MAGNETIC™

Platform with Brains. Data with Soul.

Algorithms to Sample From Streams

Jonathan Arfa, Data Scientist @ Magnetic

What if you want to store a representative sample of data from a stream, in order to understand the distribution on-the-fly?

Data Streams

- Continuous
- Unknown length
- Hard to process with algorithms designed for batch data

Reservoir Sampling

Get a uniformly random, fixed-size sample from a stream of events of unknown length

Motivation for Reservoir Sampling

We want:

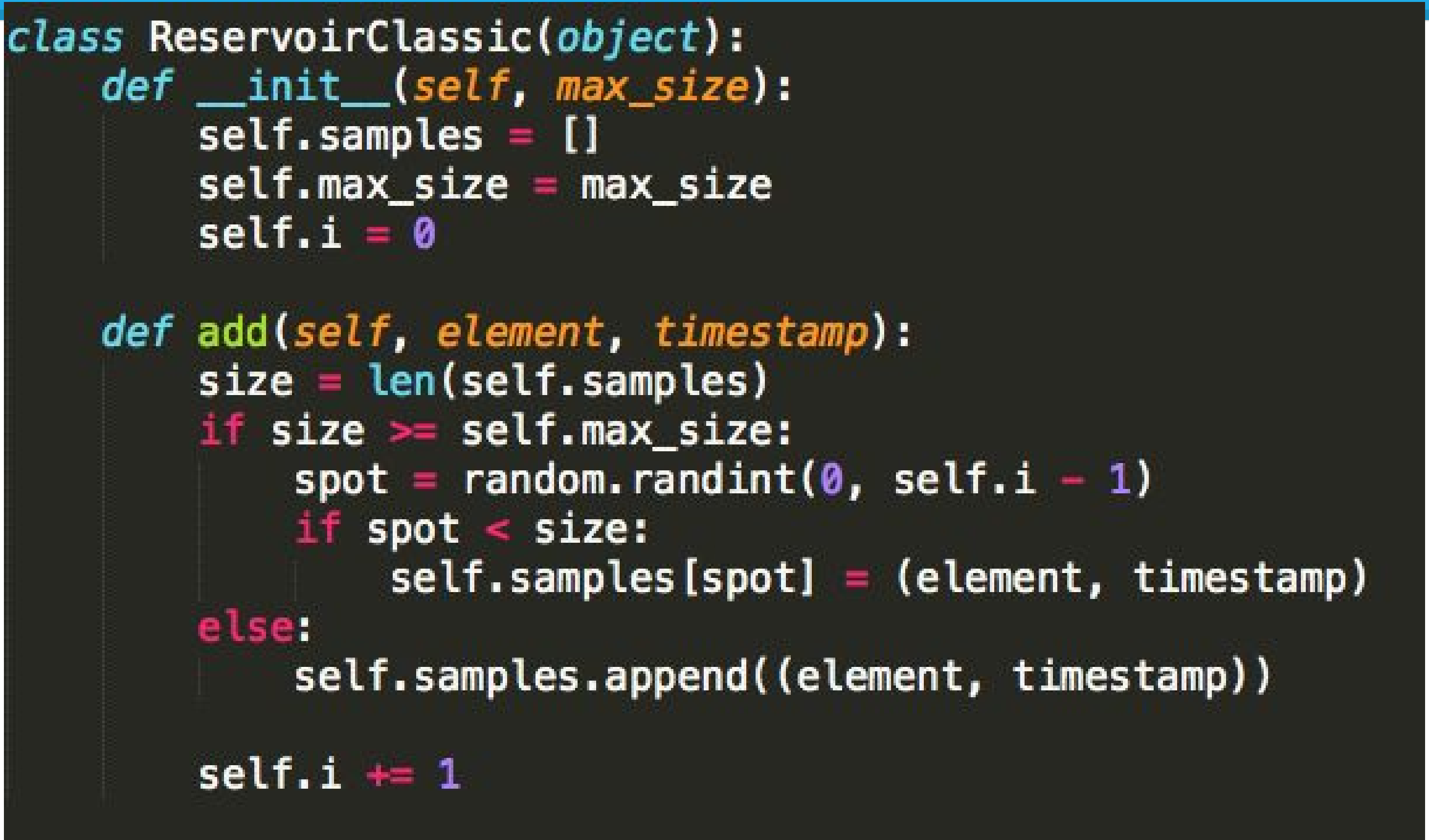
1. Exactly *K* samples
2. Unbiased samples - every event in an *N*-length stream (*N* could be unknown) should have an equal chance of being in our sample
3. Fast: an extra O(1) per event in the steam
4. Low Storage: only *K* events at any point

Reservoir Sampling

1. The first *K* events in the stream automatically enter the reservoir

2. For the *i*th event, if *i* > *K*: there's a *K* / *i* probability that it enters the reservoir. If so, it replaces a randomly selected event that's already there

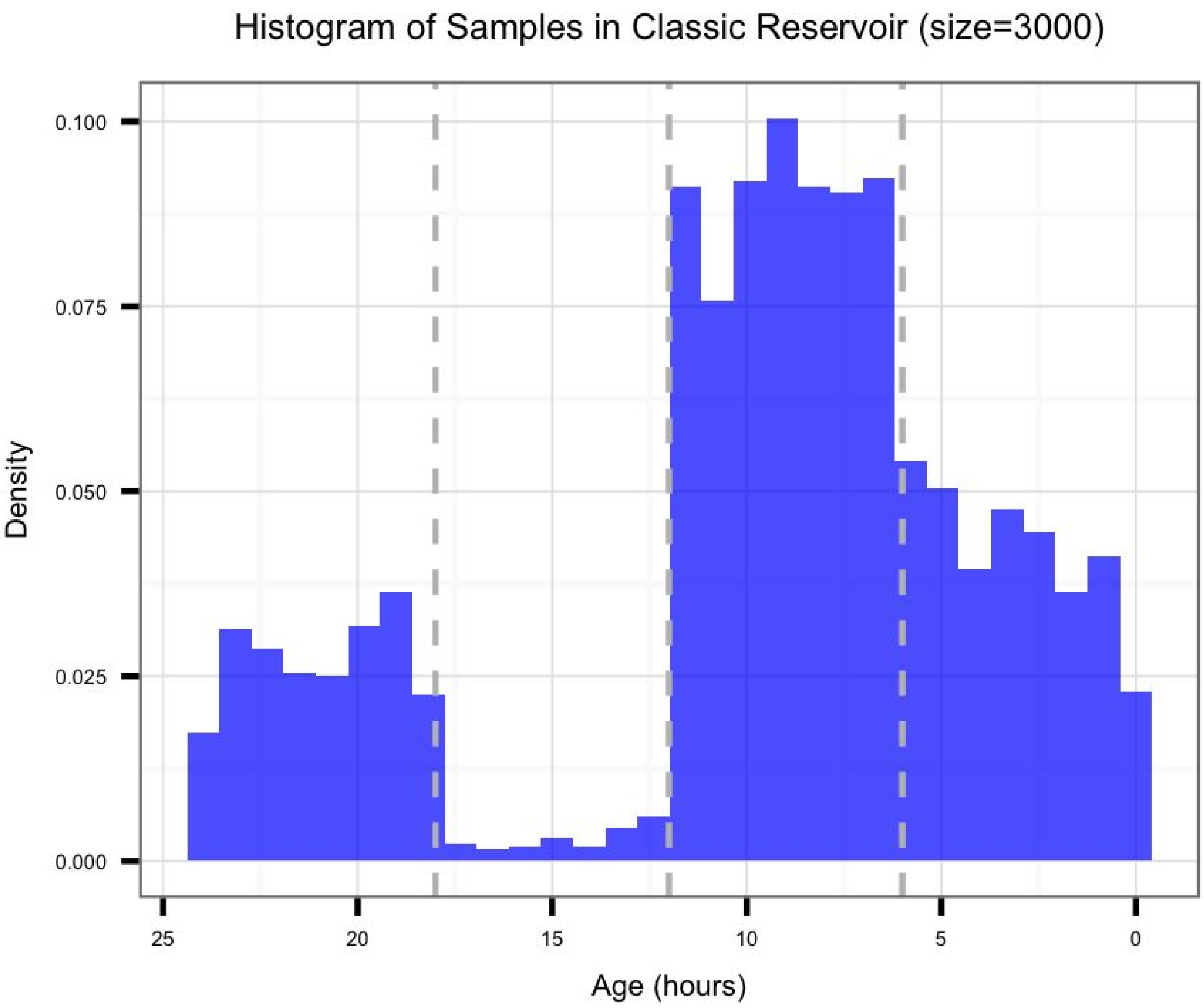
Reservoir Sampling

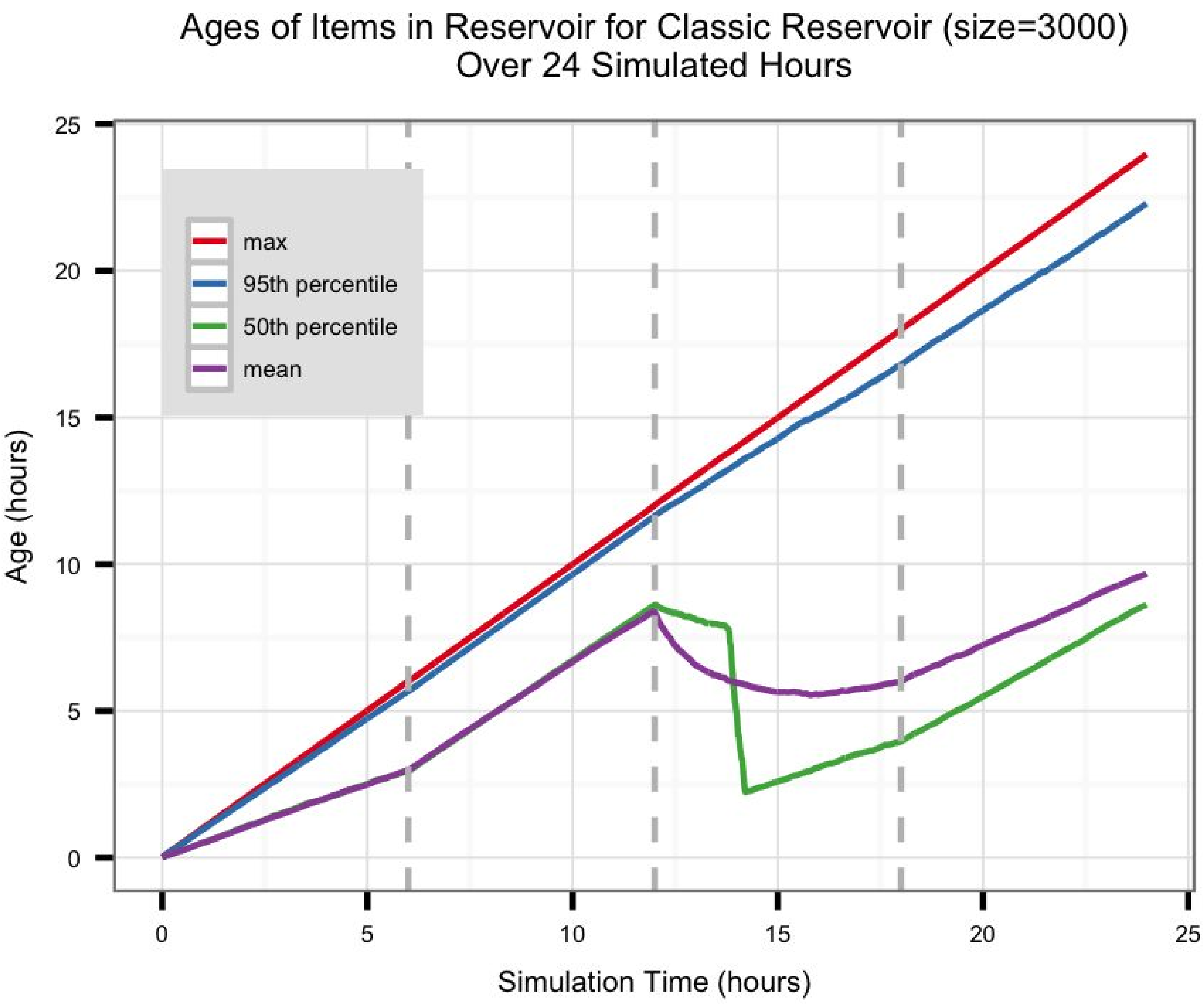


```python
class ReservoirClassic(object):
    def __init__(self, max_size):
        self.samples = []
        self.max_size = max_size
        self.i = 0

    def add(self, element, timestamp):
        size = len(self.samples)
        if size >= self.max_size:
            spot = random.randint(0, self.i - 1)
            if spot < size:
                self.samples[spot] = (element, timestamp)
        else:
            self.samples.append((element, timestamp))

        self.i += 1
```


Histogram of Samples in Classic Reservoir (size=3000)
Density
0.100
0.075
0.050
0.025
0.000
25
20
15
10
5
0
Age (hours)


Ages of Items in Reservoir for Classic Reservoir (size=3000)
Over 24 Simulated Hours
max
95th percentile
50th percentile
mean
Age (hours)
0
5
10
15
20
25
Simulation Time (hours)
0
5
10
15
20
25

But what if you don’t want unbiased samples?

VIRBs

Variable **I**ncoming **R**ate **B**iased Samplers

Collaborators: Jonathan Arfa, Dan Crosta, Sam Steingold, Vladimir Vladimirov (formerly Magnetic)

VIRBs

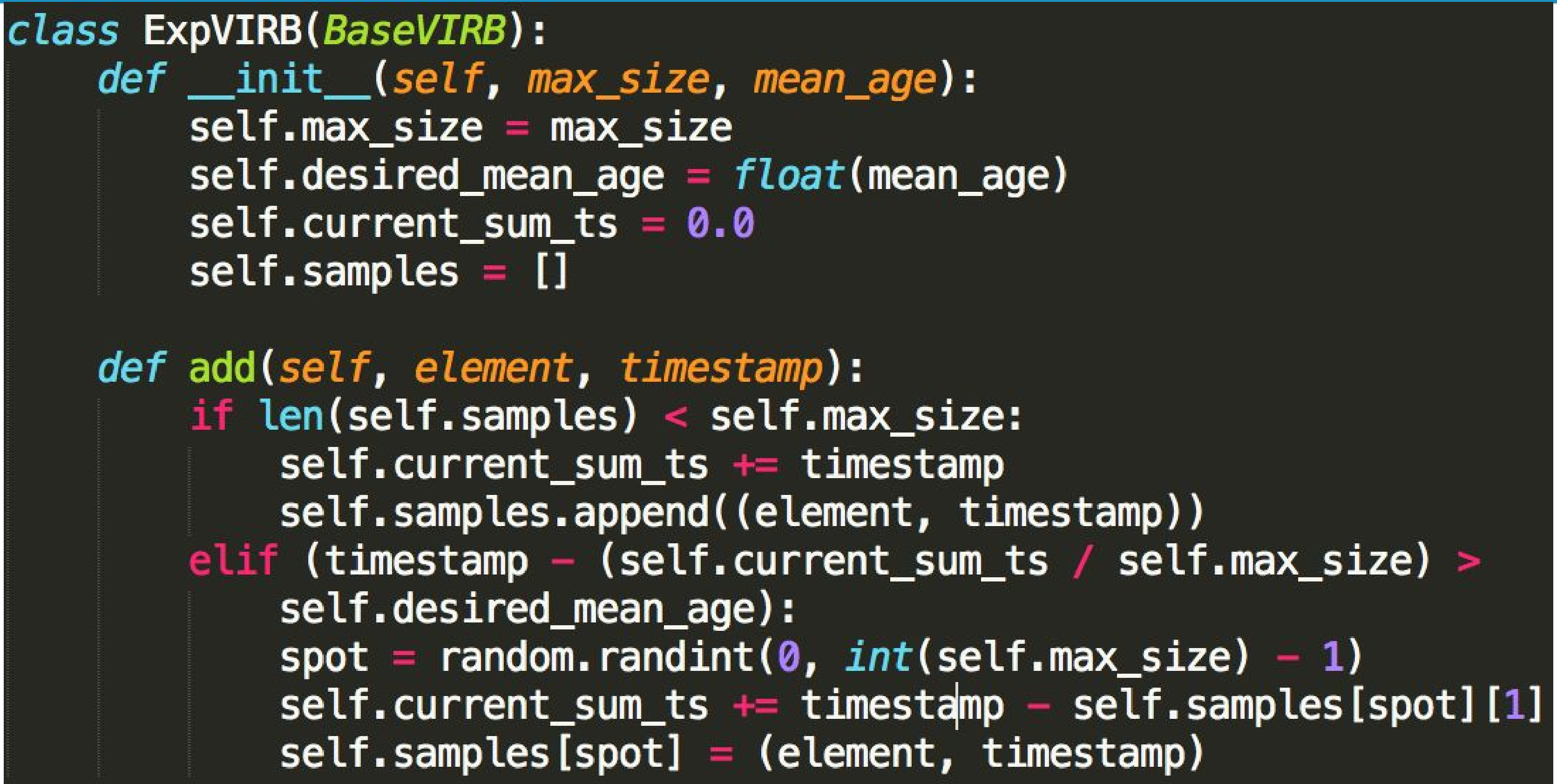
1. Specify both *K* (max_size) and the desired mean_age

2. The first *K* events in the stream automatically enter the reservoir

3. For any subsequent event: enter the reservoir only if the current mean age of events in the reservoir is older than the desired mean age
   a. But what event does it replace? Two versions

Exponential VIRB

Replace a random event

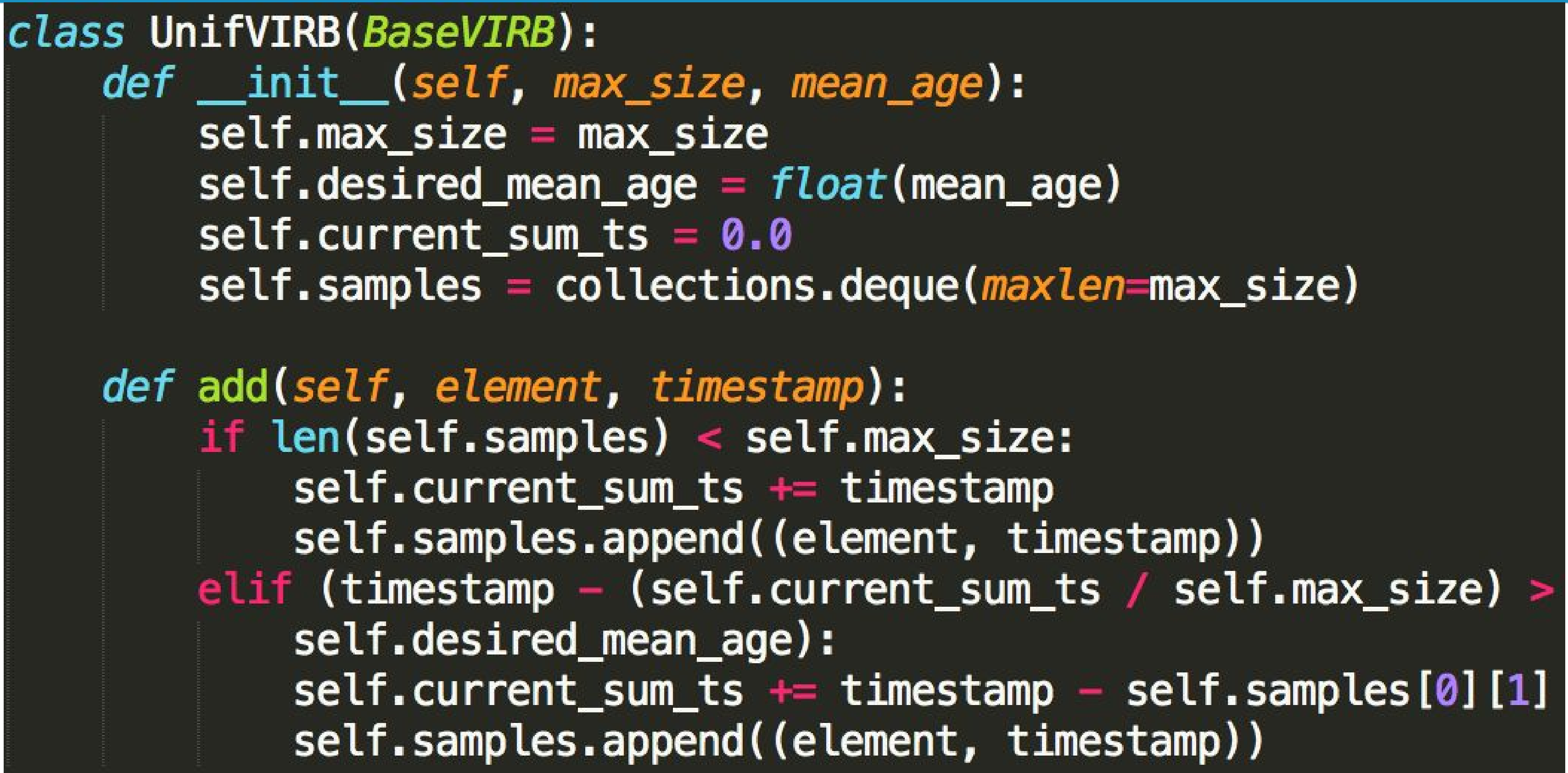


```
class ExpVIRB(BaseVIRB):
    def __init__(self, max_size, mean_age):
        self.max_size = max_size
        self.desired_mean_age = float(mean_age)
        self.current_sum_ts = 0.0
        self.samples = []

    def add(self, element, timestamp):
        if len(self.samples) < self.max_size:
            self.current_sum_ts += timestamp
            self.samples.append((element, timestamp))
        elif (timestamp - (self.current_sum_ts / self.max_size) >
            self.desired_mean_age):
            spot = random.randint(0, int(self.max_size) - 1)
            self.current_sum_ts += timestamp - self.samples[spot][1]
            self.samples[spot] = (element, timestamp)
```

Uniform VIRB

Replace the oldest event



```
class UnifVIRB(BaseVIRB):
    def __init__(self, max_size, mean_age):
        self.max_size = max_size
        self.desired_mean_age = float(mean_age)
        self.current_sum_ts = 0.0
        self.samples = collections.deque(maxlen=max_size)

    def add(self, element, timestamp):
        if len(self.samples) < self.max_size:
            self.current_sum_ts += timestamp
            self.samples.append((element, timestamp))
        elif (timestamp - (self.current_sum_ts / self.max_size) >
            self.desired_mean_age):
            self.current_sum_ts += timestamp - self.samples[0][1]
            self.samples.append((element, timestamp))
```

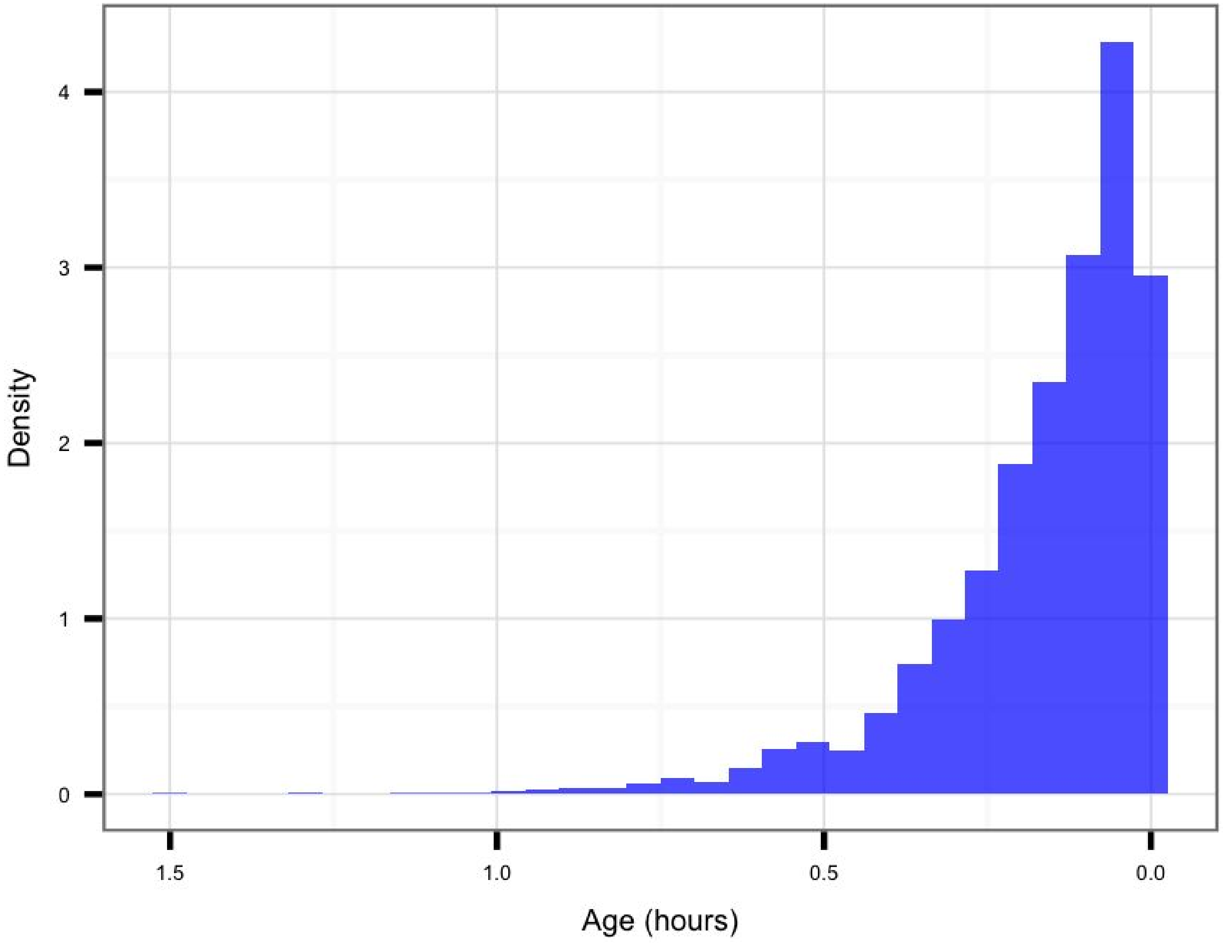
VIRBs

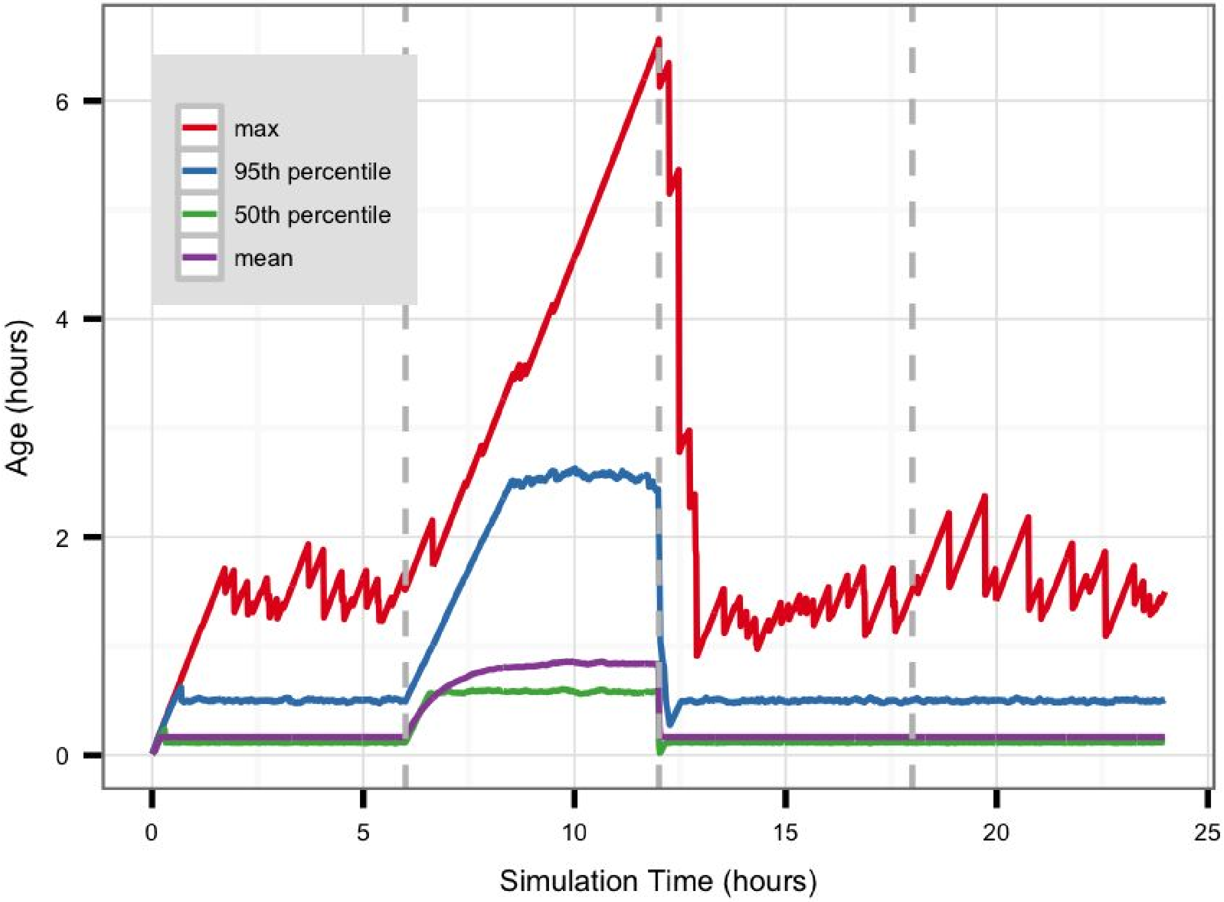
Questions

1. To what extent are these random samples?

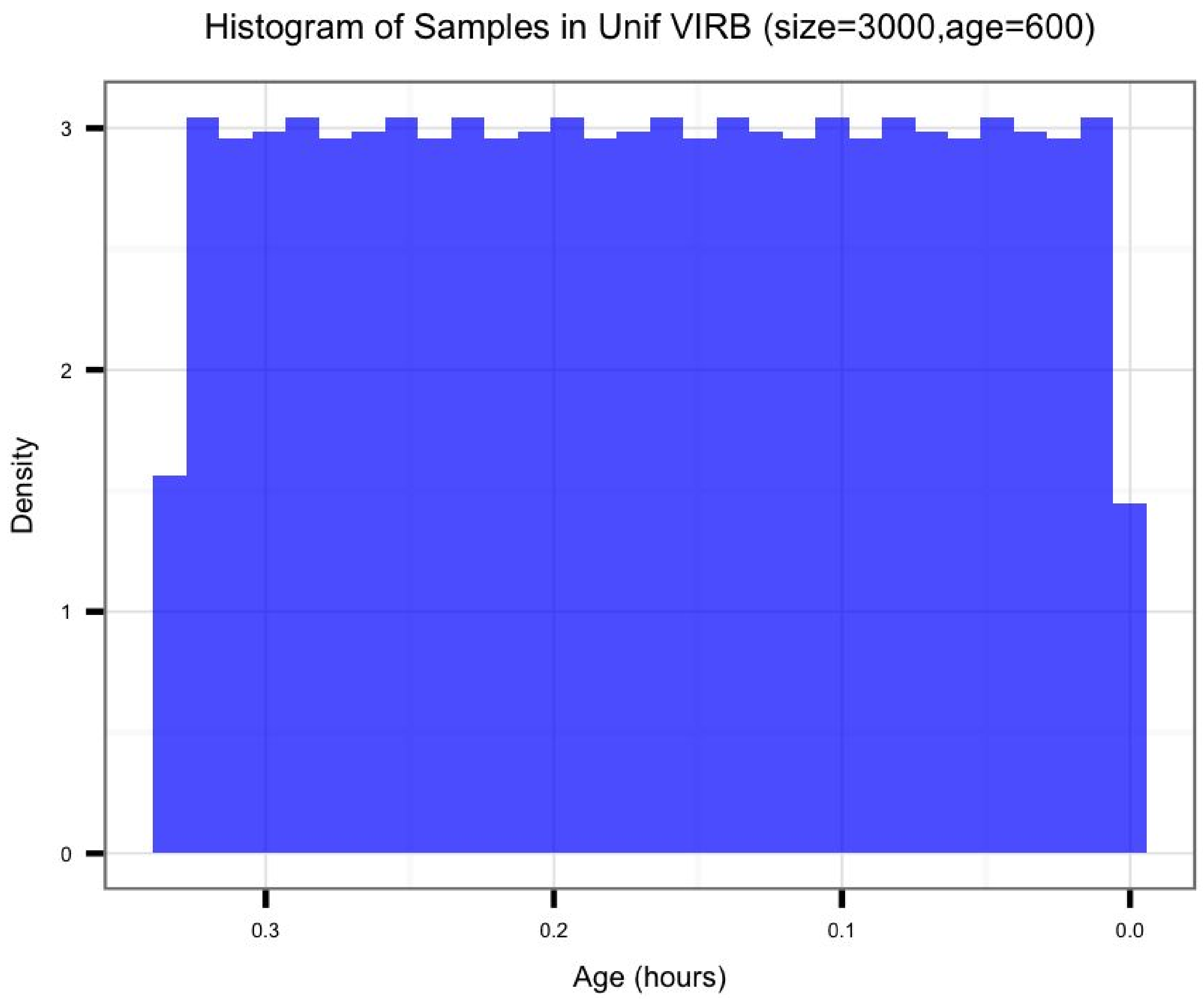
2. What happens if the incoming rate is too low to keep *K* events at a defined *mean_age* ?

Histogram of Samples in Exp VIRB (size=3000,age=600)

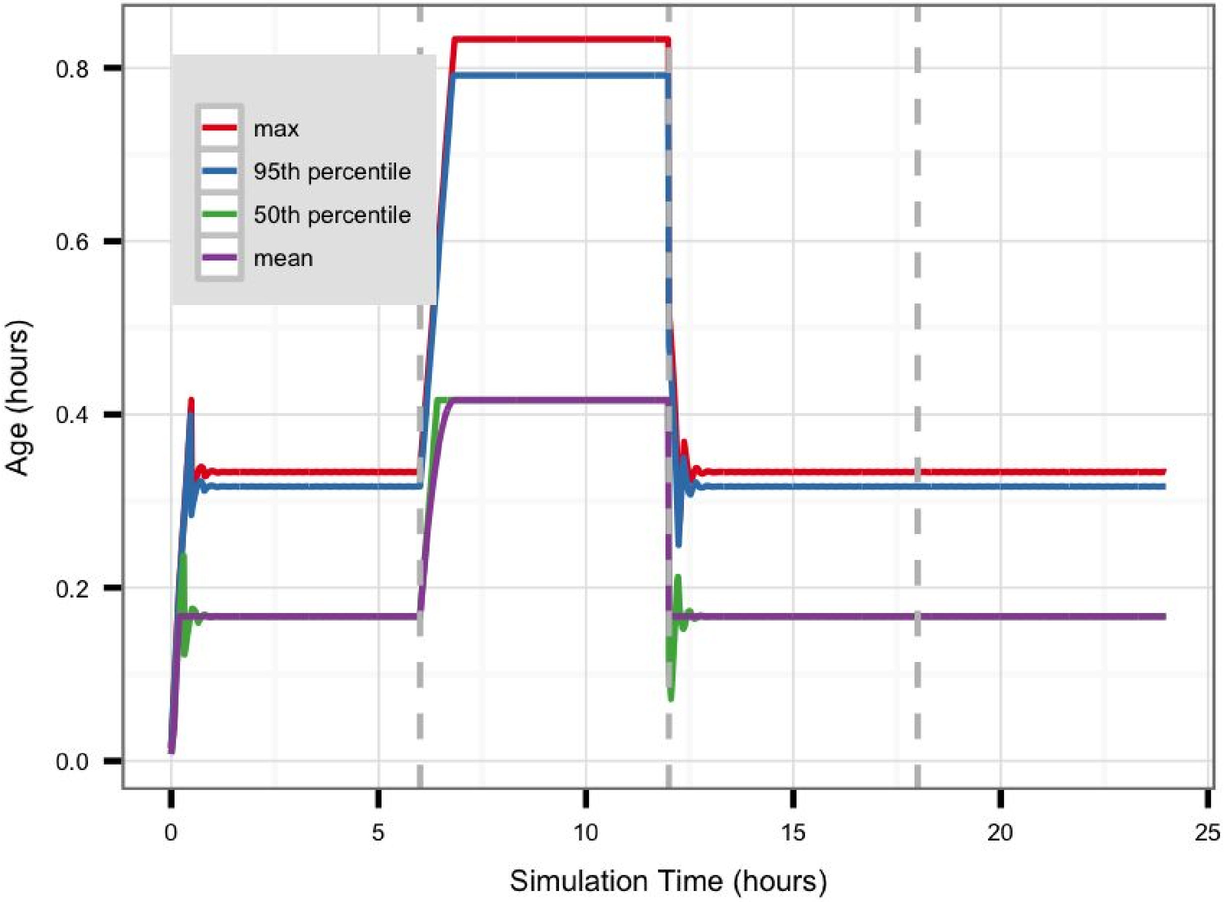




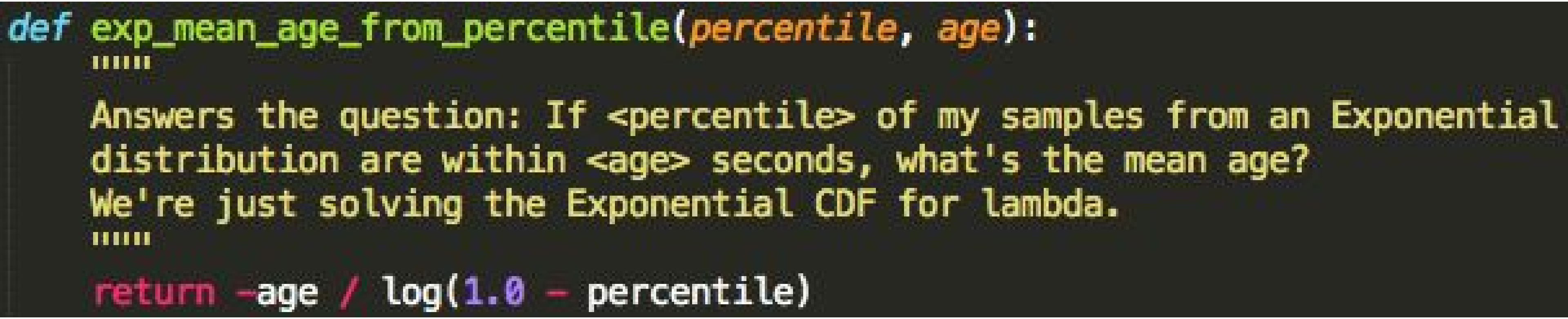
Ages of Items in Reservoir for Exp VIRB (size=3000,age=600)
Over 24 Simulated Hours


Histogram of Samples in Unif VIRB (size=3000,age=600)
Density
3
2
1
0
0.3
0.2
0.1
0.0
Age (hours)

Ages of Items in Reservoir for Unif VIRB (size=3000,age=600)
Over 24 Simulated Hours

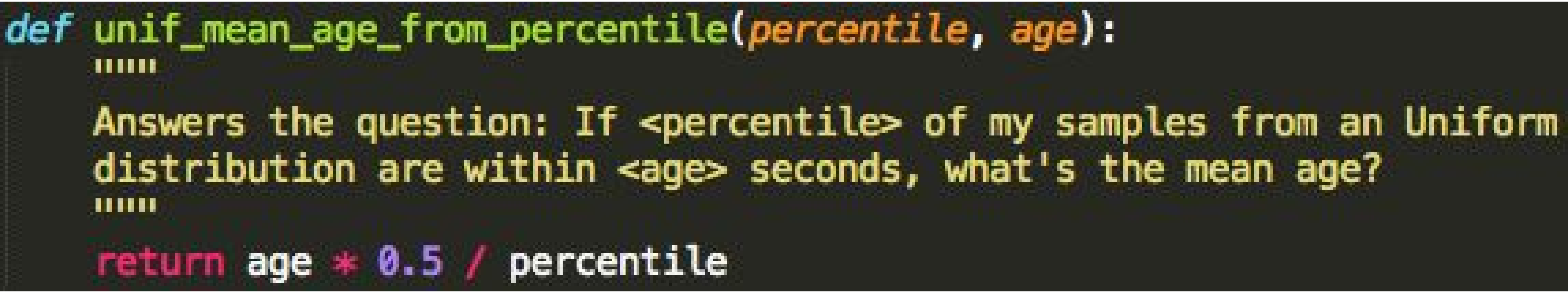


Flexible Age Specification



```
def exp_mean_age_from_percentile(percentile, age):
    """
    Answers the question: If <percentile> of my samples from an Exponential
    distribution are within <age> seconds, what's the mean age?
    We're just solving the Exponential CDF for lambda.
    """
    return -age / log(1.0 - percentile)
```

Flexible Age Specification



```python
def unif_mean_age_from_percentile(percentile, age):
    """
    Answers the question: If <percentile> of my samples from an Uniform
    distribution are within <age> seconds, what's the mean age?
    """
    return age * 0.5 / percentile
```

Aggarwal’s Reservoir Sampler

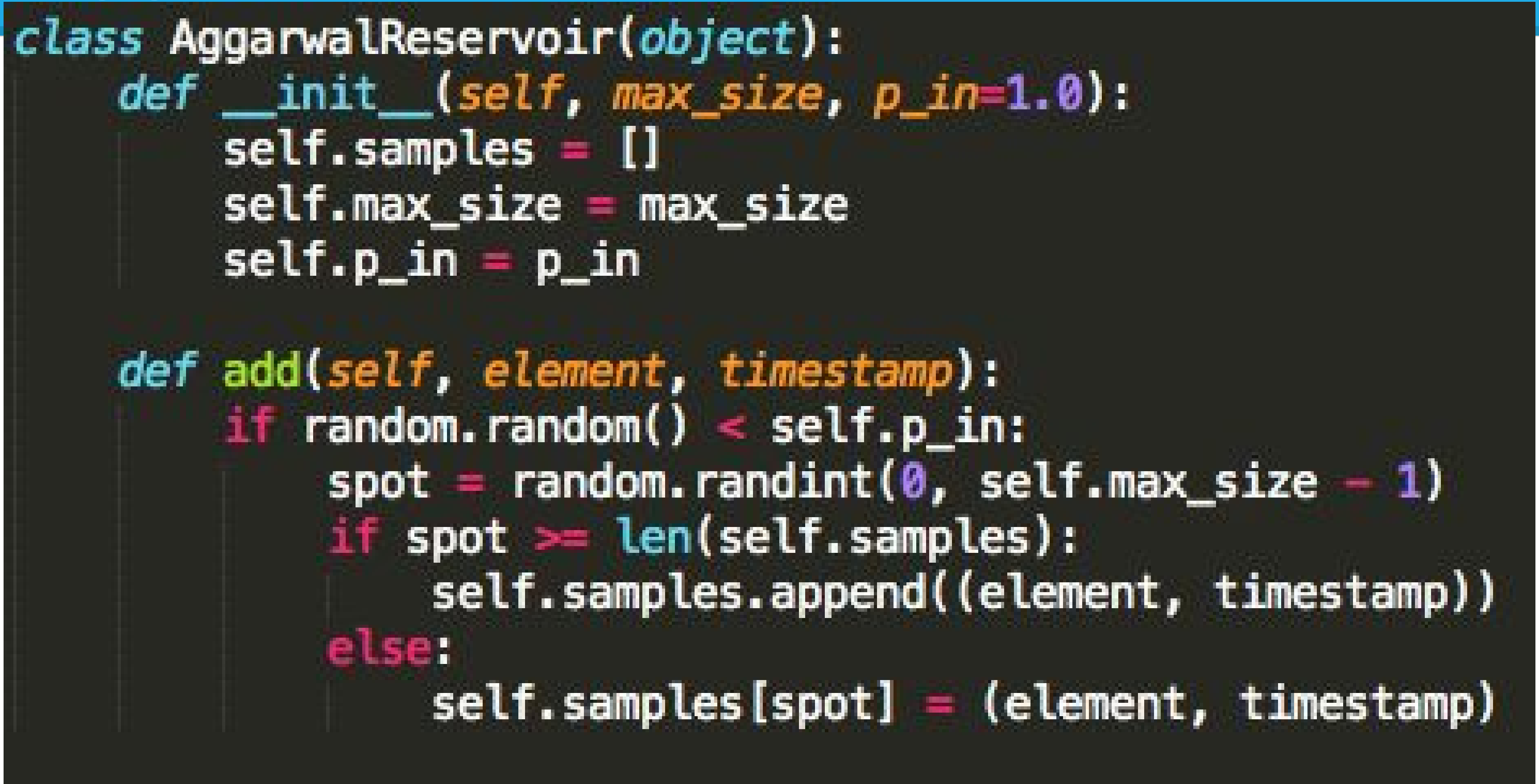
“On Biased Reservoir Sampling in the Presence of Stream Evolution” - Charu C. Aggarwal
Algorithm 3.1

Aggarwal’s Reservoir Sampler

1. The event enters the reservoir with probability *p_in*, otherwise it’s discarded

2. If the current size of the reservoir is *N* out of a maximum of *K*,
   a. the event replaces a random pre-existing event with probability *N* / *K*.
   b. Otherwise, it’s added to the end of the reservoir, making it bigger.

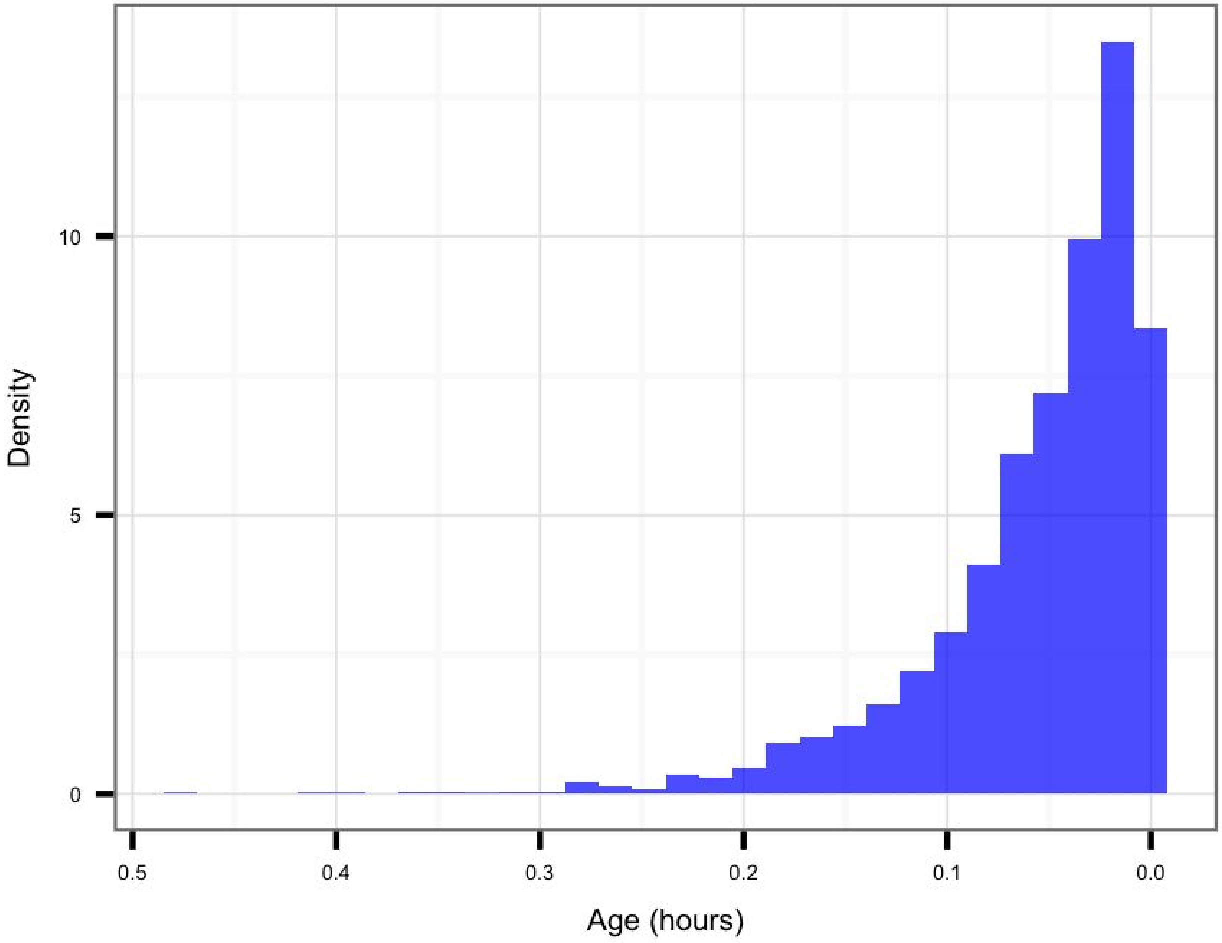
Aggarwal’s Reservoir Sampler

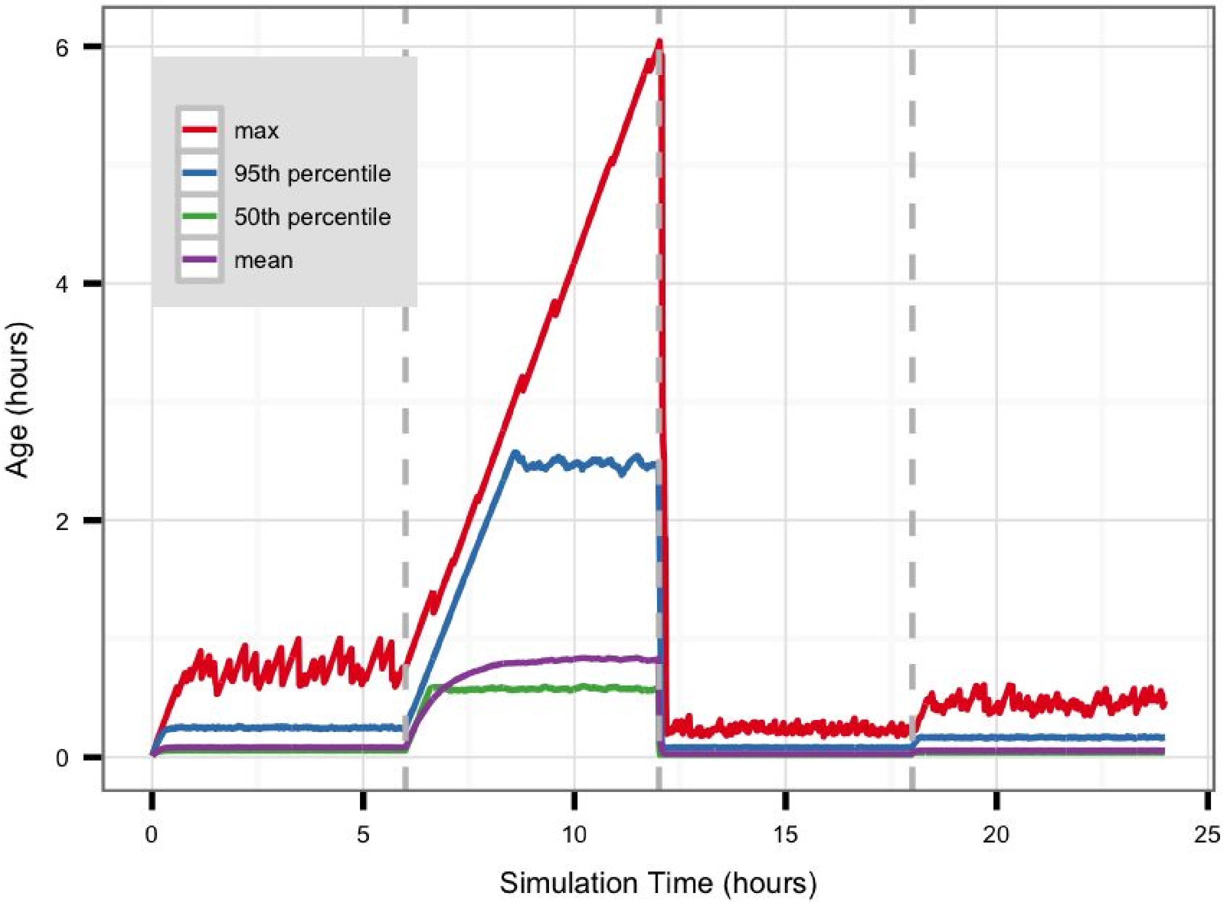


```
class AggarwalReservoir(object):
    def __init__(self, max_size, p_in=1.0):
        self.samples = []
        self.max_size = max_size
        self.p_in = p_in

    def add(self, element, timestamp):
        if random.random() < self.p_in:
            spot = random.randint(0, self.max_size - 1)
            if spot >= len(self.samples):
                self.samples.append((element, timestamp))
            else:
                self.samples[spot] = (element, timestamp)
```

Histogram of Samples in Aggarwal (size=3000)





Ages of Items in Reservoir for Aggarwal (size=3000)
Over 24 Simulated Hours

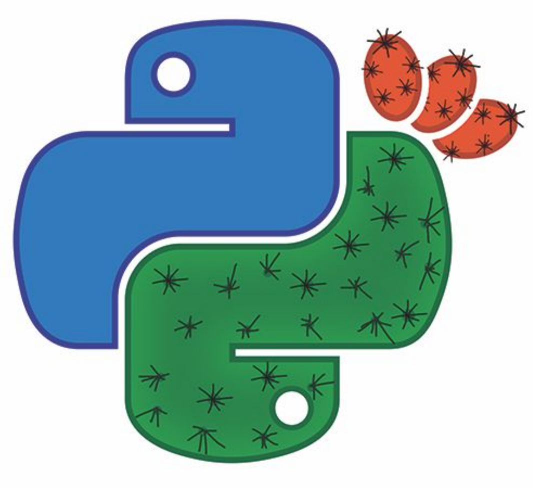
- Want to sample uniformly over **all** events?
    - Old-school reservoir sampling

- Want to sample from a defined period of time with a defined shape?
    - VIRBs, courtesy of team Magnetic

Overly Complicated Table

Algorithm	Parameters	Add new event if:	New events replace:	Samples over	Time till full reservoir	Shape
Reservoir Sampling	max size	random() < (max_size / i)	random event	events (all)	K events seen	Uniform
Aggarwal's 3.1	max size, p_in	random() < p_in	random (it's complicated)	events (recent)	Longer	Exponential
Uniform VIRB	max size, mean age	current age > desired age	oldest event	time (recent)	K events seen	Uniform
Exponential VIRB	max size, mean age	current age > desired age	random event	time (recent)	K events seen	Exponential

MAGNETIC™



http://tech.magnetic.com/2016/04/virbs-sampling-events-from-streams.html