Identifying and Visualizing Probabilistic and Statistical Relationships

Janani Ravi
CO-FOUNDER, LOONYCORN
www.loonycorn.com

Overview

Seaborn for statistical visualizations

Understanding kernel density estimation and KDE plots

Univariate analysis using histograms, KDE plots and rug plots

Visualizing pairwise relationships in data

Visualizing multivariate relationships using the facet grid

Visualizing Data with Seaborn

Seaborn

Built on top of matplotlib and tightly integrated with the PyData stack, including support for numpy and pandas data structures and statistical routines from scipy and statsmodels.

seaborn.pydata.org

Seaborn For “Production Plots”

Matplotlib

Part of “Pydata” - open data science stack

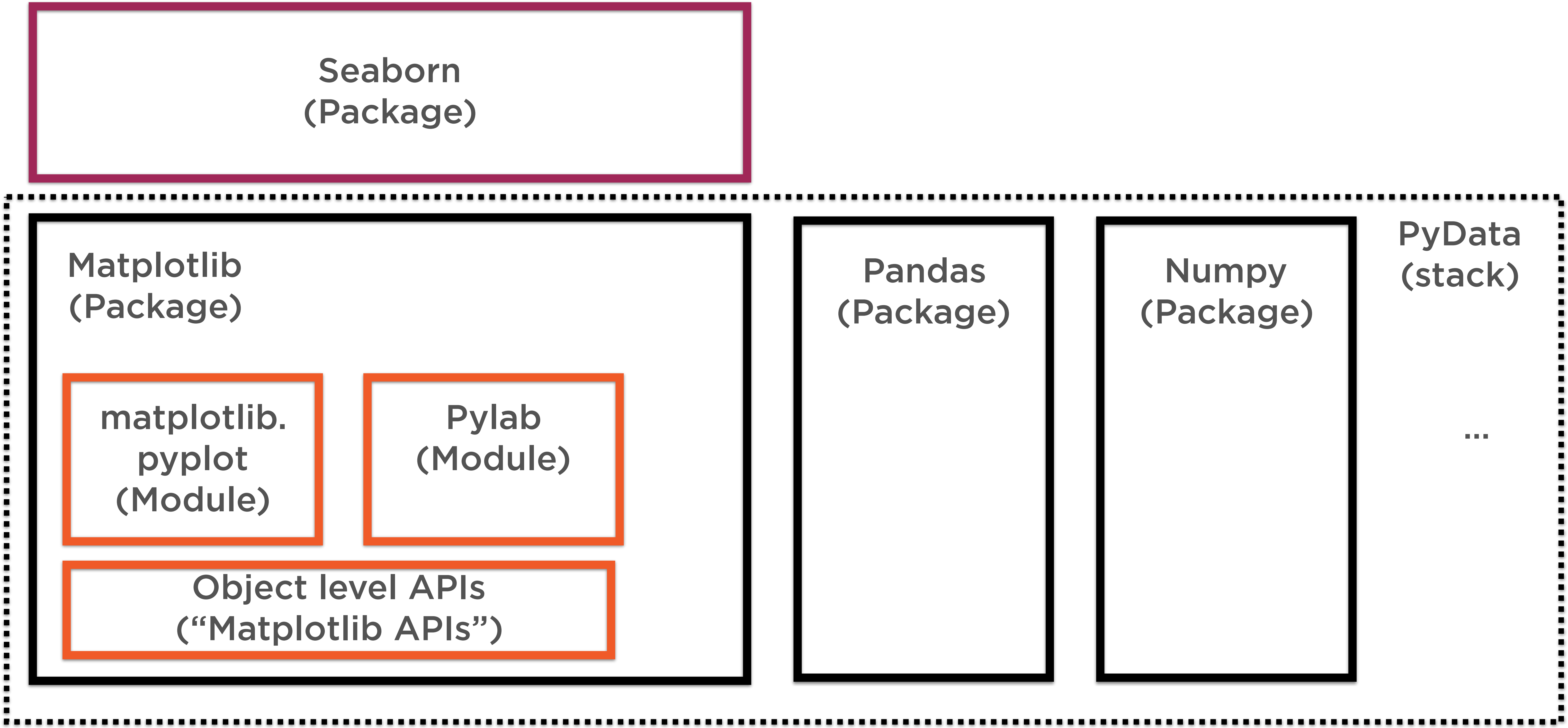
Provides fine-grained control so that pretty much everything is possible

Seaborn

Built atop Matplotlib and tightly integrates with Pydata

High level, easy-to-use abstractions for common use cases

Matplotlib and Seaborn



Seaborn
(Package)

Matplotlib
(Package)

matplotlib.
pyplot
(Module)

Pylab
(Module)

Object level APIs
("Matplotlib APIs")

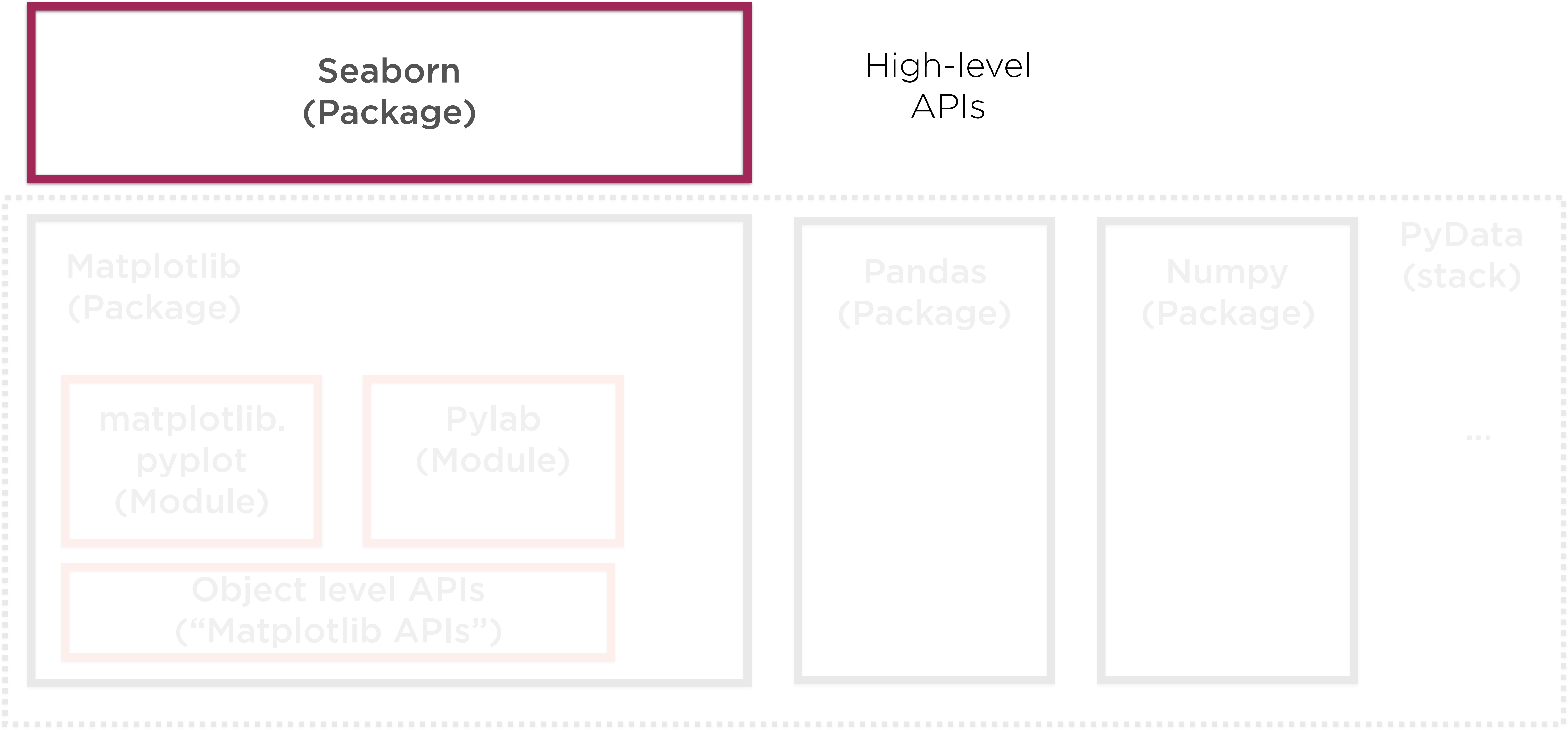
Pandas
(Package)

Numpy
(Package)

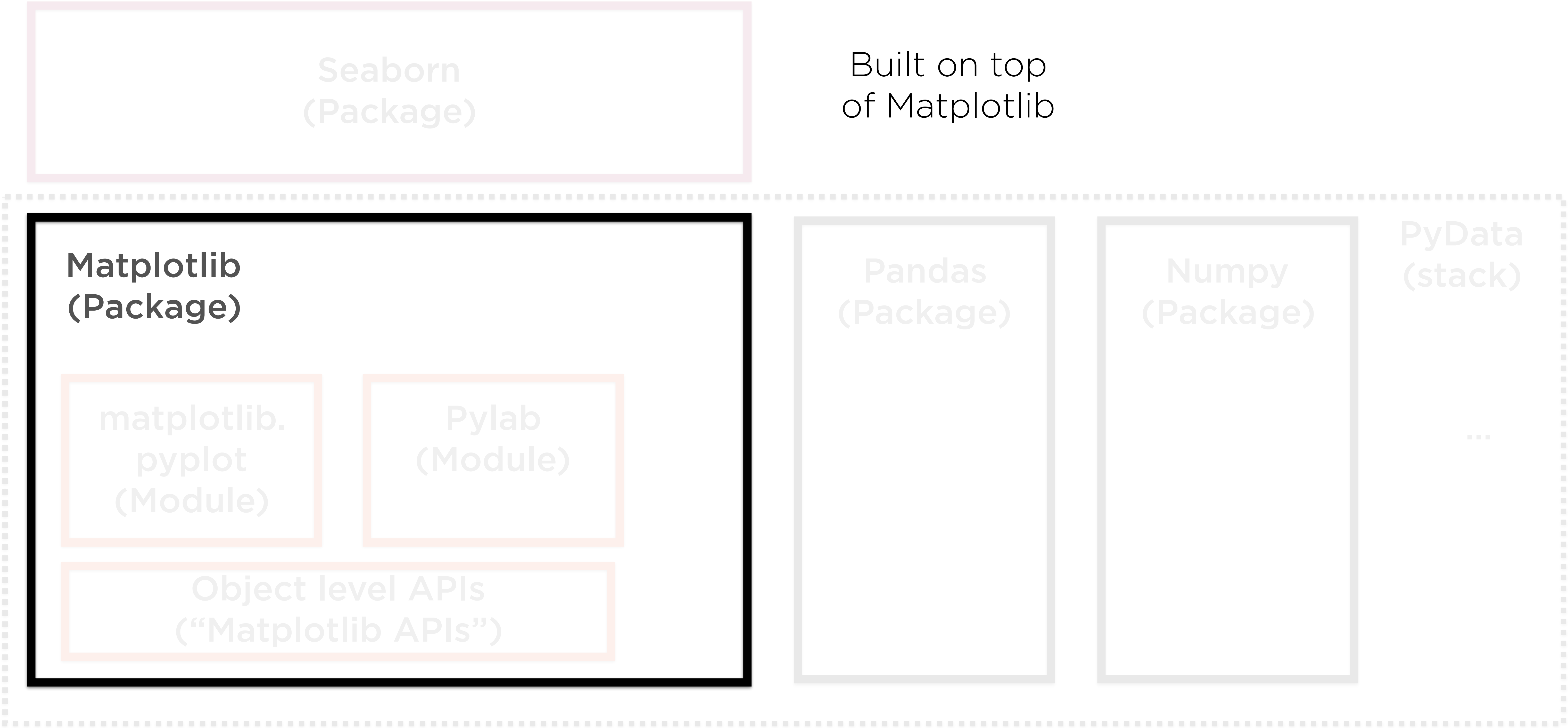
PyData
(stack)

...

Matplotlib and Seaborn



Matplotlib and Seaborn

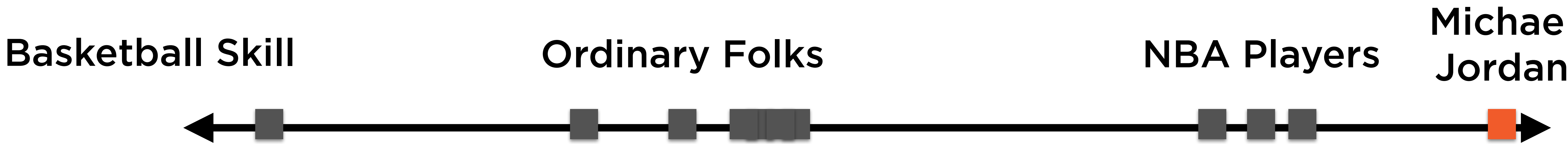


Built on top of Matplotlib

Understanding KDE Plots

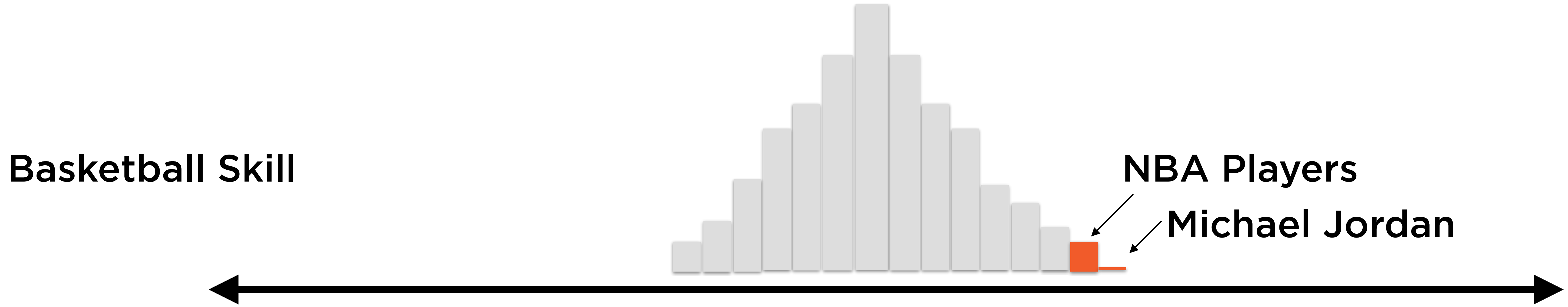
“Michael Jordan is a once-in-a-lifetime player”

Outliers



A once-in-a-lifetime player is an outlier, a point far from the pack

Outliers

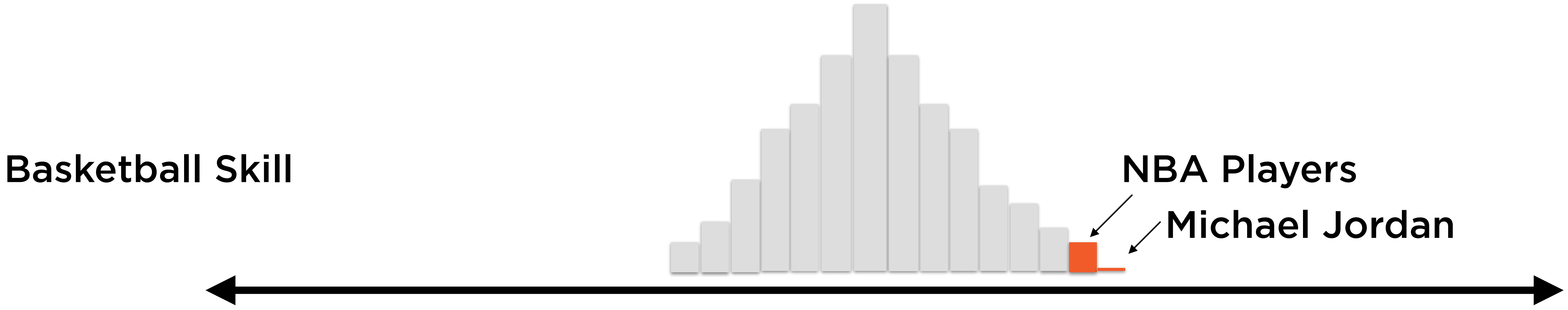


In reality, most ordinary folks would be clustered around an average level of skill

The NBA players would be outliers

Michael Jordan would be an even greater outlier

Outliers

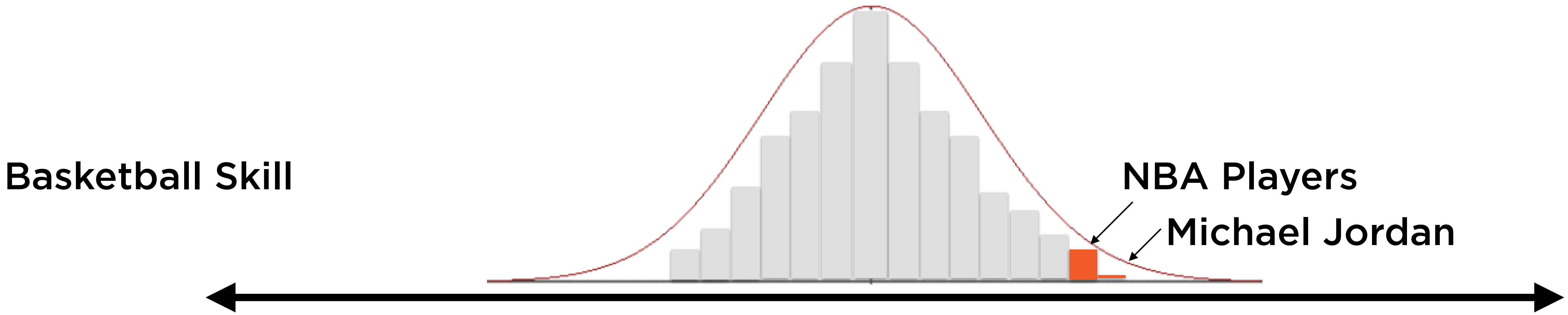


This chart above tells us how common a specific level of skill is

The shape of this chart resembles a bell

This is a Normal Probability Distribution

Outliers

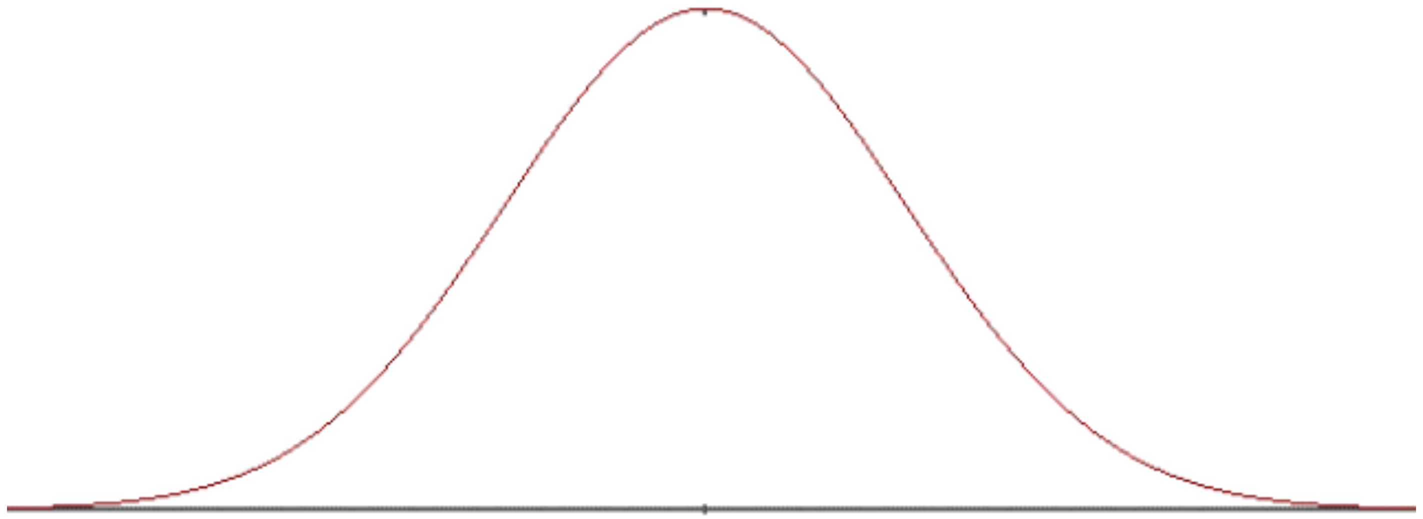


This chart above tells us how common a specific level of skill is

The shape of this chart resembles a bell

This is a Normal Probability Distribution

Outliers

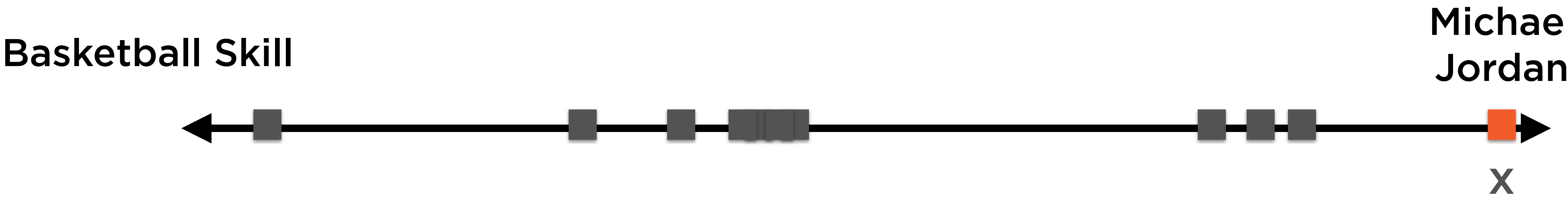


Average is common

Very high and very low are both unusual

The bell curve occurs everywhere in nature

Outliers



What is the probability of any specific value x occurring in the data?

The answer lies in a **probability distribution function**

Kernel Density Estimation

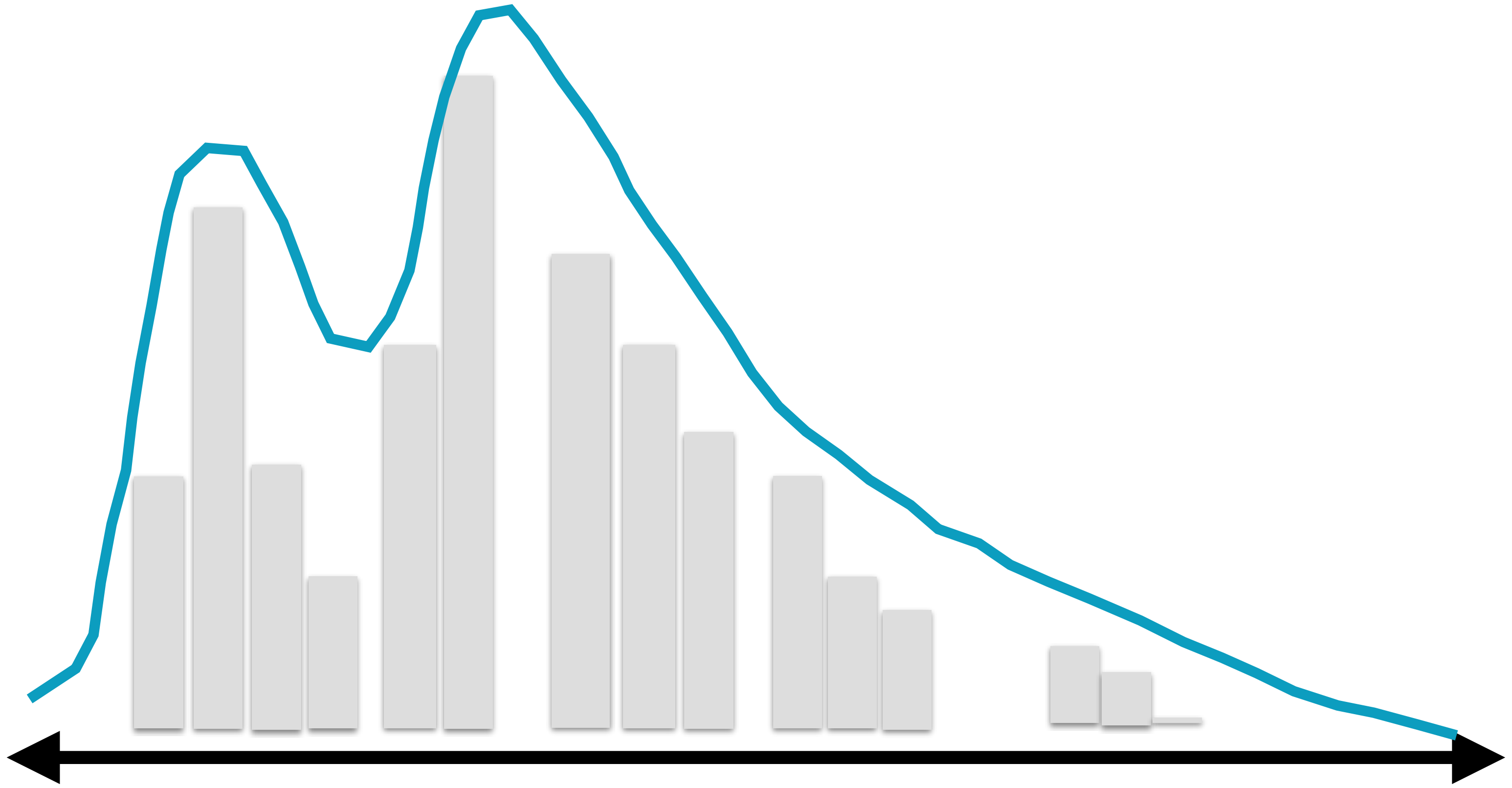
A mathematical technique used to get a smooth probability distribution from a histogram of raw data

Kernel Density Estimation

Given a set of points

Figure out their probability distribution

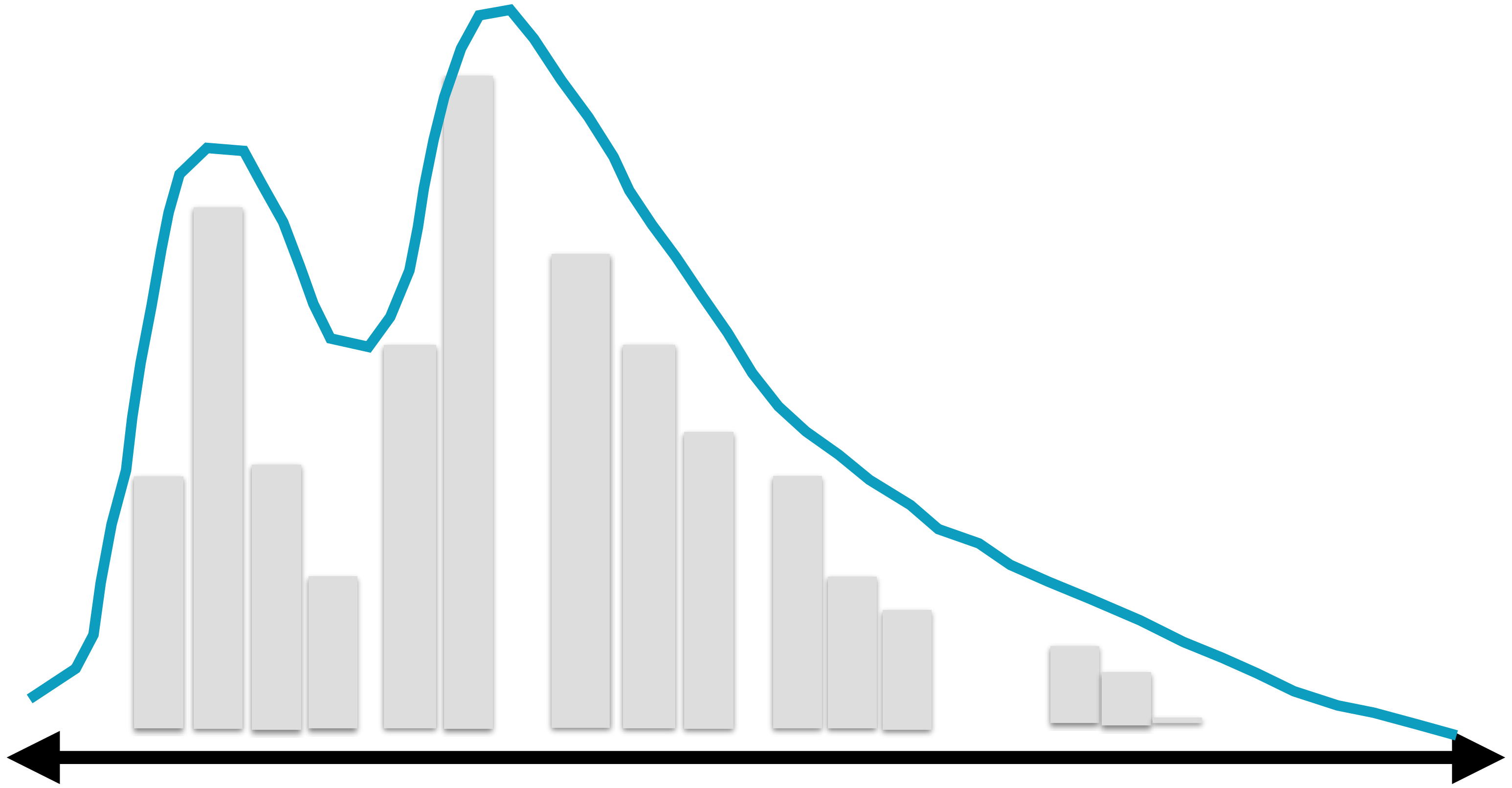
Area under curve must sum to 1



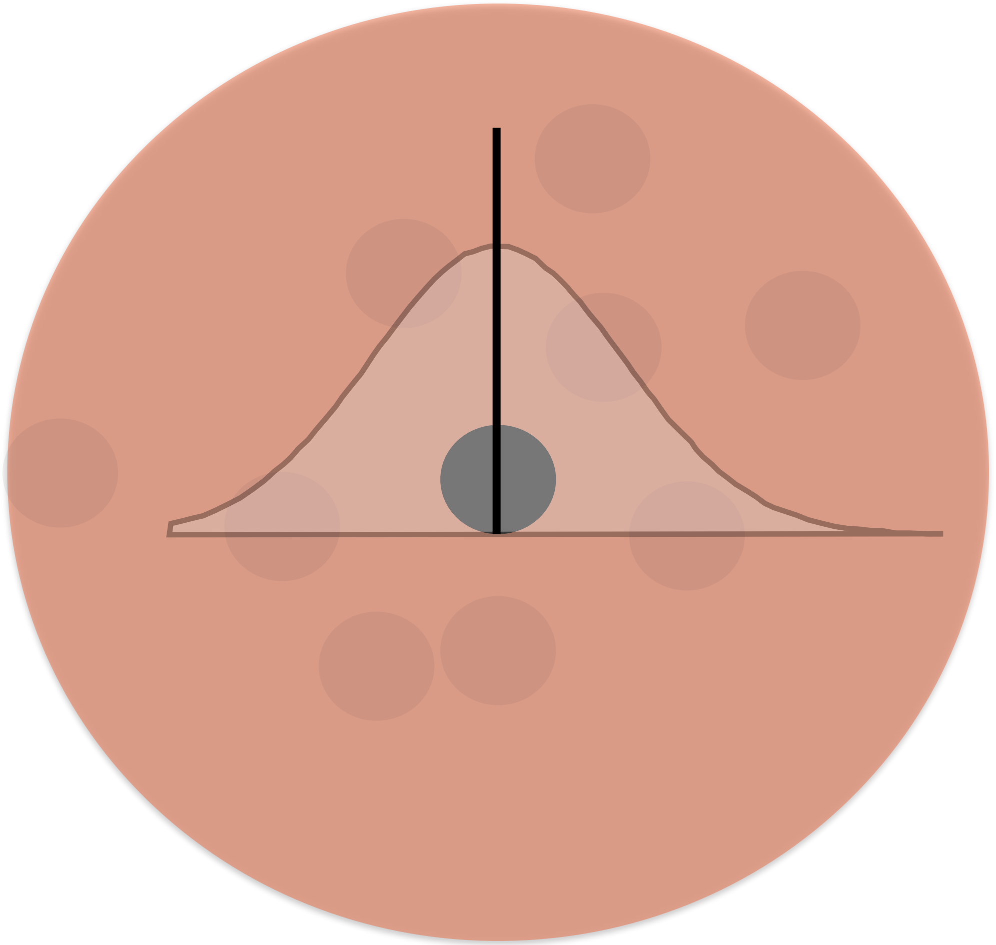
Kernel Density Estimation

KDE is a standard technique

Non-parametric "smoothing" technique



Gaussian Kernel

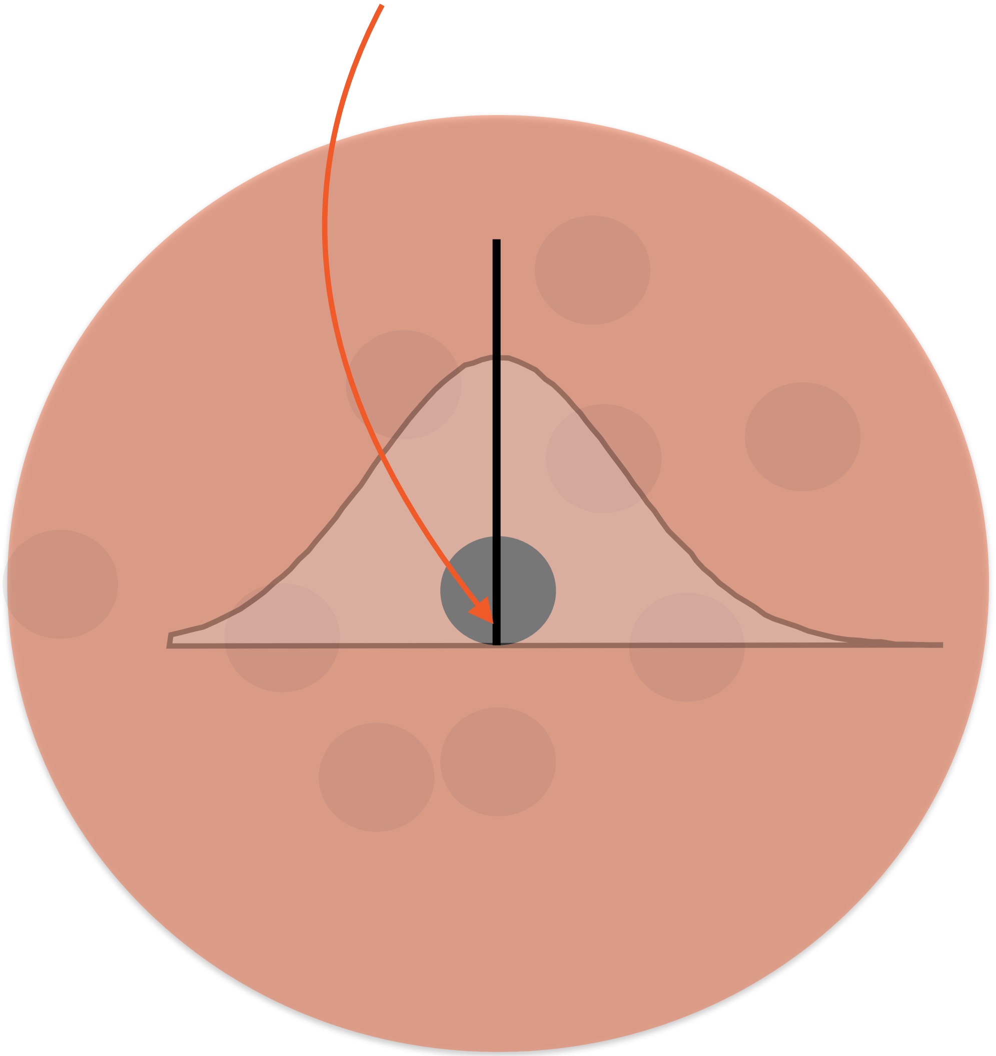


Gaussian probability distribution

Defined by

- mean μ
- standard deviation σ

Gaussian Kernel



Mean μ = center point

Standard deviation σ ~ bandwidth

(Bandwidth is a hyperparameter)

Demo

Visualizing univariate data using histograms, KDE plots and Rug plots

Visualizing bivariate relationships using scatter plots and hex bin plots

Demo

Visualizing continuous and categorical data using different plots in Seaborn

Demo

Visualizing and customizing pairwise relationships using the PairGrid

Demo

Visualizing multiple relationships using facets

Summary

Seaborn for statistical visualizations

Understanding kernel density estimation and KDE plots

Univariate analysis using histograms, KDE plots and rug plots

Visualizing pairwise relationships in data

Visualizing multivariate relationships using the facet grid