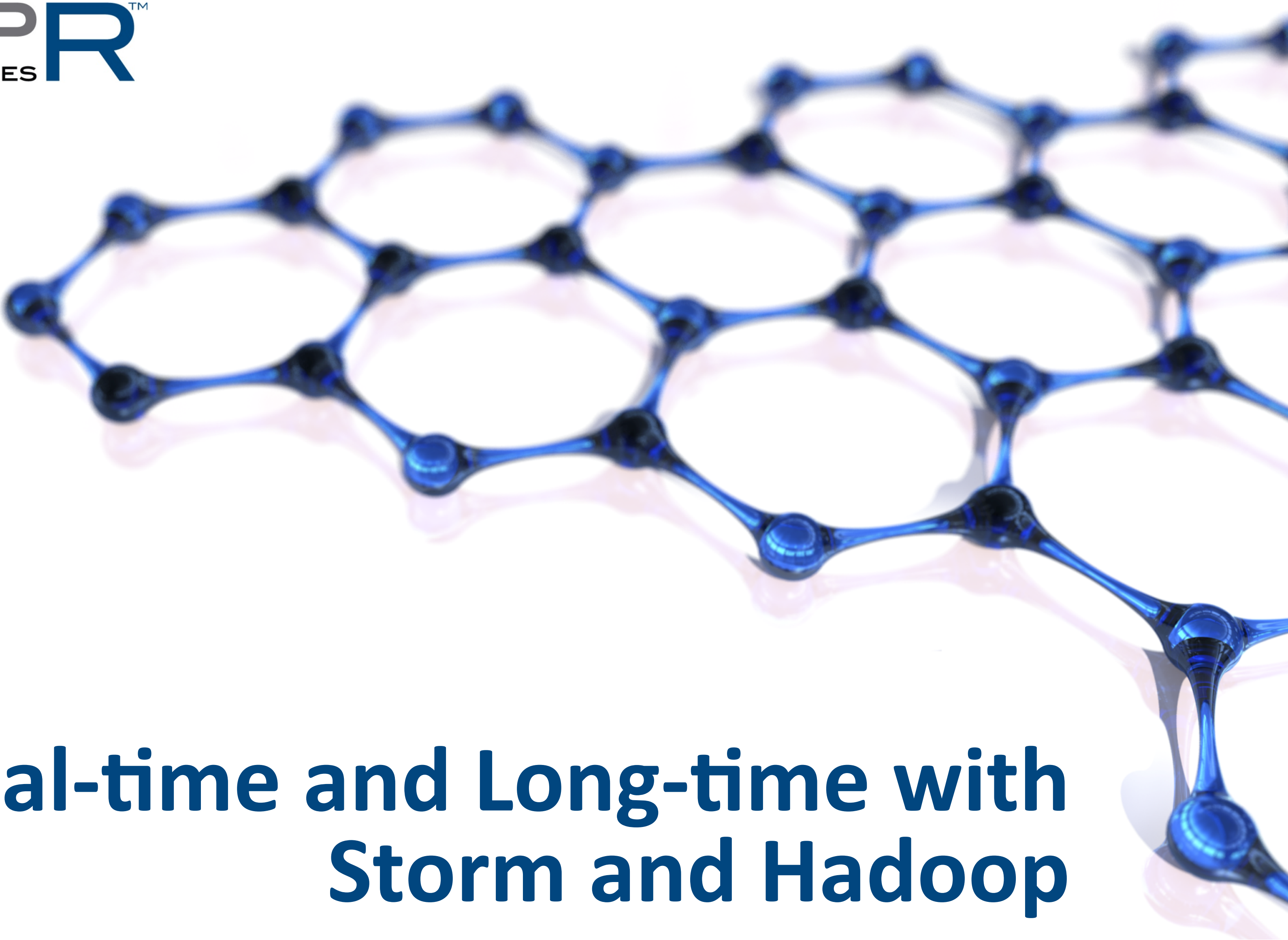
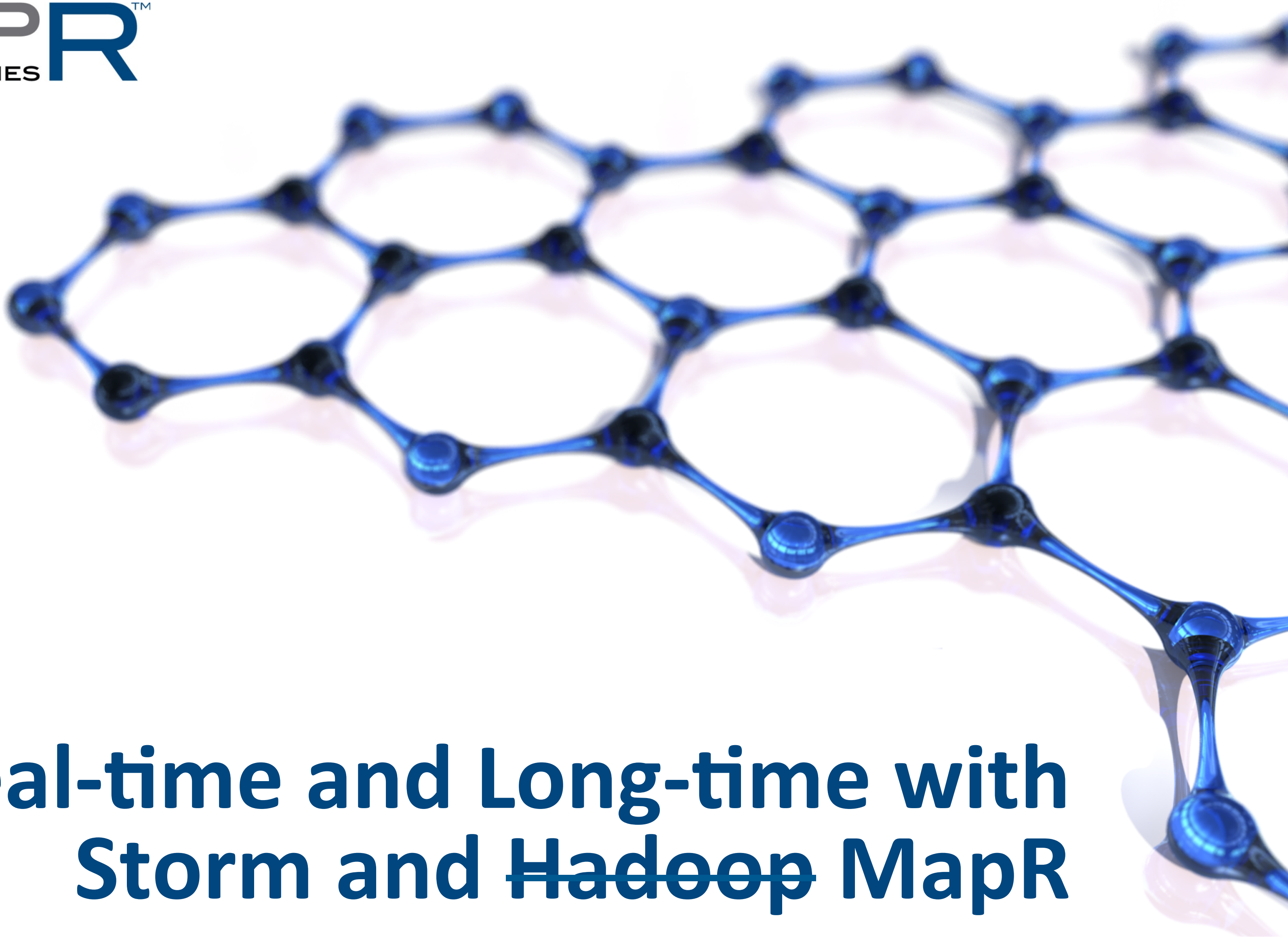
MAPR™
TECHNOLOGIES



Real-time and Long-time with Storm and Hadoop



Real-time and Long-time with Storm and ~~Hadoop~~ MapR

The Challenge

- Hadoop is great of processing vats of data
  - But sucks for real-time (by design!)

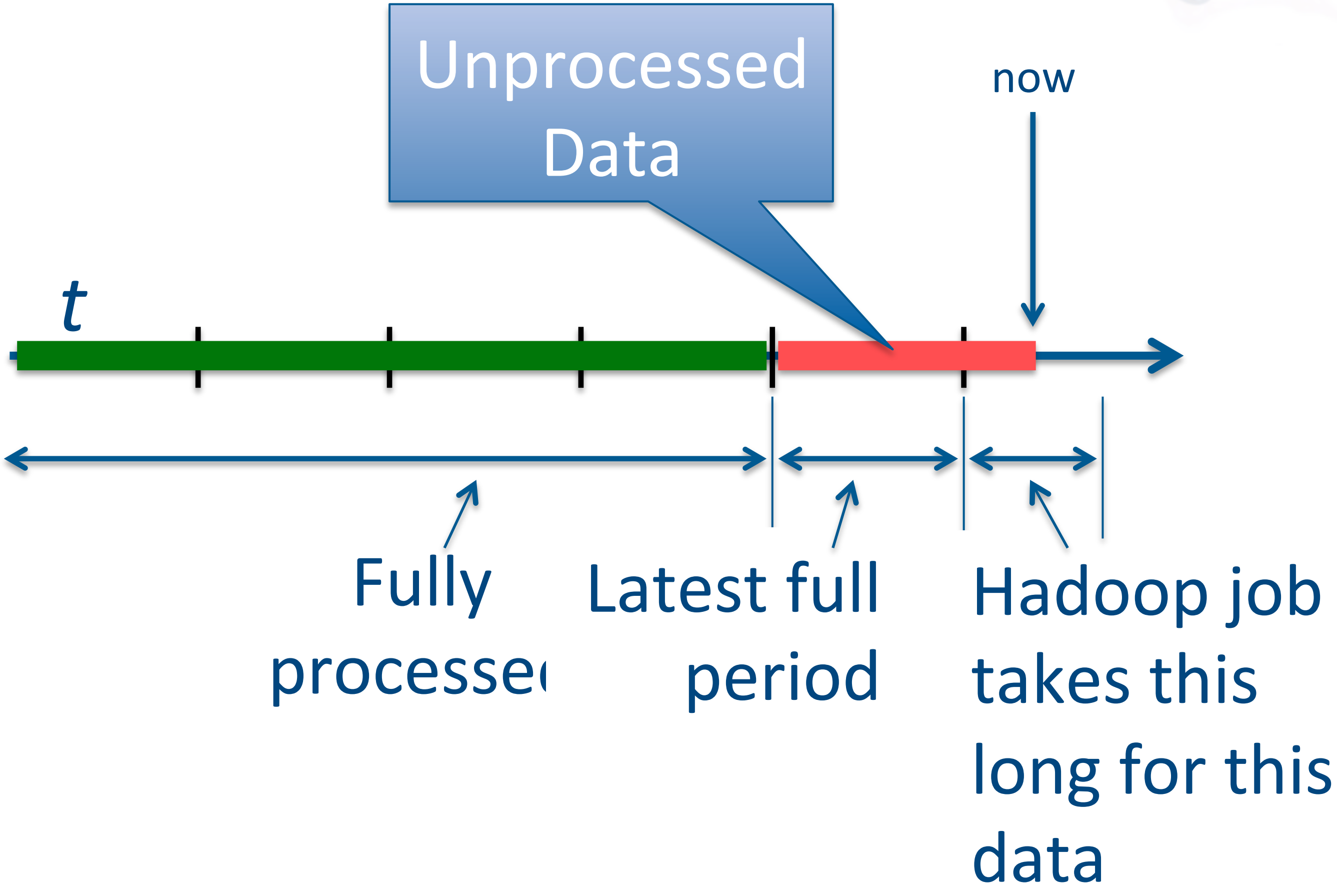
- Storm is great for real-time processing
  - But lacks any way to deal with batch processing

- It sounds like there isn’t a solution
  - Neither fashionable solution handles everything

This is not a problem.

It’s an opportunity!

Hadoop is Not Very Real-time



Unprocessed Data

now

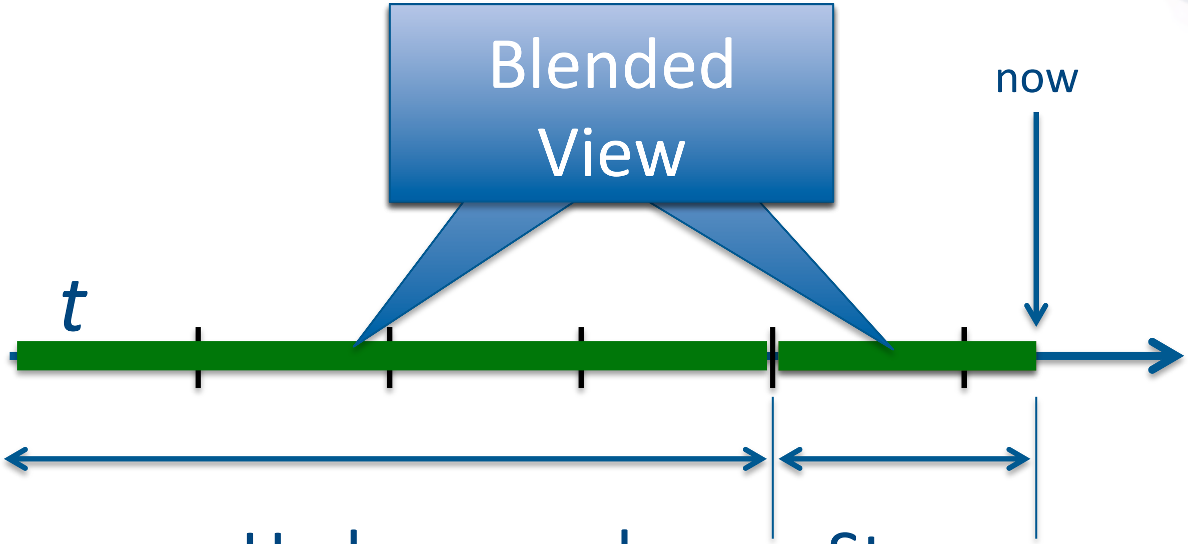
t

Fully processe

Latest full period

Hadoop job takes this long for this data

Real-time and Long-time together



Blended View

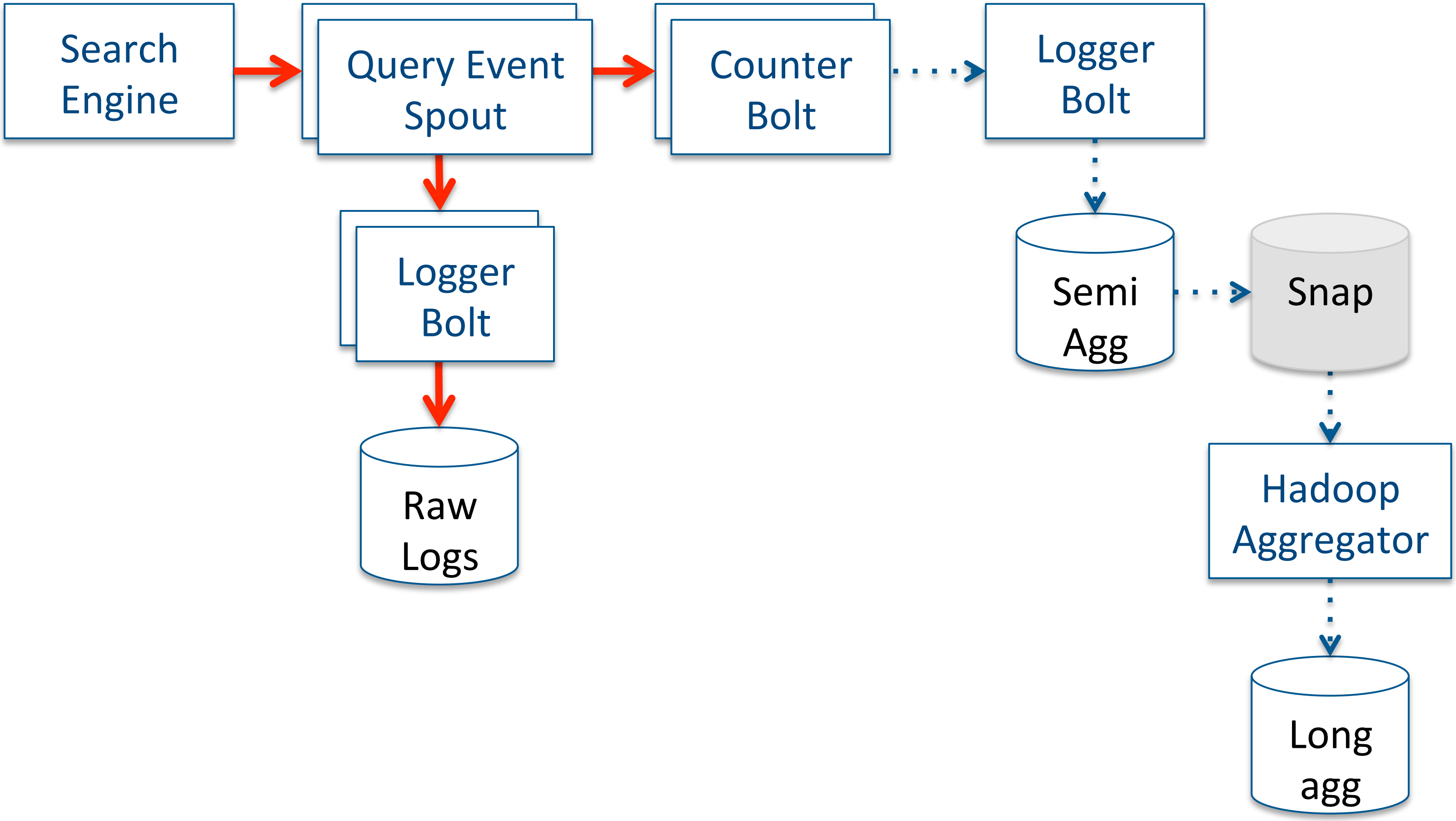
now

t

Hadoop works great back here

Storm works here

Rough Design – Data Flow



Guarantees

- Counter output volume is small-ish
  - the greater of *k* tuples per 100K inputs or *k* tuple/s
  - 1 tuple/s/label/bolt for this exercise
- Persistence layer must provide guarantees
  - distributed against node failure
  - must have either readable flush or closed-append
- HDFS is distributed, but provides no guarantees and strange semantics

- MapRfs is distributed, provides all necessary guarantees

Presentation Layer

- Presentation must
  - read recent output of Logger bolt
  - read relevant output of Hadoop jobs
  - combine semi-aggregated records
- User will see
  - counts that increment within 0-2 s of events
  - seamless meld of short and long-term data

Example 2 – AB testing in real-time

- I have 15 versions of my landing page
- Each visitor is assigned to a version
  - Which version?
- A conversion or sale or whatever can happen
  - How long to wait?
- Some versions of the landing page are horrible
  - Don't want to give them traffic

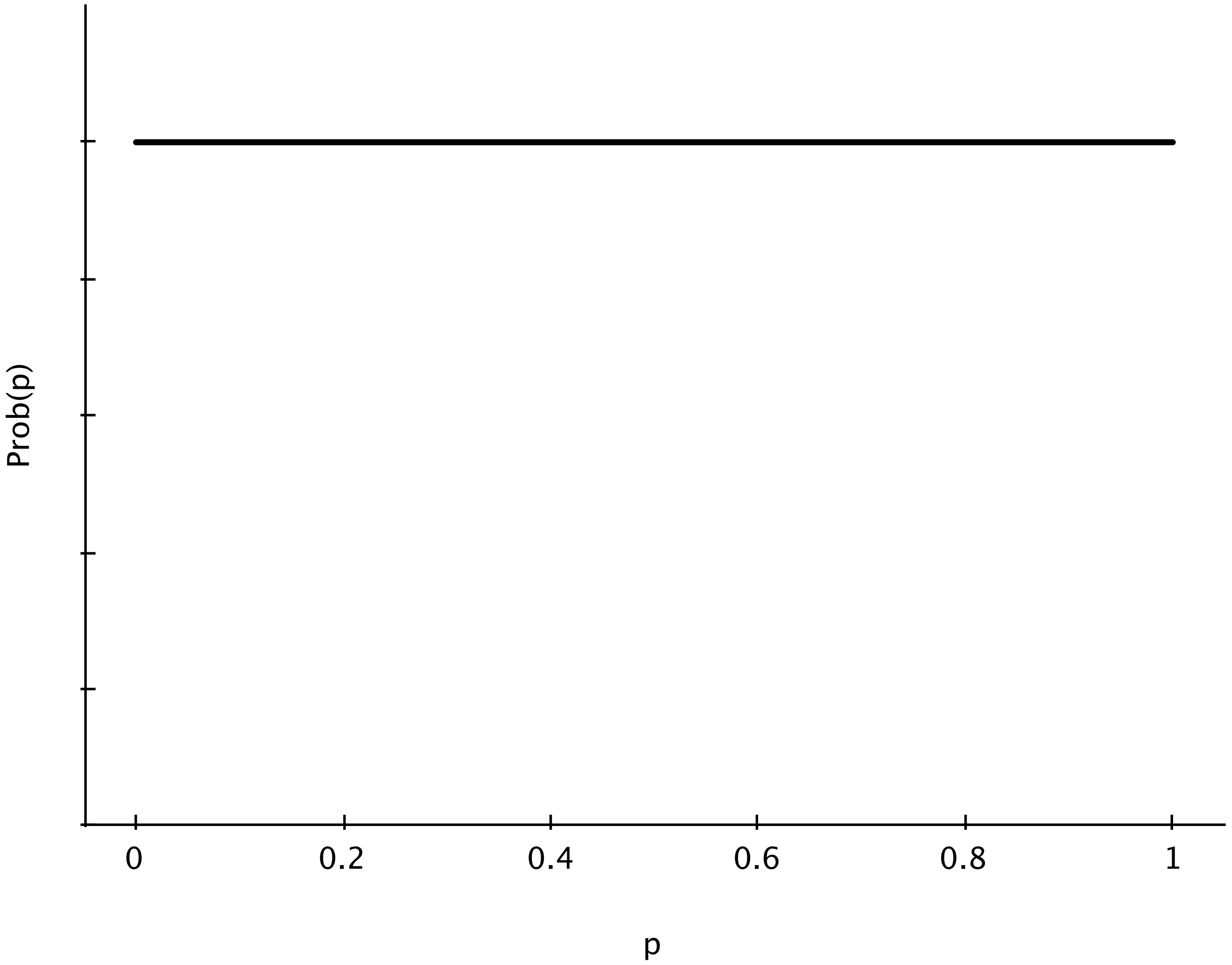
A Quick Diversion

- You see a coin
  - What is the probability of heads?
  - Could it be larger or smaller than that?
- I flip the coin and while it is in the air ask again
- I catch the coin and ask again
- I *look* at the coin (and you don't) and ask again
- Why does the answer change?
  - And did it ever have a single value?

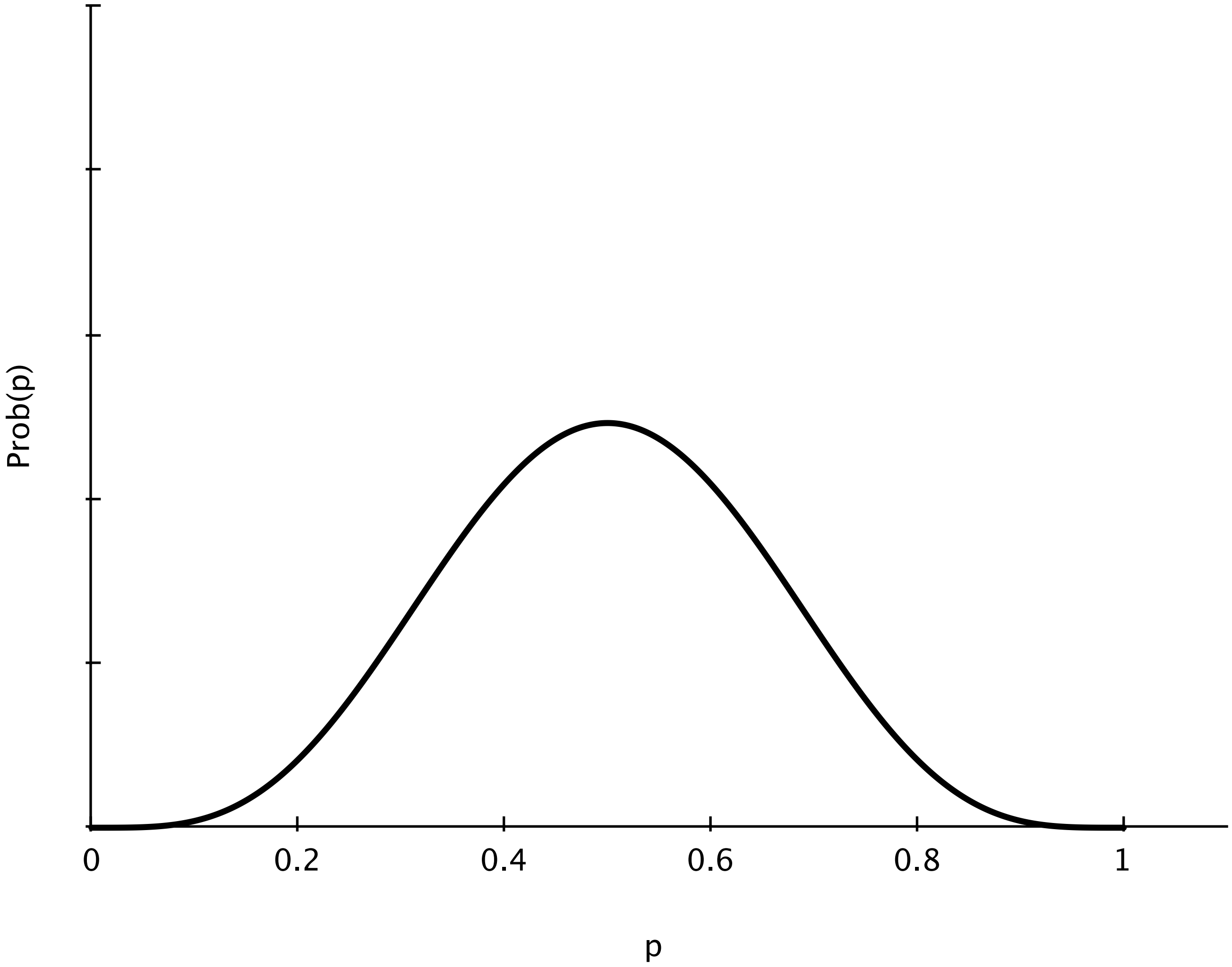
A Philosophical Conclusion

- Probability as expressed by humans is subjective and depends on information and experience

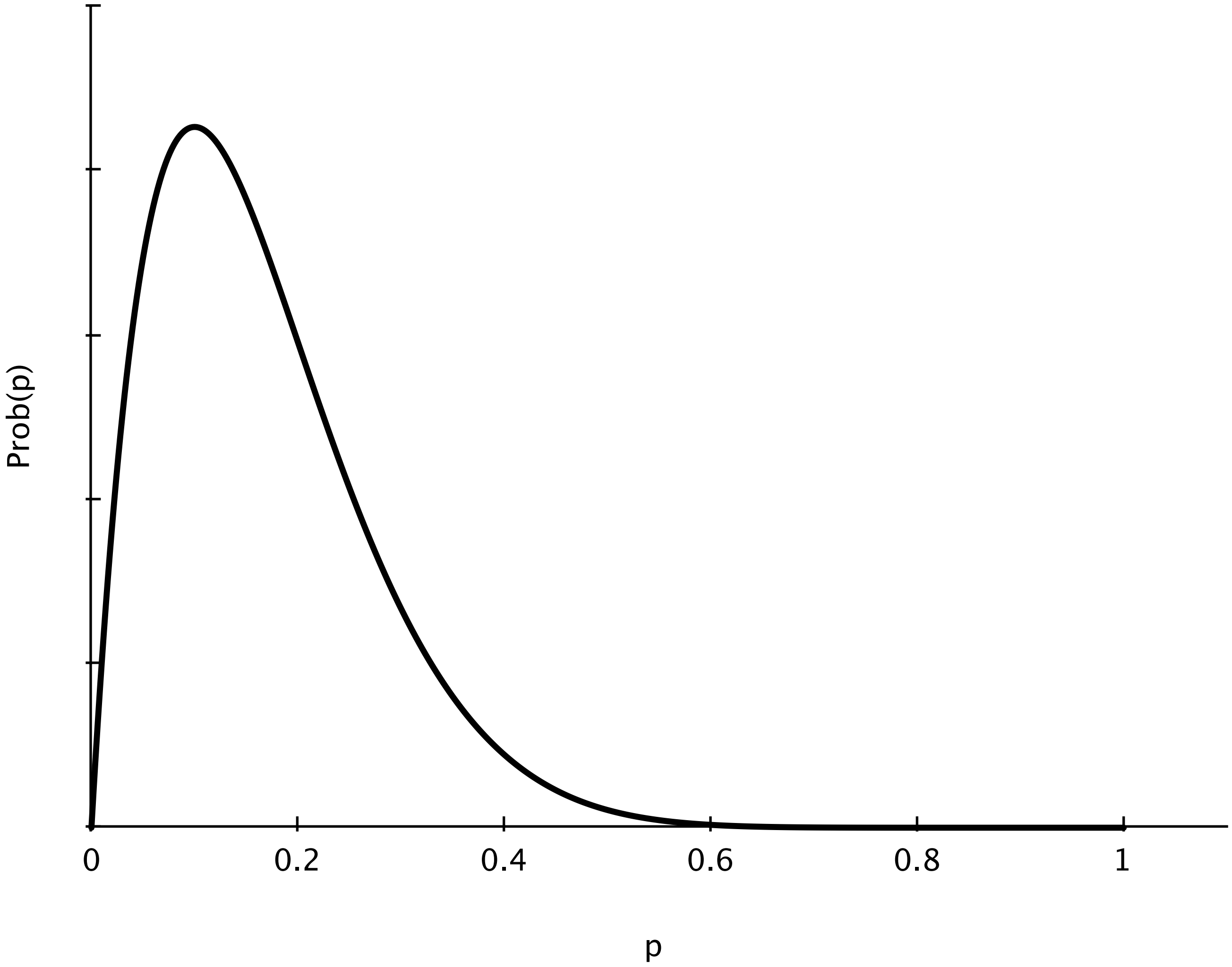
I Dunno



5 heads out of 10 throws



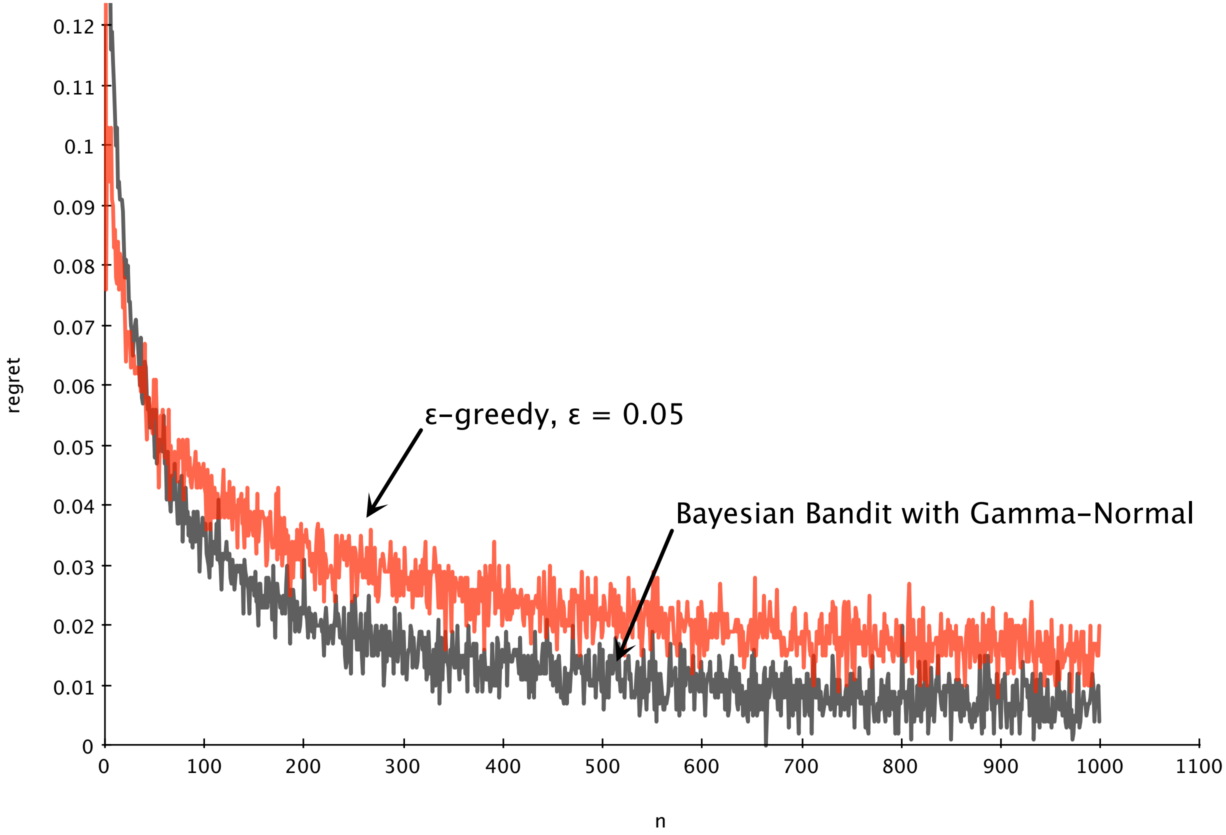
2 heads out of 12 throws



Bayesian Bandit

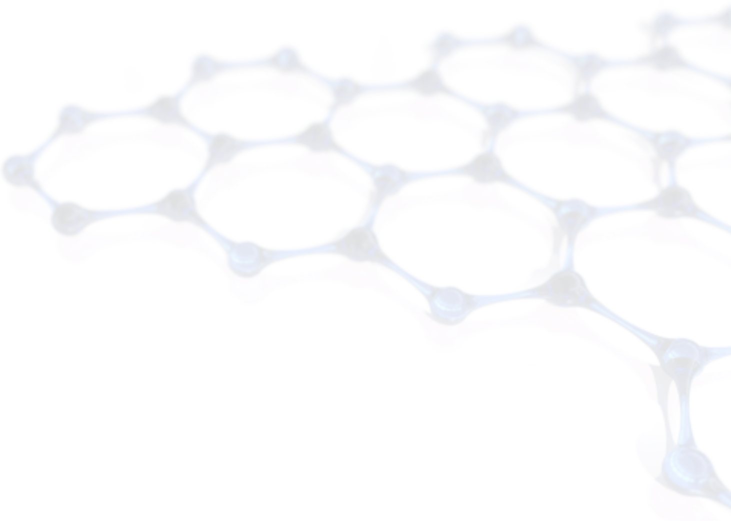
- Compute distributions based on data
- Sample p_1 and p_2 from these distributions
- Put a coin in bandit 1 if $p_1 > p_2$
- Else, put the coin in bandit 2

And it works!


0.12
0.11
0.1
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
regret
0
100
200
300
400
500
600
700
800
900
1000
1100
n
ε-greedy, ε = 0.05
Bayesian Bandit with Gamma-Normal

Video Demo

The Code



- Select an alternative

```
n = dim(k)[1]
p0 = rep(0, length.out=n)
for (i in 1:n) {
  p0[i] = rbeta(1, k[i,2]+1, k[i,1]+1)
}
return (which(p0 == max(p0)))
```

- Select and learn

```
for (z in 1:steps) {
   i = select(k)
   j = test(i)
   k[i,j] = k[i,j]+1
 }
 return (k)
```

- But we already know how to *count!*

The Basic Idea

- We can encode a distribution by sampling
- Sampling allows unification of exploration and exploitation

- Can be extended to more general response models

- Contact:
  - tdunning@maprtech.com
  - @ted_dunning

- Slides and such (available late tonight):
  - http://info.mapr.com/ted-bbuzz-2012

- Hash tags: #mapr #bbuzz

Collective notes: http://bit.ly/JDCRhc

Thank You