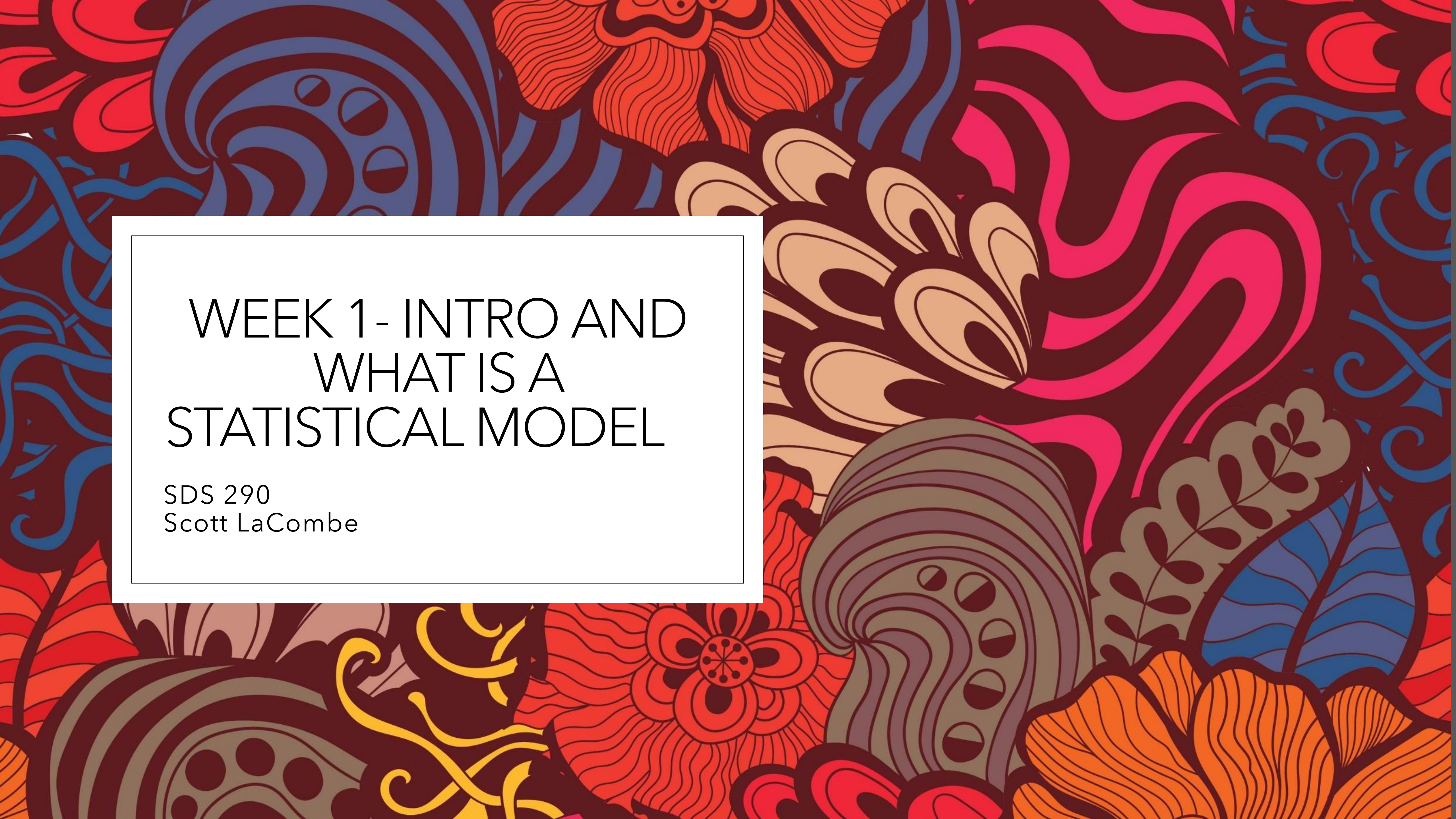


WEEK 1- INTRO AND WHAT IS A STATISTICAL MODEL

SDS 290
Scott LaCombe

Todays Plan

- Introductions
- Course overview
- CITI training set up

Teaching Assistant: Gollum LaCombe

A Little About Me

JOINT GOV/SDS POSITION

GRADUATE FROM UNIVERSITY OF IOWA

ORIGINALLY FROM KANSAS CITY, MO

I FOCUS ON STATE POLITICS AND NETWORKS OF PUBLIC POLICIES

Introductions (in small groups)

- Name
- Major
- Year at Smith
- Most recent show/movie you’ve been obsessed with

Goals of this course

- Understand fundamentals of experimentation and observational research
- Learn about how to design and implement survey experiments
- Implementing ANOVAs and similar models
- Use software and data to answer real world questions about the world around us

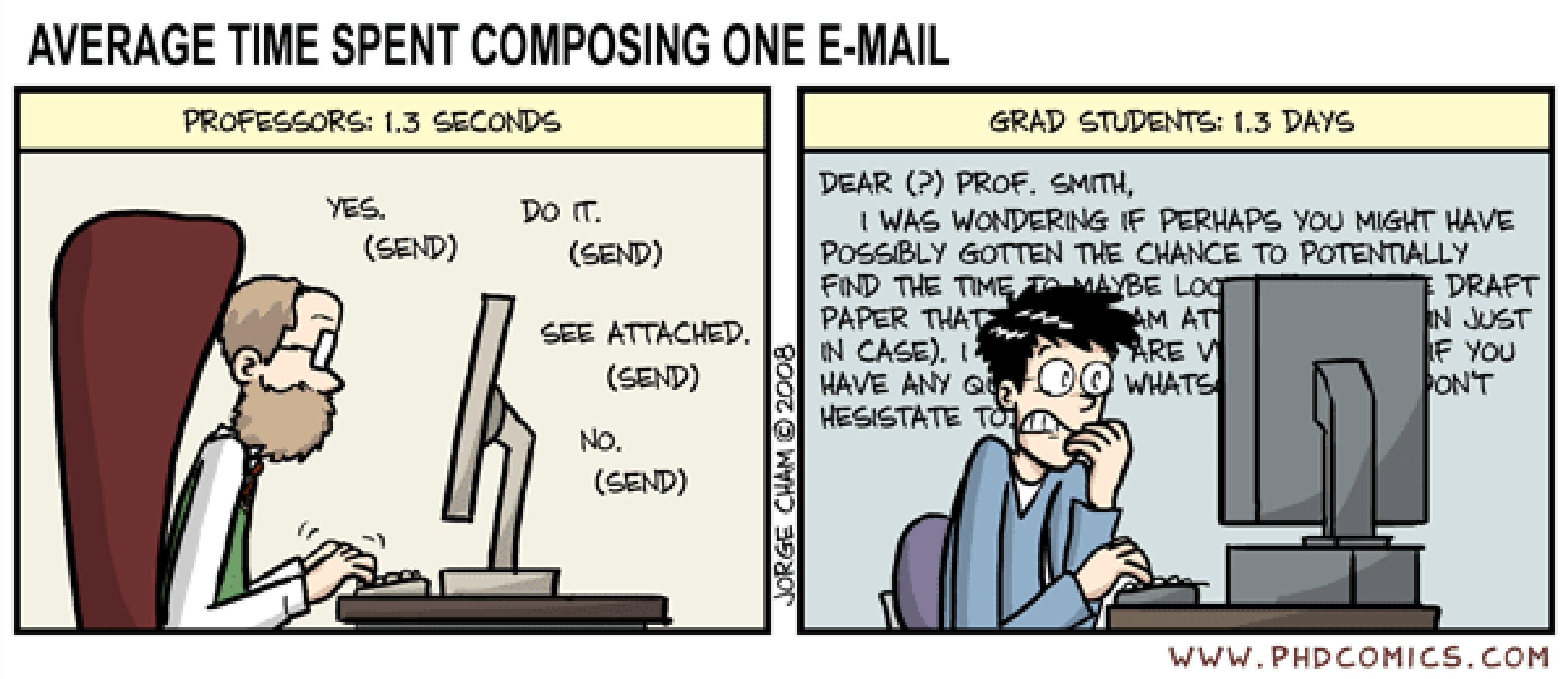
Quick note on R and Stats Background

- Assumption- you’ve taken an introductory stats class
  - Demonstrate familiarity with descriptive statistics, normal/t distribution, hypothesis testing, p-values, and confidence intervals
- We will be using R extensively in this class
  - Will start slow, but quickly build
  - If you are unfamiliar with R, I **strongly suggest** working through first 4 chapters of ModernDive (see syllabus)
- Talk with me, go to stats Tas
- SDS 100

Tips for Learning this semester

- Office hours:
  - Mondays: 2-3, Wednesdays 11-12, Thursdays 4:15-5:15
- Complete readings before class
- Use office hours and tutors
- Post on slack
  - If you have a question, someone else probably does too
  - Also counts toward participation
- Keep me in the loop if you are struggling inside/outside class
  - Much easier to give extensions **before** due date than after

Slack Chanel & Moodle


AVERAGE TIME SPENT COMPOSING ONE E-MAIL
PROFESSORS: 1.3 SECONDS
YES. (SEND)
DO IT. (SEND)
SEE ATTACHED. (SEND)
NO. (SEND)
JORGE CHAM © 2008
GRAD STUDENTS: 1.3 DAYS
DEAR (?) PROF. SMITH,
I WAS WONDERING IF PERHAPS YOU MIGHT HAVE POSSIBLY GOTTEN THE CHANCE TO POTENTIALLY FIND THE TIME TO MAYBE LOO E DRAFT PAPER THAT AM AT N JUST IN CASE). I ARE V IF YOU HAVE ANY Q WHATS ON'T HESISTATE TO
WWW.PHDCOMICS.COM

A note on course Delivery and Participation

- Will record lectures, no remote option
  - Welcome to “zoom-in” classmate
- Participation and attendance contribute to course participation grade
  - If you can’t make it to class, email me and post something on slack.

- In person R labs are critical to your learning
- I’m trying to be as flexible as possible, extending the same to you
    - In person attendance expected, but if you are feeling sick, close exposure, watch recording and get notes from a friend

Syllabus Walkthrough

Basic Structure of Course

- Lecture with periodic group discussion/prompts
- Weekly(ish) homework assignments, due Fridays at 11:59 PM
- Periodic R workshops to build programming skills
- 2 exams
- 2 mini projects- will talk about more next week
  - 1 mini project solo
  - 1 in groups of 3
  - Design and implement survey

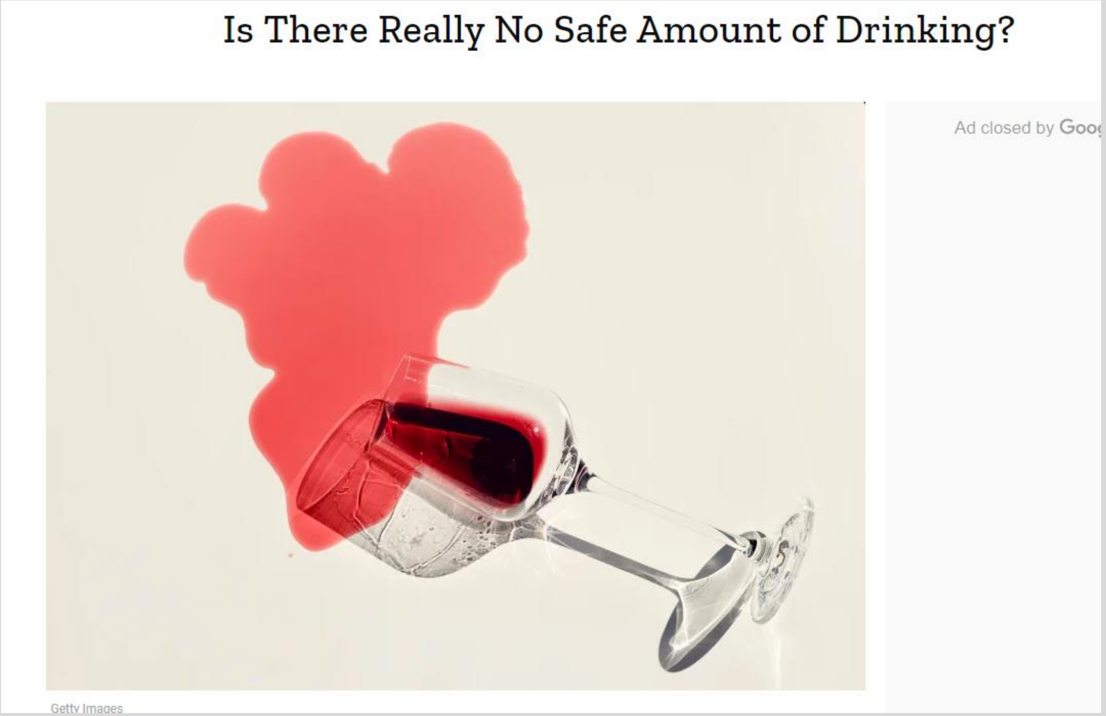
QUESTIONS?

Before we get started with content…

- By tomorrow- Fill out introductory survey
- For next Friday
  - CITI training
  - Intro to R lab
- Don’t put off! Citi training takes a bit of time

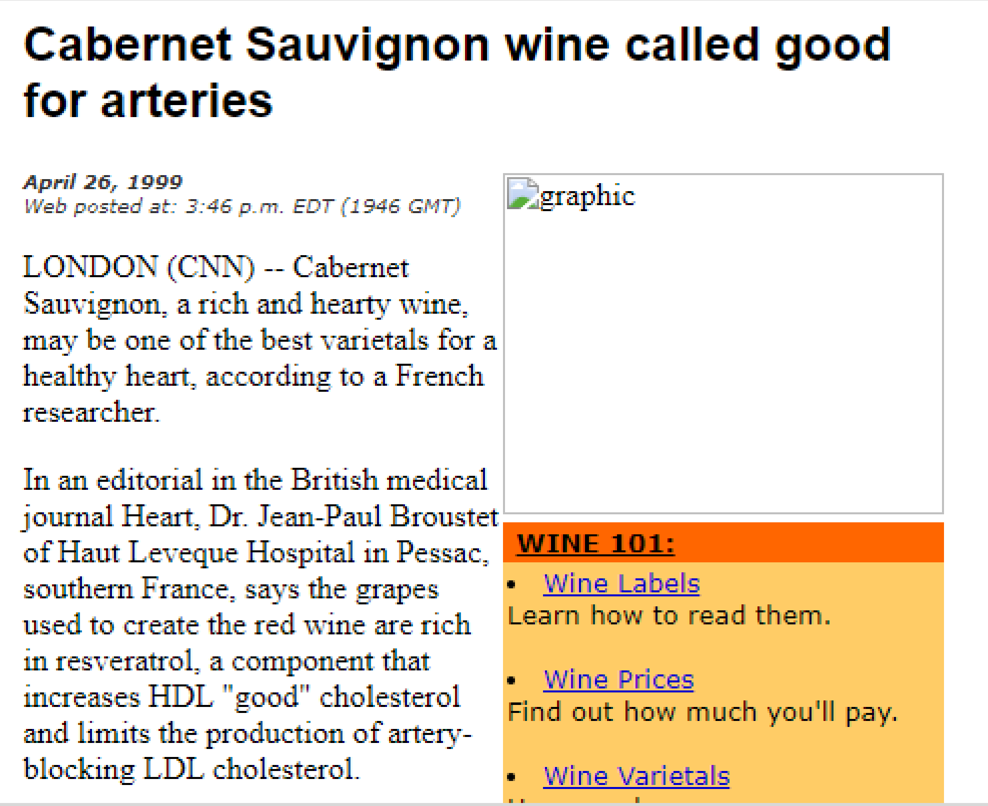
WHAT IS A STATISTICAL MODEL?

Problem we face today- so much data!



Is There Really No Safe Amount of Drinking?

Getty Images



Cabernet Sauvignon wine called good for arteries

April 26, 1999
Web posted at: 3:46 p.m. EDT (1946 GMT)

LONDON (CNN) -- Cabernet Sauvignon, a rich and hearty wine, may be one of the best varietals for a healthy heart, according to a French researcher.

In an editorial in the British medical journal Heart, Dr. Jean-Paul Broustet of Haut Leveque Hospital in Pessac, southern France, says the grapes used to create the red wine are rich in resveratrol, a component that increases HDL "good" cholesterol and limits the production of artery-blocking LDL cholesterol.

WINE 101:

- Wine Labels

Learn how to read them.

- Wine Prices

Find out how much you'll pay.

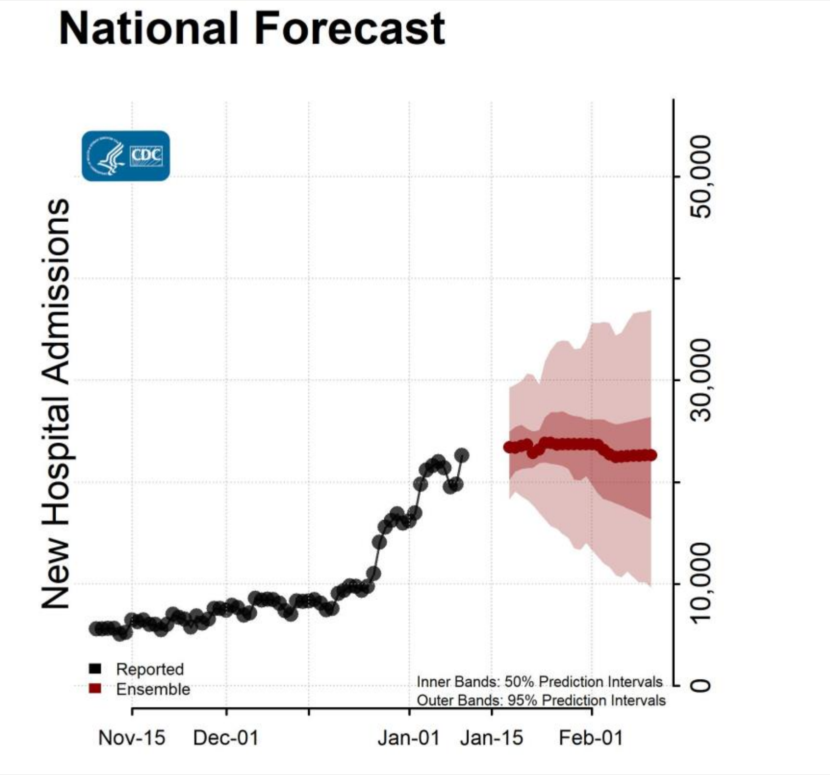
- Wine Varietals

Our goal

- With so much info, how to separate out the signal from the noise?

- Our approach- Modeling
  - Simplification of complex processes to use data to better understand the world around us

  - All models are wrong, some are better than others
    - World is complex, shouldn't forget that
    - Uncertainty is central



Goal of modeling

- Prediction
- Classification
- Evaluating a treatment
- Testing a theory
- Summarizing a pattern
- Improving a process
- Making a decision

Word Cloud- What do you see as most important goal of modeling?

Powered by Poll Everywhere

Start the presentation to see live content. For screen share software, share the entire screen. Get help at **pollev.com/app**

What **should** our goal be?- small groups

YOU WERE SO PREOCCUPIED WITH WHETHER OR NOT YOU COULD
YOU DIDN'T STOP TO THINK IF YOU SHOULD

Model basics

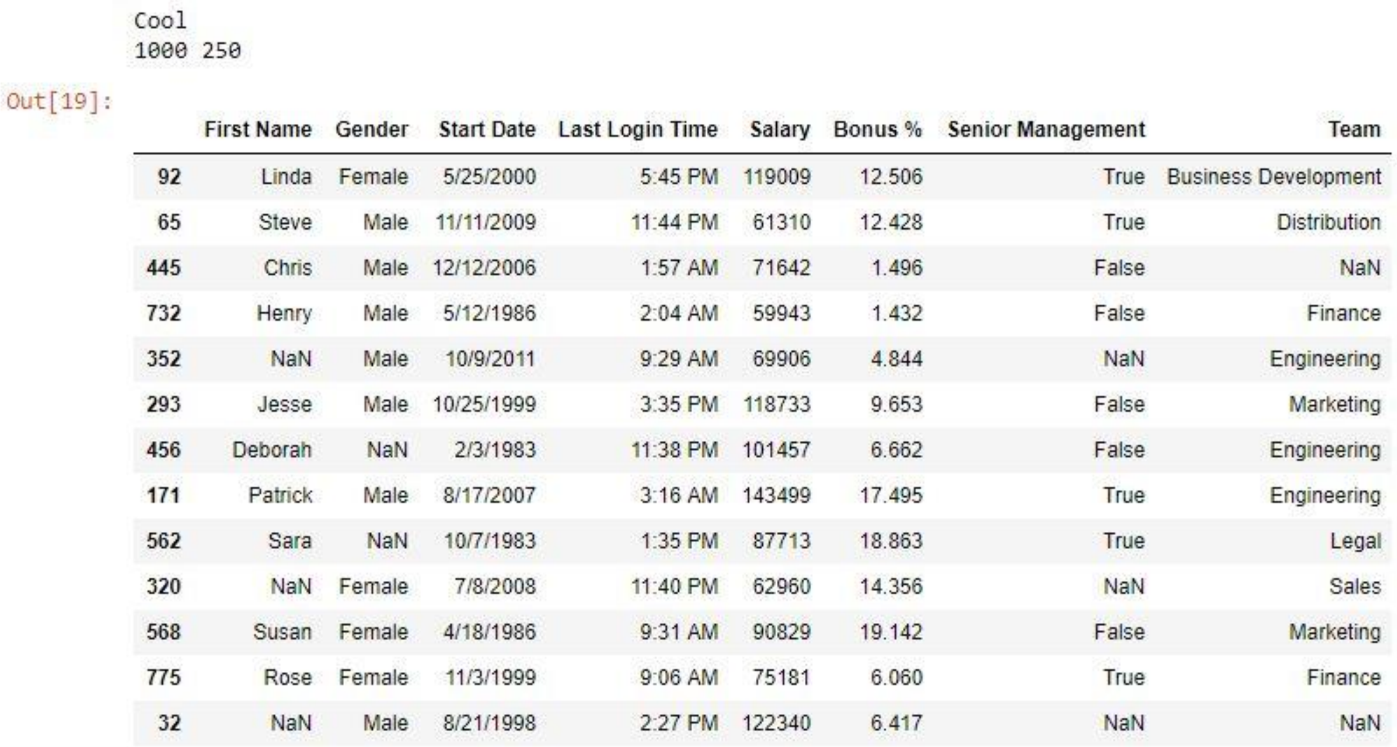
- Y=model+error
- Y
  - Dependent variable, response variable, outcome
  - Thing we are trying to explain/model
- Model
  - Explanatory variables, independent variables
- Error
  - Residuals, difference between predicted and observed

Other important terms

- Sample vs Population
- Statistic vs Parameter
- Inference
    - Parameter estimate
    - Causal inference
        - Role of experimentation vs observation
- Covariates

Data

- Cases
  - Unit of analysis
- Variables
- Types of variables
  - Quantitative/continuous
  - Categorical
    - Ordinal vs nominal
  - binary



```
Cool
1000 250
```

Out[19]:

	First Name	Gender	Start Date	Last Login Time	Salary	Bonus %	Senior Management	Team
92	Linda	Female	5/25/2000	5:45 PM	119009	12.506	True	Business Development
65	Steve	Male	11/11/2009	11:44 PM	61310	12.428	True	Distribution
445	Chris	Male	12/12/2006	1:57 AM	71642	1.496	False	NaN
732	Henry	Male	5/12/1986	2:04 AM	59943	1.432	False	Finance
352	NaN	Male	10/9/2011	9:29 AM	69906	4.844	NaN	Engineering
293	Jesse	Male	10/25/1999	3:35 PM	118733	9.653	False	Marketing
456	Deborah	NaN	2/3/1983	11:38 PM	101457	6.662	False	Engineering
171	Patrick	Male	8/17/2007	3:16 AM	143499	17.495	True	Engineering
562	Sara	NaN	10/7/1983	1:35 PM	87713	18.863	True	Legal
320	NaN	Female	7/8/2008	11:40 PM	62960	14.356	NaN	Sales
568	Susan	Female	4/18/1986	9:31 AM	90829	19.142	False	Marketing
775	Rose	Female	11/3/1999	9:06 AM	75181	6.060	True	Finance
32	NaN	Male	8/21/1998	2:27 PM	122340	6.417	NaN	NaN

What types of variables are these?

- Race
- Education level
- Income (in dollars)
- left-handed/non-left handed
- Voter turnout
- Letter grade in a course

Modeling Process- 4 steps

- Choose a form for the model
- Fit the model
- Assess the Model
- Address research question

- **Theory comes first**

Our plan

- Anova- Analysis of Variance
  - Response variable Quantitative
  - Explanatory variable typically categorical
  - Fundamentals of experimentation
- Later in semester- Causal inference with observational data