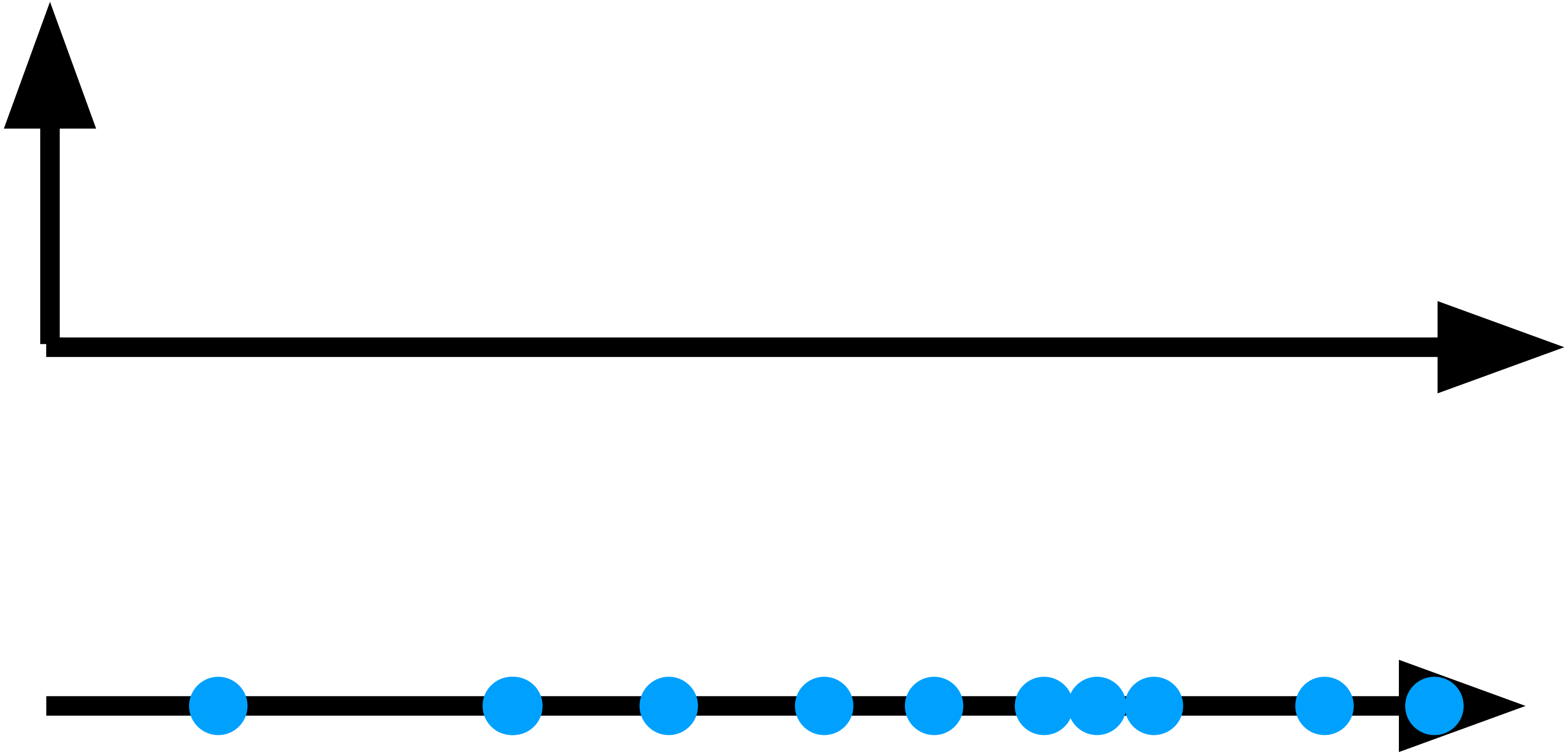
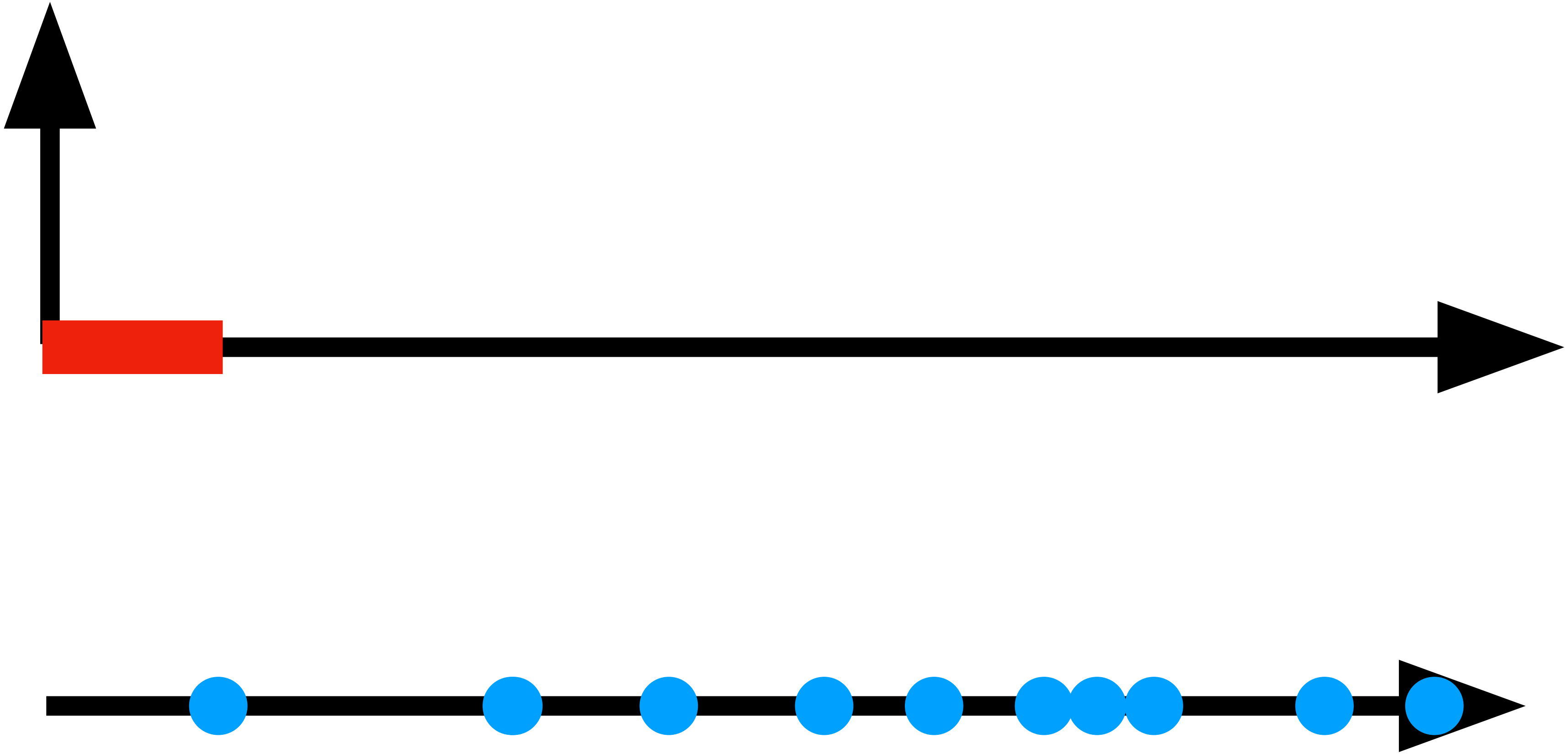
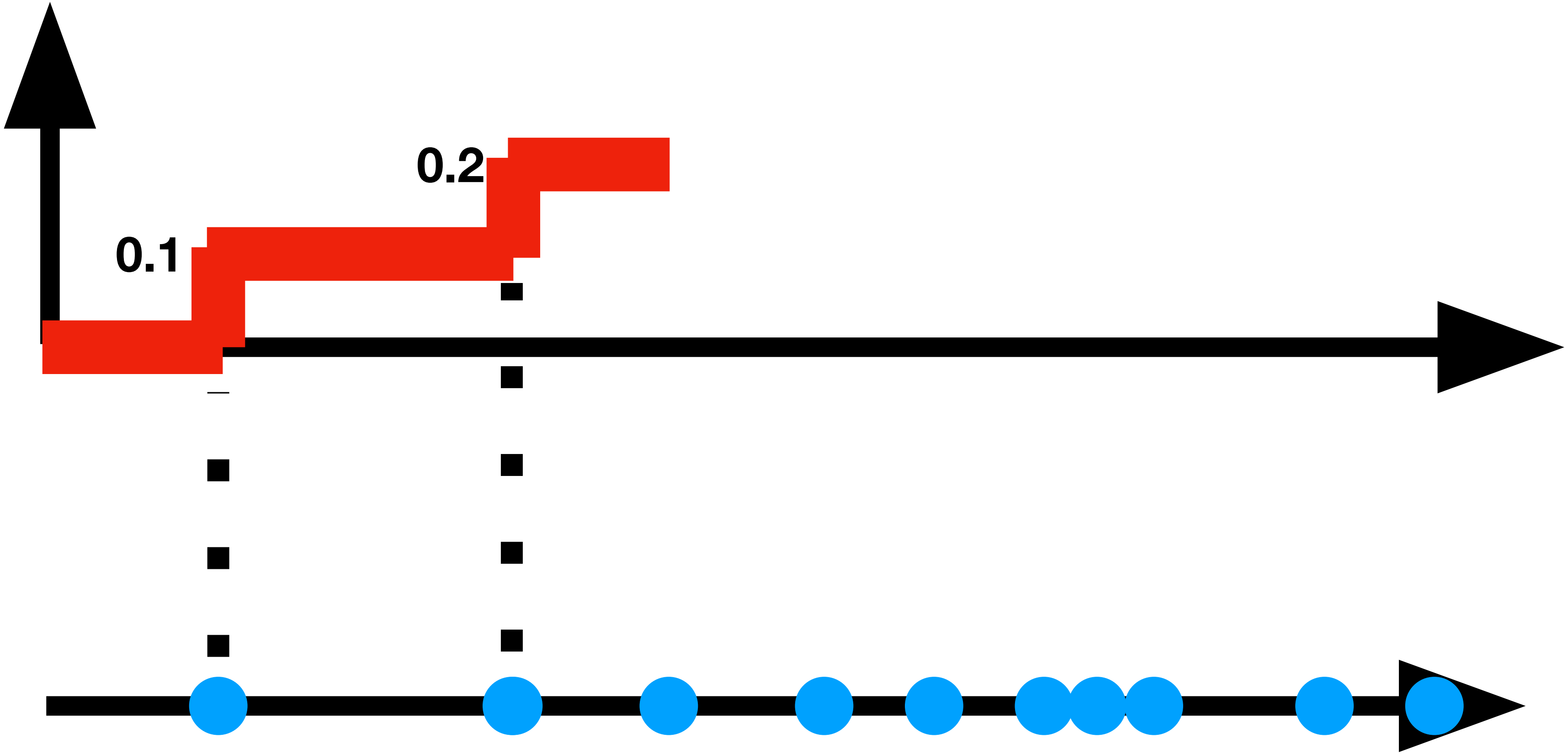
# Data Visualization
# W7-2

# Quiz

- What do you find interesting in today's VotW?
- Box plot vs. Histogram
  - You want to understand your data distribution and you have no idea how your data is distributed. Which method would you want to use first to examine the data? Why?
  - You have a dataset that documents the distribution of individual yearly soda consumption across several regions. You are mainly interested in comparing these regions. Which method do you want to use and why?
- Draw an empirical CDF (cumulative distribution function) of the following data: [-3, 5, 6, 6, 10]



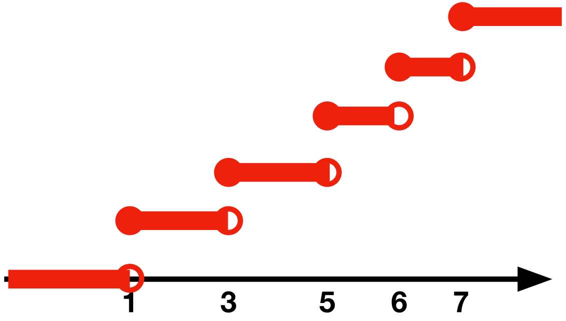


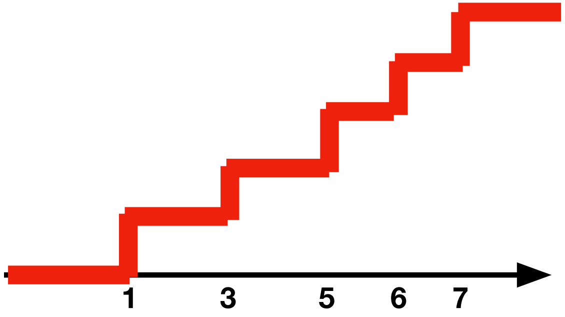

0.1
0.2

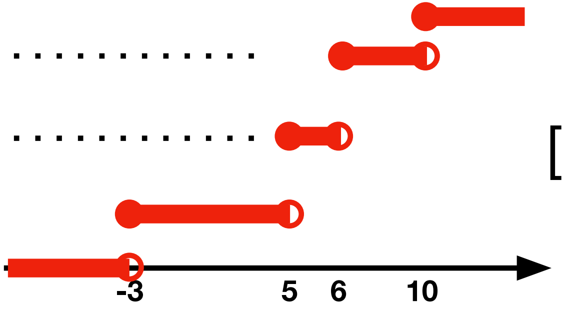
# CDF (Cumulative Distribution Function)

[1, 3, 5, 6, 7]

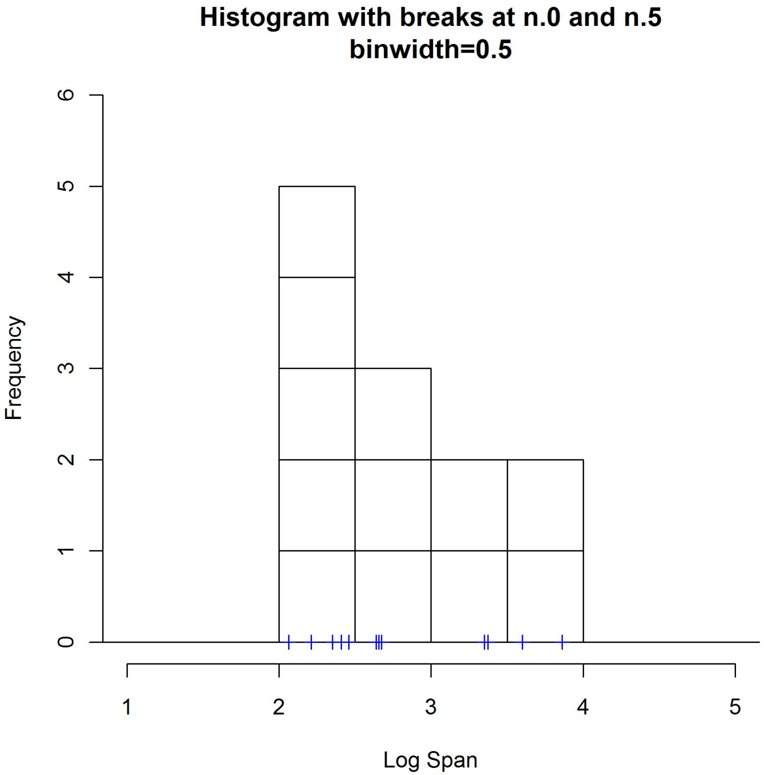
$$F_X(x) = P(X \leq x)$$

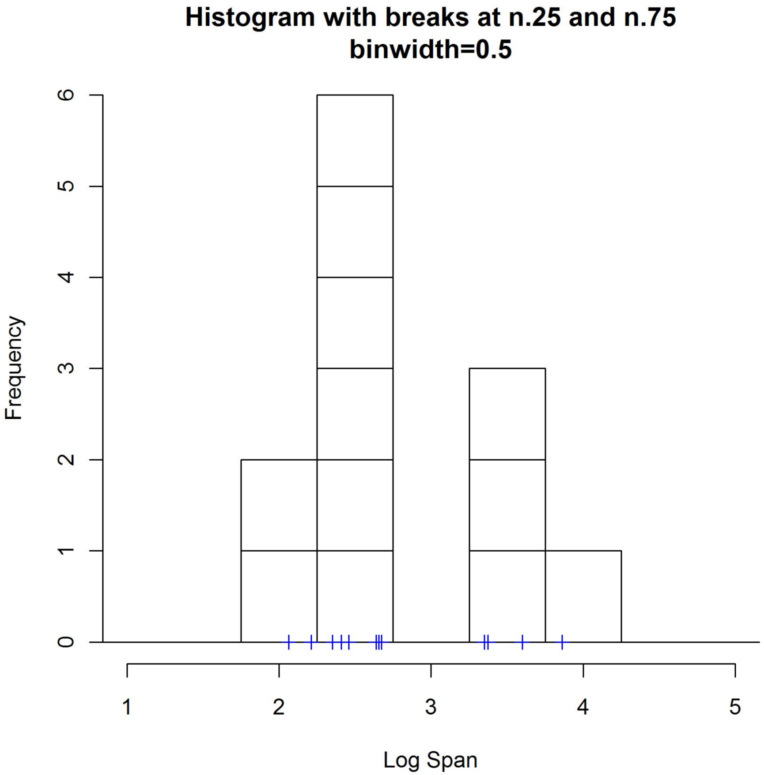




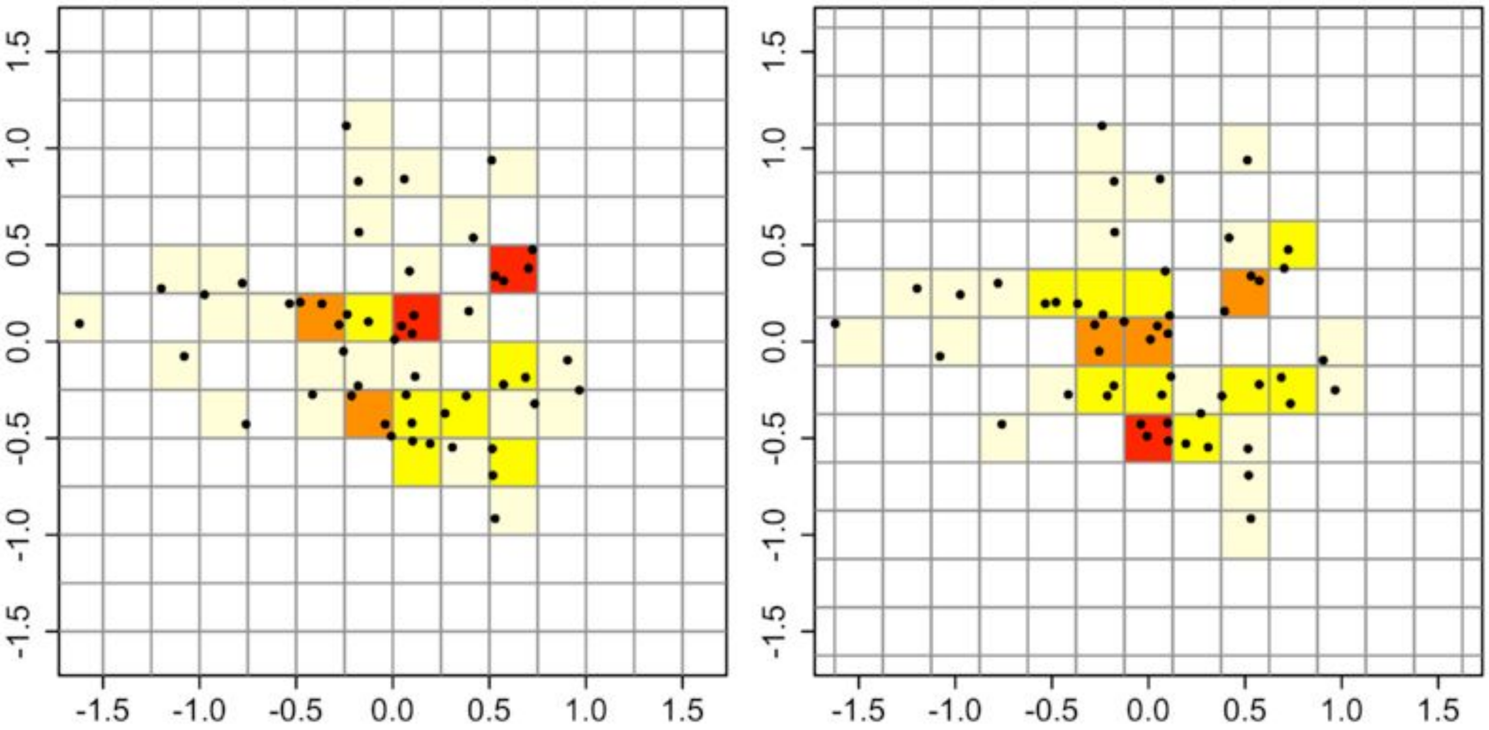


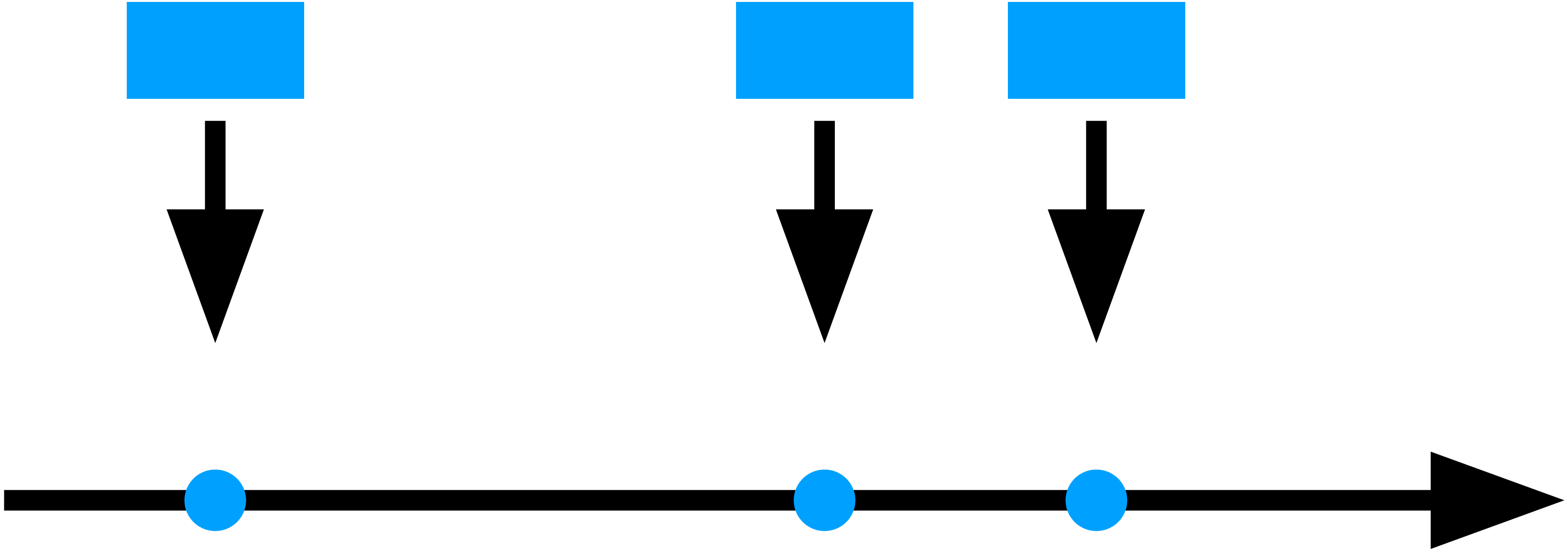
[-3, 5, 5, 6, 6, 10]

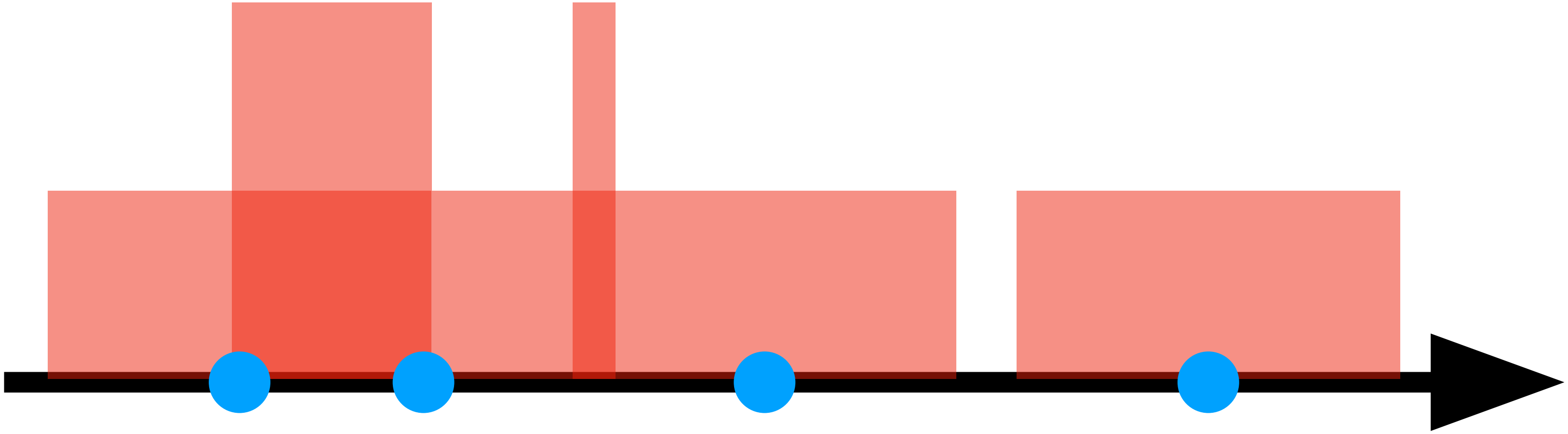


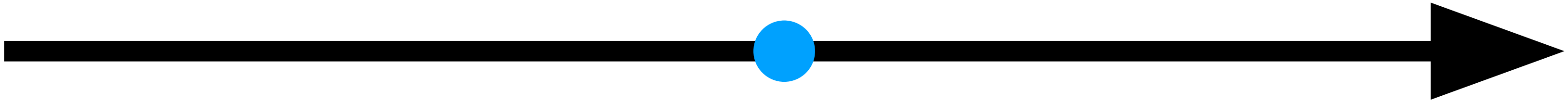


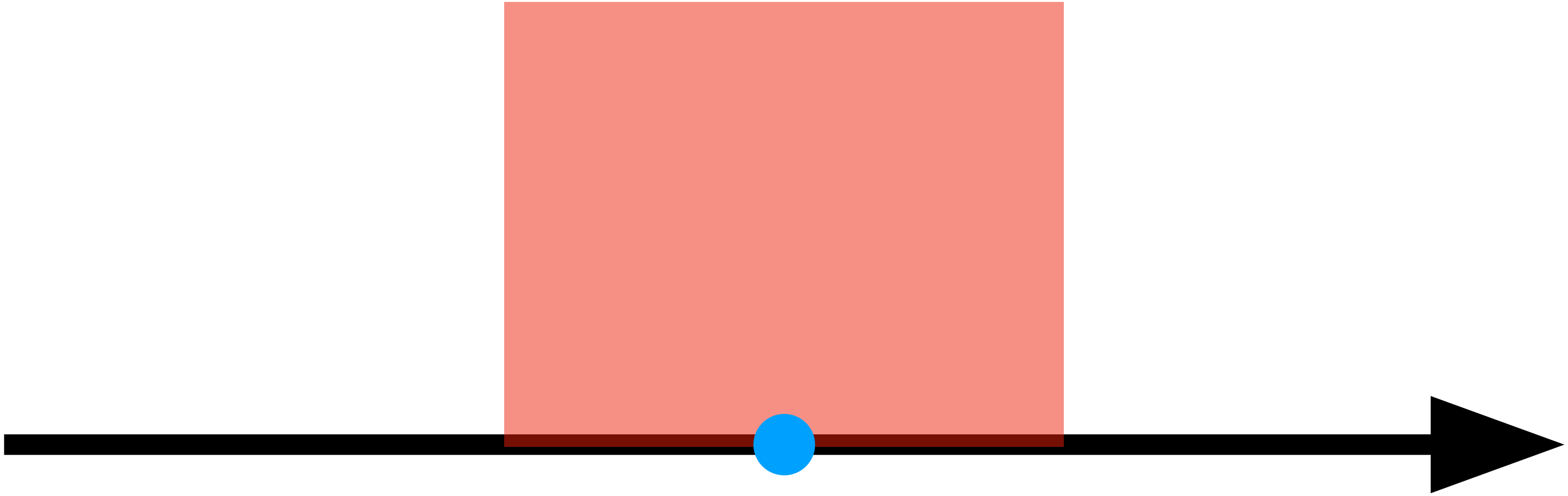
Tarn Duong, An introduction to kernel density estimation, http://www.mvstat.net/tduong/research/seminars/seminar-2001-05/

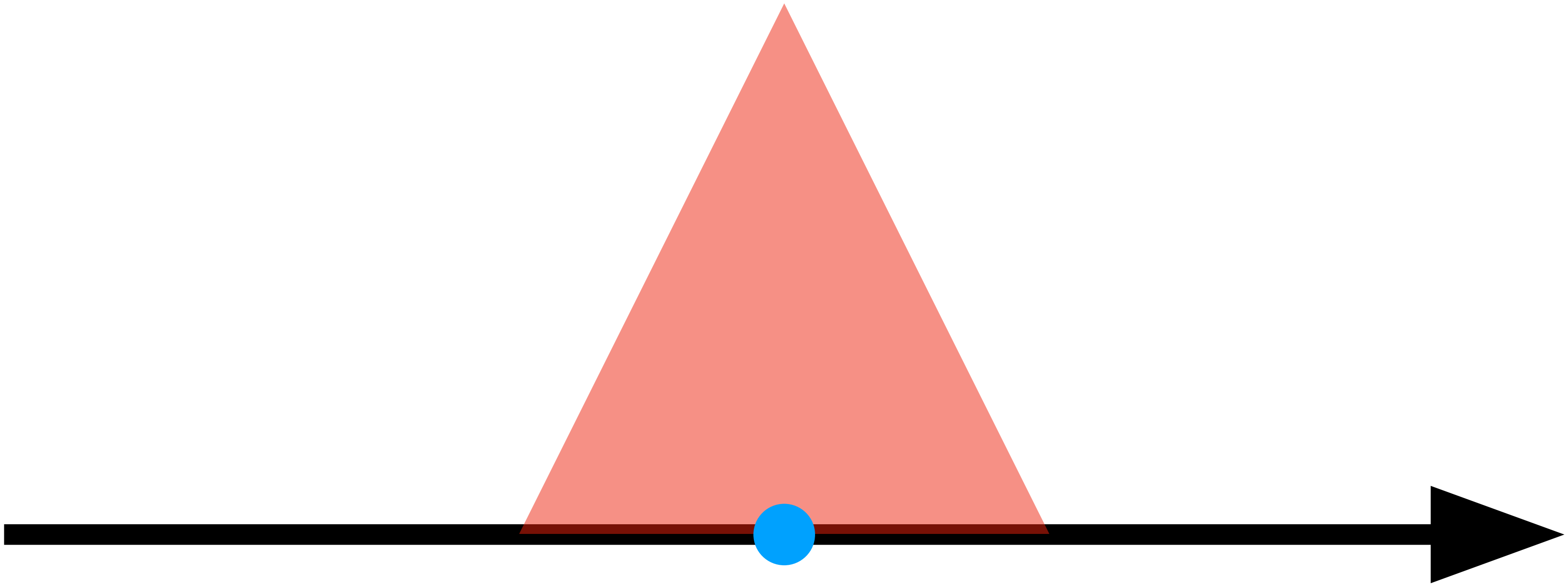

1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5

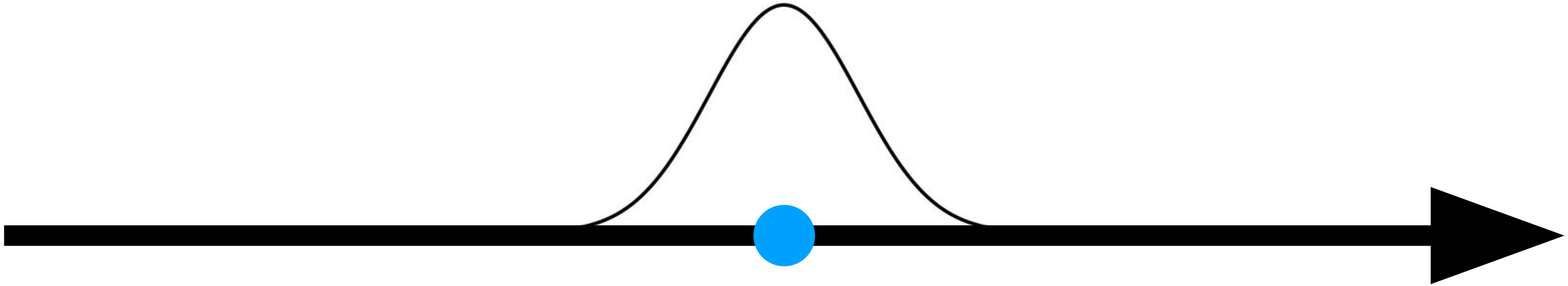


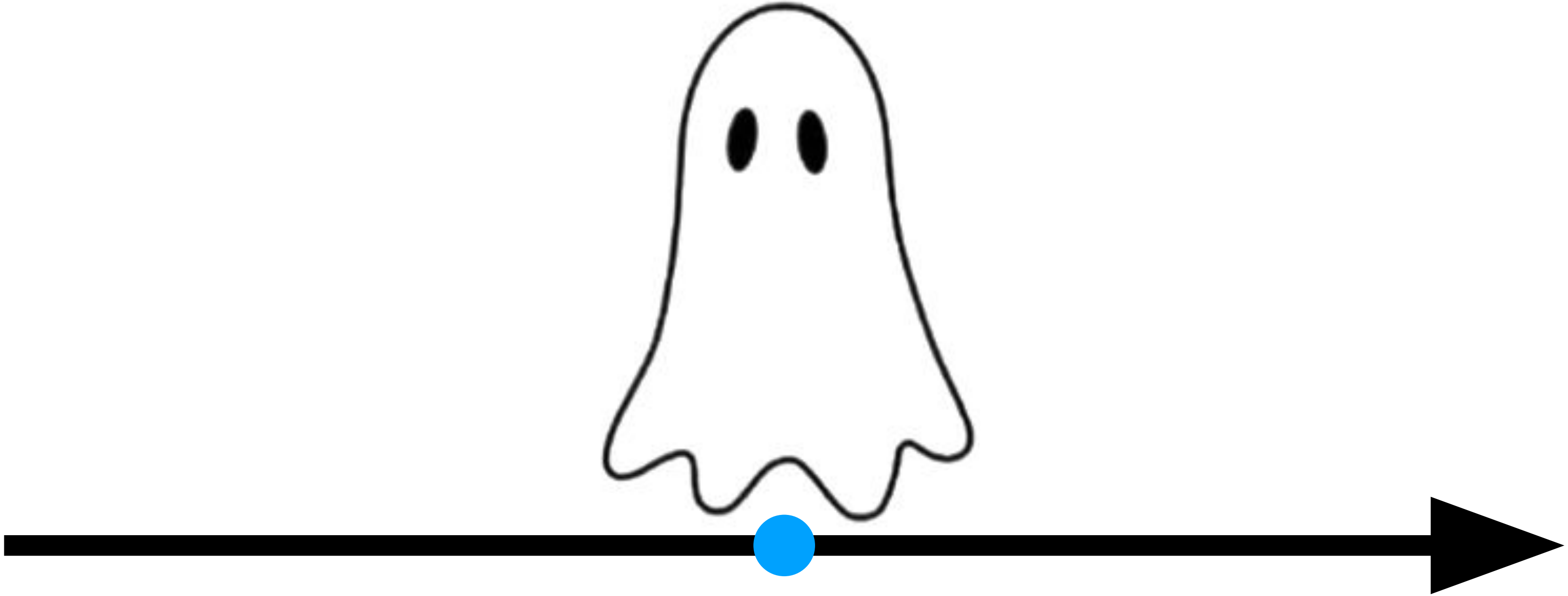




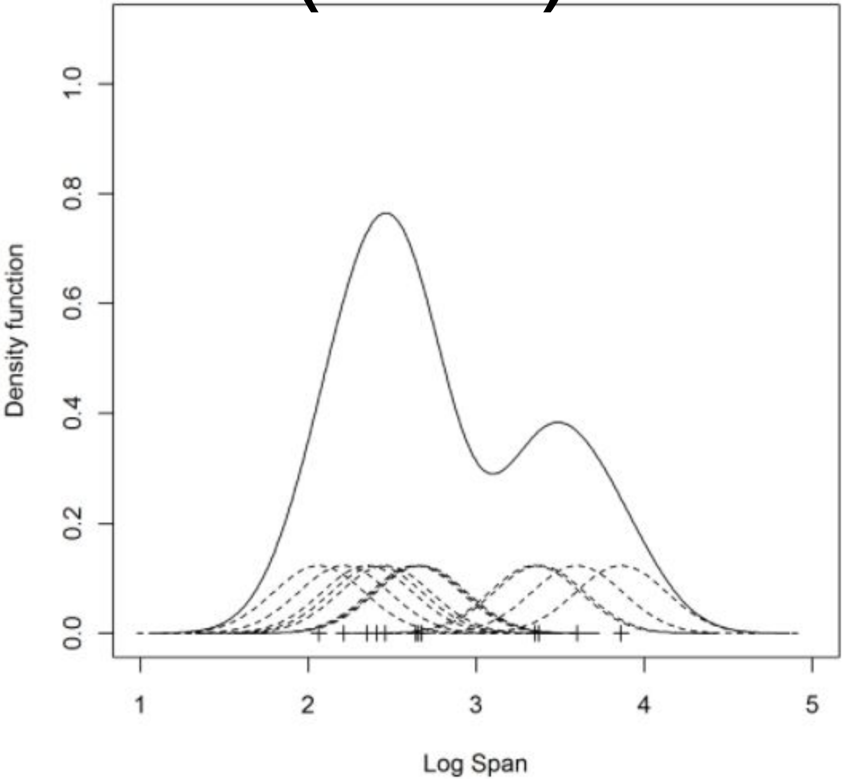






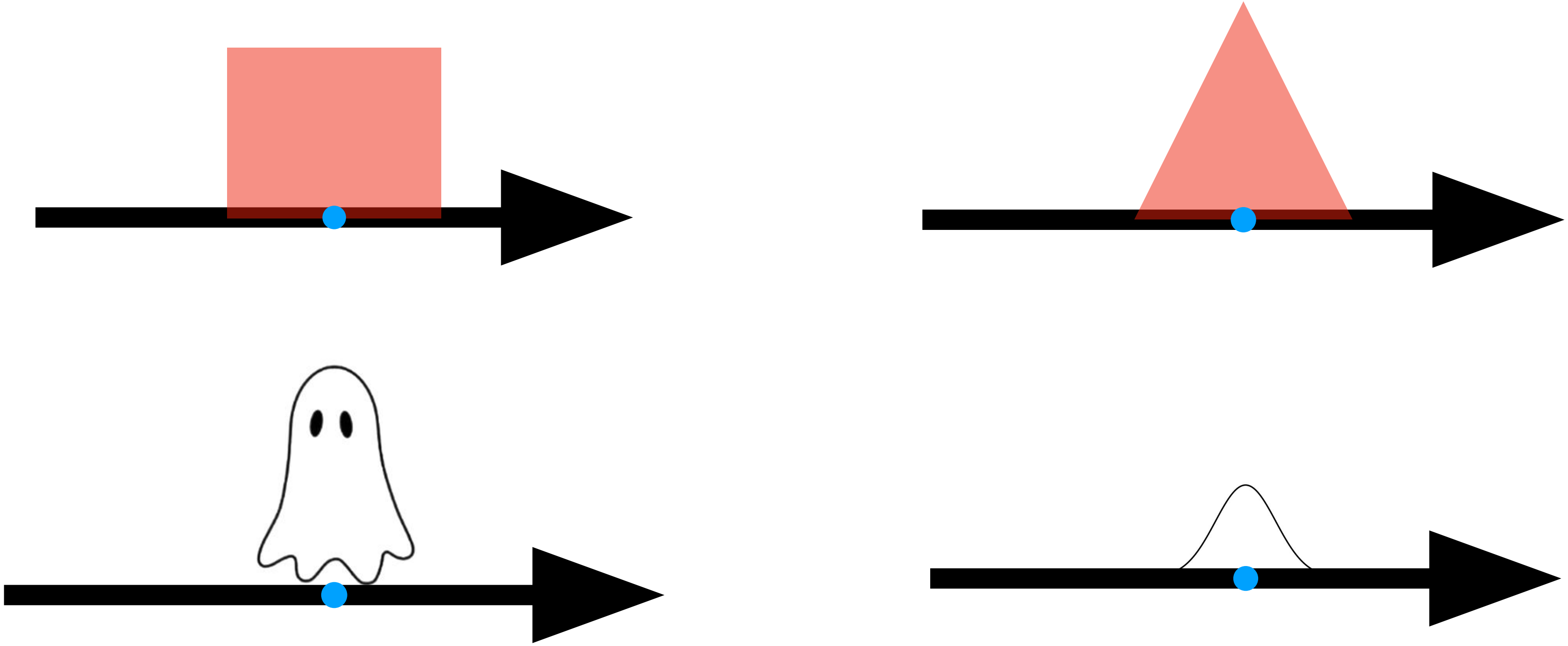


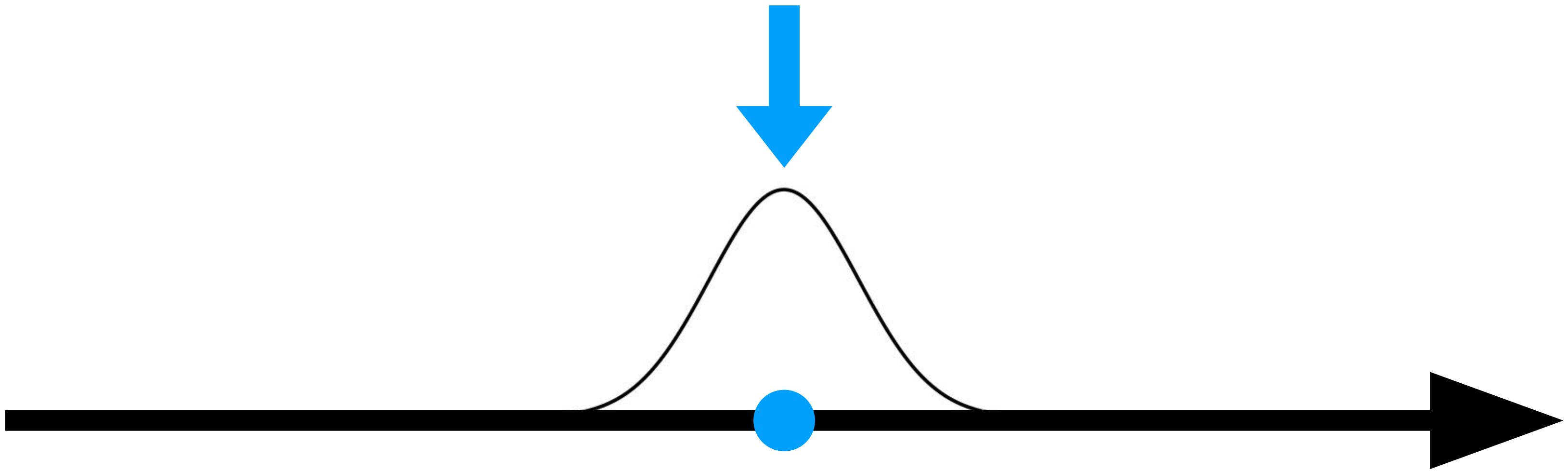
# **K**ernel **D**ensity **E**stimation (KDE)



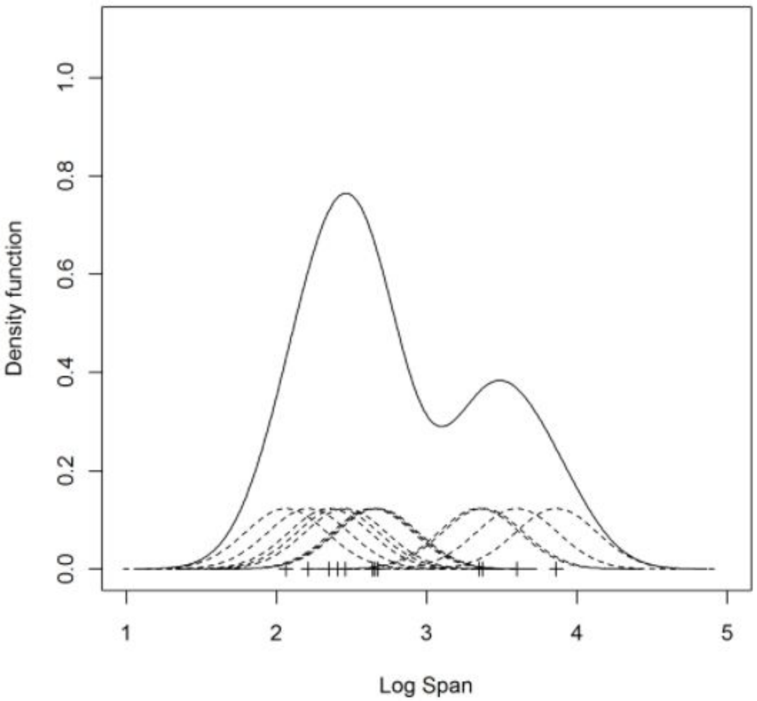
Tarn Duong, An introduction to kernel density estimation, http://www.mvstat.net/tduong/research/seminars/seminar-2001-05/

# "Kernels"



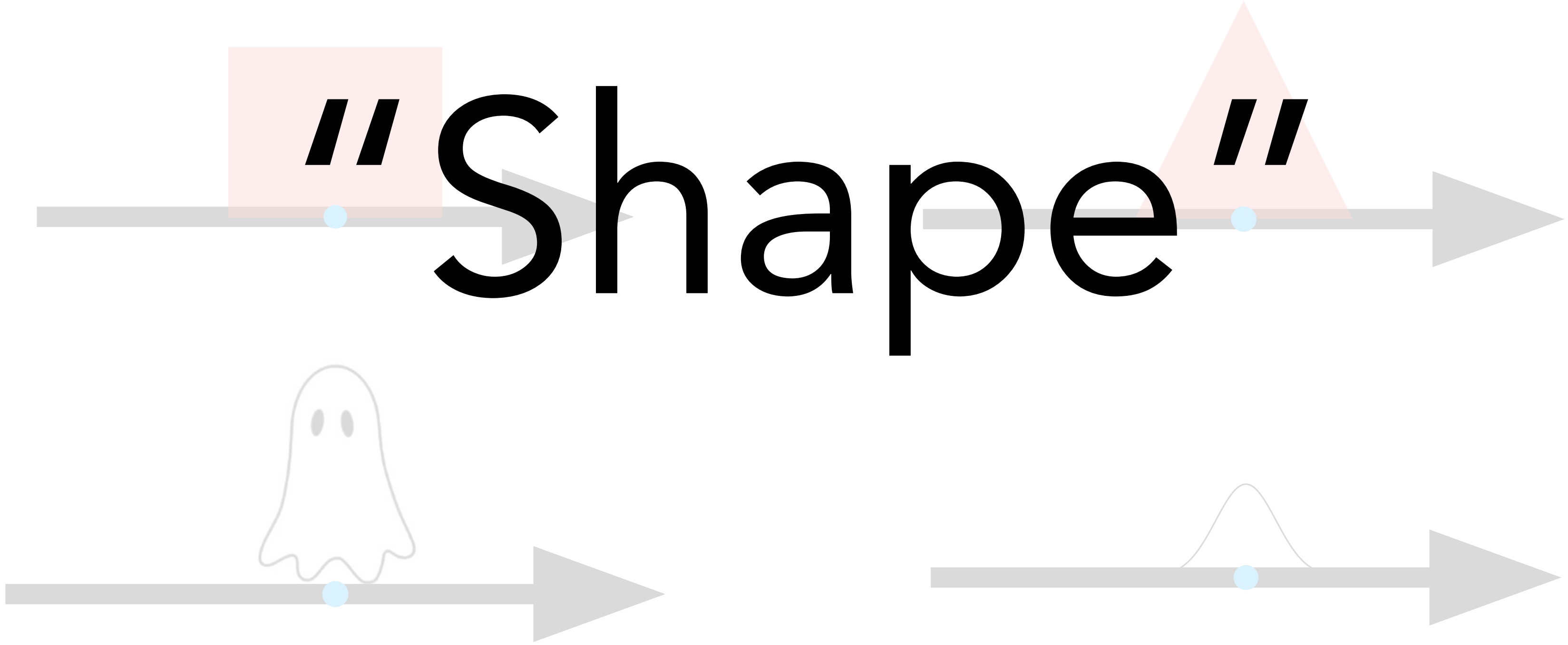


# Choices?

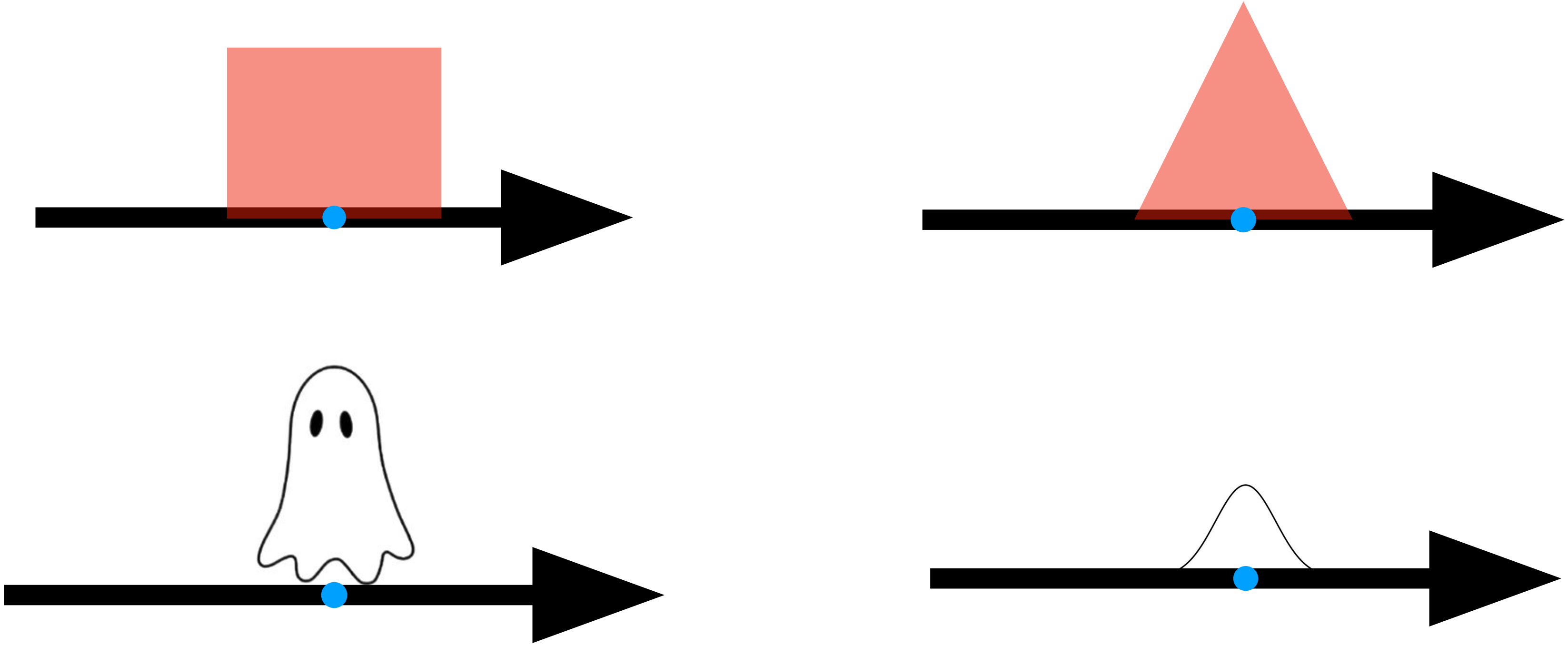


# “Kernels”

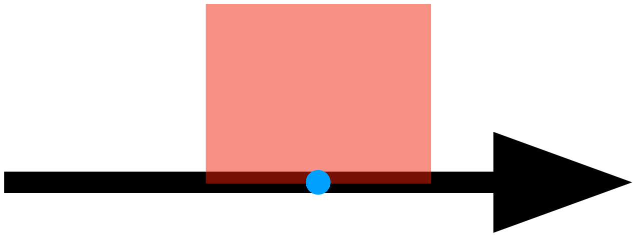
# “Shape”

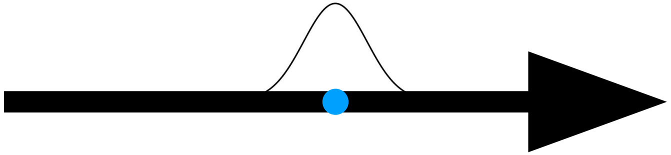


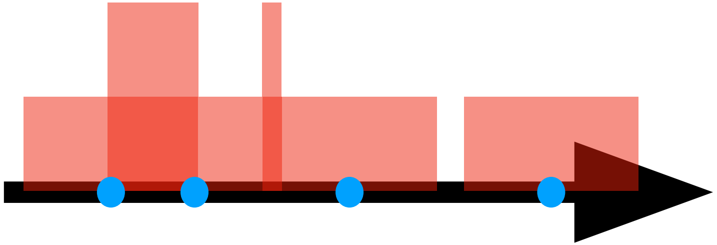
# "Kernels"

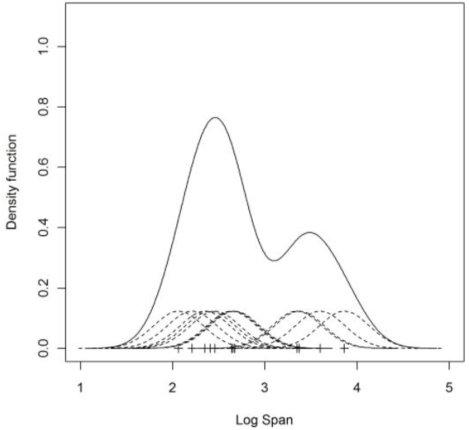


# Efficiency vs. Smoothness

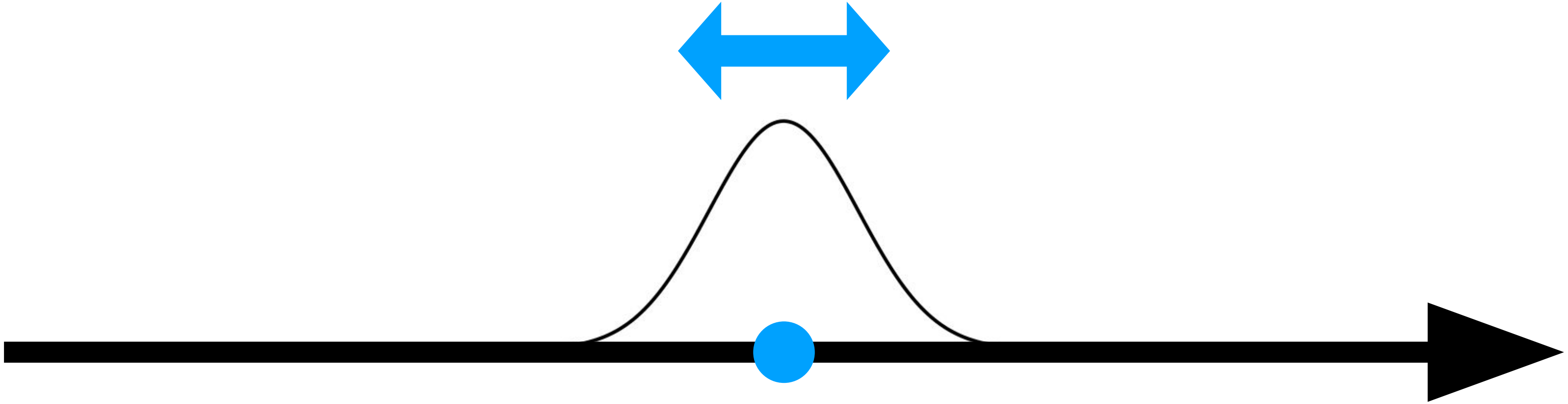


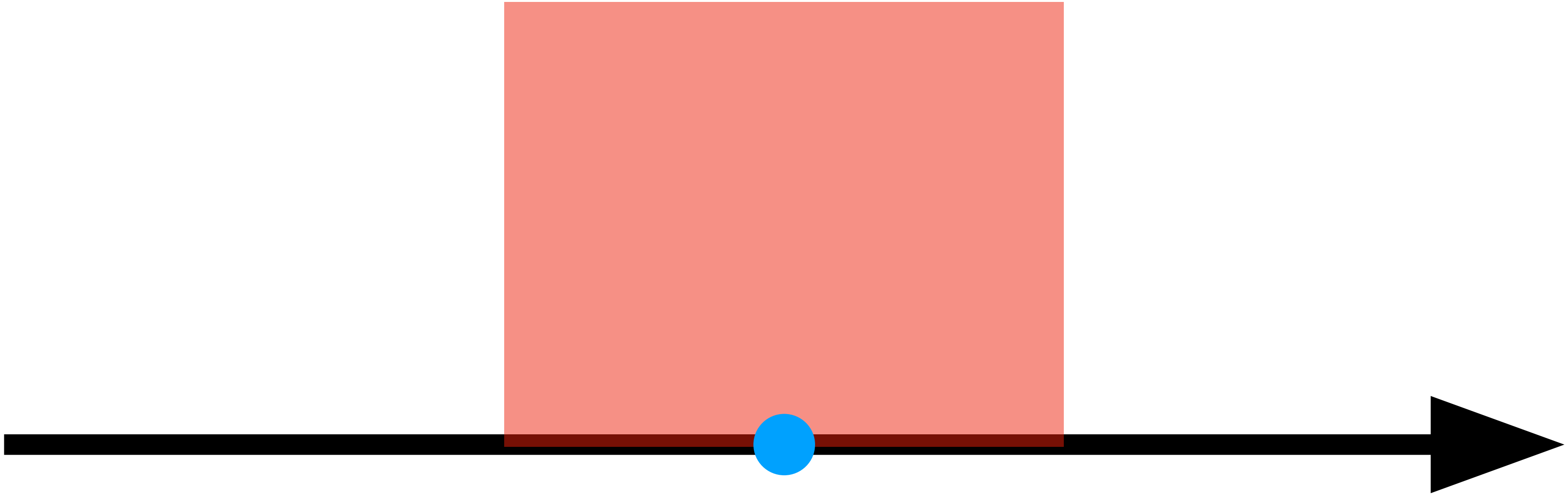


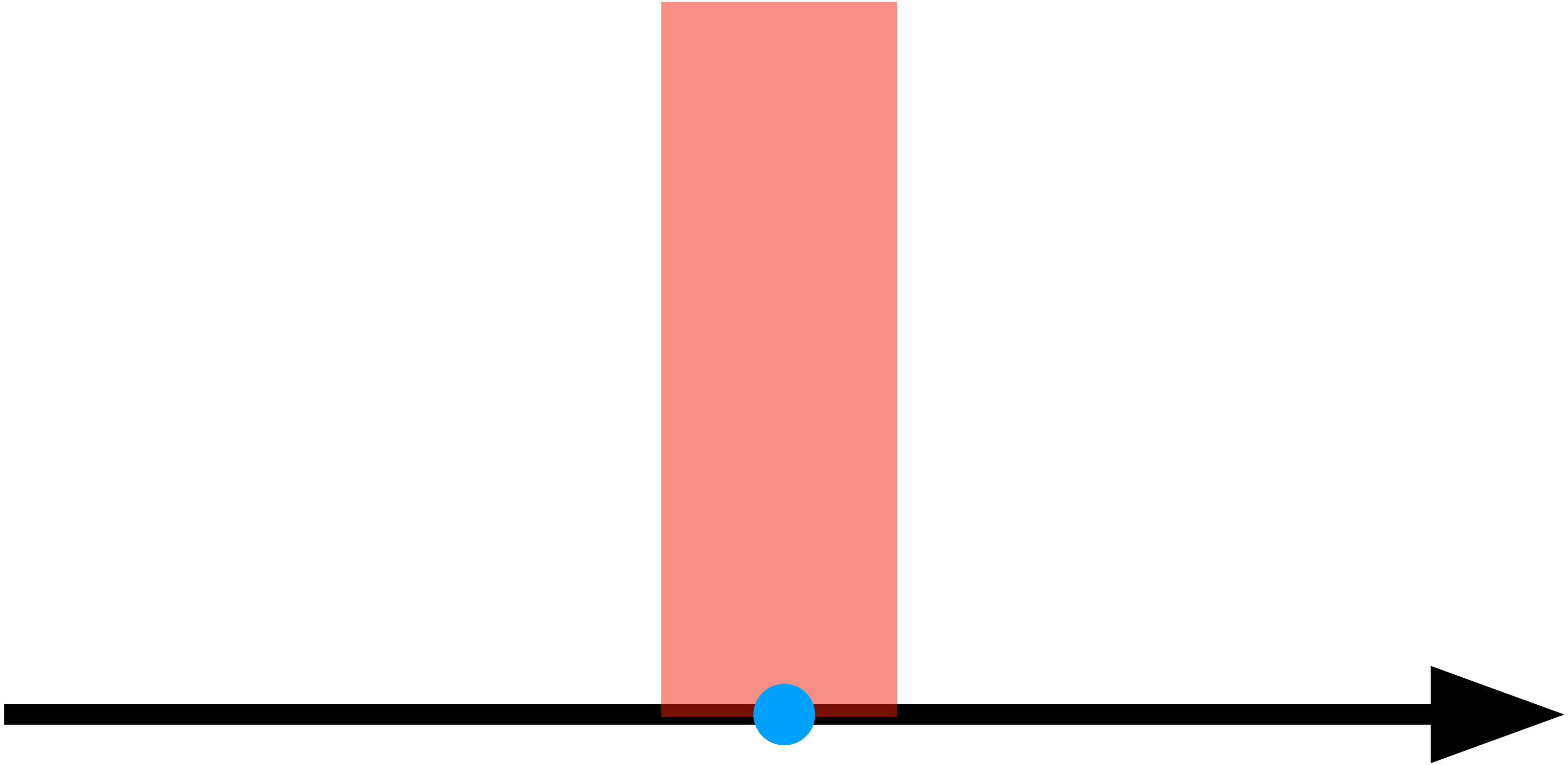


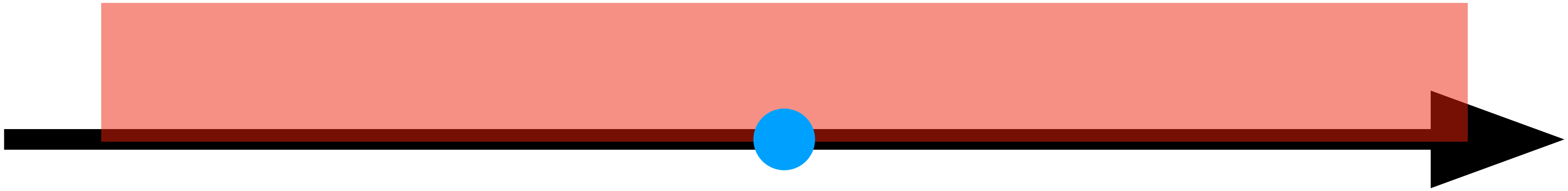


# Bandwidth





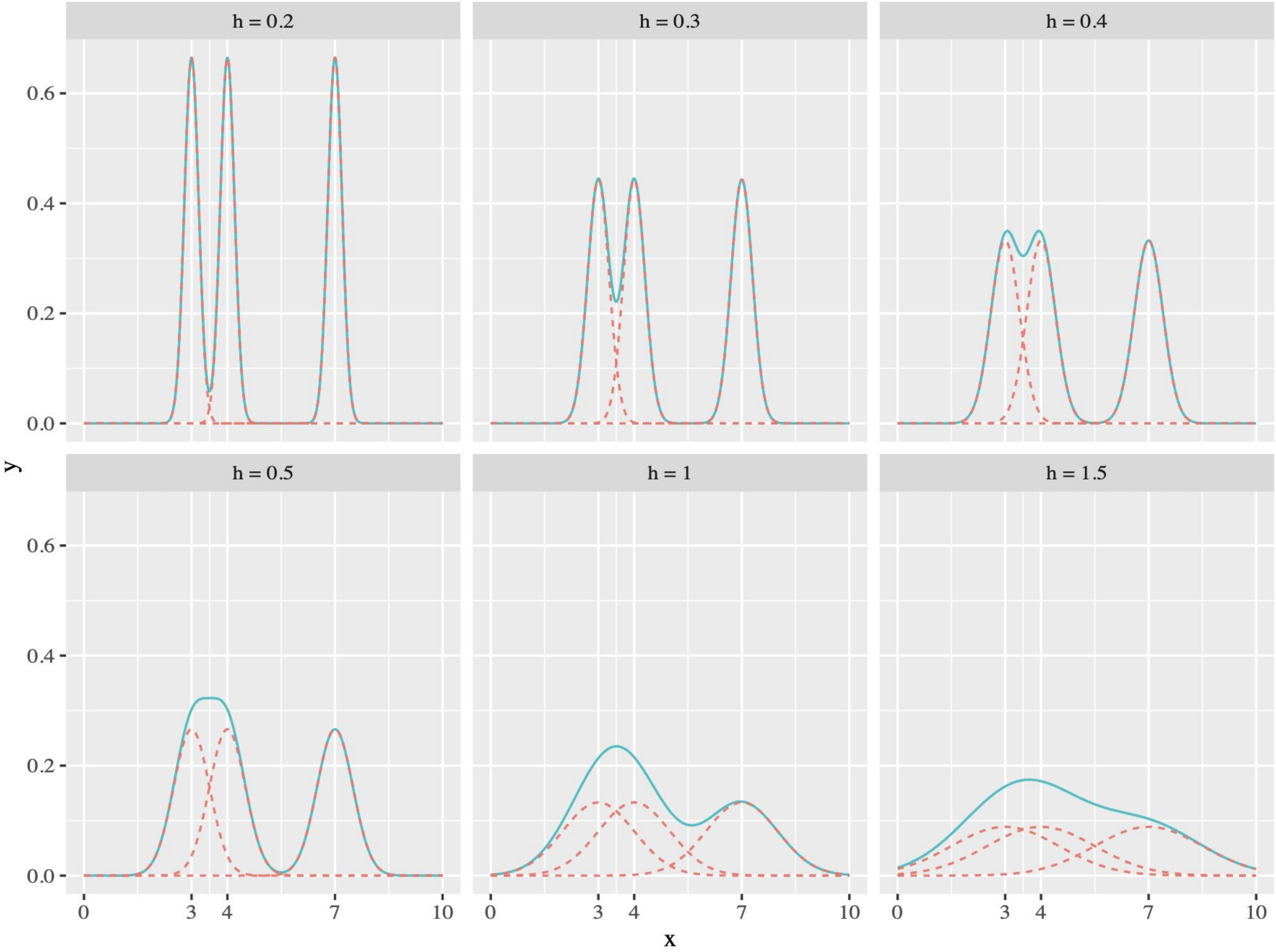


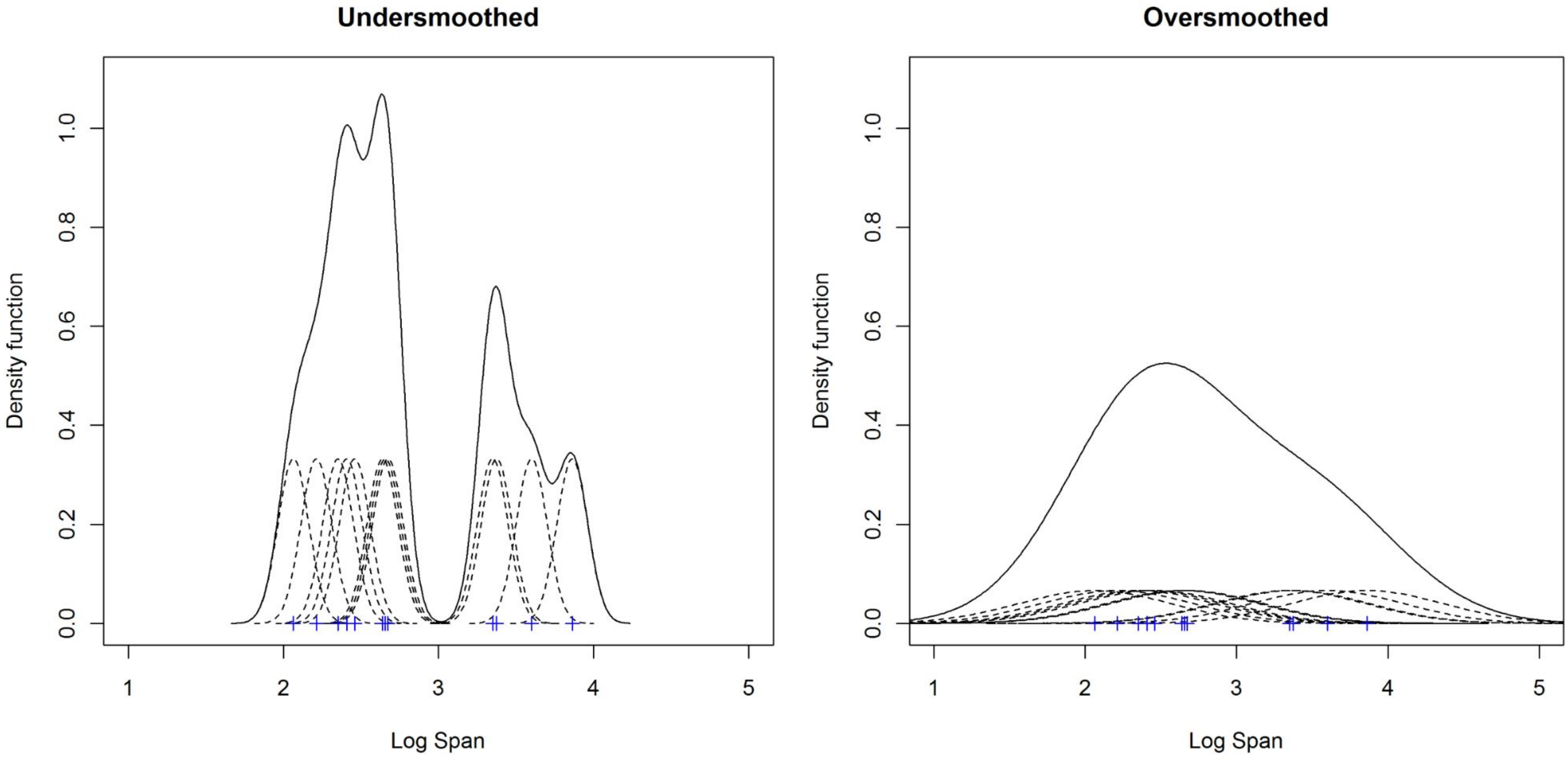


KDE

Histogram

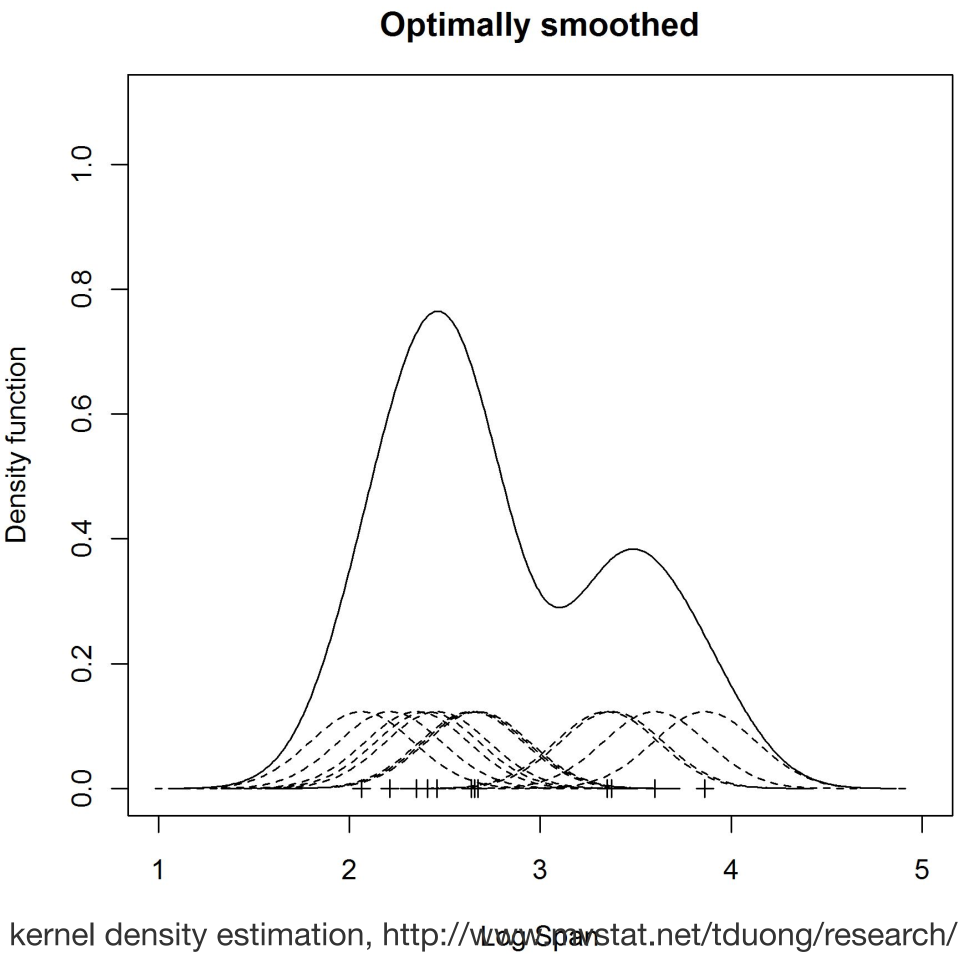
Choice of **bandwidth** ~ Choice of **bin size**


h = 0.2
h = 0.3
h = 0.4
h = 0.5
h = 1
h = 1.5
y
0.0
0.2
0.4
0.6
x
0
3
4
7
10



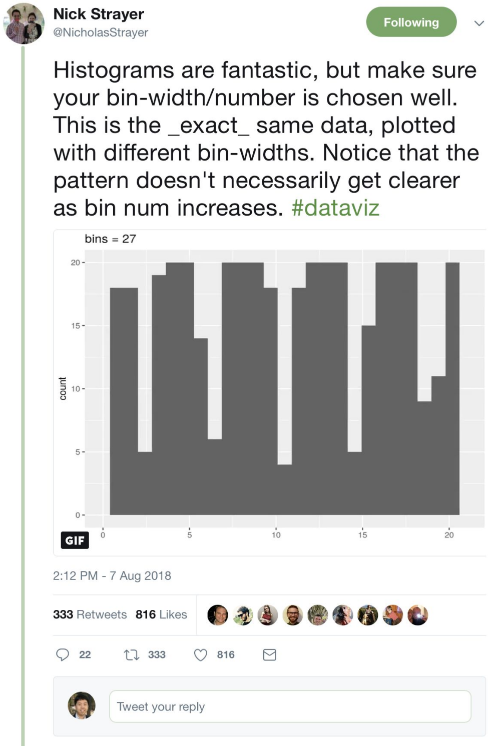
Undersmoothed

Oversmoothed



Tarn Duong, An introduction to kernel density estimation, http://www.mvstat.net/tduong/research/seminars/seminar-2001-05/

https://twitter.com/NicholasStrayer/status/1026893778404225024



Nick Strayer
@NicholasStrayer

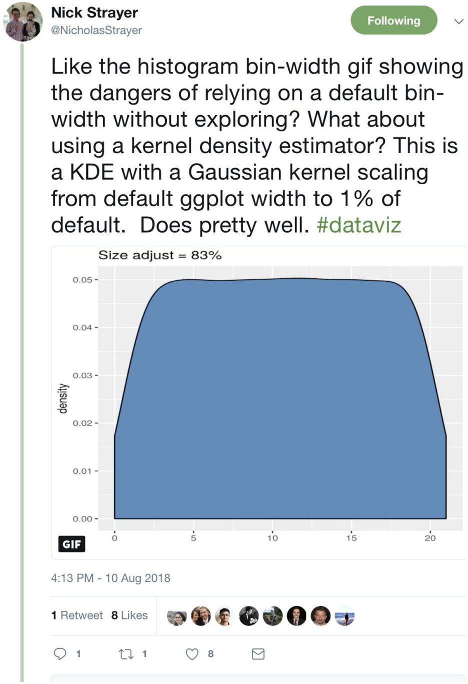
Histograms are fantastic, but make sure your bin-width/number is chosen well. This is the _exact_ same data, plotted with different bin-widths. Notice that the pattern doesn't necessarily get clearer as bin num increases. #dataviz

2:12 PM - 7 Aug 2018

333 Retweets 816 Likes

22 333 816

https://twitter.com/NicholasStrayer/status/1028011396448116736



Nick Strayer
@NicholasStrayer

Like the histogram bin-width gif showing the dangers of relying on a default bin-width without exploring? What about using a kernel density estimator? This is a KDE with a Gaussian kernel scaling from default ggplot width to 1% of default. Does pretty well. #dataviz

4:13 PM - 10 Aug 2018

1 Retweet 8 Likes

1 1 8

# Discussions for Team Project

# Project idea presentation (10 Oct, in class)

- Check “Proposal” in prof. yy’s project page for what you need to include in your presentation (Intro, Questions or objectives, Datasets and methods, References).
- Submit your presentation slides (PDF) by 9 Oct (one per team)
    - To save the transition time, all presentation slides will be played on the instructor’s laptop.
- 6 minutes per team (elevator pitch!)