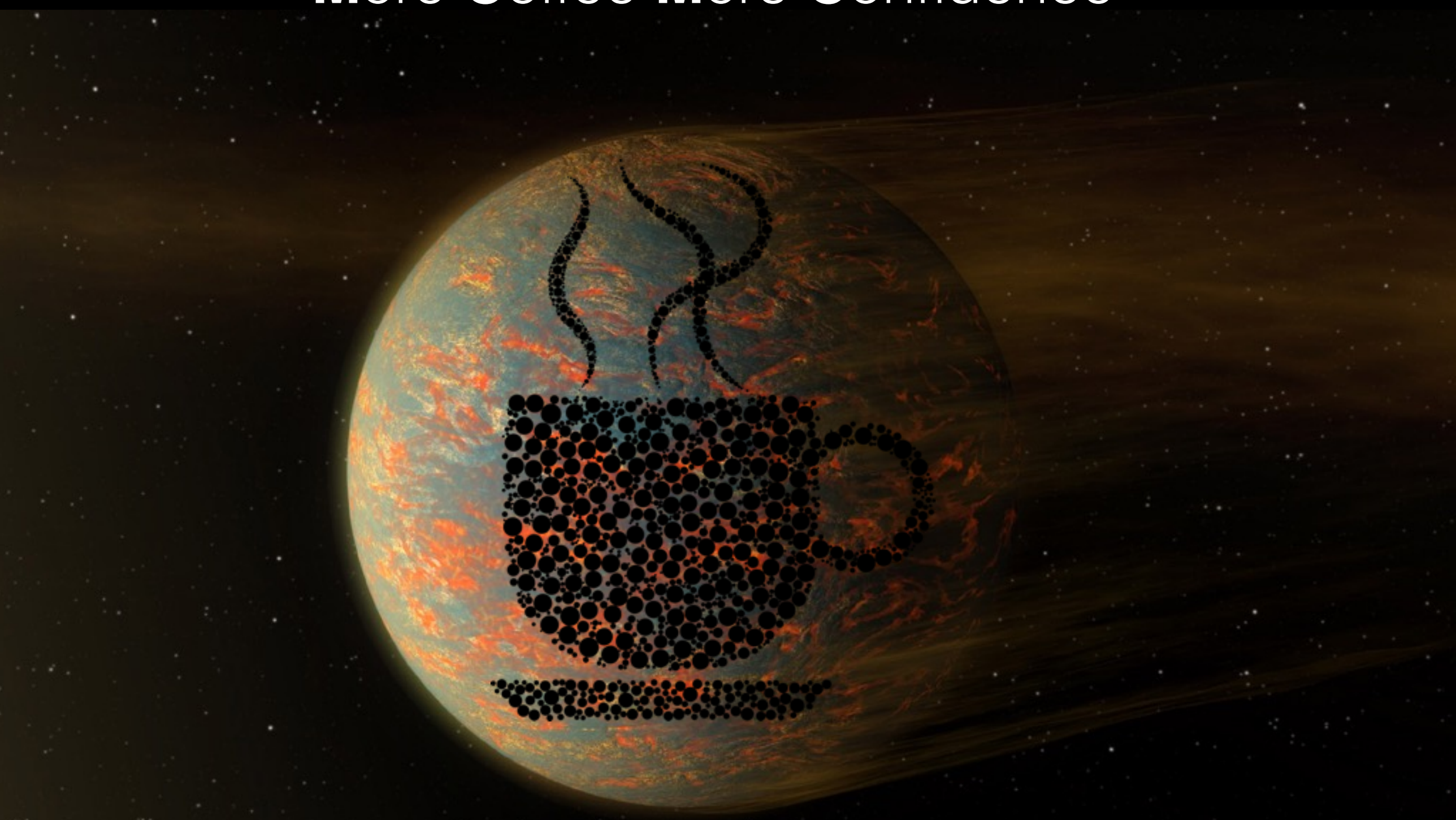
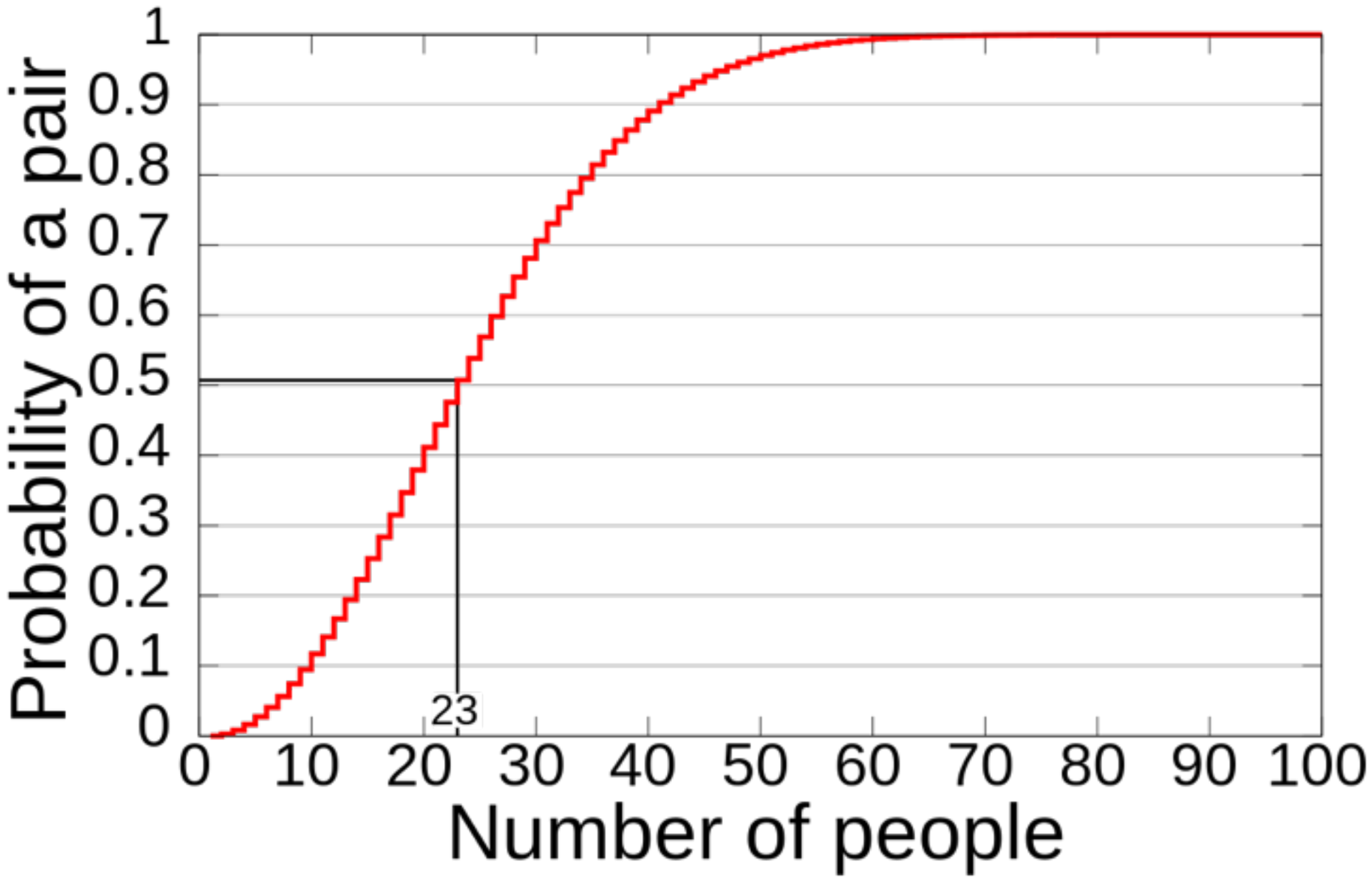
MCMC Coffee

More **C**offee **M**ore **C**onfidence

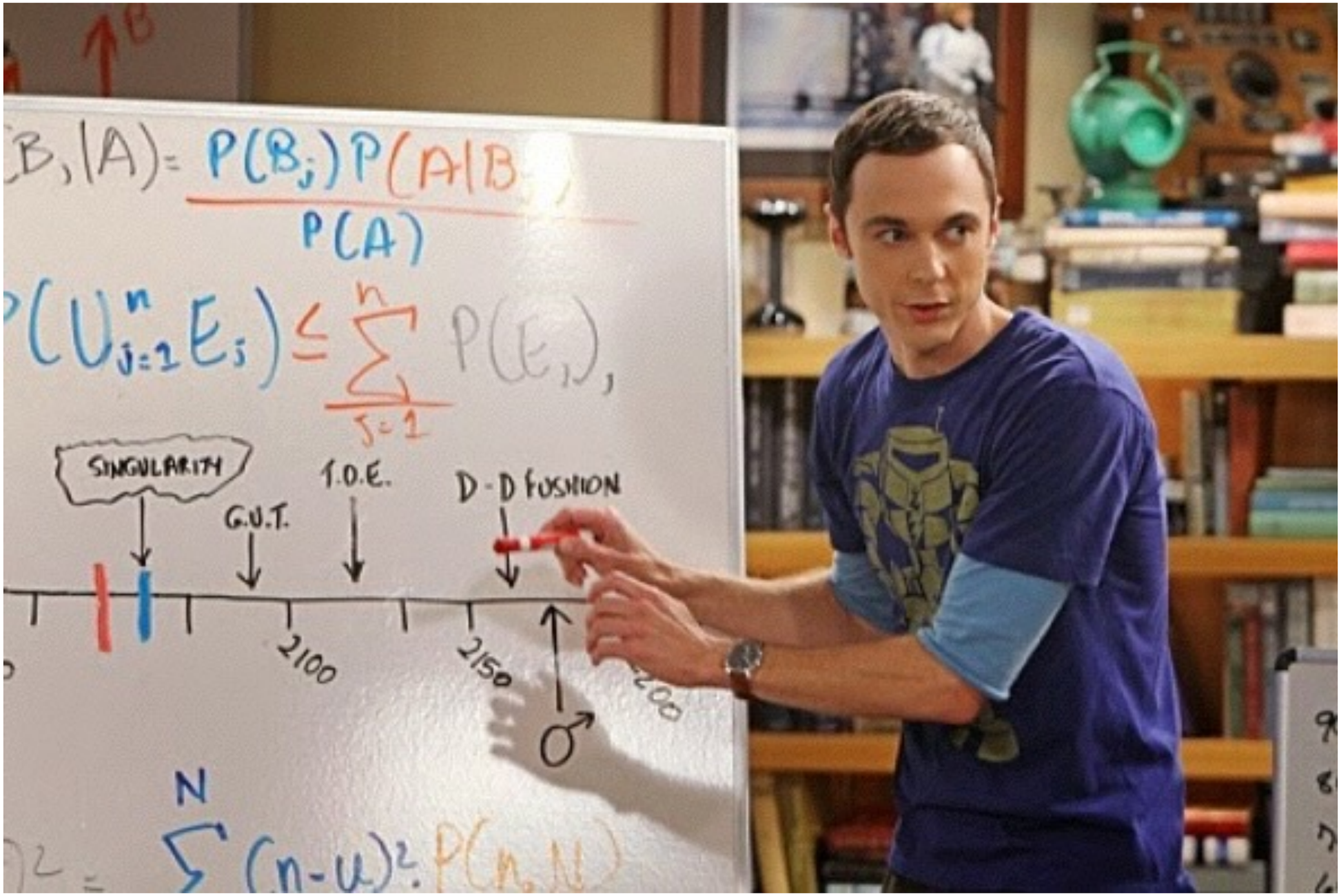


Jorge Lillo-Box & Adele Plunkett

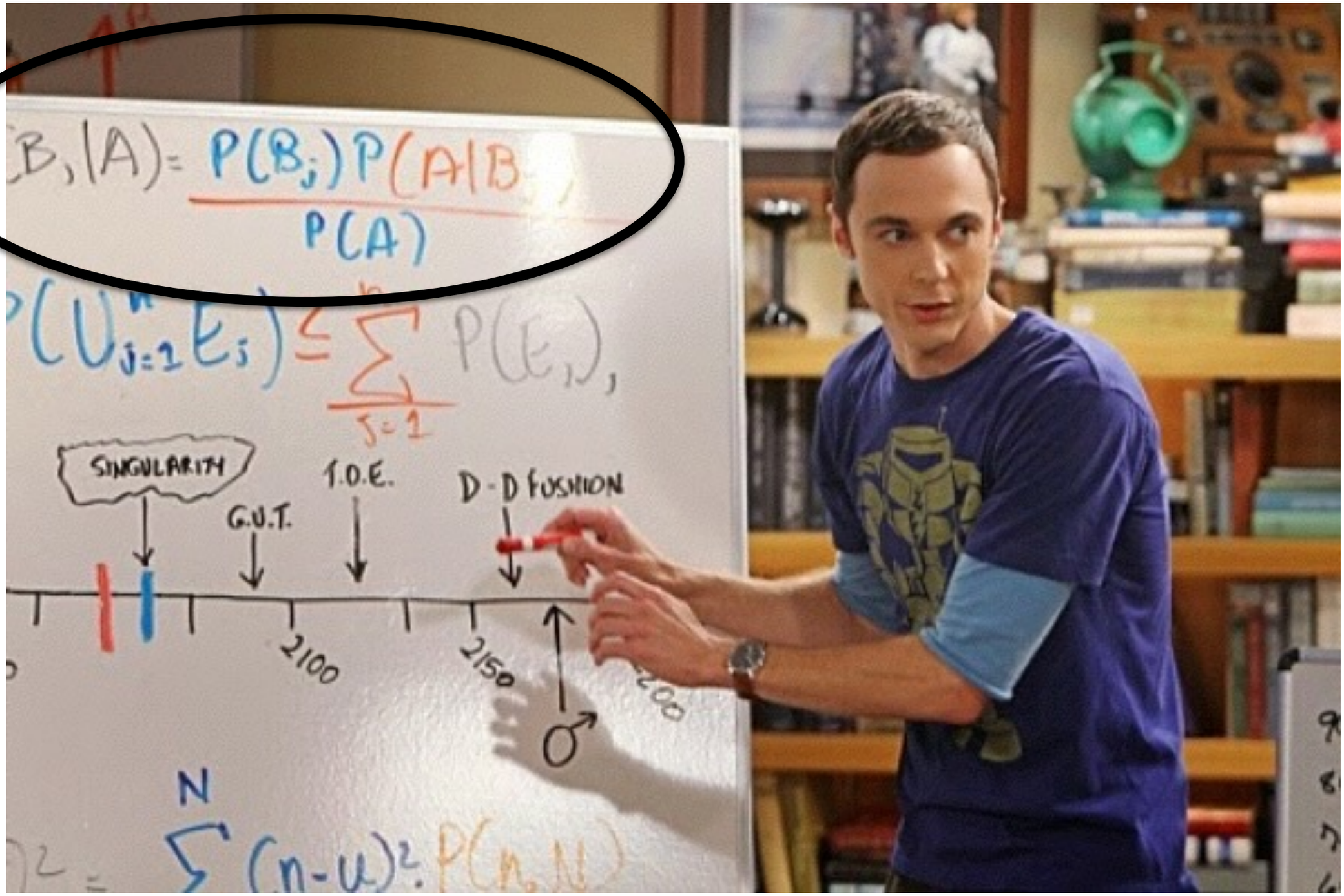
Statistics are fun!



Statistics are fun!



Statistics are fun!



Statistics are fun!

Evidence

Hypothesis

Likelihood of the evidence if the hypothesis is true

Prior probability of the hypothesis

$$P(H|E) = \frac{P(E|H) \times P(H)}{P(E)}$$

Posterior probability of the hypothesis “H” given the evidence “E”

Prior probability that the evidence is true

Statistics are fun!

Pick bowl #1 or bowl #2…

…close your eyes…

pick a **random cookie** from that bowl

Statistics are fun!

Evidence: you picked a **plain cookie**

Hypothesis: you picked the cookie from **bowl #1**

Probability of picking a plain cookie from bowl #1

Prior probability of picking bowl #1

$$P(H|E) = \frac{P(E|H) \times P(H)}{P(E)}$$

Probability that you picked the cookie from bowl #1 given that it is plain

Prior probability of getting a plain cookie from any bowl

Statistics are fun!

Evidence: you picked a **plain cookie**

Hypothesis: you picked the cookie from **bowl #1**

Probability of picking a plain cookie from bowl #1

Prior probability of picking bowl #1

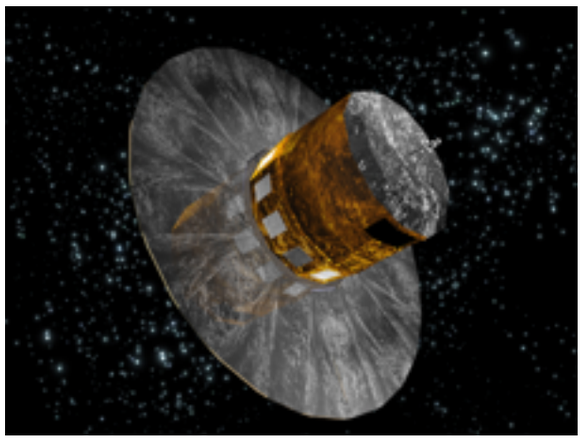
$$P(H|E) = \frac{15/20 \times 1/2}{20/40} = 75\%$$

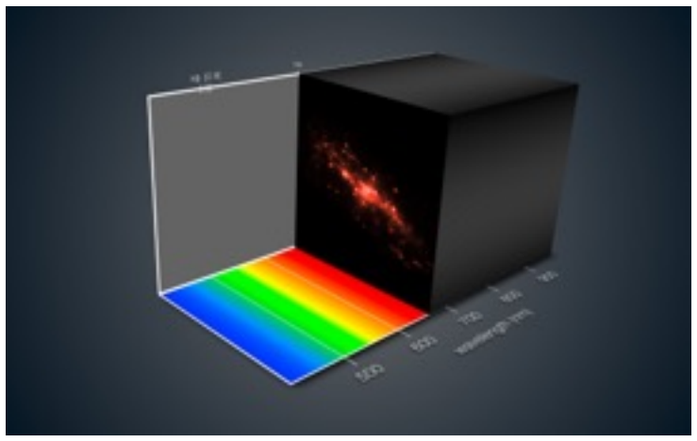
Probability that you picked the cookie from bowl #1 given that it is plain

Prior probability of getting a plain cookie from any bowl

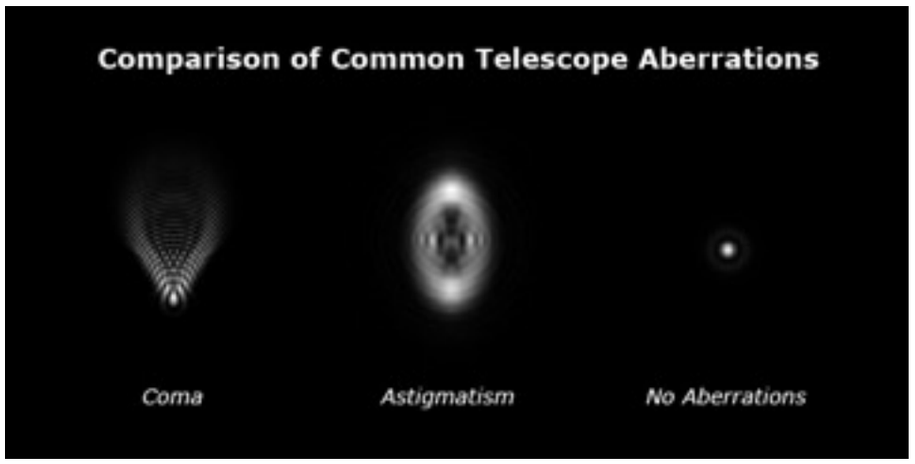
Statistics are useful (and critical in science)!

Large datasets

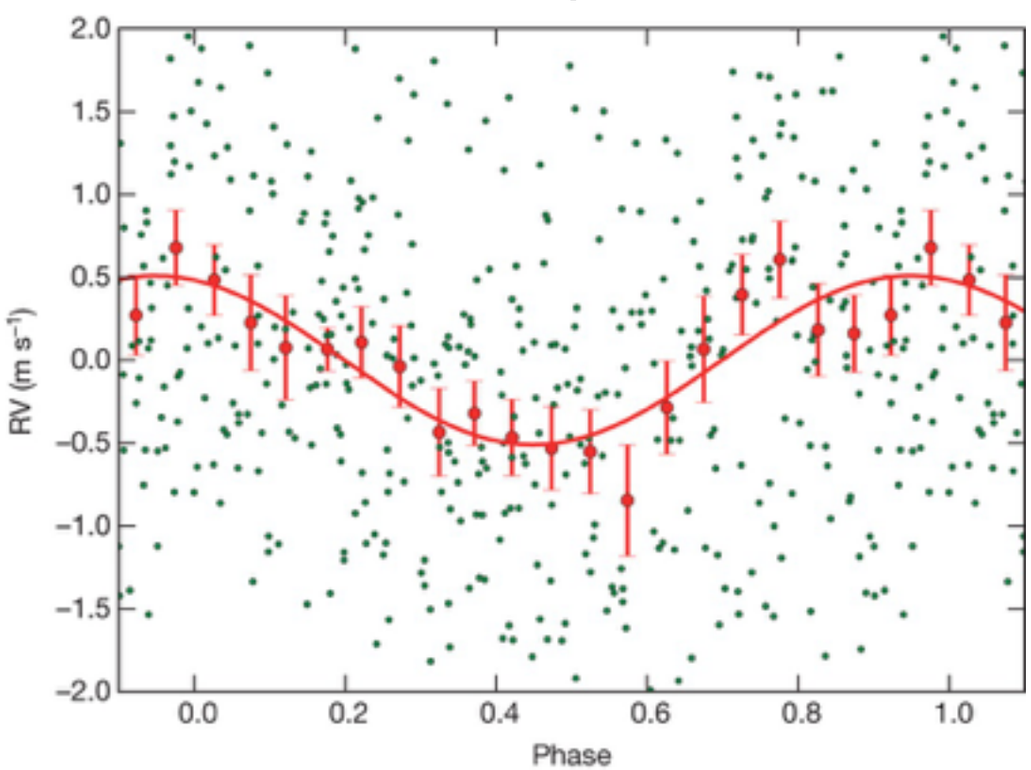




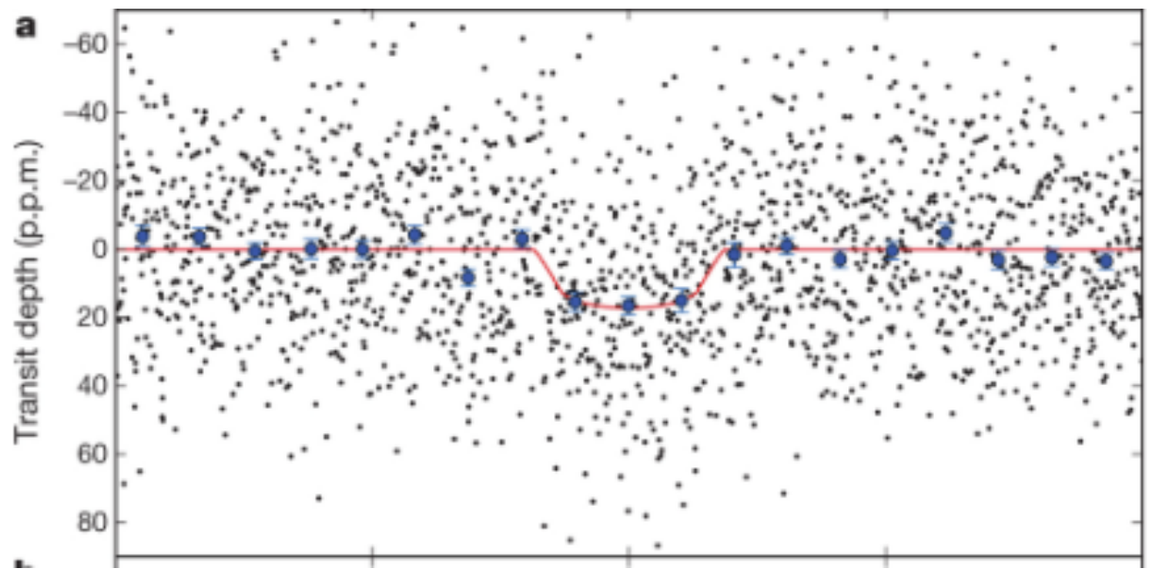
Non-ideal instrumentation



Need for more precise data

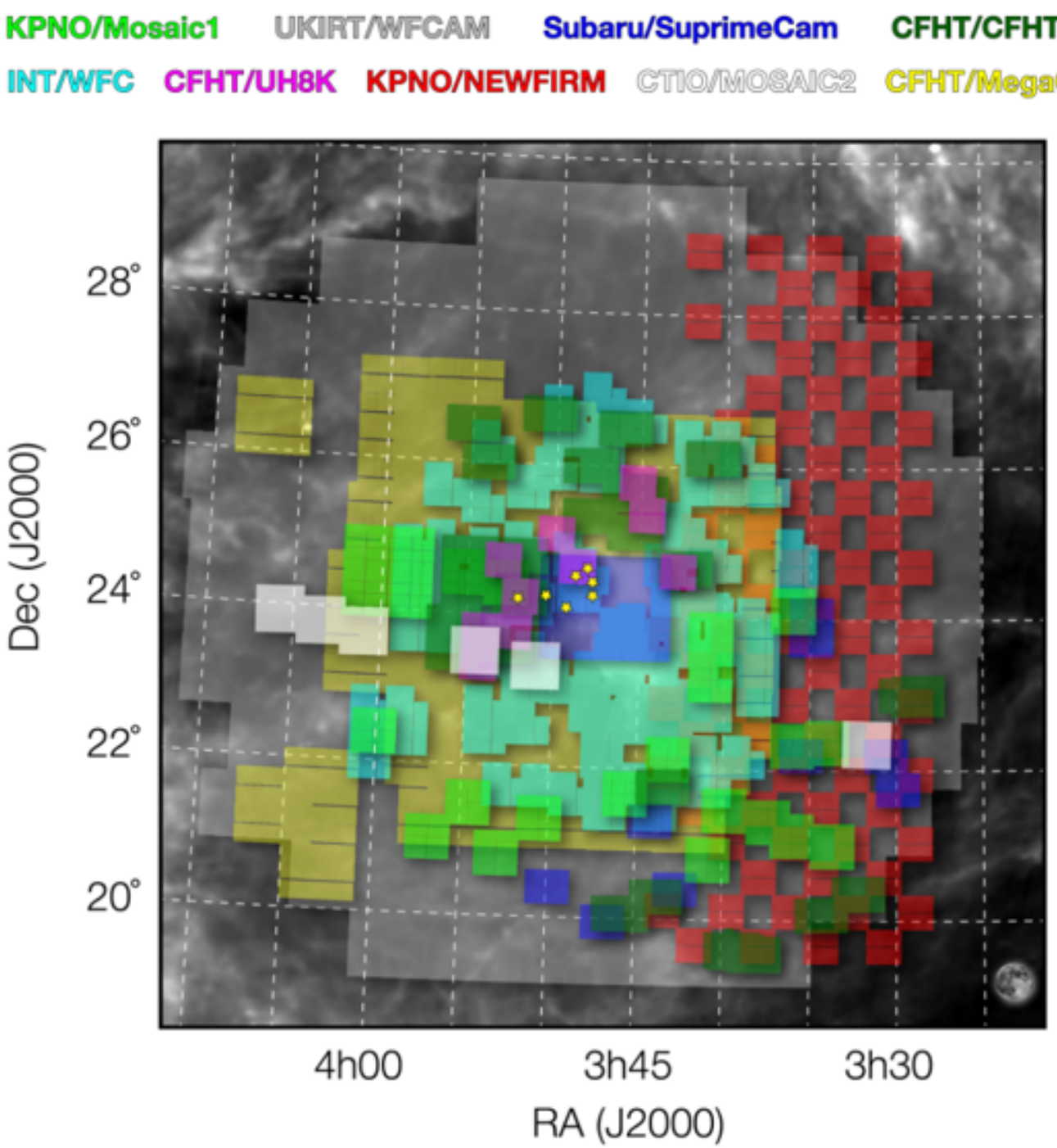


Need to get the most out of our data



Statistics are useful (and critical in science)!

An example: the DAnCe project

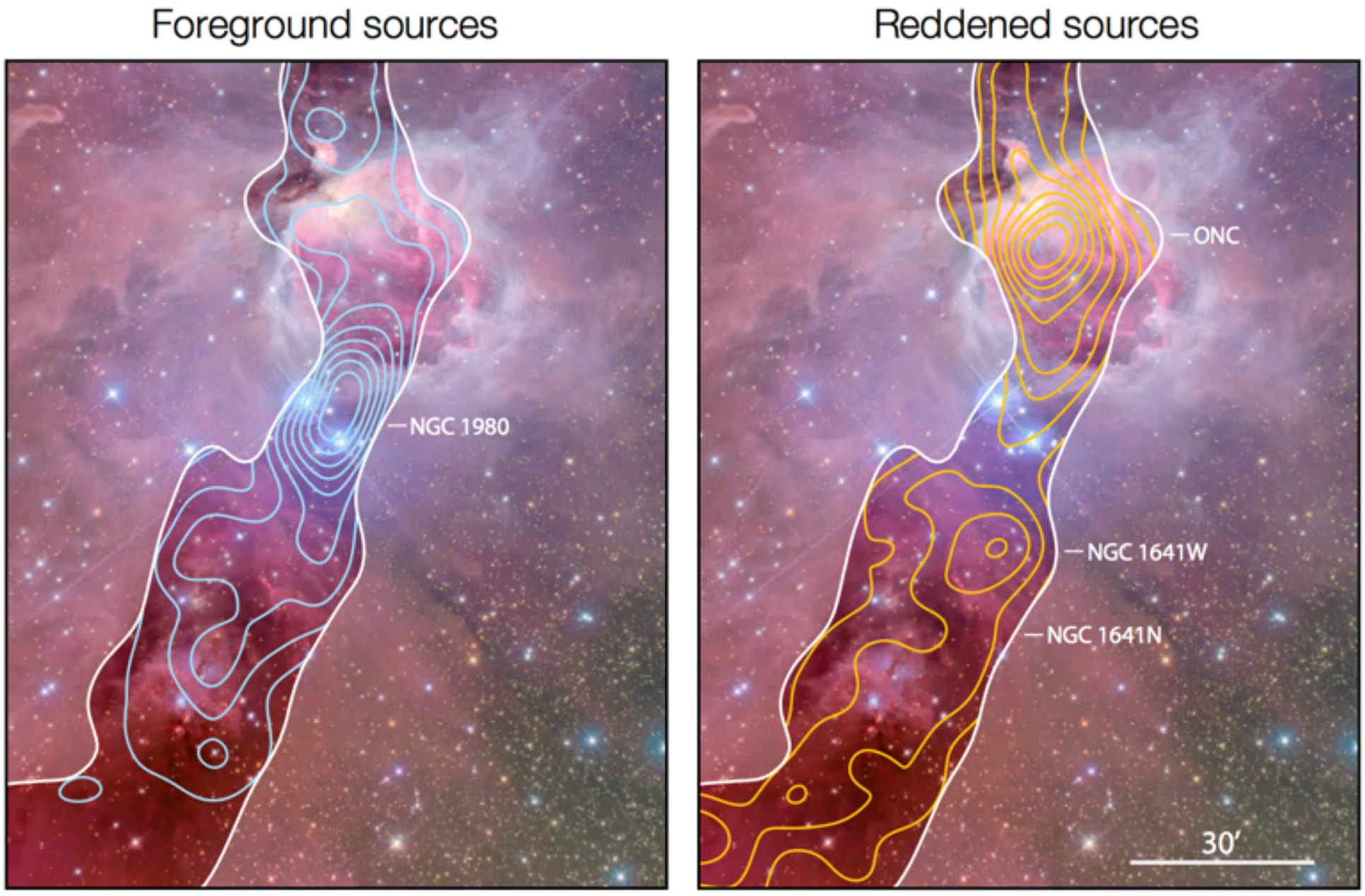


3.35×10^6 sources studied

Bouy et al. (2012)

Statistics are useful (and critical in science)!

An example: the DAnCe project



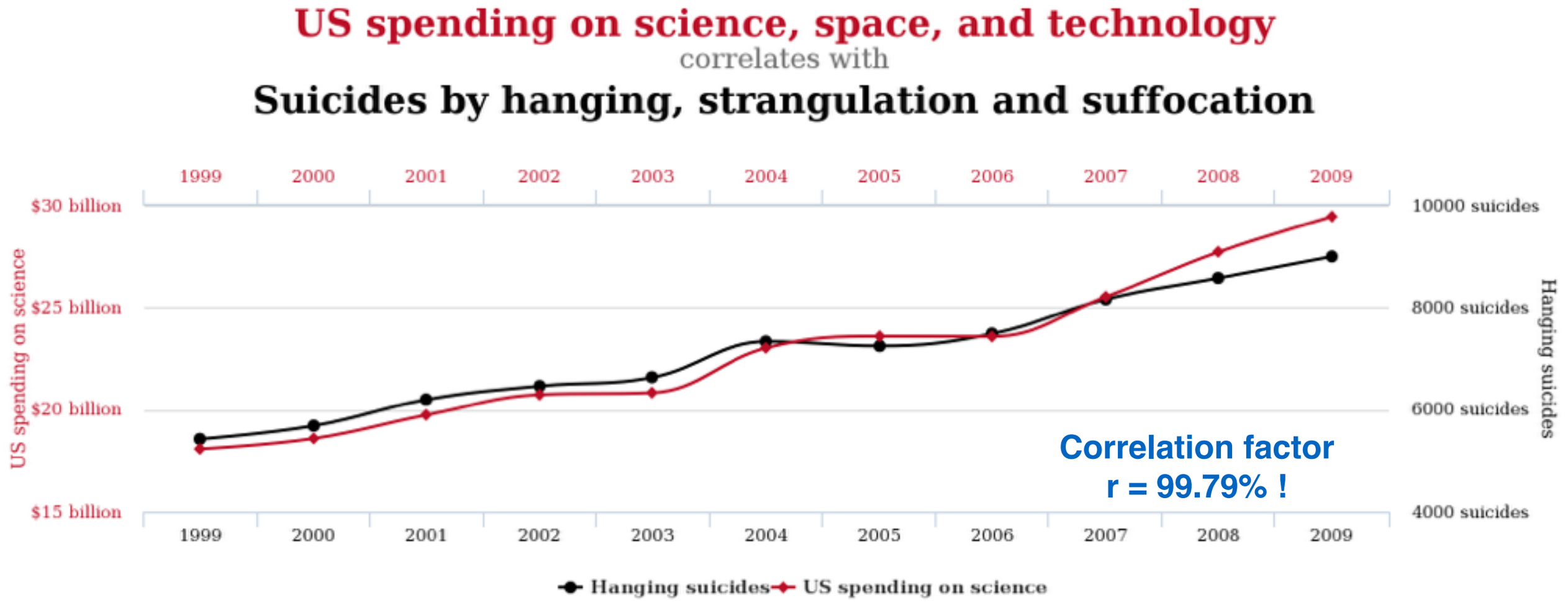
Alves & Bouy (2015)

But...

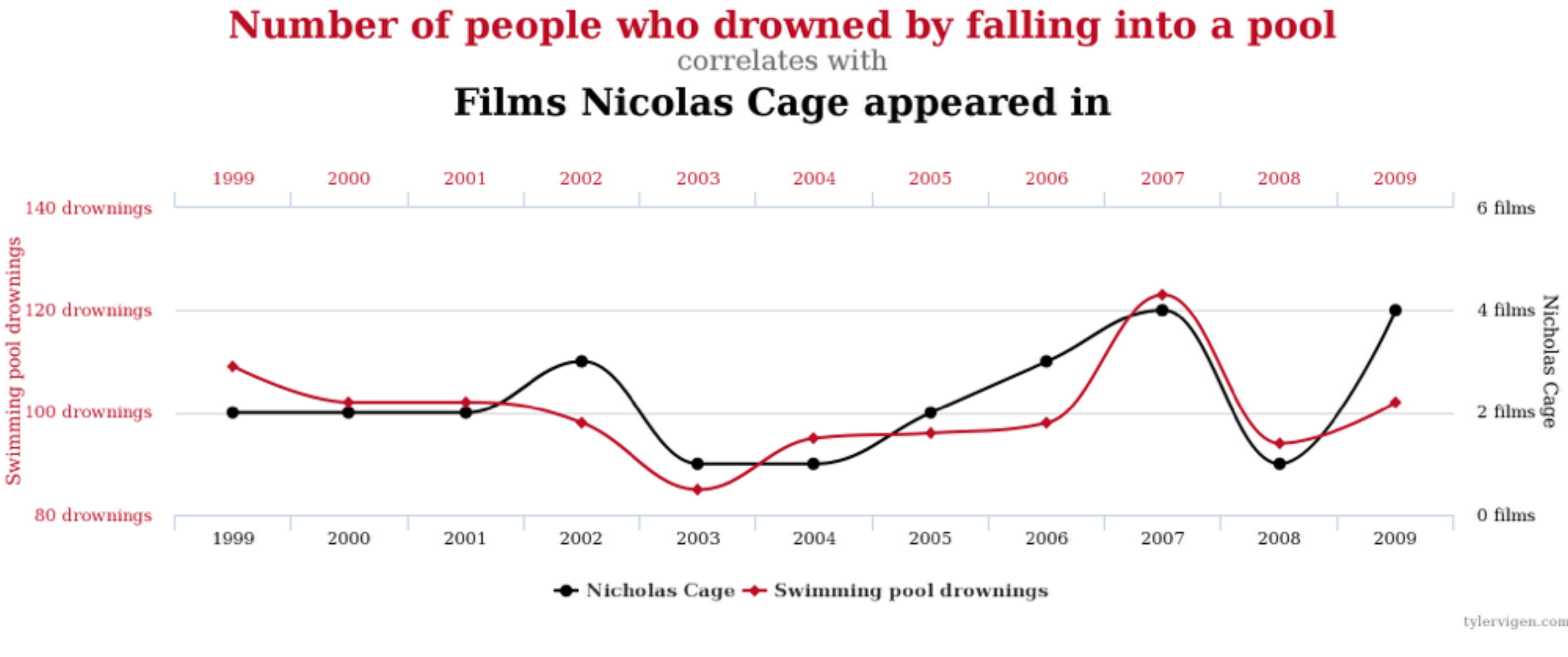
Statistics is just a tool for **extracting knwoledge from our data**

Interpreting the conclusions and validity of the results **needs human interaction**

For example…


US spending on science, space, and technology
correlates with
Suicides by hanging, strangulation and suffocation
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
$30 billion
$25 billion
$20 billion
$15 billion
US spending on science
10000 suicides
8000 suicides
6000 suicides
4000 suicides
Hanging suicides
Correlation factor
r = 99.79% !
Hanging suicides
US spending on science
tylervigen.com

For example…


Number of people who drowned by falling into a pool
correlates with
Films Nicolas Cage appeared in
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
140 drownings
120 drownings
100 drownings
80 drownings
Swimming pool drownings
6 films
4 films
2 films
0 films
Nicholas Cage
Nicholas Cage
Swimming pool drownings
tylervigen.com

For example…

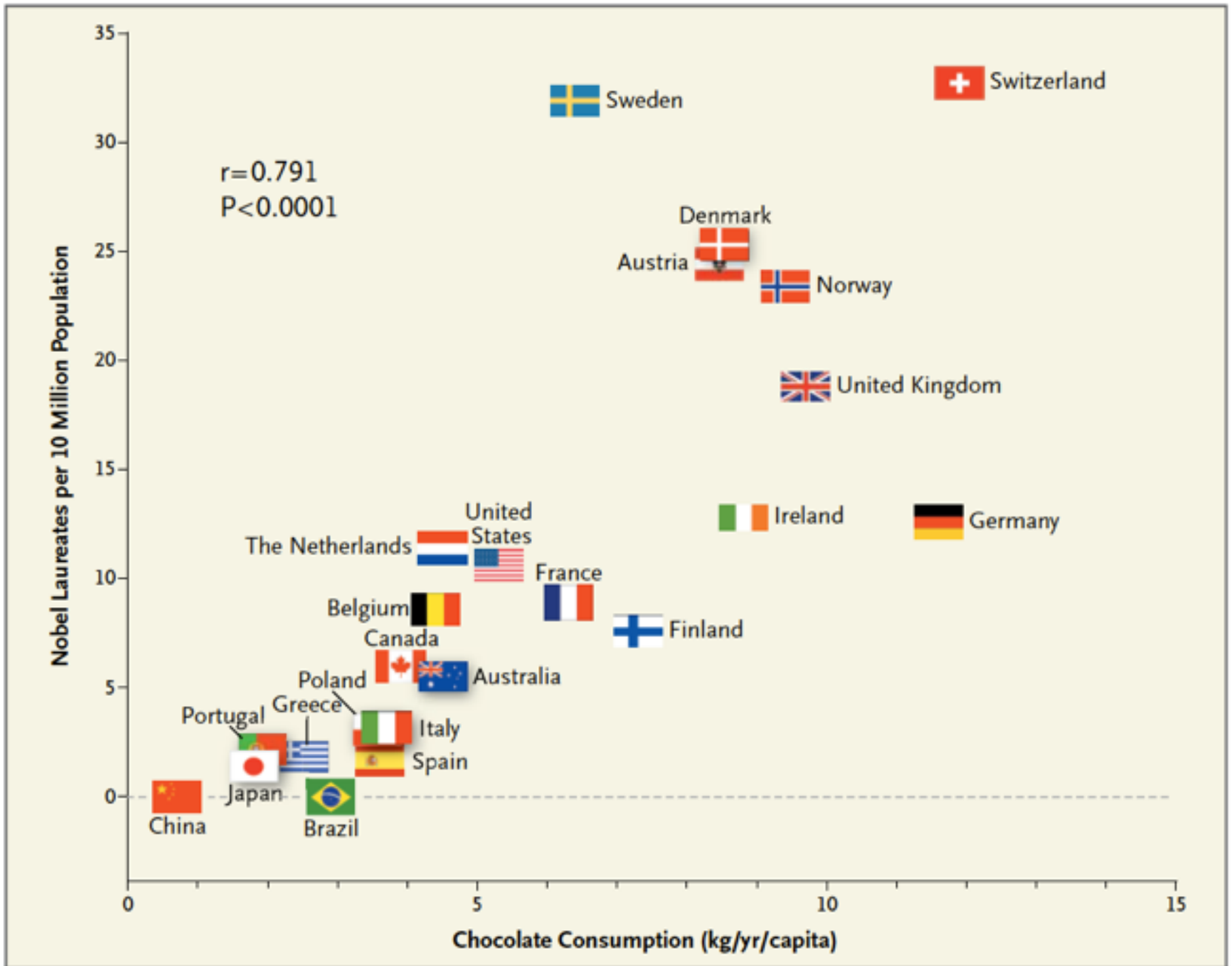


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

So we need to (and we will) learn…

(At least) the **basic theory** of statistics

e.g.,

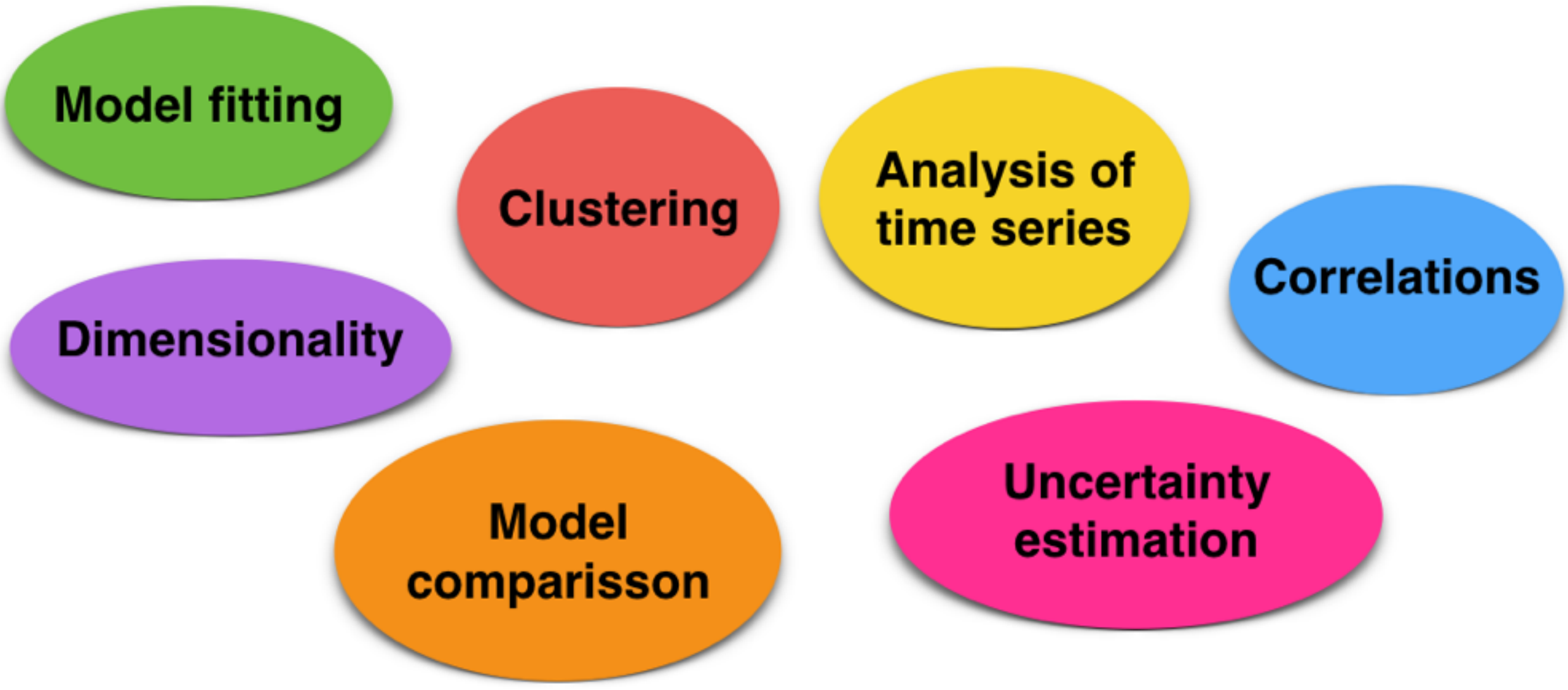
$$P(H|E) = \frac{P(E|H) \times P(H)}{P(E)}$$

So we need to (and we will) learn…

(At least) the **basic theory** of statistics

$$P(H|E) = \frac{P(E|H) \times P(H)}{P(E)}$$

The **statistic tools** to apply to our data

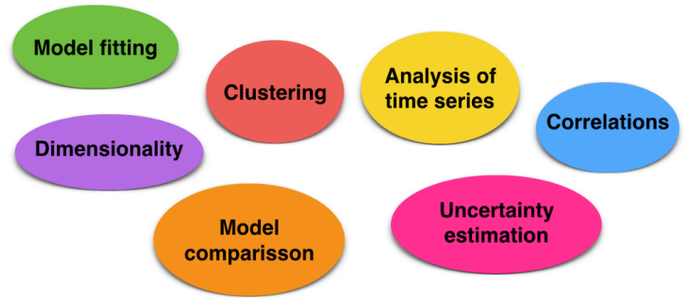


So we need to (and we will) learn…

(At least) the **basic theory** of statistics

$$P(H|E) = \frac{P(E|H) \times P(H)}{P(E)}$$

The **statistic tools** to apply to our data



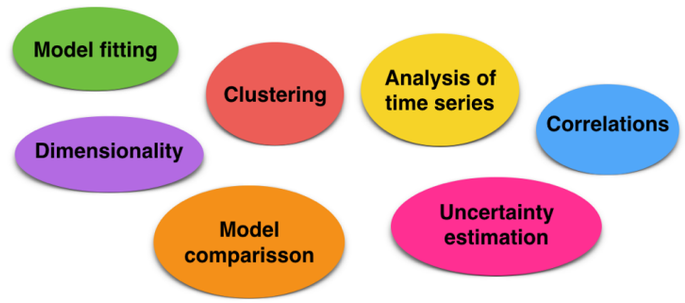
How to **decide** which tool/s should be applied

So we need to (and we will) learn…

(At least) the **basic theory** of statistics

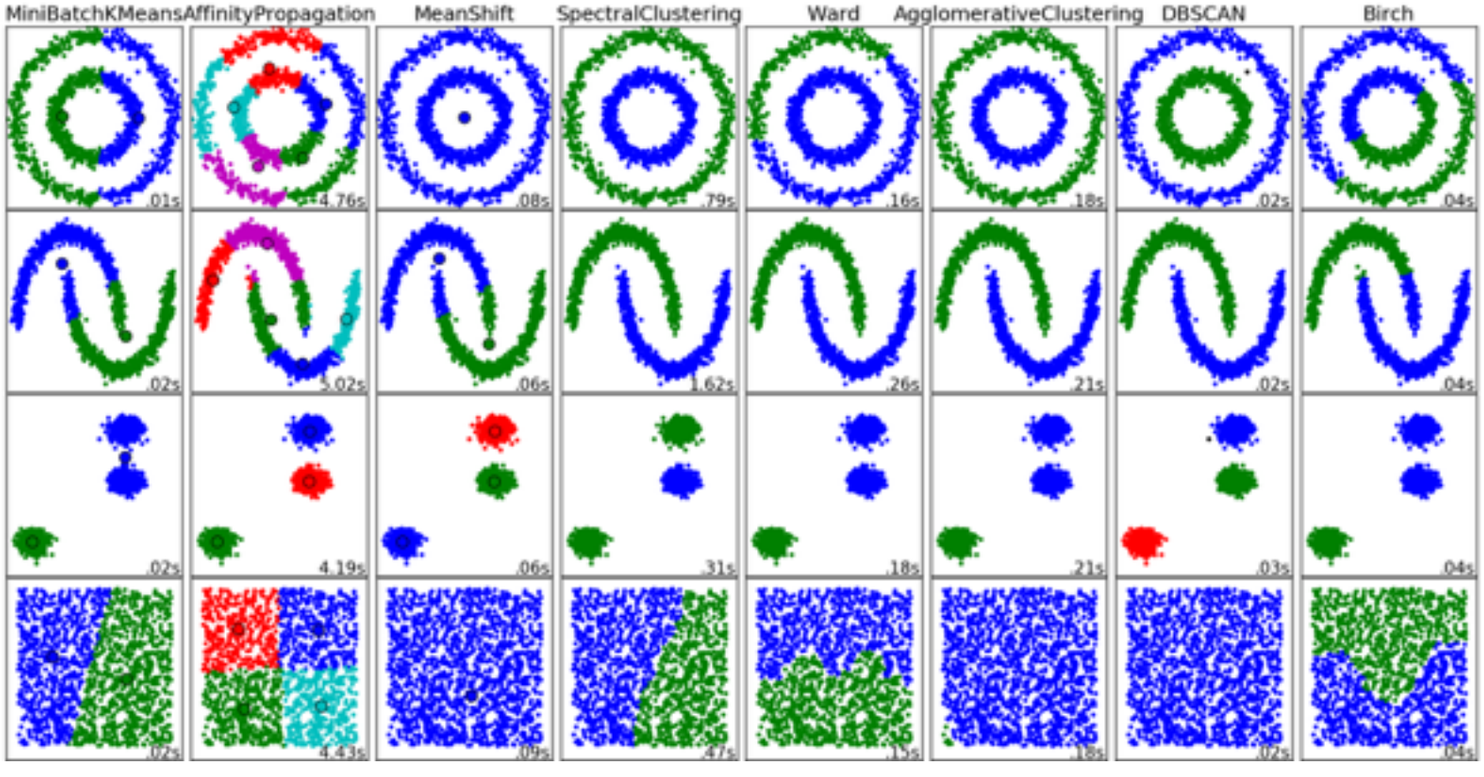
$$P(H|E) = \frac{P(E|H) \times P(H)}{P(E)}$$

The **statistic tools** to apply to our data



How to **decide** which tool/s should be applied

How to **compare** the results obtained from the different tools

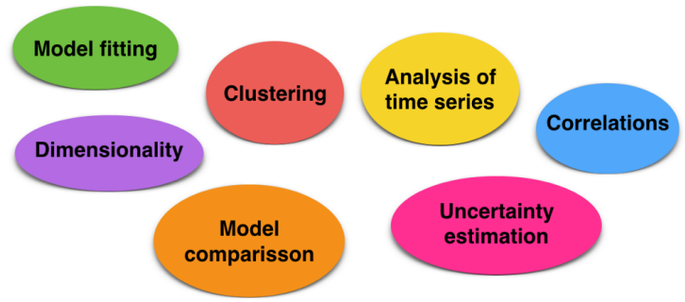


So we need to (and we will) learn…

(At least) the **basic theory** of statistics

$$P(H|E) = \frac{P(E|H) \times P(H)}{P(E)}$$

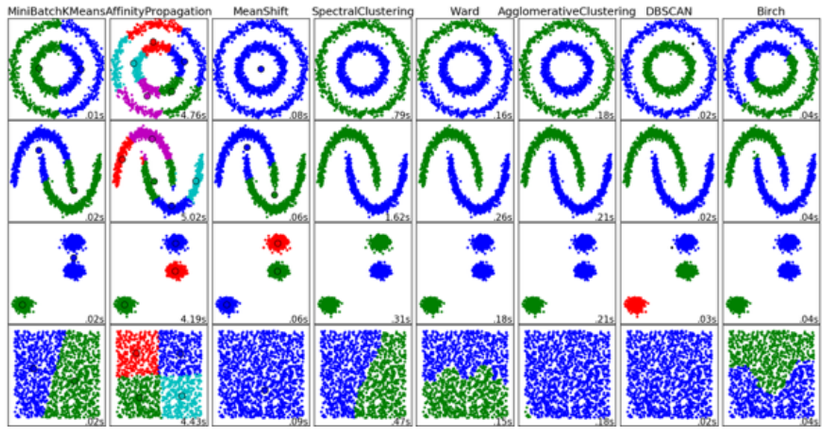
The **statistic tools** to apply to our data



How to **decide** which tool/s should be applied

How to **compare** the results obtained from the different tools

How to **interpret** the results from our statistical analysis

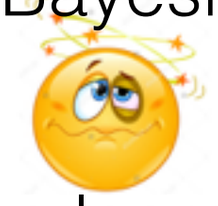


Philosophy of the *MCMC Coffee*

Ideal paper outline
(advice by E. Feigelson)

1. Introduction
2. Observations and data reduction
3. Non-parametric exploration of my data
4. Maximum likelihood analysis
5. Bayesian analysis
   (aka, include priors and see what happens)
6. Model comparison
7. Scientific discussion

Wrong use of statistical techniques (and overuse of unefficient but popular tools)

- K-S test —> Anderson-Darling test
- Overuse of histograms for inference (e.g., fit a gaussian to a histogram to get the median…). 
- Overuse of linear and power-law regressions —> use local regressions!
- Underuse of poisson regression
- Insufficient examination of regression results: R^2, analysis of the residuals, autocorrelation, outliers via Cook distance, etc. 
- Overuse of Bayesian inference with uninformative priors (that's MLE!) —> Just use Bayesian inference when you have informative priors. 
- Underuse of machine learning methods.

Format of the sessions

i. Quick overview of the **scientific context**

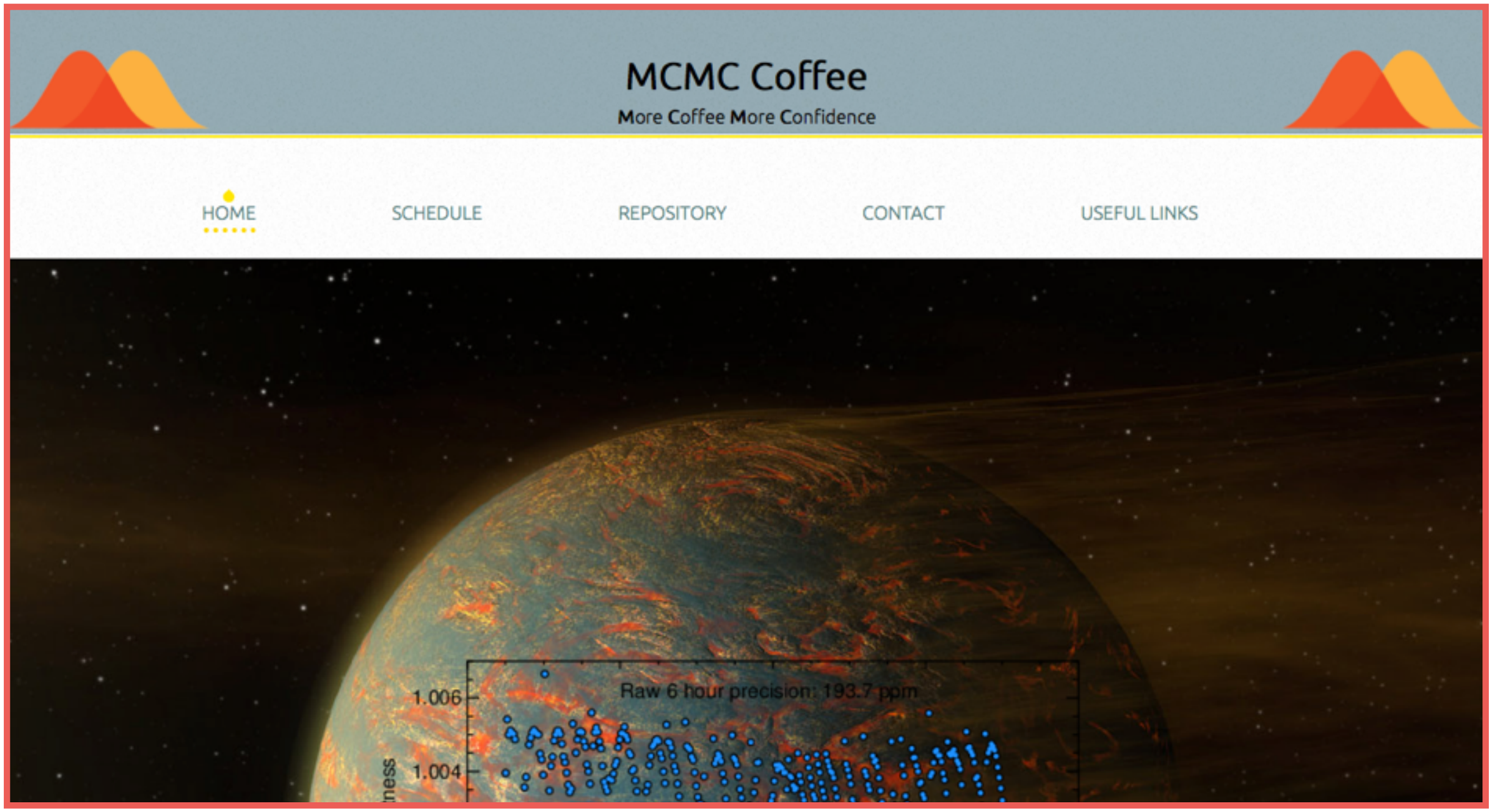
ii. **Statistical challenge:** question you want to answer

iii. **Technique/s used** and reasons for not using other.

iv. Present the **code** and final solution that you found

MCMC Coffee website

http://www.sc.eso.org/~jlillobo/mcmc_coffee/index.html



Thank you!