

Welcome to IS507!

“Statistics is the study of how best to collect, analyze, and draw conclusions from data...” - OpenIntro Statistics

Intro to Moodle class website

Quick plug: IS400

<https://ischool.illinois.edu/degrees-programs/courses/is400>

Outline of lectures:

- Lecture ~60 minutes on concepts**
- Coding follow along**
- Occasional group activities**

Note: today will be a little lecture heavy

Optional: Lab “HW” hour right after class

(I will leave if nobody shows up - email me if you need to come late)

Orientation

Course Description

An introduction to statistical and probabilistic models as they pertain to quantifying information, assessing information quality, and principled application of information to decision-making. The increasing prevalence of massive data sets and falling computational barriers have rendered statistical modeling an integral part of contemporary information management. With this in mind, this class prepares students to select and properly undertake commonly encountered modeling tasks. The course reviews relevant results from probability theory, emphasizing the merits and limitations of familiar probability distributions as vehicles for modeling information. Subsequent consideration includes parametric and non-parametric predictive models, as well as a discussion of extensions of these models for unsupervised learning. Throughout these discussions, the course focuses on model selection and gauging model quality. Applications of statistical and probabilistic models to tasks in information management (e.g. prediction, ranking, and data reduction) are emphasized.

Learning Objectives

Students will demonstrate an understanding of probability theory and statistical learning by building and evaluating models of a diverse range of data sets. By the end of the course students will have basic concepts of what constitutes a “good” statistical question, what one can feasibly learn and predict with data, and an overview of toolsets and methods to answer elementary statistical questions. In particular, each student will be able to:

- * Articulate the role of marginal, joint, and conditional probability in modeling processes involving information.
- * Select, parameterize, and compare probability distributions as vehicles for modeling information.
- * Specify, estimate and evaluate elementary parametric statistical models.
- * Specify, estimate and evaluate elementary non-parametric statistical models.
- * Articulate professional responsibilities with respect to creating, describing and using models built from data.

Orientation

Course Description

An introduction to statistical and probabilistic models as they pertain to quantifying information, assessing information quality, and principled application of information to decision-making. The increasing prevalence of massive data sets and falling computational barriers have rendered statistical modeling an integral part of contemporary information management. With this in mind, this class prepares students to select and properly undertake complex statistical analyses, emphasizing the merits and limitations of familiar parametric and non-parametric predictive models. The course focuses on how well we can know anything, really? But these discussions, which are integral parts of the course, will be applied to tasks in information management (e.g., predicting consumer behavior).

Learning Objectives

Students will demonstrate an understanding of data sets. By the end of the unit, students will learn how to analyze data sets and predict with confidence what they will find if they were to collect more data. They will be able to:

- * Articulate the role
 - * Select, parameterise
 - * Specify, estimate
 - * Specify, estimate
 - * Articulate professional

How well can we know anything, really?

What kinds of questions can we ask with data?

How accurately can we answer those questions from a particular dataset? How does this depend on features of this dataset (e.g. how it was procured)?

How can we make predictions from collected data? What is the “accuracy” of those predictions?

How can we use computational tools to answer statistical questions.

* Specify, estimate and evaluate elementary parametric statistical models.

Orientation

Course Description

An introduction to statistical and probabilistic models as they pertain to quantifying information, assessing information quality, and principled application of information to decision-making. The increasing prevalence of massive data sets and falling computational barriers have rendered statistical modeling an integral part of contemporary information management. With this in mind, this class prepares students to select and properly undertake common and limitations of familiar parametric predictive models. The course focuses on management (e.g. prediction, estimation, inference) of data sets.

Learning Objectives

Students will demonstrate an understanding of data sets. By the end of the course, students will learn and predict with confidence. They will be able to:

- * Articulate the role of statistical models in decision making.
- * Select, parameterize and estimate elementary parametric statistical models.
- * Specify, estimate and evaluate elementary non-parametric statistical models.
- * Articulate professional responsibilities with respect to creating, describing and using models built from data.

How well can we know anything, really?

What

data?

How accurate
dataset? How

from a particular
dataset (e.g.

How can we

? What is the

How can we

statistical

questions.



Orientation

Required Texts

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013) *An Introduction to Statistical Learning*. New York: Springer. [**abbreviated ISL**]
<http://www-bcf.usc.edu/~gareth/ISL/>

Diez, D., Barr, C., and Cetinkaya-Rundel, M. (2015) *OpenIntro Statistics* Fourth Edition, [available online, https://www.openintro.org/stat/textbook.php?stat_book=os, **abbreviated OIS**]

Venables, W.N., Smith, D.M and the R Core Team (2012) *An Introduction to R*. [available online, <http://cran.r-project.org/doc/manuals/R-intro.pdf>, **abbreviated ITR**]

| Week | Topic | Reading |
|------|---|-----------------------|
| 1 | <ul style="list-style-type: none"> • Data, Models, and Information • Elementary statistics: Definitions • Overview of R | OIS 1 (ISL 1) |
| 2 | <ul style="list-style-type: none"> • Elementary statistics: Applications & Plots | OIS 1 (ISL 1) |
| 3 | <ul style="list-style-type: none"> • Introduction to data analysis with R • Review of tabular and graphical displays of data | ITR 1, 2, 5, 6, 7, 12 |
| 4 | <ul style="list-style-type: none"> • Random variables: expectation and variance • Joint and conditional probability • Bayes rule | OIS 2 |
| 5 | <ul style="list-style-type: none"> • Random variables: distributions (normal, binomial, poisson) | OIS 3 |



Definitions, basic concepts, R practice

| Week | Topic | Reading |
|------|---|-----------------------|
| 1 | <ul style="list-style-type: none"> • Data, Models, and Information • Elementary statistics: Definitions • Overview of R | OIS 1 (ISL 1) |
| 2 | <ul style="list-style-type: none"> • Elementary statistics: Applications & Plots | OIS 1 (ISL 1) |
| 3 | <ul style="list-style-type: none"> • Introduction to data analysis with R • Review of tabular and graphical displays of data | ITR 1, 2, 5, 6, 7, 12 |
| 4 | <ul style="list-style-type: none"> • Random variables: expectation and variance • Joint and conditional probability • Bayes rule | OIS 2 |
| 5 | <ul style="list-style-type: none"> • Random variables: distributions (normal, binomial, poisson) | OIS 3 |



Probability basics,
typical distributions

| | | |
|----|---|------------------|
| 6 | <ul style="list-style-type: none"> • Modeling data with probability distributions • Foundations for inference | OIS 4 |
| 7 | <ul style="list-style-type: none"> • Inference for numerical data • Inference for categorical data | OIS 5, OIS 6 |
| 8 | <ul style="list-style-type: none"> • Linear regression | OIS 7 (ISL 3) |
| 9 | <ul style="list-style-type: none"> • Multiple linear regression | OIS 8 (ISL 3) |
| 10 | <ul style="list-style-type: none"> • Logical regression | OIS 8 (ISL 4) |
| 12 | <ul style="list-style-type: none"> • <i>k</i>-Nearest neighbor classification and regression | ISL 2.2.3, 4.6.5 |
| 13 | <ul style="list-style-type: none"> • Intro to Unsupervised linear models: Principle component analysis | ISL 10.0-10.2 |



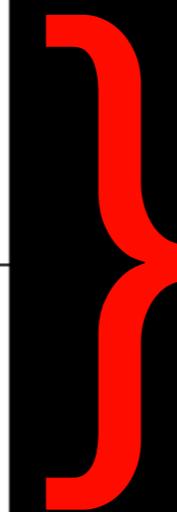
How well can we answer questions with our data?

| | | |
|----|---|------------------|
| 6 | <ul style="list-style-type: none"> • Modeling data with probability distributions • Foundations for inference | OIS 4 |
| 7 | <ul style="list-style-type: none"> • Inference for numerical data • Inference for categorical data | OIS 5, OIS 6 |
| 8 | <ul style="list-style-type: none"> • Linear regression | OIS 7 (ISL 3) |
| 9 | <ul style="list-style-type: none"> • Multiple linear regression | OIS 8 (ISL 3) |
| 10 | <ul style="list-style-type: none"> • Logical regression | OIS 8 (ISL 4) |
| 12 | <ul style="list-style-type: none"> • <i>k</i>-Nearest neighbor classification and regression | ISL 2.2.3, 4.6.5 |
| 13 | <ul style="list-style-type: none"> • Intro to Unsupervised linear models: Principle component analysis | ISL 10.0-10.2 |



Making predictions
from data

| | | |
|----|---|------------------|
| 6 | <ul style="list-style-type: none"> • Modeling data with probability distributions • Foundations for inference | OIS 4 |
| 7 | <ul style="list-style-type: none"> • Inference for numerical data • Inference for categorical data | OIS 5, OIS 6 |
| 8 | <ul style="list-style-type: none"> • Linear regression | OIS 7 (ISL 3) |
| 9 | <ul style="list-style-type: none"> • Multiple linear regression | OIS 8 (ISL 3) |
| 10 | <ul style="list-style-type: none"> • Logical regression | OIS 8 (ISL 4) |
| 12 | <ul style="list-style-type: none"> • <i>k</i>-Nearest neighbor classification and regression | ISL 2.2.3, 4.6.5 |
| 13 | <ul style="list-style-type: none"> • Intro to Unsupervised linear models: Principle component analysis | ISL 10.0-10.2 |



Classification and
unsupervised
learning

| | | |
|----|---|------------------|
| 6 | <ul style="list-style-type: none"> • Modeling data with probability distributions • Foundations for inference | OIS 4 |
| 7 | <ul style="list-style-type: none"> • Inference for numerical data • Inference for categorical data | OIS 5, OIS 6 |
| 8 | <ul style="list-style-type: none"> • Linear regression | OIS 7 (ISL 3) |
| 9 | <ul style="list-style-type: none"> • Multiple linear regression | OIS 8 (ISL 3) |
| 10 | <ul style="list-style-type: none"> • Logical regression | OIS 8 (ISL 4) |
| 12 | <ul style="list-style-type: none"> • k-Nearest neighbor classification and regression | ISL 2.2.3, 4.6.5 |
| 13 | <ul style="list-style-type: none"> • Intro to Unsupervised linear models: Principle component analysis | ISL 10.0-10.2 |

**Midterms + Final will
be a guided application
of these principles
(more on that in the
next few weeks)**

Pre- and Co-requisite

IS452 Foundations of Information Processing is strongly recommended as a prerequisite. A highly motivated student could pass this course without IS452 (so the prerequisite is not enforced), but programming will not be covered in this course. Students who have not completed an introductory course on statistics will need to come up to speed quickly on material covered early in the semester.

Assignment

Weekly homework
Midterm exam
Final exam

55%
20%
25%



See Syllabus for late HW policies, and we will drop your lowest HW score.

HW and Exam Formats

File name structure: lastname-first-module.ext
(e.g., naiman-jill-assignment1.pdf).

The submission must include:

1) A narrative document as a PDF file (to be read by a human). To preserve the natural flow of the narrative, figures (e.g., screenshots, code snippets) and tables should be embedded into the document near their first mention. Any supplementary files containing R programs or data should be referenced in the text and separately uploaded.

AND

2) All R code as separate files with an .R extension (to be read by a computer).

HW — Self Grading (We will go over this again in week 3)

| Assignment #3 self-grading | | | |
|--|--------|--------------------|----------------------------------|
| Fill in the orange area. Explanations for incorrect answers are optional, but will help your grade if they are good. | | | |
| | | | |
| | | Correct: | 0.00 |
| | | Incorrect: | 0.00 |
| | | Explanation score: | 0.00 |
| | | Estimated score: | 0.0% |
| | | | |
| | | | |
| Exercise | Weight | Correct (Y or N) | Explanation for incorrect answer |
| OIS 1.2a | 0.25 | | |
| OIS 1.2b | 0.25 | | |
| OIS 1.2c | 0.25 | | |

Are you correct?

If incorrect, why?

HW — Self Grading (We will go over this again in week 3)

| Assignment #3 self-grading | | | |
|--|--------|--------------------|----------------------------------|
| Fill in the orange area. Explanations for incorrect answers are optional, but will help your grade if they are good. | | | |
| | | Correct: | 0.00 |
| | | Incorrect: | 0.00 |
| | | Explanation score: | 0.00 |
| | | Estimated score: | 0.0% |
| | | | |
| | | | |
| Exercise | Weight | Correct (Y or N) | Explanation for incorrect answer |
| OIS 1.2a | 0.25 | | |
| OIS 1.2b | 0.25 | | |
| OIS 1.2c | 0.25 | | |

Timeline



General Adulting Policies

- Come to class
- Participate in class
- Don't get in the way of the learning process of others/keep a Growth Mindset
- Do your own work

Course Strategies:

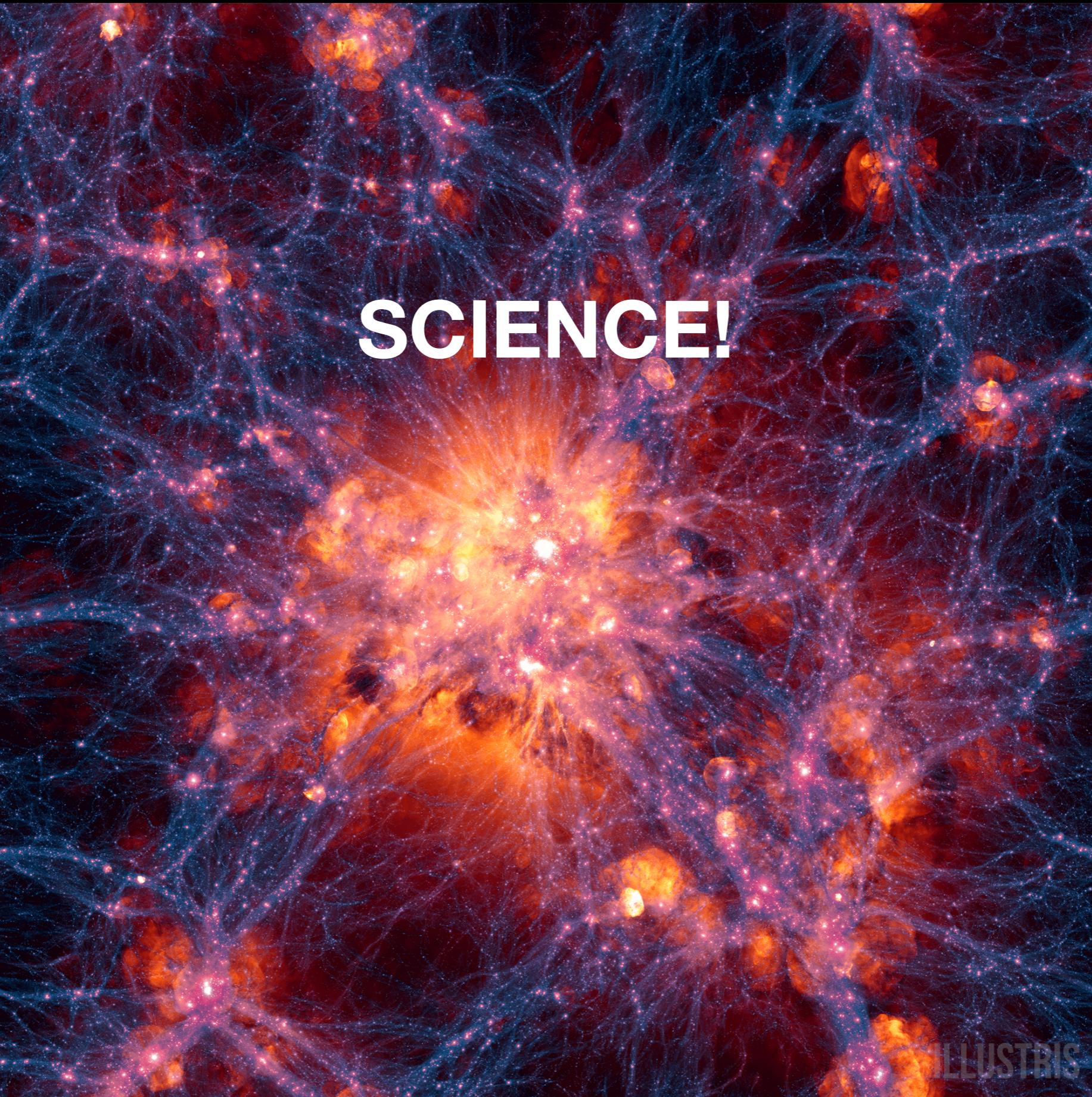
- Come to class & Participate in class
- New to coding: read the prep notebooks before class!

~~What are we doing?~~

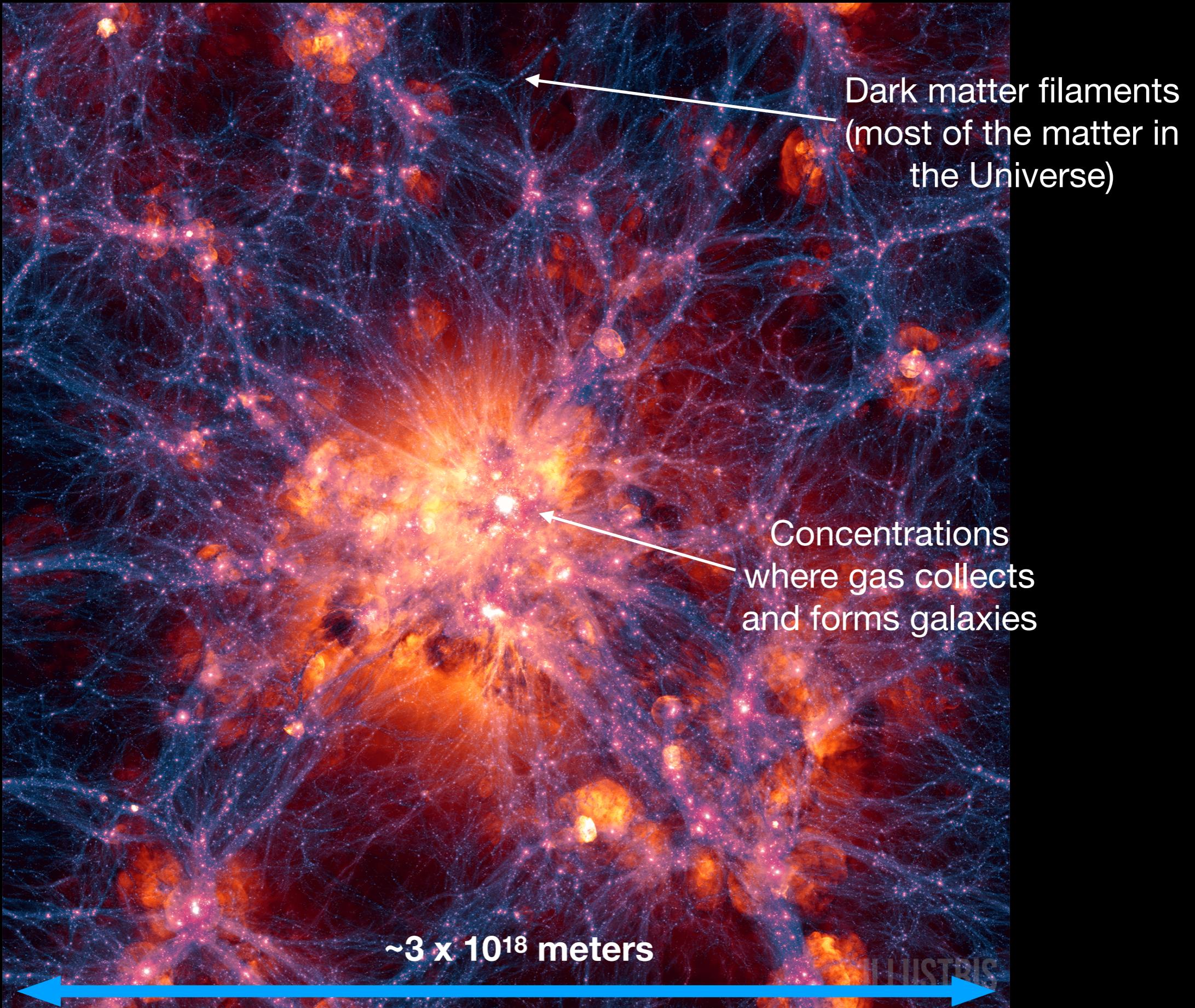
~~How are we going to do it?~~

Who are you?

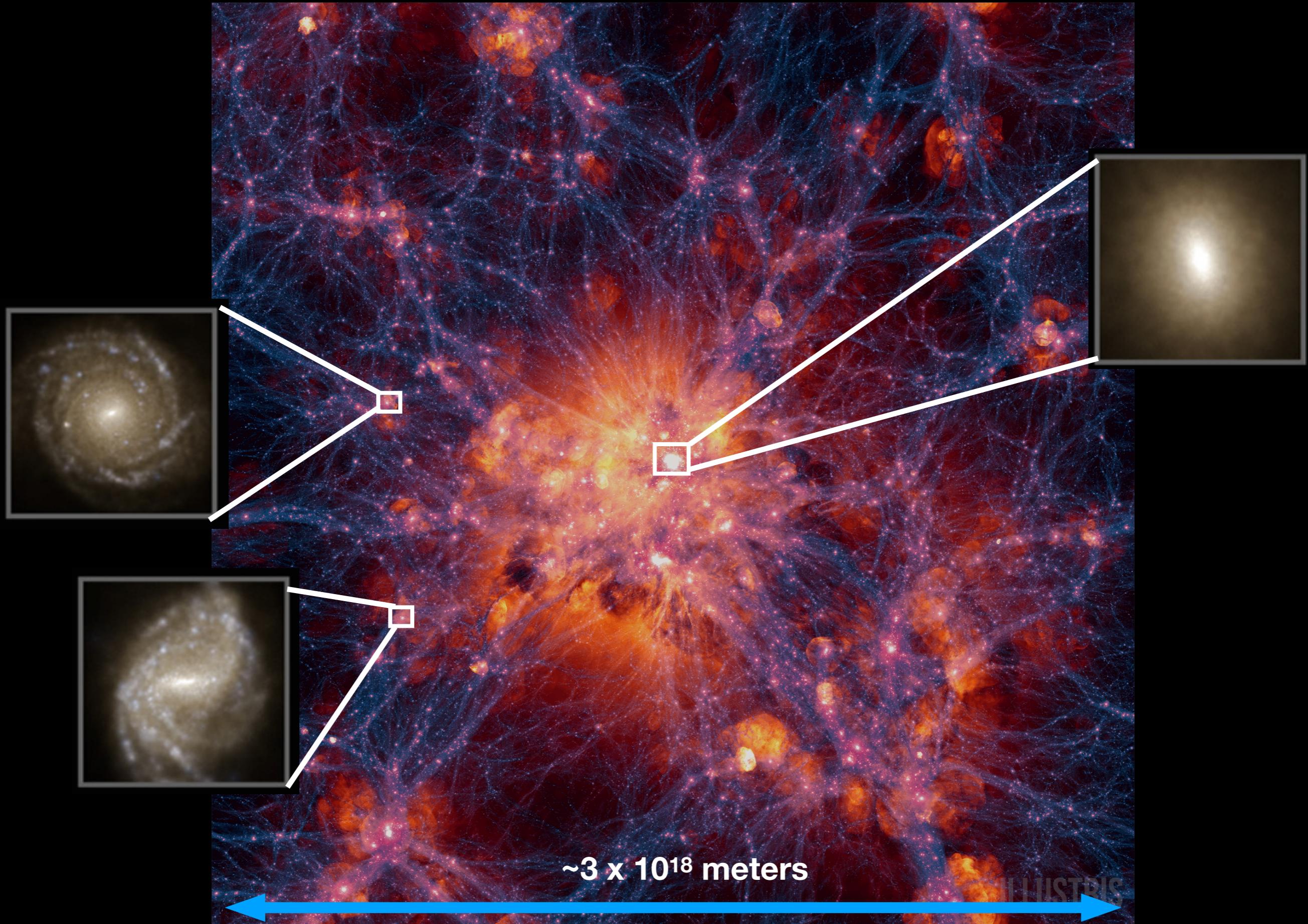
My background



My background

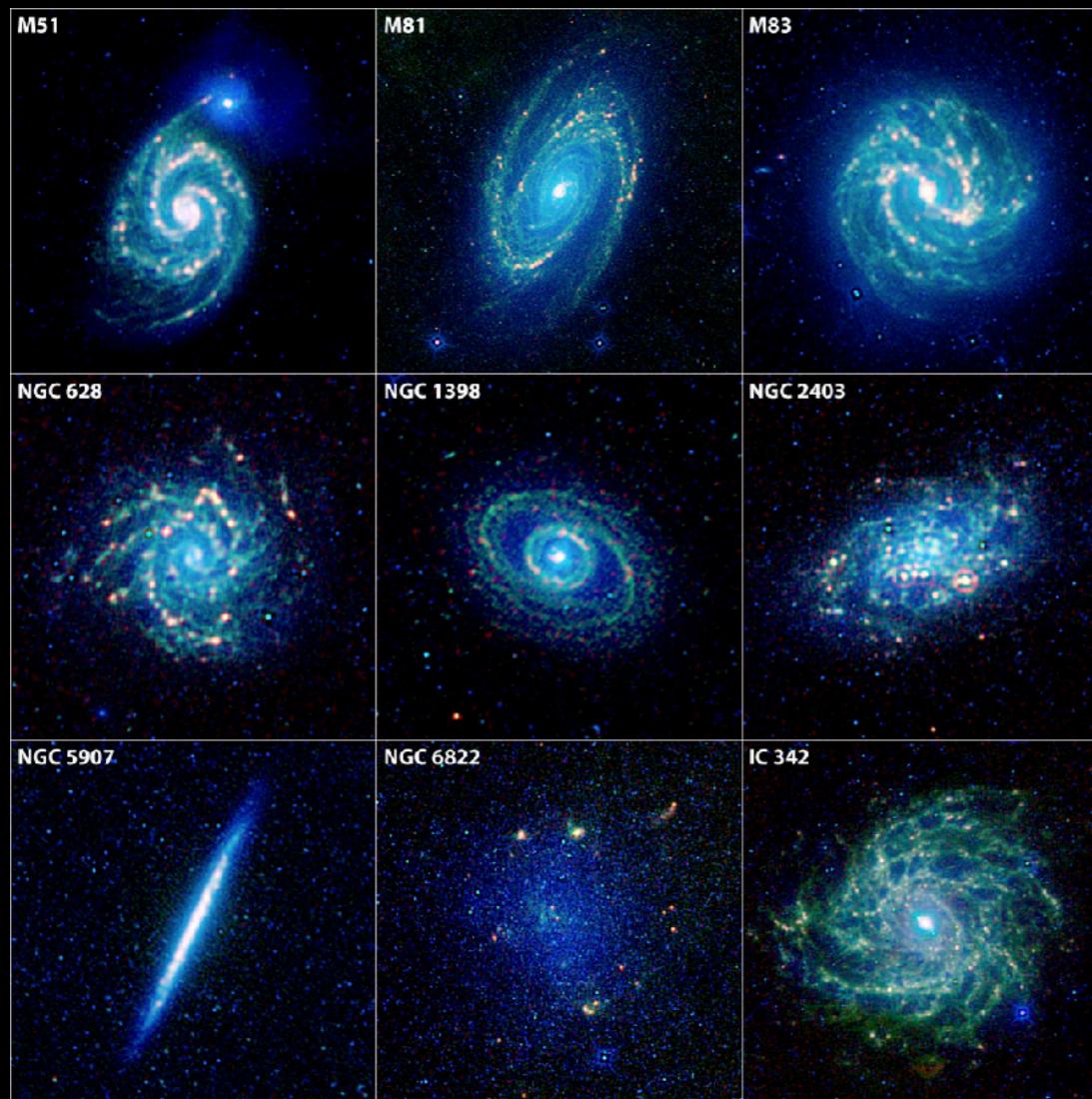


My background

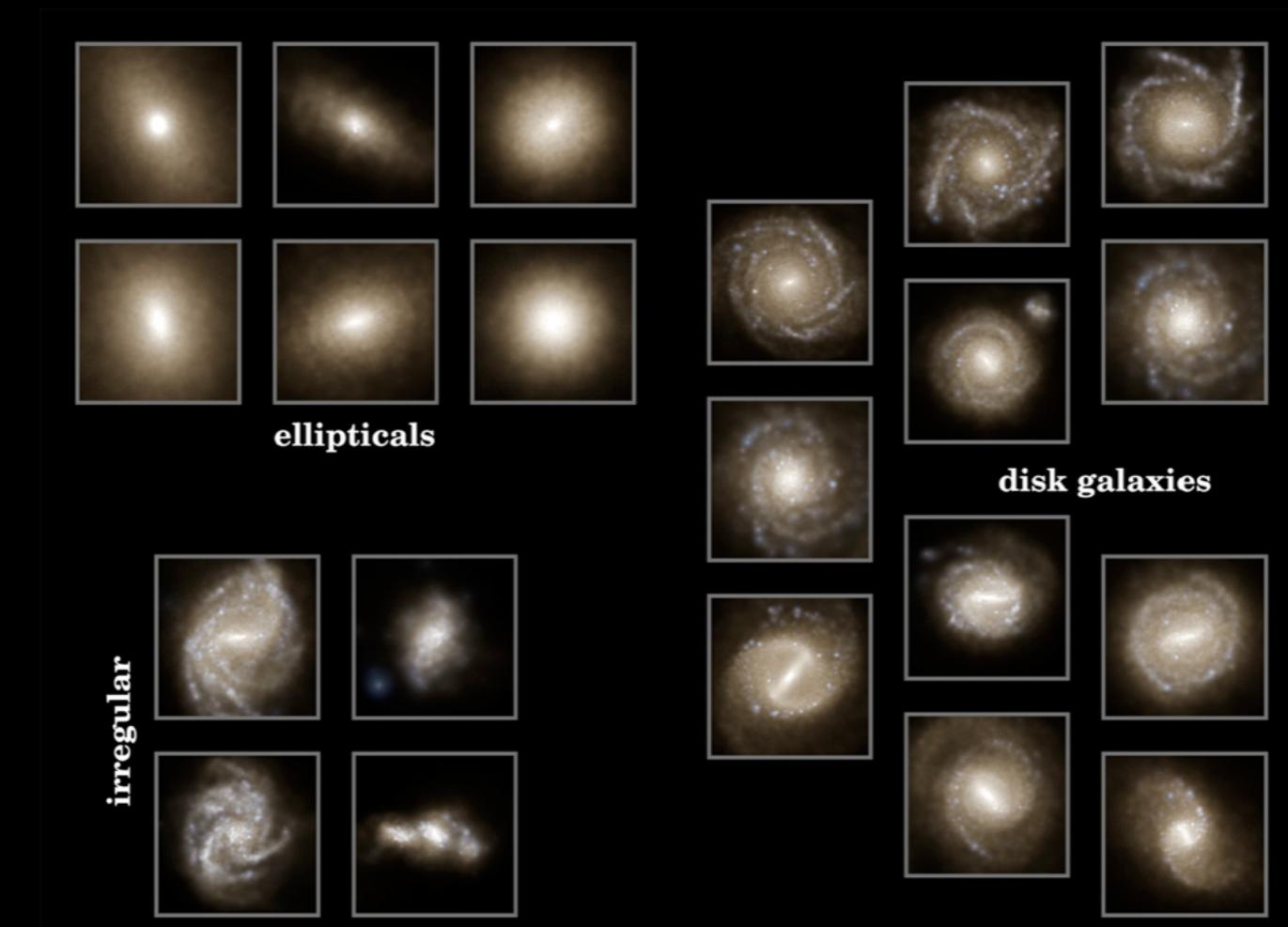


Statistical comparison between Observations & Models

Images of real galaxies from NASA's Wide-field Infrared Survey Explorer



Images of “fake” galaxies simulated in a super computer



(differences in colors due to color schemes chosen for each image)

Recent Machine Learning Work

240

WIGGS & GIES

Vol 396

Stickland, we define phase 0.0 to be superior conjunction of the primary star.

The spectra were reduced by standard techniques (see Paper I) which included the removal of atmospheric telluric lines. The final step yields rectified spectra transformed to a uniform heliocentric wavelength grid. The resulting spectra appear in Figures 1 and 2 for H α and He I $\lambda 6678$, respectively. In these figures we plot the spectra as a function of radial velocity assuming rest wavelengths of 6562.682 Å and 6678.148 Å for H α and He I, respectively. The spectra are placed in the upper portion of each diagram so that their continua are aligned with the orbital phase of observation. The bar at the upper right indicates the intensity scale relative to the continuum. The lower portion of the figure shows the same spectra portrayed as a gray-scale image in which spectral intensity is transformed

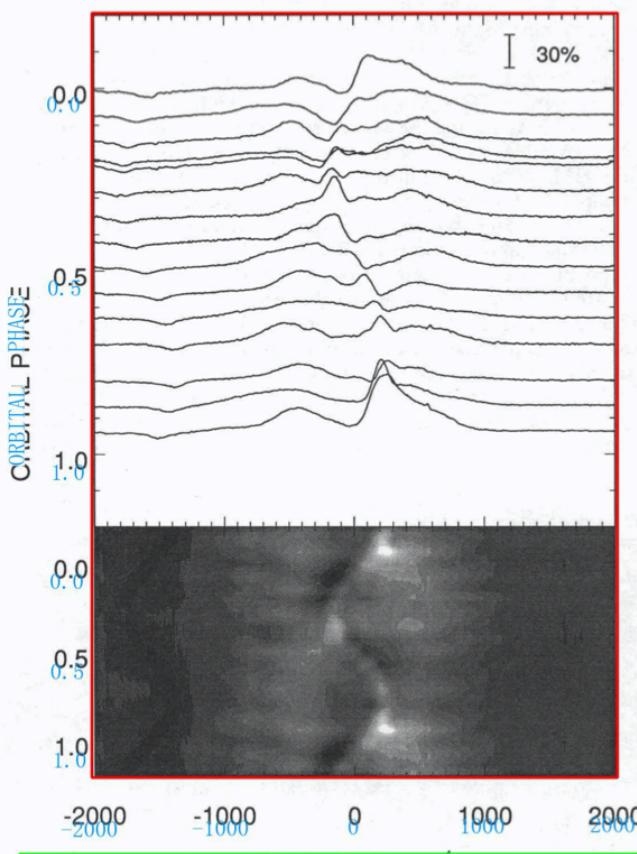


FIG. 1.—Top: H α line profiles plotted against heliocentric radial velocity. The profiles are arranged in order of increasing orbital phase and each spectrum is placed in the y-ordinate so that the continuum equals the phase of observation (phase 0.0 = superior conjunction of the primary). The bar in the upper right gives the spectrum intensity scale relative to a unit continuum. The fourth spectrum from the top was obtained two years after the other spectra. Bottom: A gray-scale representation of the profile variations shown above. Here each spectral intensity is assigned one of 16 gray levels based on its value between the minimum (deepest absorption) and maximum (highest emission) observed values. The spectrum at each phase in the image is calculated by a linear interpolation between the closest observed phases. The spectral image for the first and last 20% of the orbit are reproduced at the bottom and top of the image, respectively, to improve the sense of phase continuity.

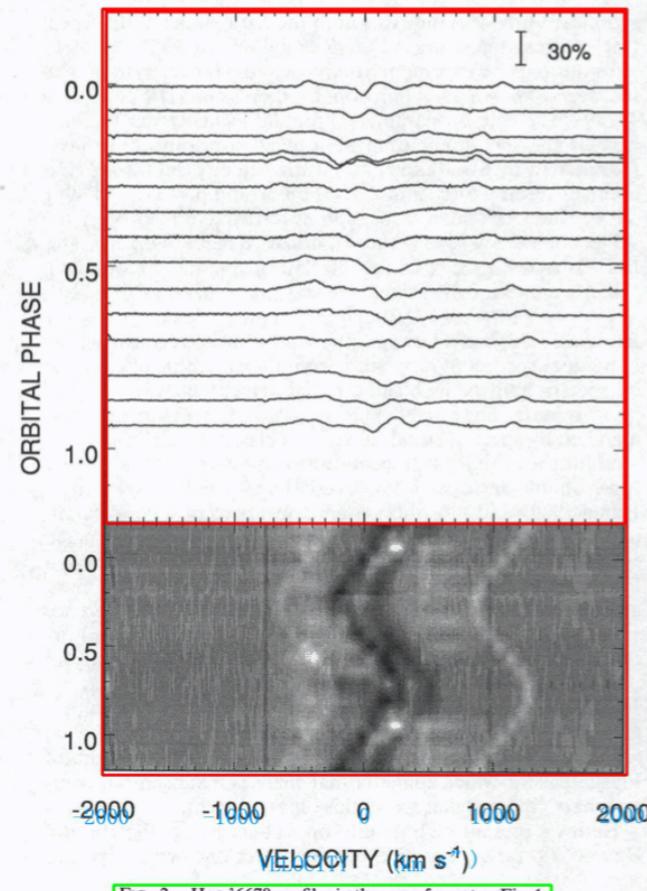


FIG. 2.—He I $\lambda 6678$ profiles in the same format as Fig. 1

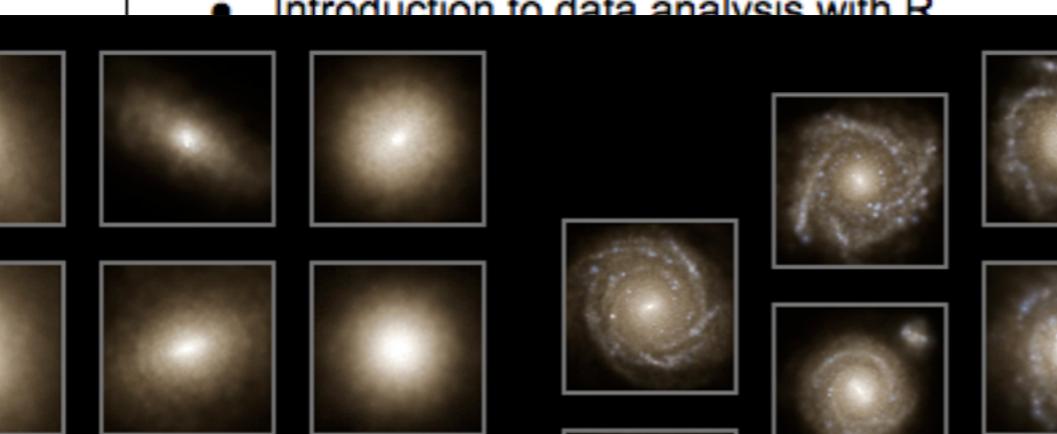
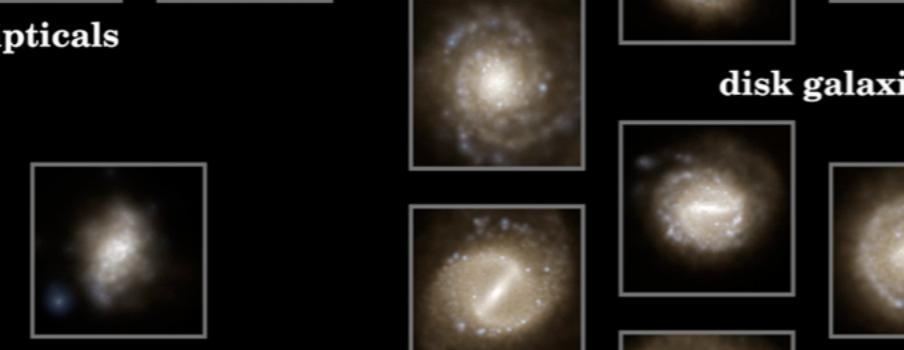
into a gray-scale intensity (dark in the absorption cores and bright at the emission peaks). The spectrum at each orbital phase in the image was calculated by a linear interpolation between observations with the closest phases. In some cases, the interpolation method causes narrow features to appear as discrete circles near conjunction phases (0.0 and 0.5) where the velocity changes rapidly between observations (see, for example, Si IV $\lambda 6701$ located near $V_r = 1050 \text{ km s}^{-1}$ in Fig. 2).

In the following subsections, we first describe the orbital phase-related variations in features associated with the primary star and then those in lines associated with circumstellar gas. The last subsection deals with variations unrelated to orbital phase in the H α emission line.

2.1. Spectral Lines Associated with the Primary Star

There are several lines in Figures 1 and 2 which show the familiar “s” shape corresponding to the velocity curve of the primary star. These include the absorption lines He I $\lambda 6678$ and He II $\lambda\lambda 6527.13, 6683.20$ and the emission lines Si IV $\lambda\lambda 6667.56, 6701.21$ (Fullerton 1989). All of these lines are typically found in mid-O supergiants (Conti 1974; Ninkov, Walker, & Yang 1987a; Ninkov et al. 1987b). In addition, we observed a weak emission line (maximum intensity $\approx 4\%$ of the continuum) near 6610 Å which we tentatively identify as N II $\lambda 6610.58$. Meinel, Aveni, & Stockton (1981) offer this iden-

Fig. 1—Top: H α line profiles plotted against heliocentric radial velocity. The profiles are arranged in order of increasing orbital phase and each spectrum is placed in the y-ordinate so that the continuum equals the phase of observation (phase 0.0 = superior conjunction of the primary). The bar in the upper right gives the spectrum intensity scale relative to a unit continuum. The fourth spectrum from the top was obtained two years after the other spectra. Bottom: A gray-scale representation of the profile variations shown above. Here each spectral intensity is assigned one of 16 gray levels based on its value between the minimum (deepest absorption) and maximum (highest emission) observed values. The spectrum at each phase in the image is calculated by a linear interpolation between the closest observed phases. The spectral image for the first and last 20% of the orbit are reproduced at the bottom and top of the image, respectively, to improve the sense of phase continuity.

| Week | Topic | Reading |
|-----------|--|-----------------------|
| 1 | <ul style="list-style-type: none"> • Data, Models, and Information • Elementary statistics: Definitions • Overview of R | OIS 1 (ISL 1) |
| 2 | <ul style="list-style-type: none"> • Elementary statistics: Applications & Plots | OIS 1 (ISL 1) |
| | <ul style="list-style-type: none"> • Introduction to data analysis with R | ITR 1, 2, 5, 6, 7, 12 |
| |  <p>ellipticals</p> | OIS 2 |
| irregular |  <p>disk galaxies</p> | OIS 3 |

Definitions, basic concepts, R practice

What is the “typical” value of a dataset? (median)

**What is the typical deviation of any model around this typical value?
(standard deviation)**

| | | |
|----|--|-----------------------|
| 6 | <ul style="list-style-type: none"> Modeling data with probability distributions Foundations for inference | OIS 4 |
| 7 | <ul style="list-style-type: none"> Inference for numerical data Inference for categorical data | OIS 5, OIS 6 |
| 8 | <ul style="list-style-type: none"> Linear regression | OIS 7 (ISL 3) |
| |  <p>ellipticals</p> <p>disk galaxies</p> <p>irregular</p> | L 3) L 4) 4.6.5 |
| 13 | <ul style="list-style-type: none"> Intro to Unsupervised linear models: Principle component analysis | ISL 10.0-10.2 |



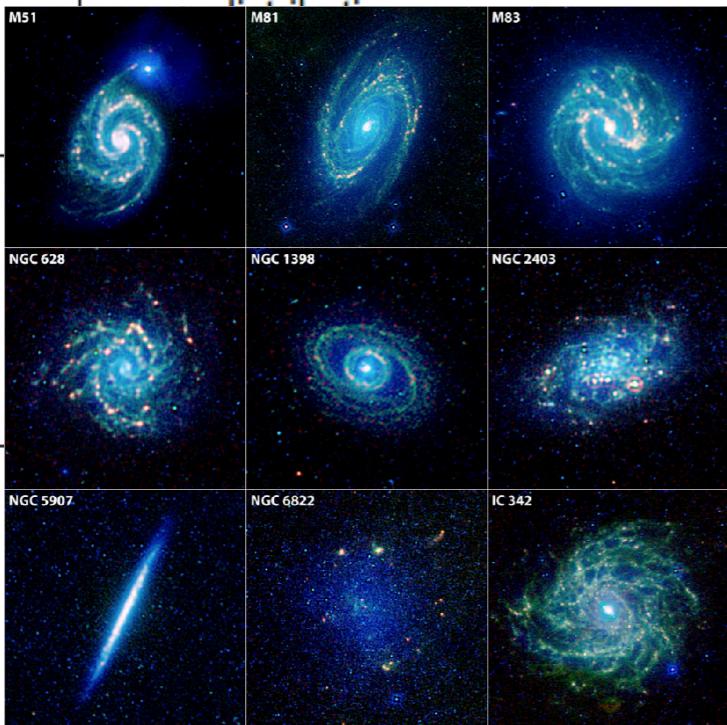
How well can we answer questions with our data?

How do different “variables” in our dataset (like supernova rate and galaxy mass) depend on one another?

How sure can we be that one variable depends on another?

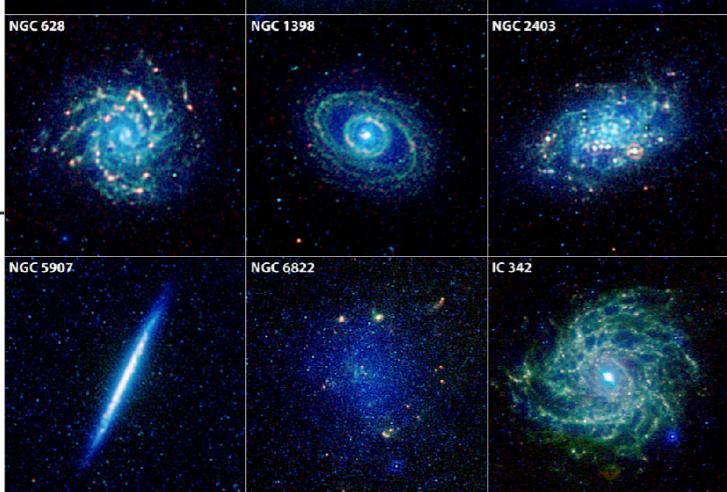
6

- Modeling data with probability



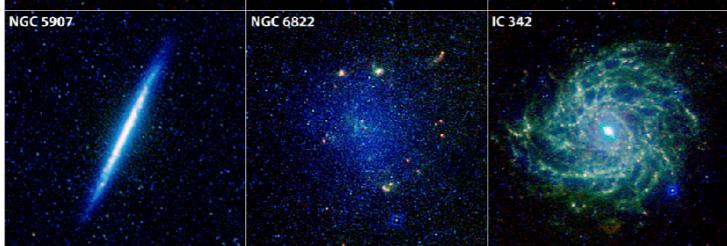
OIS 4

7



OIS 5, OIS 6

8



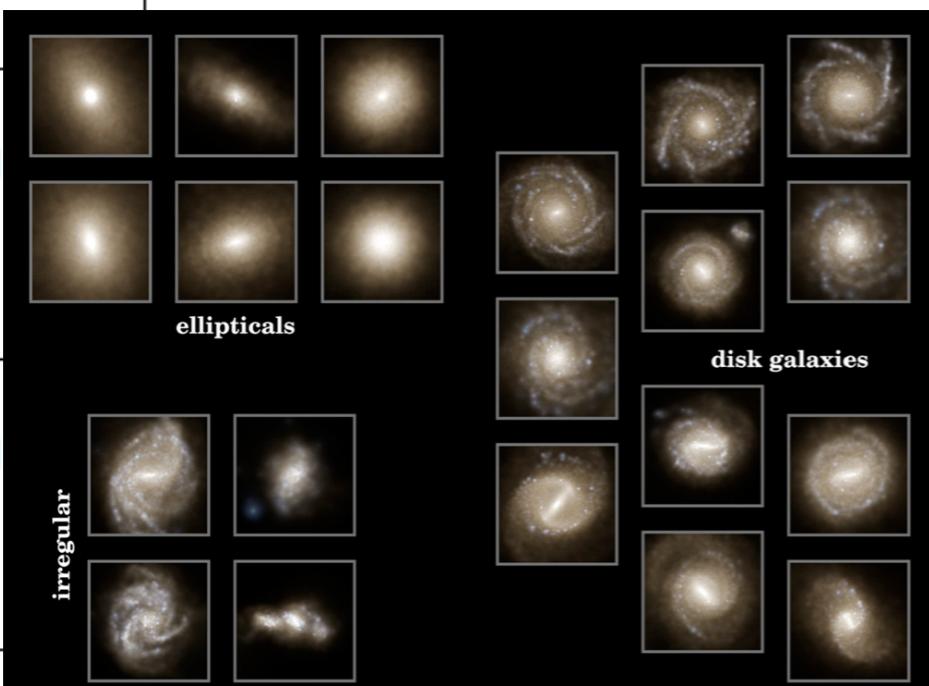
OIS 7 (ISL 3)

9

- Multiple linear regression

OIS 8 (ISL 3)

10



OIS 8 (ISL 4)

11

ISL 2.2.3, 4.6.5

13

- Intro to Unsupervised linear models:
Principle component analysis

ISL 10.0-10.2



Making predictions
from data

Stickland, we define phase 0.0 to be superior conjunction of the primary star.

The spectra were reduced by standard techniques (see Paper I) which included the removal of atmospheric telluric lines. The final step yields rectified spectra transformed to a uniform heliocentric wavelength grid. The resulting spectra appear in Figures 1 and 2 for H α and He I $\lambda\lambda 6678$, respectively. In these figures we plot the spectra as a function of radial velocity assuming rest wavelengths of 6562.682 Å and 6678.148 Å for H α and He I, respectively. The spectra are placed in the upper portion of each diagram so that their continua are aligned with the orbital phase of observation. The bar at the upper right indicates the intensity scale relative to the continuum. The lower portion of the figure shows the same spectra portrayed as a gray-scale image in which spectral intensity is transformed

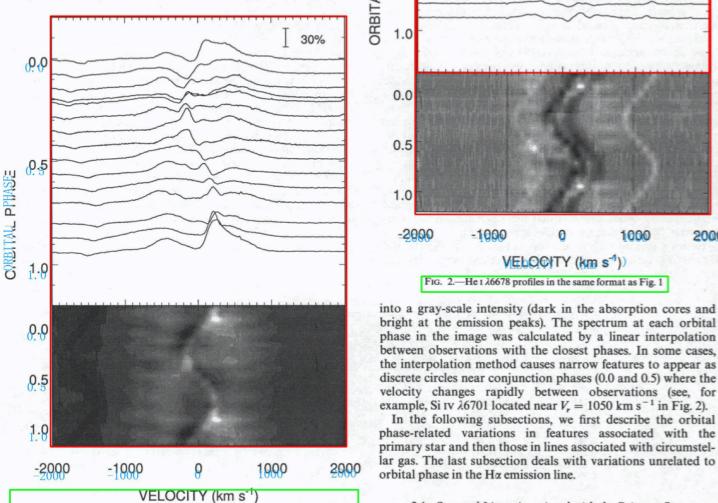


FIG. 1.—Top: H α line profiles plotted against heliocentric radial velocity. The profiles are arranged in order of increasing orbital phase and each spectrum is placed in the y-ordinate so that the continuum equals the phase of observation (phase 0.0 = superior conjunction of the primary). The bar in the upper right gives the spectrum intensity scale relative to a unit continuum. The fourth spectrum from the top was obtained two years after the other spectra. Bottom: A gray-scale representation of the profile variations shown above. Here each spectral intensity is assigned one of 16 gray levels based on its value between the minimum (deepest absorption) and maximum (highest emission) observed values. The spectrum at each phase in the image is calculated by a linear interpolation between the closest observed phases. The spectral image for the first and last 20% of the orbit are reproduced at the bottom and top of the image, respectively, to improve the sense of phase continuity.

Fig. 1—Top: H α line profiles plotted against heliocentric radial velocity. The profiles are arranged in order of increasing orbital phase and each spectrum is placed in the y-ordinate so that the continuum equals the phase of observation (phase 0.0 = superior conjunction of the primary). The bar in the upper right gives the spectrum intensity scale relative to a unit continuum. The fourth spectrum from the top was obtained two years after the other spectra. Bottom: A gray-scale representation of the profile variations shown above. Here each spectral intensity is assigned one of 16 gray levels based on its value between the minimum (deepest absorption) and maximum (highest emission) observed values. The spectrum at each phase in the image is calculated by a linear interpolation between the closest observed phases. The spectral image for the first and last 20% of the orbit are reproduced at the bottom and top of the image, respectively, to improve the sense of phase continuity.

FIG. 2.—He I $\lambda\lambda 6678$ profiles in the same format as Fig. 1.

into a gray-scale intensity (dark in the absorption cores and bright at the emission peaks). The spectrum at each orbital phase in the image was calculated by a linear interpolation between observations with the closest phases. In some cases, the interpolation method causes narrow features to appear as discrete circles near conjunction phases (0.0 and 0.5) where the velocity changes rapidly between observations (see, for example, Si IV 26701 located near $V_r = 1050$ km s $^{-1}$ in Fig. 2).

In the following subsections, we first describe the orbital phase-related variations in features associated with the primary star and then those in lines associated with circumstellar gas. The last subsection deals with variations unrelated to orbital phase in the H α emission line.

2.1. Spectral Lines Associated with the Primary Star

There are several lines in Figures 1 and 2 which show the familiar "s" shape corresponding to the velocity curve of the primary star. These include the absorption lines He I $\lambda\lambda 6678$ and He II $\lambda\lambda 6527.13, 6683.20$ and the emission lines Si IV $\lambda\lambda 6667.56, 6701.21$ (Fullerton 1989). All of these lines are typically found in mid-O supergiants (Conti 1974; Ninkov, Walker, & Yang 1987a; Ninkov et al. 1987b). In addition, we observed a weak emission line (maximum intensity $\approx 4\%$ of the continuum) near 6610 Å which we tentatively identify as N II $\lambda 6610.58$. Meinel, Aveni, & Stockton (1981) offer this iden-

© American Astronomical Society • Provided by the NASA Astrophysics Data System

| | | |
|----|---|------------------|
| 10 | <ul style="list-style-type: none"> • Logical regression | OIS 8 (ISL 4) |
| 12 | <ul style="list-style-type: none"> • k-Nearest neighbor classification and regression | ISL 2.2.3, 4.6.5 |
| 13 | <ul style="list-style-type: none"> • Intro to Unsupervised linear models: Principle component analysis | ISL 10.0-10.2 |



Making predictions
from data = intro to
machine learning
methods

Quick activity!

On a piece of paper or in notes on your computer:

- What are the most memorable movies you saw over the last year?
- Do you prefer cats or dogs?
- How would you quantify your experience in statistics?
- How many chairs or tables are there in your room?

Breakout group & Jamboard - if you were to summarize your datasets with 1 or 2 numbers (or something else...?) how would you do it?

Use jamboard to draw the approximate distributions in the survey and show visually what this measurement would look like. Actual calculations not required!

Don't forget to say hi to each other!

Quick activity!

On a piece of paper or in notes on your computer:

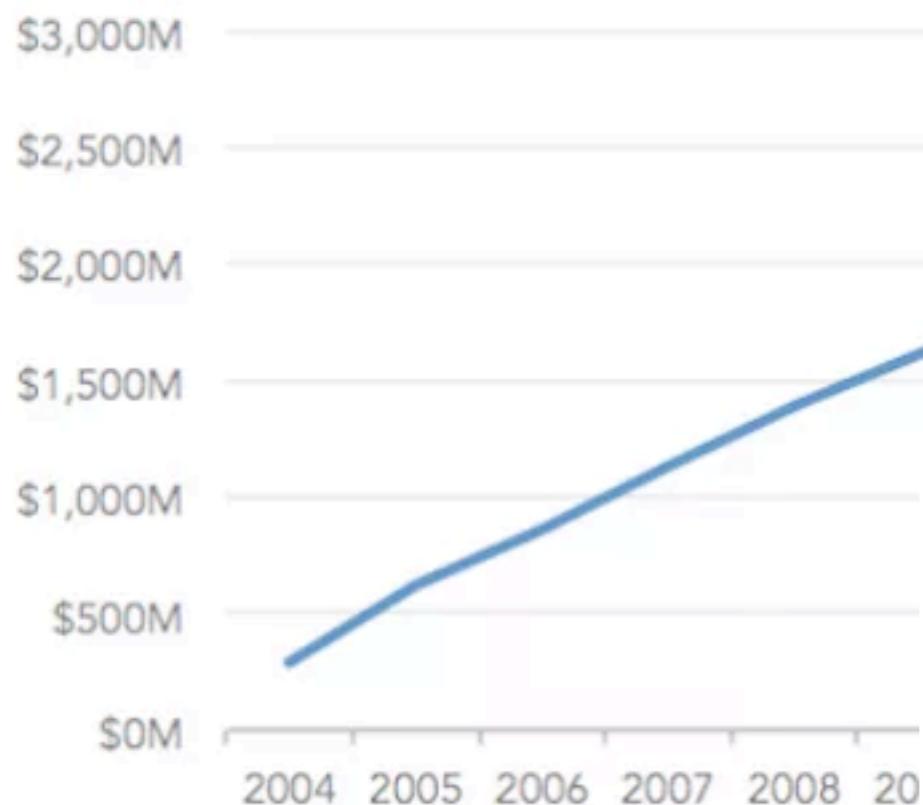
- What are the most memorable movies you saw over the last year?
- Do you prefer cats or dogs?
- How would you quantify your experience in statistics?
- How many chairs or tables are there in your room?

As a group: what was easy/hard about summarizing your datasets in this way? How confident are you of your summaries?

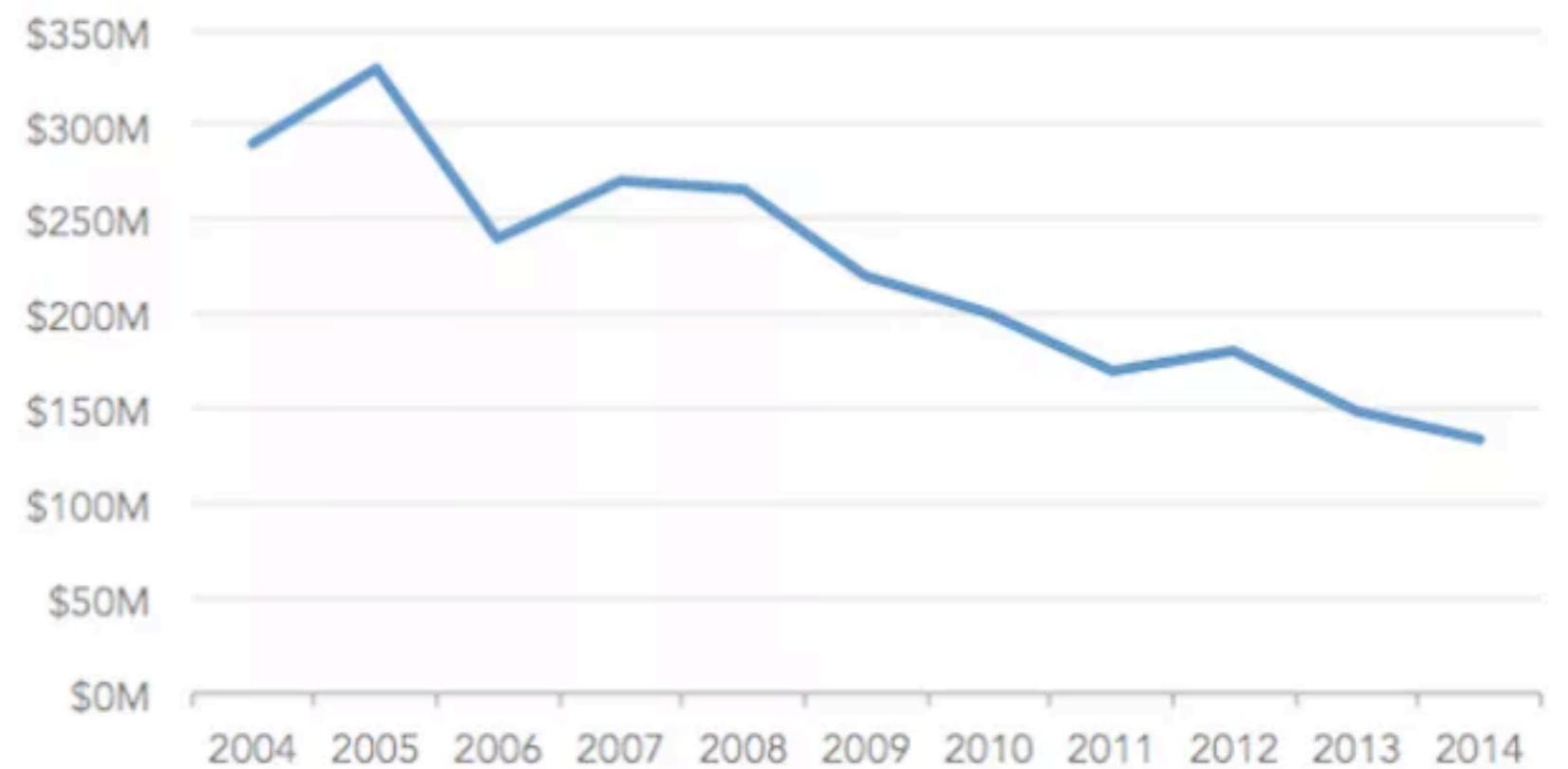
Lying... with Data!

How about if I ask about how profitable your company?

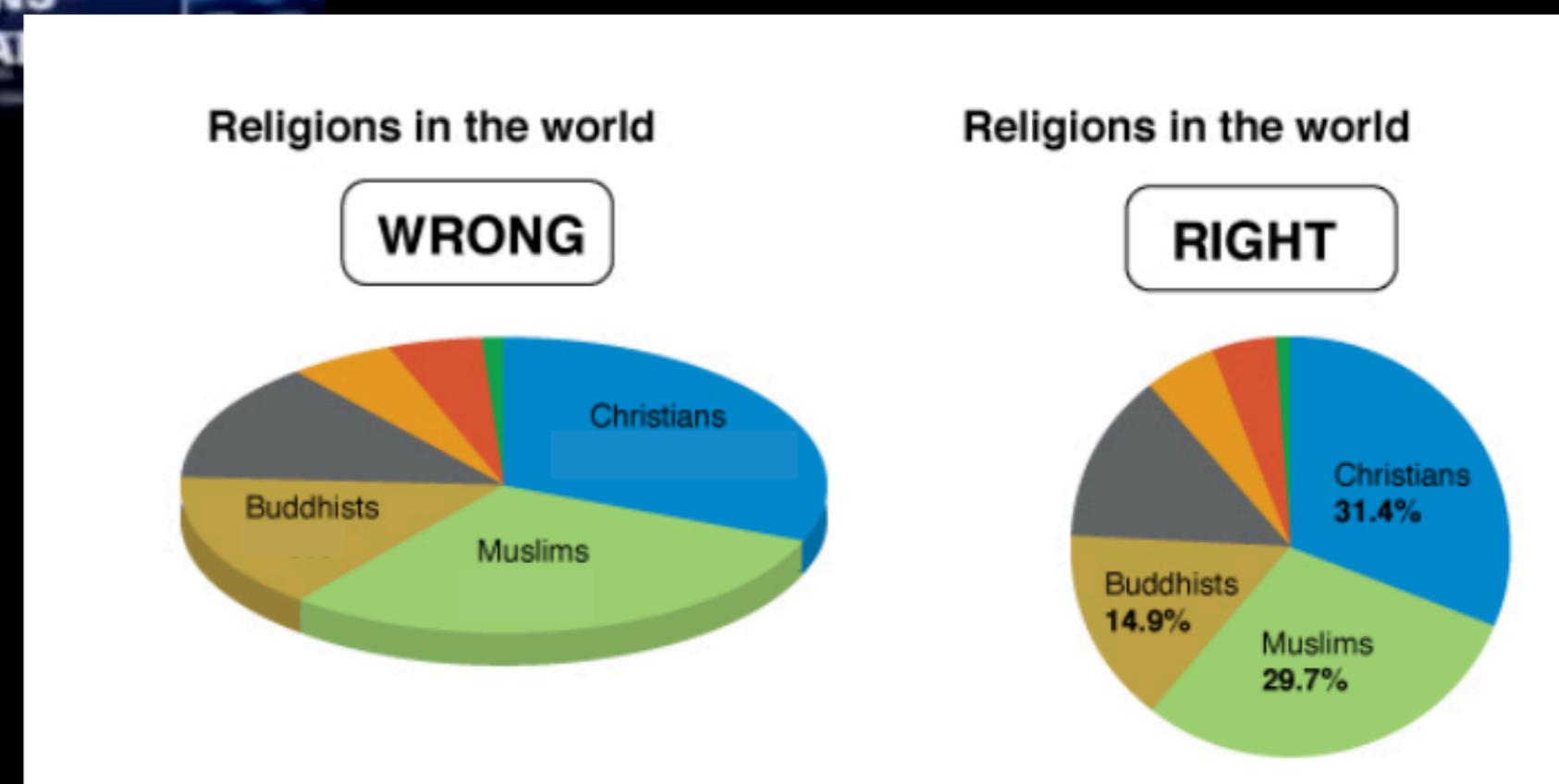
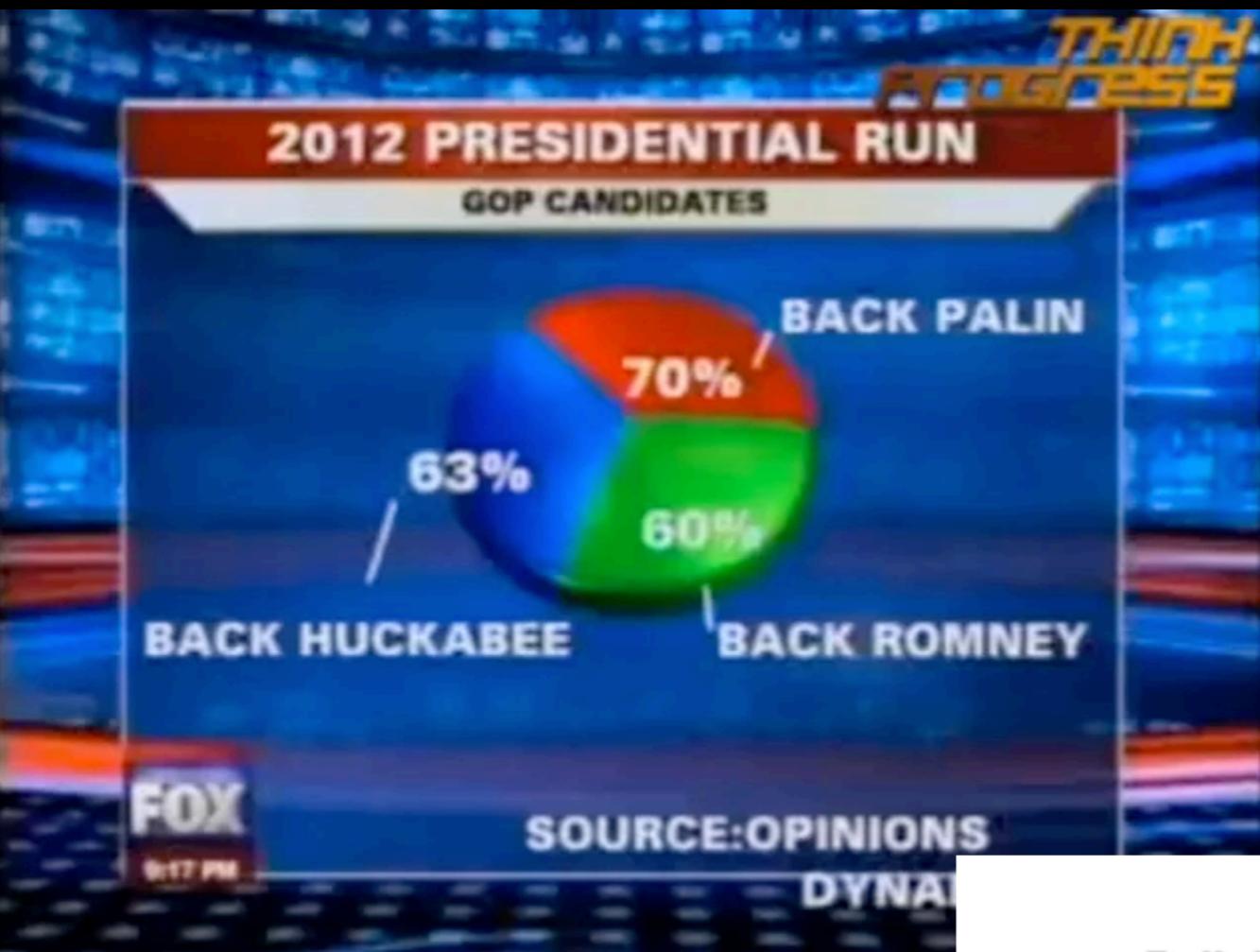
Cumulative Annual Revenue



Annual Revenue



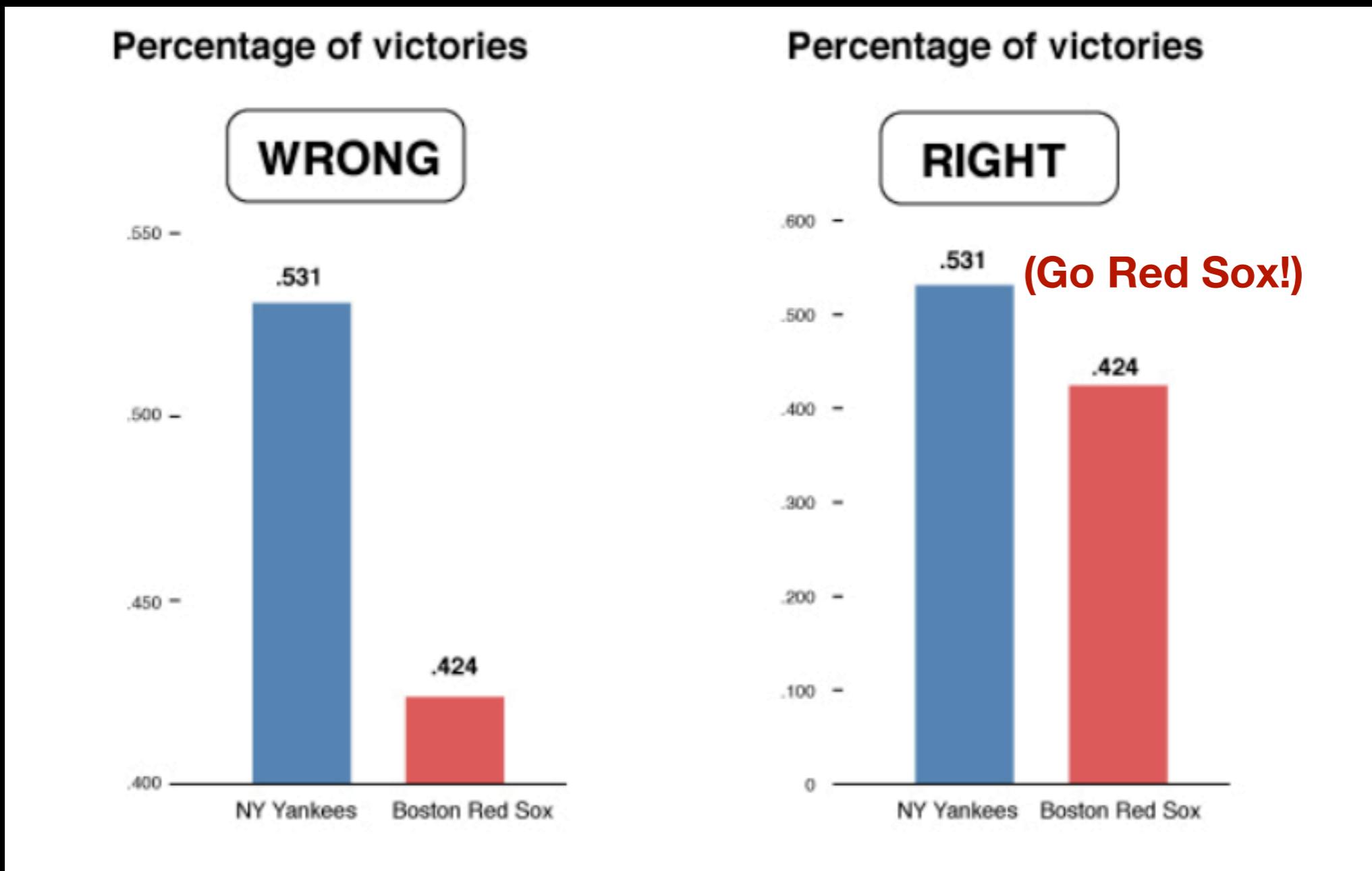
Lying... with Data!



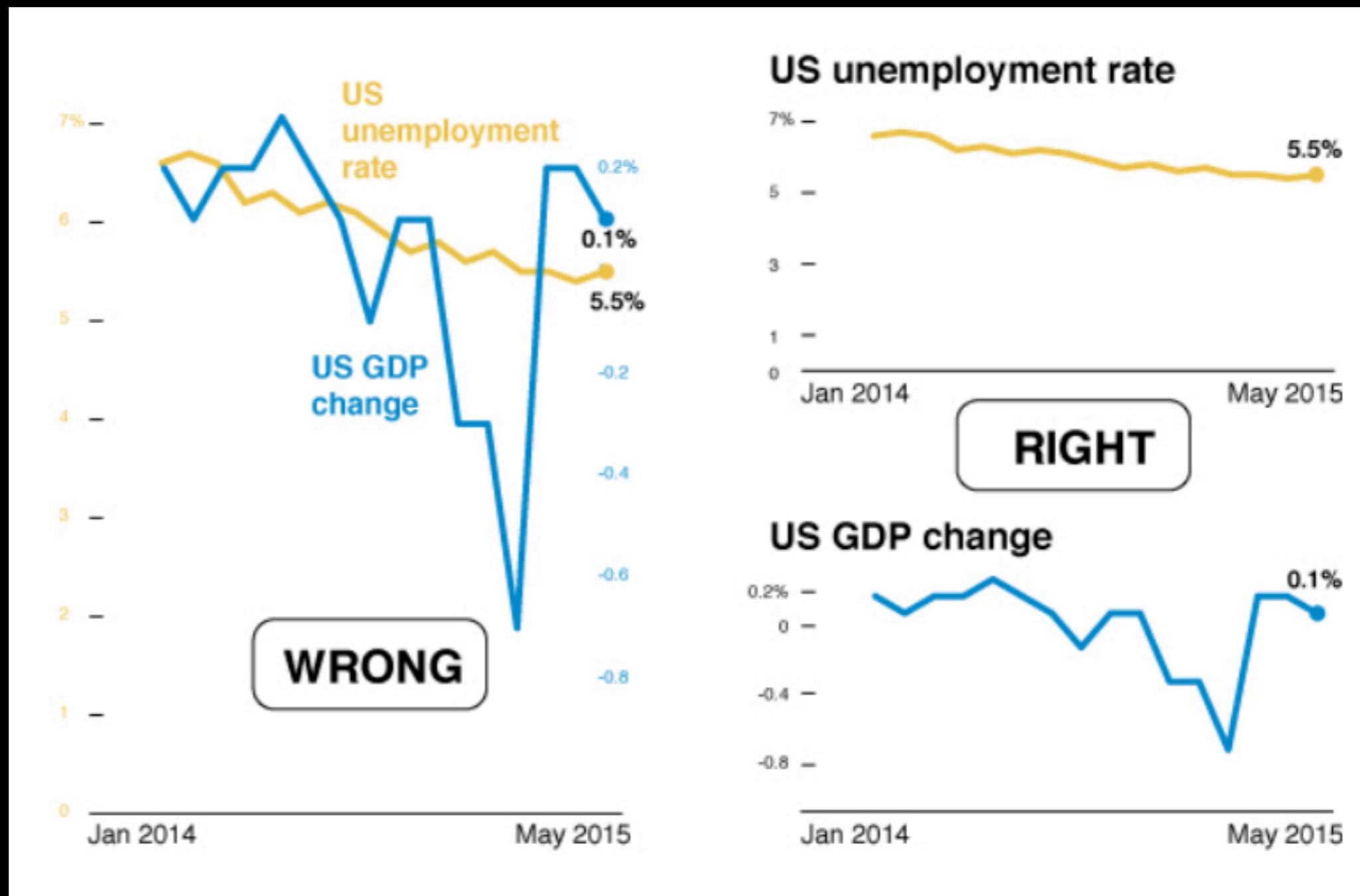
<https://gizmodo.com/how-to-lie-with-data-visualization-1563576606>

<https://news.nationalgeographic.com/2015/06/150619-data-points-five-ways-to-lie-with-charts/>

Lying... with Data!



Lying... with Data!



<http://www.tylervigen.com/spurious-correlations>

Plots can mislead people!

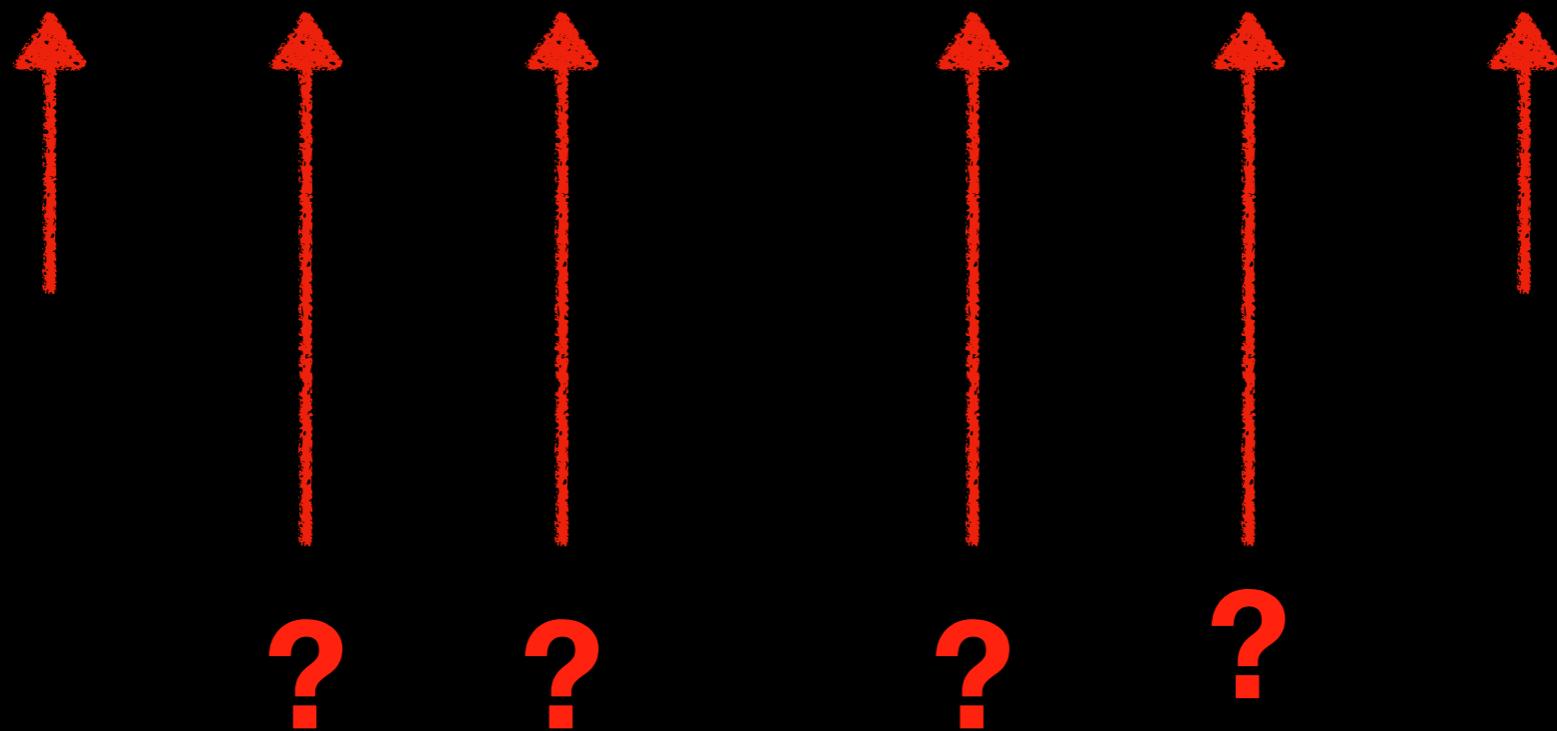
Proceed with caution.

Let's get plotting!

... but first, let's all make sure R & R-Studio are installed!

summary(after)

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|-------|---------|--------|-------|---------|-------|
| 39.00 | 41.25 | 44.50 | 44.21 | 46.75 | 50.00 |



Summary Statistic Definitions!

Mean (Sample) = sum of all data values divided by number of data points

$$\text{Mean} = \frac{\text{Sum of all values}}{\text{Number of values}}$$

$$\text{Symbolically, } \bar{x} = \frac{\sum x}{n}$$

where \bar{x} (read as 'x bar') is the mean of the set of x values,
 $\sum x$ is the sum of all the x values, and
 n is the number of x values.

(note - only works with “numerical” data types... more about data types later)

Median = if we order the data from smallest to largest, this is the observation in the middle (splits the data in 2 halves)

First/Third Quartiles = where 25% of the data falls below/above

Standard Deviation = this is the square root of the variance, where the variance is roughly the average distance of data values from the mean

$$\text{Standard Deviation (sample)} = \sqrt{\frac{\sum_{i=1}^n (x_i - \text{mean})^2}{n-1}}$$

Summary Statistic Definitions!

Mean (Sample) = sum of all data values divided by number of data points

$$\text{Mean} = \frac{\text{Sum of all values}}{\text{Number of values}}$$

$$\text{Symbolically, } \bar{x} = \frac{\sum x}{n}$$

where \bar{x} (read as 'x bar') is the mean of the set of x values,
 $\sum x$ is the sum of all the x values, and
 n is the number of x values.

(note - only works with “numerical” data types... more about data types later)

Median = if we order the data from smallest to largest, this is the observation in the middle (splits the data in 2 halves)

First/Third Quartiles = where 25% of the data falls below/above

Standard Deviation = this is the square root of the variance, where the variance is roughly the average distance of data values from the mean

$$\text{Standard Deviation (sample)} = \sqrt{\frac{\sum_{i=1}^n (x_i - \text{mean})^2}{n-1}}$$