COGS138: Neural Data Science

Lecture 12

C. Alex Simpkins, PhD
UCSD Dept. of Cognitive Science, Spring 2023
RDPRobotics, LLC
http://casimpkinsjr.radiantdolphinpress.com/pages/cogs138_sp23

rdprobotics@gmail.com | csimpkinsjr@ucsd.edu

(Based on a course created by Prof. Bradley Voytek)

Plan for today

- Announcements
- Previous project review document
- Project overview
- Assignment 3 release
- Review - Last time
- Time series, sampling, aliasing, filtering, signal processing continued,

Announcements

- **Mid-quarter check-in survey assignment (required)**
- **Grade check-in - survey upcoming, to address any issues now and any concerns**
  - Check for missing quizzes (canvas)
  - Check for missing class participation (look over assignments page for completion of everything assigned, will import to canvas over the weekend if possible)
  - Check for data hub assignments (imported to canvas by weekend, check on data hub for now)
- **Group formation** - check canvas for empty groups if you want to self add
  - We have assigned everyone who did not say they did not want to be assigned, please connect with the team and quickly decide if you want to stay together or move
  - Contact Siddhant to move if needed, contact me if other issues or he doesn't get back to you
- Previous project review released when we get the groups together (this week)

Last time

Course links

Website	http://casimpkinsjr.radiantdolphinpress.com/pages/cogs138_sp23	Main face of the course and everything will be linked from here. Lectures, Readings, Handouts, Files, links
GitHub	https://github.com/drsimpkins-teaching	files/data, additional materials & final projects
datahub	https://datahub.ucsd.edu	assignment submission
Piazza	https://piazza.com/ucsd/spring2023/cogs138_sp23_a00/home (course code on canvas home page)	questions, discussion, and regrade requests
Canvas	https://canvas.ucsd.edu/courses/44897	grades, lecture videos
Anonymous Feedback	Will be able to submit via google form	If I ever offend you, use an example you are uncomfortable with, or to provide general feedback. Please remain constructive and polite

About the final projects

Finding the project files

- Blank starter document for report, handouts and info to start with (draft): https://github.com/drsimpkins-teaching/cogs138/tree/main/main_project
- Links to old projects: https://github.com/NeuralDataScience/Projects

Where will you turn in, work from?

- Github - we will create a repository for each group with the files in previous slides as a starter directory
- You'll add your data, notebook file etc there
- This will be the final turning location

Final Project Overview

1. Identify **questions** that you can answer by using publicly available datasets
2. Integrate different datasets
3. Implement **technical skills** to answer your scientific questions
4. Work effectively with a **team**

What will you be turning in??

Proposal

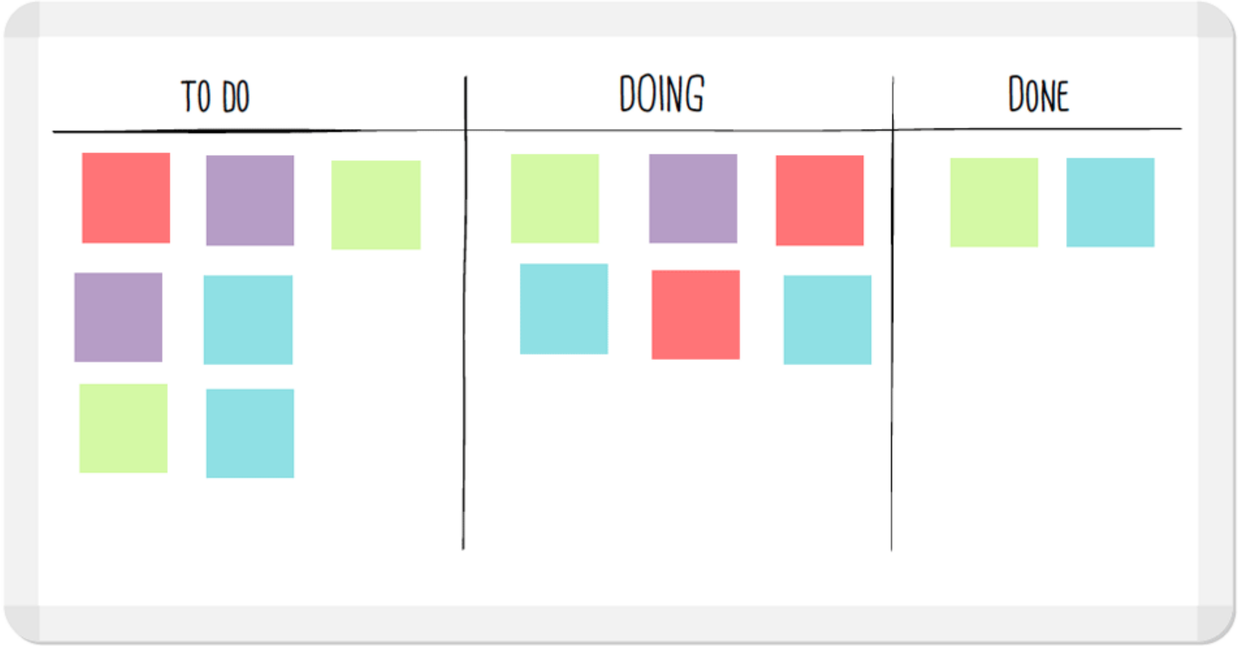
- 1 page document
- Pitching your question & approach

Final Project

- Jupyter notebook
- Steps of analysis that answers your main question
- Background and discussion sections

Proposal

1. Group member names
2. Experimental question
3. Background
4. Approach

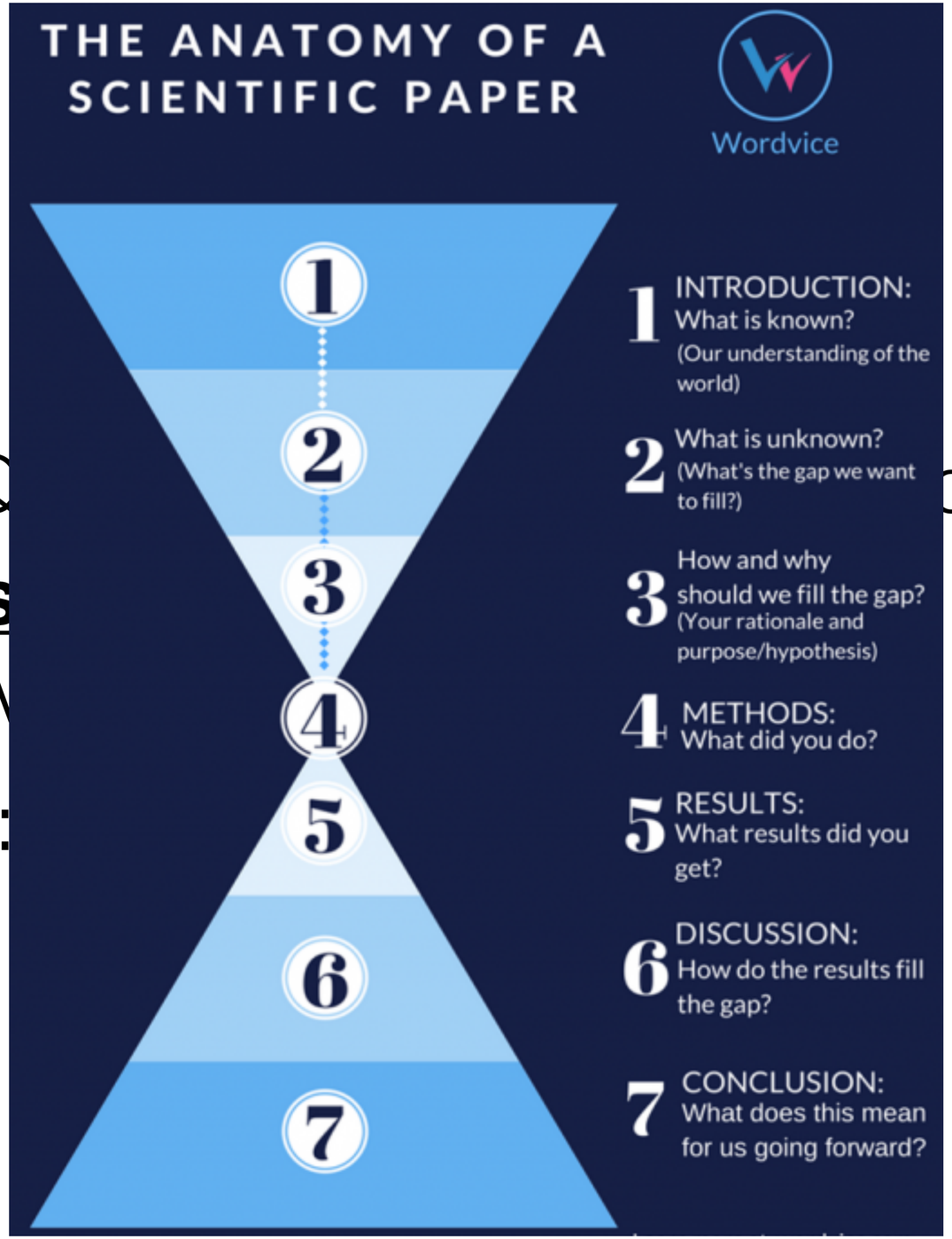


Final Project

1. **Intro**:
   a. Overview, Question, Background, Hypothesis
2. **Data Analysis**:
   a. Wrangling, Viz, Results
3. **Conclusion**:
   a. Discussion, Limitation, Future Steps

Final Project

1. **Intro**:
   a. Overview, Q… …othesis
2. **Data Analys…**
   a. Wrangling, …
3. **Conclusion**:…
   a. Discussion, …



Groups

Distributions

- http://localhost:8888/notebooks/Documents/teaching/cogs138/old/Tutorials-master/Distributions.ipynb
- http://localhost:8888/notebooks/Documents/teaching/cogs138/old/Tutorials-master/Central%20Limit%20Theorem.ipynb
- http://localhost:8888/notebooks/Documents/teaching/cogs138/old/Tutorials-master/Correlation%20resampling.ipynb

Correlations analysis, covariance and Time series analysis

PIP package manager

- pip (package manager) – Wikipedia
- Written in python
- Used to install, remove, manage software packages
- Connects to online package repository of public software (Python Package Index)
- Most python packages come with PIP installed
- Home page:pip documentation v23.1.2 (pypa.io)

Usage 99%

- ***pip install some_package_name***
- ***pip uninstall some_package_name***

Central tendency - Mean

- Balance point
- "Expected value" (population mean)
- Computed by
  - Sum scores,
  - Divide by number of scores

$$M = \left(\sum_{i=1}^{N} x_i \right) \Big/ N$$

$$\{1.0, 1.0, 2.0, 3.0, 4.0, 4.0, 4.0, 4.0, 8.0, 8.0, 8.0, 8.0, 8.0, 8.0, 9.0, 0.0, 0.0, 0.0\}$$

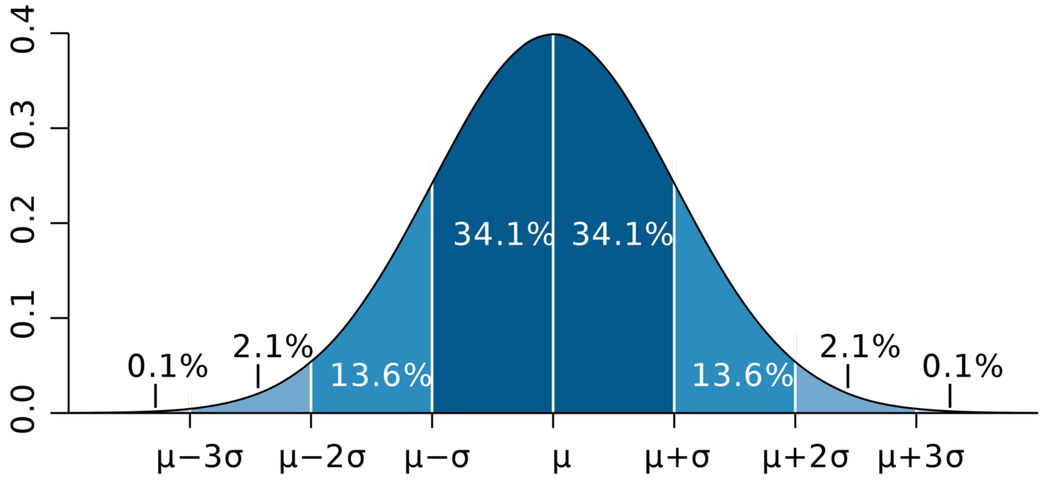
$$N = 18$$

$$\sum x_i = 80.0$$

$$M = \left(\sum x_i \right) \Big/ N = 80.0 / 18 = 4.4$$

Central Limit Theorem and Law of Large Numbers

- If X is taken independently from the same distribution, then X_i is said to be a random sample from that distribution
- X_i are said to be independent identically distributed (i.i.d.)
- **Law of large numbers** (LLN)- sample mean approaches population mean as n approaches infinity
- **Central limit theorem** (CLT) - the distribution of the sample mean approaches a normal distribution for n approaching infinity



Mean in neural data science

- Calculation in python
  - `import statistics`
  - `statistics.mean([data])`
- Application
  - DC or AC eeg?
    - How do you remove a DC bias?
  - Mean number of responses
  - Mean movement
  - Mean amplitude of oscillation in stroke, parkinson’s, etc patience
  - Where else do we see the mean in the brain or neural data science?

Mean, variance, standard deviation review

- Sample mean

- Population mean (“expected value”)

$$\mu = E(X_i)$$

Central tendency - Mode

- Most common number of a distribution
- Tells you which value has the highest frequency
- **What if there are ties?**
  - **More than one mode!**
  - **Which of the following is the mode?**

$$\{1,2,2,2,2,2,2,3,4,5,6,7,8,8,8,9,9\}$$

Mode in neural data science

- Calculation
  - `import statistics`
  - `statistics.mode(data)`
- Application
  - Some examples in NDS
  - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4059688/

Central tendency - Median

- The middle number of a distribution when the numbers have been ordered (sorted)
- Each score is counted separately, so if you have repeating scores such as 50 and 50, each one becomes part of the count
- Order the scores from low to high or high to low
- Count from both ends to the middle position

Central tendency - Median

- If odd number of scores, there will be one median

$$\{1,2,3,10,50\}$$

$$Median = 3$$

- If an even number of scores, count to the two closest to the middle (ie count from low towards high, high towards low) and take their average (add them up and divide by two)

$$\{1,2,2,3,3,4\}$$

$$2,3$$

$$Median = (2+3)/2 = 2.5$$

Median in neural data science

- Calculation
  - `import statistics`
  - `statistics.median(data)`
- Application
  - Median and MAD: The median and median absolute deviation (MAD)

$$x'_k = \frac{x_k - median}{MAD}$$

$$\text{where MAD} = \text{median}(|x_k - median|)$$

https://www.cs.ccu.edu.tw/~wylin/BA/Fusion_of_Biometrics_II.ppt

Mean, variance, standard deviation review

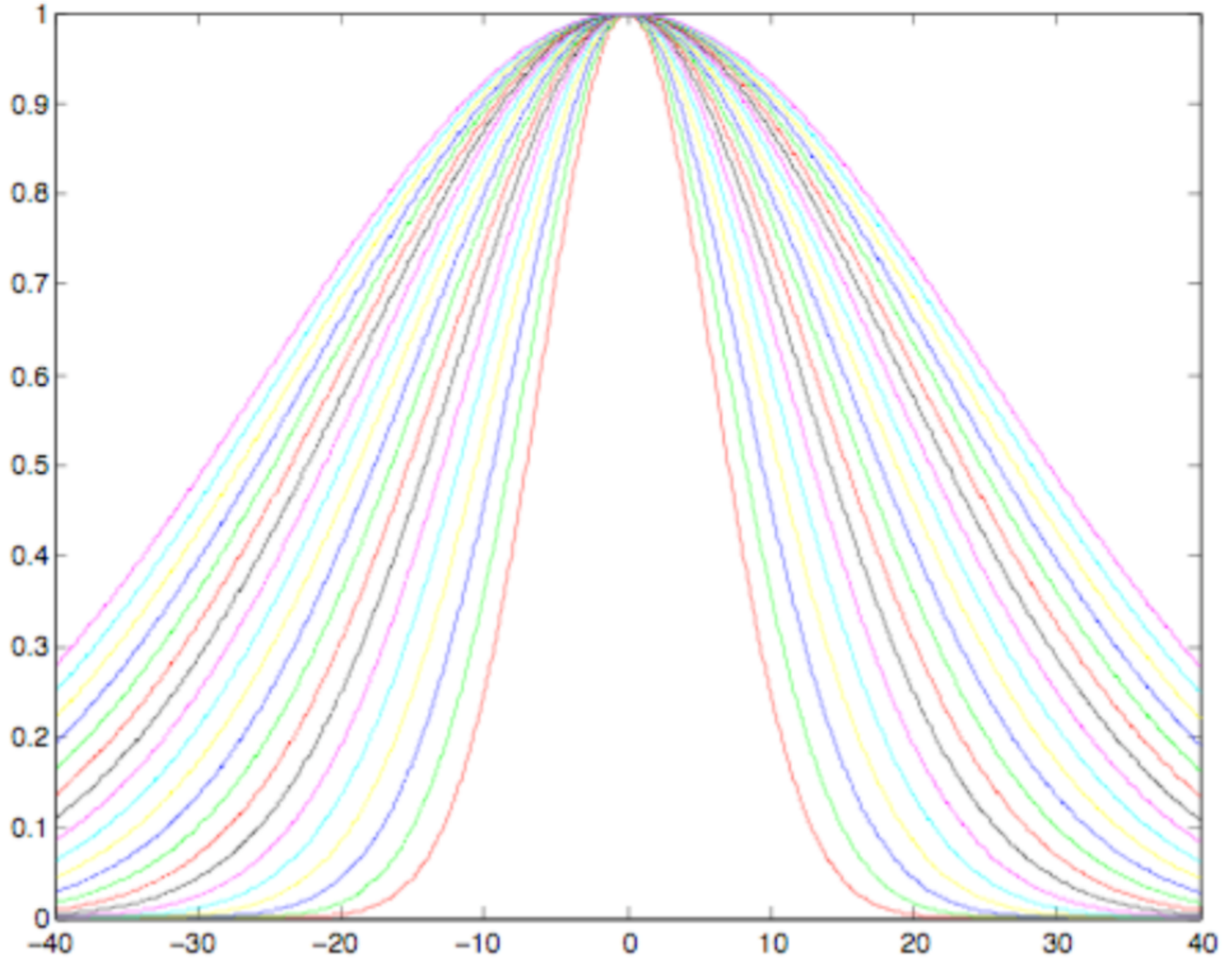
Standard deviation

How are they related?

- If you have a…
  - **Normal distribution,**
    - Mean=Median=Mode
  - **Symmetric distribution**
    - Median = Mean
  - **Skew distribution**
    - Median towards the body, mean towards the tail
      - **+skew: mean>median**
      - **-skew: mean<median**
- But this doesn't seem to be saying everything…

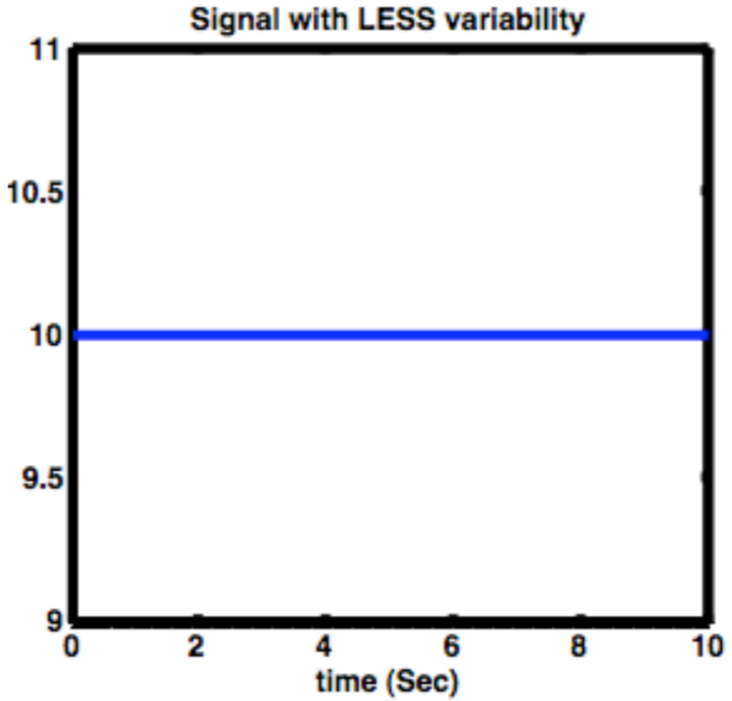
The mean isn't everything!

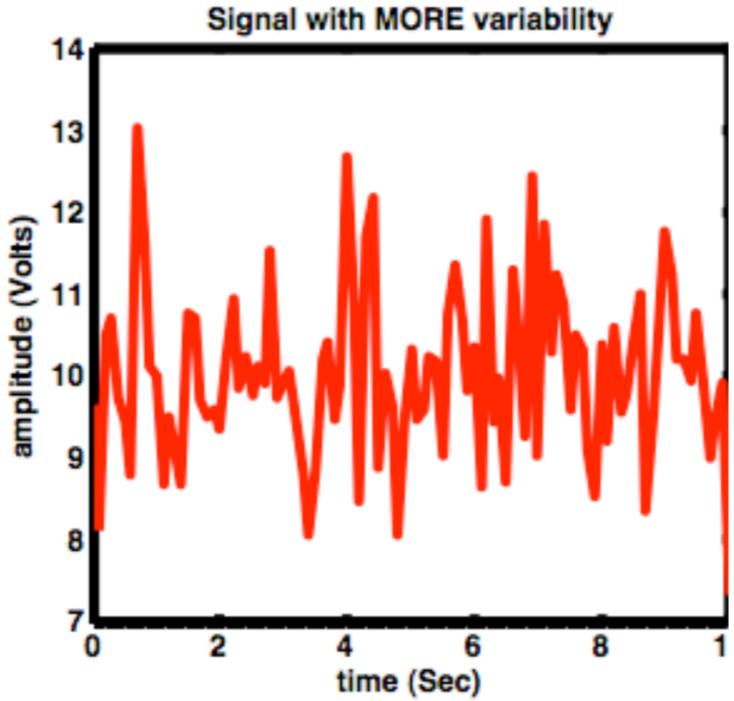
These all have the same mean



Why we need a measure of variability

Same means, different variability of the signal





We need a measure of Variability, here are a few…

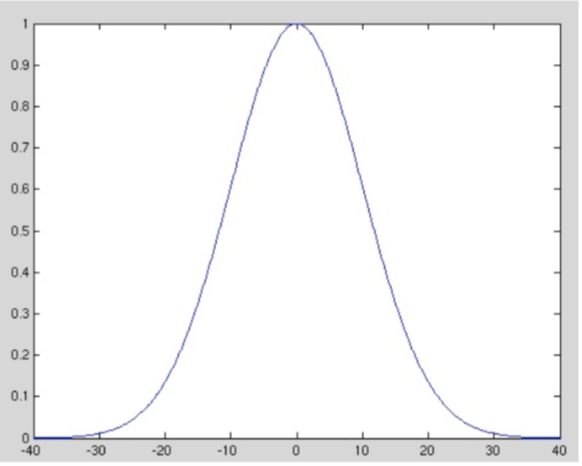
- Range
  - From math review, difference between max and min values of the data

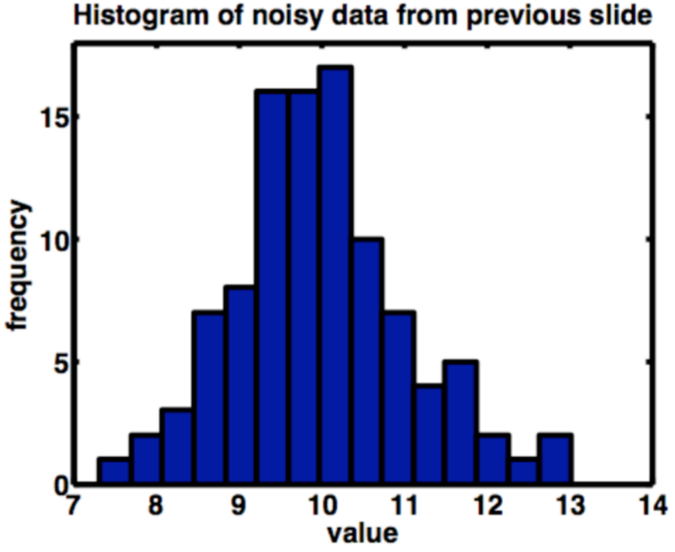
$$Range(x) = Max(x) - Min(x)$$

- Variance
  - Mean of squared deviations from the mean
  - In square units of the sample variable
- Standard deviation
  - Square root of variance
  - In units of the sample variable - sometimes easier to interpret

Returning to the normal distribution…and considering our data in terms of a histogram…

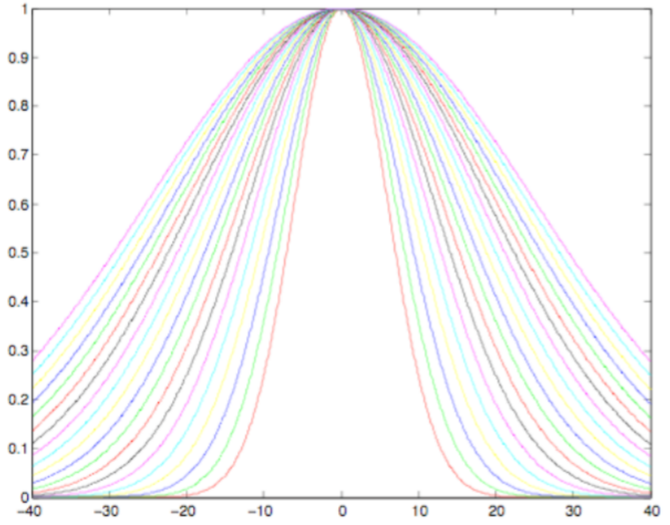
- The distribution of points about the mean can be considered in terms of probabilities
- How likely is a point to deviate from the mean?
- We call the normal distribution a *probability density function (PDF)* because it allows us to predict the likelihood that a sample will take on a particular value





Variance

- Whereas the mean defines a measure for the most likely point in state space (the center ‘location’ of a normal distribution)
- We can define the spread of the normal distribution about the mean by its *variance*



Variance (part II)

- Steps to compute the variance
  - **Compute the deviations from the mean for all the data**

$$d_i = \left(x_i - \bar{x}\right)$$

  - **Compute the square of each of the deviations**

$$sd_i = \left(d_i\right)^2$$

  - **Sum up all these squared deviations**

$$ssqd = \sum_{i=1}^{N} \left(sd_i\right)$$

  - **Divide the mean squared deviations by N, the number of observations**

$$Var = \frac{ssqd}{N}$$

Standard Deviation

- Typical ‘deviation’ from the mean
- Ie how far on average scores depart on either side from the mean
- Easy to compute after the variance - just take the square root of the variance

$$SD = \sqrt{Var} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{N}}$$

$$\bar{x} = \frac{\sum x_i}{N}$$

Z scores

- A Z score is simply a measure of how many standard deviations away from the mean a score is
- Units are standard deviations

$$Z_i = \frac{X_i - \mu}{SD}$$

Covariance

- Covariance is very commonly used in statistical analysis as the basis for advanced statistics
- Gives a quantitative measure of the relationship between two variables

$$Cov(X,Y) = E\left[\left(X - \mu_x\right)\left(Y - \mu_y\right)^T\right]$$

$E = \text{exp}ectation$

$\mu = mean$

More Covariance

- If the two variables are independent, the covariance is 0
  - **(BUT IF COVARIANCE IS 0 THAT DOESN'T MEAN THE VARIABLES ARE INDEPENDENT!!!)**
- If they are totally dependent the covariance of data, can be arbitrarily large
  - **(AGAIN THE CONVERSE IS NOT NECESSARILY TRUE)**
- The diagonals are the variance of each variable
- If each row is an observation, and each column a variable…

$$\text{cov}(X) = \left(\frac{1}{N-1}\right)\left(X - mean(X)\right)\left(X - mean(X)\right)^T$$

Correlation coefficient motivation

- We want to define a measure of how related our dependent and independent variables are
  - Variance, STD – variation of a single variable
  - Covariance – how two things vary in relation to each other
  - How do we compute the linear dependence of one variable to another?
- Correlation coefficient!

Intuitive arrival at the Correlation Coefficient

- Many kinds (we are going to discuss Pearson's product moment coefficient by Galton)
- A test for linear independence
- We want to measure how two things co-vary
  - We observe one thing varying (e.g. sunset)
  - We observe another thing varying (e.g. air temp. decrease)

Intuitive arrival at the correlation coefficient (II)

- **Positive Correlation** - When one thing's magnitude varies positively, and another thing's magnitude varies positively
  - **and if both vary negatively, also this is referred to as positive correlation**
- **Negative correlation** - When one thing's magnitude varies positively, and another thing's magnitude varies negatively
  - **And if one varies positively while the other varies negatively, this is also referred to as negative correlation**

Intuitive arrival at the correlation coefficient (III)

- We want our measure to be a single number
- In some way we'll need to scale the calculations so that the number is unitless
  - **The variables we're comparing may be in different units**
  - **We also don't care about bias - we're interested in variations, so we make our measures about zero, and normalize each**
  - **Remember when we presented z-scores as a normalized measure of how far from the mean a particular sample is in a dataset?**

$$Z_i = \frac{X_i - \mu}{SD}$$

Intuitive arrival at the correlation coefficient (IV)

- We arrive at the correlation coefficient by multiplying each z-score from one variable by the z-score from the other variable, then averaging all those results
  - **Thus if both tend to vary positively?**
    - Positive correlation
  - **If both tend to vary negatively?**
    - Positive correlation
  - **If one varies positively, and the other negatively?**
    - Negative correlation
  - **If sometimes they both vary positively or negatively, sometimes they vary oppositely?**
    - Small or near zero correlation

Correlation coefficient

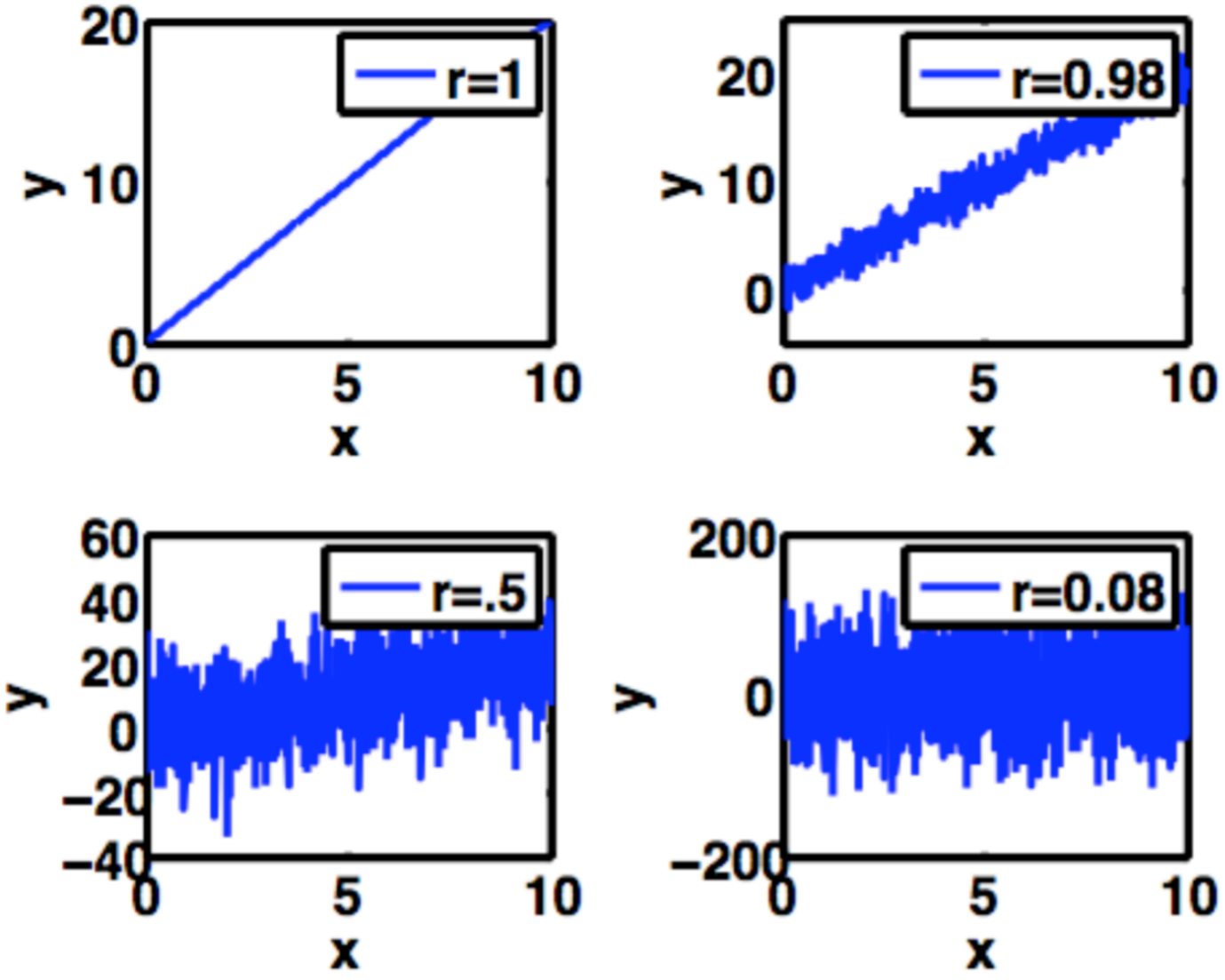
$$\rho(j,k) = \frac{\sum_{i=1}^{N} Z_{ij} Z_{ik}}{N}$$

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

$$\rho(X,Y) = \frac{E((X-\mu_X)(Y-\mu_Y))}{\sigma_X \sigma_Y}$$

Characteristics

- Range
  - **-1<=r<=1**
- Interpretation - independence
  - **Statistical independence**
    - The more distinct and unrelated the covariation, the closer to zero the correlation coefficient
    - Statistically independent if their correlation is zero
  - **Linear independence**
    - Two things varying perfectly together are linearly dependent, variables with less than perfect correlation are linearly independent


r=1
r=0.98
r=.5
r=0.08
x
y
0
5
10
20
60
40
−20
−40
200
−200

On to today…

Math and symbol review

- http://casimpkinsjr.radiantdolphinpress.com/pages/cogs138_sp23/handouts/greek_letters_review.pdf
- http://casimpkinsjr.radiantdolphinpress.com/pages/cogs138_sp23/handouts/math_review.pdf
- Handouts page on website:
  - http://casimpkinsjr.radiantdolphinpress.com/pages/cogs138_sp23/handouts.html

Demo in python…
Central tendency for neural data

Python docs on statistics

- Individual stats:
  - https://docs.python.org/3/library/statistics.html
- Comparisons:
  - https://github.com/drsimpkins-teaching/cogs138/blob/main/Tutorials-master/12-StatisticalComparisons.ipynb

Correlations and pitfalls in neural data science

Perspective taking, painting a picture with data, pitfalls on both sides

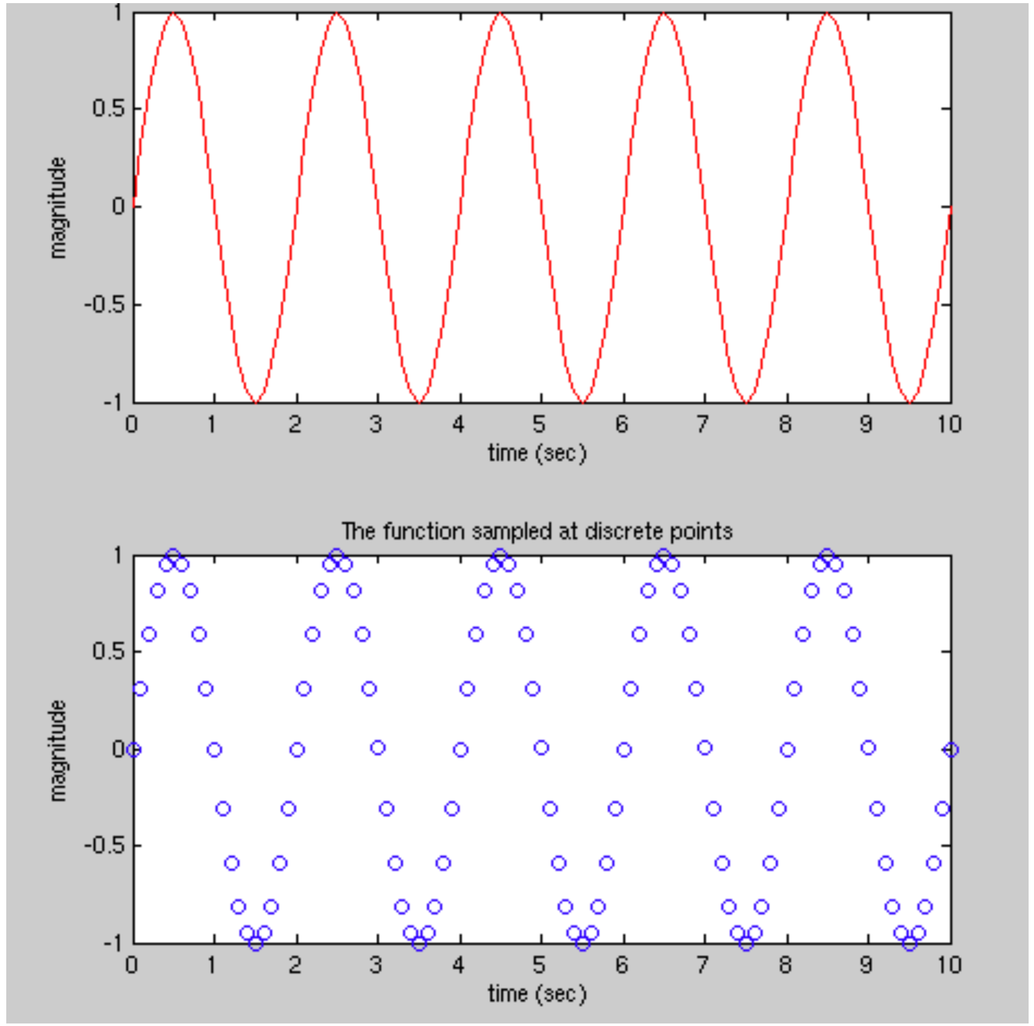
- Churchland's paper(s)
  - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3393826/
  - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6908571/
- Criticisms of Churchland's paper
  - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6908571/

More on sampling, discretization, filtering

- Review of continuous vs. discrete quantities
- Analog vs. Digital
- Discretization, sampling, aliasing
- Filter theory, frequency response, filter types
- Linearity

Continuous vs. Discrete quantities

- Information storage
  - **Continuous** signals have information at every point in time
  - **Discrete** signals have info only at specified intervals (fixed or variable)



Examples of continuous and discrete systems

- Continuous or discrete?
  - –# of people in this class •Discrete
  - –# of Time zones •Discrete
  - –Time •Continuous
  - –Answers on multiple choice tests •Discrete
  - –A Sound •Continuous
  - –Body temperature •Continuous

Analog vs. Digital quantities

- Information storage
  - **Analog** contains infinite information
  - **Digital** contains limited information, depending on the number of bits of information the digital value can store
    - 0 or 1 in each bit means each bit multiplies the possible combinations of numbers by 2
    - 2^4 = 0-15 (a 4-bit number, 16 different values)
    - 2^8 = 0-255 (an 8-bit number, 256 different values)
    - 2^16 = 0-65535 (a 16-bit number, 65536 different values)

More on digital quantities

- Measuring an EEG boils down to recording a sequence of numbers into computer memory, stored in values of a specific size, such as 8 bit numbers.
  - i.e. signal is 0-5V, digitized with 8 bit ***precision*** would yield a ***resolution*** of 5V/256 = 0.020V, or 20mV (mV = 'milli-Volts')
  - **Resolution -** defined as the smallest quantity which can be reliably measured
  - **Digital Precision -** The number of bits of information contained in a digital quantity
- Also important for computations
  - Round off errors can accumulate
    - Example
      - 2.245+3.432+1.234 = 6.911
      - 2+3+1 = 6, and that's only 3 samples! Imagine 1000/sec (1kHz) !
  - More on this later

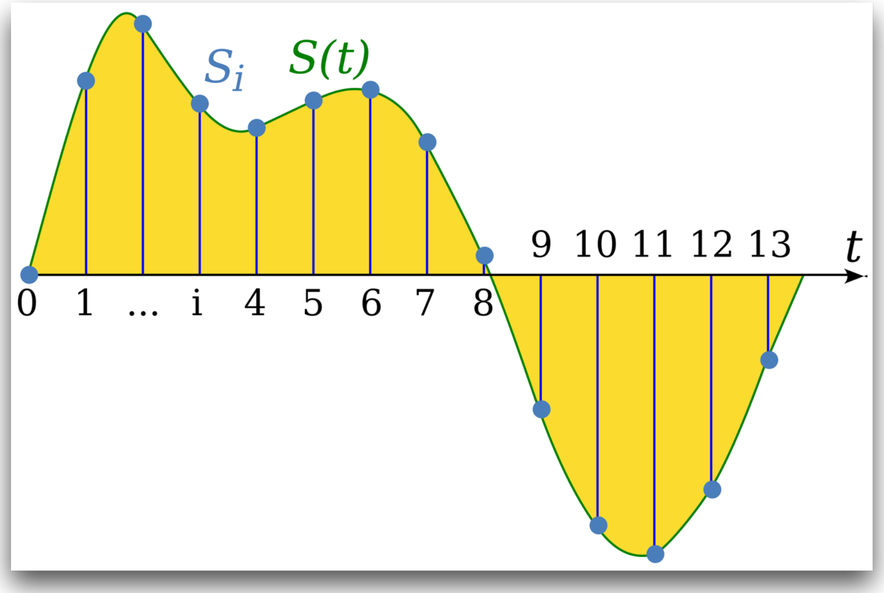
Discretization

- Measuring a continuous (analog) signal means capturing information at specified (fixed or variable) intervals
  - **Sampling frequency -** the frequency at which data is recorded from a signal (Typically in Hz, ie 5kHz)
- When capturing data, or when manipulating data which has been discretized, there are several issues to consider
  - Aliasing (not the TV show:)
  - Sampling rates
  - Post-processing – filtering data to remove unwanted information while retaining desired information

Sampling

- **Sample -** We record data at specific points in time
- **Period -** The time between samples, ***T*** **[sec]**
- **Sample frequency -** The frequency of sampling, *f* [Hz]

$$f = \frac{1}{\Delta T}$$



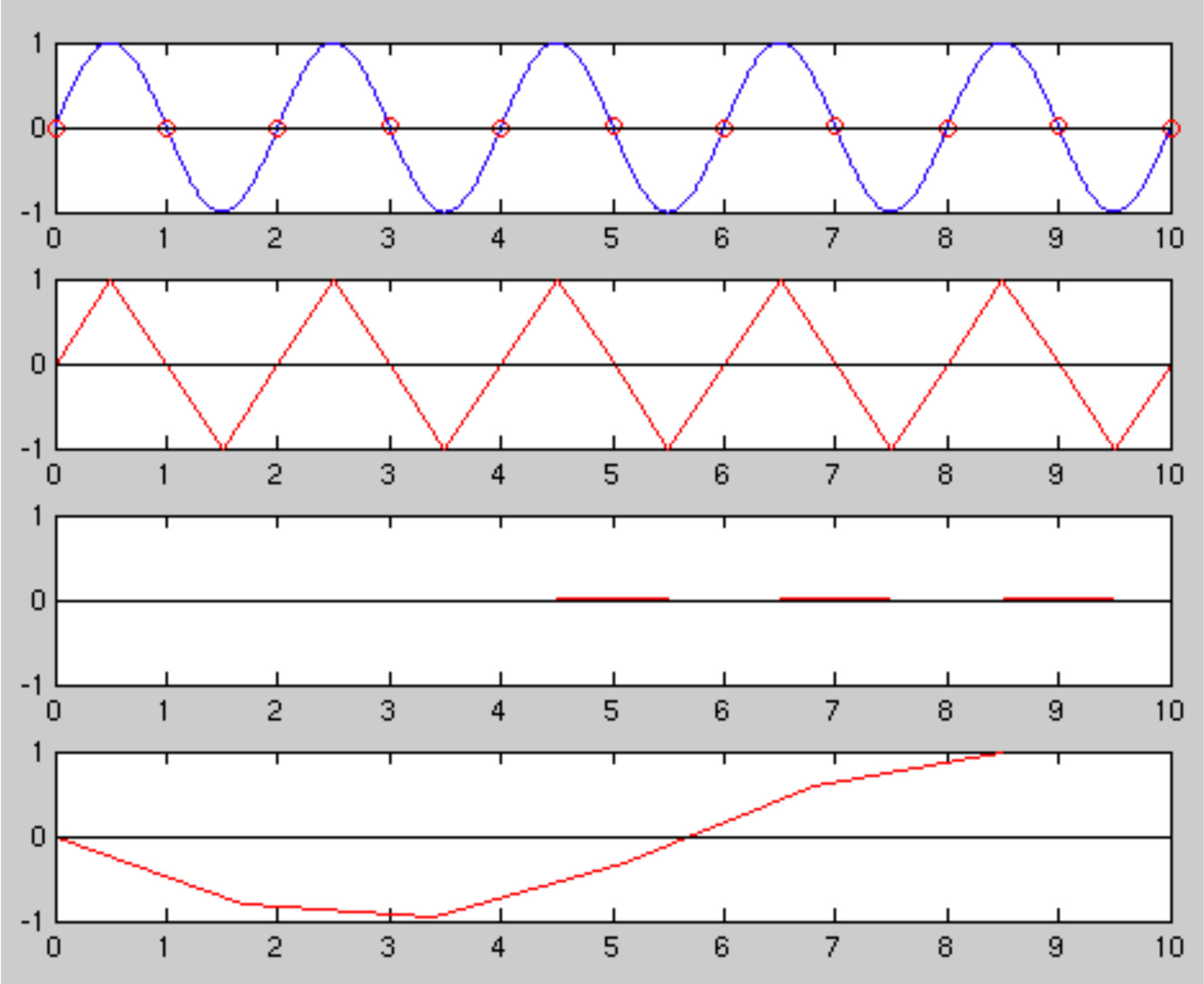
https://en.wikipedia.org/wiki/Sampling_%28signal_processing%29

Nyquist and Sampling

- Stories
  - Running in the dark with periodic lights on the ground, with sharp turns
  - Ping pong (no sound, periodic view of the system)
- As a rule of thumb, you must sample AT LEAST twice as fast as the highest frequency you want to measure
  - **Nyquist frequency -** max freq. that can be measured [Hz]
  - **Nyquist rate -** sampling frequency (which is 2x the nyquist frequency) required to sample at the nyquist frequency
  - 20 times as fast is better
  - Filter out higher frequency components

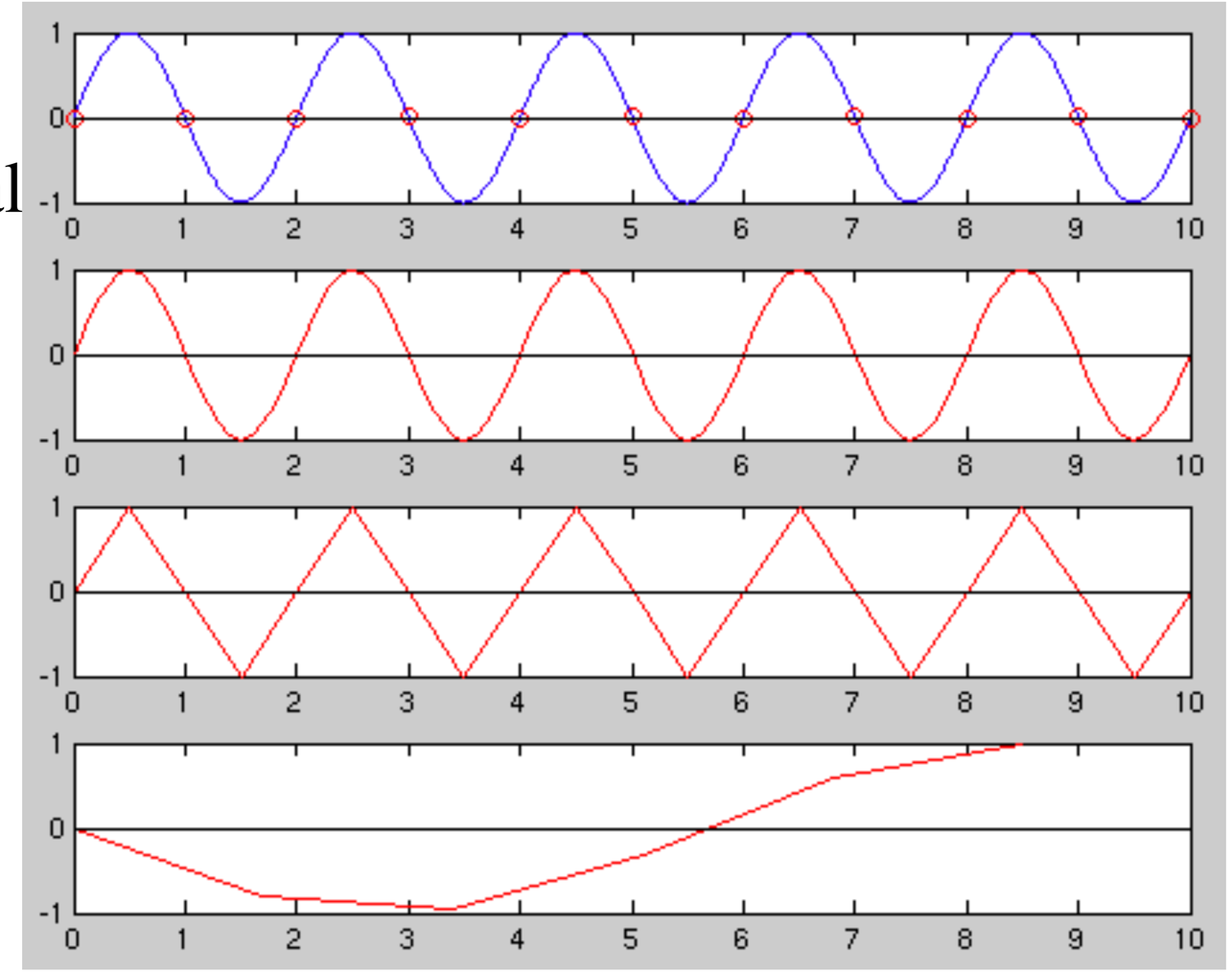
Nyquist frequency

$$f_n = \frac{1}{2\Delta T}$$


1
0
-1
0 1 2 3 4 5 6 7 8 9 10
1
0
-1
0 1 2 3 4 5 6 7 8 9 10
1
0
-1
0 1 2 3 4 5 6 7 8 9 10
1
0
-1
0 1 2 3 4 5 6 7 8 9 10

What do we see in this picture?

- **Aliasing -** the corrupting of a signal by components of higher frequencies overlapping into the lower frequency



How do we solve this?

- Filter out the frequencies we don’t want
  - Low pass filter
  - High pass filter

Examples: Visual discretization

- Color shading

- Color and visual boundaries:

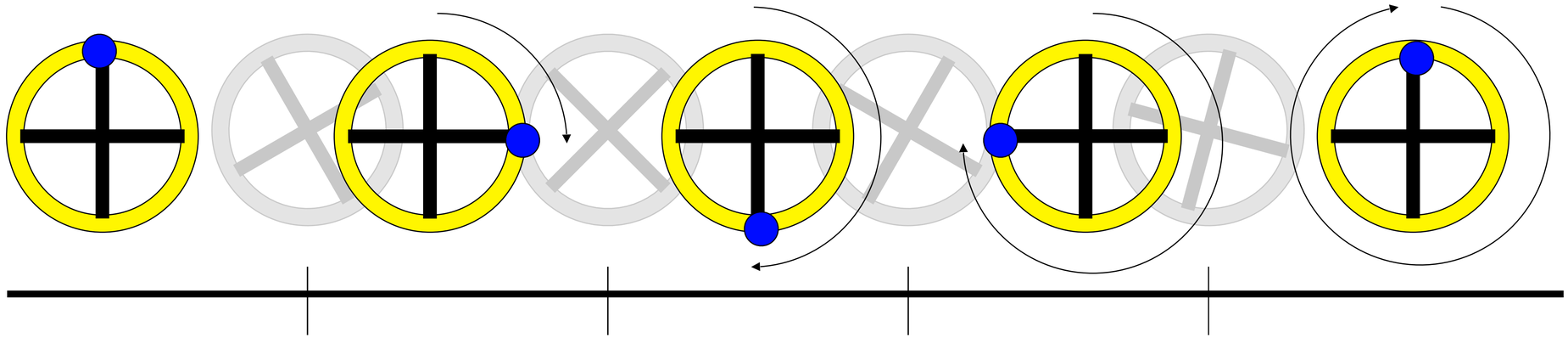
Few colors and low spatial resolution

Low spatial resolution only

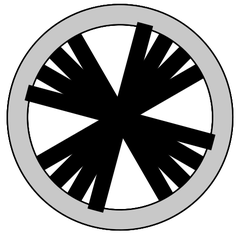
High spatial resolution and colors

Example: Sampling and Aliasing

- The wheel spokes example…<Live demo>



- We're sampling at too slow a rate to accurately see the spokes rotate, and at a particular rotational velocity of the wheel, we see an 'aliased' reverse rotation!

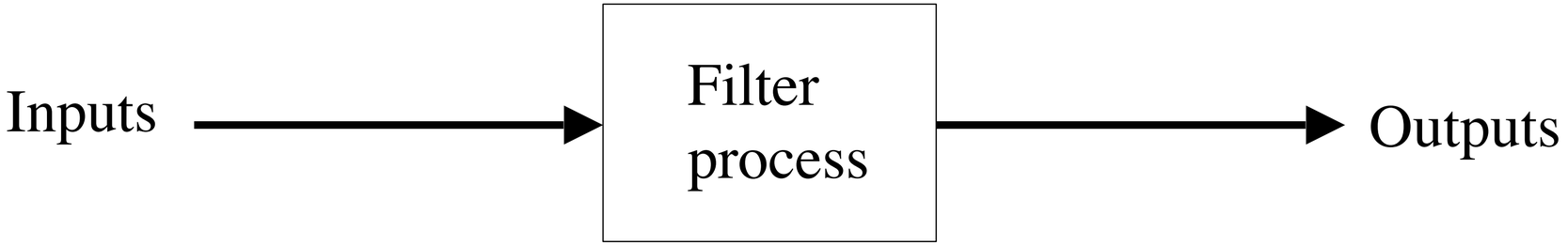


Obviously aliasing can be bad…

- Aliasing can lead to improper interpretations of data
  - **So what do we do about it?**
    - We must first sample at twice the rate of the fastest signal we care about
    - Filter our data (humans do this, and so do cognitive scientists!)

Thus we filter our data…

- **Filter -** an operation or process which alters input data according to some mathematical relationship or heuristic rule to produce output data which is more desirable



Computational filtering

- *Noisy auditory data can be filtered to remove undesired signals*
- *EEG signals can be filtered to remove 60Hz noise from AC lines nearby*
- *Other sensor signals can be filtered to improve results*

Frequency Response

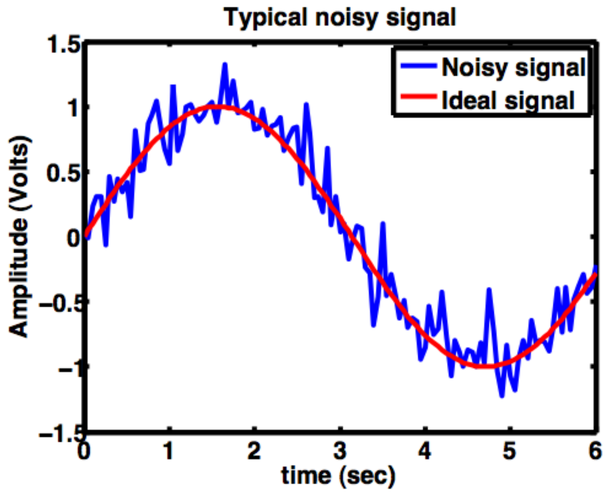
- Linearity of systems vs. nonlinearity
- The response of a linear system to a sinusoidal input is a sinusoidal output with the amplitude and phase shifted in some way
- This is useful for characterizing the behavior of some signal over a range of possible input frequencies
- Example with the chalk

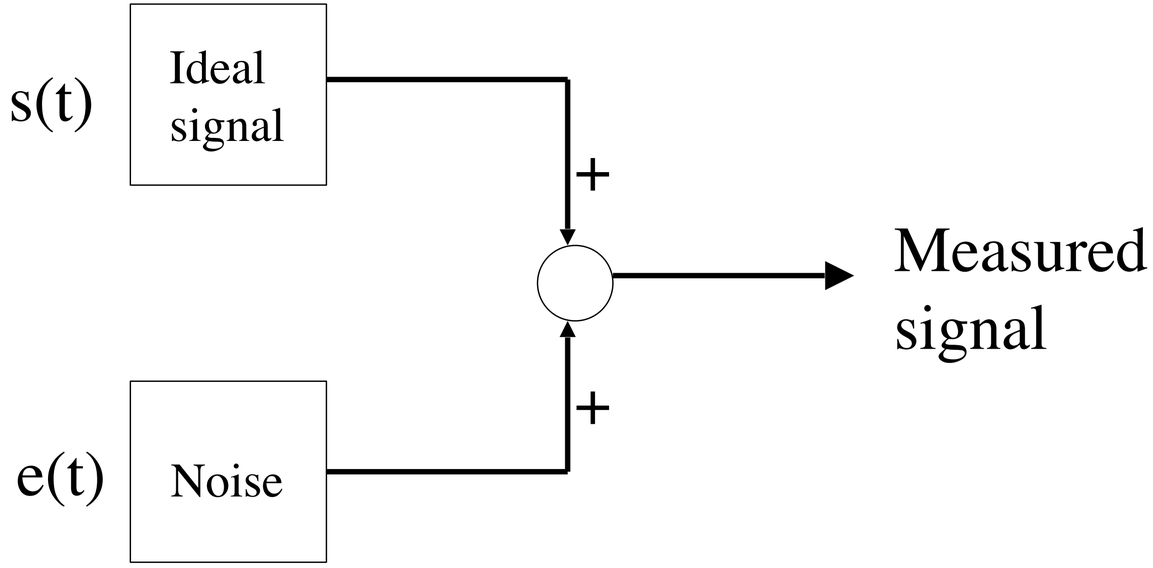
Common filter types in signal processing

- **Low-pass filter** - (ideal) attenuates high frequency data, while allowing low frequency data to pass unchanged
- **High-pass filter** - (ideal) attenuates low frequency data, while allowing high frequency data to pass unchanged
- **Band-pass filter** - (ideal) attenuates all frequencies except a particular frequency band (or bands)
- **Band-stop filter** - (ideal) attenuates one or a selection of frequency ranges of data, allowing all the rest to pass unchanged
- Actual filters are not exactly ideal…which we will discuss

Signals and noise…

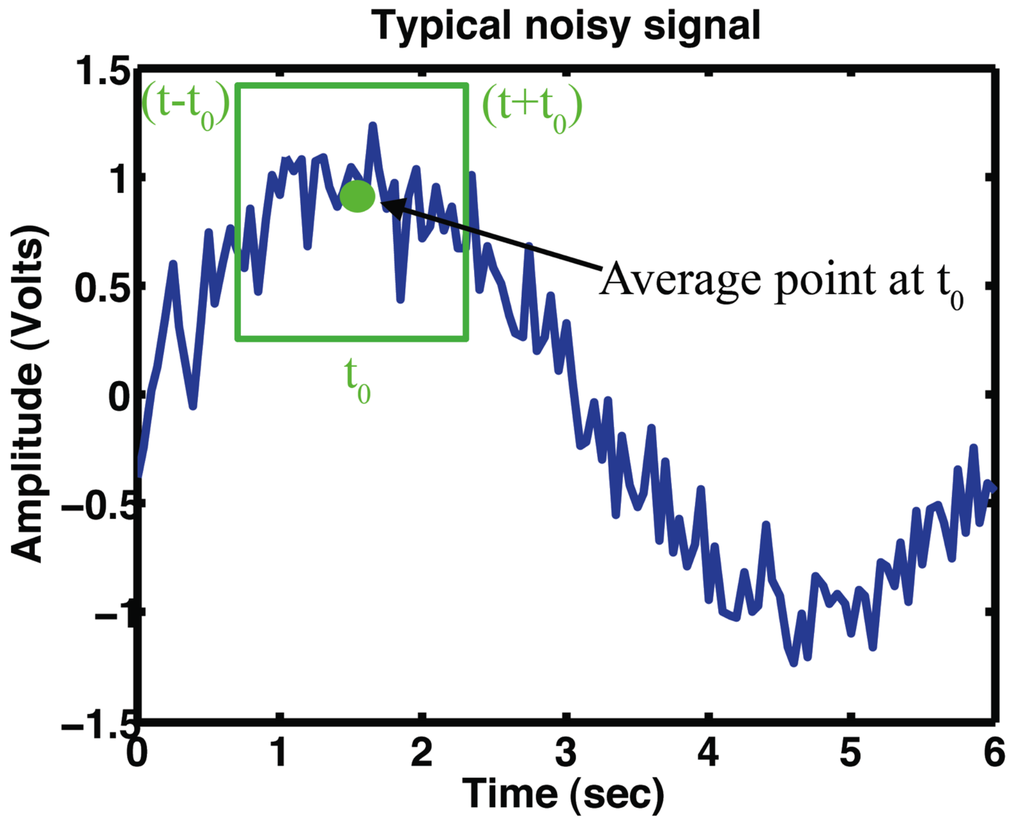
- By making assumptions about the properties of the unwanted 'noise' *e(t),* we can reconstruct an appropriate *estimate* of the original signal *s(t)*
  - **Noise - any unwanted portion of a signal, lumped together. It may come from multiple sources but tends toward some statistically predictable properties**





Low-pass filtering

- So the effect is this



More on linearity vs. nonlinearity

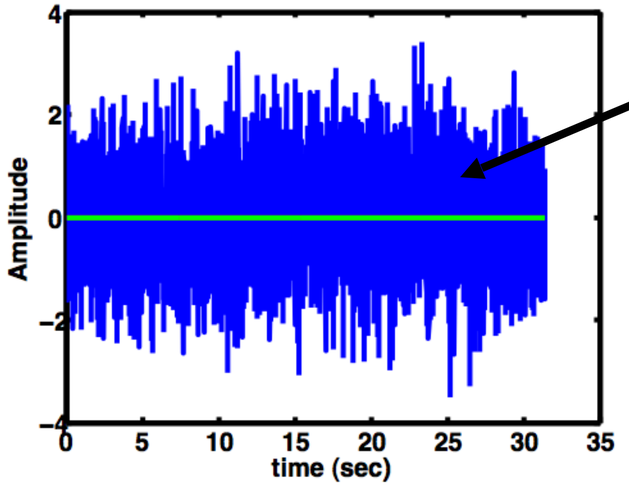
- Power
  - **A linear system is a system whose dependent variables are related to its independent variables by a power of one**
- Linear systems have these particular properties (and they are very favorable)
  - **Additive** $$T[x_1(n) + x_2(n)] = T[x_1(n)] + T[x_2(n)]$$
  - **Homogeneous** $$T[cx(n)] = cT[x(n)]$$
- Linear differential equations are more well-understood than nonlinear differential equations

Fourier transforms

- Frequency domain example : Musical note vs. the sound
  - **More parsimonious to describe a song in terms of its notes than time domain signal (when creating a ‘model’ for a song which can be communicated)**

We return to noisy data which we want to ‘clean up’

- We do this by removing undesired components of the signal
- One way to do this is *averaging* out the noise
- If it’s Gaussian and additive…

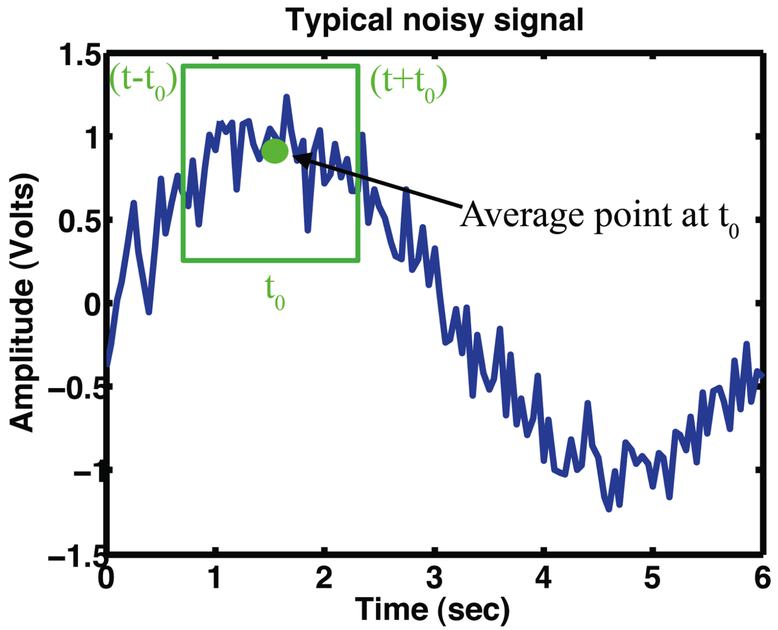


This is gaussian noise, and the average of this is approximately the green line, 0

$$-5 + 5 = 0$$

How to do it

- **Decide on a ‘window’ of data to average over, which is narrower than the fastest component to your changing signal**
- **Sum up over that window of points and divide by the number of points (average)**



Continuous form

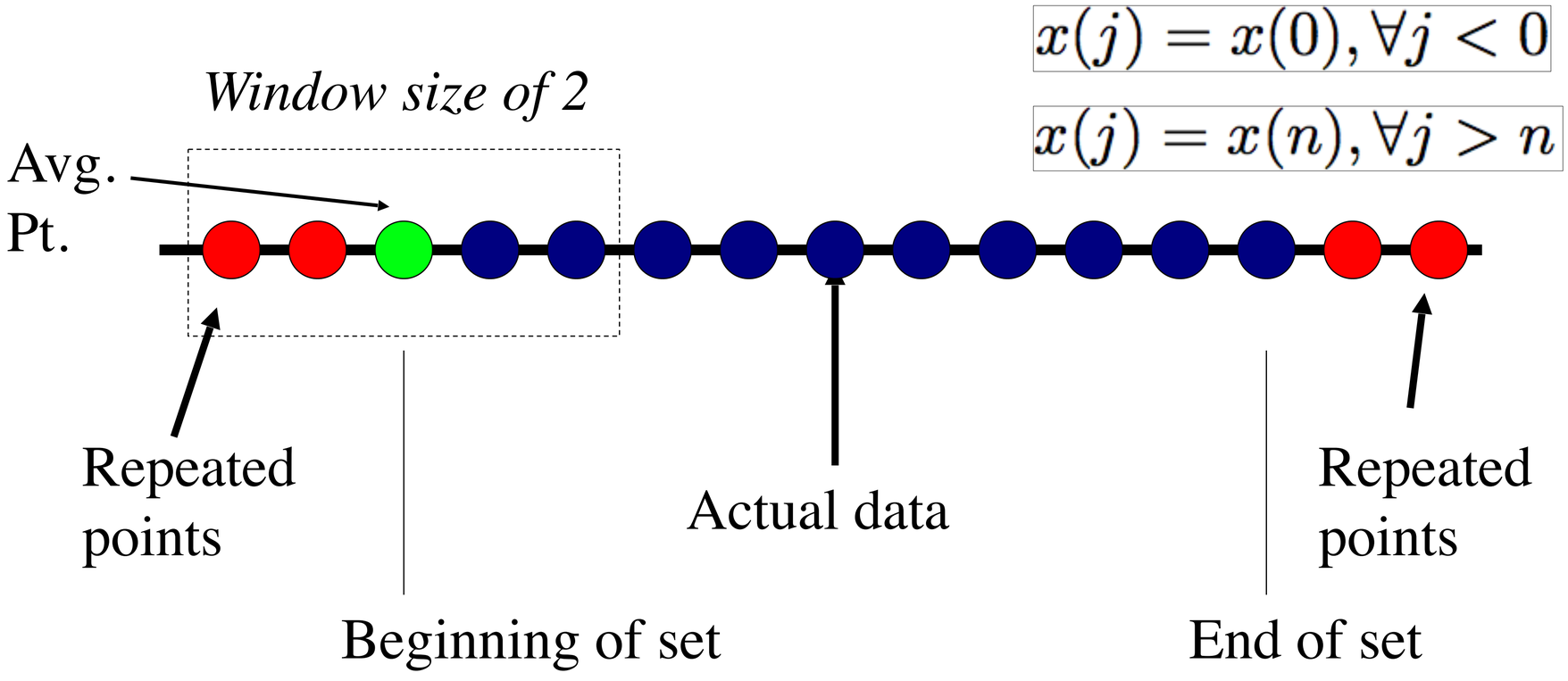
$$x_f(t) = \int_{t-t_0}^{t+t_0} x(\tau)d\tau$$

Discrete form

$$x_f(i) = \frac{1}{2k+1} \sum_{j=i-k}^{i+k} x(j)$$

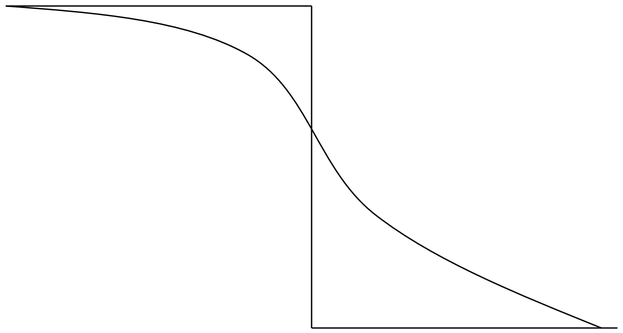
A few details

- What about at the ends of the data where we don't have information before (at the beginning of the data set) or after (at the end of the data set)?
  - **Copy the first or last point and repeat as necessary**



Disadvantages…

- Need to have all data in memory already, so it isn't an 'online' filter
- Causality
  - **If we care about an exact event timing, this is a poor filter to use:**



Signal anticipates changes!

Solution

- Recursive filter
  - Solves causality issue
  - Easy to implement as we saw last time

Field potential data

- let’s do some loading processing and filtering of field potential data