SFU SIMON FRASER UNIVERSITY
THINKING OF THE WORLD

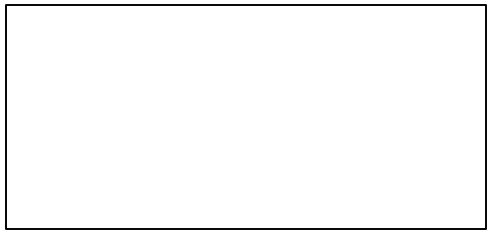
STAT 452/652
Statistical Learning and Prediction

Lecture 1: Introduction

What is “Statistical Learning”?

Lecture 1: Introduction

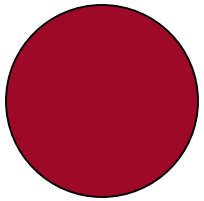
Statistical Problems



Statistics: 90* Years Ago

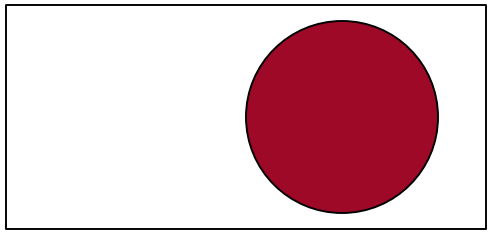
*I made this number up.

Methods taught in Intro Stat Course



Statistics: 90 Years Ago

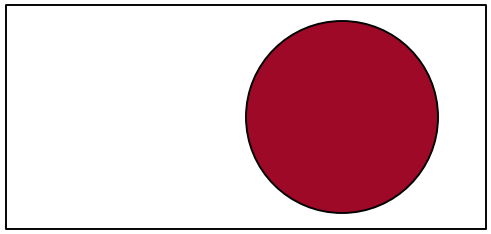
…which was a pretty good deal



Statistics: 90 Years Ago

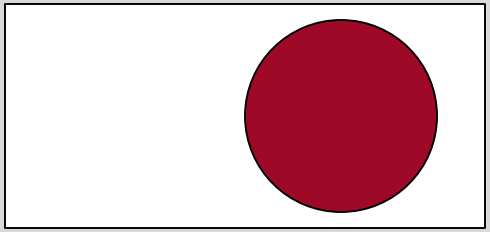
Biggest Complaint:

"I wish I had more data!"



Statistics: 90 Years Ago

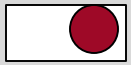
(Not to scale)



Statistics: Today

(To scale*)

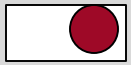
*I may be exaggerating slightly**



**Maybe more than “slightly”

Statistics: Today

- Real-time monitoring of internet traffic
- Patient records from everyone in BC
- Billions of credit card transactions
- GENES! ("You" = 3 billion+ base pairs)

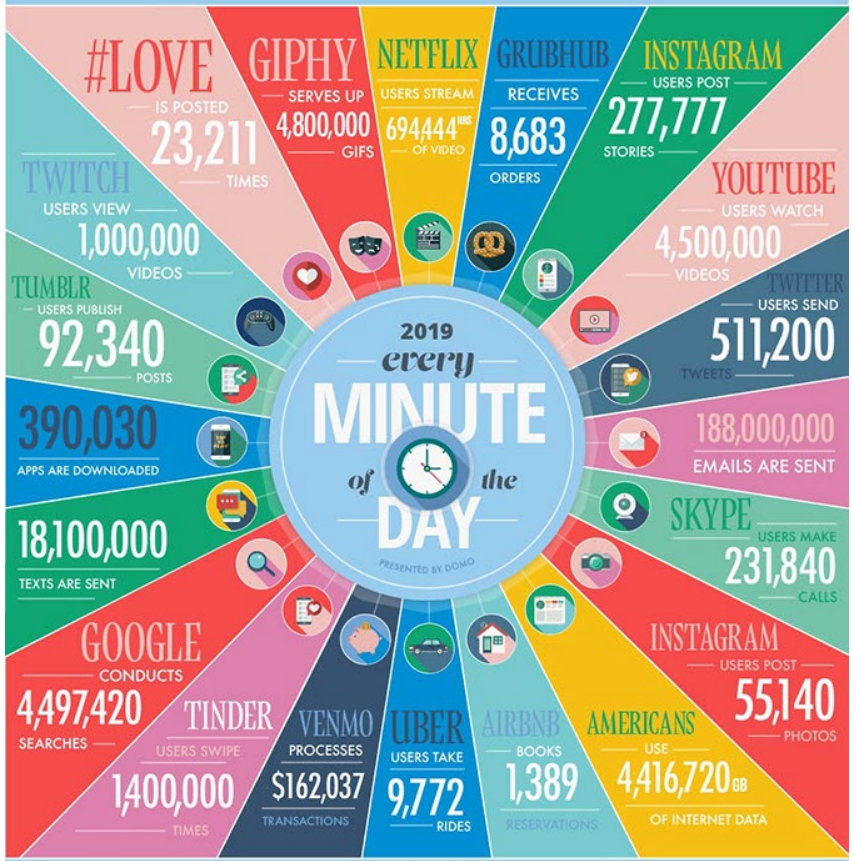


Statistics: Today

DATA NEVER SLEEPS 7.0

How much data is generated *every minute?*

There's no way around it: big data just keeps getting bigger. The numbers are staggering, and they're not slowing down. By 2020, there will be 40x more bytes of data than there are stars in the observable universe. In our 7th edition of Data Never Sleeps, we bring you the latest stats on how much data is being created in every digital minute — and the numbers are staggering.



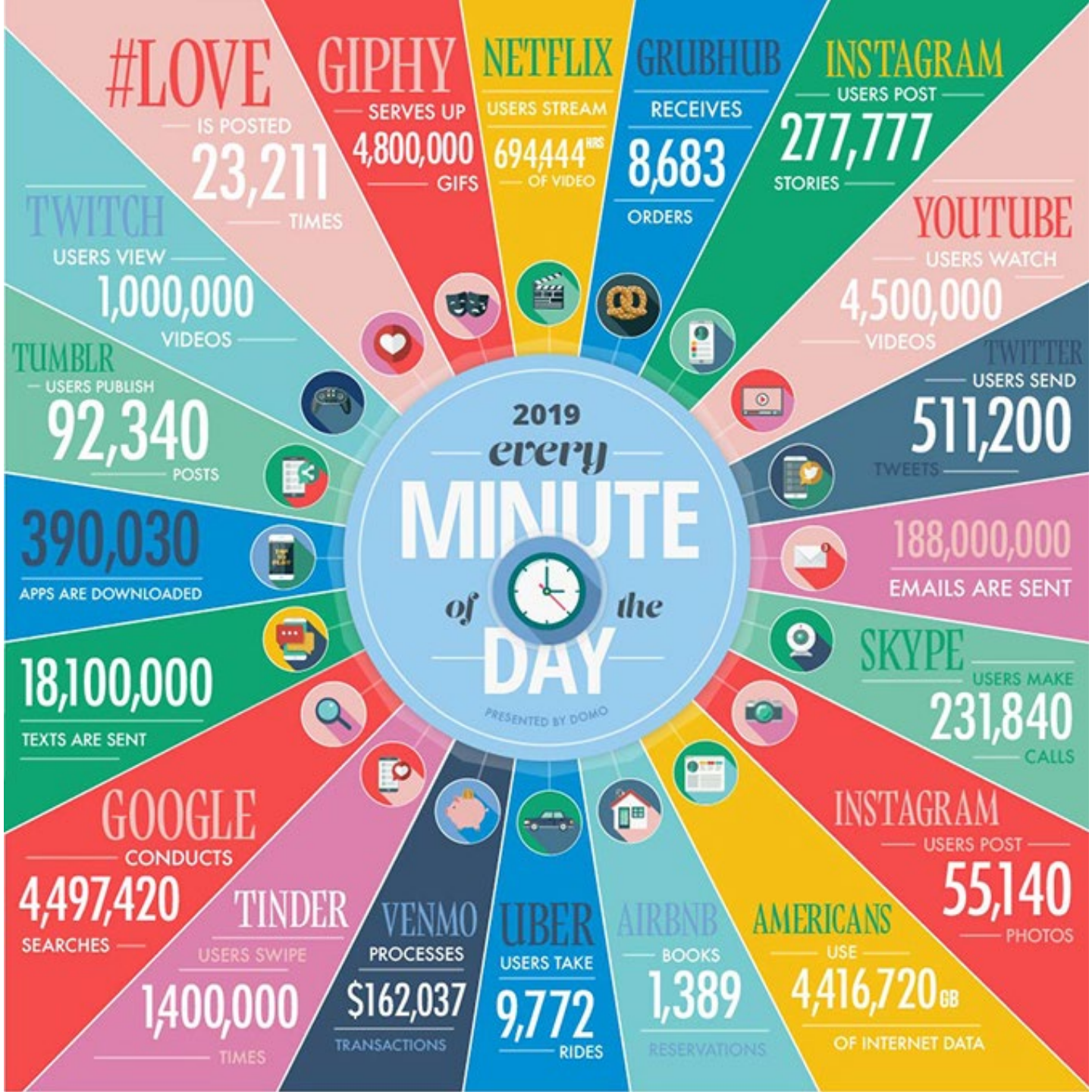
The world's internet population is growing significantly year-over-year. As of January 2019, the internet reaches 56.1% of the world's population and now represents 4.39 billion people — a 9% increase from January 2018.

GLOBAL INTERNET POPULATION GROWTH 2012–2018
(IN BILLIONS)

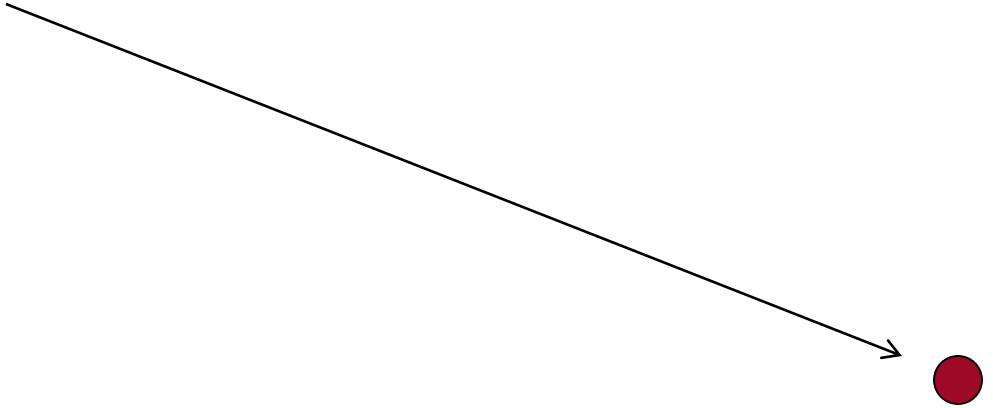
The ability to make data-driven decisions is crucial to any business. With each click, swipe, share, and like, a world of valuable information is created. Domo puts the power to make those decisions right into the palm of your hand by connecting your data and your people at any moment, on any device, so they can make the kind of decisions that make an impact.

Learn more at domo.com

SOURCES: STATISTA, INTERNET LIVE STATS, EXPANDED RAMBLINGS, NATIONAL ASSOCIATION OF CITY TRANSPORTATION OFFICIALS, WIRED

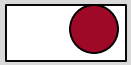

#LOVE
IS POSTED
23,211
TIMES
GIPHY
SERVES UP
4,800,000
GIFS
NETFLIX
USERS STREAM
694,444 HRS
OF VIDEO
GRUBHUB
RECEIVES
8,683
ORDERS
INSTAGRAM
USERS POST
277,777
STORIES
TWITCH
USERS VIEW
1,000,000
VIDEOS
YOUTUBE
USERS WATCH
4,500,000
VIDEOS
TUMBLR
USERS PUBLISH
92,340
POSTS
TWITTER
USERS SEND
511,200
TWEETS
2019
every
MINUTE
of the
DAY
PRESENTED BY DOMO
390,030
APPS ARE DOWNLOADED
188,000,000
EMAILS ARE SENT
18,100,000
TEXTS ARE SENT
SKYPE
USERS MAKE
231,840
CALLS
GOOGLE
CONDUCTS
4,497,420
SEARCHES
INSTAGRAM
USERS POST
55,140
PHOTOS
TINDER
USERS SWIPE
1,400,000
TIMES
VENMO
PROCESSES
$162,037
TRANSACTIONS
UBER
USERS TAKE
9,772
RIDES
AIRBNB
BOOKS
1,389
RESERVATIONS
AMERICANS
USE
4,416,720 GB
OF INTERNET DATA

- Oddly enough, it's still pretty much the same Intro Stat Course

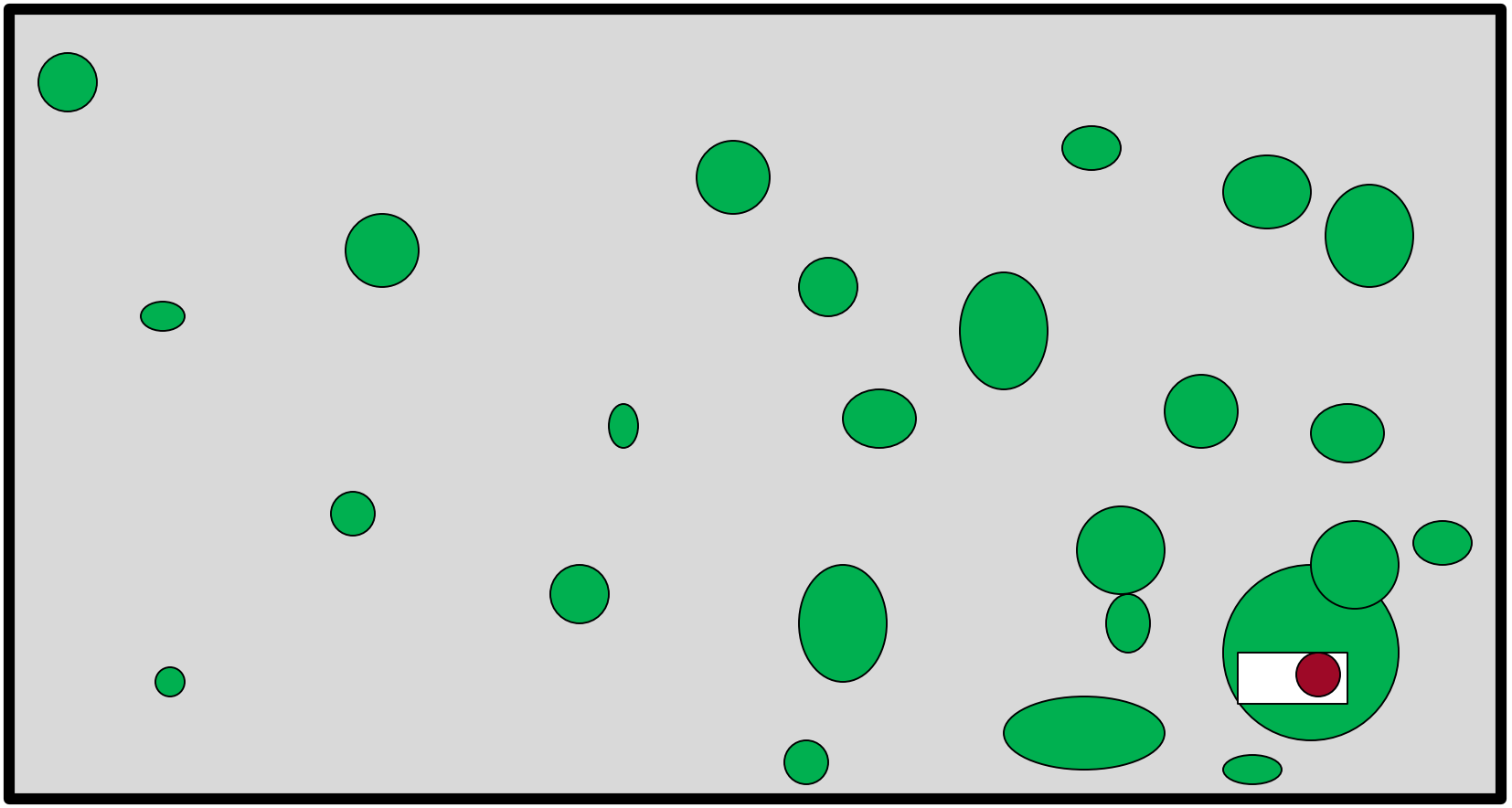


Statistics: Today

It's just a lot less useful than it used to be



Statistics: Today

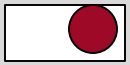


Statistics Research Fills in Gaps

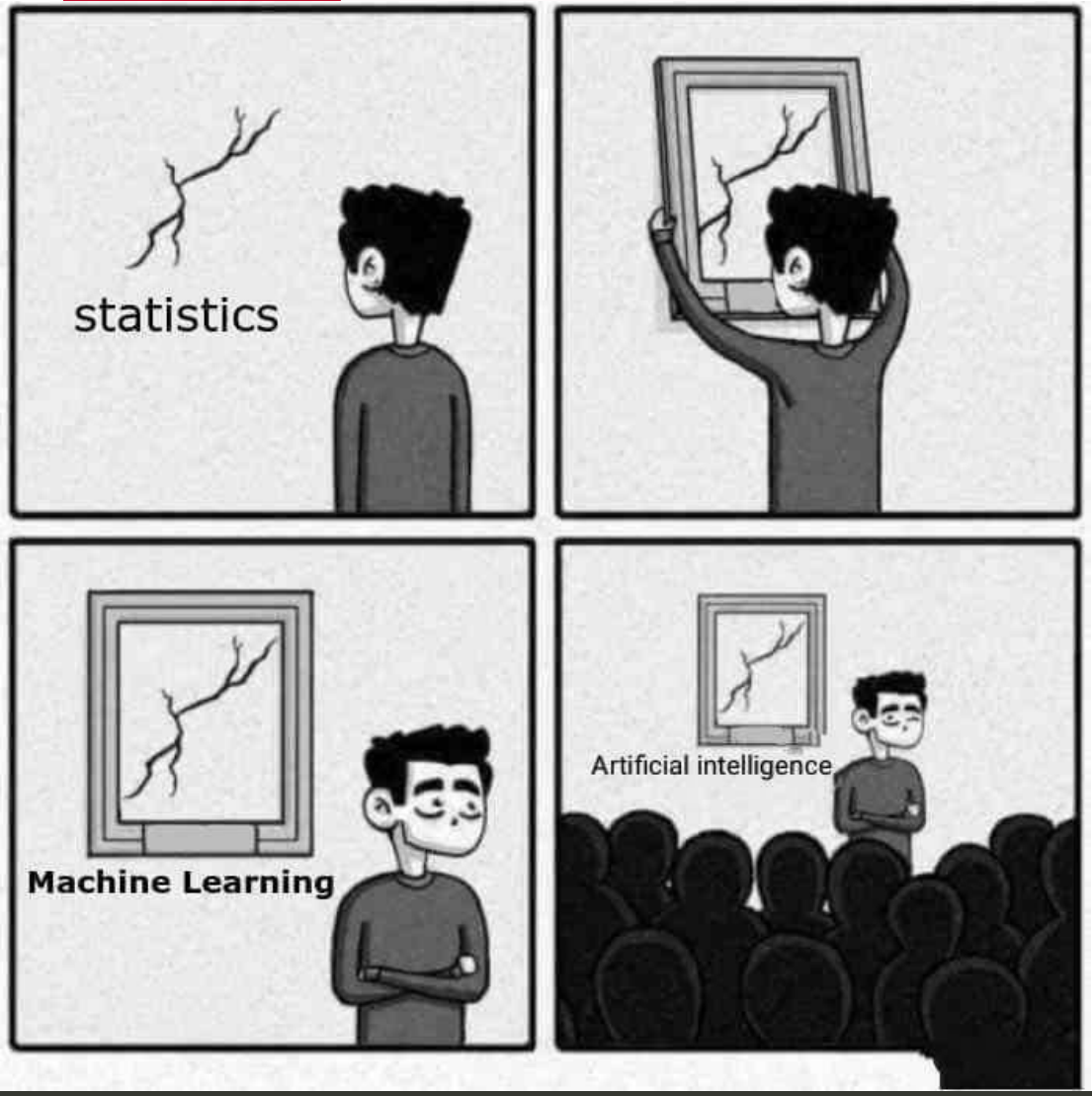
Statistical learning (SL) provides new formulas and algorithms for analyzing certain complex problems

Related names/subjects:

- Data Mining
- Knowledge Discovery
- Machine Learning
- Artificial Intelligence



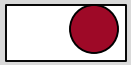
Statistical Learning



Meme source unknown, based on original comic by sandserif

Statistical Learning

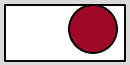
- Prediction
  - Regression (predict a numerical response)
  - Classification (predict a categorical response)
    - Especially predict a binary (2-group) response
- Dimension reduction
  - Variable selection (features relating to response)
  - "Subspaces": data lie mostly in smaller dimension
- Clustering



  - Find groups of similar objects from their properties

Statistical Learning

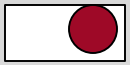
- Classical statistics starts with an assumed distributional model for a problem
  - T-tests and linear regression assume normality
- Provides a solid foundation for inference
  - Tests and confidence intervals
  - Measures of uncertainty
- Can fall apart when data source is more complex than model



Statistical Learning

SL takes a different approach from classical statistics

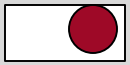
- Focus on the goal, and optimize a measure of closeness to the goal
  - Prediction starts with the goal of being “close” to future measurements.
  - Clustering starts with the goal of creating “tight” groups of similar individuals



- Use math and computing to find optimum

Statistical Learning

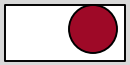
- Little use of testing, confidence intervals,
  - No probability foundation for notion of "α"
- Need for computational efficiency
  - Often need to scale up to huge numbers of observations and variables!
- Algorithmic, rather than foundational basis
  - Accept that different data sets from same population may yield different answers
  - Usually no measure of HOW different!



Statistical Learning

SL Problem I:

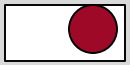
- Predictive
  - Regression and classification
  - Build models empirically
    - Maybe not care about what is in model: black box!
  - Rarely assess expected accuracy of predictions!
- Called *Supervised learning*
  - Response variable is target output
  - Explanatory variables are inputs



Statistical Learning

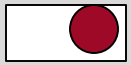
SL Problem II:

- Descriptive
  - Want to know properties of data
  - Relationships, structures, patterns
  - Groupings/clusters, outliers, special cases
- Called *Unsupervised learning*
  - No response variable as target
  - All variables contribute in same ways



Statistical Learning

- Both approaches have advantages over other
  - We will learn about these.
  - Understanding both helps you to understand when each one might be more appropriate
- There is definitely room for both in your toolkit.

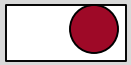


Statistical Learning

SFU SIMON FRASER UNIVERSITY
THINKING OF THE WORLD

Reading Assignment

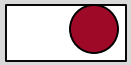
- ISLR Chapter 1



Statistical Learning

Everything we do is an approximation

(We want to find the “best” possible approximation for the “appropriate” amount of effort)



Statistical Learning