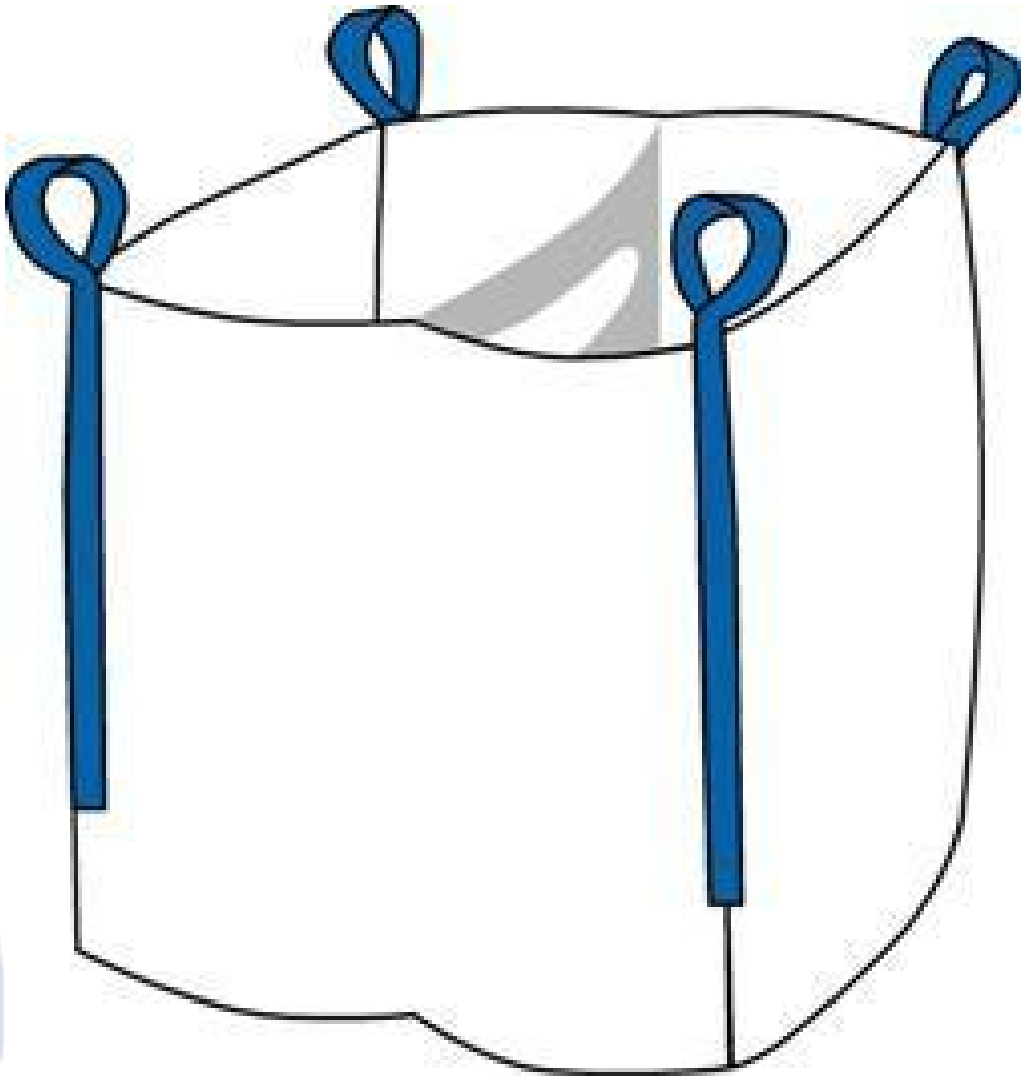
Estimating Uncertainty of Categorical Web Data

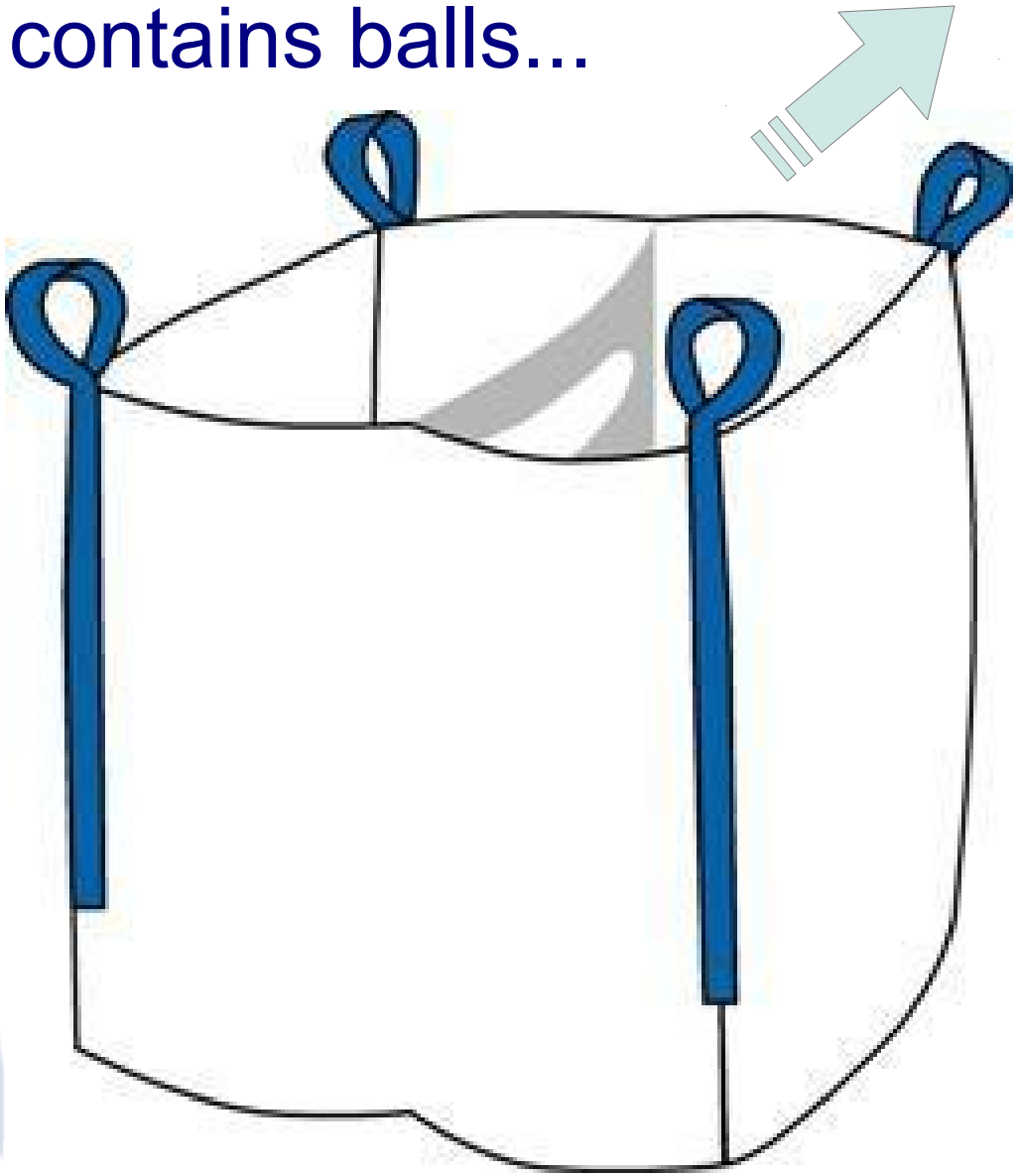
Davide Ceolin, Willem Robert van Hage,
Wan Fokkink, Guus Schreiber

VU University Amsterdam

Suppose we know that this bag contains balls...

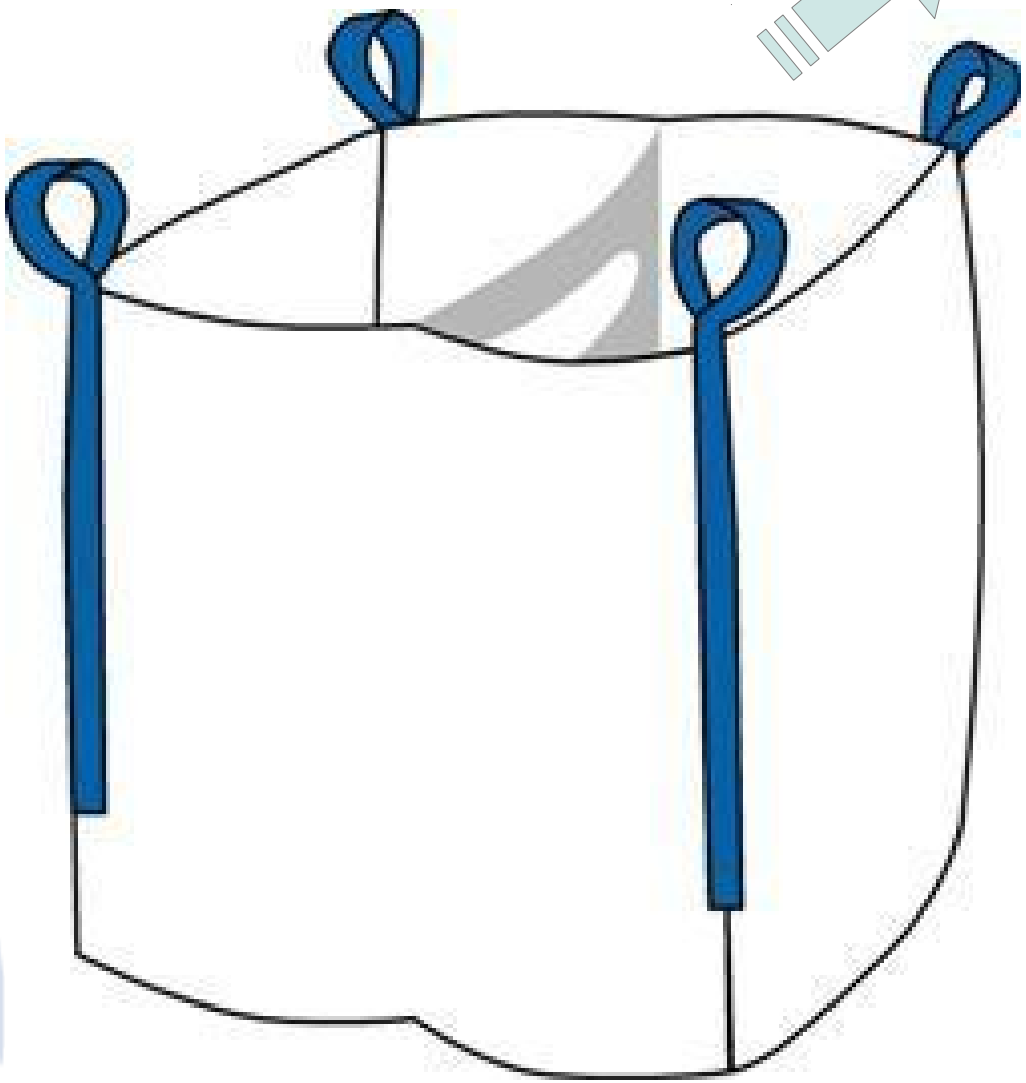


Suppose we know that this bag contains balls...



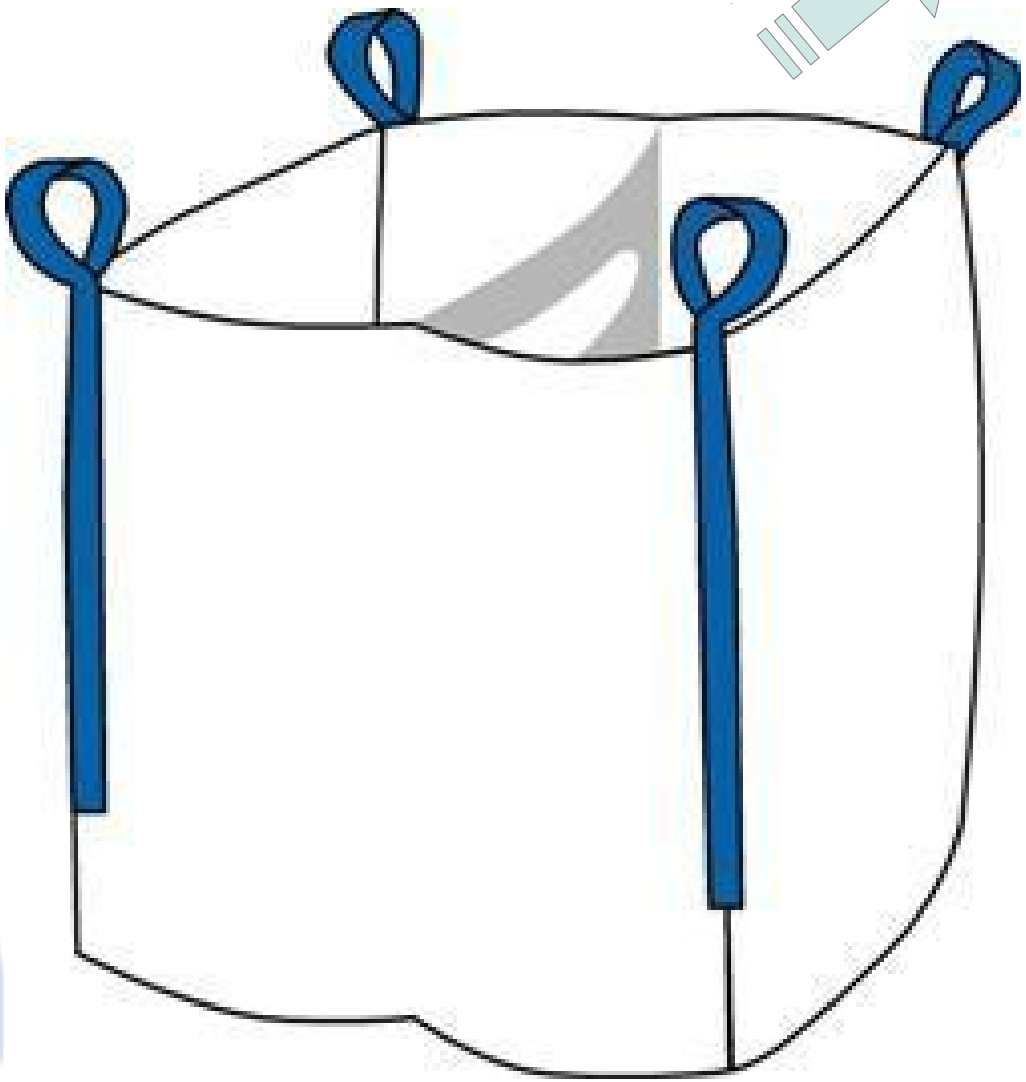
We draw some of them...

Suppose we know that this bag contains balls...



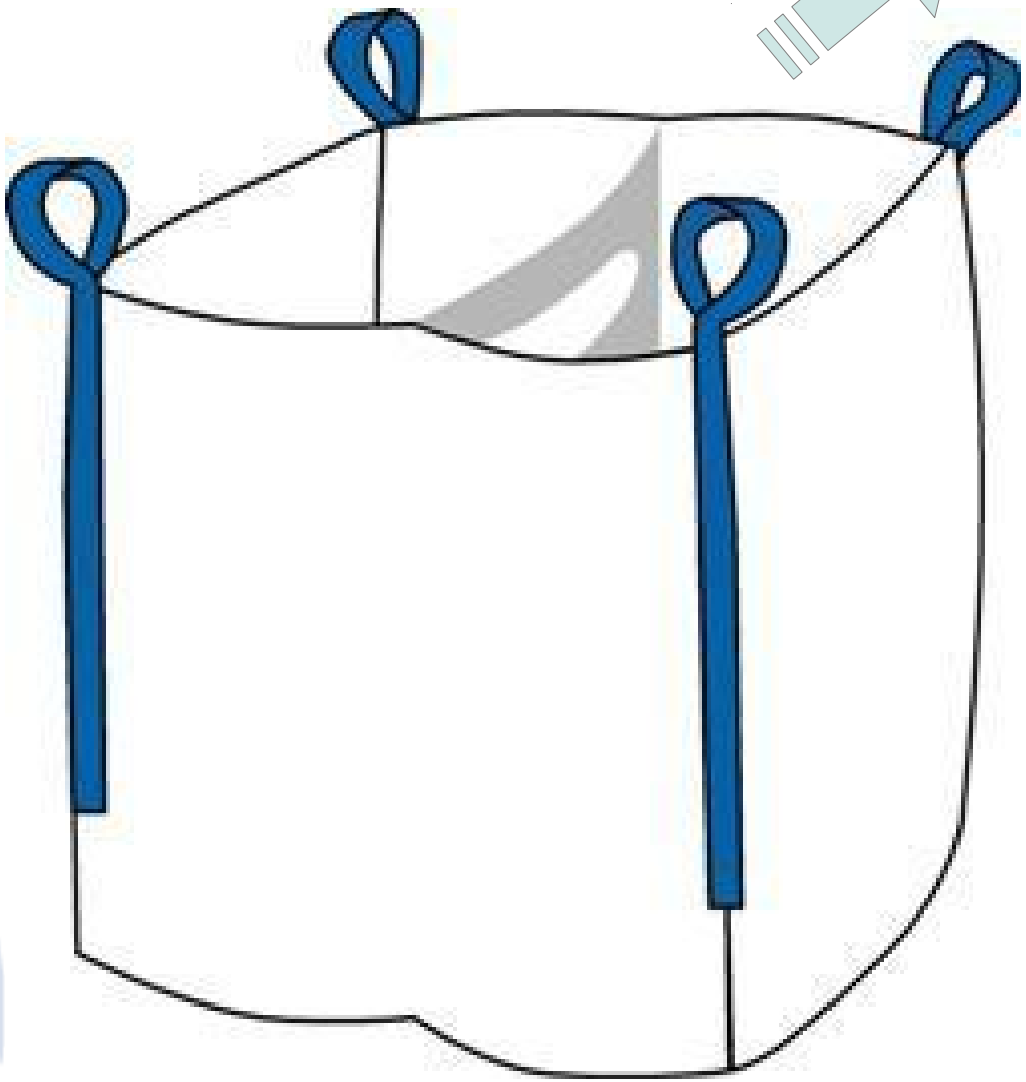
We draw some of them...

Suppose we know that this bag contains balls...



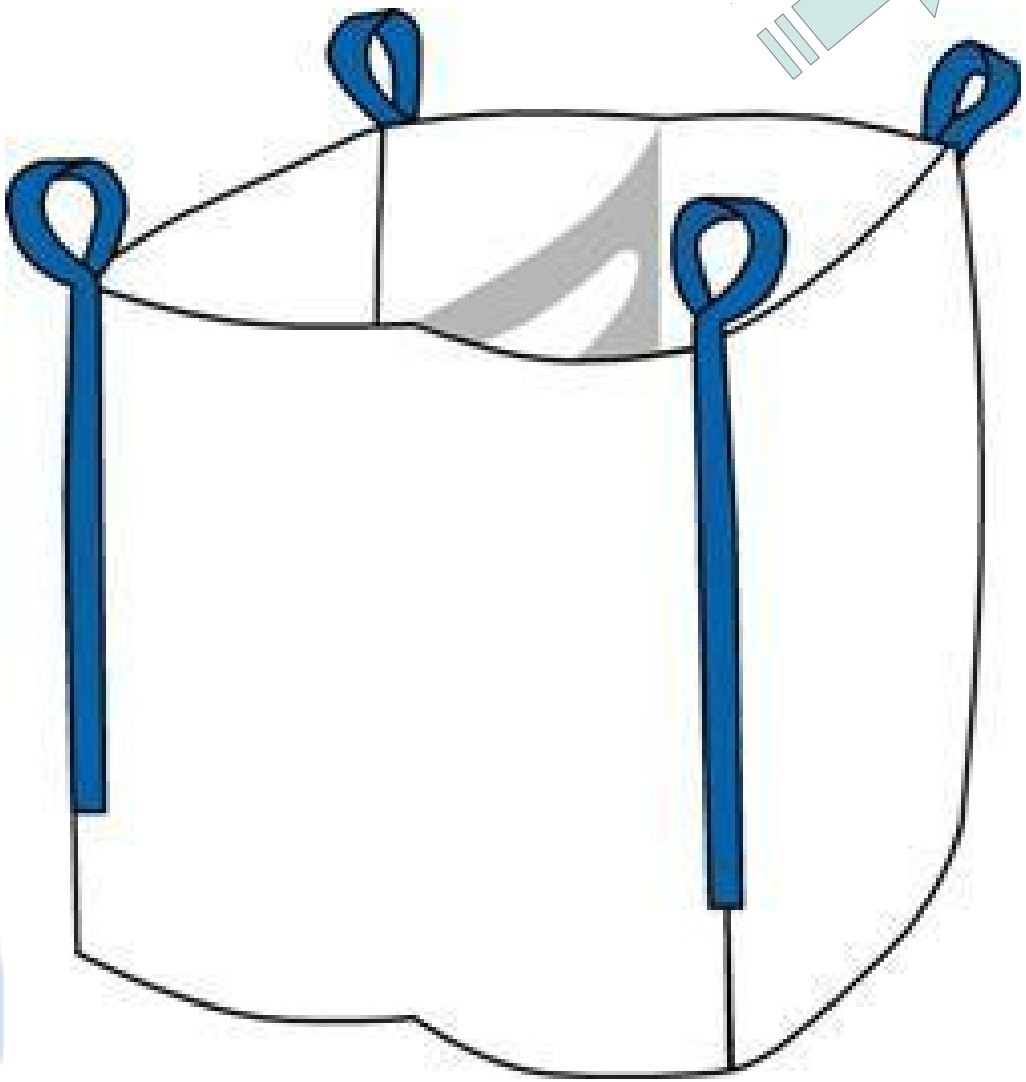
We draw some of them...

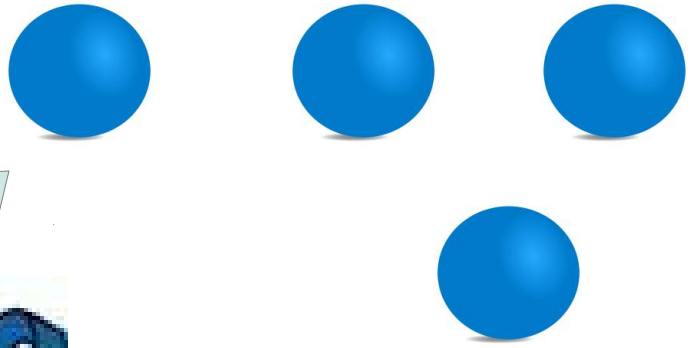
Suppose we know that this bag contains balls...



We draw some of them...

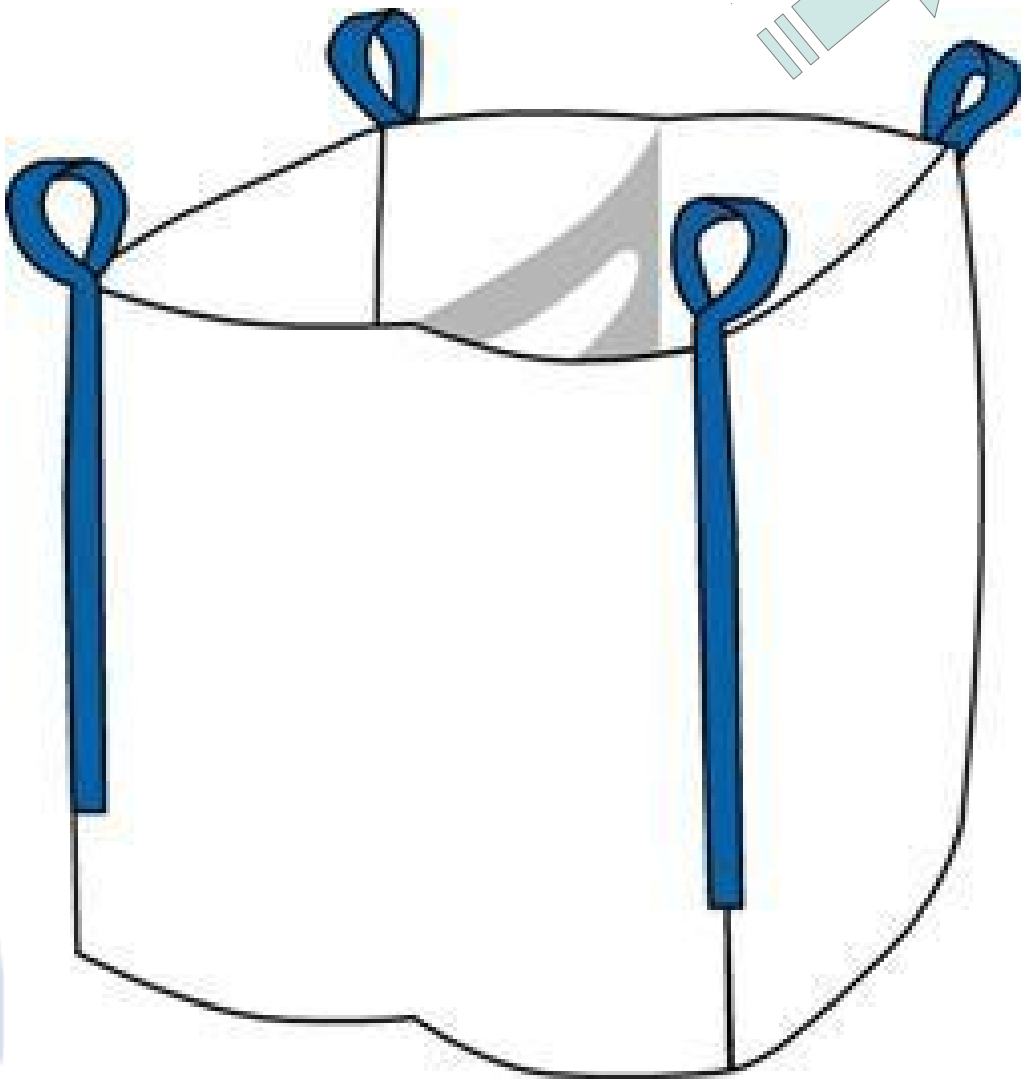
Suppose we know that this bag contains balls...

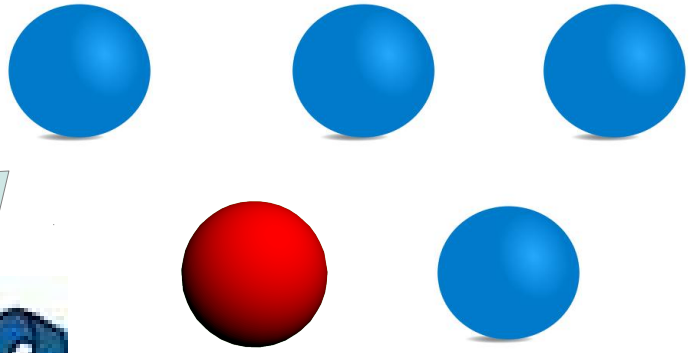




We draw some of them...

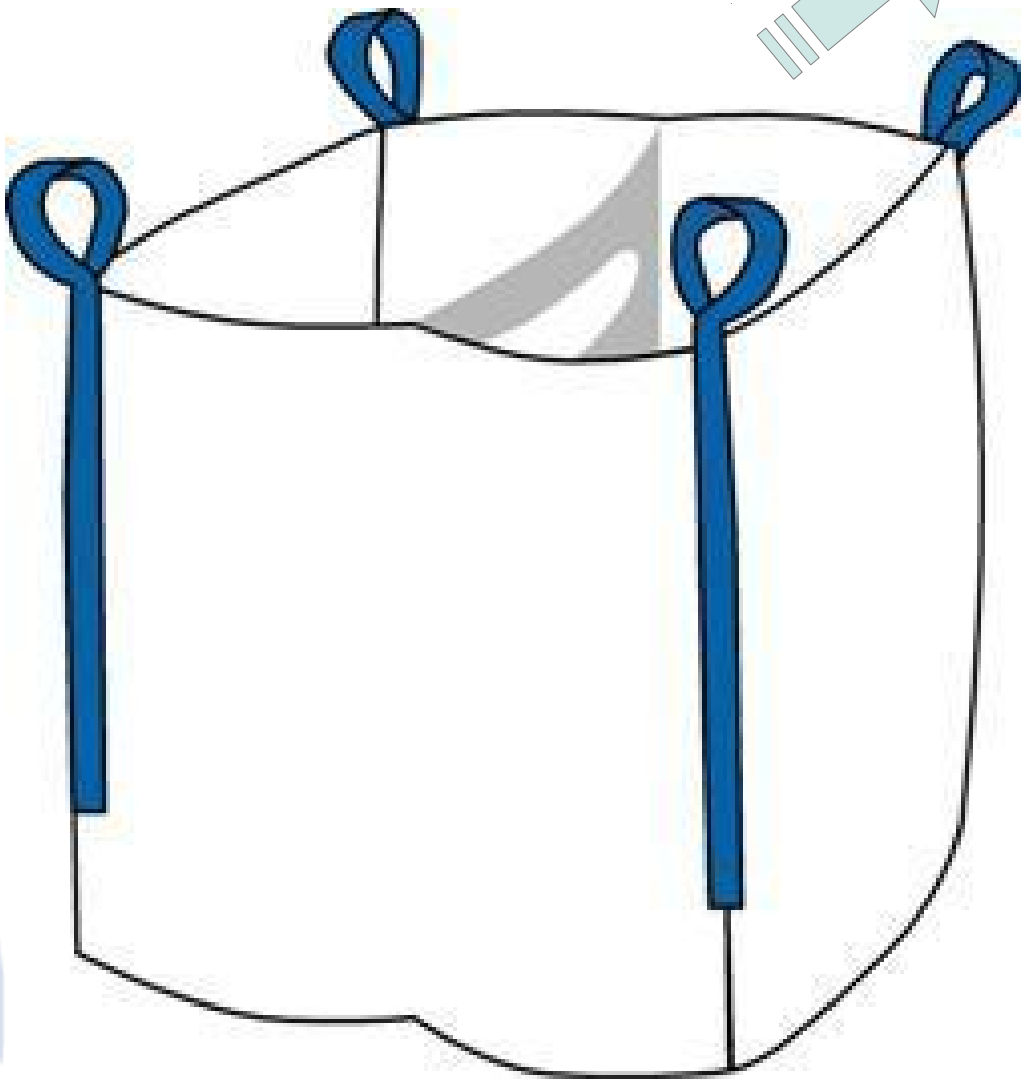
Suppose we know that this bag contains balls...

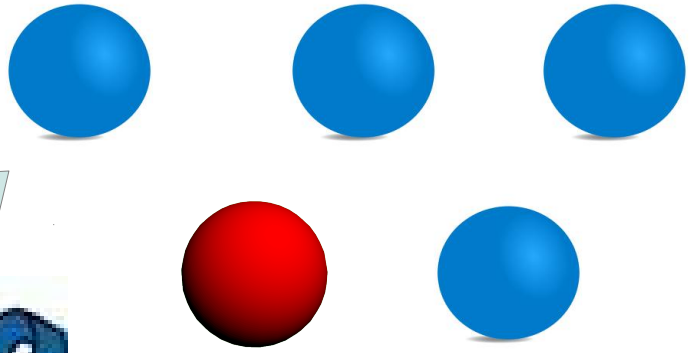




We draw some of them...

Suppose we know that this bag contains balls...

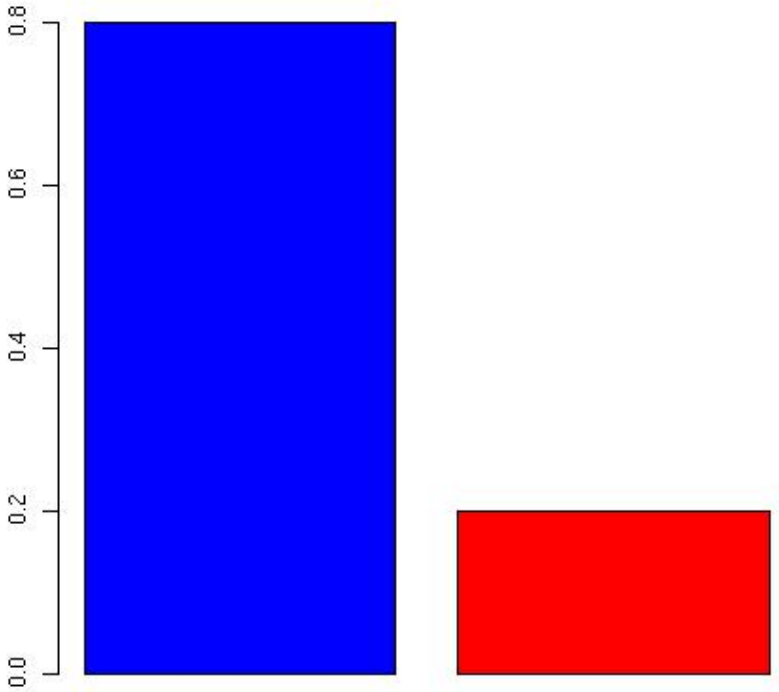




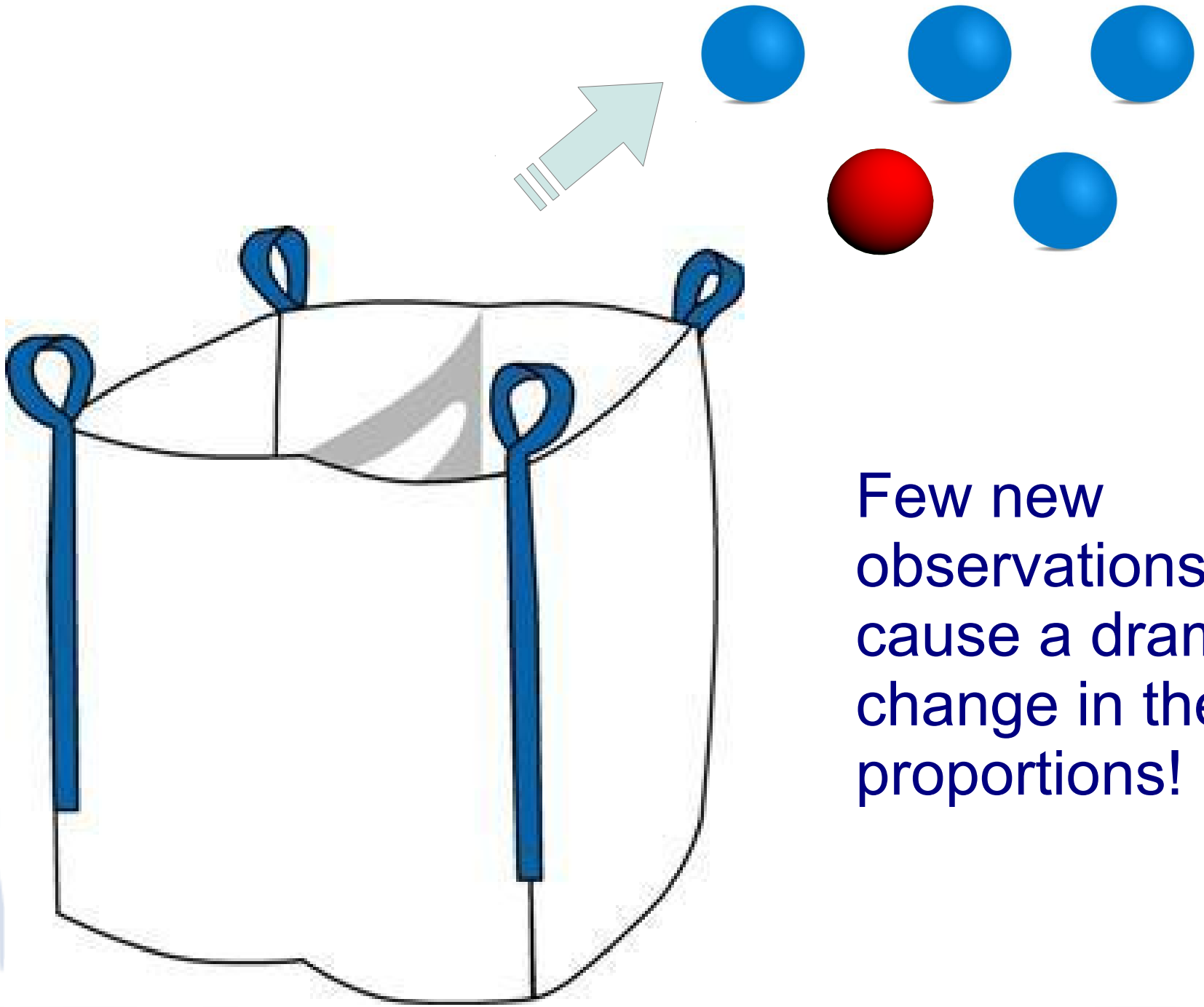
We draw some of them...

What can say about the bag content?

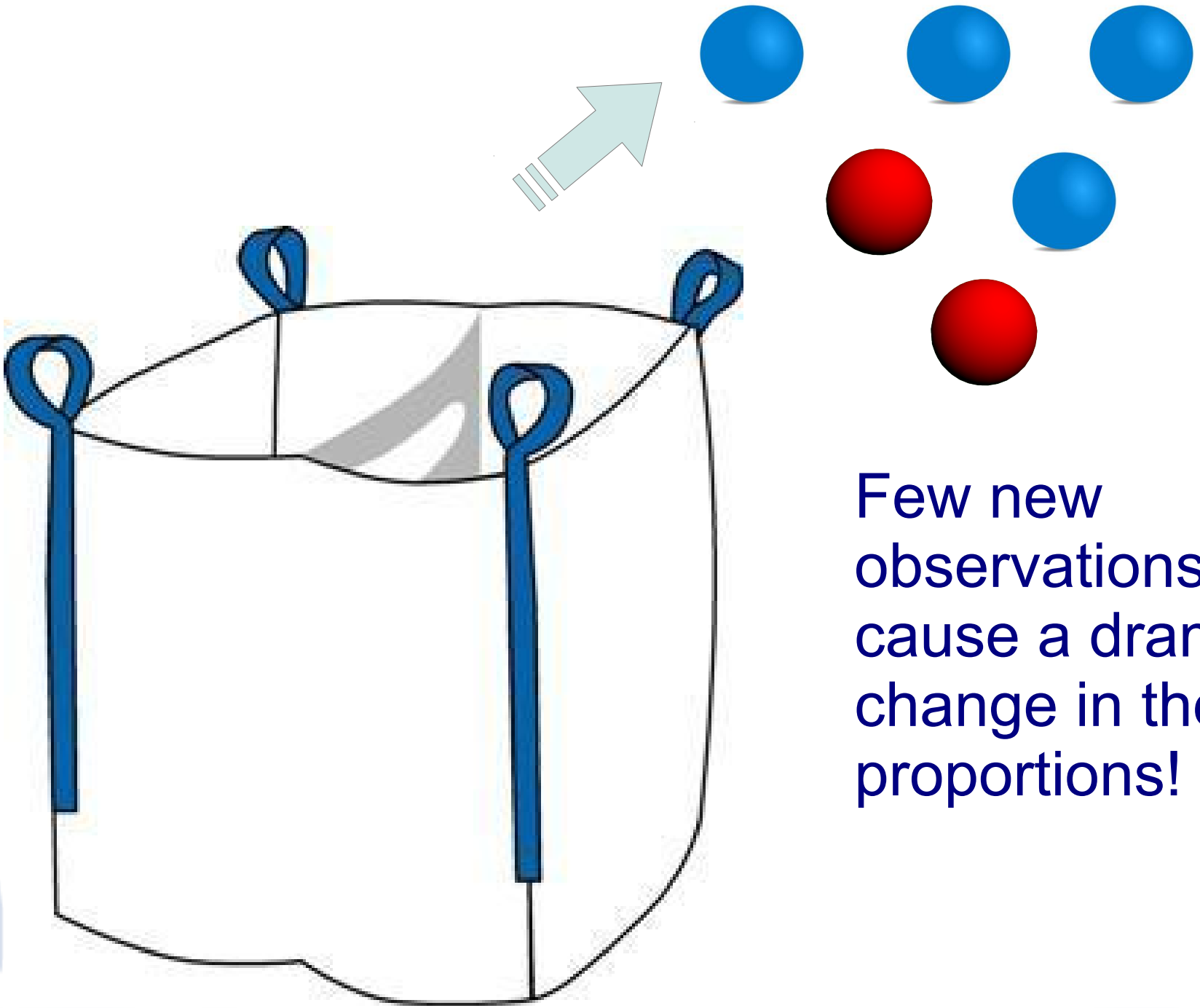
Bag content



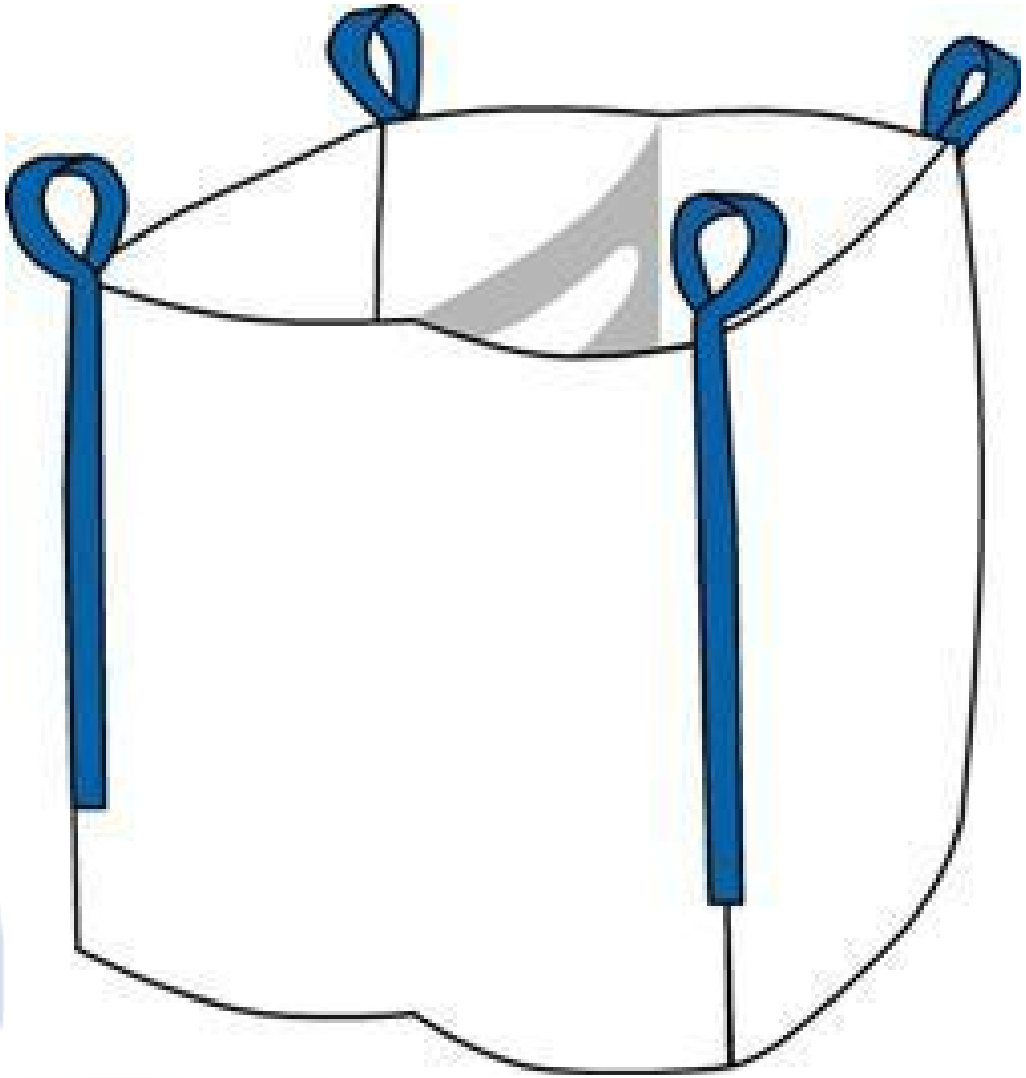
- A binomial distribution can represent the sample
- But, does it represent also the entire bag content?

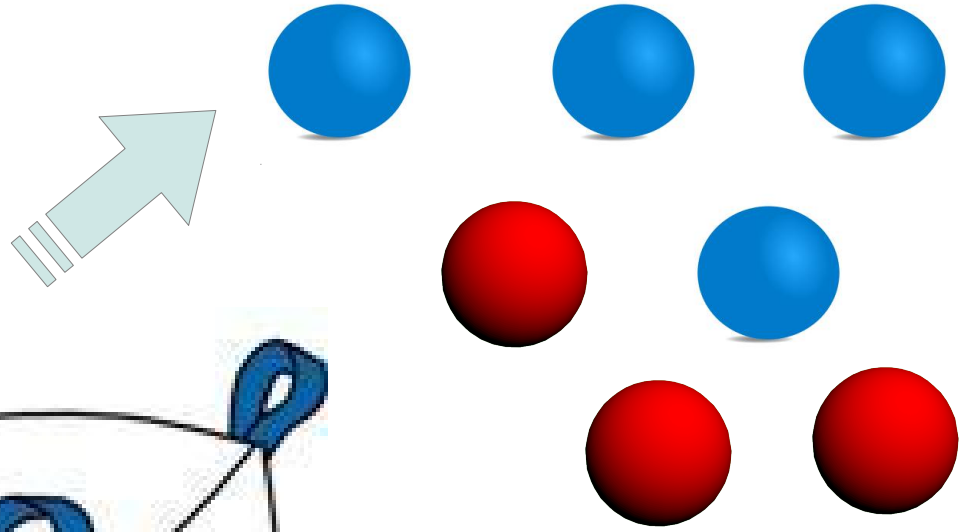


Few new observations can cause a dramatic change in the proportions!

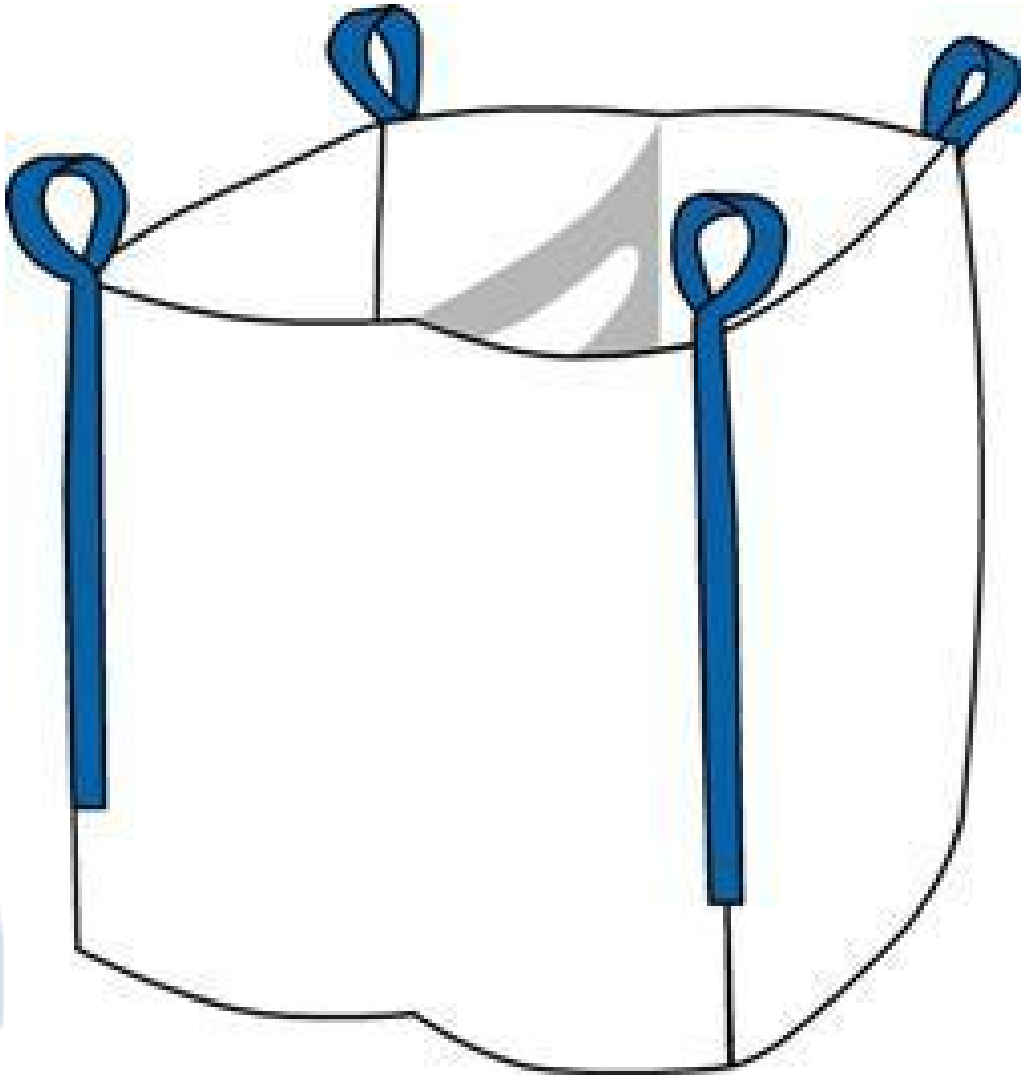


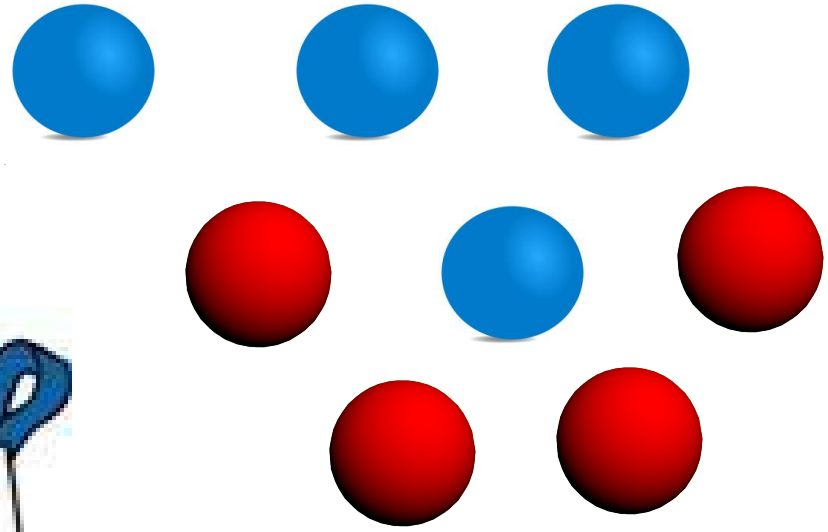
Few new observations can cause a dramatic change in the proportions!



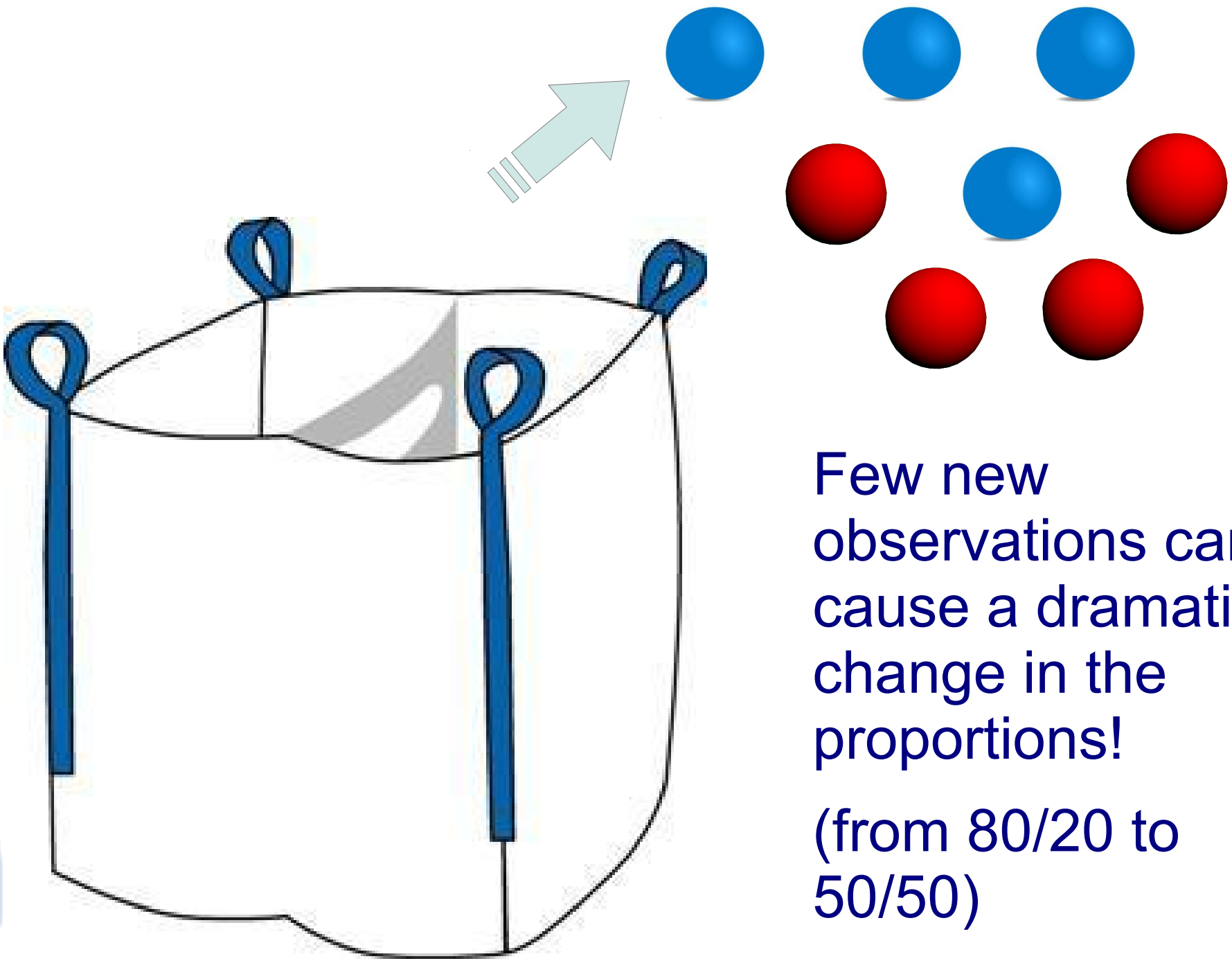


Few new observations can cause a dramatic change in the proportions!





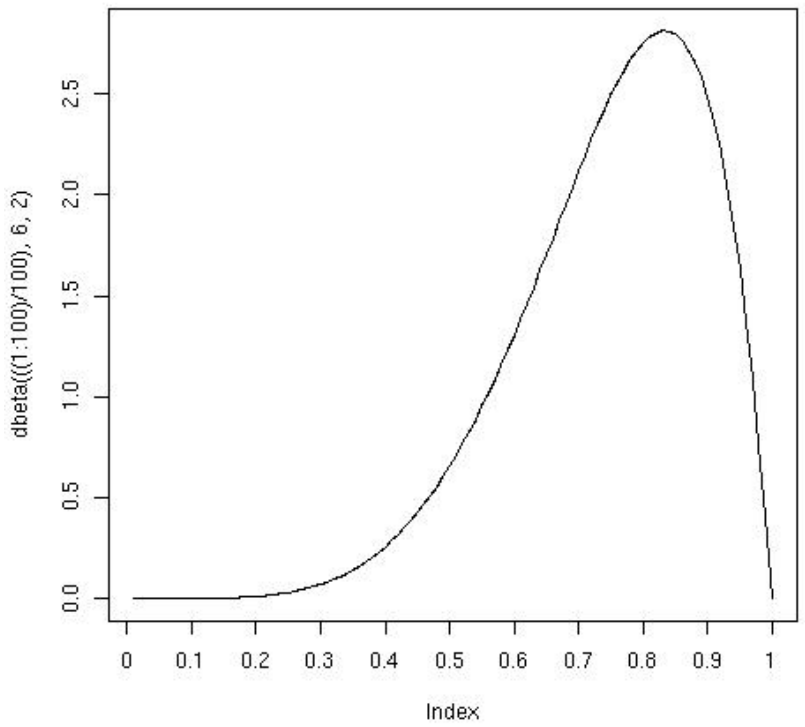
Few new observations can cause a dramatic change in the proportions!



Few new observations can cause a dramatic change in the proportions!

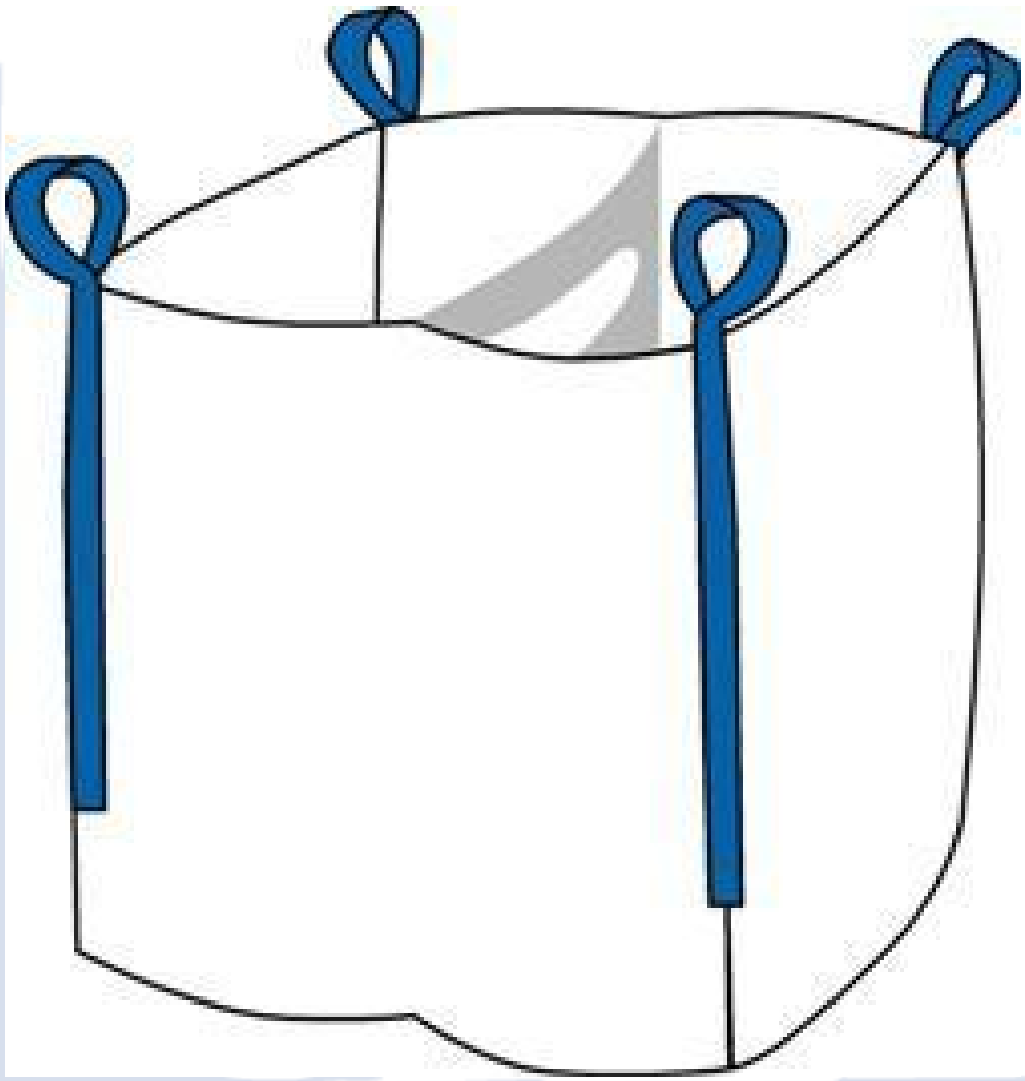
(from 80/20 to 50/50)

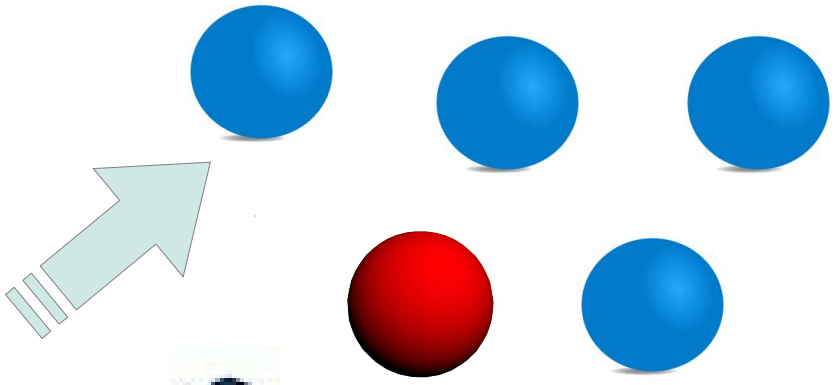
Estimating the second order probability



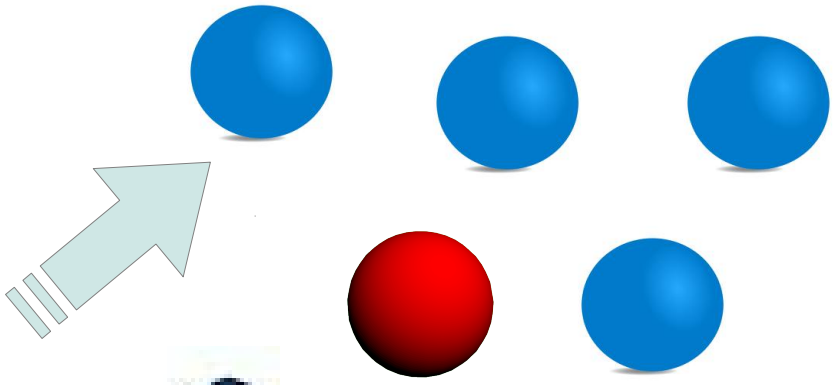
- We should estimate the uncertainty about the ratio *p*.
- The Beta is the best candidate to describe *p* (because it is conjugated to the multinomial).

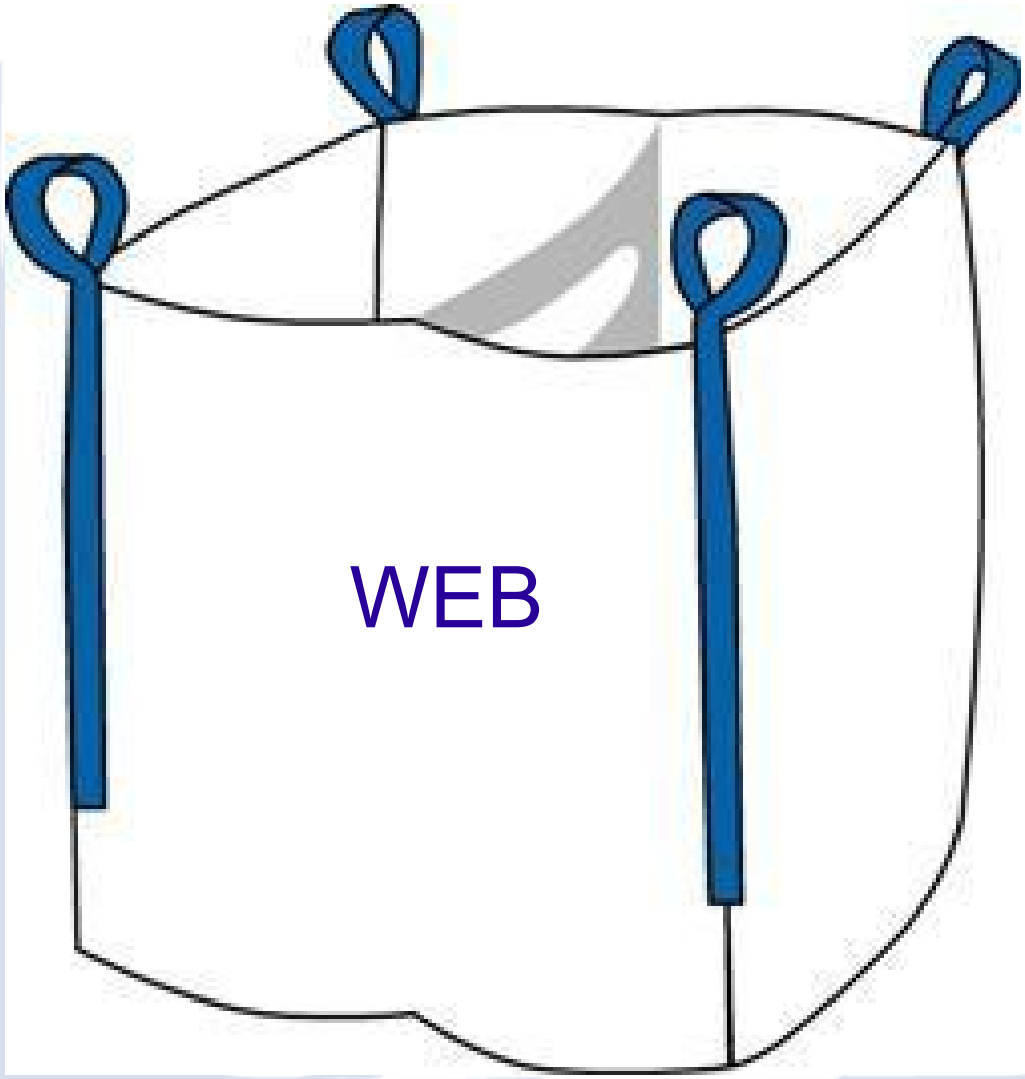
Of course, this was a metaphore...



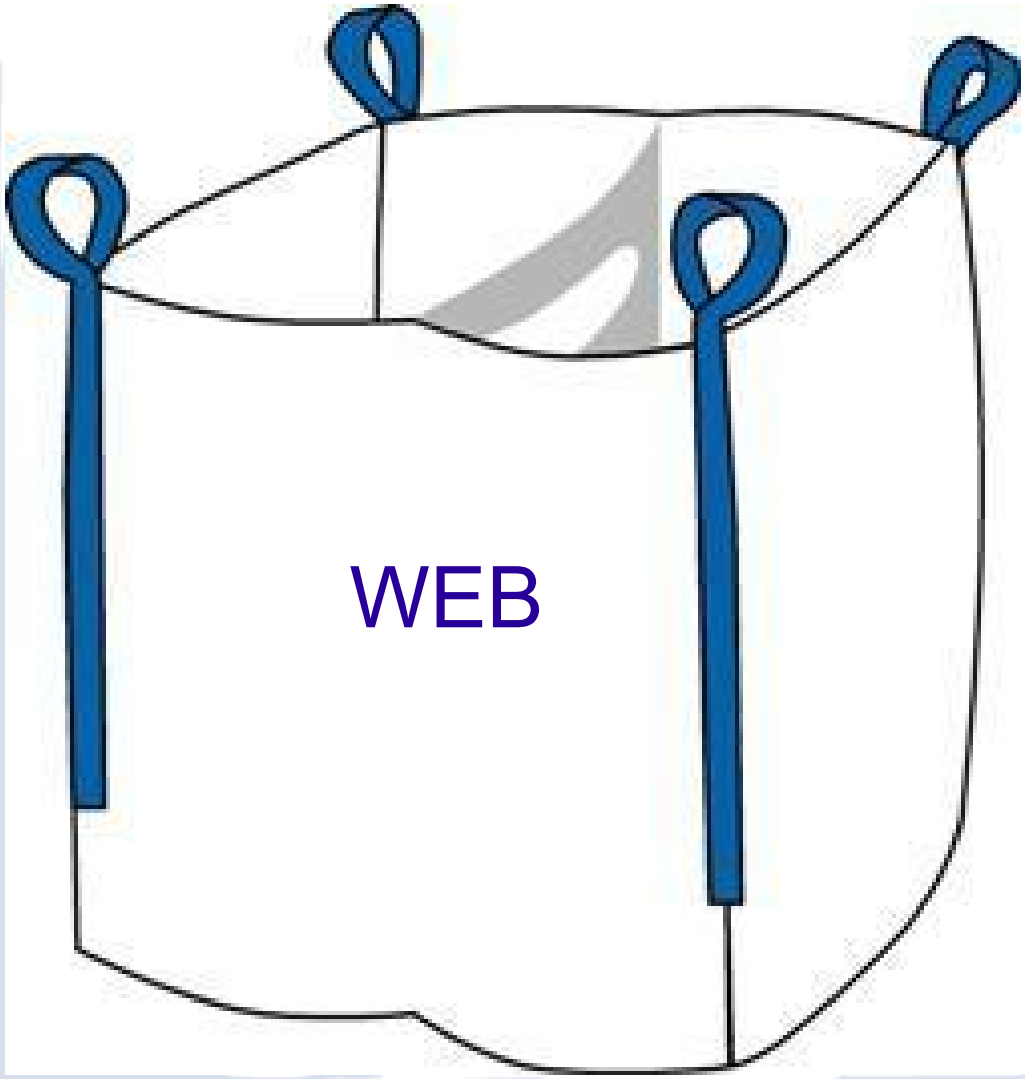


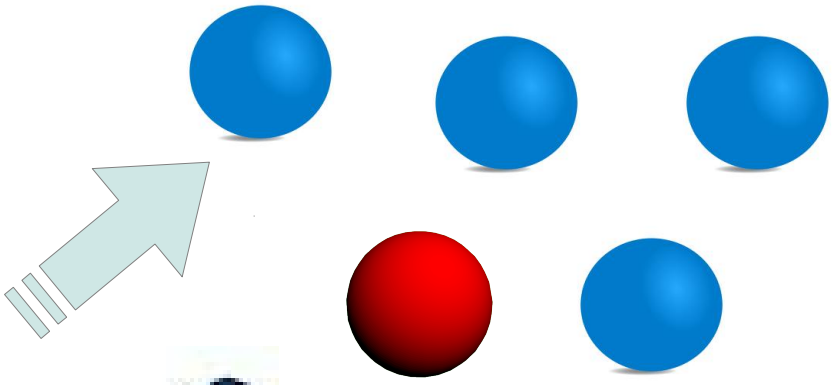
Of course, this was a metaphore...





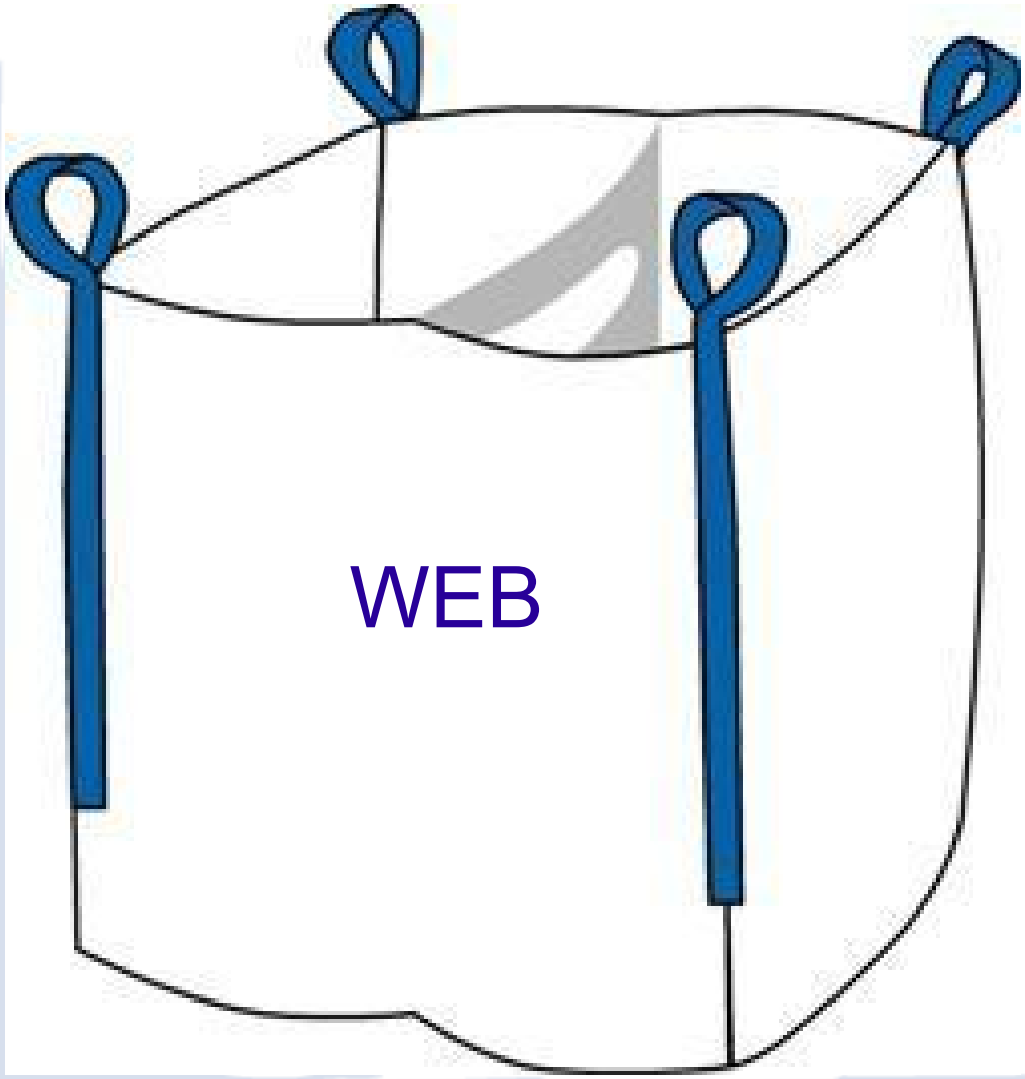
Of course, this was a metaphore...

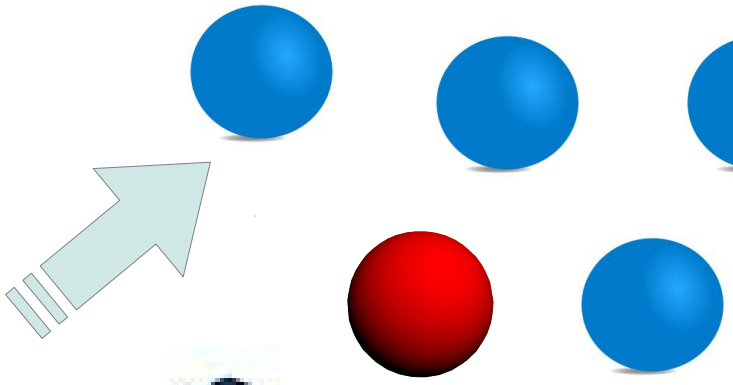




Classes of URIs / Web pages / Links / ...

Of course, this was a metaphore...





Classes of URIs / Web pages / Links / ...

Does this change something?

Deal with Web Samples

The Web makes the situation more complicated:

- Samples can be biased;
- The Web evolves over time;
- Different domains imply distinct subpopulations;
- Data are accessed incrementally, by crawling.

Deploying second order probabilities

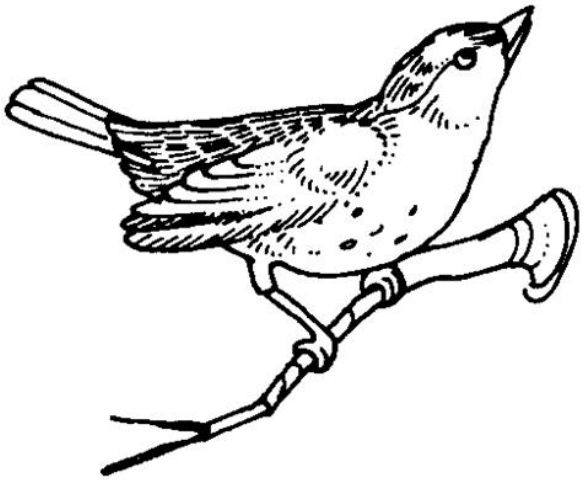
Why? Possible bias, time variability, sub-populations increase uncertainty.

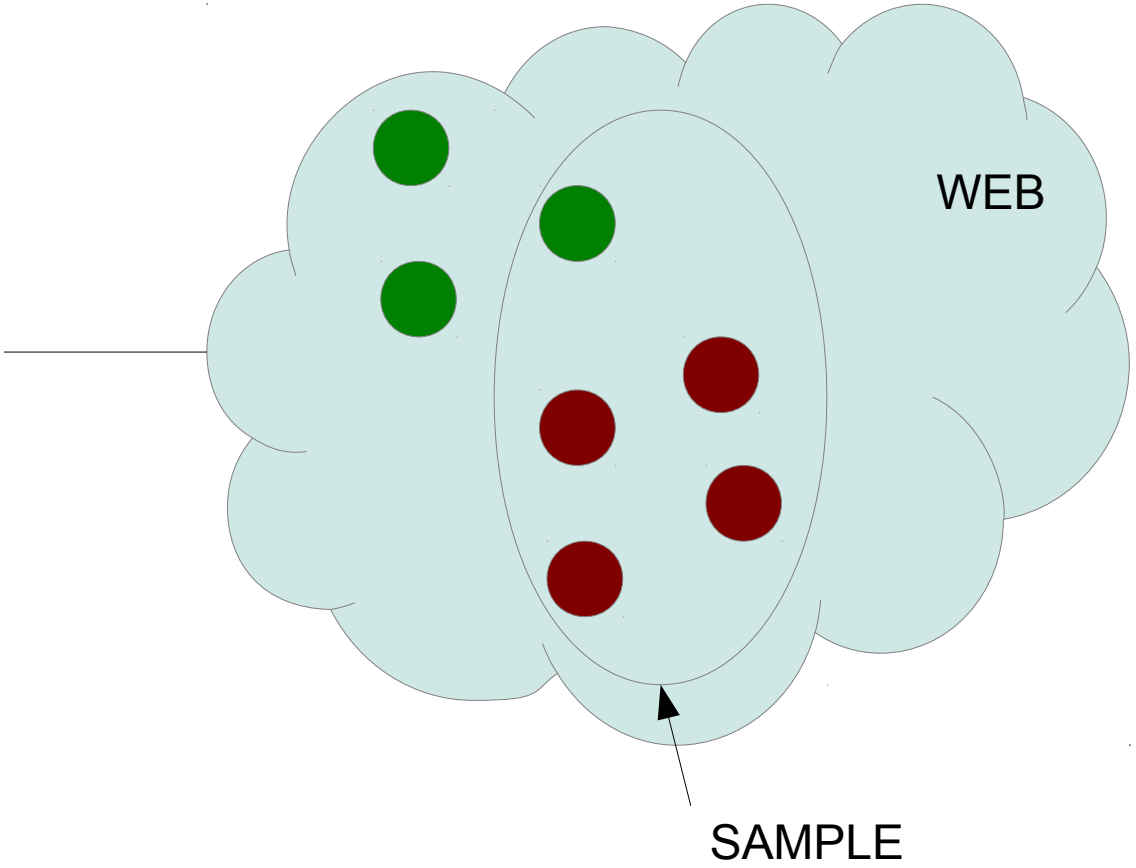
Rationale: Instead of trying to estimate the correct proportion among categories, we compute a set of candidate values.

Over time: More evidence makes the set smaller.

Soundness: Conjugacy guarantees correct choice and update of probability distributions.

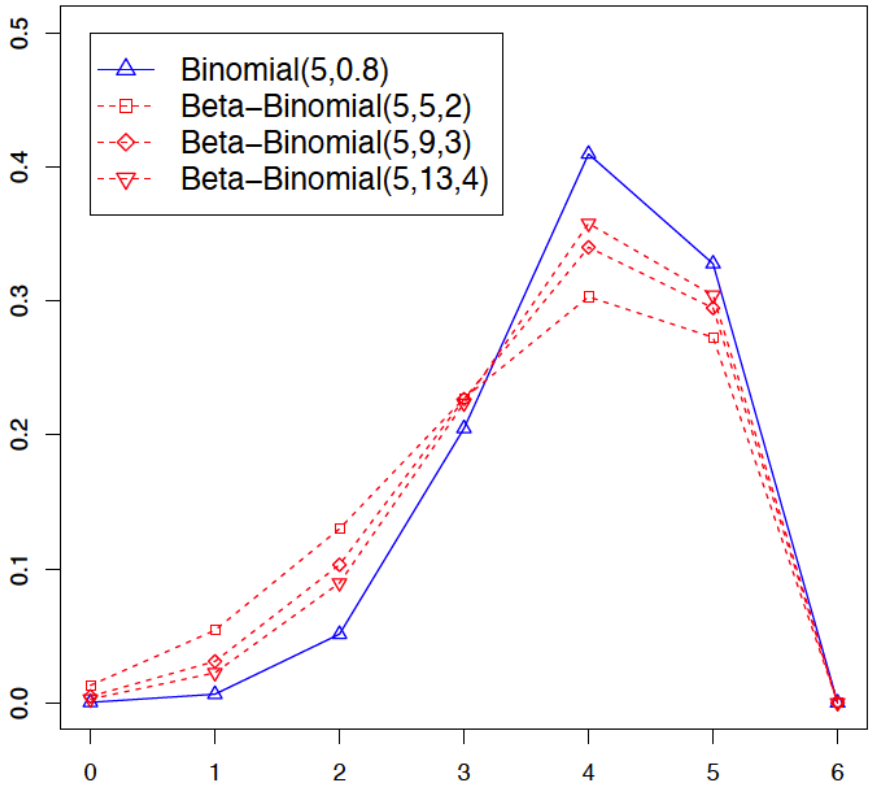
A natural history example





What is the right annotation for this specimen?

The distribution may help us



- The Beta-Binomial is a Binomial which parameter *p* is randomly drawn from a Beta distribution.
- Beta-binomial is more smoothed than Binomial. As data size grows, it tends to the Binomial.

Linked Open Piracy

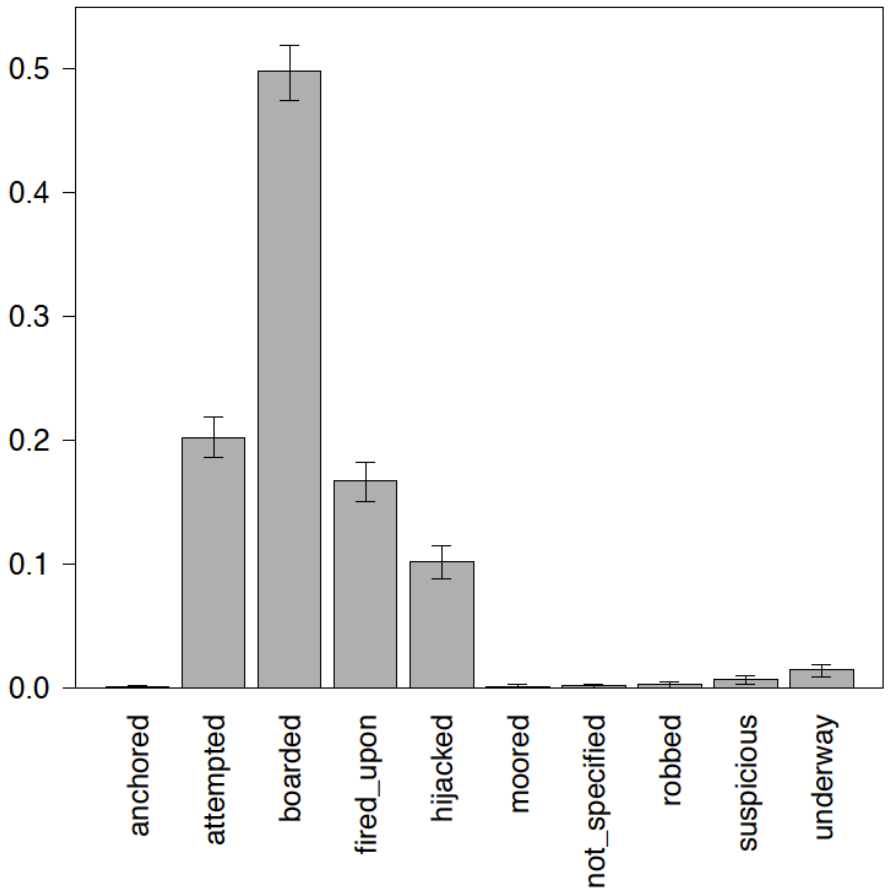
Linked Open Piracy is a repository about piracy attacks.

Time, place, attack type and ship type of each attack are recorded.

The repository is known to be accurate, but incomplete.

Let us see how to deal with this issue.

Estimating attack type proportions

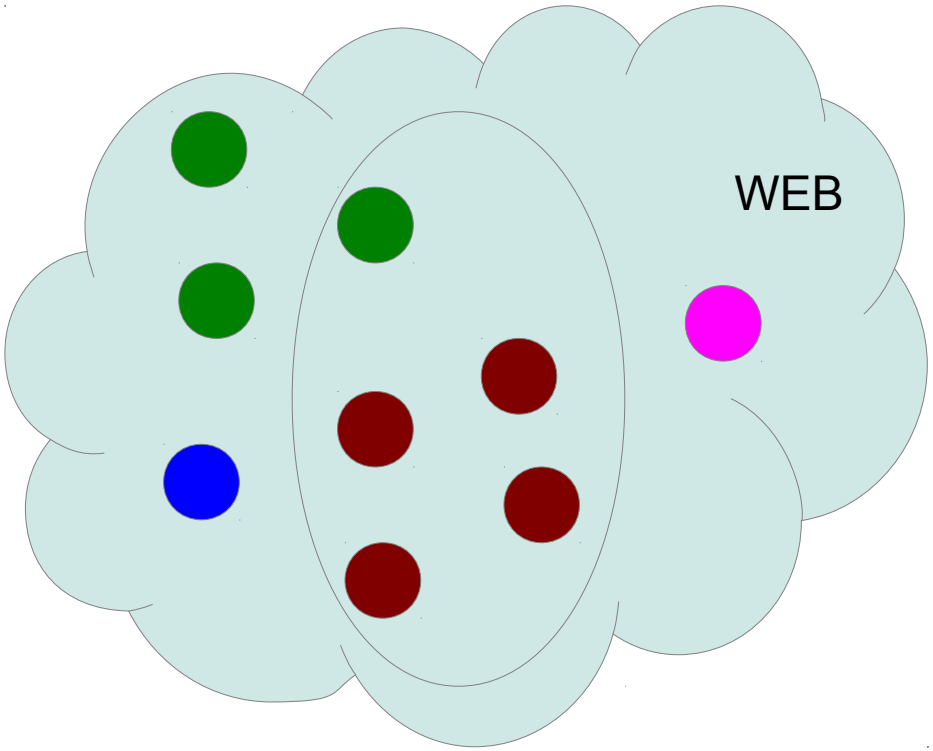


- Unknown population size.
- Our variables are not necessarily iid.
- Data are represented by a multinomial distribution.
- Using a Dirichlet prior we can estimate their uncertainty.

New attack type prediction

In many regions, new attack types show up over time.

How is is possible to estimate future proportions in this situation?



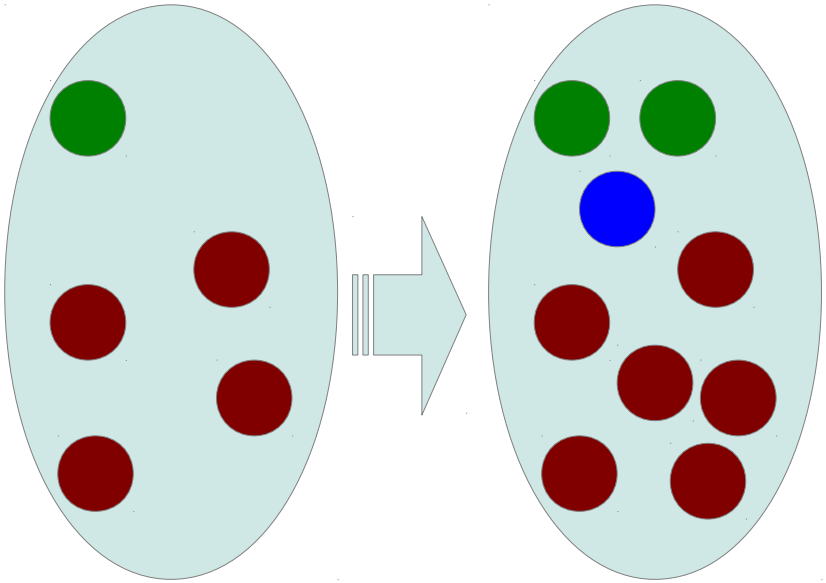
Dirichlet Process can help us!

- First: mapping. Attack types → [0..1]
- A priori, U[0…1] (events equally likely).
- Class of new observation can be
  - Drawn from U[0..1] (names are mapped manually);
  - Proportional to already observed data.
- The weight of observations increases as more data are seen.

Dirichlet processes as generalized Dirichlet distributions

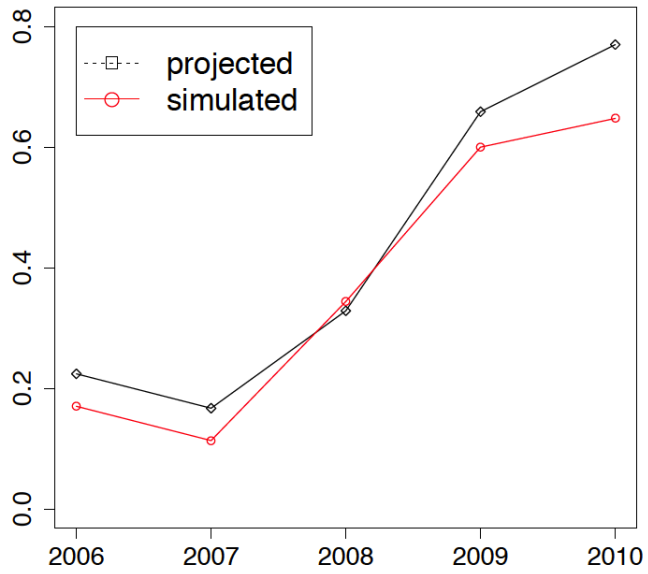
- Uncertainty about the proportions and uncertainty about the classes.

- Simulations driven by Dirichlet Process can provide good estimates.



Results

	Simulation	Projection
Average error	0.29	0.35
Variance	0.09	0.21



- Per region, we predict year *n+1* proportions, based on year *n* data.
- Dirichlet process performs better than a projection of the current proportions.

Conclusions

Web data are characterized by more layers of uncertainty.

Second order probabilities help handling part of these layers.

Dirichlet process helps to compensate when not all categories are known.

There is still much to do! Consider concrete domain data, integrate with logics, etc...

Thank you!

Questions?

d.ceolin@vu.nl
http://www.few.vu.nl/~dceolin

http://www.cs.vu.nl/lop