Introduction, Notation, and Overview

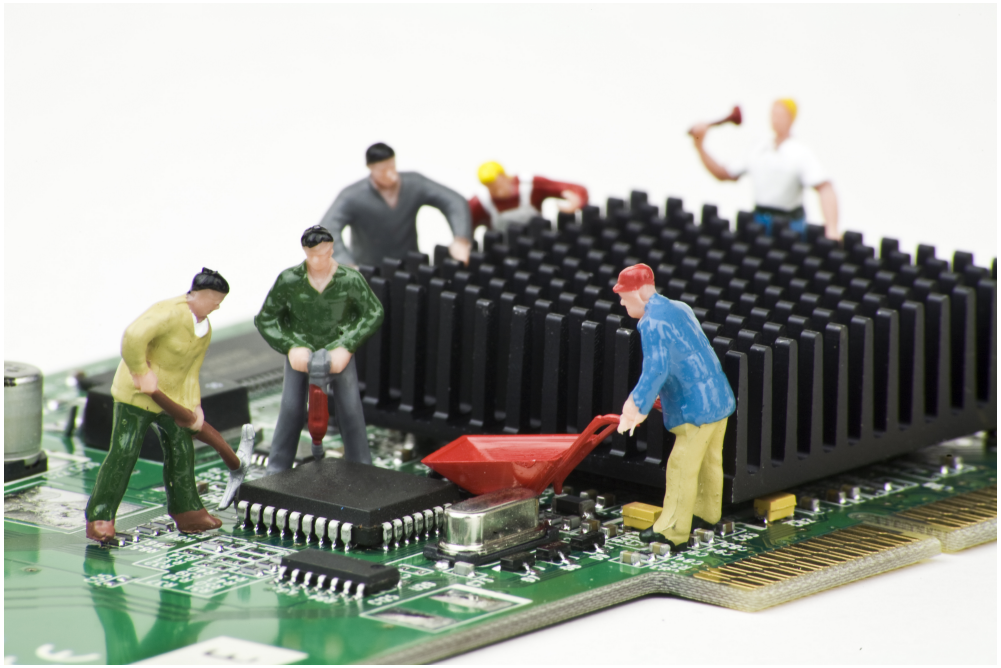
-Applied Analytics-

APM: Chapters 1.3 & 1.6

Lecturer: Darren Homrighausen, PhD

Preamble:

- Define applied analytics
- Begin an introductory example
- Go over terminology and introduce notation



Applied Analytics is about using data to make ...

- ... predictions about unknown quantities
- ... actionable insights
- ... convert data into data sets

References:

Main references:

- "Applied Predictive Modeling" (Kuhn & Johnson)
- "R for Data Science" (Grolemund & Wickham)

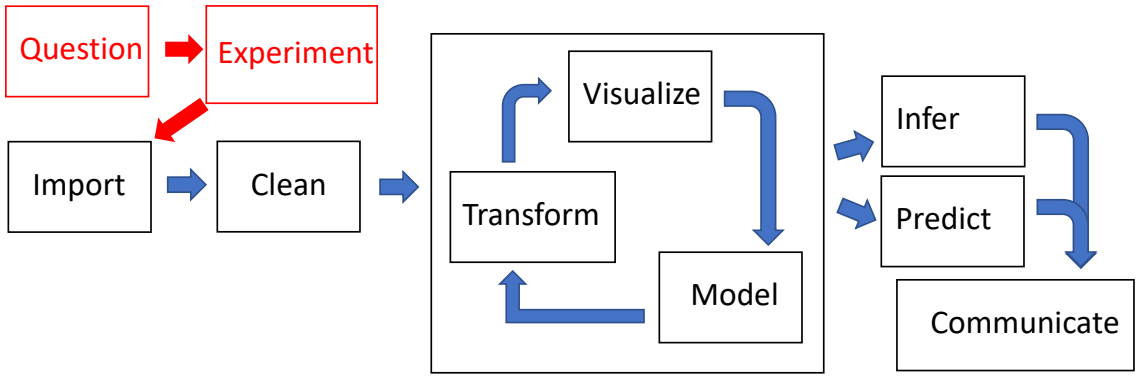
Secondary references:

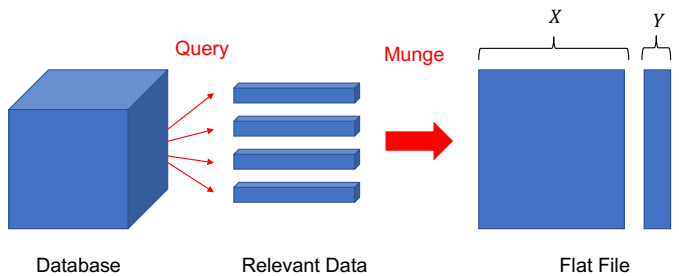
- *An Introduction to Statistical Learning w/ Applications in R* (ISL) (James, Witten, Hastie, Tibshirani)
- Topic specific notes or lectures

Wait, statistical learning & data science; what are those?

Modern data analysis techniques are found in many related topics:

- data science
- statistical (machine) learning
- statistics
- data mining





Introduction

Some common tasks we will encounter:

Tidying/transforming/wrangling: Taking real world data and converting it into a usable data set

Prediction: Making a statement about what an unobserved value is likely to be

Quantifying uncertainty: Making accurate (but not necessarily precise) ranges of values

Introductory example

Flight delays

Scientific question: Is there evidence that there is a difference in flight delays out of IAH or DFW?

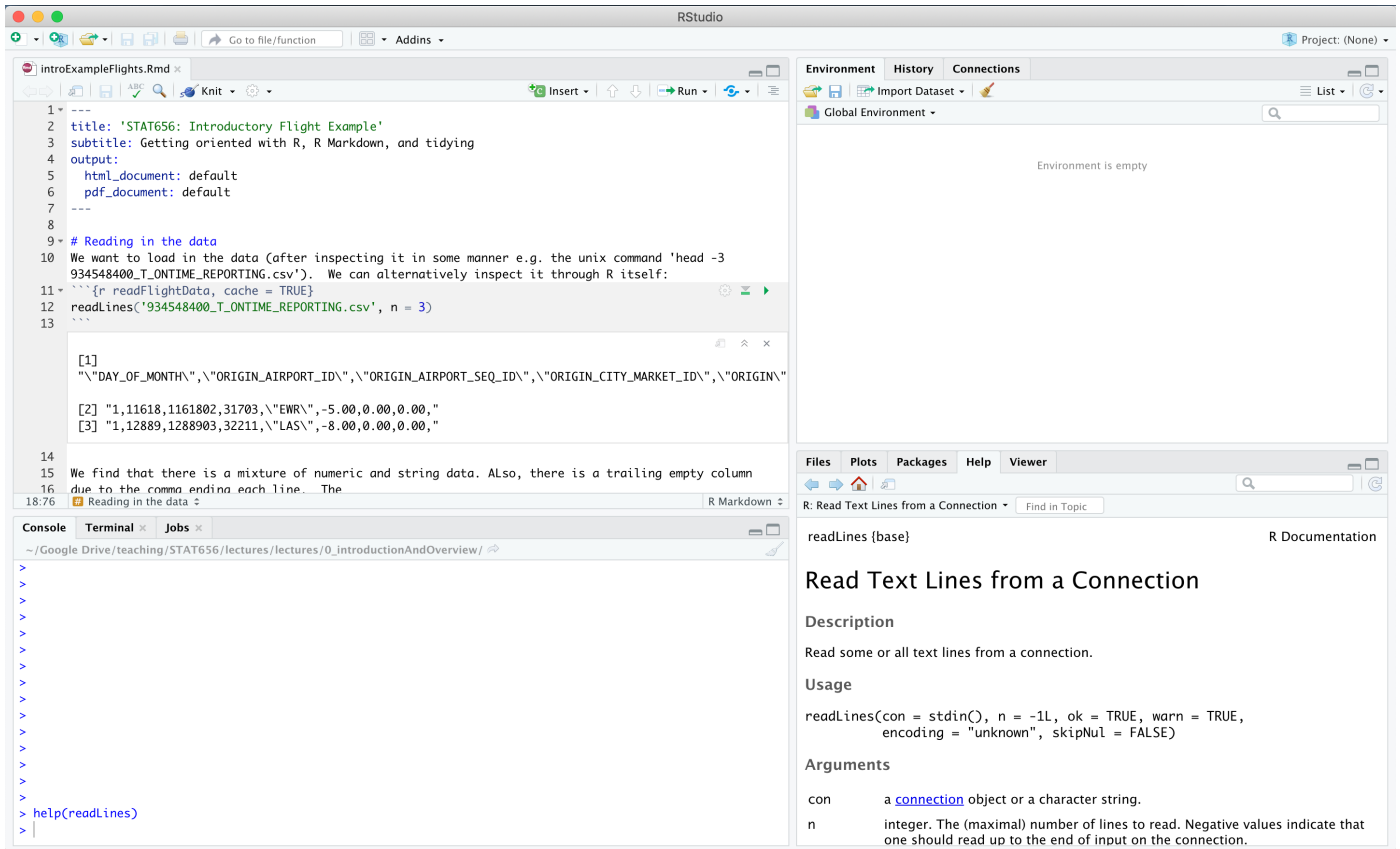
Flight information is maintained by the department of transportation:

`https://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236&DB_Short_Name=On-Time`

Flight delays

Before next class: Install R, Rstudio, and knit the "R Markdown file" introExampleFlights.Rmd into an .html file

(Hint: Go to the link on previous slide, download the .csv file (for now, choose January 2018), and make sure that file is in the same directory as the .Rmd file)



Terminology & notation

Terminology

In the introductory example, we are treating:

- 'ORIGIN' or 'DAY_OF_MONTH' as the input variables
- 'DEP_DELAY_NEW' as the output variable

In many analyses, it is a crucial and nontrivial step

Terminology

We will use notation related to X to indicate the input variable(s)

Some equivalent terms:

- predictor
- explanatory variables
- feature
- covariate
- independent variable

Likewise, notation related to Y is used for the output variable(s)

Some equivalent terms:

- response
- supervisor
- dependent variable

APM chapter 1.6

Notation summary

- We have data $\mathcal{D} = \{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\}$
  (The training data)
- X is a length p vector of measurements for each subject
  (Example: $X_i = [1, \text{income}_i, \text{education}_i]$)
- x is a length n vector of subjects for each measurement
  (Example: $x_j = [\text{income}_1, \text{income}_2, \ldots, \text{income}_n]$)
- X_{ij} is the j^{th} measurement on the i^{th} subject
  (Example: $X_{ij} = \text{income}_i$)

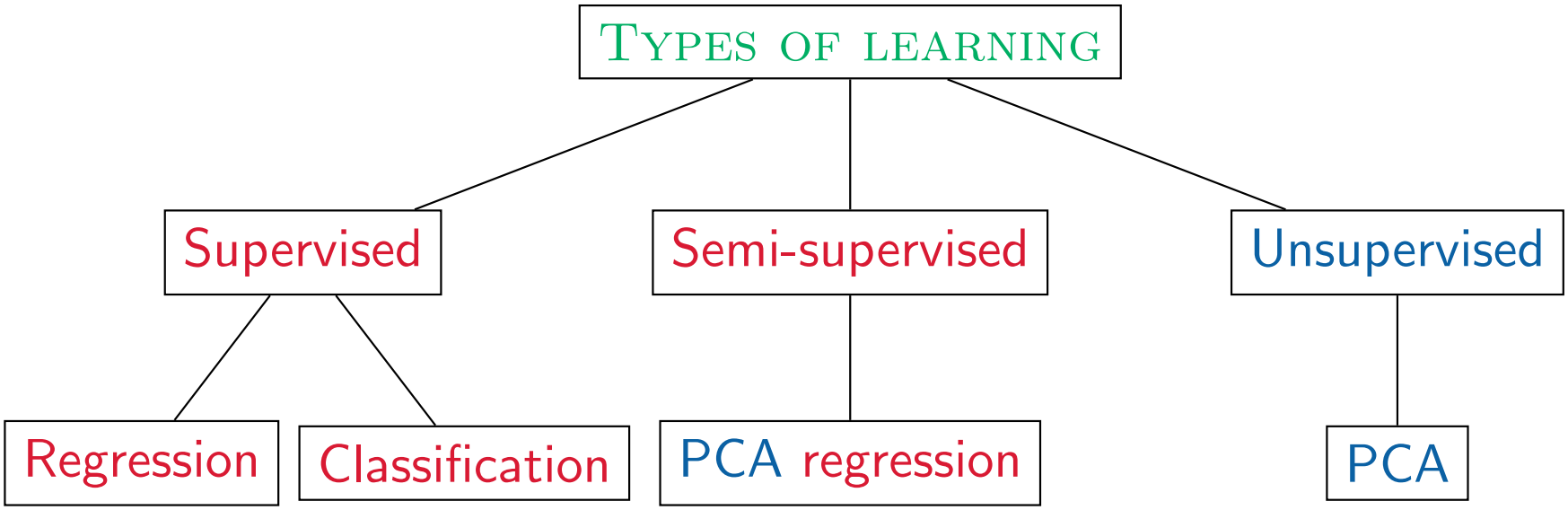
This notation is very slightly different from APM but it shouldn't cause issue

Notation

We will concatenate the features into the design or feature matrix $\mathbb{X}$, and the supervisors into the supervisor vector $\mathbb{Y}$

$$\mathbb{X} = \begin{bmatrix} | & & | \\ x_1 & \cdots & x_p \\ | & & | \end{bmatrix} = \begin{bmatrix} -X_1- \\ -X_2- \\ \vdots \\ -X_n- \end{bmatrix} \quad \text{and} \quad \mathbb{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$$

This makes $\mathbb{X}$ an $n \times p$ matrix and $\mathbb{Y}$ a length n vector.

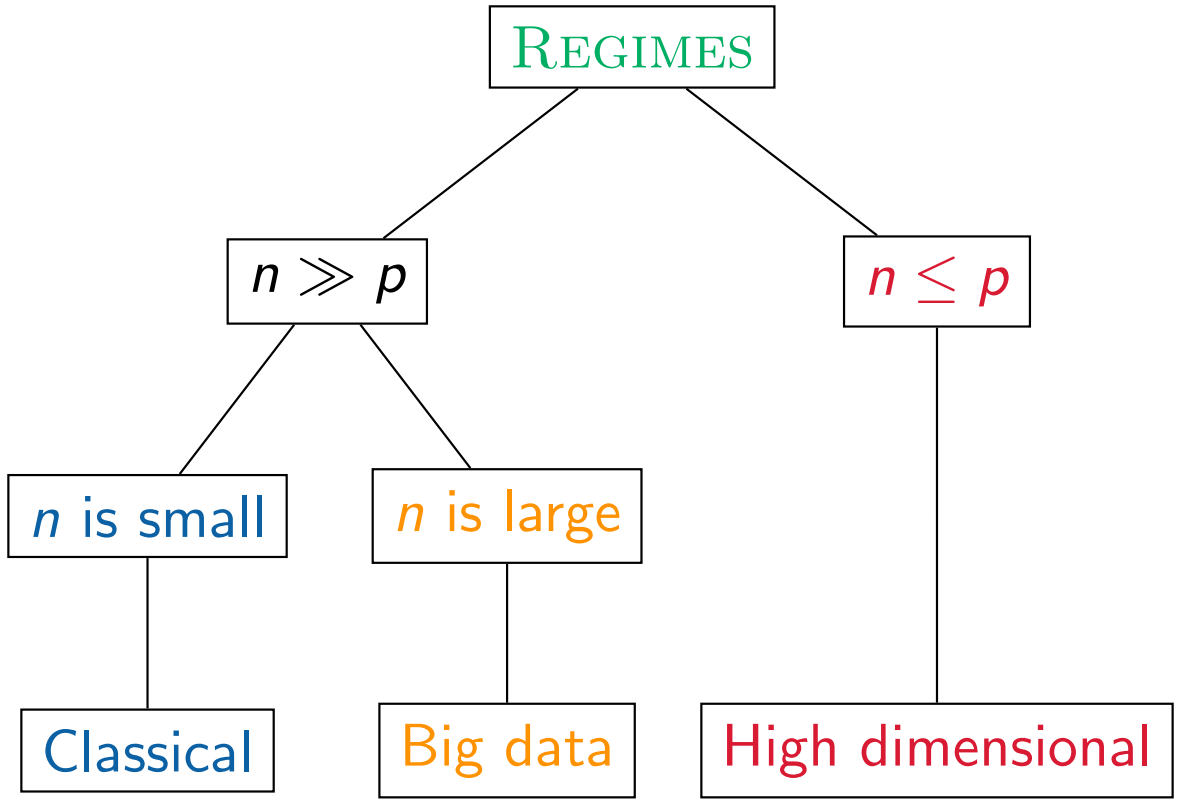


Some comments:

- Comparing predictions to Y gives a natural notion of prediction accuracy
- Much more heuristic, unclear what a good solution would be

Areas of emphasis

There are roughly three regimes of interest, assuming $\mathbb{X} \in \mathbb{R}^{n \times p}$



(We will return to these in more detail)

Postamble:

- Define applied analytics

  (We gave other related terms e.g. Data Science)
- Begin an introductory example

  (We will return to this next time)
- Go over terminology and introduce notation

  (Defined supervisor & features, along with supervised & unsupervised learning and types of data size.)