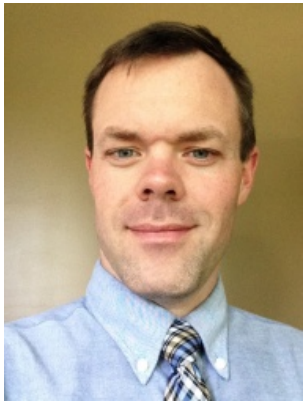
CQS Summer Institute: Machine Learning and Statistics in R

Matthew S. Shotwell, Ph.D.

Department of Biostatistics
Vanderbilt University Medical Center
Nashville, TN, USA

August 13, 2018

My Bio



- Matthew (Matt) S. Shotwell, Ph.D.
- Assoc. Prof. in Biostatistics
- 8 years at VU/VUMC
- 85% Biomed. Research / 15% Teaching
- R user 10+ years
- Teach "Statistical Learning" (BIOS 8362); 4 years
- Hastie et al. *Elements of Statistical Learning*

Why study Machine Learning and Statistics?

Data science is **HOT**. From `glassdoor.com`:

Data Scientist Salaries in Nashville, TN Area

About This Data

14 Salaries Updated Feb 20, 2018

Industries | Company Sizes | Years of Experience

Average Base Pay

$96,751 /yr

20% below national average

Not enough reports to show salary distribution

$73K	$97K	$123K
Low	Average	High

Additional Cash Compensation

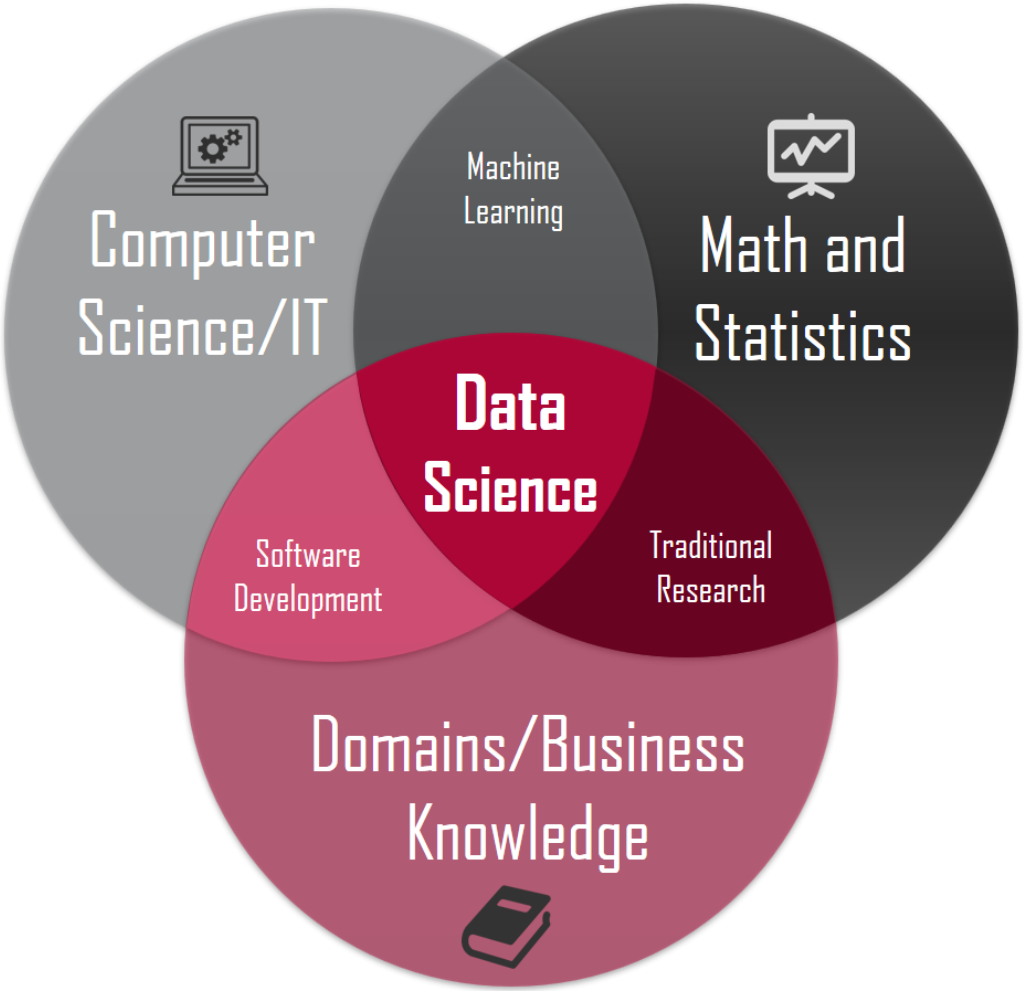
Average	$10,100
Range	$3,441 - $23,078

How much does a Data Scientist make in Nashville, TN?
The average salary for a Data Scientist is $96,751 in Nashville, TN. Salaries estimates are based on 14... More

Salaries for Related Job Titles

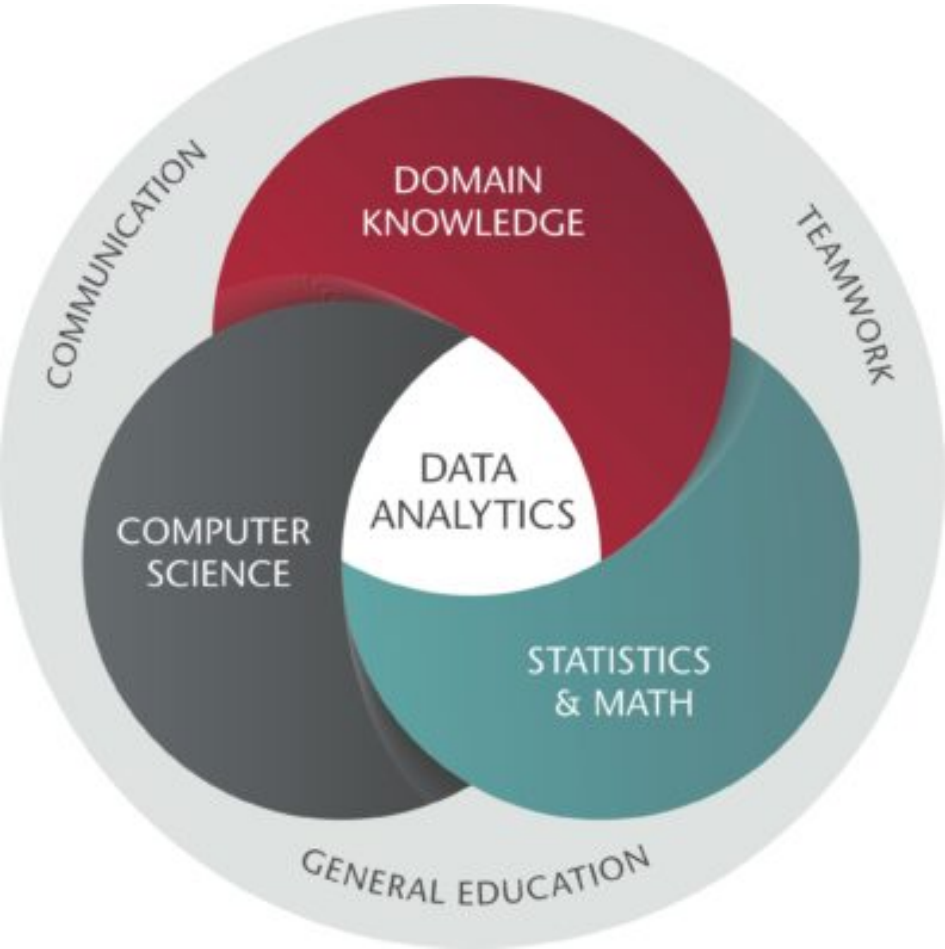
Data Analyst	$58K
Data Scientist Intern	$72K
Quantitative Analyst	$74K
Senior Data Scientist	$114K

Why study Machine Learning and Statistics?



source: https://towardsdatascience.com/introduction-to-statistics-e9d72d818745

Why study Machine Learning and Statistics?



source: https://everett.wsu.edu/majorsdegrees/data_analytics/

Why study Machine Learning and Statistics?

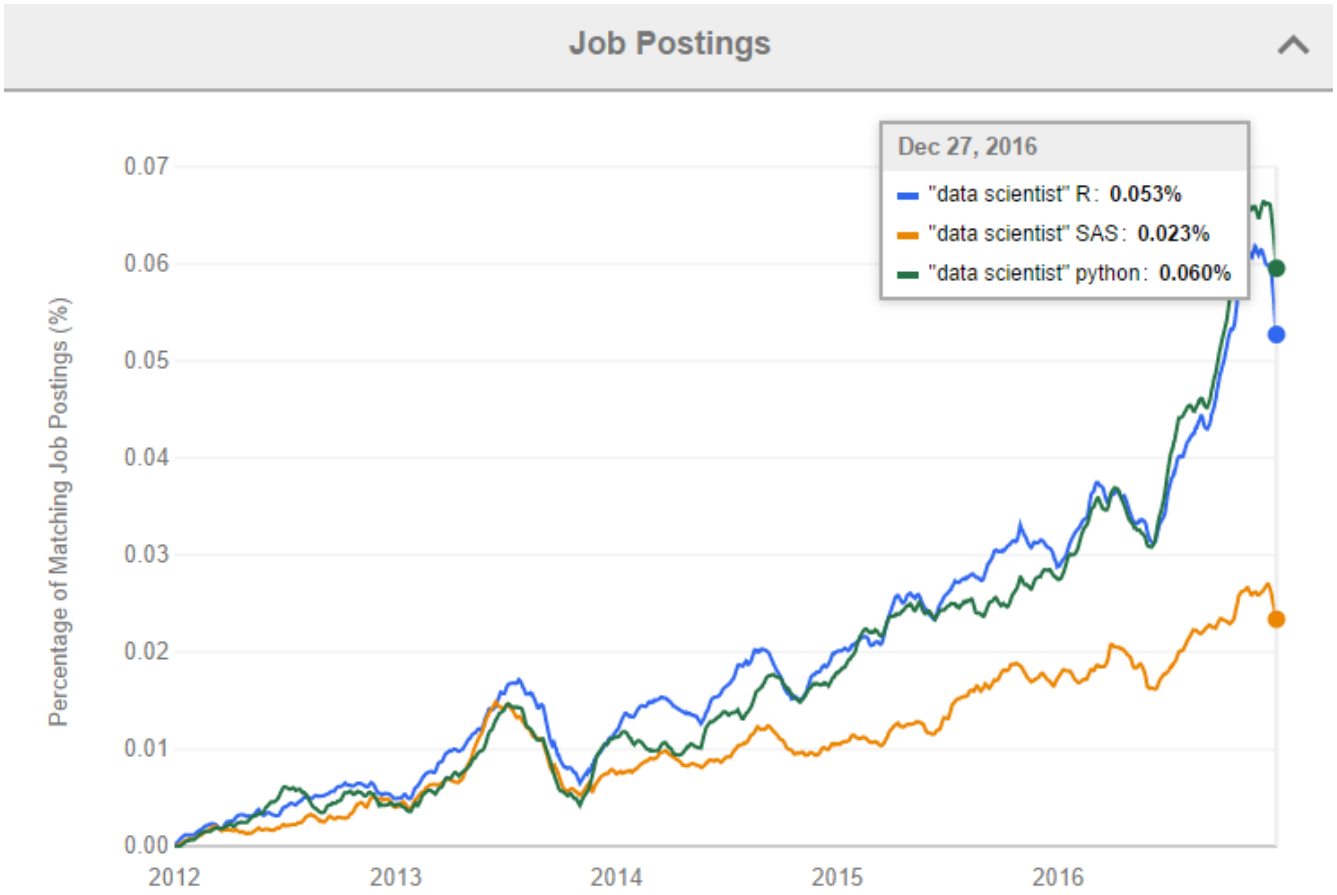
Impact!

Data scientists:

- Can contribute to almost any worthwhile effort
- Can have large-scale impact
- Are the first to "know"
- Provide crucial interpretation

Why study Machine Learning and Statistics in R?

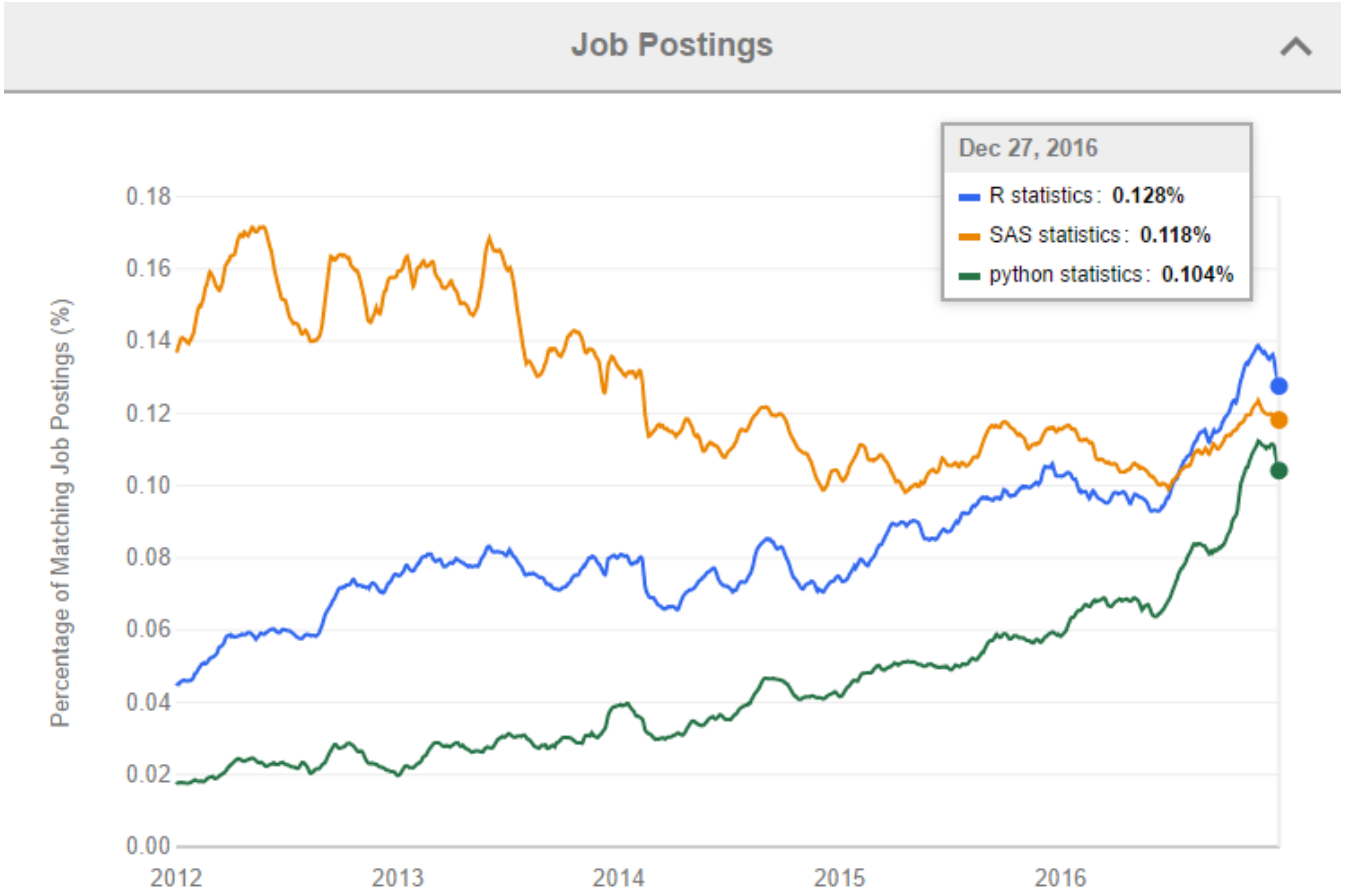
Employers are increasingly looking for data scientists and statisticians with experience using R. From `indeed.com`:



source: http://blog.revolutionanalytics.com/2017/02/job-trends-for-r-and-python.html

Why study Machine Learning and Statistics in R?

Employers are increasingly looking for data scientists and statisticians with experience using R. From `indeed.com`:



source: http://blog.revolutionanalytics.com/2017/02/job-trends-for-r-and-python.html

Course Info

- Date: Mon. Aug. 13 - Fri. Aug. 17
- Time: 1pm - 4pm
- Location: Kissam Center, Room C216

Course Structure

- Each 3h session: 3-4 modules
- Each module:
  - 20-30min presentation
  - 20-30min laboratory ("hands on")
  - 5-10min break

Course Overview

- Syllabus and R code:
- https://github.com/biostatmatt/cqs-ml-stat-r
- Monday: Intro and Data Management
- Tuesday: Supervised Learning Part 1
- Wednesday: Supervised Learning Part 2
- Thursday: Unsupervised Learning
- Friday: Statistical Inference

Intro to R and Data Management: Monday (today)

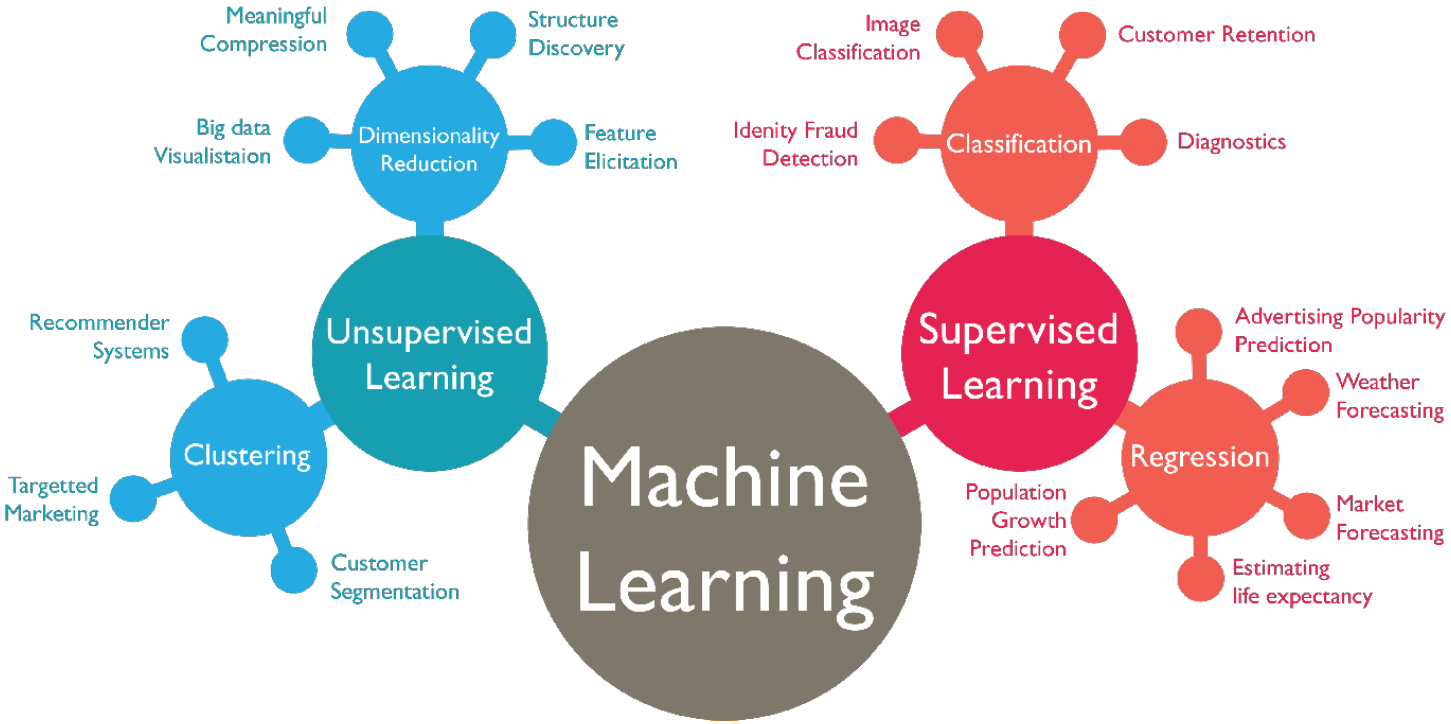
- R/RStudio
- variables and data types
- Reading/writing data
- Manipulating data (e.g., reshaping wide-to-long)

ID	T	P.1	P.2	P.3
1	24.3	10.2	5.5	2.1
2	23.4	10.4	5.7	2.8
3	22.1	10.5	5.9	3.1
4	19.9	10.2	5.2	2.4

ID	Channel	T	P
1	1	24.3	10.2
2	1	23.4	10.4
3	1	22.1	10.5
4	1	19.9	10.2
1	2	24.3	5.5
2	2	23.4	5.7
3	2	22.1	5.9
4	2	19.9	5.2
1	3	24.3	2.1
2	3	23.4	2.8
3	3	22.1	3.1
4	3	19.9	2.4

source: https://stackoverflow.com/questions/29844056/

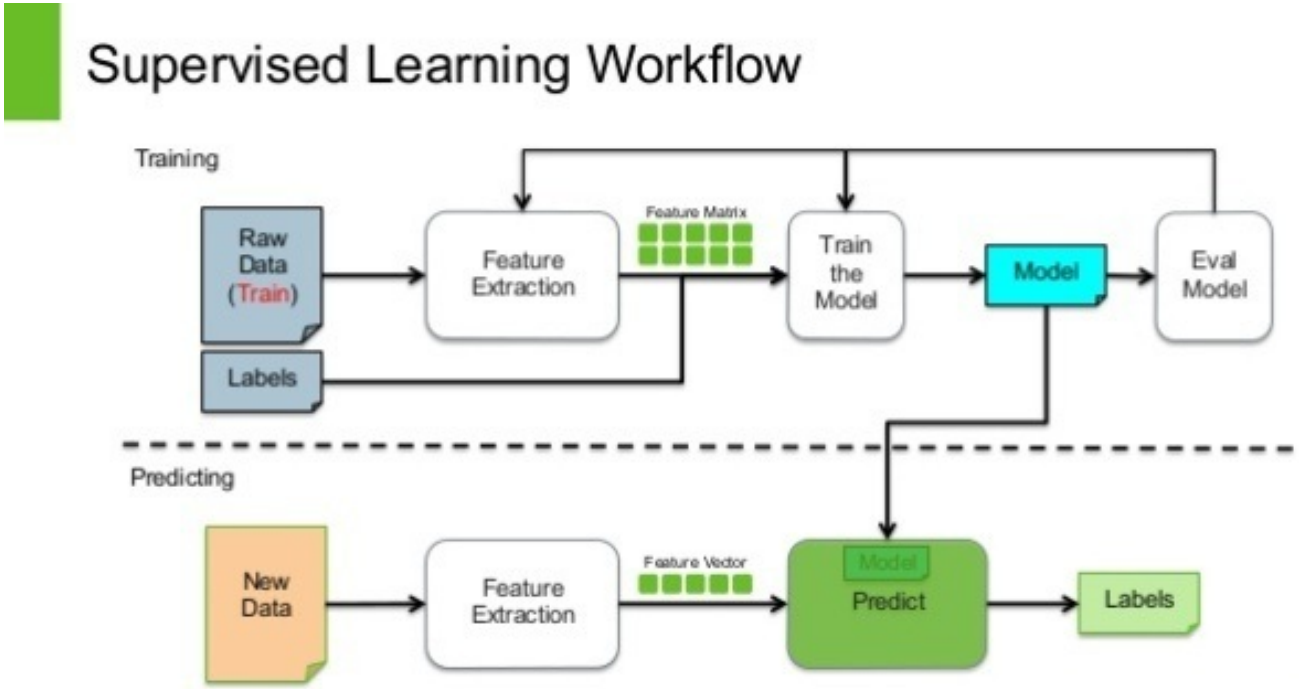
Machine learning



source: https://www.wordstream.com/blog/ws/2017/07/28/machine-learning-applications

Supervised learning: Tuesday & Wednesday

- Have input ('features') AND output ('target')
- Create a model ('learner') using observed inputs and outputs
- Goal is to predict outputs from new inputs
- "Supervised" because both inputs *and outputs* to guide model

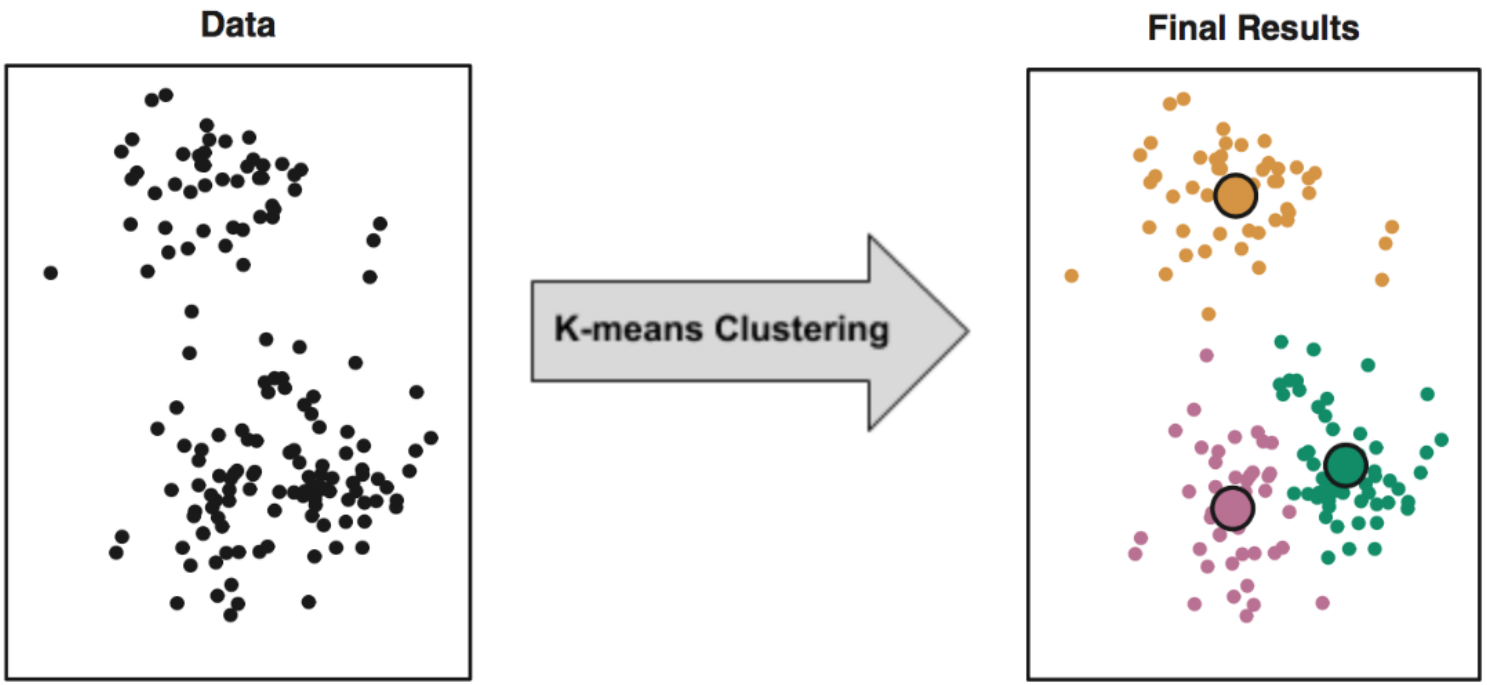


source: https://www.quora.com/What-is-pattern-recognition

FIGURE 1.1. *Scatterplot matrix of the prostate cancer data. The first row shows the response against each of the predictors in turn. Two of the predictors,* `svi` *and* `gleason`*, are categorical.*

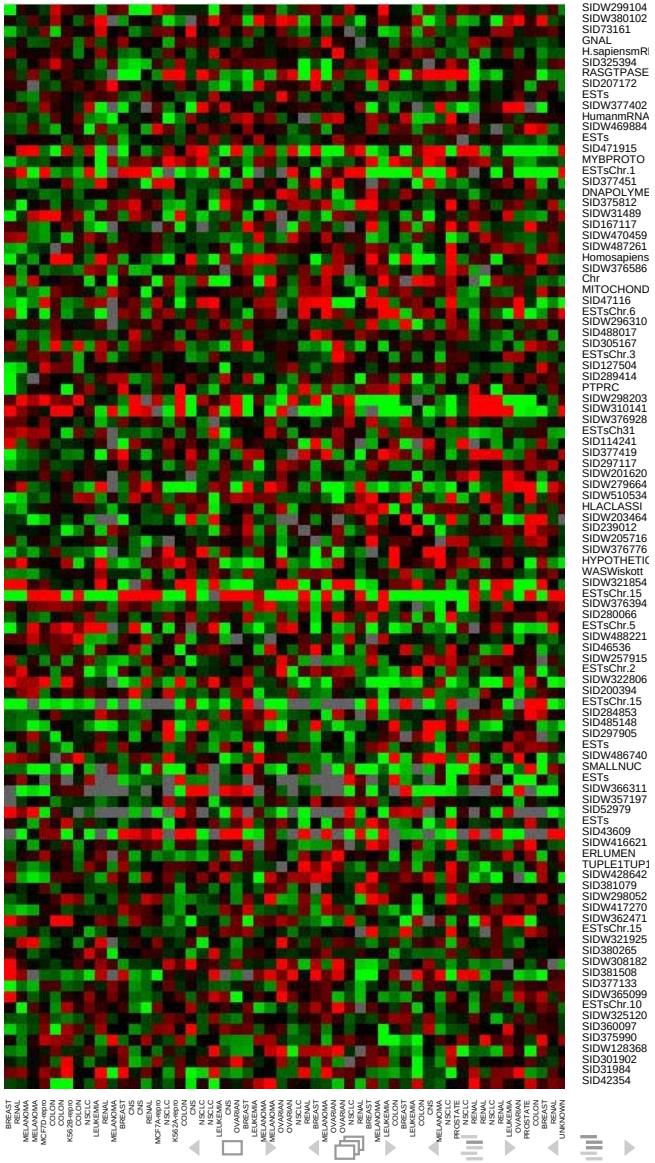
Unsupervised learning: Thursday

- Have only input, no output
- Discover organization or clustering of input



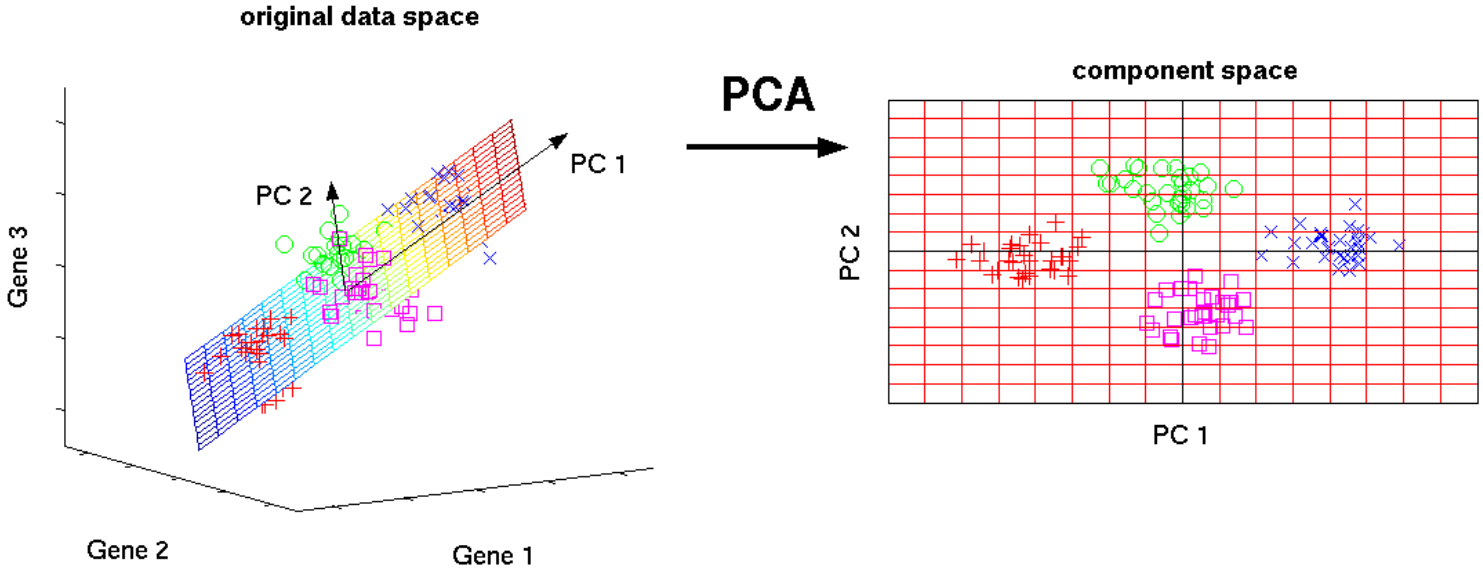
source: https://www.leverege.com/blogpost/machine-learning-course-iot

- Gene expression array
- Rows - tumor samples
- Cols - genes
- Green - overexpressed
- Red - underexpressed
- Similar samples?
- Similar genes?



Unsupervised learning: Thursday

- Have only input, no output
- Discover organization or clustering of input

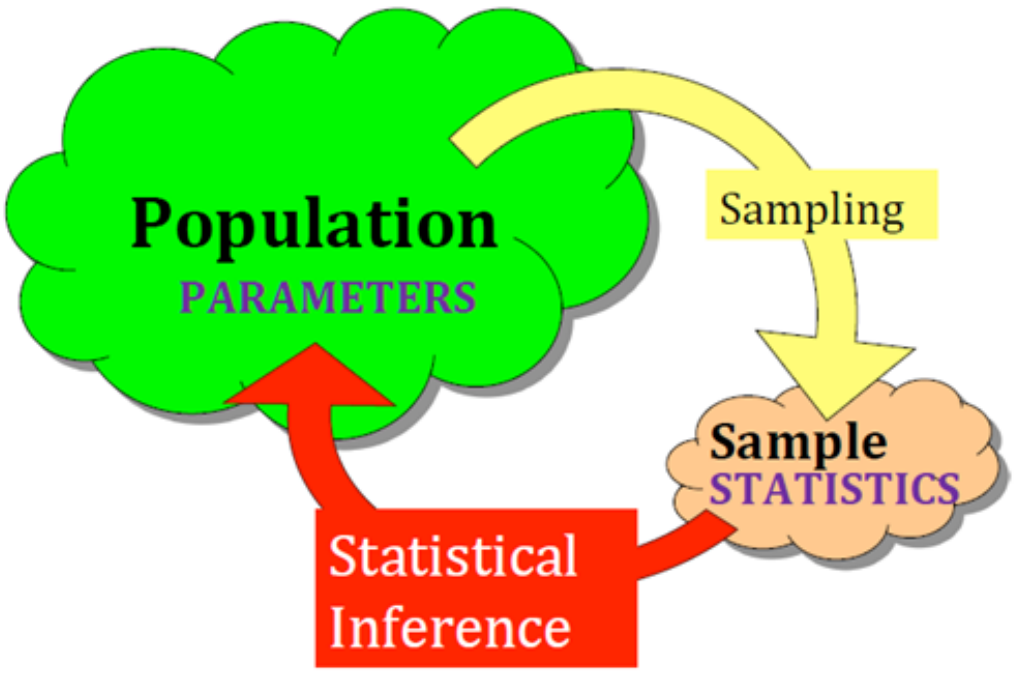


source:

https://hackernoon.com/a-laymans-introduction-to-principal-components-2fca55c19fa0

Statistical Inference: Friday

- Populations, samples, sampling biases
- How methods and tools for inference relate to those for ML
- Fundamentals of frequentist (and maybe Bayesian) statistics



source: https://mahritaharahap.wordpress.com/teaching-areas/inferential-statistics/