ML-Based COVID-19 Pathogen Classification with Genomics Signatures & Chaos-Inspired Methods

By: Esha Ananth & Anoushka Bhat

https://tinyurl.com/SLIDES99

Project Overview

Breakdown of Project Goal for Genomic Signatures

Pathogen Classification

Finding the “ID” of a certain virus

Taxonomy: method of classification

Domain → Kingdom → ……. Family → Genus → Species

Example: Location

Method of Classification: Scale

Planet → Continent → Country → Province/State → City

Earth → North America → U.S. → California

Project Goal

Use Machine Learning models to find an **classification for COVID-19** through the use of genomic signatures

Genomic signatures

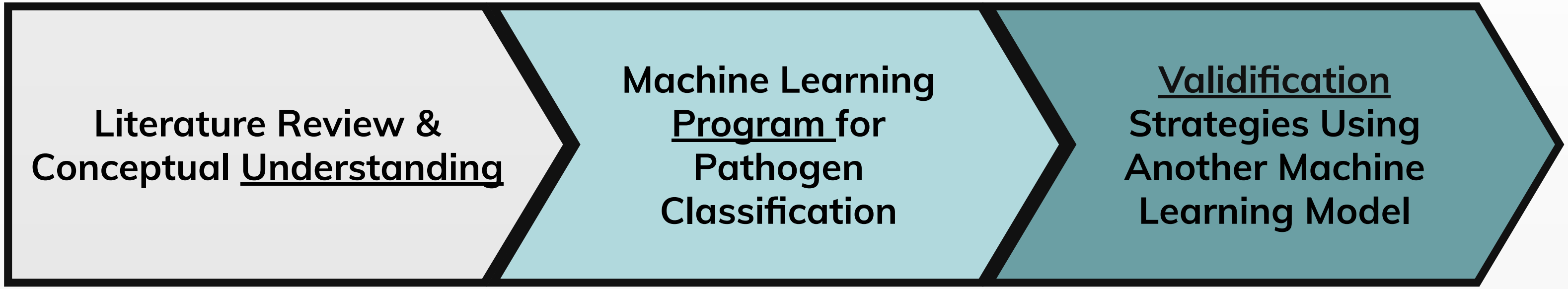
Genomic signature: frequency of oligonucleotides in a dna sequence

Oligonucleotides: synthetically made short DNA sequences (8-50 base pairs) that bind to complementary DNA of the species being tested

- Genomic signature = amount of the oligonucleotide that binds to the number of base pairs in the DNA sequence divided by the total number of base pairs

Since each genomic signature is unique for a species, it can be used to differentiate and classify between species

Project Plan



Literature Review & Conceptual Understanding

Machine Learning Program for Pathogen Classification

Validification Strategies Using Another Machine Learning Model

Literature Review

Research already existing on COVID-19 Classification

PLOS paper: COVID-19 case study

Research Objective: Use a machine learning tool to find the taxonomy of COVID-19 genomes through its genomic signatures

Procedure:

1. Turn genomes into numerical values, apply Discrete Fourier Transforms
2. Find Pairwise Distance Matrix between 2 magnitude spectra using Pearson's Correlation Coefficient
3. Run genomes through the machine learning with digital signal processing (MLDSP)
    a. Do 10 fold cross validation by splitting the data into testing and training data
    b. Apply to 6 classification models: Subspace KNN, Average Accuracy,Fine KNN, Linear Discriminant, Quadratic SVM, Linear SVM
4. Validate results using Spearman's Rank Correlation Coefficient

Results: all 29 COVID-19 sequences ran through tool classified the virus as *sarbecovirus*

- closely relating to SARS-CoV-2 & 2 coronaviruses found in bats

Machine Learning with Digital Signal Processing

Taking a closer look at the machine learning tool used in the COVID-19 case study

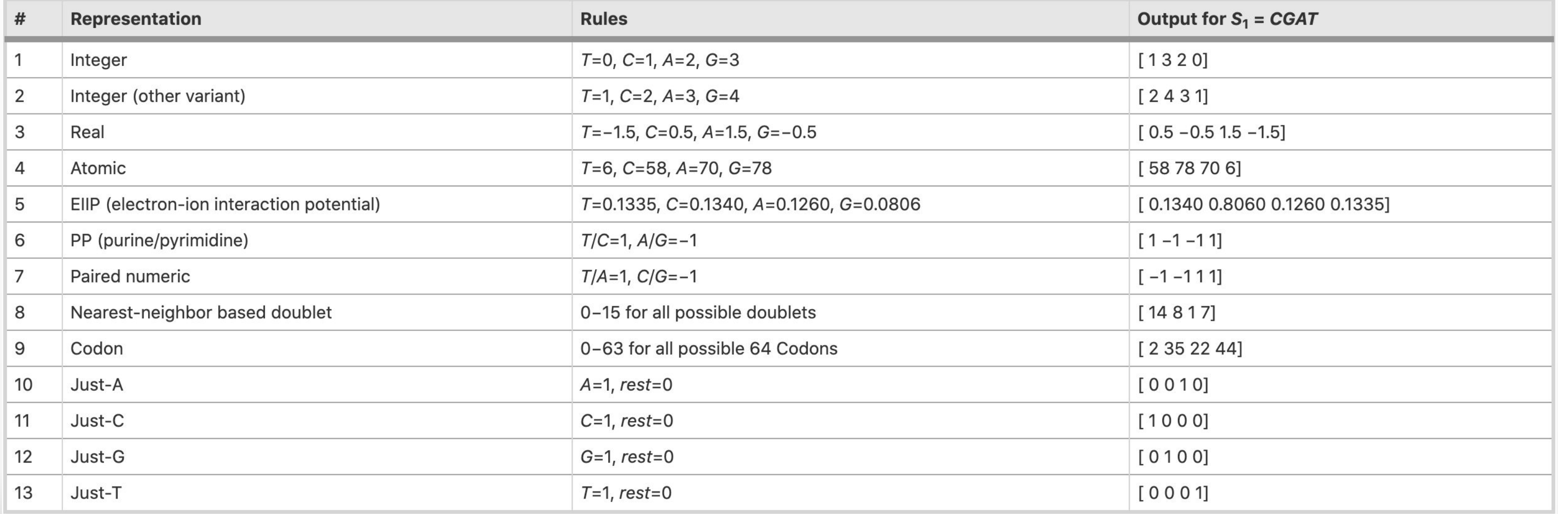
Why MLDSP?

- Alignment free = not comparing genomes through exact alignment
  - better measure of similarity
- Compares data using feature vectors
  - faster processing time, especially with larger data sets

Preparing Data for MLDSP

MLDSP was tested with training data prior to the COVID-19 pandemic

1. Express the Genomic Sequence Numerically
   - best methods were “PP”, “Just A”, and “Real”

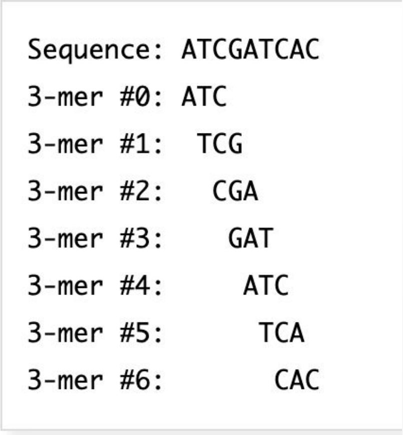


#	Representation	Rules	Output for S_1 = *CGAT*
1	Integer	*T*=0, *C*=1, *A*=2, *G*=3	[1 3 2 0]
2	Integer (other variant)	*T*=1, *C*=2, *A*=3, *G*=4	[2 4 3 1]
3	Real	*T*=−1.5, *C*=0.5, *A*=1.5, *G*=−0.5	[0.5 −0.5 1.5 −1.5]
4	Atomic	*T*=6, *C*=58, *A*=70, *G*=78	[58 78 70 6]
5	EIIP (electron-ion interaction potential)	*T*=0.1335, *C*=0.1340, *A*=0.1260, *G*=0.0806	[0.1340 0.8060 0.1260 0.1335]
6	PP (purine/pyrimidine)	*T*/*C*=1, *A*/*G*=−1	[1 −1 −1 1]
7	Paired numeric	*T*/*A*=1, *C*/*G*=−1	[−1 −1 1 1]
8	Nearest-neighbor based doublet	0−15 for all possible doublets	[14 8 1 7]
9	Codon	0−63 for all possible 64 Codons	[2 35 22 44]
10	Just-A	*A*=1, *rest*=0	[0 0 1 0]
11	Just-C	*C*=1, *rest*=0	[1 0 0 0]
12	Just-G	*G*=1, *rest*=0	[0 1 0 0]
13	Just-T	*T*=1, *rest*=0	[0 0 0 1]

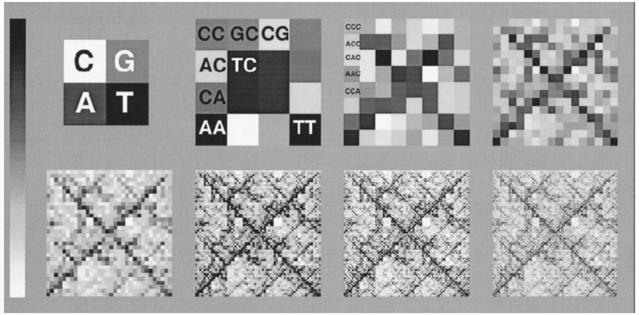
Preparing Data for MLDSP (cont)

Alternately, in the COVID-19 case study, the genomic sequence was represented in 2D

- 2D Chaos Game Representation obtained through the use of k-mer values



```
Sequence: ATCGATCAC
3-mer #0: ATC
3-mer #1:  TCG
3-mer #2:   CGA
3-mer #3:    GAT
3-mer #4:     ATC
3-mer #5:      TCA
3-mer #6:       CAC
```



k-mer values: short sections of the larger genome

2D Chaos Game Representation: takes k-mer values in 2D format to create fractal image

- numerical representation = point corresponding to k-mer on plane

We have not tested this numerical representation yet.

Length Normalization

1. Before you can do Discrete Fourier Transform, or Pearson’s Correlation, you need to make the lengths of the two sequences that you are comparing the same

2. You can do that by either using minimum length, zero padding, or antisymmetric padding(which yielded the most accurate results)

Initial Sequence 1 = [1, 2, 3, 4]	Final Sequence 1 = [-2, -1, 1, 2, 3, 4, -4, -3]
Initial Sequence 2 = [1,4, 3, 2, 4, 3, 2, 2]	Final Sequence 2 = [1, 4, 3, 2, 4, 3, 2, 2]

Preparing Data for MLDSP (cont)

2. Calculate a “feature” vector using Discrete Fourier Transform (DFT)

- DFT used to find the frequency of samples in a function
  - frequency of nucleotide **signals** in the given genomic sequence

Equation for DFT

$$F_i(k) = \sum_{j=0}^{p-1} f\left(S_i(j)\right) \cdot e^{(-2\pi i/p)kj}$$

3. Find the magnitude of the feature vector (absolute value)

- reflects nucleotide **distribution** in original genome sequence

Preparing Data for MLDSP (cont)

4. Use Pearson's Correlation Coefficient to find the correlation between 2 sequences

- measured between -1 and 1, with negative PCC indicating negative correlation
- absolute value of PCC closer to 1 = stronger correlation

Significance of PCC in context

- Measures the "distance" between two different genomic signatures
- Less distance = more correlated

P-Value

In our code, the PCC function gave two values: the PCC & a p-value

- P-value tells if the correlation value is statistically significant or occured to chance
- P-value < 0.05 = statistically significant

10 fold cross validation using MLDSP

1. Choose a test to work with
2. Let’s say you choose test 4 which contains datsets:

   Alphacoronavirus, Betacoronavirus, Gammacoronavirus, Deltacoronavirus

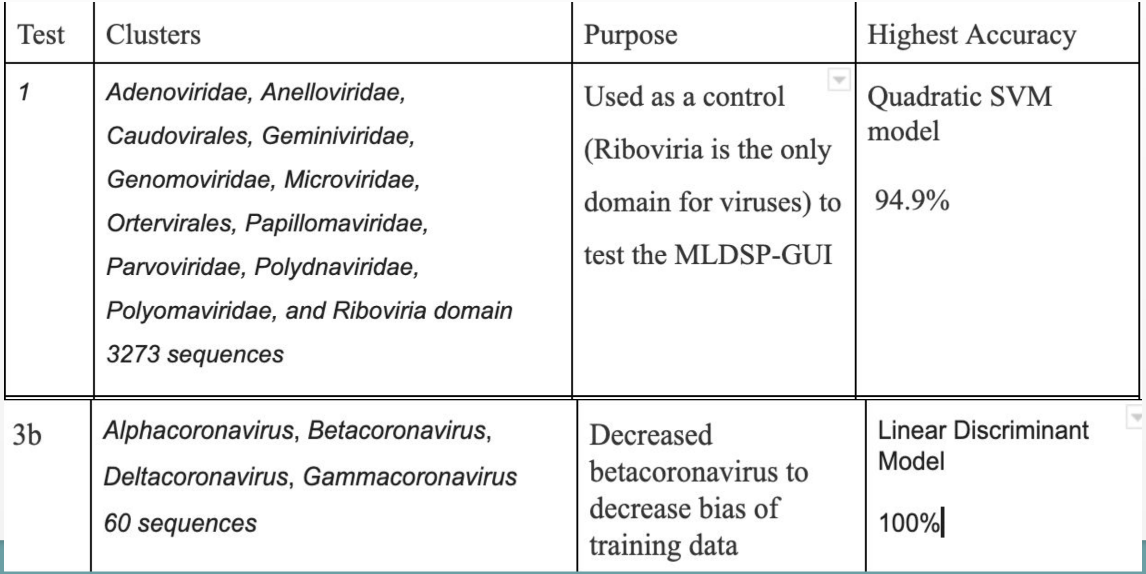
3. Shuffle all the data in the test
4. Split the data into 10 groups
5. Choose a classification model
6. Train the data on 9 groups, test on 1 group
7. Average all the accuracies from the 10 tests from the specific model

Training Data	Testing Data
1, 2, 3, 4, 5, 6, 7, 8, 9	10
1, 2, 3, 4, 5, 6, 7, 8, 10	9
1, 2, 3, 4, 5, 6, 7, 9, 10	8
1, 2, 3,4, 5, 6, 8, 9, 10	7
1, 2, 3, 4,5, 7, 8, 9, 10	6
1, 2, 3, 4, 6, 7, 8, 9, 10	5
1, 2, 3, 5, 6, 7, 8, 9, 10	4
1, 2, 4,5, 6, 7, 8, 9, 10	3
1, 3, 4, 5, 6, 7, 8, 9, 10	2
2, 3, 4, 5, 6,7, 8, 9, 10	1

Concerns

1. **How did the tool get 100% accuracy?**
   a. possible case of overfitting with the training data

2. **Some tests had very accurate results because of the minimal amount of data provided**



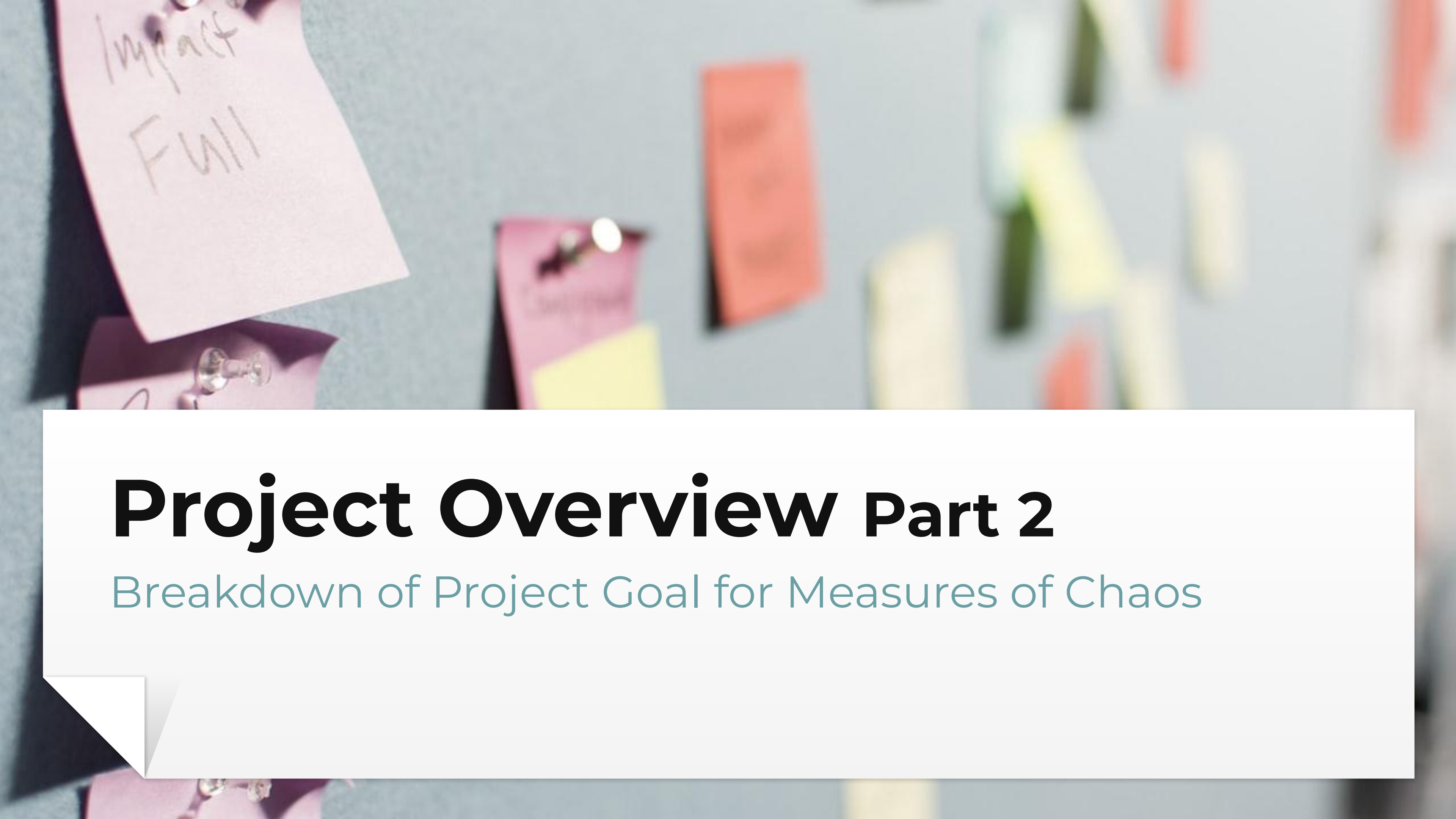
Test	Clusters	Purpose	Highest Accuracy
1	*Adenoviridae, Anelloviridae, Caudovirales, Geminiviridae, Genomoviridae, Microviridae, Ortervirales, Papillomaviridae, Parvoviridae, Polydnaviridae, Polyomaviridae, and Riboviria domain* *3273 sequences*	Used as a control (Riboviria is the only domain for viruses) to test the MLDSP-GUI	Quadratic SVM model 94.9%
3b	*Alphacoronavirus, Betacoronavirus, Deltacoronavirus, Gammacoronavirus* *60 sequences*	Decreased betacoronavirus to decrease bias of training data	Linear Discriminant Model 100%

Concerns (cont)

3. **Using 2d representations to get accurate data is a sign that the data may not be accurate with 1d representations that are simpler like Integer Representation.**
   a. A 2016 study on BMC Bioinformatics showed that closely related nuclear DNA signatures cannot always be differentiated

4. They only used 6 classification models, but there are many more that exist and can be tested on. How does this fact possibly skew results in their favor?

5. What defines success in classification? Are the accuracies enough to give a good measure of classification, or should they also look at other variables?



Project Overview Part 2

Breakdown of Project Goal for Measures of Chaos

Project Goal

Use Machine Learning models to find an **identification for COVID-19** through the use of chaos-inspired features

Measures of Chaos

Chaos refers to randomness

Entropy: measure of unpredictability/uncertainty in a certain system

Informational Entropy: increase in uncertainty as a result of the increase of information (same concept as Shannon's entropy)

Application to Pathogen Classification

Shannon's entropy can be applied to genetic information

- indicate genetic diversity
- implicate similarity/variance from previously classified pathogens

Shannon’s Entropy

measures the information/uncertainty of a random process

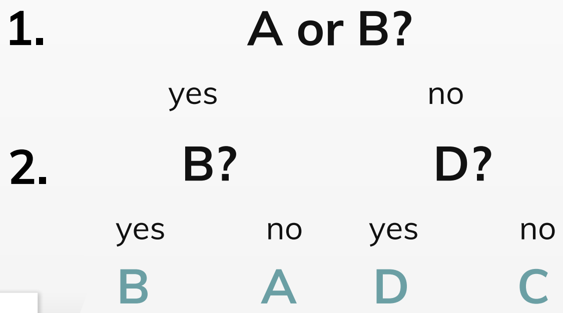
Machine 1: BACDDCBA

All letters occurring 25% of the time

Machine 2: ABCDAABA

A: 50%, B: 25%, C: 12.5%, and D: 12.5%

For which machine can you accurately predict the next letter more quickly?



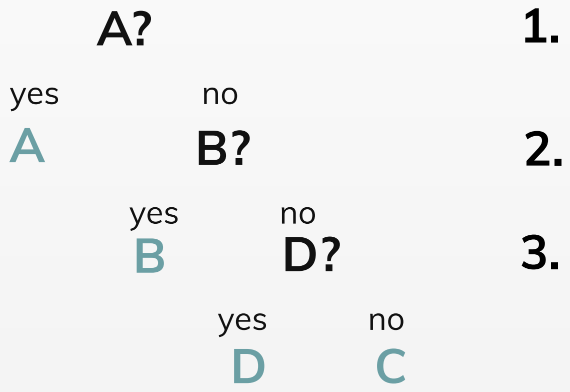
1. **A or B?**

yes no

2. **B?** **D?**

yes no yes no

B A D C



A? **1.**

yes no

A **B?** **2.**

yes no

B **D?** **3.**

yes no

D C

Shannon’s Entropy (continued)

For which machine can you accurately predict the next letter more quickly?

Machine 1:

Regardless of what the next letter in the sequence is, two questions will always need to be asked to determine the next letter.

On average, it takes **2 questions** to accurately predict the next letter in the sequence.

Machine 2:

Compute weighted average:

Avg = # of questions x probability of letter occuring

$= (1 \times P_A) + (2 \times P_B) + (3 \times P_C) + (3 \times P_D)$

$= 0.5 + 0.5 + 0.375 + 0.375 = 1.75$

On average, it takes **1.75 questions** to accurately predict the next letter in the sequence.

Shannon's Entropy (continued)

The next letter can be predicted more efficiently for Machine 2 compared to Machine 1.

What does that mean?

Machine 1 has more information, so it takes more time to process

↑ information to sort through = ↑ uncertainty of accuracy = ↑ entropy

Shannon's Entropy Equation

$$H(X) = -\sum_{i=1}^{n} p_i \log_2 p_i$$

Significance of Shannon's Entropy

Application

Shannon's entropy can be applied to genetic information

- Compare genome of unclassified pathogen with previously classified pathogen
- Similar entropic value = similarity in sequence information

Significance in Classification

- find different factors that affect genetic variability
- see if these factors' entropic levels can help determine the similarity among genomes

Progress Check

Finished tasks & upcoming goals

Current Progress

Finished Tasks

✓ Reviewed COVID-19 case study paper
✓ Reviewed methods of MLDSP
✓ Code for calculating PCC
✓ Downloaded data from COVID-19 case study
✓ Entropy code

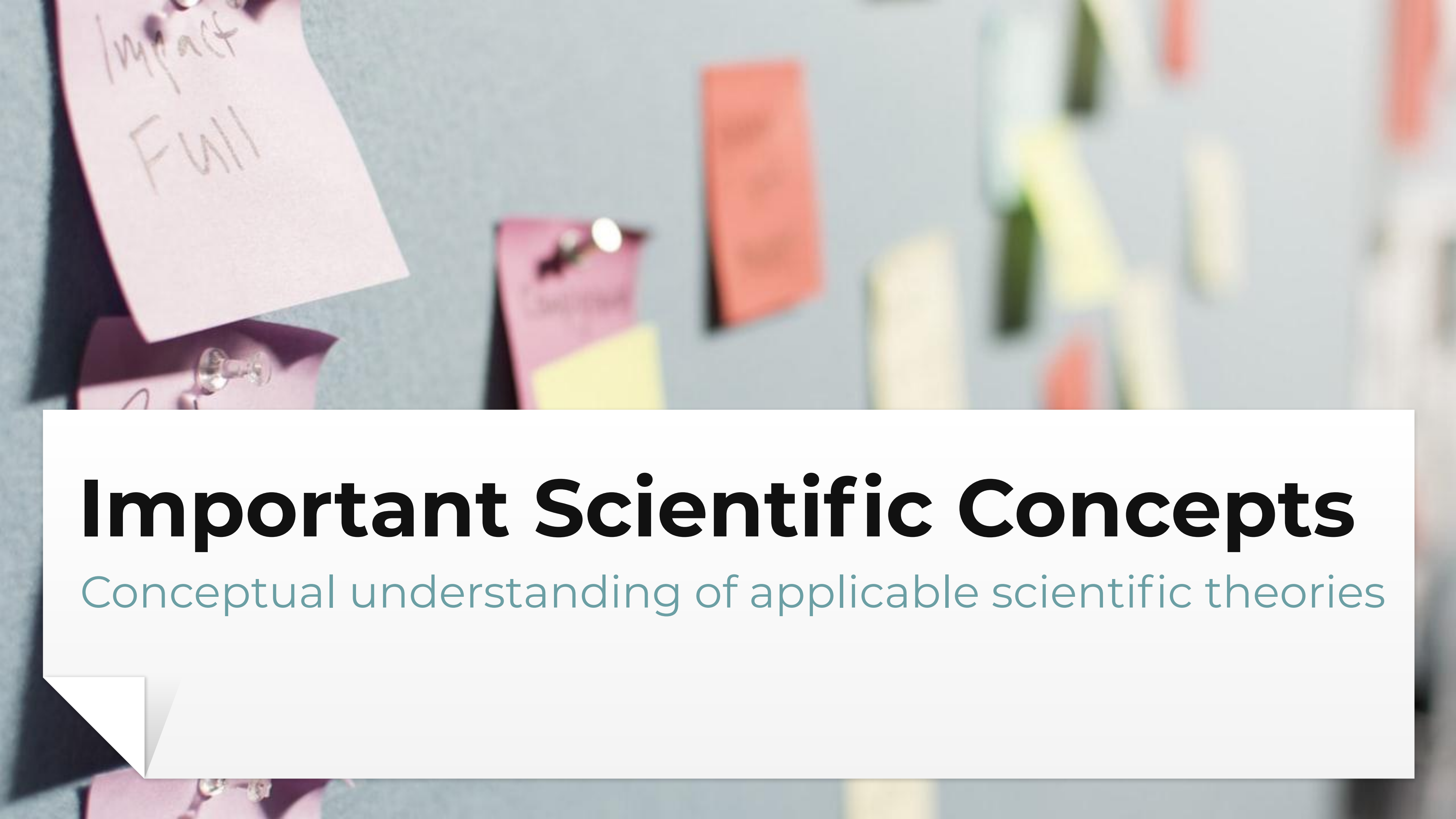
Goals

❏ Apply 2D Chaos Game Representation
❏ Run through 10-fold validation
❏ Analyze entropy code for genetic variance
❏ Look at other methods of chaos
❏ Finding our own ML algorithms to yield better accuracy results

Within the next two weeks, we hope to complete the literature review and conceptual understanding section of our project plan.

Thanks!

Any questions?



Important Scientific Concepts

Conceptual understanding of applicable scientific theories

Instructions for use

EDIT IN GOOGLE SLIDES

Click on the button under the presentation preview that says "Use as Google Slides Theme".

You will get a copy of this document on your Google Drive and will be able to edit, add or delete slides.

You have to be signed in to your Google account.

EDIT IN POWERPOINT®

Click on the button under the presentation preview that says "Download as PowerPoint template". You will get a .pptx file that you can edit in PowerPoint.

Remember to download and install the fonts used in this presentation (you'll find the links to the font files needed in the Presentation design slide)

More info on how to use this template at www.slidescarnival.com/help-use-presentation-template

This template is free to use under Creative Commons Attribution license. You can keep the Credits slide or mention SlidesCarnival and other resources used in a slide footer.

Hello!

I am Jayden Smith

I am here because I love to give presentations.

You can find me at @username

Transition headline

Let’s start with the first set of slides

“

Quotations are commonly printed as a means of inspiration and to invoke philosophical thoughts from the reader.

This is a slide title

- Here you have a list of items
- And some text
- But remember not to overload your slides with content

Your audience will listen to you or read the content, but won’t do both.

Big concept

Bring the attention of your audience over a key concept using icons or illustrations

You can also split your content

White

Is the color of milk and fresh snow, the color produced by the combination of all the colors of the visible spectrum.

Black

Is the color of ebony and of outer space. It has been the symbolic color of elegance, solemnity and authority.

In two or three columns

Yellow

Is the color of gold, butter and ripe lemons. In the spectrum of visible light, yellow is found between green and orange.

Blue

Is the colour of the clear sky and the deep sea. It is located between violet and green on the optical spectrum.

Red

Is the color of blood, and because of this it has historically been associated with sacrifice, danger and courage.

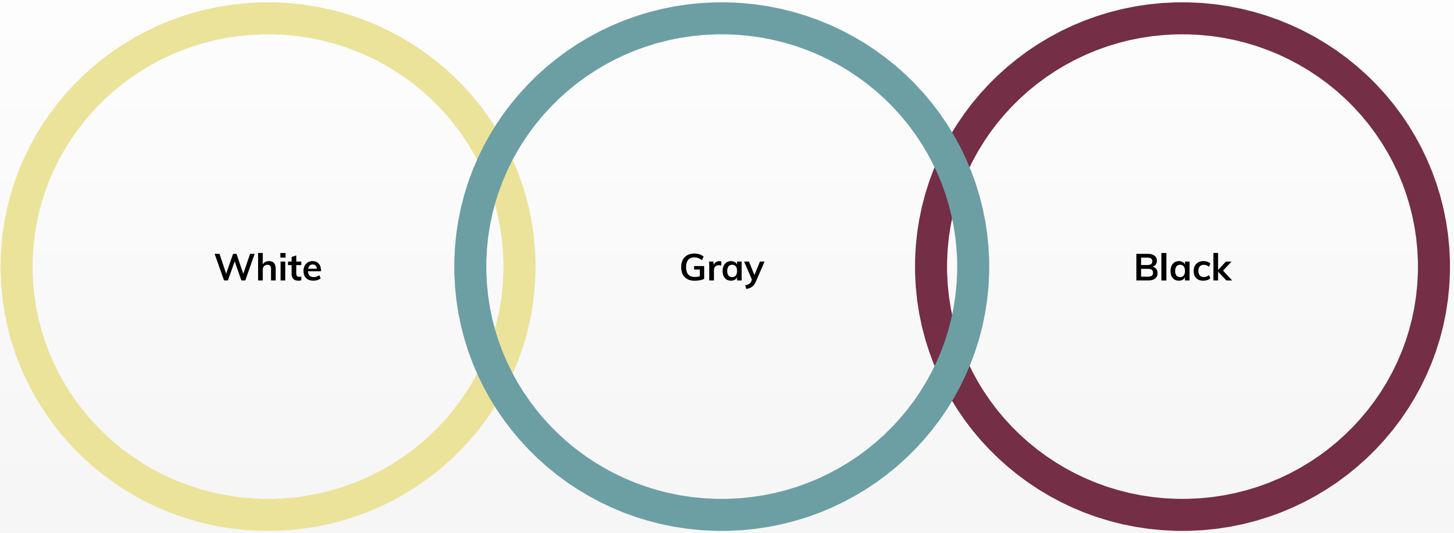
A picture is worth a thousand words

A complex idea can be conveyed with just a single still image, namely making it possible to absorb large amounts of data quickly.

Want big impact?

Use big image.

Use charts to explain your ideas



And tables to compare data

	A	B	C
Yellow	**10**	**20**	**7**
Blue	**30**	**15**	**10**
Orange	**5**	**24**	**16**

Maps

89,526,124

Whoa! That’s a big number, aren’t you proud?

89,526,124$

That’s a lot of money

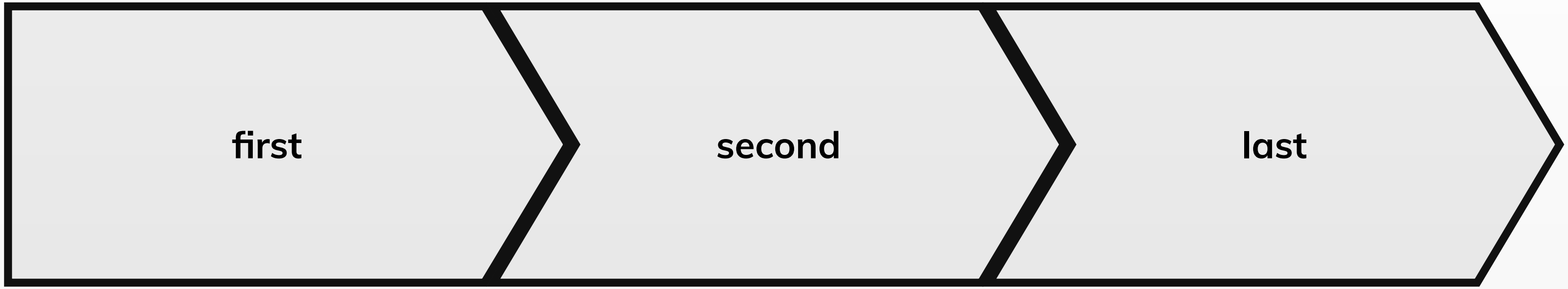
185,244 users

And a lot of users

100%

Total success!

Our process is easy



first

second

last

Let’s review some concepts

Yellow

Is the color of gold, butter and ripe lemons. In the spectrum of visible light, yellow is found between green and orange.

Blue

Is the colour of the clear sky and the deep sea. It is located between violet and green on the optical spectrum.

Red

Is the color of blood, and because of this it has historically been associated with sacrifice, danger and courage.

Yellow

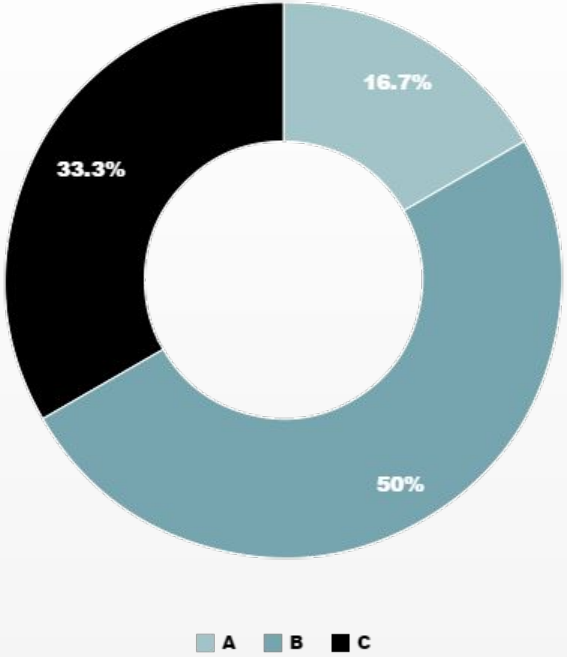
Is the color of gold, butter and ripe lemons. In the spectrum of visible light, yellow is found between green and orange.

Blue

Is the colour of the clear sky and the deep sea. It is located between violet and green on the optical spectrum.

Red

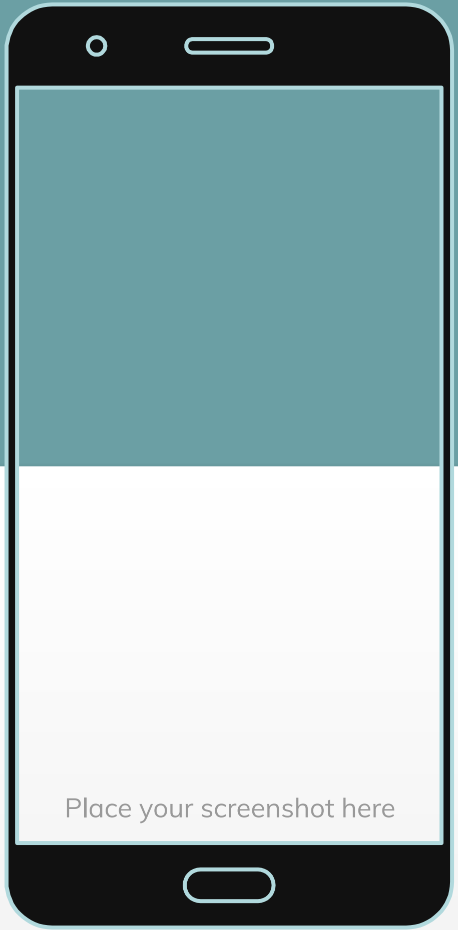
Is the color of blood, and because of this it has historically been associated with sacrifice, danger and courage.



You can copy&paste graphs from Google Sheets

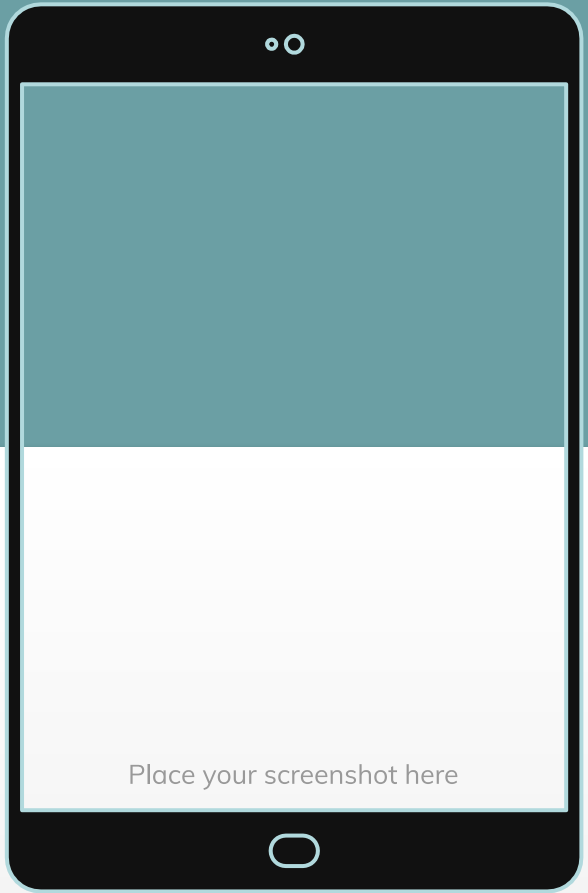
Mobile project

Show and explain your web, app or software projects using these gadget templates.



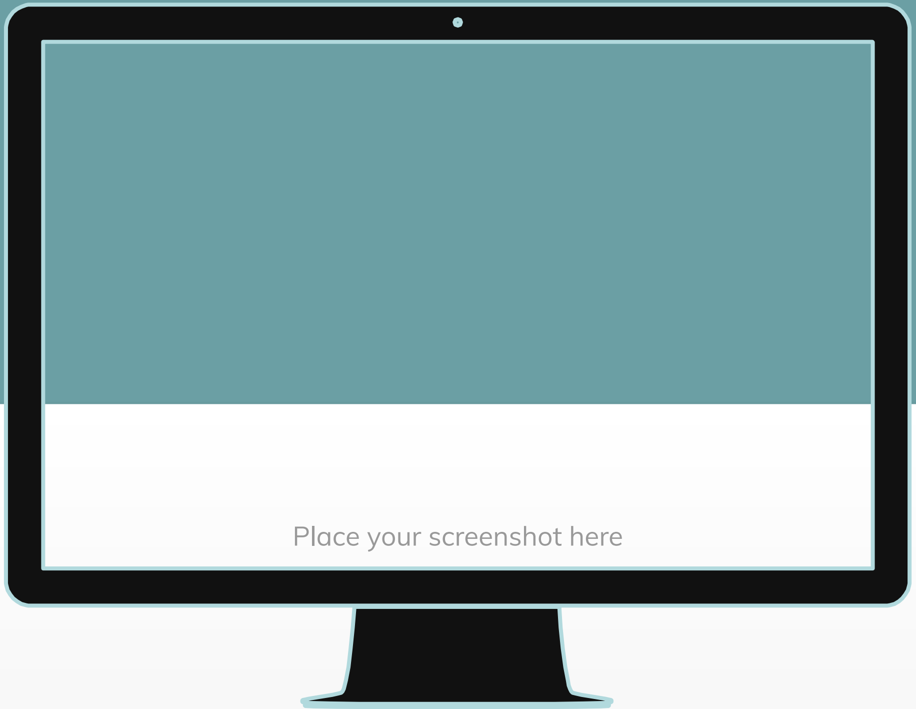
Tablet project

Show and explain your web, app or software projects using these gadget templates.



Desktop project

Show and explain your web, app or software projects using these gadget templates.



Thanks!

Any questions?

You can find me at:

@username

user@mail.me

Credits

Special thanks to all the people who made and released these awesome resources for free:

- Presentation template by SlidesCarnival
- Photographs by Unsplash

Presentation design

This presentations uses the following typographies and colors:

- Titles: Montserrat
- Body copy: Muli

You can download the fonts on these pages:

https://www.fontsquirrel.com/fonts/montserrat

https://www.fontsquirrel.com/fonts/muli

- Dark gray **#111111**
- Teal **#6b9fa4**

You don’t need to keep this slide in your presentation. It’s only here to serve you as a design guide if you need to create new slides or download the fonts to edit the presentation in PowerPoint®

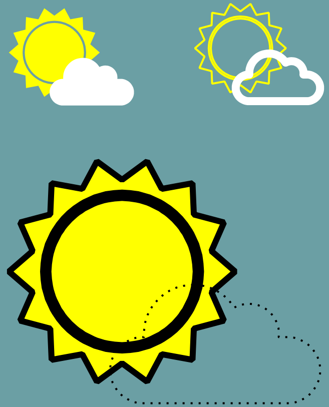
SlidesCarnival icons are editable shapes.

This means that you can:

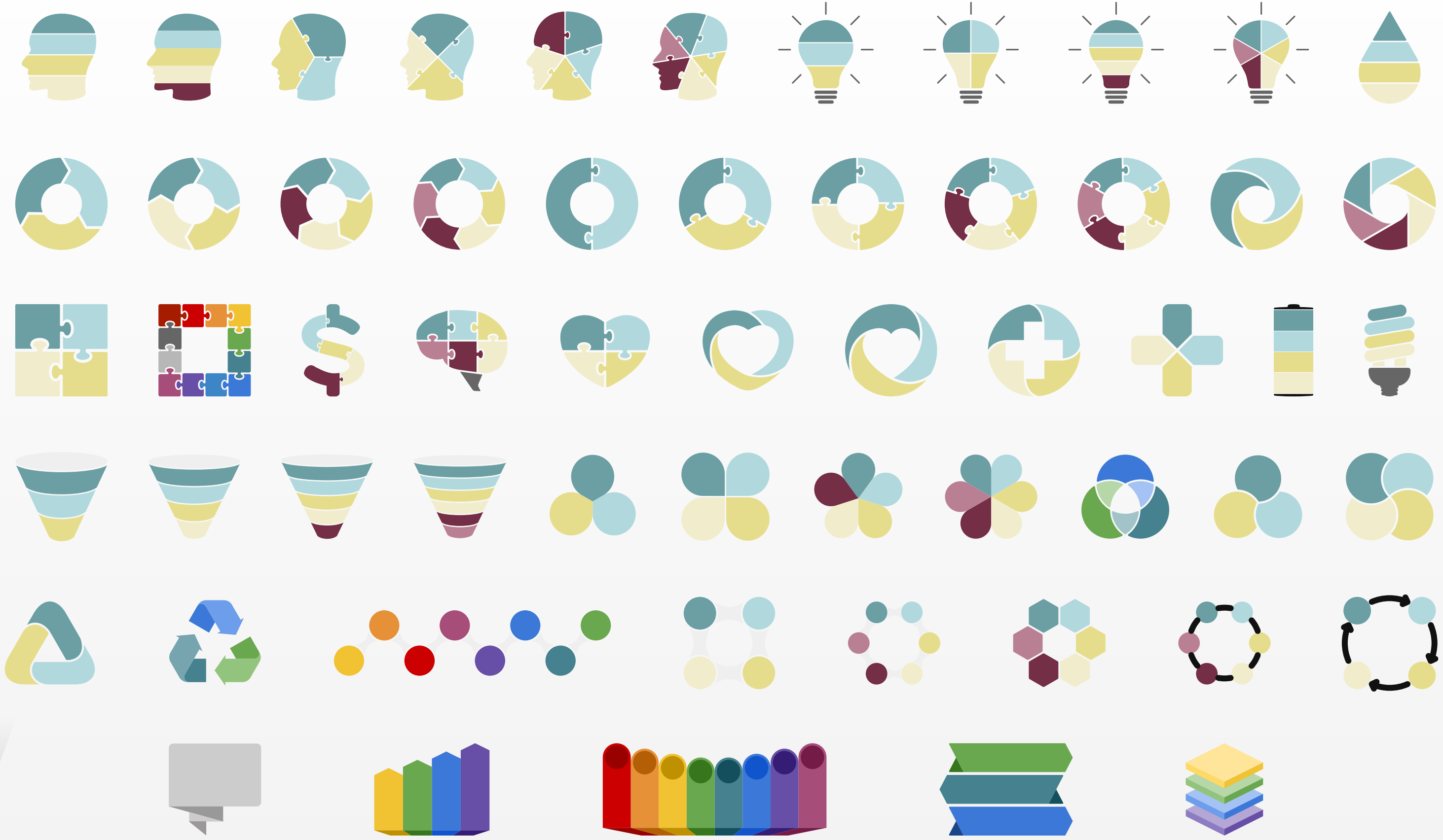
- Resize them without losing quality.
- Change fill color and opacity.
- Change line color, width and style.

Isn’t that nice? :)

Examples:



Diagrams and infographics



Now you can use any emoji as an icon!
And of course it resizes without losing quality and you can change the color.

How? Follow Google instructions
https://twitter.com/googledocs/status/730087240156643328

and many more…

Slides Carnival

Free templates for all your presentation needs

For PowerPoint and Google Slides

100% free for personal or commercial use

Ready to use, professional and customizable

Blow your audience away with attractive visuals