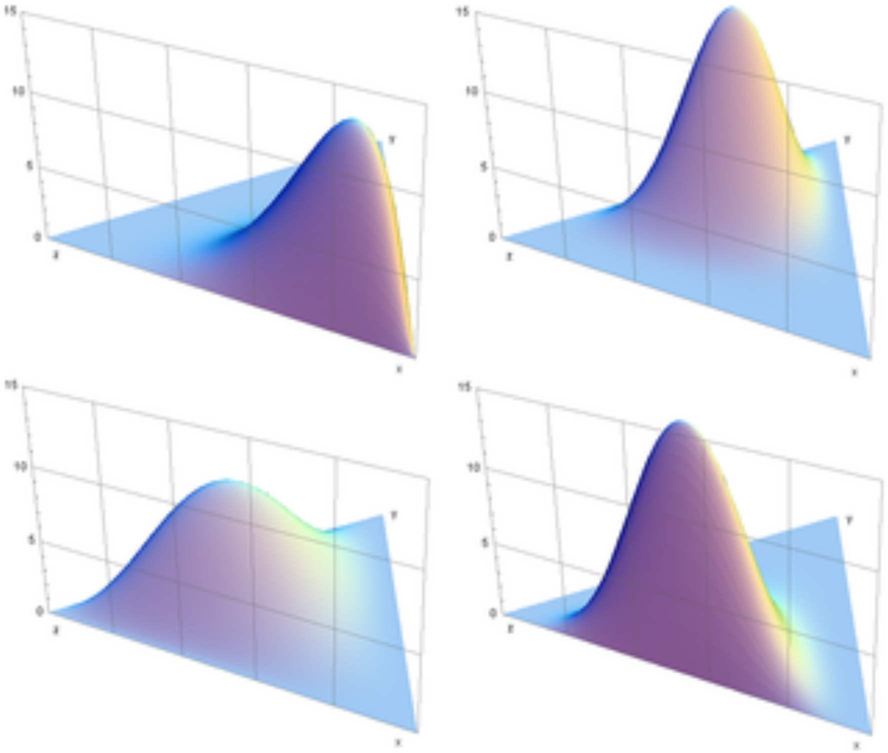
Digging into the Dirichlet

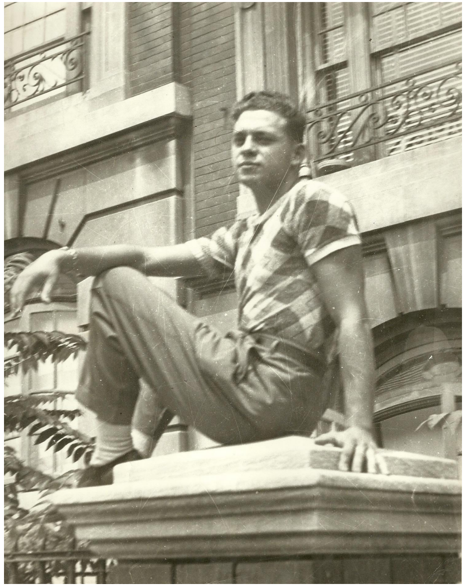


Max Sklar
@maxsklar

foursquare

New York Machine Learning Meetup
December 19th, 2013

Dedication



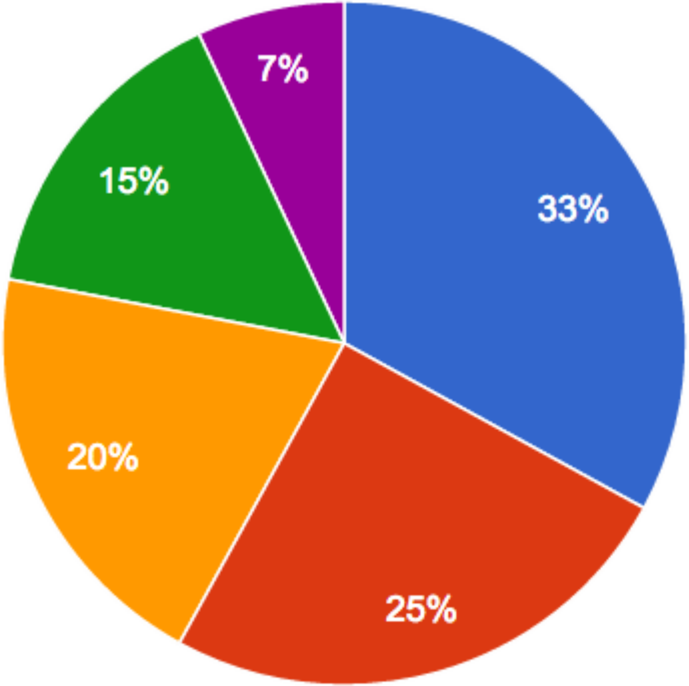
Meyer Marks
1925 - 2013

The Dirichlet Distribution

$$P(p|\alpha) = \frac{\Gamma\left(\sum_{k=0}^{K-1}\alpha_k\right)}{\prod_{k=0}^{K-1}\Gamma(\alpha_k)} \prod_{k=0}^{K-1} p_k^{\alpha_k - 1}$$

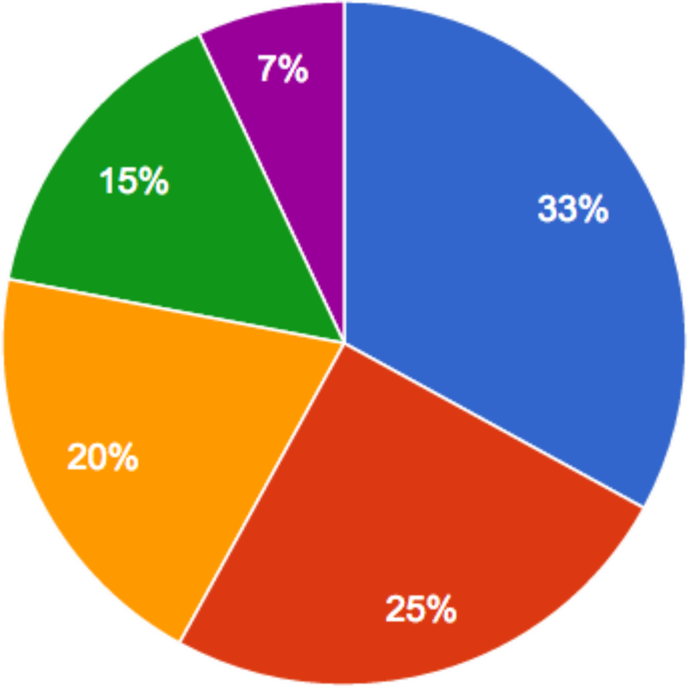
Let’s start with something simpler

Let’s start with something simpler



A Pie Chart!

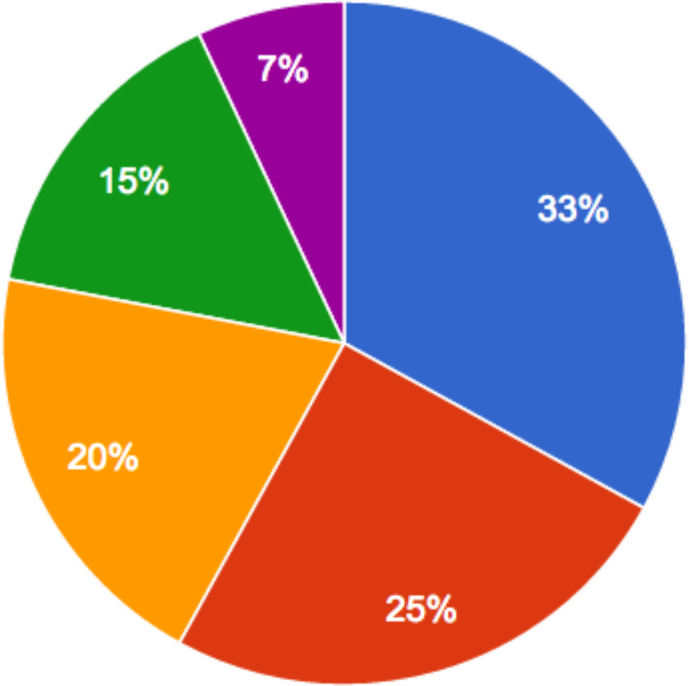
Let’s start with something simpler



A Pie Chart!

AKA
Discrete Distribution
Multinomial Distribution

Let’s start with something simpler



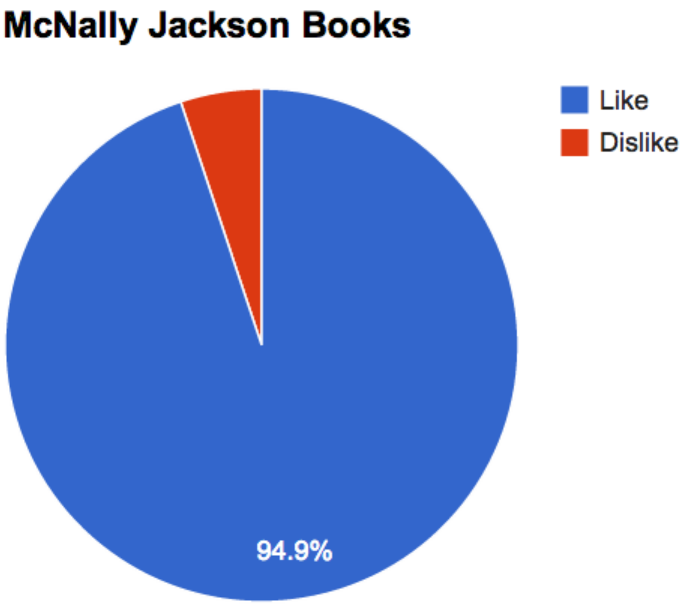
A Pie Chart!

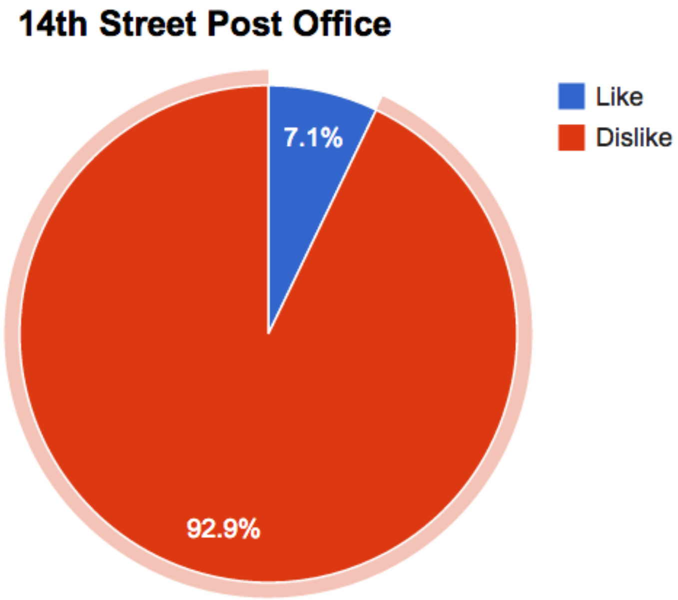
K = The number of categories.

K = 5

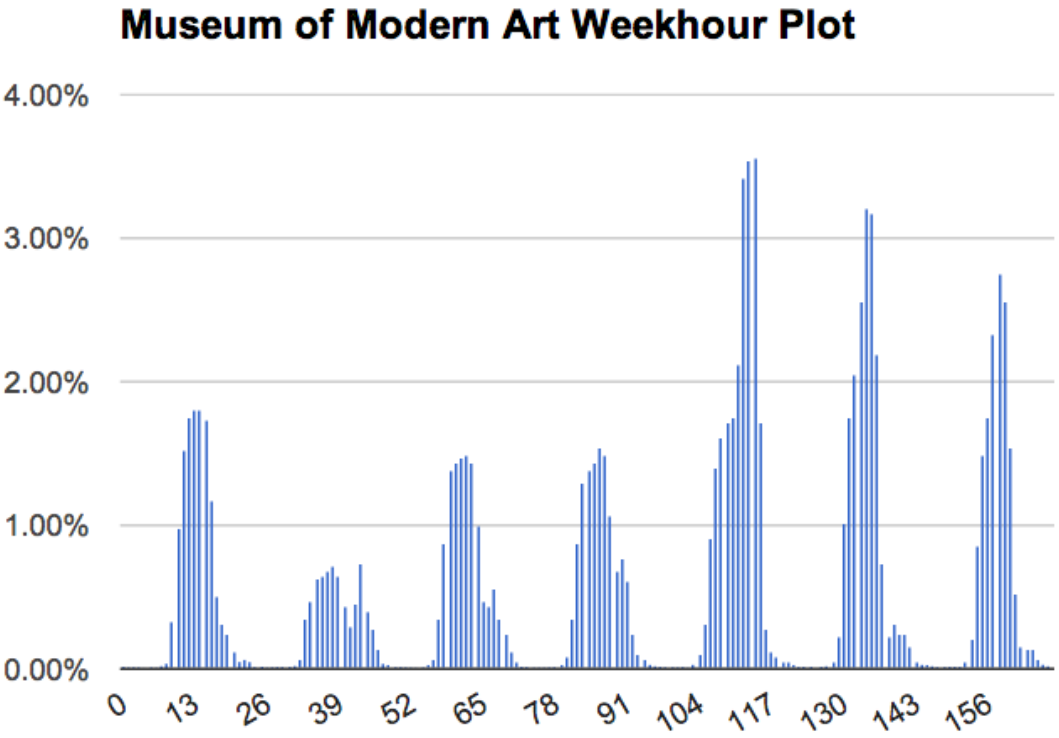
Examples of Multinomial Distributions

Examples of Multinomial Distributions

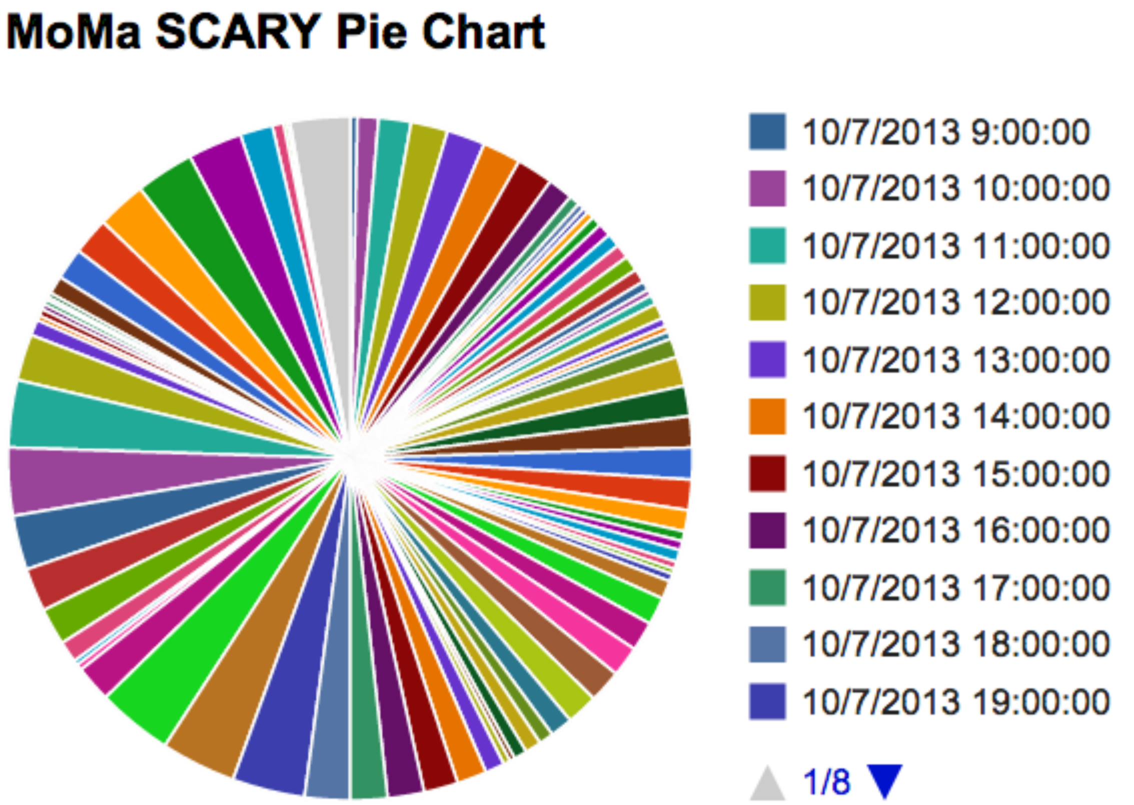




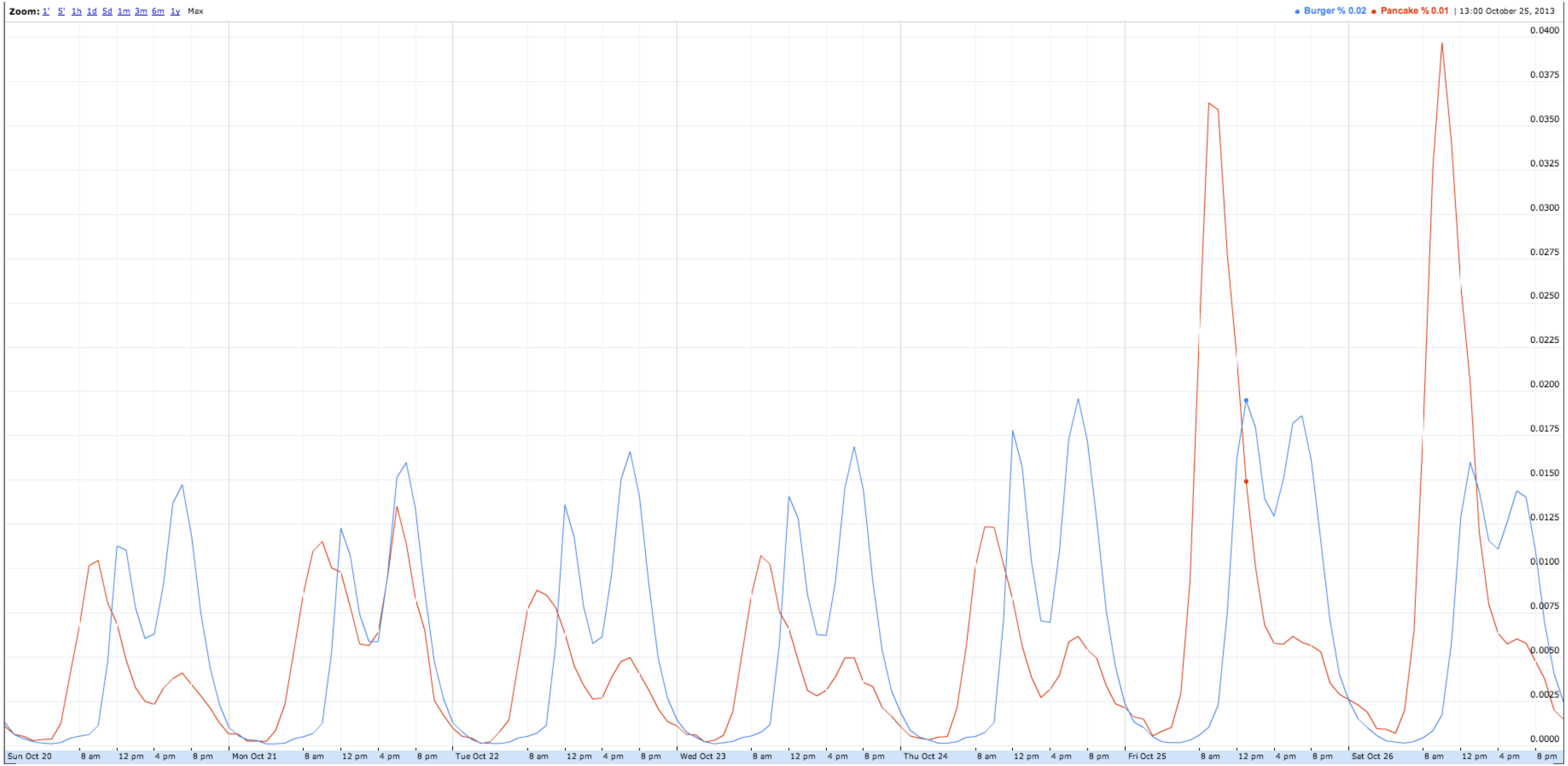
Examples of Multinomial Distributions



Examples of Multinomial Distributions



Examples of Multinomial Distributions



What does the raw data look like?

What does the raw data look like?

Counts!

id	# likes	# dislikes
1	231	23
2	81	40
3	67	9
4	121	14
5	9	31
6	18	0
7	1	1

What does the raw data look like?

More specifically:
- K columns of counts
- N rows of data

id	# likes	# dislikes
1	231	23
2	81	40
3	67	9
4	121	14
5	9	31
6	18	0
7	1	1

BUT...

Counts != Multinomial Distribution

BUT...

We can estimate the multinomial distribution with the counts, using the maximum likelihood estimate

366	181	203

BUT...

We can estimate the multinomial distribution with the counts, using the maximum likelihood estimate

Sum =
366 + 181 + 203 =
750

366	181	203

BUT...

We can estimate the multinomial distribution with the counts, using the maximum likelihood estimate

366 / 750
181 / 750
203 / 750

366	181	203

BUT...

We can estimate the multinomial distribution with the counts, using the maximum likelihood estimate

48.8%
24.1%
27.1%

366	181	203

BUT...

Uh Oh

366	181	203
1	2	1

BUT...

This column will be all Yellow right?

366	181	203
1	2	1
0	1	0

BUT...

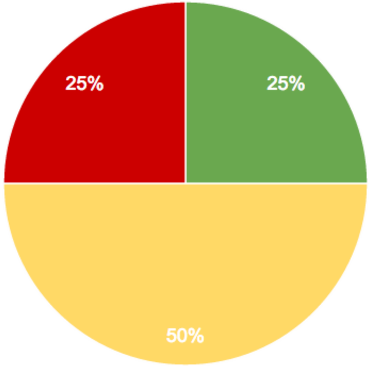
Panic!!!!

366	181	203
1	2	1
0	1	0
0	0	0

Bayesian Statistics to the Rescue

Bayesian Statistics to the Rescue

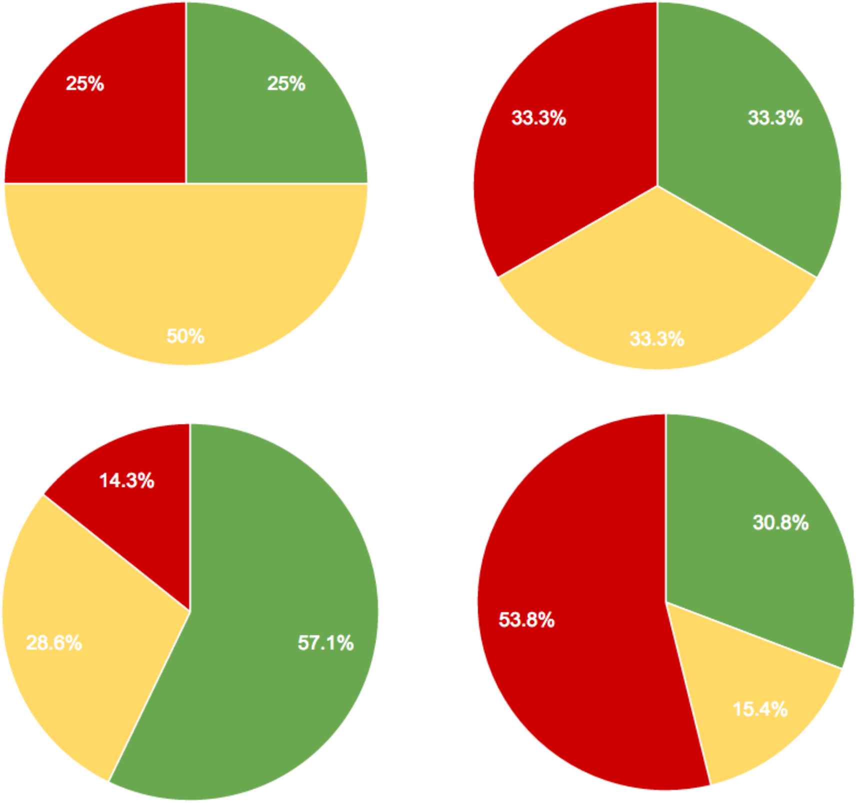
Still assume each row was generated by a multinomial distribution



Bayesian Statistics to the Rescue

Still assume each row was generated by a multinomial distribution

We just don't know which one!



The Dirichlet Distribution

Is a probability distribution over all possible multinomial distributions, p.

The Dirichlet Distribution

?

Represents our uncertainty over the actual distribution that created the row.

?

The Dirichlet Distribution

$$P(p|\alpha) = \frac{\Gamma\left(\sum_{k=0}^{K-1} \alpha_k\right)}{\prod_{k=0}^{K-1} \Gamma(\alpha_k)} \prod_{k=0}^{K-1} p_k^{\alpha_k - 1}$$

p: represents a multinomial distribution
alpha: the parameters of the dirichlet
K: the number of categories

Bayesian Updates

$$P(p|data) = \frac{P(data|p) * P(p|\alpha)}{P(p)}$$

Bayesian Updates

$$P(p|data) = \frac{P(data|p) * P(p|\alpha)}{P(p)}$$

Also a Dirichlet!

Bayesian Updates

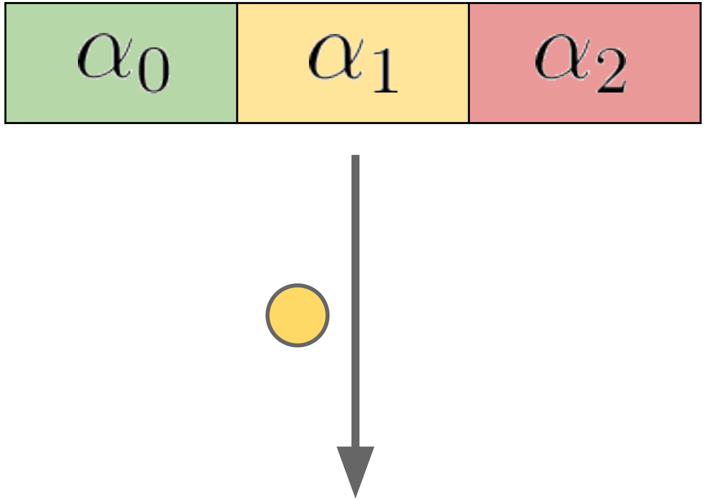
$$P(p|data) = \frac{P(data|p) * P(p|\alpha)}{P(p)}$$

Also a Dirichlet!
(Conjugate Prior)

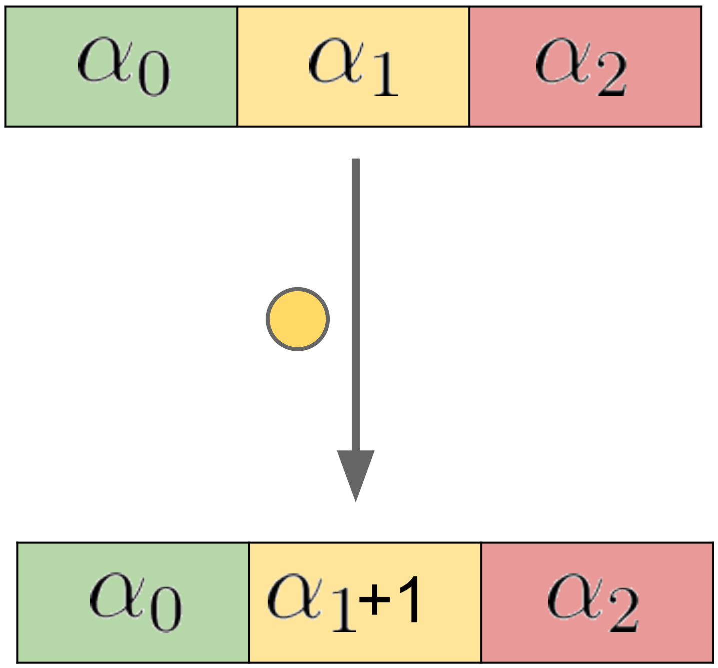
Bayesian Updates

α_0	α_1	α_2

Bayesian Updates

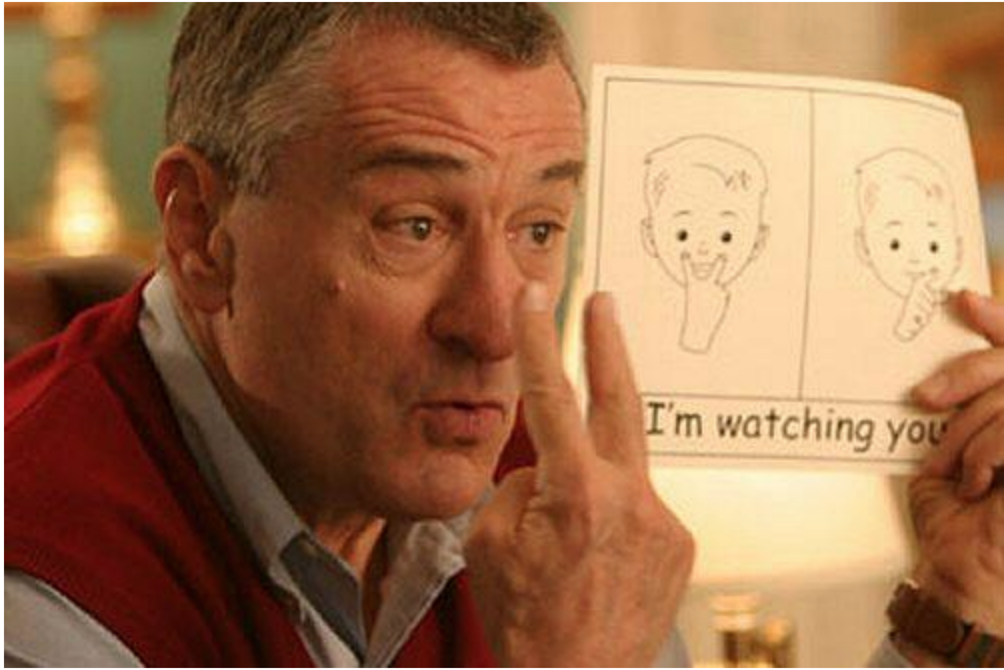


Bayesian Updates



Why Does this Work?

Let’s look at it again.



Entropy

Entropy
Information Content

Entropy

Information Content

Energy

Entropy
Information Content
Energy
Log Likelihood

Entropy

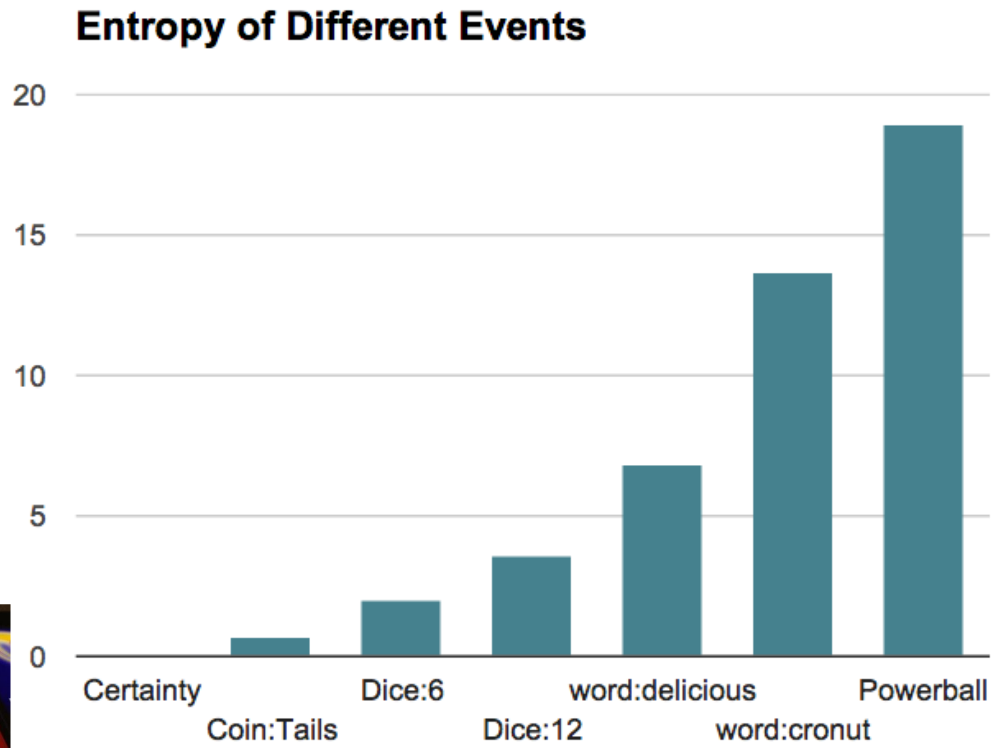
Information Content

Energy

Log Likelihood

$$-ln(p)$$


UNITED STATES OF AMERICA
E PLURIBUS UNUM
ONE CENT


Entropy of Different Events
20
15
10
5
0
Certainty
Coin:Tails
Dice:6
Dice:12
word:delicious
word:cronut
Powerball


UNITED STATES OF AMERICA
E PLURIBUS UNUM
ONE CENT


POWERBALL
ESTIMATED JACKPOT
500
MILLION

The Dirichlet Distribution

$$P(p|\alpha) = \frac{\Gamma\left(\sum_{k=0}^{K-1} \alpha_k\right)}{\prod_{k=0}^{K-1} \Gamma(\alpha_k)} \prod_{k=0}^{K-1} p_k^{\alpha_k - 1}$$

The Dirichlet Distribution

$$P(p|\alpha) = \frac{\Gamma\left(\sum_{k=0}^{K-1}\alpha_k\right)}{\prod_{k=0}^{K-1}\Gamma(\alpha_k)} \prod_{k=0}^{K-1} p_k^{\alpha_k - 1}$$

Normalizing Constant

The Dirichlet Distribution

$$P(p|\alpha) = \boxed{} \prod_{k=0}^{K-1} p_k^{\alpha_k - 1}$$

Normalizing Constant

The Dirichlet Distribution

$$E(p|\alpha) = -ln\left(\prod_{k_0}^{K-1} p_k^{\alpha_k - 1}\right)$$

The Dirichlet Distribution

$$E(p|\alpha) = \sum_{k_0}^{K-1} (\alpha_k - 1)\,(-ln(p_k))$$

The Dirichlet Distribution

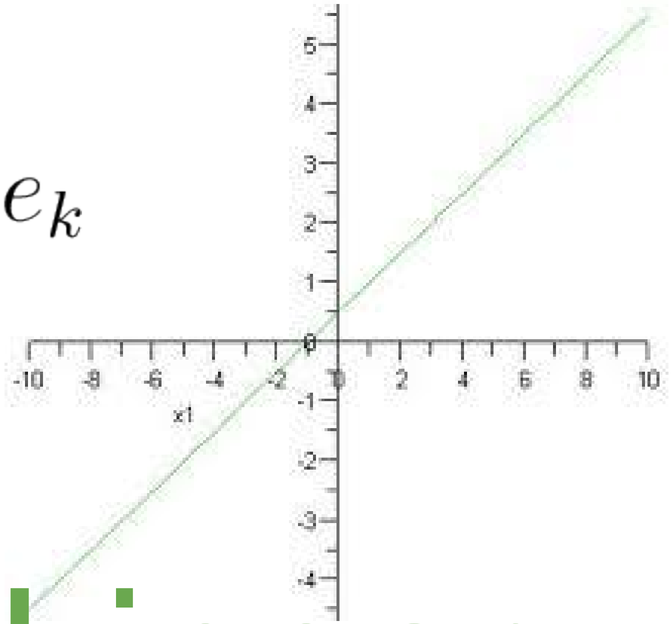
$$E(p|\alpha) = \sum_{k_0}^{K-1} (\alpha_k - 1)\, e_k$$

$$e_k = -ln(p_k)$$

The Dirichlet Distribution

$$E(p|\alpha) = \sum_{k_0}^{K-1} (\alpha_k - 1)\, e_k$$

$$e_k = -ln(p_k)$$



Linear

The Dirichlet MACHINE

Prior	1.2	3.0	0.3

The Dirichlet MACHINE

Prior	1.2	3.0	0.3

The Dirichlet MACHINE

Update	2.2	3.0	0.3

The Dirichlet MACHINE

Prior	2.2	3.0	0.3

The Dirichlet MACHINE

Update	2.2	3.0	1.3

The Dirichlet MACHINE

Prior	2.2	3.0	1.3

The Dirichlet MACHINE

Update

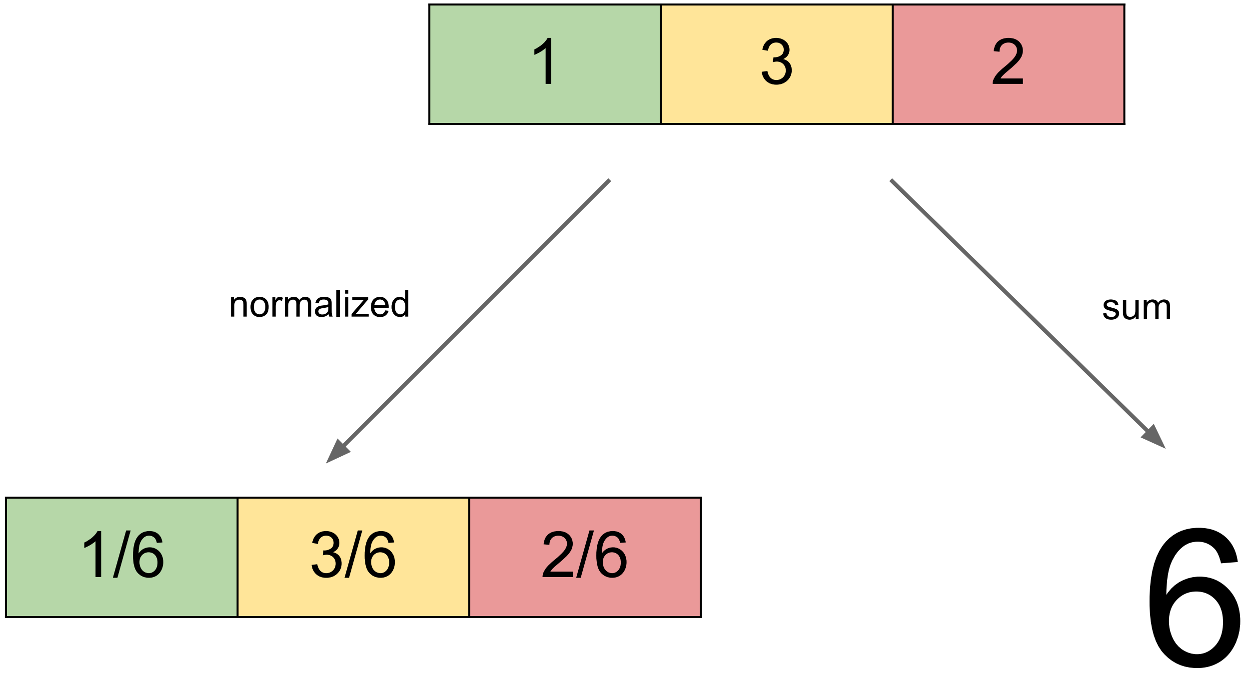
2.2	3.0	2.3

Interpreting the Parameters

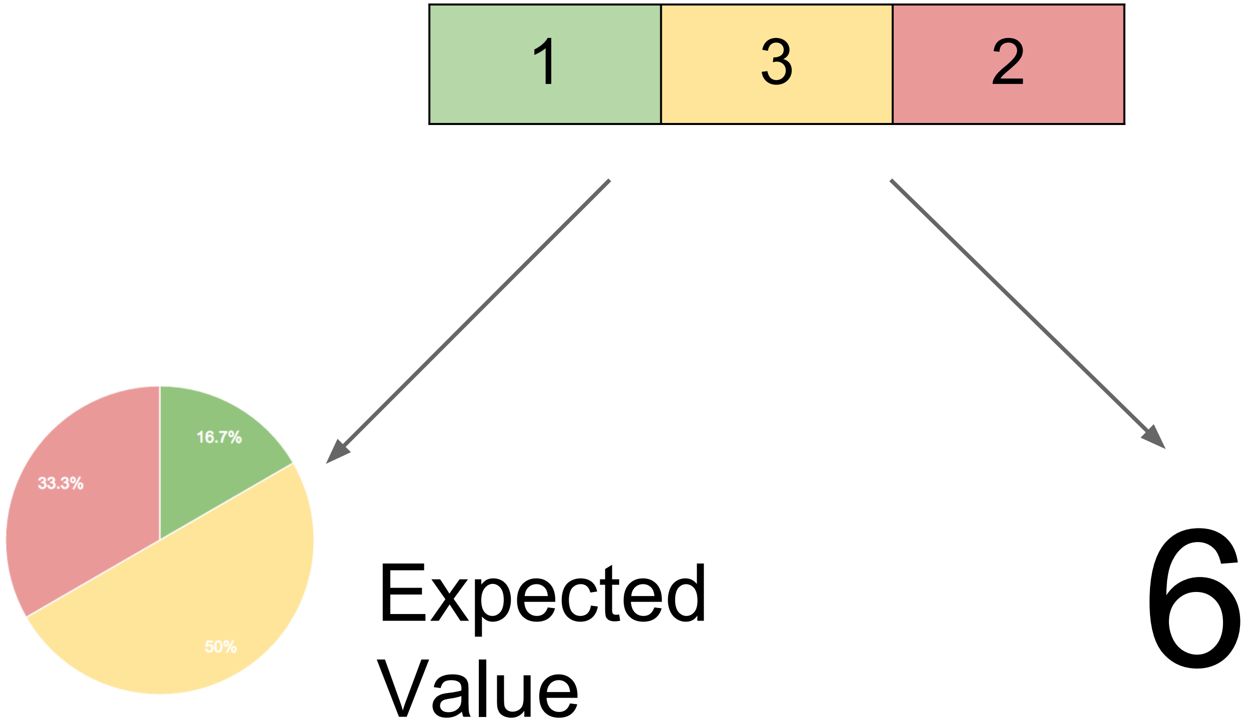
1	3	2

What does this alpha vector really mean?

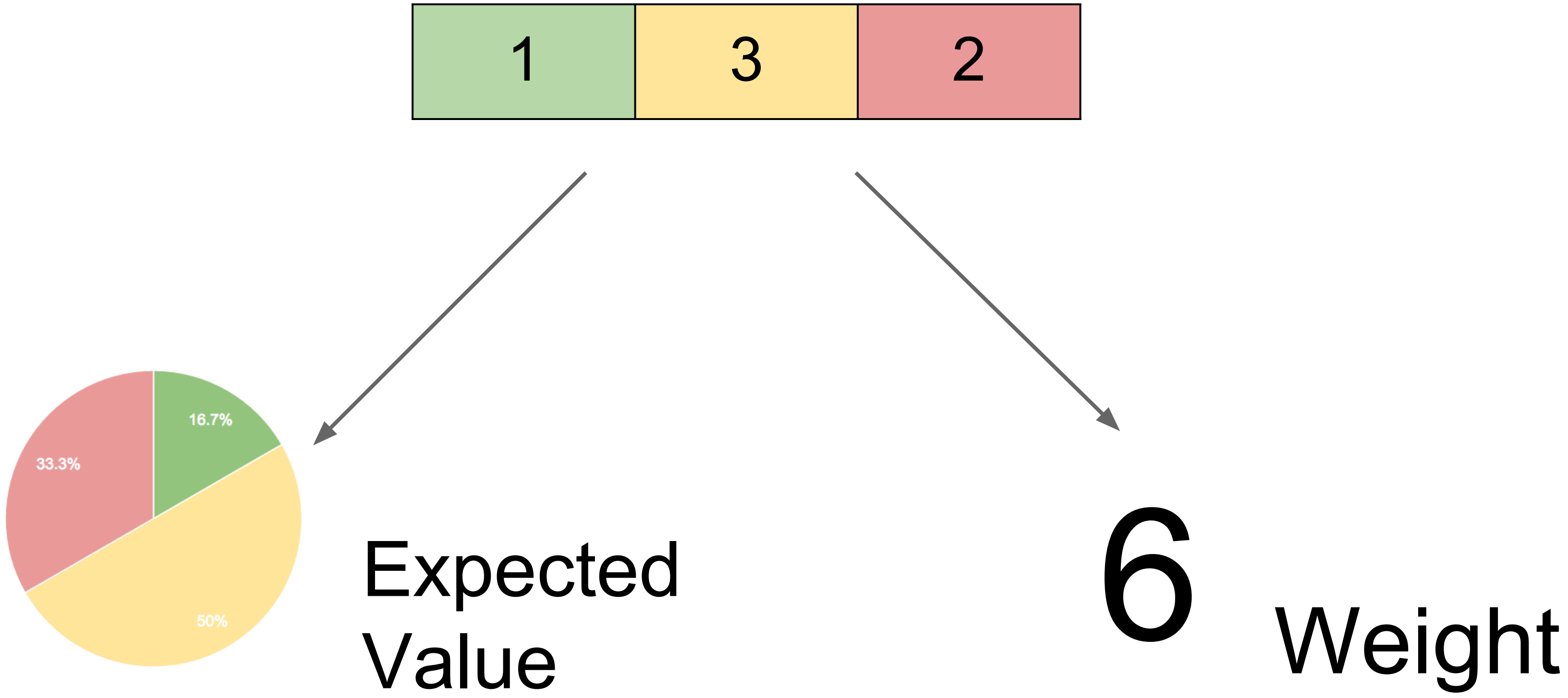
Interpreting the Parameters



Interpreting the Parameters

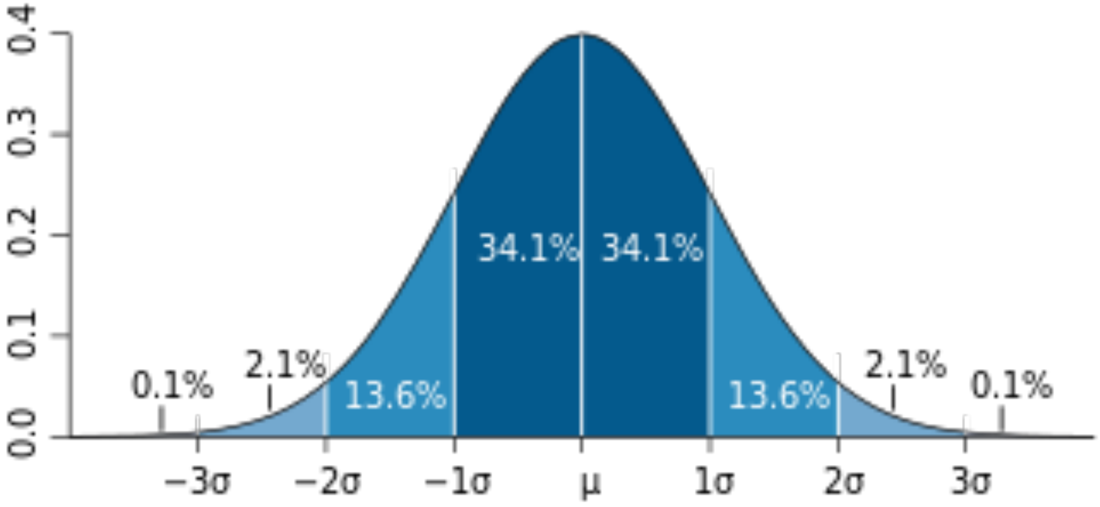


Interpreting the Parameters



ANALOGY: Normal Distribution

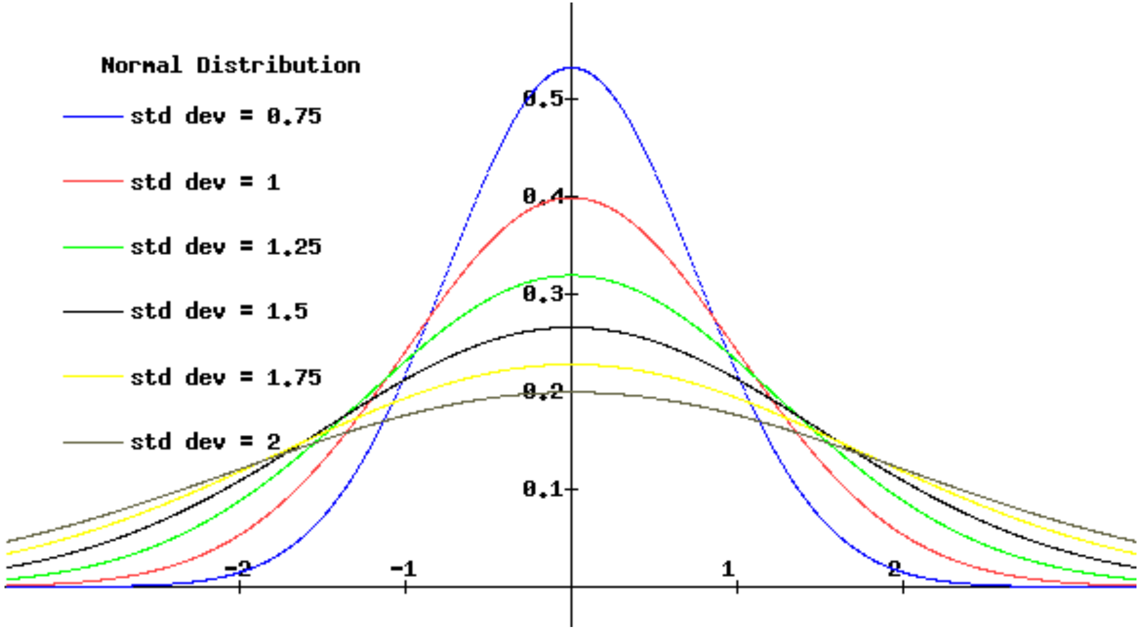
Precision =
1 / variance



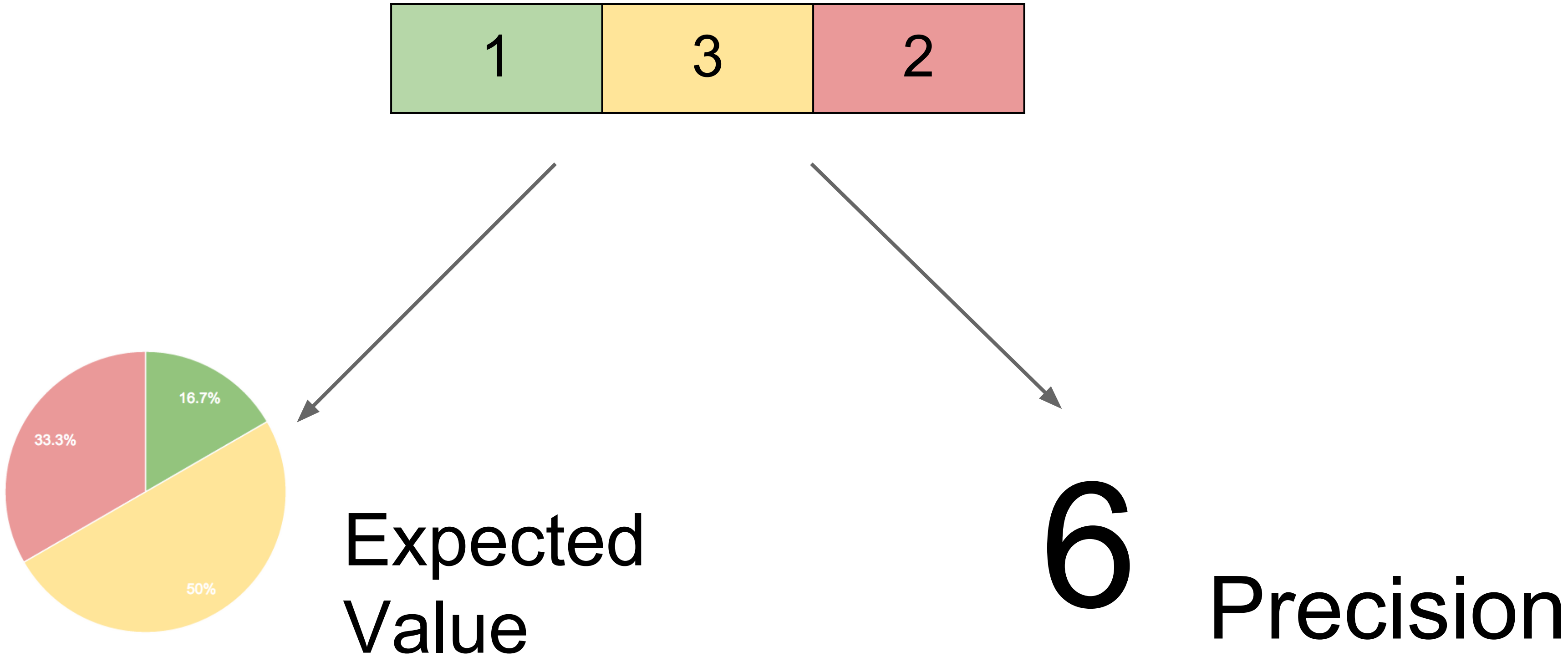
ANALOGY: Normal Distribution

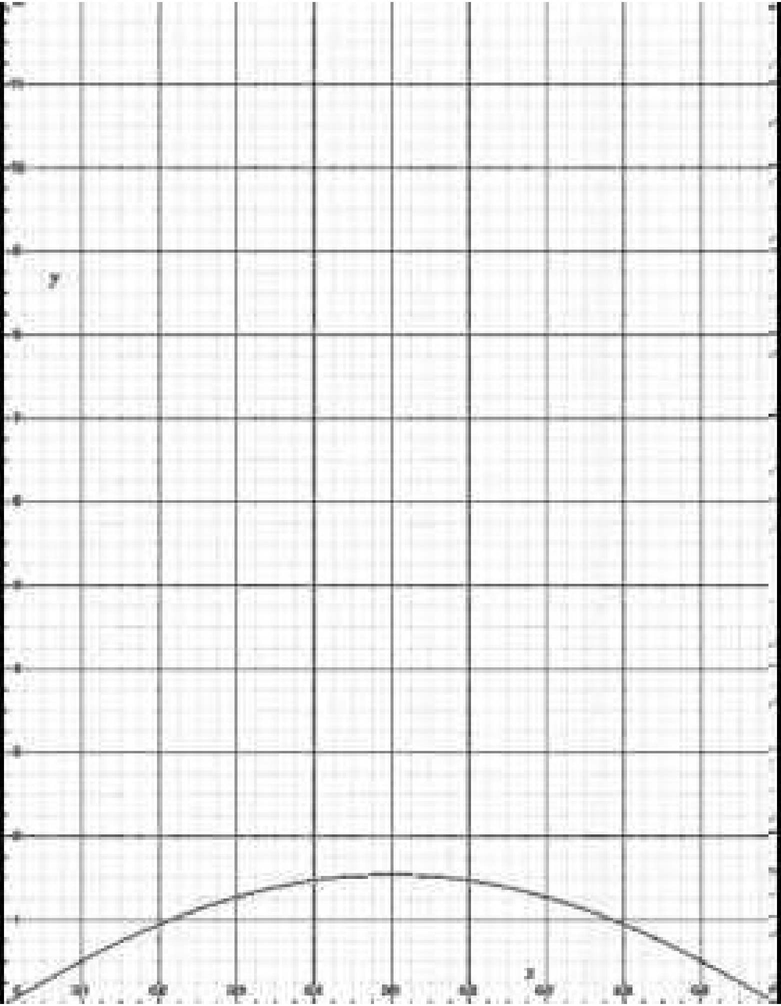
High precision: data is close to the mean

Low precision: far away from the mean

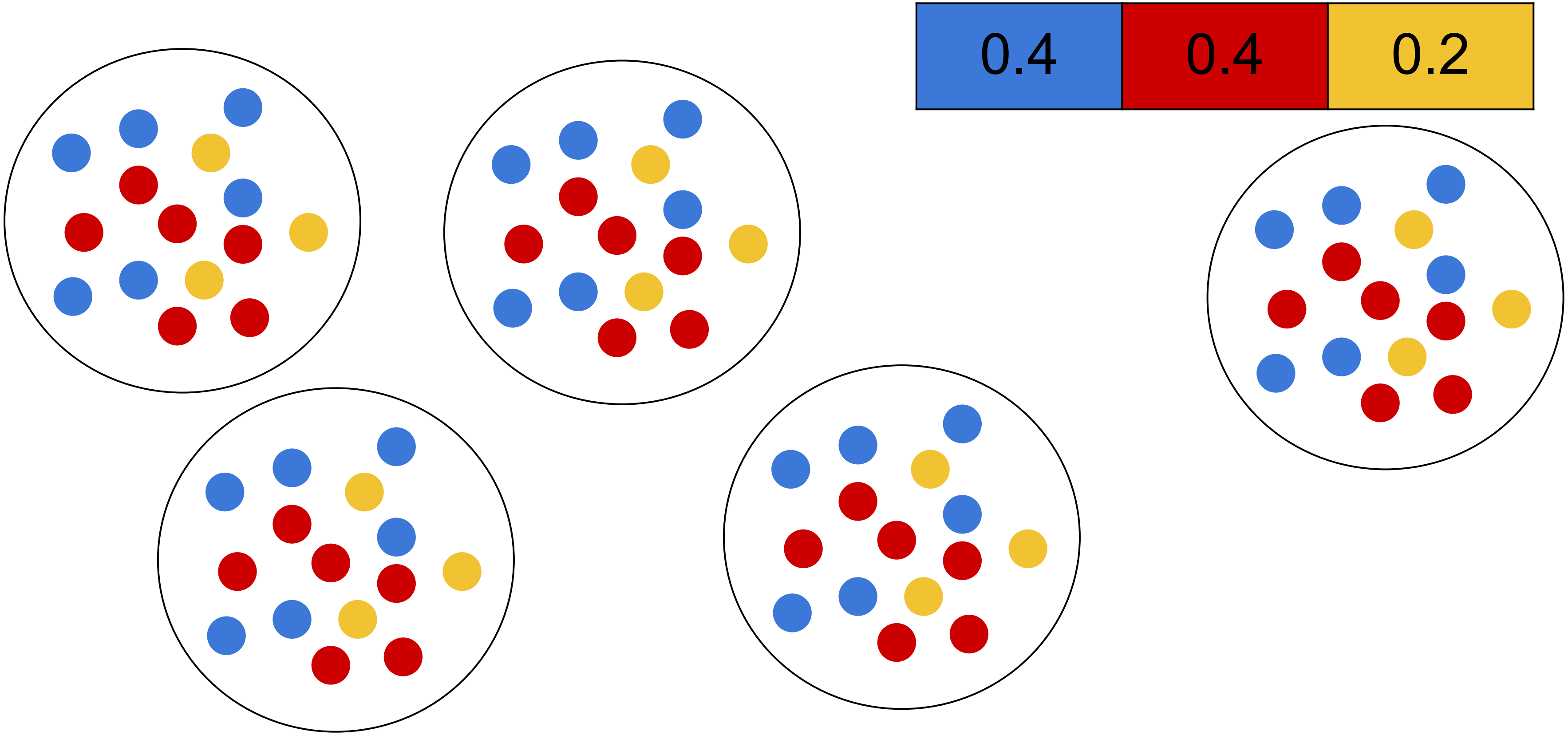


Interpreting the Parameters

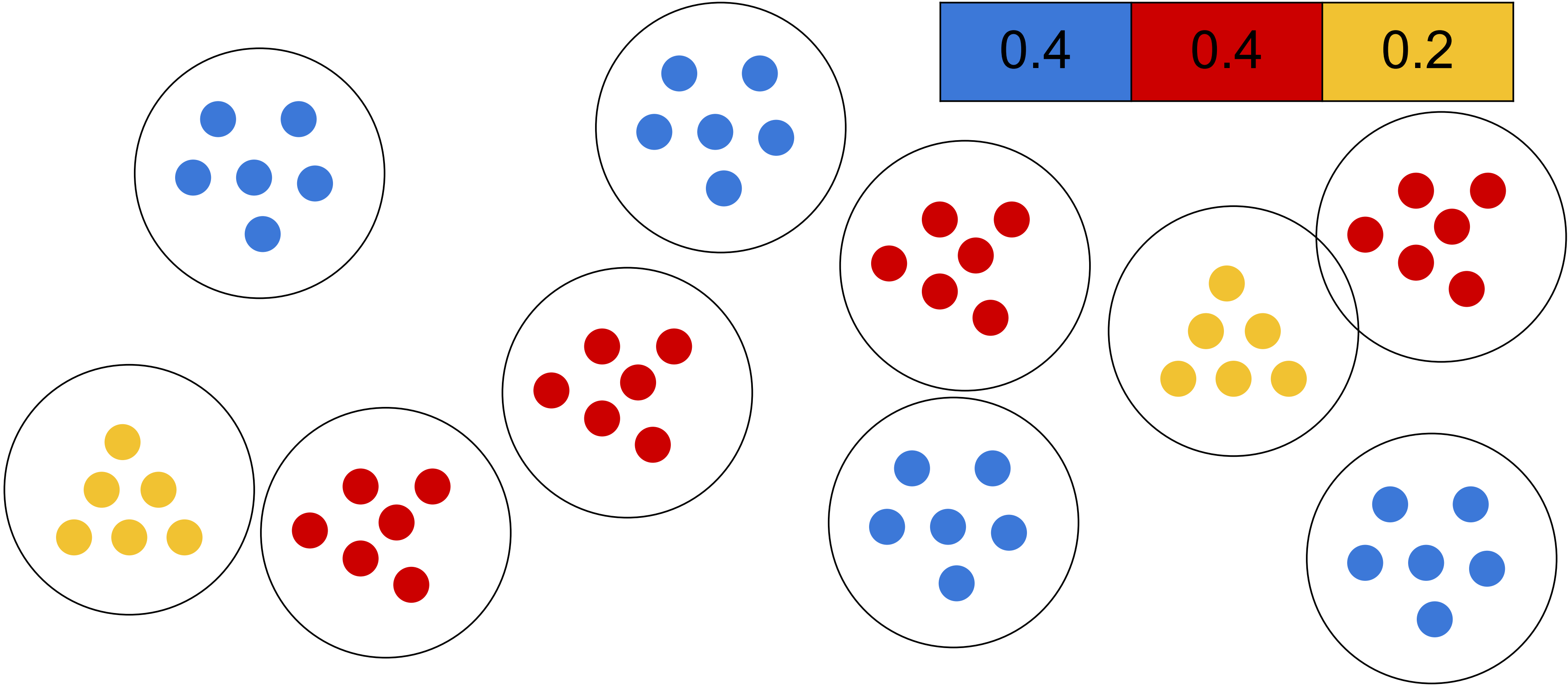



y
x

High Weight Dirichlet

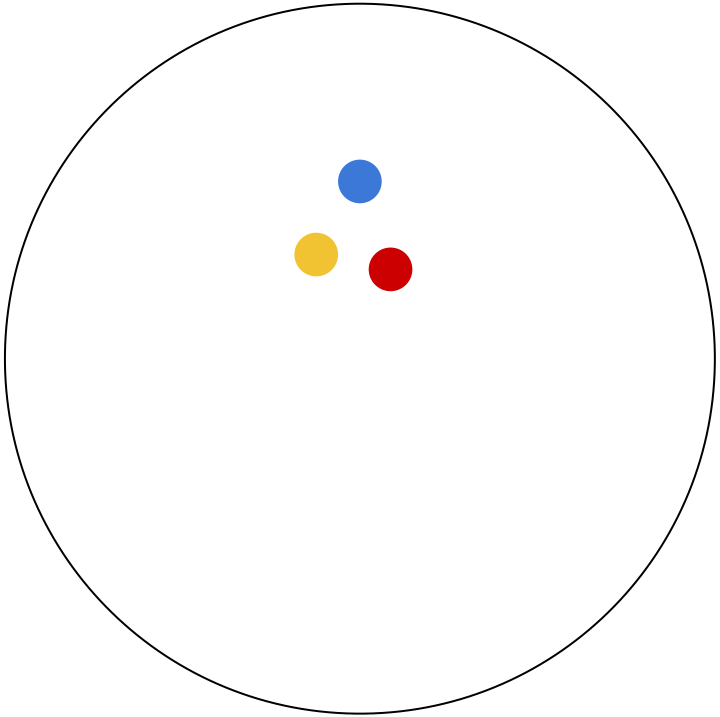


Low Weight Dirichlet



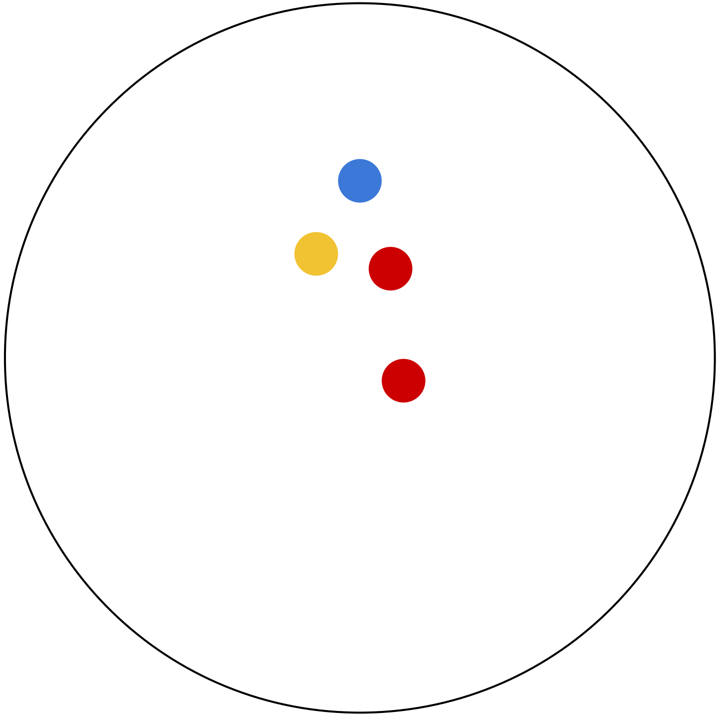
Urn Model

At each step, pick a ball from the urn.. Replace it, and add another ball of that color into the urn



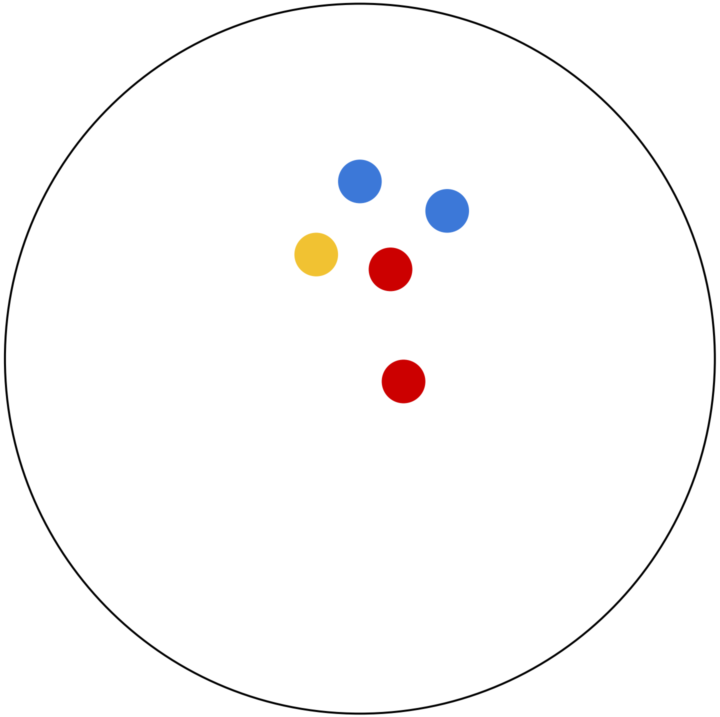
Urn Model

At each step,
pick a ball from
the urn..
Replace it, and
add another ball
of that color into
the urn



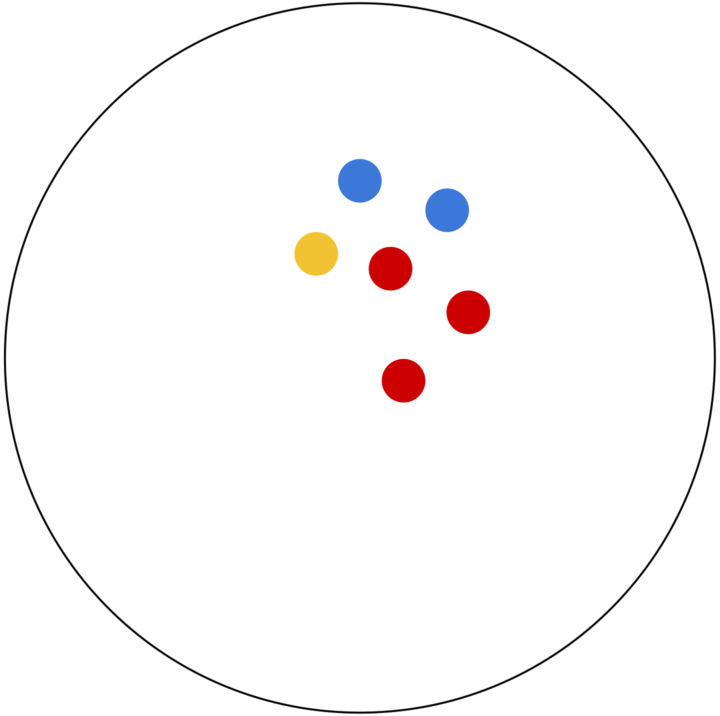
Urn Model

At each step, pick a ball from the urn.. Replace it, and add another ball of that color into the urn



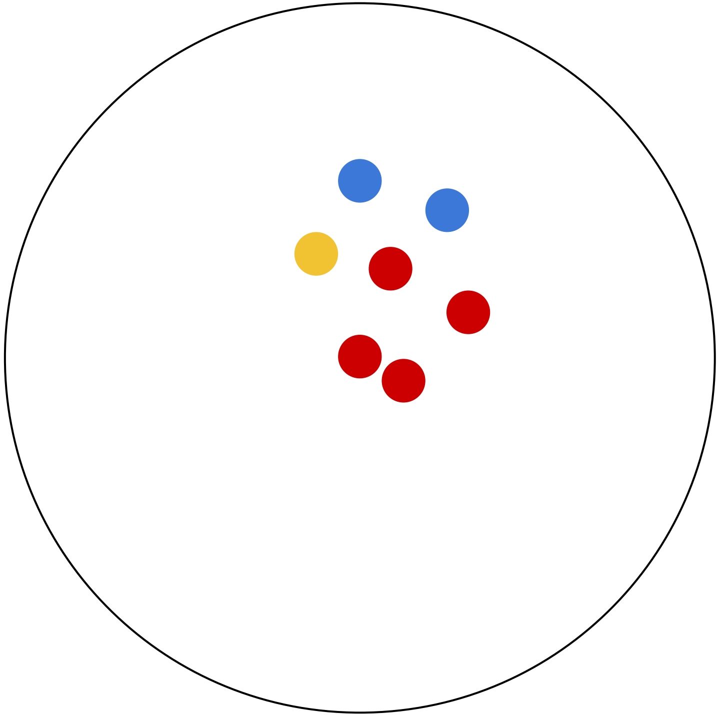
Urn Model

At each step, pick a ball from the urn.. Replace it, and add another ball of that color into the urn



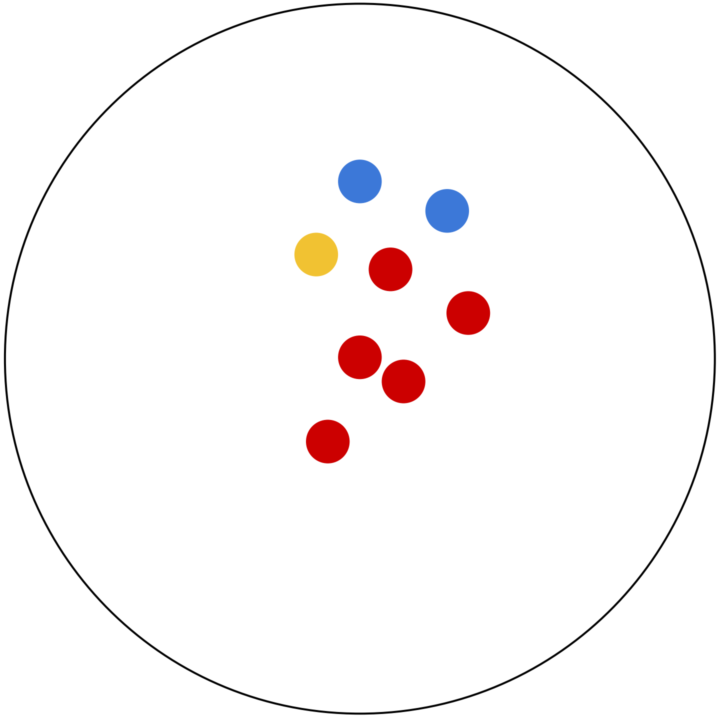
Urn Model

At each step, pick a ball from the urn.. Replace it, and add another ball of that color into the urn



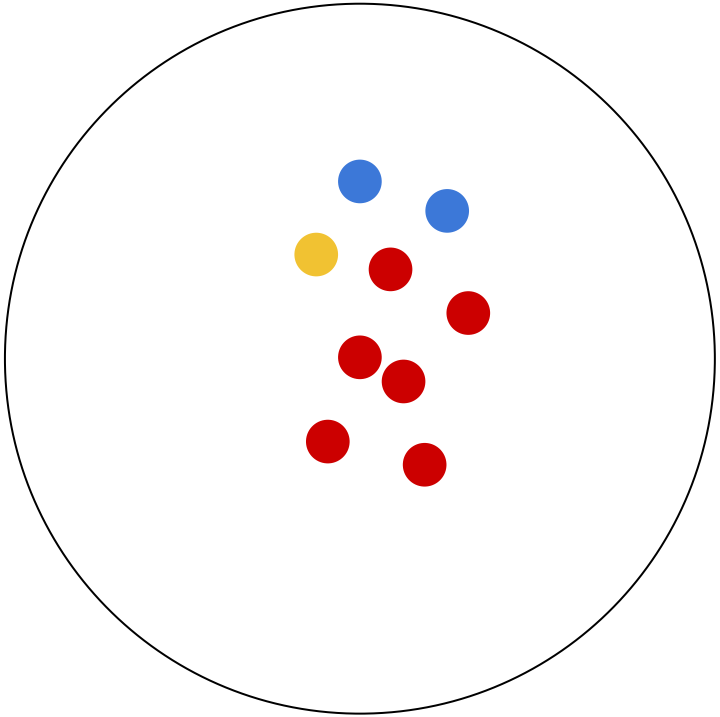
Urn Model

Rich get richer...



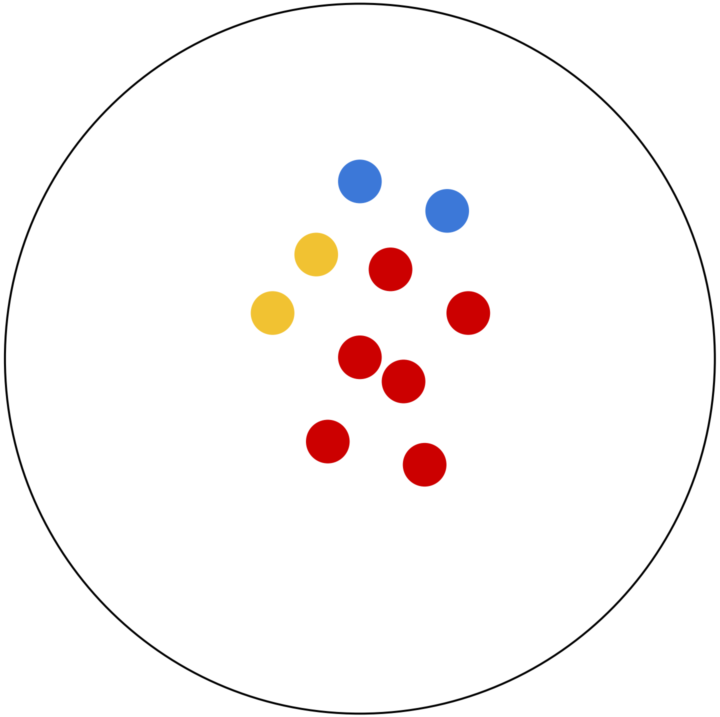
Urn Model

Rich get richer...



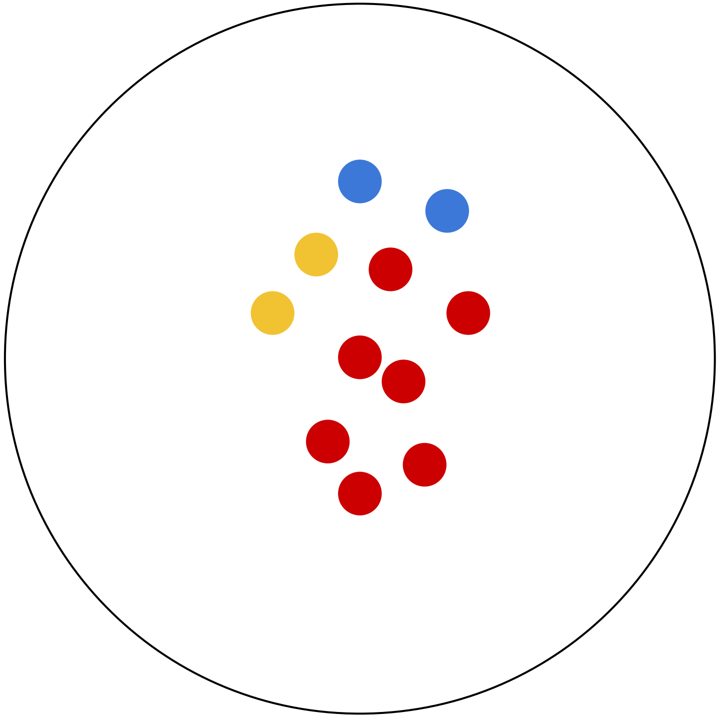
Urn Model

Finally yellow
catches a break



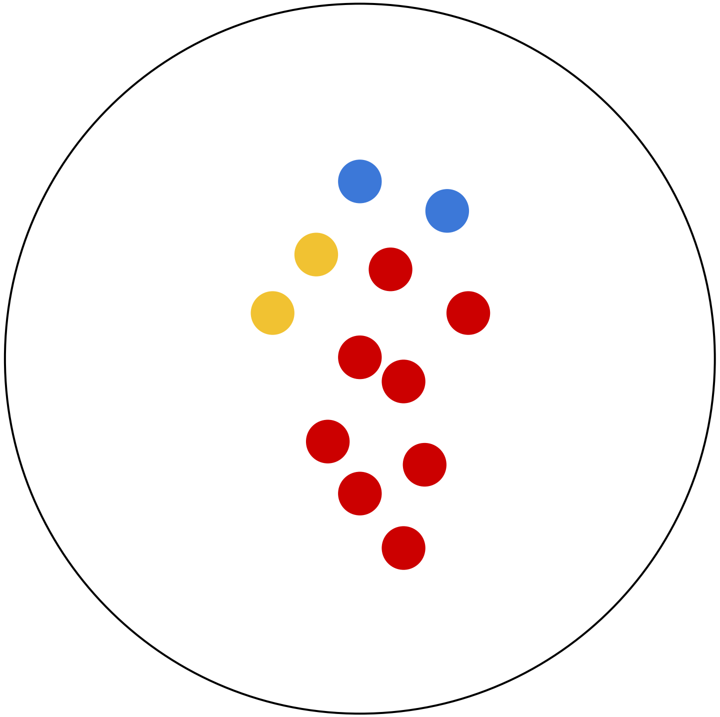
Urn Model

Finally yellow
catches a break



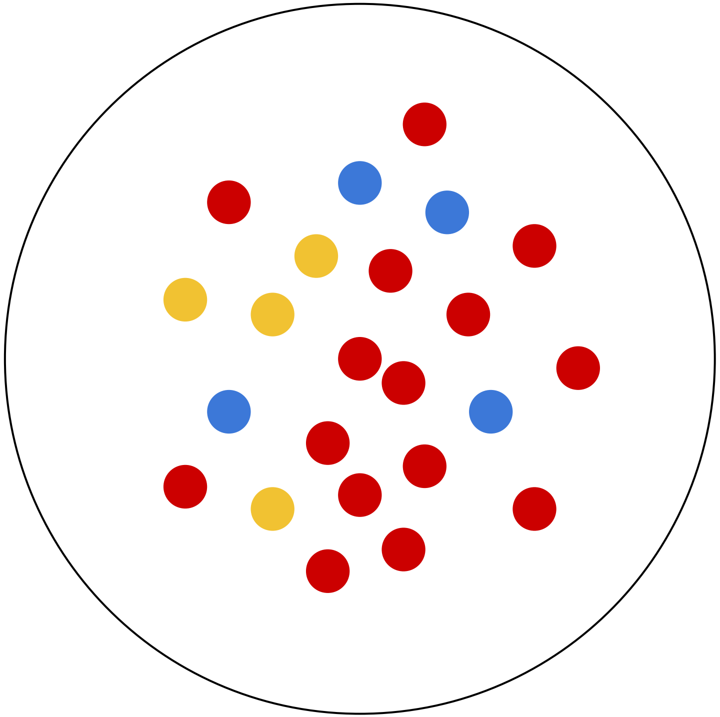
Urn Model

But it's too late...



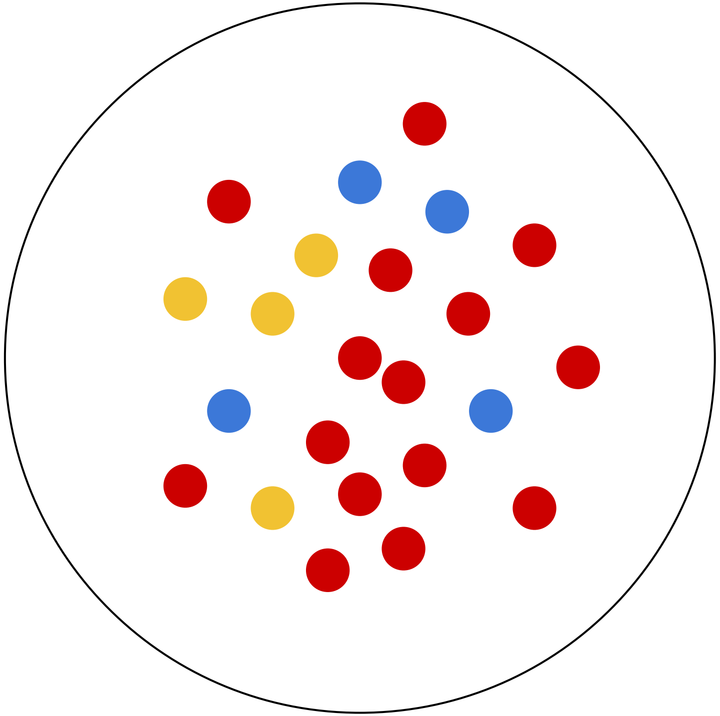
Urn Model

As the urn gets more populated, the distribution gets “stuck” in place.



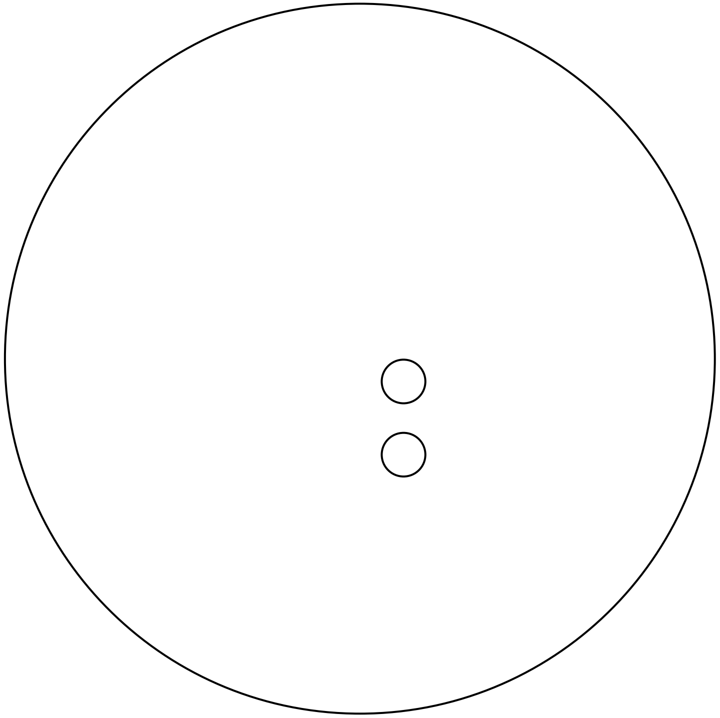
Urn Model

Once lots of data has been collected, or the dirichlet has high precision, it's hard to overturn that with new data



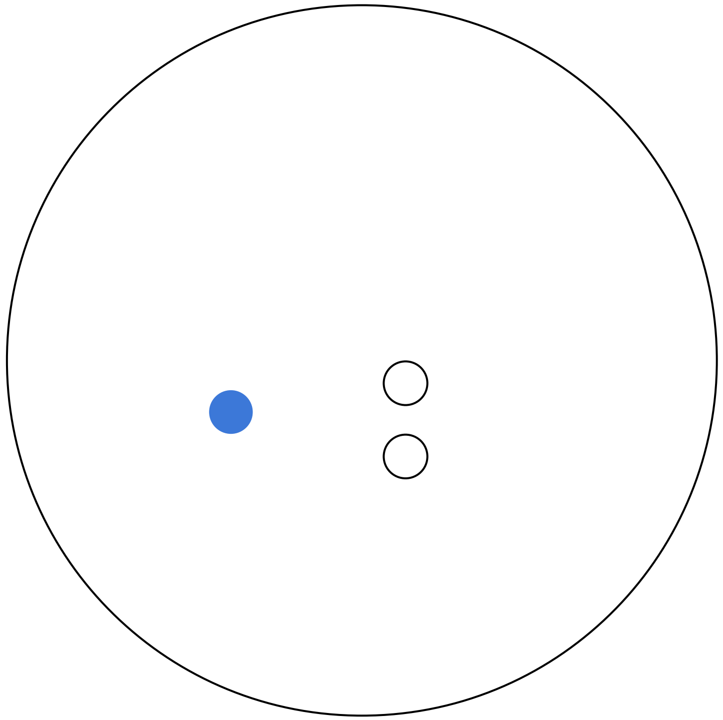
Chinese Restaurant Process

When you find the white ball, throw a new color into the mix.



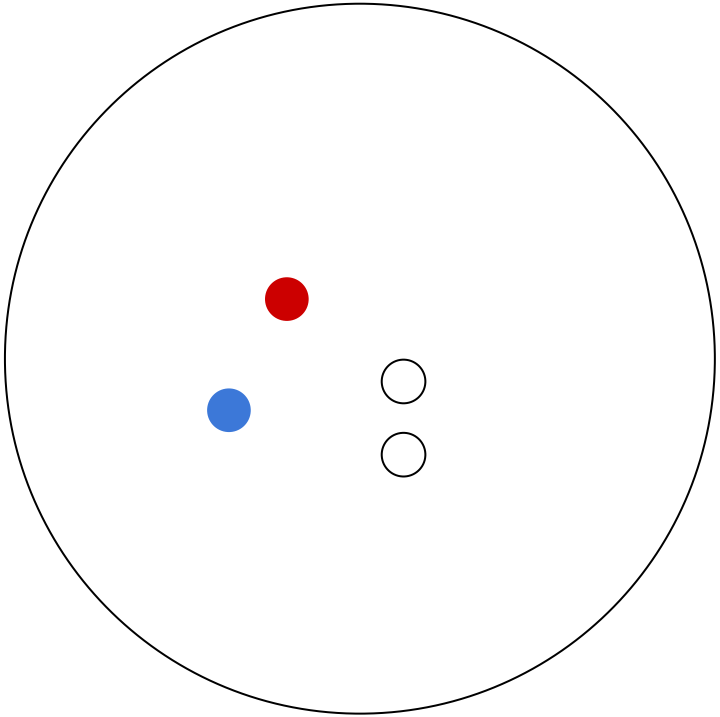
Chinese Restaurant Process

When you find the white ball, throw a new color into the mix.



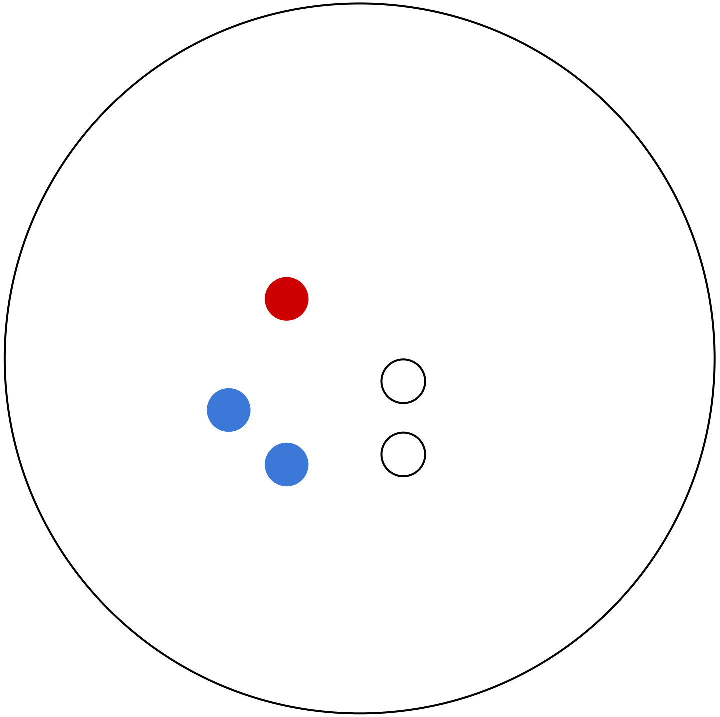
Chinese Restaurant Process

When you find the white ball, throw a new color into the mix.



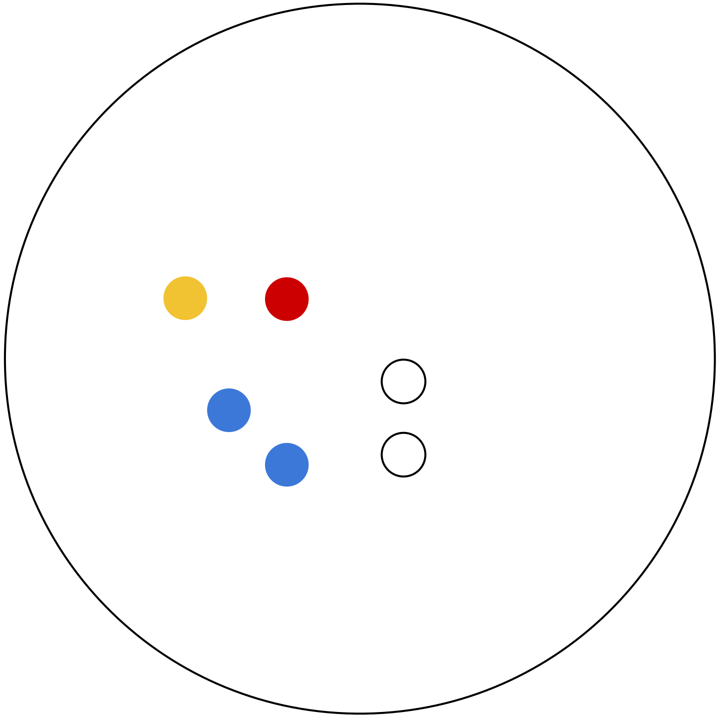
Chinese Restaurant Process

When you find the white ball, throw a new color into the mix.



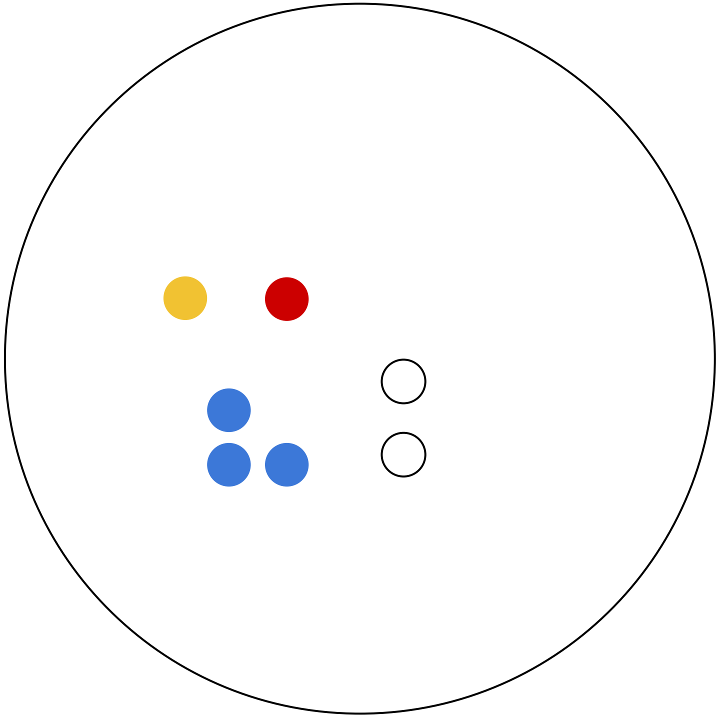
Chinese Restaurant Process

When you find the white ball, throw a new color into the mix.



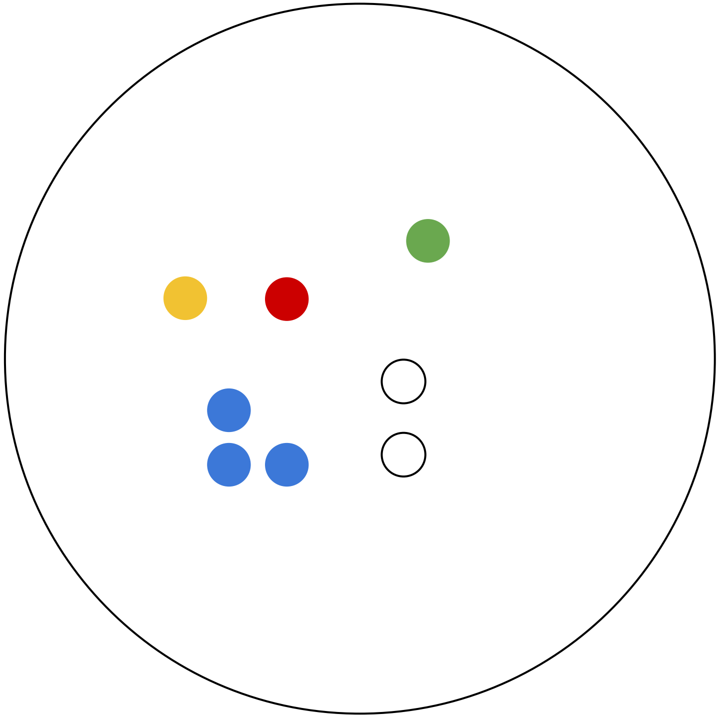
Chinese Restaurant Process

When you find the white ball, throw a new color into the mix.



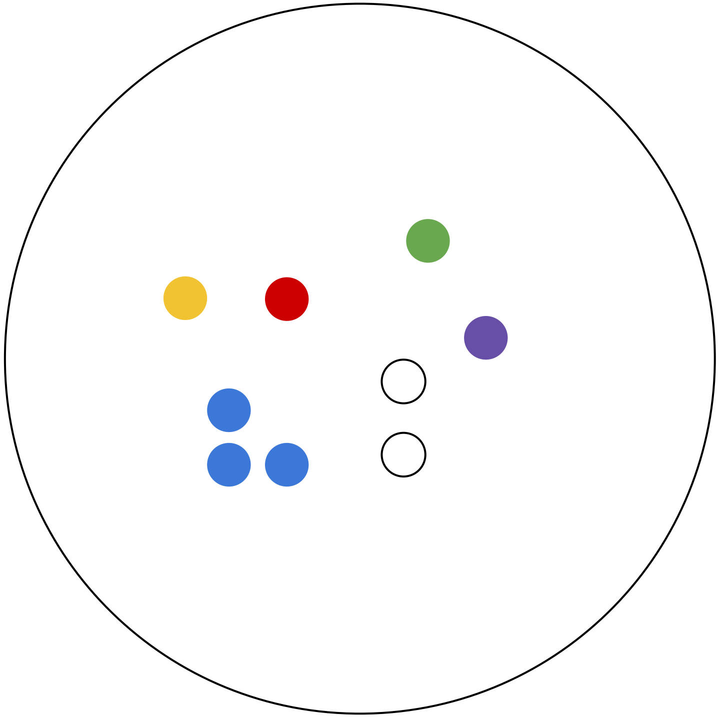
Chinese Restaurant Process

When you find the white ball, throw a new color into the mix.



Chinese Restaurant Process

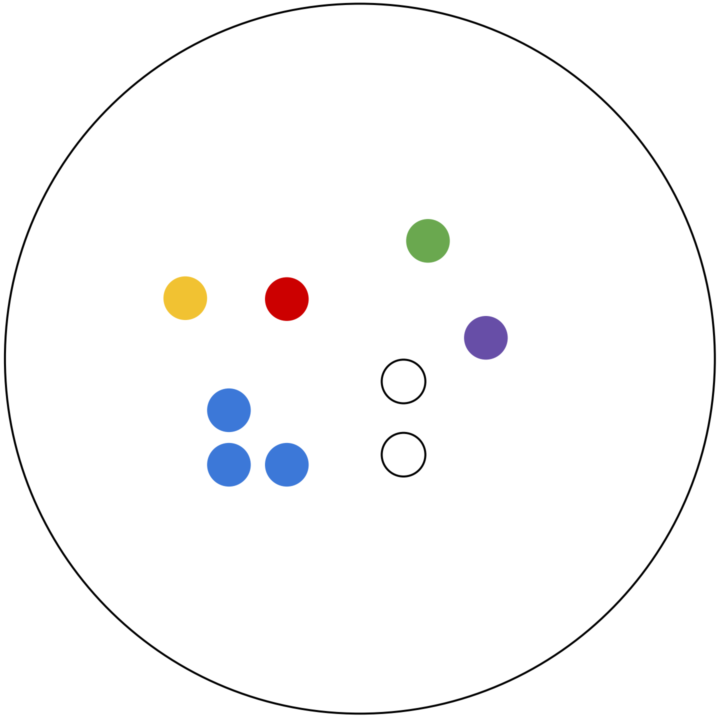
When you find the white ball, throw a new color into the mix.



Chinese Restaurant Process

The expected infinite distribution (mean) is exponential.

# of white balls controls the exponent



So coming back to the count data:

What Dirichlet parameters explain the data?

20	0	0
2	1	17
14	6	0
15	5	0
0	20	0
0	14	6

So coming back to the count data:

Newton's Method:
Requires Gradient
+ Hessian

20	0	0
2	1	17
14	6	0
15	5	0
0	20	0
0	14	6

So coming back to the count data:

Reads all of the data...

20	0	0
2	1	17
14	6	0
15	5	0
0	20	0
0	14	6

So coming back to the count data:

https://github.com/maxsklar/BayesPy/tree/master/ConjugatePriorTools

20	0	0
2	1	17
14	6	0
15	5	0
0	20	0
0	14	6

So coming back to the count data:

Compress the data into a Matrix and a Vector:

Works for lots of sparsely populated rows

20	0	0
2	1	17
14	6	0
15	5	0
0	20	0
0	14	6

The Compression MACHINE

0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0

The Compression MACHINE

K = 3
(the 4th row is a special, total row)

M = 6
The maximum # samples per input

0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0

The Compression MACHINE

1	0	0

K = 3
(the 4th row is a special, total row)

M = 6
The maximum # samples per input

0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0

The Compression MACHINE

K = 3
(the 4th row is a special, total row)

M = 6
The maximum # samples per input

1	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
1	0	0	0	0	0

The Compression MACHINE

1	3	2

K = 3
(the 4th row is a special, total row)

M = 6
The maximum # samples per input

1	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
1	0	0	0	0	0

The Compression MACHINE

K = 3
(the 4th row is a special, total row)

M = 6
The maximum # samples per input

2	0	0	0	0	0
1	1	1	0	0	0
1	1	0	0	0	0
2	1	1	1	1	1

The Compression MACHINE

0	6	0

K = 3
(the 4th row is a special, total row)

M = 6
The maximum # samples per input

2	0	0	0	0	0
1	1	1	0	0	0
1	1	0	0	0	0
2	1	1	1	1	1

The Compression MACHINE

K = 3
(the 4th row is a special, total row)

M = 6
The maximum # samples per input

2	0	0	0	0	0
2	2	2	1	1	1
1	1	0	0	0	0
3	2	2	2	2	2

The Compression MACHINE

2	1	1

K = 3
(the 4th row is a special, total row)

M = 6
The maximum # samples per input

2	0	0	0	0	0
2	2	2	1	1	1
1	1	0	0	0	0
3	2	2	2	2	2

The Compression MACHINE

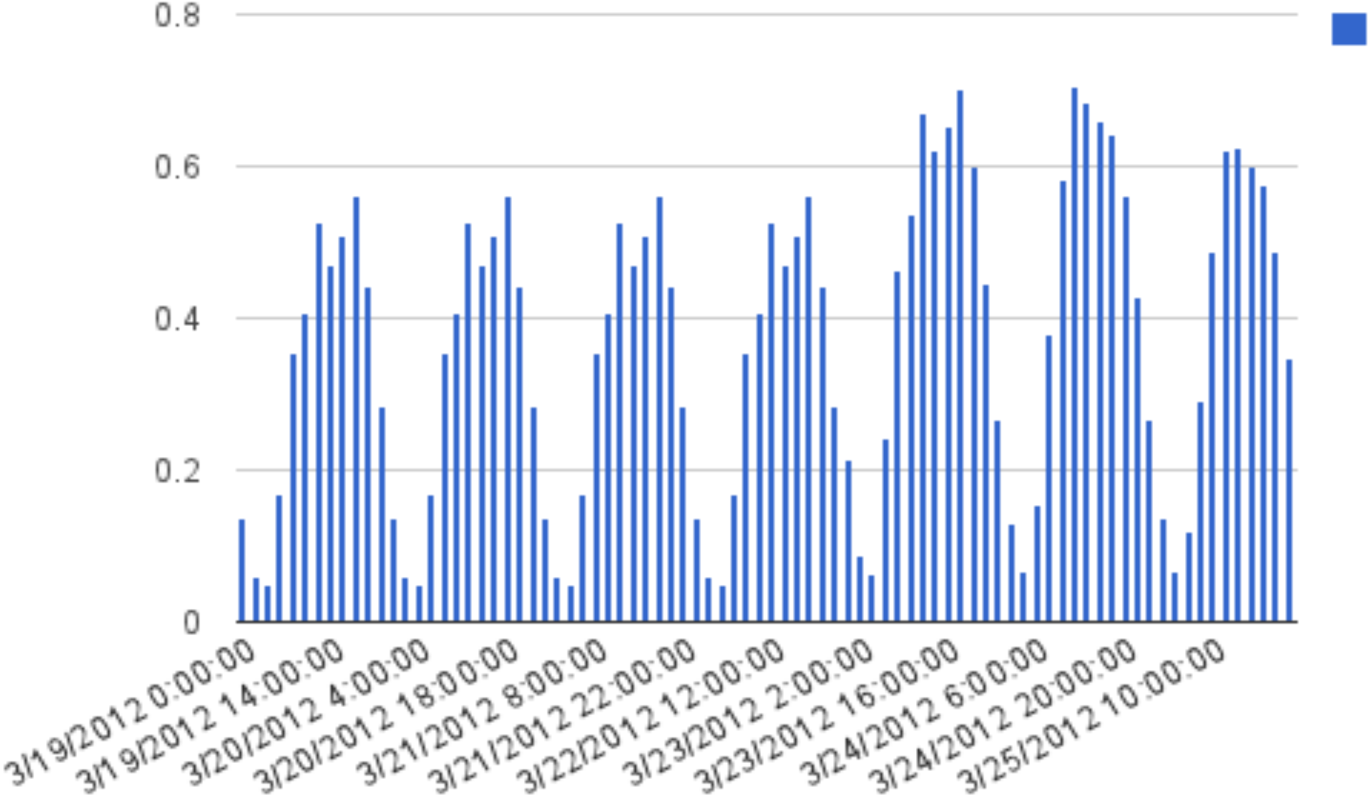
K = 3
(the 4th row is a special, total row)

M = 6
The maximum # samples per input

3	1	0	0	0	0
3	2	2	1	1	1
2	1	0	0	0	0
4	3	3	3	2	2

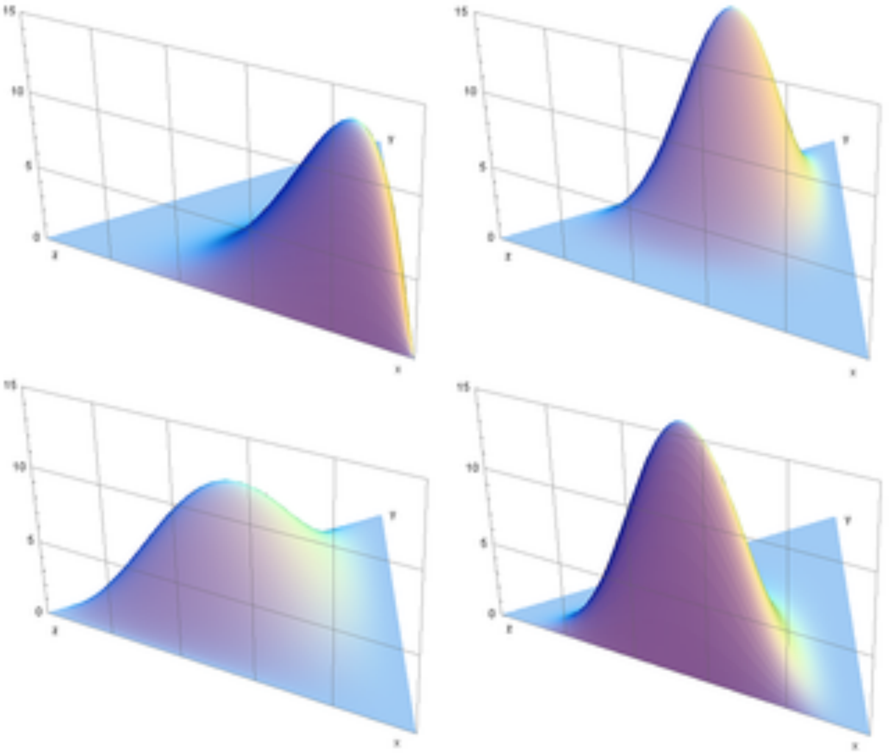
DEMO

Our Popularity Prior



Dirichlet Mixture Models

Anything you can go with a Gaussian, you can also do with a Dirichlet



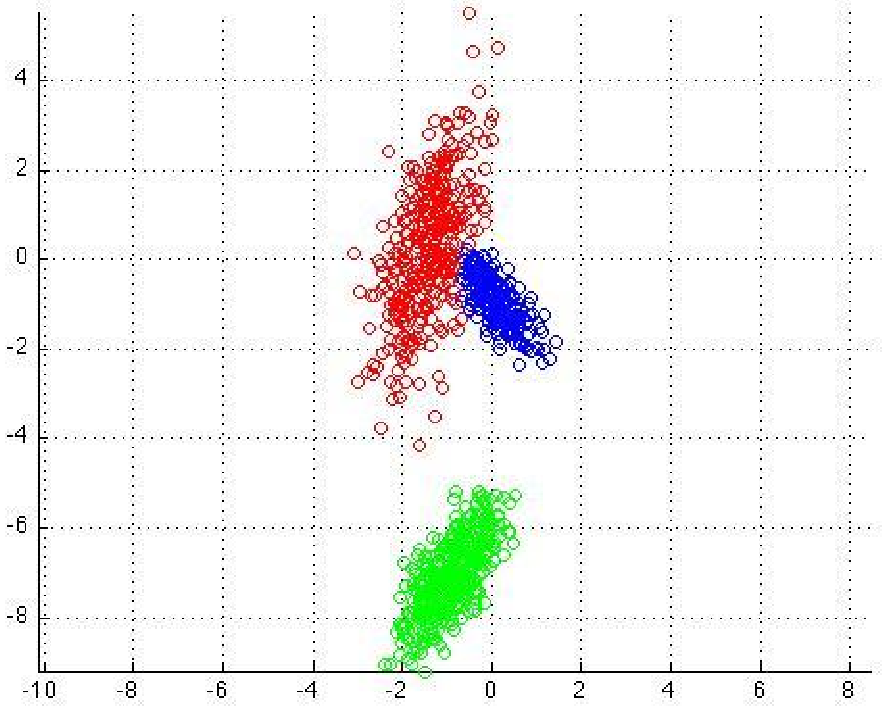
Dirichlet Mixture Models

Example:

Mixture of Gaussians using

Expectation-

Maximization



Dirichlet Mixture Models

Assume each row is a mixture of **multinomials.**

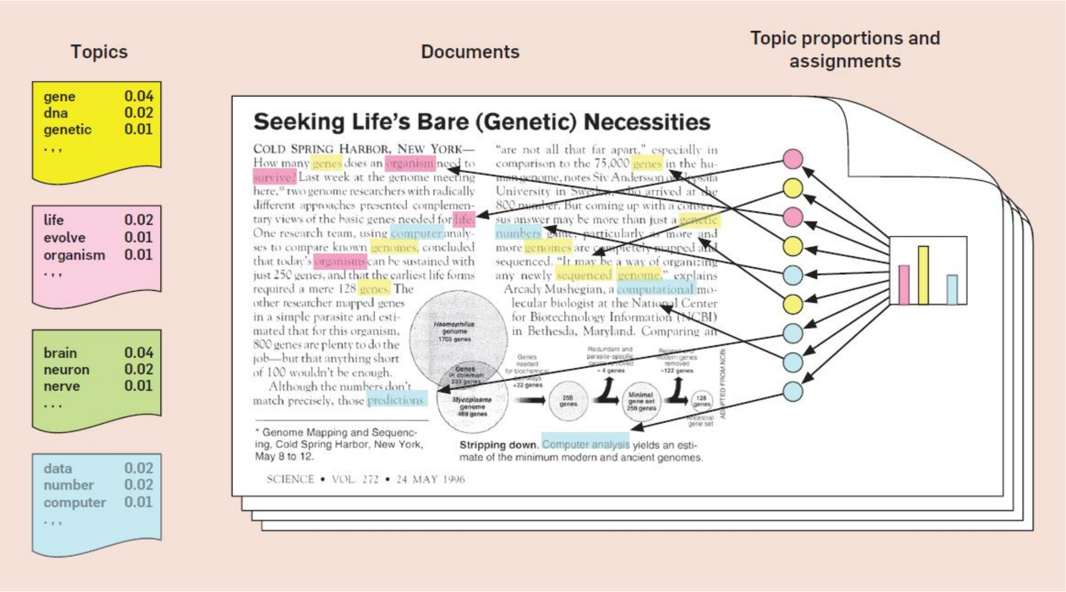
And the parameters of that mixture are pulled from a Dirichlet.

?

Dirichlet Mixture Models

Latent Dirichlet Allocation

Topic Model



Questions