Approximation of Distributions via T-Digests

Dr. Alexander Weiß

ALGORITHMS DATA CHALLENGES

Two algorithms walks into bar... part I

Zalando

March 5, 2015


Improved Formula
Health King康帝®
Digest
(Caffeine–Free)
Herb Tea™
助消化茶
Help digestion, clear liver heat & increase acidity
100% NATURAL
20 Tea bags
Net Wt: 1.26 oz (36 g)

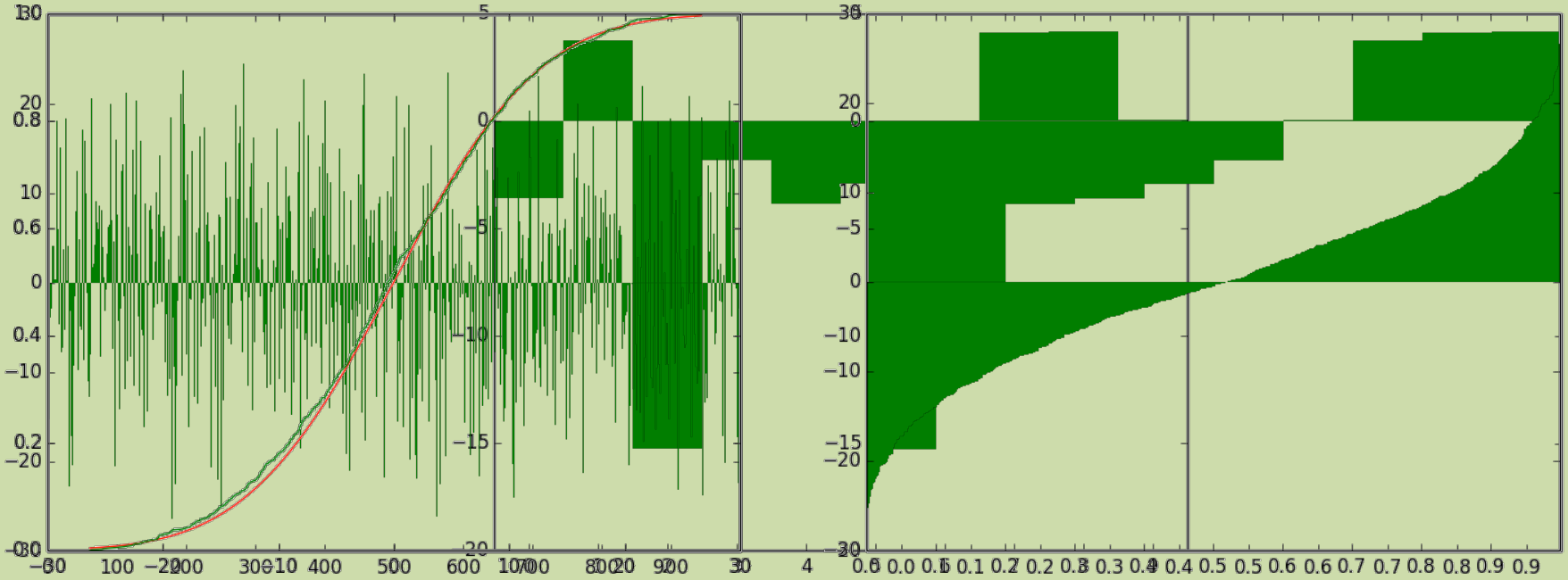

independent
identically
distributed
random
variables
0.03
4.18
4.12
-10.33
-2.92
-3.83
-1.82
-15.26
3.77
-3.58
©2009 FreeBigPictures.com

What do I want to know?

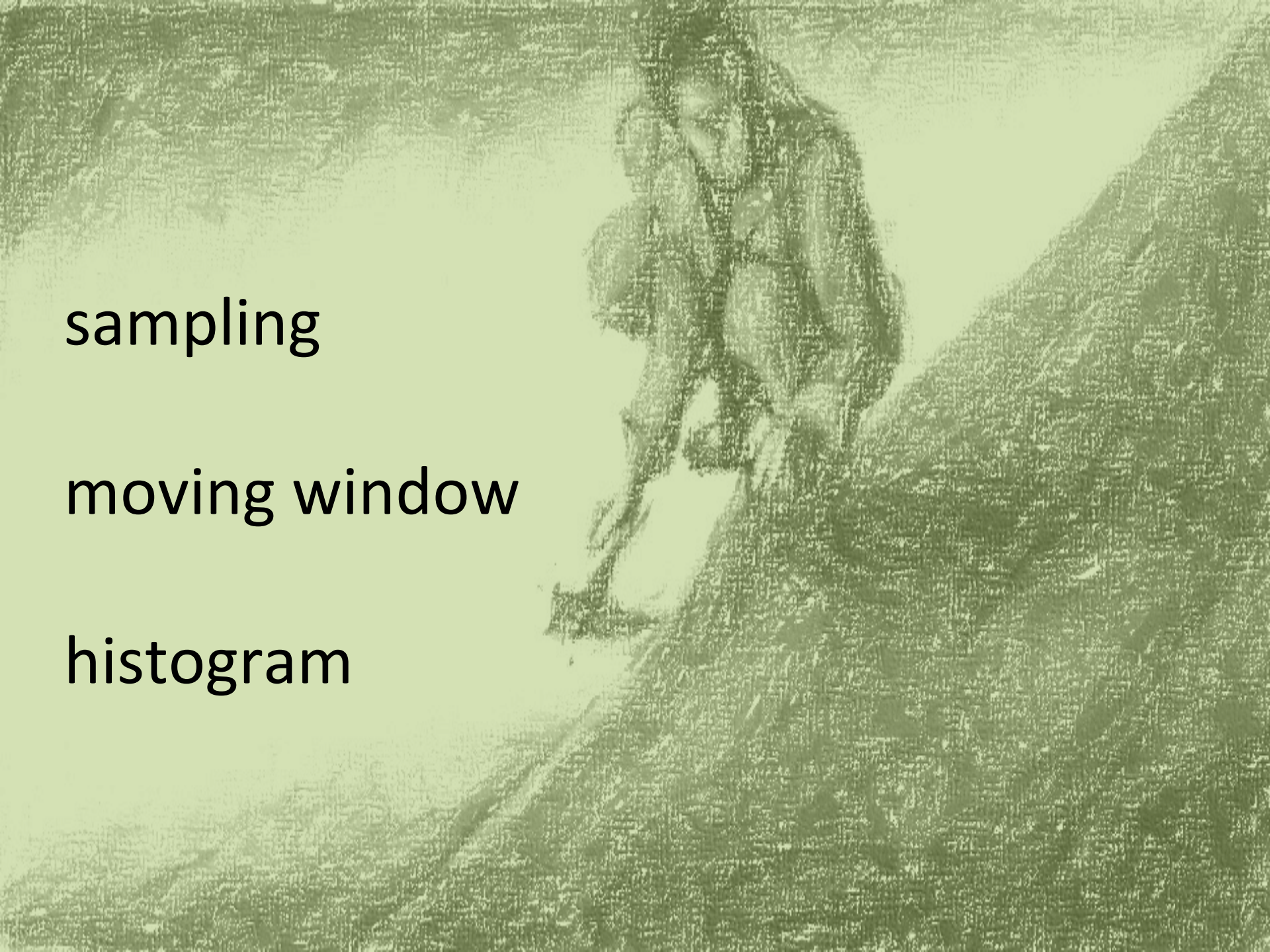
Moments? Quantiles? Distribution!

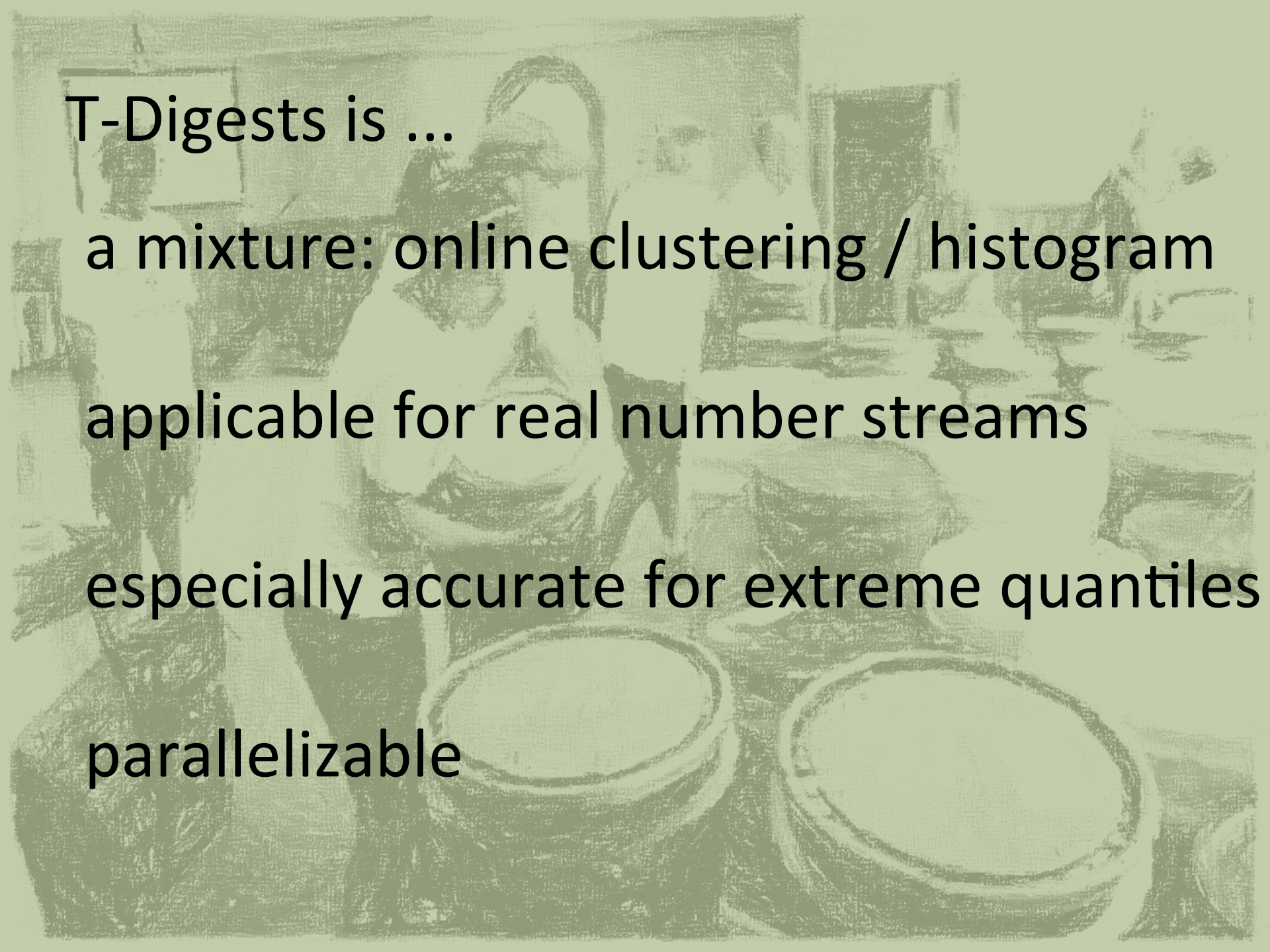
Lossless approach

-3.58, 3.77, -15.26, -1.82, -3.83, -2.92, -10.33, 4.12, 4.18, 0.03, ...




KEEP
CALM
AND
USE
ESS COMPRE


sampling
moving window
histogram



T-Digests is ...

a mixture: online clustering / histogram

applicable for real number streams

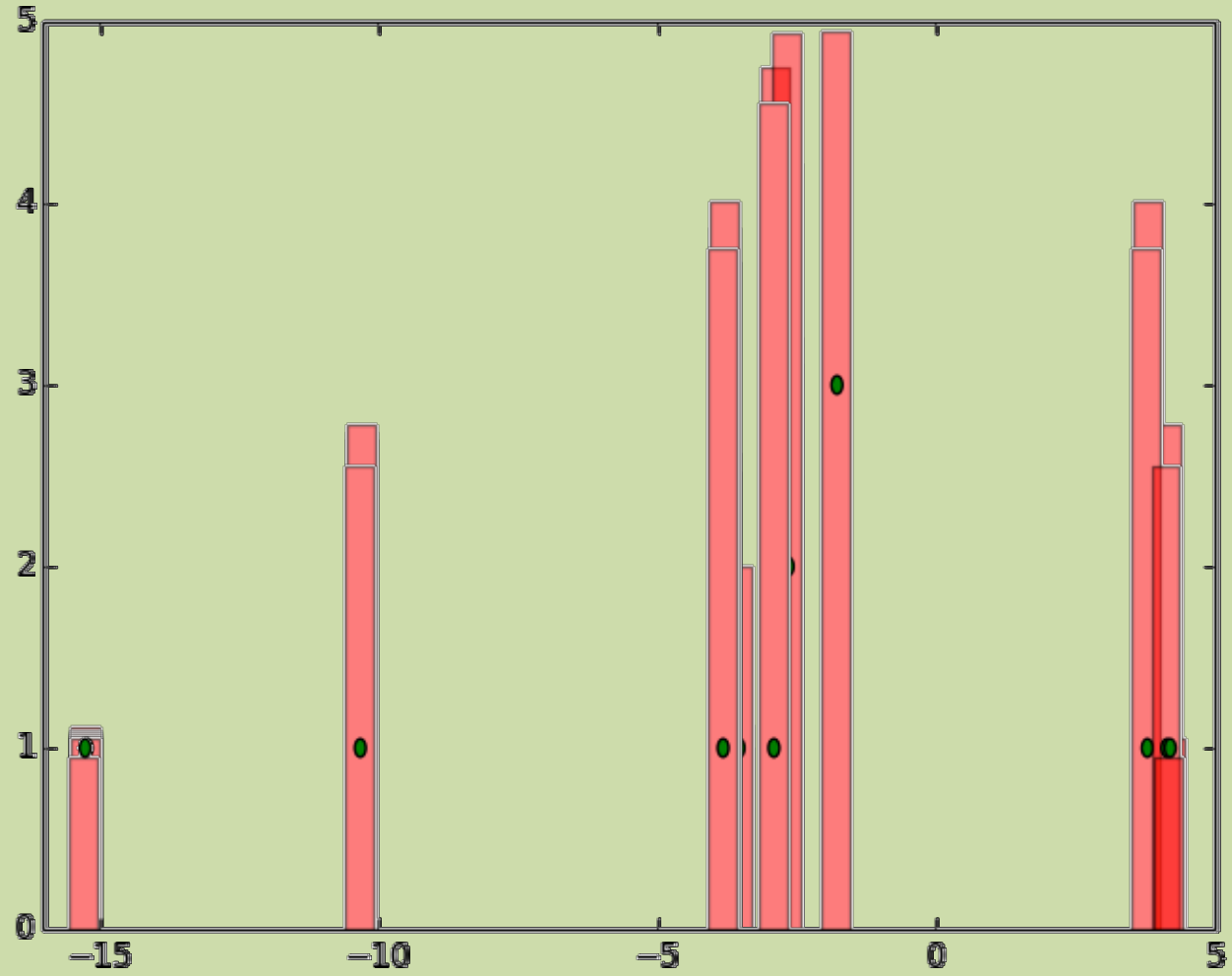
especially accurate for extreme quantiles

parallelizable

t-Digests step by step

-3.58, 3.77, -15.26, -1.82, -3.83, -2.92, -10.33, 4.12, 4.18, 0.03

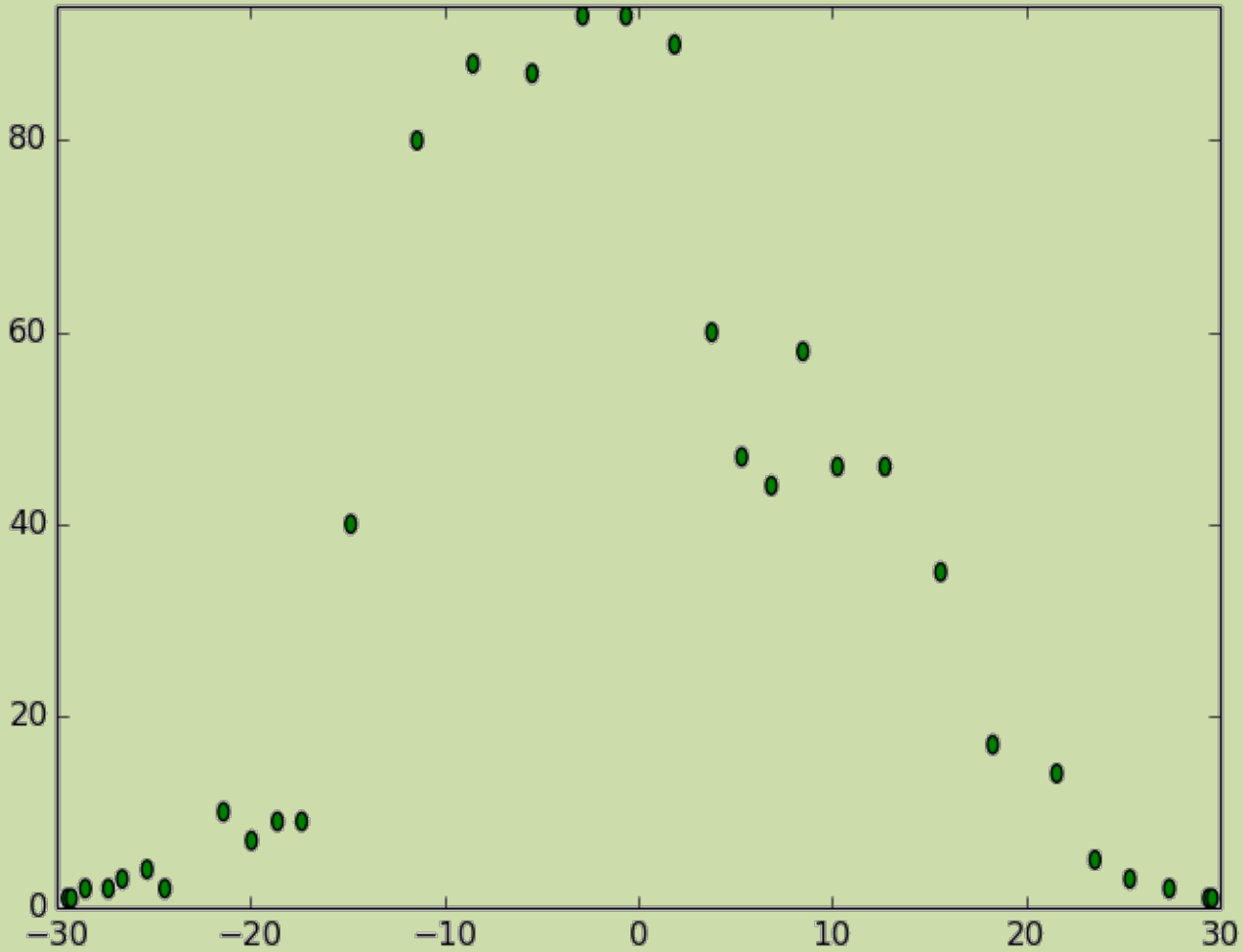
(-15.26, 1), (-10.33, 1), (-3.83, 1), (-3.58, 1), (-2.92, 1), (-1.82, 1), (3.77, 1), (4.12, 1), (4.18, 1)



t-Digests step by step

32 centroids

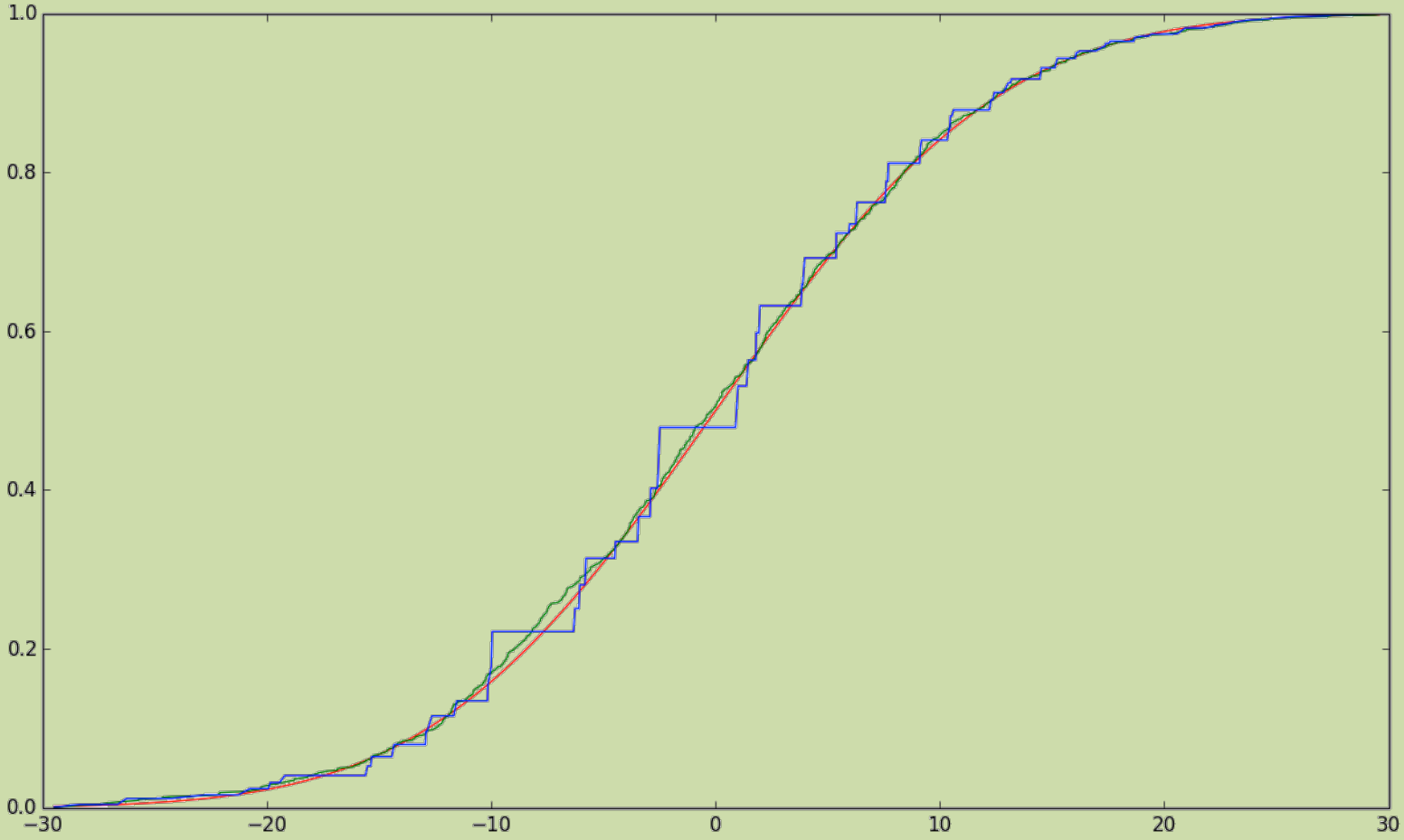
-3.58, 3.77, -15.26, -1.82, -3.83, -2.92, -10.33, 4.12, 4.18, 0.03, ...



t-Digests step by step

32 centroids

-3.58, 3.77, -15.26, -1.82, -3.83, -2.92, -10.33, 4.12, 4.18, 0.03, ...



Capacity of a centroid c

$$4 * n * \delta * q(c) * (1 - q(c))$$

free parameter

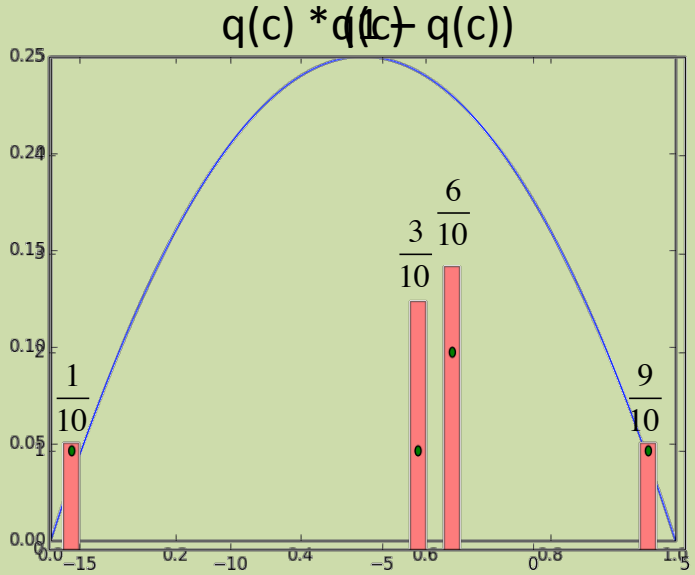
number of sample

quantile of centroid

Quantile of centroid c

$$q(c) = \frac{\frac{1}{2} c.count + \sum_{d:d.mean<c.mean} d.count}{\sum_{d} d.count}$$

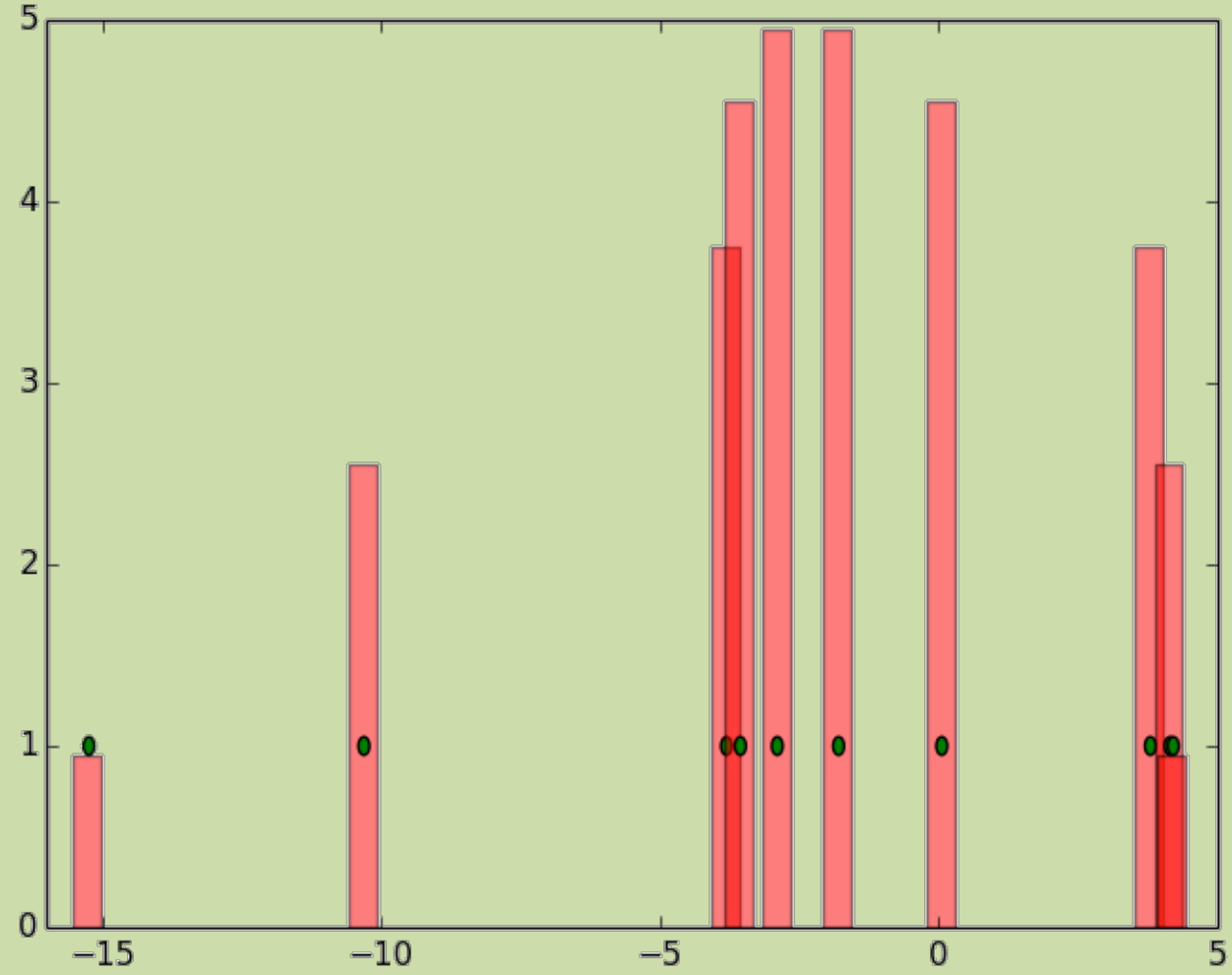
$$\Rightarrow q(c) \in (0,1)$$



t-Digests with ordered input

-15.26, -10.33, -3.83, -3.58, -2.92, -1.82, 0.03, 3.77, 4.12, 4.18

(-15.26, 1), (-10.33, 1), (-3.83, 1), (-3.58, 1), (-2.92, 1),
(-1.82, 1), (0.03, 1), (3.77, 1), (4.12, 1), (4.18, 1)



t-Digests with ordered input: shuffle!

(-15.26, 1), (-10.33, 1), (-3.83, 1), (-3.58, 1), (-2.92, 1),
(-1.82, 1), (0.03, 1), (3.77, 1), (4.12, 1), (4.18, 1)

10 centroids

(-3.58, 1), (3.77, 1), (-15.26, 1), (-1.82, 1), (-3.83, 1),
(-2.92, 1), (-10.33, 1), (4.12, 1) , (4.18, 1), (0.03, 1)

8 centroids

shuffling typically reduces number of centroids by 20% - 40% without loss of accuracy

taking a centroid list as input could cause weights <> 1

Algorithm 1: Construction of a t-Digest

Input: Ordered set of weighted centroids $C = \{\}$, sequence of real-valued, weighted points $X = \{(x_1, 1), \ldots (x_N, 1)\}$, and accuracy tolerance δ

Output: final set $C = [c_1 \ldots c_m]$ of weighted centroids

for $(x_n, 1) \in X$ **:**
  $z = \min |c_i.\texttt{mean} - x|$;
  $S = \{c_i : |c_i.\texttt{mean} - x| = z \wedge c_i.\texttt{count} + 1 \leq 4n\delta q(c_i)(1 - q(c_i))\}$;
  while $S \neq \{\}$
    Sample $c_j \sim \text{Uniform}(S)$;

    $c_j.\texttt{count} \leftarrow c_j.\texttt{count} + 1$
    $c_j.\texttt{mean} \leftarrow c_j.\texttt{mean} + (x_n - c_j.\texttt{mean})/c_j.\texttt{count}$;
    $w_n \leftarrow w_n - 1$;

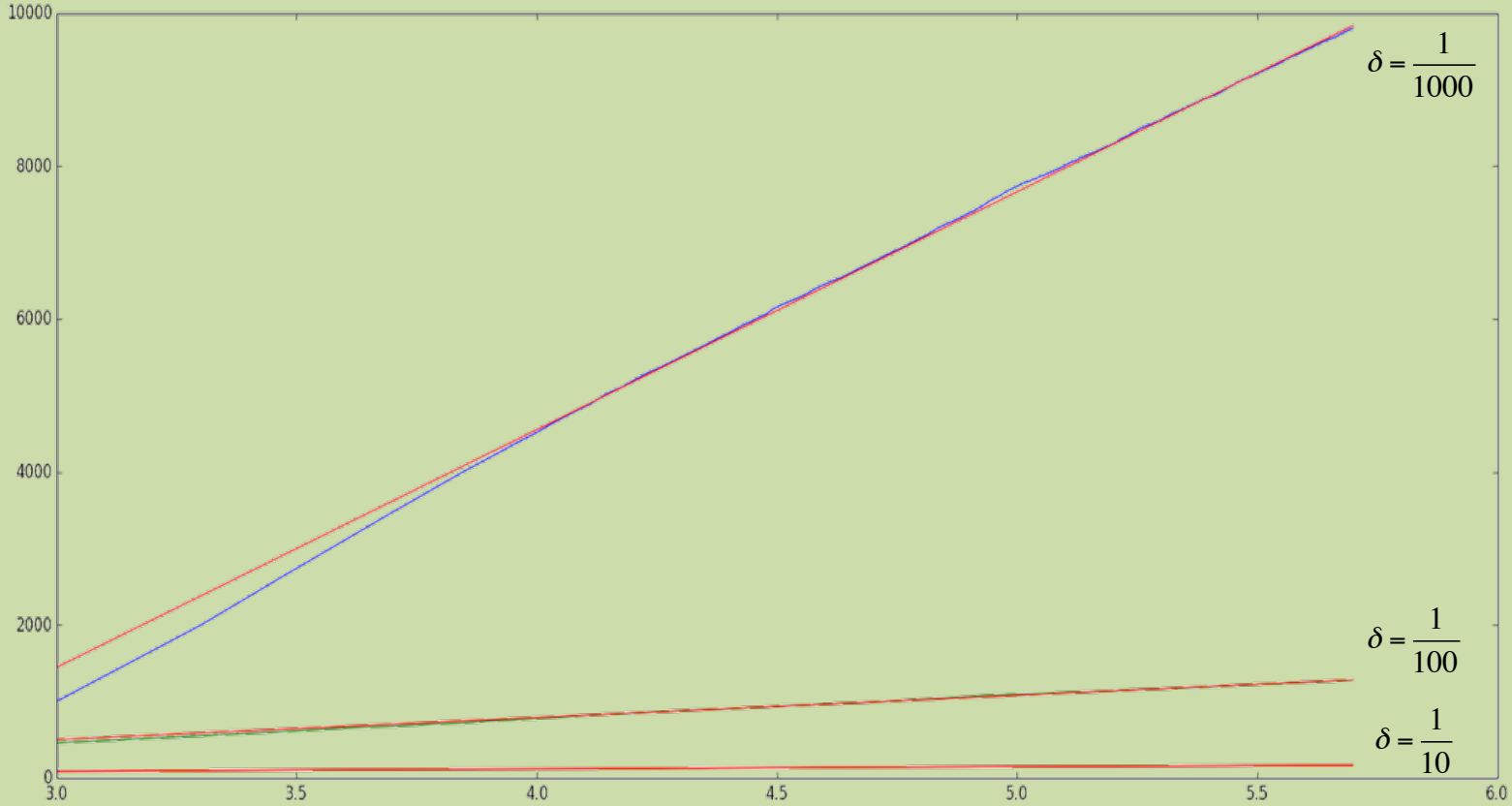
  if $w_n > 0$ **:**
    $C \leftarrow C + (x_n, 1)$;
  if $|C| > K/\delta$ **:**
    $C \leftarrow \texttt{cluster}(\texttt{permute}(C))$;
$C \leftarrow \texttt{cluster}(\texttt{permute}(C))$;
return C

special case:
input weights are always 1

taken from *https://github.com/tdunning/t-digest/blob/master/docs/t-digest-paper/histo.pdf*

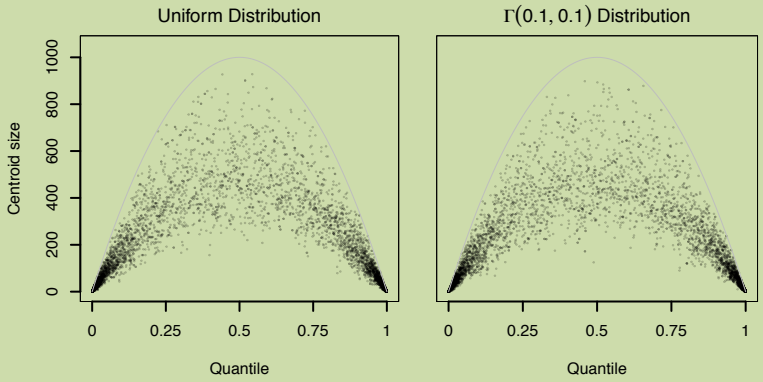
Number of centroids

- 500K samples
- x axis has exponential scale (-> straight line has logarithmic growth)

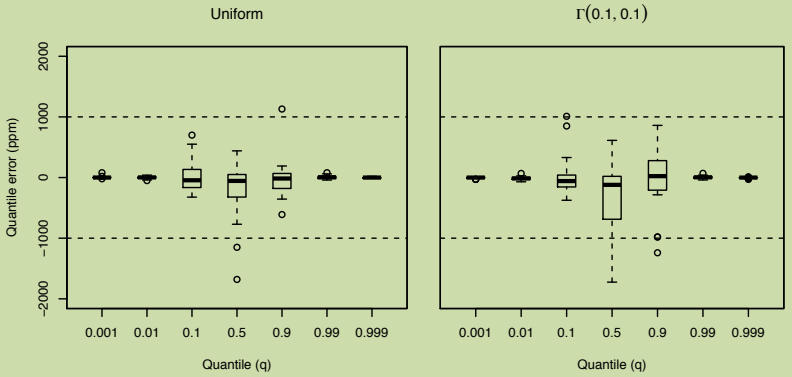


Estimation qualtiy for quantiles

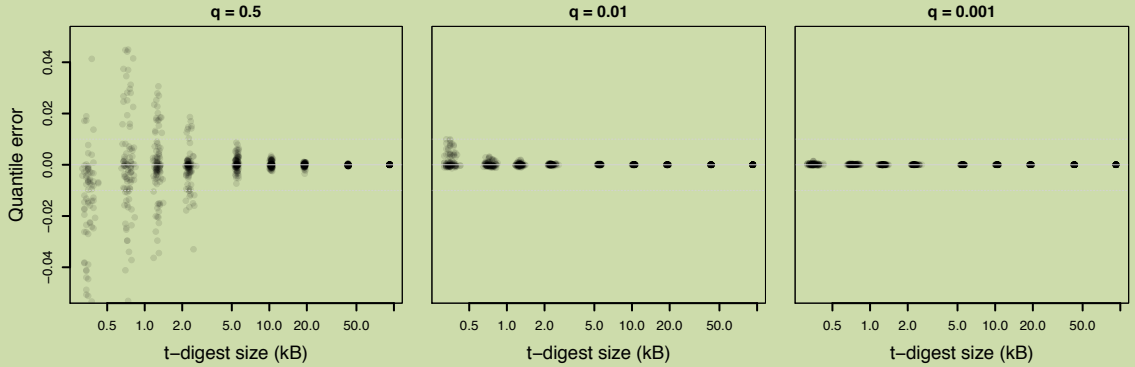
size of centroid



absolute quantile error



absolute quantile error



taken from *https://github.com/tdunning/t-digest/blob/master/docs/t-digest-paper/histo.pdf*

Ressources

Ted Dunning‘s GitHub rep: github.com/tdunning/t-digest

- research paper
- JAVA implementation

Integrated in

- Mahout
- ElasticSearch

Python implementation

- Trademob‘s working on open source package


When?
Where?
Who?
What?
Why?
Questions & Answers
Yeah, Maybe