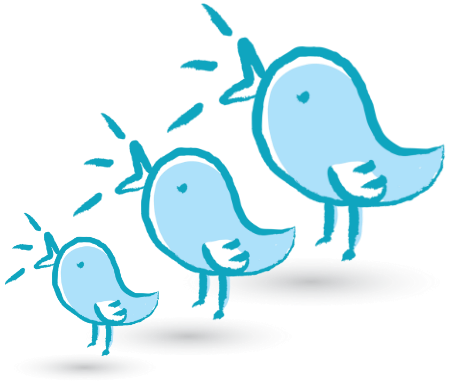
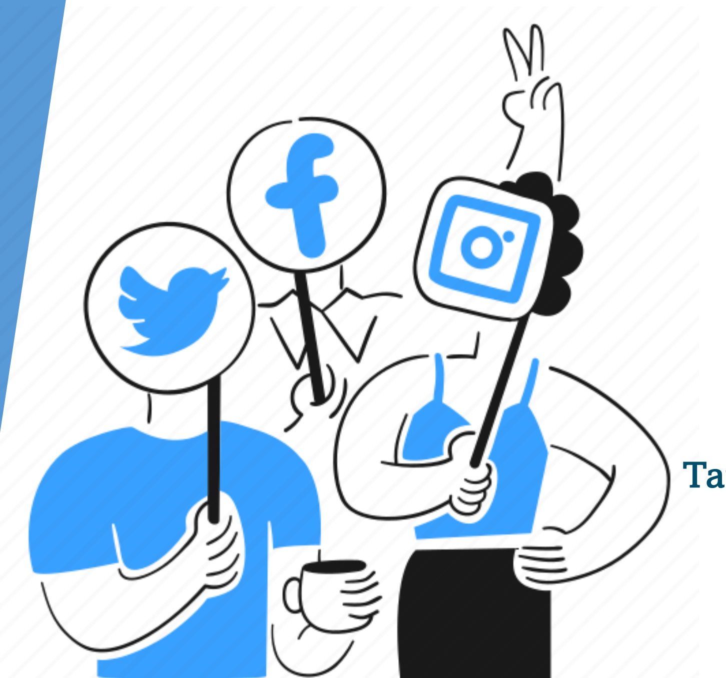
Building Predictive Models for Social Media Data





Dr. Geetha Raju

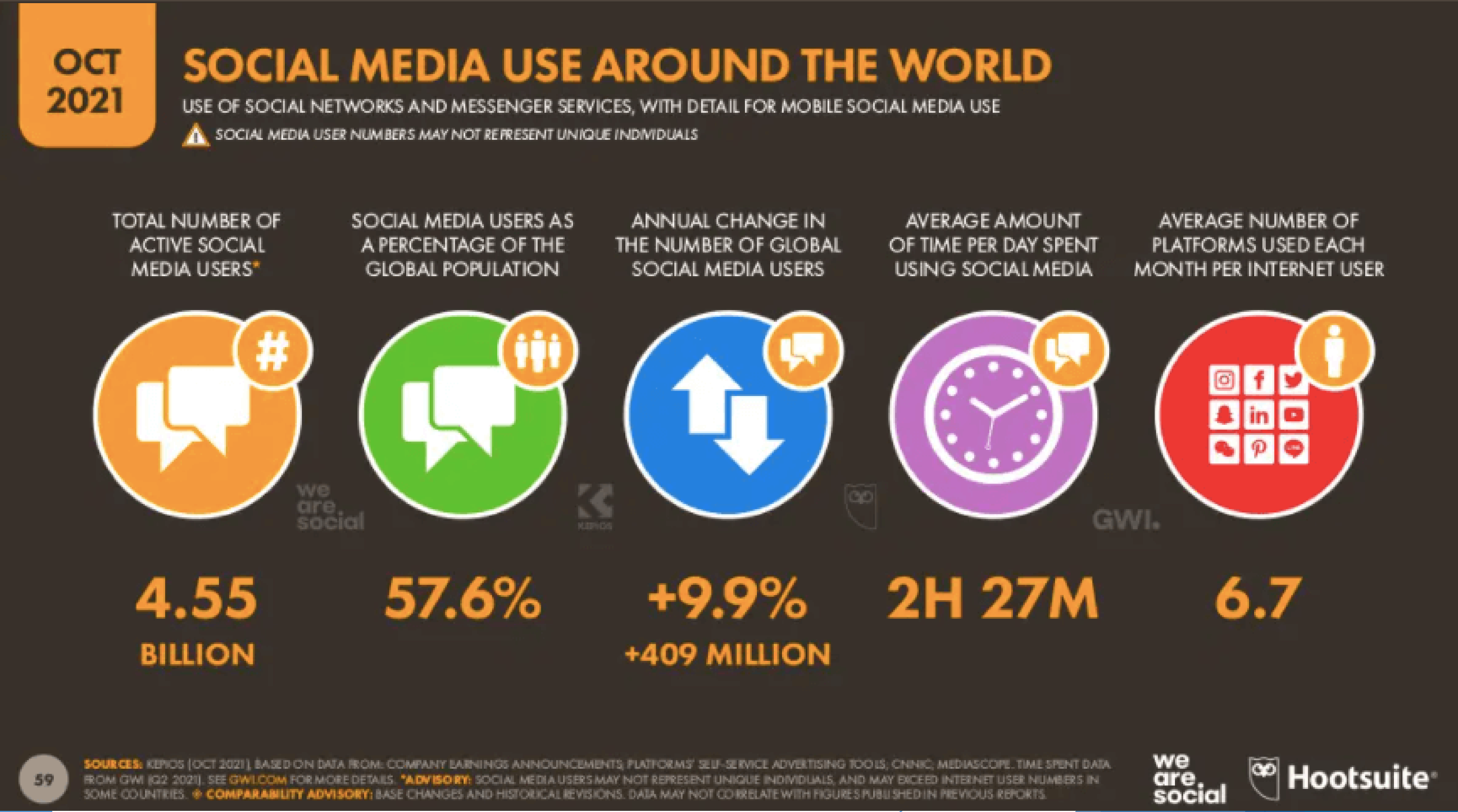
ML Engineer

Tamil Nadu e-Governance Agency, Chennai

Social Media and Data Analytics

- Social Media
  - Reinforce social bonds; manage social identities
  - Easy and open access
  - Social networking (Facebook, LinkedIn, Google+)
  - Microblogging (Twitter, Tumblr)
  - Photo sharing (Instagram, Snapchat, Pinterest)
  - Video sharing (YouTube, Facebook Live, Periscope, Vimeo)
- Practitioners, researchers and analyst – rich resources – social media data
- What they do?
  - Gather data
  - Find meaning / context
  - Derive insights that support decision making
  - Analyze / predict performance
- What is data in SM?
  - Post specific data - likes, reactions, comments, clicks, previews, etc.,
  - User specific data – name, DOB, followers, friends, etc.,
  - Network specific data – followers, following, friends, community / group, etc.,

What's Happening ?!?



OCT 2021

SOCIAL MEDIA USE AROUND THE WORLD

USE OF SOCIAL NETWORKS AND MESSENGER SERVICES, WITH DETAIL FOR MOBILE SOCIAL MEDIA USE

⚠ *SOCIAL MEDIA USER NUMBERS MAY NOT REPRESENT UNIQUE INDIVIDUALS*

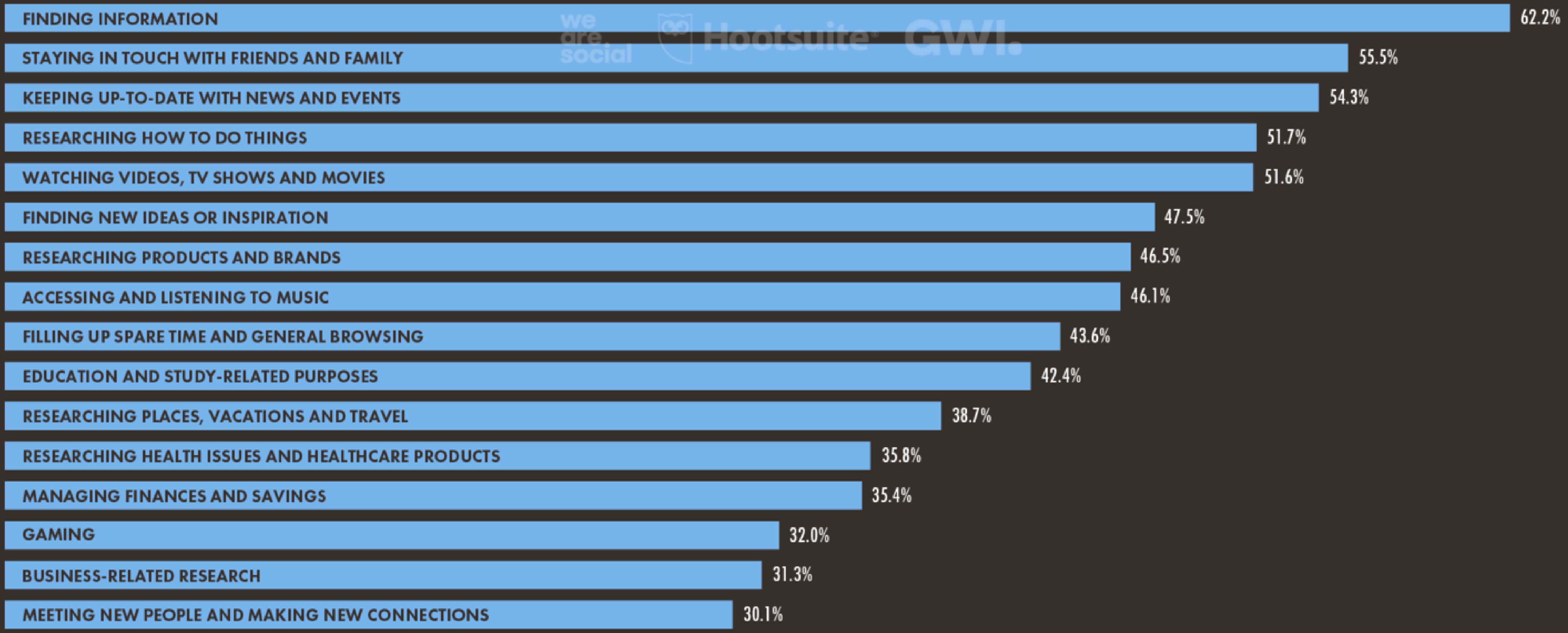
TOTAL NUMBER OF ACTIVE SOCIAL MEDIA USERS*	SOCIAL MEDIA USERS AS A PERCENTAGE OF THE GLOBAL POPULATION	ANNUAL CHANGE IN THE NUMBER OF GLOBAL SOCIAL MEDIA USERS	AVERAGE AMOUNT OF TIME PER DAY SPENT USING SOCIAL MEDIA	AVERAGE NUMBER OF PLATFORMS USED EACH MONTH PER INTERNET USER
4.55 BILLION	57.6%	+9.9% +409 MILLION	2H 27M	6.7

SOURCES: KEPIOS (OCT 2021), BASED ON DATA FROM: COMPANY EARNINGS ANNOUNCEMENTS; PLATFORMS' SELF-SERVICE ADVERTISING TOOLS; CNNIC; MEDIASCOPE. TIME SPENT DATA FROM GWI (Q2 2021). SEE GWI.COM FOR MORE DETAILS. *ADVISORY: SOCIAL MEDIA USERS MAY NOT REPRESENT UNIQUE INDIVIDUALS, AND MAY EXCEED INTERNET USER NUMBERS IN SOME COUNTRIES. ◈ COMPARABILITY ADVISORY: BASE CHANGES AND HISTORICAL REVISIONS. DATA MAY NOT CORRELATE WITH FIGURES PUBLISHED IN PREVIOUS REPORTS.

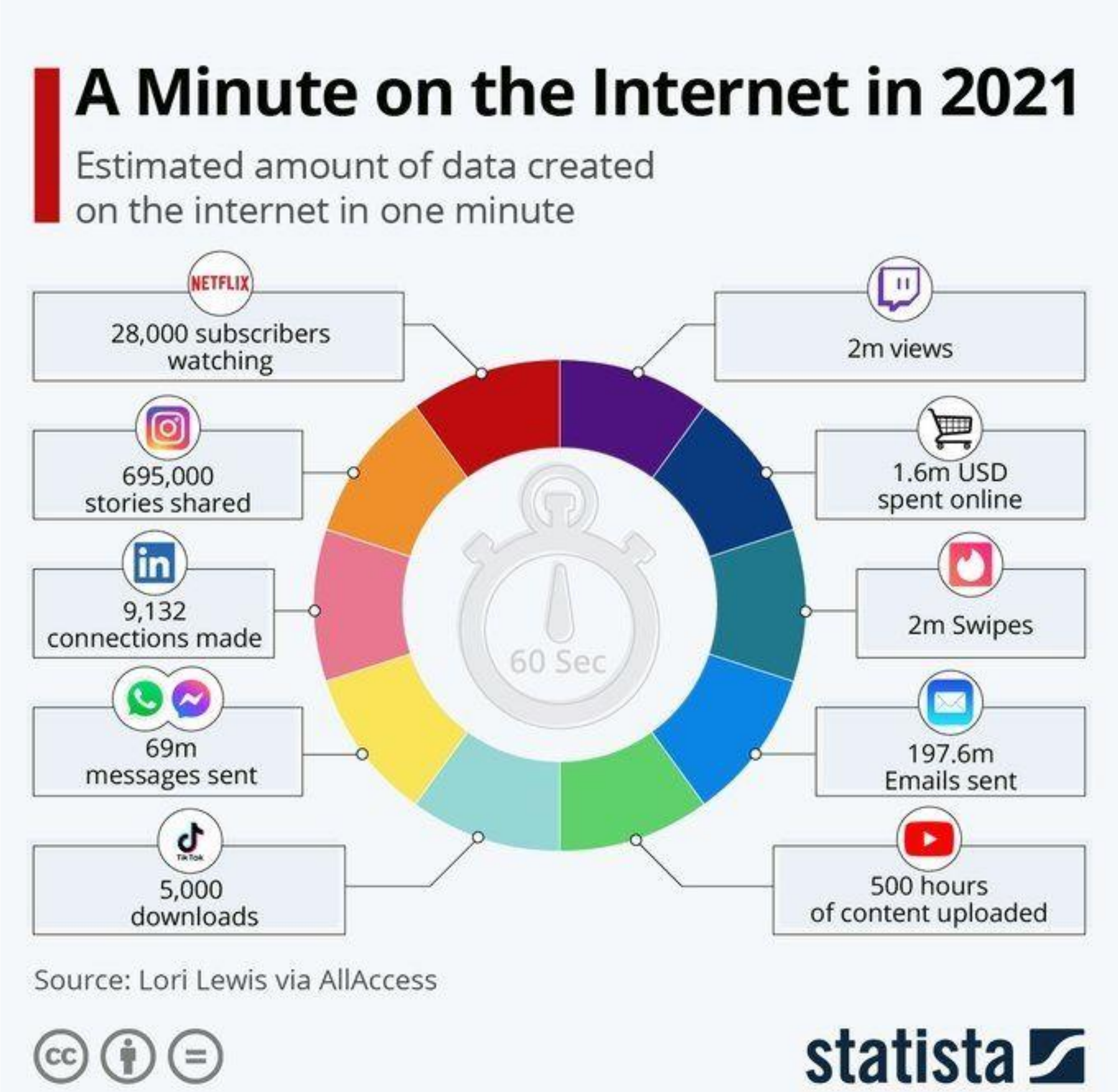
JUL
2021

REASONS FOR USING THE INTERNET

PRIMARY REASONS WHY GLOBAL INTERNET USERS AGED 16 TO 64 USE THE INTERNET



SOURCE: GWI (Q1 2021). FIGURES REPRESENT THE FINDINGS OF A BROAD GLOBAL SURVEY OF INTERNET USERS AGED 16 TO 64. SEE GWI.COM FOR MORE DETAILS.



A Minute on the Internet in 2021

Estimated amount of data created
on the internet in one minute

NETFLIX
28,000 subscribers watching

695,000 stories shared

9,132 connections made

69m messages sent

5,000 downloads

2m views

1.6m USD spent online

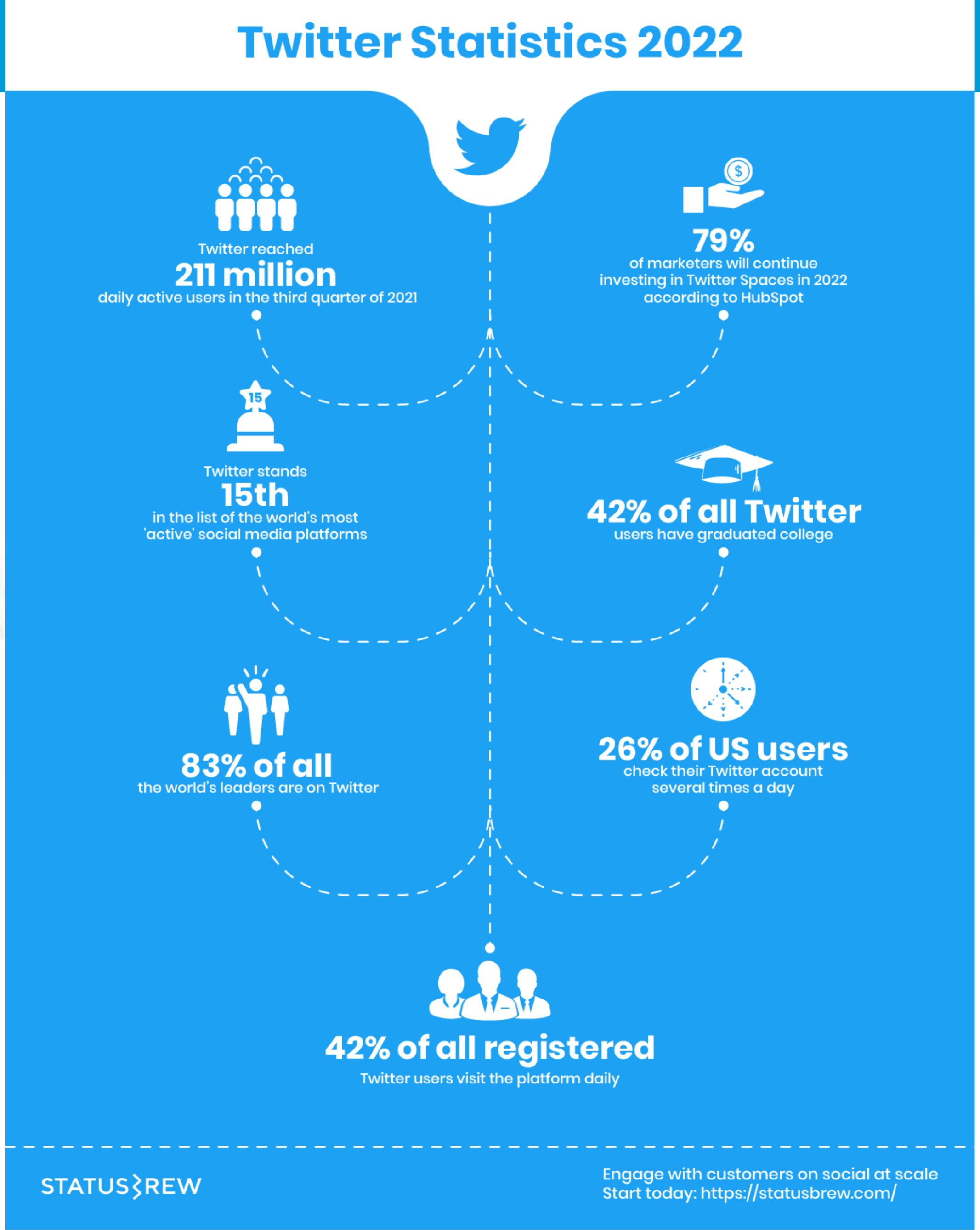
2m Swipes

197.6m Emails sent

500 hours of content uploaded

Source: Lori Lewis via AllAccess

statista



Twitter Statistics 2022

Twitter reached **211 million** daily active users in the third quarter of 2021

79% of marketers will continue investing in Twitter Spaces in 2022 according to HubSpot

Twitter stands **15th** in the list of the world's most 'active' social media platforms

42% of all Twitter users have graduated college

83% of all the world's leaders are on Twitter

26% of US users check their Twitter account several times a day

42% of all registered Twitter users visit the platform daily

STATUSBREW

Engage with customers on social at scale
Start today: https://statusbrew.com/

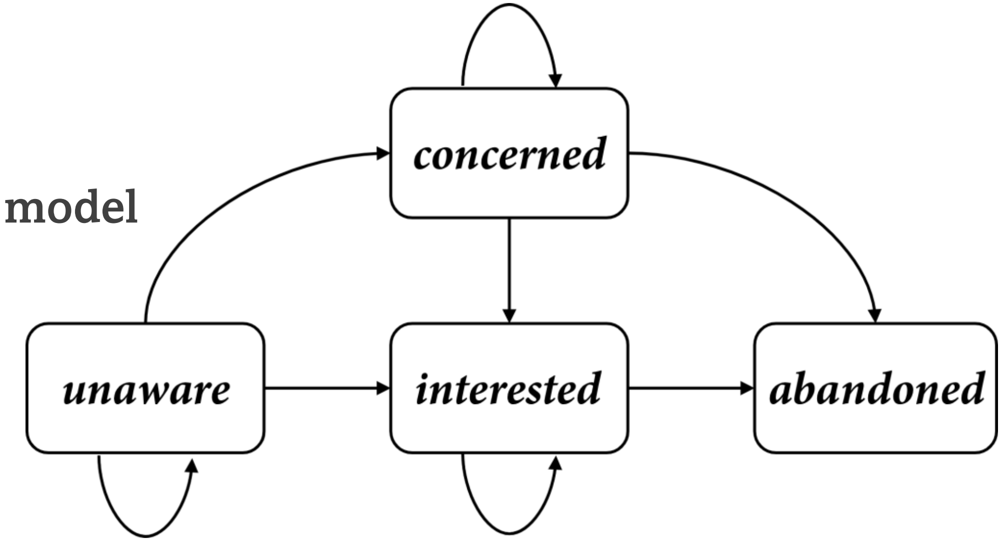
Social Media and Data Analytics

- Challenges in social media data
  - Time Sensitivity
  - Short length
  - Unstructured form
- More than 7 million web pages of text are being added to our collective repository, daily
- Processing speed
  - 15,000- 250,000 pages an hour – TM software
  - 60 pages for humans

Real Life Vs Online Social Media

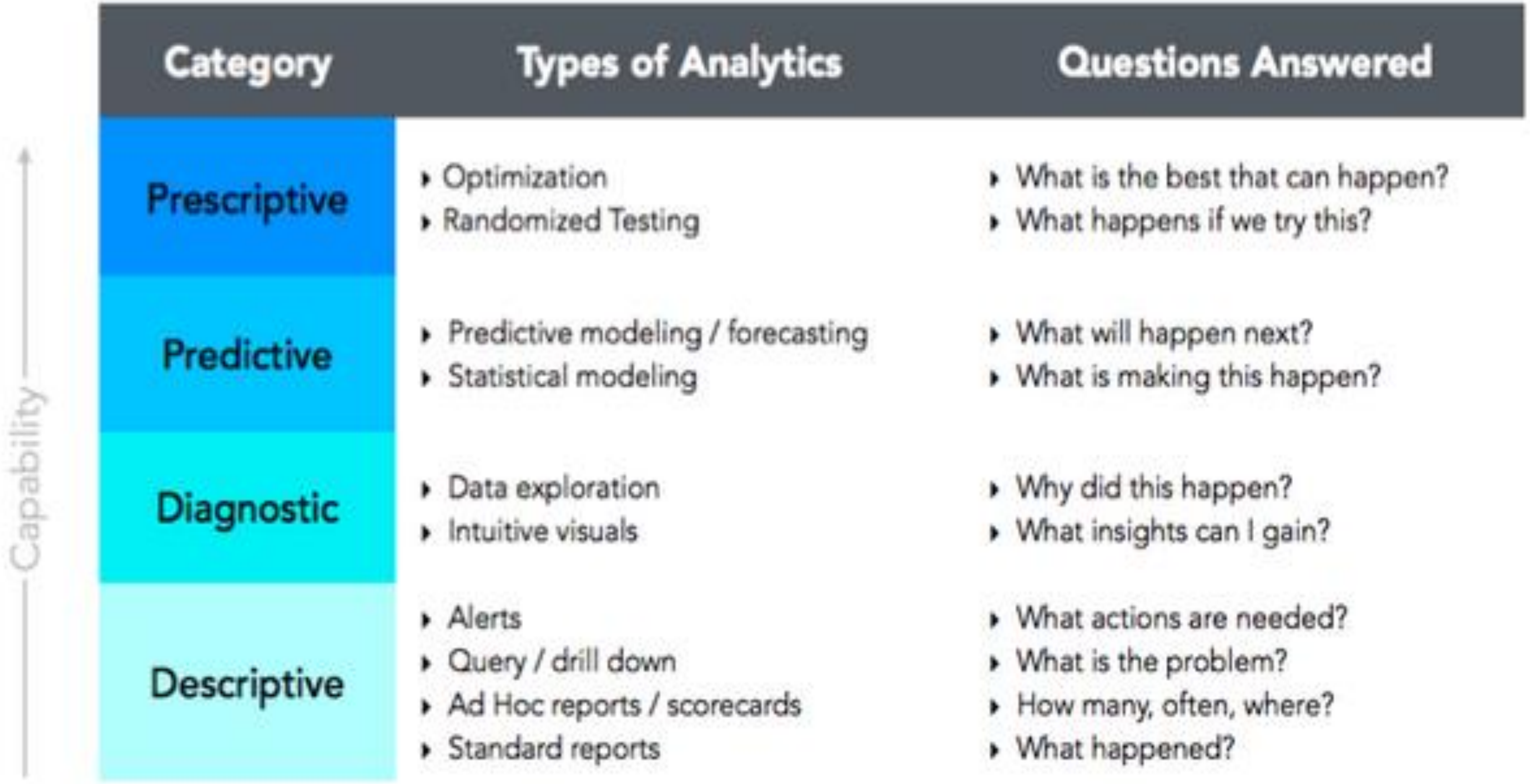
Information Spread and Flow in Social Media

- Information sharing
  - Why - Pleasure and helps in gaining public attention
  - What - Profile, Status, Location and shred content
  - How - Noisy and unstructured
- Online social network information spreading (OSIS) model



- Information amplification
- Influenced by psychological and social factors
- False news
- How to analyze information flows?
  - Subgraph constructions
  - Activity and degree distributions
  - Network analysis

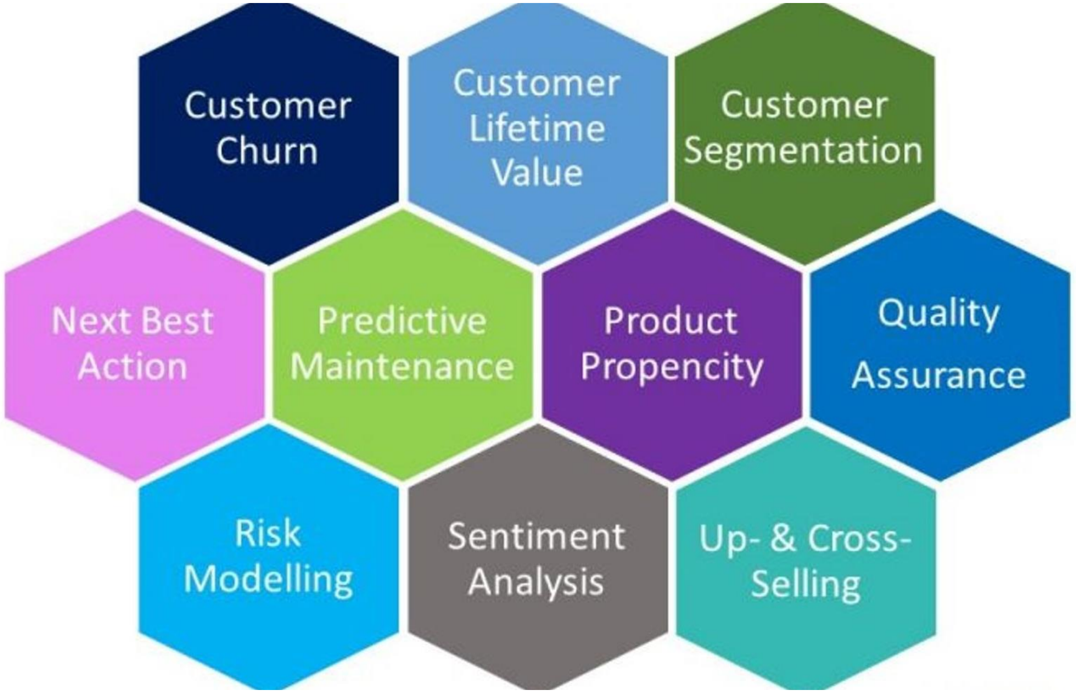
What ML / AL can do?


Capability

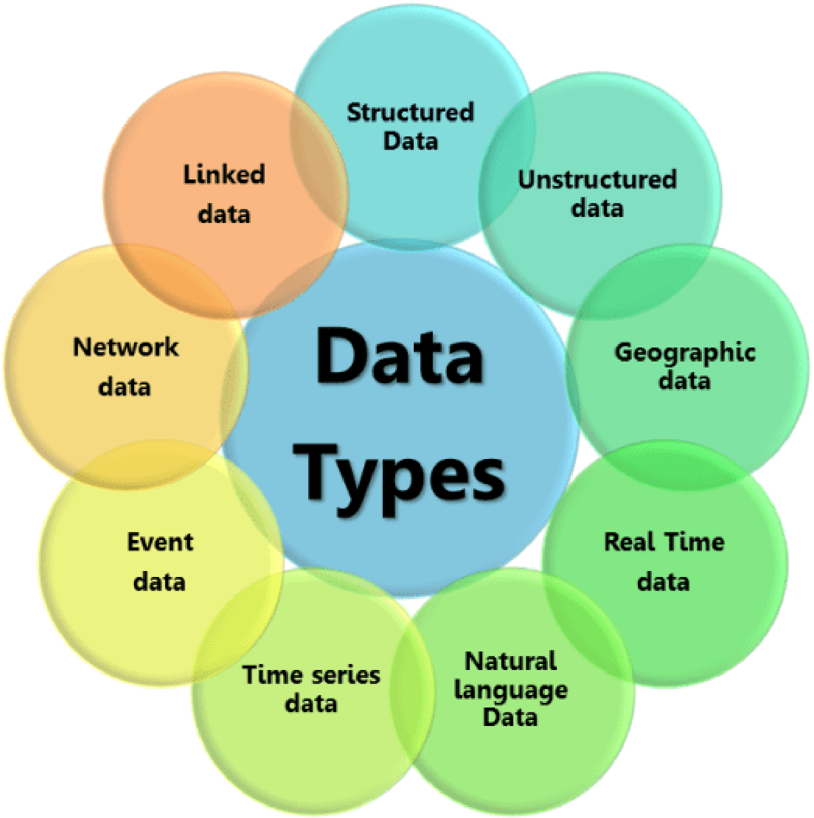
Category	Types of Analytics	Questions Answered
Prescriptive	Optimization Randomized Testing	What is the best that can happen? What happens if we try this?
Predictive	Predictive modeling / forecasting Statistical modeling	What will happen next? What is making this happen?
Diagnostic	Data exploration Intuitive visuals	Why did this happen? What insights can I gain?
Descriptive	Alerts Query / drill down Ad Hoc reports / scorecards Standard reports	What actions are needed? What is the problem? How many, often, where? What happened?

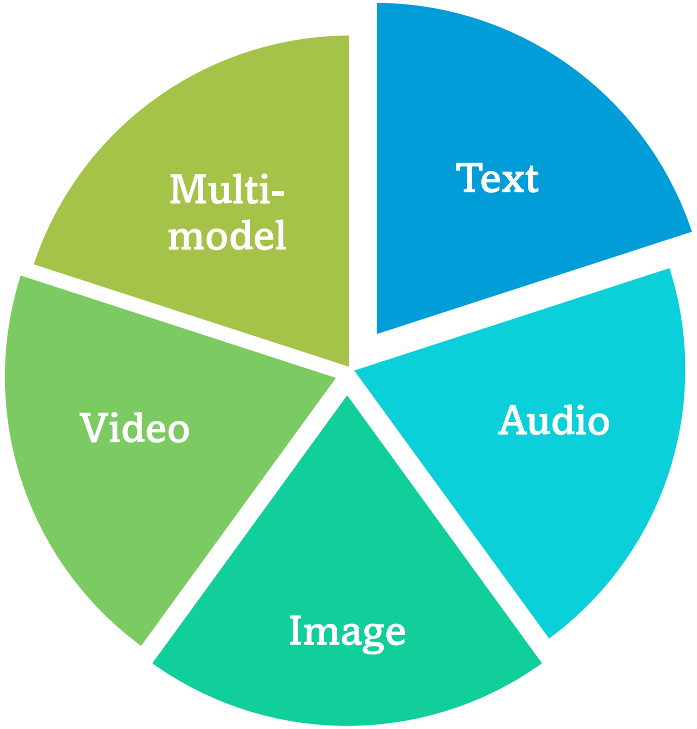
Scope of Predictive analytics on Social Media Data

- Predictive systems
  - Stock Market Predictions
  - Spam Detection
- Content based Recommender Systems
  - News, Movie, Product reviews / suggestions.
- Social media and Text Analytics
  - Sentiment Analysis – Product / Brand / Topic
  - Topic / Keyword / Phrase Identification
  - Cybercrimes and Cyberbullying
- Linguistic Rules and Machine Learning Analysis

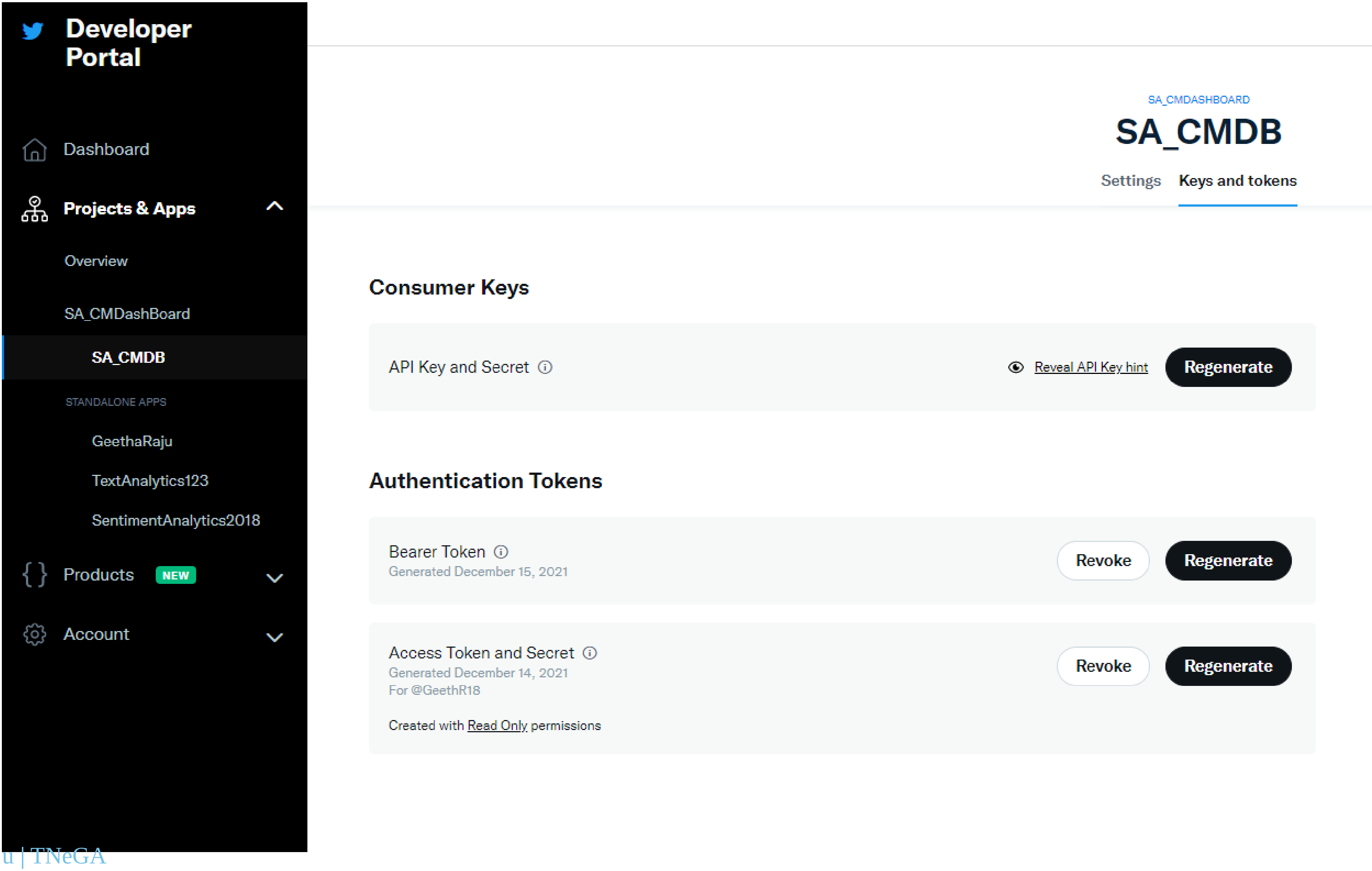


So, Lets Start with Data Preparation





Data Collection Pipeline – Twitter Developer Account



SA_CMDASHBOARD

SA_CMDB

Consumer Keys

API Key and Secret

Authentication Tokens

Bearer Token

Generated December 15, 2021

Access Token and Secret

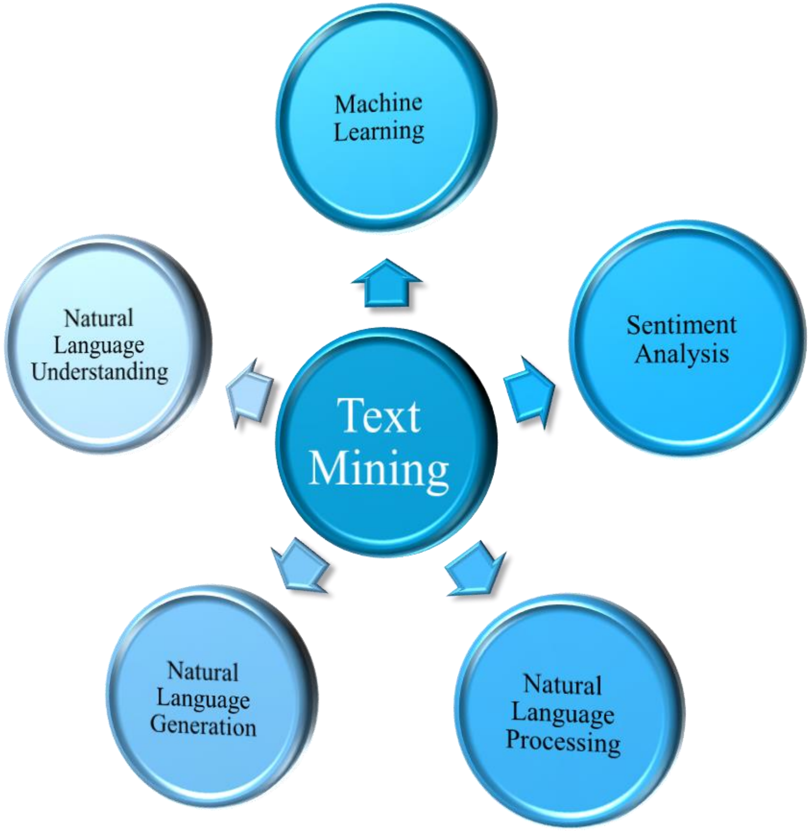
Generated December 14, 2021

For @GeethR18

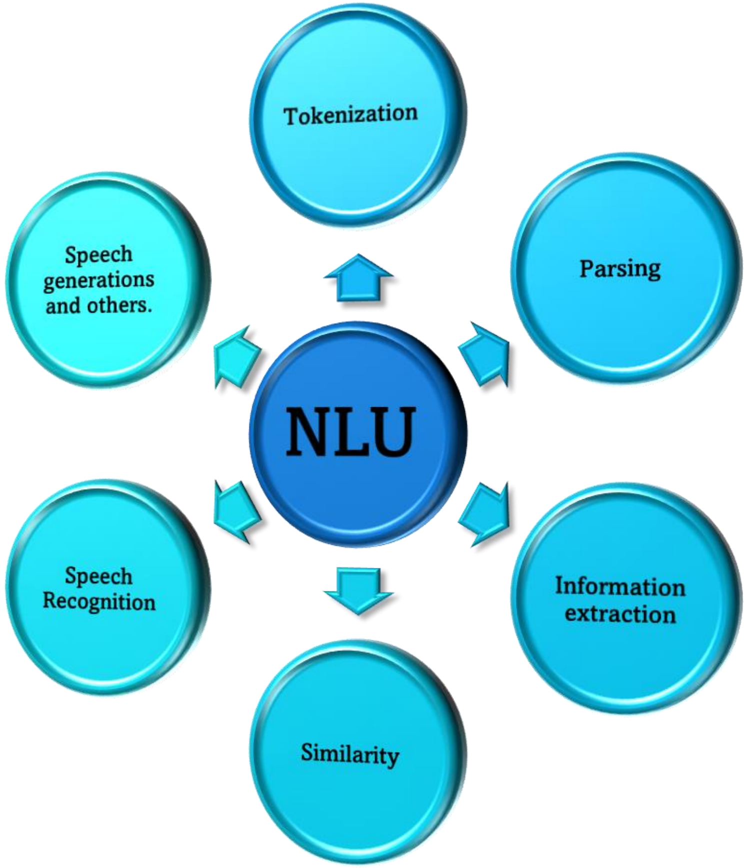
Created with Read Only permissions

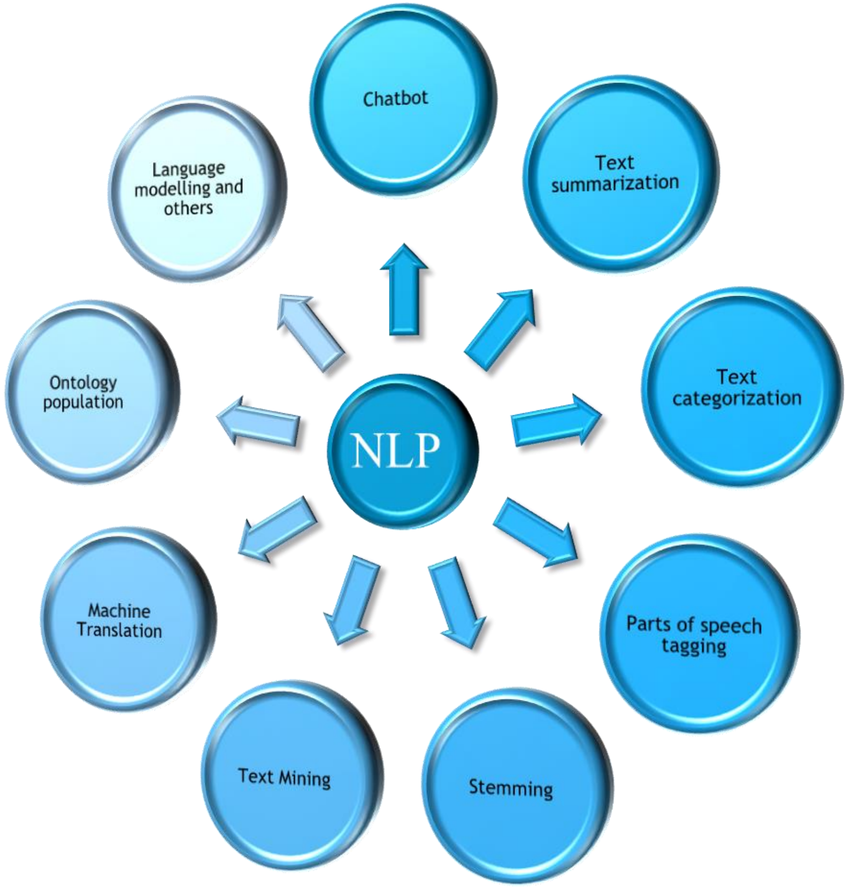
Text Data Features

- Generate structure where there is no structure
- Challenges
  - unstructured textual form (80% - 90% of world’s data)
  - large text data
  - high dimension but sparse
  - word / phrase types in various languages
  - complex relationship between concepts in text
  - word ambiguity and context sensitivity
  - noisy data

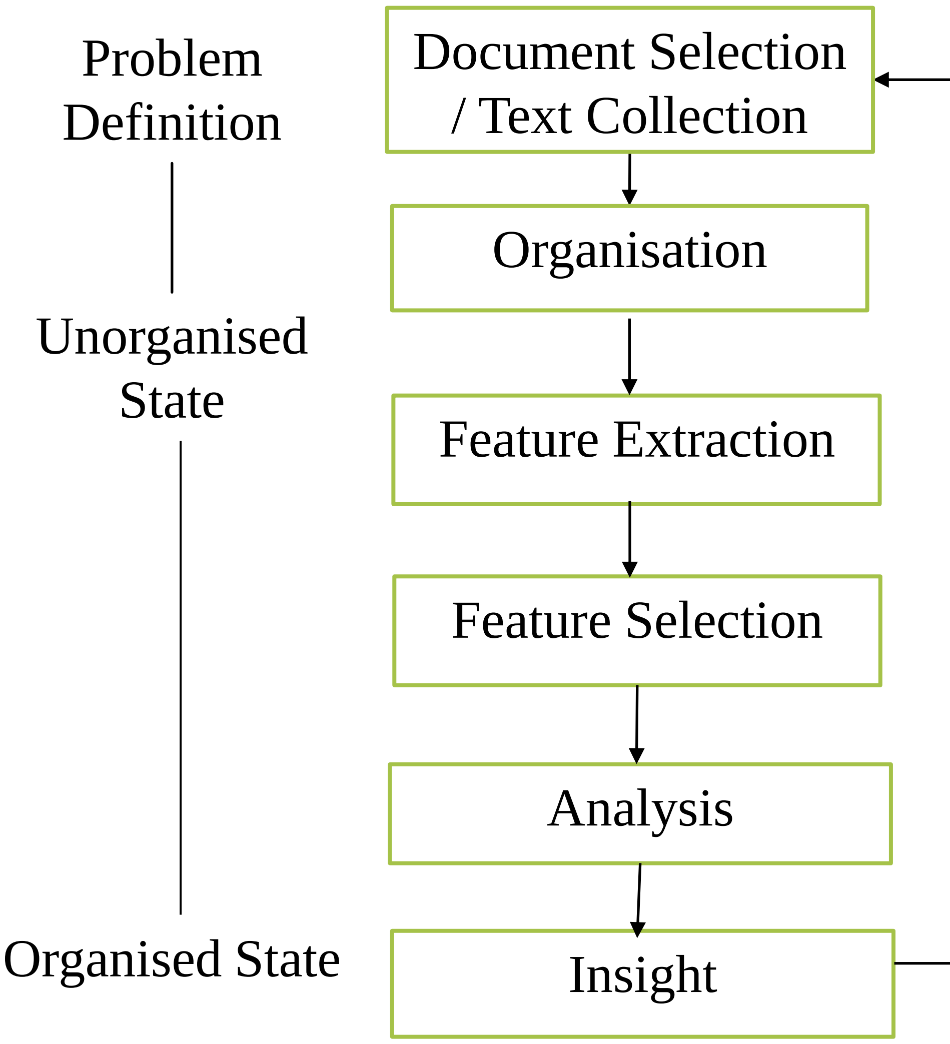

Machine Learning
Natural Language Understanding
Sentiment Analysis
Text Mining
Natural Language Generation
Natural Language Processing

NLU Vs NLP


Tokenization
Speech generations and others.
Parsing
NLU
Speech Recognition
Information extraction
Similarity


Chatbot
Language modelling and others
Text summarization
Ontology population
NLP
Text categorization
Machine Translation
Parts of speech tagging
Text Mining
Stemming

Overview of building a Text based Predictive Model



Problem Definition

Unorganised State

Organised State

Document Selection / Text Collection

Organisation

Feature Extraction

Feature Selection

Analysis

Insight

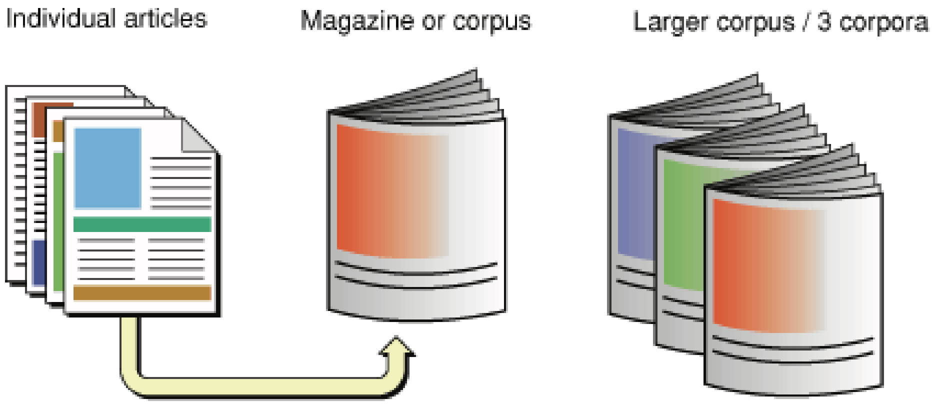
Predictive Modelling

Text Analytics - Terminologies

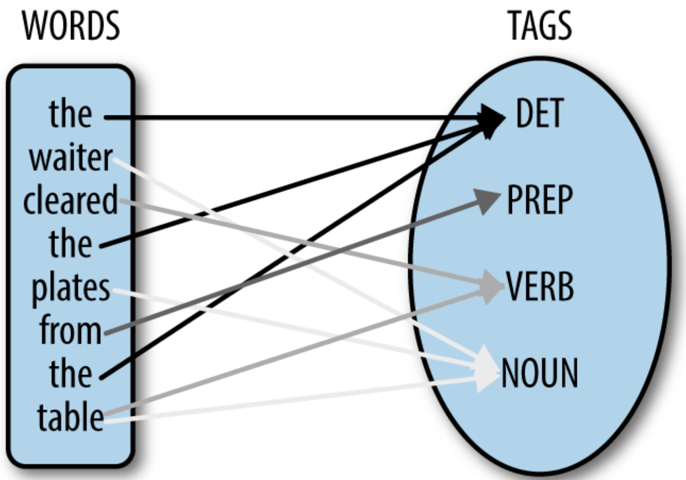
Document

Tokens / Terms

Stopwords



Corpus



Parts of Speech Tagging

Textual Features – Extraction Methods

- Bag-of-words
- n-grams
- Scoring words – counts, frequencies, binary
    - Histogram
    - Document Term Matrix
    - Term Frequency – Inverse Document Frequency
- Word hashing
- Word embeddings

Document Term Matrix

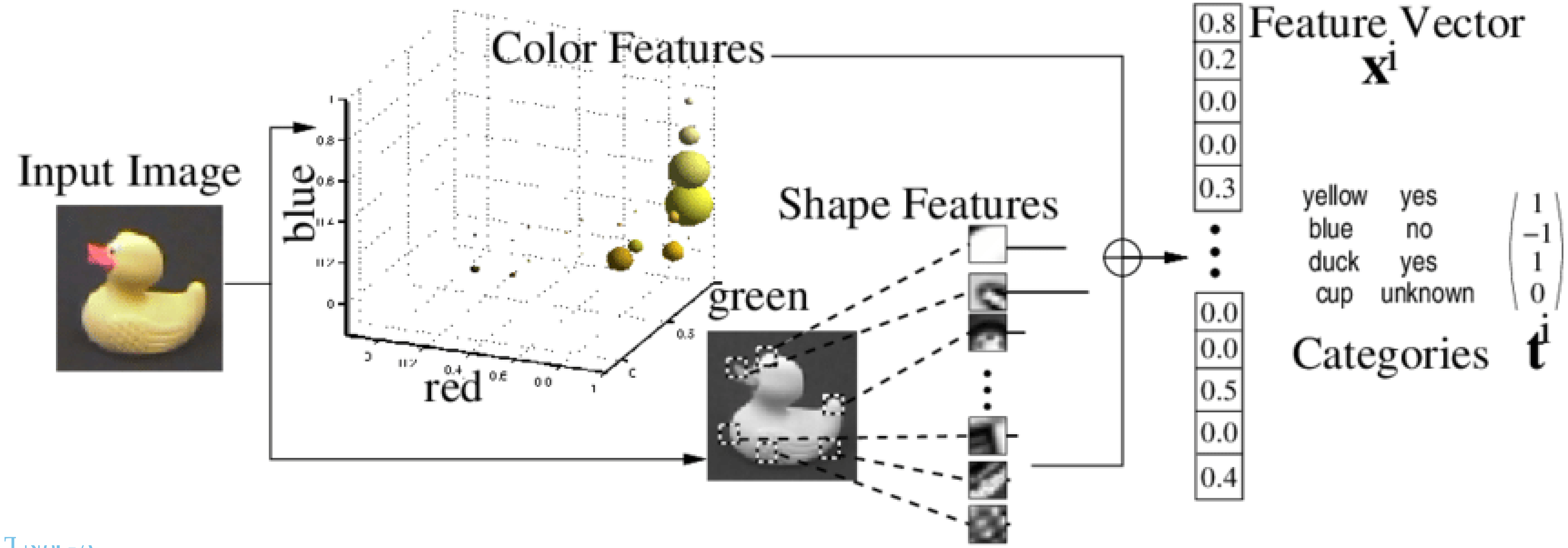
	Term 1	Term 2	.	.	Term n
Document 1					
Document 2					
.					
.					
Document m					

Term Frequency – Inverse Document Frequency (TF-IDF)

- Importance / Significance of a word.
- $TF = \frac{Number\ of\ times\ Term\ t\ occurs}{Total\ number\ of\ terms}$
- $IDF = Log(\frac{Total\ number\ of\ docs}{Number\ of\ docs\ containing\ Term\ t})$
- $TF - IDF = TF * IDF$

Image Feature

- Matrix of numbers
- Size of this matrix depends on the number of pixels of the input image.
- Pixel Values
  - Intensity and Brightness - how bright that pixel is?
  - what color it should be?
  - Important shape / objects / edges



Audio Features

Abstraction Level

High – Understood and enjoyed by humans

- instrumentation, key, chords, melody, harmony, rhythm, genre, mood, etc

Mid – Perceived by humans

- pitch, beat-related descriptors, note onsets, fluctuation patterns, s, etc

Low - statistical features which sense to the machine, but not to humans

- amplitude envelope, energy, spectral centroid, spectral flux, zero-crossing rate,

Temporal scope

Instantaneous

- Range in milliseconds – 10ms

Segment-level

- Wider range

Global

- Aggregate feature for whole word / sentence

Signal domains

Time zone

- waveforms of the raw audio.

Frequency zone

- Frequency component

Time-Frequency zone

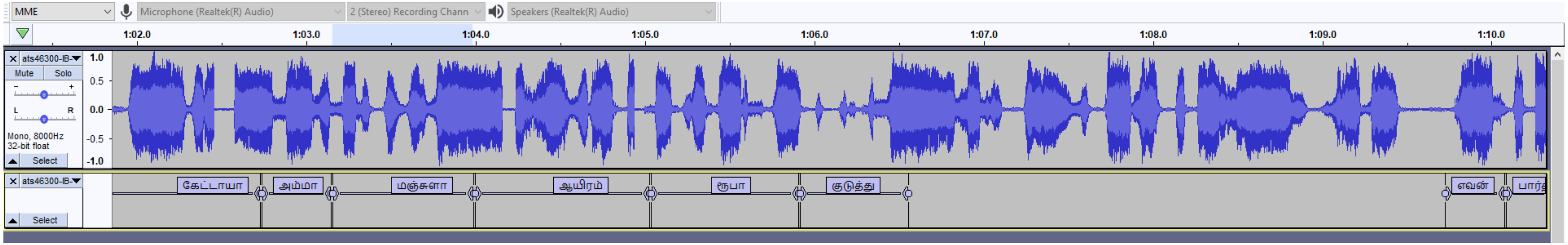
- Combination of Time and Frequency

Time domain

Amplitude

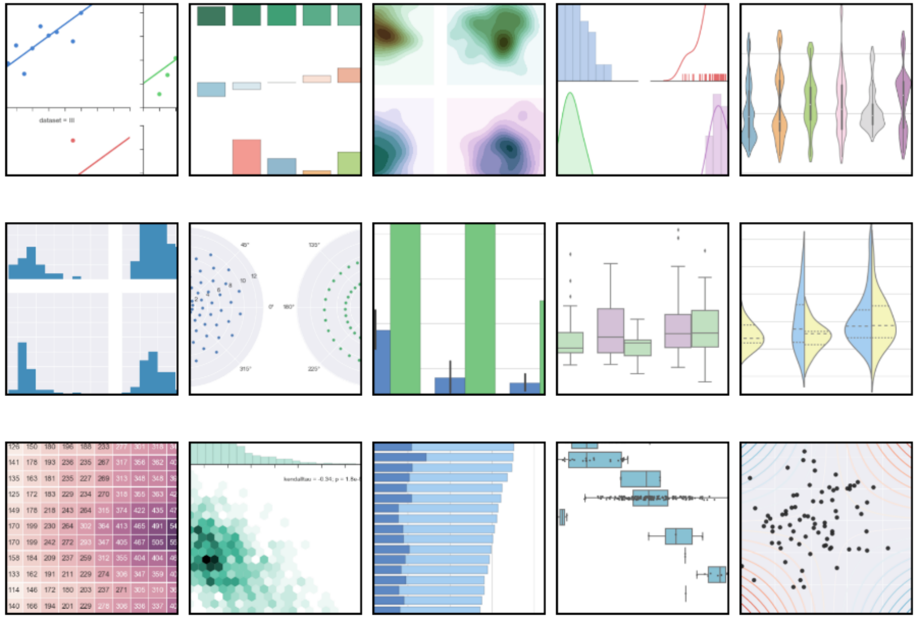
Root mean square energy

Zero crossing rate



Exploratory Data Analytics

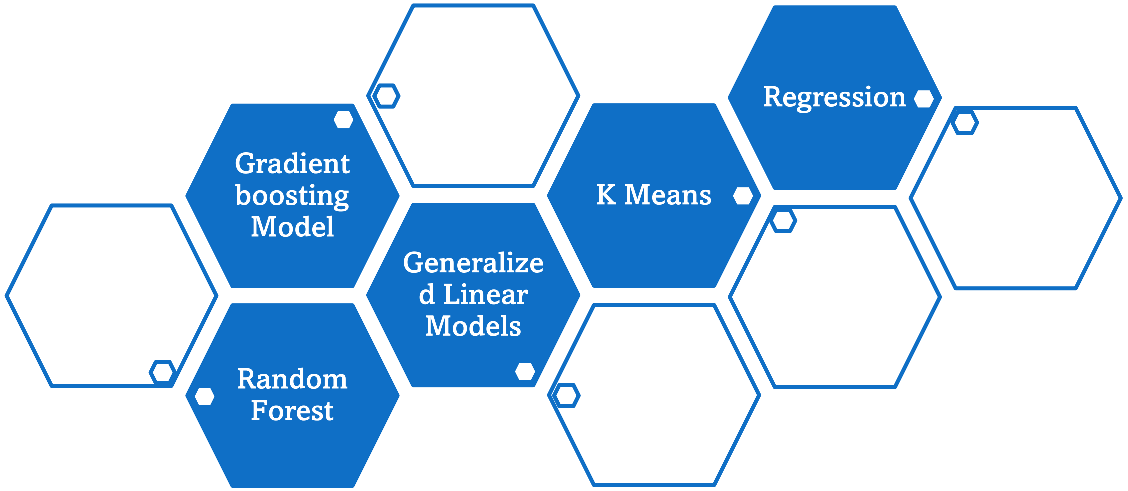
- Examine the data distribution
- Handling missing values of the dataset
- Handling the outliers
- Removing duplicate data
- Encoding the categorical variables
- Normalizing and Scaling



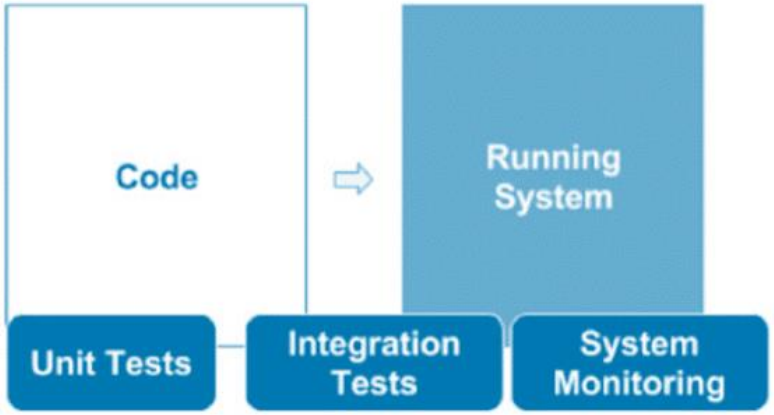
DATA
SORTED
ARRANGED
PRESENTED VISUALLY
EXPLAINED WITH A STORY

Predictive Modeling

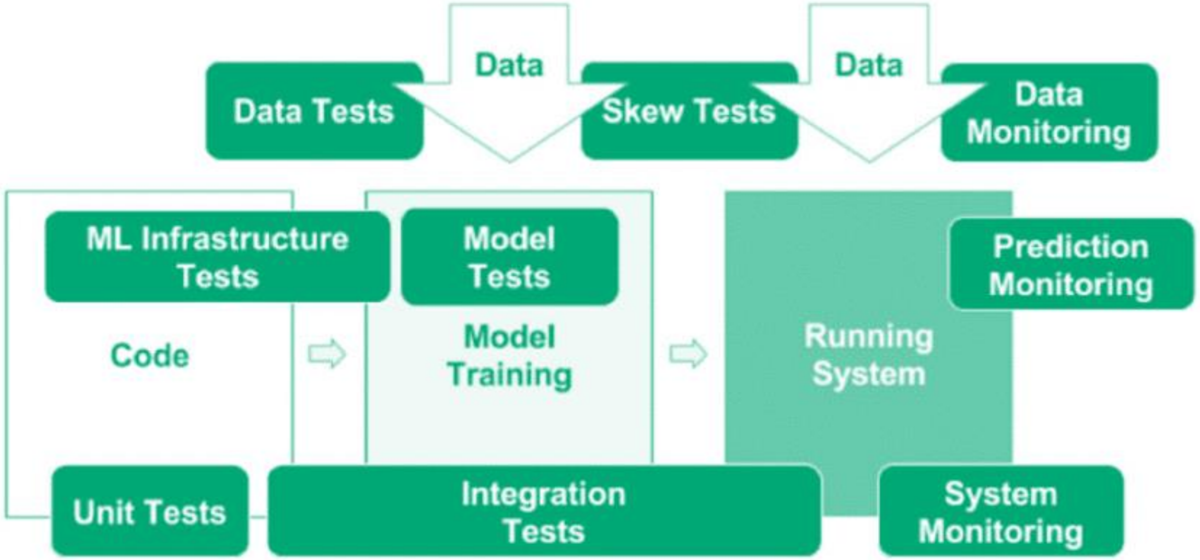
- Statistical technique using machine learning and data mining
- Requires historical and existing data
- Forecast future based on historical and existing data
- Types of Predictive Modelling
  - Classification Model
  - Clustering Model
  - Time Series Model
  - Forecast Model
  - Language Model



Traditional Vs ML System



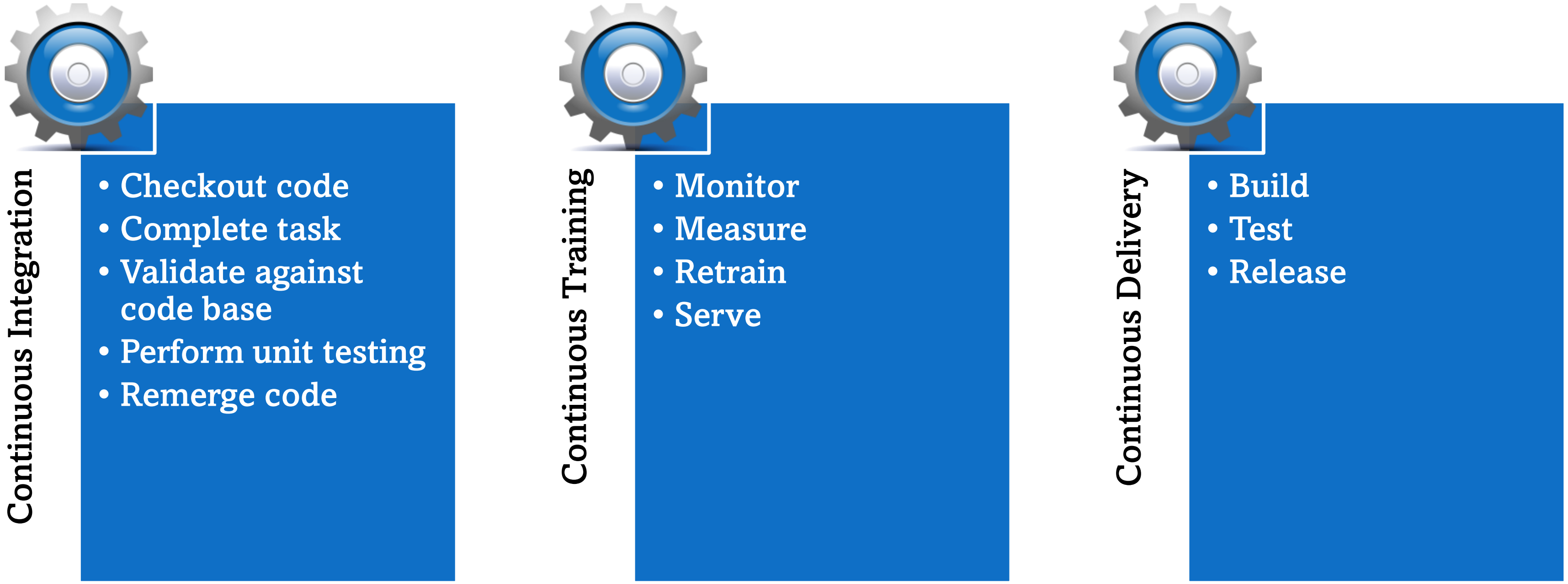
Traditional System Testing and Monitoring



ML-Based System Testing and Monitoring

ML-Ops – System Pipeline

- Major challenges in ML Models and SM Data
  - Tracability – What ideas have been tried? What are successful?
  - Reproducibility – How to reproduce successful ideas?
- ML Lifecyle
  - Manage resources, data, code, time, and quality to meet objectives.



Continuous Integration

- Checkout code
- Complete task
- Validate against code base
- Perform unit testing
- Remerge code

Continuous Training

- Monitor
- Measure
- Retrain
- Serve

Continuous Delivery

- Build
- Test
- Release

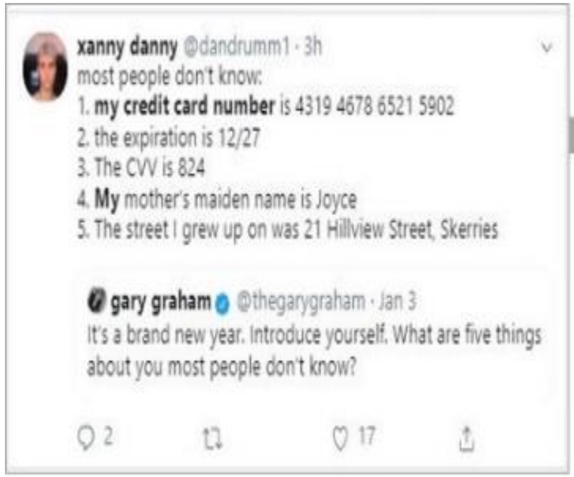
Predictive Modeling – Case Study

Predictive Modeling
Emotion Modeling

Identifiable Entities in Social Media Data

Entities	Description
ORG	referring to institutions, organizations, companies, agencies etc.,
GPE	denoting countries, states, and cities
PERSON	representing people names and fictional characters
DATE	symbolizing relative or absolute dates or time periods
TIME	pertaining to time periods that are shorter than a day
NORP	mapping to nationalities, political and religious groups or communities
LOC	signifying any non-GPE entities like water bodies, mountains, etc.,
PRODUCT	Characterizing things, objects, food, vehicles and other non-service entities
EVENT	for data indicating any disasters like battles, hurricanes, earthquakes, wars, sports and well-established happenings
PERCENT	implying any values represented in percentage format (%)

Text Data Preprocessing



xanny danny @dandrumm1 · 3h

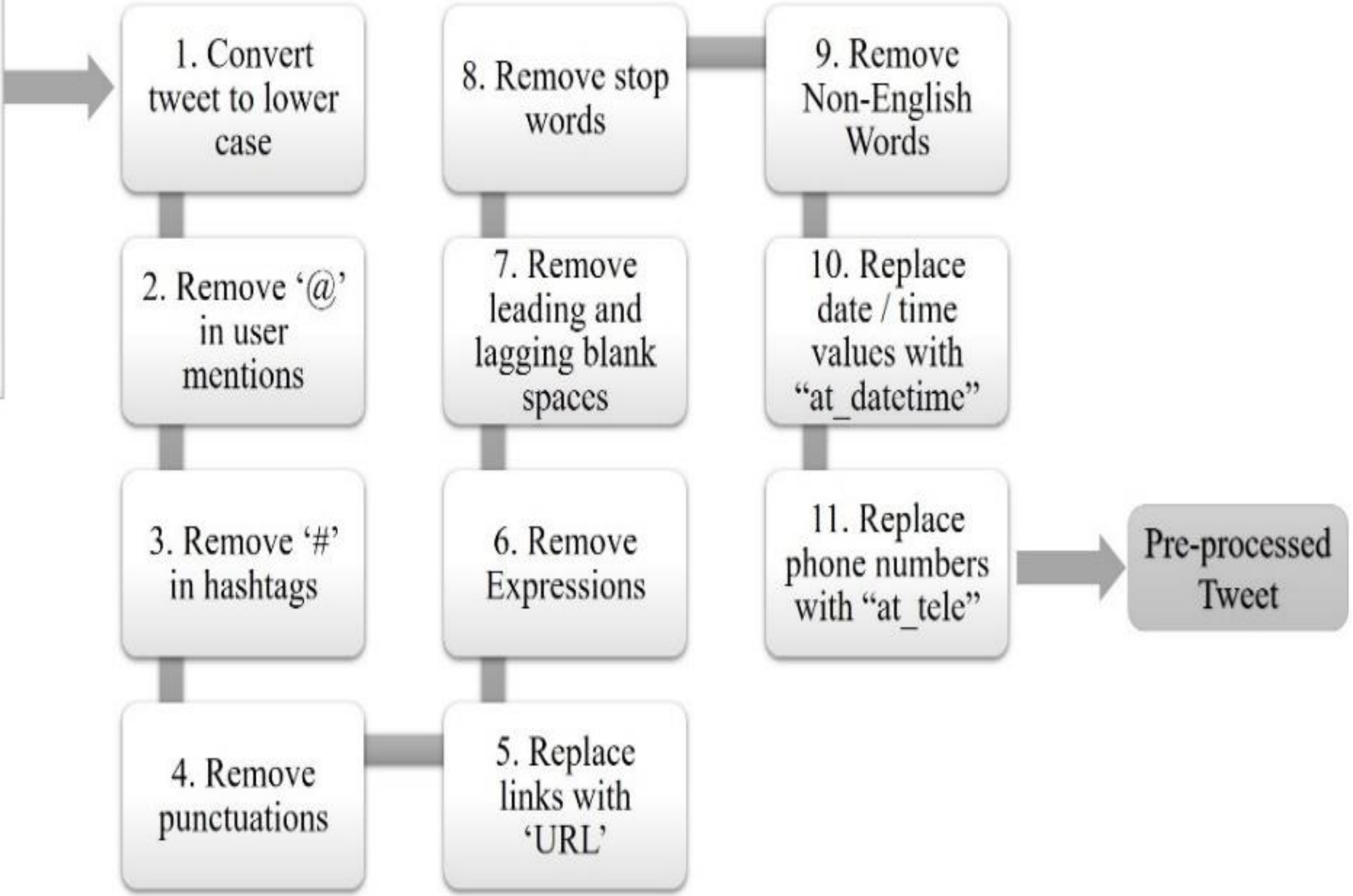
most people don't know:

1. **my credit card number** is 4319 4678 6521 5902
2. the expiration is 12/27
3. The CVV is 824
4. **My** mother's maiden name is Joyce
5. The street I grew up on was 21 Hillview Street, Skerries

gary graham @thegarygraham · Jan 3

It's a brand new year. Introduce yourself. What are five things about you most people don't know?

Raw Tweet



1. Convert tweet to lower case

2. Remove '@' in user mentions

3. Remove '#' in hashtags

4. Remove punctuations

5. Replace links with 'URL'

6. Remove Expressions

7. Remove leading and lagging blank spaces

8. Remove stop words

9. Remove Non-English Words

10. Replace date / time values with "at_datetime"

11. Replace phone numbers with "at_tele"

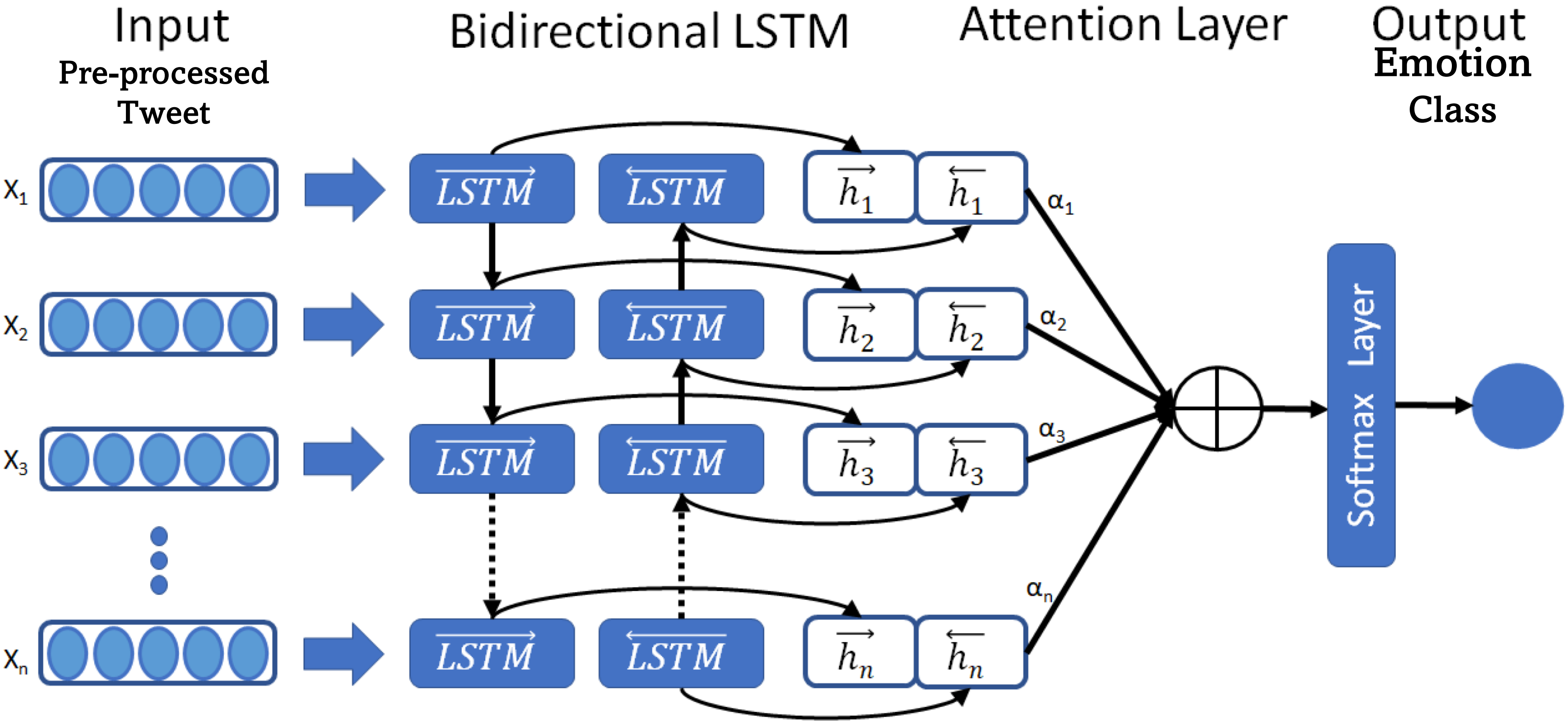
Pre-processed Tweet

Entity – Emotion Modeling

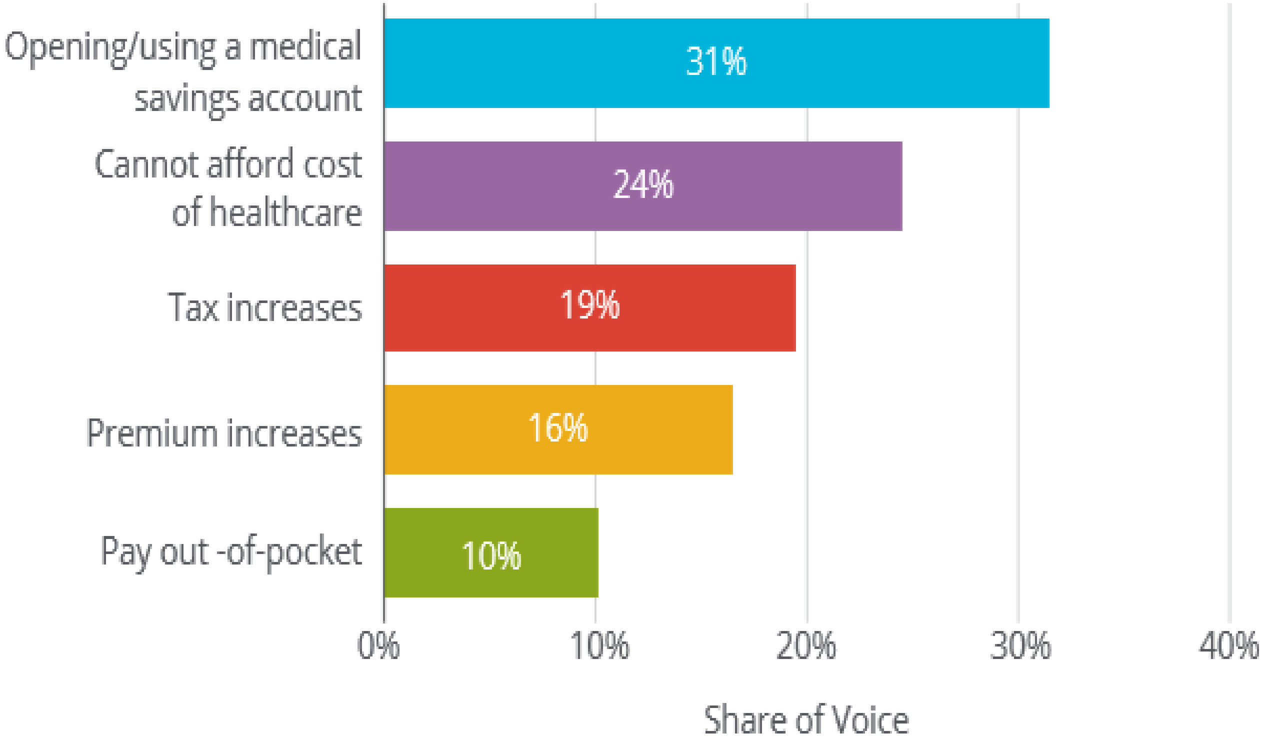
- Identification of entities through NER – Sequence labelling task
- Understanding the emotions through emotion mining mechanism
- Build features based on discovered entities and emotion
- Study ‘Correlations’ by joint probability between adjacent features
- Interdependency modelling by statistical analysis on univariate and multi-variate entity correlation
- Rank based correlation – Spearman’s correlation coefficient

$$\rho = 1 - \frac{6\sum_{i=1}^{n}(u_i - v_i)^2}{n(n^2-1)}$$

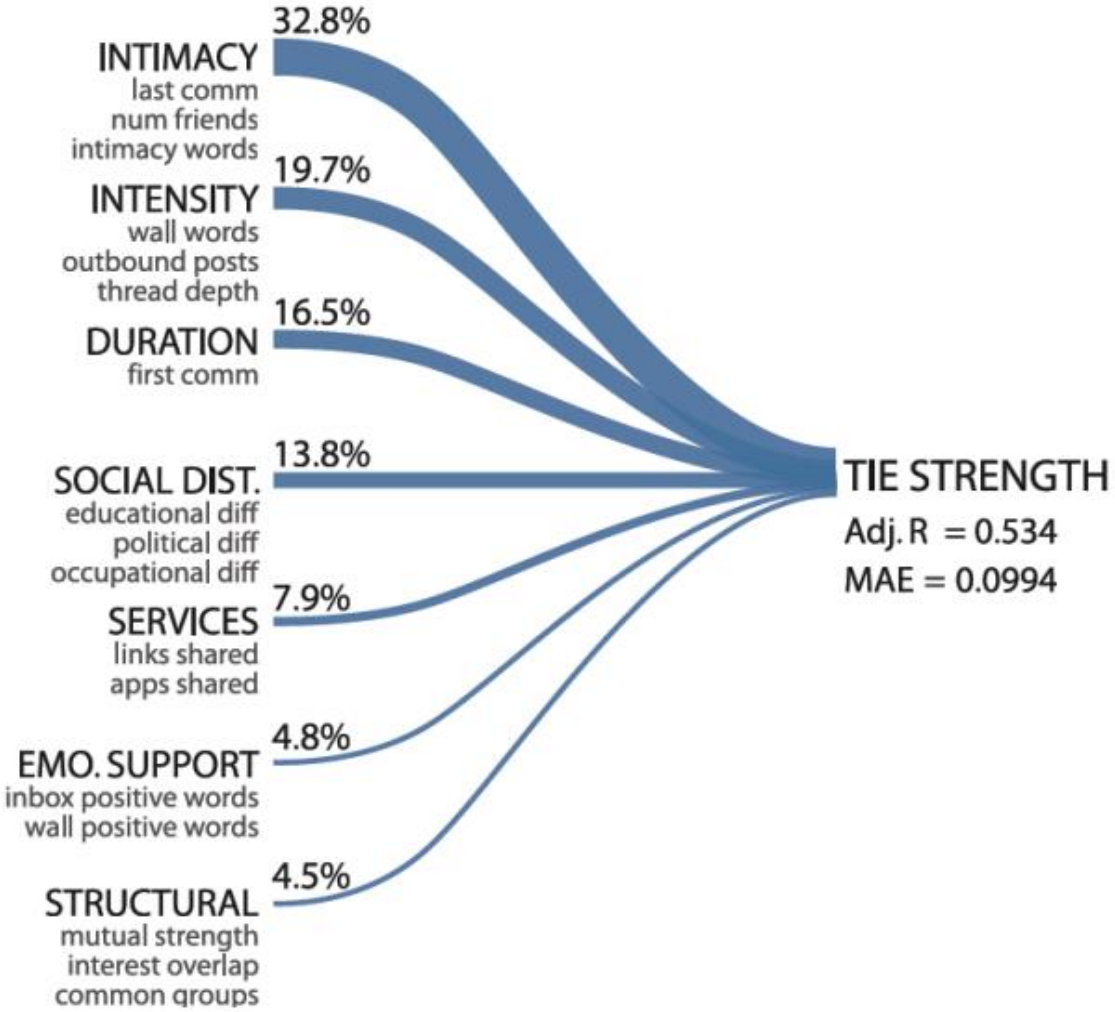
Predictions – Insights on Citizen opinions



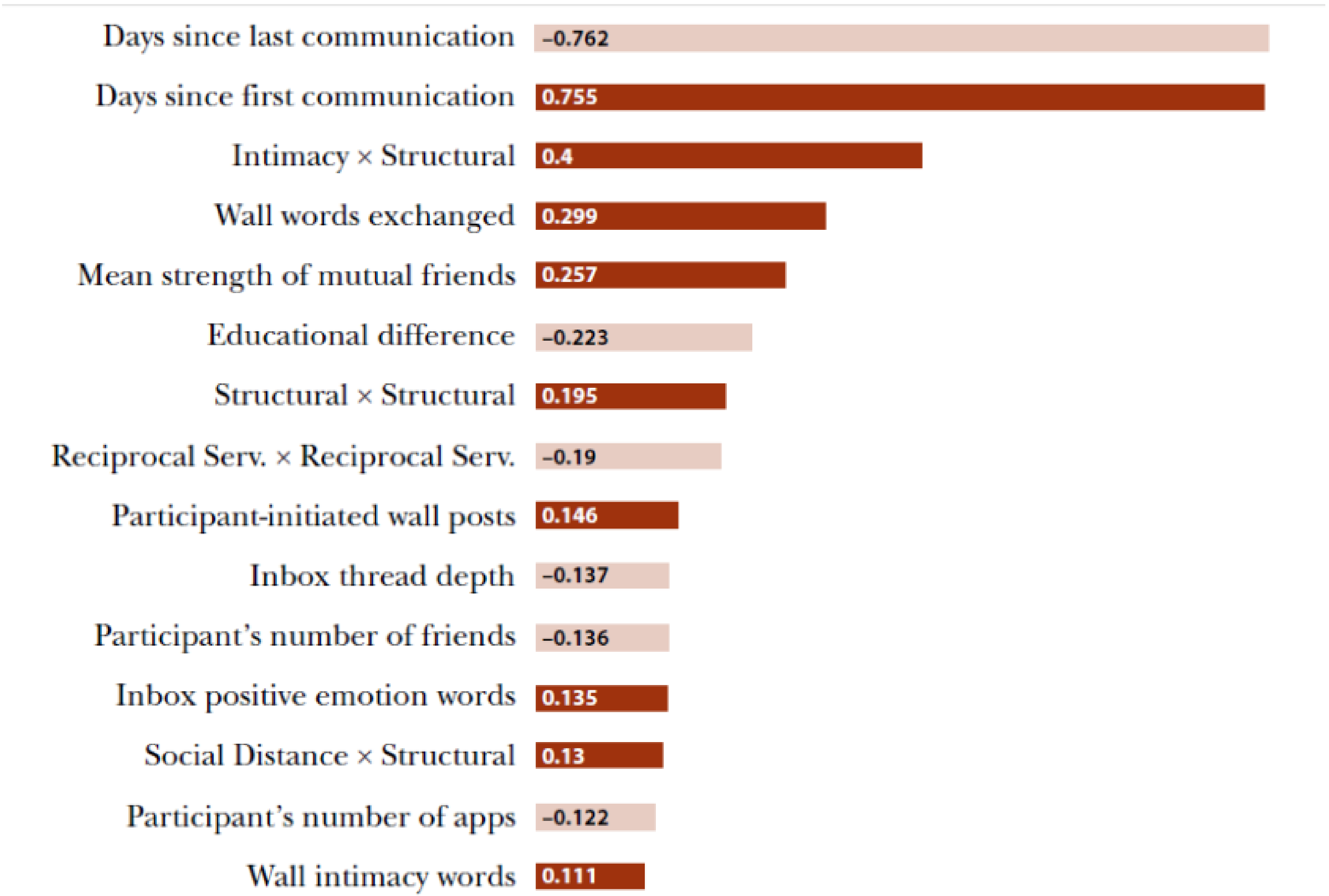
Predictions – Insights on Citizen opinions



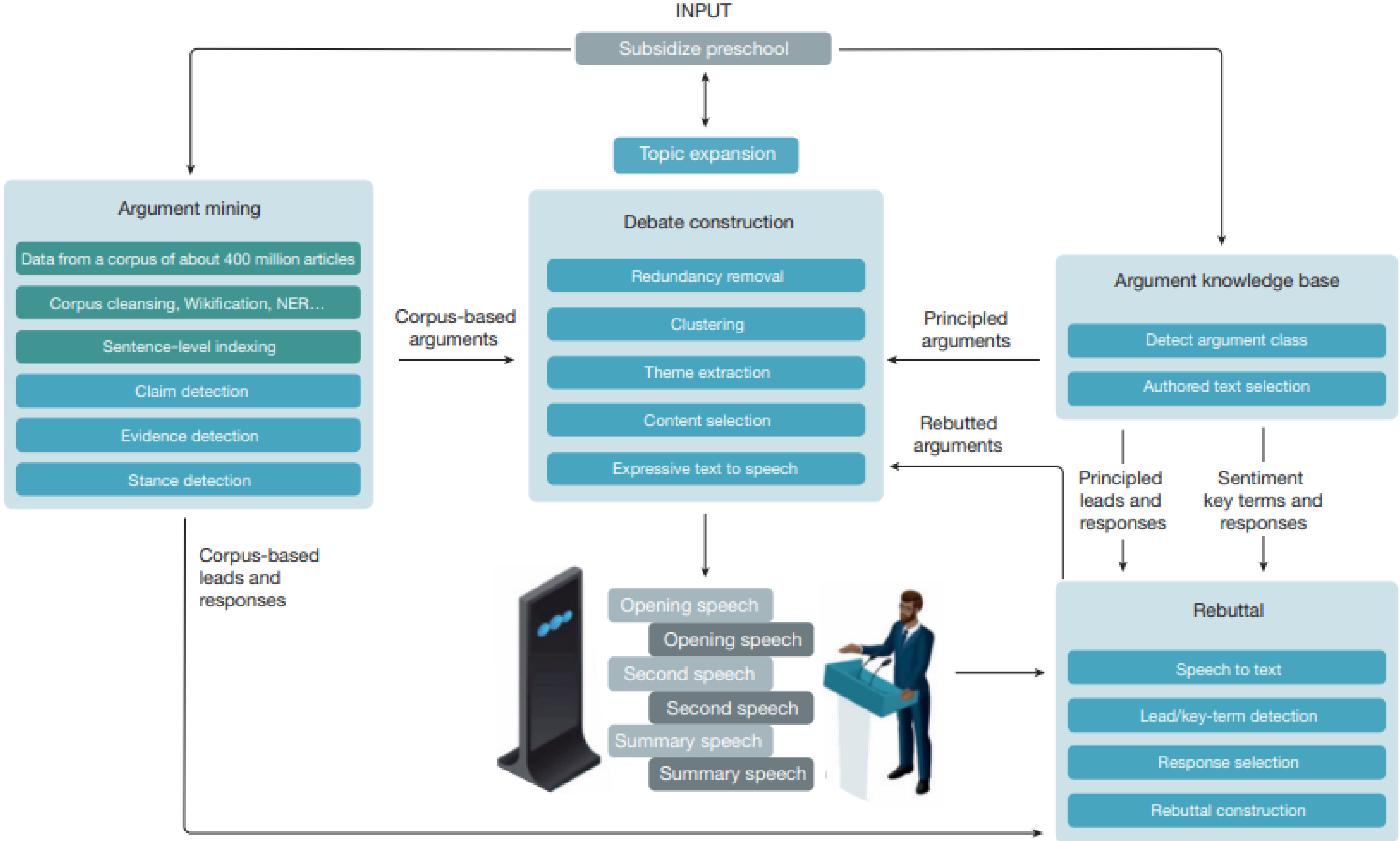
Linear Regression model
How strong is your relationship?



Predictable Network Features in Social Media



Audio – Debater - An automated Debating System



INPUT

Subsidize preschool

Topic expansion

Argument mining

- Data from a corpus of about 400 million articles
- Corpus cleansing, Wikification, NER…
- Sentence-level indexing
- Claim detection
- Evidence detection
- Stance detection

Corpus-based arguments

Debate construction

- Redundancy removal
- Clustering
- Theme extraction
- Content selection
- Expressive text to speech

Principled arguments

Rebutted arguments

Argument knowledge base

- Detect argument class
- Authored text selection

Principled leads and responses

Sentiment key terms and responses

Corpus-based leads and responses

Opening speech

Opening speech

Second speech

Second speech

Summary speech

Summary speech

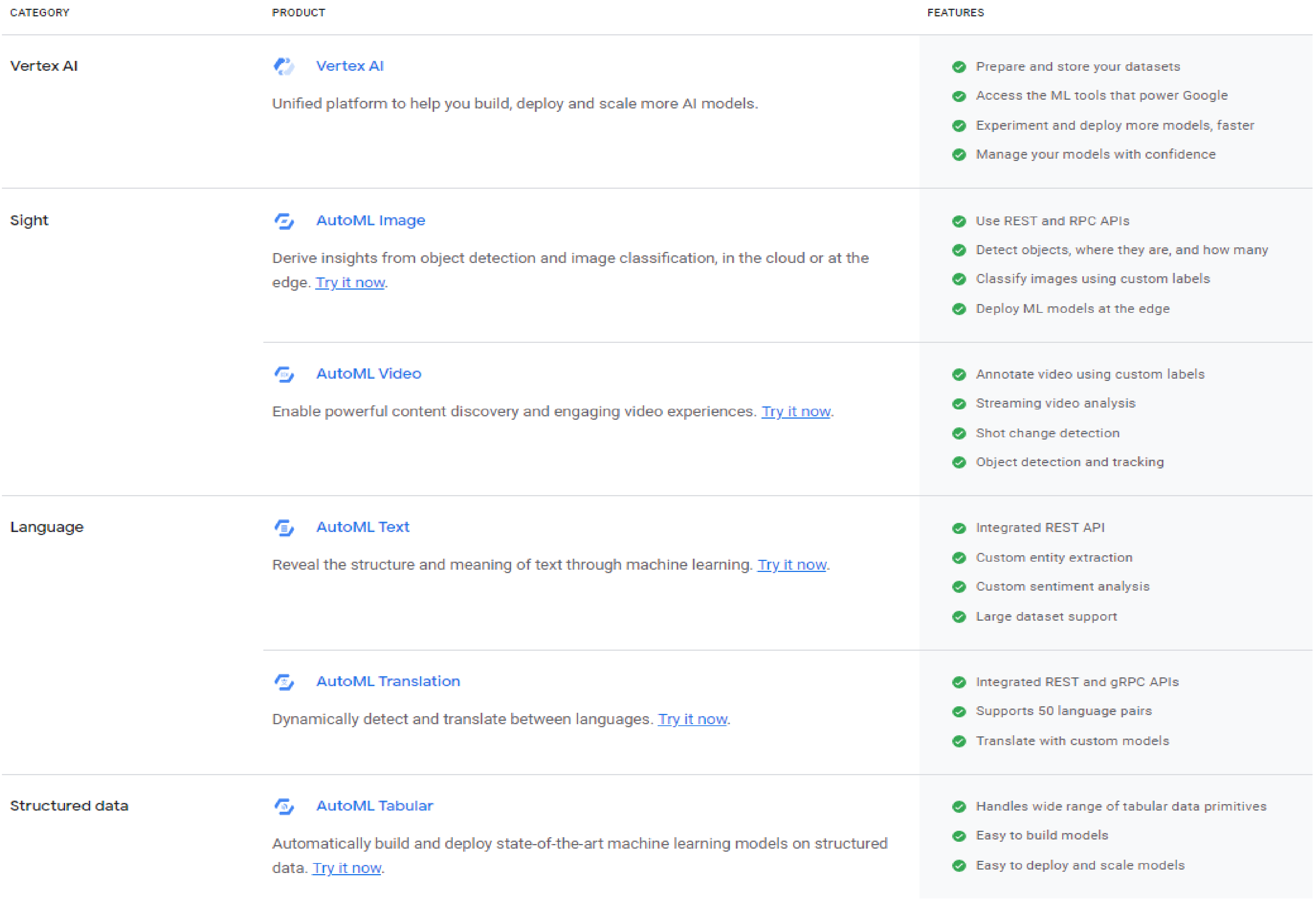
Rebuttal

- Speech to text
- Lead/key-term detection
- Response selection
- Rebuttal construction

Where are we now…?!

Google Cloud Platform

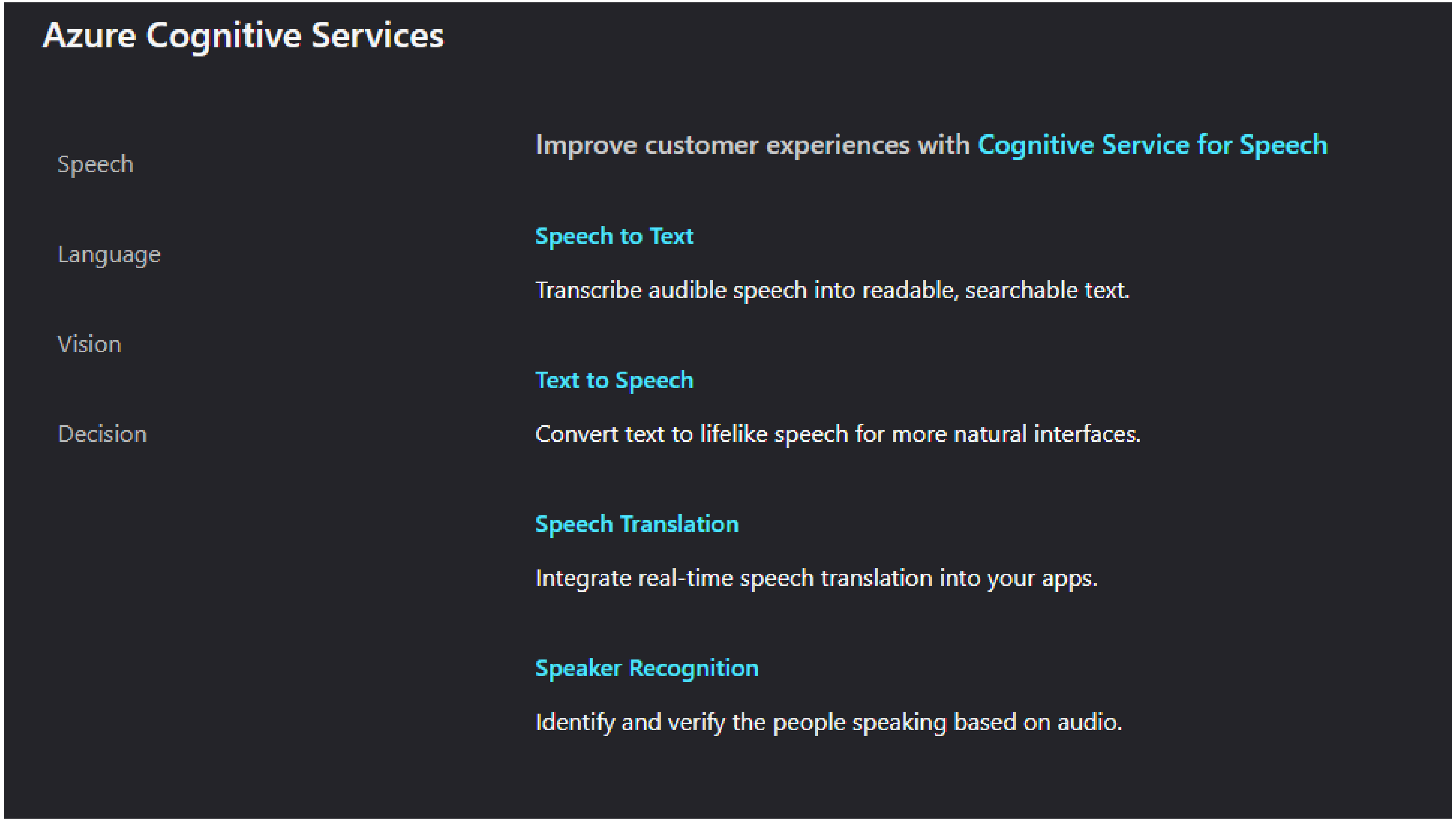
https://cloud.google.com/automl



CATEGORY	PRODUCT	FEATURES
Vertex AI	Vertex AI Unified platform to help you build, deploy and scale more AI models.	✓ Prepare and store your datasets ✓ Access the ML tools that power Google ✓ Experiment and deploy more models, faster ✓ Manage your models with confidence
Sight	AutoML Image Derive insights from object detection and image classification, in the cloud or at the edge. Try it now.	✓ Use REST and RPC APIs ✓ Detect objects, where they are, and how many ✓ Classify images using custom labels ✓ Deploy ML models at the edge
	AutoML Video Enable powerful content discovery and engaging video experiences. Try it now.	✓ Annotate video using custom labels ✓ Streaming video analysis ✓ Shot change detection ✓ Object detection and tracking
Language	AutoML Text Reveal the structure and meaning of text through machine learning. Try it now.	✓ Integrated REST API ✓ Custom entity extraction ✓ Custom sentiment analysis ✓ Large dataset support
	AutoML Translation Dynamically detect and translate between languages. Try it now.	✓ Integrated REST and gRPC APIs ✓ Supports 50 language pairs ✓ Translate with custom models
Structured data	AutoML Tabular Automatically build and deploy state-of-the-art machine learning models on structured data. Try it now.	✓ Handles wide range of tabular data primitives ✓ Easy to build models ✓ Easy to deploy and scale models

Microsoft Azure

https://azure.microsoft.com/en-in/services/cognitive-services/#overview



Azure Cognitive Services

Improve customer experiences with Cognitive Service for Speech

Speech to Text

Transcribe audible speech into readable, searchable text.

Text to Speech

Convert text to lifelike speech for more natural interfaces.

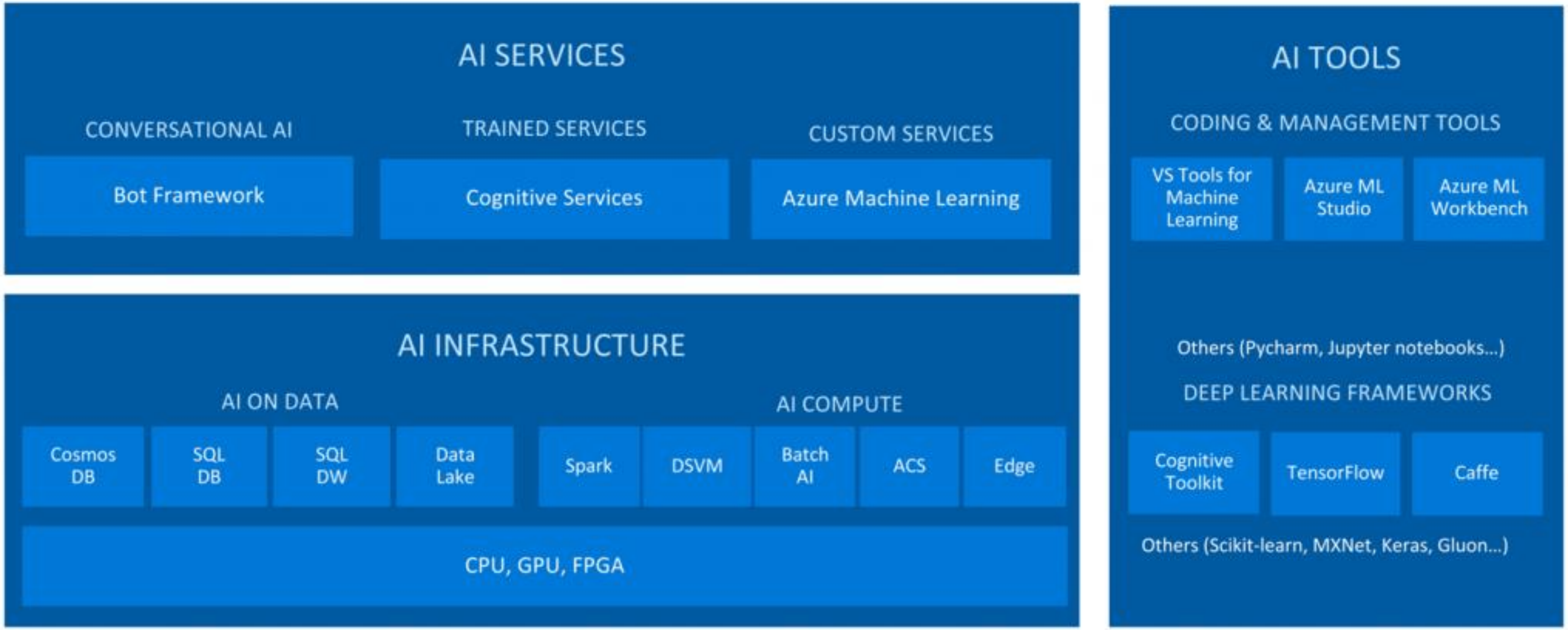
Speech Translation

Integrate real-time speech translation into your apps.

Speaker Recognition

Identify and verify the people speaking based on audio.

Microsoft Azure



AI SERVICES

CONVERSATIONAL AI

Bot Framework

TRAINED SERVICES

Cognitive Services

CUSTOM SERVICES

Azure Machine Learning

AI INFRASTRUCTURE

AI ON DATA

Cosmos DB

SQL DB

SQL DW

Data Lake

AI COMPUTE

Spark

DSVM

Batch AI

ACS

Edge

CPU, GPU, FPGA

AI TOOLS

CODING & MANAGEMENT TOOLS

VS Tools for Machine Learning

Azure ML Studio

Azure ML Workbench

Others (Pycharm, Jupyter notebooks...)

DEEP LEARNING FRAMEWORKS

Cognitive Toolkit

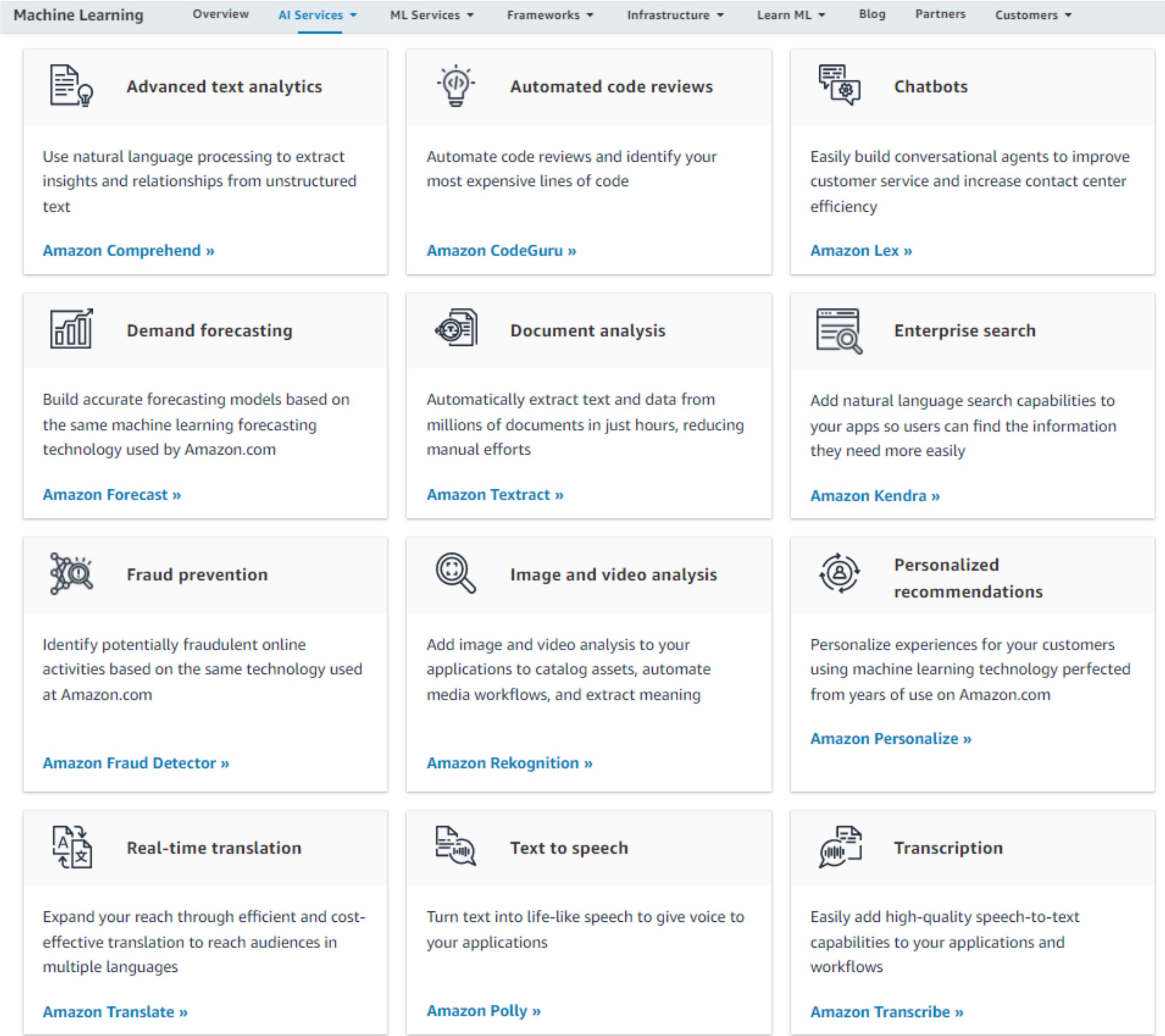
TensorFlow

Caffe

Others (Scikit-learn, MXNet, Keras, Gluon...)

Amazon Web Services – AI / ML Services

https://aws.amazon.com/machine-learning/



Advanced text analytics

Use natural language processing to extract insights and relationships from unstructured text

Amazon Comprehend »

Automated code reviews

Automate code reviews and identify your most expensive lines of code

Amazon CodeGuru »

Chatbots

Easily build conversational agents to improve customer service and increase contact center efficiency

Amazon Lex »

Demand forecasting

Build accurate forecasting models based on the same machine learning forecasting technology used by Amazon.com

Amazon Forecast »

Document analysis

Automatically extract text and data from millions of documents in just hours, reducing manual efforts

Amazon Textract »

Enterprise search

Add natural language search capabilities to your apps so users can find the information they need more easily

Amazon Kendra »

Fraud prevention

Identify potentially fraudulent online activities based on the same technology used at Amazon.com

Amazon Fraud Detector »

Image and video analysis

Add image and video analysis to your applications to catalog assets, automate media workflows, and extract meaning

Amazon Rekognition »

Personalized recommendations

Personalize experiences for your customers using machine learning technology perfected from years of use on Amazon.com

Amazon Personalize »

Real-time translation

Expand your reach through efficient and cost-effective translation to reach audiences in multiple languages

Amazon Translate »

Text to speech

Turn text into life-like speech to give voice to your applications

Amazon Polly »

Transcription

Easily add high-quality speech-to-text capabilities to your applications and workflows

Amazon Transcribe »

And now, build 'problem statements' from what you know / want…!

Dr. Geetha Raju | TNeGA