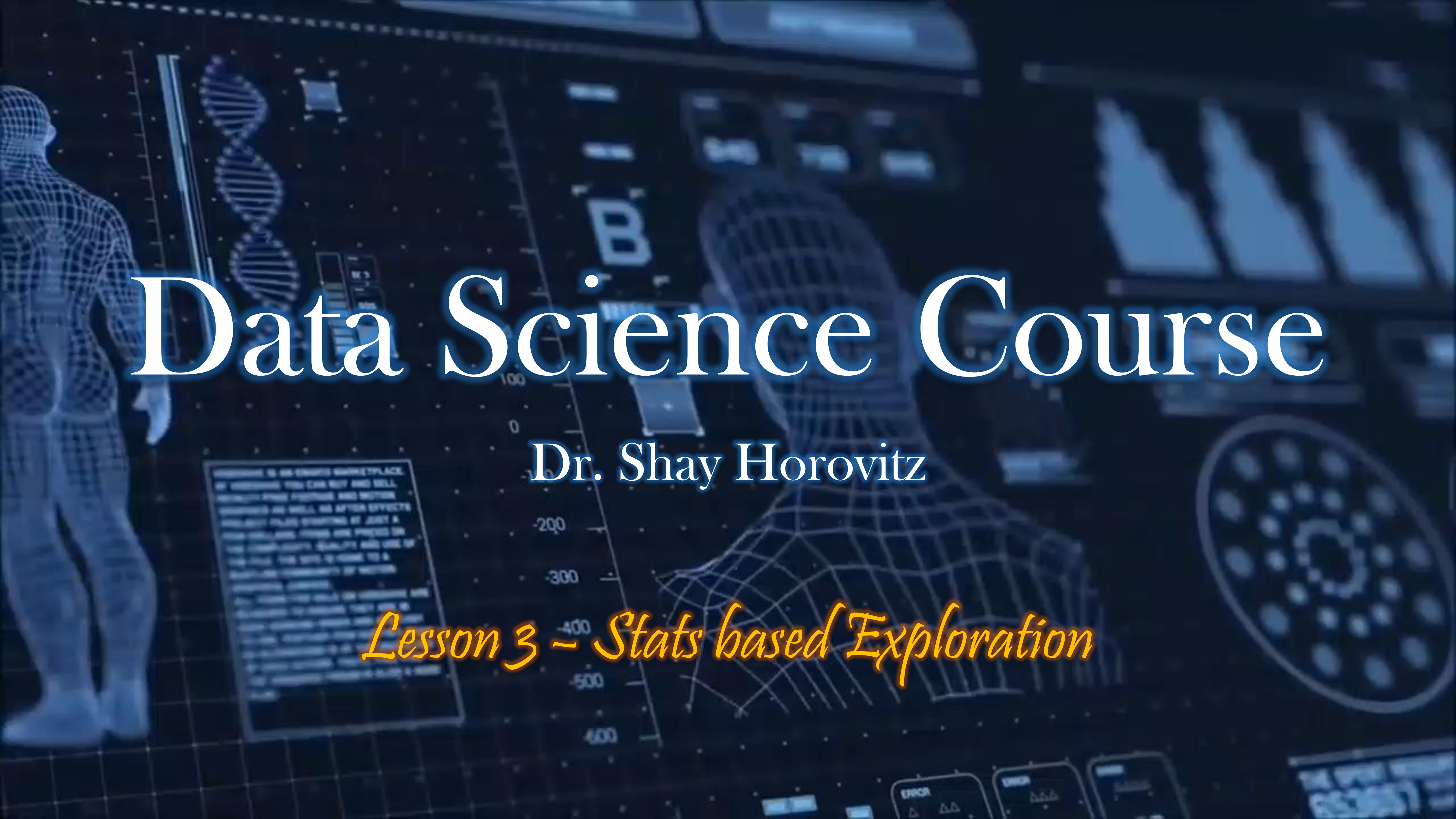
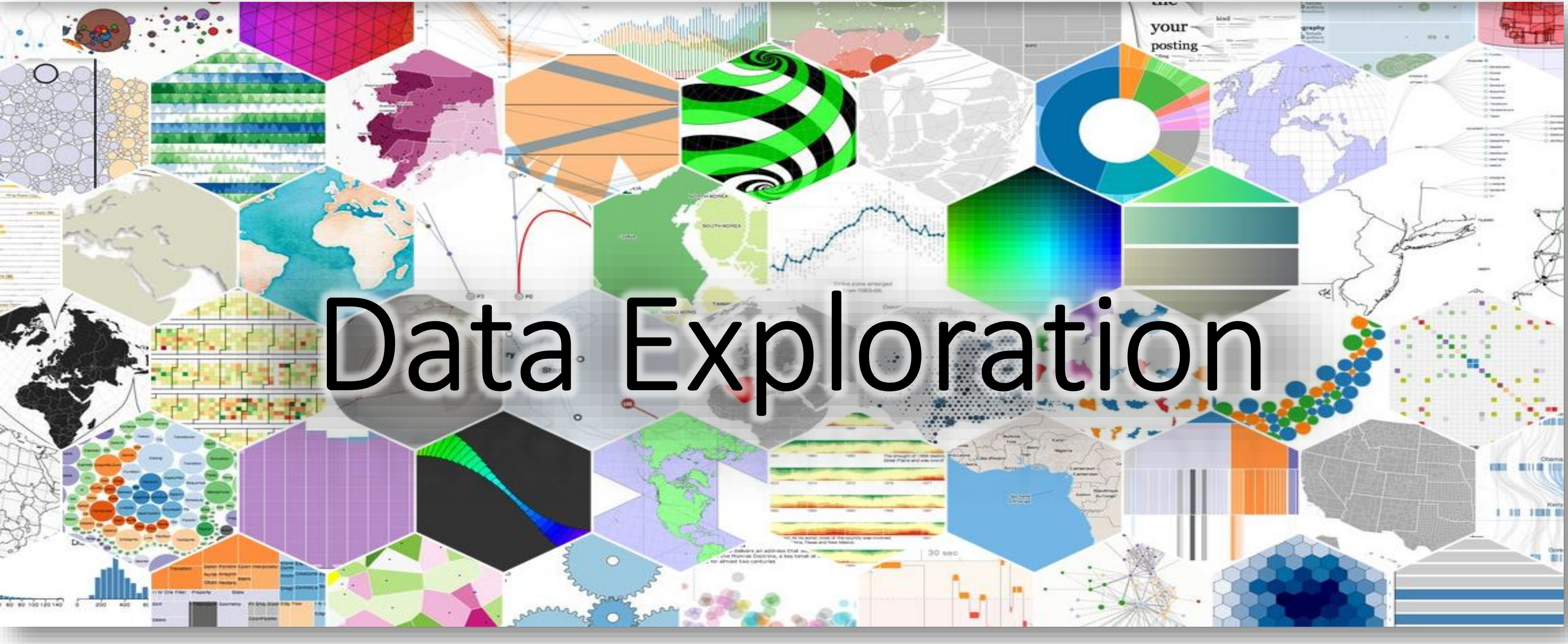


Data Science Course

Dr. Shay Horovitz

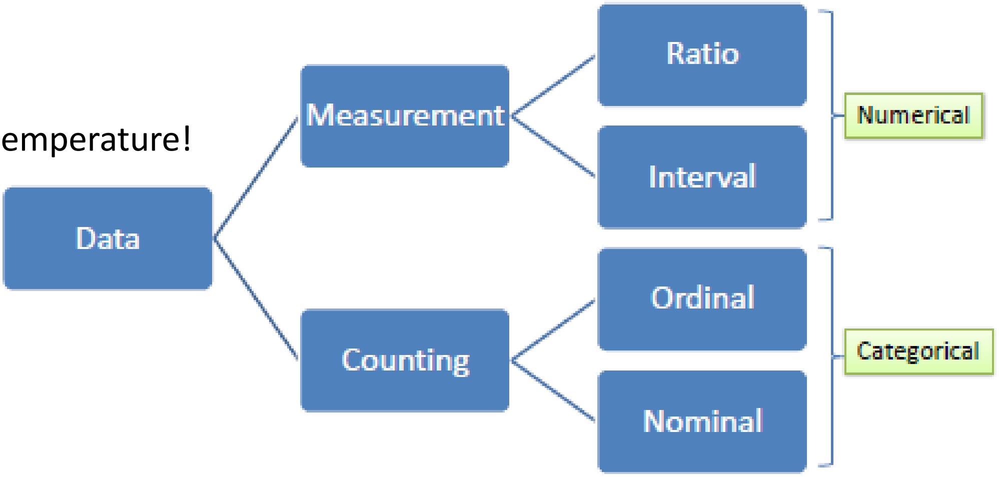
Lesson 3 – Stats based Exploration

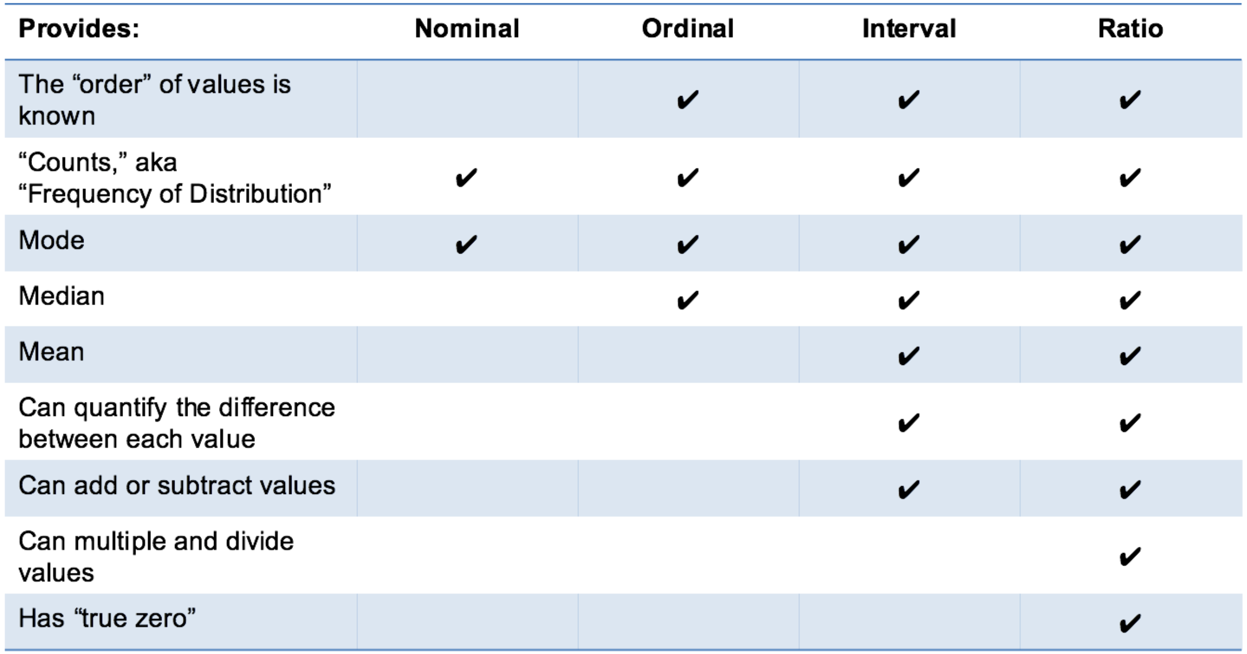
Data Exploration



Data Types

- Numerical/Continuous Data
  - Interval
    - **can + - , can't * /** *(no true Zero) – example: temperature*
    - Difference between values is meaningful. Zero Celsius doesn't mean that there's no temperature!
  - Ratio
    - **can + - * /** *(has true Zero) – example: weight*
    - 0 means that there's no variable
- Categorical/Discrete Data
  - Nominal
    - No order is defined between categories. *example – (Male, Female)*
  - Ordinal
    - Order between categories is defined *example - level of energy, movie rating 1-5*





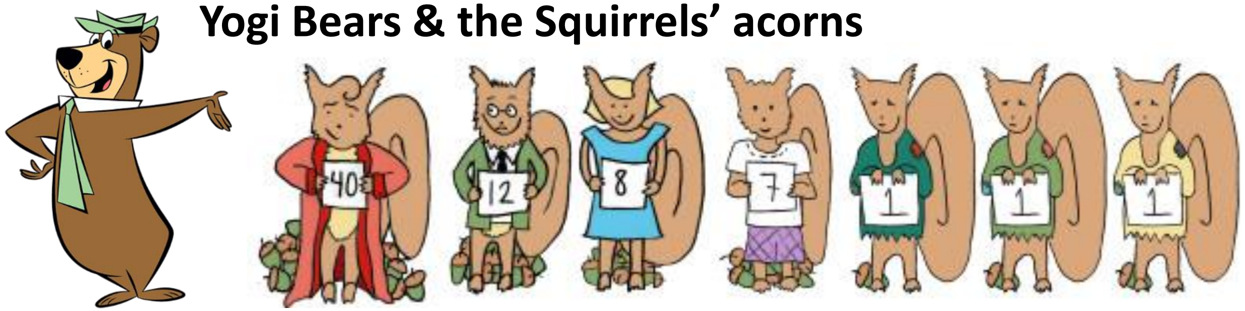
Provides:	Nominal	Ordinal	Interval	Ratio
The "order" of values is known		✔	✔	✔
"Counts," aka "Frequency of Distribution"	✔	✔	✔	✔
Mode	✔	✔	✔	✔
Median		✔	✔	✔
Mean			✔	✔
Can quantify the difference between each value			✔	✔
Can add or subtract values			✔	✔
Can multiple and divide values				✔
Has "true zero"				✔

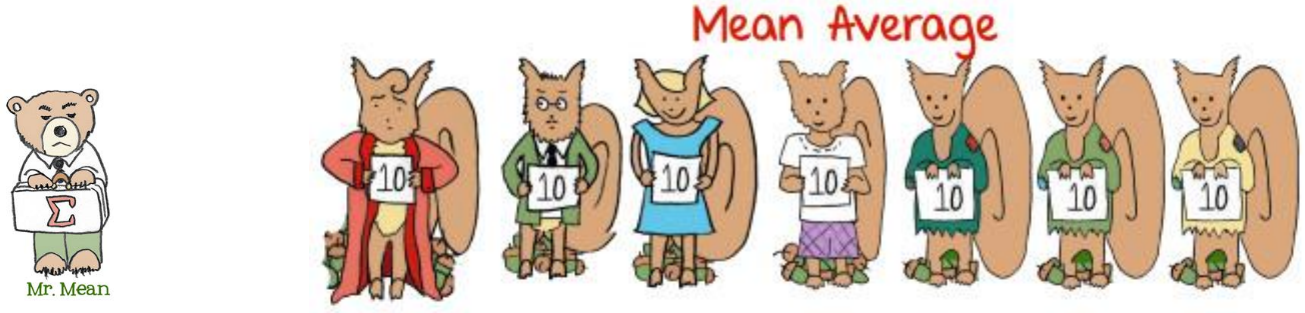
Data Types Quiz

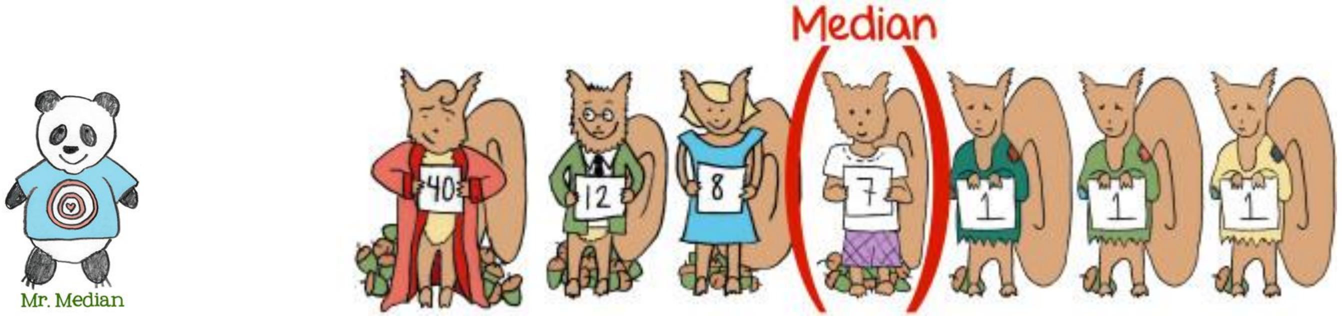
• How much gas in your gas tank?	Continuous **Ratio**
• A rating of your health: “poor”, ”moderate”, “good”, “excellent”	Discrete **Ordinal**
• The race of your classmates	Discrete **Nominal**
• Ages in years	Discrete **Ordinal**
• Money spent in a store	Continuous **Ratio**

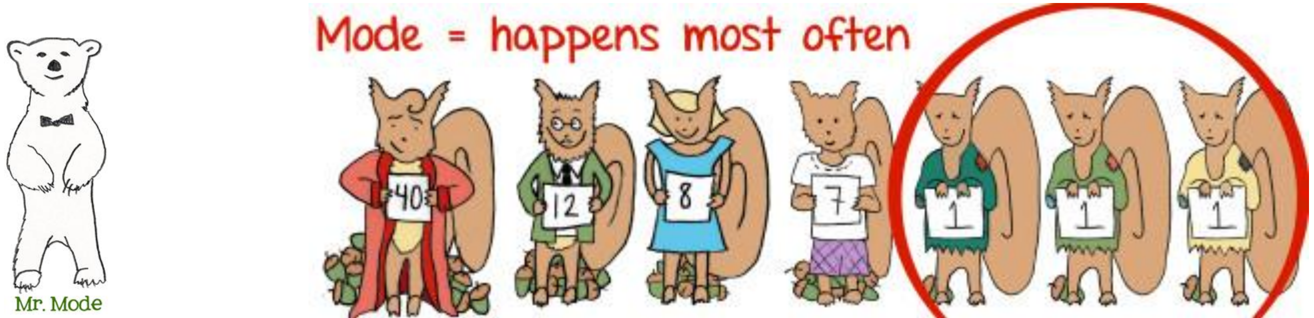
Statistics – Mean, Median, Mode

Yogi Bears & the Squirrels' acorns









4. MeanMedianMode.ipynb

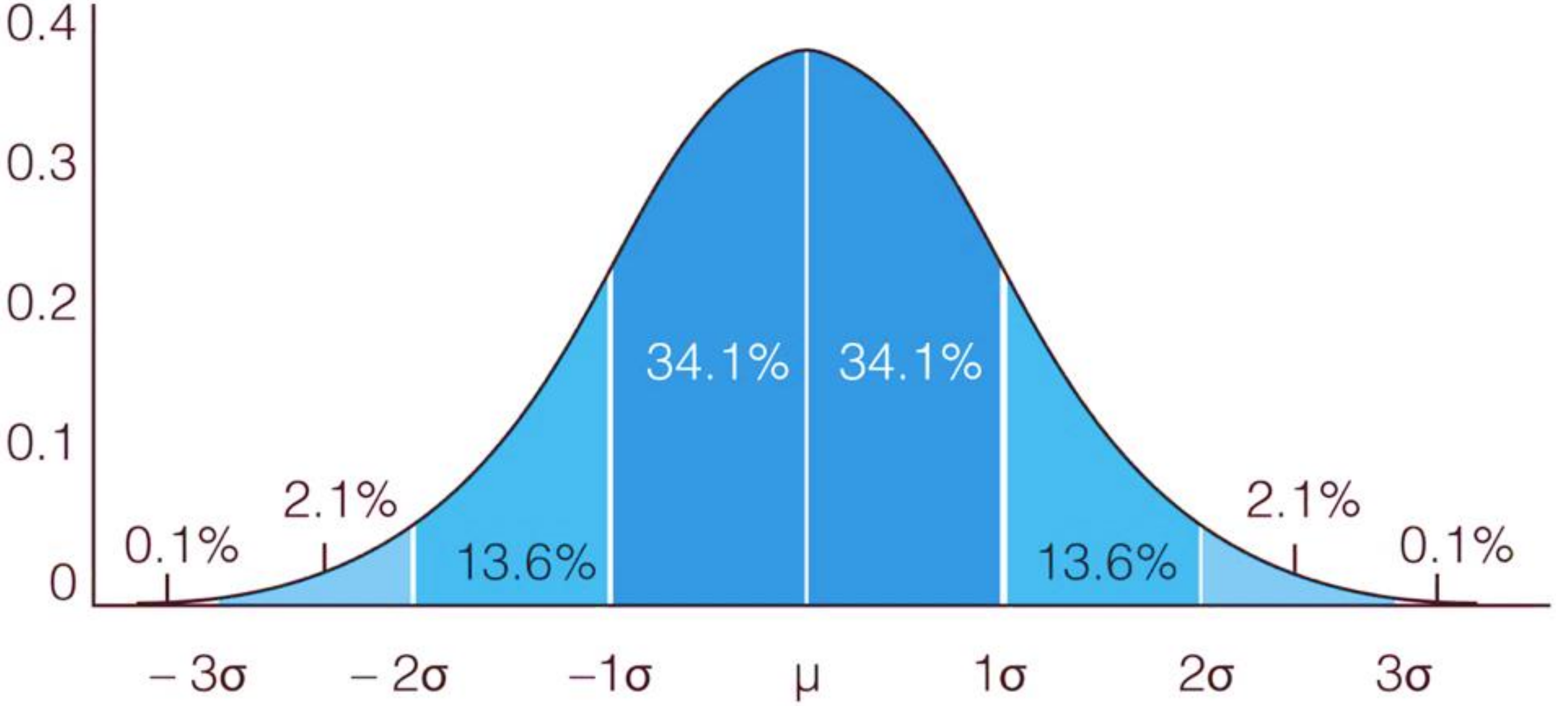
Statistics – Std.Dev, Variance

Variance is simply the **average of the squared differences from the mean**

$$\text{variance} = \sigma^2 = \frac{\Sigma (x_r - \mu)^2}{n}$$

$$\text{standard deviation} \quad \sigma = \sqrt{\frac{\Sigma (x_r - \mu)^2}{n}}$$

μ = mean

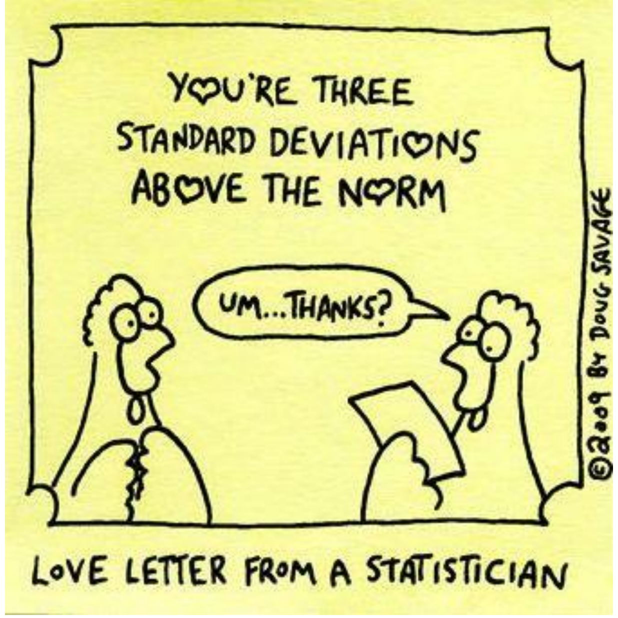


μ = Expected Value

-1σ to 1σ = 1 Standard Deviation (ie: ~2/3 of the time, your results/variance will fall within this range)

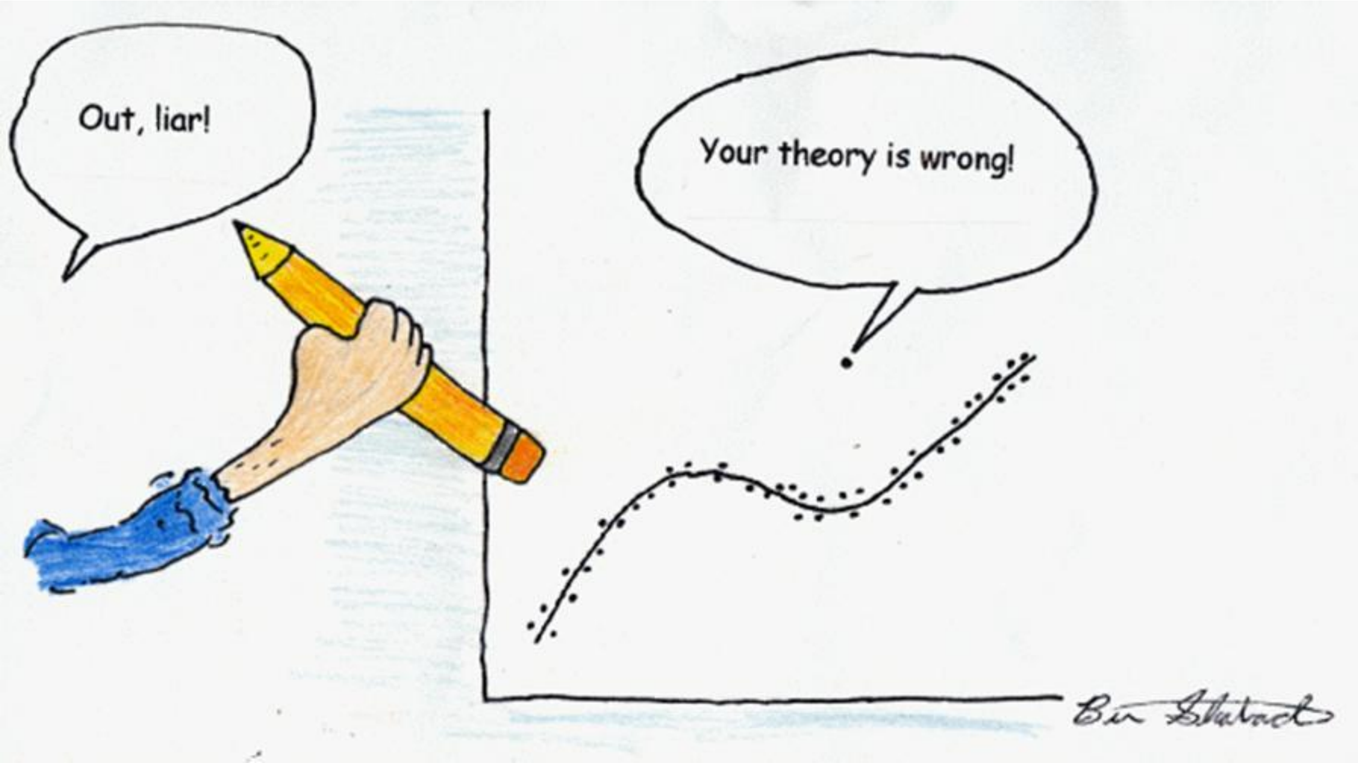
-2σ to 2σ = 2 Standard Deviations (ie: 95% of the time, your results/variance will fall within this range)

-3σ to 3σ = 3 Standard Deviations (ie: 99.7 of the time, your results/variance will fall within this range)


YOU'RE THREE STANDARD DEVIATIONS ABOVE THE NORM
UM...THANKS?
©2009 BY DOUG SAVAGE
LOVE LETTER FROM A STATISTICIAN

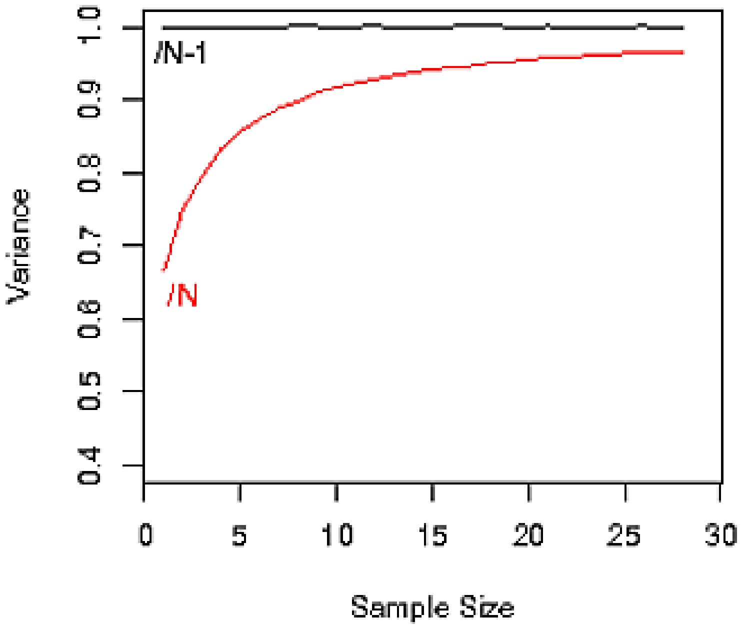
Statistics – Std.Dev, Variance

- **Std.Dev** is usually used as a way to **identify outliers**/anomalies
- You can talk about how extreme a data point is by talking about "**how many sigmas**" away from the mean it is



Population vs Sample

- **N-1 based Sample variance is a much better unbiased estimate of the population variance**



Population Variance

$$\sigma^2 = \frac{\sum_{i=1}^{n}(X_i - X_{avg})^2}{n}$$

Population Standard Deviation

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(X_i - X_{avg})^2}{n}}$$

Sample Variance

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - X_{avg})^2}{n-1}$$

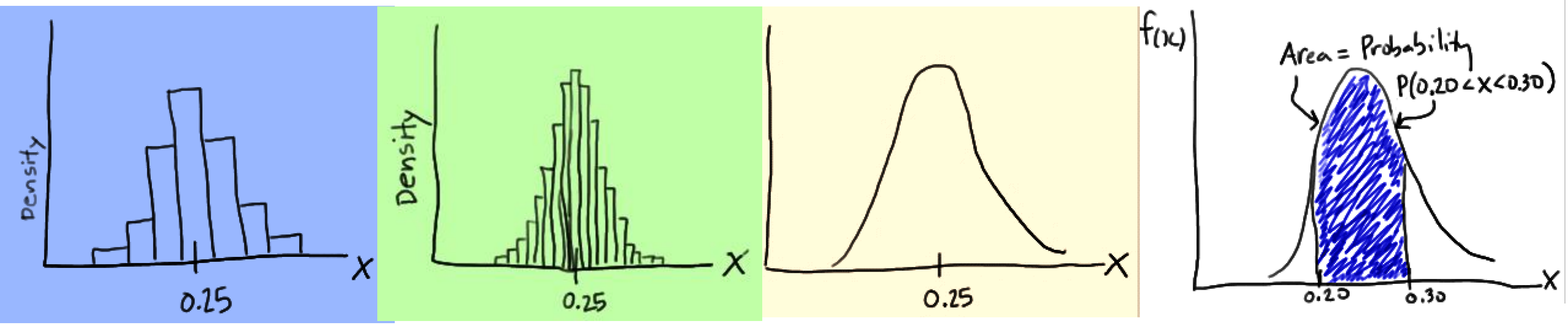
Sample Standard Deviation

$$s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - X_{avg})^2}{n-1}}$$

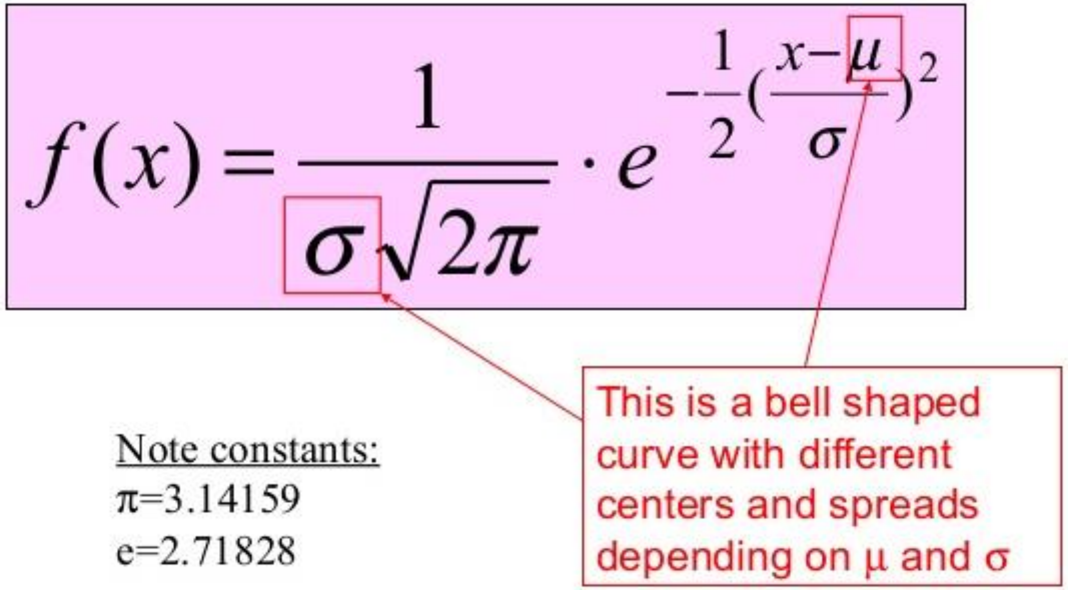
5. StdDevVariance.ipynb

Probability Density Function (PDF)

- The **probability density function** ("**p.d.f.**") is a function of a **continuous random** variable, whose **integral** across an **interval** gives the **probability that the value** of the variable lies within the same interval.



Normal Distribution as a function (PDF)

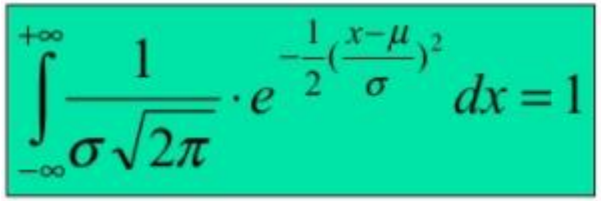


$$f(x) = \frac{1}{\boxed{\sigma}\sqrt{2\pi}} \cdot e^{-\frac{1}{2}(\frac{x-\boxed{\mu}}{\sigma})^2}$$

Note constants:
π=3.14159
e=2.71828

This is a bell shaped curve with different centers and spreads depending on μ and σ

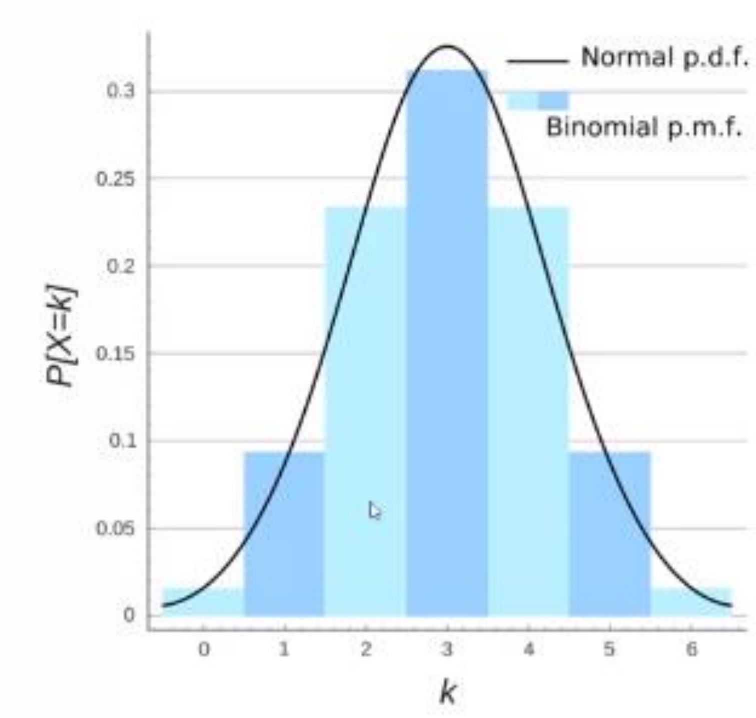
- Surely,Its probability sums up to 1:



$$\int_{-\infty}^{+\infty} \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx = 1$$

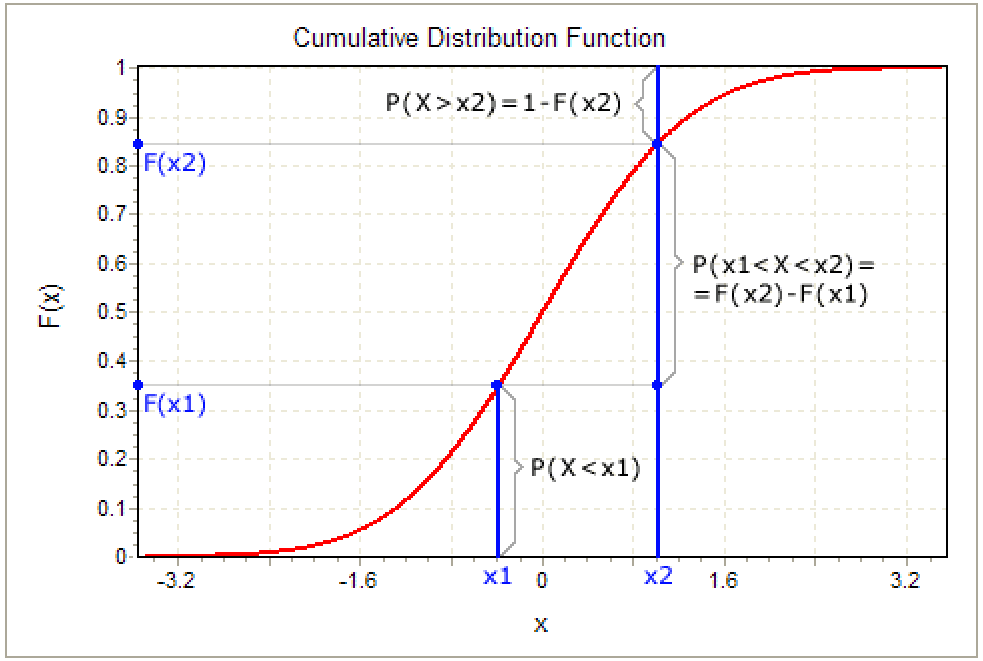
Probability Mass Function (PMF)

- a **probability mass function** (pmf) is a **function** that gives the **probability** that a **discrete** random variable is exactly equal to some value.

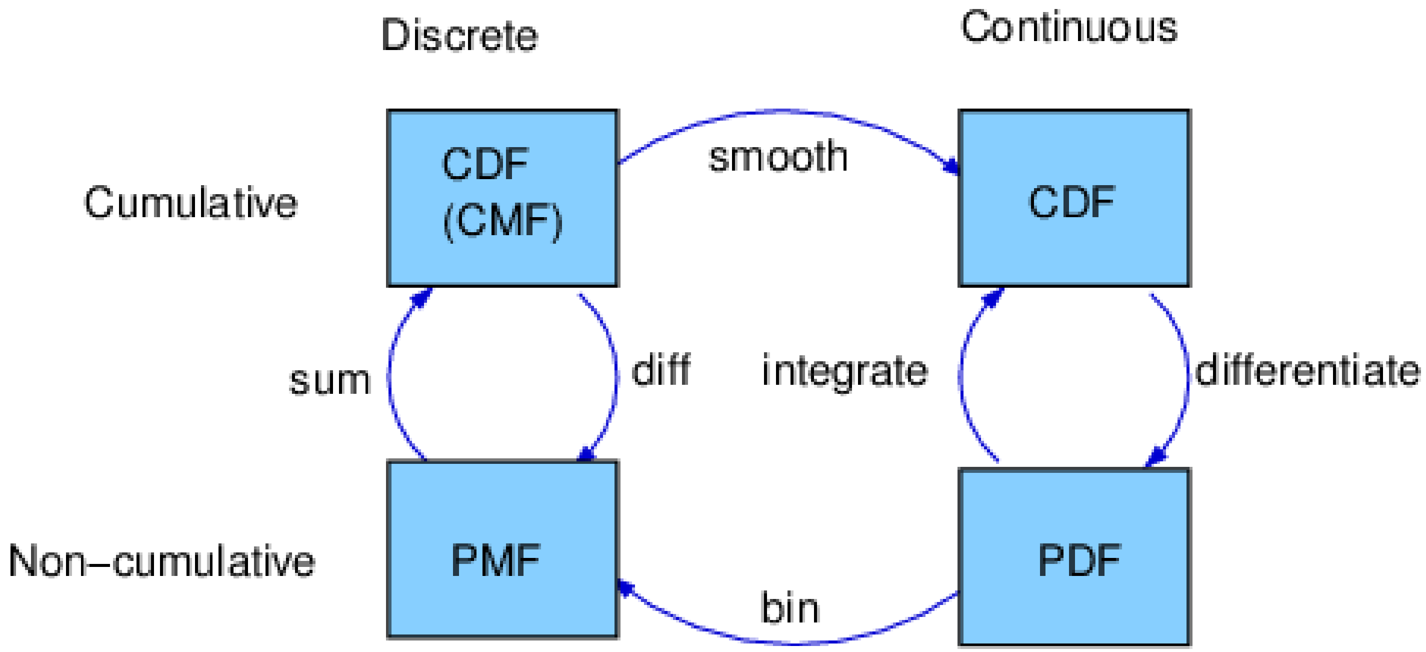


Cumulative Distribution Function (CDF)

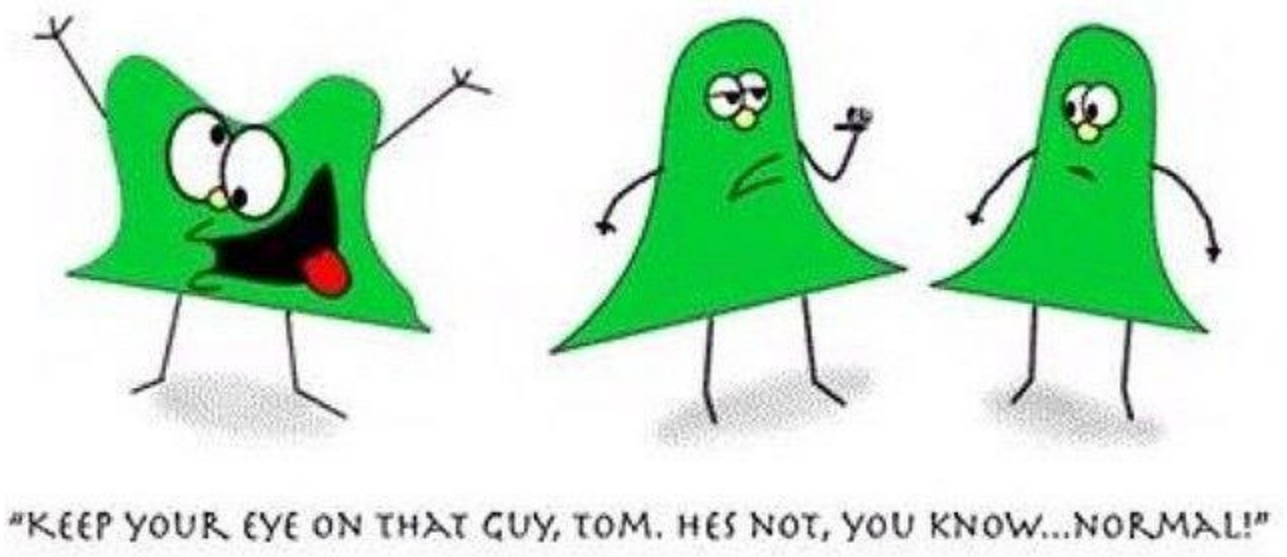
- the **cumulative distribution function** (**CDF**) of a real-valued random variable X, or just **distribution function** of X, evaluated at x, is the **probability** that **X will take a value less than or equal to x**.



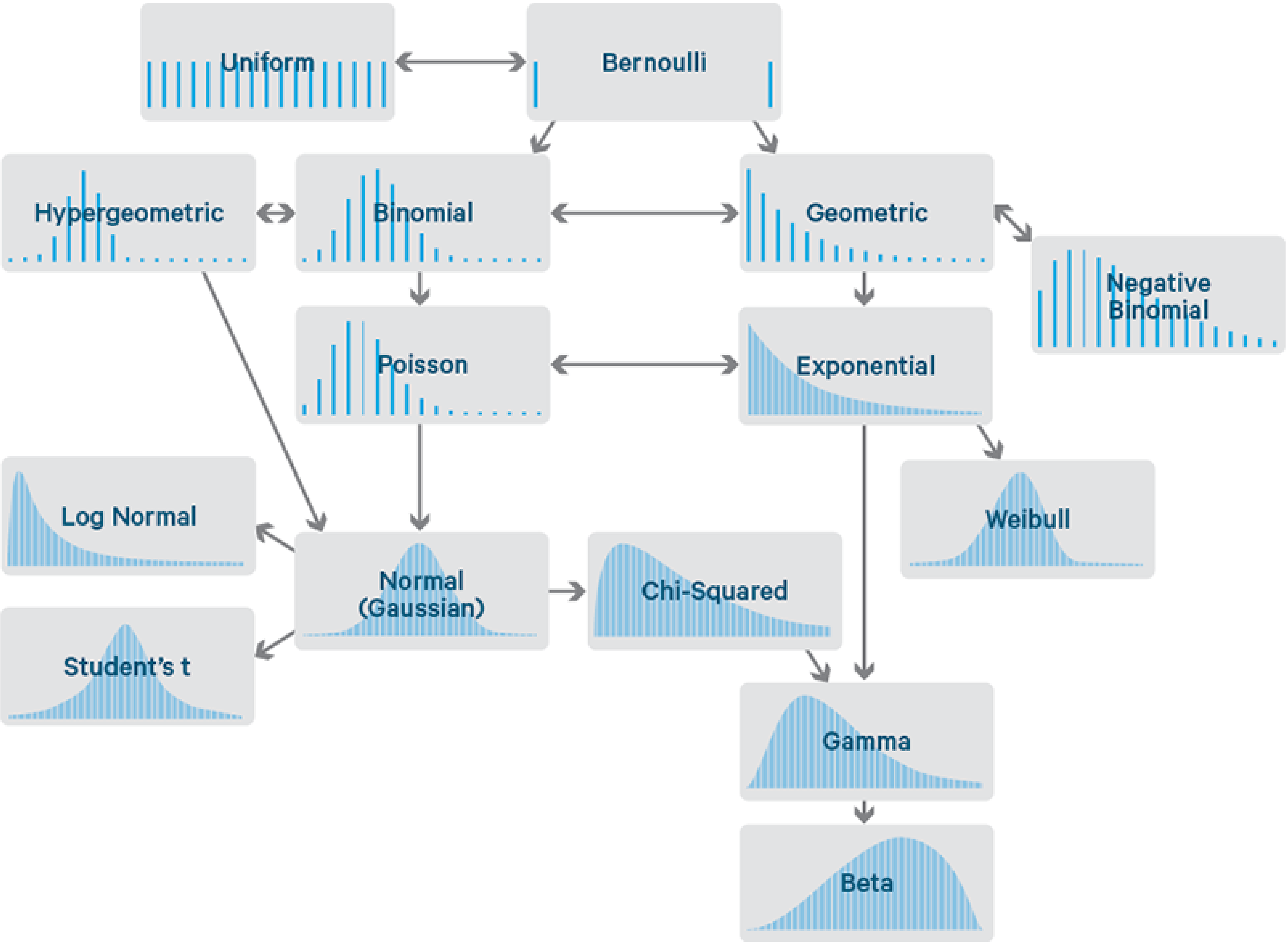
CDF, PDF, PMF,…

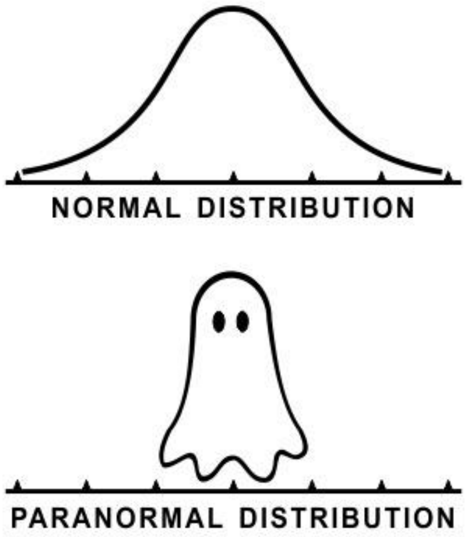


Discrete
Continuous
Cumulative
CDF (CMF)
smooth
CDF
sum
diff
integrate
differentiate
Non–cumulative
PMF
bin
PDF

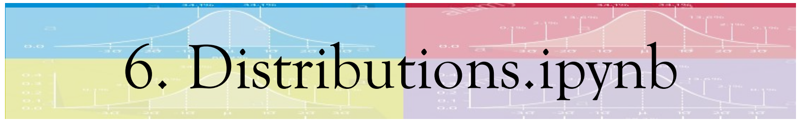

"KEEP YOUR EYE ON THAT GUY, TOM. HES NOT, YOU KNOW...NORMAL!"

Distributions





- https://blog.cloudera.com/blog/2015/12/common-probability-distributions-the-data-scientists-crib-sheet/



6. Distributions.ipynb

END