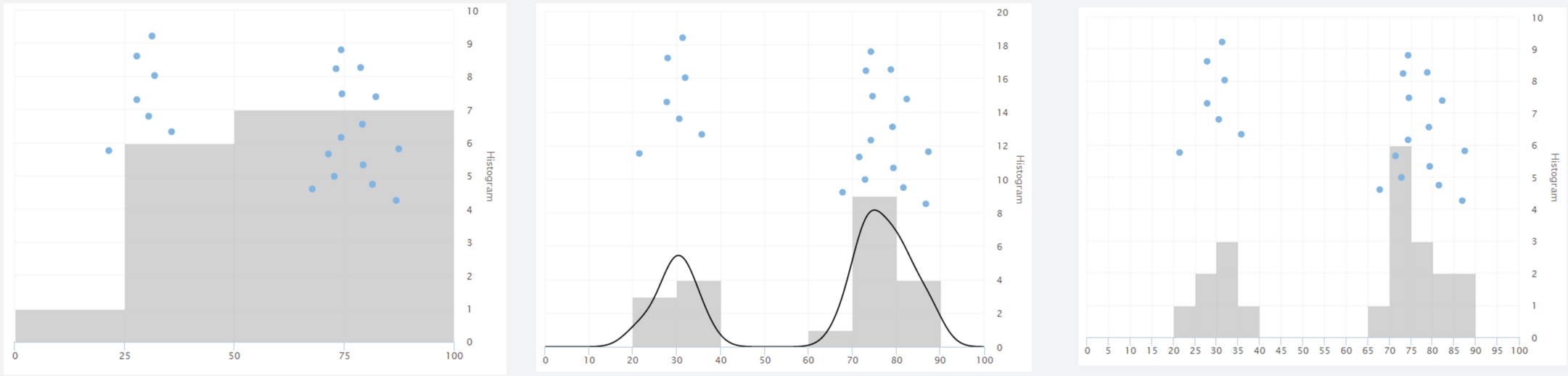
Kernel Density Estimation

http://kde.tume-maailm.pri.ee

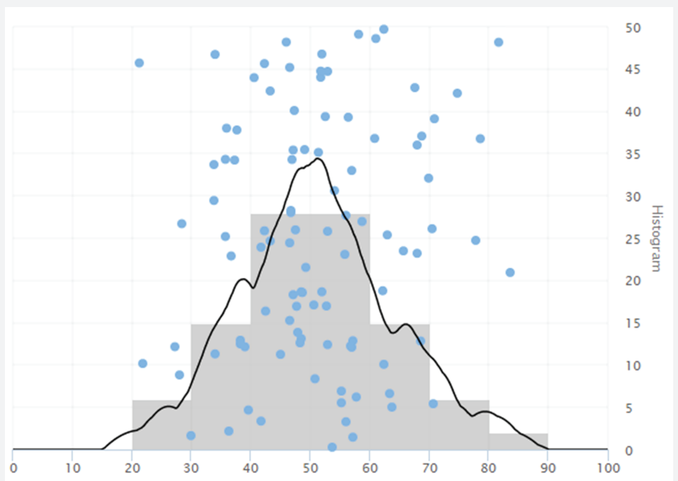
Data collected from some events can be described in many ways. Usually it has some underlying distribution that data scientists are interested in finding out. Given a finite number of samples from the data, histogram is one of the basic methods to describe what the shape of the distribution might be. Unfortunately there are cases where incorrect bin width or shift results in misrepresentation of the data.



Another way to guess the underlying distribution is to do kernel density estimation on the data (black line on the middle chart above). This approach uses one of several kernels in order to create a continuous function that will resemble the unknown distribution and thus describes the event in question more accurately.

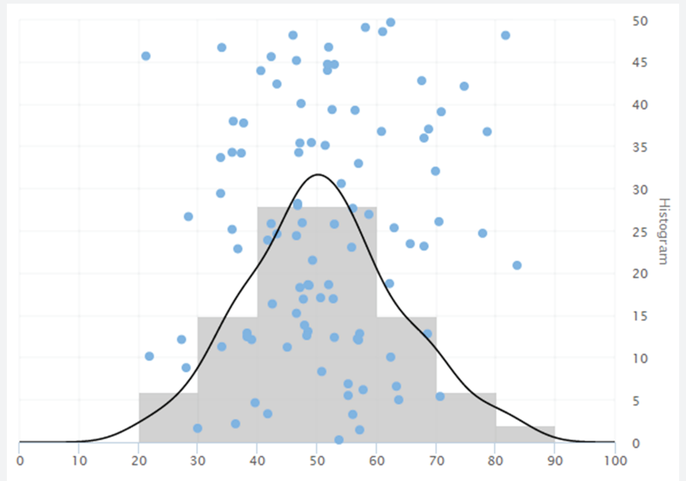
We created an online tool that demonstrates this technique and allows to investigate the results on either user inputted data or on samples from one of several implemented distributions. Below are examples of that using three of the most common kernel functions.

Epanechnikov



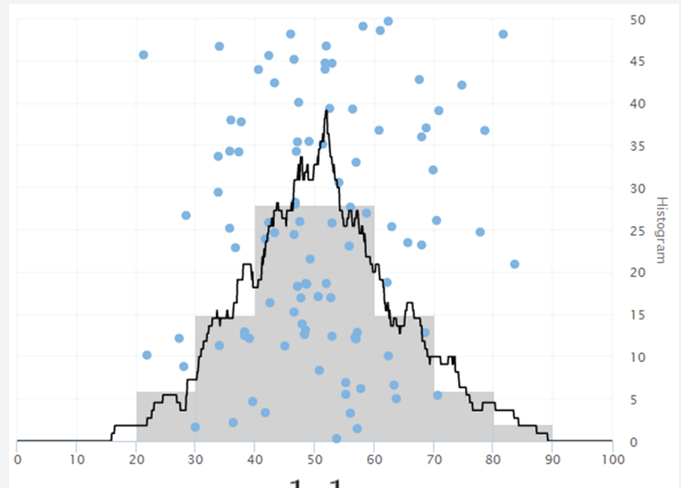
$$K(u) = \frac{1}{s}\frac{3}{4}(1-u^2)\cdot I(|u| \le 1)$$

Gaussian



$$K(u) = \frac{1}{s}\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}$$

Uniform



$$K(u) = \frac{1}{s}\frac{1}{2}\cdot I(|u| \le 1)$$

Where *u* is distance from kernel center to a neighbouring data point divided by the scale parameter *s* of the kernel.

Authors

Raimond Tunnel
Master's curriculum of Computer Science
Institute of Computer Science
Faculty of Mathematics and Computer Science
University of Tartu

Lauri Hämarik
Bachelor's curriculum of Physics
Institute of Physics
Faculty of Science and Technology
University of Tartu

Project is managed in Assembla.com http://assembla.com/spaces/kernel-density-estimation

TARTU ÜLIKOOL · UNIVERSITAS TARTUENSIS · 1632

Study IT in .ee