BIOST 546: Machine Learning for Biomedical Big Data

Ali Shojaie

Lecture 1: Introduction
Spring 2017

Course objectives

- An introduction to statistical machine learning methods for analysis of Biomedical Big Data
- Supervised Learning: high dimensional regression and classification, variable selection, support vector machines and random forests
- Unsupervised Learning: clustering and dimension reduction methods
- Pitfalls and challenges of statistical learning methods for analysis of Biomedical Big Data
- We will *not* cover semi-supervised learning, reinforcement learning etc

Course structure

- Homeworks: combination of applied and conceptual questions (30% of total grade)
- You are allowed (and encouraged) to work together on homework problems, but the final solution (codes, implementation, writeup) should be yours
- Project: (40% of total grade)
  - Applications of machine learning methods for analysis of biomedical data, or development of new ideas (more later)
  - Teams of 2, ideally, working on data from your own research
  - Abstract (team members, description of data, project idea etc):
  - Proposal presentation (project description and preliminary results):
  - Final presentation:
- Exam: Last day of class, or exam week, depending on how much we get to cover (30% of total grade)

Resources I

- Instructor: Ali Shojaie, PhD, Department of Biostatistics, UW
  - Office: HSB F642
  - Email: ashojaie@uw.edu
  - Phone: 616-5323
- TA: Asad Haris, Department of Biostatistics, UW
  - Email: aharis@uw.edu
- Office hours:
  - Ali: Tue 12:30-1:30pm, or by appointment.
  - Asad: Wed 10:30am - 11:50am in South Campus Center (SCC) 303, except for the following days:
    - ⋆ on 4/5, 5/3, 5/31 TA OH will be in Foege (Genome Sciences) S060
- Questions are *very* welcome during the class (please interrupt!)
- Class website:

  www.biostat.washington.edu/∼ashojaie/teaching/ML.html

  Handouts, datasets, R code etc will be provided on the website *only*

Resources II

- Recommended reading;
  - Introduction to Statistical Learning: James et al (2013), Springer (free online, includes R labs)
  - Elements of Statistical Learning: Hastie et al (2009) Springer (free online)
  - Machine Learning: Murphy (2012)

Topics

- Linear, and generalized linear regression analysis (logistic regression and survival analysis)
- Resampling methods (bagging, bootstrap)
- Multiple comparison adjustment (family wise error rates, and false discovery rate control)
- Tree-based methods (CART, random forests, and deep learning (time permitting)
- Clustering (hierarchical, k-means, model based, bi-clustering)
- Dimension reduction (principal component analysis, multi-dimensional scaling)

Throughout the course, we will discuss challenges and remedies for biomedical big data

This course

- We will cover the big ideas in statistical machine learning for biomedical big data
  - will not discuss some of the details on theory or formulations
  - will not cover implementation details (i.e. how to solve the optimization problems or code-up the algorithms)
  - will cover practical issues regarding the use of algorithms, particularly related to biomedical big data
- We will focus on using `R`

Why use R?

- Limited point-and-clickability
- Raw output is not what your co-authors (or professors) want to see
- Data manipulation non-trivial
- Formal language, can do almost anything
- Core is written by experts, also contributed packages
- Free! - and available on any sensible platform
- New ML methods implemented as R packages
- Many existing packages for processing and analyzing biomedical data
- You are welcome to use other software/programming languages, but only R is supported in class

Case Studies

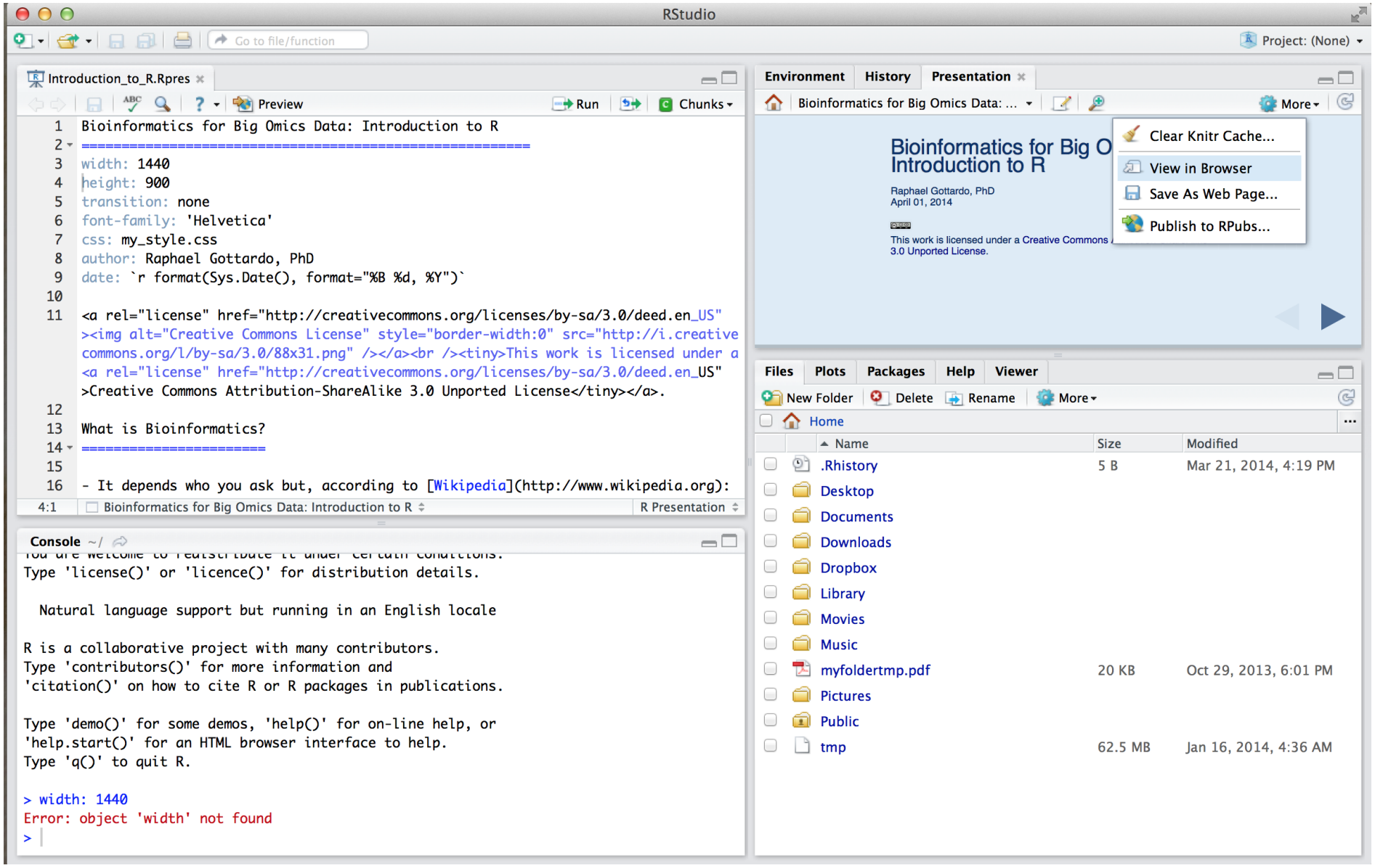
- Bring your laptop, with R installed.
- We will try out some of the labs in Introduction to Statistical Learning (ISL) as well as application cases focusing on biomedical data
- To learn more... go through the labs on your own!
  - To make sure you're ready, take a look at Lab 1 (end of Chapter 2) of ISL, and try the commands if needed!!

Prerequisites

- Two main prereqs for the class:
  - Basic statistics and probability: you need to know simple linear regression and hypothesis testing
  - Familiarity with computing: you need to be able to prepare your data for analyses in this class (data wrangling)

Prerequisites

- Two main prereqs for the class:
  - Basic statistics and probability: you need to know simple linear regression and hypothesis testing
  - Familiarity with computing: you need to be able to prepare your data for analyses in this class (data wrangling)
- This course focuses on statistical machine learning methods for analysis of biomedical data, *after the data has been preprocessed*
  - We do not focus on pre-processing of biomedical big data
  - BIOST 544 gives an *Intro to Biomedical Data Science*
  - BIOST 545 focuses on preprocessing of *omics* data. The material (lecture notes, `R` codes, etc) for BIOST 545 are available at `https://github.com/raphg/Biostat-578`
    - ⋆ You need to open the slides in `RStudio`: download the complete folder from the website, and 'open' each of the slides in `RStudio`; you can then choose either *Preview* or choose *Open in Browser* from the *More* menu in the upper right corner.


RStudio
Go to file/function
Project: (None)
Introduction_to_R.Rpres
Preview
Run
Chunks
1 Bioinformatics for Big Omics Data: Introduction to R
2 ==========
3 width: 1440
4 height: 900
5 transition: none
6 font-family: 'Helvetica'
7 css: my_style.css
8 author: Raphael Gottardo, PhD
9 date: `r format(Sys.Date(), format="%B %d, %Y")`
10
11
<tiny>This work is licensed under a Creative Commons Attribution-ShareAlike 3.0 Unported License</tiny>.
12
13 What is Bioinformatics?
14 ==========
15
16 - It depends who you ask but, according to [Wikipedia](http://www.wikipedia.org):
4:1
Bioinformatics for Big Omics Data: Introduction to R
R Presentation
Console ~/
Type 'license()' or 'licence()' for distribution details.
Natural language support but running in an English locale
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
> width: 1440
Error: object 'width' not found
>
Environment
History
Presentation
Bioinformatics for Big Omics Data: ...
More
Clear Knitr Cache...
View in Browser
Save As Web Page...
Publish to RPubs...
Bioinformatics for Big O
Introduction to R
Raphael Gottardo, PhD
April 01, 2014
This work is licensed under a Creative Commons
3.0 Unported License.
Files
Plots
Packages
Help
Viewer
New Folder
Delete
Rename
More
Home
Name
Size
Modified
.Rhistory 5 B Mar 21, 2014, 4:19 PM
Desktop
Documents
Downloads
Dropbox
Library
Movies
Music
myfoldertmp.pdf 20 KB Oct 29, 2013, 6:01 PM
Pictures
Public
tmp 62.5 MB Jan 16, 2014, 4:36 AM

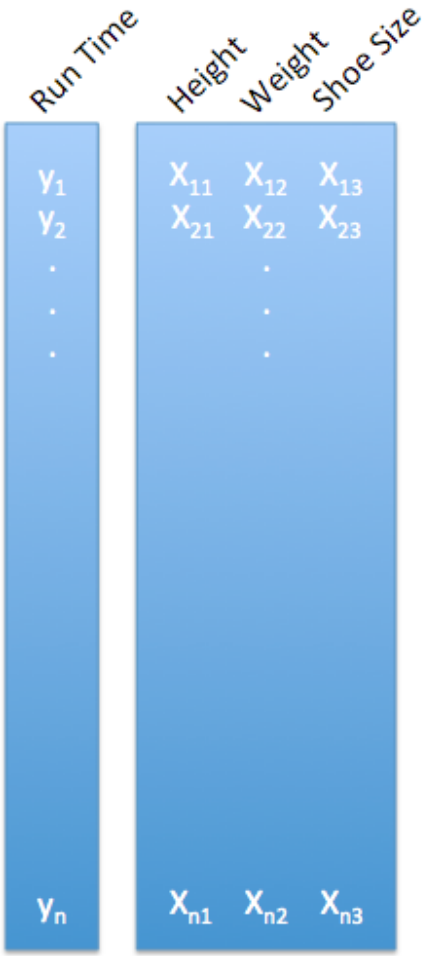
Today's lecture

- What is statistical learning?
- Supervised vs unsupervised learning
- Low-dimensional vs high-dimensional learning

A Simple Example

- Suppose we have $n = 500$ kids for whom we have $p = 3$ measurements: height, weight, and shoe size.
- We wish to predict these kids' 1600-meter run times using these measurements.

A Simple Example



Notation:

- n is the number of observations.
- p the number of variables/features/predictors.
- y is a n-vector containing response/outcome for each of n observations.
- X is a $n \times p$ data matrix.

Linear Regression on a Simple Example

- You can perform linear regression to develop a model to predict run time using height, weight, and shoe size:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

where y is run time, X_1, X_2, X_3 are height, weight, and shoe size, and ε is a noise term.

Linear Regression on a Simple Example

- You can perform linear regression to develop a model to predict run time using height, weight, and shoe size:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

  where y is run time, X_1, X_2, X_3 are height, weight, and shoe size, and ε is a noise term.

- You can look at the coefficients, p-values, and t-statistics for your linear regression model in order to interpret your results.

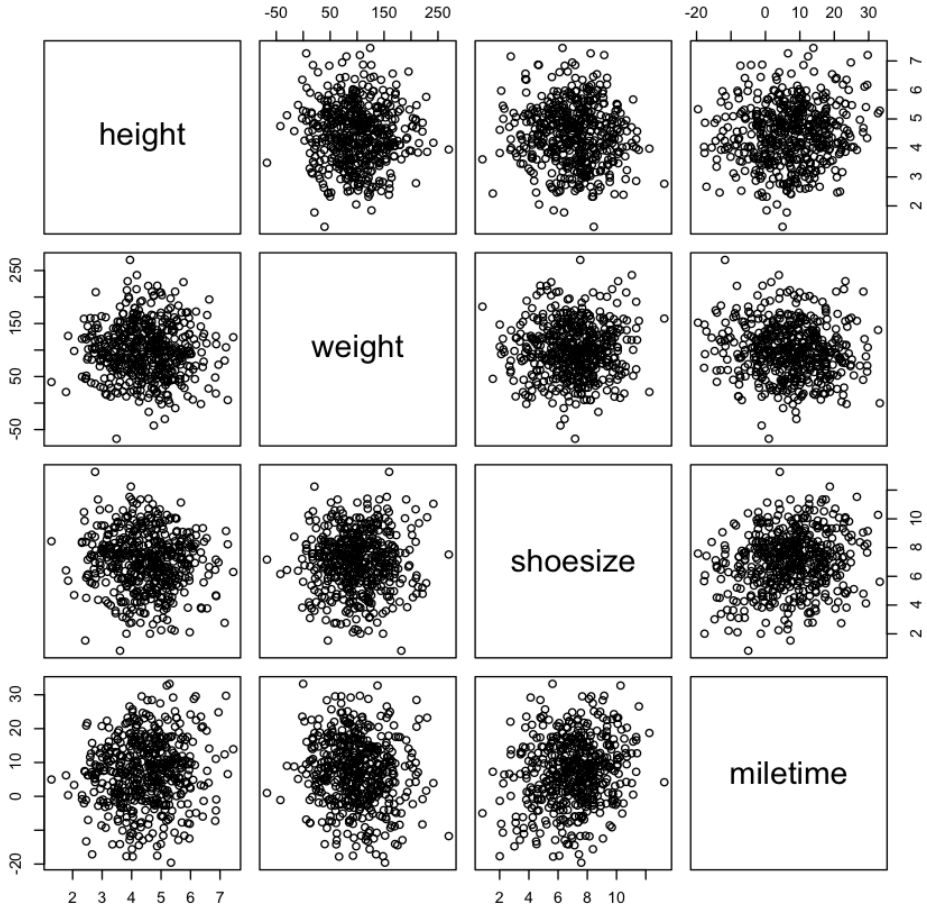
Linear Regression on a Simple Example

- You can perform linear regression to develop a model to predict run time using height, weight, and shoe size:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

  where y is run time, X_1, X_2, X_3 are height, weight, and shoe size, and ε is a noise term.
- You can look at the coefficients, p-values, and t-statistics for your linear regression model in order to interpret your results.
- You learned everything (or most of what) you need to analyze this data set in AP Statistics!

A Relationship Between the Variables?



Linear Model Output

	Estimate	Std. Error	T-Stat	P-Value
Intercept	-2.265831	2.644654	-0.857	0.39199
height	1.074814	0.414789	2.591	0.00985 **
weight	-0.021155	0.008482	-2.494	0.01295 *
shoesize	0.955222	0.214449	4.454	1.04e-05 ***

$$\text{RunTime} \approx -2.27 + 1.07 \times \text{Height} - 0.021 \times \text{Weight} + 0.96 \times \text{ShoeSize}.$$

More General Models

- The linear regression above was quite simple:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

More General Models

- The linear regression above was quite simple:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

- In general, we don't have to use linear regression, and can consider any model:

$$y = f(X_1, X_2, \ldots, X_p) + \varepsilon$$

- Here, f is any general function relating the covariates (predictors, independent variables) X to response (outcome, dependent variable) y

More General Models

- The linear regression above was quite simple:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

- In general, we don't have to use linear regression, and can consider any model:

$$y = f(X_1, X_2, \ldots, X_p) + \varepsilon$$

- Here, f is any general function relating the covariates (predictors, independent variables) X to response (outcome, dependent variable) y
- In both cases, ε is the noise term: we cannot perfectly determine y from $X_1, X_2, \ldots, X_p$, because y *is also a function of* ε, which is not observable

More General Models

- The linear regression above was quite simple:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

- In general, we don't have to use linear regression, and can consider any model:

$$y = f(X_1, X_2, \ldots, X_p) + \varepsilon$$

- Here, f is any general function relating the covariates (predictors, independent variables) X to response (outcome, dependent variable) y
- In both cases, ε is the noise term: we cannot perfectly determine y from $X_1, X_2, \ldots, X_p$, because y *is also a function of* ε, which is not observable
- We are usually not interested in the single data set that we have available: the data set is an *example* of other data sets (e.g. run times of other kids)

More General Models

More General Models

- Our goal is usually to build "models" $\hat{y} = \hat{f}(X_1, X_2, \ldots, X_p)$ to

More General Models

- Our goal is usually to build "models" $\hat{y} = \hat{f}(X_1, X_2, \ldots, X_p)$ to
  - predict the run time for other kids, *not in our sample*

More General Models

- Our goal is usually to build "models" $\hat{y} = \hat{f}(X_1, X_2, \ldots, X_p)$ to
  - predict the run time for other kids, *not in our sample*
  - make inference about covariates that are associated with the response, or their relationship *beyond our sample*

More General Models

- Our goal is usually to build "models" $\hat{y} = \hat{f}(X_1, X_2, \ldots, X_p)$ to
  - predict the run time for other kids, *not in our sample*
  - make inference about covariates that are associated with the response, or their relationship *beyond our sample*
- To find a good choice of f, then we usually try to minimize

$$E(y - \hat{y})^2$$

More General Models

- Our goal is usually to build "models" $\hat{y} = \hat{f}(X_1, X_2, \ldots, X_p)$ to
  - predict the run time for other kids, *not in our sample*
  - make inference about covariates that are associated with the response, or their relationship *beyond our sample*
- To find a good choice of f, then we usually try to minimize

$$E(y - \hat{y})^2$$

- However,

$$E(y - \hat{y})^2 = E\left(\hat{f} - f\right)^2 + \text{Var}(\varepsilon)$$

More General Models

- Our goal is usually to build "models" $\hat{y} = \hat{f}(X_1, X_2, \ldots, X_p)$ to
  - predict the run time for other kids, *not in our sample*
  - make inference about covariates that are associated with the response, or their relationship *beyond our sample*
- To find a good choice of f, then we usually try to minimize

$$E(y - \hat{y})^2$$

- However,

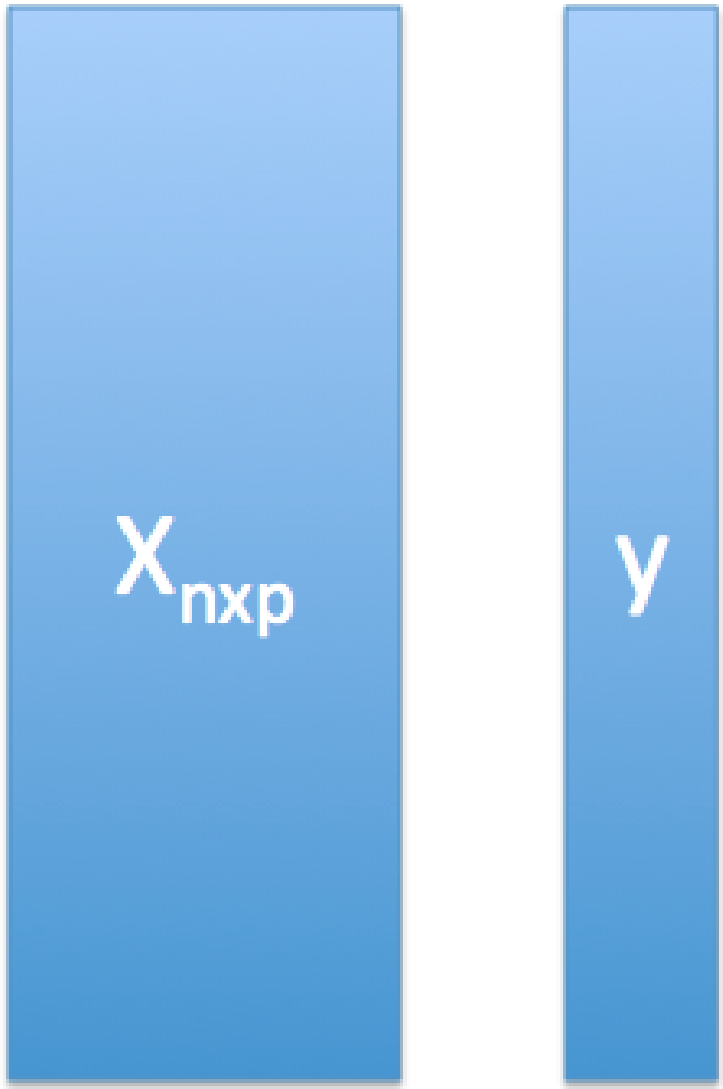
$$E(y - \hat{y})^2 = E\left(\hat{f} - f\right)^2 + \mathrm{Var}(\varepsilon)$$

  - $E\left(\hat{f} - f\right)^2$ is called the reducible error
  - $\mathrm{Var}(\varepsilon)$ is called the irreducible error
- These errors are with respect to the 'distribution of data' ...

Low-Dimensional Versus High-Dimensional

- The data set that we just saw is low-dimensional: $n \gg p$.
- Lots of the data sets coming out of modern biological techniques are high-dimensional: $n \approx p$ or $n \ll p$.
- This poses statistical challenges! AP Statistics no longer applies.

Low Dimensional



High Dimensional

What Goes Wrong in High Dimensions?

- Suppose that we include many more predictors in our model, such as
  - 50-yard dash time
  - Age
  - Zodiac symbol
  - Favorite color
  - Mother's birthday, in base 2

What Goes Wrong in High Dimensions?

- Suppose that we include many more predictors in our model, such as
  - 50-yard dash time
  - Age
  - Zodiac symbol
  - Favorite color
  - Mother's birthday, in base 2
- Some of these predictors are useful, others aren't.

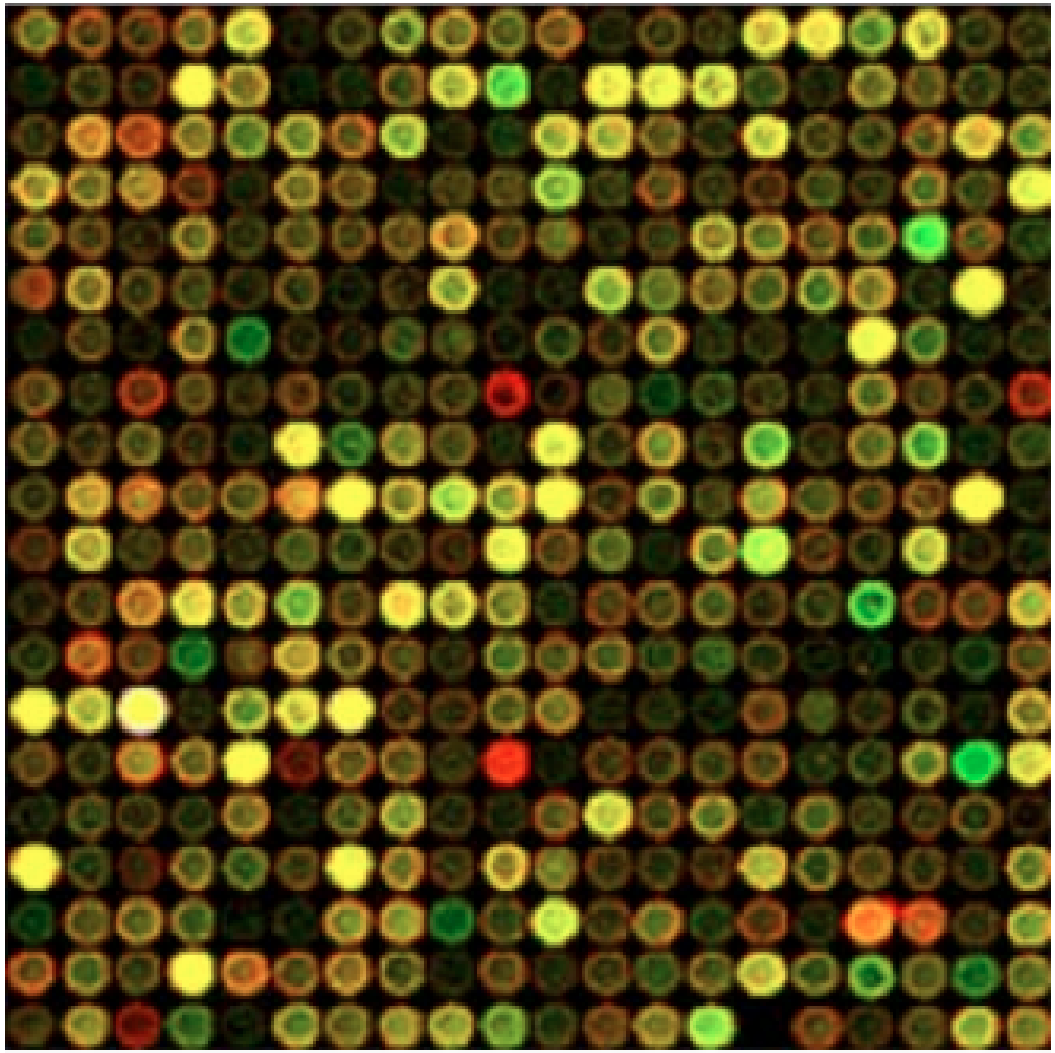
What Goes Wrong in High Dimensions?

- Suppose that we include many more predictors in our model, such as
  - 50-yard dash time
  - Age
  - Zodiac symbol
  - Favorite color
  - Mother's birthday, in base 2
- Some of these predictors are useful, others aren't.
- If we include too many predictors, we will overfit the data.
- Overfitting: Model looks great on the data used to develop it, but will perform very poorly on future observations.

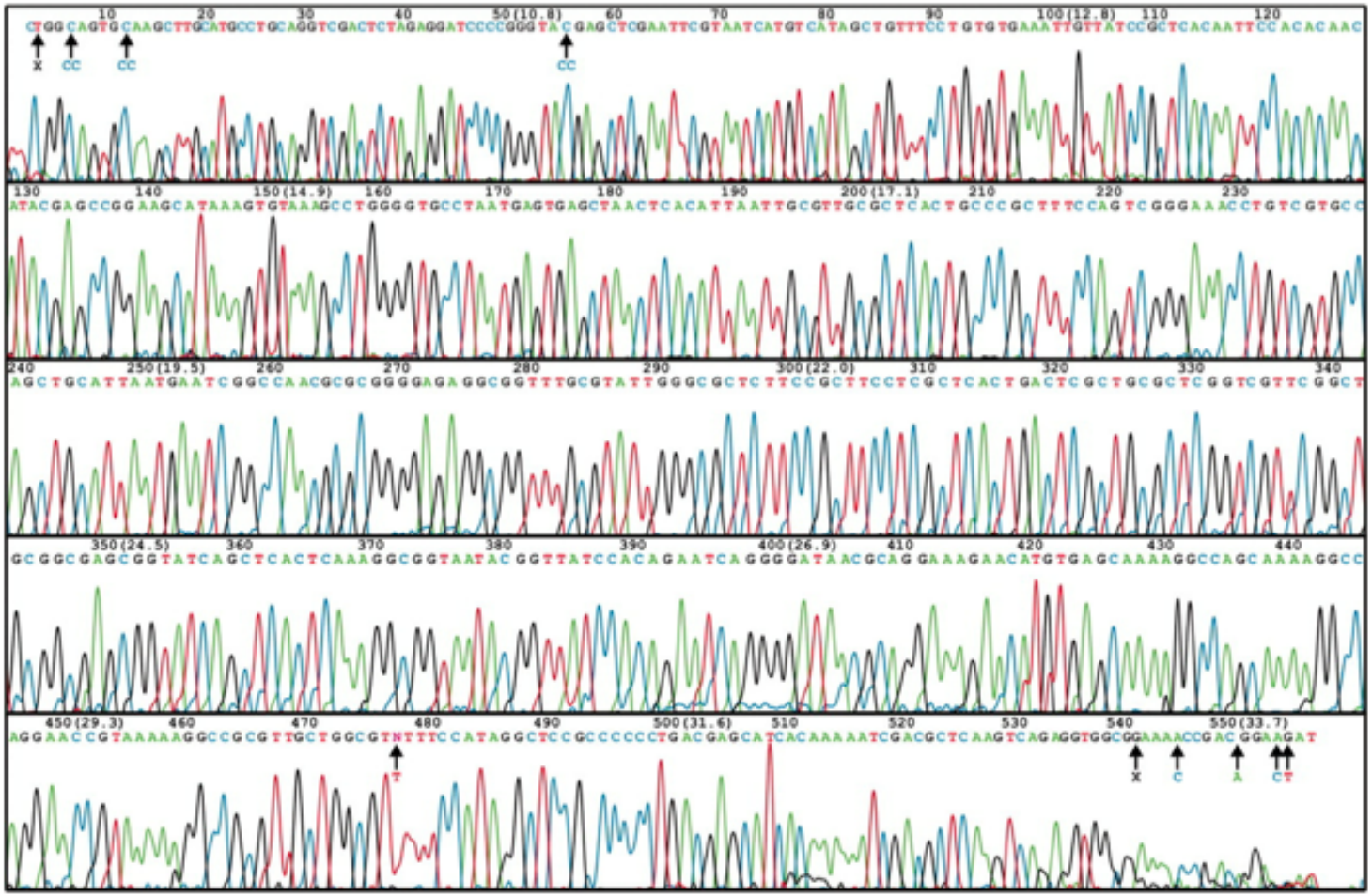
What Goes Wrong in High Dimensions?

- Suppose that we include many more predictors in our model, such as
  - 50-yard dash time
  - Age
  - Zodiac symbol
  - Favorite color
  - Mother's birthday, in base 2
- Some of these predictors are useful, others aren't.
- If we include too many predictors, we will overfit the data.
- Overfitting: Model looks great on the data used to develop it, but will perform very poorly on future observations.
- When $p \approx n$ or $p > n$, overfitting is guaranteed unless we are very careful.

Gene Expression Data

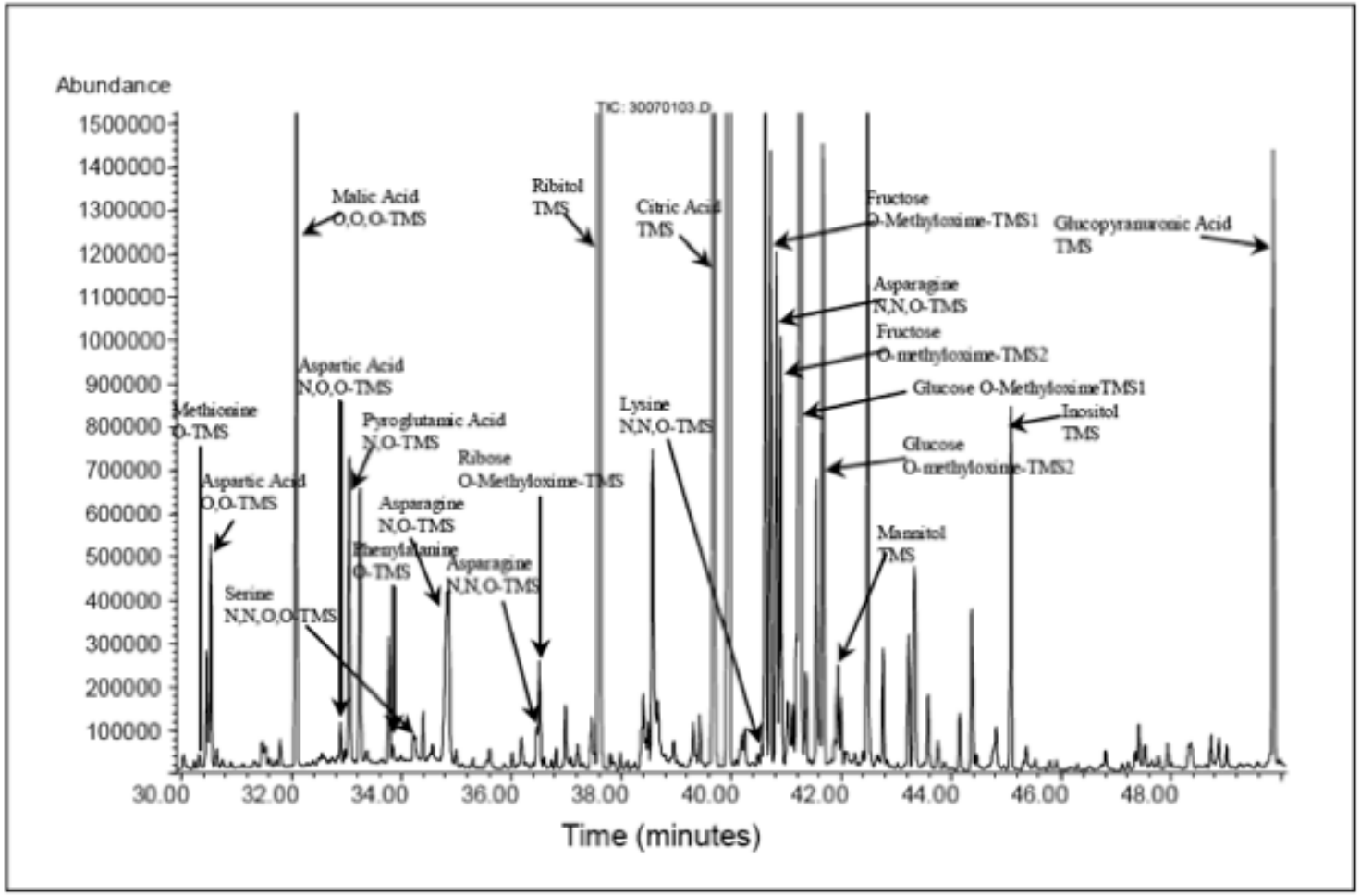


DNA Sequence Data

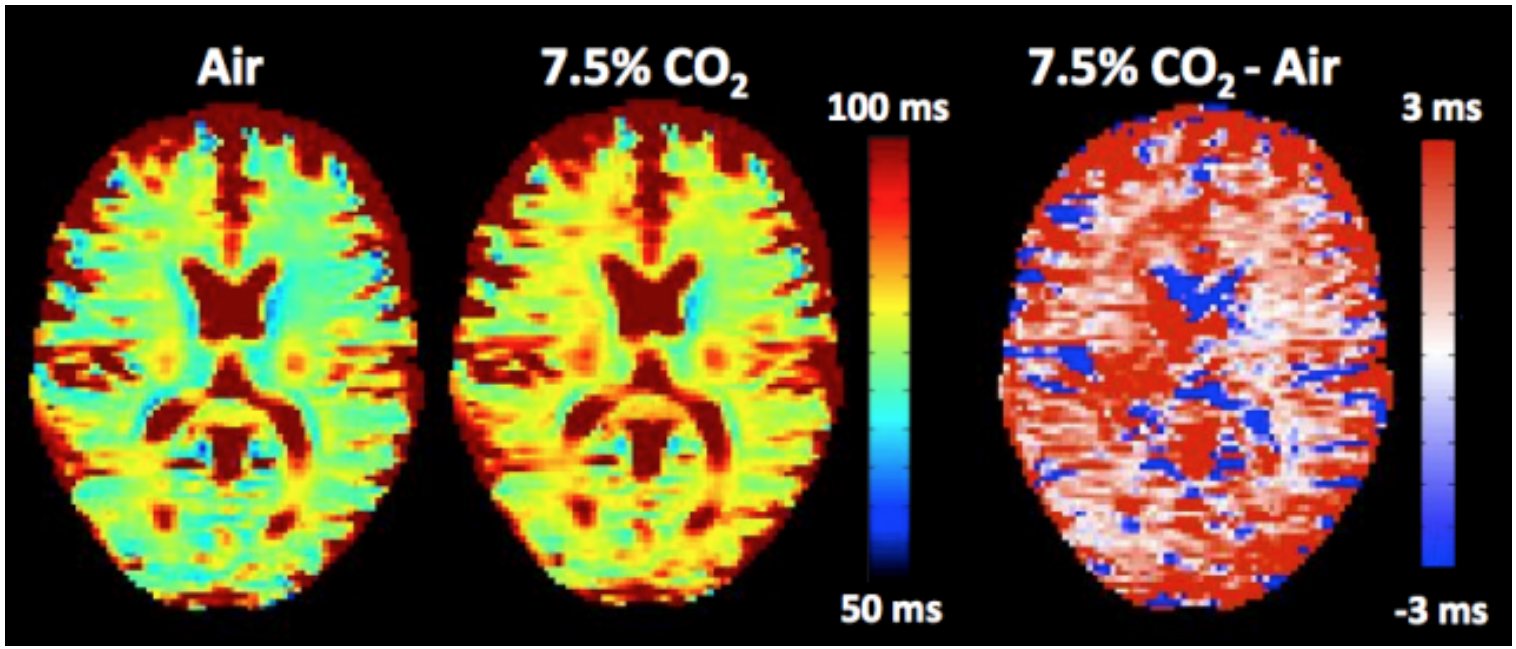


DNAse Hypersensitivity Data

Metabolomic Data



Brain Imaging Data



Electronic Health Records

Biomedical Big Data Analyses

For most biomedical data analyses (e.g. omics data, imaging etc), we have many more variables than observations.... i.e. $p \gg n$.

Biomedical Big Data Analyses

For most biomedical data analyses (e.g. omics data, imaging etc), we have many more variables than observations.... i.e. $p \gg n$.

- Predict risk of diabetes on the basis of DNA sequence data.... using $n = 1000$ patients and $p = 3,000,000$ variables.

Biomedical Big Data Analyses

For most biomedical data analyses (e.g. omics data, imaging etc), we have many more variables than observations.... i.e. $p \gg n$.

- Predict risk of diabetes on the basis of DNA sequence data.... using $n = 1000$ patients and $p = 3,000,000$ variables.
- Cluster tissue samples on the basis of DNase hypersensitivity... using $n = 200$ cell types and $p = 1,000,000,000$ variables.

Biomedical Big Data Analyses

For most biomedical data analyses (e.g. omics data, imaging etc), we have many more variables than observations.... i.e. $p \gg n$.

- Predict risk of diabetes on the basis of DNA sequence data.... using $n = 1000$ patients and $p = 3,000,000$ variables.
- Cluster tissue samples on the basis of DNase hypersensitivity... using $n = 200$ cell types and $p = 1,000,000,000$ variables.
- Identify subset of $p = 20,000$ brain regions (variables) whose activities are associated with onset of Alzheimer's disease... using images from $n = 250$ subjects (healthy and diseased).

Why Does Dimensionality Matter?

- Classical statistical techniques, such as linear regression, *cannot* be applied.
- Even very simple tasks, like identifying variables that are associated with a response, must be done with care.
- High risks of overfitting, false positives, and more.

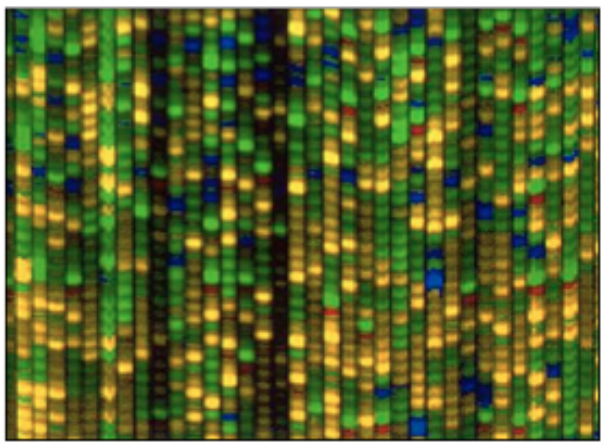
Why Does Dimensionality Matter?

- Classical statistical techniques, such as linear regression, *cannot* be applied.
- Even very simple tasks, like identifying variables that are associated with a response, must be done with care.
- High risks of overfitting, false positives, and more.

This course: Statistical machine learning tools to obtain generalizable insight from Biomedical Big Data.

Statistical Machine Learning

Google™



Supervised and Unsupervised Learning

- Statistical machine learning can be divided into two main areas: supervised and unsupervised.

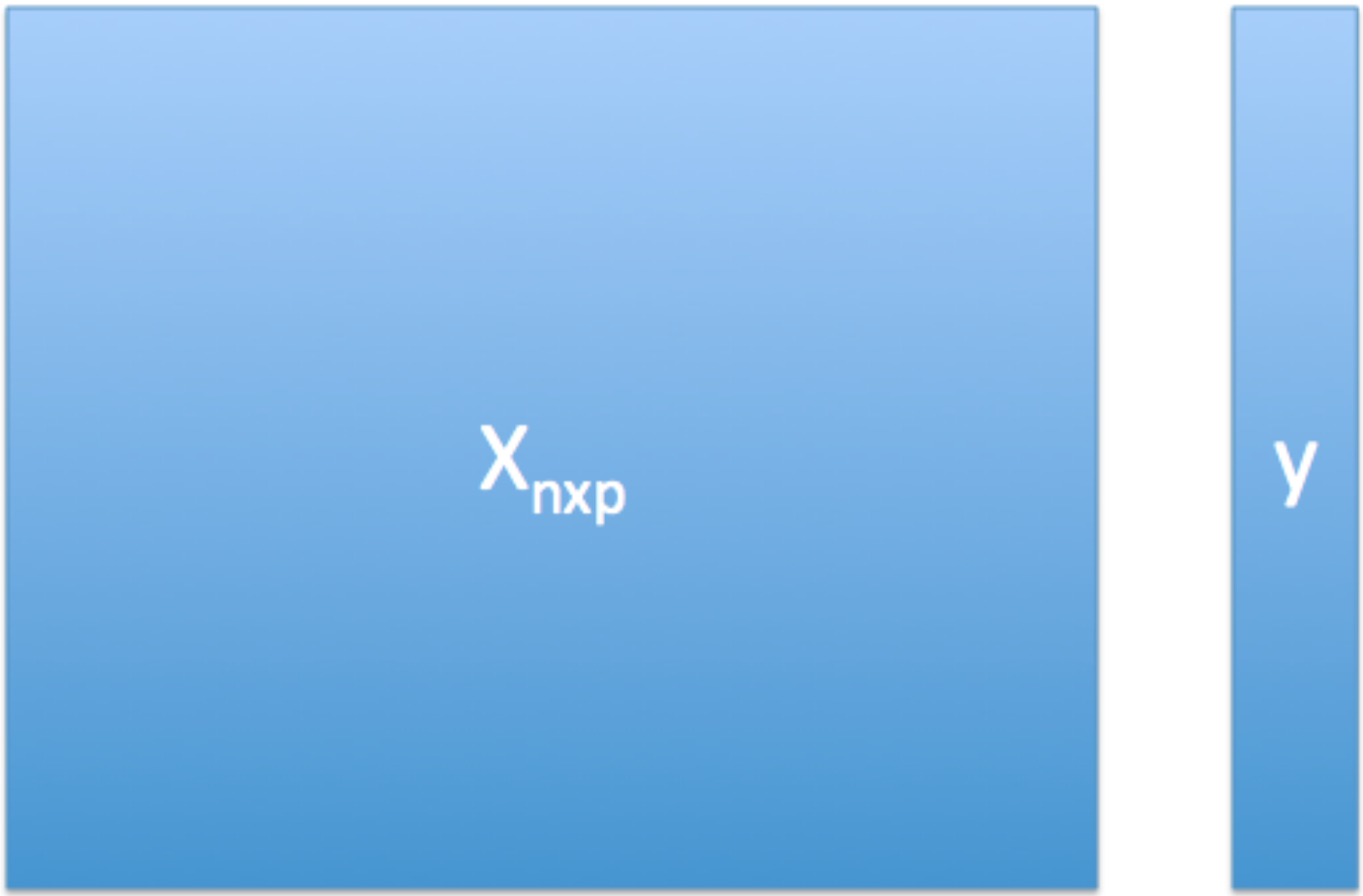
Supervised and Unsupervised Learning

- Statistical machine learning can be divided into two main areas: supervised and unsupervised.
- Supervised Learning: Use a data set X to predict or detect association with a response y.
  - Regression
  - Classification
  - Hypothesis Testing

Supervised and Unsupervised Learning

- Statistical machine learning can be divided into two main areas: supervised and unsupervised.
- Supervised Learning: Use a data set X to predict or detect association with a response y.
  - Regression
  - Classification
  - Hypothesis Testing
- Unsupervised Learning: Discover the signal/patterns in X, or detect associations within X.
  - Dimension Reduction
  - Clustering

Supervised Learning



Unsupervised Learning

Next Lecture

- A review of regression
- Training and test errors
- Cross validation