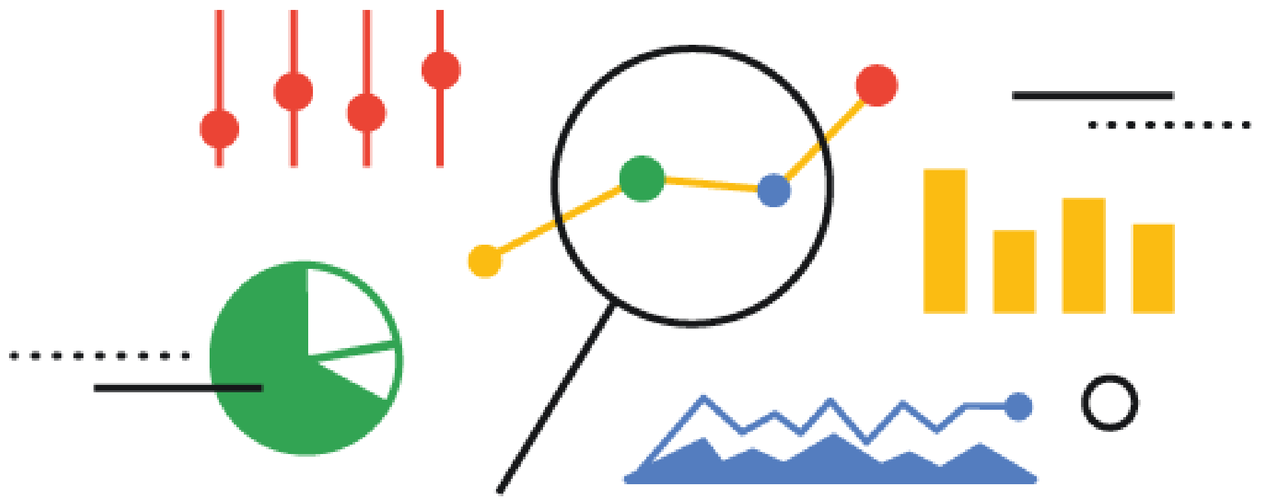
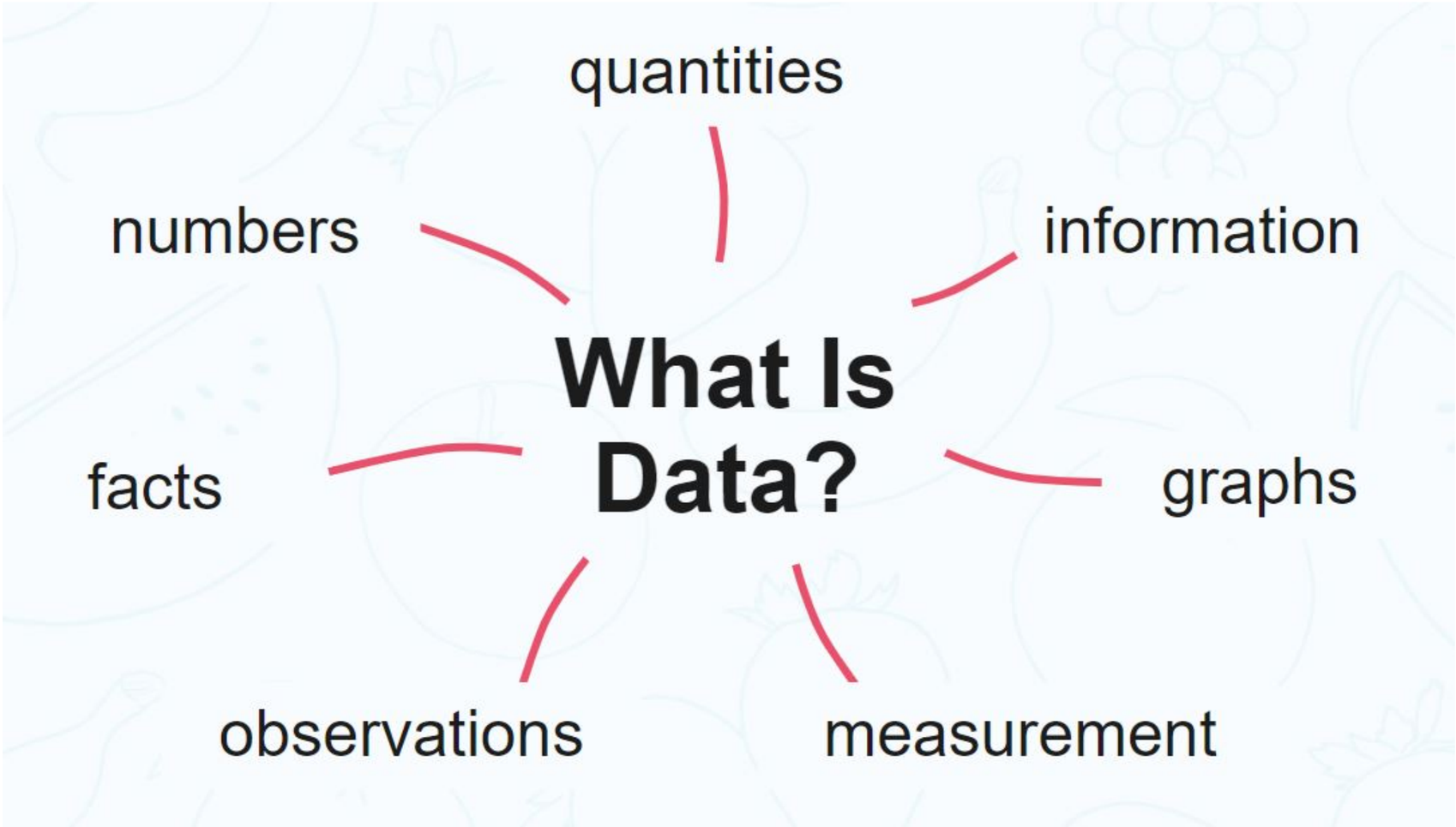
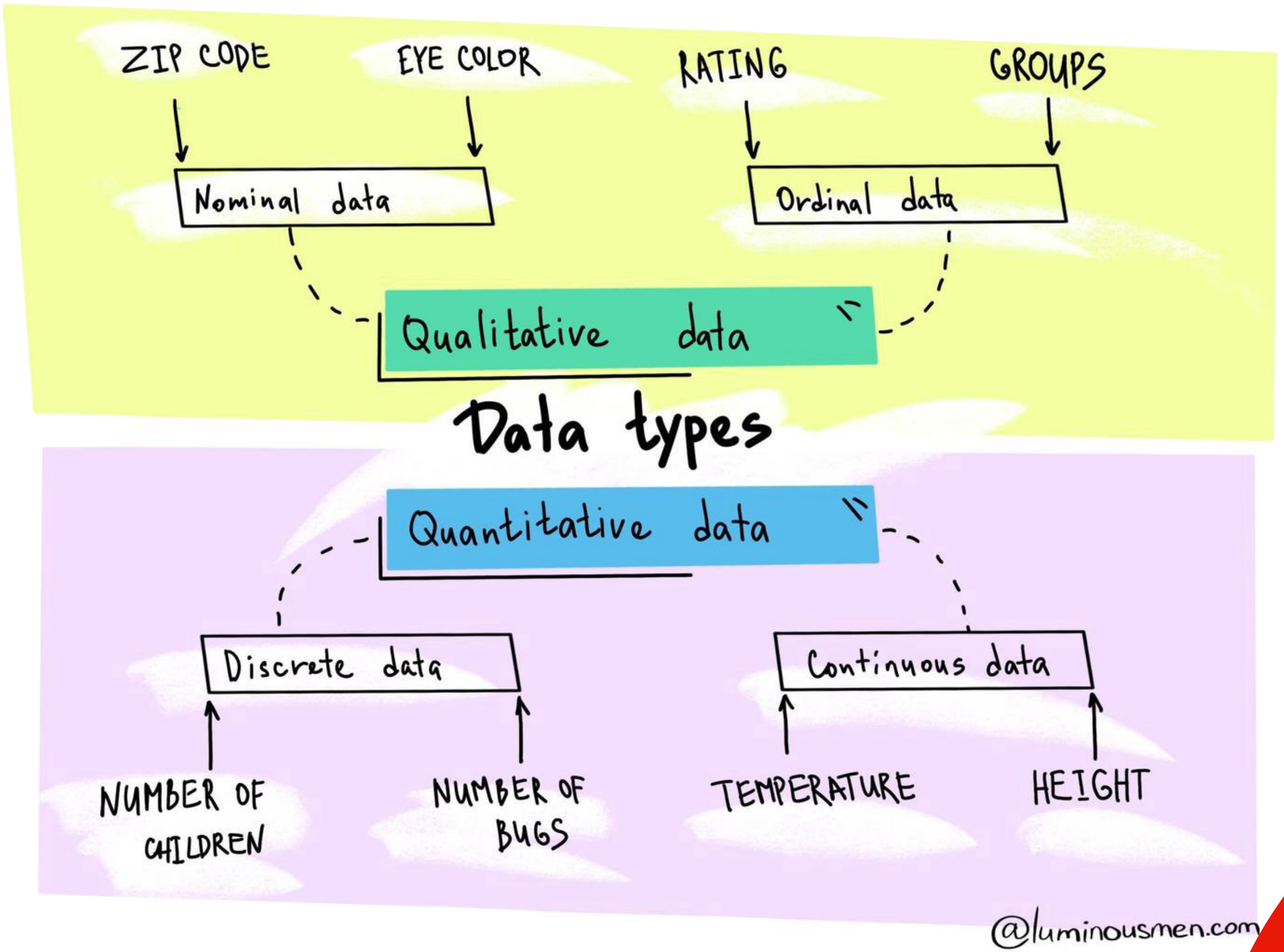
DATA PREPROCESSING



WHAT IS DATA ?

FACTS AND STATISTICS COLLECTED TOGETHER FOR REFERENCE OR ANALYSIS.


quantities
numbers
information
What Is Data?
facts
graphs
observations
measurement


ZIP CODE
EYE COLOR
RATING
GROUPS
Nominal data
Ordinal data
Qualitative data
Data types
Quantitative data
Discrete data
Continuous data
NUMBER OF CHILDREN
NUMBER OF BUGS
TEMPERATURE
HEIGHT
@luminousmen.com

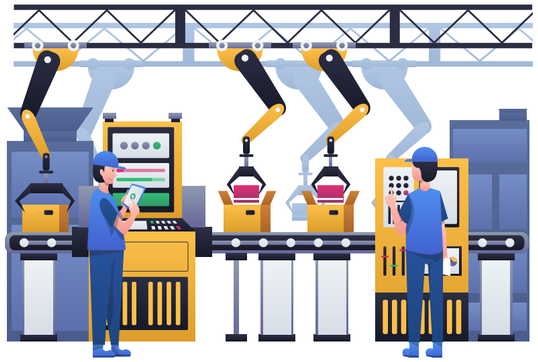
DIFFERENT FIELDS WHERE WE USE DATA

BANKING
AND SECURITIES

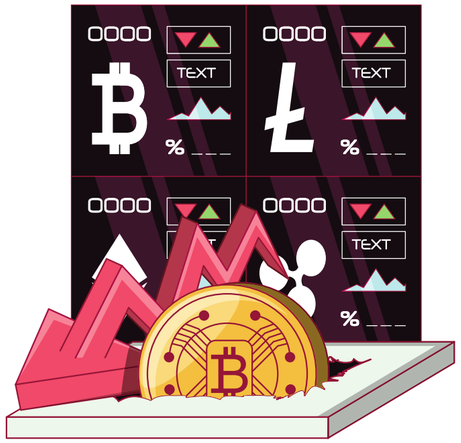
COMMUNICATIONS
AND MEDIA

HEALTHCARE
PROVIDERS

EDUCATION



MANUFACTURING



TRADE

DATA ANALYST

DATA ANALYST ANALYZES NUMERIC DATA AND USES IT TO HELP COMPANIES MAKE BETTER DECISIONS.

DATA ENGINEER

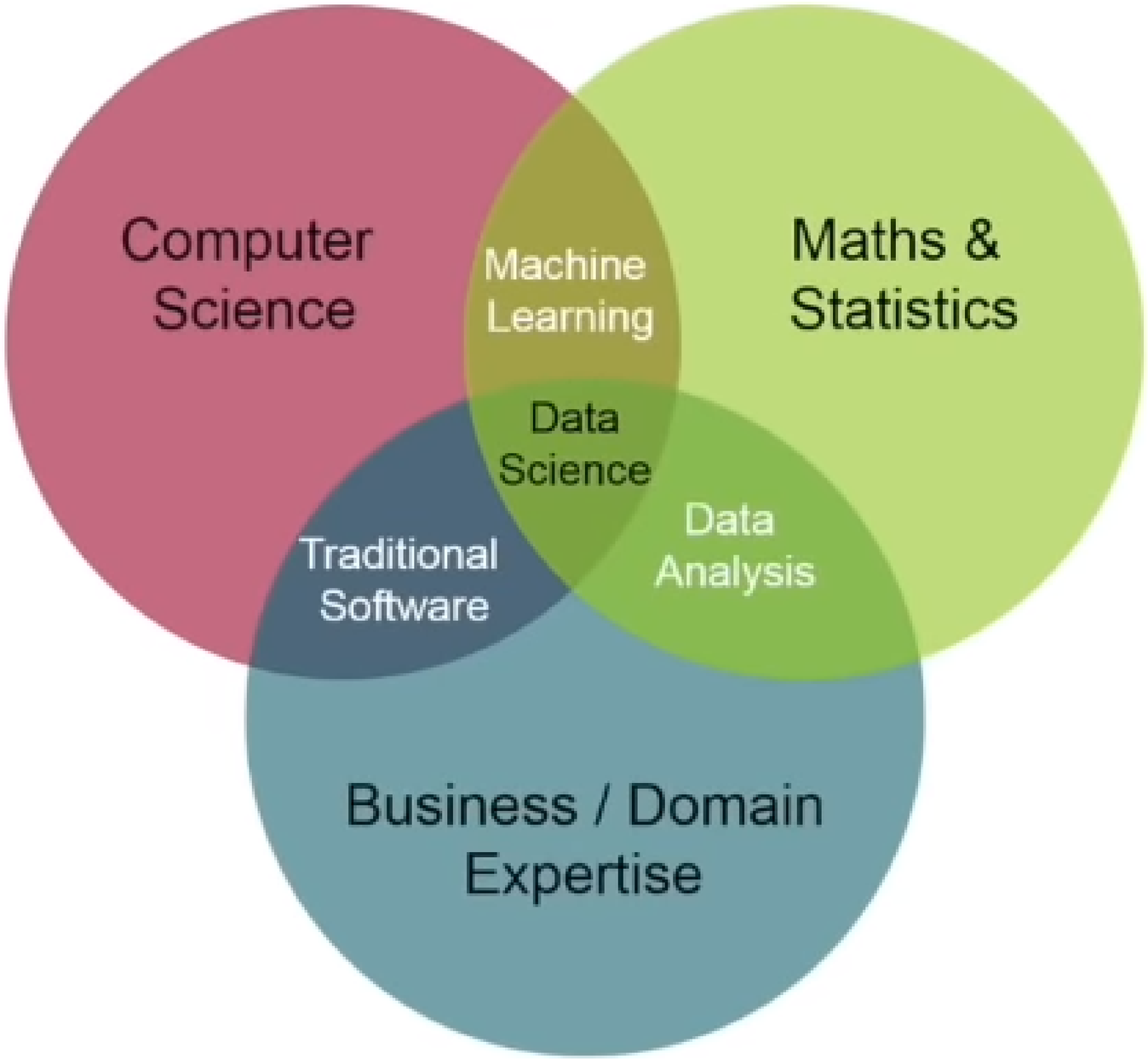
DATA ENGINEER INVOLVES IN PREPARING DATA. THEY DEVELOP, CONSTRUCTS, TESTS & MAINTAIN COMPLETE ARCHITECTURE.

BUSINESS ANALYSTS

BUISNESS ALAYSTS USE DATA TO MAKE STRATEGIC BUSINESS DECISIONS.

DATA SCIENTIST

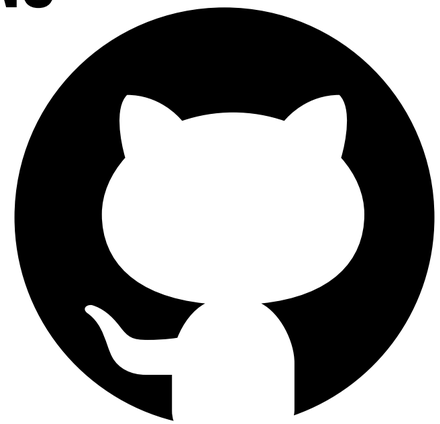
A DATA SCIENTIST ANALYZES AND INTERPRET COMPLEX DATA. THEY ARE DATA WRANGLERS WHO ORGANIZE (BIG) DATA.


Computer Science
Machine Learning
Maths & Statistics
Data Science
Traditional Software
Data Analysis
Business / Domain Expertise

DATA SOURCES

WEB SCRAPPING

FORMS



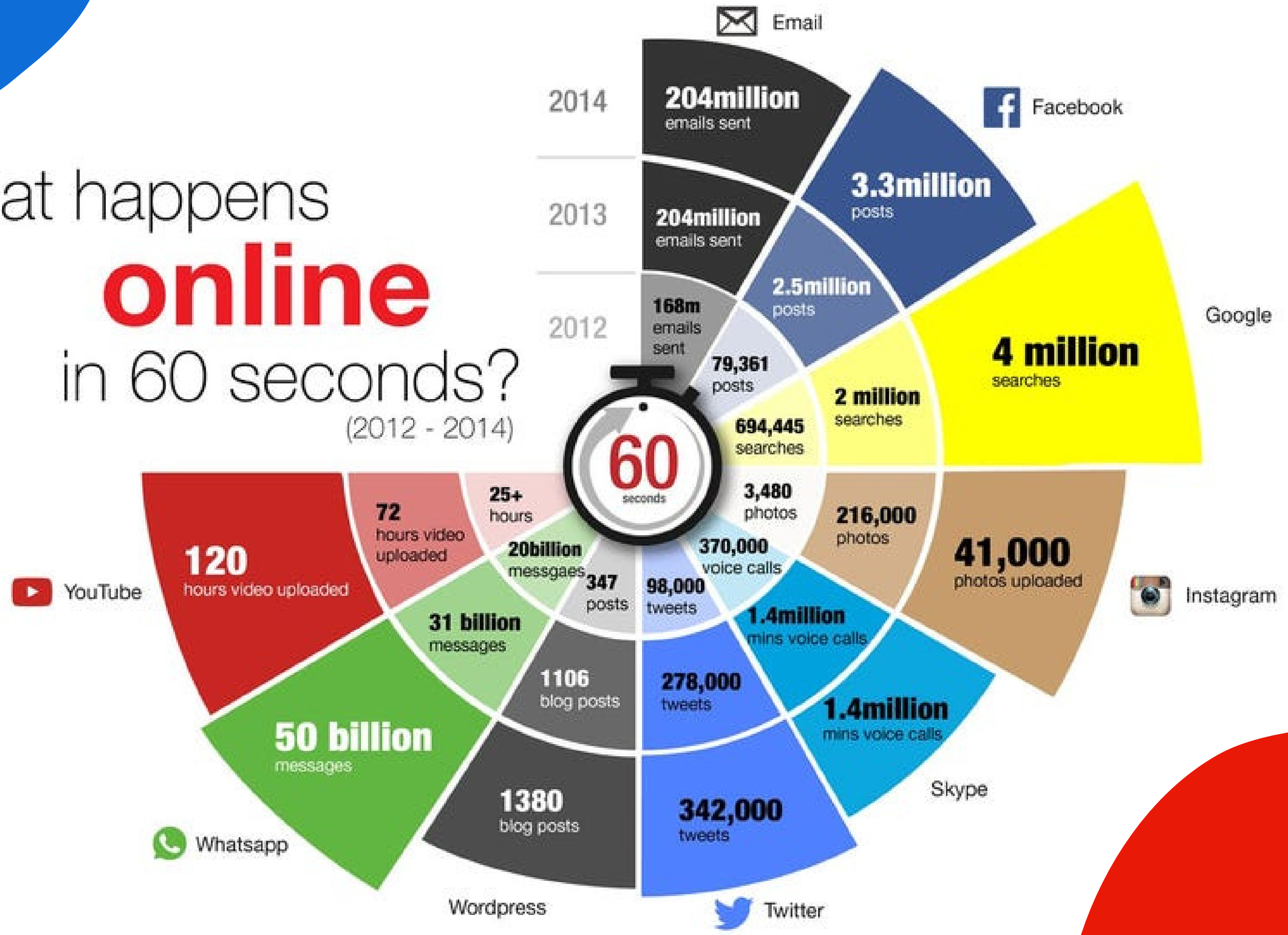
GITHUB

GENERATE IT

DATA SIZE

What happens online in 60 seconds?

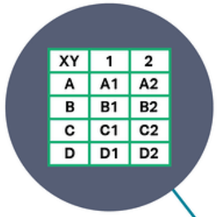
(2012 - 2014)

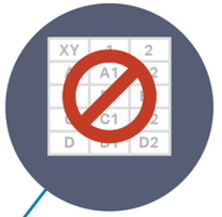


DATA :
STRUCTURED VS UNSTRUCTURED

Unstructured VS Structured Data

Structured Data VS Unstructured Data





Structured Data	Unstructured Data
Can be displayed in rows, columns and relational databases	Cannot be displayed in rows, columns and relational databases
Numbers, dates and strings	Images, audio, video, word processing files, e-mails, spreadsheets
Estimated 20% of enterprise data *(Gartner)*	Estimated 80% of enterprise data *(Gartner)*
Requires less storage	Requires more storage
Easier to manage and protect with legacy solutions	More difficult to manage and protect with legacy solutions

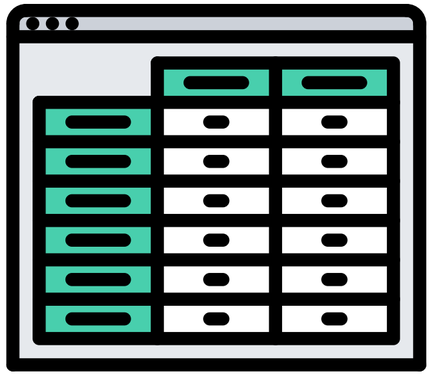
TYPES OF DATA

TEXTE

IMAGE

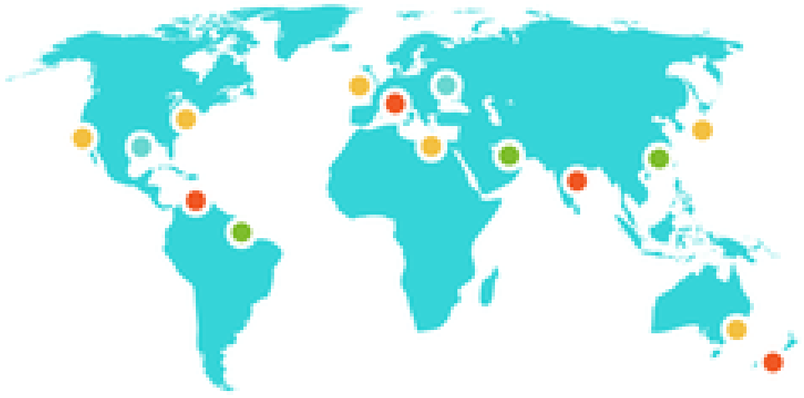
AUDIO

VIDEO



STATISTICS

VISUALIZE YOUR DATA



56298

DUIS AUTE IRURE DO REPREHENDERIT IN VELIT ESSE CILLUM FUGIAT NULLA PAR

56298

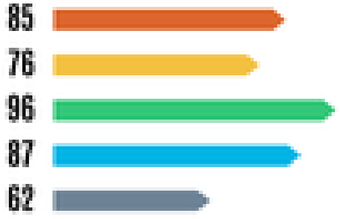
LOREM IPSUM DOLOR CONSECTETUR ADIPIS SED DO EIUSMOD TEMPOR UT LABORE ET DOLORE

56298

SINT OCCAECAT CUP PROIDENT, SUNT IN OFFICIA DESERUNT ANIM ID EST LABORUM.

56298

IPSUM DOLOR LOREM CONSECTETUR ADIPIS SED DO EIUSMOD TEM OFFICIA DESERUNT



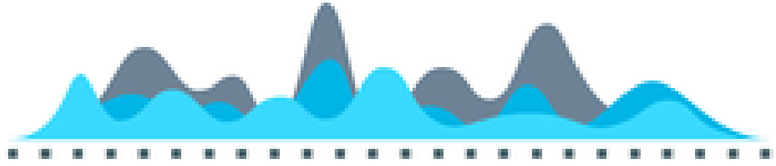
DUIS AUTE IRURE DOLOR REPREH ENDERIT IN VOL VELIT ESSE CILLUM

6-12
NISI UT ALIQUIP EX EA

16-25
COMMODO CONSEQUAT

36-55
ULLAMCO LABORIS

75-90
NOSTRUD EXERCITATION



SIT AMET, CONSECTETUR ADIPISCING ELIT, SED DO EIUSMOD TEMPOR INCIDIDUNT UT ENIM AD MINIM LOREM IPSUM DOLOR CONSECTETUR ADIPIS SED

DUIS AUTE IRURE DO REPREHENDERIT IN

LOREM IPSUM DOLOR CONSECTETUR ADIPIS

SINT OCCAECAT CUP PROIDENT, SUNT IN

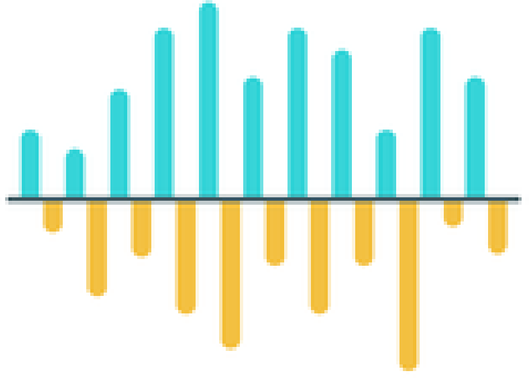
IPSUM DOLOR LOREM CONSECTETUR ADIPIS

65496

98713

12386

49513



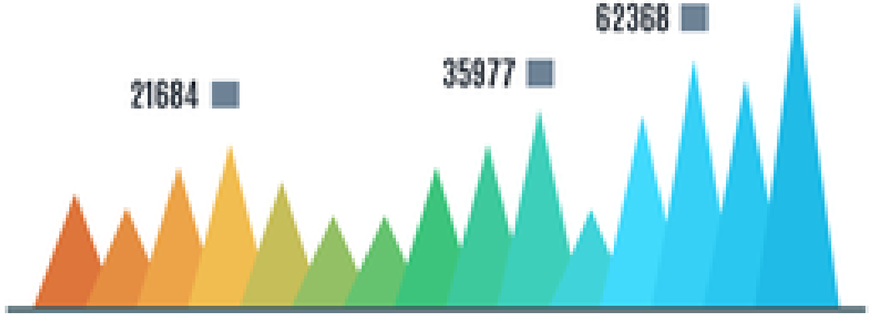
2576 DUIS AUTE IRURE DO

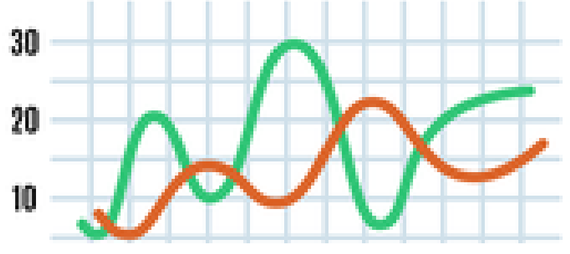
7852 CONSECTETUR ADIPIS

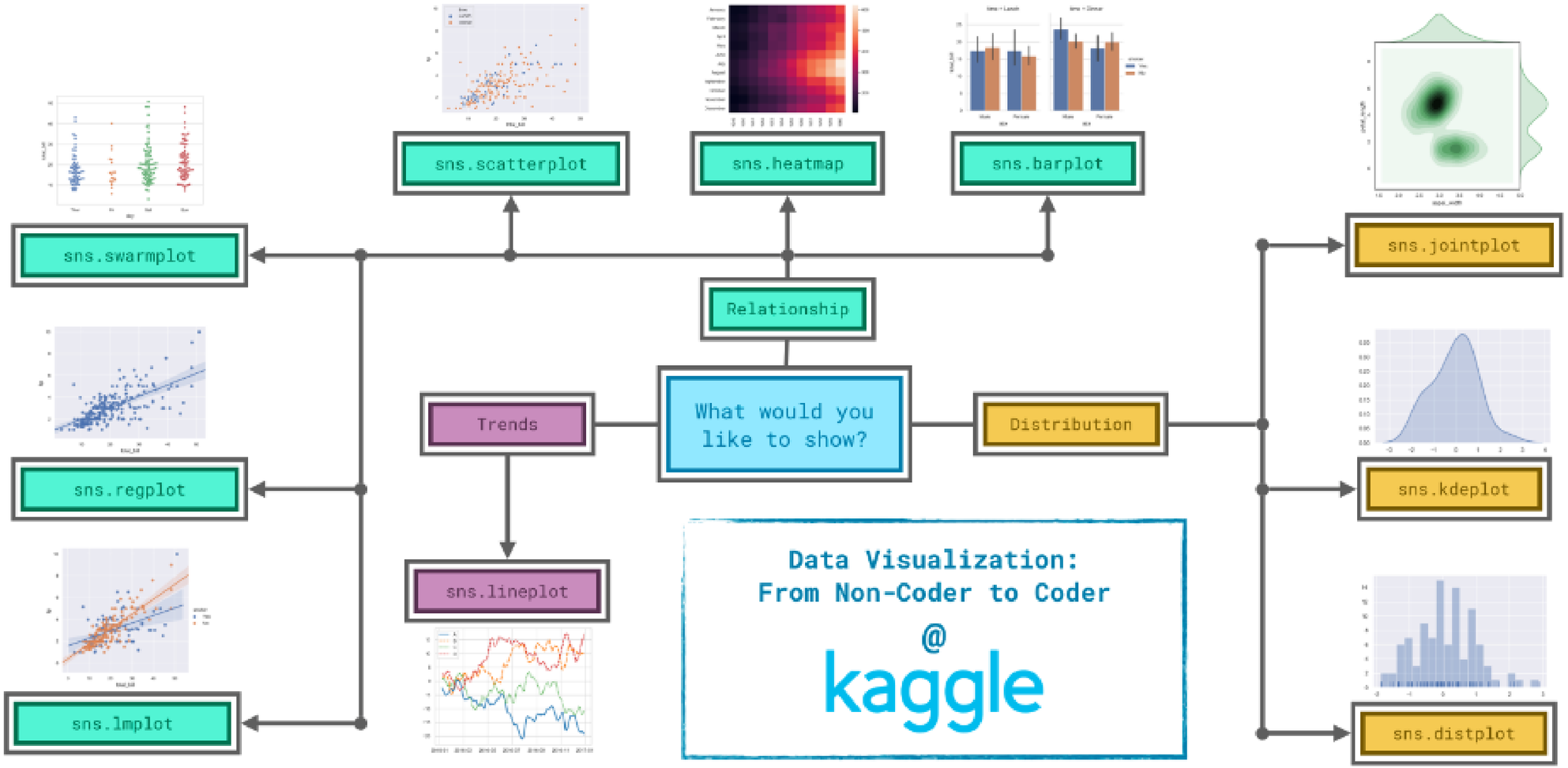
1638 SINT OCCAECAT CUP

5287 IPSUM DOLOR LOREM

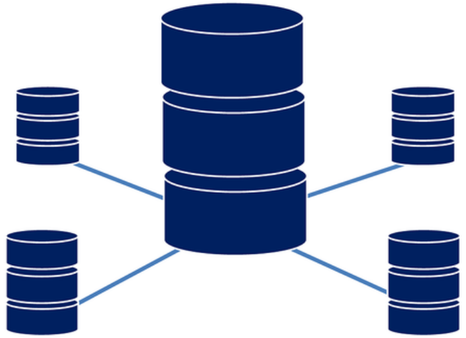
2787 PROIDENT SUNT IN






sns.scatterplot
sns.heatmap
sns.barplot
sns.swarmplot
Relationship
sns.jointplot
sns.regplot
Trends
What would you like to show?
Distribution
sns.kdeplot
sns.lineplot
Data Visualization:
From Non-Coder to Coder
@
kaggle
sns.lmplot
sns.distplot

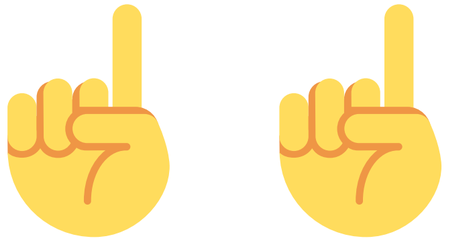
WHERE TO STORE DATA



FEATURES SELECTION

Gender	Age	Married	BankCustomer	EducationLevel	Employed	CreditScore	DriversLicense	City	ZipCode	ApprovalStatus.
M	28	Y	Y	B+3	N	10	Y	N	2034	YES
M	28	Y	Y	B+4	N	10	Y	N	2034	YES
M	28	Y	Y	B+5	N	10	Y	N	2034	YES
M	28	Y	Y	B+6	N	10	Y	N	2034	NO
M	28	Y	Y	B+7	N	10	Y	N	2034	YES
M	28	Y	Y	B+8	N	10	Y	N	2034	YES
M	28	Y	Y	B+9	N	10	Y	N	2034	NO

Gender	Age	Married	BankCustomer	EducationLevel	Employed	CreditScore	DriversLicense	City	ZipCode	ApprovalStatus.
M	28	Y	Y	B+3	N	10	Y	N	2034	YES
M	28	Y	Y	B+4	N	10	Y	N	2034	YES
M	28	Y	Y	B+5	N	10	Y	N	2034	YES
M	28	Y	Y	B+6	N	10	Y	N	2034	NO
M	28	Y	Y	B+7	N	10	Y	N	2034	YES
M	28	Y	Y	B+8	N	10	Y	N	2034	YES
M	28	Y	Y	B+9	N	10	Y	N	2034	NO



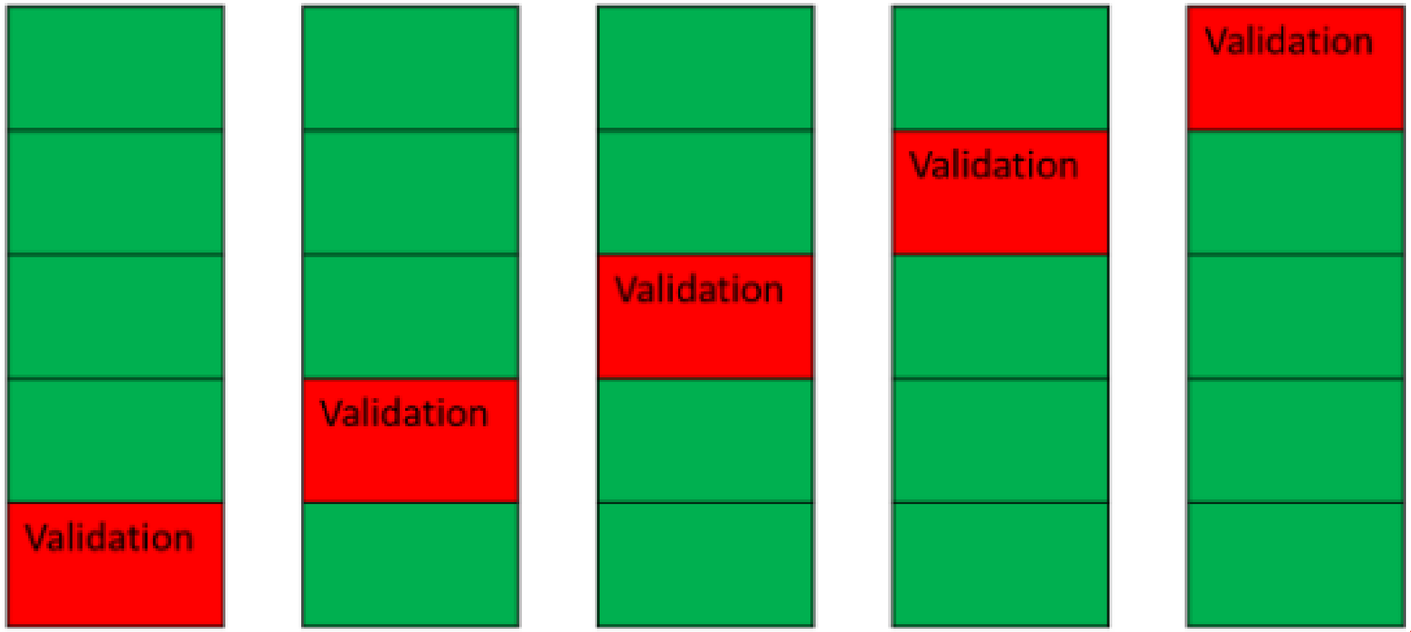
Gender	Age	Married	BankCustomer	EducationLevel	Employed	CreditScore	DriversLicense	City	ZipCode	ApprovalStatus.
M	28	Y	Y	B+3	N	10	Y	N	2034	YES
M	28	Y	Y	B+4	N	10	Y	N	2034	YES
M	28	Y	Y	B+5	N	10	Y	N	2034	YES
M	28	Y	Y	B+6	N	10	Y	N	2034	NO
M	28	Y	Y	B+7	N	10	Y	N	2034	YES
M	28	Y	Y	B+8	N	10	Y	N	2034	YES
M	28	Y	Y	B+9	N	10	Y	N	2034	NO

Gender	Age	Married	BankCustomer	EducationLevel	Employed	CreditScore	ZipCode	ApprovalStatus.
M	28	Y	Y	B+3	N	10	2034	YES
M	28	Y	Y	B+4	N	10	2034	YES
M	28	Y	Y	B+5	N	10	2034	YES
M	28	Y	Y	B+6	N	10	2034	NO
M	28	Y	Y	B+7	N	10	2034	YES
M	28	Y	Y	B+8	N	10	2034	YES
M	28	Y	Y	B+9	N	10	2034	NO

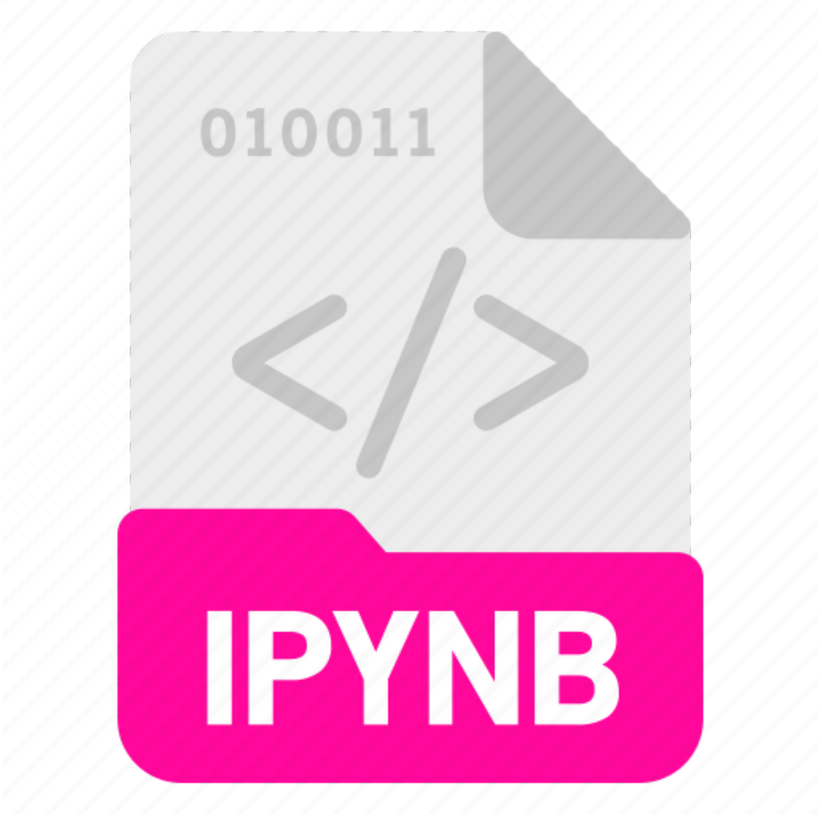
SPLIT YOUR DATA

Training
Test

Training
Validation
Test

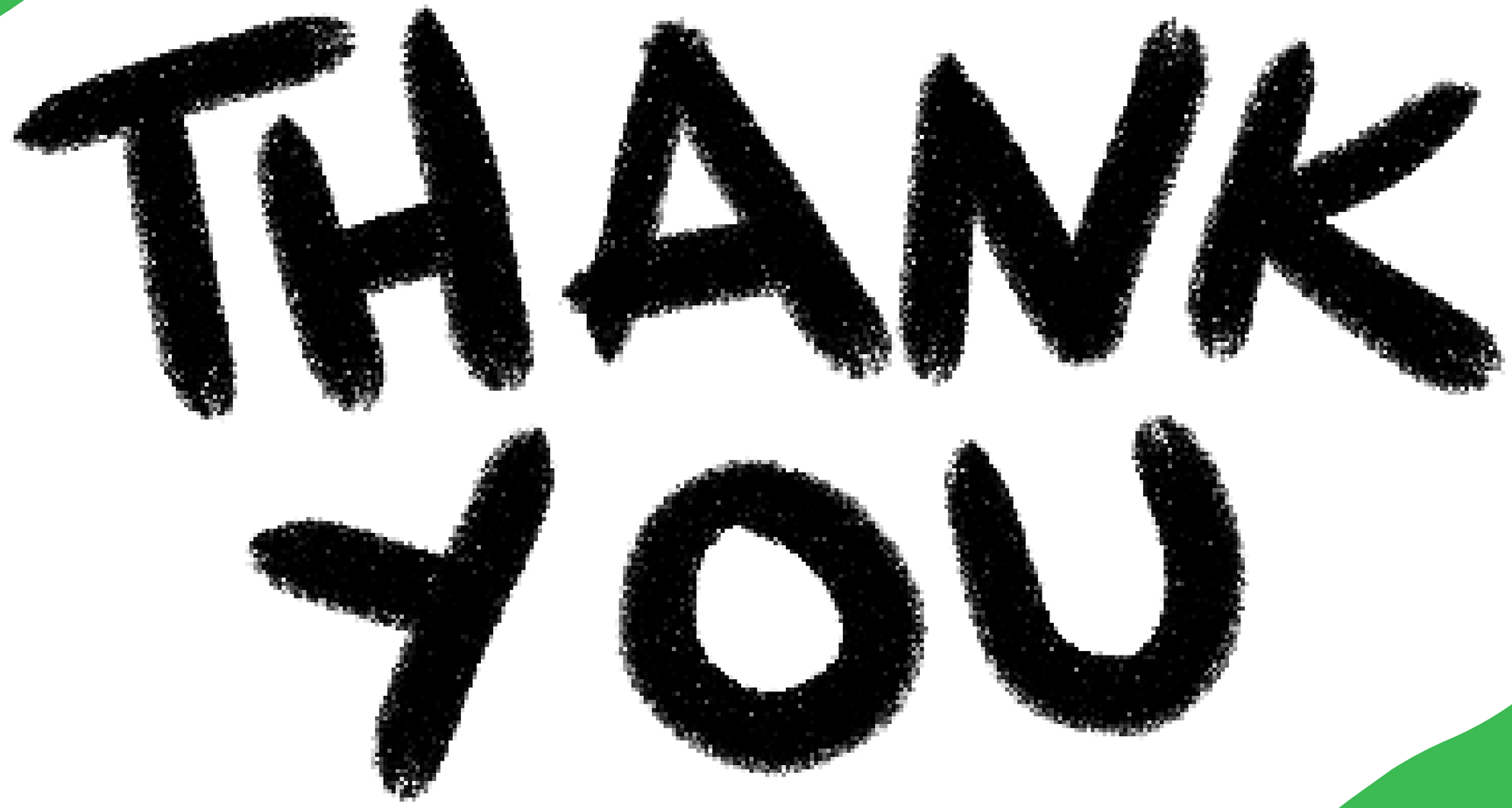

Validation
Validation
Validation
Validation
Validation

WHAT'S NEXT ?


010011
IPYNB

kaggle

Q
A


THANK
YOU