Welcome to IS542!

Outline of lectures:

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

Note: today will be a little lecture heavy

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 con mphasizing the merits and limitations of famili s parametric and non-parametric predictive m ut these discussions, the course focuses on sks in information management (e.g. prec

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.

Learning Objectives

Students will demonstr s of a diverse range of data sets. By the en what one can feasibly learn and predict with c cular, each student will be able to:

- Articulate the rol
- Select, paramete
- 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 con mphasizing the merits and limitations of famili s parametric and non-parametric predictive m ut these discussions, the course focuses on sks in information management (e.g. prec

How well can we know anything, really?

What lata?

How accurate m a particular dataset? How s dataset (e.g.

How can we ? What is the

How can w statistical questions.

Learning Objectives

Students will demonstr s of a diverse range of data sets. By the en what one can feasibly learn and predict with c cular, each student will be able to:

- Articulate the rol
- Select, paramete
- 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

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* Third 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	• Data, Models, and Information • Elementary statistics: Definitions • Overview of R	OIS 1 (ISL 1)
2	• Elementary statistics: Applications & Plots	OIS 1 (ISL 1)
3	• Introduction to data analysis with R • Review of tabular and graphical displays of data	ITR 1, 2, 5, 6, 7, 12
4	• Random variables: expectation and variance • Joint and conditional probability • Bayes rule	OIS 2
5	• Random variables: distributions (normal, binomial, poisson)	OIS 3

Week	Topic	Reading
1	• Data, Models, and Information • Elementary statistics: Definitions • Overview of R	OIS 1 (ISL 1)
2	• Elementary statistics: Applications & Plots	OIS 1 (ISL 1)
3	• Introduction to data analysis with R • Review of tabular and graphical displays of data	ITR 1, 2, 5, 6, 7, 12
4	• Random variables: expectation and variance • Joint and conditional probability • Bayes rule	OIS 2
5	• Random variables: distributions (normal, binomial, poisson)	OIS 3

Definitions, basic concepts, R practice

Week	Topic	Reading
1	• Data, Models, and Information • Elementary statistics: Definitions • Overview of R	OIS 1 (ISL 1)
2	• Elementary statistics: Applications & Plots	OIS 1 (ISL 1)
3	• Introduction to data analysis with R • Review of tabular and graphical displays of data	ITR 1, 2, 5, 6, 7, 12
4	• Random variables: expectation and variance • Joint and conditional probability • Bayes rule	OIS 2
5	• Random variables: distributions (normal, binomial, poisson)	OIS 3

Probability basics, typical distributions

6	• Modeling data with probability distributions • Foundations for inference	OIS 4
7	• Inference for numerical data • Inference for categorical data	OIS 5, OIS 6
8	• Linear regression	OIS 7 (ISL 3)
9	• Multiple linear regression	OIS 8 (ISL 3)
10	• Logical regression	OIS 8 (ISL 4)
12	• *k*-Nearest neighbor classification and regression	ISL 2.2.3, 4.6.5
13	• Intro to Unsupervised linear models: Principle component analysis	ISL 10.0-10.2

How well can we answer questions with our data?

6	• Modeling data with probability distributions • Foundations for inference	OIS 4
7	• Inference for numerical data • Inference for categorical data	OIS 5, OIS 6
8	• Linear regression	OIS 7 (ISL 3)
9	• Multiple linear regression	OIS 8 (ISL 3)
10	• Logical regression	OIS 8 (ISL 4)
12	• *k*-Nearest neighbor classification and regression	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

6	• Modeling data with probability distributions • Foundations for inference	OIS 4
7	• Inference for numerical data • Inference for categorical data	OIS 5, OIS 6
8	• Linear regression	OIS 7 (ISL 3)
9	• Multiple linear regression	OIS 8 (ISL 3)
10	• Logical regression	OIS 8 (ISL 4)
12	• *k*-Nearest neighbor classification and regression	ISL 2.2.3, 4.6.5
13	• Intro to Unsupervised linear models: Principle component analysis	ISL 10.0-10.2

Classification and unsupervised learning

6	• Modeling data with probability distributions • Foundations for inference	OIS 4
7	• Inference for numerical data • Inference for categorical data	OIS 5, OIS 6
8	• Linear regression	OIS 7 (ISL 3)
9	•	
10	• Logical regression	OIS 8 (ISL 4)
12	• *k*-Nearest neighbor classification and regression	ISL 2.2.3, 4.6.5
13	• Intro to Unsupervised linear models: Principle component analysis	ISL 10.0-10.2

14	• Case studies	NA
15	• Case studies and review	NA

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	50%
Midterm exam	15%
Final exam	25%
Class engagement	10%

No late HW or exams, but we will drop your lowest HW score.

Assignment	
Weekly homework	50%
Midterm exam	15%
Final exam	25%
Class engagement	10%

No late HW or exams, but 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).

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

What are we doing?

How are we going to do it?

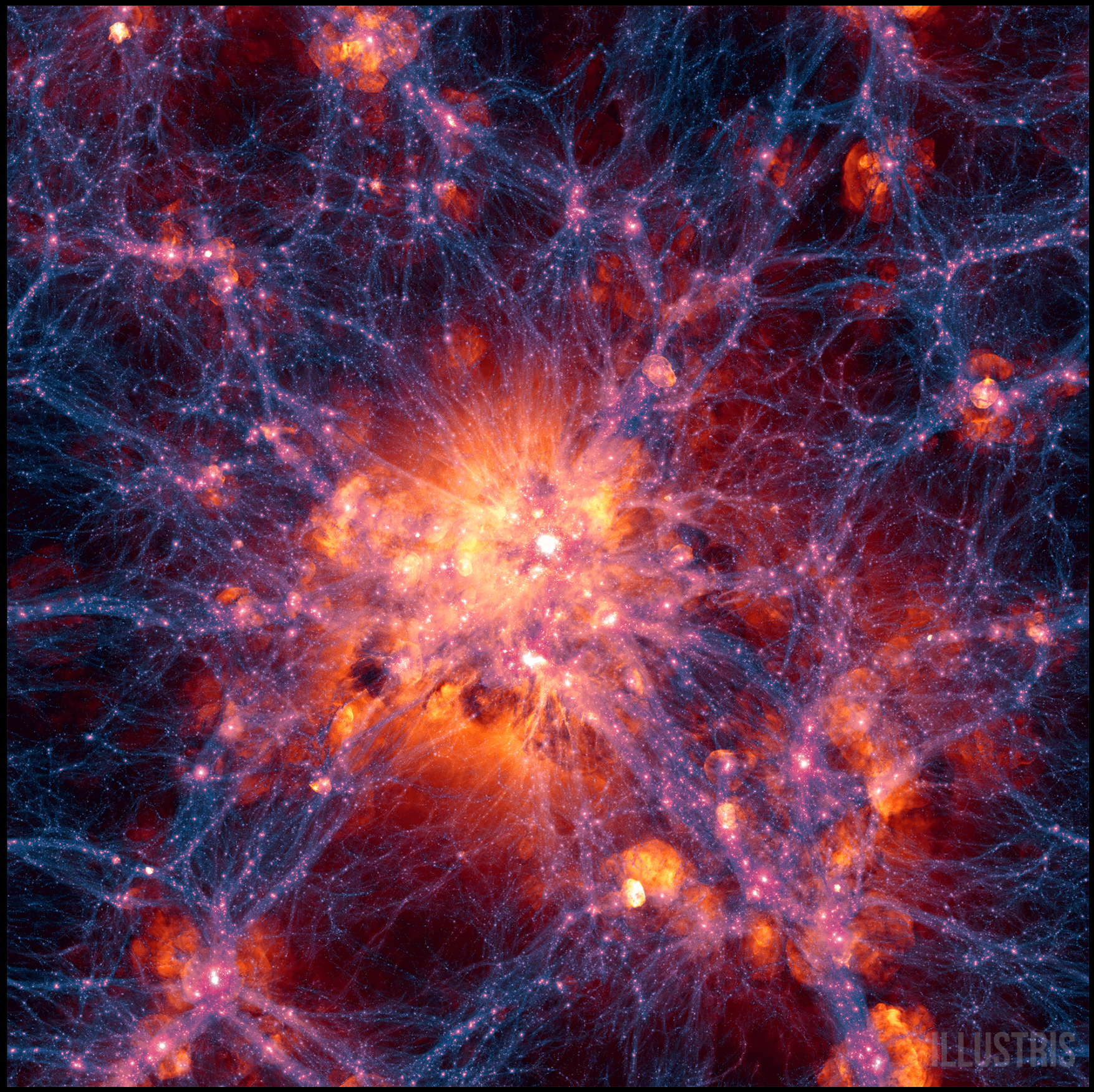
Who are you?

~~What are we doing?~~

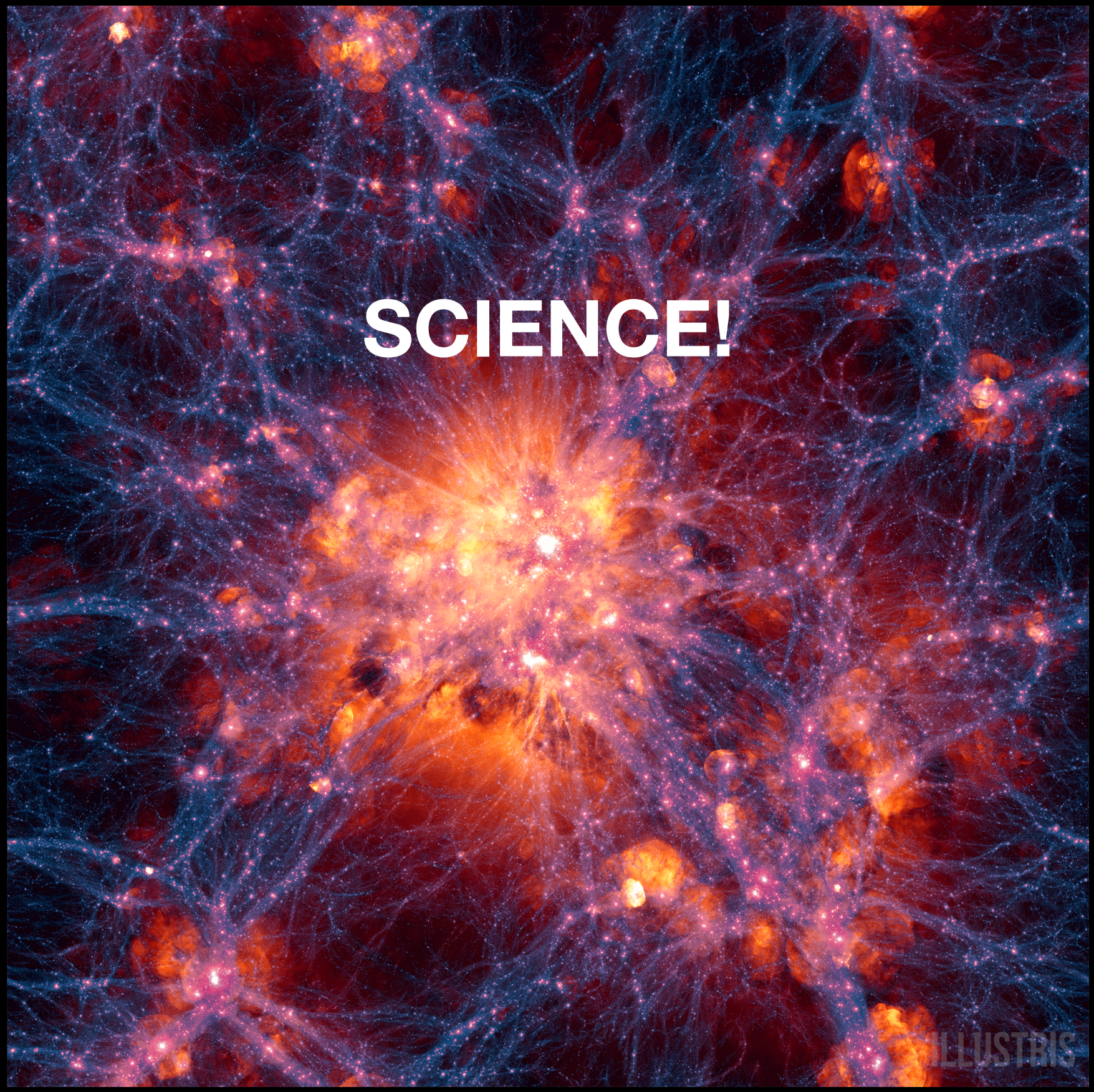
~~How are we going to do it?~~

Who are you?

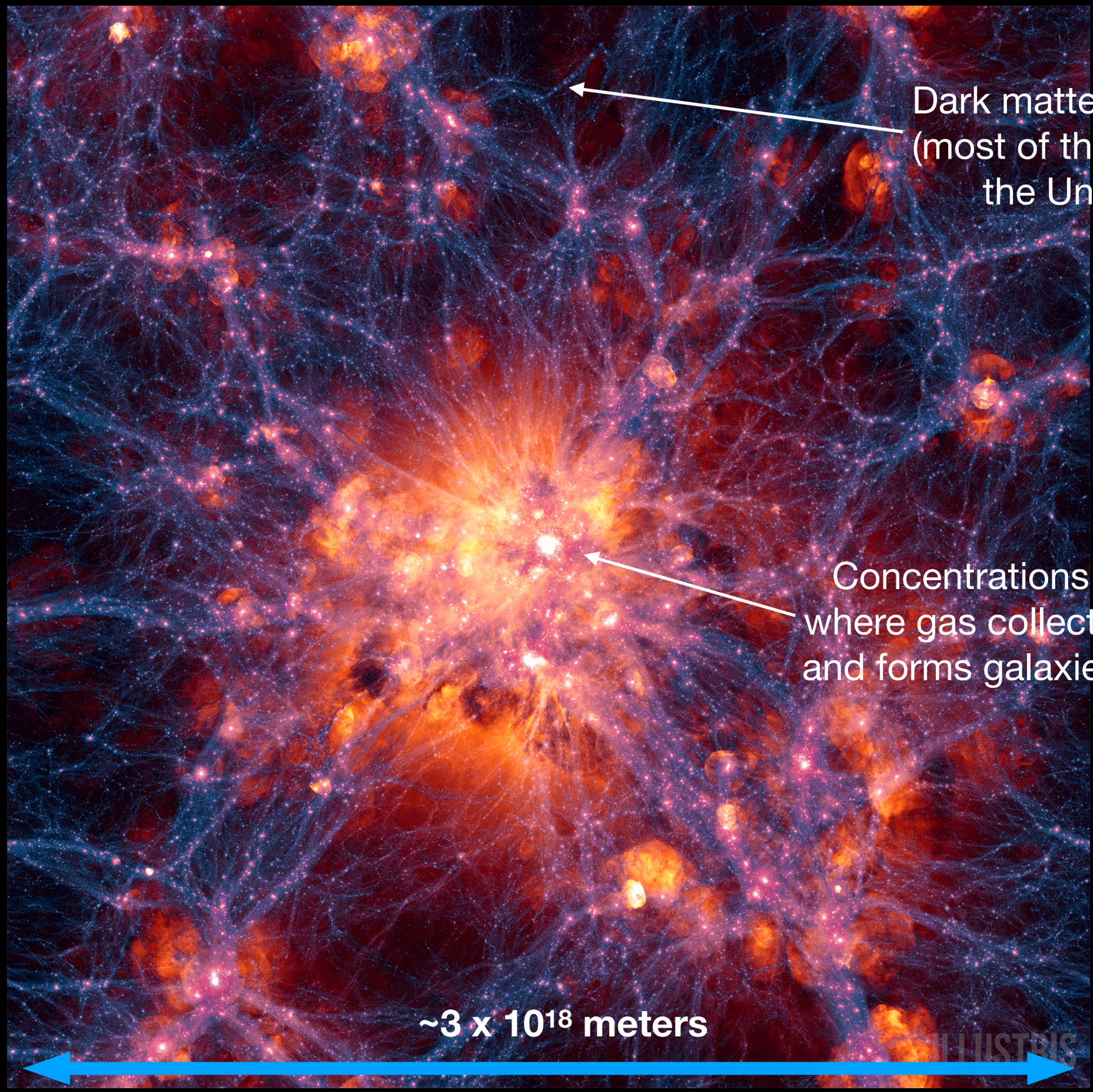
My background



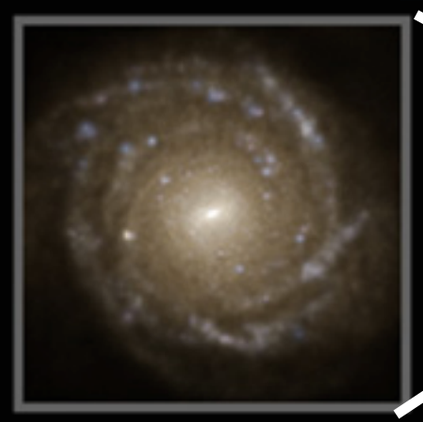
My background

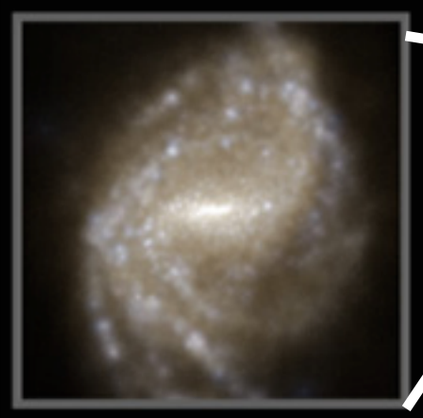


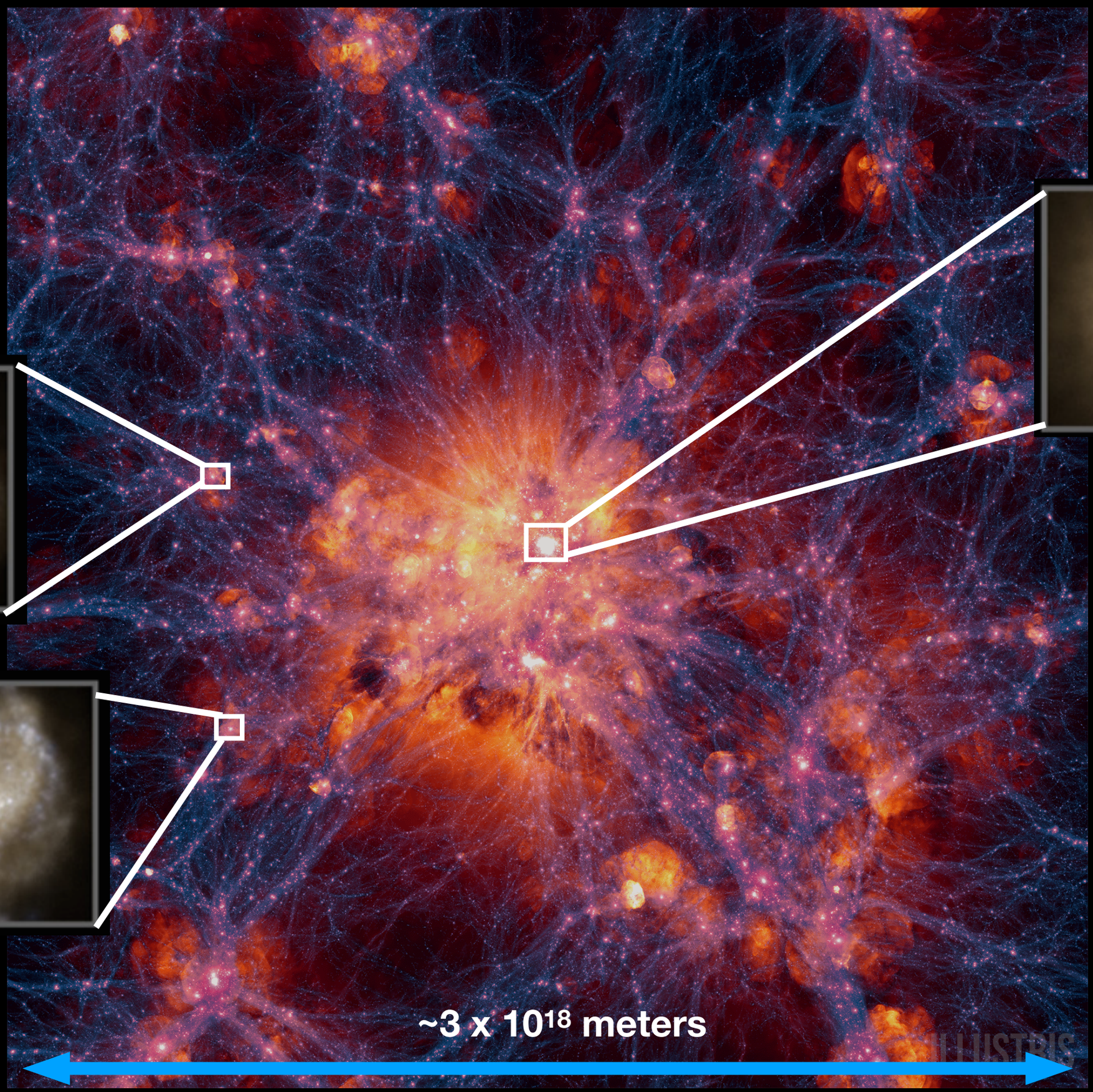
My background

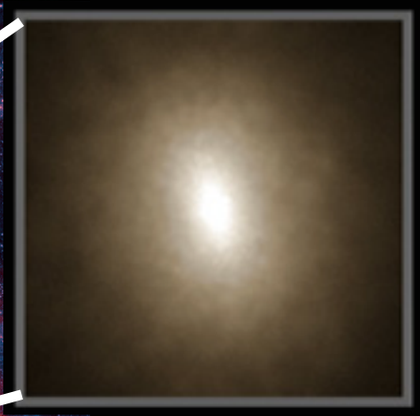


My background









Example: Supernova Explosions in the Milky Way

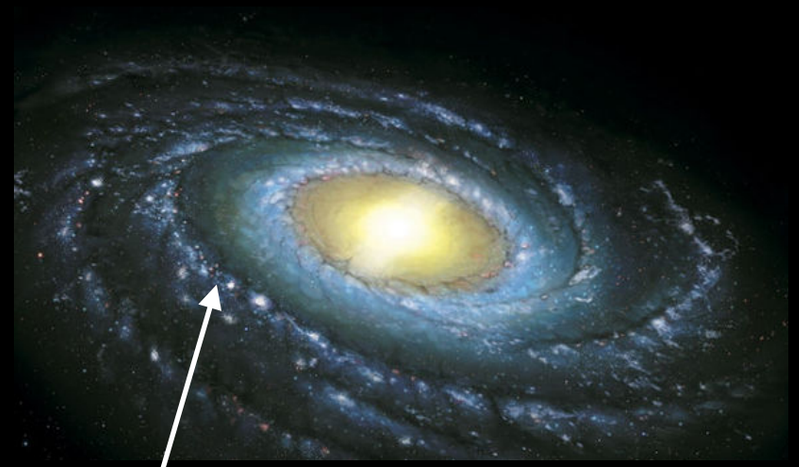
Example: Supernova Explosions in the Milky Way

Our galaxy

(artist concept)

Example: Supernova Explosions in the Milky Way

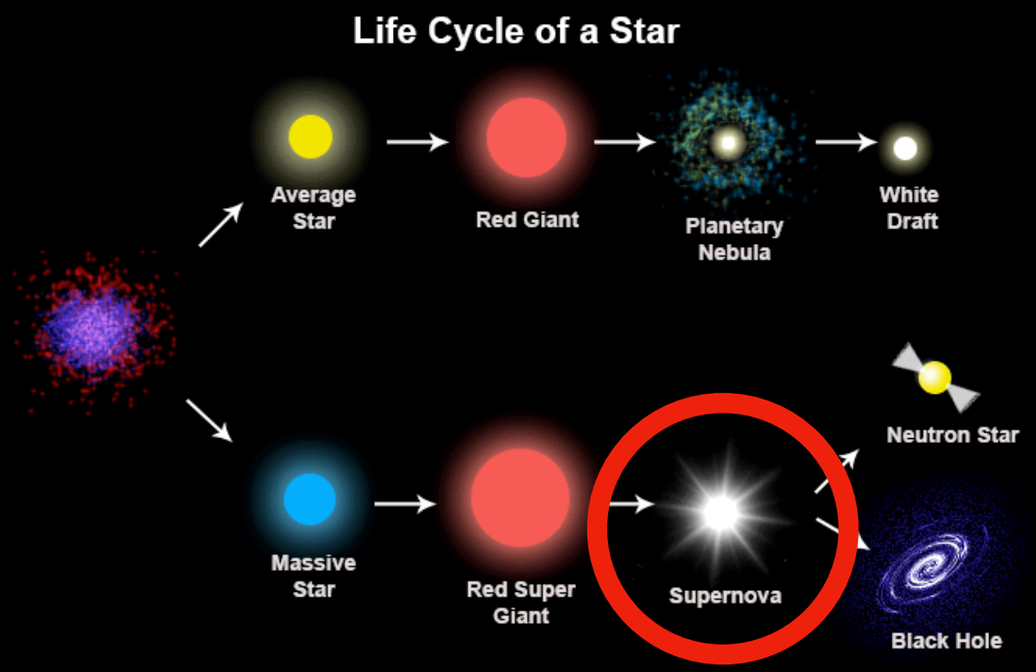
Our galaxy



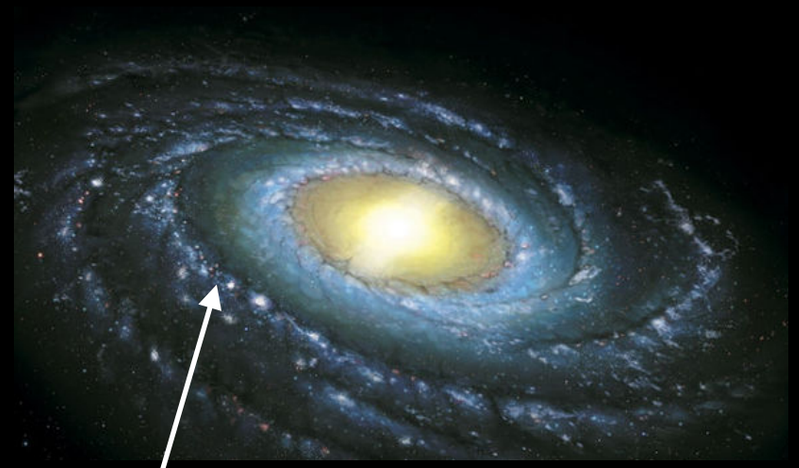
(artist concept)

You are here

Example: Supernova Explosions in the Milky Way



Our galaxy

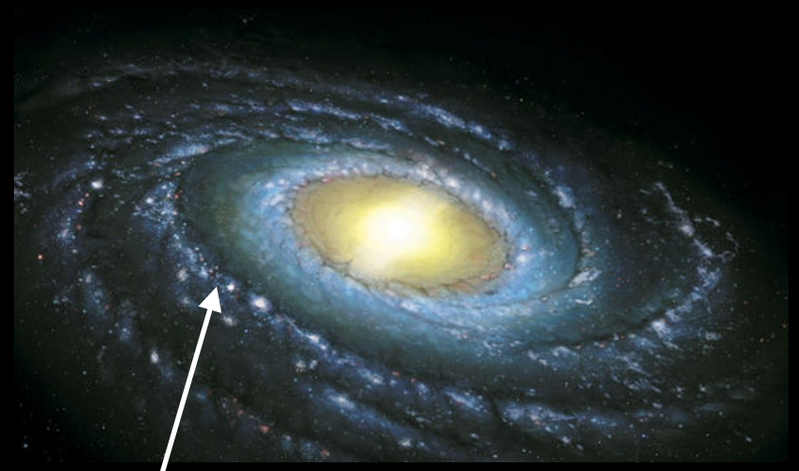


(artist concept)

You are here

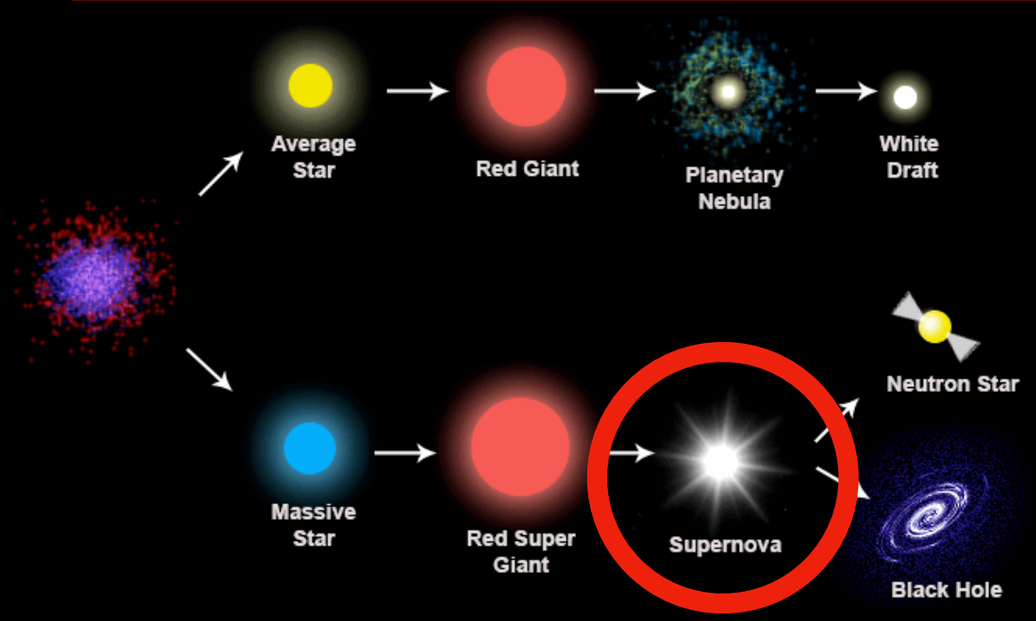
Example: Supernova Explosions in the Milky Way

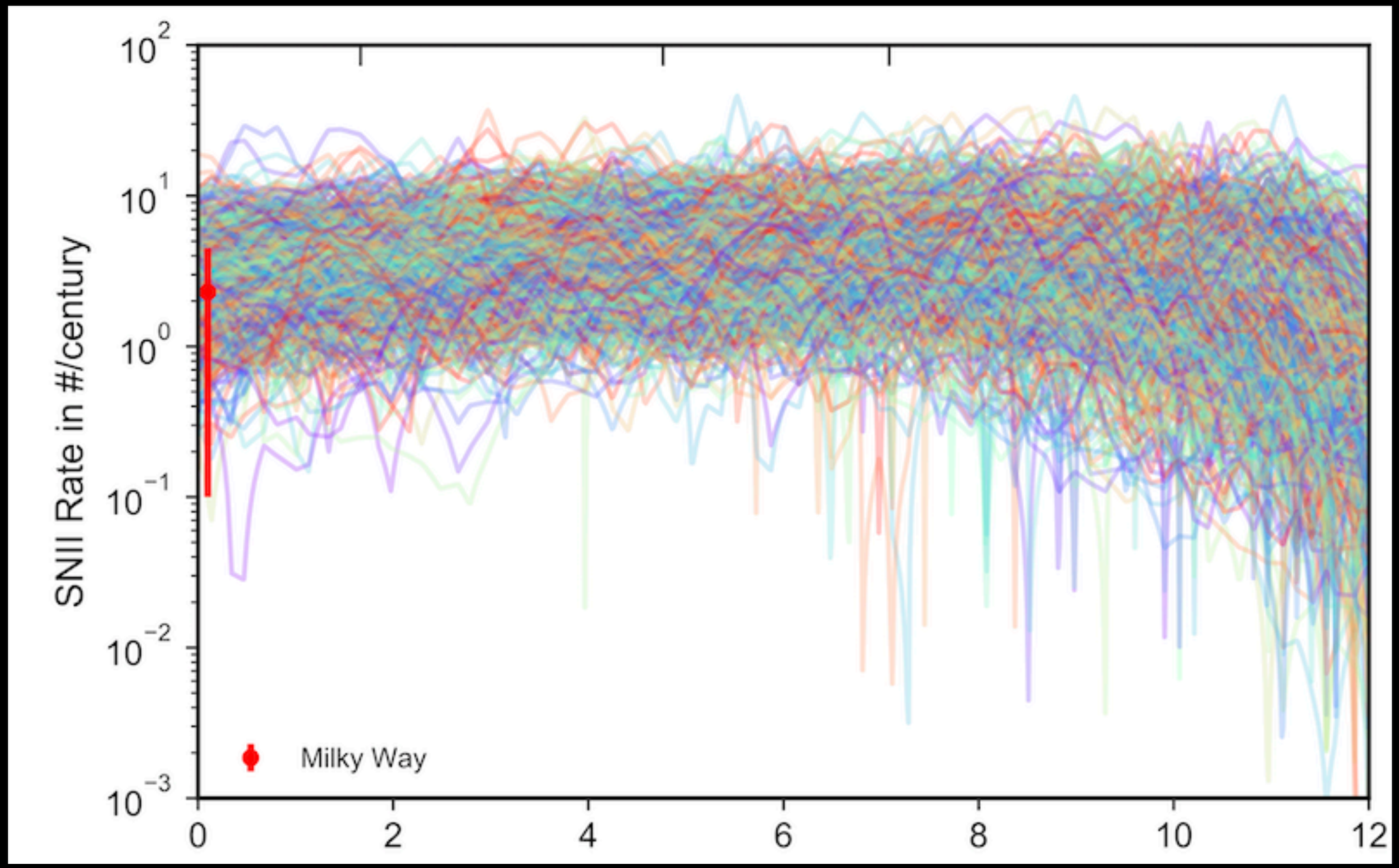
Our galaxy



What are typical supernovae rates (#/year) in galaxies that are similar to our Milky Way?
What determines the number of supernovae which happen in a Milky-Way like galaxy?

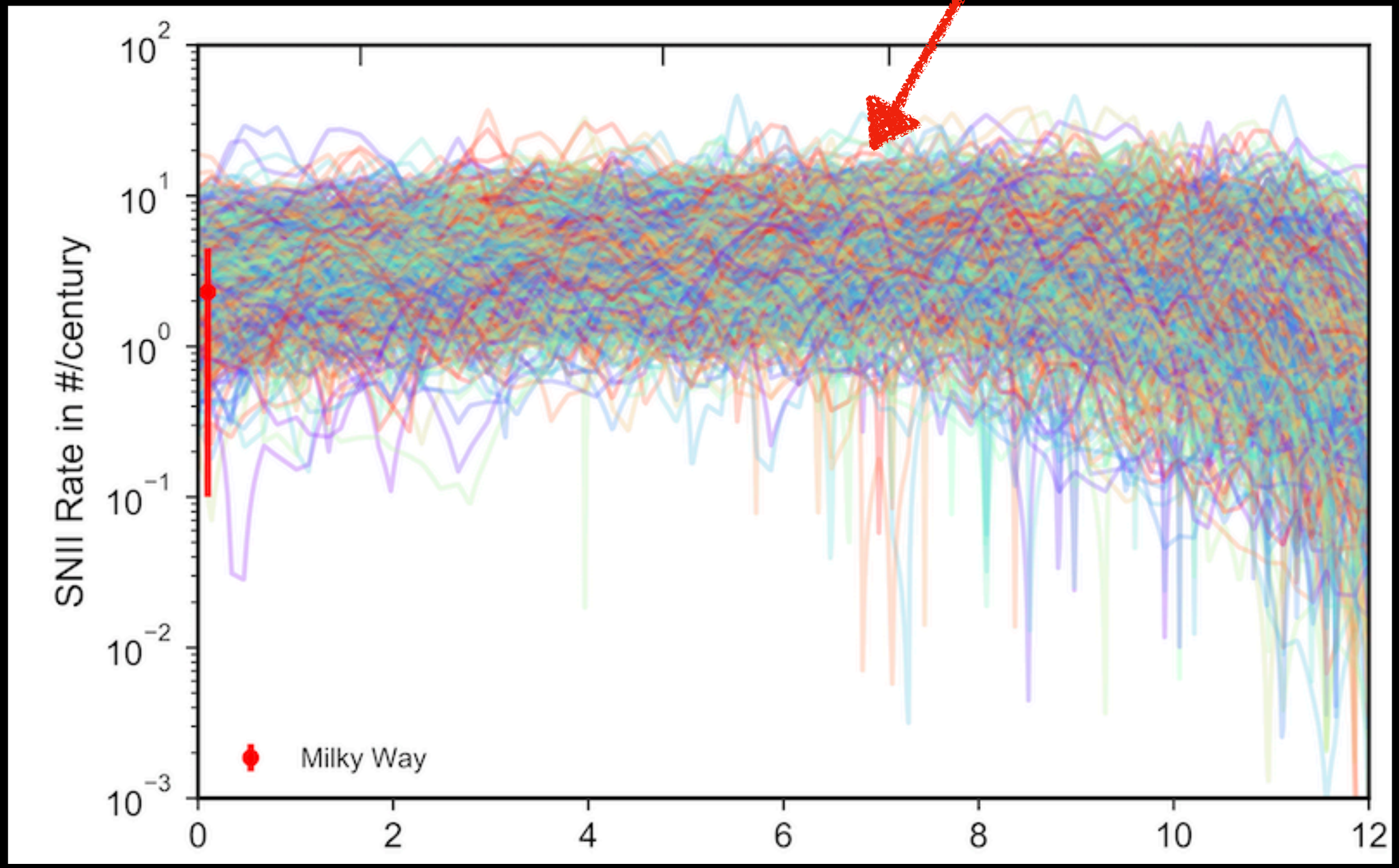
You are here





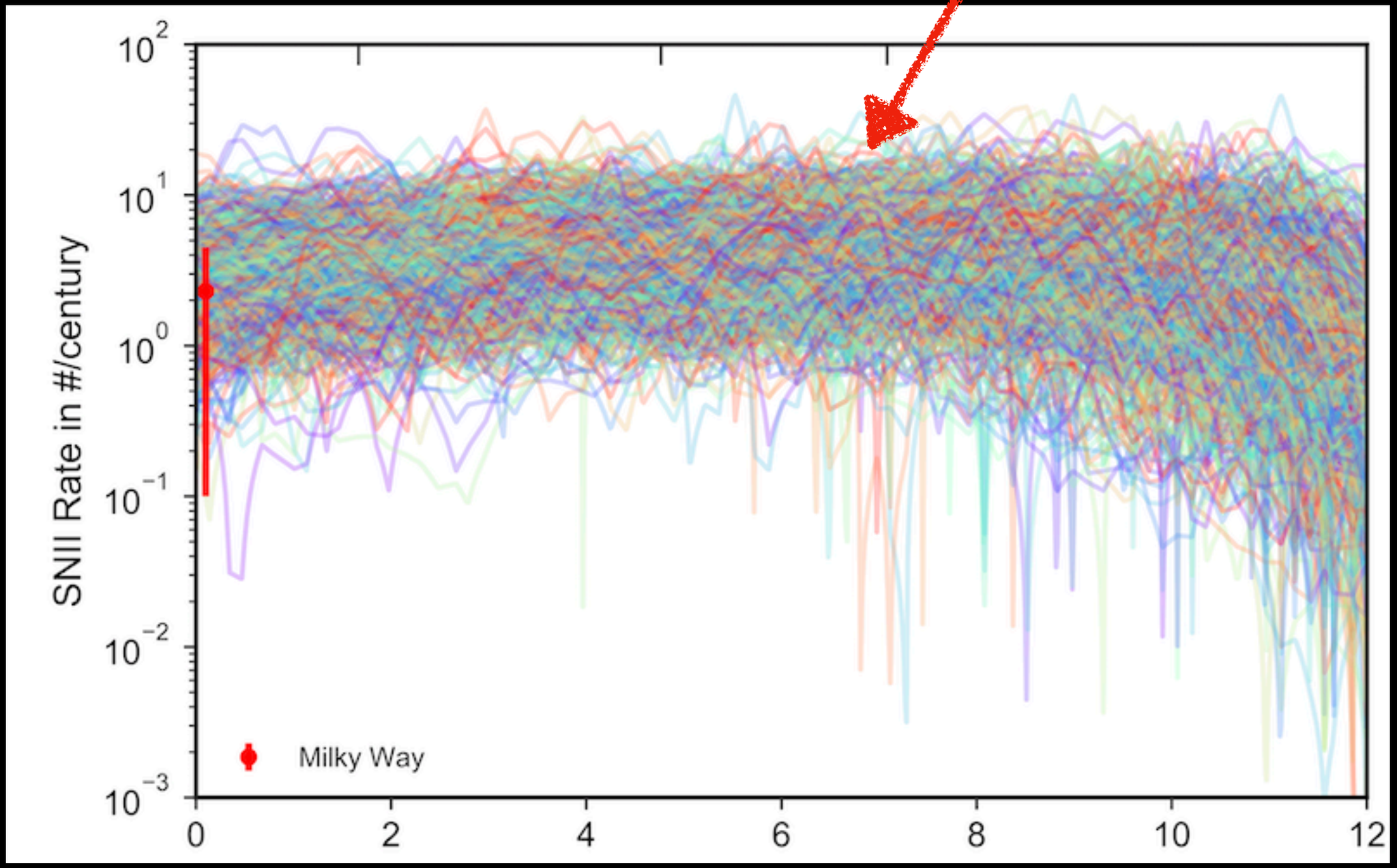
Time in Gyrs (10^9 years)

Each line is the # of supernova explosions in a Milky Way-like galaxy from when it was “born” to present day.



Time in Gyrs (10^9 years)

Each line is the # of supernova explosions in a Milky Way-like galaxy from when it was “born” to present day.

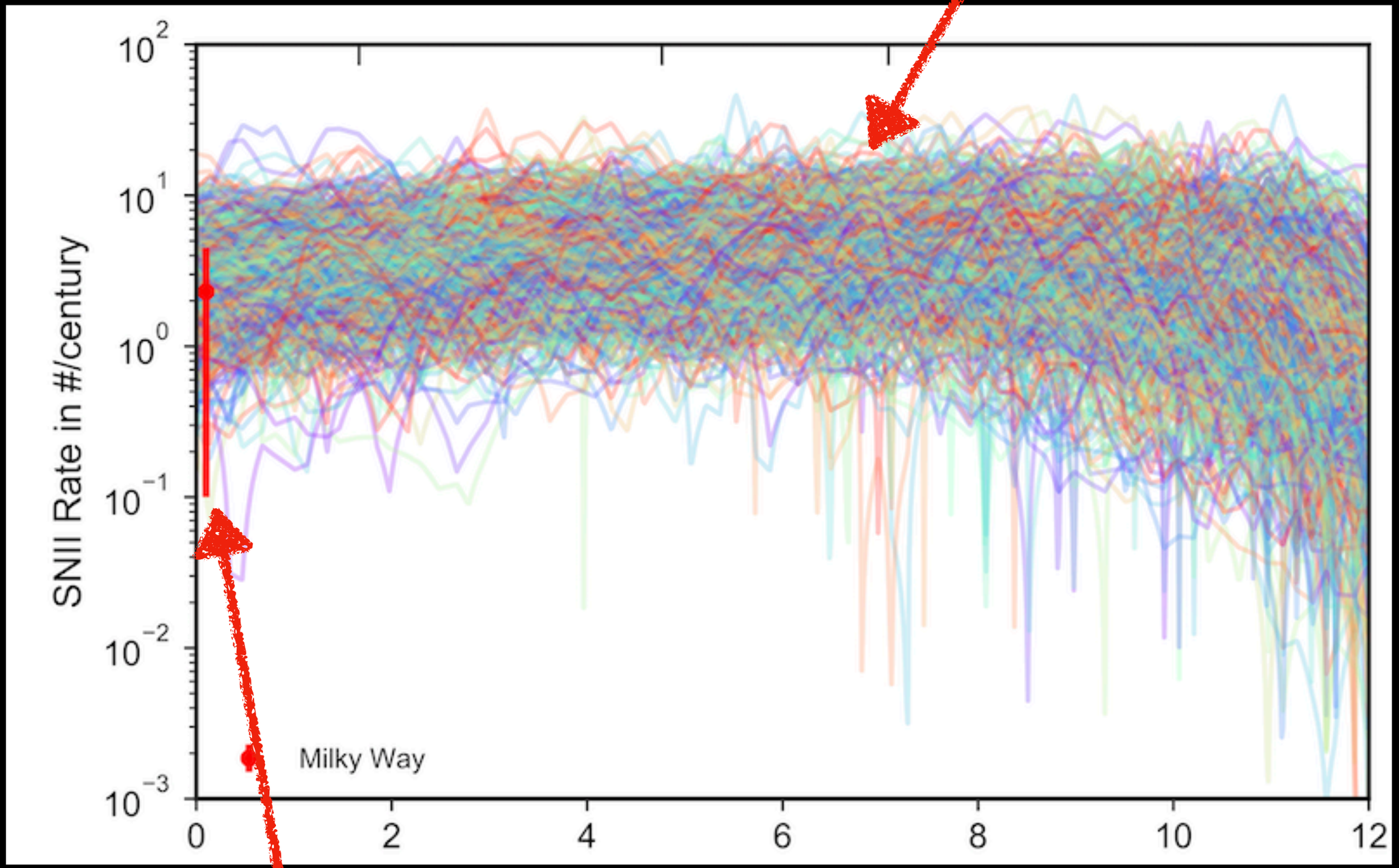


present-day

Time in Gyrs (10^9 years)

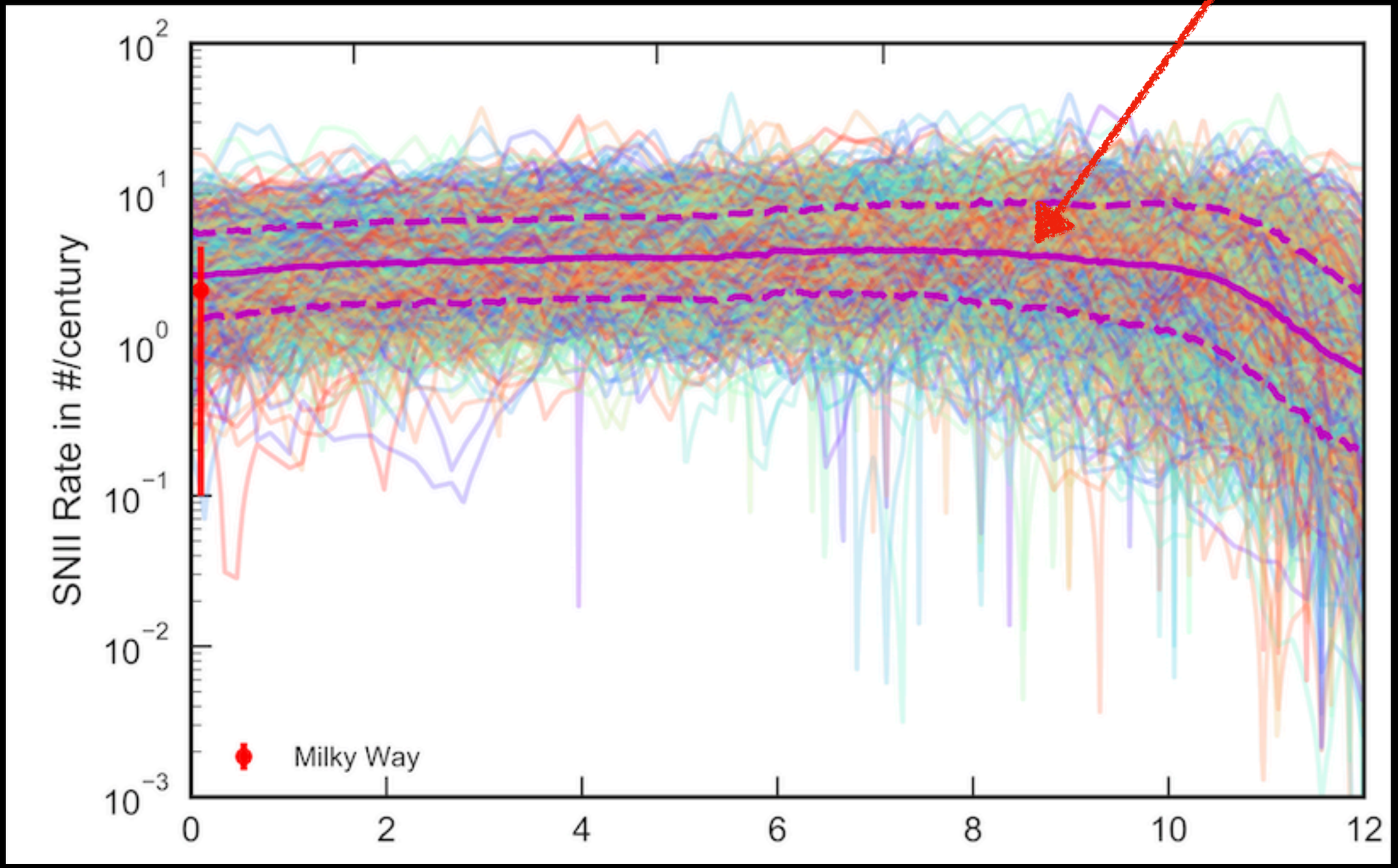
~Start of our Universe

Each line is the # of supernova explosions in a Milky Way-like galaxy from when it was “born” to present day.



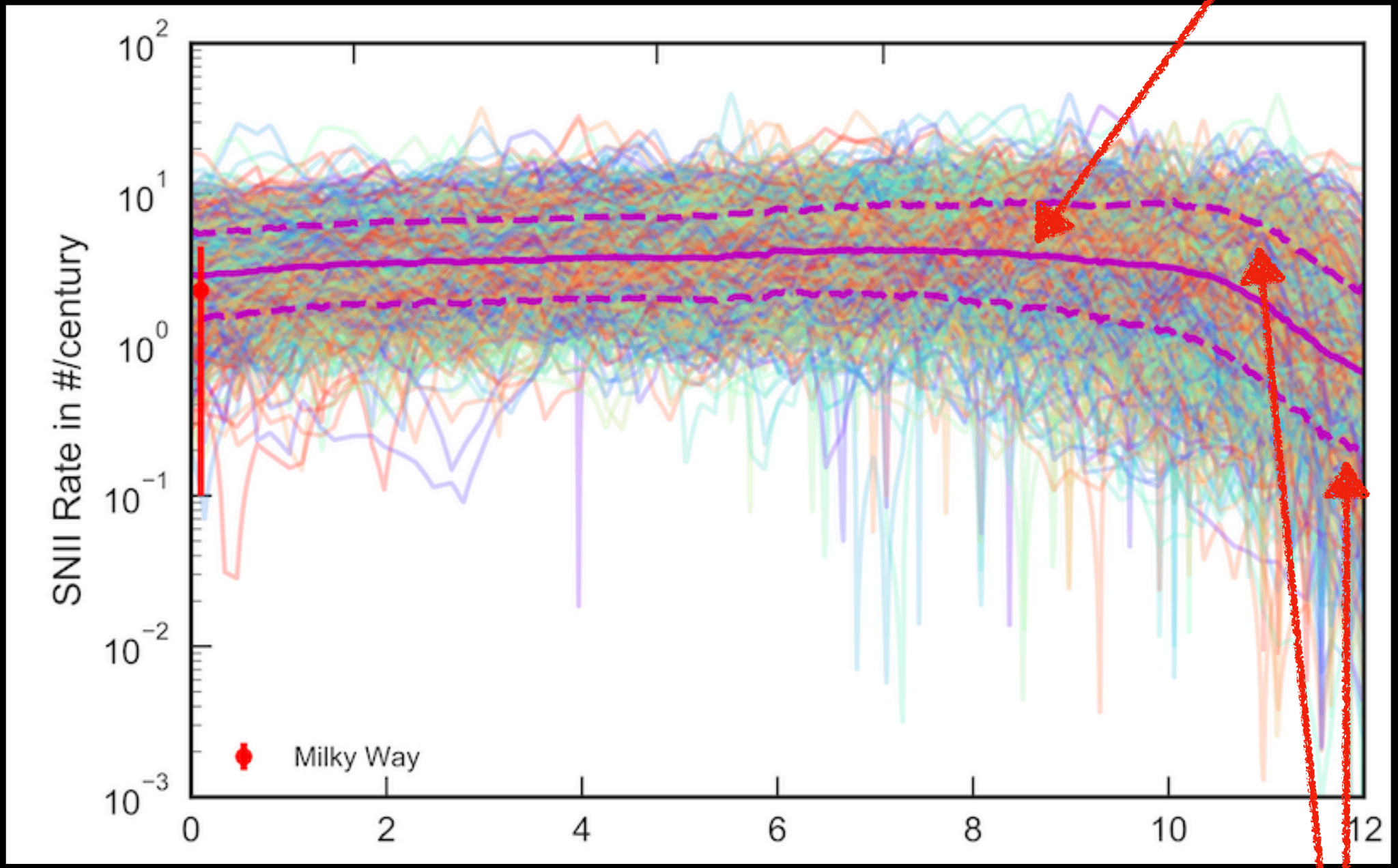
Observed in our own Milky Way Galaxy

“Typical" value - Median



Time in Gyrs (10^9 years)

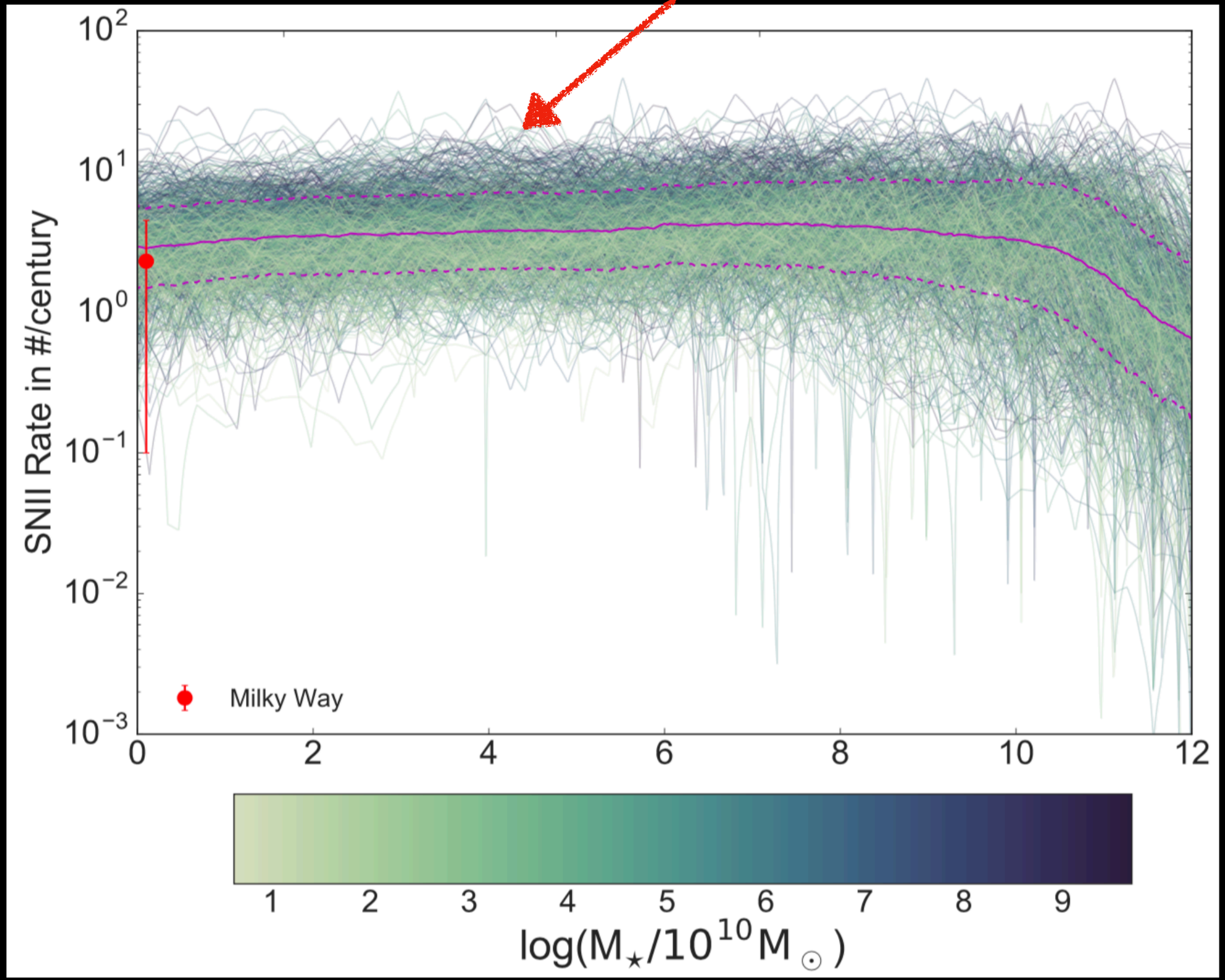
“Typical" value - Median



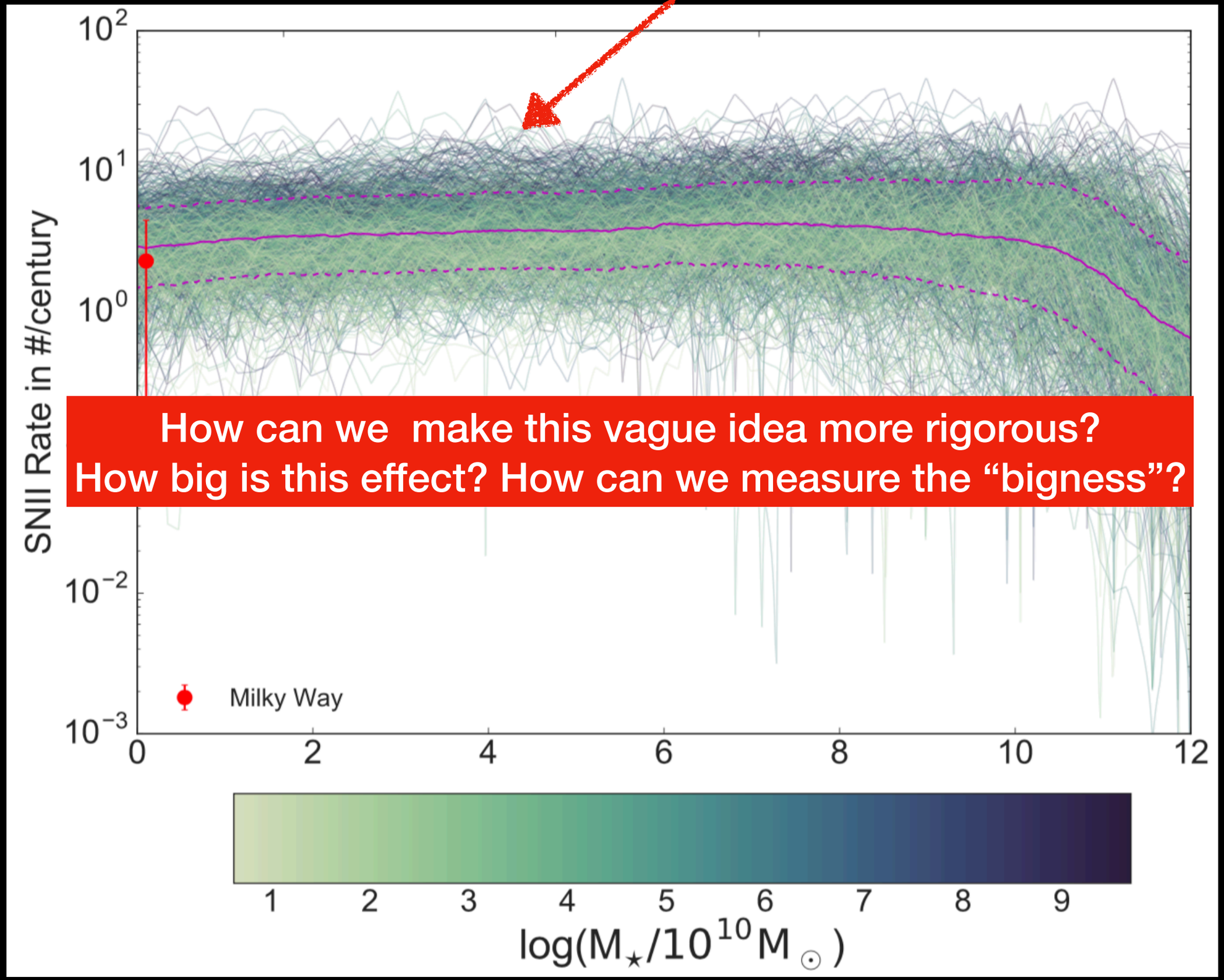
Time in Gyrs (10^9 years)

Standard Deviations around the median

More stars in a galaxy more supernovae explosions?

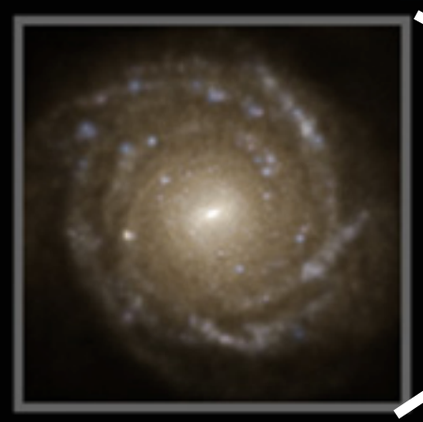


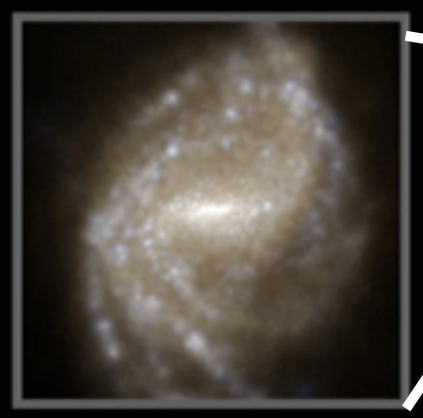
More stars in a galaxy more supernovae explosions?

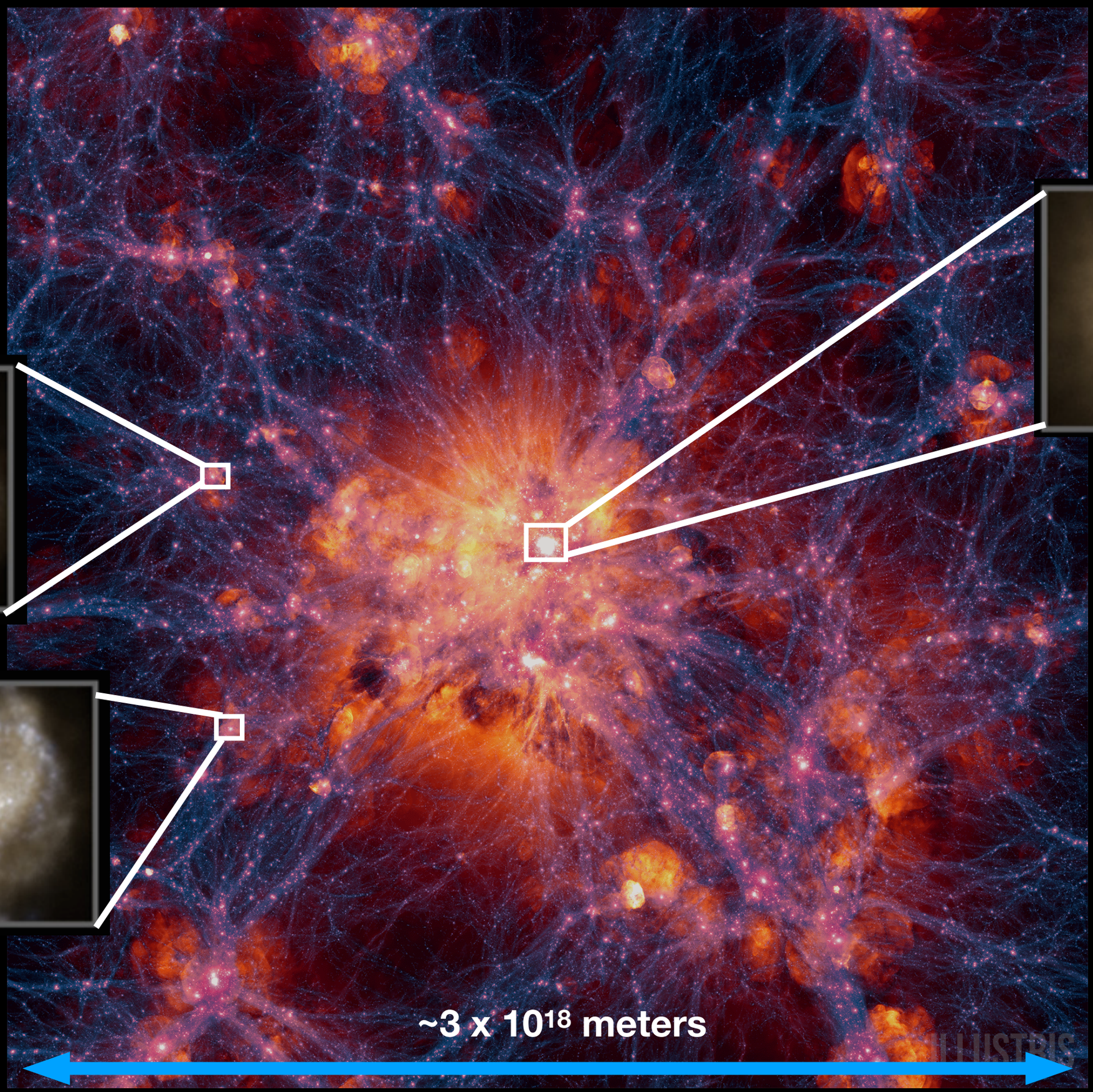


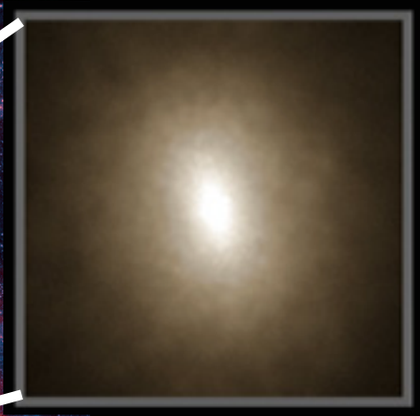
How can we make this vague idea more rigorous?
How big is this effect? How can we measure the “bigness”?

My background

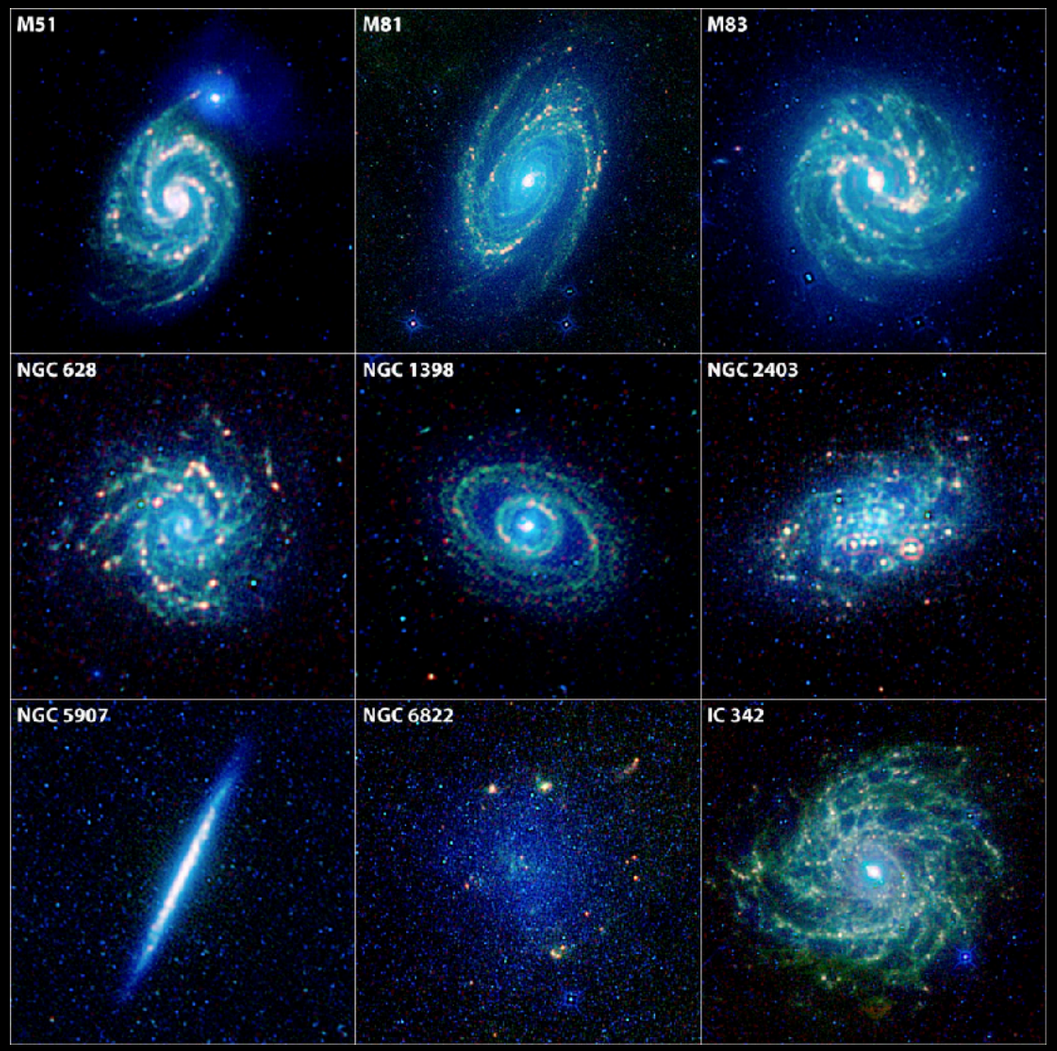








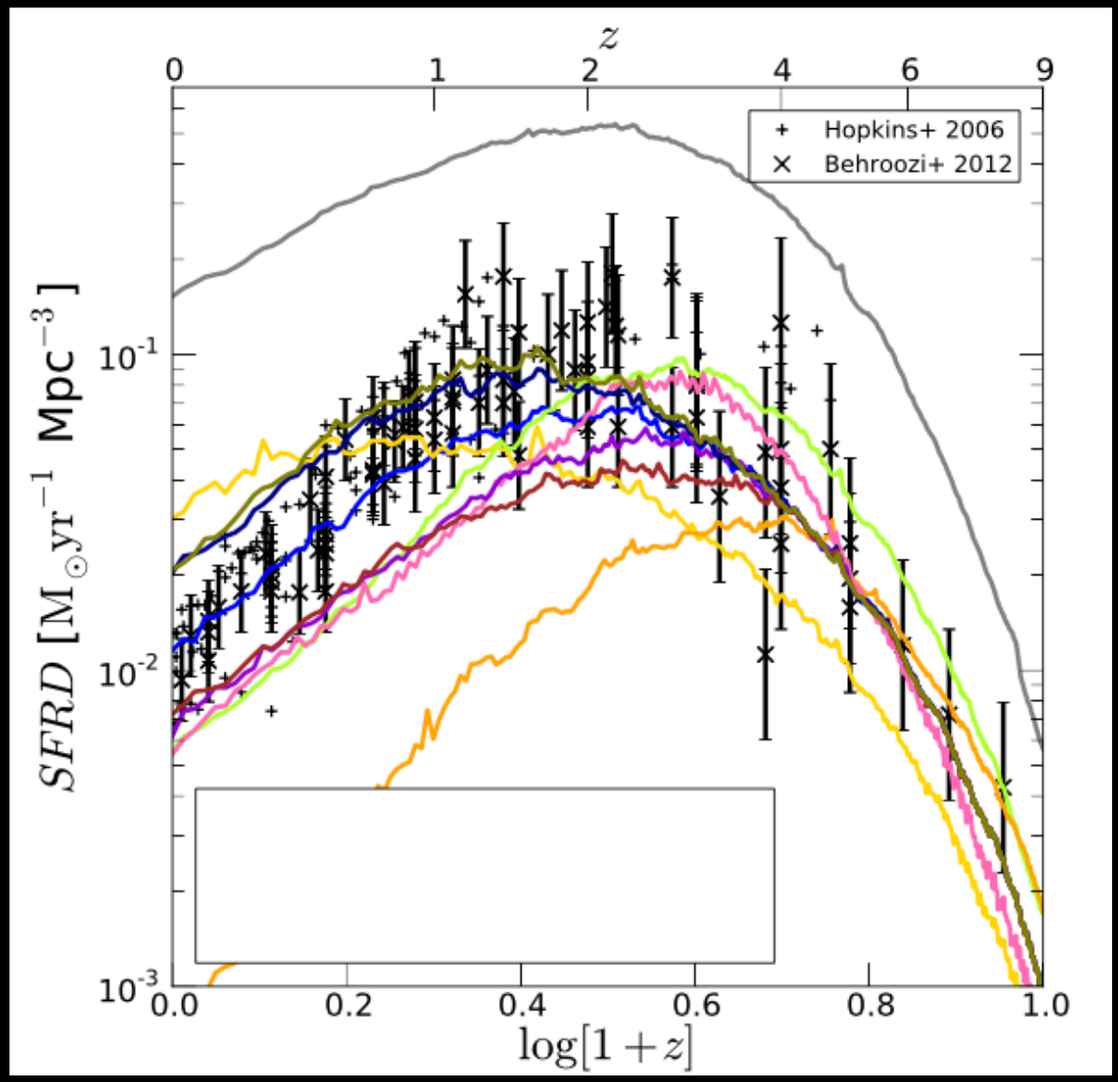
Images of real galaxies from NASA's Wide-field Infrared Survey Explorer



Images of "fake" galaxies simulated in a super computer

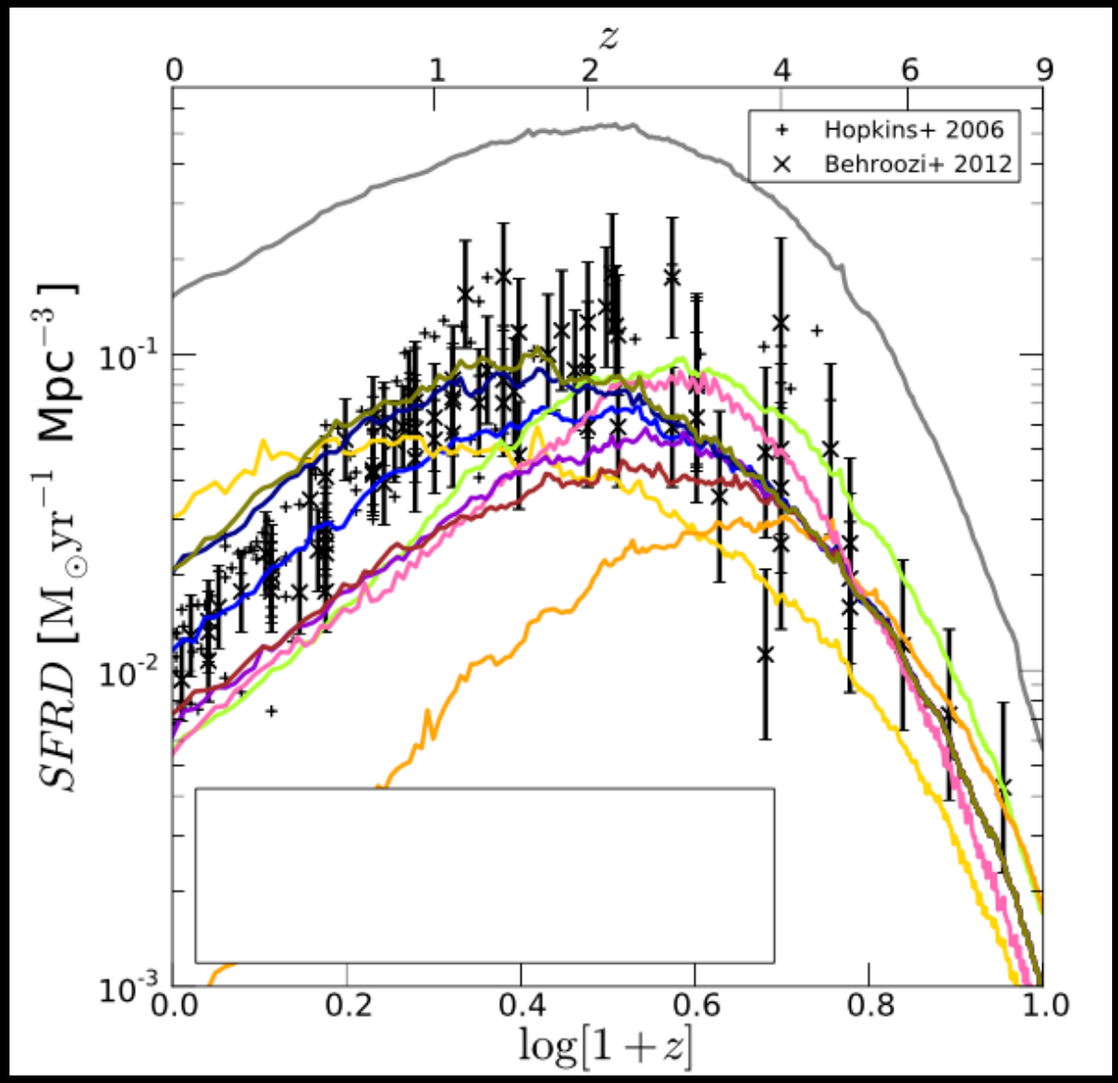
(differences in colors due to color schemes chosen for each image)

Example: Fitting Models to Data



Vogelsberger et al. 2013

Fitting Models to Data



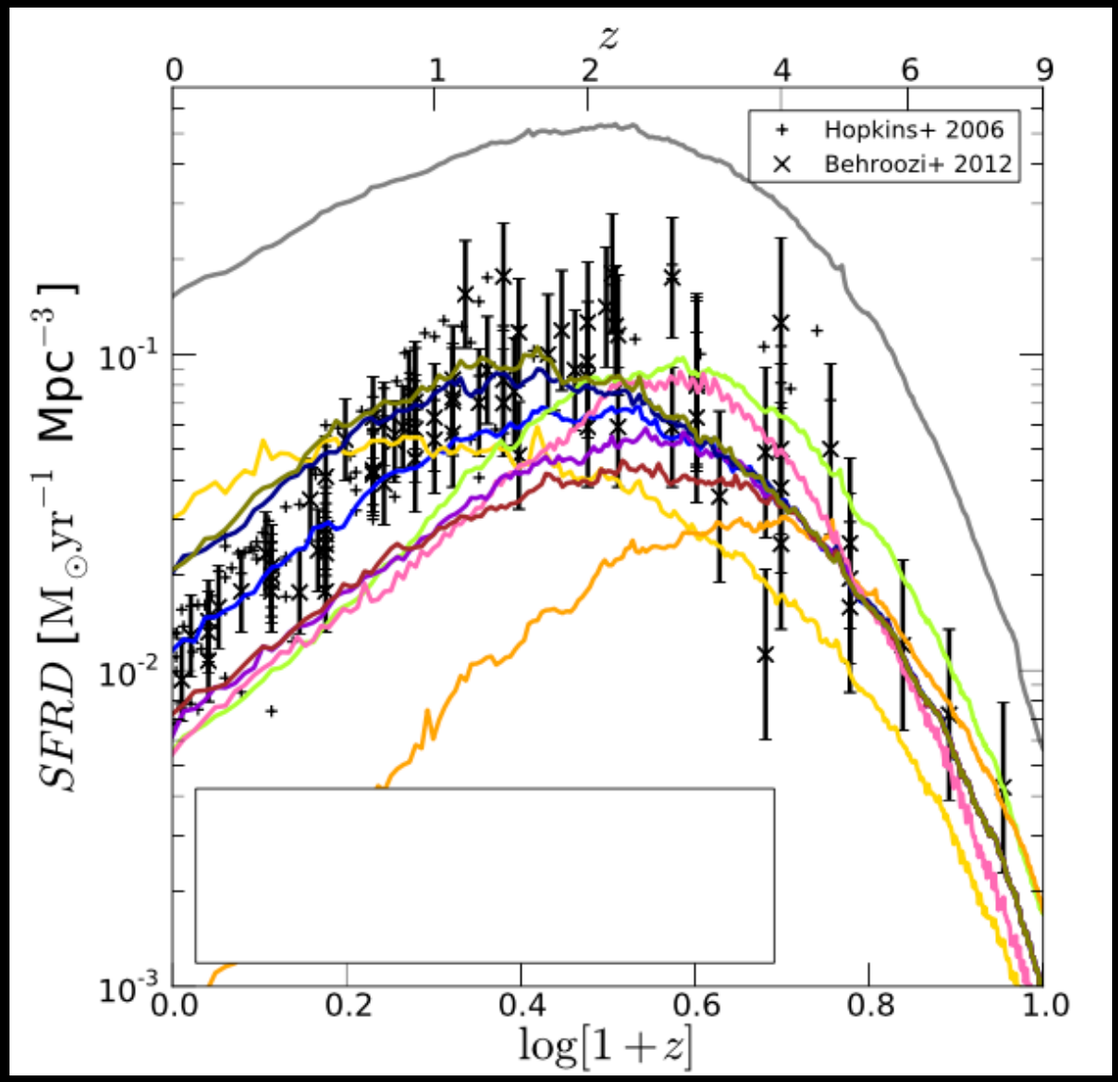
Time Since Universe Started (present time at 0.0)

(Redshift)

Vogelsberger et al. 2013

Fitting Models to Data

How many stars form
(Cosmic Star Formation Rate Density)

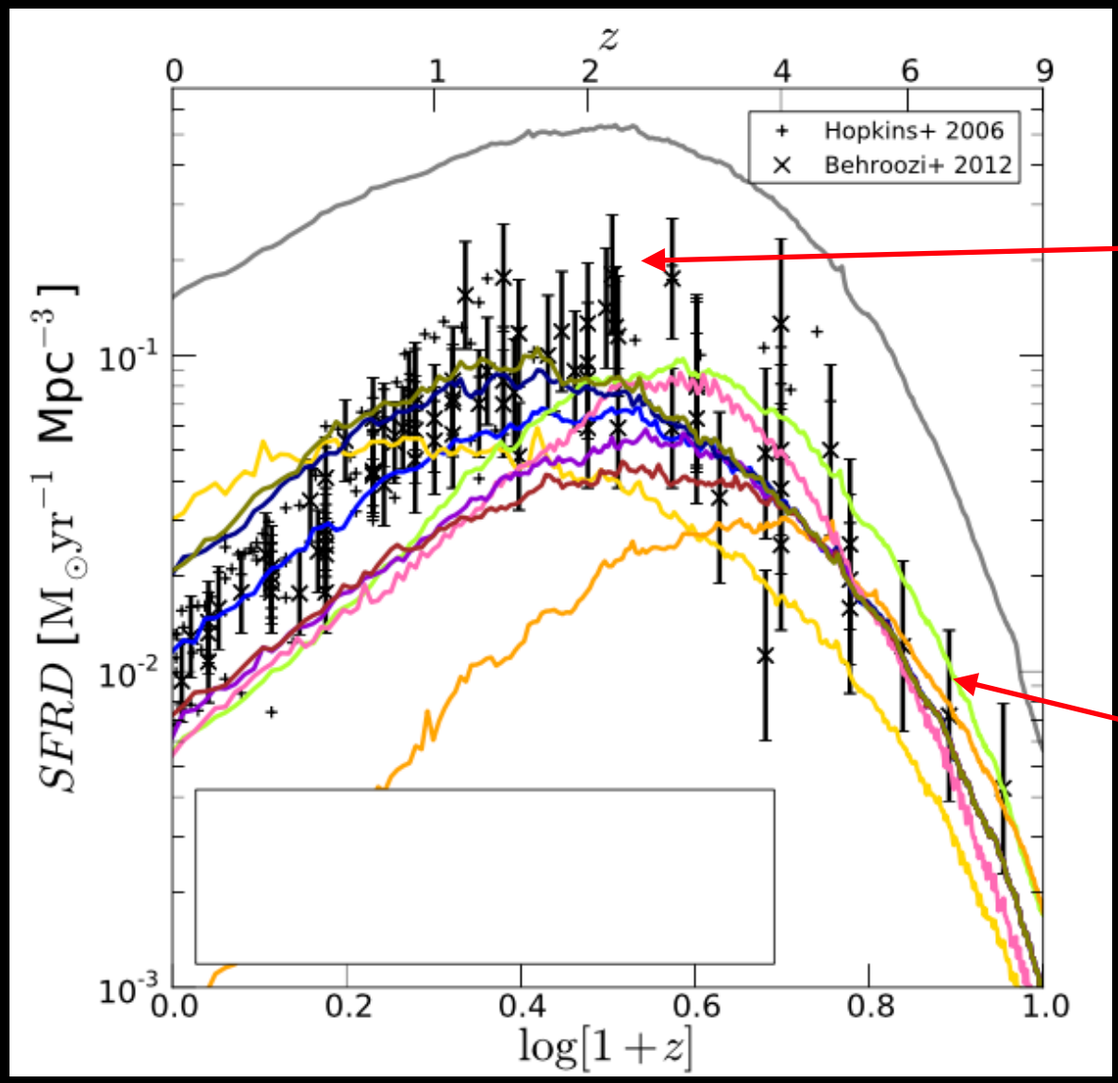


Time Since Universe Started (present time at 0.0)
(Redshift)

Vogelsberger et al. 2013

Fitting Models to Data

How many stars form
(Cosmic Star Formation Rate Density)



Observations

Different simulations

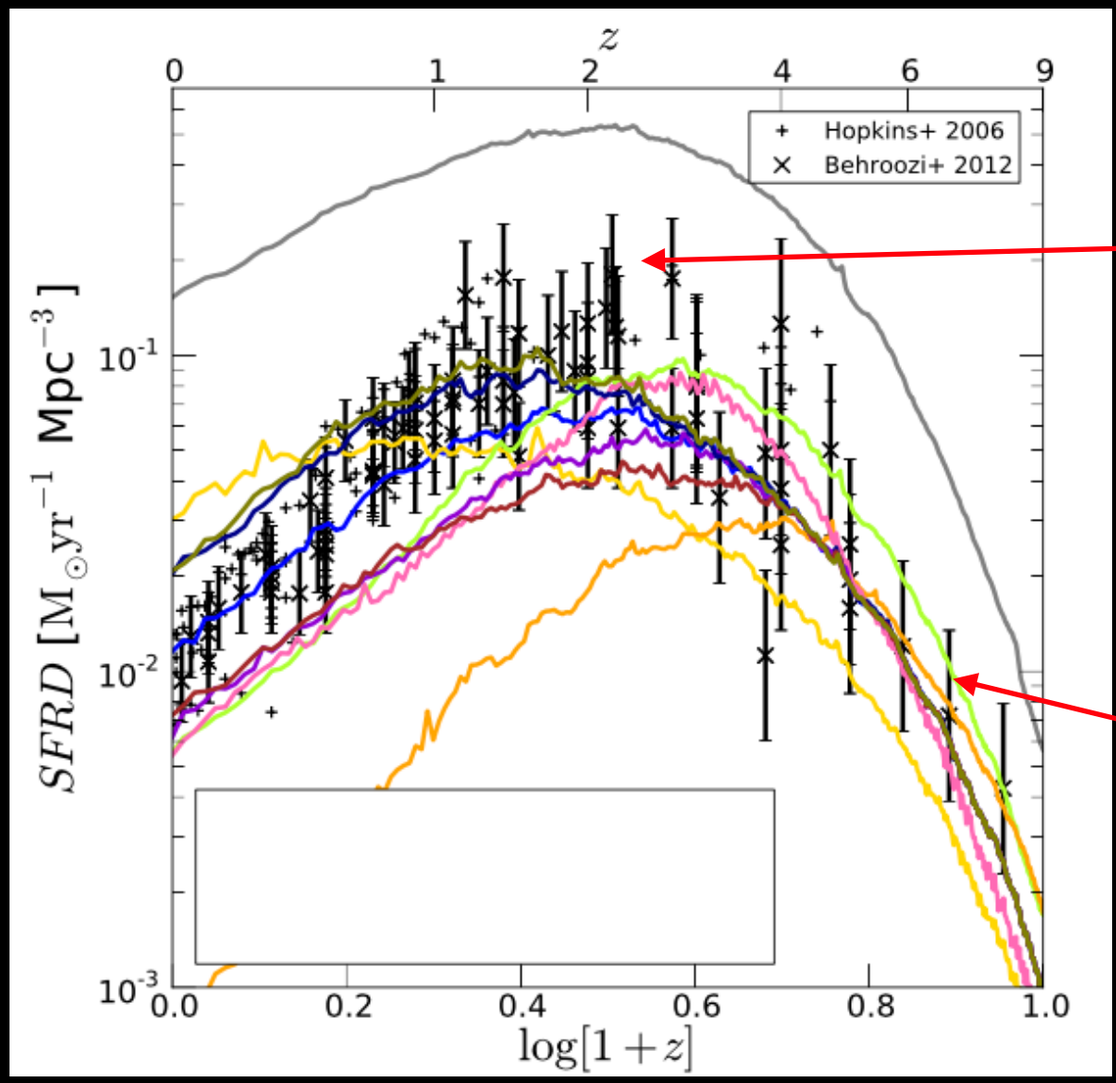
Time Since Universe Started (present time at 0.0)
(Redshift)

Vogelsberger et al. 2013

Fitting Models to Data

What simulation best reproduces (fits) the observed data?
How "good" is the fit to the data for our best-fit model?

How many stars form
(Cosmic Star Formation Rate Density)



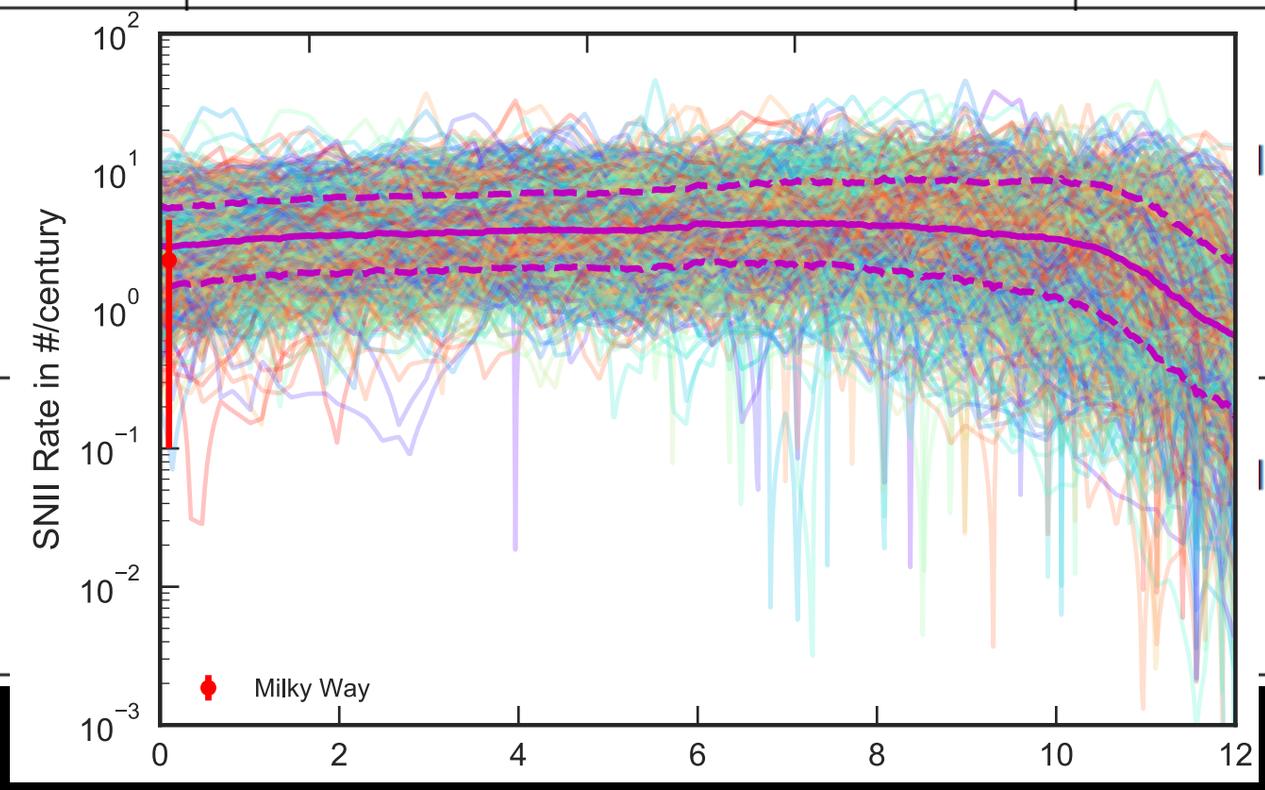
Observations

Different simulations

Time Since Universe Started (present time at 0.0)
(Redshift)

Vogelsberger et al. 2013

Week	Topic	Reading
1	• Data, Models, and Information • Elementary statistics: Definitions • Overview of R	OIS 1 (ISL 1)
2	• Elementary statistics: Applications & Plots	OIS 1 (ISL 1)
3	• Introduction to data analysis with R • Review of tabular and graphical displays of data	ITR 1, 2, 5, 6, 7, 12
		IS 2
		IS 3



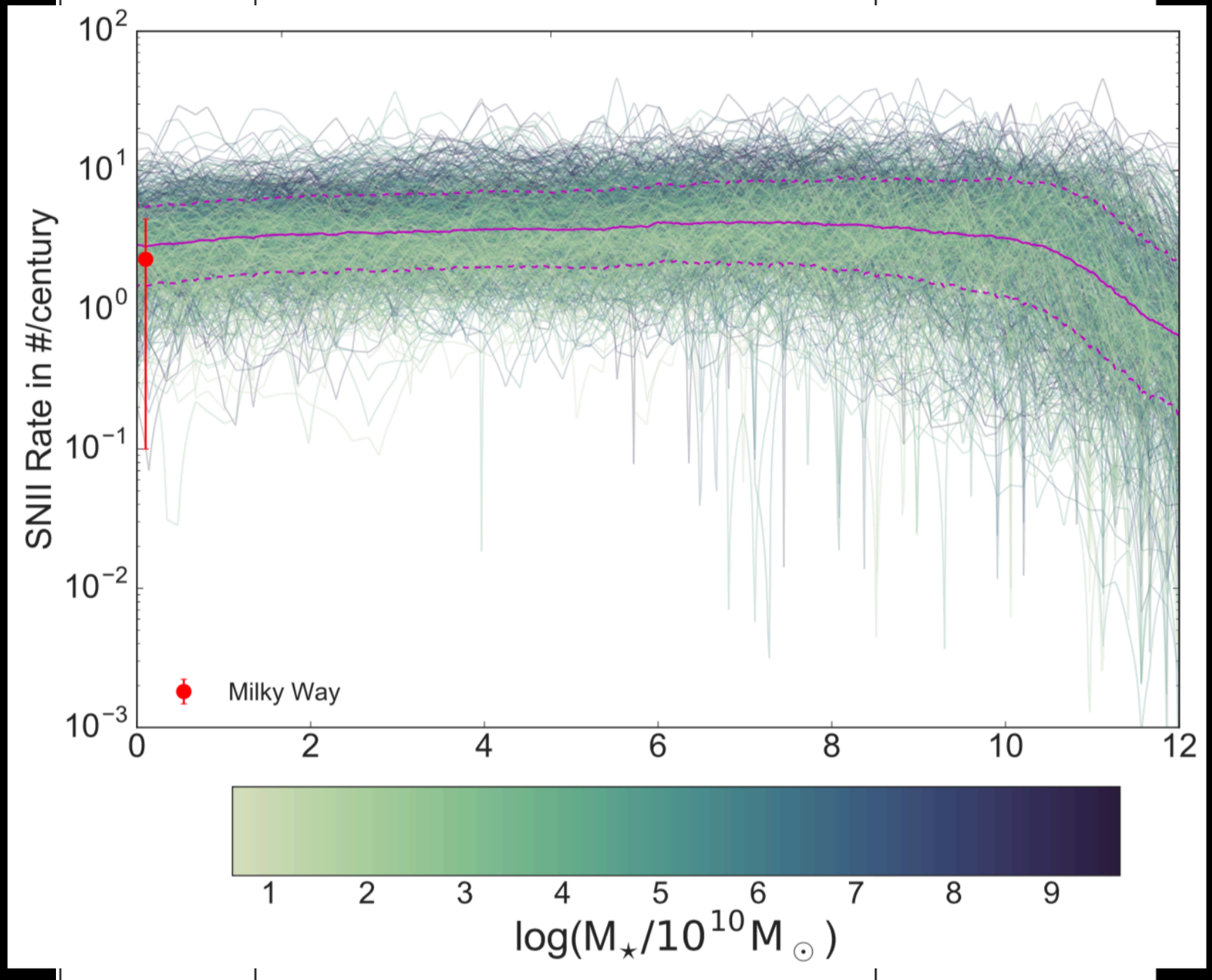
Definitions, basic concepts, R practice

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

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

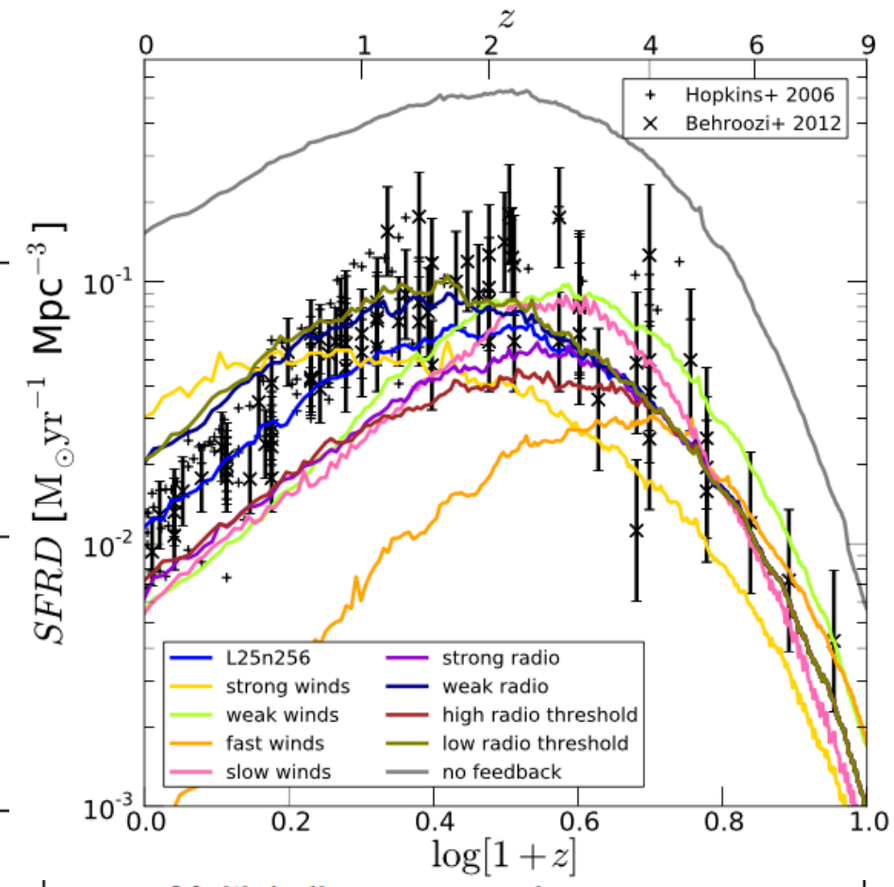
6	• Modeling data with probability distributions • Foundations for inference	OIS 4
7	• Inference for numerical data • Inference for categorical data	OIS 5, OIS 6

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		OIS 4
7		OIS 5, OIS 6
8		OIS 7 (ISL 3)
9	• Multiple linear regression	OIS 8 (ISL 3)
10	• Logical regression	OIS 8 (ISL 4)
12	• *k*-Nearest neighbor classification and regression	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

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?

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 #1 - if you were to summarize your datasets with 1 or 2 numbers (or something else…?) how would you do it?

Quick activity!

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

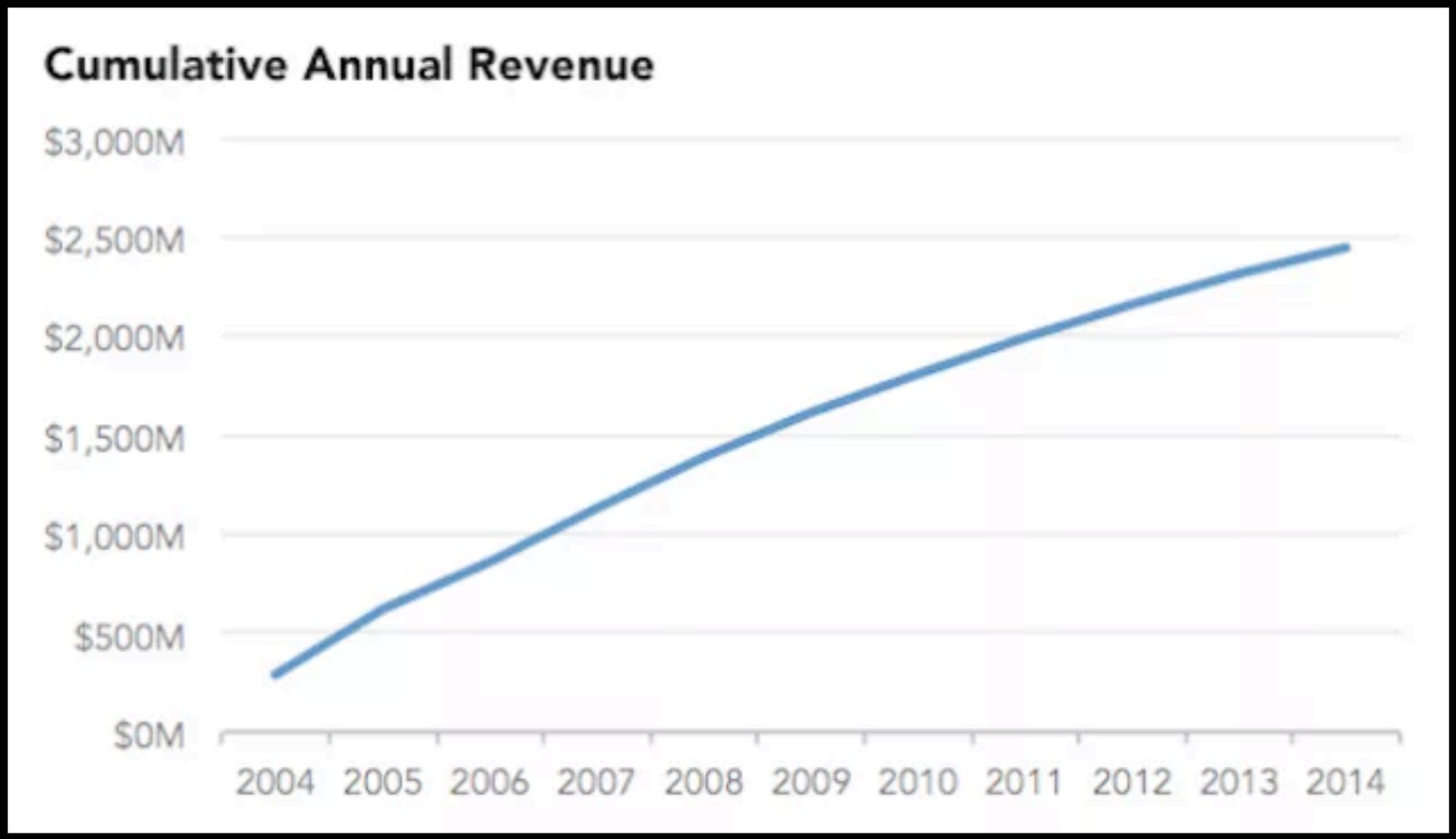
- What are the most memorable movies you saw over the last year?
- Do you prefer cats or dogs?
- How would you quantify your experience in statistics?
- How many chairs or tables are there in your room?

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

Breakout group #2 - what was easy/hard about summarizing your datasets in this way? How confident are you of your summaries?

A quick aside: Lying… with Data!

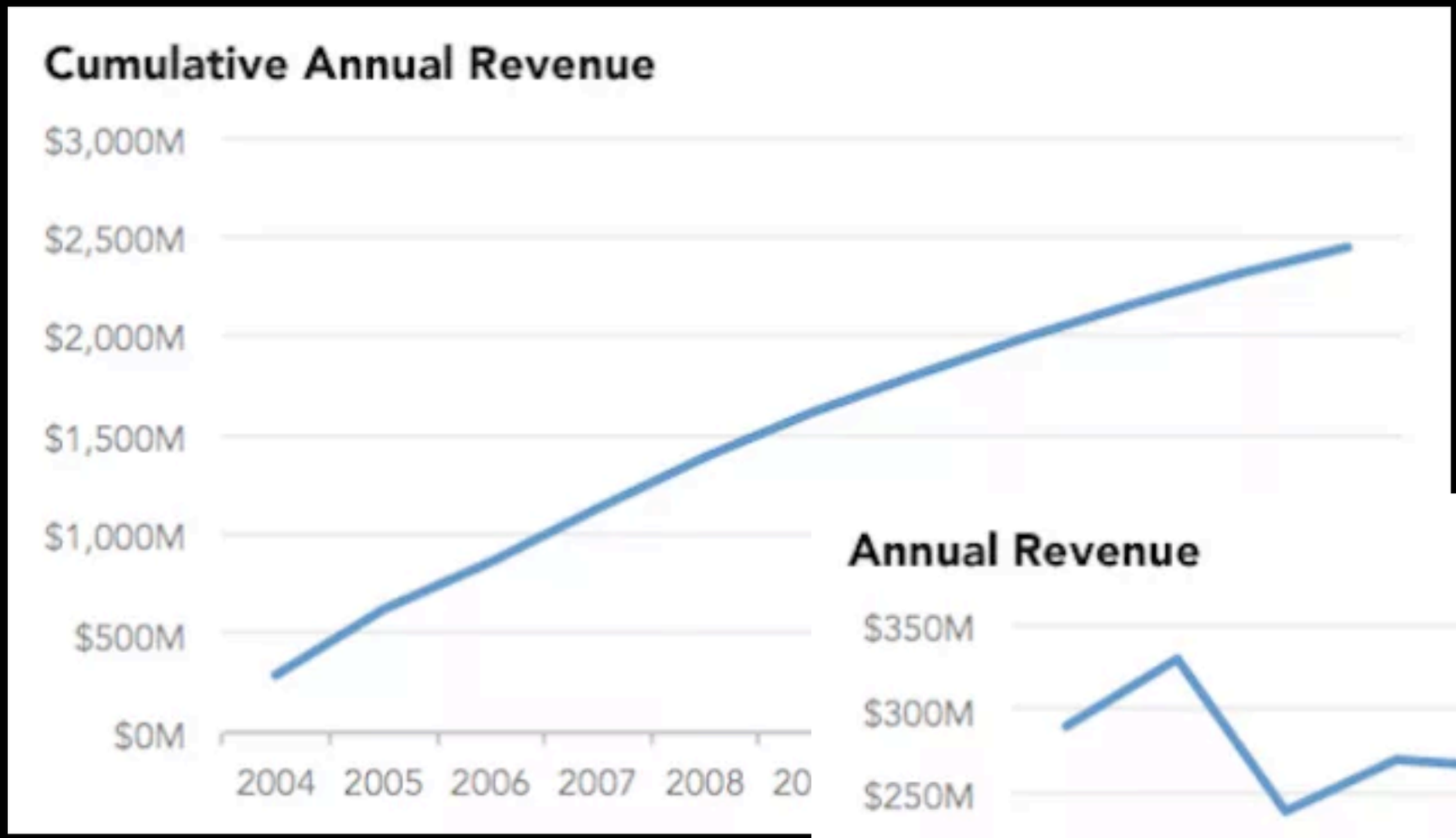
How about if I ask about how profitable your company?

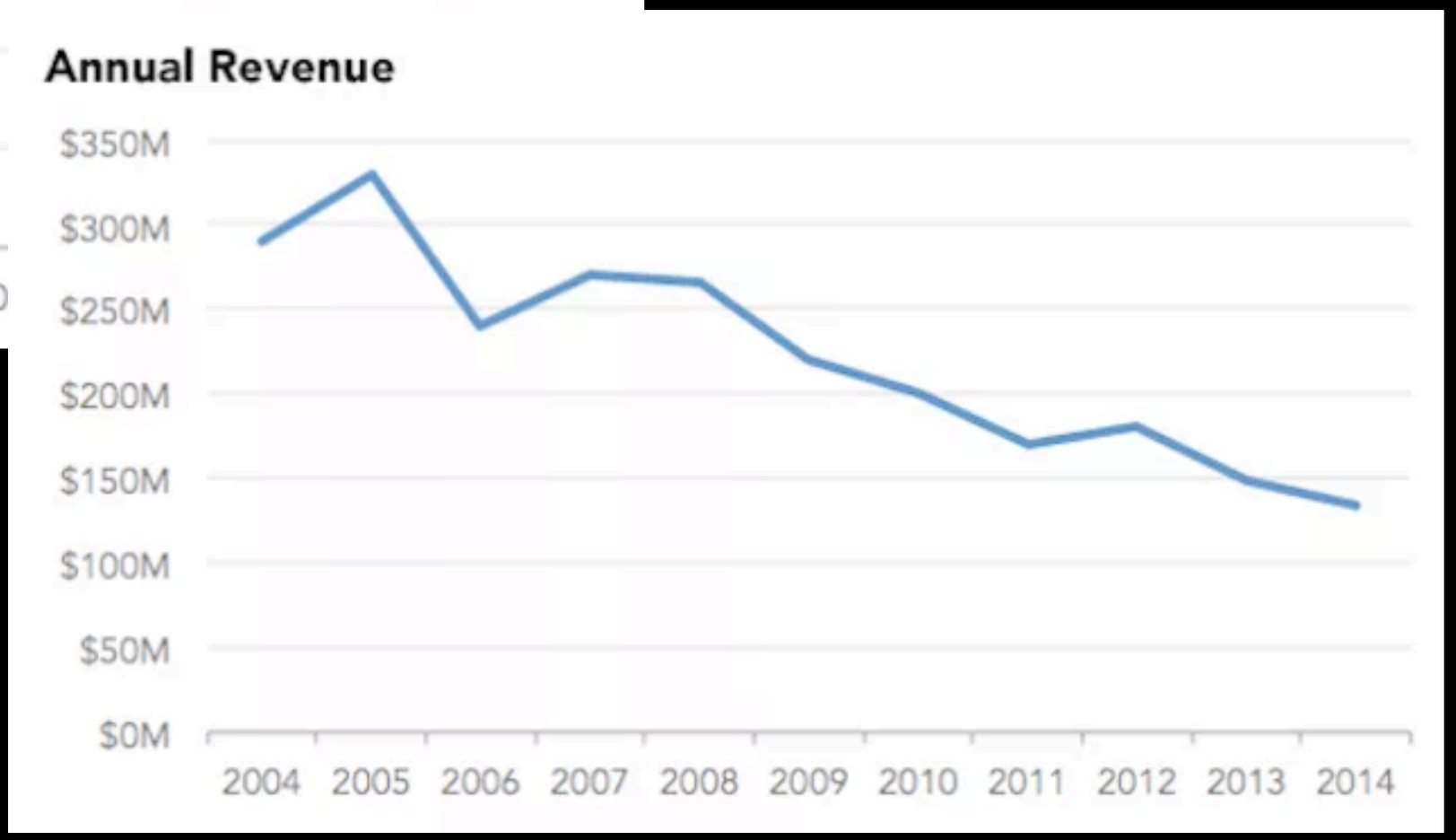


https://gizmodo.com/how-to-lie-with-data-visualization-1563576606

Lying… with Data!

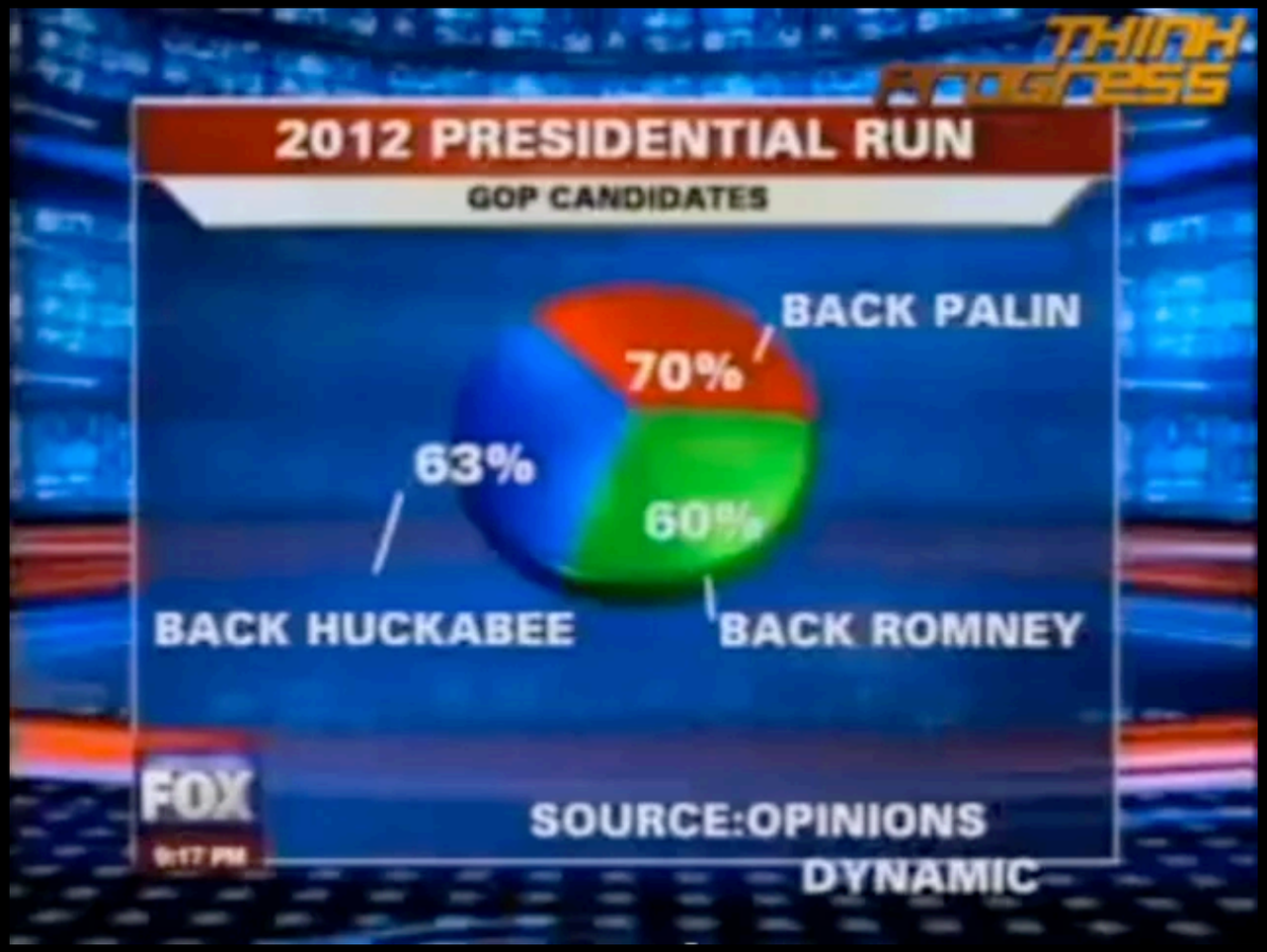
How about if I ask about how profitable your company?





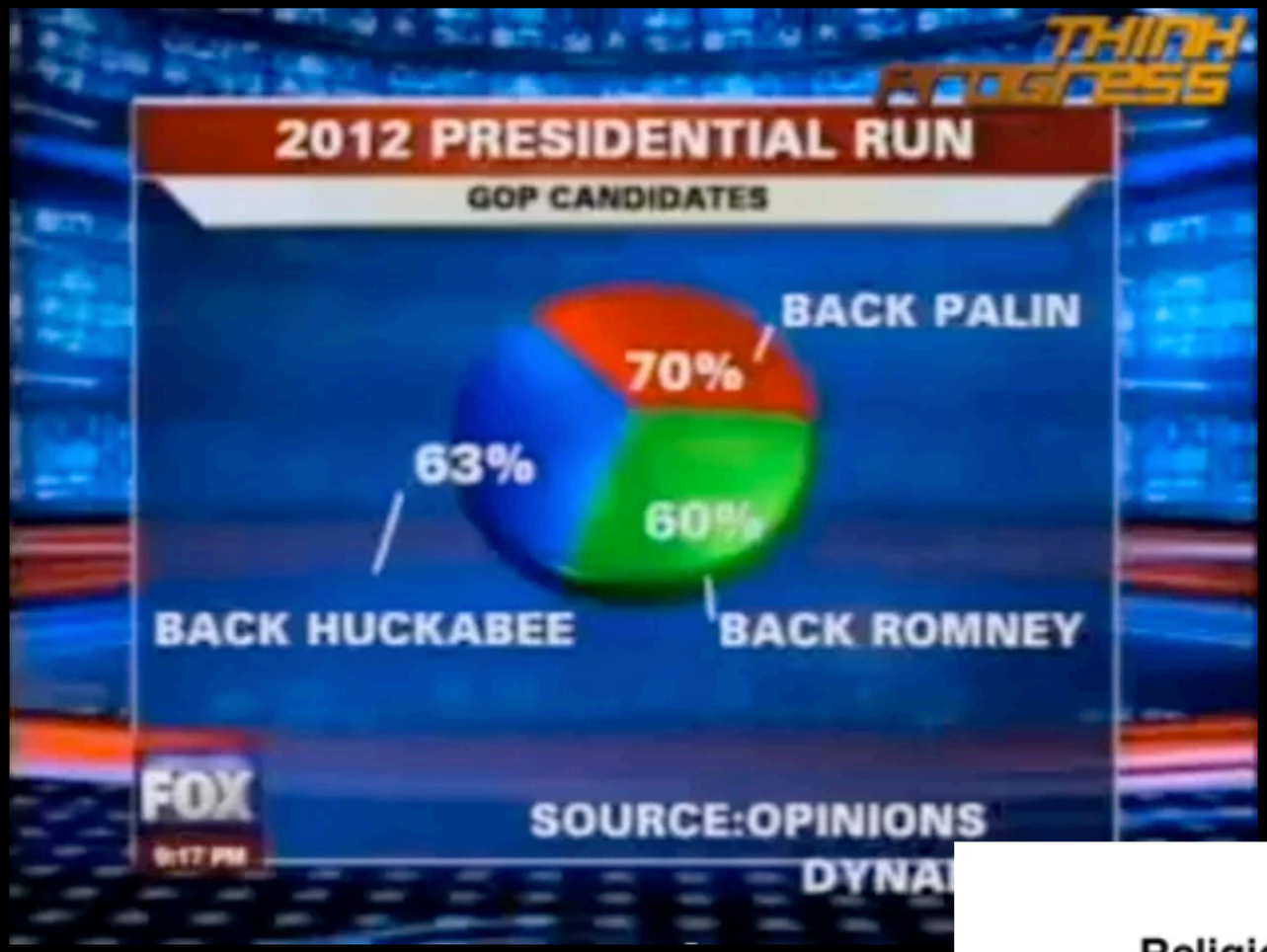
https://gizmodo.com/how-to-lie-with-data-visualization-1563576606

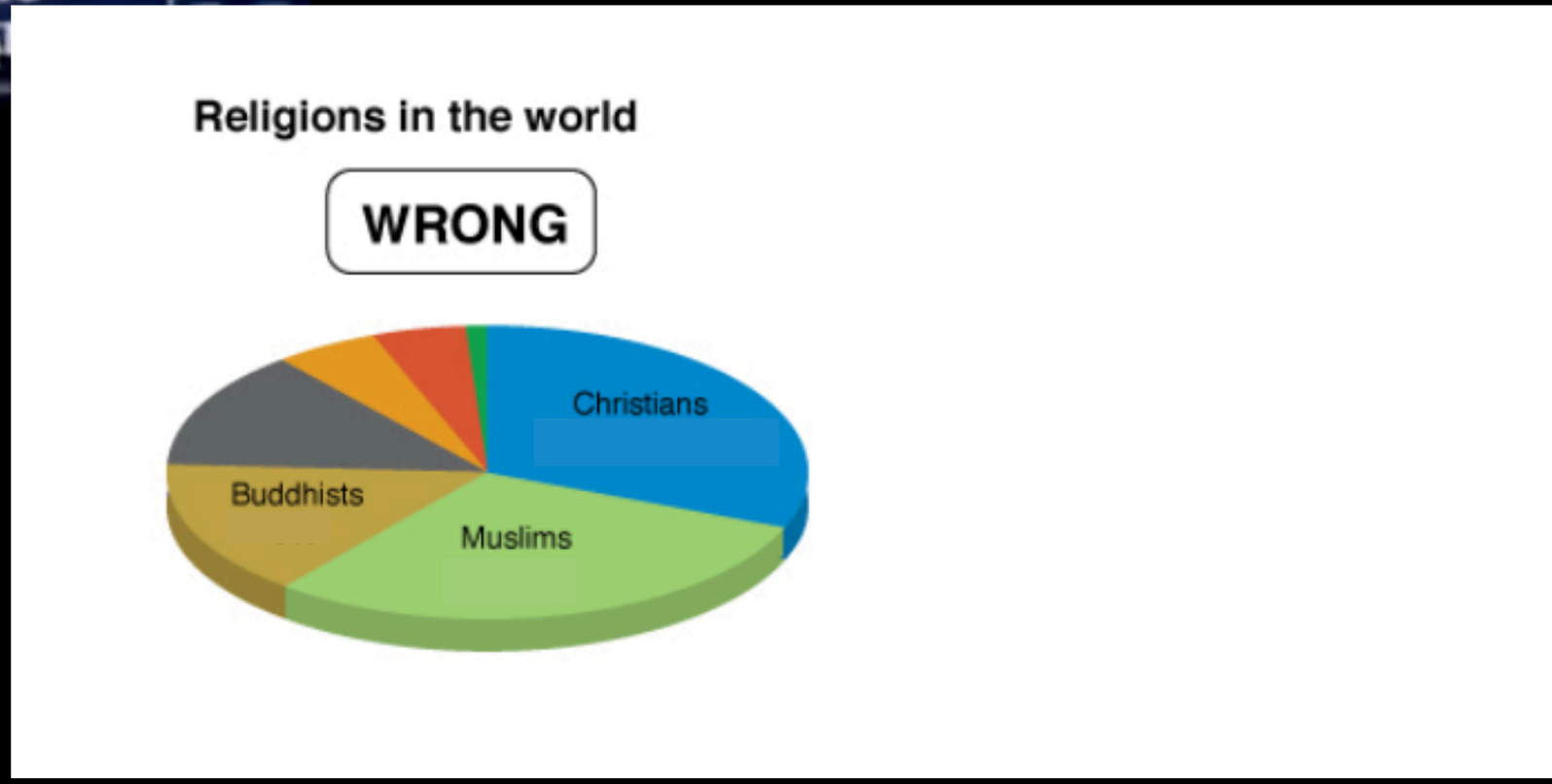
Lying… with Data!



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/

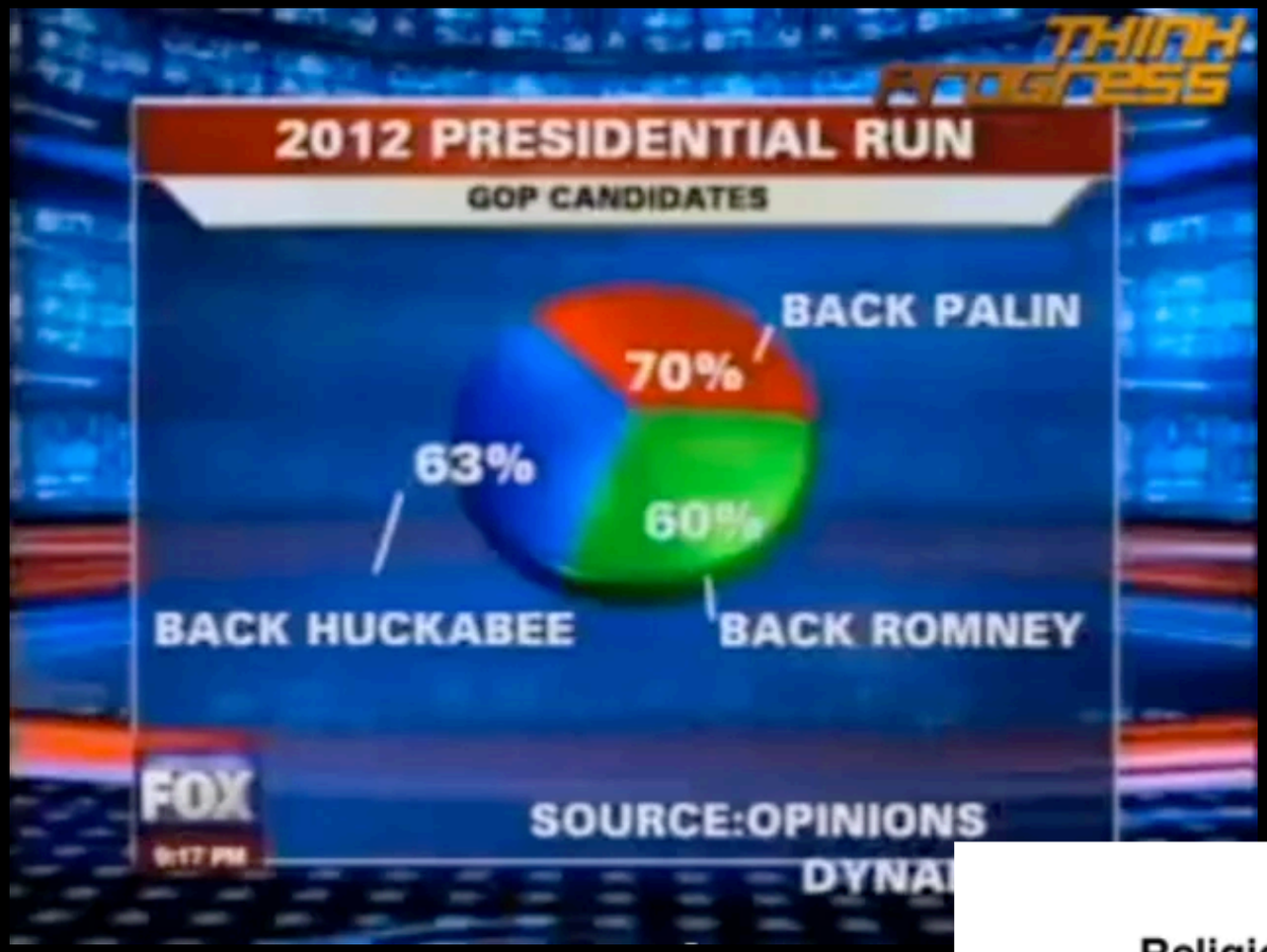
Lying… with Data!

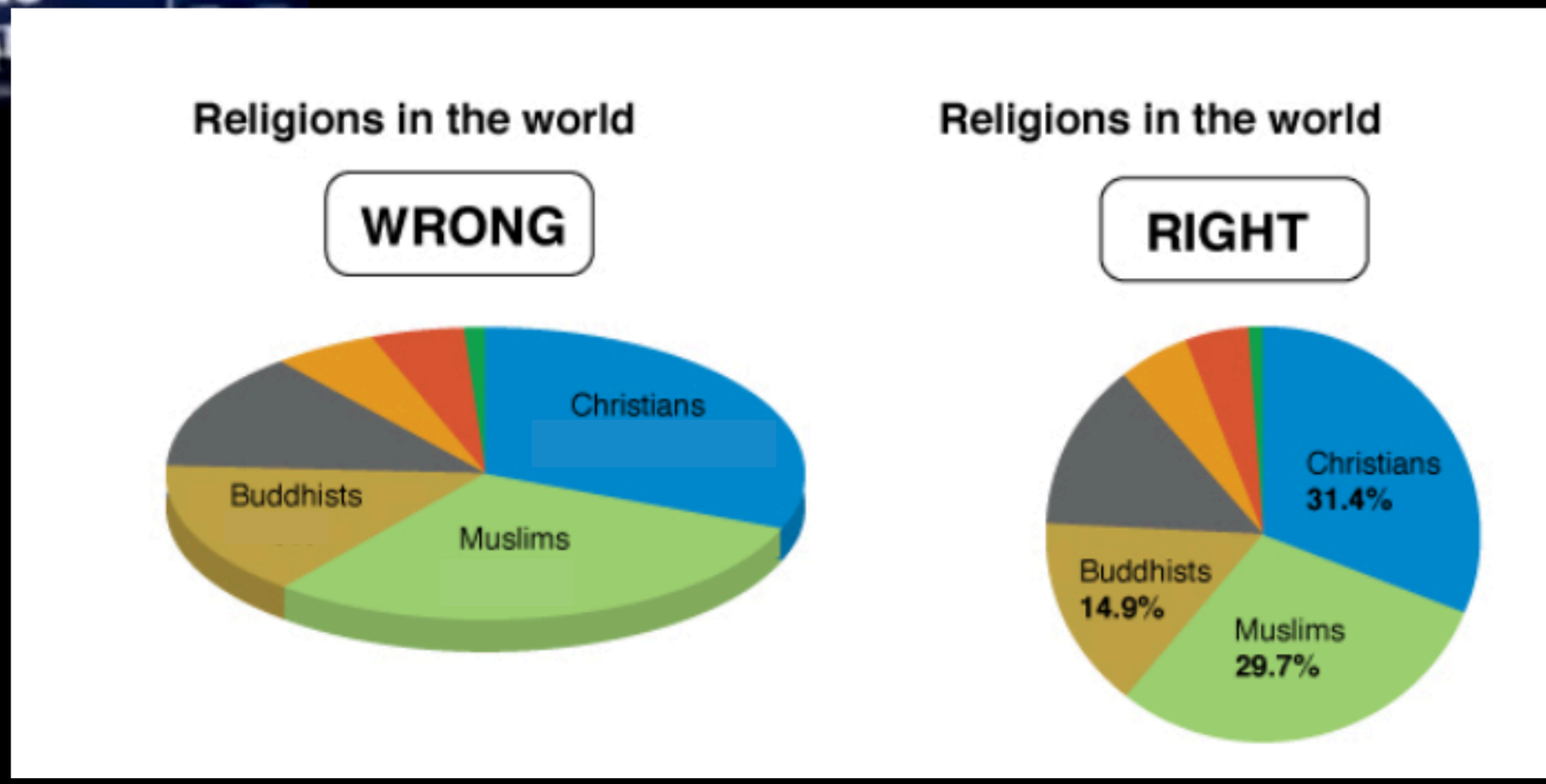




https://gizmodo.com/how-to-lie-with-data-visualization-1563576606
https://news.nationalgeographic.com/2015/06/150619-data-points-five-ways-to-lie-with-charts/

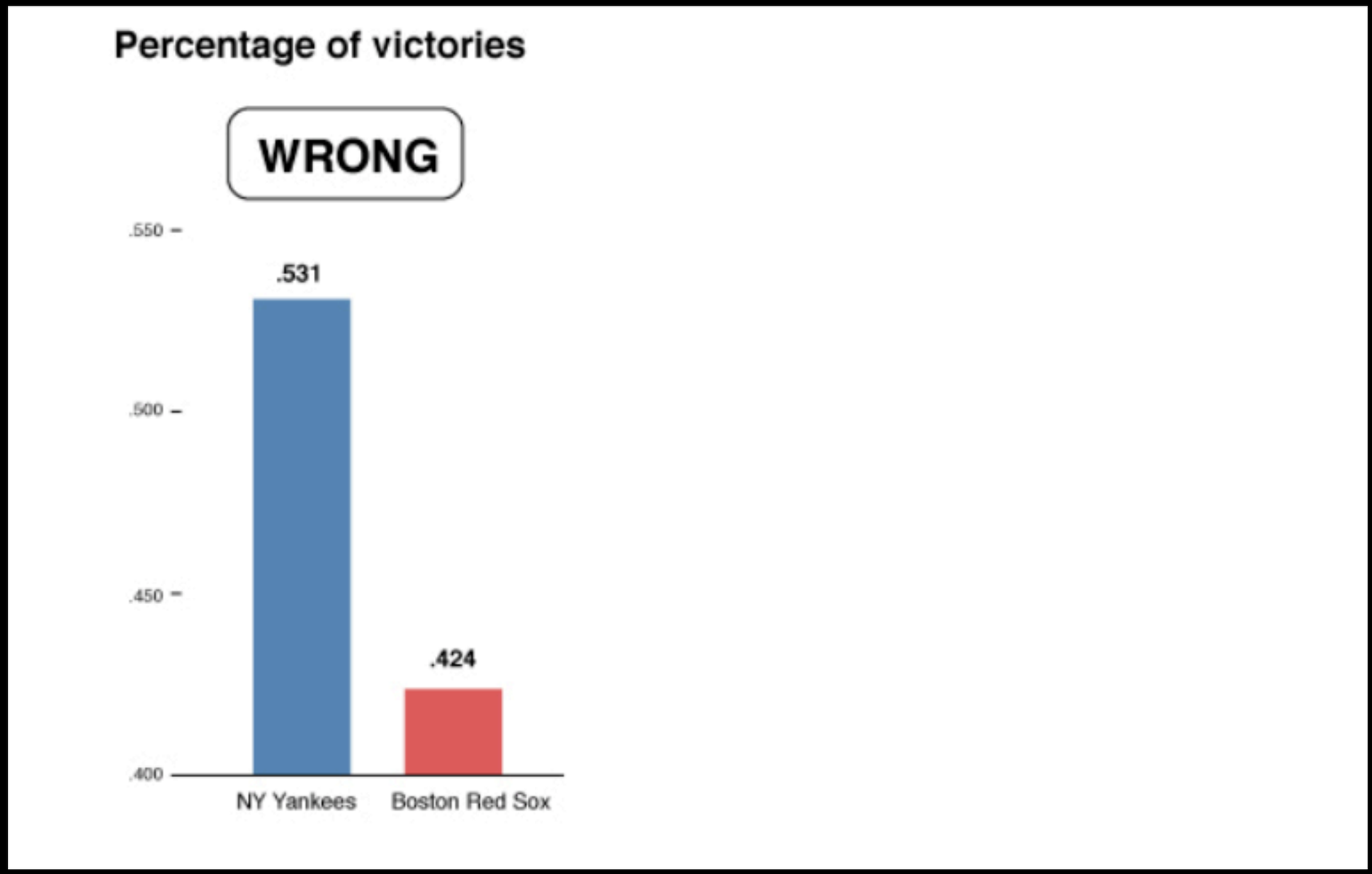
Lying… with Data!





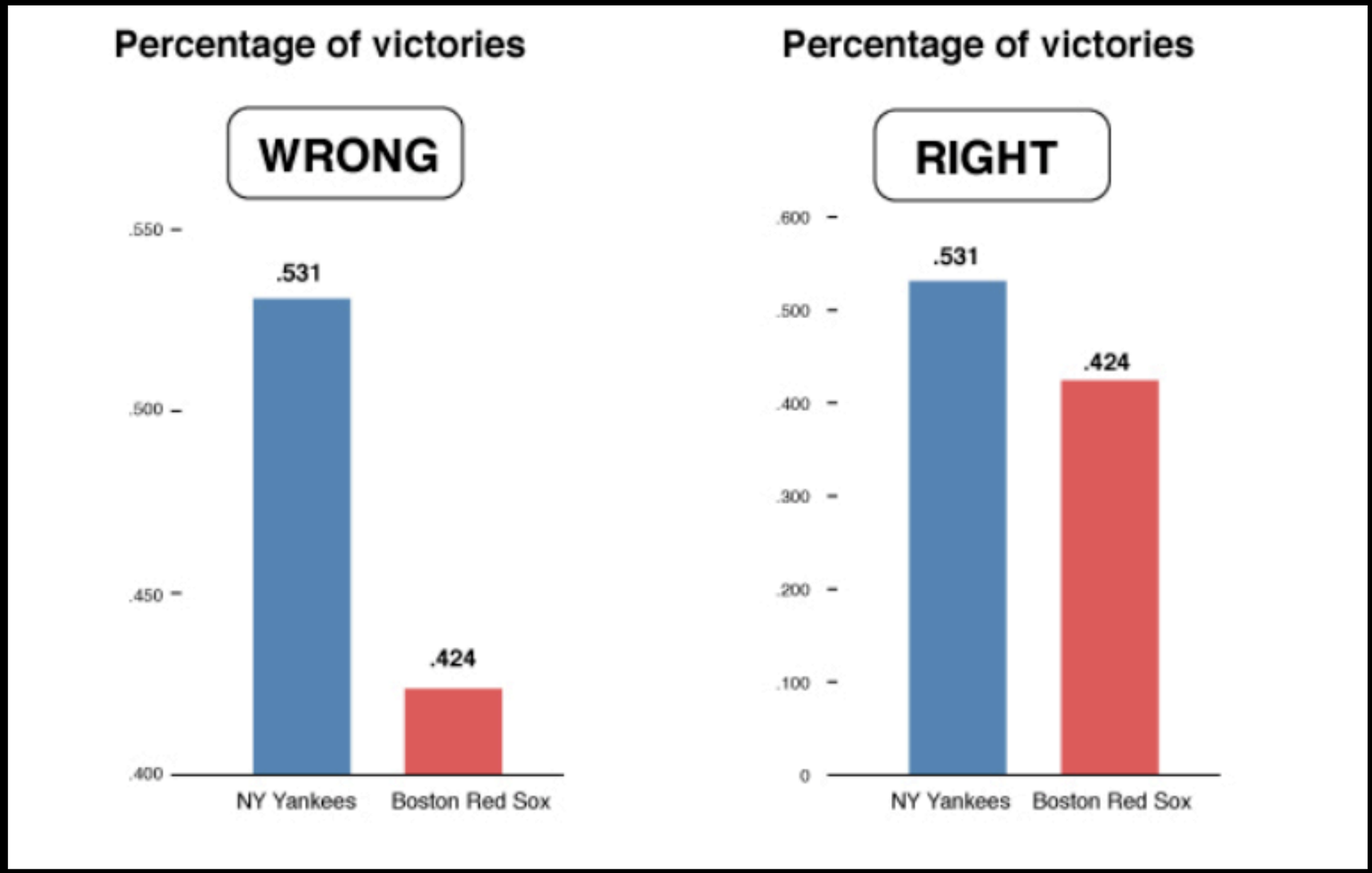
https://gizmodo.com/how-to-lie-with-data-visualization-1563576606
https://news.nationalgeographic.com/2015/06/150619-data-points-five-ways-to-lie-with-charts/

Lying… with Data!



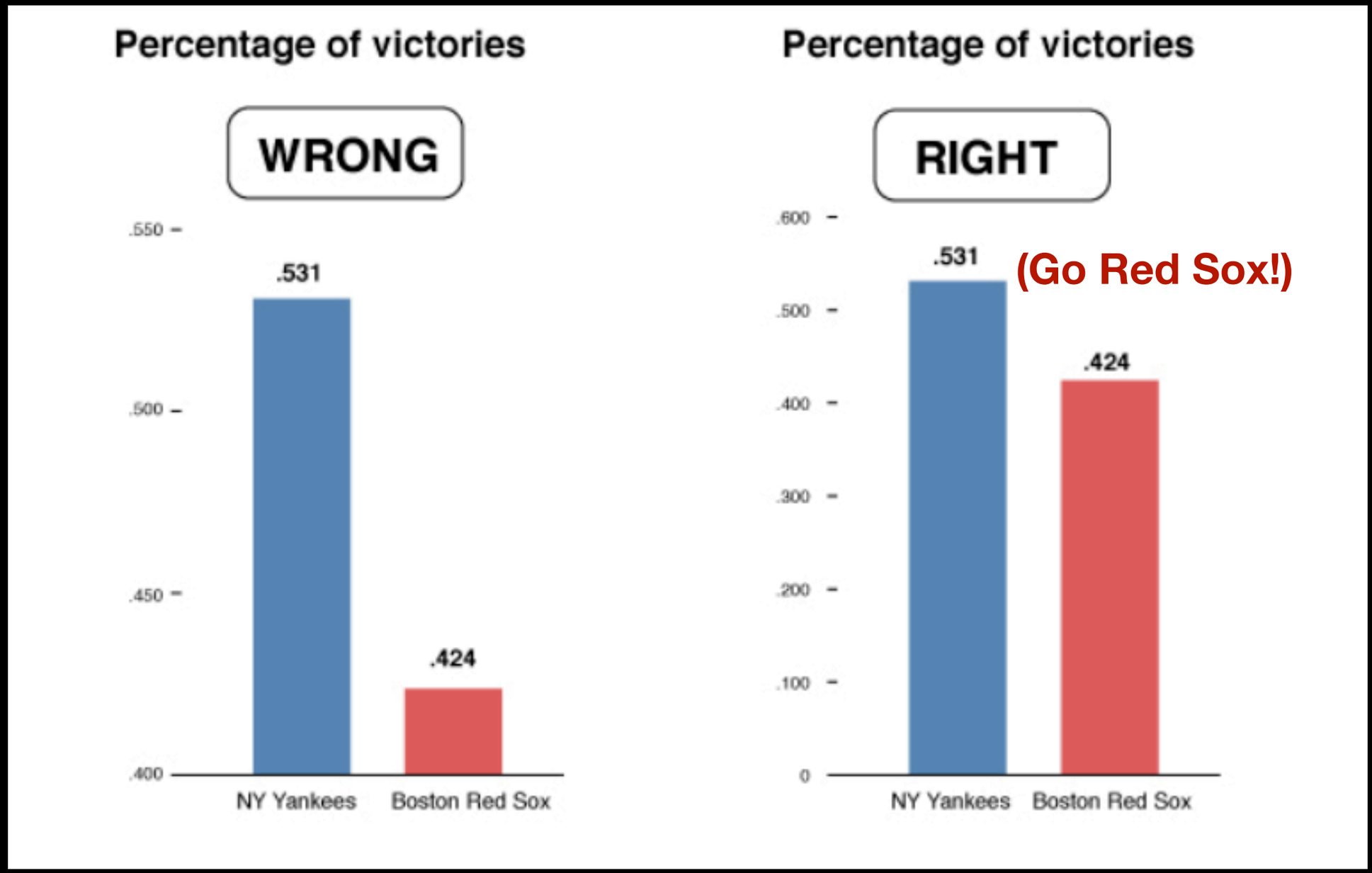
https://news.nationalgeographic.com/2015/06/150619-data-points-five-ways-to-lie-with-charts/

Lying… with Data!



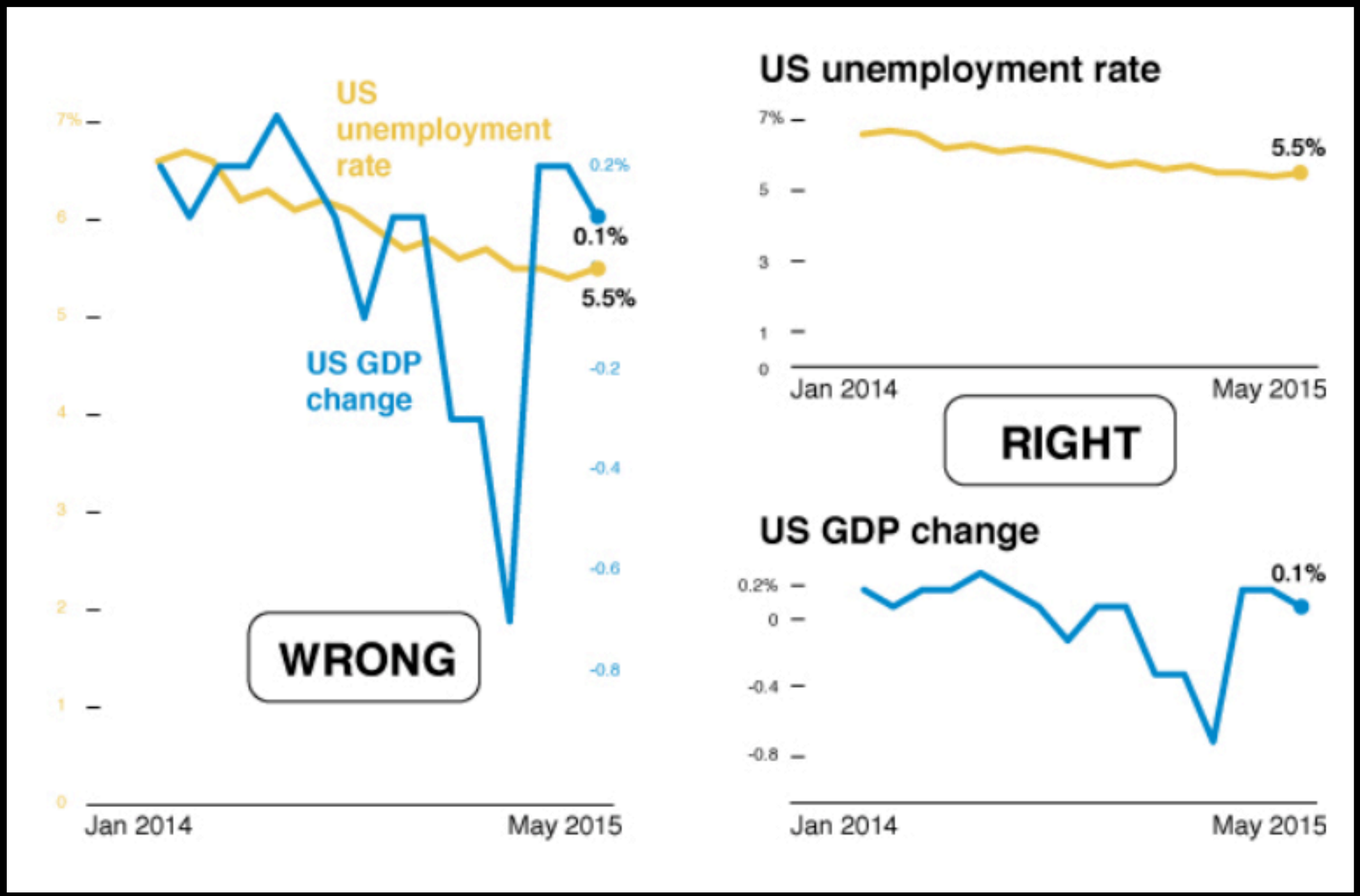
https://news.nationalgeographic.com/2015/06/150619-data-points-five-ways-to-lie-with-charts/

Lying… with Data!



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

Lying… with Data!



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

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

Plots can mislead people!

Proceed with caution.

Plots can mislead people!

Proceed with caution.

Lets get plotting!

… but first, lets all install R!

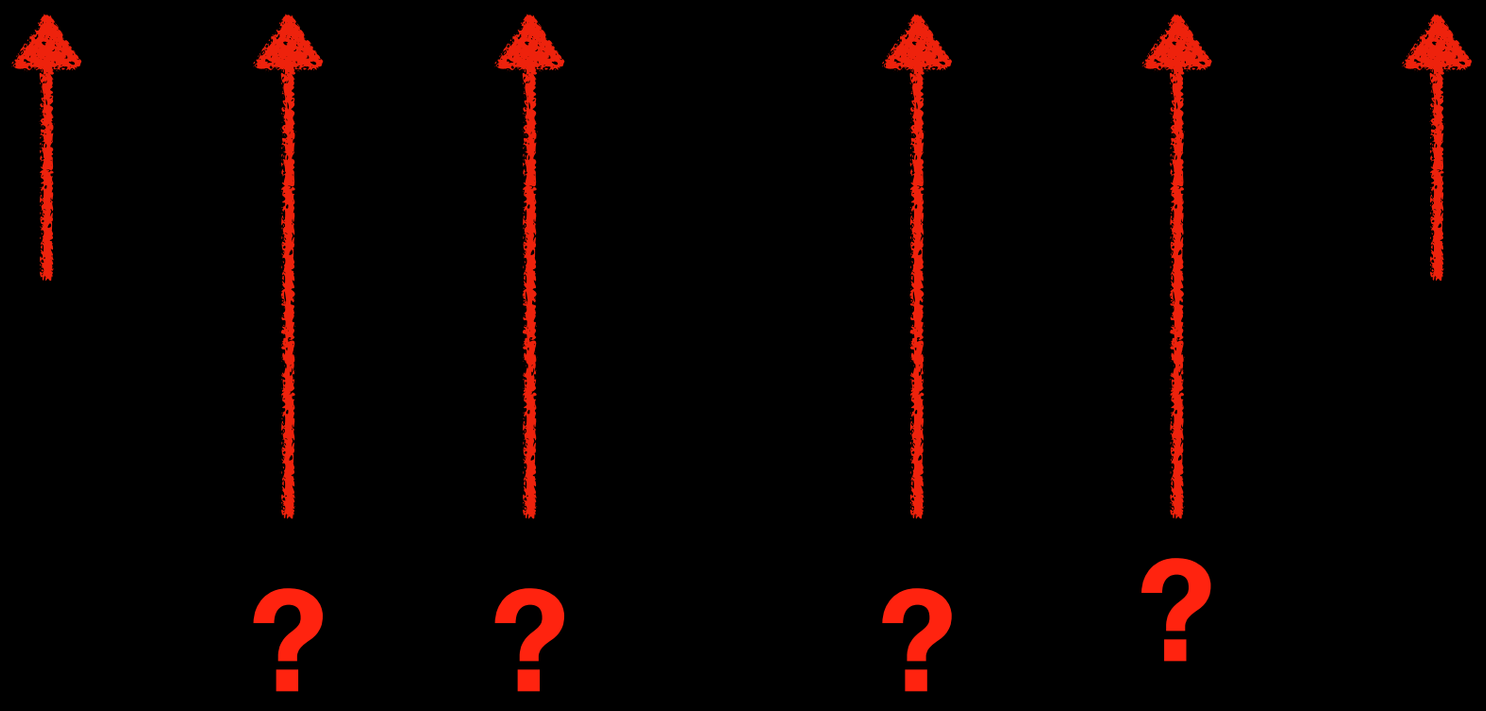
```
summary(after)
   Min.   1st Qu.  Median    Mean   3rd Qu.    Max. 
  39.00    41.25    44.50    44.21    46.75    50.00 
```

summary(after)

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

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!

Summary Statistic Definitions!

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

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)

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)

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

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

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

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

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