



GAUTHAM NARAYAN

---

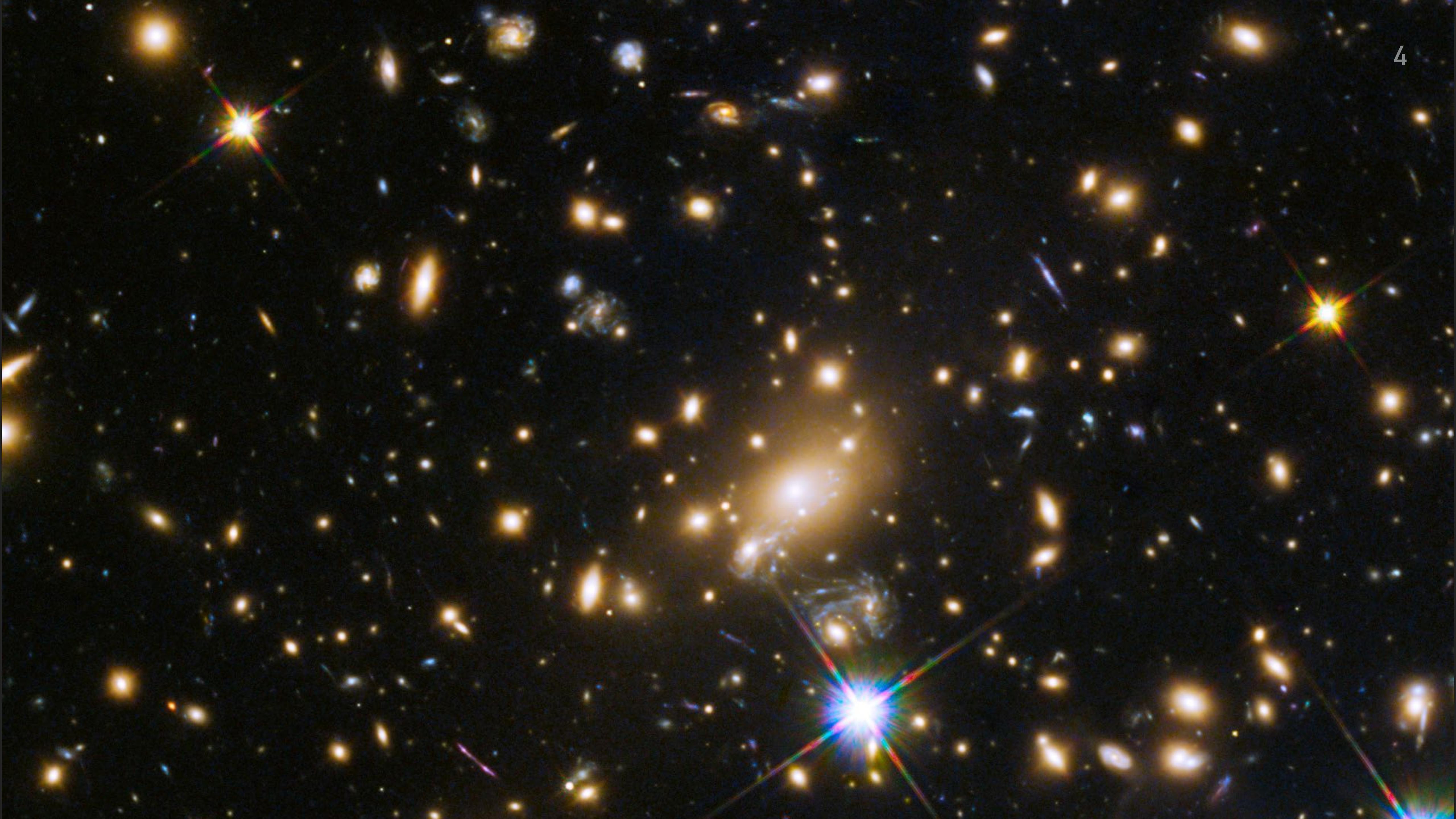
**ASTR496: FOUNDATIONS OF DATA SCIENCE  
DISTRIBUTIONS,**

1.0

---

# DATA

- ▶ To do some science you:
  - ▶ Hypothesis building: determine a question you'd like to answer
  - ▶ Experimental Design: figure out what observables you can use to answer that question
  - ▶ Observations: take some measurements (i.e. data)
  - ▶ Analysis: explore that data, summarize, visualize it and build a hypothesis (i.e. a model)
  - ▶ Inference: evaluate the model vs the alternatives given the data
  - ▶ and actually you'd better also write up the results.



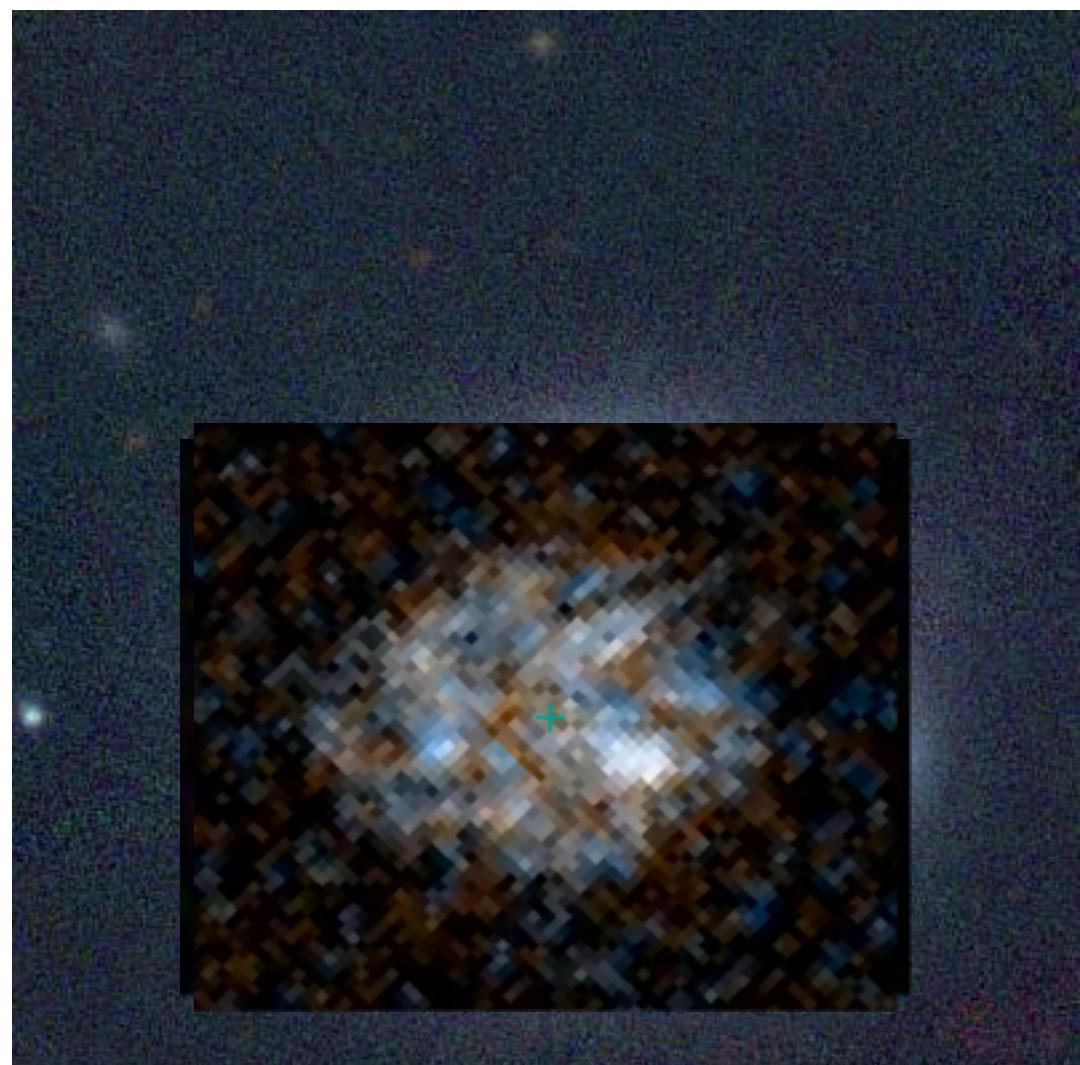
# WHAT ASTRONOMERS CAN MEASURE

- ▶ Angular separations
- ▶ Time differences
- ▶ Energy differences

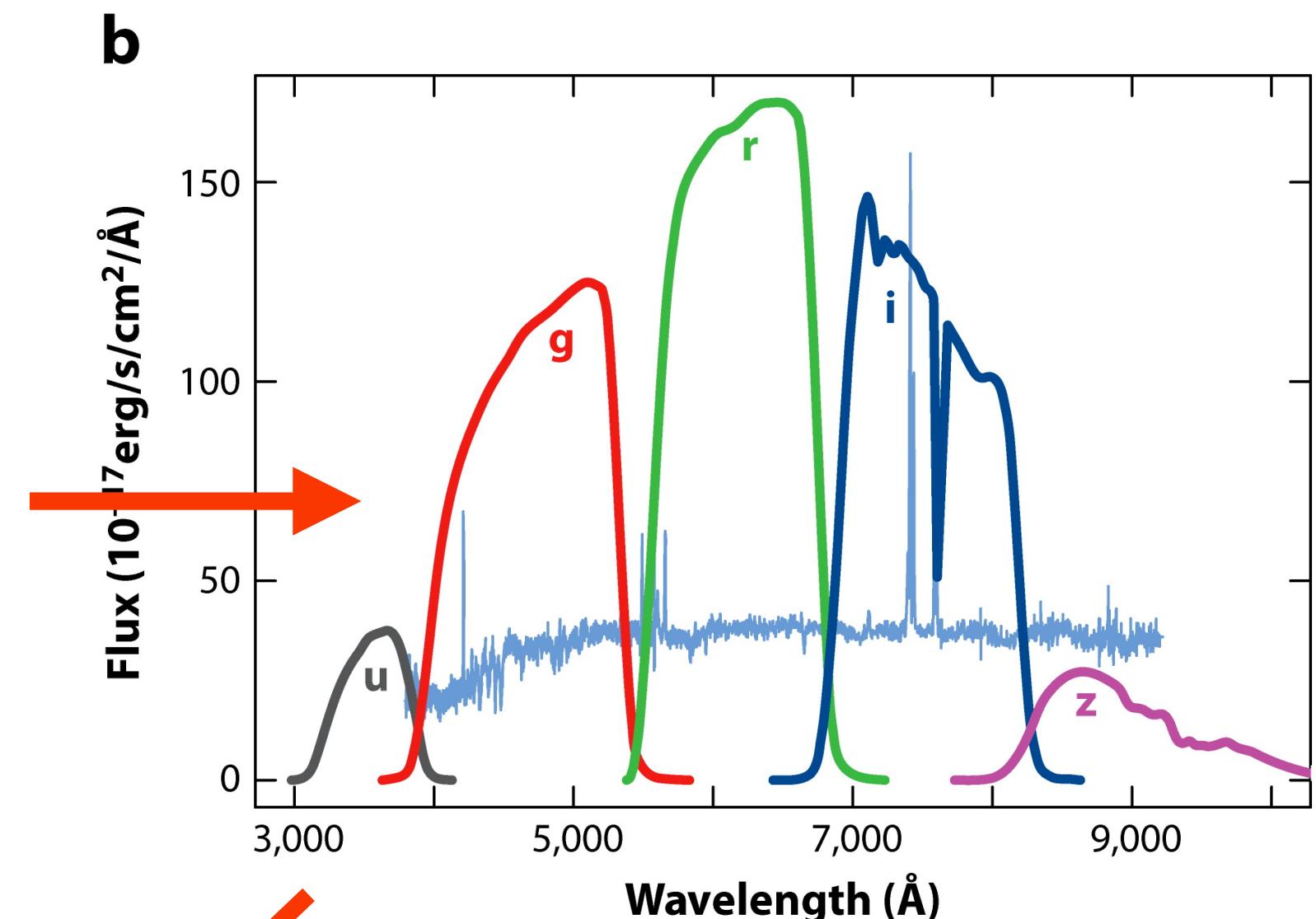
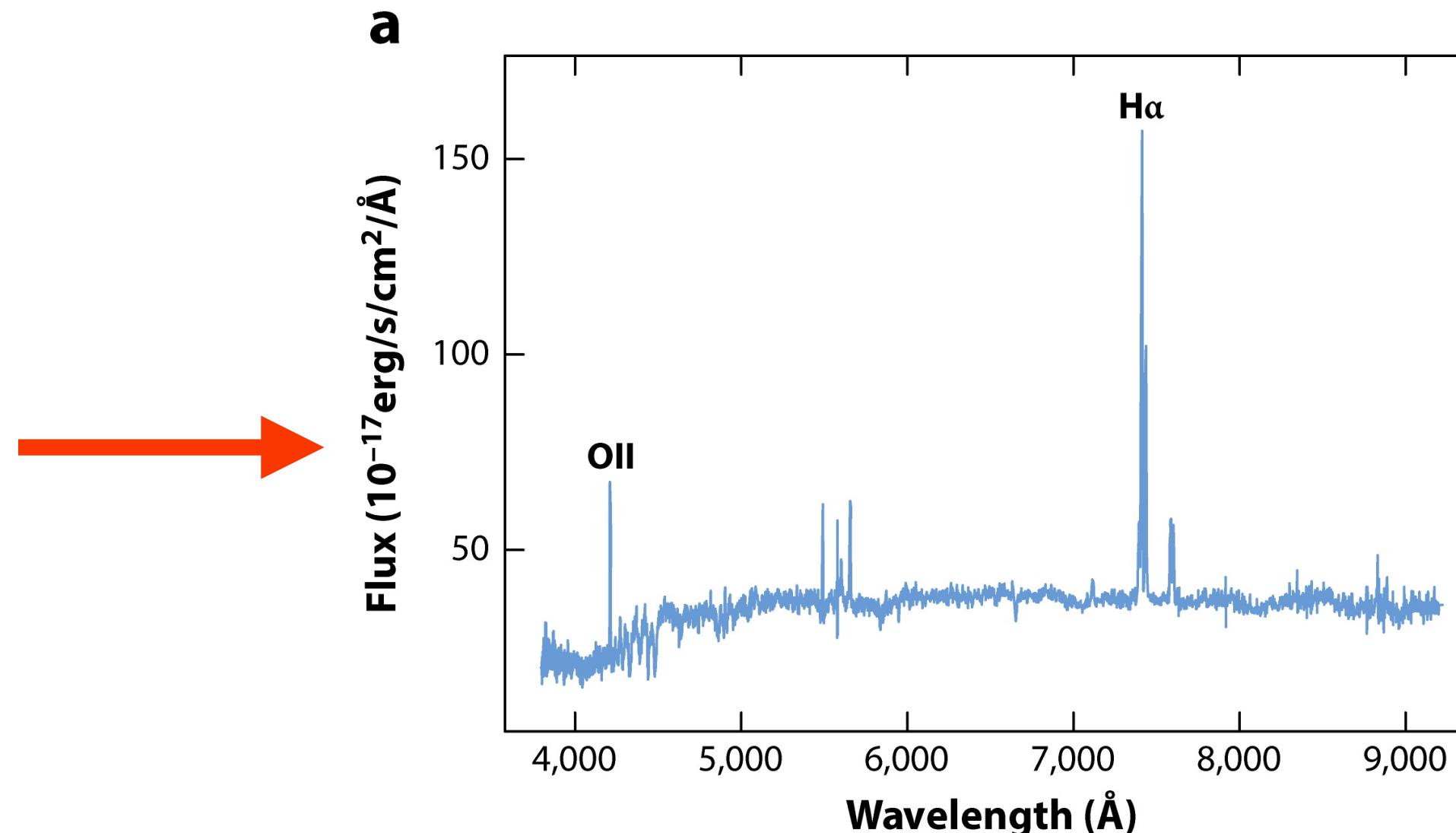
That's it.

# WHAT ASTRONOMERS CAN MEASURE

- ▶ **Astrometry** (angular position on the sky) - arcseconds
  - ▶ Related definition: 1 parsec (pc) = distance at which a distance of 1 AU (i.e. Earth-Sun) subtends an angle of 1 arcsecond  
i.e.  $1 \text{ pc} = 1 \text{ AU}/\tan(1'') \sim 31 \text{ trillion kilometers or 3.26 light years (ly)}$
- ▶ **Photometry** (how bright something is)
  - ▶ Flux = photons (or energy in ergs)/sec/cm<sup>2</sup>
  - ▶ (Apparent) Magnitude =  $-2.5 \log_{10}(\text{Flux}) + \text{const}$
  - ▶ (Absolute) Magnitude =  $-2.5 \log_{10}(\text{Luminosity}) + \text{const} = \text{magnitude you'd measure if you could move the source to 10 pc}$
- ▶ **Light curves** = photometry vs time
  - ▶ Evolution in source brightness either because of intrinsic (supernovae, AGN) or extrinsic (asteroids, eclipsing binaries)
- ▶ **Spectroscopy** = Energy vs wavelength/frequency
- ▶ **Images/maps** = Energy vs position on the sky (clustering, spatial correlation functions)
- ▶ **Proper Motion** = Astrometry vs time (e.g. stars, satellite galaxies, asteroids...)



Optical/IR/UV



**A** Schafer CM. 2015.  
**R** Annu. Rev. Stat. Appl. 2:141–62

Galaxy Photometry

Galaxy Spectrum

(Brightness: Flux / Magnitude)

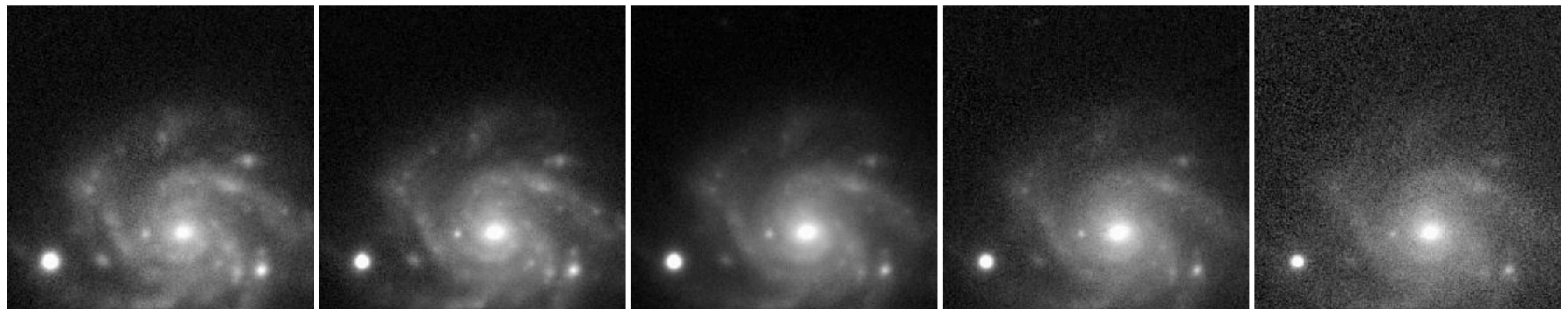
stack 920.038 g  
[Display FITS FITS-cutout](#)

stack 920.038 r  
[Display FITS FITS-cutout](#)

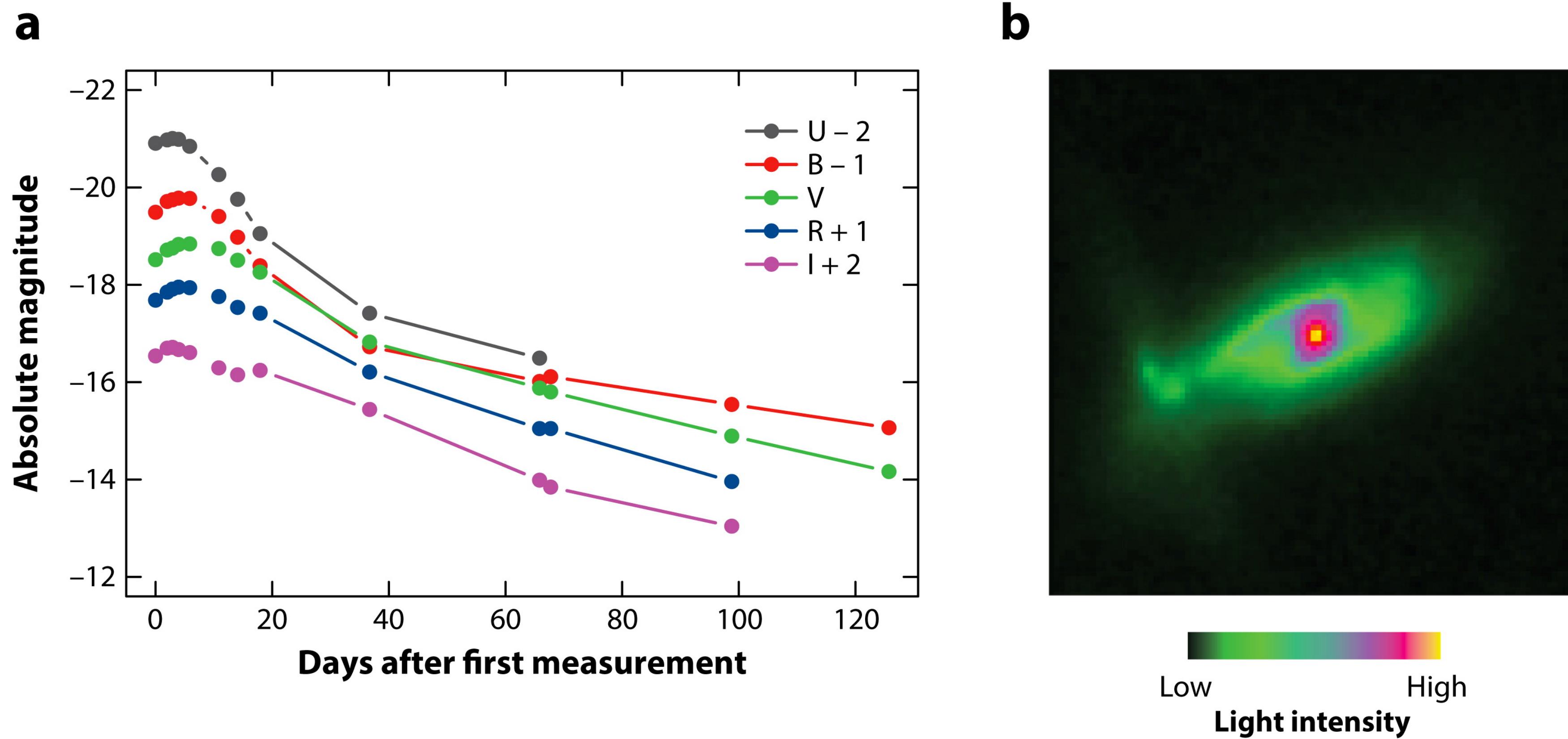
stack 920.038 i  
[Display FITS FITS-cutout](#)

stack 920.038 z  
[Display FITS FITS-cutout](#)

stack 920.038 y  
[Display FITS FITS-cutout](#)



# Temporal & Spatial Variation



Schafer CM. 2015.

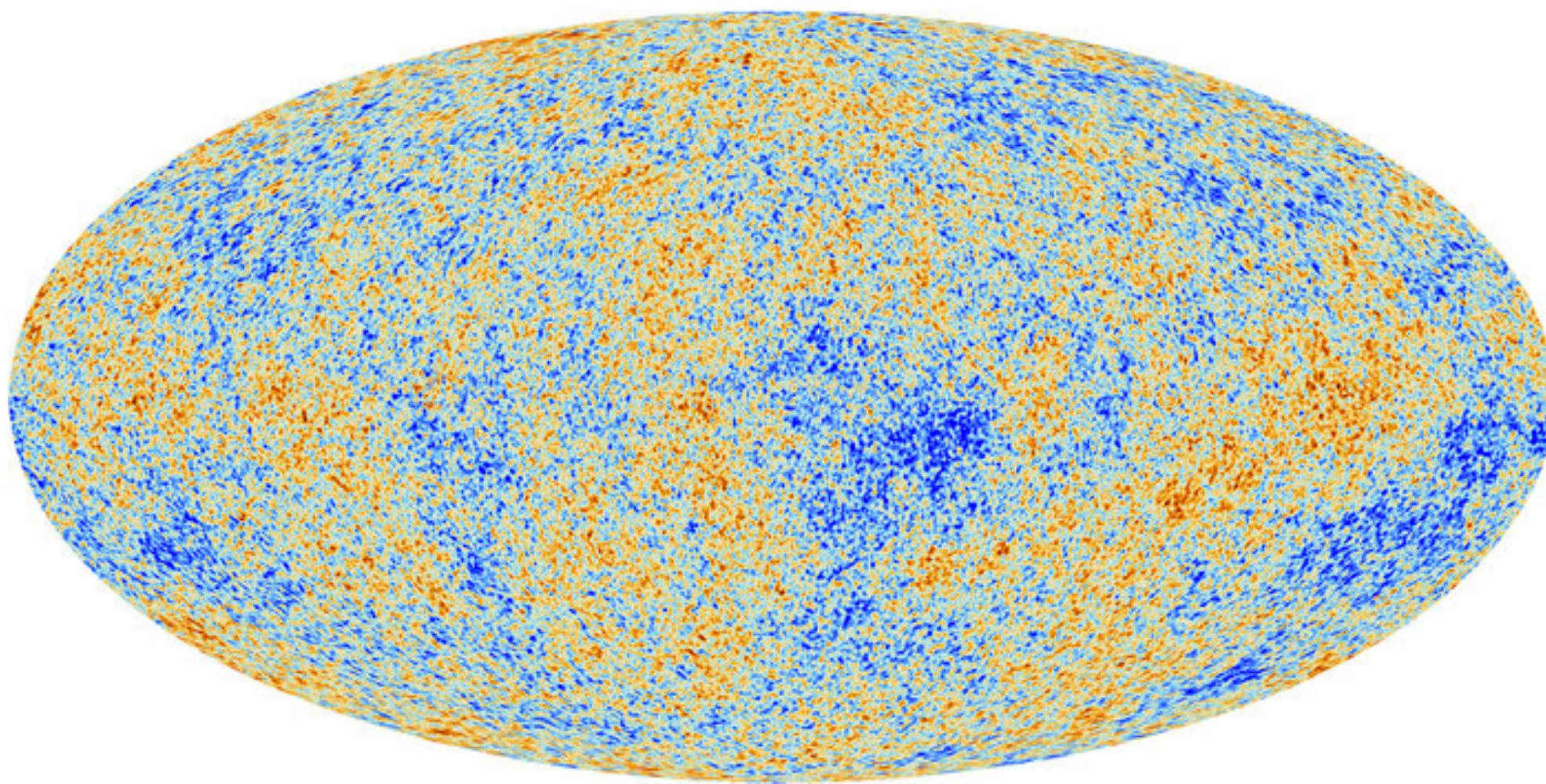
Annu. Rev. Stat. Appl. 2:141–62

Time Series (Light Curve)  
Supernova

Galaxy Image  
(Intensity Map)

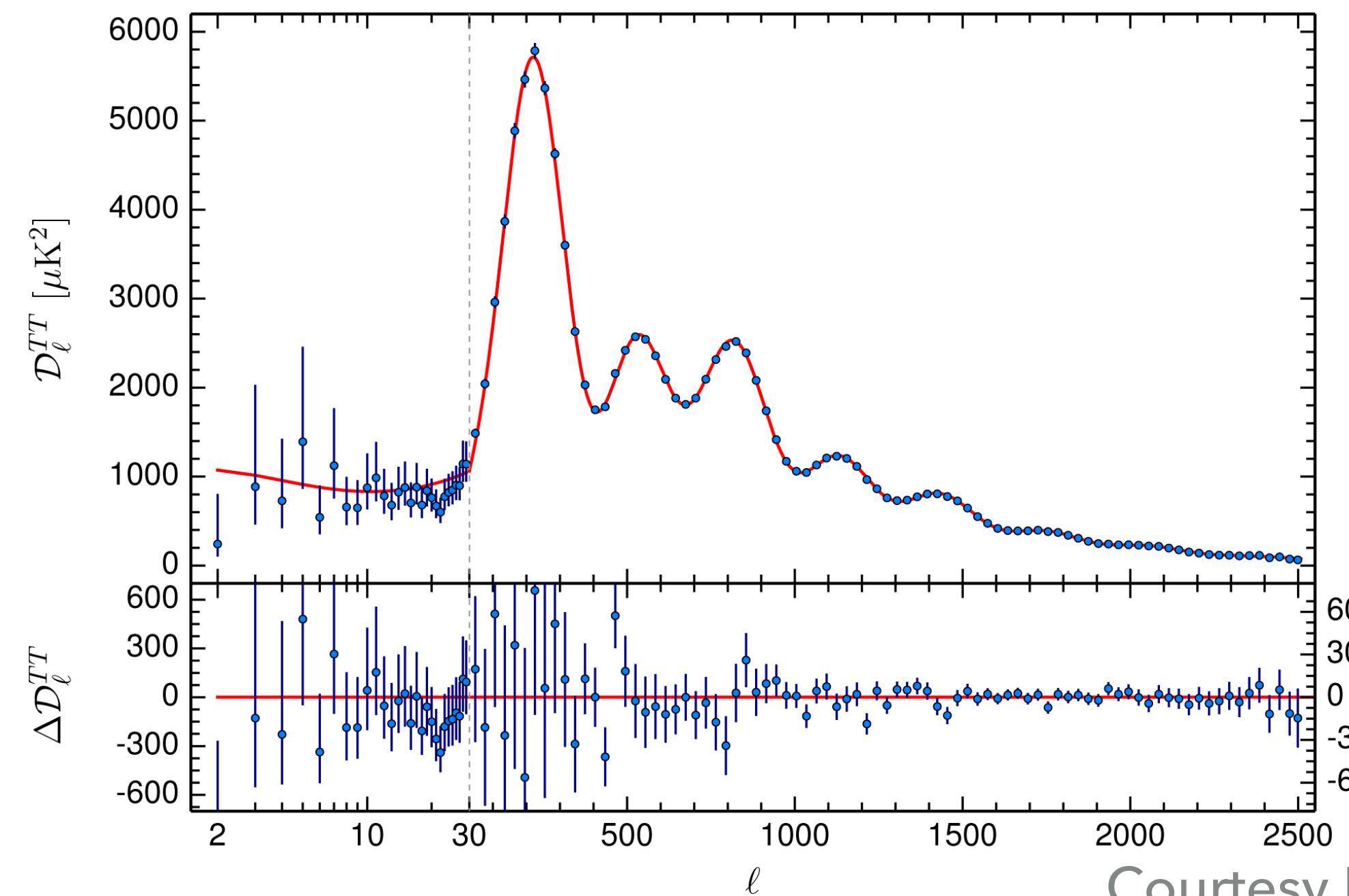
Courtesy Kaisey Mandel

# Spatial Variation



Cosmic Microwave  
Background (Planck)  
~ Gaussian Random Field  
(mean = 2.7 K,  
std dev  $\sim 10^{-5}$ )

Power Spectrum  
(~Fourier Transform of  
Correlation Function)  
sensitive to cosmological  
parameters



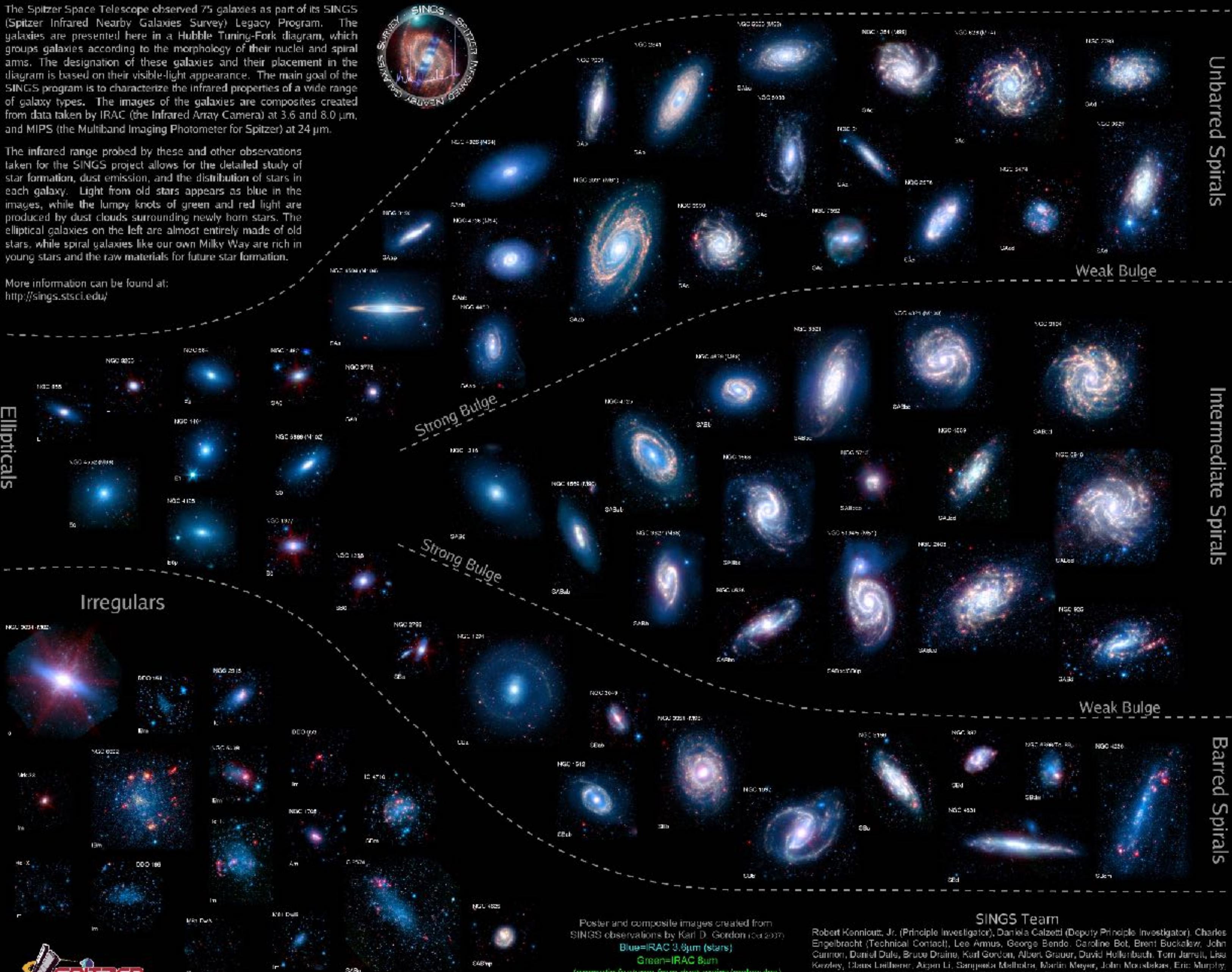
Courtesy Kaisey Mandel

The Spitzer Infrared Nearby Galaxies Survey (SINGS) Hubble Tuning-Fork

The Spitzer Space Telescope observed 75 galaxies as part of its SINGS (Spitzer Infrared Nearby Galaxies Survey) Legacy Program. The galaxies are presented here in a Hubble Tuning-Fork diagram, which groups galaxies according to the morphology of their nuclei and spiral arms. The designation of these galaxies and their placement in the diagram is based on their visible-light appearance. The main goal of the SINGS program is to characterize the infrared properties of a wide range of galaxy types. The images of the galaxies are composites created from data taken by IRAC (the Infrared Array Camera) at 3.6 and 8.0  $\mu$ m and MIPS (the Multiband Imaging Photometer for Spitzer) at 24  $\mu$ m.

The infrared range probed by these and other observations taken for the SINGS project allows for the detailed study of star formation, dust emission, and the distribution of stars in each galaxy. Light from old stars appears as blue in the images, while the lumpy knots of green and red light are produced by dust clouds surrounding newly born stars. The elliptical galaxies on the left are almost entirely made of old stars, while spiral galaxies like our own Milky Way are rich in young stars and the raw materials for future star formation.

More information can be found at  
<http://sings.stsci.edu/>



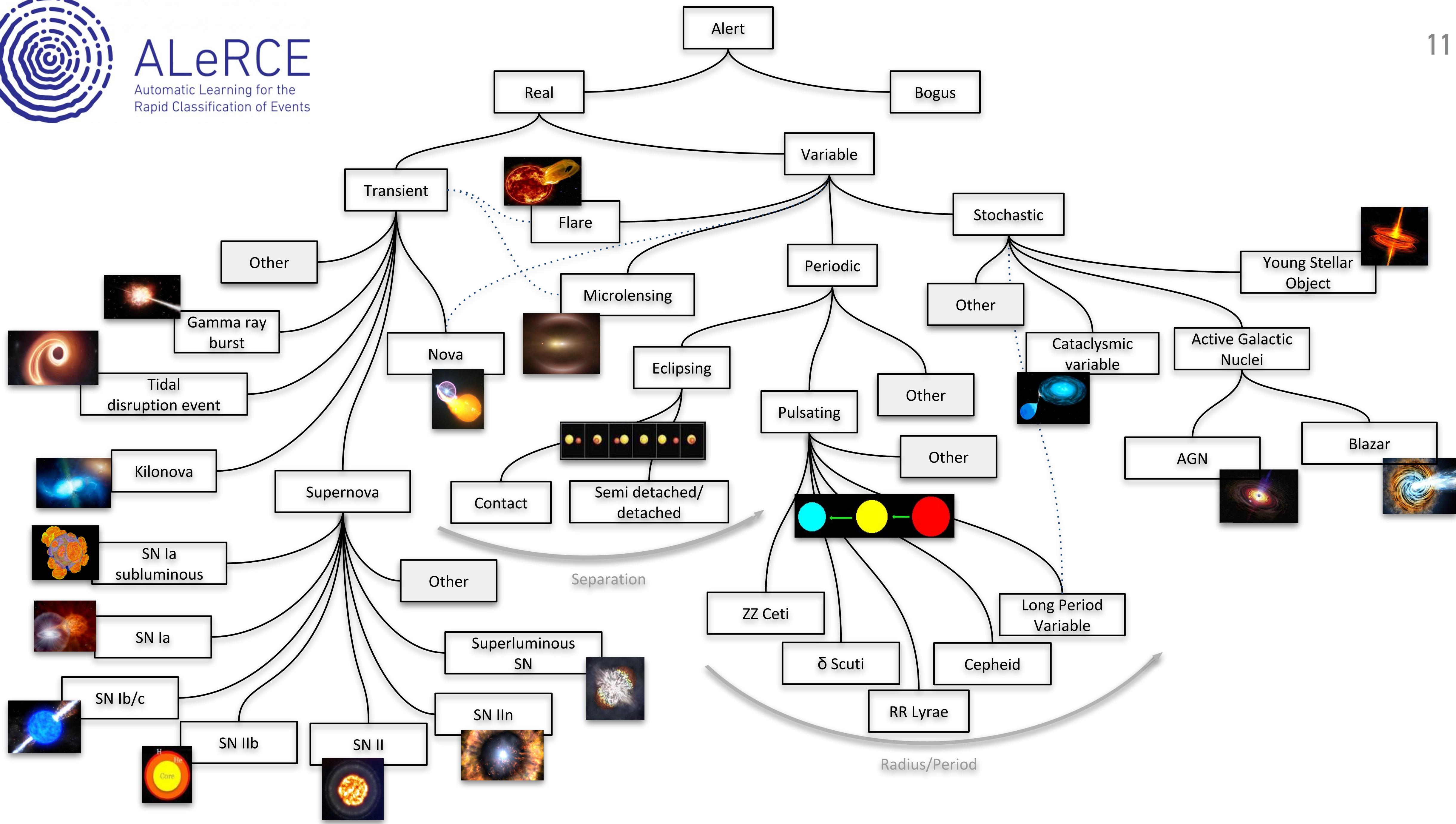
Poster and composite images created from SINGS observations by Karl D. Gordon (caltech)  
 Blue=IRAC 3.6 $\mu$ m (stars)  
 Green=IRAC 8 $\mu$ m  
 (aromatic features from dust grains/molecules)  
 Red=MIPS 24 $\mu$ m (warm dust)

**SINGS Team**  
Robert Kennicutt, Jr. (Principle Investigator), Daniela Calzetti (Deputy Principle Investigator), Charles Engelbracht (Technical Contact), Lee Armus, George Bendo, Caroline Bot, Brent Buckalew, John Cannon, Daniel Dale, Bruce Draine, Karl Gordon, Albert Grauer, David Hellenbach, Tom Jarrett, Lissa Keeleley, Claus Leitherer, Aiguo Li, Sanjeeva Mathuram, Merlin Meyer, John Moosleykao, Eric Murphy, Michael Regan, Georges Rieke, Marcia Rieke, Helene Roussel, Kartik Sheth, J.D. Smith, Michelle Thornley, Fabian Walter & George Helou.



# ALeRCE

Automatic Learning for the  
Rapid Classification of Events



Statistical inference is a logical framework  
with which to test our beliefs of a noisy world  
against data.

We formalize our beliefs in a probabilistic model.

## 1.1

---

# AXIOMS OF PROBABILITY

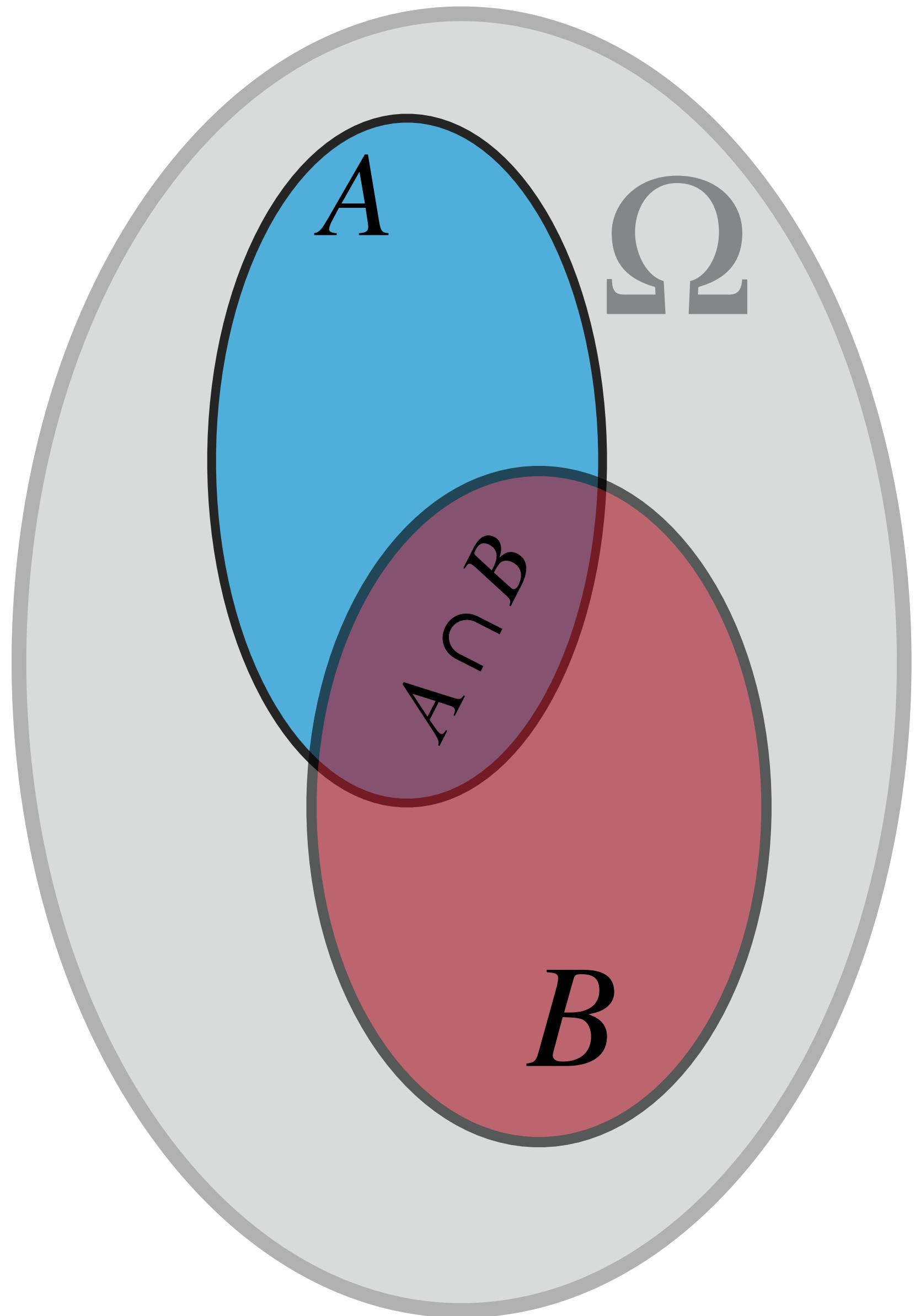
$x$  is a scalar quantity, measured  $N$  times

$x_i$  is a single measurement with  $i = 1, \dots, N$

$x_i$  refers to the set of all  $N$  measurements

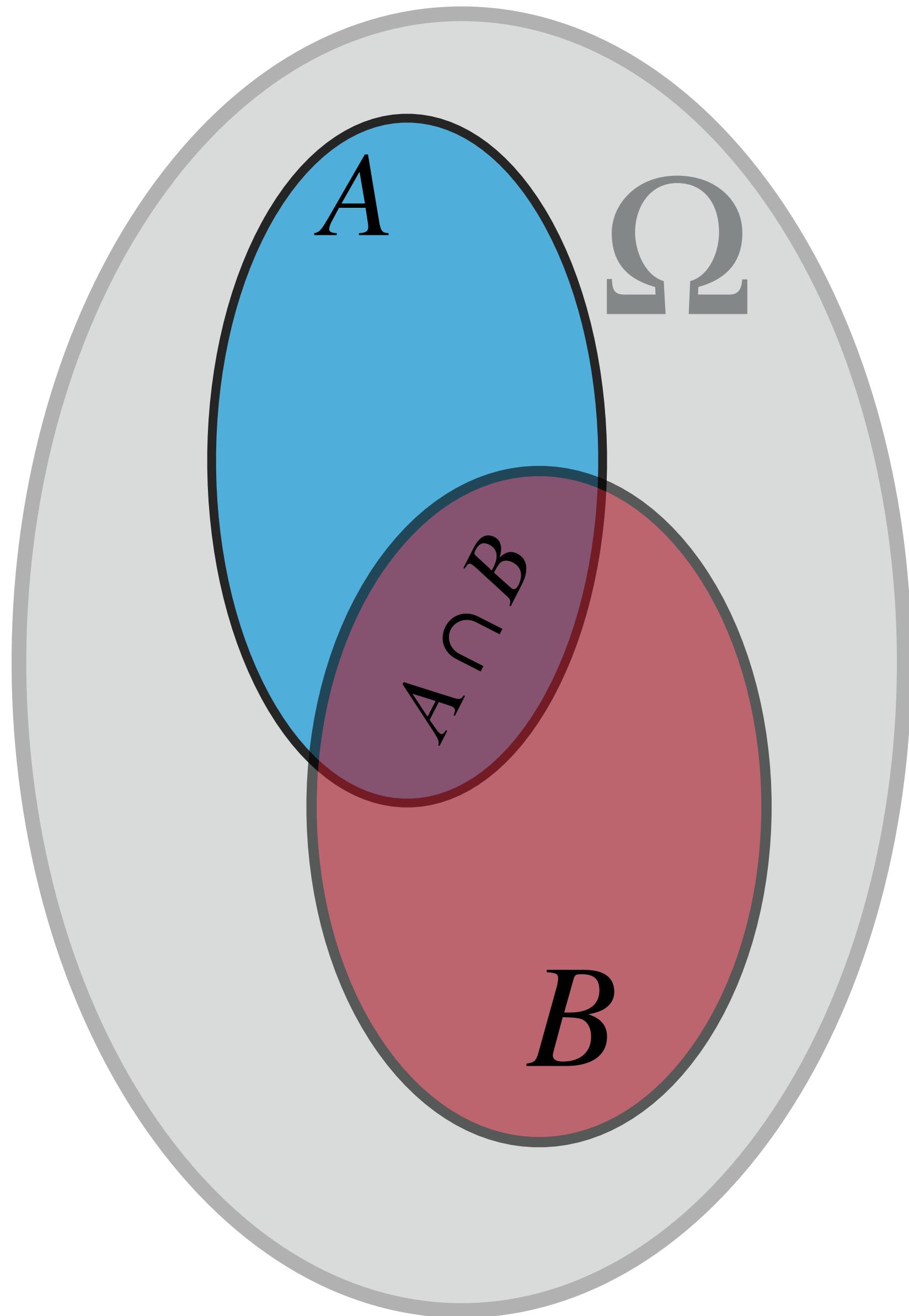
Let  $\Omega$  be a collection of possible elementary events, and  $A$  and  $B$  events such that  $A, B \in \Omega$

- ▶  $P(A) \geq 0$  for all  $A$ ;
- ▶  $P(\Omega) = 1$ ;
- ▶  $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$   
for all countable disjoint sets  $A_1, A_2 \dots \in \Omega$



Let  $\Omega$  be a collection of possible elementary events, and  $A$  and  $B$  events such that  $A, B \in \Omega$

- ▶  $P(A) \geq 0$  for all  $A$ ;
- ▶  $P(\Omega) = 1$ ;
- ▶  $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$   
for all countable disjoint sets  $A_1, A_2 \dots \in \Omega$
- ▶ Some consequences:  $P(A) + P(A^C) = 1$

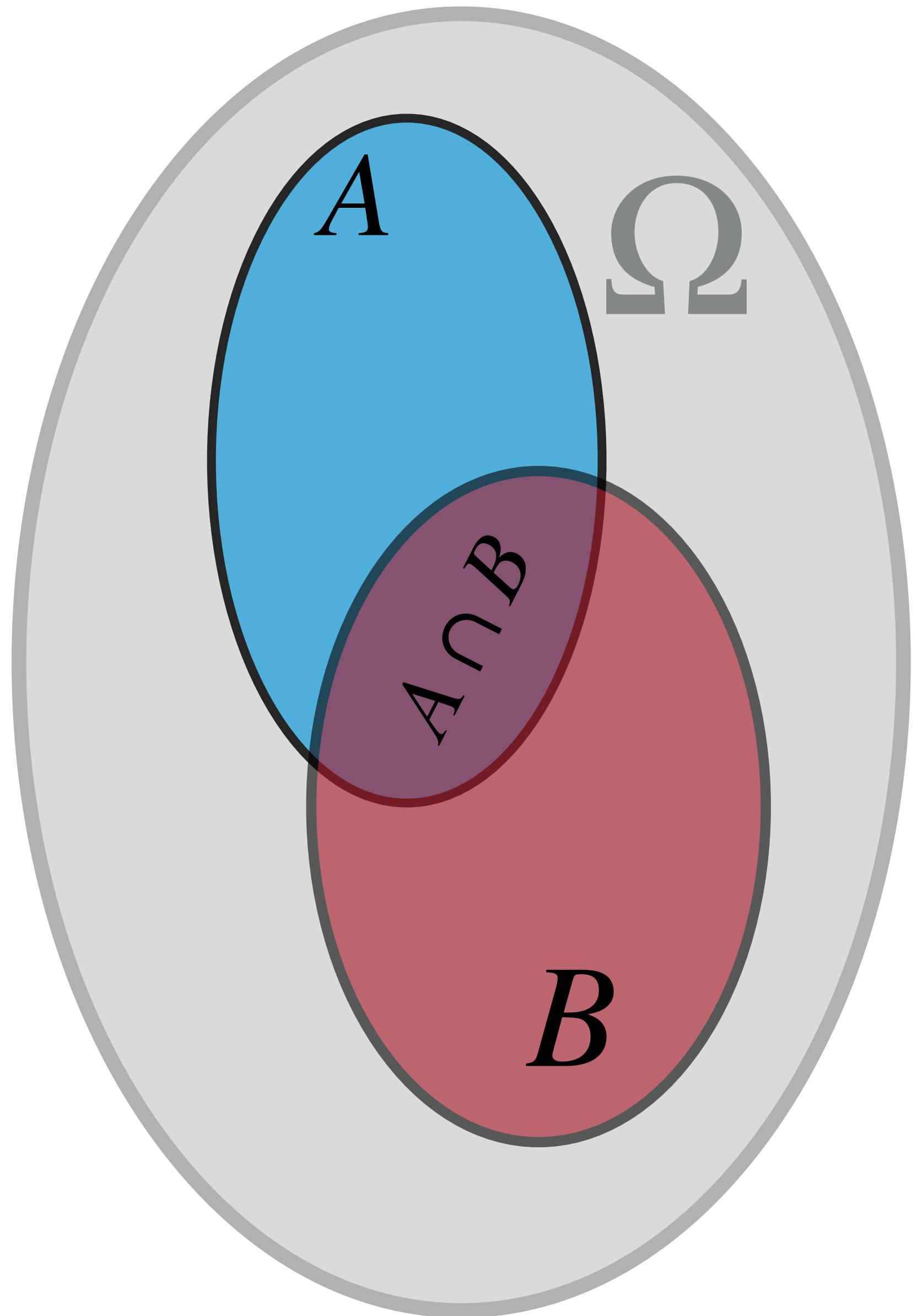


Largely because a 100 years is not enough time for people to agree on very much, there's a few different notations used in the readings.

$$P(A \cap B) = P(A, B) = P(A \text{ and } B)$$

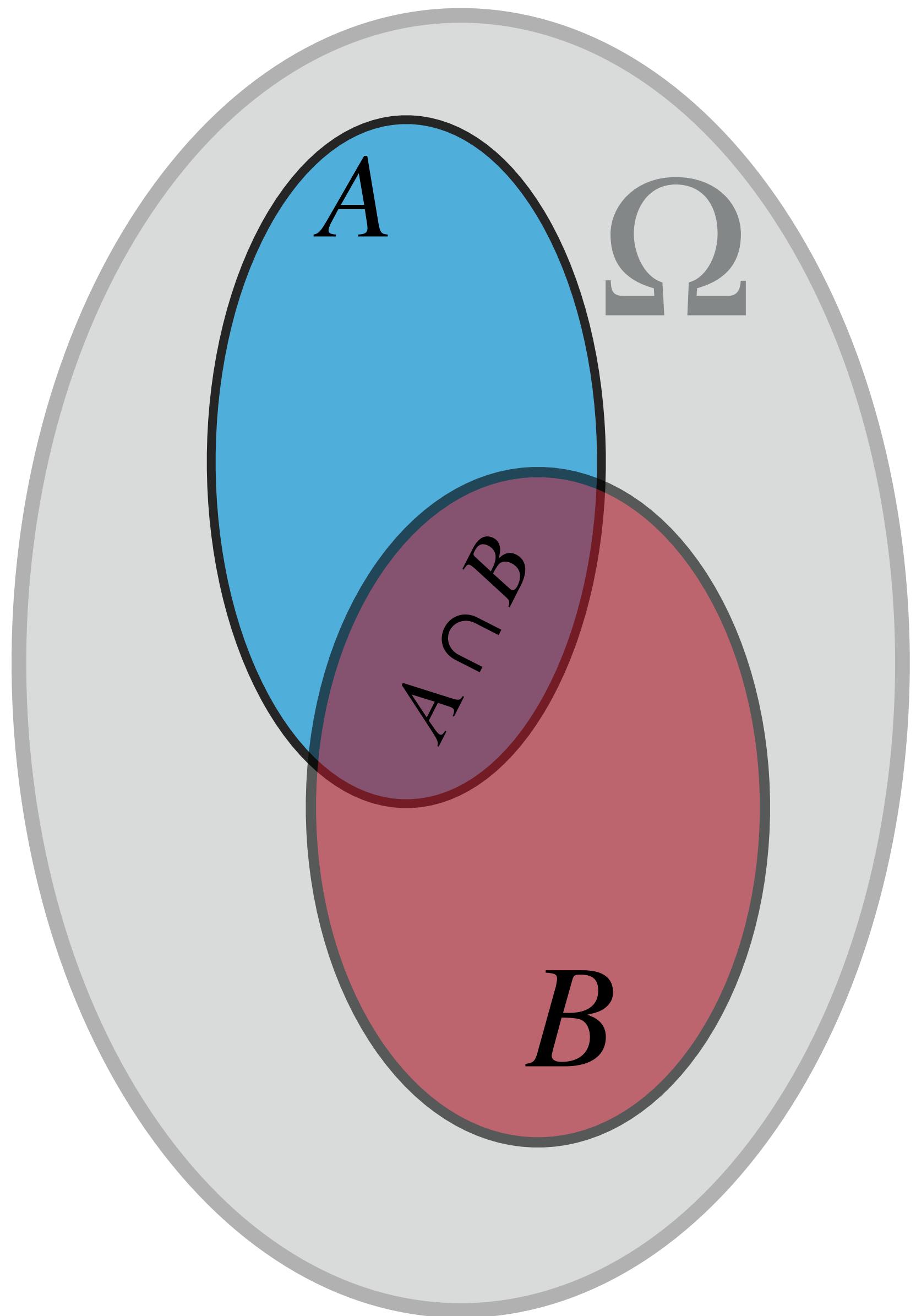
Let  $\Omega$  be a collection of possible elementary events, and  $A$  and  $B$  events such that  $A, B \in \Omega$

- ▶  $P(A) \geq 0$  for all  $A$ ;
- ▶  $P(\Omega) = 1$ ;
- ▶ 
$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$
for all countable disjoint sets  $A_1, A_2 \dots \in \Omega$
- ▶  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$



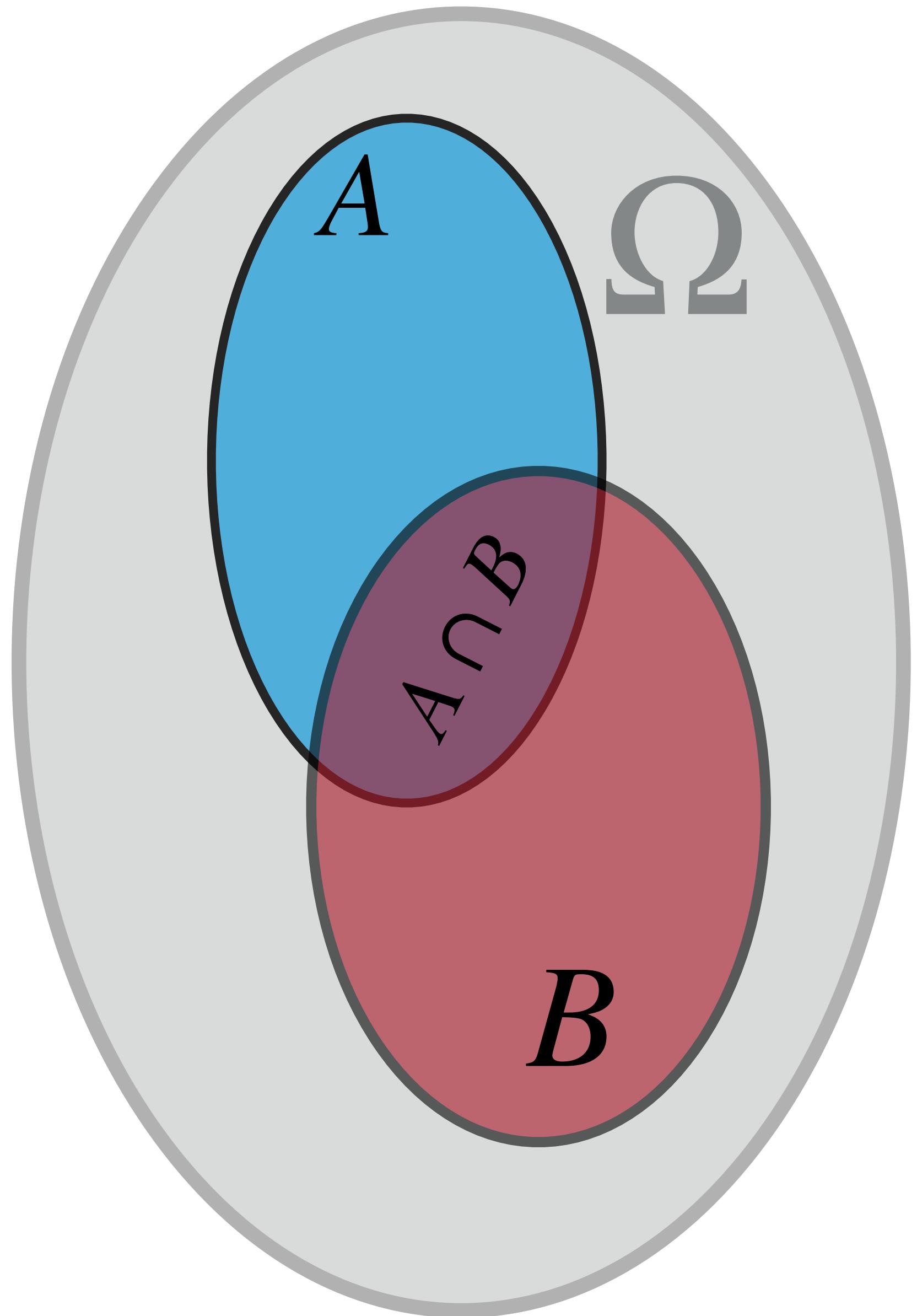
Let  $\Omega$  be a collection of possible elementary events, and  $A$  and  $B$  events such that  $A, B \in \Omega$

- ▶  $P(A) \geq 0$  for all  $A$ ;
- ▶  $P(\Omega) = 1$ ;
- ▶  $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$   
for all countable disjoint sets  $A_1, A_2 \dots \in \Omega$
- ▶  $P(A | B) = \frac{P(A \cap B)}{P(B)}$       Conditional Probability



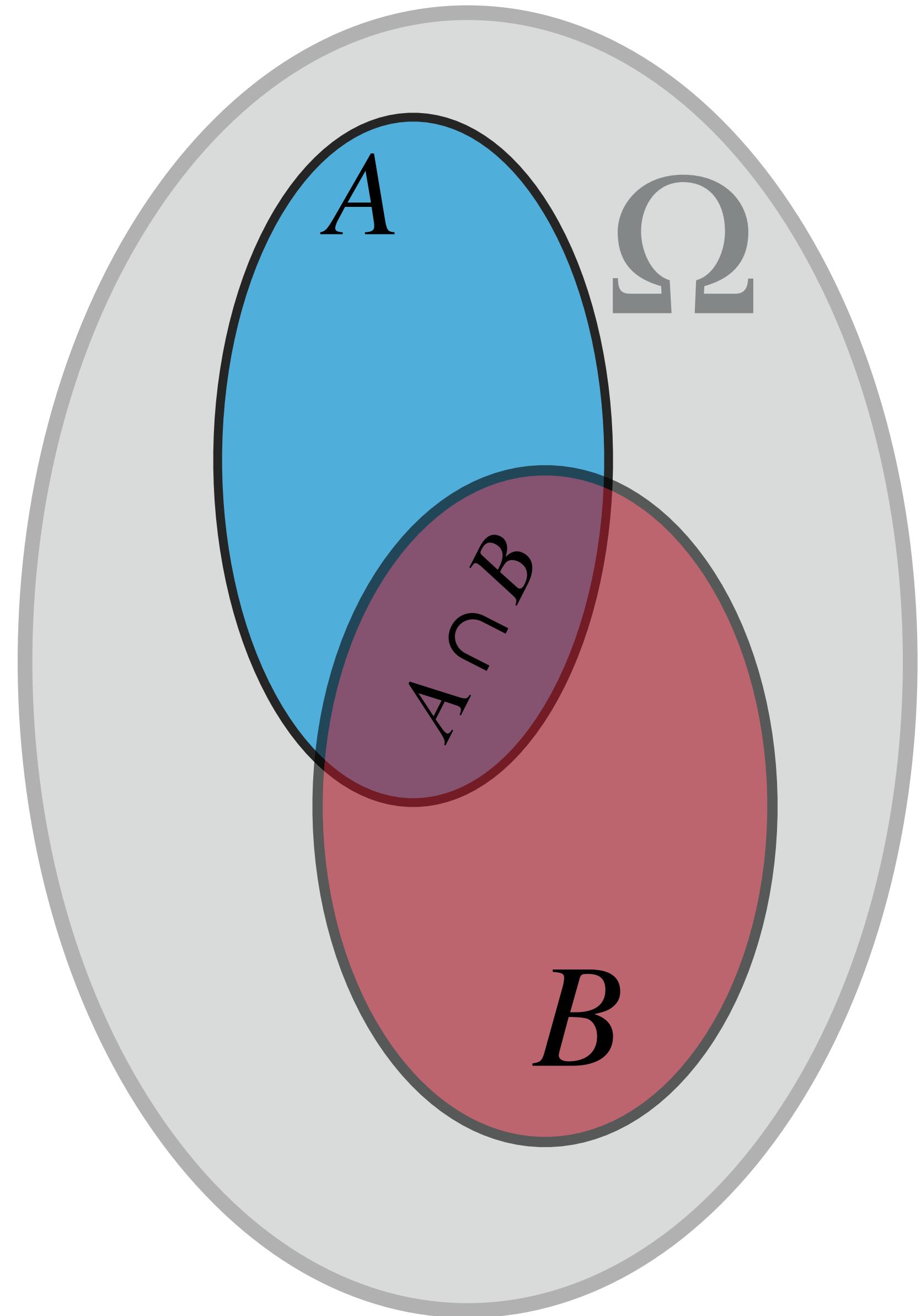
Let  $\Omega$  be a collection of possible elementary events, and  $A$  and  $B$  events such that  $A, B \in \Omega$

- ▶  $P(A) \geq 0$  for all  $A$ ;
- ▶  $P(\Omega) = 1$ ;
- ▶  $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$   
for all countable disjoint sets  $A_1, A_2 \dots \in \Omega$
- ▶  $P(A | B) \cdot P(B) = P(B | A) \cdot P(A) = P(A \cap B)$



Let  $\Omega$  be a collection of possible elementary events, and  $A$  and  $B$  events such that  $A, B \in \Omega$

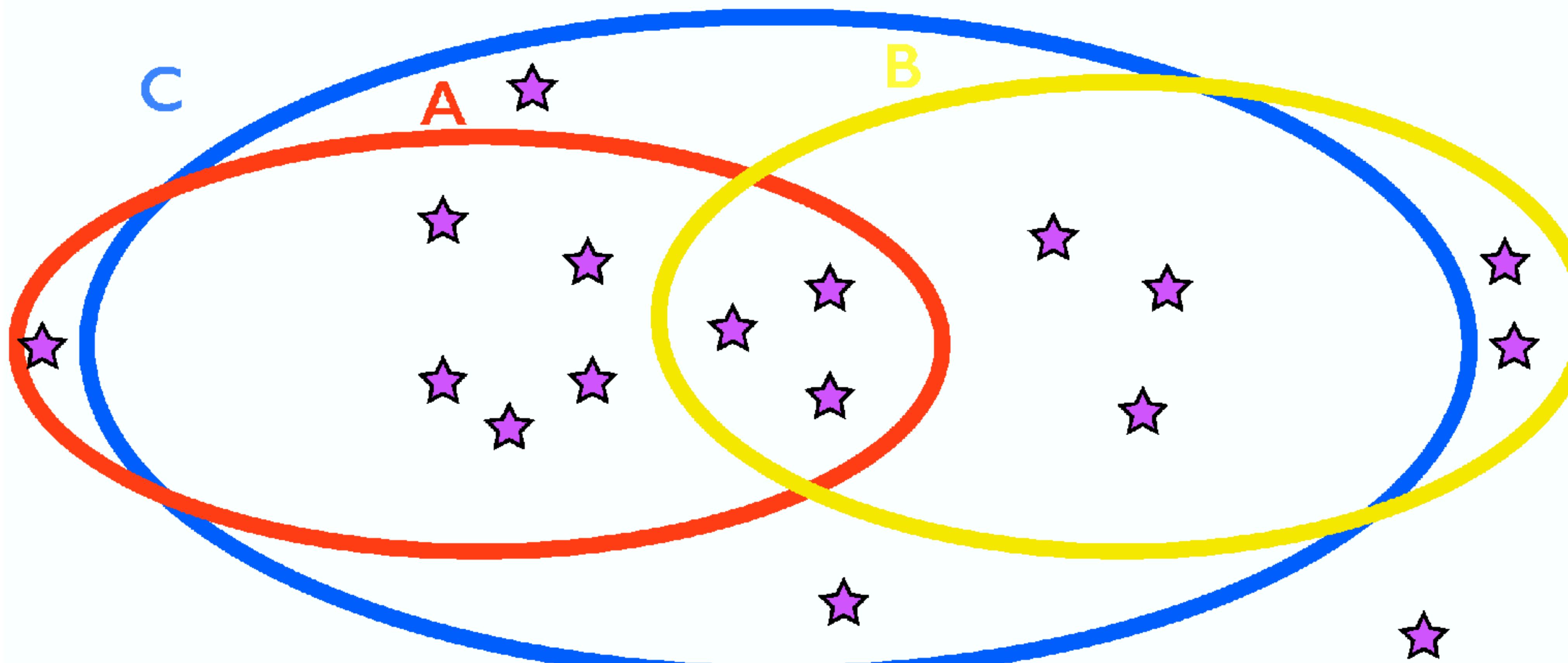
- ▶  $P(A) \geq 0$  for all  $A$ ;
- ▶  $P(\Omega) = 1$ ;
- ▶  $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$   
for all countable disjoint sets  $A_1, A_2 \dots \in \Omega$
- ▶  $P(A) = \sum_{i=i}^{\infty} P(A | B_i) \cdot P(B_i)$



The law of total probability

# Rules of probability

- $P(A \cup B | C) = P(A|C) + P(B|C) - P(A \cap B | C)$
- $11/13 = 8/13 + 6/13 - 3/13$



## A real world example:

Schawinski et al.

“The Green Valley is a Red Herring: GalaxyZoo reveals two evolutionary pathways towards quenching of star formation in early- and late-type galaxies”

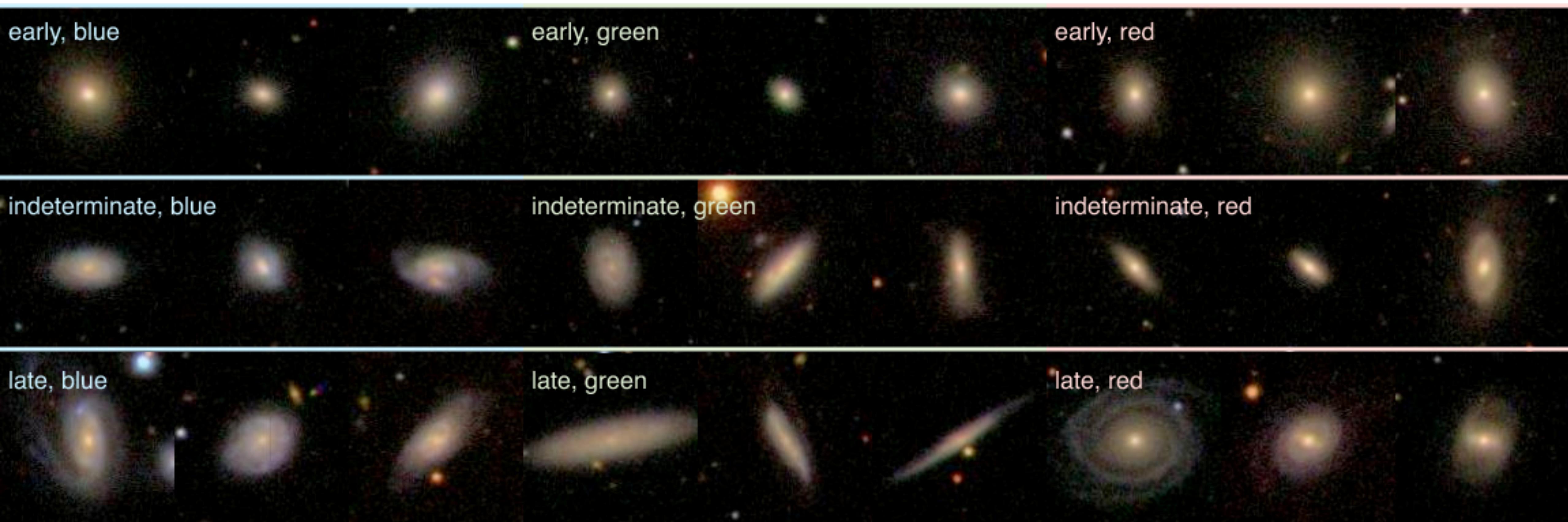
MNRAS 440, no. 1, 889-907 (2014)

arXiv 1402.4814

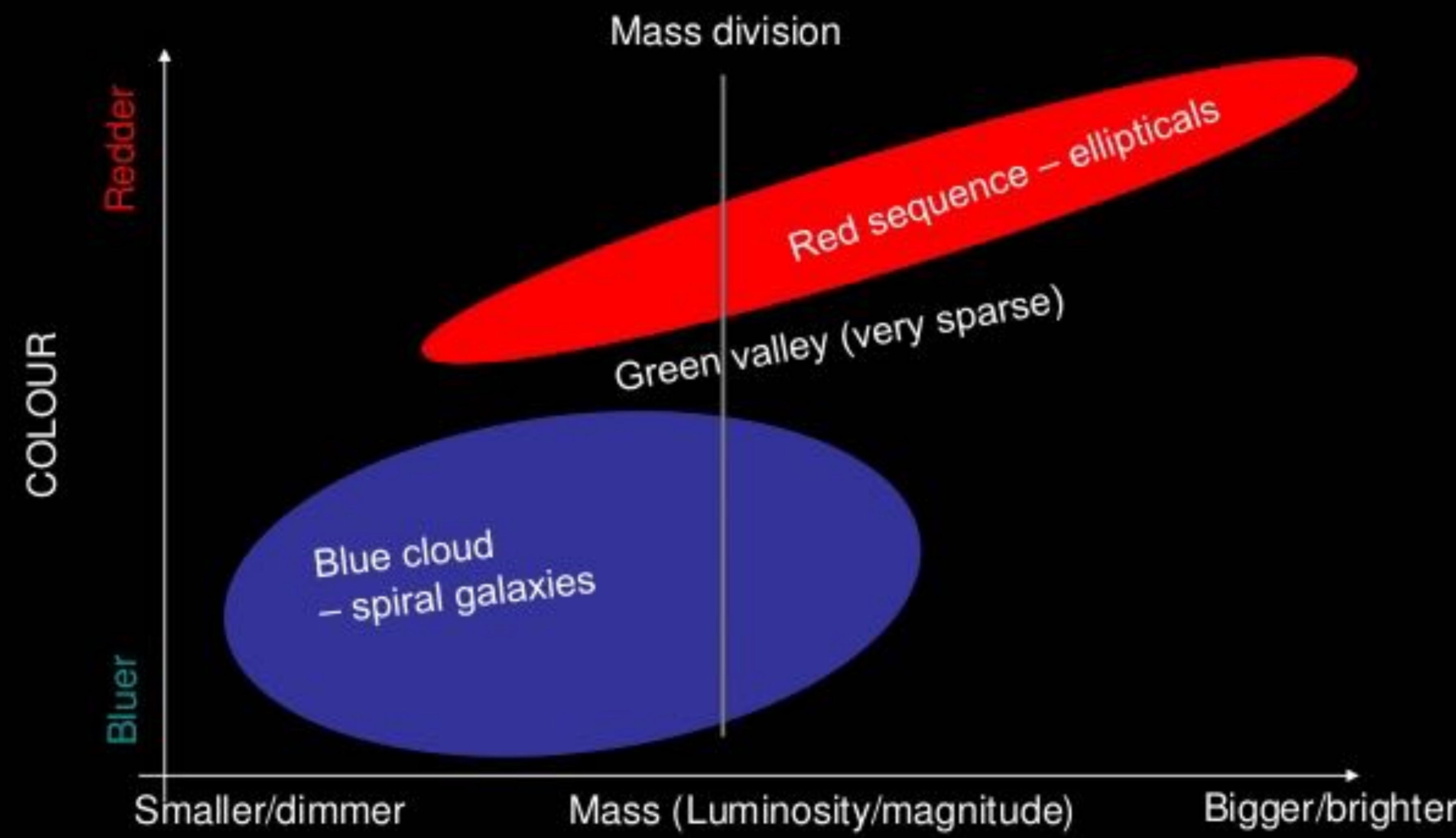
## The Green Valley is a Red Herring: Galaxy Zoo reveals two evolutionary pathways towards quenching of star formation in early- and late-type galaxies

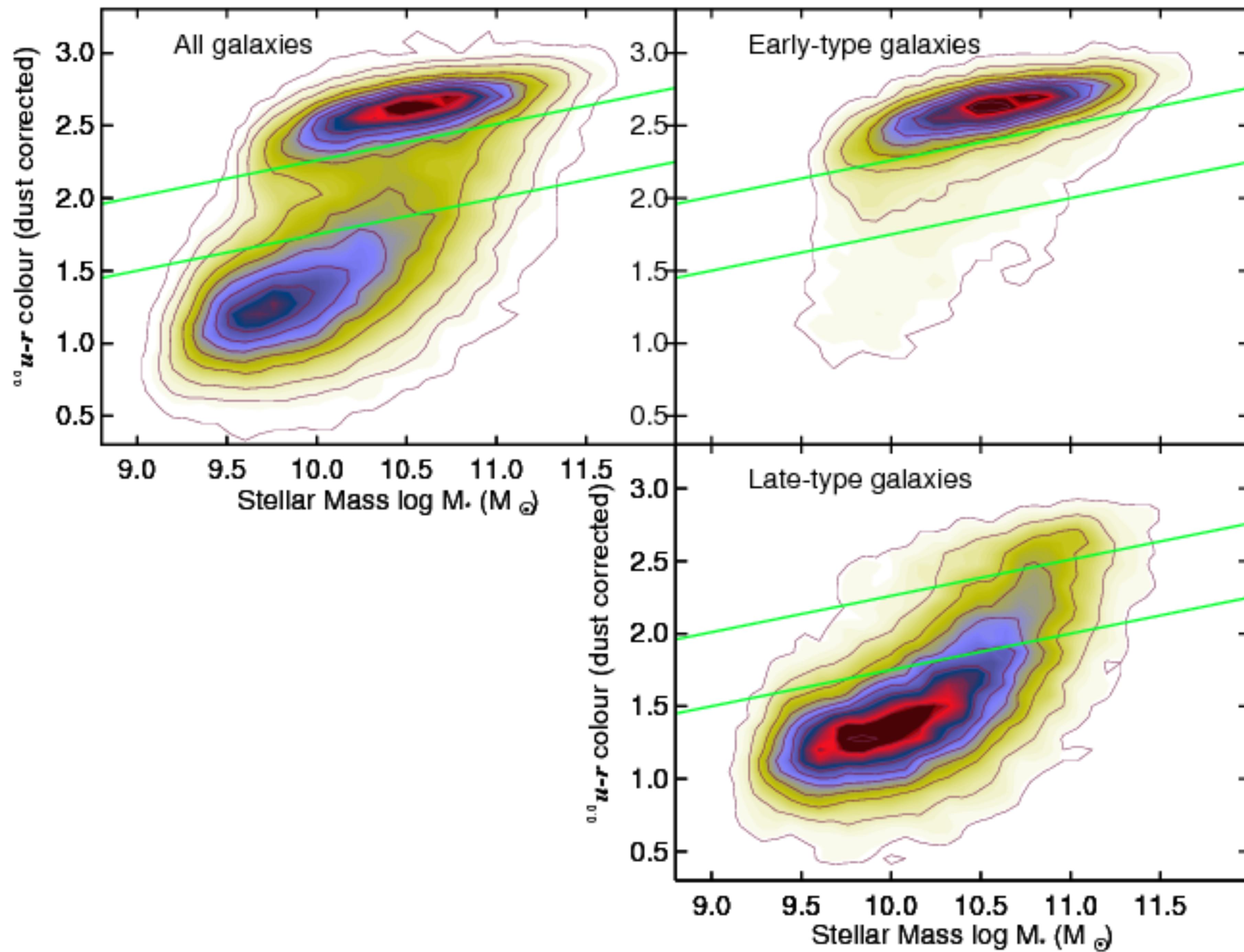
Kevin Schawinski, C. Megan Urry, Brooke D. Simmons, Lucy Fortson, Sugata Kaviraj, William C. Keel, Chris J. Lintott, Karen L. Masters, Robert C. Nichol, Marc Sarzi, Ramin Skibba, Ezequiel Treister, Kyle W. Willett, O. Ivy Wong, Sukyoung K. Yi

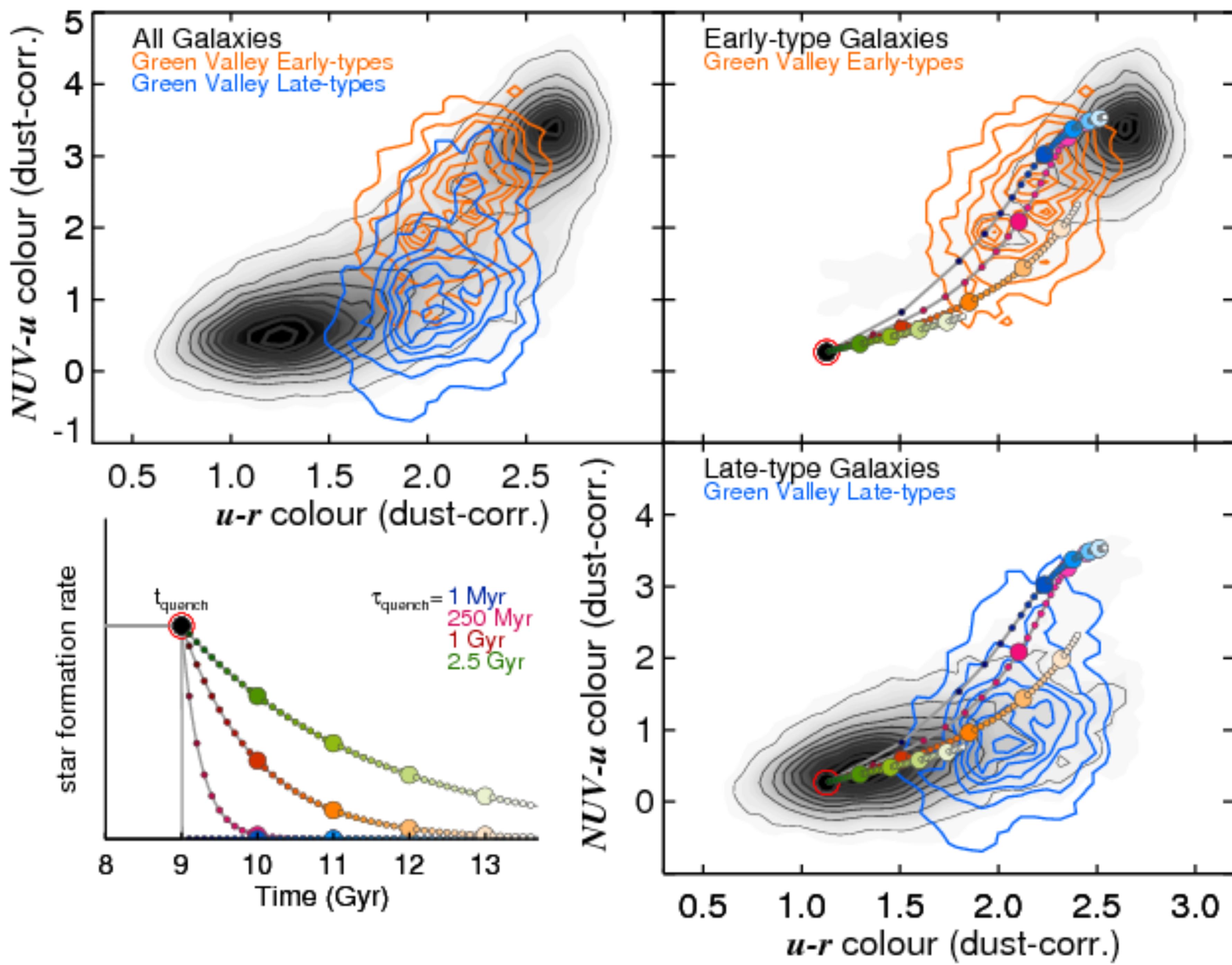
We use SDSS+\textit{GALEX}+Galaxy Zoo data to study the quenching of star formation in low-redshift galaxies. We show that the green valley between the blue cloud of star-forming galaxies and the red sequence of quiescent galaxies in the colour-mass diagram is not a single transitional state through which most blue galaxies evolve into red galaxies. Rather, an analysis that takes morphology into account makes clear that only a small population of blue early-type galaxies move rapidly across the green valley after the morphologies are transformed from disk to spheroid and star formation is quenched rapidly. In contrast, the majority of blue star-forming galaxies have significant disks, and they retain their late-type morphologies as their star formation rates decline very slowly. We summarize a range of observations that lead to these conclusions, including UV-optical colours and halo masses, which both show a striking dependence on morphological type. We interpret these results in terms of the evolution of cosmic gas supply and gas reservoirs. We conclude that late-type galaxies are consistent with a scenario where the cosmic supply of gas is shut off, perhaps at a critical halo mass, followed by a slow exhaustion of the remaining gas over several Gyr, driven by secular and/or environmental processes. In contrast, early-type galaxies require a scenario where the gas supply and gas reservoir are destroyed virtually instantaneously, with rapid quenching accompanied by a morphological transformation from disk to spheroid. This gas reservoir destruction could be the consequence of a major merger, which in most cases transforms galaxies from disk to elliptical morphology, and mergers could play a role in inducing black hole accretion and possibly AGN feedback.



# Colour Magnitude Diagram







## SOME KEY IDEAS

- ▶ All of the data we collect include some degree of randomness
- ▶ Any conclusions we draw must therefore incorporate some notion of uncertainty
- ▶ There is a correct answer i.e. "Truth" - the Universe as we know it exists after all.
  - ▶ Theory gives us a useful model for it. The challenging is evaluating how likely that model is given the data
  - ▶ That "Truth" can itself include some degree of randomness - complex systems often exhibit some *intrinsic dispersion*
- ▶ Data are constants.
  - ▶ Even if they were randomly generated by the Universe, the data that we have already collected are fixed numbers.
- ▶ We describe things we don't know with perfect precision as "random"



## 1.2

---

# RANDOM VARIABLES

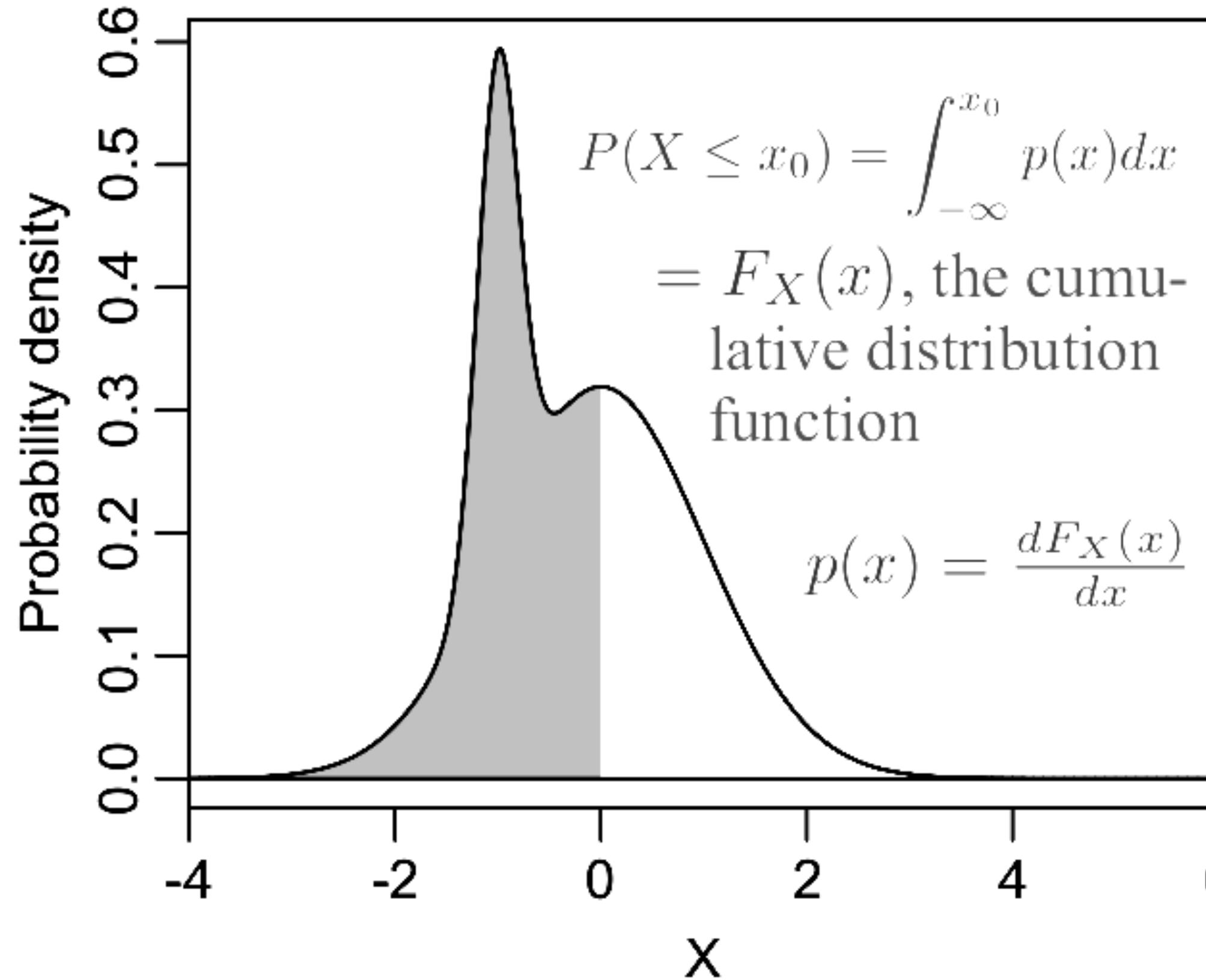
# Random variable:

Modified from Maria Suveges, Laurent Eyer  
31

the outcome of an “experiment”, with a probability for each outcome

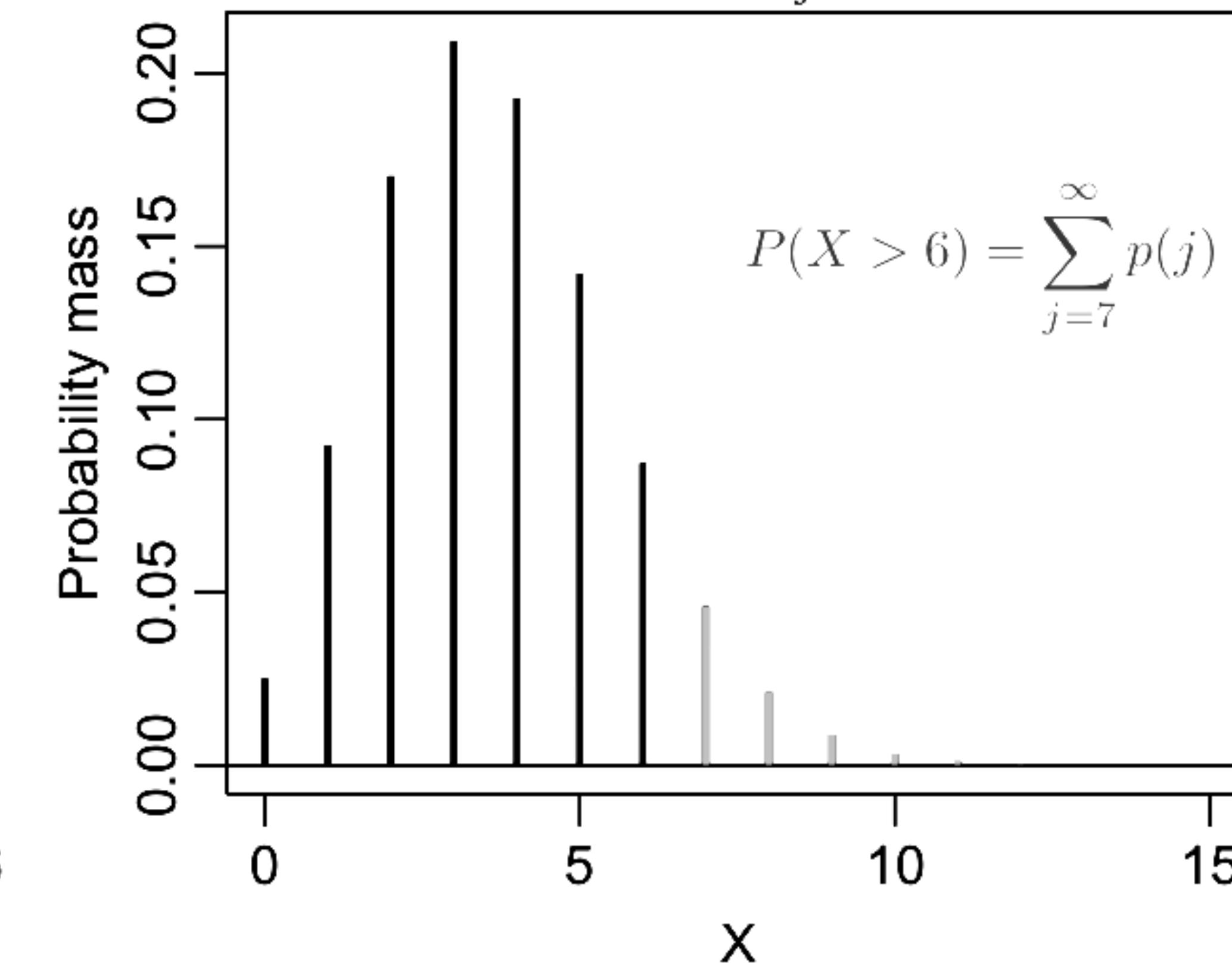
## Continuous

$$P(X \in A) = \int_A p(x)dx$$



## Discrete

$$P(X \in A) = \sum_{x_j \in A} p(x_j)$$



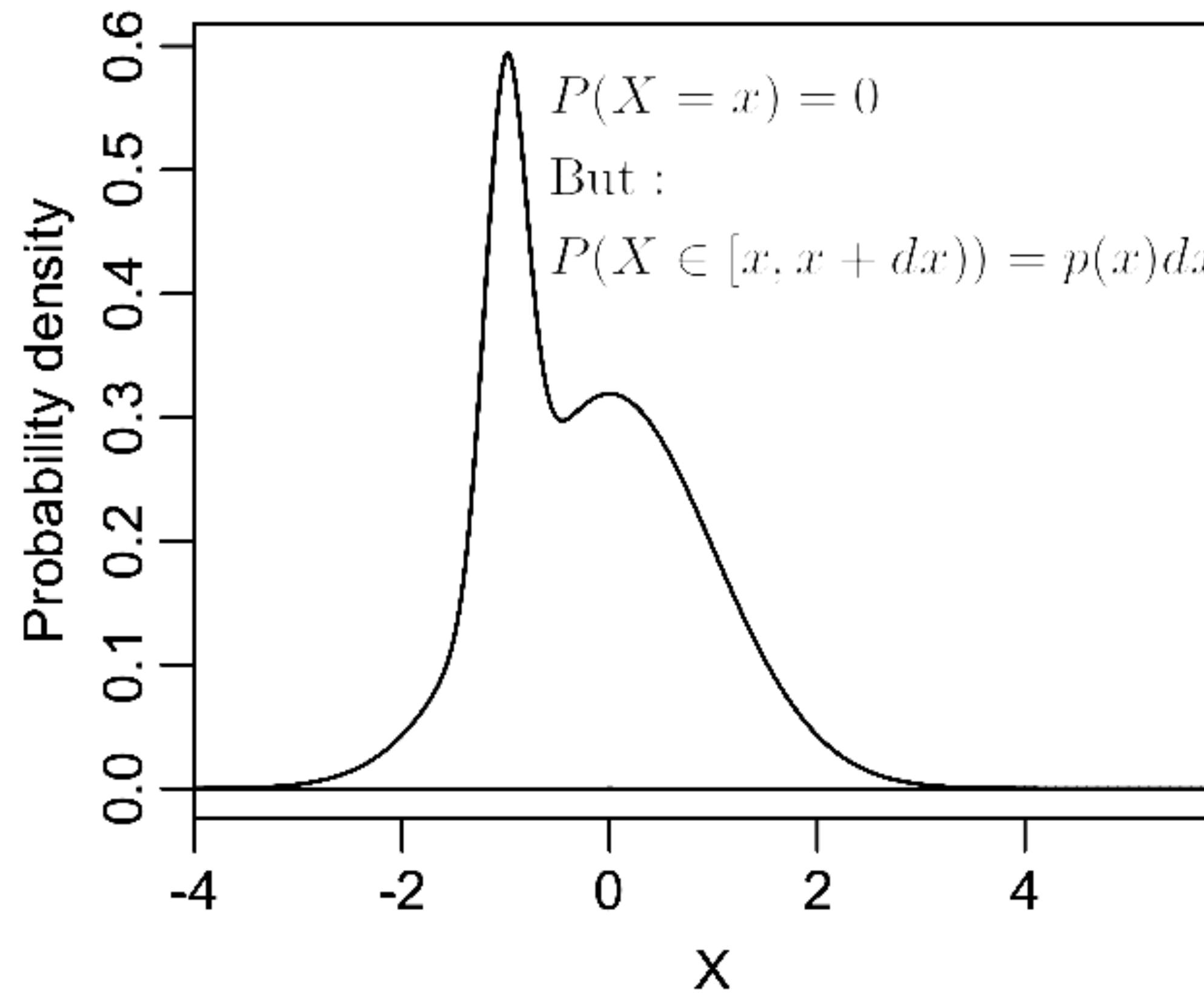
# Random variable:

Modified from Maria Suveges, Laurent Eyer  
32

the outcome of an “experiment”, with a probability for each outcome

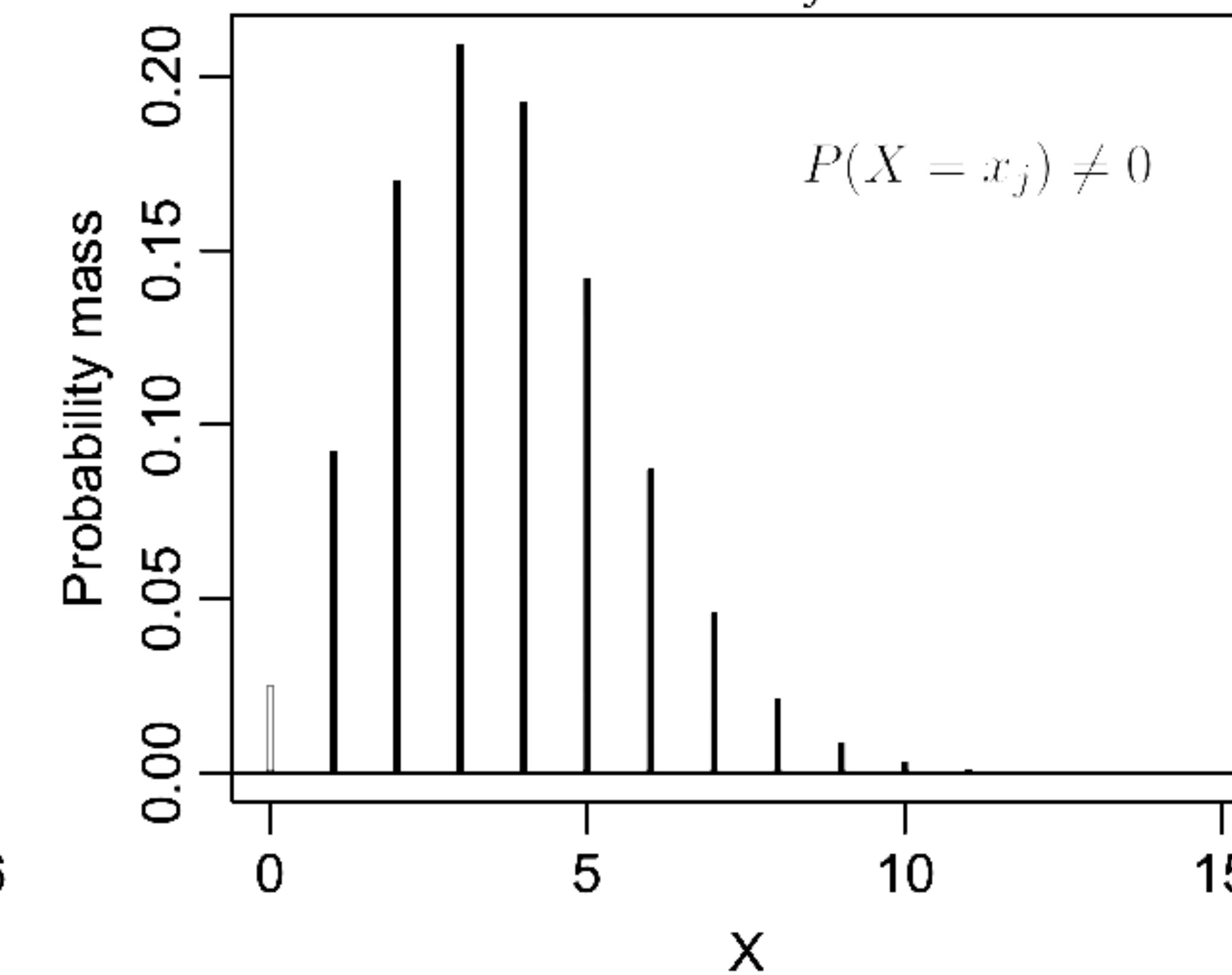
## Continuous

$$P(X \in A) = \int_A p(x)dx$$



## Discrete

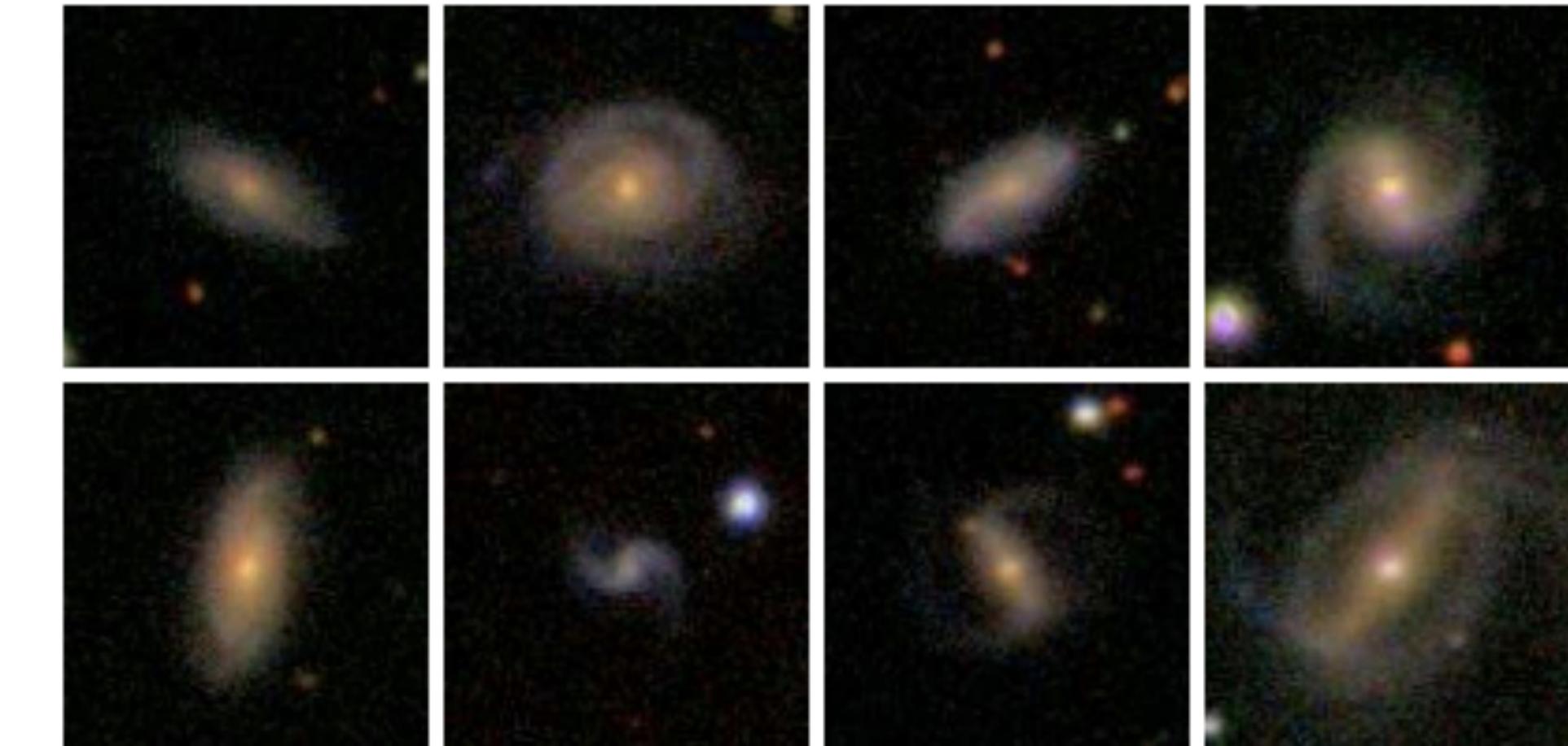
$$P(X \in A) = \sum_{x_j \in A} p(x_j)$$



# Random Variables

- Discrete:

- Spectral type (G2V, KIII)
- Galaxy type, galaxy zoo



Class	Button	Description
1	●	Elliptical galaxy
2	○	Clockwise/Z-wise spiral galaxy
3	○	Anti-clockwise/S-wise spiral galaxy
4	◐	Spiral galaxy other (e.g. edge on, unsure)
5	◆	Star or Don't Know (e.g. artefact)
6	◐	Merger

- Continuous:

- magnitude, flux, colour, radial velocity, parallax/distance, temperature, elemental abundances, magnetic field, age, etc...

We are generally trying to estimate  $p(x)$ ,  
the *true* distribution from which  $x$  is drawn.

$p(x)$  is the “Probability Density Function” of  $x$ .

$p(x) \cdot dx$  is the probability of a value  
lying between  $x$  and  $x + dx$ .

While  $p(x)$  is the true or population pdf, we don't observe it.

We measure the *empirical* pdf,  $f(x)$ .

With  
infinite data  
 $f(x) \rightarrow p(x)$ .

Not really.

So when we say:

Statistical inference is a logical framework  
with which to test our beliefs of a noisy world  
against data.

We formalize our beliefs in a probabilistic model.

Estimate  $f(x)$  from (possibly multi-D) data.

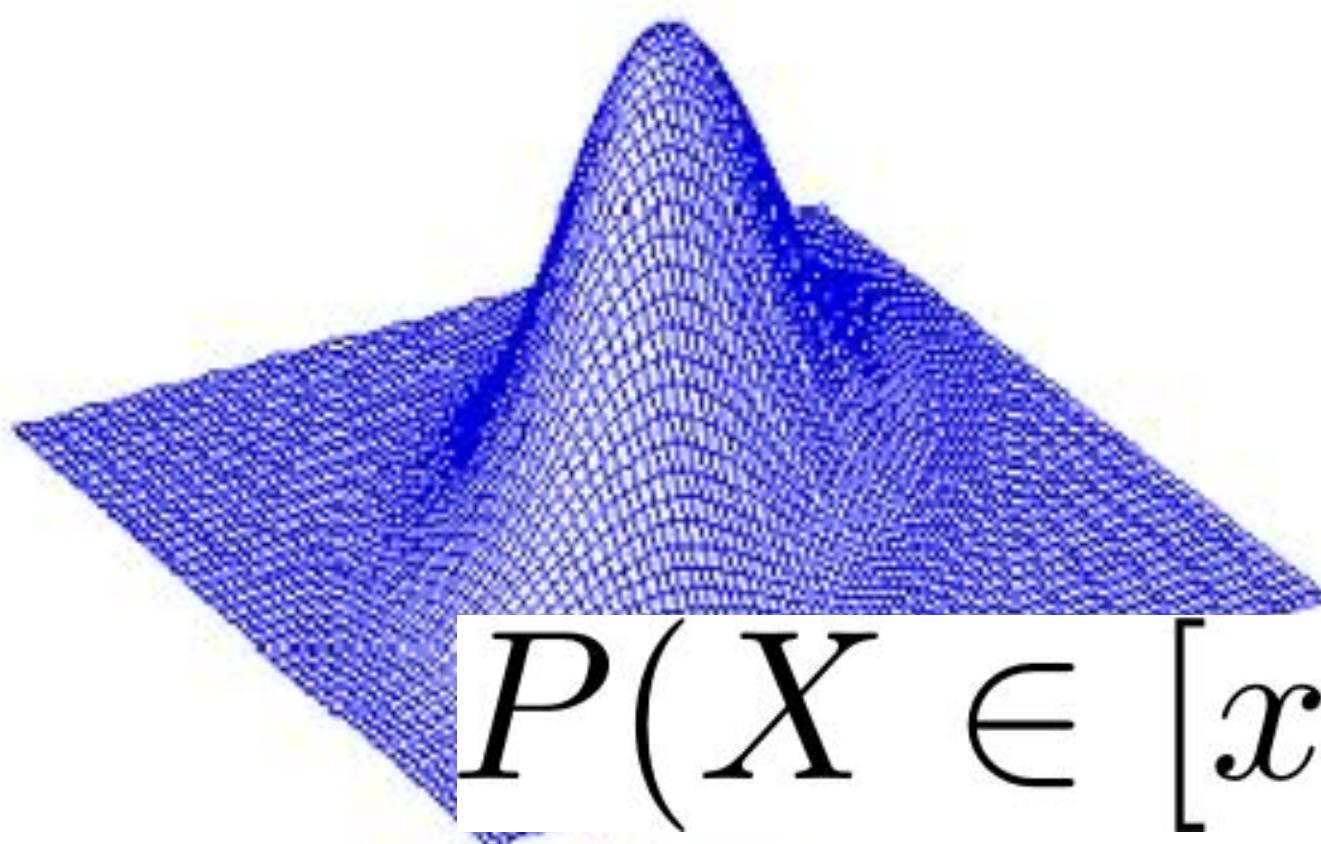
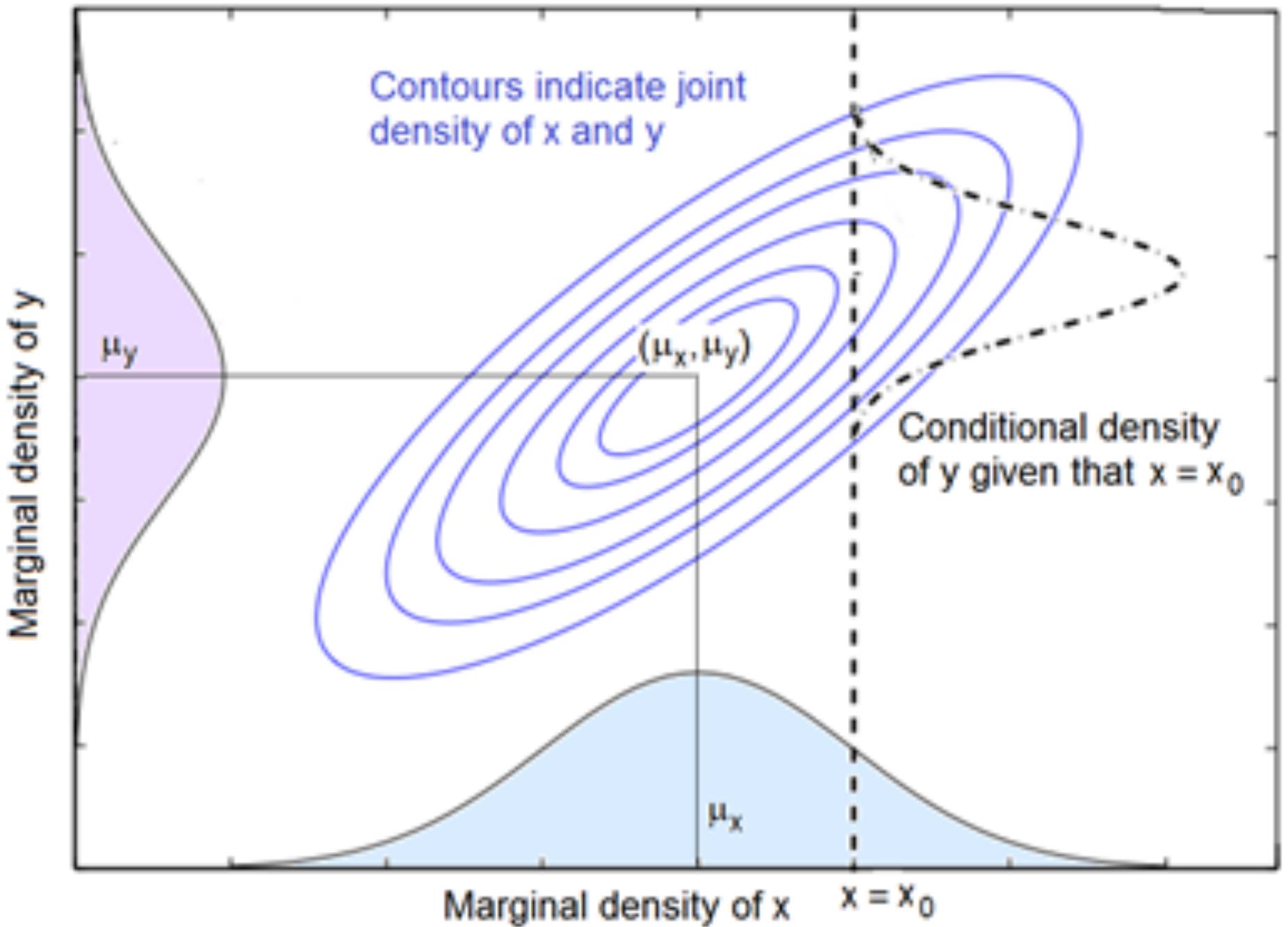
Describe  $f(x)$  and its uncertainty.

Compare it to models of  $p(x)$

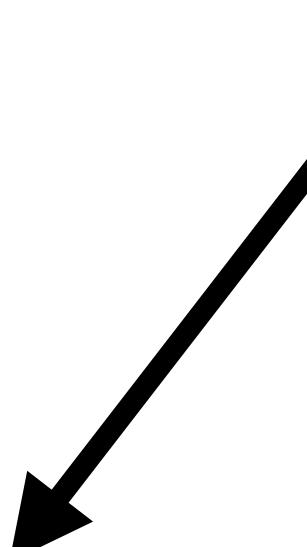
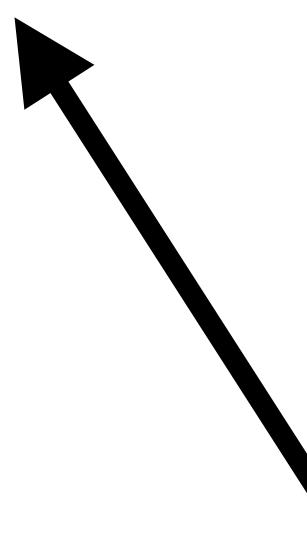
and use the knowledge gained  
to interpret new measurements

$$P(X \in A) = \int_A p(x_1, x_2, \dots, x_N) dx_1 dx_2 \dots dx_N^{38}$$

## THE MULTIVARIATE CASE

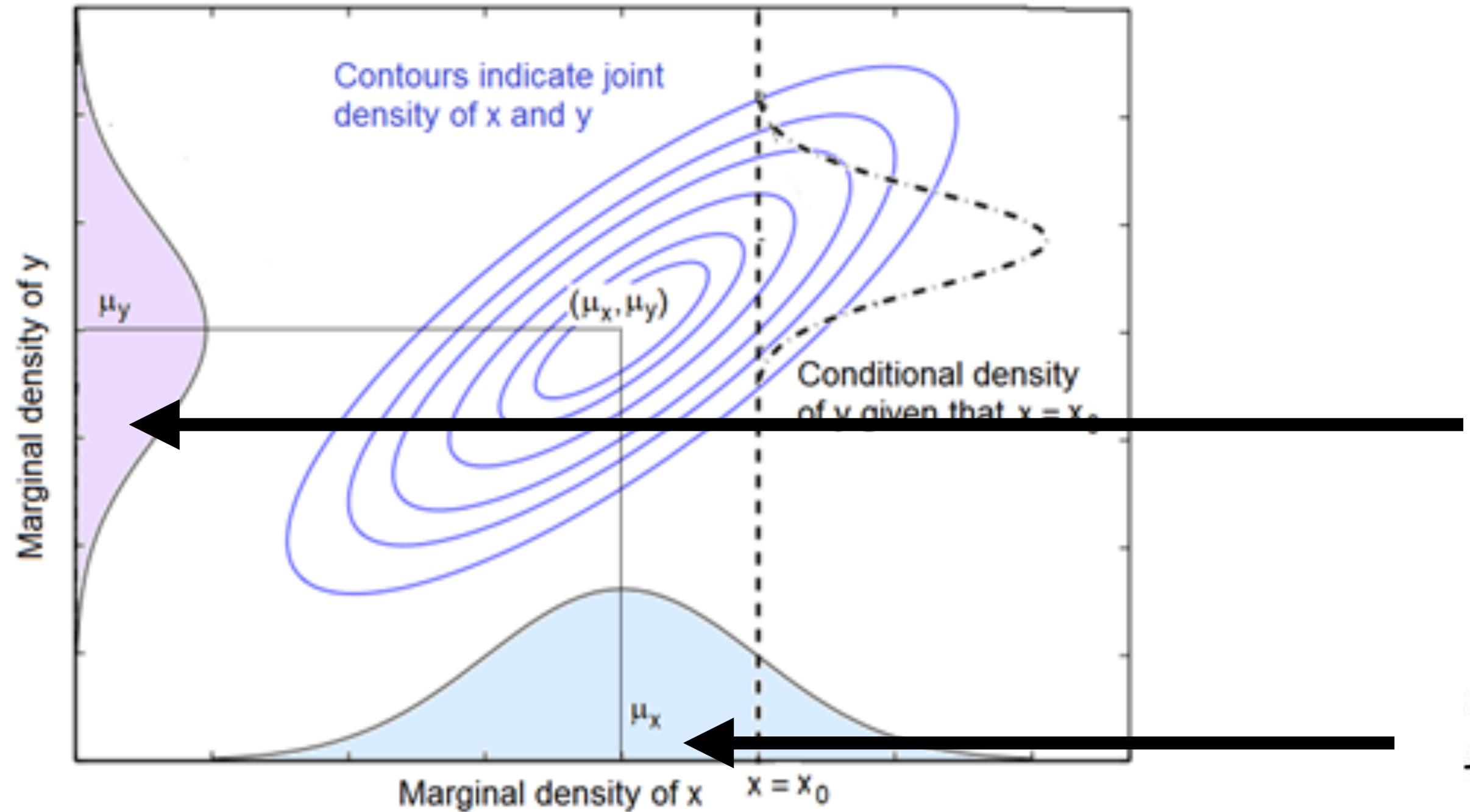


Given this in 1D,  
the probability of  $X$   
in a 2D box of area  $dx dy$  is:



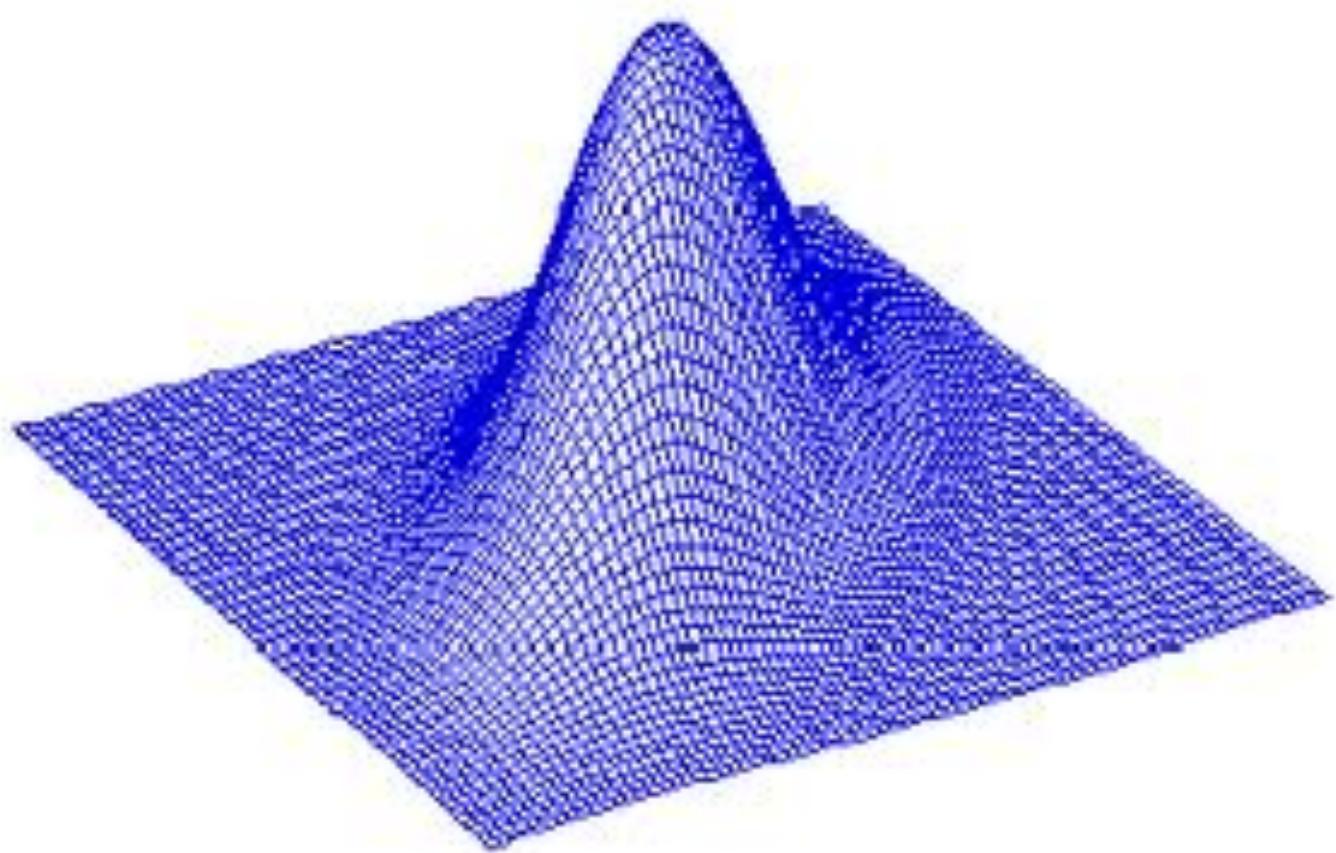
$$P(X \in [x, x + dx] \times [y, y + dy]) = p(x, y) dx dy$$

# THE MULTIVARIATE CASE



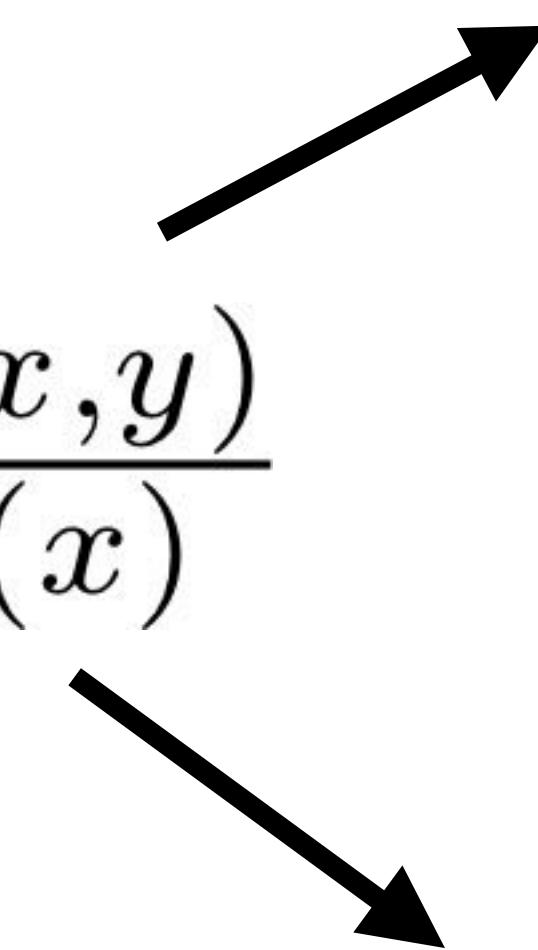
$$p(y) = \int_X p(x, y) dx$$

$$p(x) = \int_Y p(x, y) dy$$



Conditional Probability

$$p(y|x) = \frac{p(x,y)}{p(x)}$$



$$p(x,y) = p(x|y)p(y)$$

$$p(x) = \int_y p(x,y)dy$$

$$p(x) = \int_y p(x|y)p(y)dy$$

The law of total probability

Bayes Rule:

$$p(y|x) = \frac{p(x|y)p(y)}{\int_Y p(x|y)p(y)dy}$$

$$p(y|x) = \frac{p(x|y)p(y)}{\int_Y p(x|y)p(y)dy}$$

**Posterior**

How probable is the hypothesis given the data we observed

**Likelihood**

How probable is the data given the hypothesis is true

**Prior**

How probable was the hypothesis before we observed anything

$$p(\text{Hypothesis}|\text{Data}) = \frac{p(\text{Data}|\text{Hypothesis})p(\text{Hypothesis})}{p(\text{Data})}$$

**Evidence**

How probable is the data over all possible hypotheses

1.3

---

# MOMENTS AND DISTRIBUTIONS

$$E(x) = \langle x \rangle = \int_X x \cdot p(x) dx$$

**Expected Value**

$$E(f(x)) = \langle f(x) \rangle = \int_X f(x) \cdot p(x) dx$$

**Variance**

$$\text{Var}(x) = E([x - \langle x \rangle]^2)$$

**nth moment (non-central)**

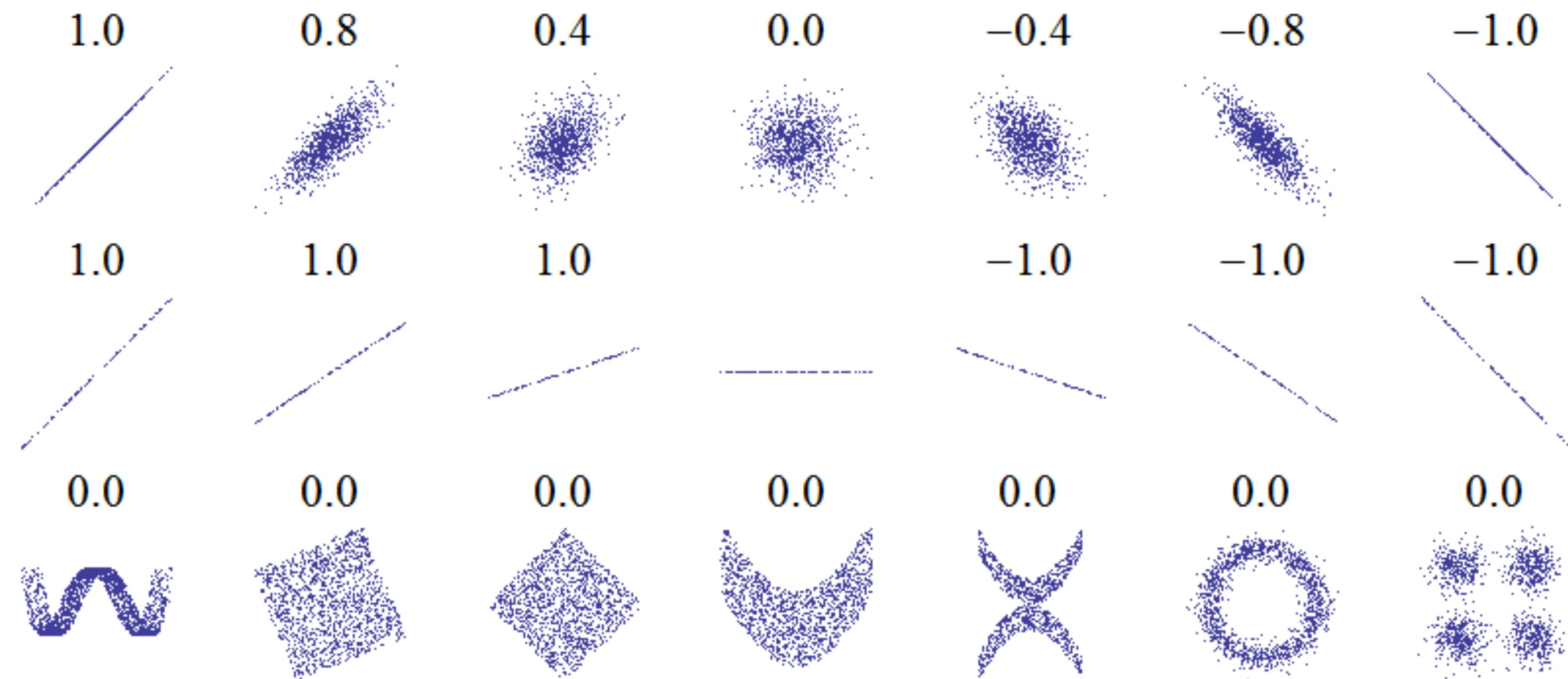
$$\mu_n(x) = E(x^n)$$

**nth moment (central)**

$$\tilde{\mu}_n(x) = E([x - \langle x \rangle]^n)$$

$$\text{Cov}(x, y) = E([x - \langle x \rangle])E([y - \langle y \rangle])$$

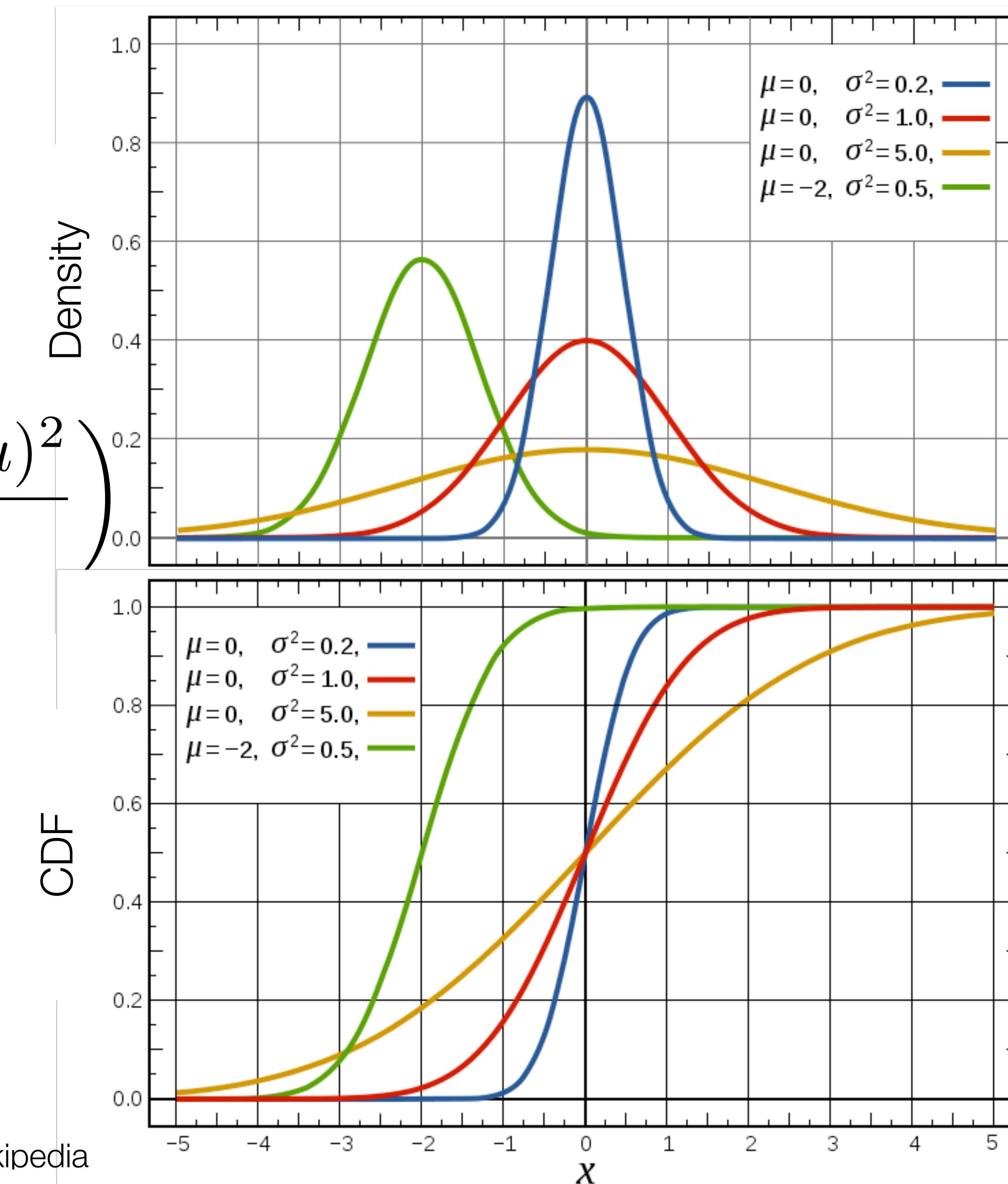
$$\text{Cor}(x, y) = \frac{\text{Cov}(x, y)}{\sqrt{\text{Var}(x)\text{Var}(y)}}$$



# Example: Gaussian / Normal distribution

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$$



from wikipedia

# Poisson distribution

46

Discrete probability distribution (no density)

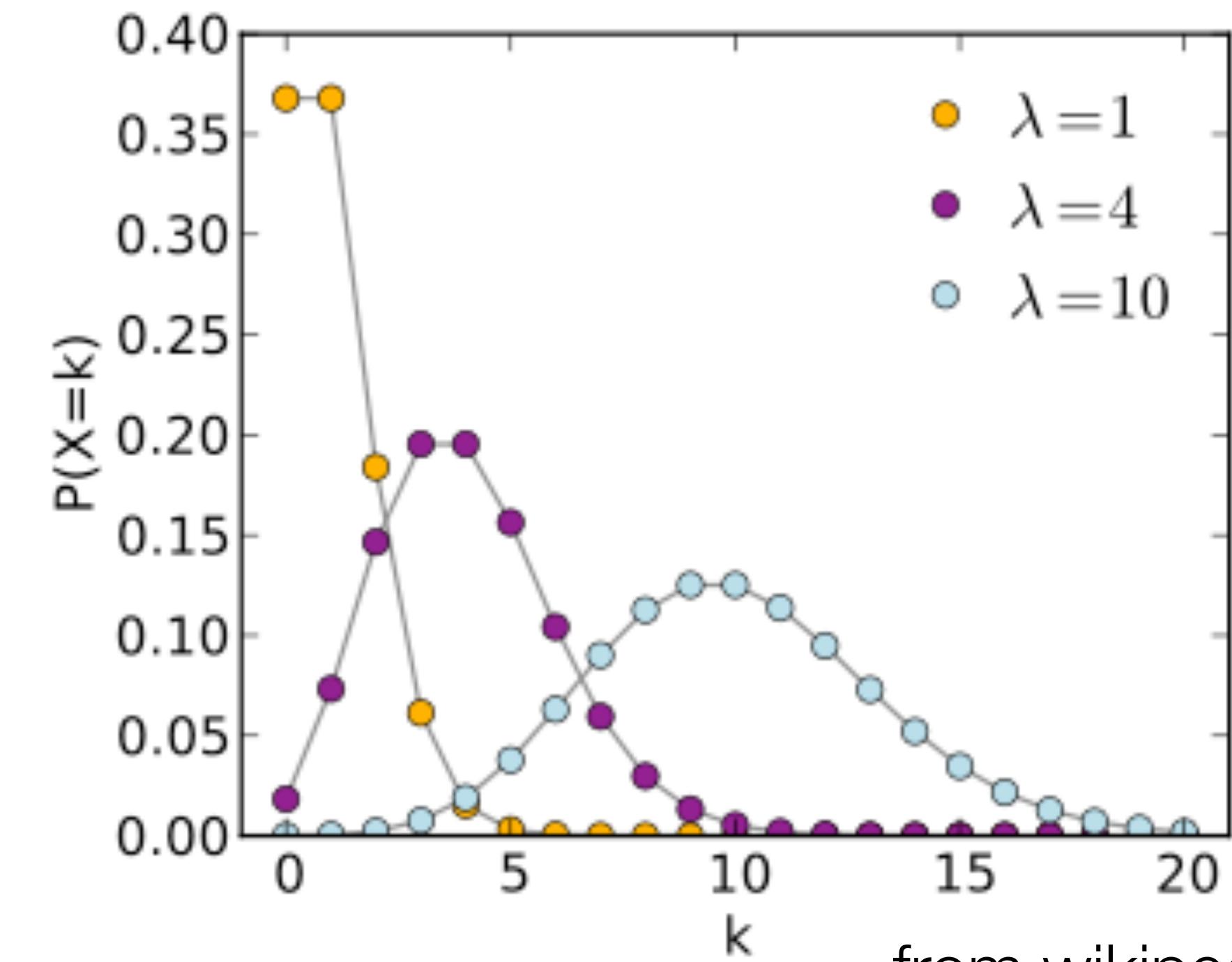
$$X \sim \text{Poisson}(\lambda)$$

Number of photons on a detector  
Number of people in a shop

$$\Pr(X = k) = \exp(-\lambda) \frac{\lambda^k}{k!}$$

For large  $\lambda$

$$\mathcal{N}(\lambda, \lambda)$$



from wikipedia

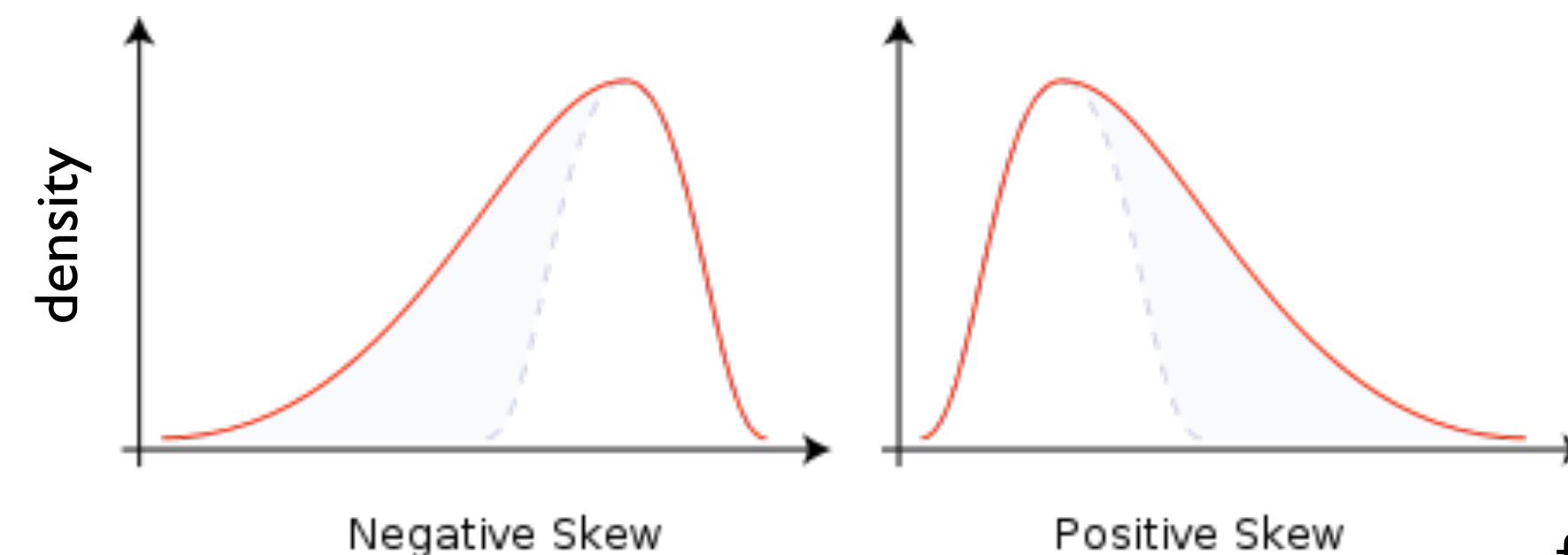
# 3d and 4th moments of a distribution

---

- Skewness, asymmetry

$$\mu_3/\sigma^3 = \int_{-\infty}^{\infty} (x - \mu)^3 f(x) dx / \sigma^3$$

**Normal: 0**  
**Poisson:  $1/\sqrt{\lambda}$**



from wikipedia

- Kurtosis

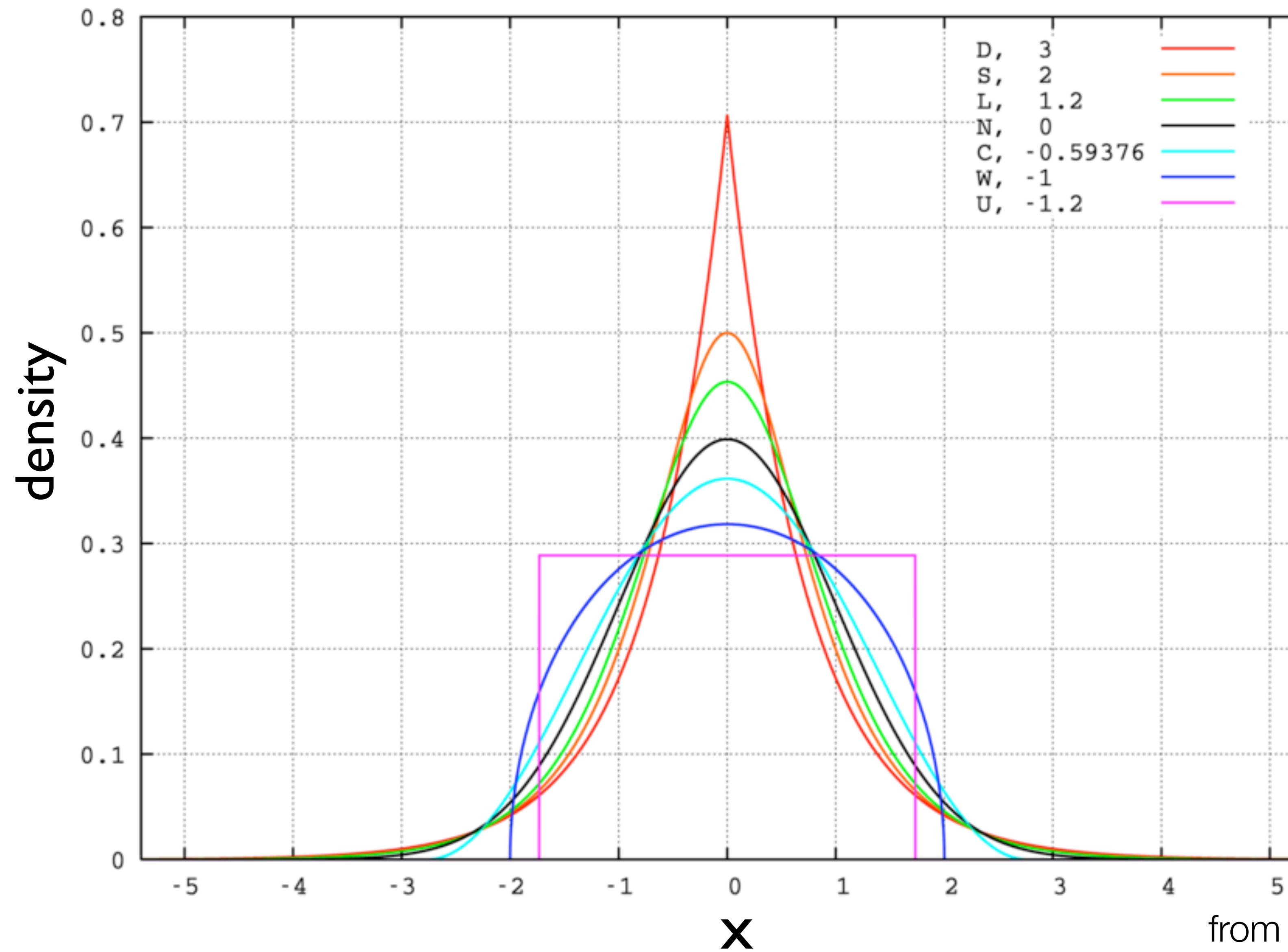


$$\mu_4/\sigma^4 = \int_{-\infty}^{\infty} (x - \mu)^4 f(x) dx / \sigma^4$$

$$\mu_4/\sigma^4 - 3$$

**Normal: 0**  
**Poisson:  $1/\lambda$**

# Example of different values of kurtosis: “boxiness” -- tail heaviness



from wikipedia

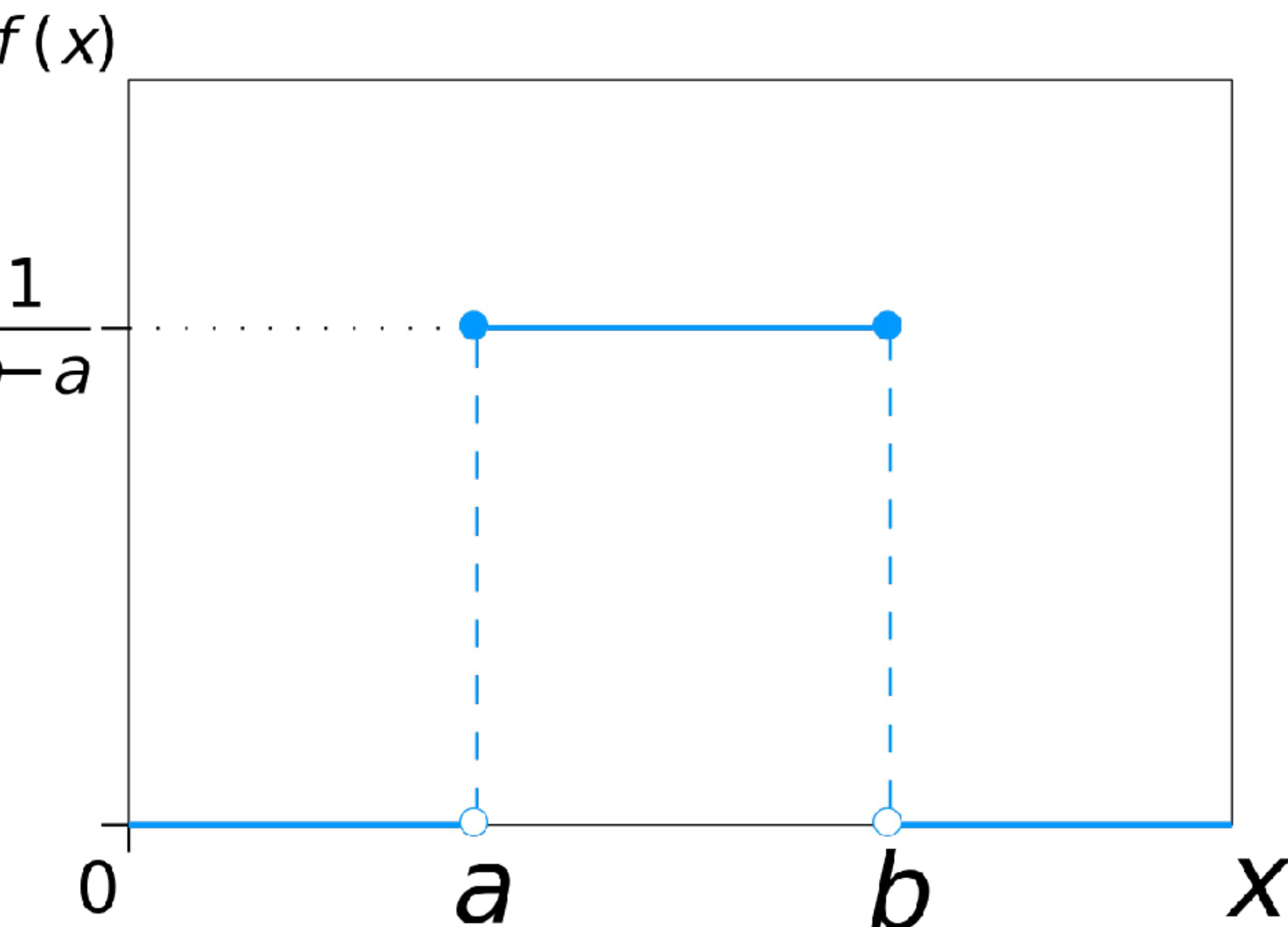
# THE UNIFORM DISTRIBUTION

$$p(x) = \frac{1}{b-a} \text{ for } x \in [a, b] \text{ else } 0$$

Frequently used as a prior when you don't know anything about the hypothesis

Used to generate random values from any distribution via probability integral transform (coming up)

Mean	$\frac{1}{2}(a + b)$
Median	$\frac{1}{2}(a + b)$
Mode	Any value between a and b
Standard deviation	$\sqrt{\frac{1}{12}(b - a)^2}$



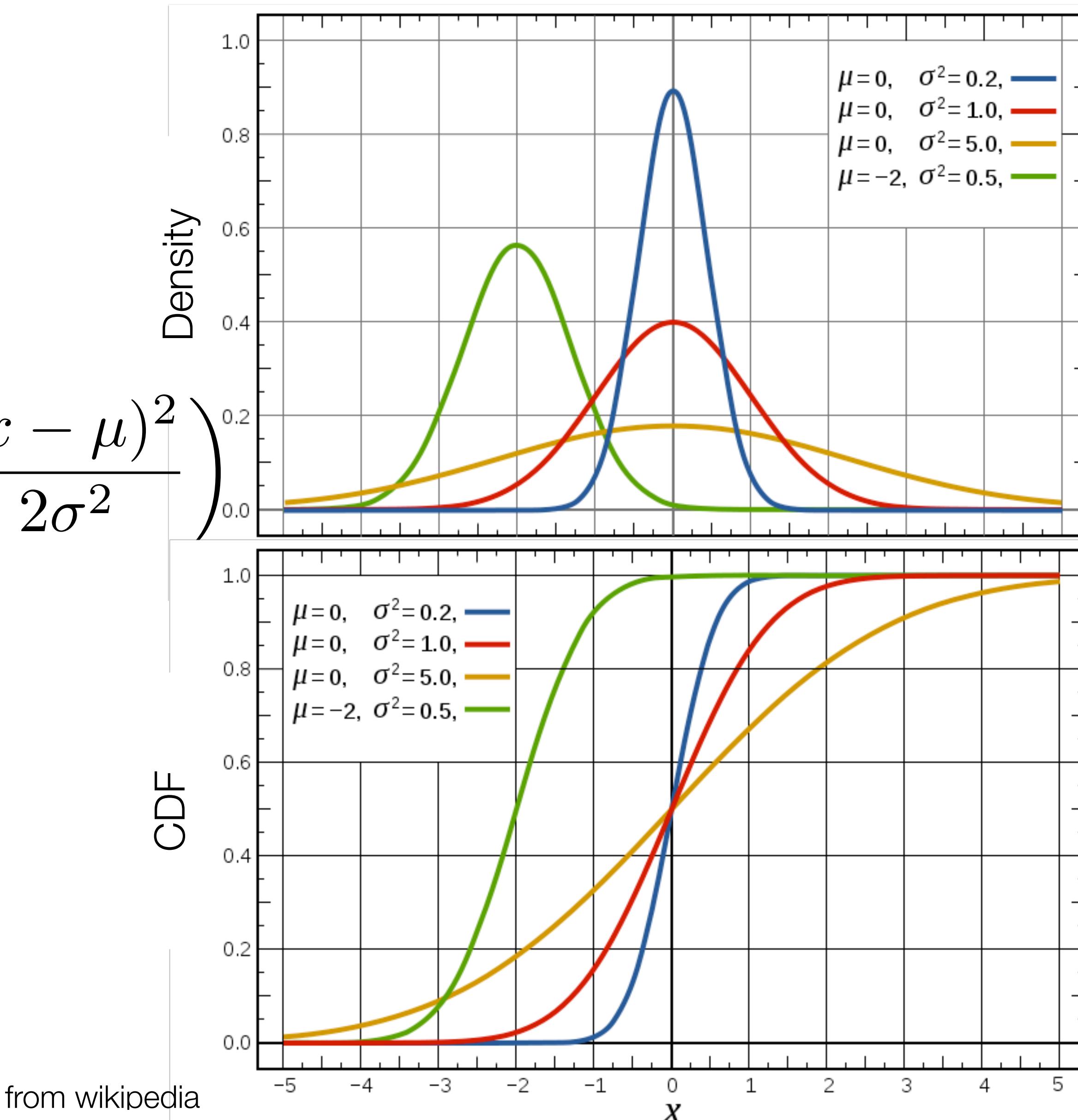
# Example: Gaussian / Normal distribution

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

$$\mu \in \mathbb{R}, \sigma > 0$$

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$$

Mean	$\mu$
Median	$\mu$
Mode	$\mu$
Standard deviation	$\sigma$



# Poisson distribution

Discrete probability distribution (no density)

$$X \sim \text{Poisson}(\lambda)$$

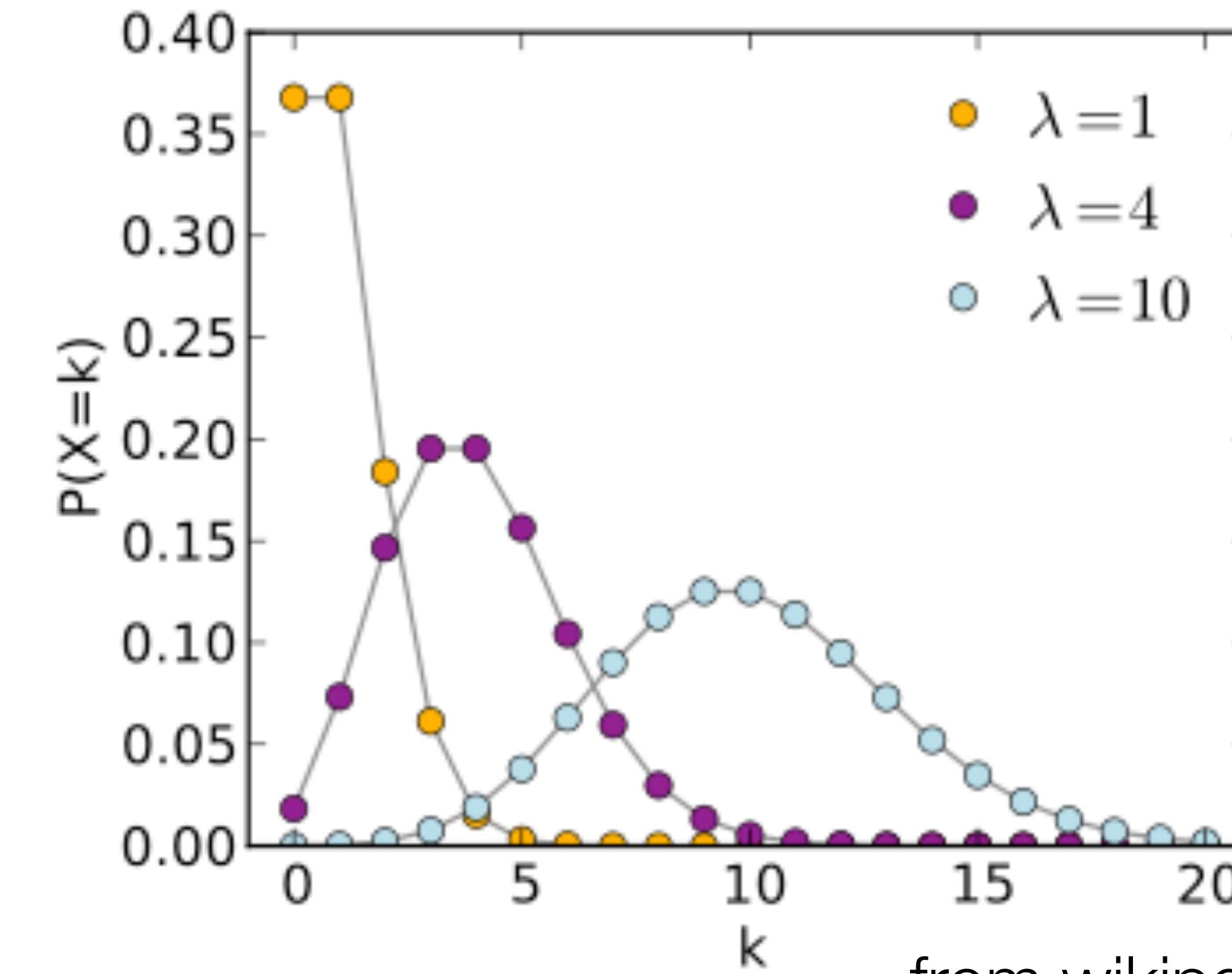
Number of photons on a detector  
Number of people in a shop

$$\Pr(X = k) = \exp(-\lambda) \frac{\lambda^k}{k!}$$

$$\lambda \in \mathbb{R}^+, k \in \mathbb{N}$$

For large  $\lambda$

$$\mathcal{N}(\lambda, \lambda)$$



from wikipedia

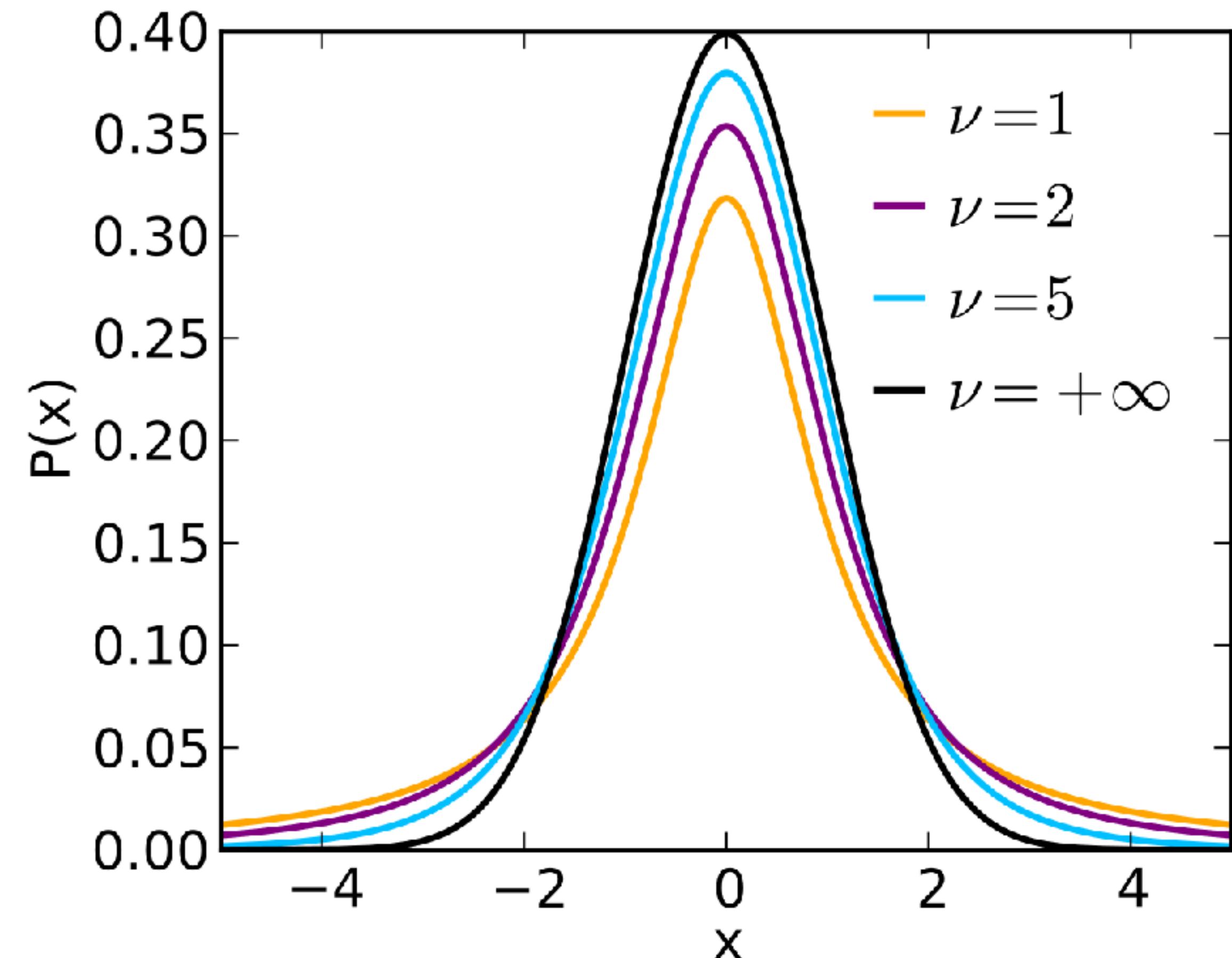
# THE STUDENT'S T-DISTRIBUTION

$$p(x; \nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\pi\nu} \cdot \Gamma(\frac{\nu}{2})} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

$$\nu \in \mathbb{R}^+, x \in \mathbb{R}$$

Mean	$0 \text{ if } \nu > 1$
Median	0
Mode	0
Standard deviation	$\frac{\sqrt{\nu - 2}}{\sqrt{\nu}} \text{ if } \nu > 2$ $\infty \text{ if } 1 < \nu < 2$

Useful for tests of significance of difference in means  
Tests for significance of slope in linear regression

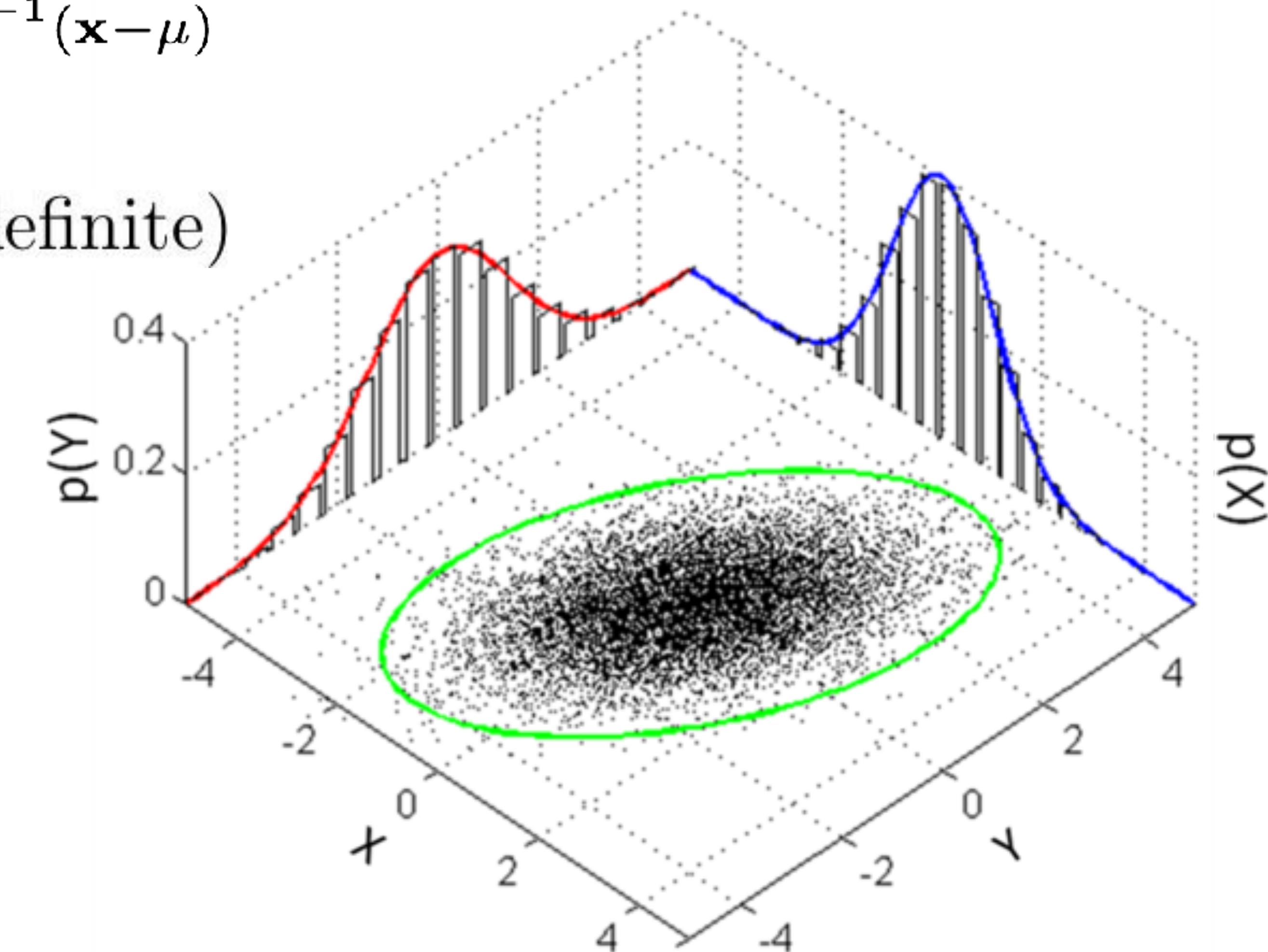


## THE MULTIVARIATE NORMAL DISTRIBUTION

$$\phi(\mathbf{x}; \mu, \Sigma) = \frac{1}{\sqrt{2\pi\det(\Sigma)}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T \Sigma^{-1} (\mathbf{x}-\mu)}$$

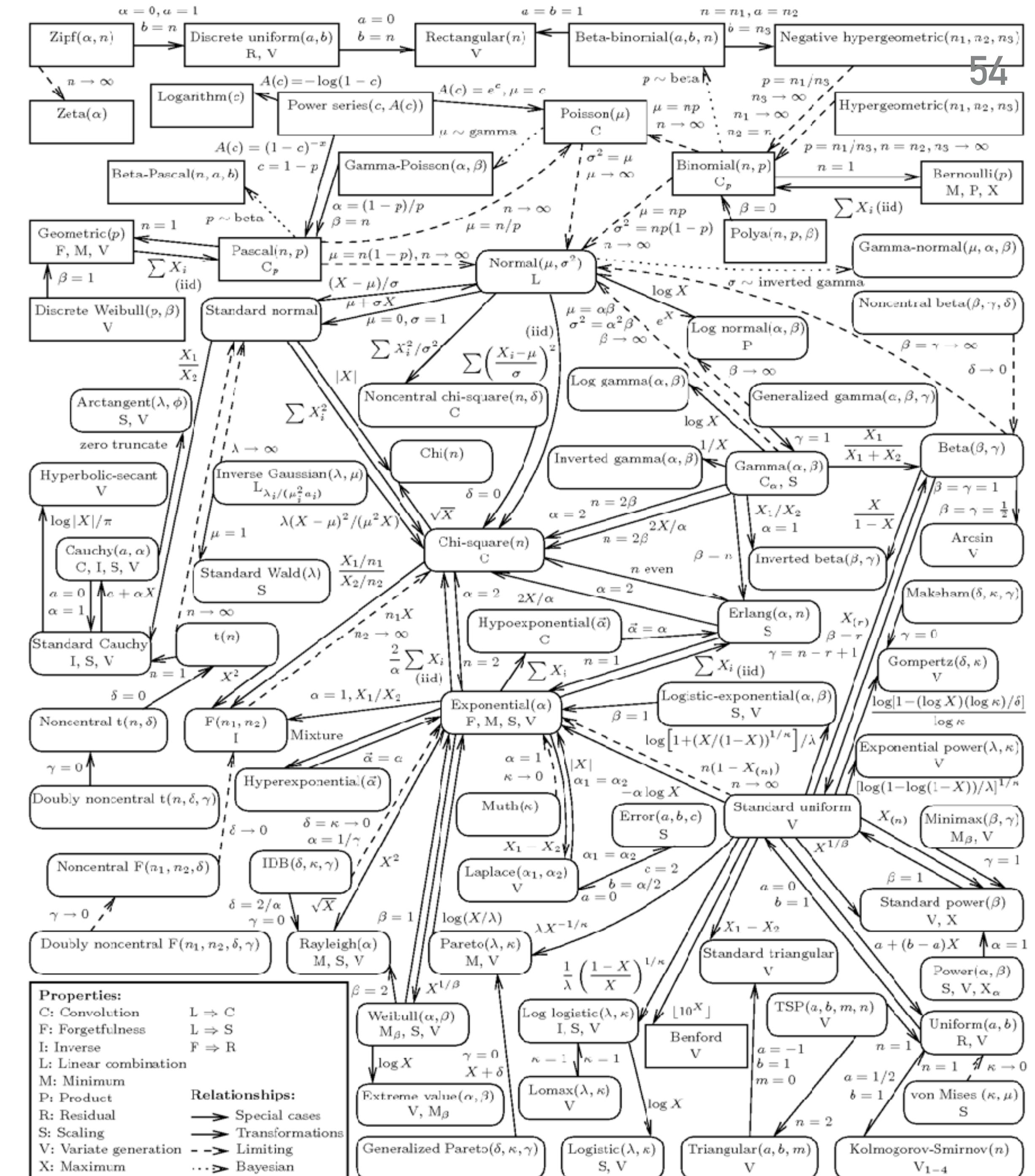
$\mu \in \mathbb{R}^n, \Sigma \in \mathbb{R}^{n \times n}$  (and is positive definite)

Mean	$\mu$
Median	$\mu$
Mode	$\mu$
Standard deviation	$\Sigma$



Lawrence M Leemis & Jacquelyn T McQueston (2008)  
 "Univariate Distribution Relationships",  
 The American Statistician, 62:1, 45-53,  
 DOI: [10.1198/000313008X270448](https://doi.org/10.1198/000313008X270448)

Chapter 2 of the textbook has a bunch of other distributions you'll encounter frequently in astrophysics (and other domains) - definitely skim this



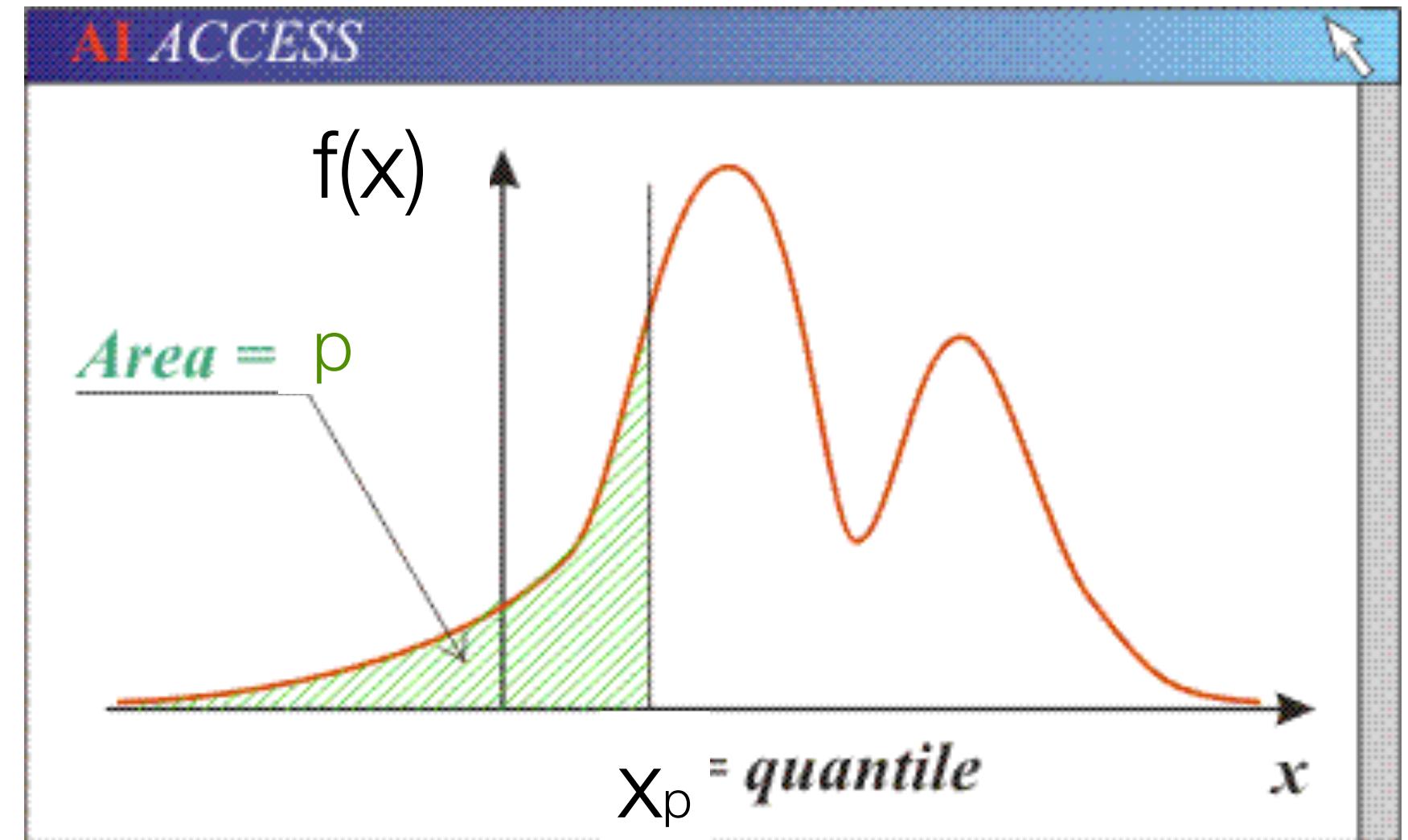
# Quantiles

- $x_p$ : p-quantiles of  $f(x)$

$$p = \int_{-\infty}^{x_p} f(x) dx$$

- Measure of location: Median

$$1/2 = \int_{-\infty}^{x_{1/2}} f(x) dx$$



from [www.aiacces.net](http://www.aiacces.net)

- Measure of dispersion: Inter-quantile range

$$\text{IQR} = x_{3/4} - x_{1/4}$$

- ▶ So given a nice distribution function, you can calculate the mean, variance and moments ...
- ▶ ... but you don't usually have a nice distribution function given to you.
- ▶ The distribution function is the thing you are trying to infer from the data!
  - ▶ i.e.  $P(H|D)$
  - ▶ The thing you have are the data - **observations,  $D$**

1.4

---

# ESTIMATORS

# Data, samples

---

- Usually we have observations, e.g. additive process

$$y_i = f(t_i) + \epsilon_i \quad i = 1, \dots, n$$

Deterministic      random variable

- We want a characterisation of the deterministic and random parts
- Suppose something about the random variable, often normality:  $\mathcal{N}(0, \sigma^2)$

- Assumption of models

- Estimate the parameters of a distribution, moments

- Exercise 1: Sample mean:  $\hat{\mu} = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$   $E(\bar{X}) = \mu$

- Exercise 2: Sample variance (bias):

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \quad E(\hat{\sigma}^2) = \frac{n}{n-1} \sigma^2$$

redefine  
→

$$\hat{\sigma}^2 = S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

- Sample quantiles are estimators of quantiles

# Distribution derived from Normal distribution

60

## 1) Chi square distribution

Modified from Maria Suveges, Laurent Eyer

If  $X_i \stackrel{\text{iid}}{\sim} \mathcal{N}(0, 1)$

iid= Independent identically distributed

**mean:**  $k$

**variance:**  $2k$

**skewness:**  $\sqrt{8/k}$

**kurtosis:**  $12/k$

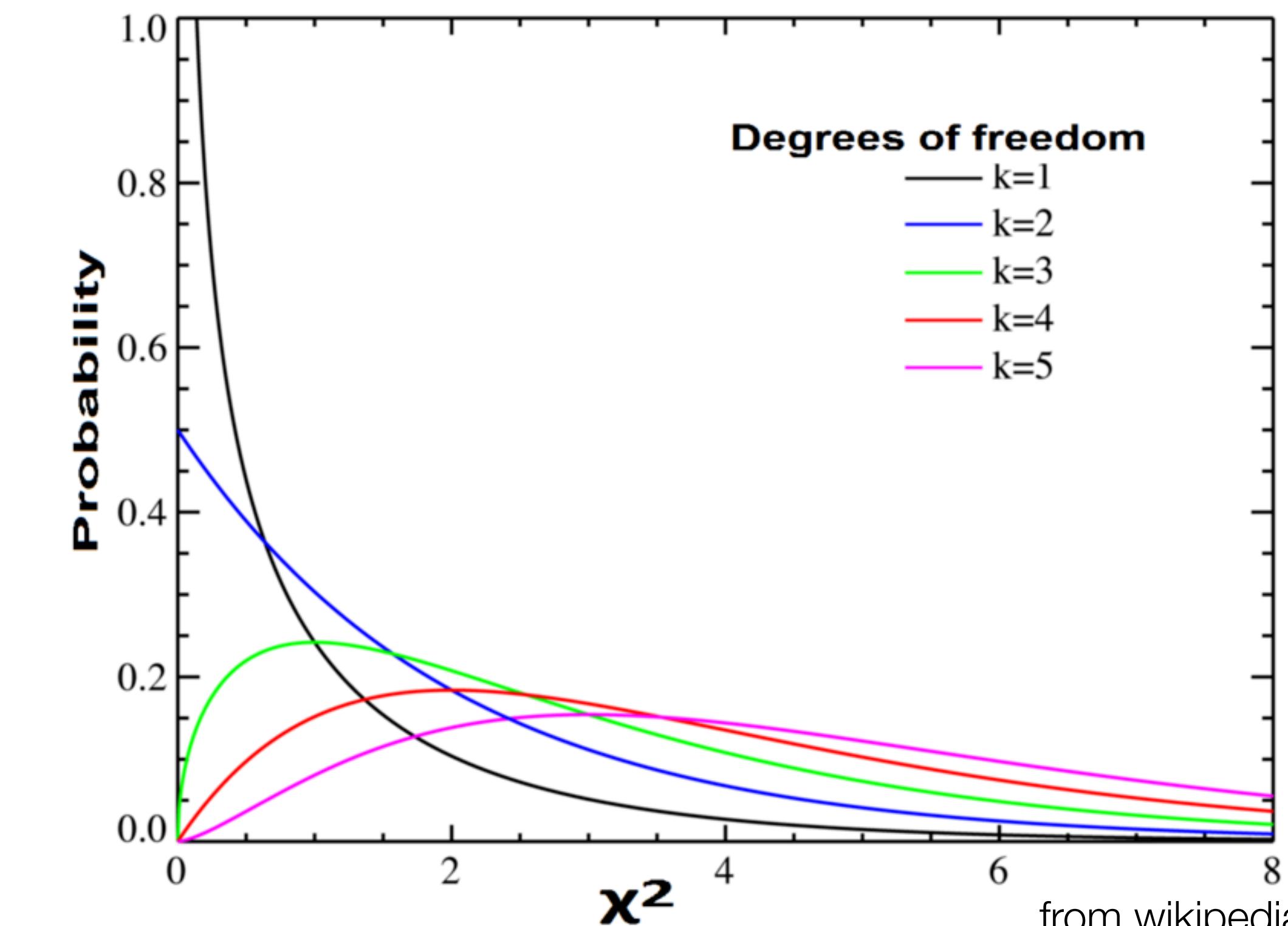
$X_i \stackrel{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma)$

$\sum_{i=1}^k (X_i - \bar{X})^2 / \sigma^2 \sim \chi_{k-1}^2$

When  $k$  is large  $\chi_k^2$  approximates a  $\mathcal{N}(k, 2k)$

$$\sum_{i=1}^k X_i^2 \sim \chi_k^2$$

$$f(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} \exp(-x/2)$$



from wikipedia

# Central limit theorem

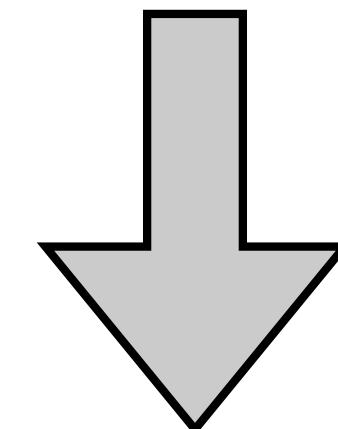
---

The distribution of the mean of a sufficiently large number of random variables can be approximated by a Gaussian distribution!

$X_i, i = 1, \dots, n$  iid with  $E(X_i) = \mu$   $\text{Var}(X_i) = \sigma^2$

iid= Independent identically distributed

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \text{ follows approximately } \mathcal{N}(0, 1)$$



**One reason why  
the Gaussian distribution is so important**

# Distribution derived from Normal distribution

## 2) Student distribution

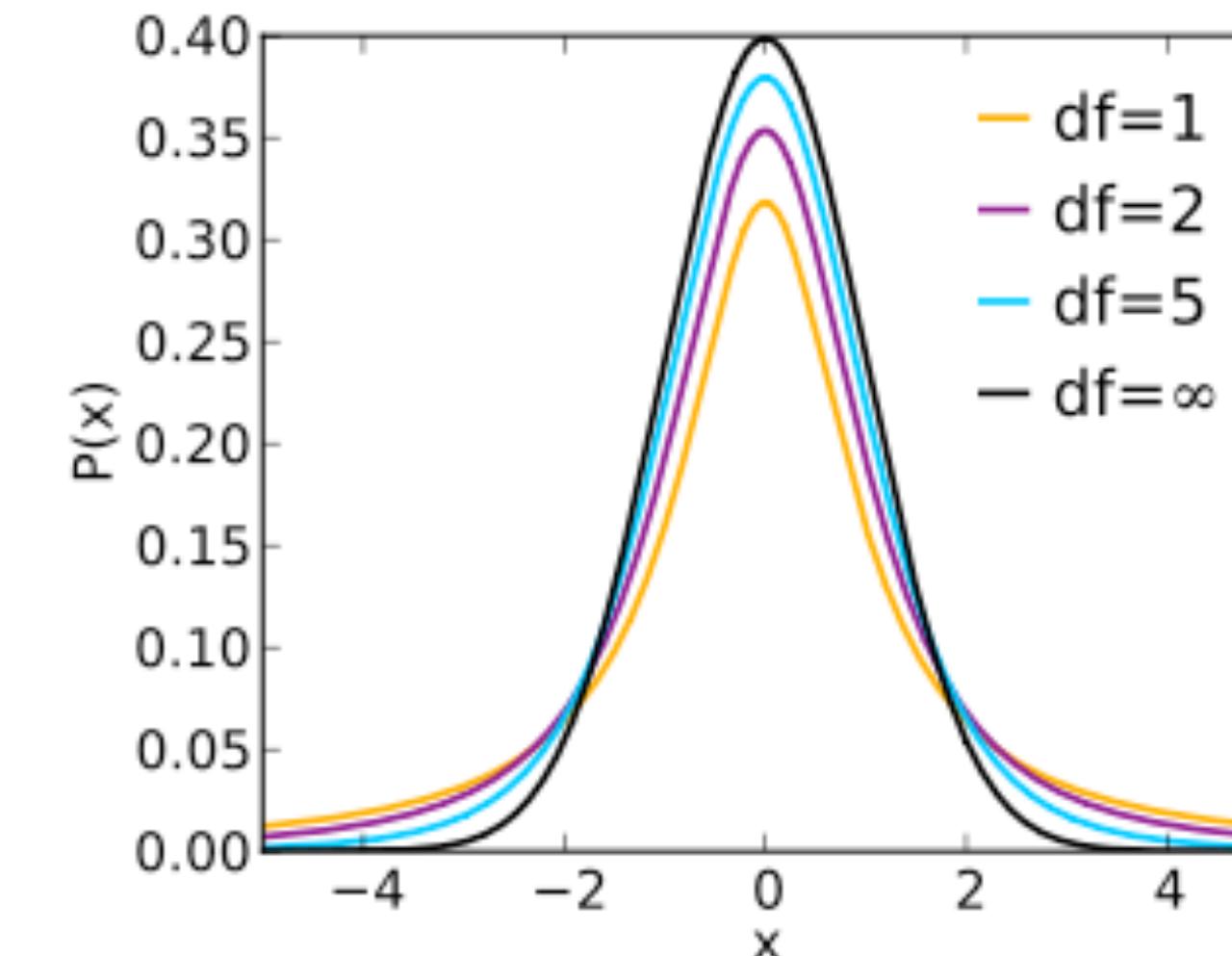
$$\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim \mathcal{N}(0, 1)$$

$$\frac{\bar{X} - \mu}{S / \sqrt{n}} \sim t_{n-1}$$

$$f(x) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(n/2)} \left(1 + \frac{x^2}{n}\right)^{-(n+1)/2}$$

Note

$$t_{\infty} = \mathcal{N}(0, 1)$$



**mean:** 0  $n > 1$

**NaN**  $n = 0, 1$

**variance:**  $n/(n-2)$   $n > 2$

$\infty$   $1 < n \leq 2$

**otherwise NaN**

**skewness:** 0  $n > 3$

**kurtosis:**  $6/(n-4)$   $n > 4$

SO WE DEFINITELY NEED A  
QUANTITATIVE WAY OF ASSESSING THE  
SIMILARITY OF TWO DISTRIBUTIONS

# Quantiles

- $x_p$ : p-quantiles of  $f(x)$

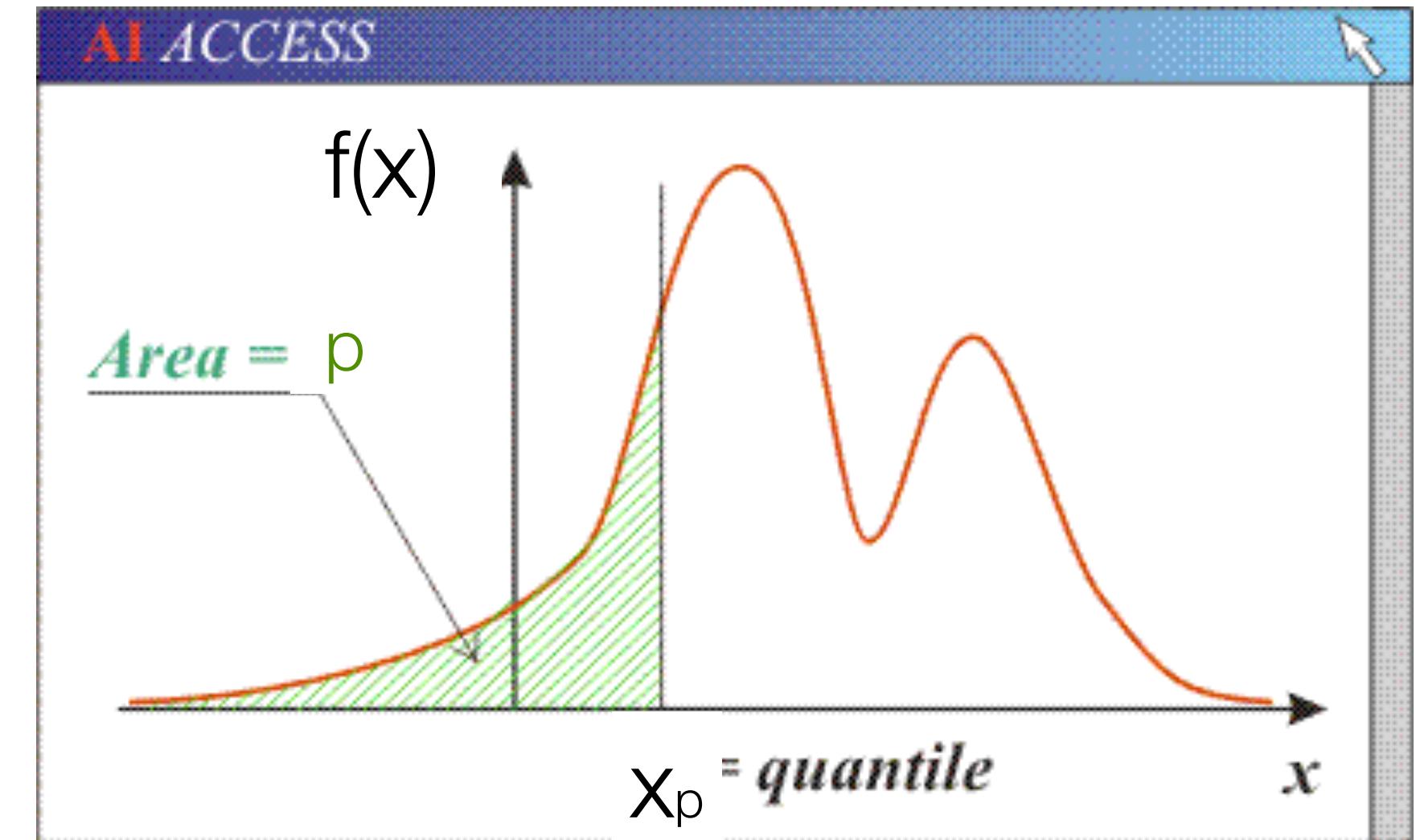
$$p = \int_{-\infty}^{x_p} f(x) dx$$

- Measure of location: Median

$$1/2 = \int_{-\infty}^{x_{1/2}} f(x) dx$$

- Measure of dispersion: Inter-quantile range

$$\text{IQR} = x_{3/4} - x_{1/4}$$

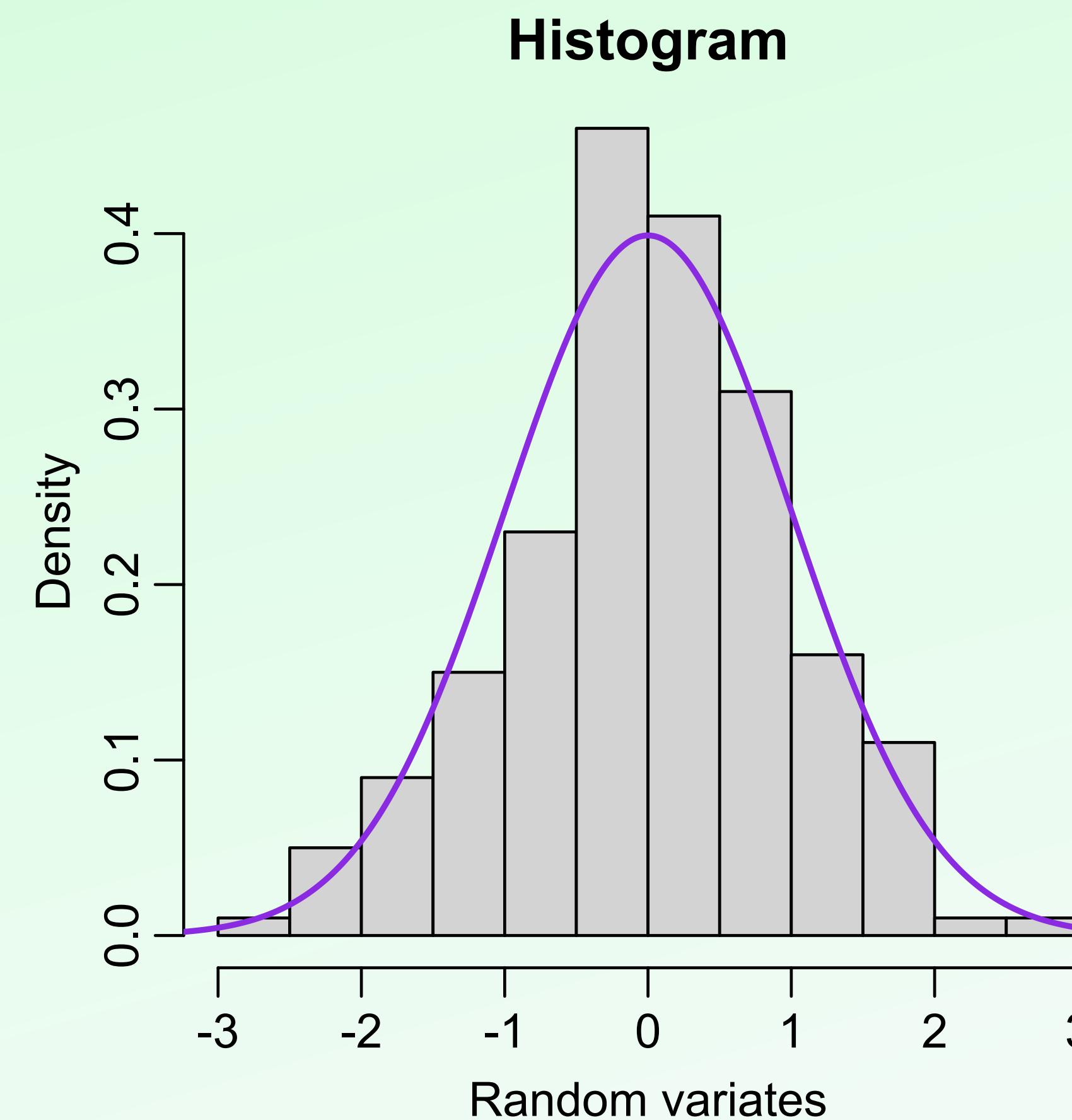


from [www.aiacces.net](http://www.aiacces.net)

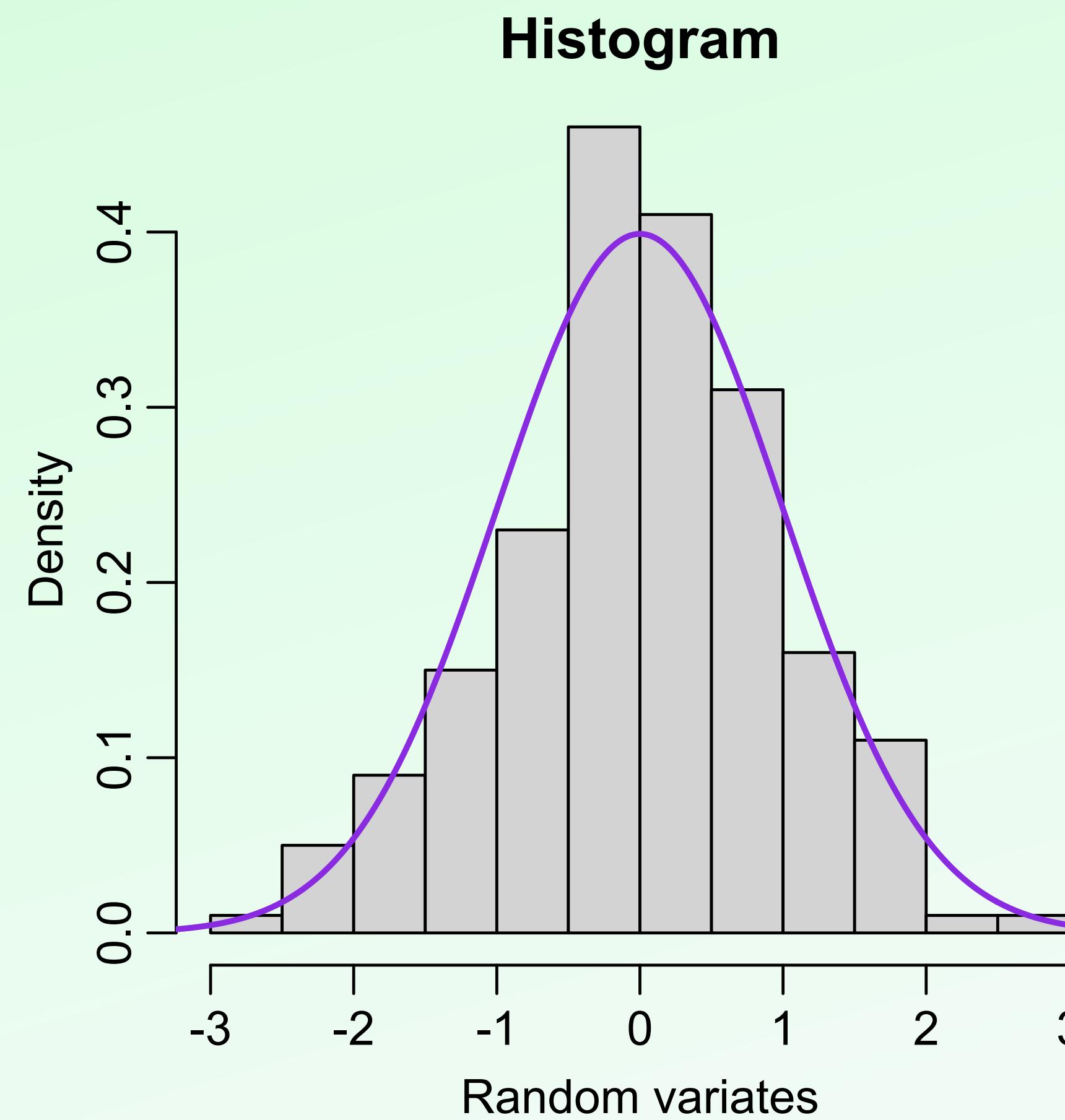
# Diagnostics: the QQ plot

Modified from Maria Suveges, Laurent Eyer  
65

- Parametric models are nearly always only approximate
- In classical statistics, CIs are based on asymptotic behaviour (that is, when  $n \rightarrow \infty$ )



- Parametric models are nearly always only approximate
- In classical statistics, CIs are based on asymptotic behaviour (that is, when  $n \rightarrow \infty$ )



Statisticians' preference:  
**quantile-quantile (QQ) plot.**

It consists of the pairs

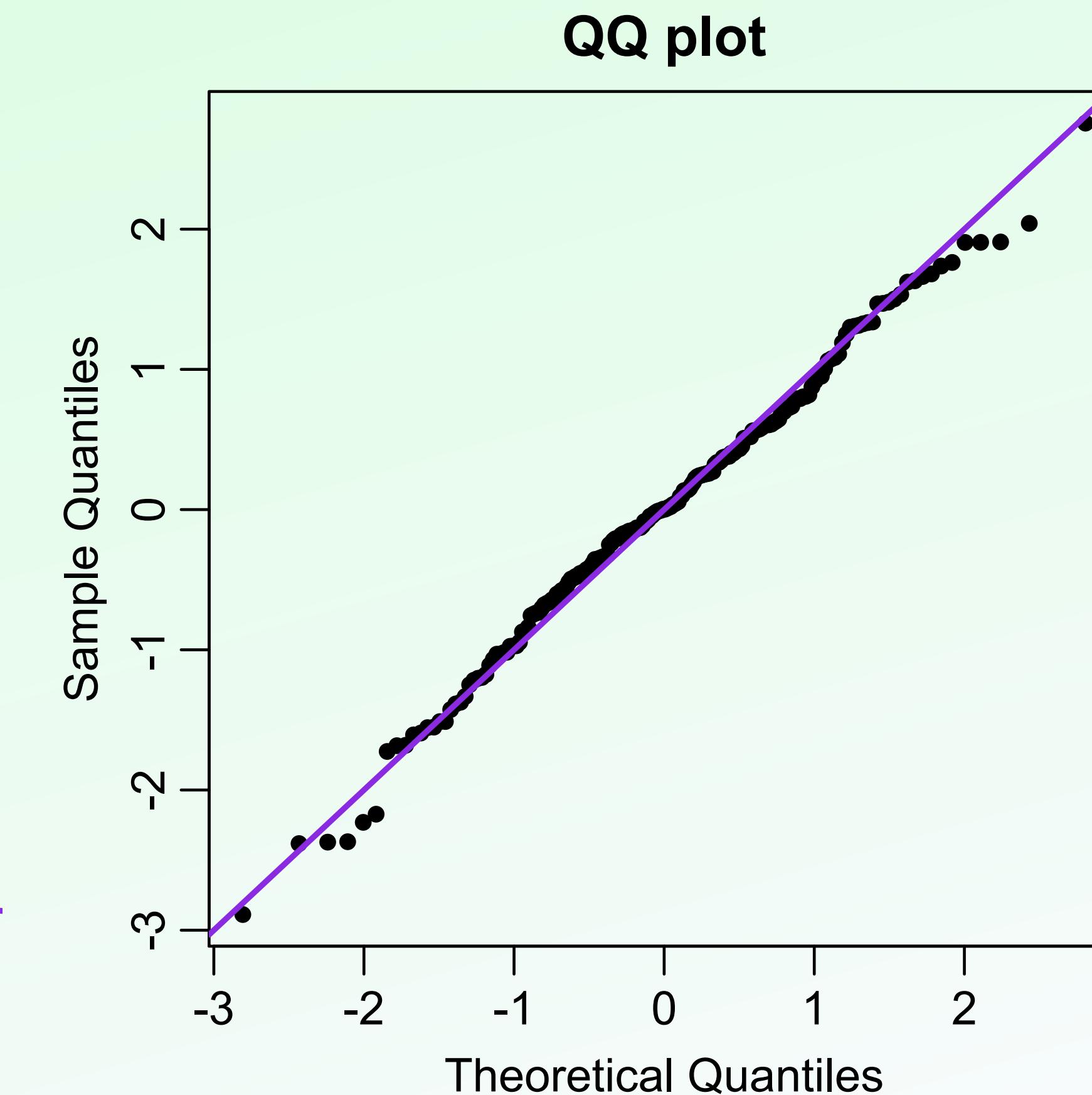
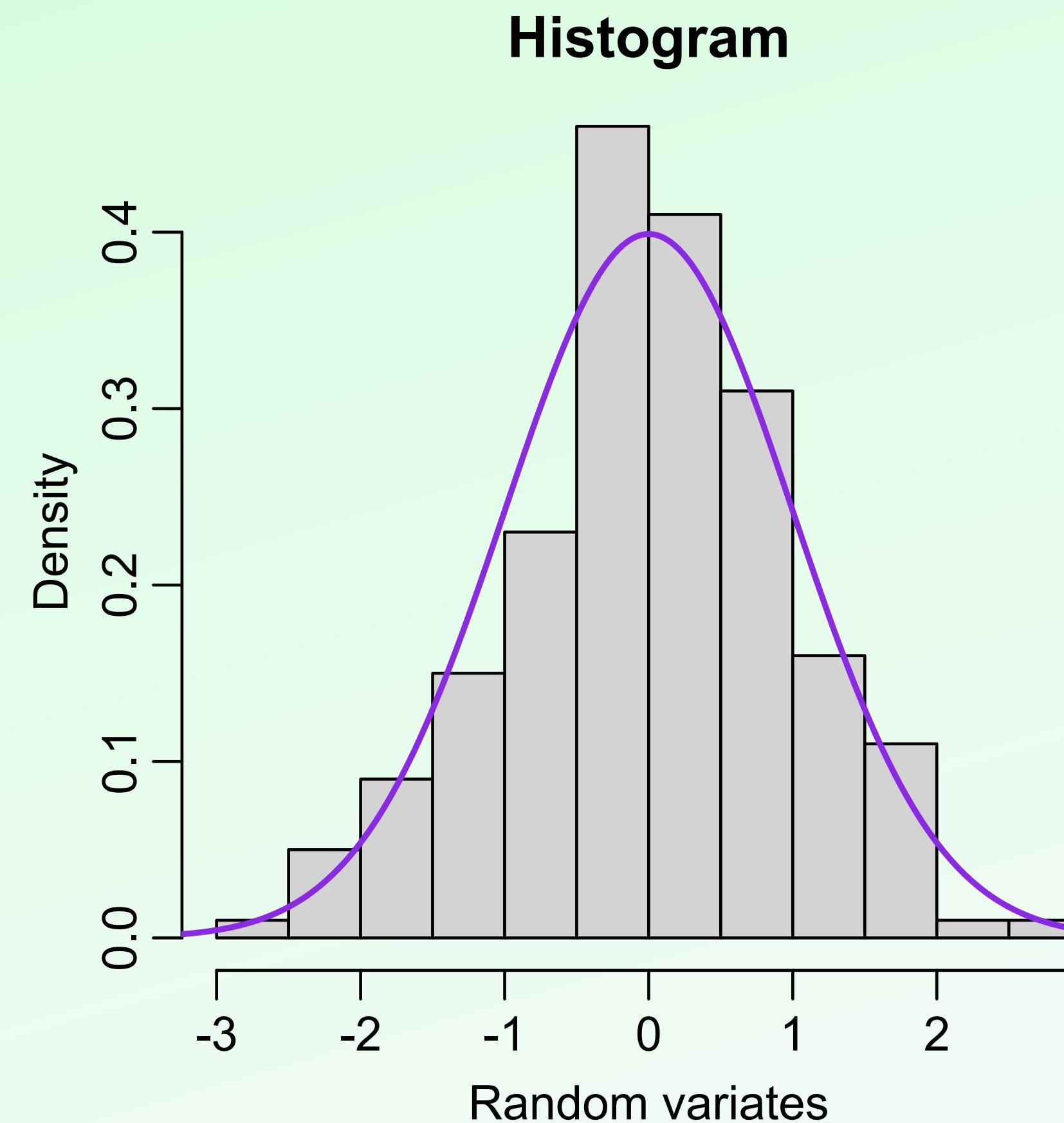
$$\left\{ \Phi^{-1} \left( \frac{j}{n+1} \right), Y_{(j)} \right\},$$

where  $\Phi^{-1} \left( \frac{j}{n+1} \right)$  is the inverse of the standard normal distribution, and  $Y_{(j)}$  is the ordered sample.

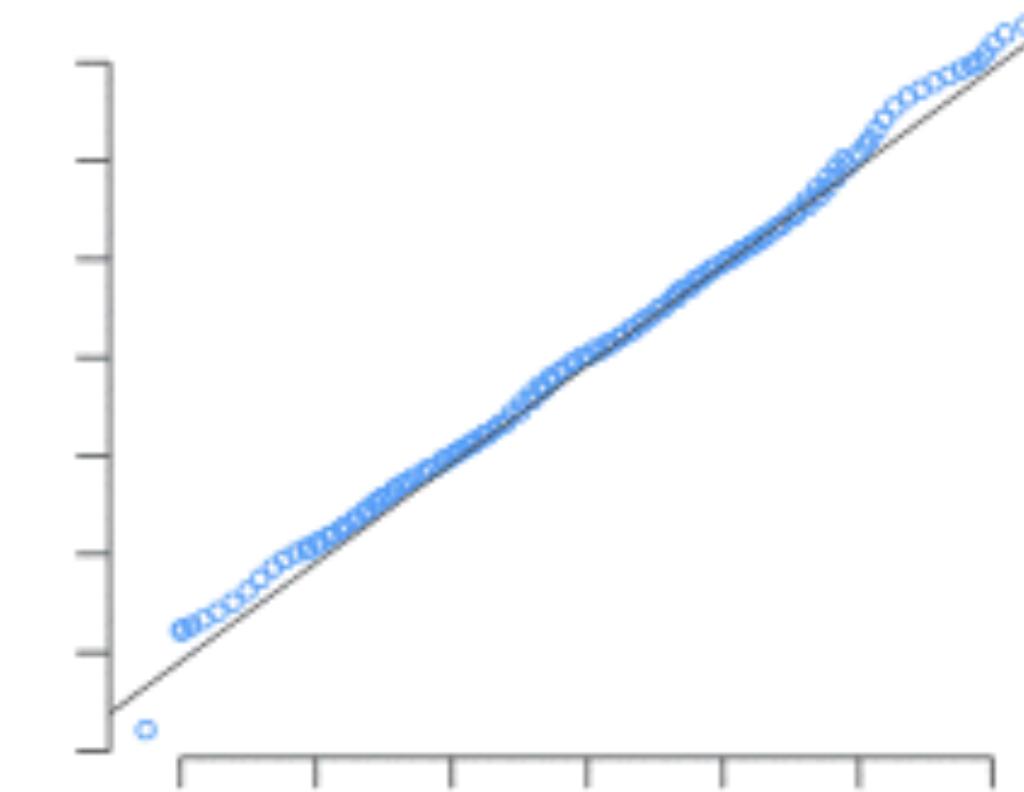
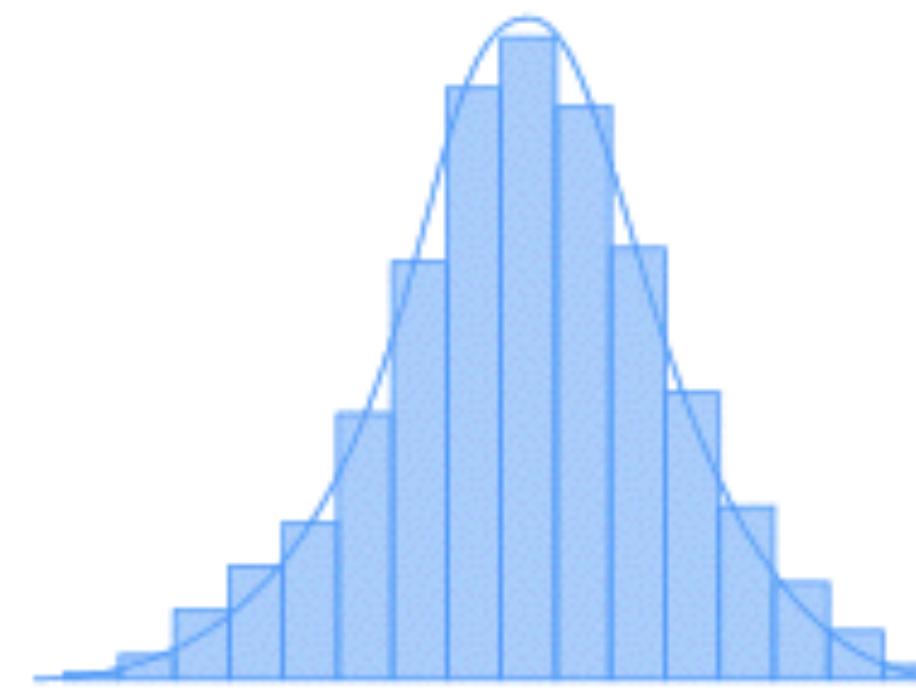
# Diagnostics: the QQ plot

Modified from Maria Suveges, Laurent Eyer  
67

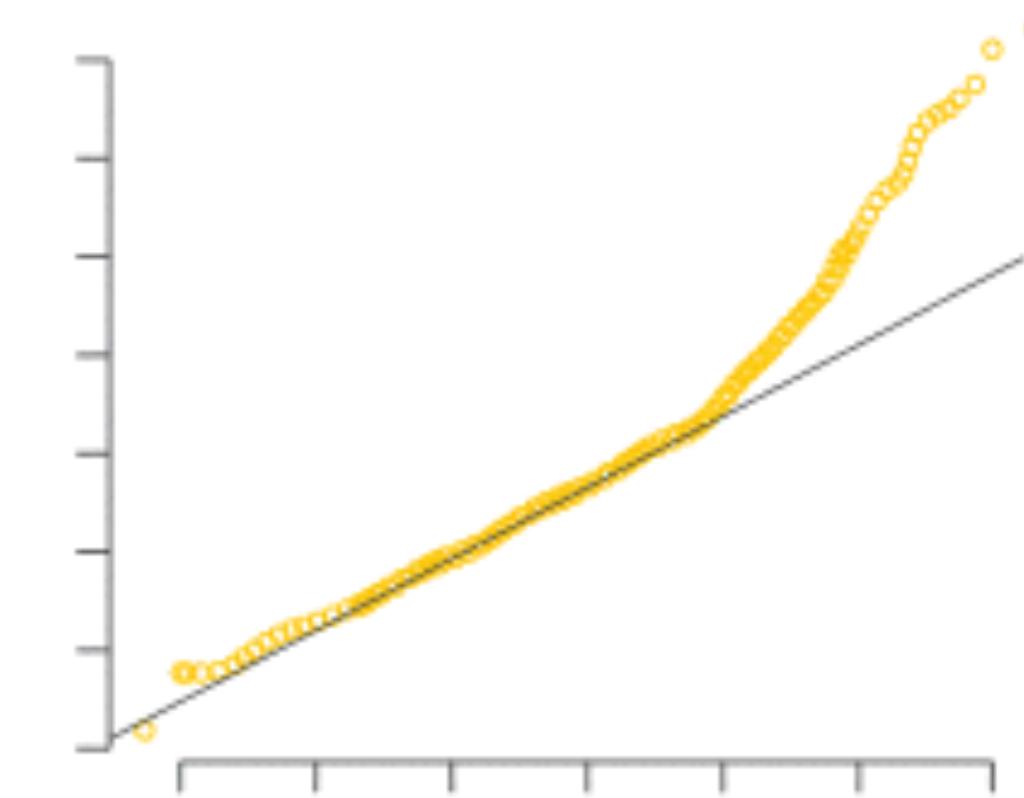
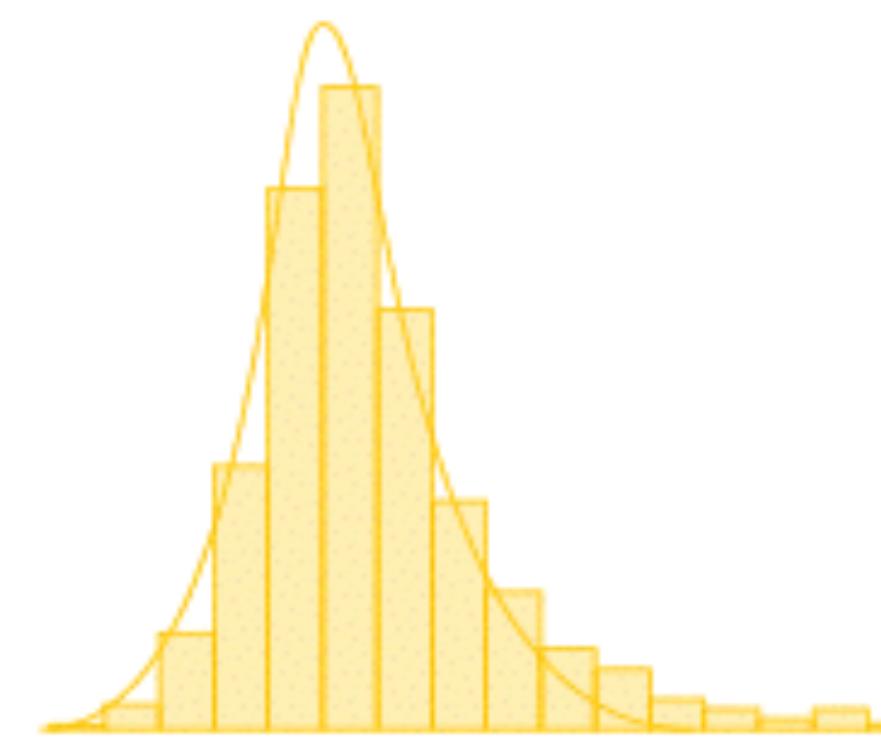
- Parametric models are nearly always only approximate
- In classical statistics, CIs are based on asymptotic behaviour (that is, when  $n \rightarrow \infty$ )



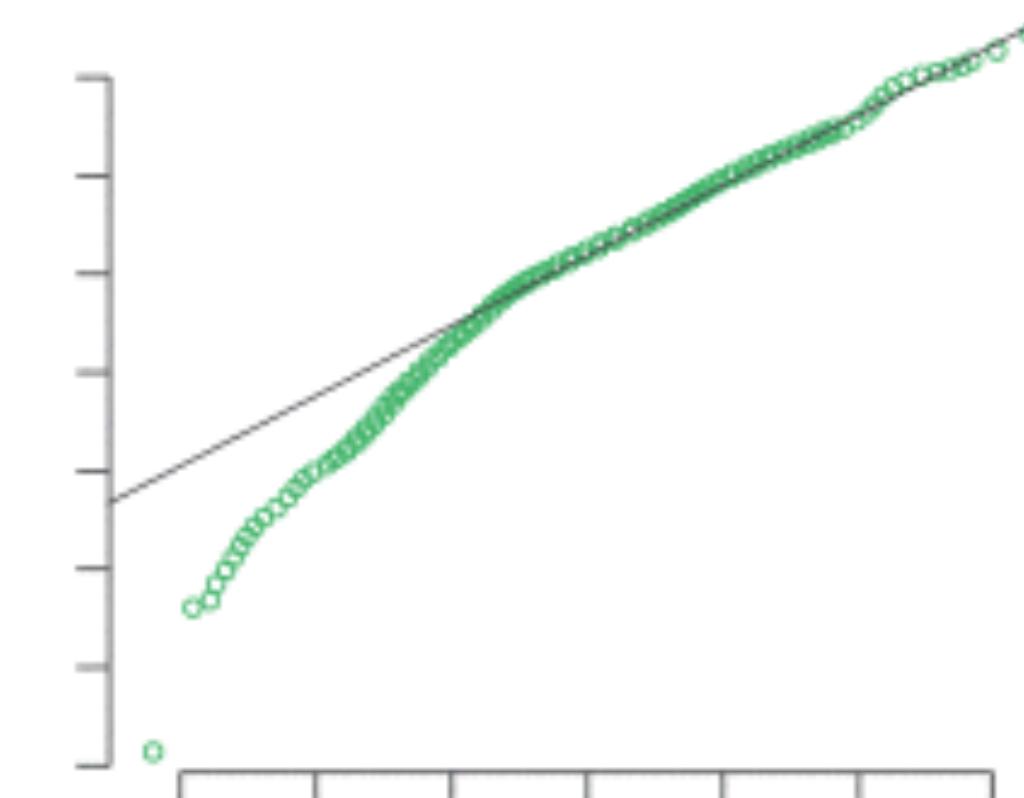
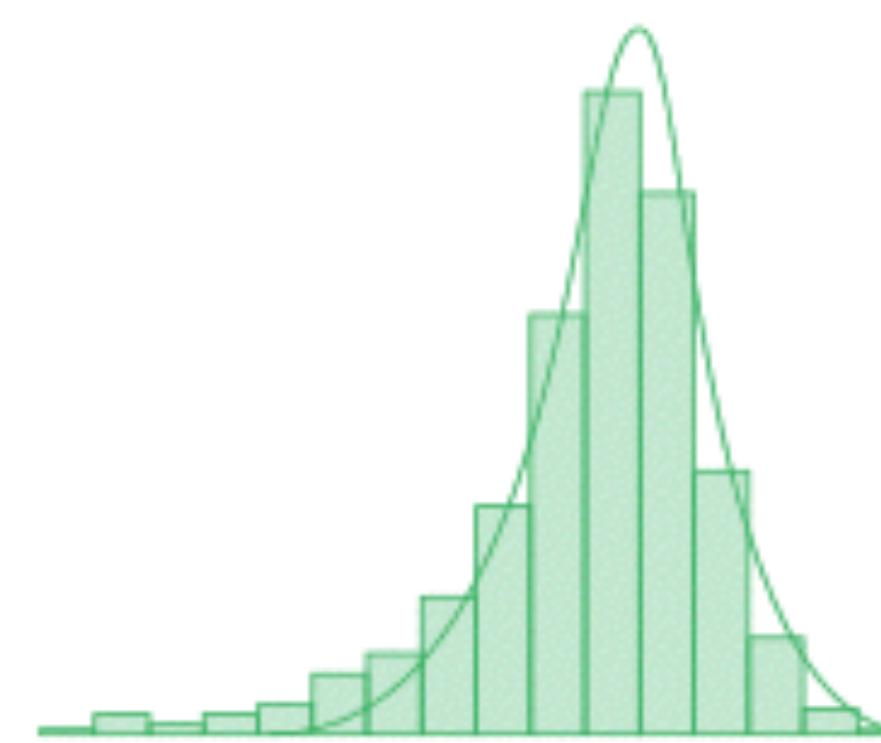
Normally distributed  
data



Right-skewed  
data



Left-skewed  
data



LII. *An Essay towards solving a Problem in the Doctrine of Chances.* By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a Letter to John Canton, A. M. F. R. S.

Dear Sir,

Read Dec. 23, 1763. I Now send you an essay which I have found among the papers of our deceased friend Mr. Bayes, and which, in my opinion, has great merit, and well deserves to be preserved. Experimental philosophy, you will find, is nearly interested in the subject of it; and on this account there seems to be particular reason for thinking that a communication of it to the Royal Society cannot be improper.

He had, you know, the honour of being a member of that illustrious Society, and was much esteemed by many in it as a very able mathematician. In an introduction which he has writ to this Essay, he says, that his design at first in thinking on the subject of it was, to find out a method by which we might judge concerning the probability that an event has to happen, in given circumstances, upon supposition that we know nothing concerning it but that, under the same circum-

## HOW YOU DO SCIENCE IS AS IMPORTANT AS WHAT SCIENCE YOU DO

# OTHER VALUABLE LESSONS FROM BAYES

# EXTRA

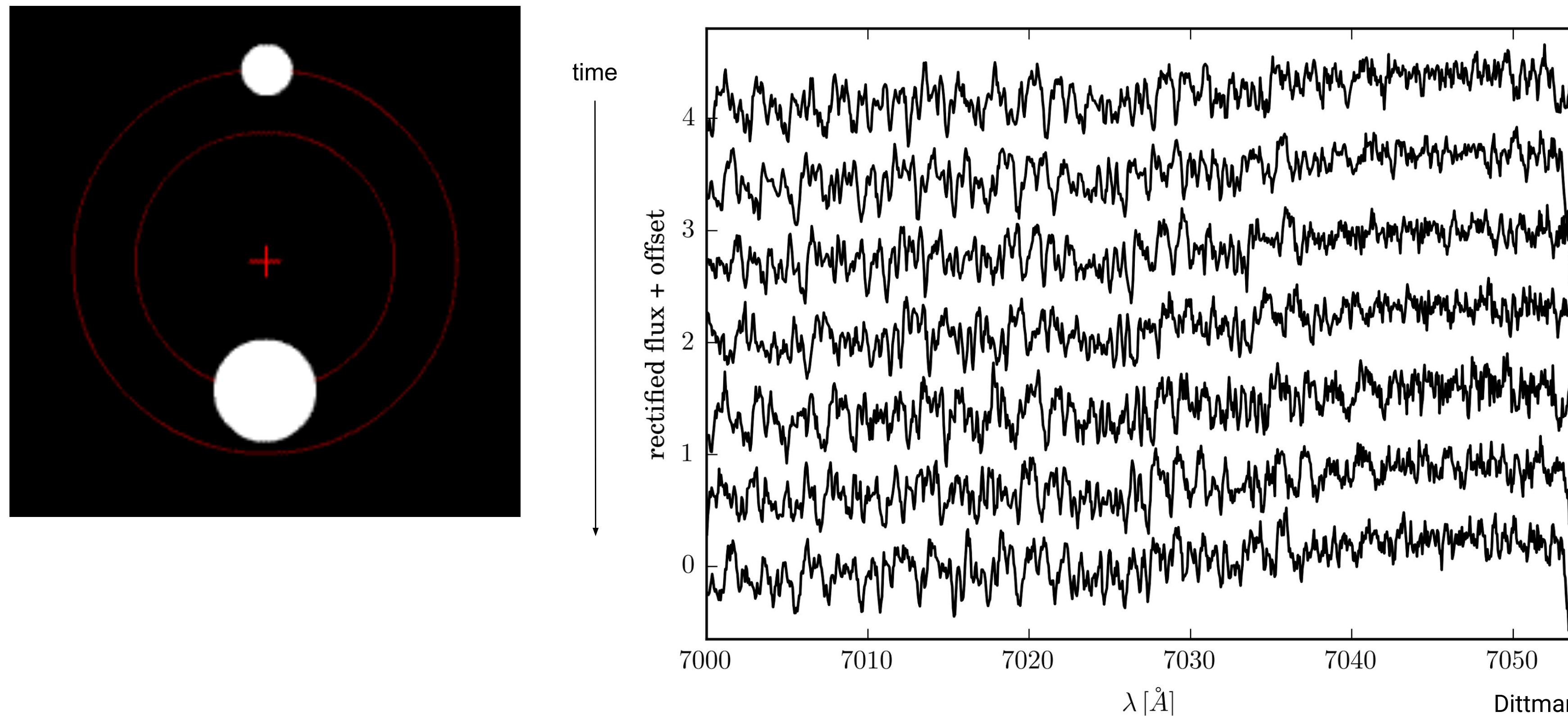
1.5

---

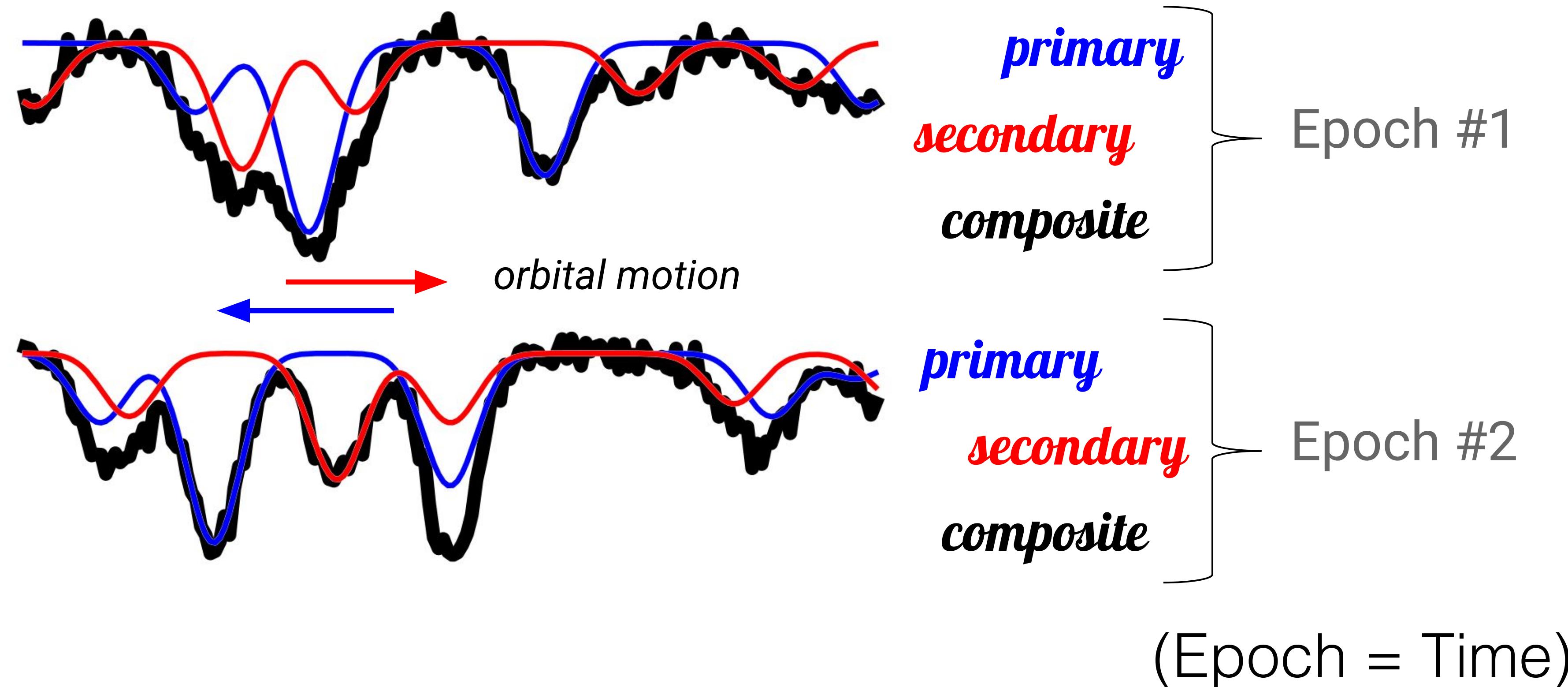
# PUTTING THINGS TOGETHER

Astrostatistics Case Studies:  
Disentangling Time Series Spectra with Gaussian  
Processes: Applications to Radial Velocity Analysis  
(Czekala et al. 2017, ApJ, 840, 49. arXiv:1702.05652)

**Raw Observations of the LP661-13 M4 Binary**



# Spectroscopic Binary Stars

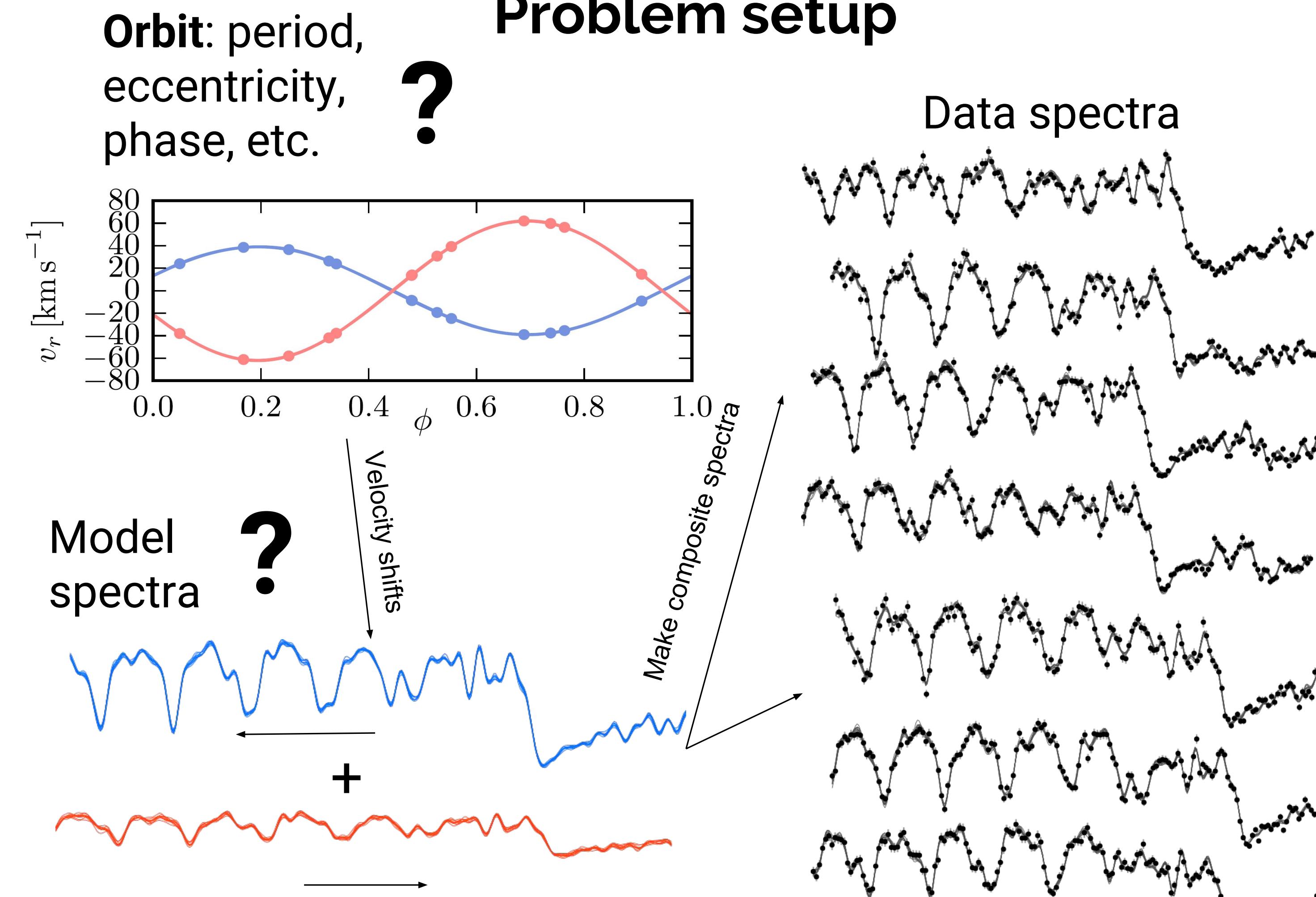


We only observe the “noisy” sum of two (latent) spectra.  
Latent (underlying) spectra are unknown functions  
Observed spectrum = Measured Data

# Forward Model = Generates Data

## Problem setup

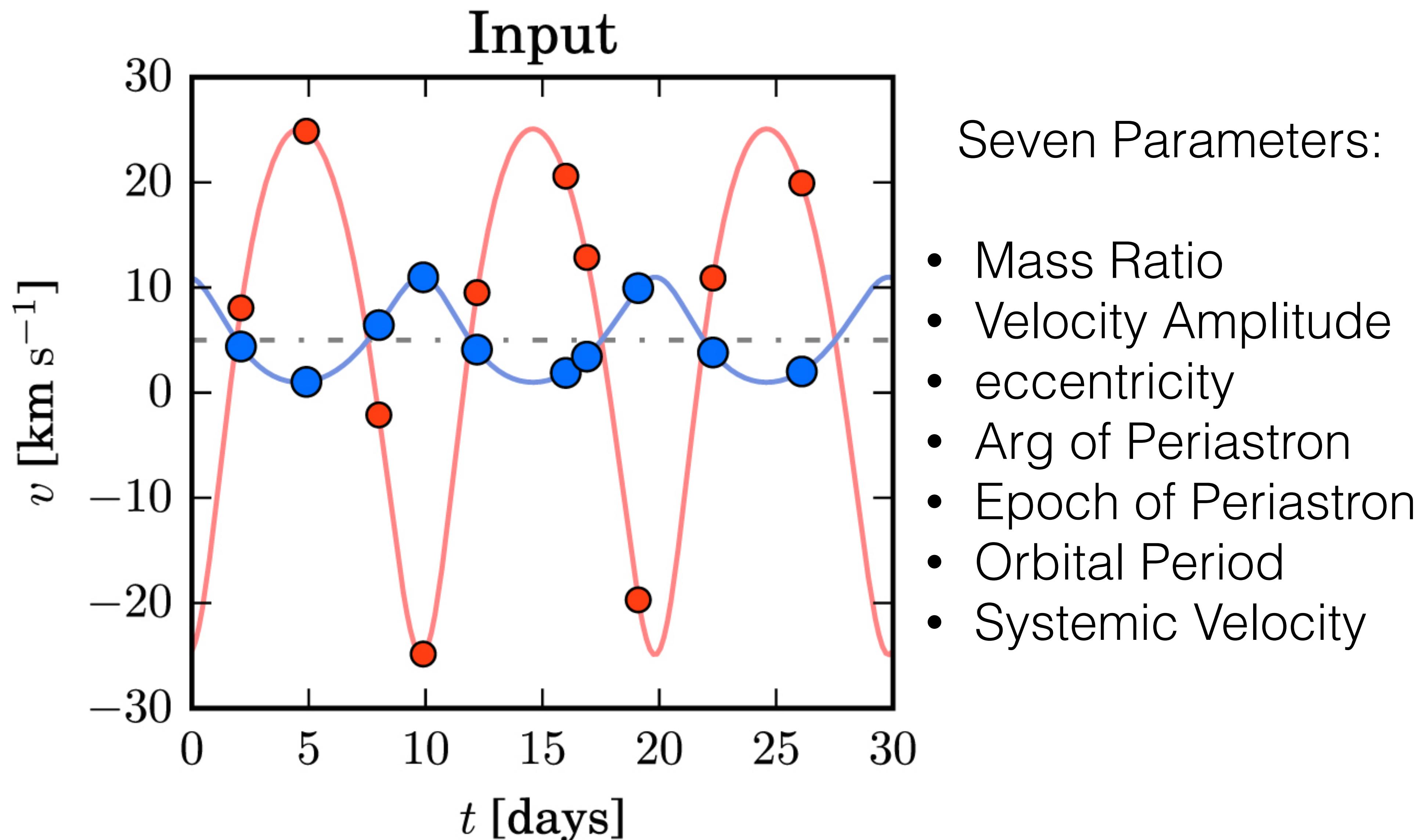
74



<https://www.youtube.com/watch?v=kHjN42ft6aU>

Goal: Go Backwards and Infer the Component Spectra & Orbital Parameters from noisy, observed (composite) spectra time series

# Orbital Parametric Model



# Nonparametric Bayes

## Gaussian processes

We will model the latent stellar spectrum  $f_\lambda$  as a Gaussian process

$$f_\lambda \sim \text{GP}(\mu(\lambda), k(\lambda, \lambda'))$$

A function is said to have a Gaussian process if for any collection of inputs the random vector  $\mathbf{f}$  has a multivariate Gaussian distribution with mean  $\mathbf{\mu}$  and covariance matrix given by  $k$  evaluated over ***lambda***

For a covariance kernel, we will use the commonly used squared exponential kernel, which relates pixels in the spectrum based upon their distance in log-wavelength ( $\propto$  velocity)

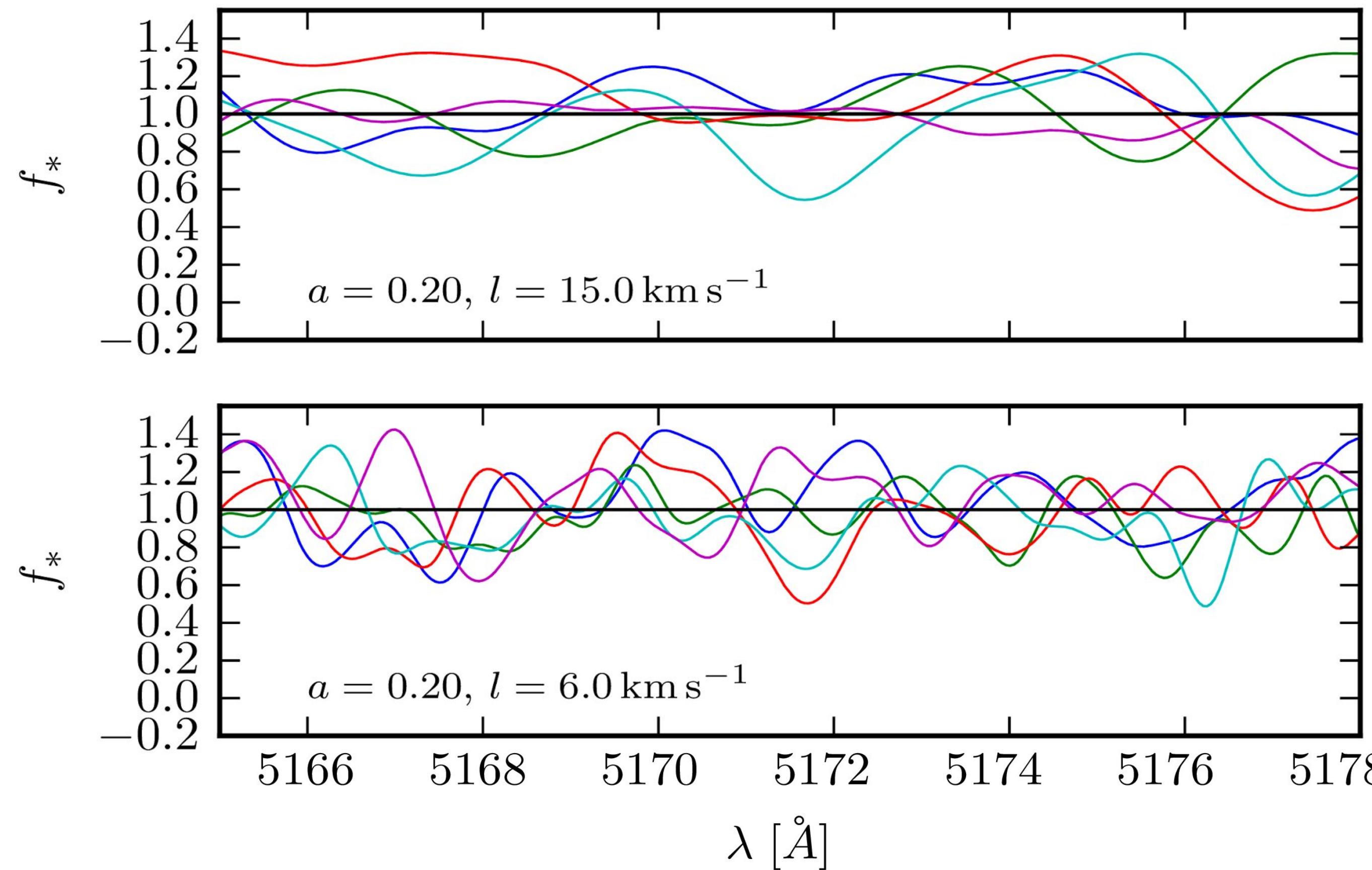
$$k_{ij}(r_{ij} | a, l) = a^2 \exp\left(-\frac{r_{ij}^2}{2l^2}\right)$$

Gaussian Process = a prior on functions (latent spectra)

77

## Gaussian Process model for a single, stationary star

(Zoomed) draws from the prior



l

l

Inference = Which function is most consistent with the data?

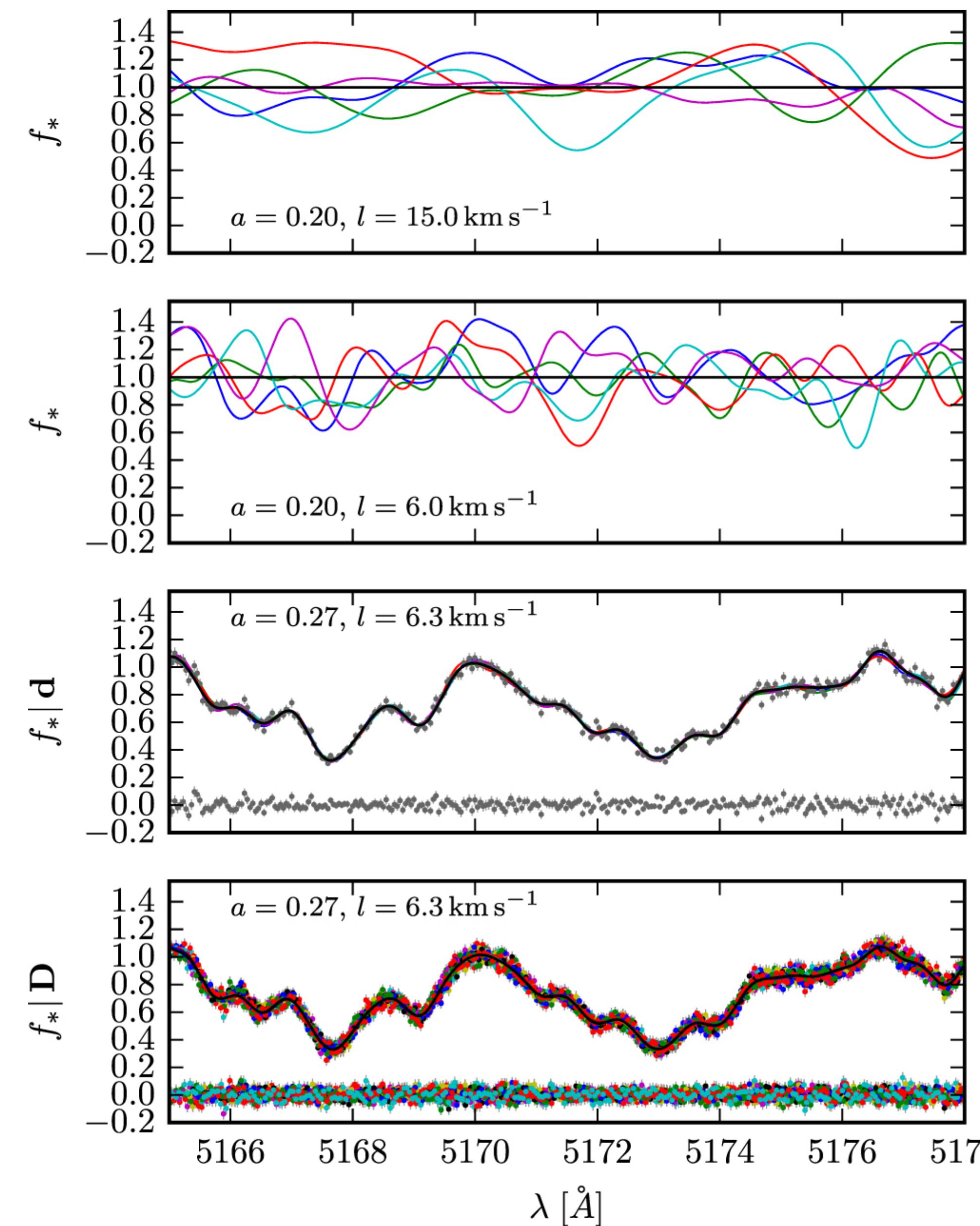
# Gaussian Process: Priors & Posteriors

78

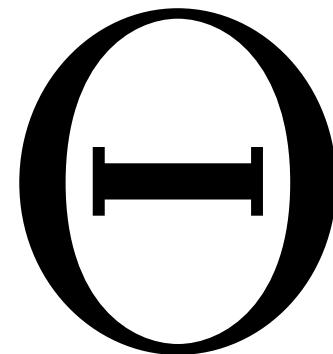
GP prior  
(long length scale)

(short length scale)

GP Posterior  
(conditioned on  
data spectrum  $\mathbf{d}$ )  
Inference of latent  
spectrum



# Known Unknowns



7-dim Orbital Parameters = Period, Phase, eccentricity, Velocity Amplitude

$$f(\lambda), g(\lambda)$$

( $\infty$ -dim) Latent Functions = the unobserved component spectra of the primary (f) and secondary (g) stars

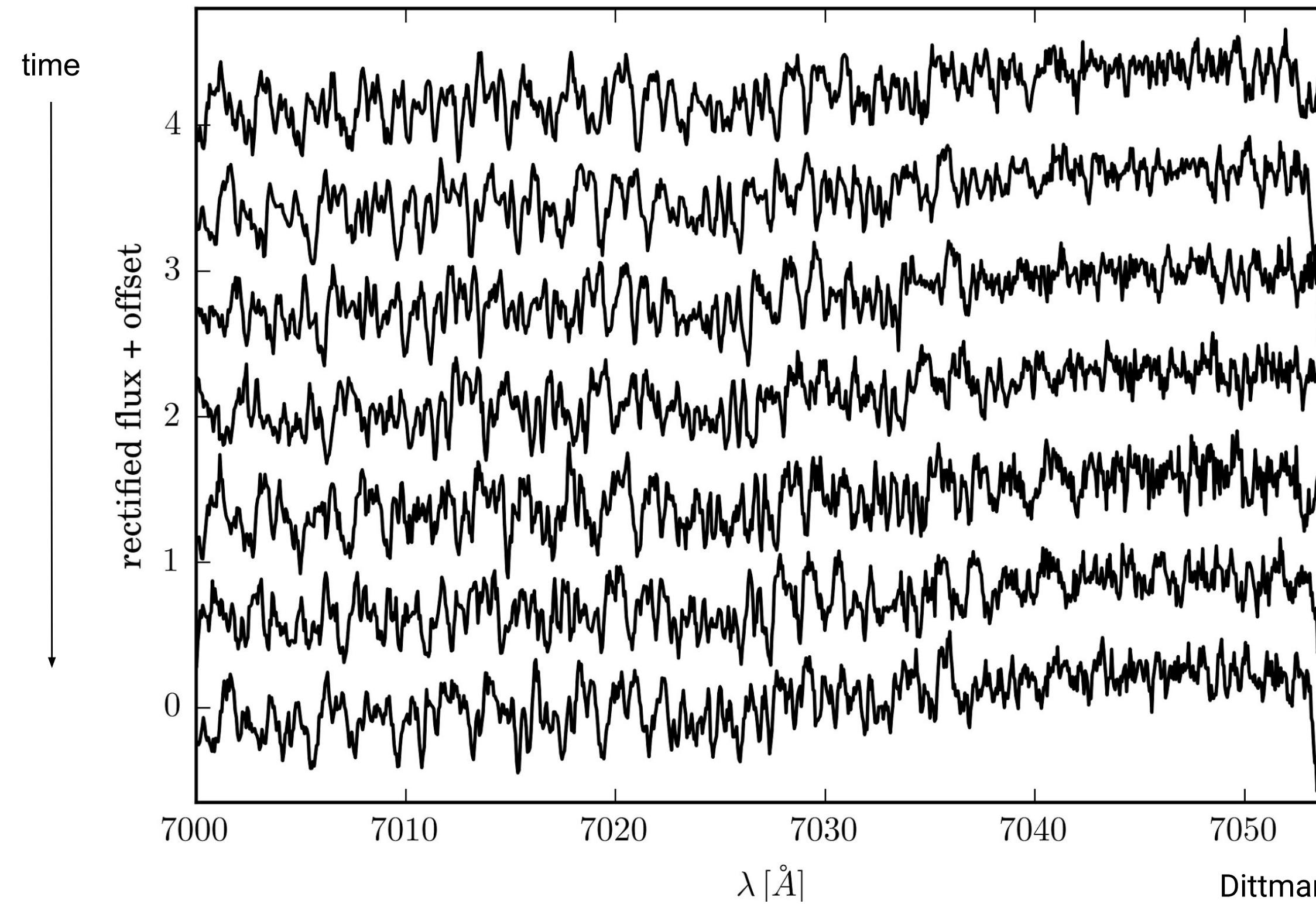
$$\alpha = (a_f, l_f, a_g, l_g)$$

4-dim GP hyperparameters = controlling the amplitude and smoothness of Gaussian Process prior on latent spectra

# Knowns (Data)

## Raw Observations of the LP661-13 M4 Binary

**D** =



# Bayesian Inference

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

In this case:

$$\begin{aligned} P(\Theta, f, g, \alpha | D) &\propto \\ P(D | \Theta, f, g, \alpha) \times P(\Theta, f, g, \alpha) \end{aligned}$$

a probability density on (4+7+ $\infty$ )-dim parameter space

# Bayesian Computation

1. Run Markov Chain Monte Carlo (MCMC)  
(e.g. *emcee* affine-invariant ensemble sampler)  
on the 4+7 small dimensional marginal posterior

$$P(\Theta, \alpha | D) = \int df \int dg P(\Theta, f, g, \alpha | D)$$

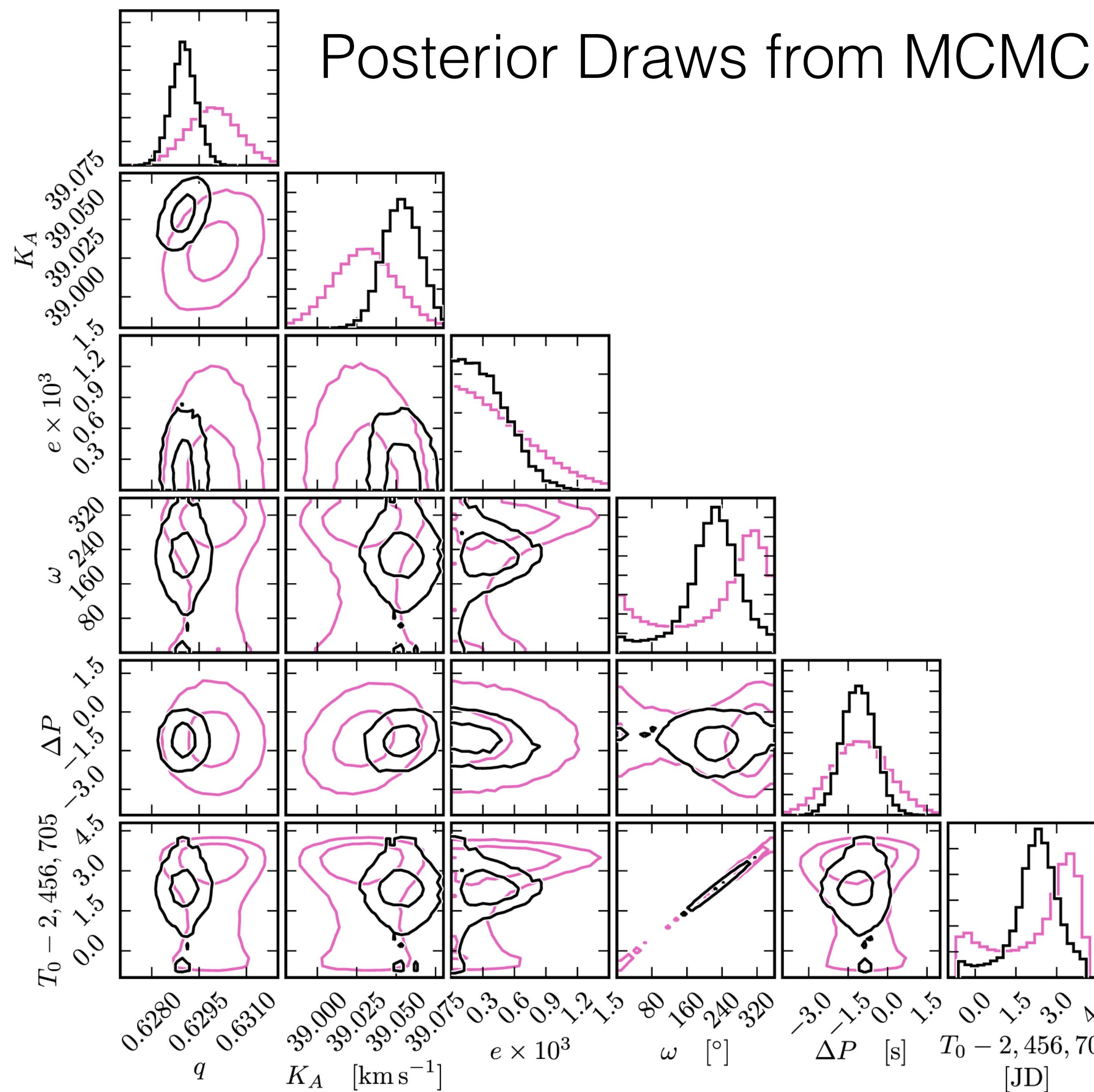
MCMC generates samples:  $\Theta_i, \alpha_i \sim P(\Theta, \alpha | D)$

2. Draw high-dim (**f**, **g**) spectra from the posterior predictive distribution

$$f_i, g_i \sim P(f, g | \Theta_i, \alpha_i, D)$$

# Application to the Mid-M-Dwarf Binary LP661-13

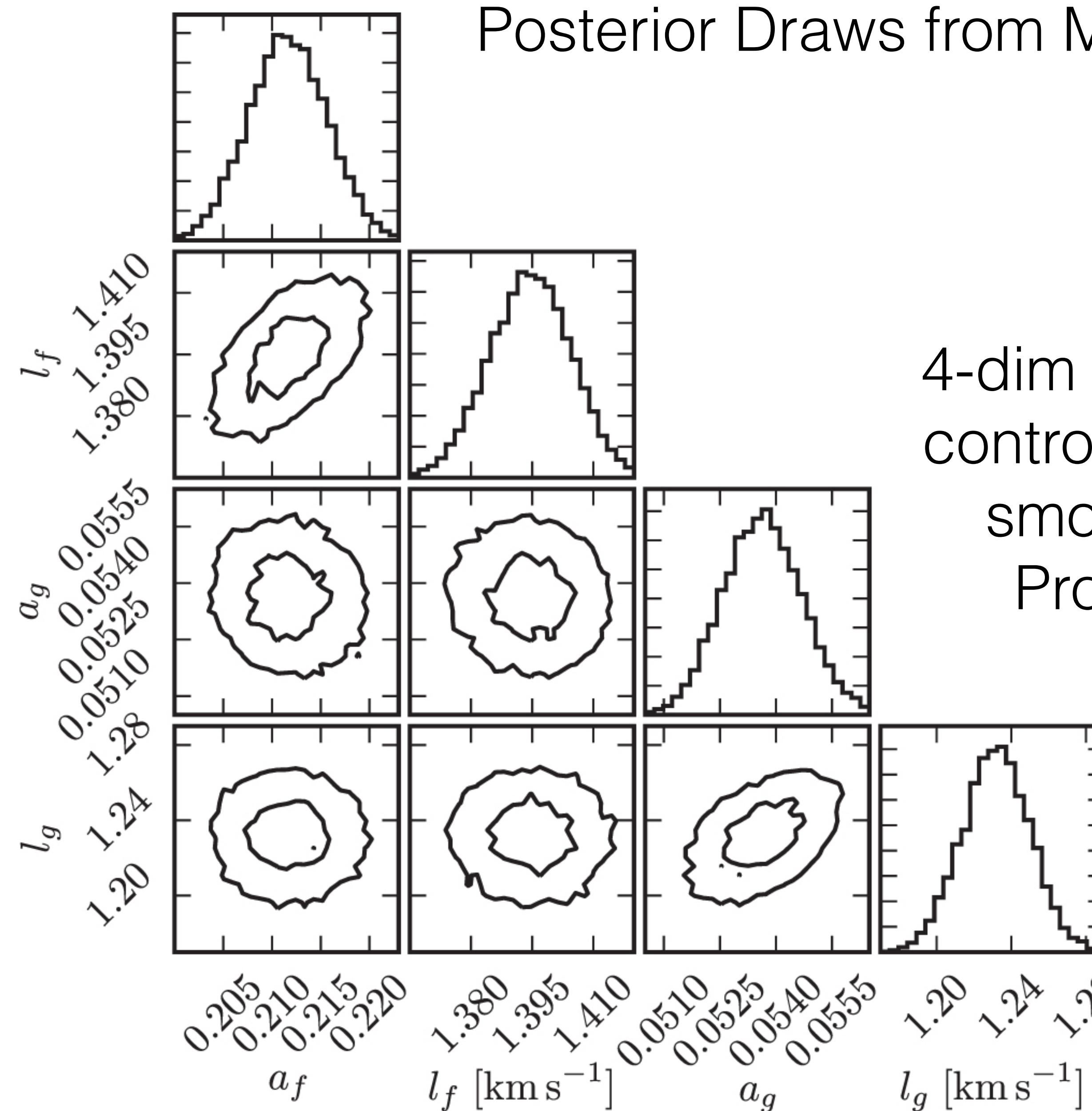
83



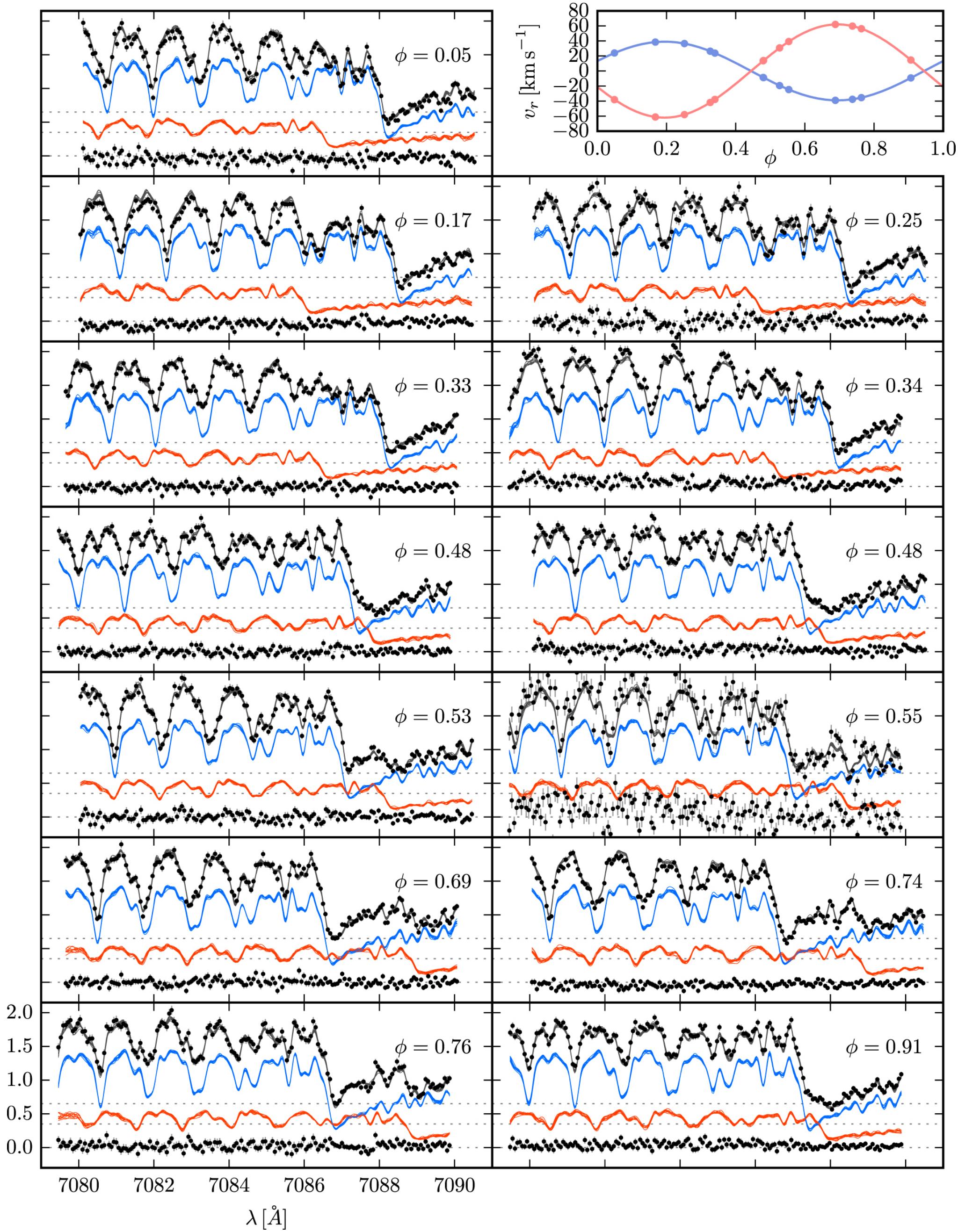
# Application to the Mid-M-Dwarf Binary LP661-13

Posterior Draws from MCMC     $\alpha =$

$$(a_f, l_f, a_g, l_g)$$



4-dim GP hyperparameters =  
controlling the amplitude and  
smoothness of Gaussian  
Process prior on latent  
spectra



## Posterior Inference of Component Spectra (f, g)

Compared to 10 epochs of observed spectra (**data**)

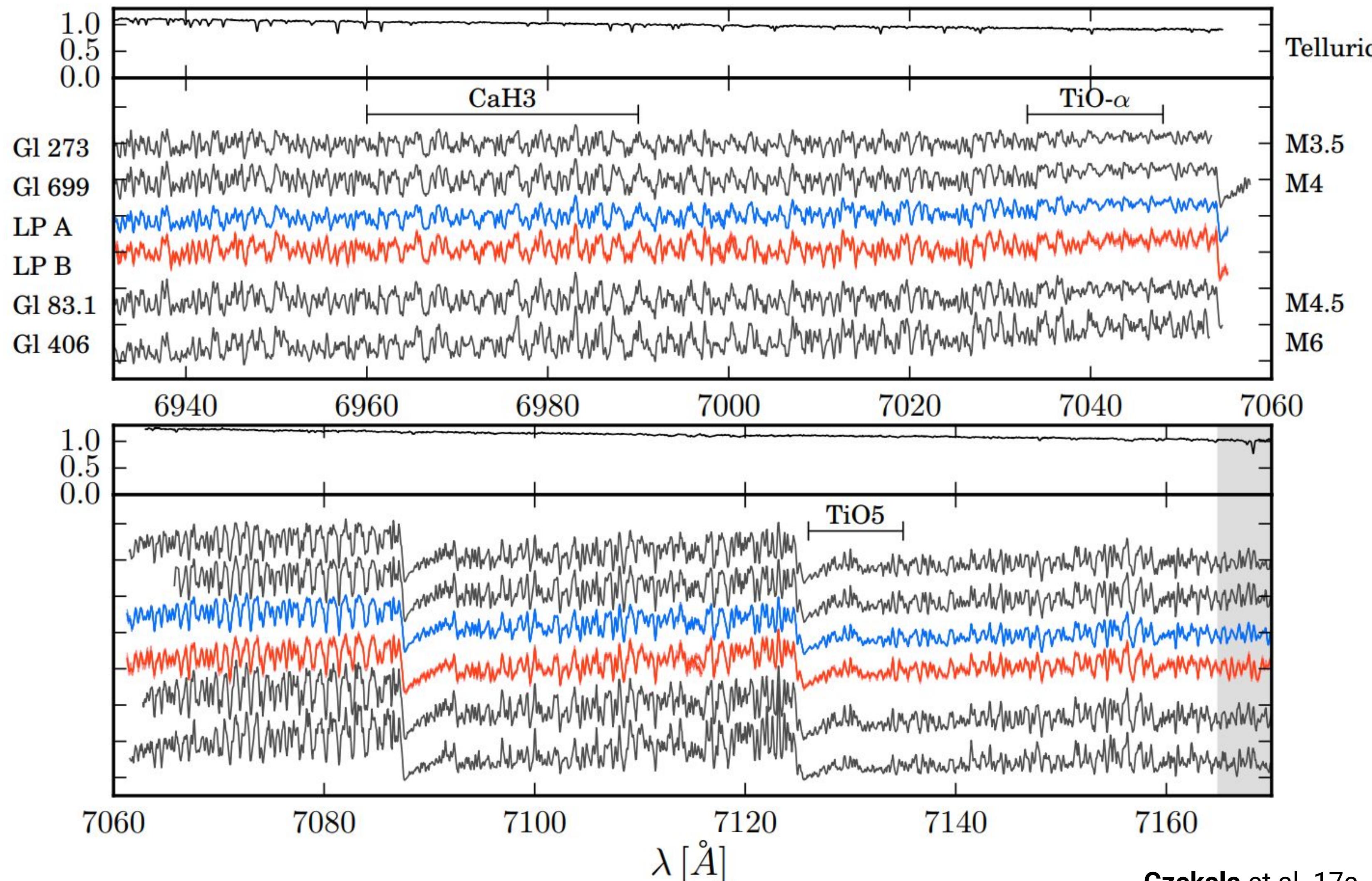
Model Checking!  
Checking Fit against Data

# Model Checking!

86

## Checking Fit against Domain Knowledge (astrophysics)!

**Disentangled spectra match other single standard stars**



*Astrostatistics Case Study:*  
Disentangling Time Series Spectra with Gaussian  
Processes: Applications to Radial Velocity Analysis  
(Czekala et al. 2017, arXiv:1702.05652)

<http://psoap.readthedocs.io/en/latest/>

- Statistics:
  - Parametric Modelling (Stellar Orbit Parameters)
  - Nonparametric Modelling (Gaussian Process Spectrum)
  - Bayesian Inference (probability of unknowns given data)
  - Markov Chain Monte Carlo (computing posterior probability)
- Astronomy:
  - Applications to Radial Velocity Analysis of Stars/Exoplanets