Bayesian Stats

Chapter 5

Pros for Bayesianism

- Symmetry and Grand Unification
  - Allows one to make probability statements about parameters. Symmetry.
- Extra information
  - Has a natural and convenient way to enter the calculation – in the prior.
- Honesty/disclosure
  - The subjective priors are good because it forces transparency.. We always have prior beliefs anyways, just make them explicit.
- More elegant in practice

Cons for Bayesianism

- Are we really being scientific?
  - "Credible regions" are not true confidence regions. Estimates based on unobservable information.
- The effect of the prior is always there
  - Even in an extremely simple model, a prior must be provided. And even an uninformative prior can effect the resulting PDF.
- Unnecessarily complicated and computationally expensive
  - Can require computationally intractable integrals
- Unnecessarily brittle and limiting
  - Sensitive to incorrect likelihood function or outliers in data.

Only **two** rules of probability

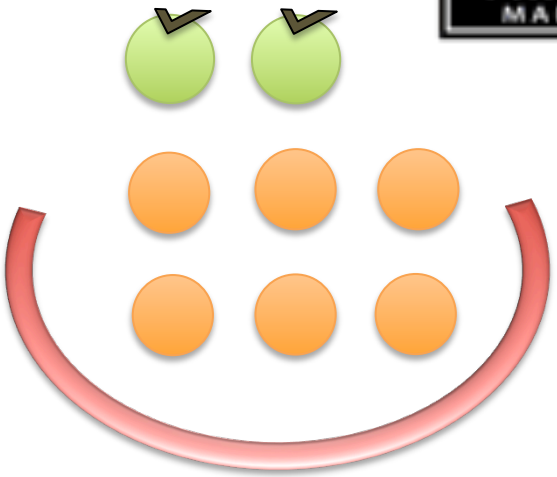
$$p(X) = \sum_{Y} P(X|Y)$$ "Sum rule"

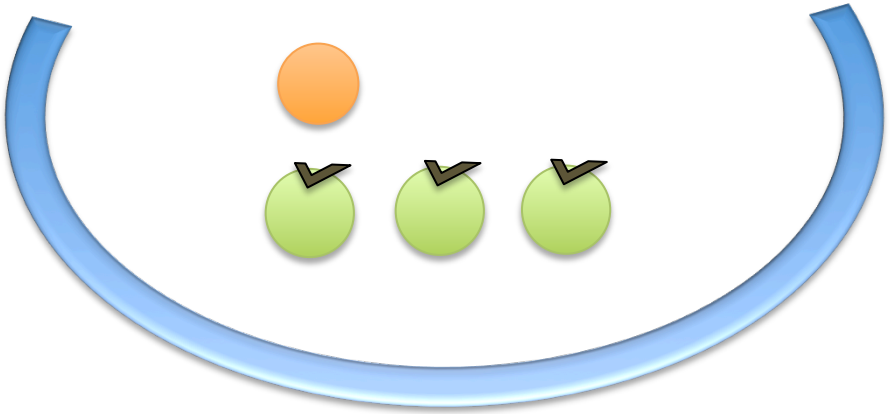
$$p(X, Y) = P(X|Y) \times P(Y)$$ "Product rule"

The rest of Bayesian stats is the **application** of these rules with the added insight that concepts of "certainty" and "belief" can be represented as probabilities, and therefore follow these rules also.

Apple

Orange





Red box

Blue box

One observation samples two random variables: The box *B*, and the fruit *F*.

$F \in \{a, o\}$ "Fruit" = apple or orange

$B \in \{r, b\}$ "Box" = red or blue

Blue box is larger: **Prior** Probabilities

$$P(B = r) = 4/10$$

$$P(B = b) = 6/10$$

Conditional probabilities

$$P(F = a|B = r) = 1/4$$
$$P(F = o|B = r) = 3/4$$
$$P(F = a|B = b) = 3/4$$
$$P(F = o|B = b) = 1/4$$

“Chances of drawing an apple from the red box”

(product rule)

$$p(X,Y) = P(X|Y) \times P(Y) \quad \text{(product rule)}$$

$$p(F = a, B = r) = P(F = a|B = r) \times P(B = r)$$

$$= \frac{1}{4}\frac{4}{10} = \frac{1}{10}$$

"What are the chances of drawing an apple at random?"

(Sum rule)

$$p(X) = \sum_Y P(X|Y)$$

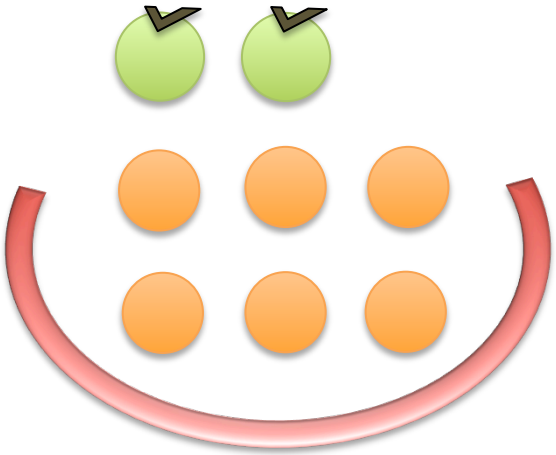
$$P(F = a) = P(F = a|B = r) + P(F = a|B = b)$$

$$= \frac{1}{4}\frac{4}{10} + \frac{3}{4}\frac{6}{10}$$

$$= \frac{11}{20}$$

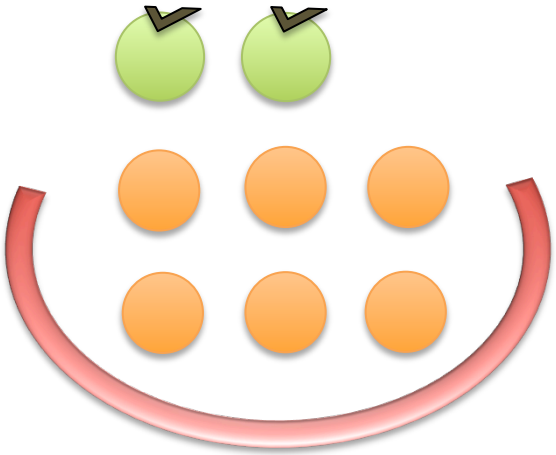
Correspondingly,

$$(p(F = o) = 9/20)$$



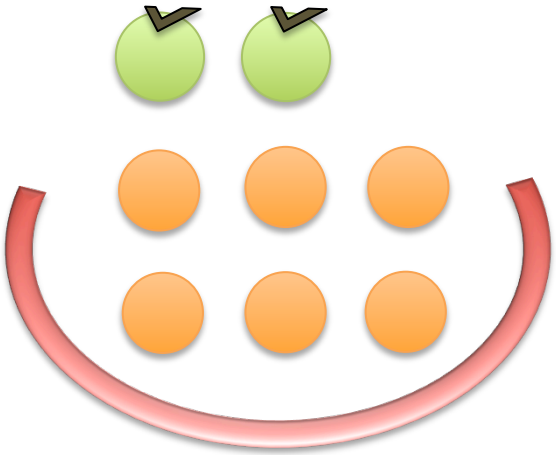
Team Work Time:

“Given that I’ve drawn an apple, what are the chances it came from the red box?”



"Given that I've drawn an orange, what's the chances it came from the red box?"

$$
\begin{aligned}
P(B|F) &= P(B,F)/P(F) = P(F,B)/P(F) \\
&= P(F|B)P(B)/P(F) \\
&= P(F=o|B=r)P(B=r)/P(F=o) \\
&= \frac{3}{4}\frac{4}{10}\frac{20}{9} = \frac{2}{3}
\end{aligned}
$$



Hypothesis testing

- Which models fits the data better?
- Need to supply alternate model – cannot rule out model without alternative in Bayesian framework
- Bayesian Evidence = Probability that the data were generated by a given model

$$E(M) \equiv p(D|M,I) = \int p(D|M,\theta,I)\, p(\theta|M,I)\, d\theta$$

Likelihood Prior

Hypothesis testing

- Odds ratio = "Which model is most likely to have generated this data?"

$$O_{21} = \frac{p(M_2|D,I)}{p(M_1|D,I)}$$

$$O_{21} = \frac{p(M_2|D,I)}{p(M_1|D,I)} = \frac{E(M_2)\,p(M_2|I)}{E(M_1)\,p(M_1|I)}$$

Hypothesis testing: example

- To test whether a Gaussian (M1) or a Box (M2) function is the correct model for given data (x, y)
- Assuming we have no prior information on the validity of box or Gaussian, then:

$$E_1 = \int\int\int e^{(y-\left[\theta_1\, e^{-(x-\theta_2)^2/2\theta_3^2}\right])^2/2\sigma^2} d\theta_1 d\theta_2 d\theta_3$$

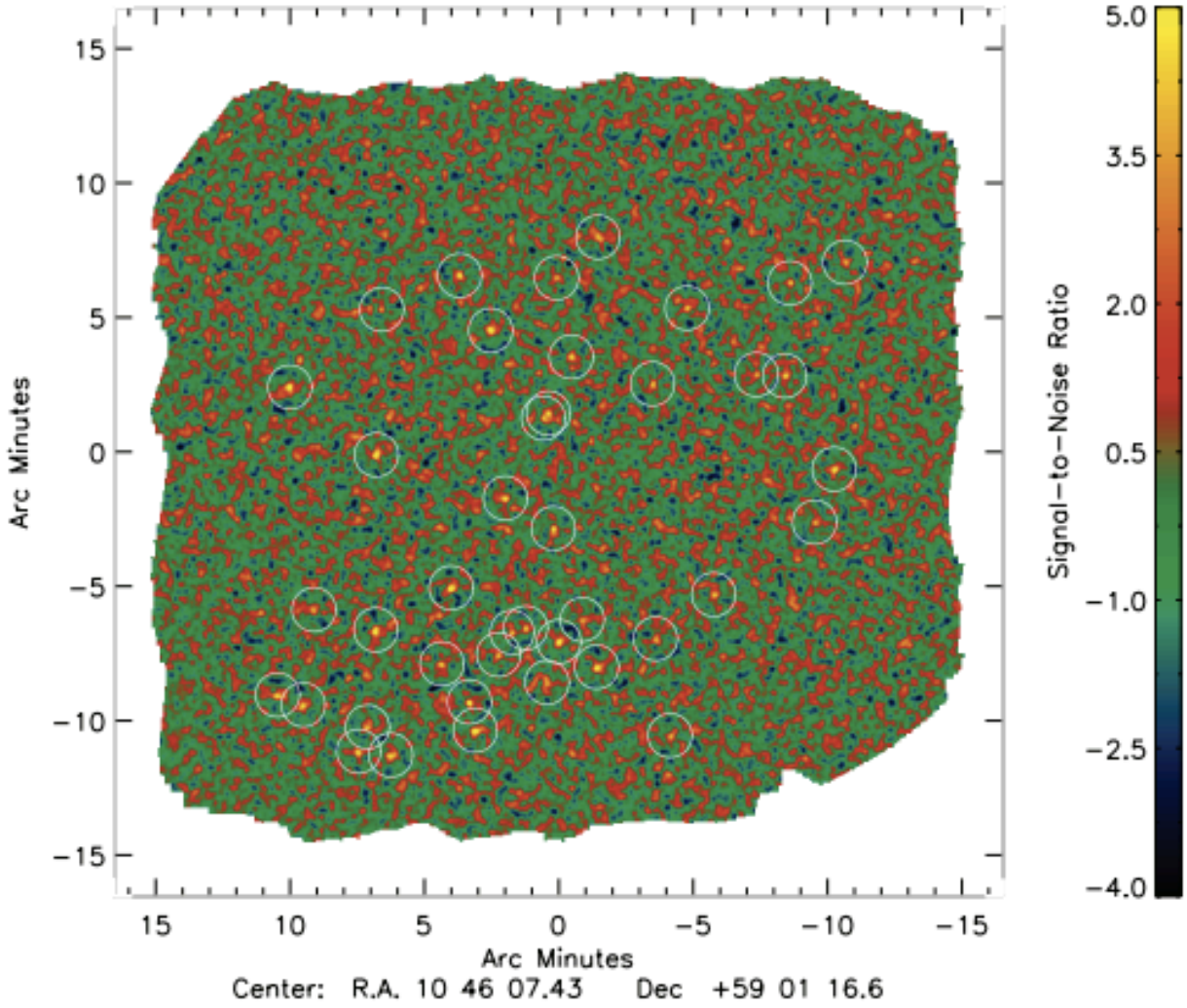
$$E_2 = \int\int\int e^{(y-[Box(\theta_1,\theta_2,\theta_3)])^2/2\sigma^2} d\theta_1 d\theta_2 d\theta_3$$

- Divide and inspect. Integral is awful for large number of parameters: use MCMC.

Realistic Astrophysics Example

Wide deep imaging.
Faint sources detected down to a given sensitivity Threshold.

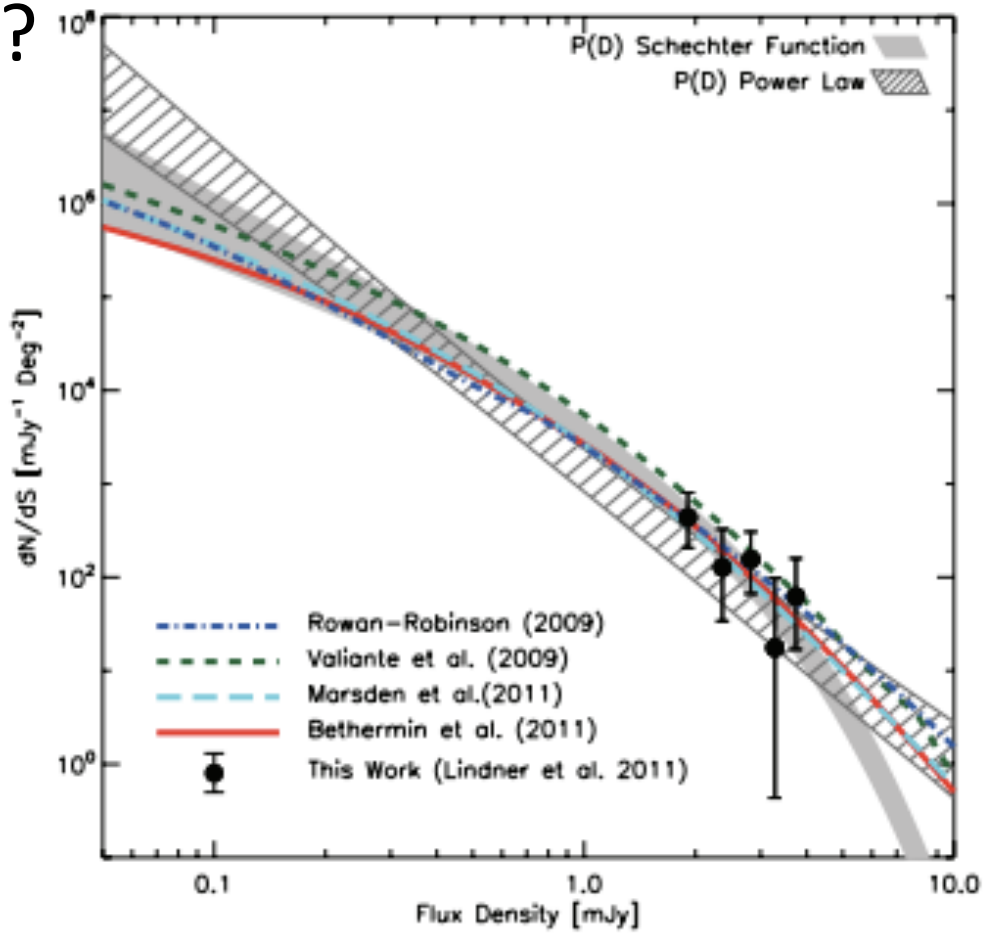
Right: Lockman Hole North Field.
1.2mm image using Max Planck Milimieter Bolometer Array



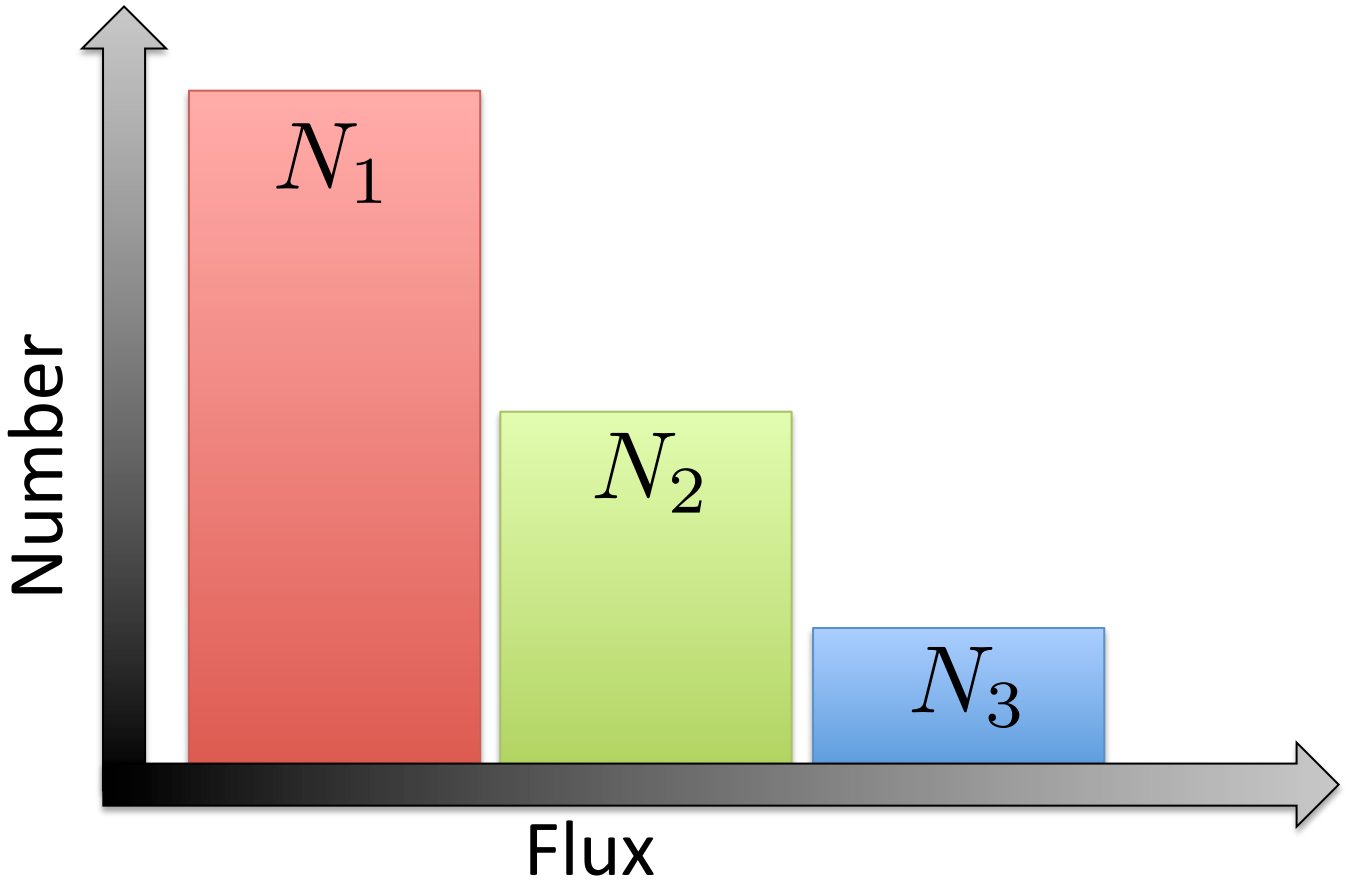
- Q: What is the true flux density of a source in my image, given its measured flux density?

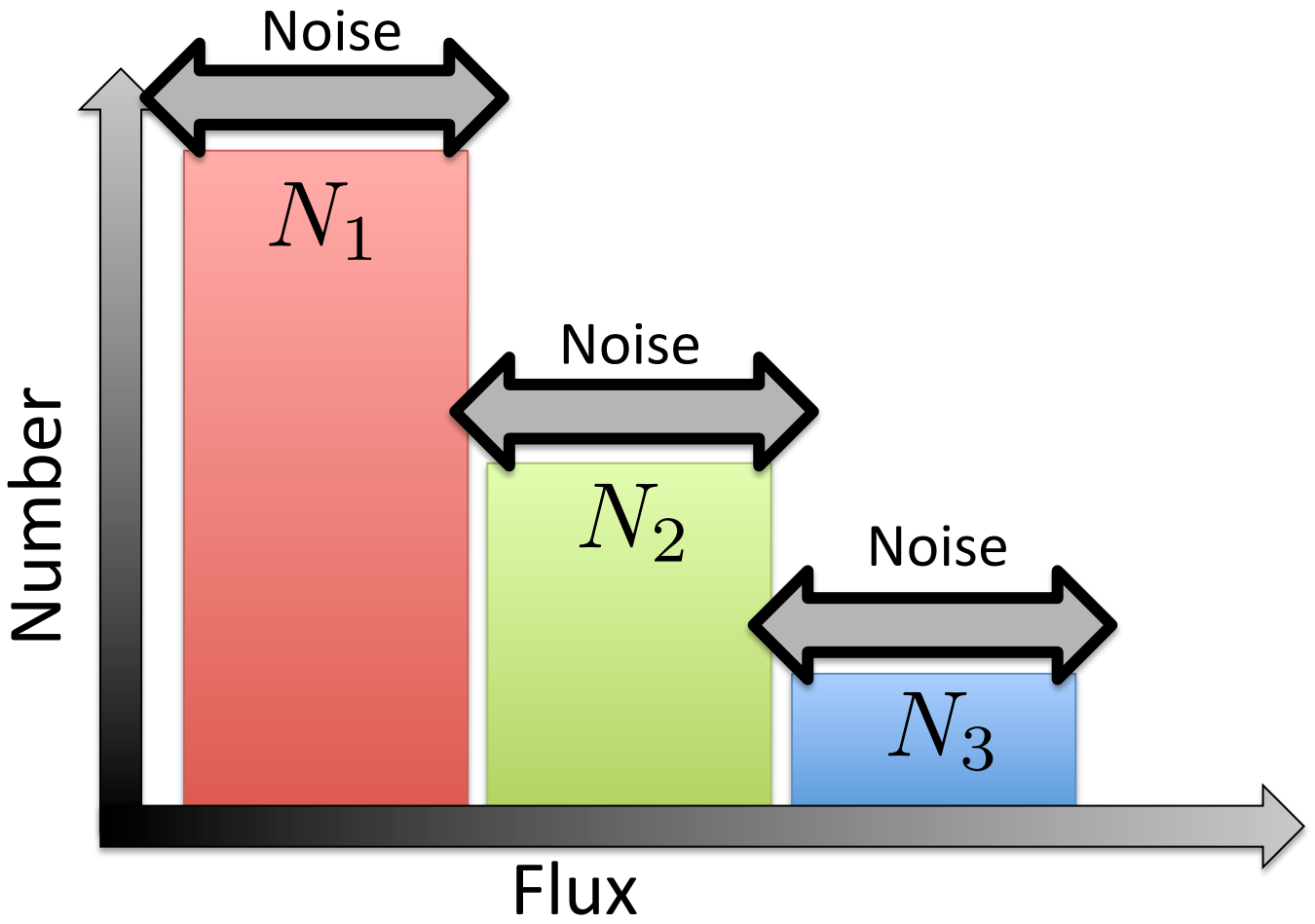
Why would they be different?

Why is that important?



- Histogram of ideal “ground truth” source fluxes

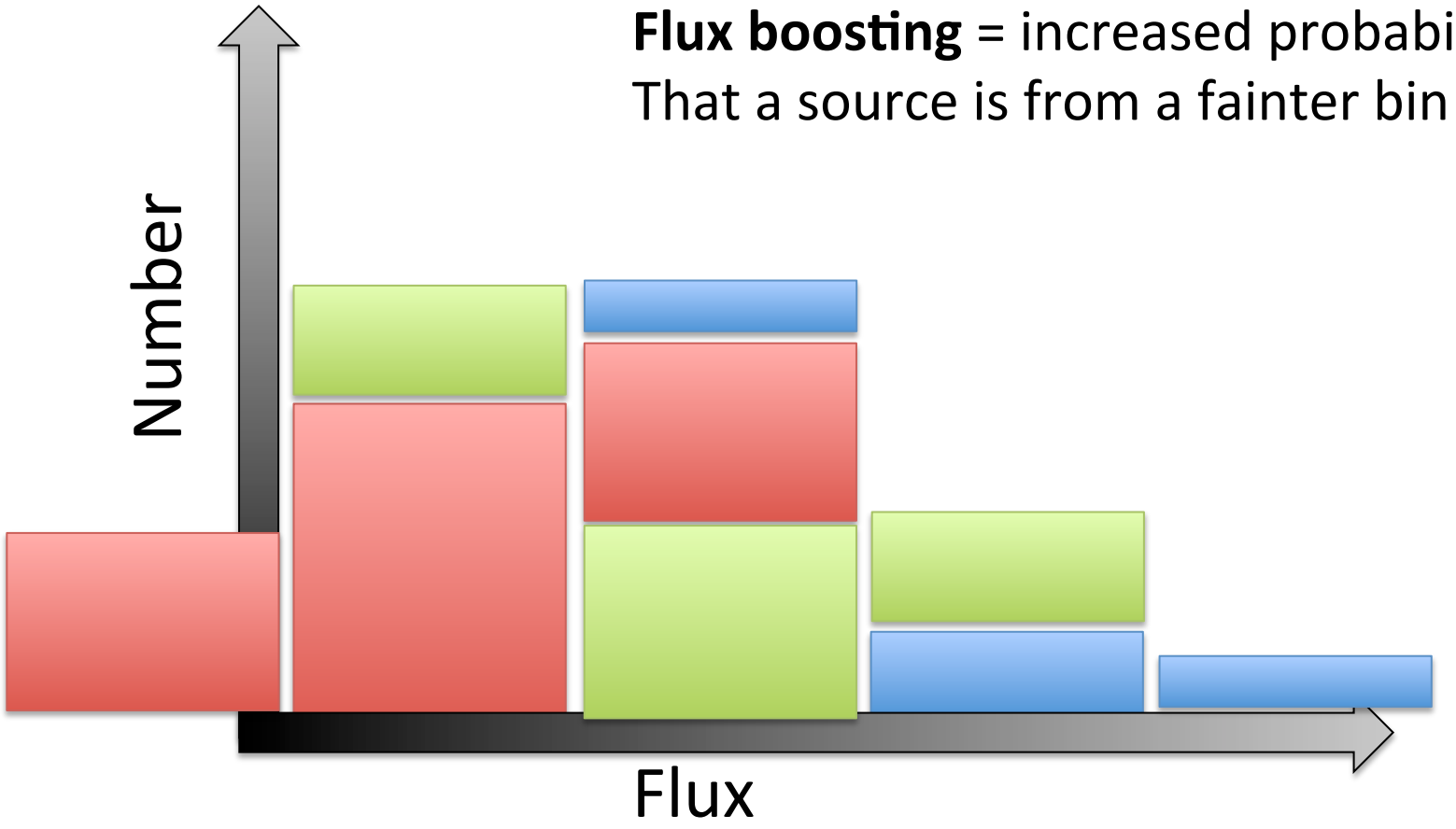



Noise
N_1
Noise
N_2
Noise
N_3
Number
Flux

Histogram after the effects of uniform image noise

Eddington bias = increased number counts

Flux boosting = increased probability
That a source is from a fainter bin



The optimal correction can be derived using Bayesian statistics

Likelihood function

Prior probability of obtaining true flux

$$P(S_{\rm true}|S_{\rm obs}) = \frac{P(S_{\rm obs}|S_{\rm true})\,P(S_{\rm true})}{P(S_{\rm obs})}$$

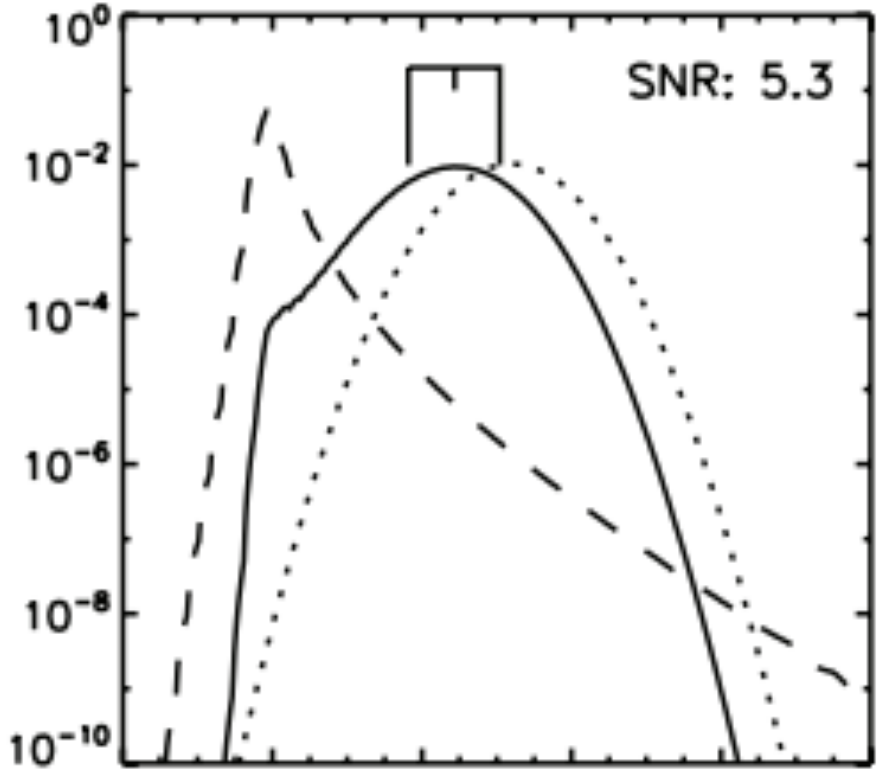
Bayesian evidence (normalization factor)

$$P(S_{\rm obs}|S_{\rm true}) = e^{\frac{-(S_{\rm obs}-S_{\rm true})^2}{2\sigma^2_{\rm obs}}}$$

$$P(S_{\rm true}) = N_0 \left(\frac{S_{\rm true}}{S_0}\right)^{-\alpha}$$

Dotted = Likelihood
Dashed = Prior (Related to number counts)
Solid = Posterior probability

True flux is probably less than observed flux



Lindner et al. (2011)

Computer time

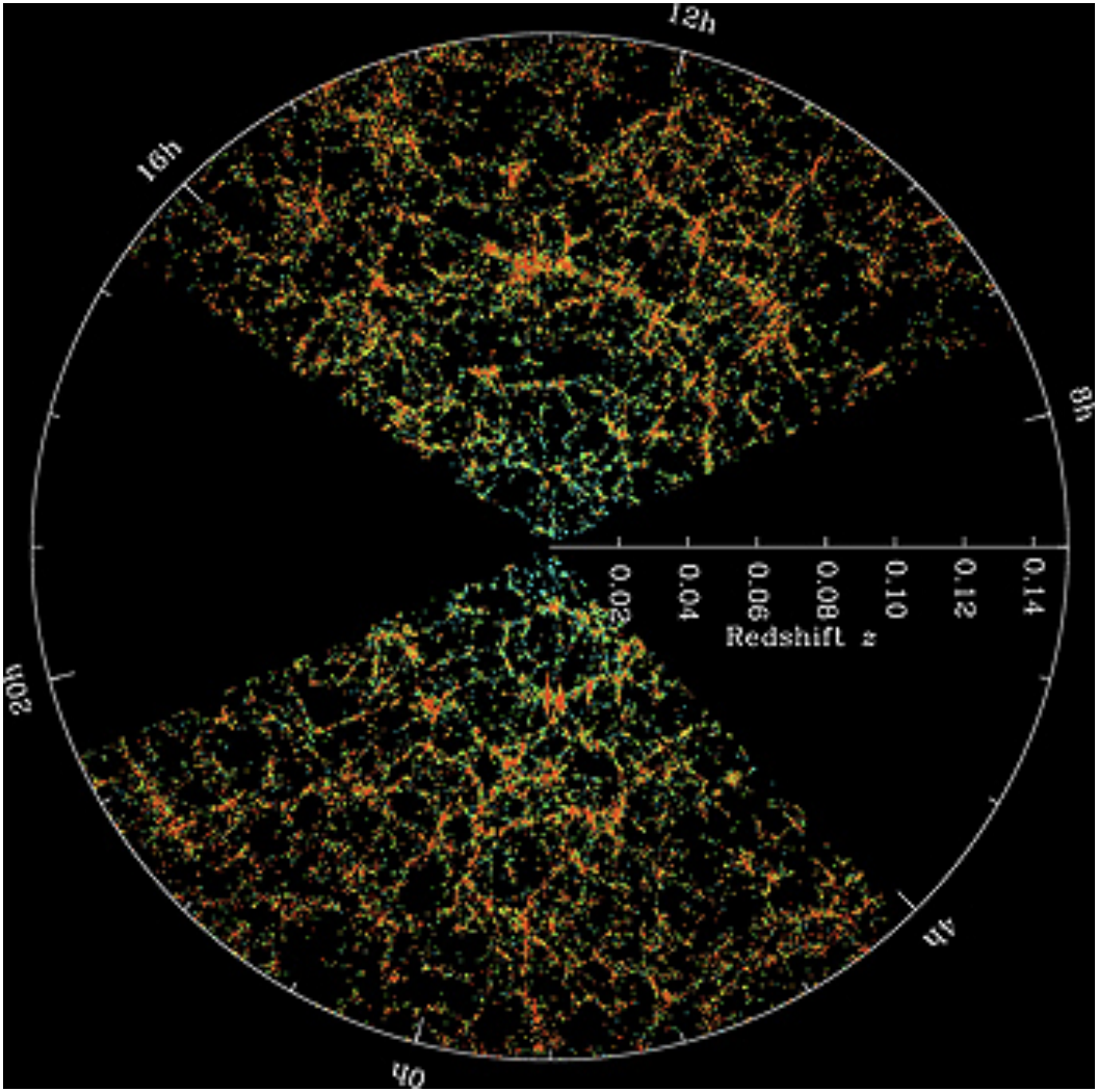
Chapter 6

Structure in point data

Point data:

SDSS Galaxy Redshifts

(ra, dec, *z*)



More point data

More point Data.
(X, Y, Z, t)

Anything that can be organized into a catalog is point data.

ID	Source Name	$S_\nu^{\rm Best}$ (mJy)	$S_\nu^{\rm Full}$ (mJy)	$S_\nu^{\rm Deboosted}$ (mJy)	$P(<0$
1	MM J104700.1+590109	3.7 ± 0.8	4.1 ± 0.6	$3.5^{+0.6}_{-0.6}$	<0.0
2	MM J104627.1+590546	4.5 ± 0.8	4.7 ± 0.7	$3.8^{+0.7}_{-0.7}$	<0.0
3	MM J104631.4+585056	6.1 ± 1.8	4.7 ± 0.7	$3.8^{+0.8}_{-0.7}$	<0.0
4	MM J104607.4+585413	2.8 ± 0.8	3.2 ± 0.5	$2.7^{+0.5}_{-0.5}$	<0.0
5	MM J104725.2+590339	4.9 ± 0.9	5.2 ± 0.8	$4.0^{+0.8}_{-0.9}$	<0.0
6	MM J104638.4+585613	3.1 ± 0.7	2.7 ± 0.5	$2.3^{+0.4}_{-0.4}$	<0.0
7	MM J104700.1+585439	2.8 ± 0.7	2.8 ± 0.5	$2.3^{+0.4}_{-0.5}$	<0.0
8	MM J104633.1+585159	4.5 ± 1.2	3.4 ± 0.6	$2.7^{+0.6}_{-0.6}$	<0.0
9	MM J104704.9+585008	5.6 ± 1.5	5.1 ± 0.9	$3.8^{+1.0}_{-0.9}$	<0.0
10	MM J104622.9+585933	3.6 ± 0.7	2.9 ± 0.5	$2.4^{+0.5}_{-0.5}$	<0.0
11	MM J104556.5+585317	3.5 ± 0.9	3.4 ± 0.6	$2.7^{+0.6}_{-0.6}$	<0.0
12	MM J104448.0+590036	5.1 ± 0.9	3.5 ± 0.6	$2.7^{+0.6}_{-0.7}$	<0.0
13	MM J104609.0+585826	2.7 ± 0.7	2.7 ± 0.5	$2.1^{+0.5}_{-0.5}$	<0.0
14	MM J104636.1+590749	4.3 ± 0.8	4.3 ± 0.8	$3.0^{+0.9}_{-0.9}$	<0.0
15	MM J104738.3+585213	5.1 ± 1.1	4.5 ± 0.9	$3.0^{+0.9}$	

Density estimation

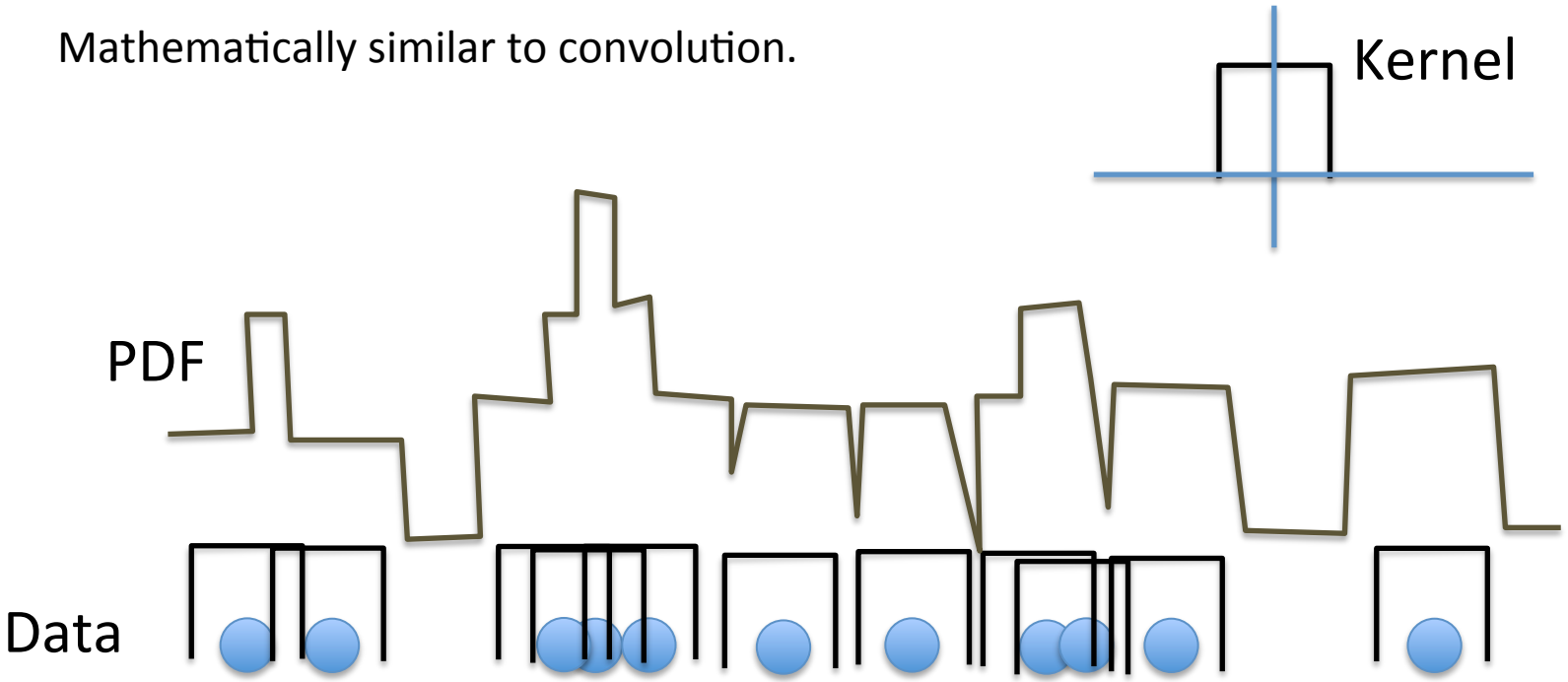
- Histograms (previously discussed)
  - Arbitrary choice of edges. Results could change when edges change. Original data can be "thrown away", as histogram contains all information.
- Kernel Density Estimators
  - No arbitrary edge choices. Specification of kernel uniquely specifies the distribution. Original data need to be "carried around" and used every time the density function is to be estimated.

Kernel Density estimators

$$\text{estimator}(x) = \frac{1}{Nh} \sum_{x'} K\left(d(x - x')/h\right)$$

kernel K All values positive, unity normalization, zero mean

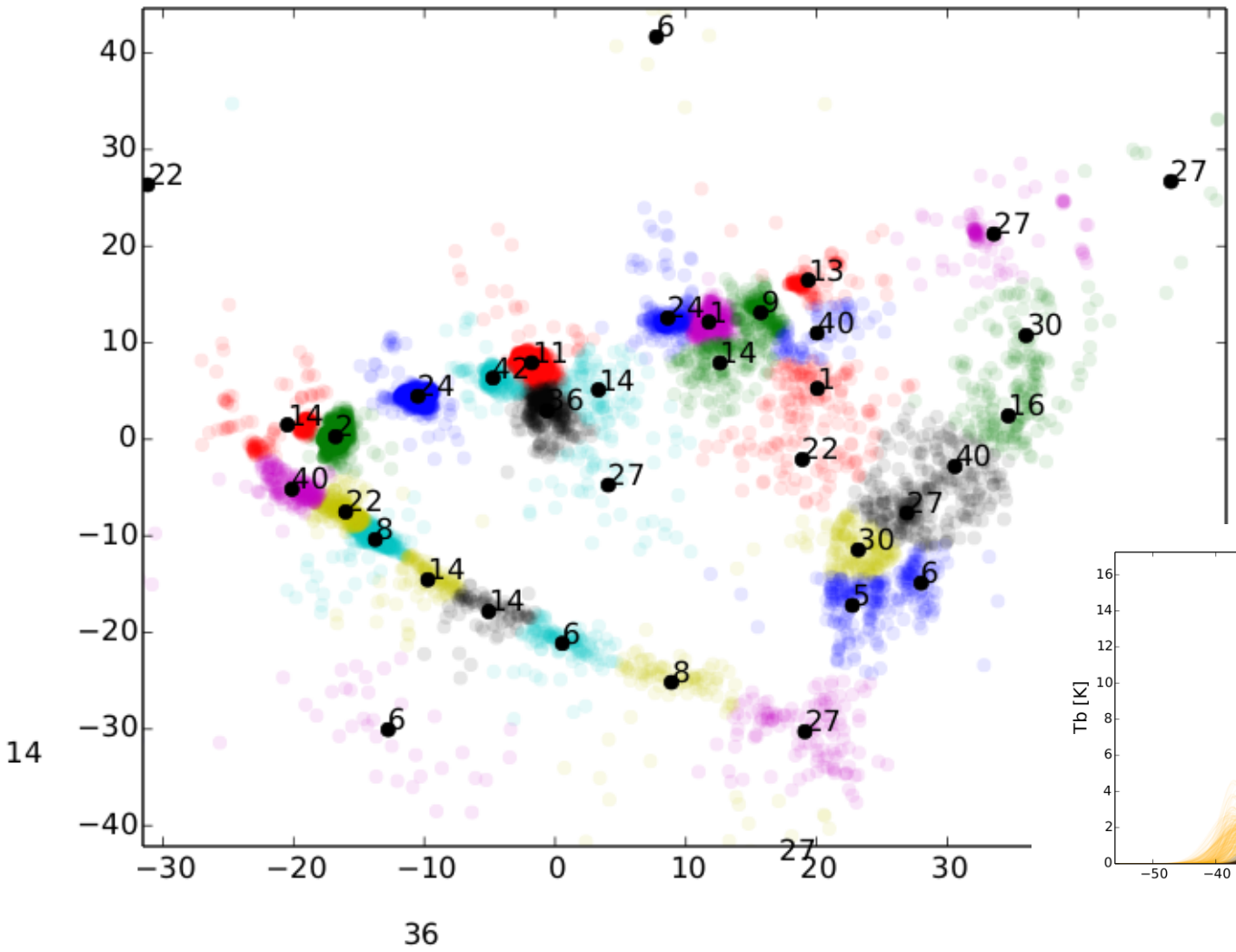
Mathematically similar to convolution.



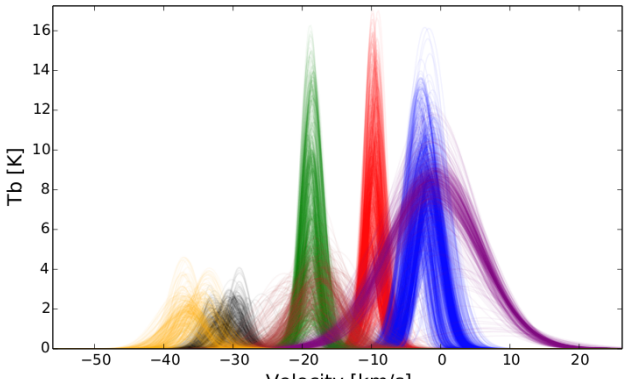
Clustering: K-means

- Expectation Maximization Algorithm:
- (0) Assign initial cluster locations
- (1) Assign labels to all points based on distance
- (2) Re-compute cluster locations using mean of all points in a given cluster
-
- http://www.bytemuse.com/post/k-means-clustering-visualization/

Example, K-means clustering



21cm Emission Data



Hierarchical clustering

- Procedural algorithm
- Initialize data of N sample to have N distinct clusters
- Merge nearest pairs of cluster until desired number of clusters is reached.
- Ability to find non-point-like morphologies

Example, Hierarchical clustering