~~Problems or opportunities in large biomedical data~~

Health
Data
Lab

CANS
Centre for New Antibacterial Strategies

TROMSØ
RESEARCH
FOUNDATION

Einar Holsbø, CANS Day 23.02.2023

Microbiome data: what's in the bag?

Health
Data
Lab

CANS
Centre for New Antibacterial Strategies

TROMSØ
RESEARCH
FOUNDATION

Einar Holsbø, CANS Day 23.02.2023

Part 1: Ken's project and some of my various confusions

Querying the microbiome

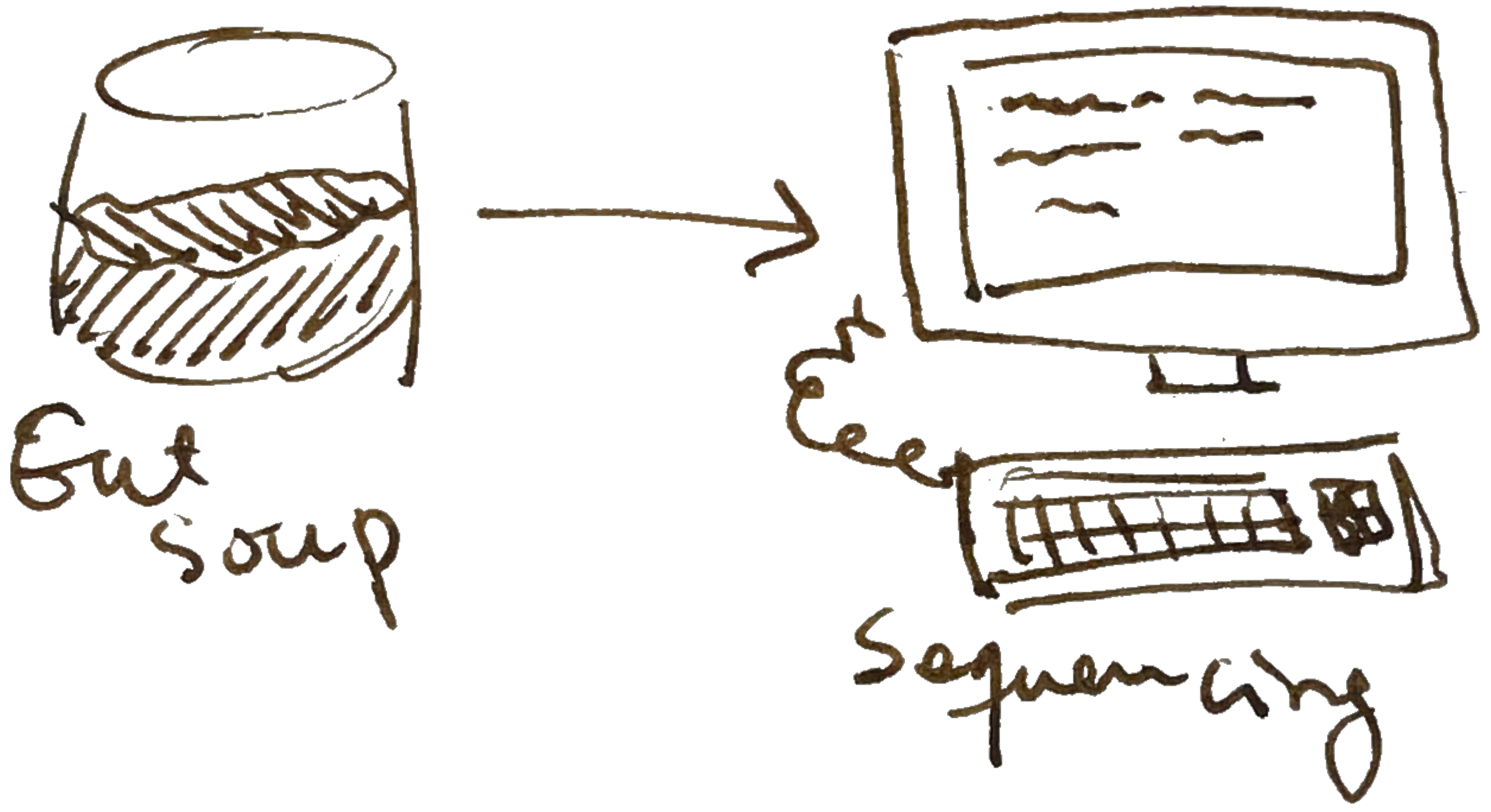
A child's understanding

Querying the microbiome

A child’s understanding

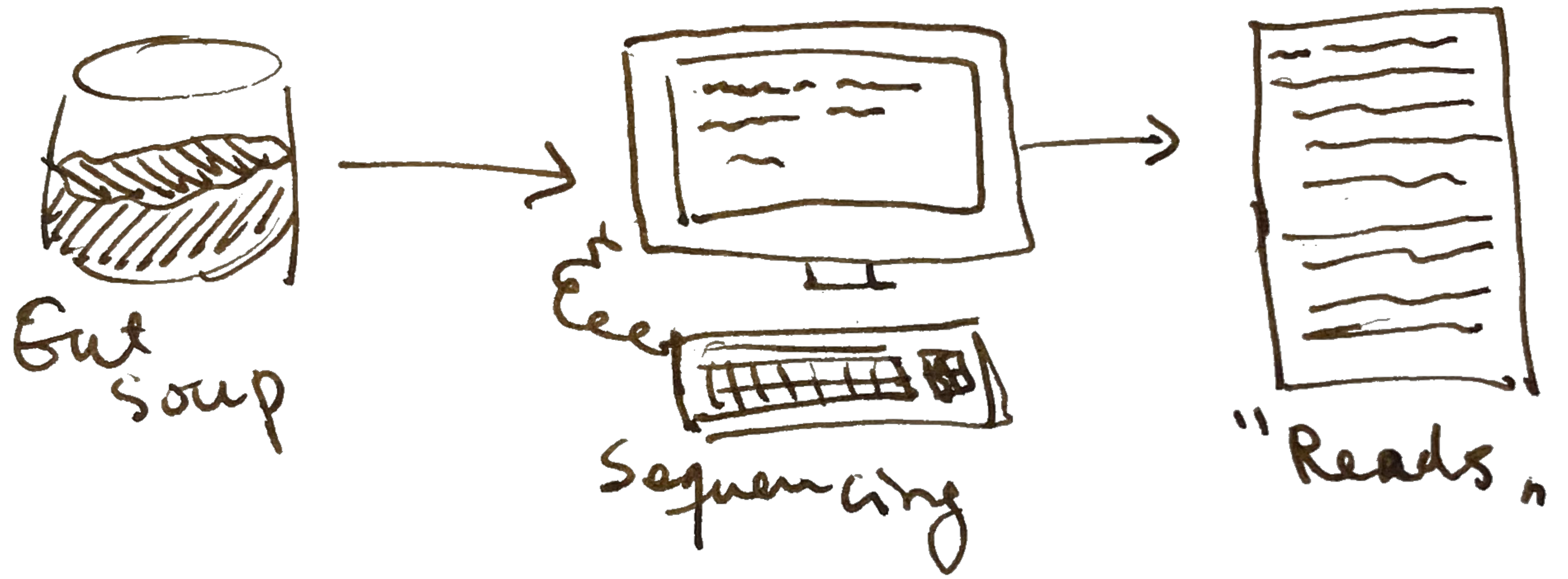
Querying the microbiome

A child's understanding



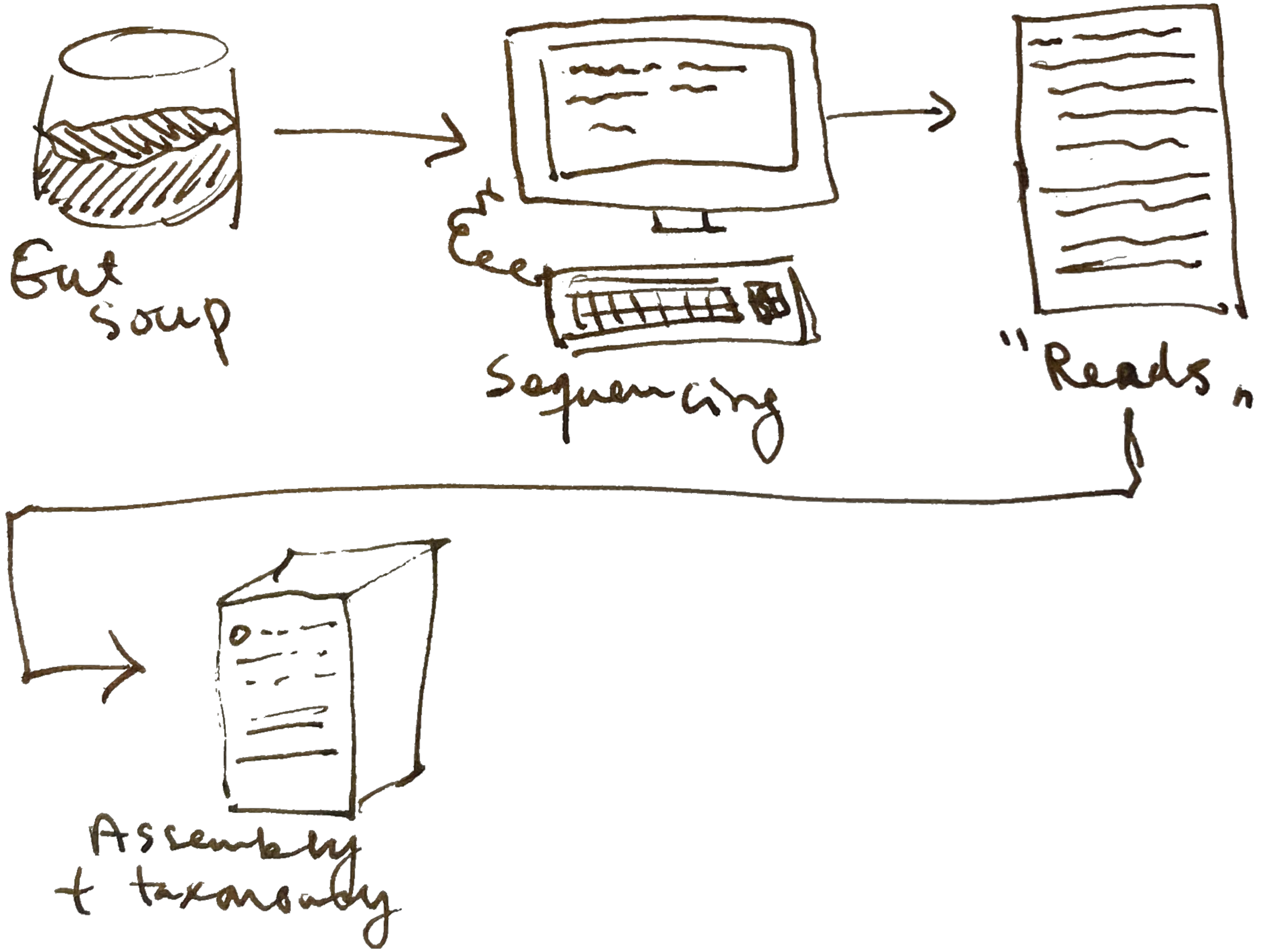
Querying the microbiome

A child's understanding



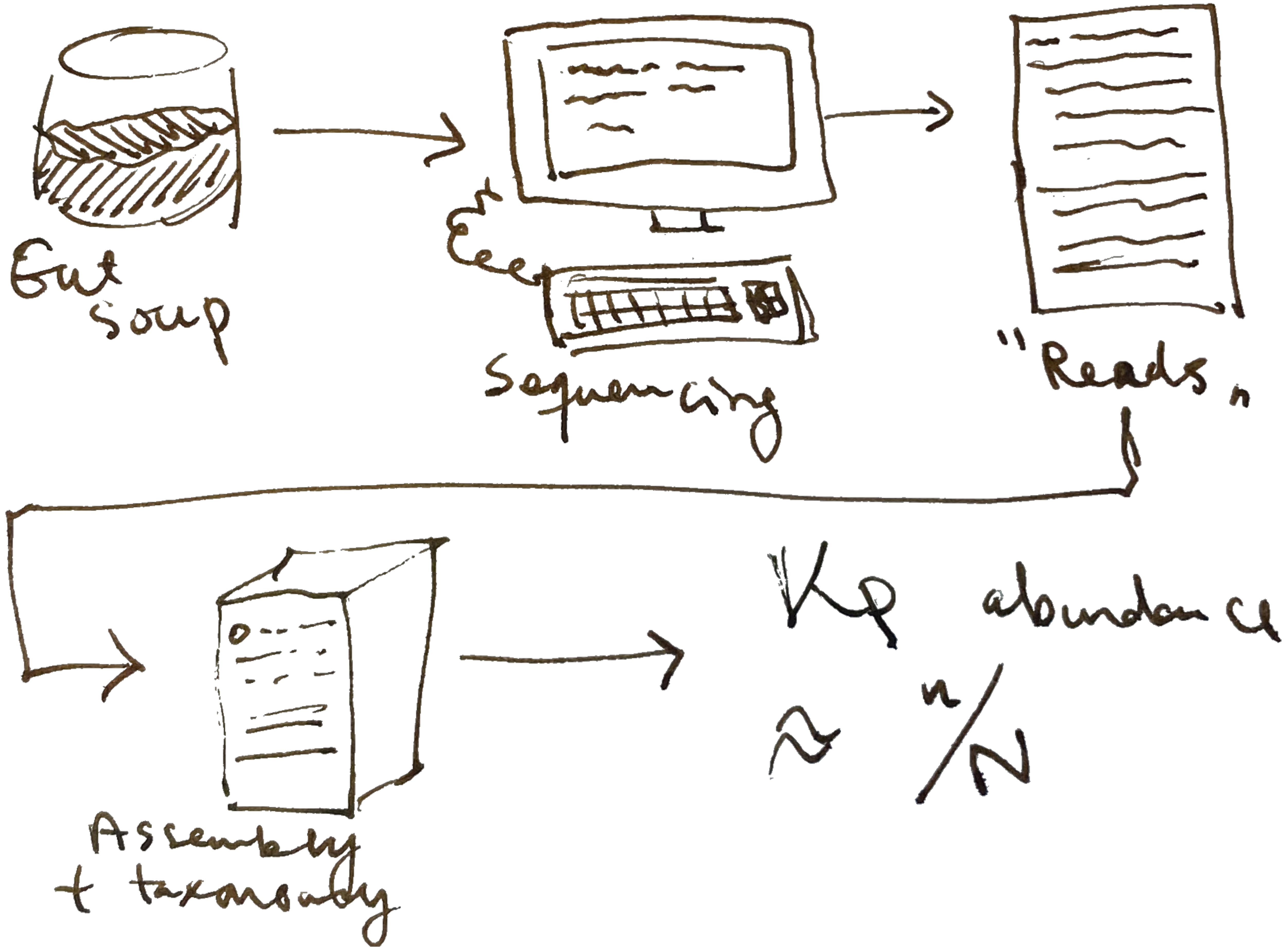
Querying the microbiome

A child's understanding



Querying the microbiome

A child's understanding



Querying the microbiome

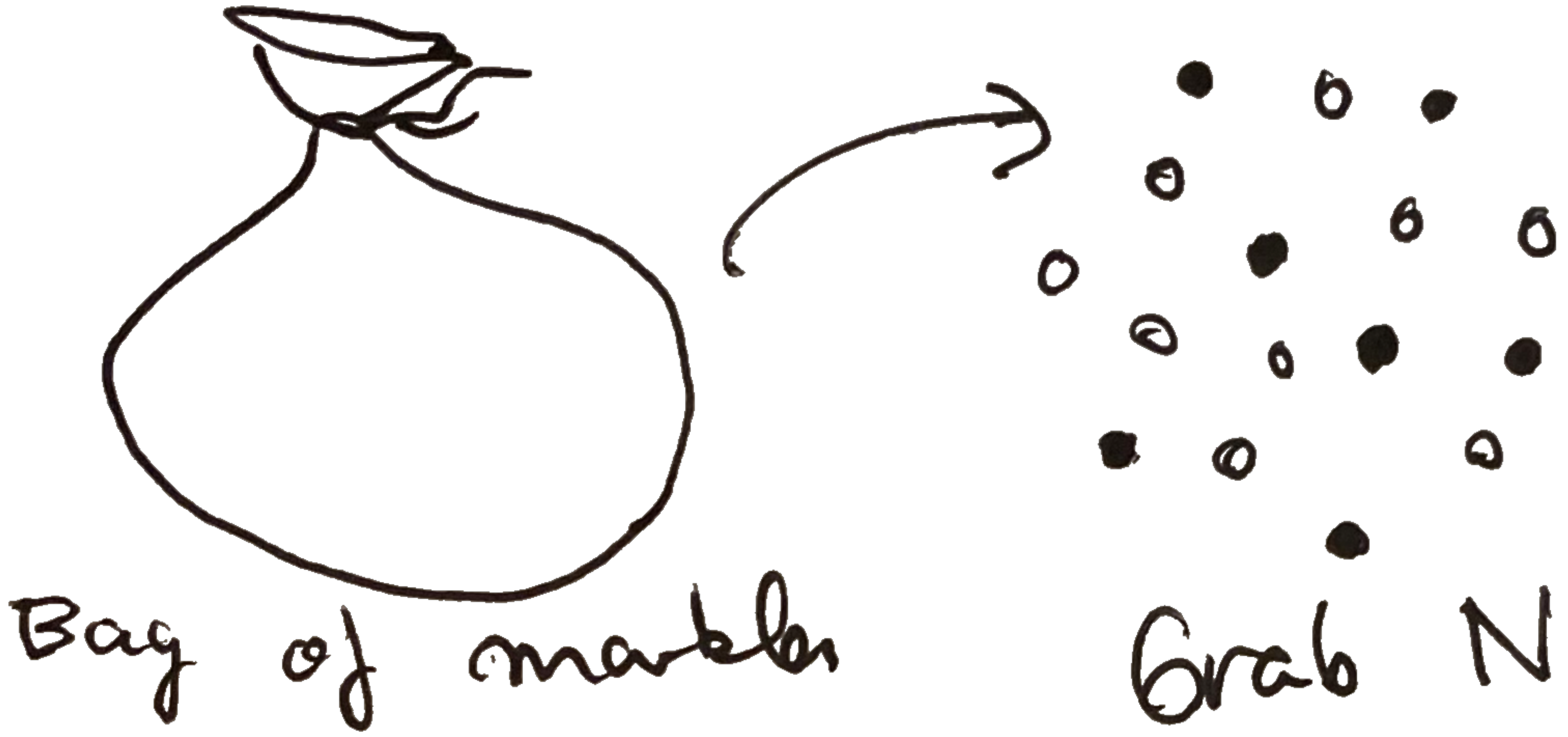
A child’s understanding

Querying the microbiome

A child's understanding

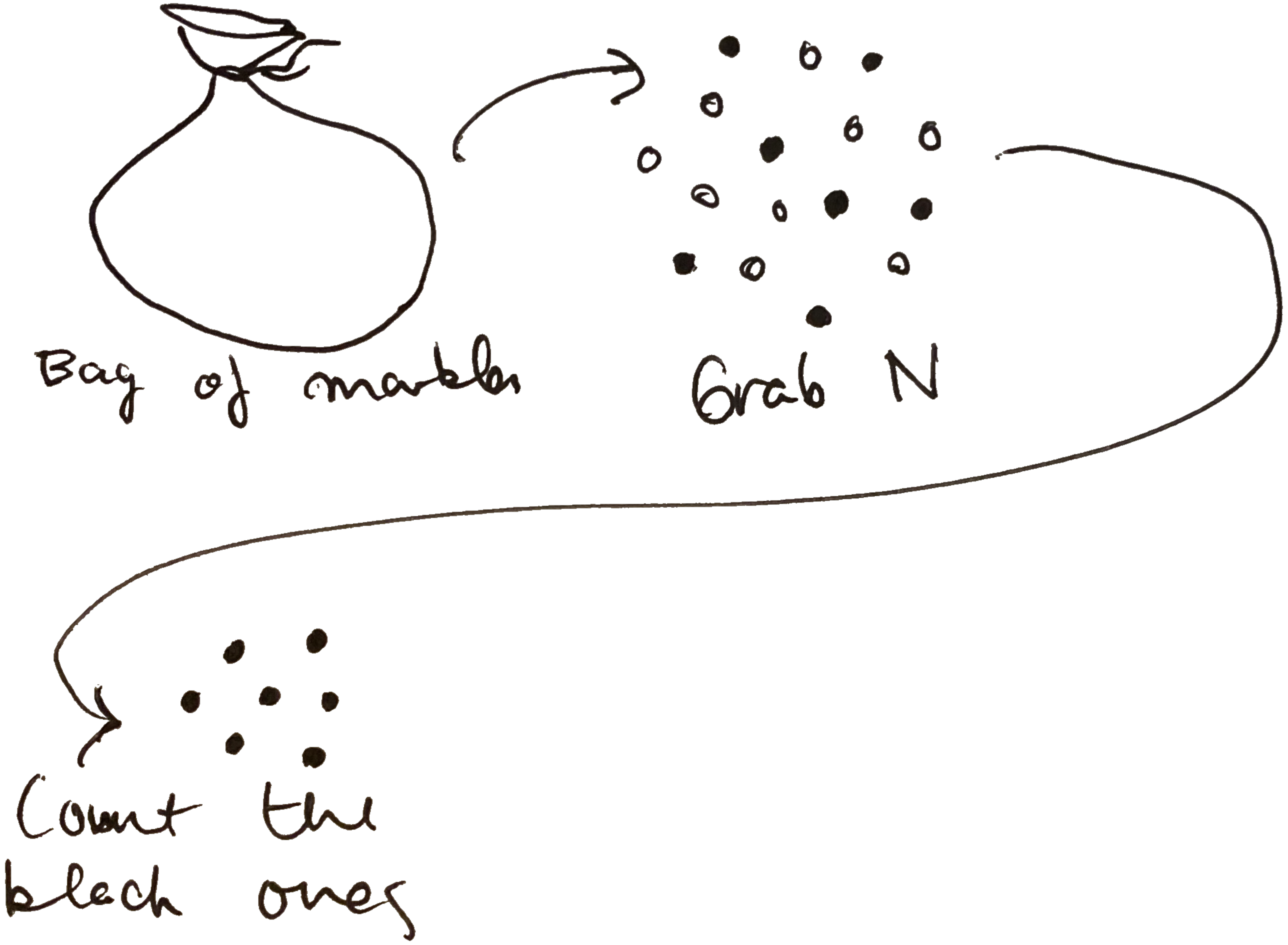
Querying the microbiome

A child's understanding



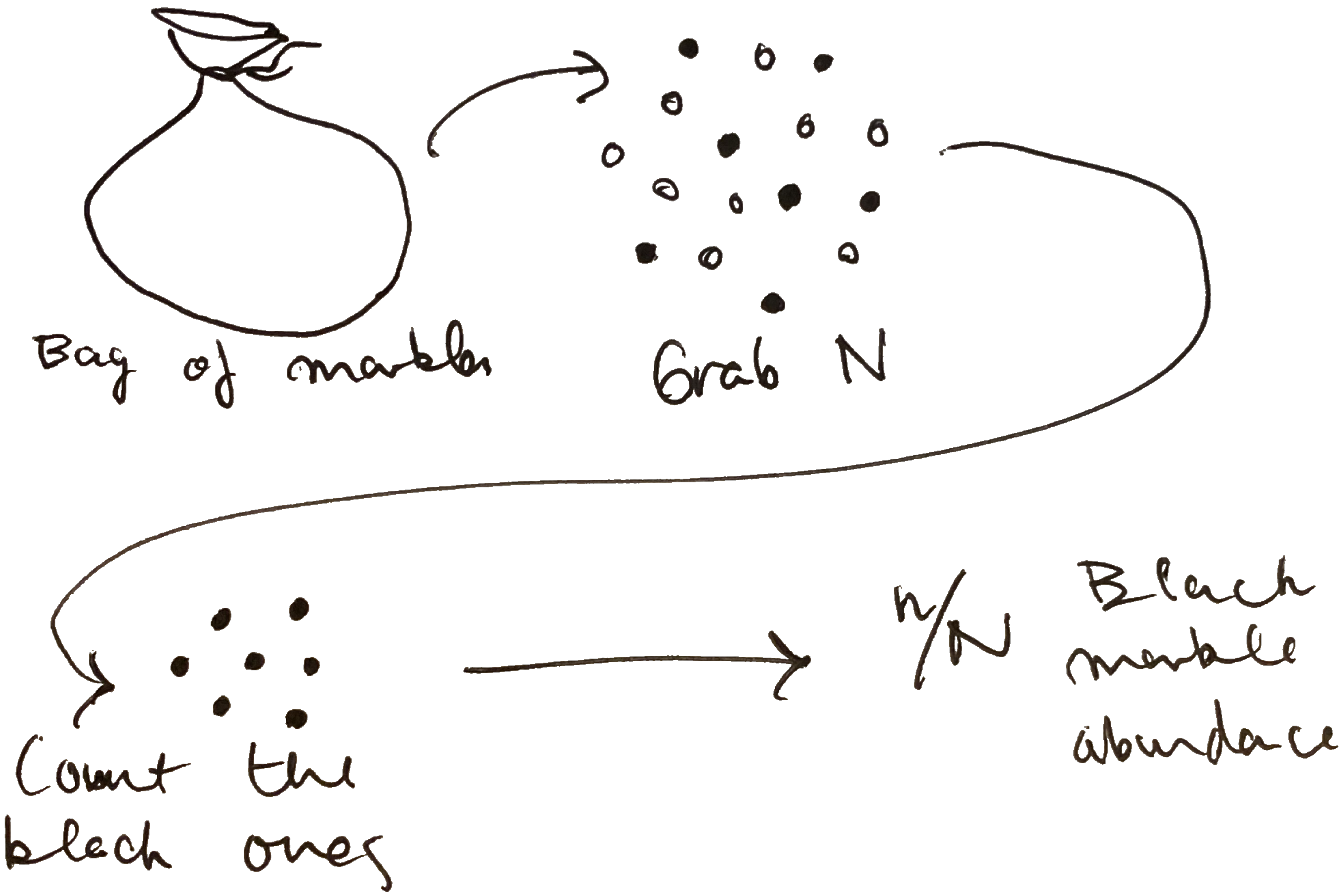
Querying the microbiome

A child's understanding



Querying the microbiome

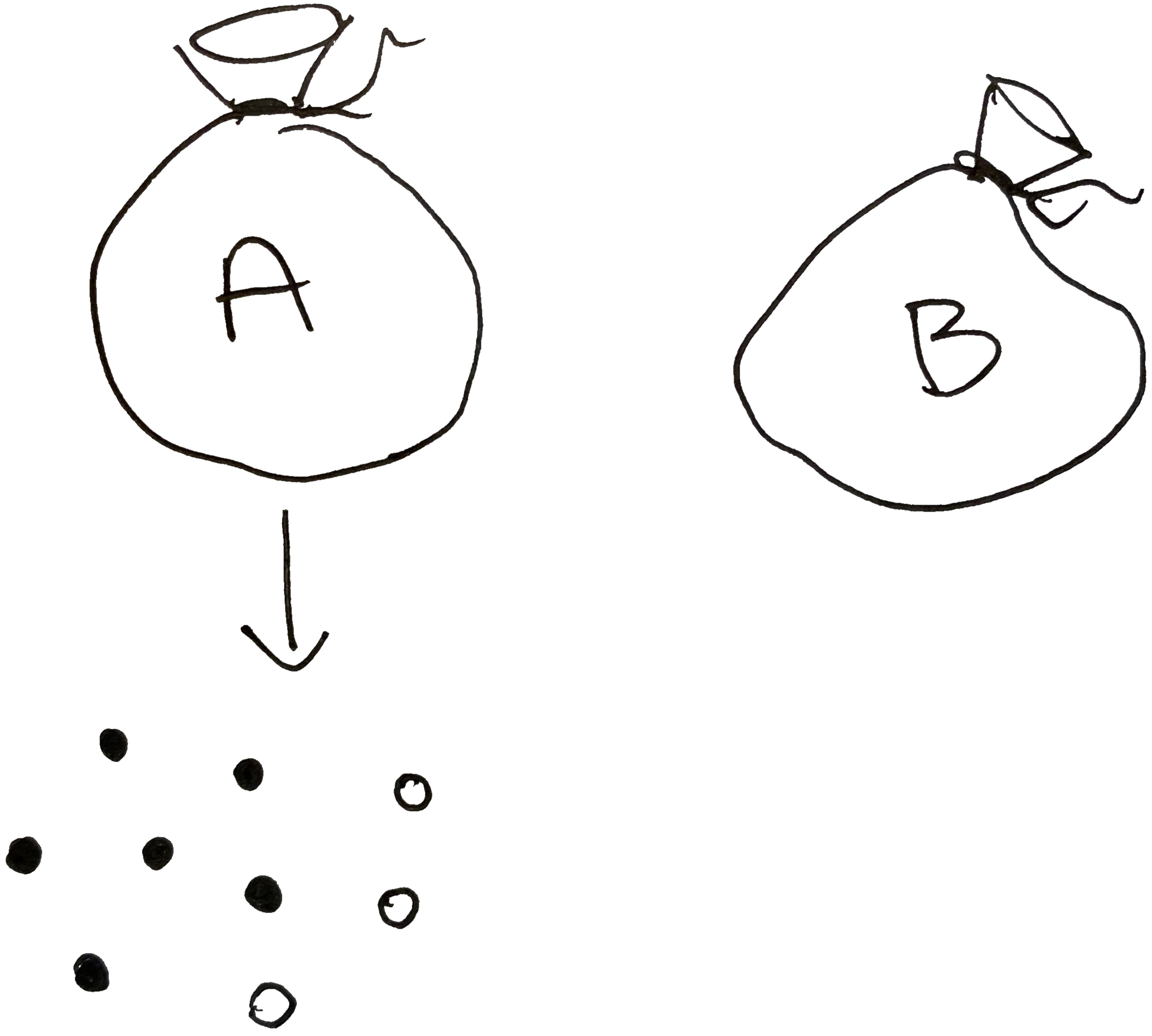
A child's understanding



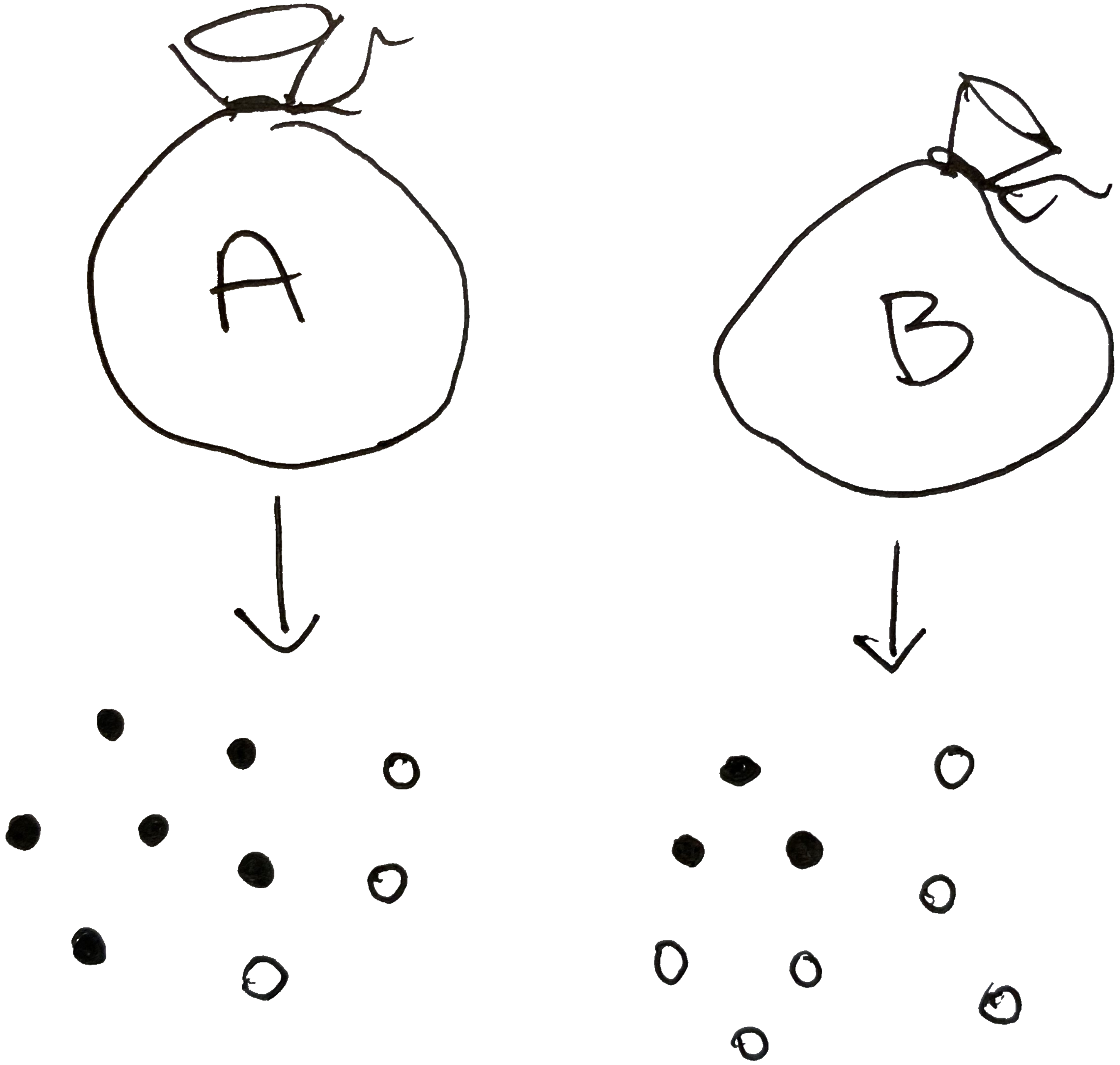
But isn’t interpretation kind of tricky?

But isn’t interpretation kind of tricky?

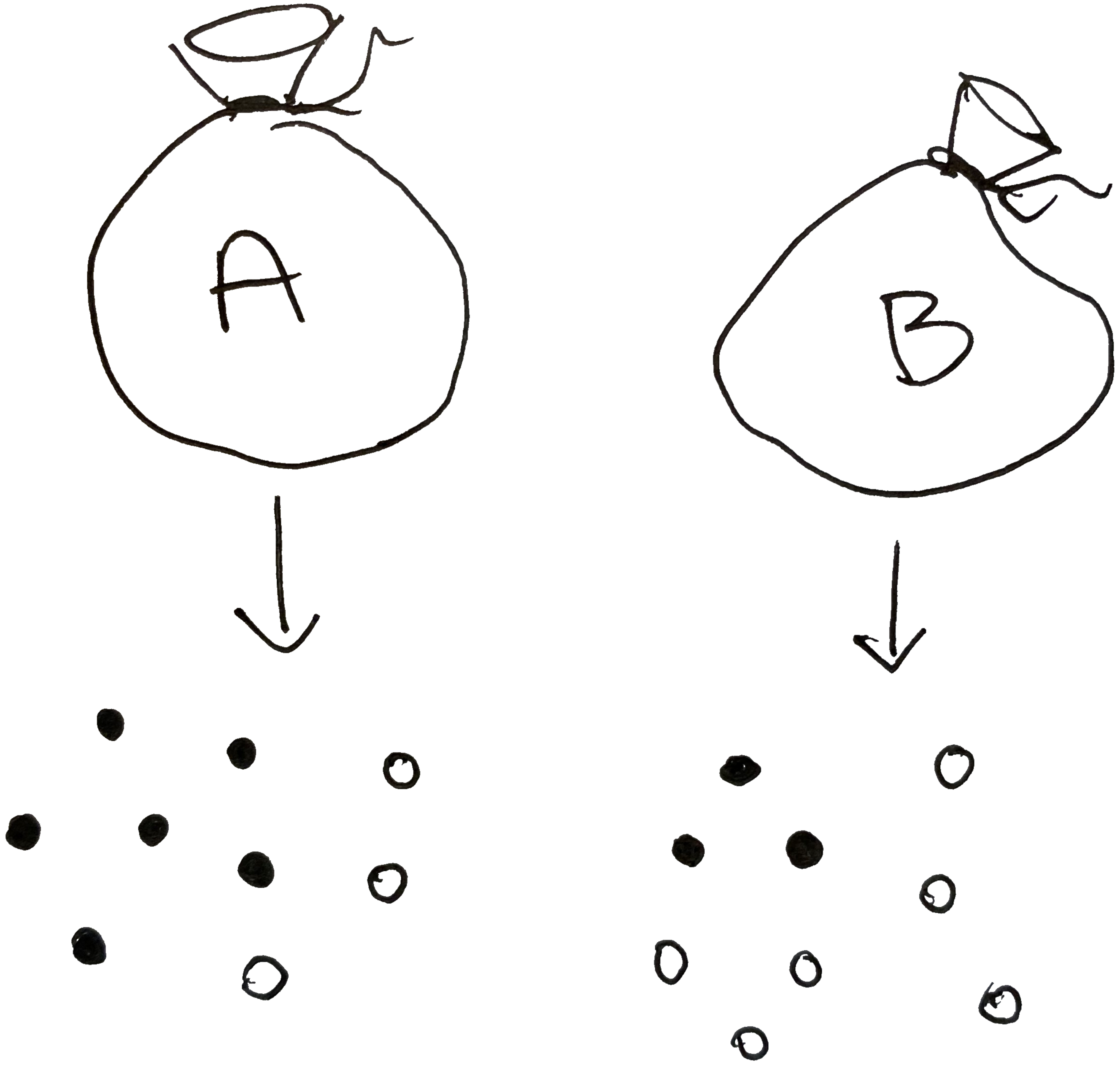
But isn’t interpretation kind of tricky?



But isn’t interpretation kind of tricky?

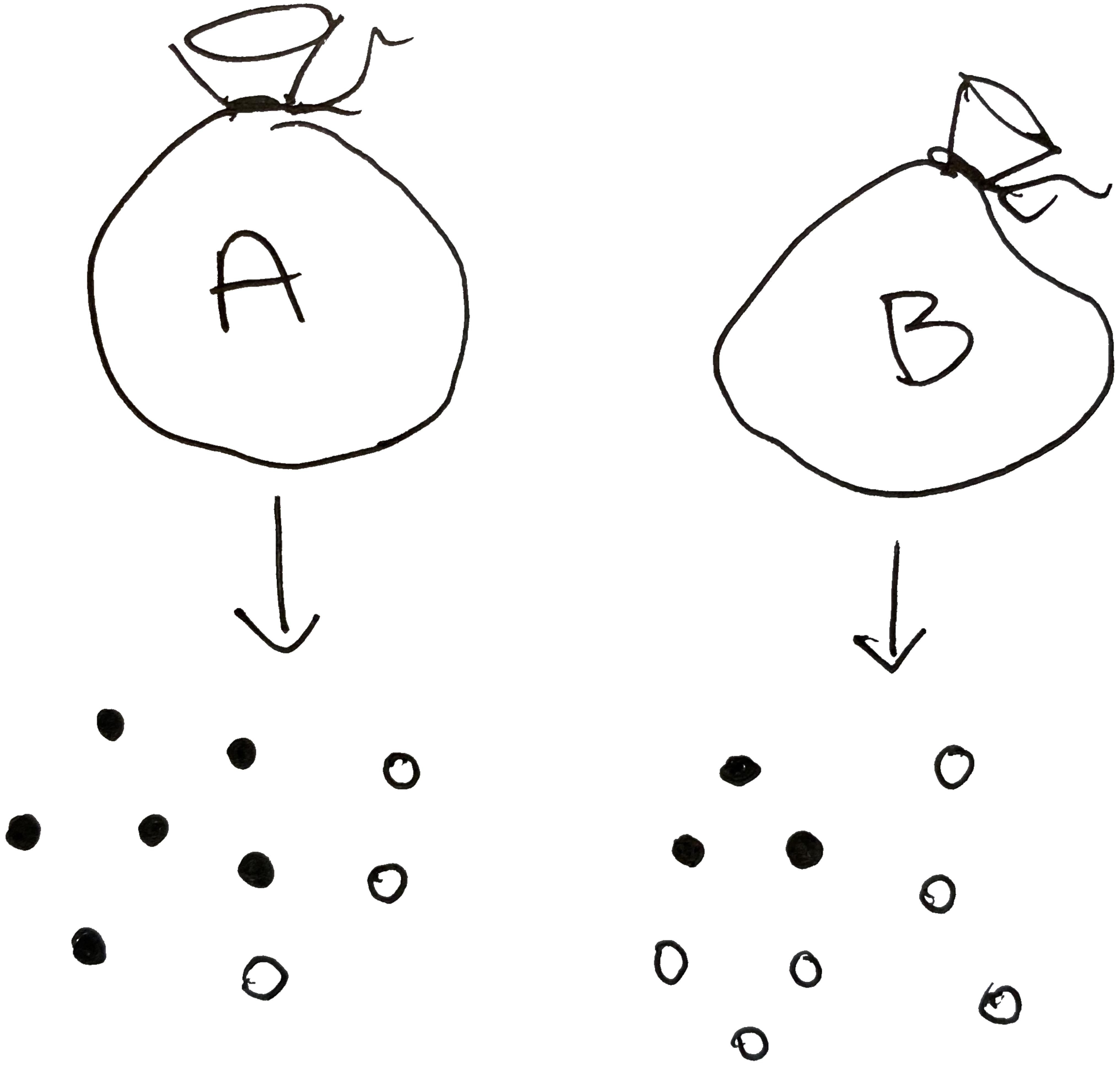


But isn't interpretation kind of tricky?



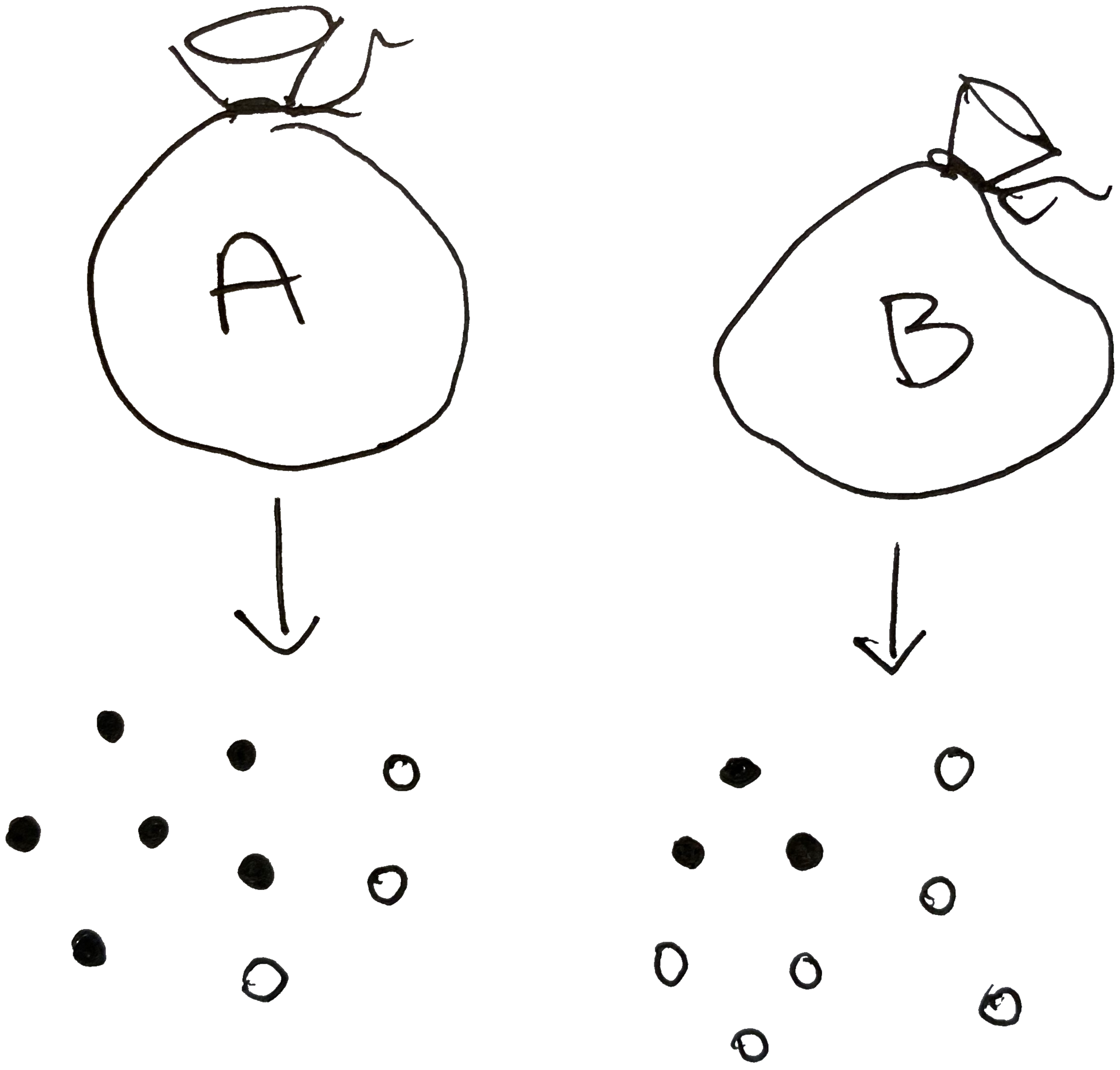
- Do we infer that there are more black marbles in A?

But isn't interpretation kind of tricky?



- Do we infer that there are more black marbles in A?
    - Don't think we can????? See next slide.

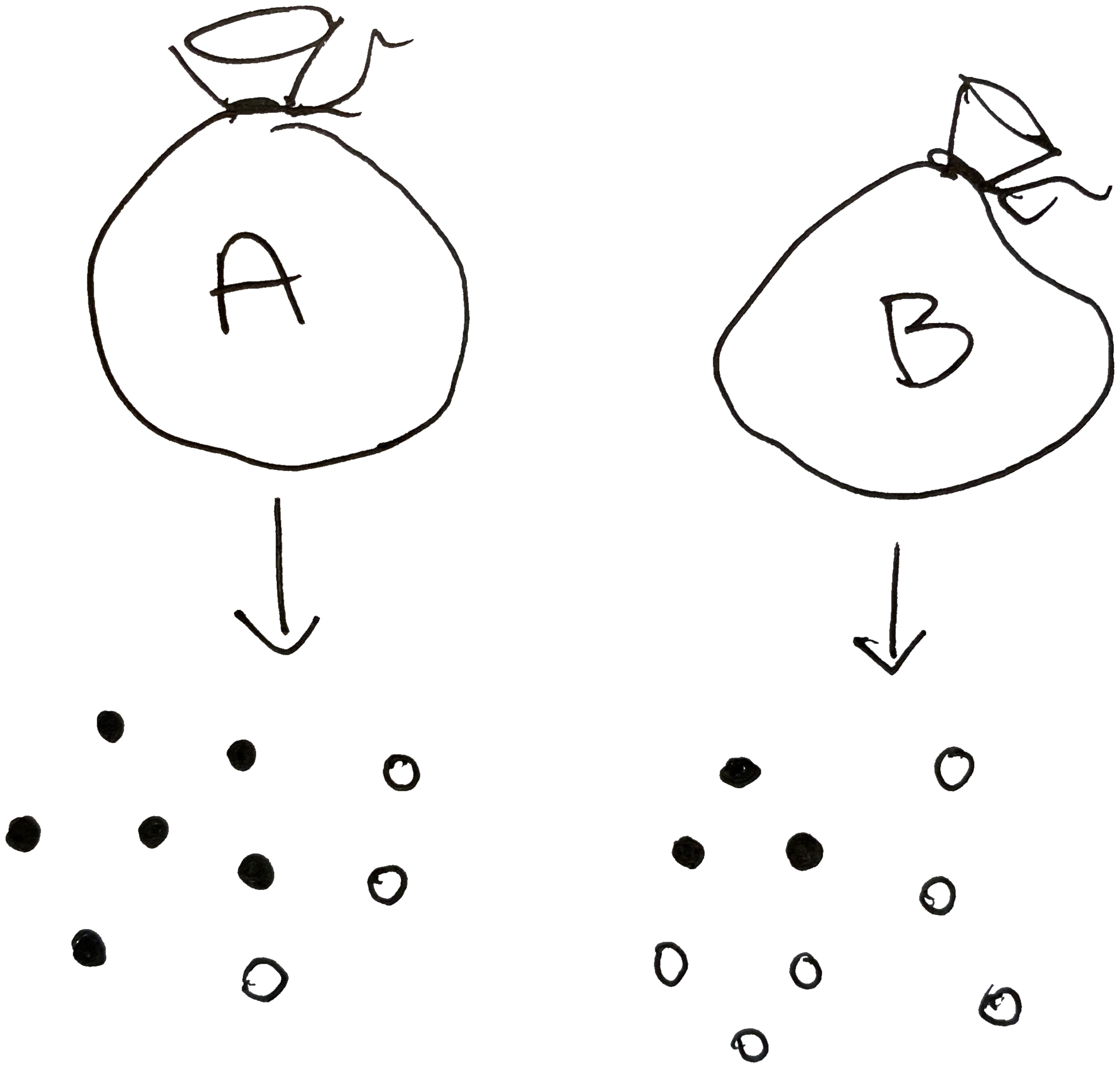
But isn't interpretation kind of tricky?



- Do we infer that there are more black marbles in A?
  - Don't think we can????? See next slide.

- Is it interesting that there are **relatively** more black marbles in B?

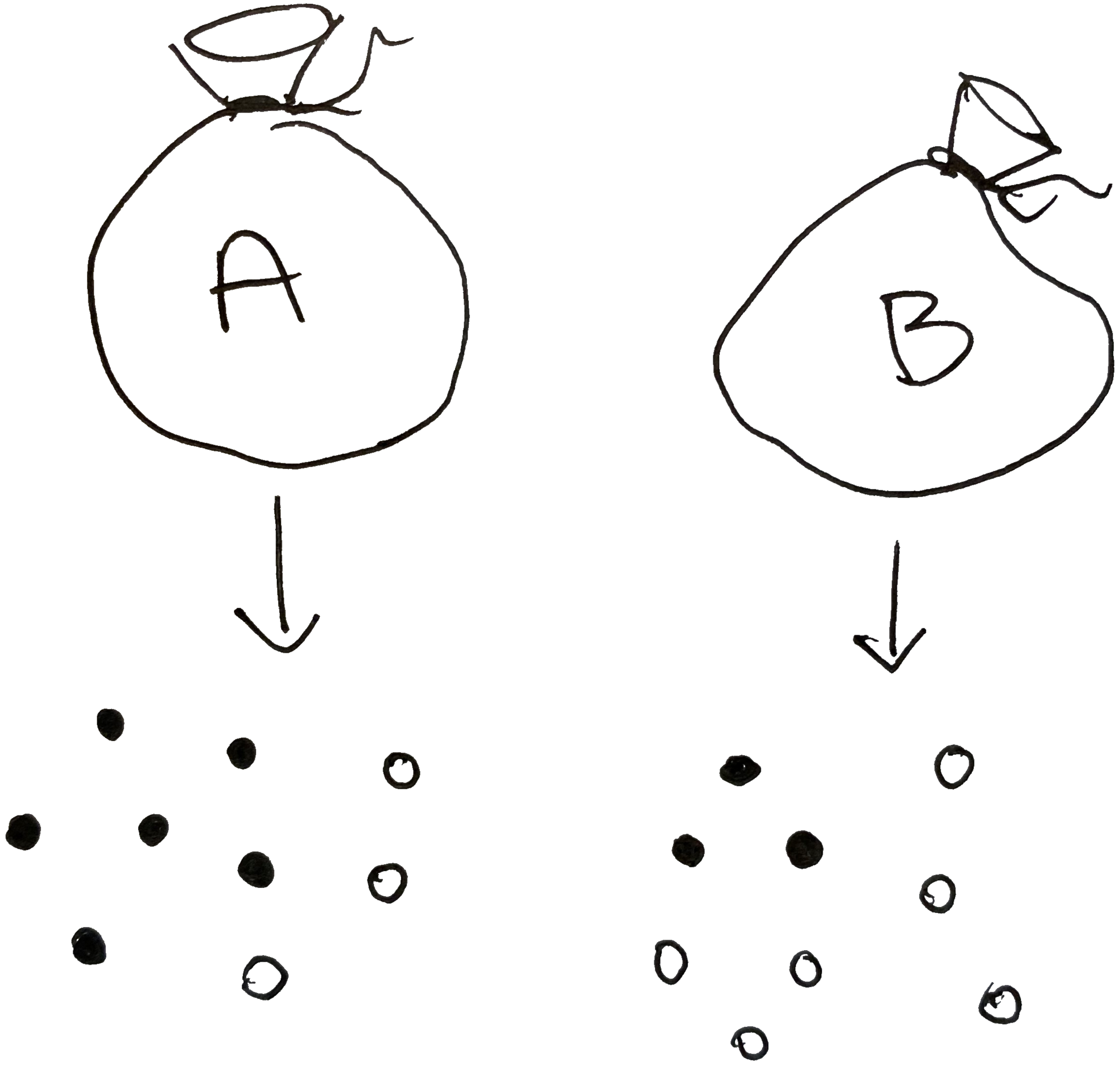
But isn’t interpretation kind of tricky?



- Do we infer that there are more black marbles in A?
  - Don’t think we can????? See next slide.

- Is it interesting that there are **relatively** more black marbles in B?
  - Difficult to say????? What if white marbles are totally unrelated to black marbles? Say black represents Kp and white represents the number of dirty mugs I have in my office right now.

But isn’t interpretation kind of tricky?

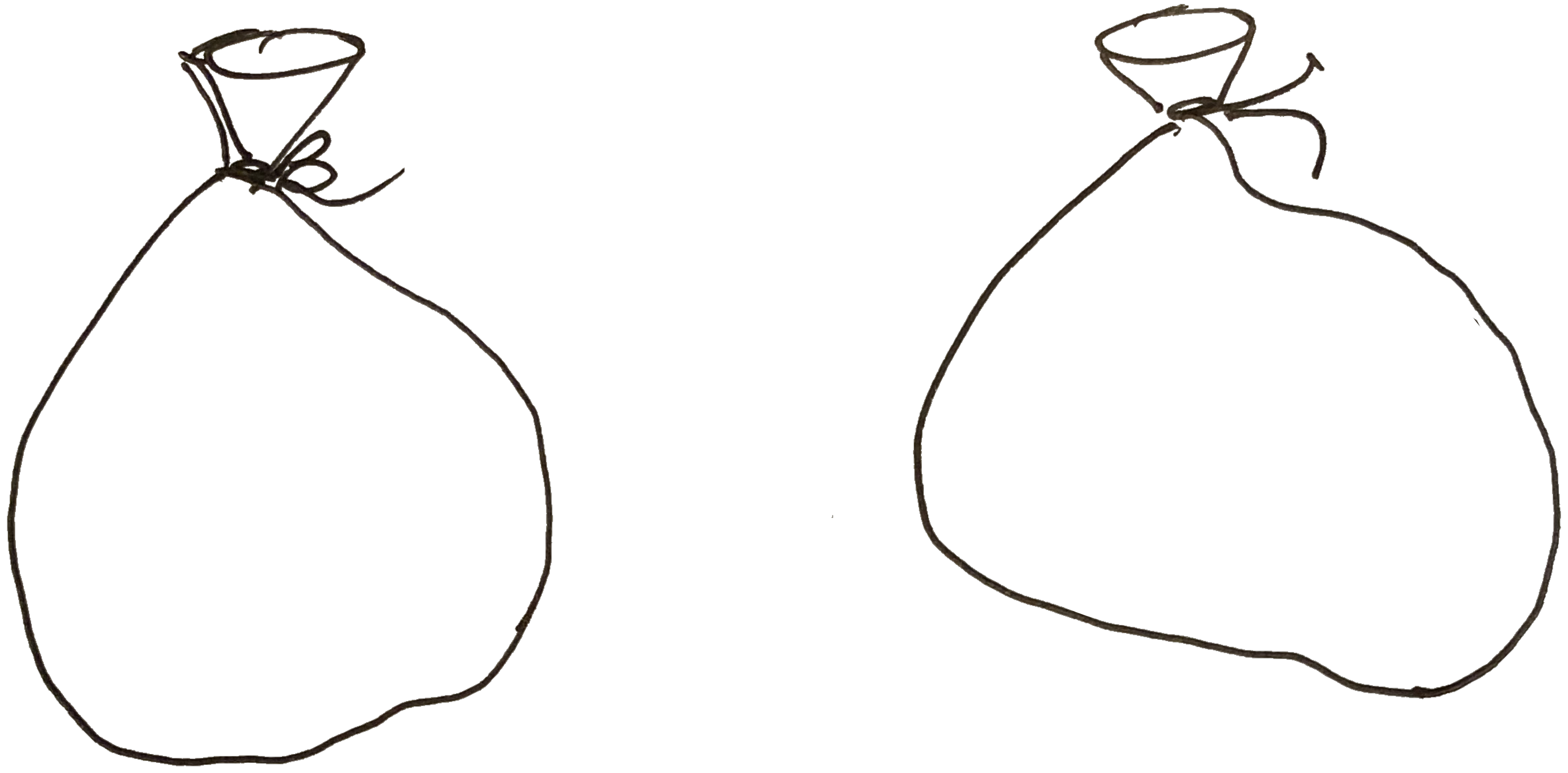


- Do we infer that there are more black marbles in A?
  - Don’t think we can????? See next slide.

- Is it interesting that there are **relatively** more black marbles in B?
  - Difficult to say????? What if white marbles are totally unrelated to black marbles? Say black represents Kp and white represents the number of dirty mugs I have in my office right now.
  - Problem: we can’t choose to spend our reads only on interesting species.

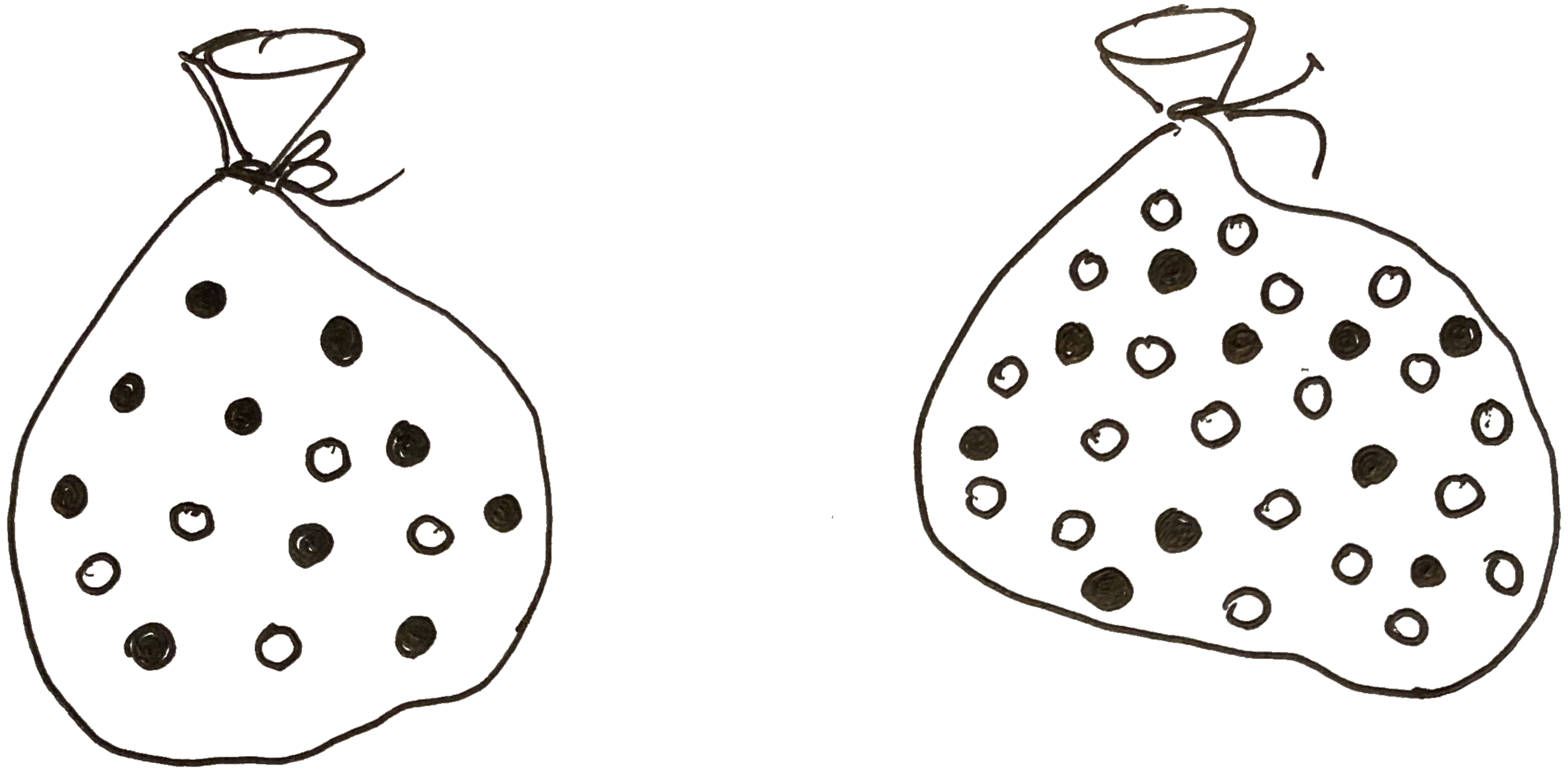
What's in the bag? We don't really know?

It's easy to contrive scenarios of alternative interpretations



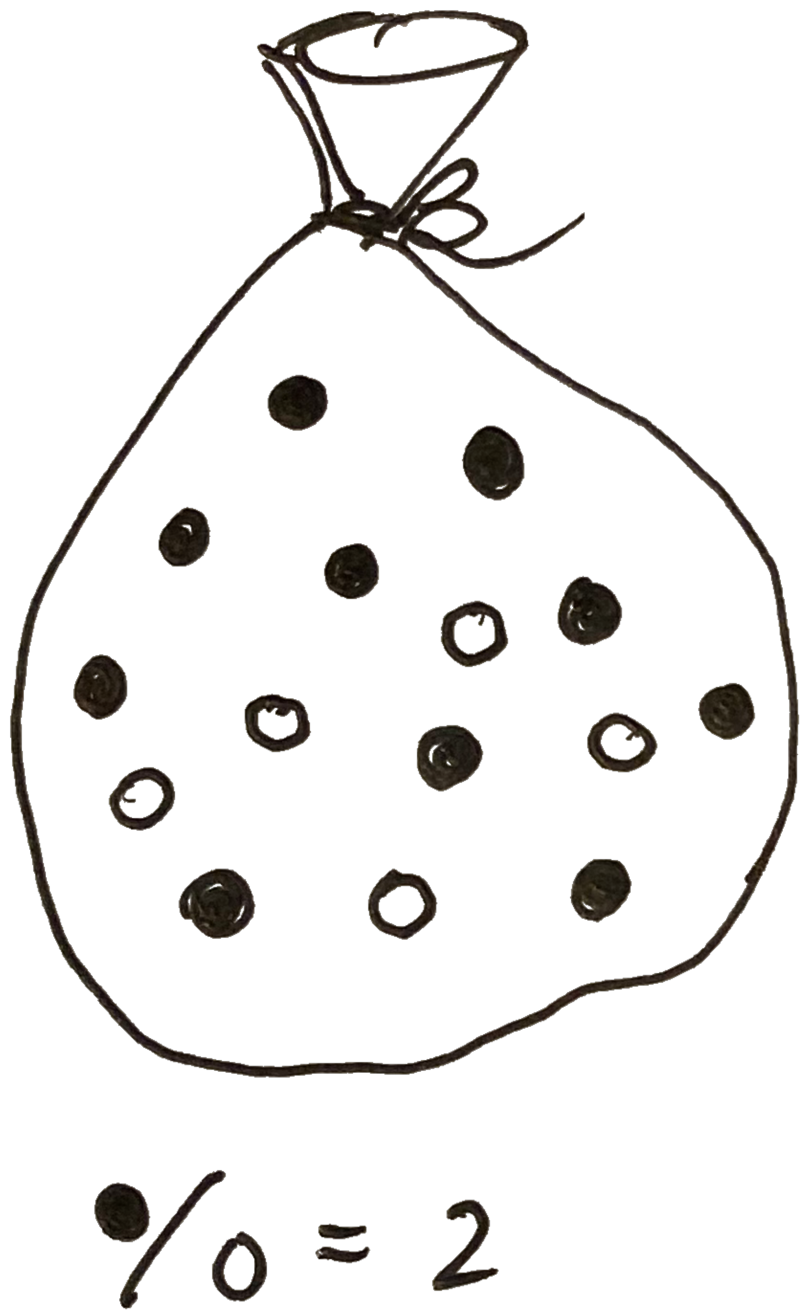
What’s in the bag? We don’t really know?

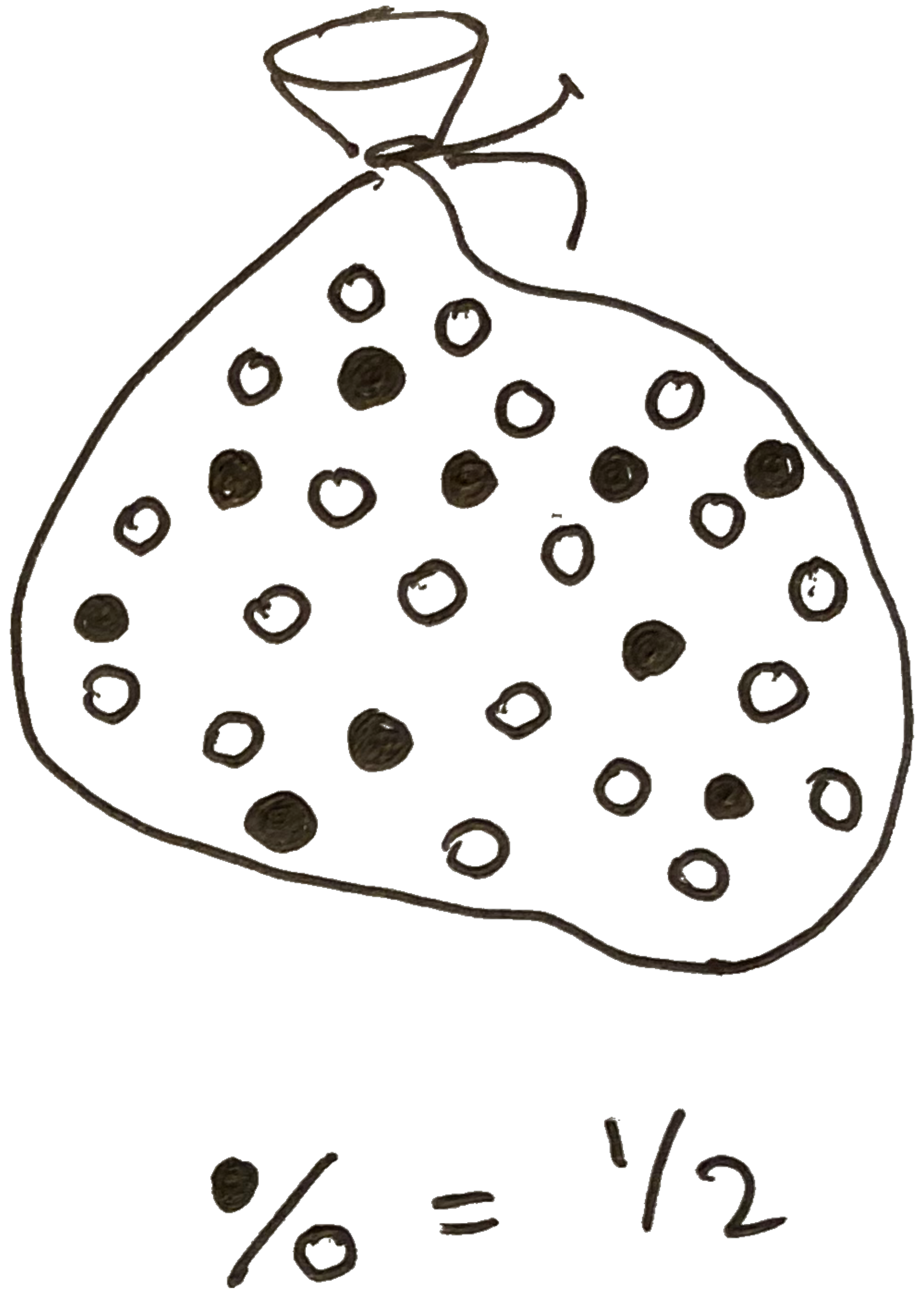
It’s easy to contrive scenarios of alternative interpretations



What's in the bag? We don't really know?

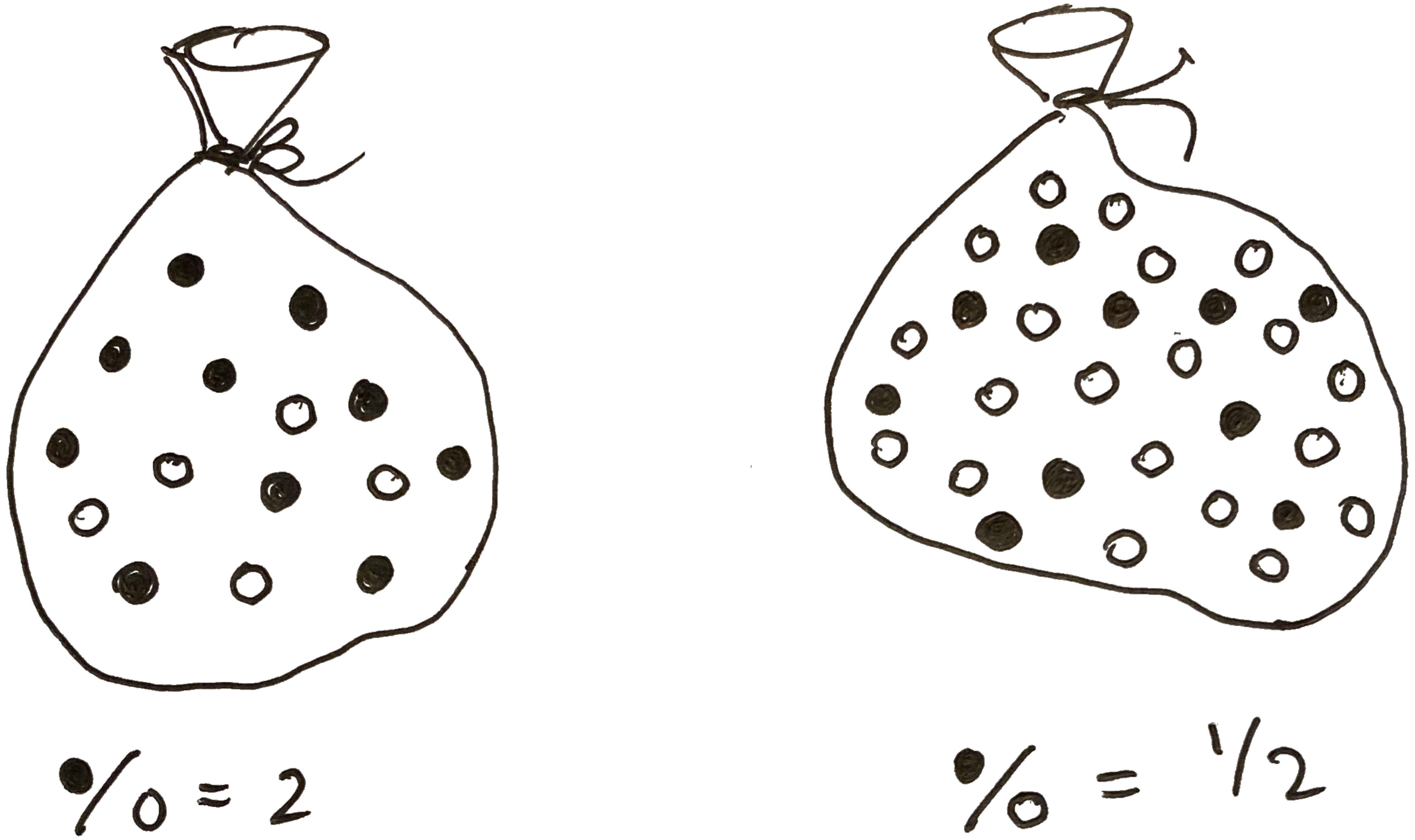
It's easy to contrive scenarios of alternative interpretations





What's in the bag? We don't really know?

It's easy to contrive scenarios of alternative interpretations

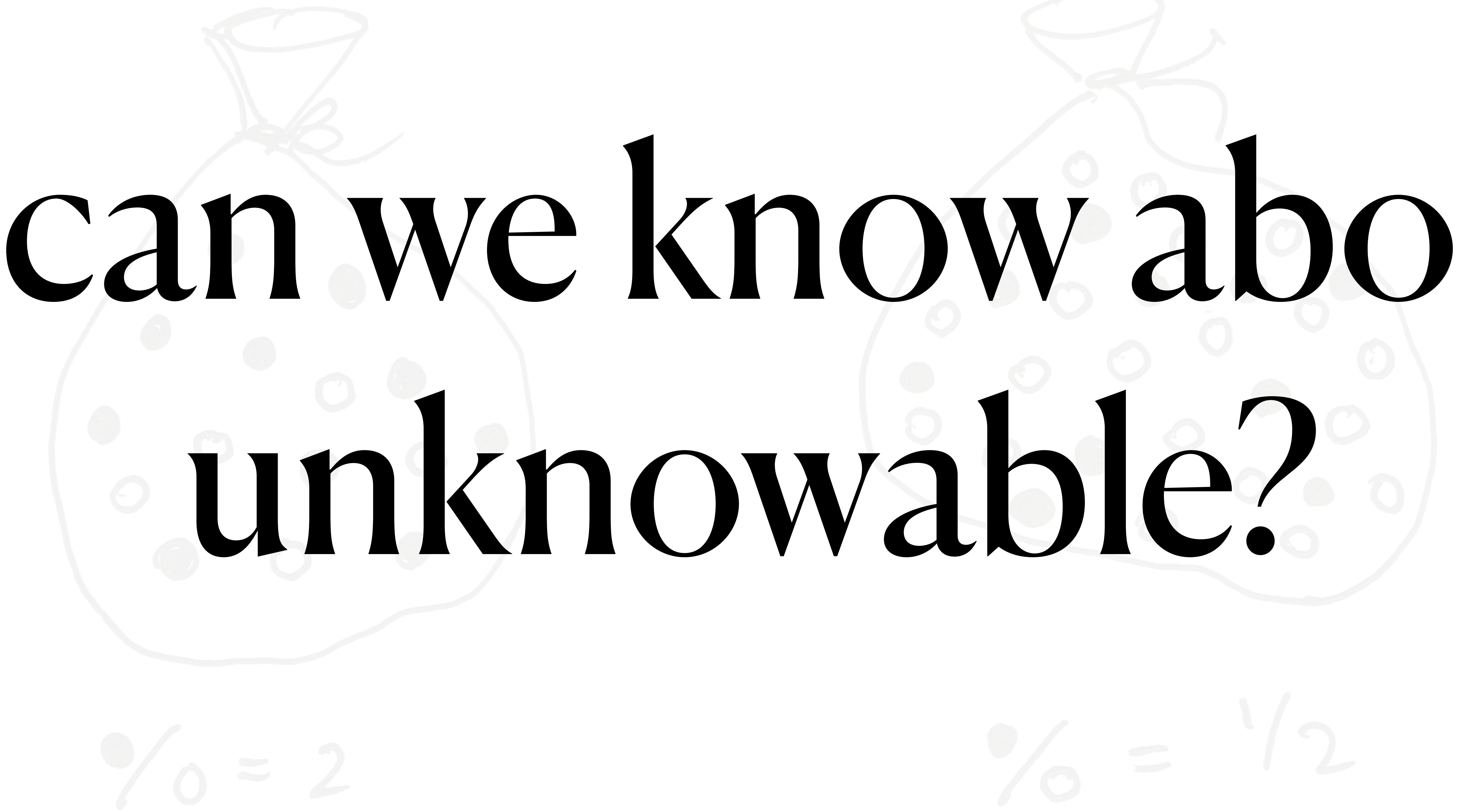


10 black marbles in both bags.

What’s in the bag? We don’t really know?

It’s easy to contrive scenarios of alternative interpretations

How can we know about the unknowable?



10 black marbles in both bags.

Part 2: Crossroads

The Robert Johnson story —
You can play the blues for a price

The Devil in your computer —
You can learn about the marbles for a price

The devil inside the computer requires assumptions and data

The devil inside the computer requires assumptions and data

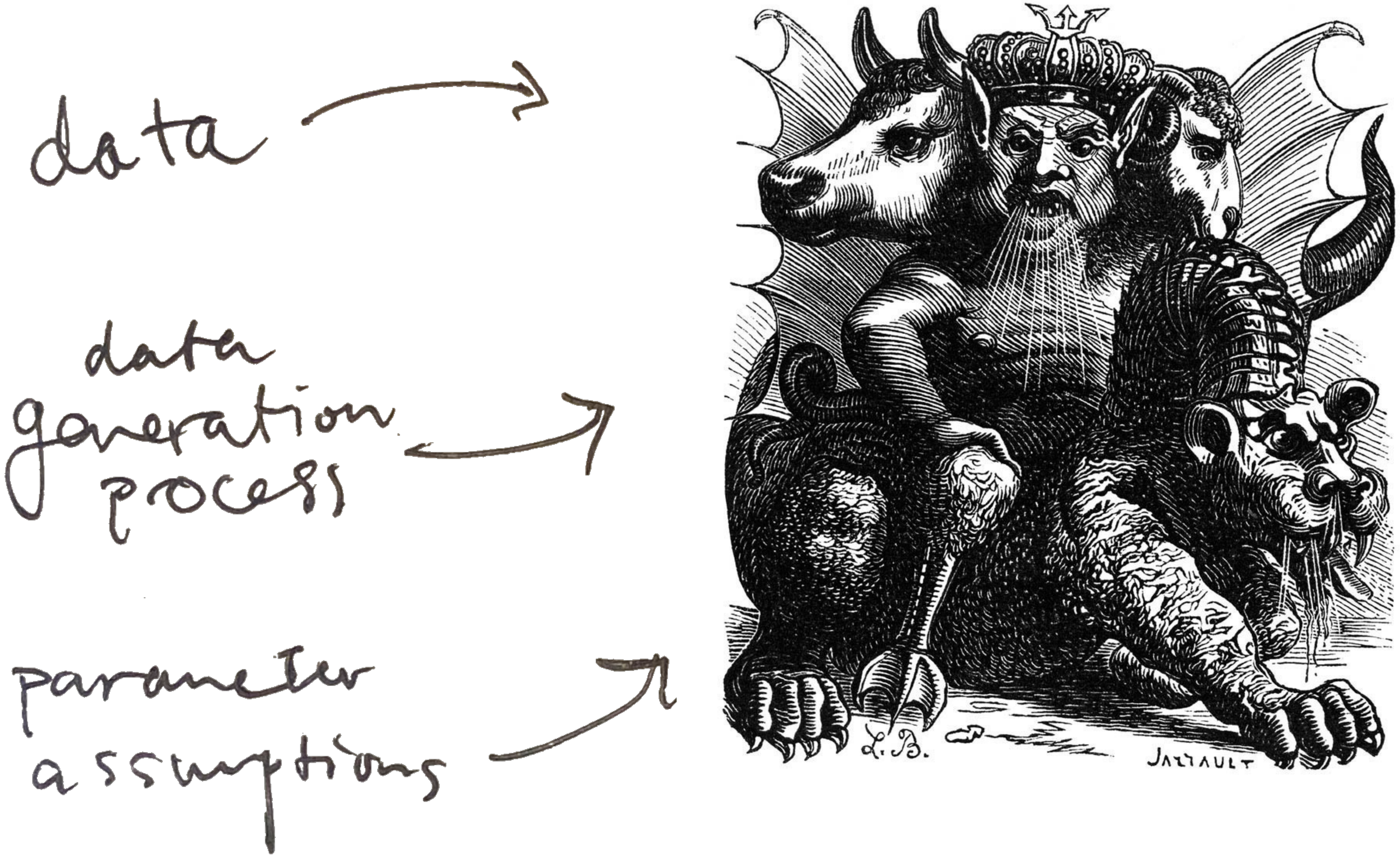
data →

The devil inside the computer requires assumptions and data

data →

data generation process →

The devil inside the computer requires assumptions and data

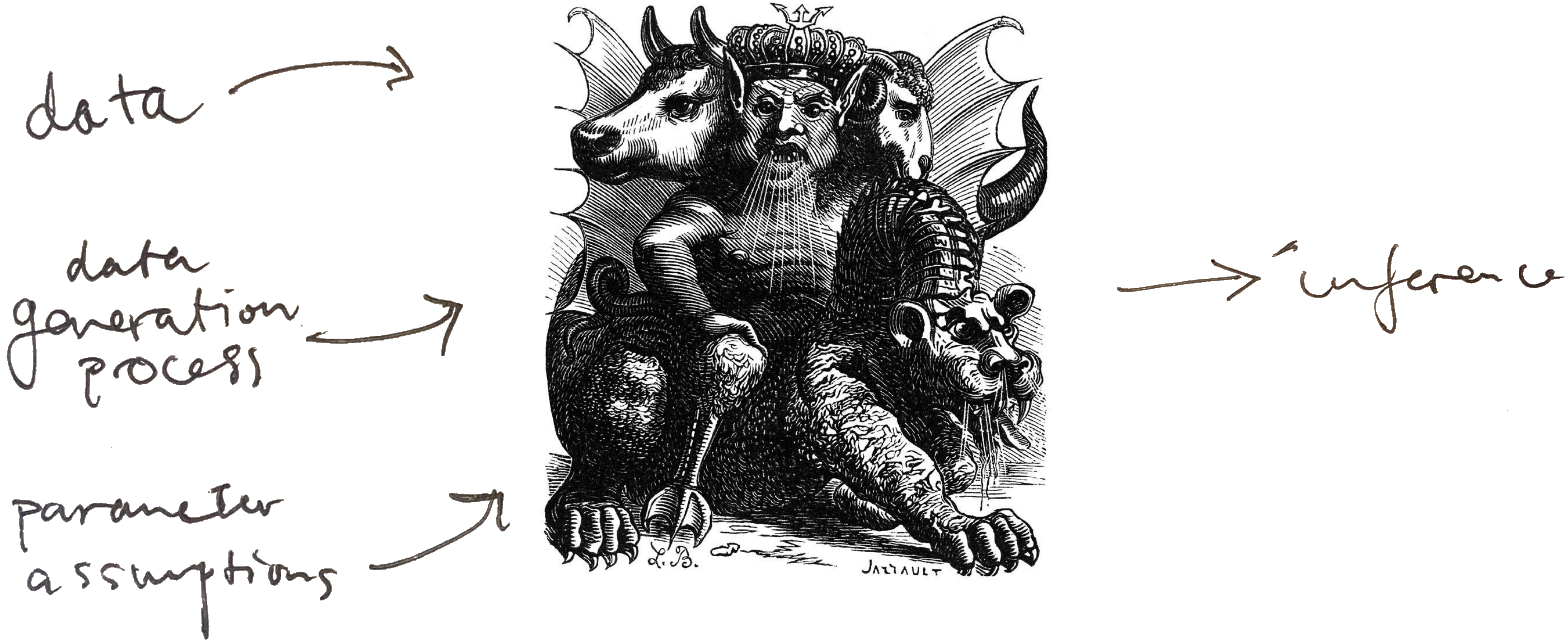


data

data generation process

parameter assumptions

The devil inside the computer requires assumptions and data



data

data generation process

parameter assumptions

inference

Experimental demonology setup

Experimental demonology setup

- 100 identical bags of marbles containing 100 000 black marbles

Experimental demonology setup

- 100 identical bags of marbles containing 100 000 black marbles
- each black marble has a 60% chance of getting recovered on drawing from a bag

Experimental demonology setup

- 100 identical bags of marbles containing 100 000 black marbles
- each black marble has a 60% chance of getting recovered on drawing from a bag
- Results in data like 59 836 (bag 1), 60 122 (bag 2), 59 952 (bag 3), etc., etc.

Experimental demonology setup

- 100 identical bags of marbles containing 100 000 black marbles
- each black marble has a 60% chance of getting recovered on drawing from a bag
- Results in data like 59 836 (bag 1), 60 122 (bag 2), 59 952 (bag 3), etc., etc.
- Now: forget that we know number of marbles and the prob. of recovery

Experimental demonology setup

```
parameters {
  real<lower=0> N;
  real<lower=0, upper=1> p;
  real<lower=0> sigma;

}

model {
  log(n) ~ normal(log(N) + log(p), sigma);

  N ~ normal(mu_n, sd_n);
  p ~ beta(alpha, beta);
  sigma ~ exponential(.1);
}
```

→

N = true number of black marbles (unknown)

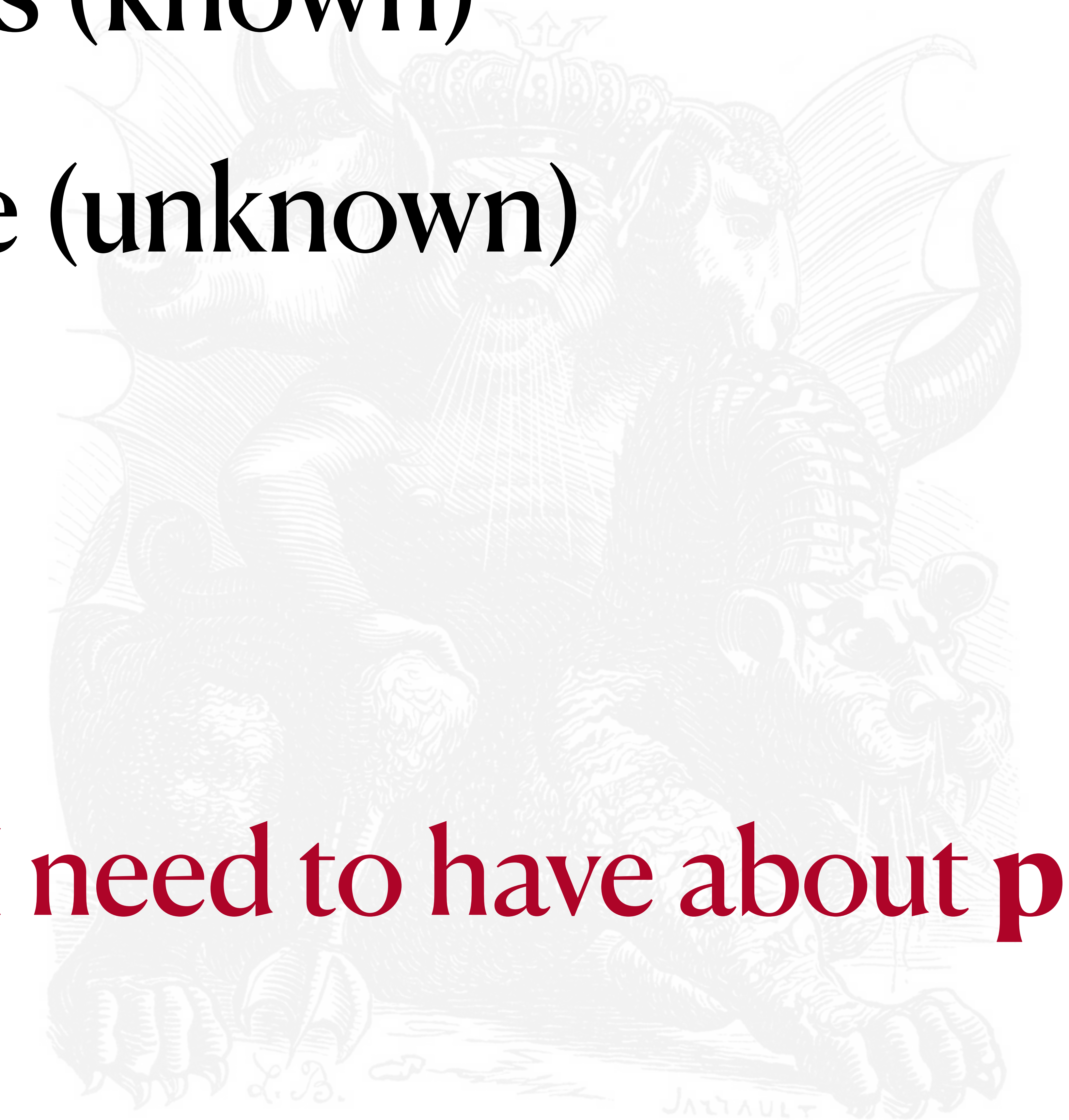
n = observed number of black marbles (known)

p = prob. of recovering a black marble (unknown)

Data model: On average **n** = **Np**

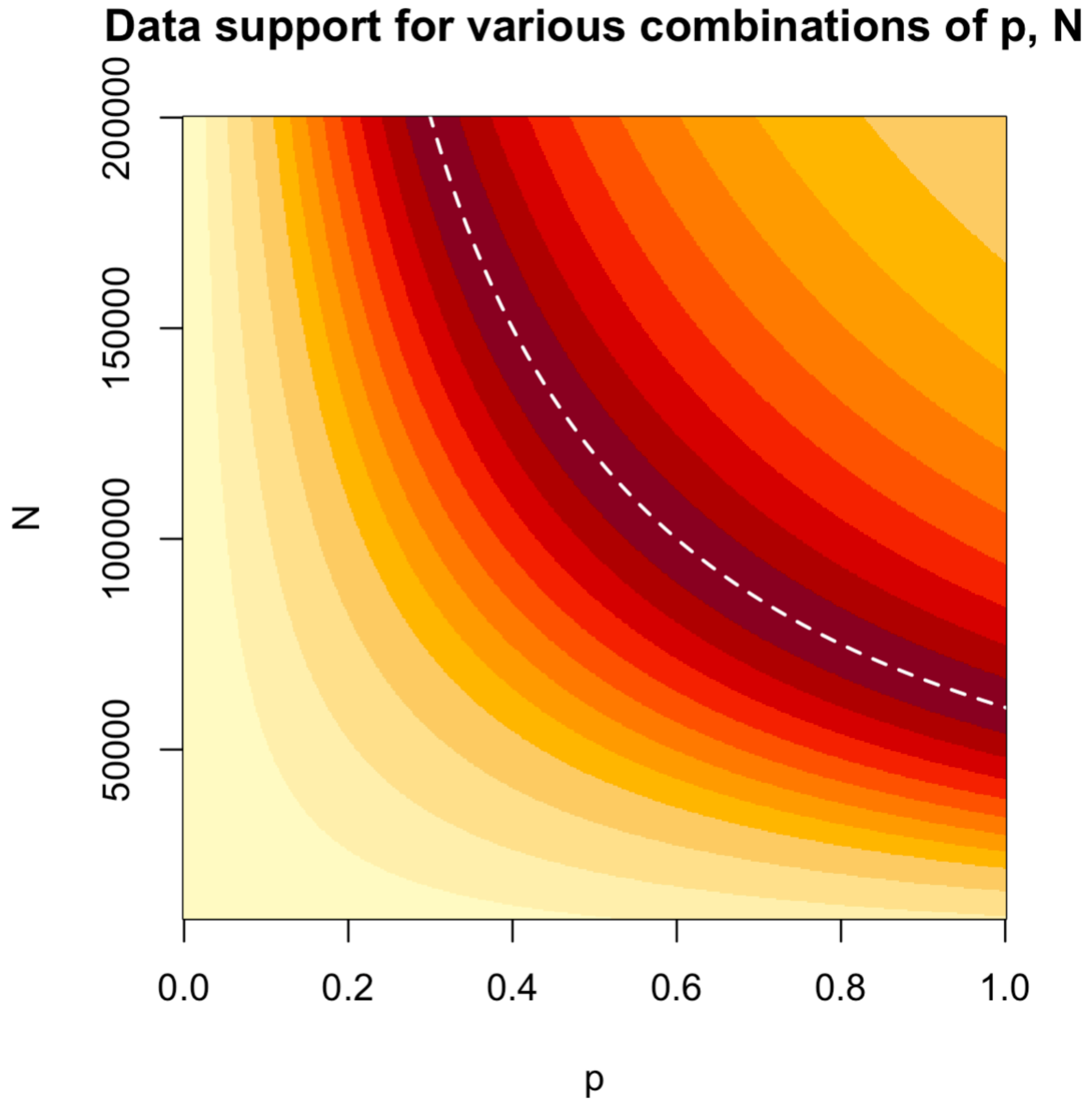
Question: how strong an opinion do I need to have about p to learn about N?

```
parameters {
  real<lower=0> N;
  real<lower=0, upper=1> p;
  real<lower=0> sigma;
}

model {
  log(n) ~ normal(log(N) + log(p), sigma);

  N ~ normal(mu_n, sd_n);
  p ~ beta(alpha, beta);
  sigma ~ exponential(.1);
}
```



Can't have both p and N for free

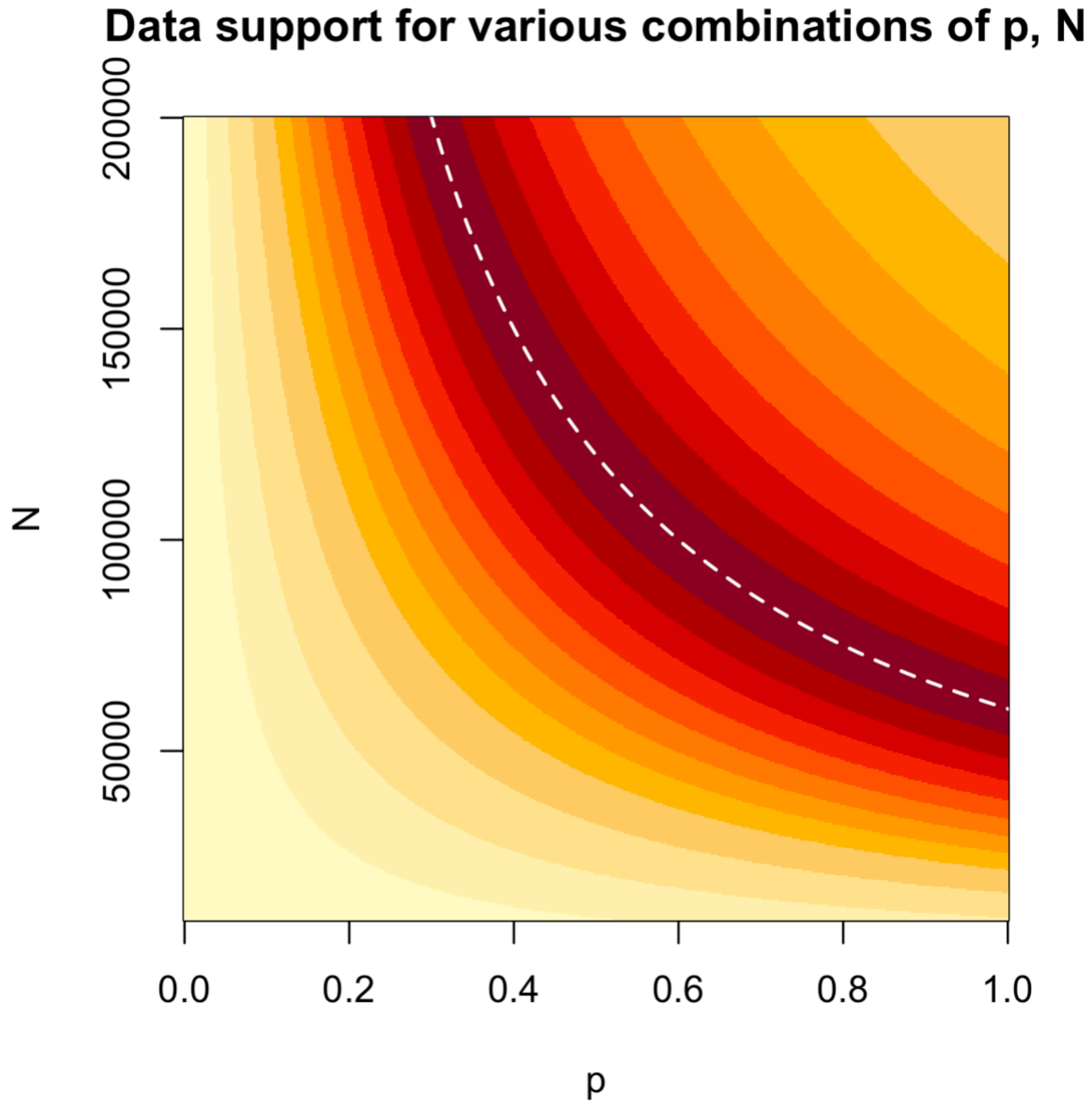
Also: higher true p better for isolating N



Dashed line describes **Np** = **60 000**. We know from grade school that if we fix one of N or p we can solve for the other. Infinite pairs fulfill the equation.

Can’t have both p and N for free

Also: higher true p better for isolating N



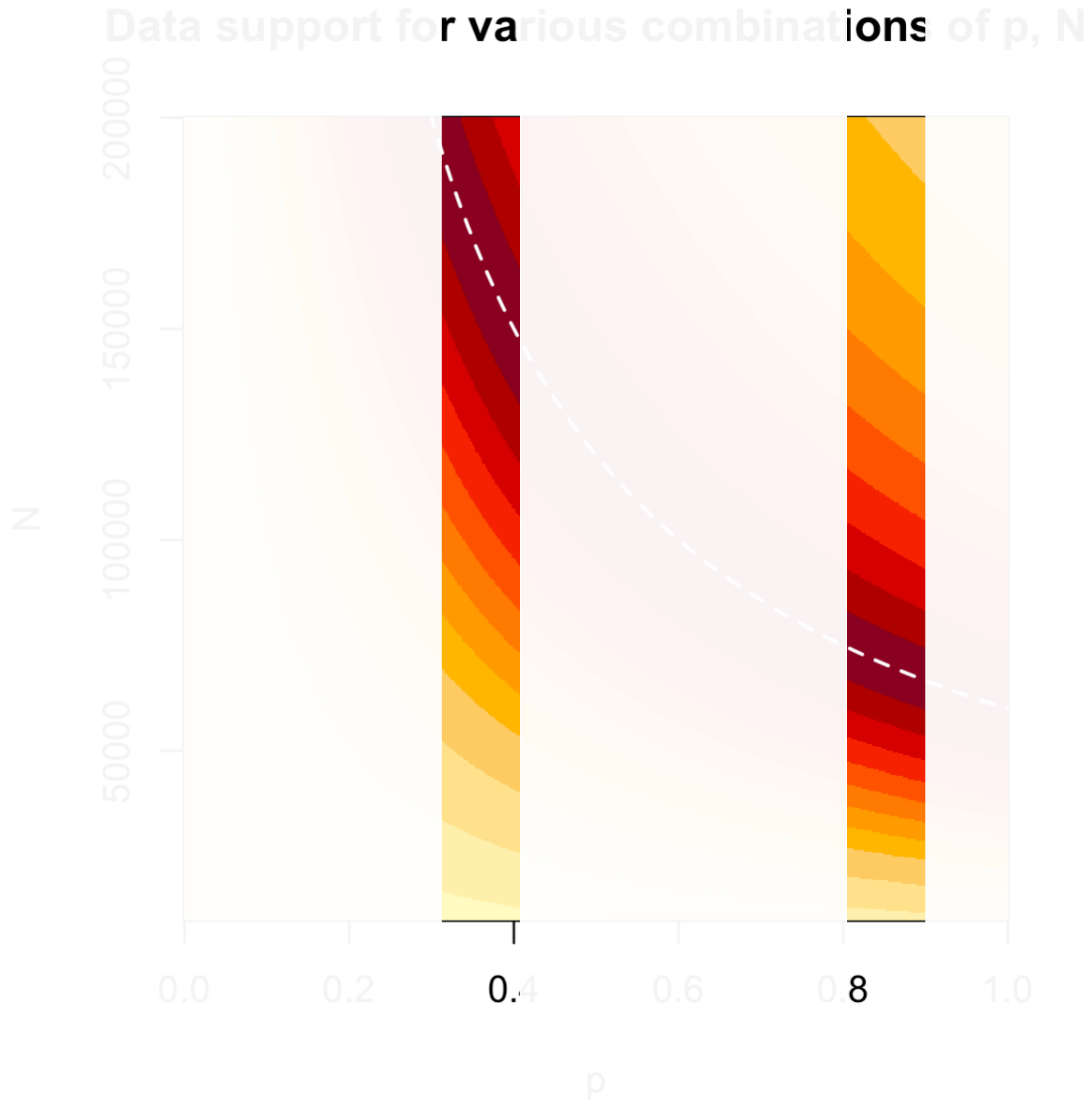
Dashed line describes **Np** = **60 000**. We know from grade school that if we fix one of N or p we can solve for the other. Infinite pairs fulfill the equation.

If **p ~ .3**, the region of uncertainty is quite wide.

But **p ~ .8** is almost comfortable

Can’t have both p and N for free

Also: higher true p better for isolating N



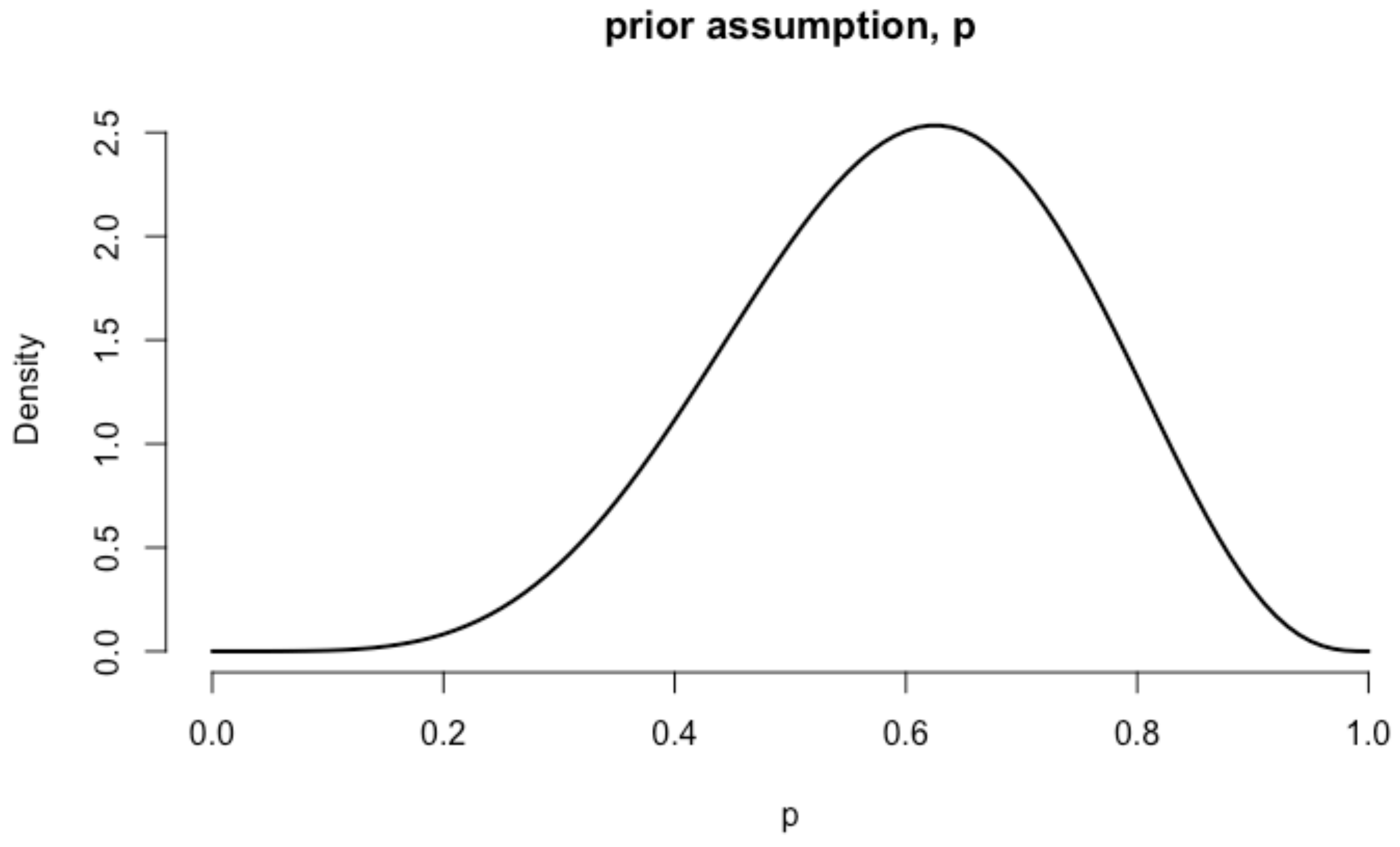
Dashed line describes **Np** = **60 000**. We know from grade school that if we fix one of N or p we can solve for the other. Infinite pairs fulfill the equation.

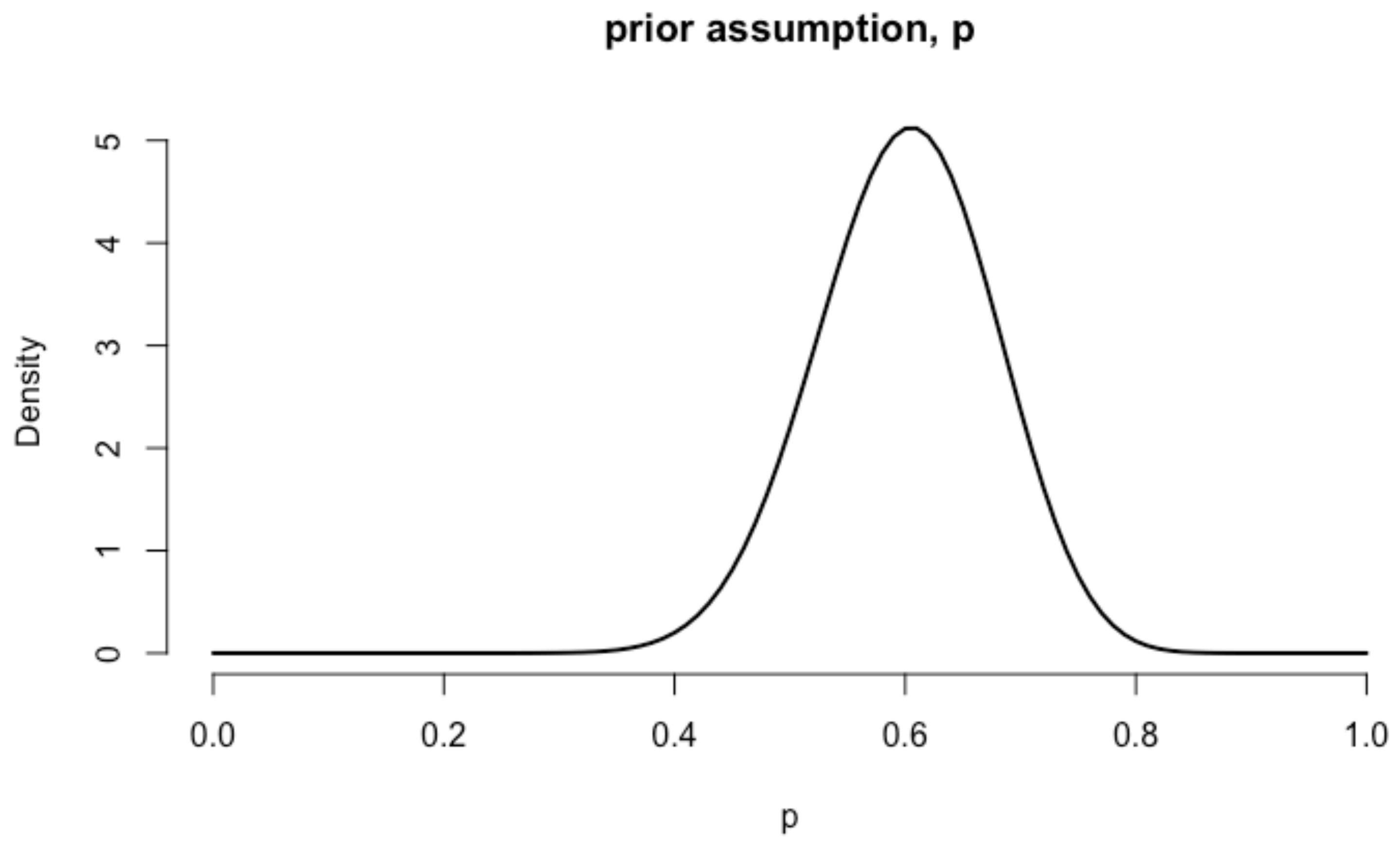
If **p ~ .3**, the region of uncertainty is quite wide.

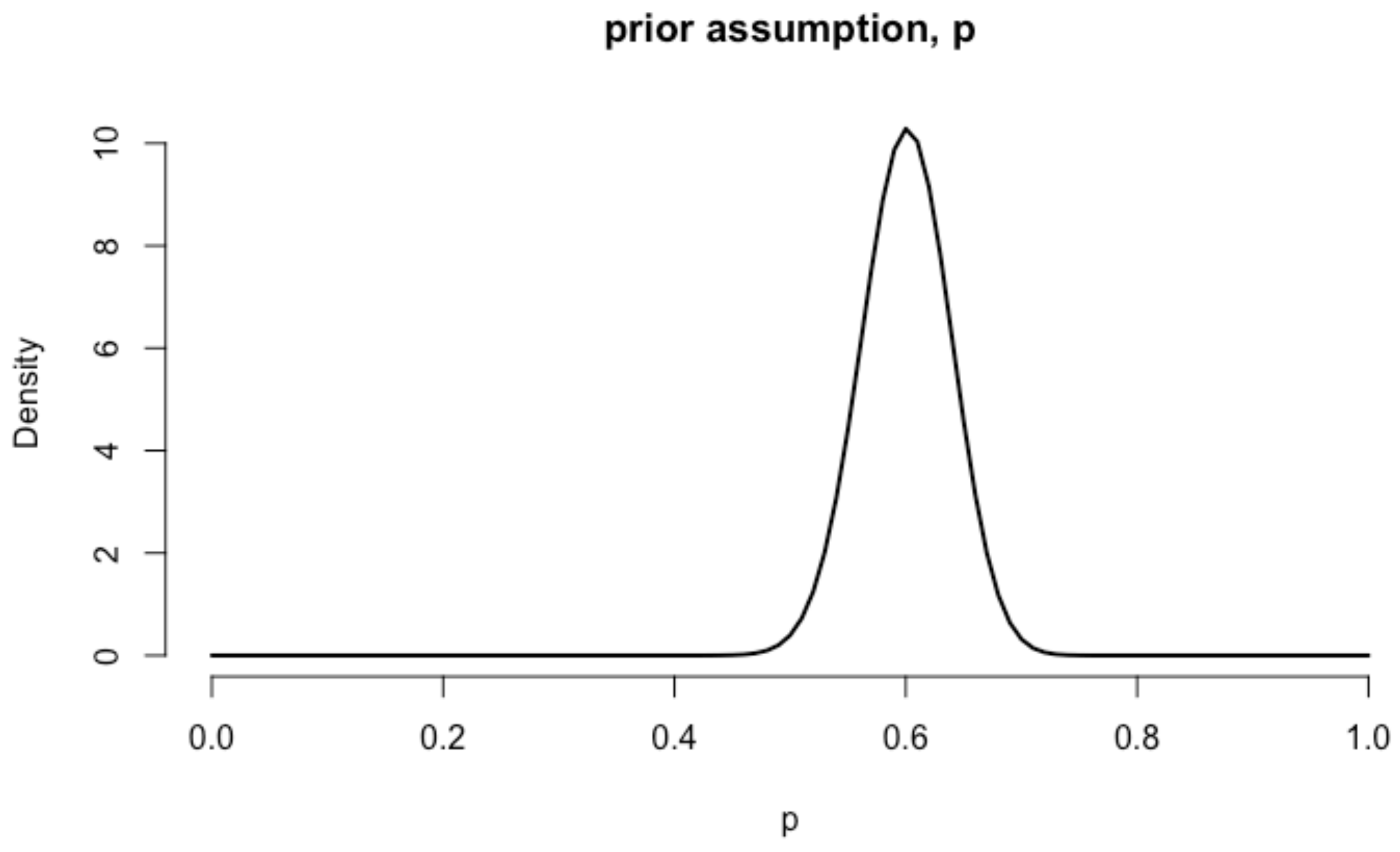
But **p ~ .8** is almost comfortable

What price the inference?

An opinion about p restricts the search space, yields a more precise opinion about N

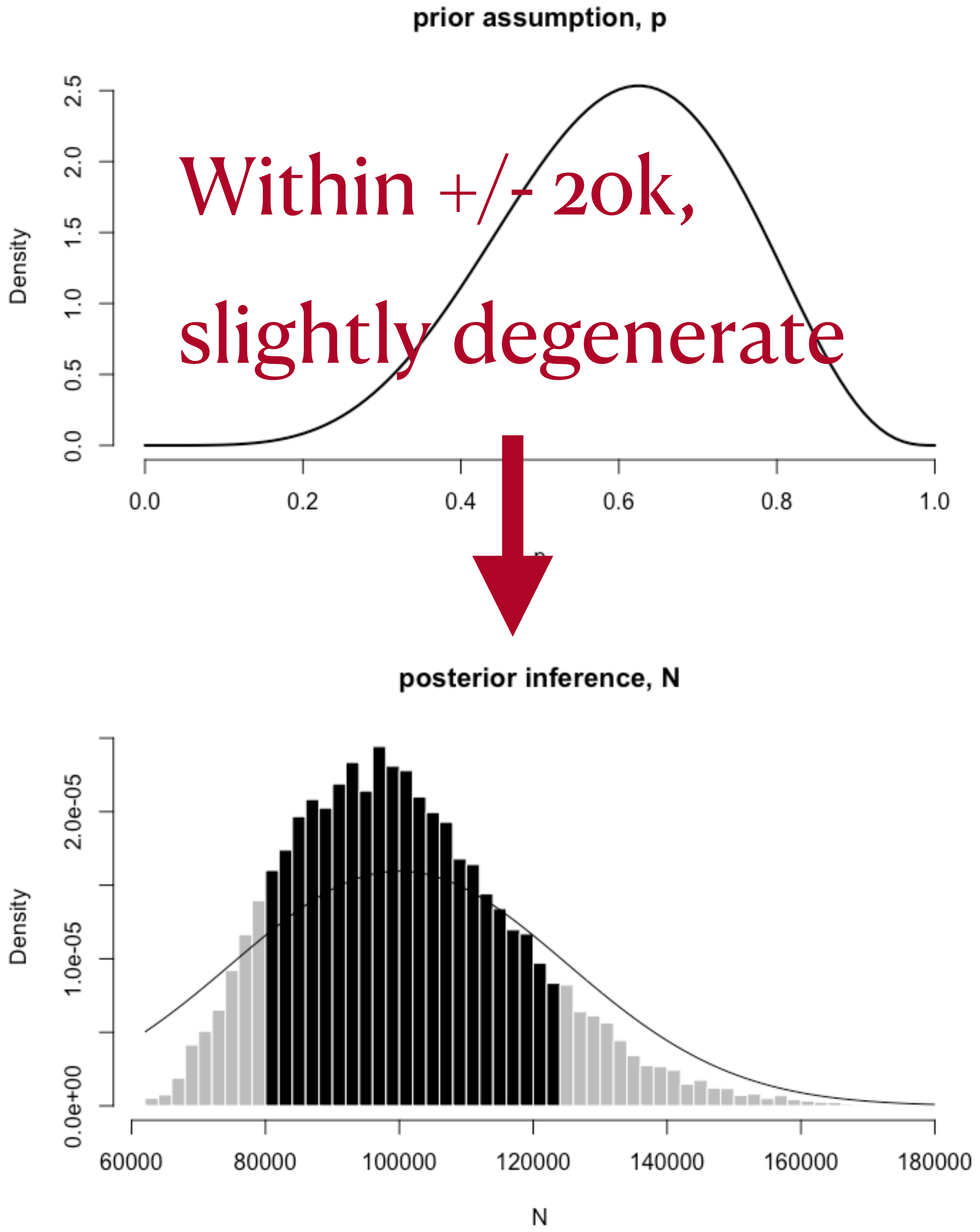

prior assumption, p


prior assumption, p


prior assumption, p

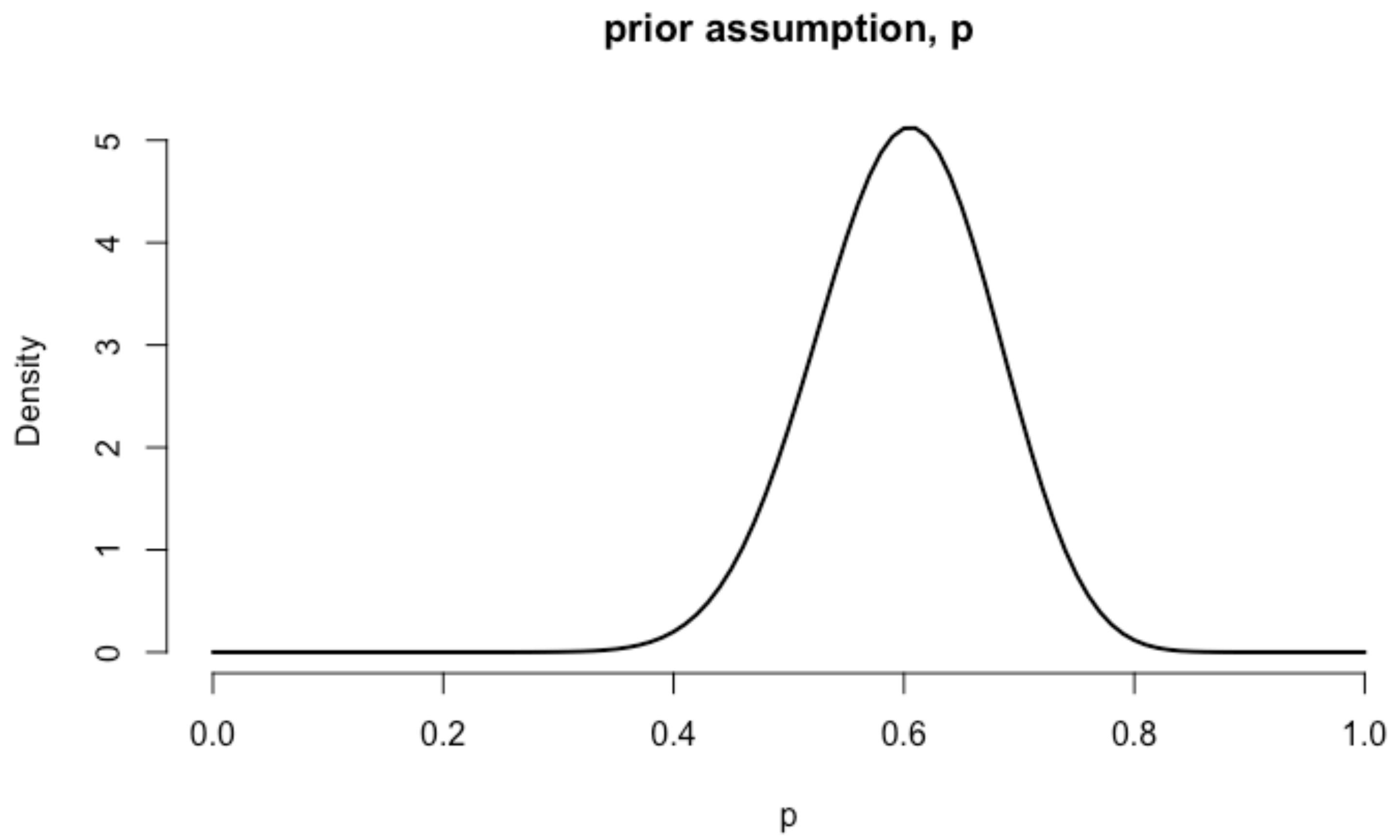
What price the inference?

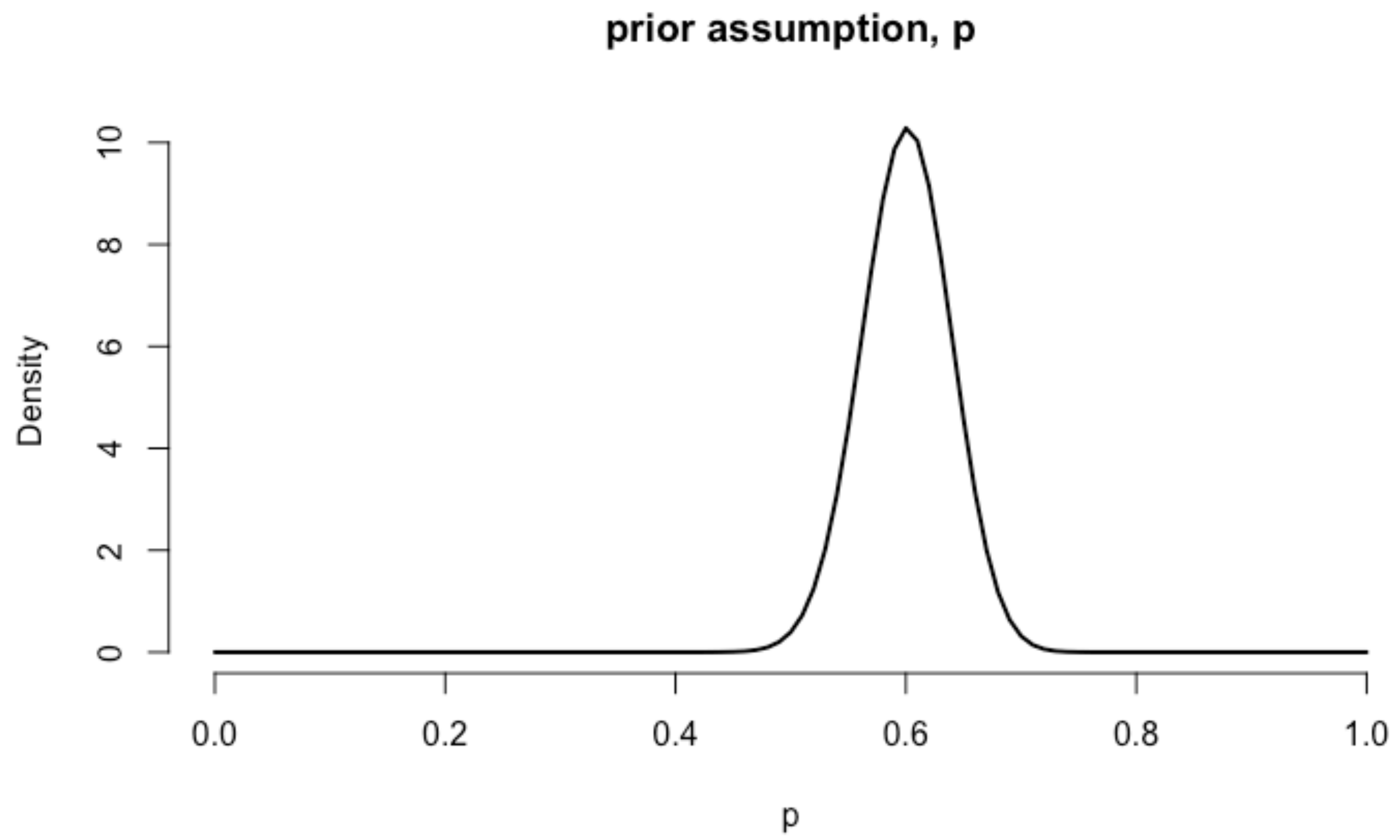
An opinion about p restricts the search space, yields a more precise opinion about N



Within +/- 20k,

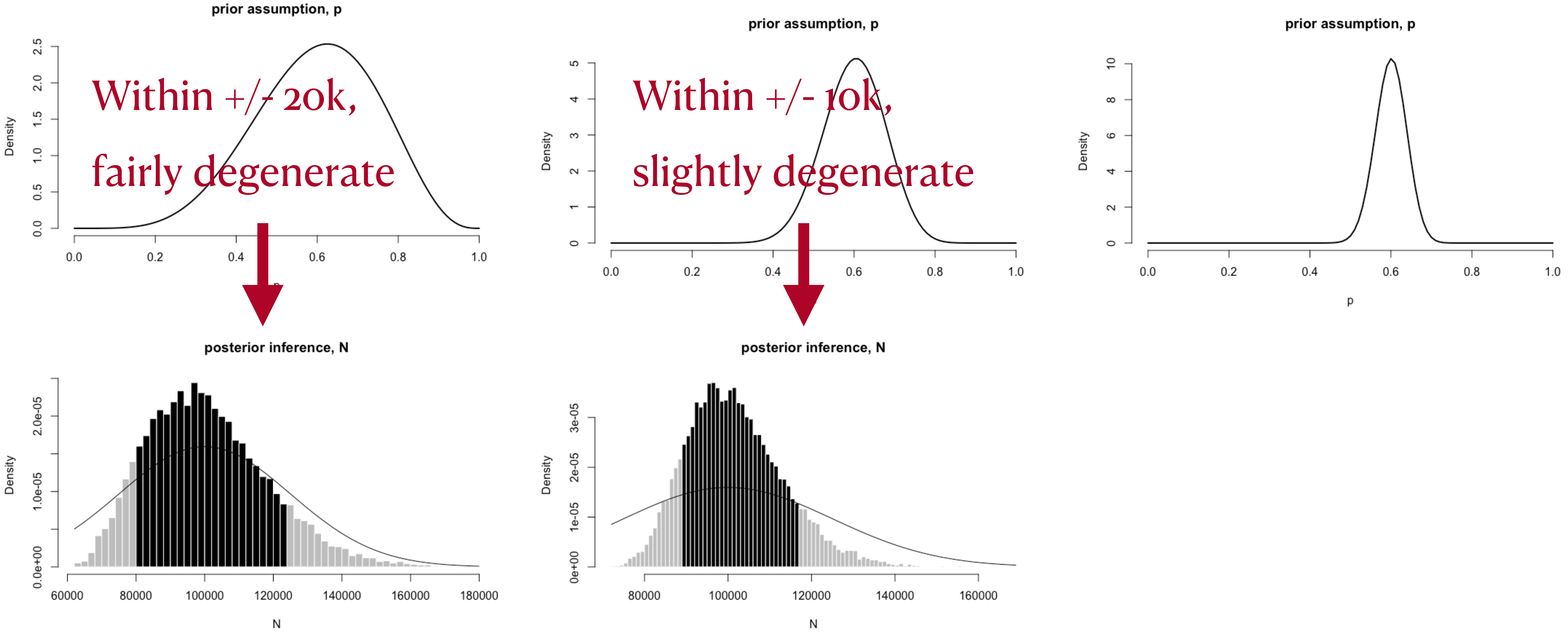
slightly degenerate





What price the inference?

An opinion about p restricts the search space, yields a more precise opinion about N


Within +/- 20k,

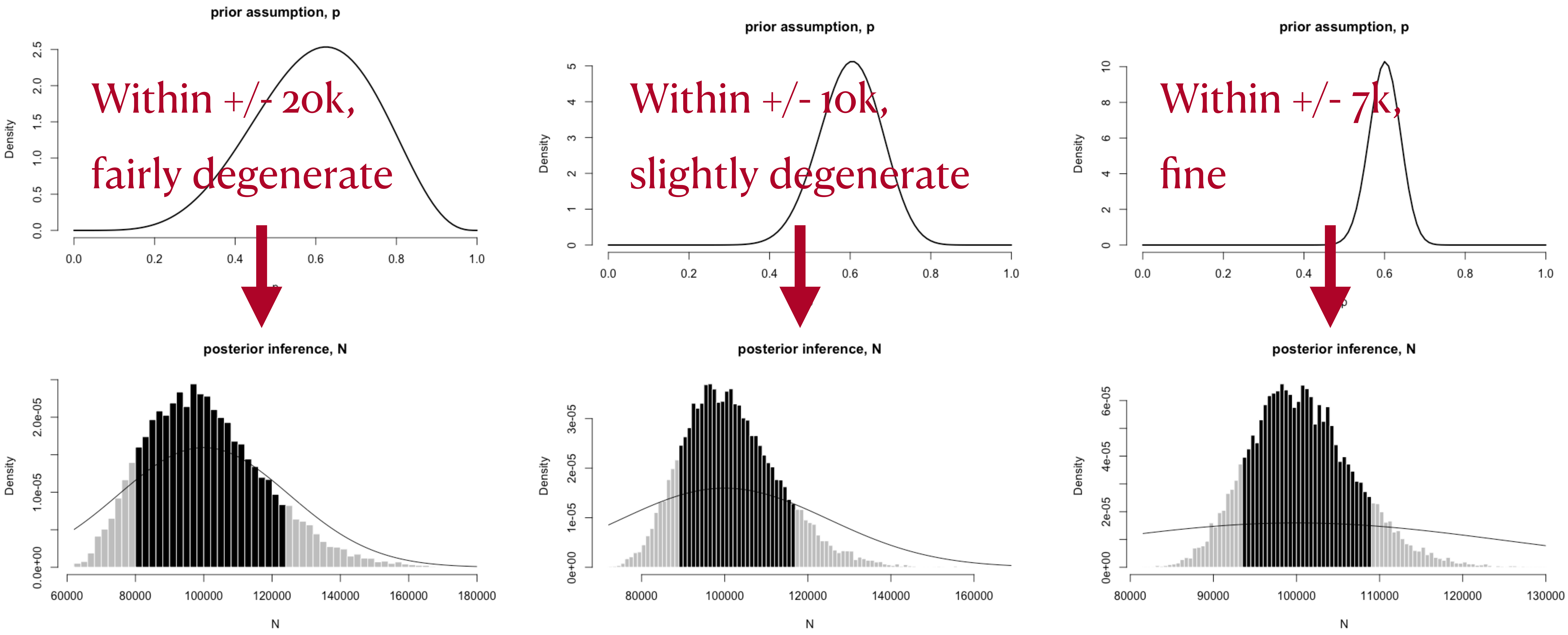
fairly degenerate

Within +/- 10k,

slightly degenerate

What price the inference?

An opinion about p restricts the search space, yields a more precise opinion about N



Within +/- 20k,

fairly degenerate

Within +/- 10k,

slightly degenerate

Within +/- 7k,

fine

The data provide very little. Mainly we're quantifying the consequences of our assumptions.

What if we can make more precise measurements?

Advanced demonology

- Same 100 bags, same 100 000 black marbles

What if we can make more precise measurements?

Advanced demonology

- Same 100 bags, same 100 000 black marbles
- Always make the vague assumption about **p**

What if we can make more precise measurements?

Advanced demonology

- Same 100 bags, same 100 000 black marbles
- Always make the vague assumption about **p**
- Separate measurements can recover ~90% of the black marbles (known)

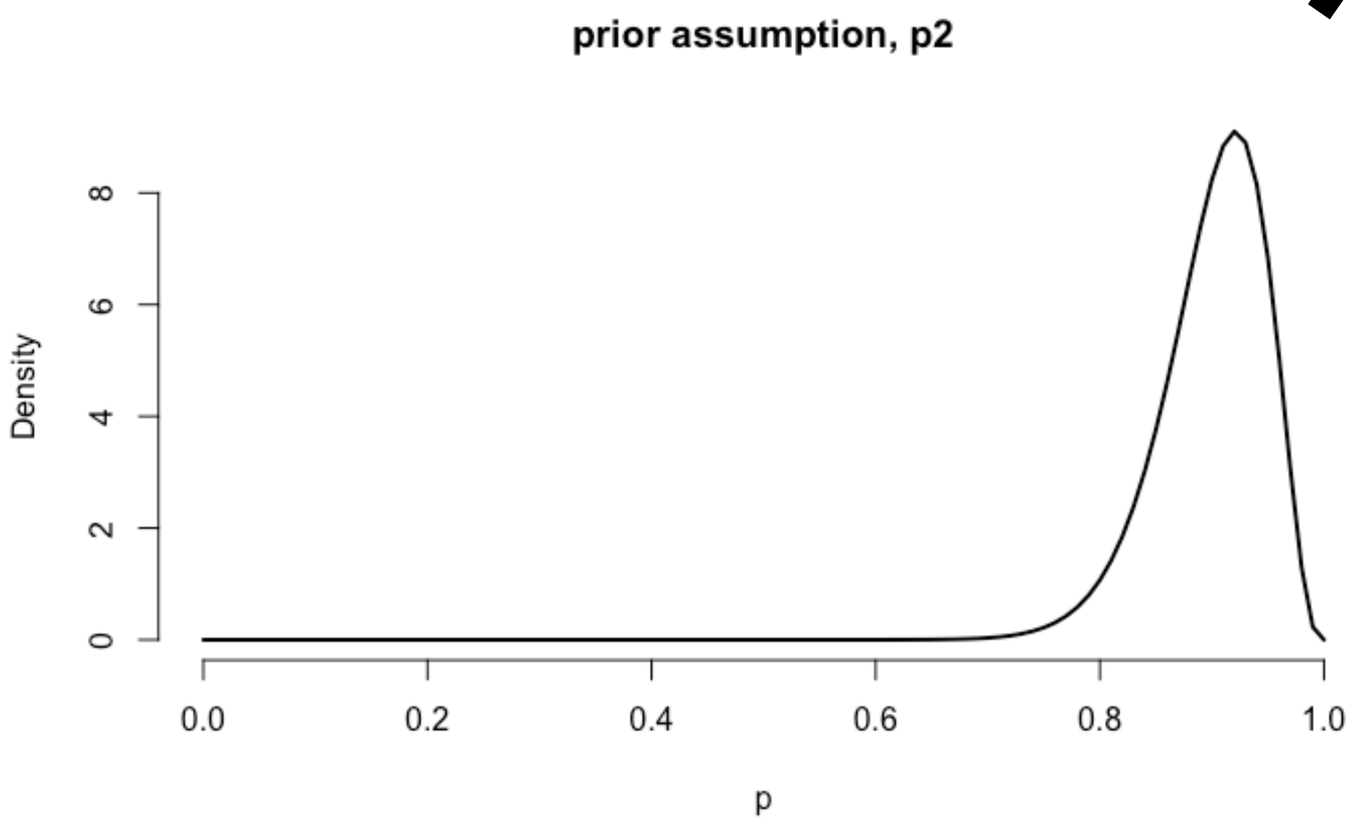
What if we can make more precise measurements?

Advanced demonology

- Same 100 bags, same 100 000 black marbles
- Always make the vague assumption about **p**
- Separate measurements can recover ~90% of the black marbles (known)
- These are more expensive and time-consuming

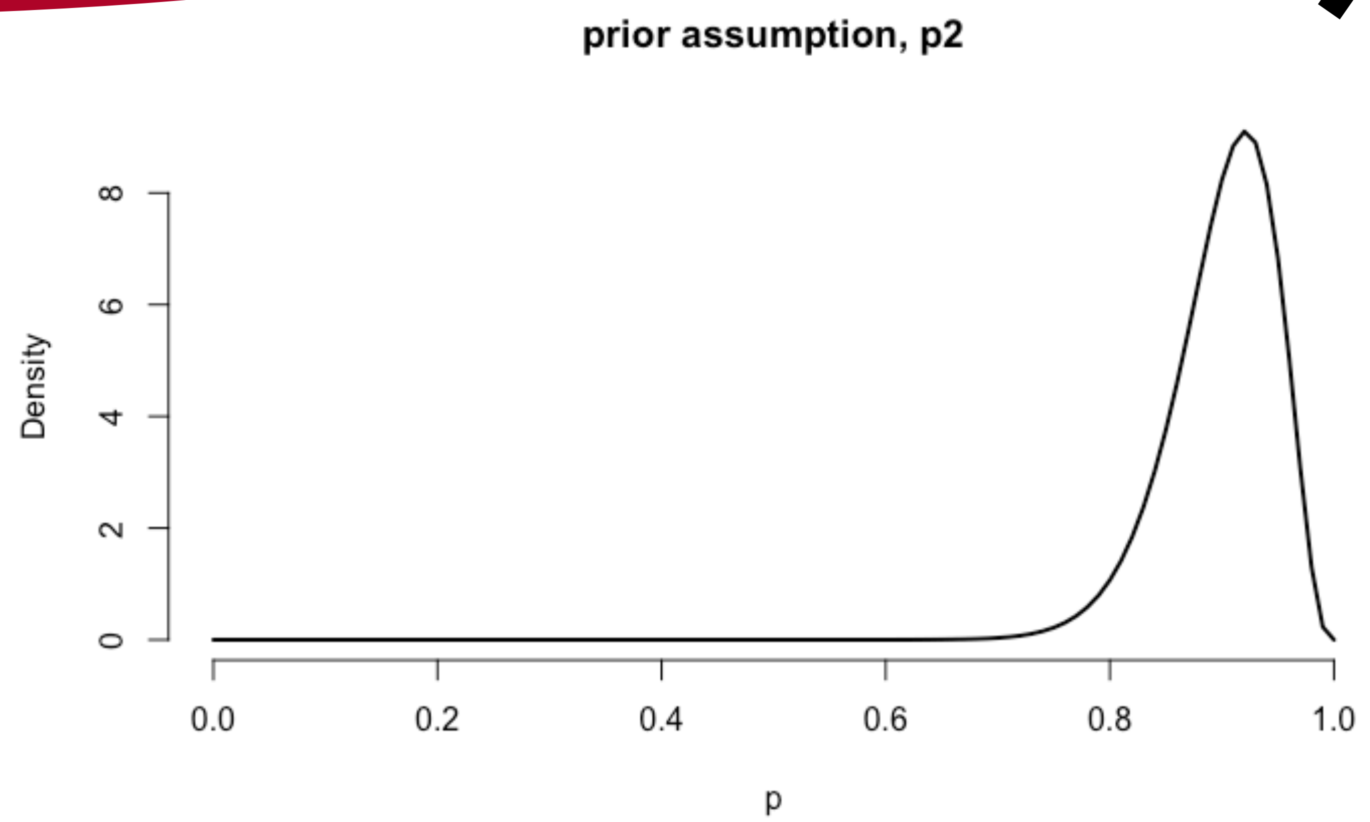
Advanced demonology

```
parameters {
  real<lower=0> N;
  real<lower=0, upper=1> p;
  real<lower=0> sigma;

  real<lower=0, upper=1> p2;
}

model {
  log(n) ~ normal(log(N) + log(p), sigma);


  log(N_obs) ~ normal(log(N) + log(p2), sigma);
  p2 ~ beta(36, 4);

  N ~ normal(mu_n, sd_n);
  p ~ beta(alpha, beta);

  sigma ~ exponential(.1);

}
```

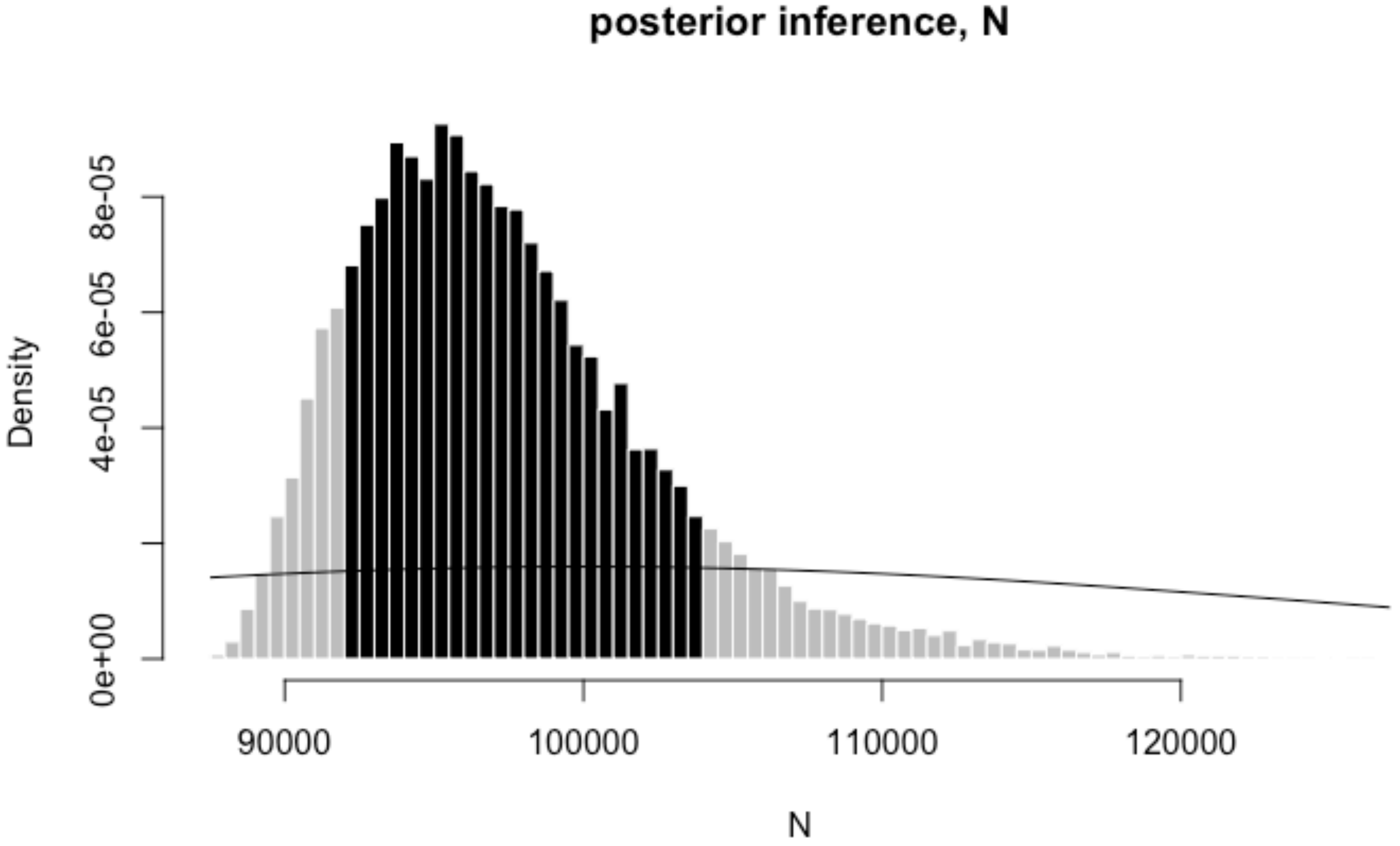


Advanced demonology

```
parameters {
  real<lower=0> N;
  real<lower=0, upper=1> p;
  real<lower=0> sigma;

  real<lower=0, upper=1> p2;
}

model {
  log(n) ~ normal(log(N) + log(p), sigma);


  log(N_obs) ~ normal(log(N) + log(p2), sigma);
  p2 ~ beta(36, 4);

  N ~ normal(mu_n, sd_n);
  p ~ beta(alpha, beta);

  sigma ~ exponential(.1);

}
```


prior assumption, p2

What price the inference?



A single extra measurement gets us within +/- 6k

(comparable with the strongest assumption earlier, no degeneracy)

Lessons

- **The Devil won’t be fooled:** no assumptions means you only get relative abundances

Lessons

- **The Devil won’t be fooled:** no assumptions means you only get relative abundances
- Need some strongish assumptions

Lessons

- **The Devil won’t be fooled:** no assumptions means you only get relative abundances
- Need some strongish assumptions
- If the base rate of recovering a given black marble (Kp transcript) is really low only your immortal soul will do

Lessons

- **The Devil won’t be fooled:** no assumptions means you only get relative abundances
- Need some strongish assumptions
- If the base rate of recovering a given black marble (Kp transcript) is really low only your immortal soul will do
- A handful of more precise measurements (qPCR? I don’t know) helps a lot

Future directions

- Really interested in bags that have different number of black marbles, **N**

Future directions

- Really interested in bags that have different number of black marbles, **N**
- Interested in seeing what analyses using inferred **N** look like (Lotka–Volterra?)

Future directions

- Really interested in bags that have different number of black marbles, **N**
- Interested in seeing what analyses using inferred **N** look like (Lotka–Volterra?)
- Models could do with refinement and debugging

Thank you.