CREATE A MODEL TO "SPOT THE SPAM"

Dustin Bracy
Paul Huggins
Grace Lang
Branum Stephans

BUSINESS OBJECTIVE

Spam is digital junk mail: impersonal, unsolicited and unnecessary.

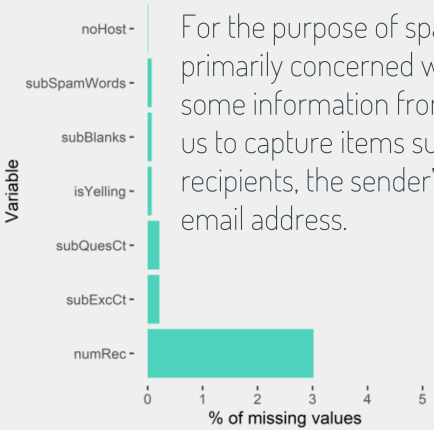
How do we create a model that accurately detects these unwanted communications and leaves us with only the important email messages to read through?

Common spam techniques have been tracked within the data set, such as the use of exclaimation marks, capital letters, and “RE:”.

Other general email practices are also aggregated into the data such as percent of spaces, capitals used, number of characters in body of message, etc.

—SPAM WAS ORIGINALLY COINED AS A REFERENCE TO A MONTY PYTHON SKIT WHERE A GROUP OF DINERS LOUDLY AND REPEATEDLY PROCLAIMED EVERYONE MUST EAT SPAM, REGARDLESS OF WHETHER THEY WANTED IT OR NOT.

DATA DESCRIPTORS



For the purpose of spam classification, we are primarily concerned with the body, but will gather some information from the header. This will allow us to capture items such as: the number of recipients, the sender's hostname and the sender's email address.

MISSING RECORDS

The highest number of missing records came from the field "numRec" which identifies number of recipients.
Due to low prevalence (~3%), all observations with missing fields have been removed.

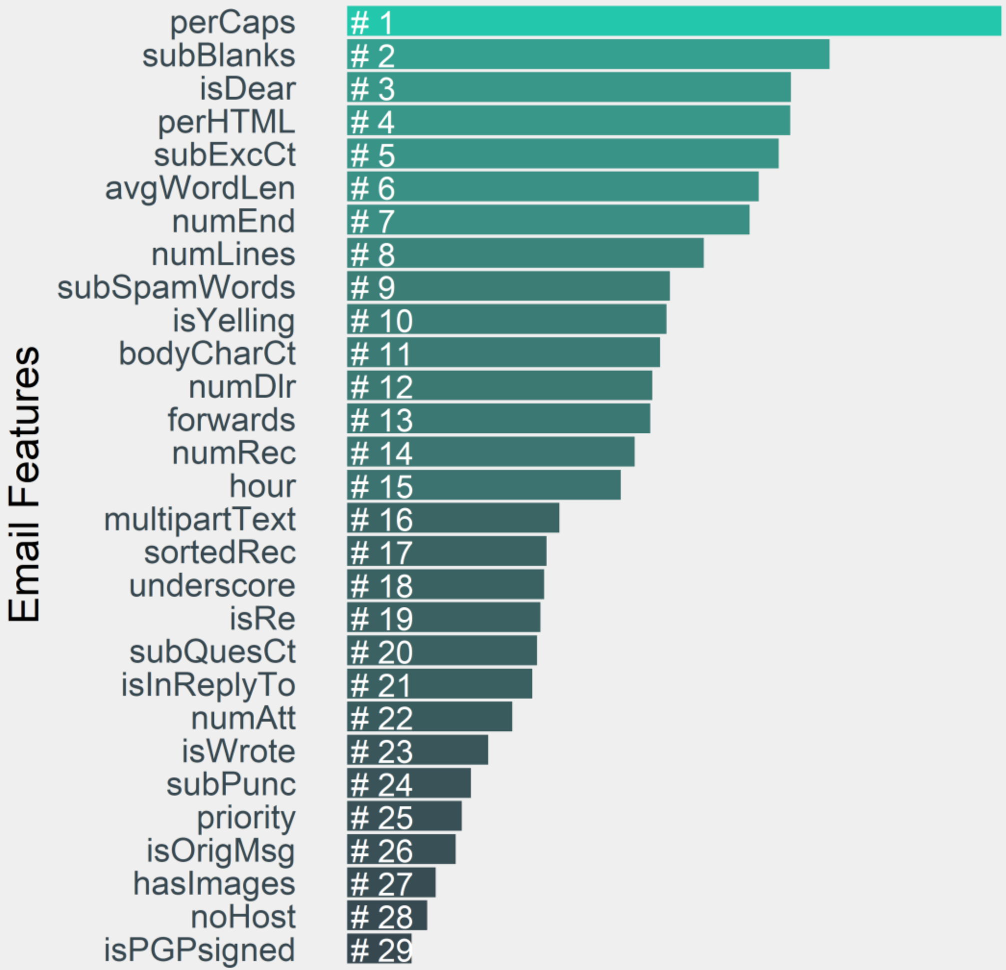
CAPITAL LETTERS

The percentage of capital letters in the email body is the number one feature in correctly predicting spam emails.

SPAM

NOT SPAM	75%
SPAM	25%

FEATURE IMPORTANCE



Features have been sorted in order of impact using the highest absolute correlation score for correctly classifying spam emails.

See Appendix for a description of features

DATA MANIPULATION

THE EMAIL DATA WAS PROVIDED TO THE TEAM POST-TEXT PROCESSING. THE RAW DATA HAD TO ORIGINALLY GO THROUGH A SERIES OF COMPLEX TEXT MINING IN ORDER TO BE IN A FORMAT THAT COULD BE DIGESTED EASILY FOR A SPAM DETECTOR MODEL.

A STANDARD SCALER WAS APPLIED TO LEVELSET THE DISTANCE BETWEEN DATA POINTS, WHICH HELPS A BINARY CLASSIFIER BY REDUCING THE AMOUNT OF WEIGHT APPLIED TO POINTS AT EITHER END OF THE SPECTRUM.

4 models were run and compared using RMSE scores to arrive at the best model for predicting if an email is spam or not:

a. XGBoost using a grid search to optimize the parameters
b. XGBoost using AzureML to determine even further optimized parameters.
c. RandomForest using AzureML optimized parameters.
d. RegressionTree – xxx

MODELING PROCEDURE

MODEL COMPARISONS

RANDOM FOREST

Random Forests build trees independently of other trees and then combine the results at the end by using majority rules.

XGBOOST

Gradient booting trees are built one at a time which are then are combined to create an optimal tree.

REGRESSION TREE

A regression tree is a single tree that evaluates each decision and subsequently moves through the tree, one decision at a time.

MEASURING SUCCESS

The model success will be measured by evlauating Precision, Recall and Accuracy.

AZUREML XGBOOST RESULTS

Confusion Matrix		
	Not Spam	**Is Spam**
Not Spam	1656	27
Is Spam	16	562

Metric	Result
Accuracy	0.981
Accuracy 95% CI	[0.974, 0.986]
Precision/Sensitivity	0.990
Recall/Specificity	0.954
Balanced Accuracy	0.972

- 5 fold cv was used
- list included params

GRID SEARCH XGBOOST RESULTS

Confusion Matrix		
	Not Spam	Is Spam
Not Spam	1635	50
Is Spam	37	539

Metric	Result
Accuracy	0.962
Accuracy 95% CI	[0.953, 0.969]
Precision/Sensitivity	0.978
Recall/Specificity	0.915
Balanced Accuracy	0.947

5 fold cv
was used

RANDOM FOREST RESULTS

Confusion Matrix		
	Not Spam	**Is Spam**
Not Spam	1672	16
Is Spam	38	535

Metric	Result
Accuracy	0.976
Accuracy 95% CI	
Precision/Sensitivity	
Recall/Specificity	
Balanced Accuracy	

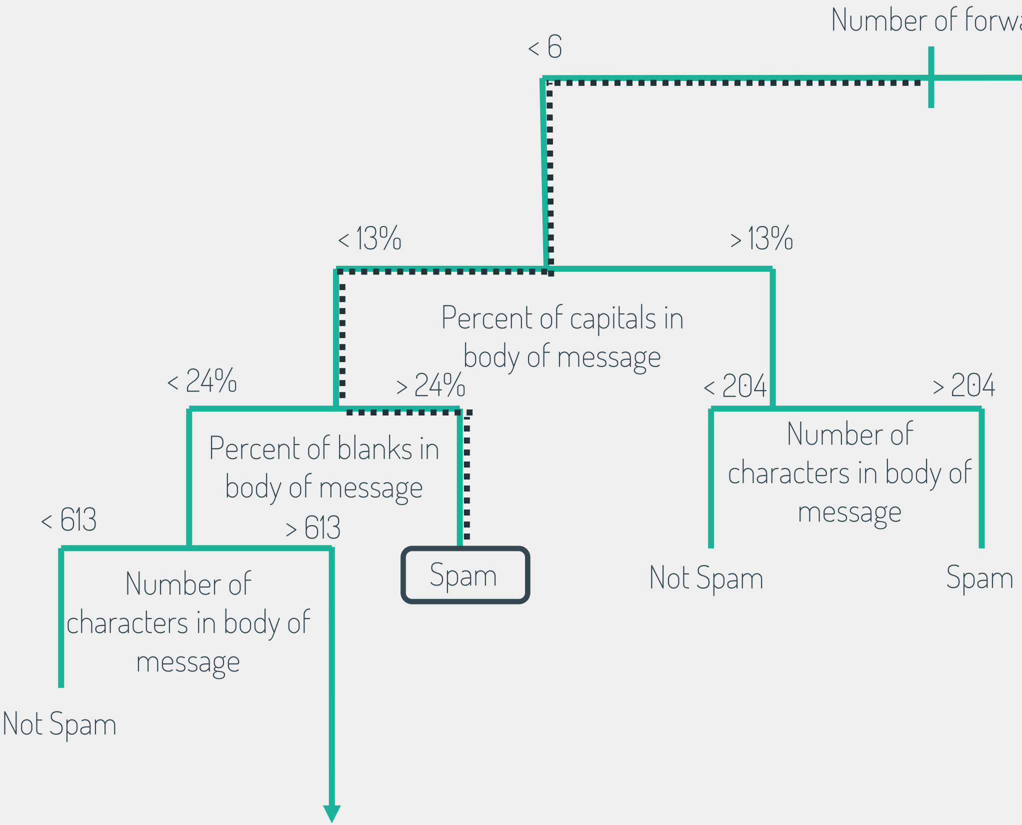
REGRESSION TREE RESULTS

Confusion Matrix		
	Not Spam	**Is Spam**
Not Spam	1656	27
Is Spam	16	562

Metric	Result
Accuracy	0.981
Accuracy 95% CI	[0.974, 0.986]
Precision/Sensitivity	0.990
Recall/Specificity	0.954
Balanced Accuracy	0.972

Baseline

REGRESSION TREE VISUALIZATION


Number of forwards
< 6
> 6
Not Spam
< 13%
> 13%
Percent of capitals in body of message
< 24%
> 24%
< 204
> 204
Percent of blanks in body of message
Number of characters in body of message
< 613
> 613
Spam
Not Spam
Spam
Number of characters in body of message
Not Spam

An email with less than 6 forwards, less than 13% of capitals in message body, and has more than 24% of blanks in the message could be considered Spam.

FUTURE

In conclusion, the __ **model** performed the best, and **is the best choice** to put into production.

Care should be taken to continue to **tune** the model **for maximum recall**: incorrectly classifying a ham email as spam is much more detrimental than classifying a spam email as ham.

APPENDIX

DATA VARIABLES

Label	Type	Description
isSpam	Logical	Whether or not the email was flagged as Spam (T/F)
isRe	Logical	TRUE if Re: appears at the start of the subject
Underscore	Logical	TRUE if email address is in the From field of the headers contains an underscore
Priority	Logical	TRUE is a Priority key is present in the header
isInReplyTo	Logical	TRUE if the In-Reply-To key is present in the header
sortedRec	Logical	TRUE if the recipients’ email addresses are sorted.
subPunc	Logical	TRUE if words in the subject have punctuation or numbers embedded in them, e.g., w!se.
multipartText	Logical	TRUE if the MIME type is multipart/text.
hasImages	Logical	TRUE if the message contains images.
isPGPsigned	Logical	TRUE if the message contains a PGP signature.
subSpamWords	Logical	TRUE if the subject contains one of the words in a spam word vector.
noHost	Logical	TRUE if there is no hostname in the Message-Id key in the header.
numEnd	Logical	TRUE if the email sender’s address (before the @) ends in a number
isYelling	Logical	TRUE if the subject is all capital letters
isOrigMsg	Logical	TRUE if the message body contains the phrase original message

DATA VARIABLES

Label	Type	Description
isDear	Logical	TRUE if the message body contains the word dear
isWrote	Logical	TRUE if the message contains the phrase wrote:.
numLines	Integer	Number of lines in the body of the message.
bodyCharCt	Integer	Number of characters in the body of the message.
subExcCt	Integer	Number of exclamation marks in the subject
subQuesCt	Integer	Number of question marks in the subject.
numAtt	Integer	Number of attachments in the message
numRec	Numeric	Number of recipients of the message, including CCs.
perCaps	Numeric	Percentage of capitals among all letters in the message body, excluding attachment
Hour	Numeric	Hour of the day in the Date field.
perHTML	Numeric	Percentage of characters in HTML tags in the message body in comparison to all characters.
subBlanks	Numeric	Percentage of blanks in the subject.
Forwards	Numeric	Number of forward symbols in a line of the body, e.g., >>> xxx contains 3 forwards.
avgWordLen	Numeric	The average length of the words in a message.
numDlr	Numeric	Number of dollar signs in the message body.

ABOUT THE PROJECT

Mercury is the closest planet to the Sun and the smallest one in the Solar System—it's only a bit larger than our Moon. The planet's name has nothing to do with the liquid metal, since it was named after the Roman messenger god

MAIN REQUIREMENTS

01

Mercury is the smallest planet in our Solar System. It’s only a bit larger than our Moon

02

Despite being red, Mars is actually a cold place. The planet is full of iron oxide

03

Saturn is the ringed one. It’s a gas giant, composed of hydrogen and helium

04

Venus has a beautiful name and is the second planet from the Sun

05

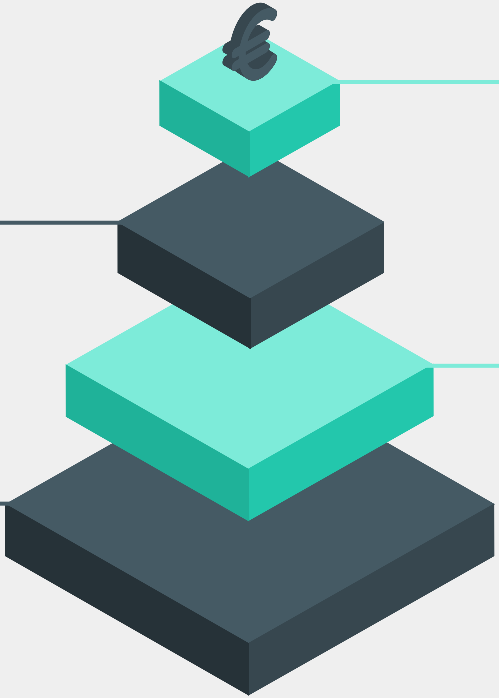
Neptune is the fourth-largest planet by diameter in our Solar System

06

Jupiter is a gas giant and the biggest planet in our Solar System

BUDGET

6,000,000



500,000

VENUS

Venus is the second planet from the Sun

1,000,000

MERCURY

Mercury is the smallest planet in our Solar System

2,000,000

SATURN

Saturn is composed of hydrogen and helium

2,500,000

JUPITER

Jupiter is the largest planet in our Solar System

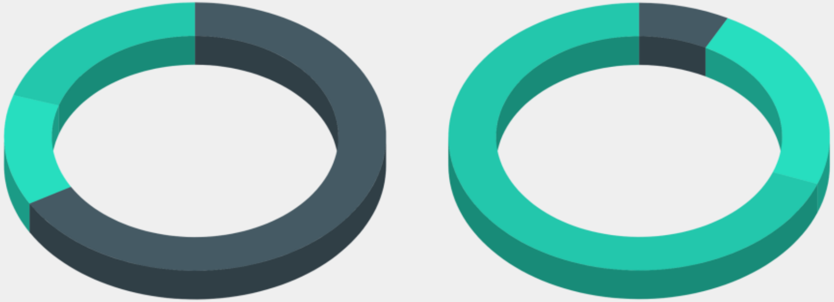
PROJECT GOALS

2017

30% MERCURY

10% JUPITER

60% VENUS



2019

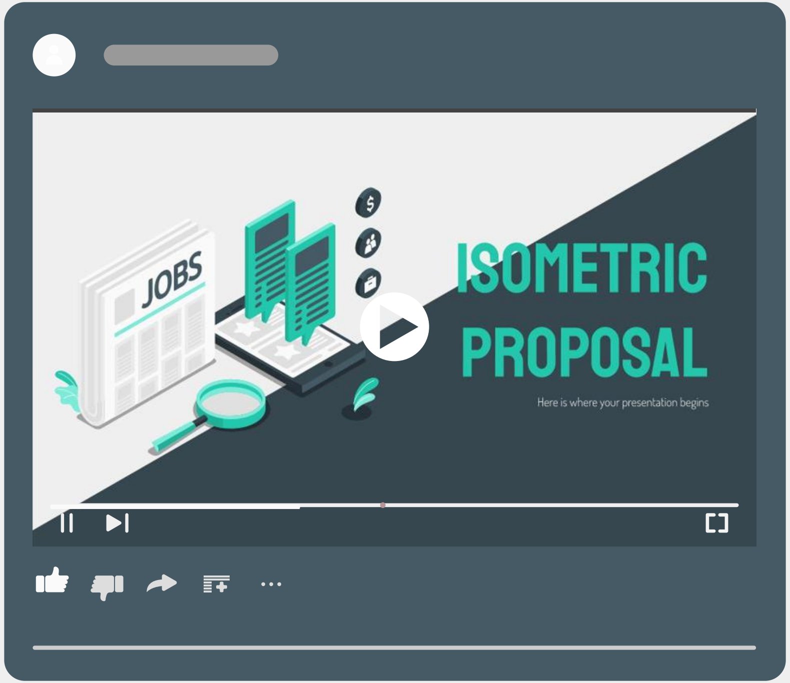
MERCURY 60%

JUPITER 35%

VENUS 5%

If you want to modify this graph, click on it, follow the link, change the data and replace it

SNEAK PEEK



Insert your multimedia content here

PROJECT STAGES

STAGE I

Despite being red, Mars is actually a cold place. It's full of iron oxide dust, which gives the planet its reddish cast

STAGE 2

Mercury is the closest planet to the Sun and the smallest one in our Solar System. It's only a bit larger than our Moon

STAGE 3

Venus has a beautiful name and is the second planet from the Sun. Its atmosphere is extremely poisonous

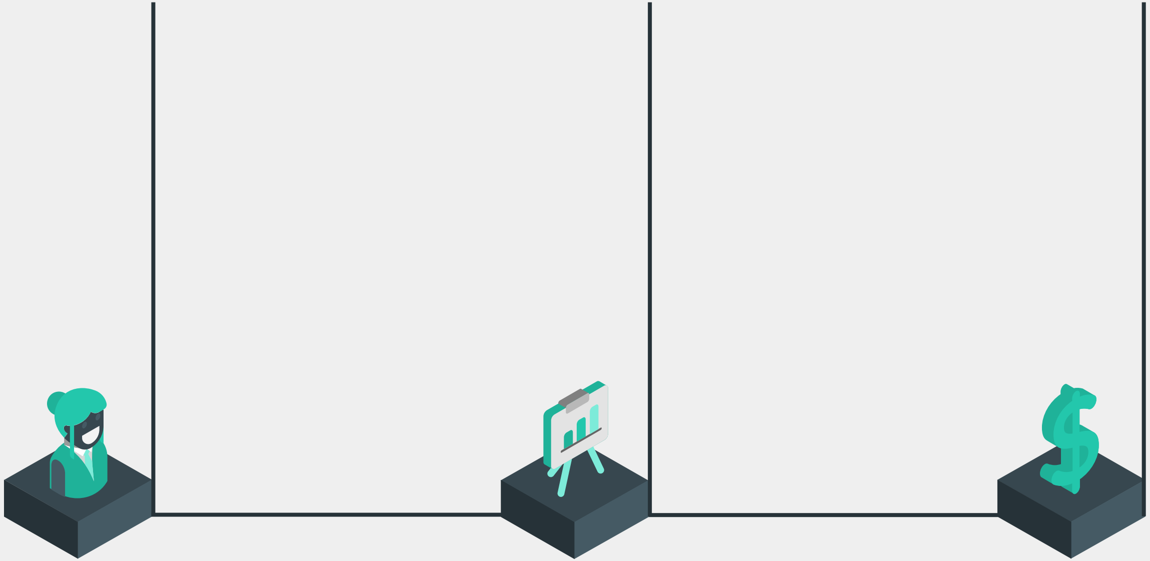
TIMELINE

STAGE 1

Jupiter is a gas giant and the biggest planet in our Solar System

STAGE 2

Saturn is a gas giant, composed mostly of hydrogen and helium



STAGE 3

Despite being red, Mars is actually a cold place. It’s full of iron oxide dust

OUR TEAM

JENNA DOE

You can replace the image on the screen with your own

RICHARD SMITH

You can replace the image on the screen with your own

THANKS

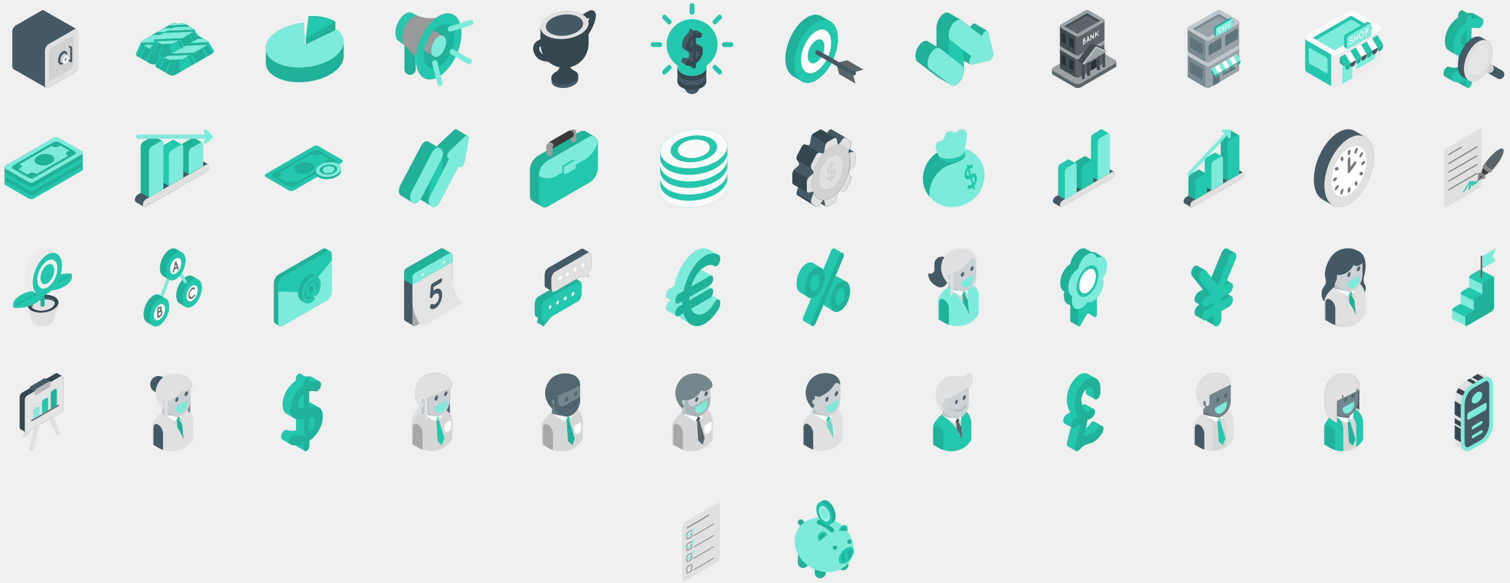
Does anyone have any questions?

youremail@freepik.com
+91 620 421 838
yourcompany.com
Follow the project updates

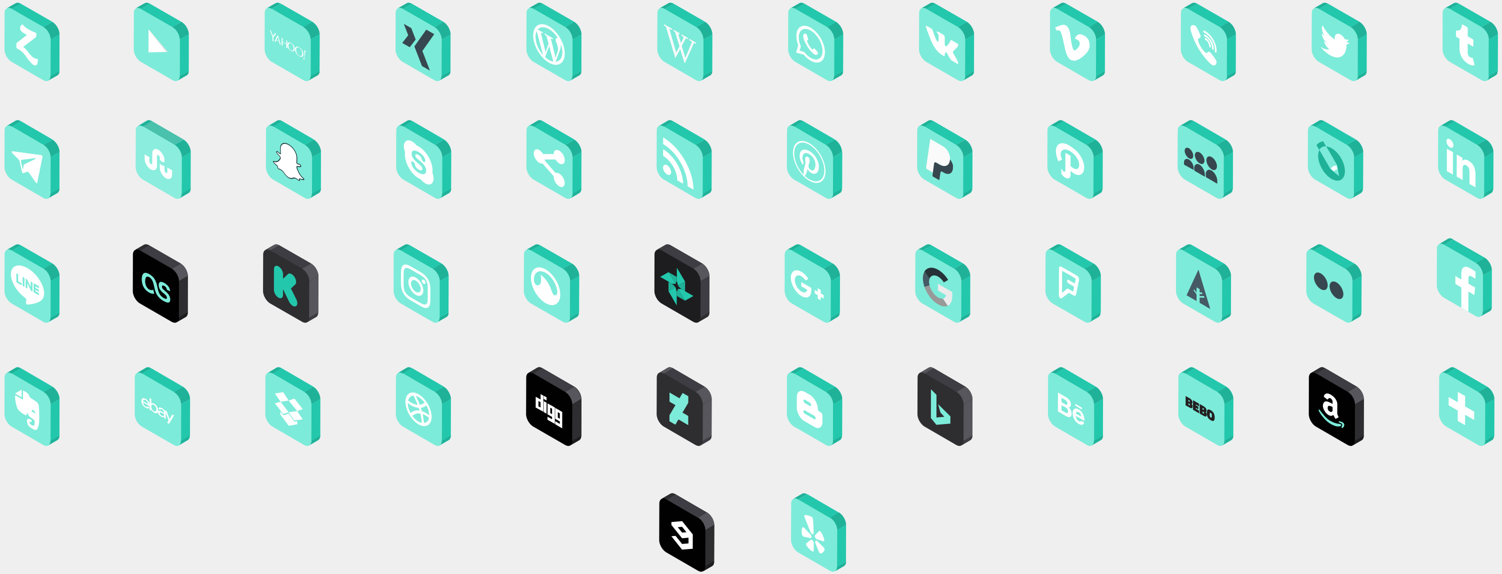
CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, and infographics & images by **Freepik** and illustrations by **Stories**

Please keep this slide for attribution.

BUSINESS ICONS

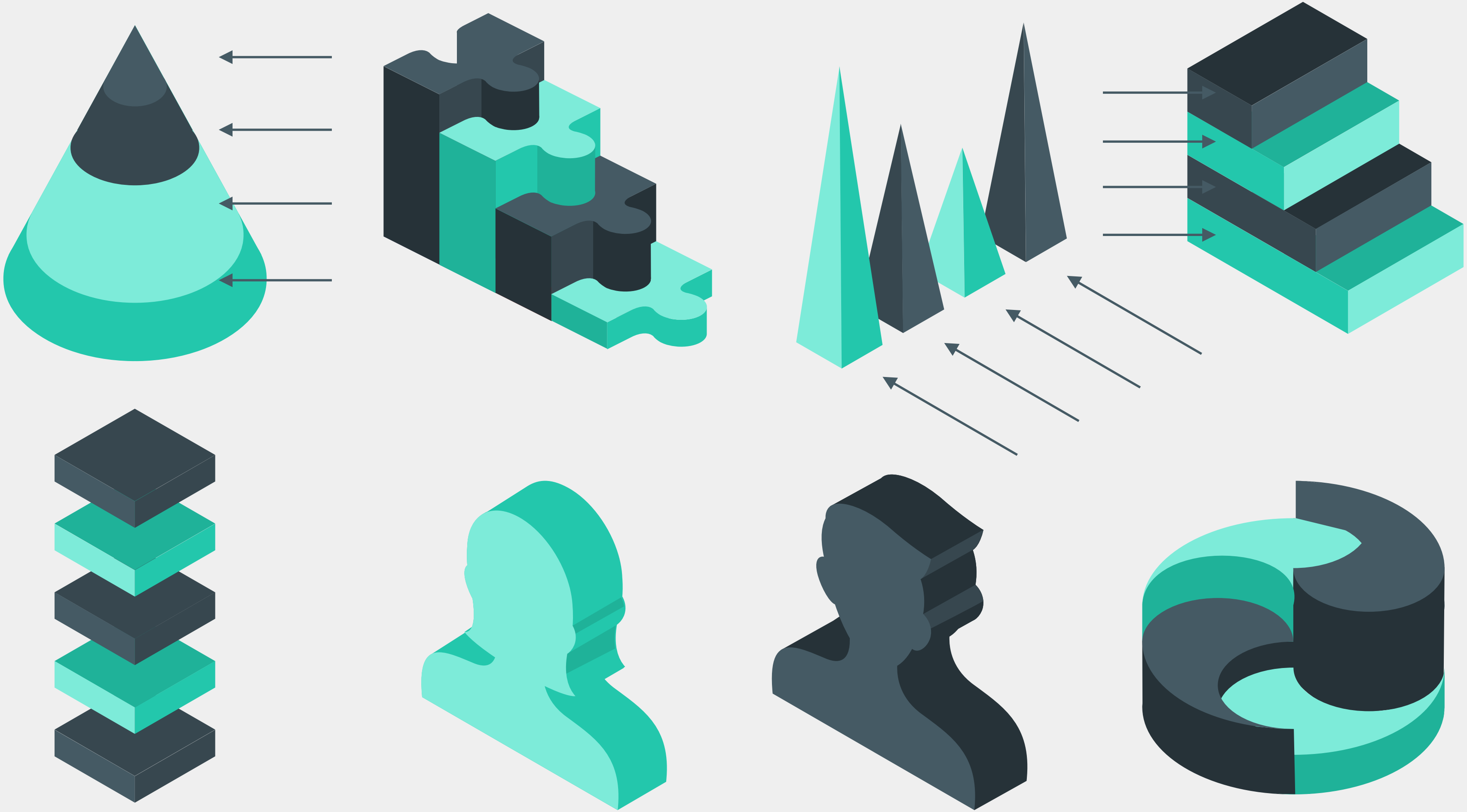


SOCIAL MEDIA ICONS



Find more illustrations like these on **Stories by Freepik**

ALTERNATIVE RESOURCES



RESOURCES

Find more illustrations like these on **Stories by Freepik**

VECTORS

- Add files
- Security concept illustration
- Experts
- Windows
- Working
- File searching
- Collaboration
- Job hunt concept illustration
- Processing
- Filter
- Account
- Mail
- Presentation
- Map dark concept illustration

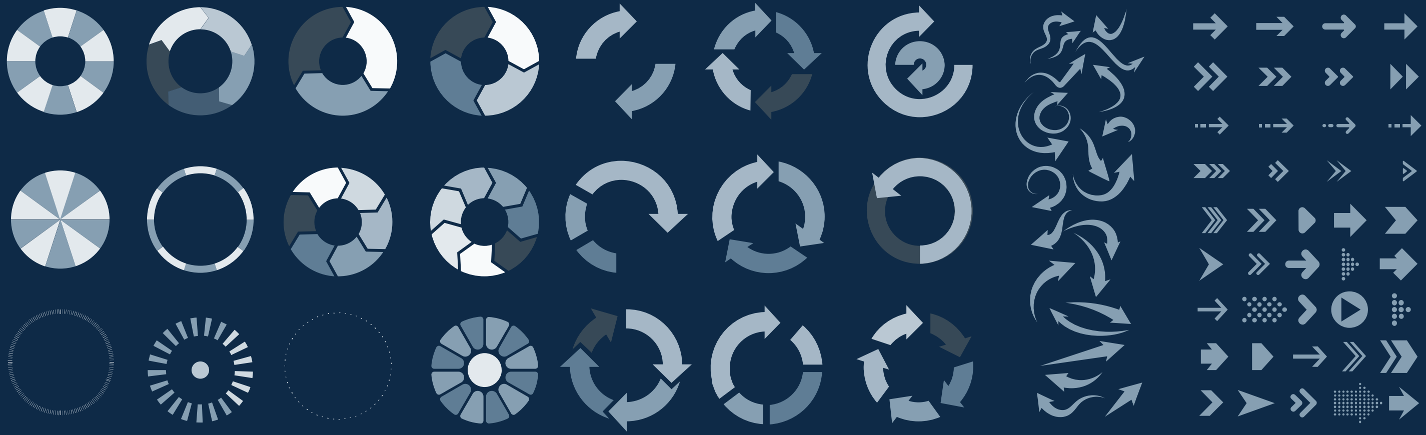
- Photo
- Isometric infographic elements template
- Isometric infographic elements
- People infographic
- Isometric infographic with timeline

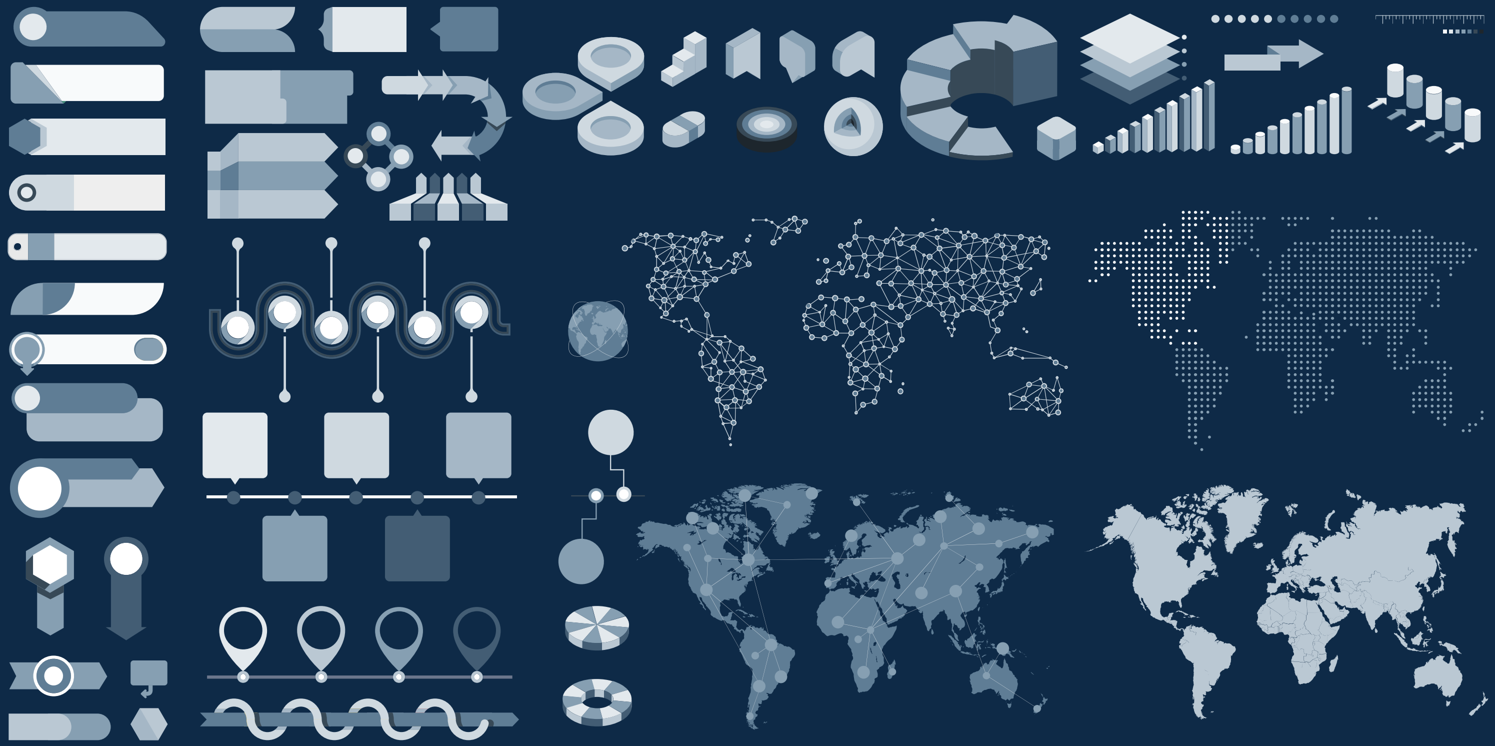
ICONS

- Logos
- Business

Use our editable graphic resources...

You can easily resize these resources, keeping the quality. To change the color, just ungroup the resource and click on the object you want to change. Then, click on the paint bucket and select the color you want. Don’t forget to group the resource again when you’re done.





...and our sets of editable icons

You can resize these icons, keeping the quality.
You can change the stroke and fill color; just select the icon and click on the paint bucket/pen.
In Google Slides, you can also use Flaticon’s extension, allowing you to customize and add even more icons.

Educational Icons

Medical Icons

Business Icons

Teamwork Icons

Help & Support Icons

Avatar Icons

Creative Process Icons

Performing Arts Icons

Nature Icons

SEO & Marketing Icons