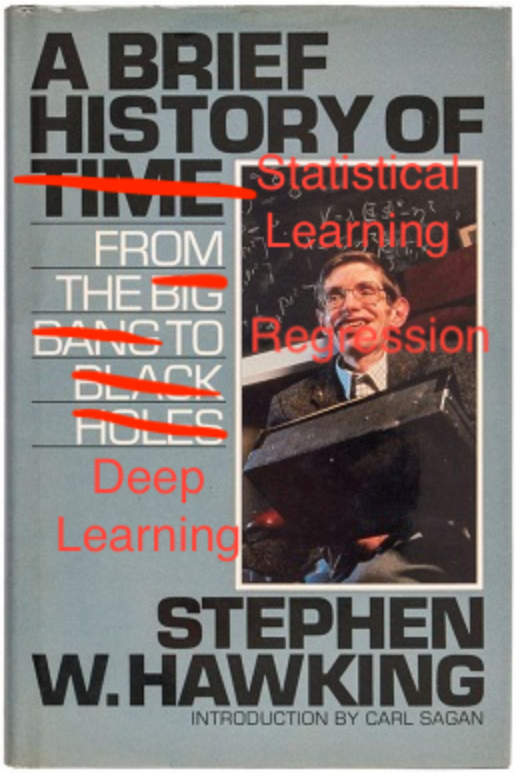
Consider the `NCI60` data, which consists of 6,830 gene expression measurements for 64 cancer lines. We are interested ind determining whether there are **groups** among the cell lines based on their gene expression measurements.

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:

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}

$\boldsymbol{X}$

$\boldsymbol{y}$

$a, \boldsymbol{A}, A$

$a \in \mathbb{R}$

Matrix multiplication