TARENTO

Data Science and Machine Learning - Capabilities

Tarento offerings

Agenda

1

AI Capabilities

- Natural Language Processing
- Computer vision and Image processing
  - Handwriting recognition
  - Scanned document text extraction
  - Face and object recognition

2

Data Architecture & Pipeline

3

Project Profiles

- Anuvaad - Judicial document translation
- Mark sheet - Handwriting recognition
- Judicial document text extraction
- Procter-less exam monitoring
- UIDAI chatbot

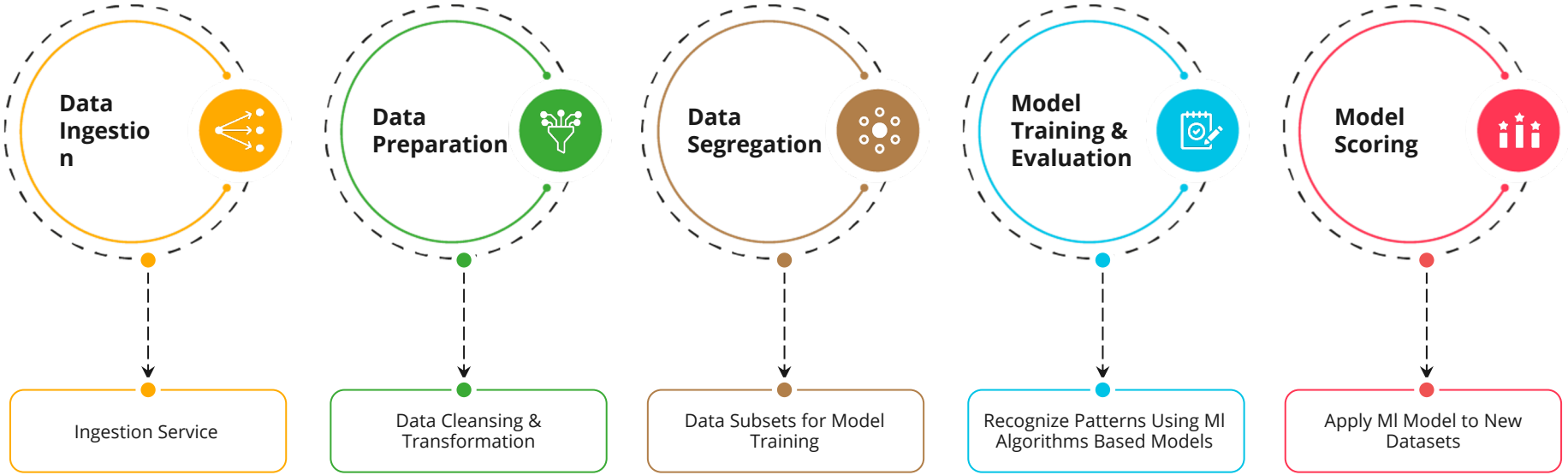
Agenda

4 Use Cases

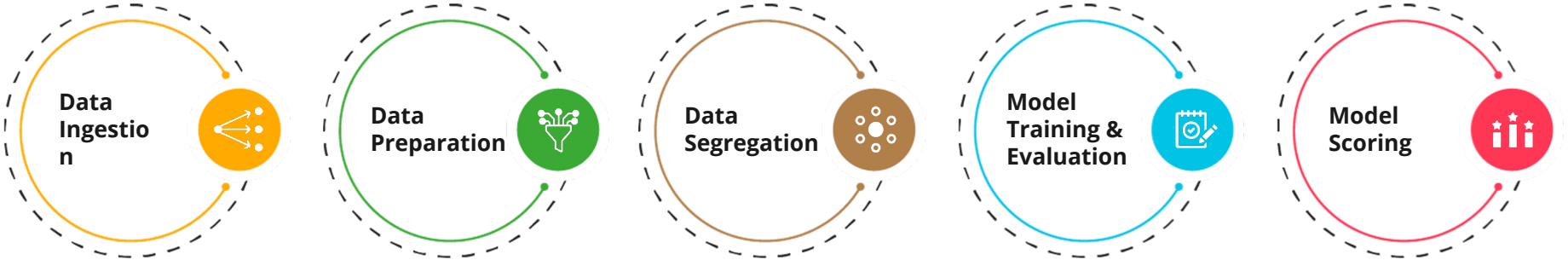
5 References

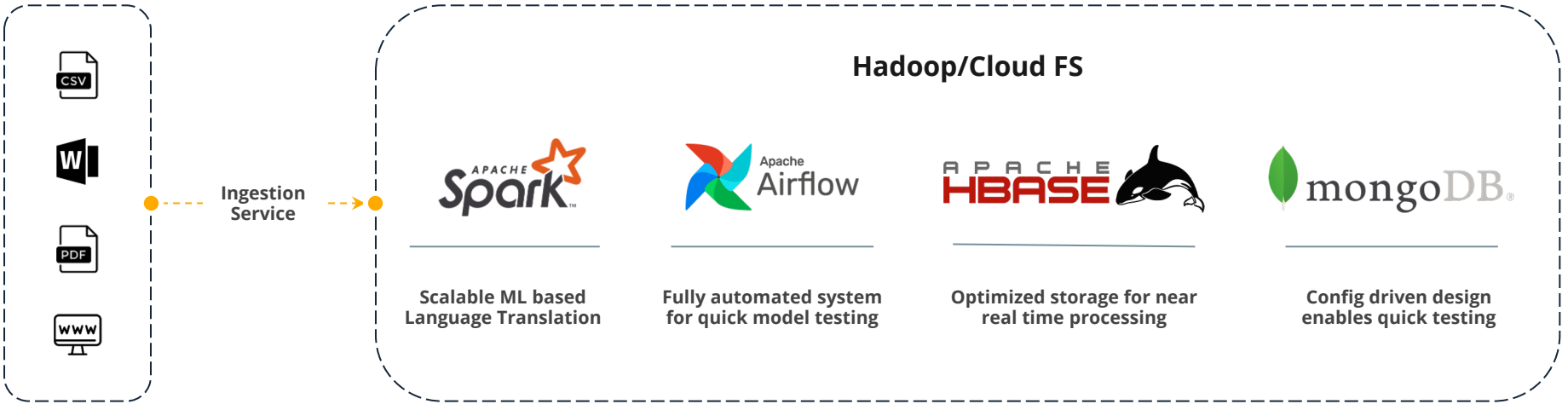
Data Architecture & Pipeline

Data Architecture & Pipeline

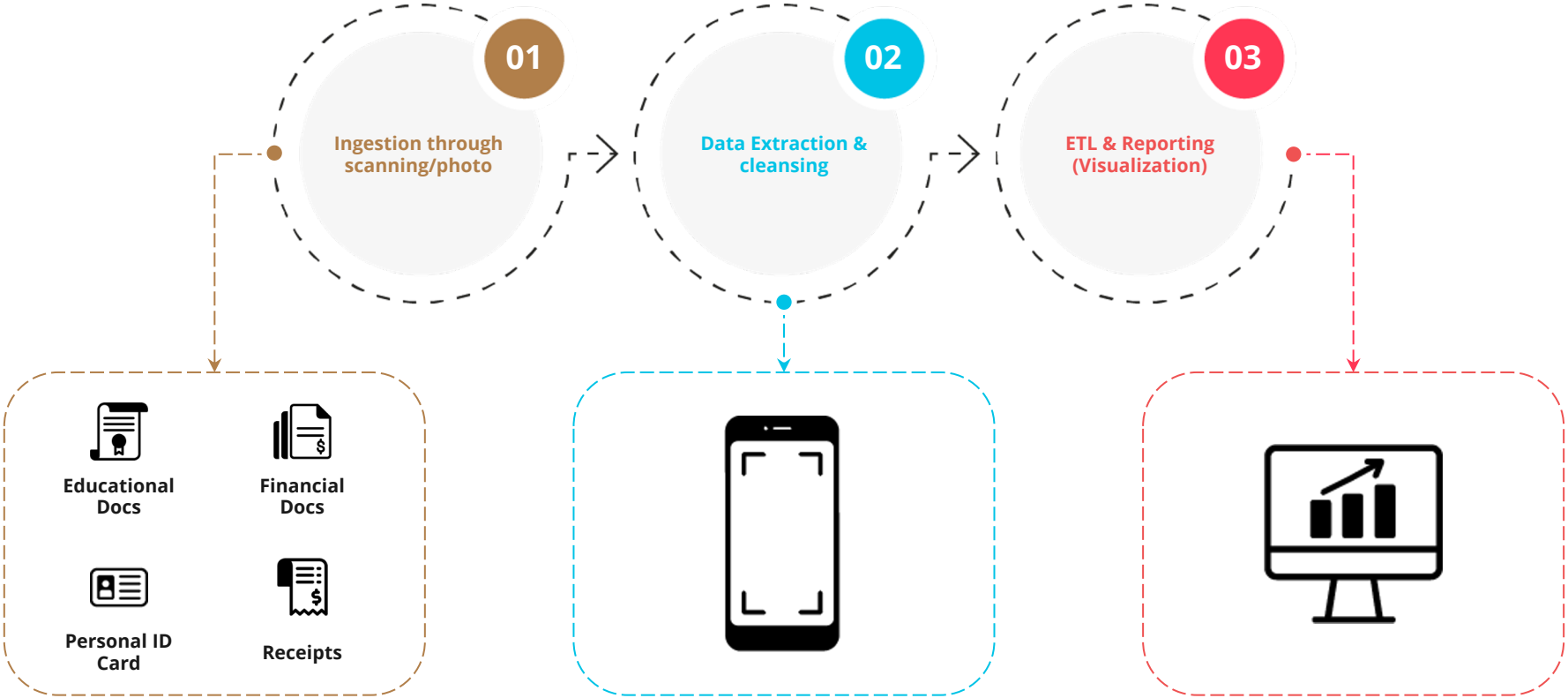


Big Data Pipeline - Anuvad





Big Data Pipeline - OCR



Capabilities - Data Architecture

Handling large volume of unstructured data

- Identify various sources like Email, Web clicks, sensor data, webpages, PDFs, docx, xlsx etc.
- OCR'd pages or images
- Existing legacy systems

Solution Stack

- **Big Data Pipeline**
  - Open sourced solutions
  - Cloud based infrastructure
- **Databases**
  - NoSQL
  - RDBMS
- **Automation**
  - MLOps

Capabilities - Data Integration

Data governance and integration

- **For downstream activities**
  - Storage
  - Analysis
  - Retrieval
- **Data transformation for other application**
- **Real time access**

Solution Stack

- **Big Data Integration**
  - Open sourced solutions (Hadoop, PIG, HBase, HIVE etc)
  - Cloud based infrastructure (Azure, AWS etc)
- **Databases**
  - NoSQL
  - RDBMS
- **Data migration**
- **Queuing system for real time access**

Capabilities - Data Processing

Processing the structured data using deep learning

- **Natural language processing**
  - Language modeling
  - Sequence prediction
  - Text classification
  - Data labeling
  - Summarization
  - Text translation
- **Image processing**
  - Handwriting recognition
  - Scanned document text extraction
  - Natural scene text detection
  - Object detection

Solution Stack

- **Deep learning backend system**
  - PyTorch
  - TensorFlow
  - OpenCV
  - Tesseract 4
- **Language Stack**
  - Python
  - JavaScript
- **Trained model integration**
  - REST'fing

Capabilities - Data Processing

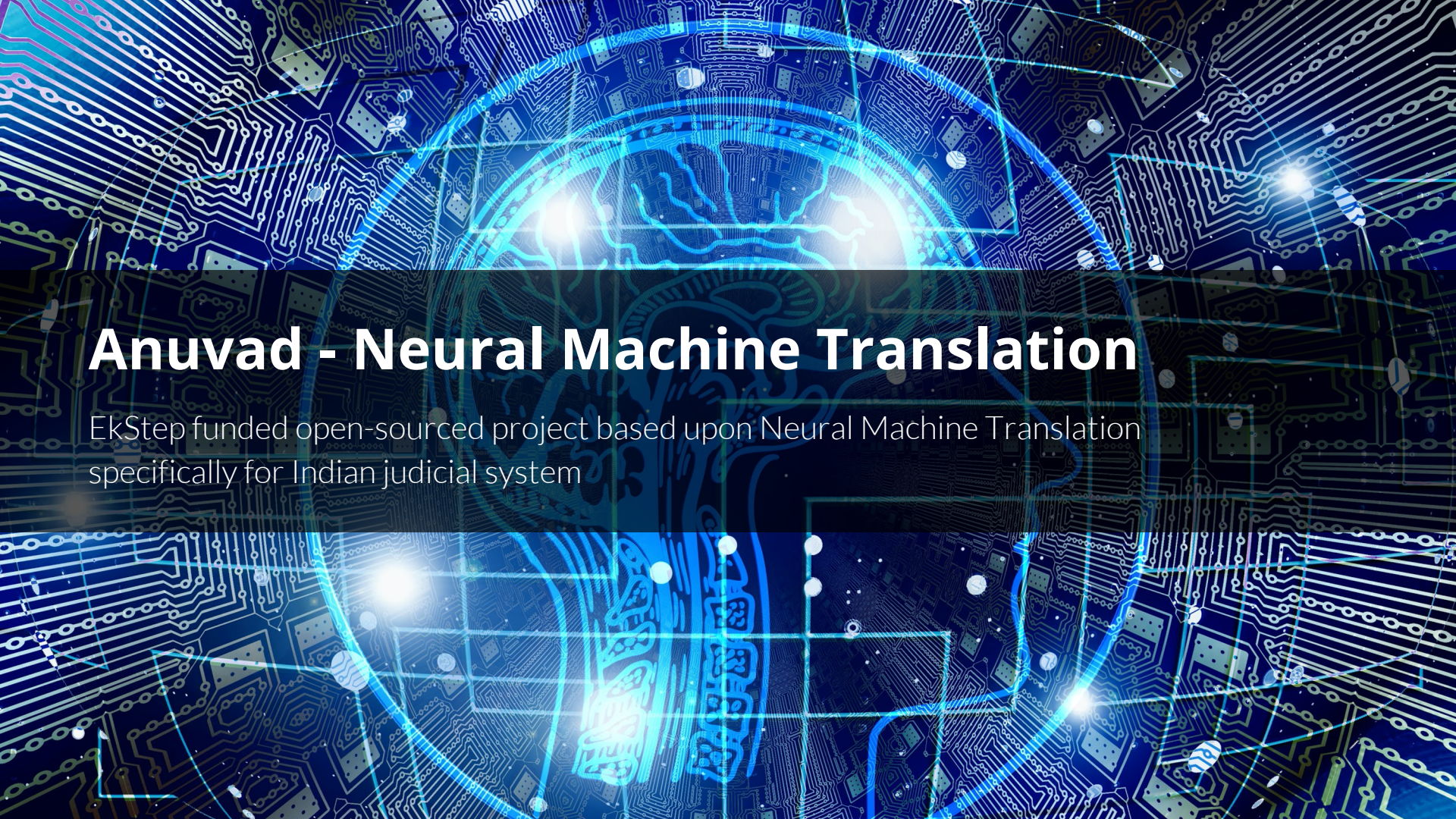
Actionable insights and representation

- **Dashboard**
- **Reporting**
- **Workflow generation and management**

Solution Stack

- **Custom visualization, open-source based**
  - ReactJS, ReactNativeJS
  - Chart library
- **Microservices based approach for backend integration**
- **Popular toolchains**
  - Tableau

Project Profiles



Anuvad - Neural Machine Translation

EkStep funded open-sourced project based upon Neural Machine Translation specifically for Indian judicial system

Project Highlights

- Anuvaad GITHUB fork is adopted by the indian judicial system
- System currently support 9 vernacular indian language, plan is to support all 22 scheduled languages.

Data Architecture

- Judgement data present in TXTs, PDFs, DOCX, HTMLs are the source of parallel corpus.
- Data pipeline is setup to ingest huge amount of parallel corpus as needed by the NMT modeling backend

Data Integration

- Various parsers to transform the received corpus data
- DOCX parser, language modeling, sequence prediction in target language
- Automated training steps
- CI/CD for model and code deployment

Visualization

- Custom build application based upon ReactJS
- Statistics dashboard
- Translation quality matrix
- Mobile application for various stakeholders

Scanned Judgement Digitization

EkStep funded, deep learning based OCR capability for indian languages

Project Highlights

- Support for wide range of fonts for all 22 scheduled languages.
- We are currently running evaluation phase for 4 indic languages in 40 font-family

Data Architecture

- Scanned document, that's a unstructured information enters the pipeline
- We are estimating millions of OCR'd pages for entire judicial system

Data Integration

- The parser and processor are based upon models and software present in public domain that have modified to suit the obtained dataset
- Training and model pipeline are automated to evaluate current phase.

We are benchmarking OCR result against Google's vision.

Handwriting Detection, Extraction and Recognition

EkStep funded, deep learning based OCR and OMR capability

Project Highlights

- **Detect and compare given face image against uploaded video or live video stream.**
- **Detect various facial motions like**
    - Facial recognition
    - Absence of person in the frame
    - Person looking to left, right or up
    - Reading (Lips moving)
    - Multiple persons in frame
    - Any objects like phone, headset etc. along with the face

Data Architecture

- **Indian face data.**
- **Segregating frame and labeling image with phone, headset for detection.**
- **Data is ingested and joins the pipeline for downstream activity**
- **REST APIs are available for integration**

Aadhaar Chatbot

A Chatbot platform to augment and enhance UIDAI's customer support and public relations

UIDAI is world's largest biometric data keeper and hence receives high user traffic on chatbot system.

Statistics:

- **Average 50,000 user queries per day**
- **System has processed 2.5 million unique citizens session**

Technology Stack

- Based upon open-source project RASA
  - Language handling
  - Intent classification
  - NER extraction

Data Architecture

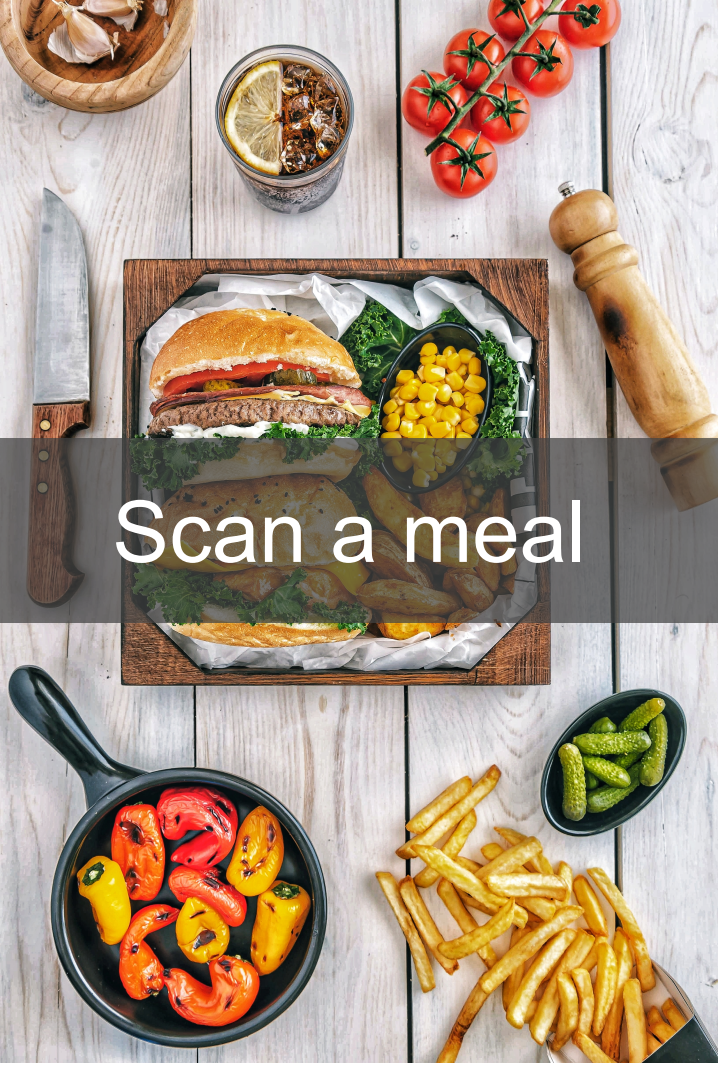
- Existing emails exchanged between citizen and support team is used to derive intents and various entities. Topic modeling methodology used to augment the data architecture team

Data Integration

- Collected user queries goes through a internally developed tool for further classification and training

Automated training pipeline

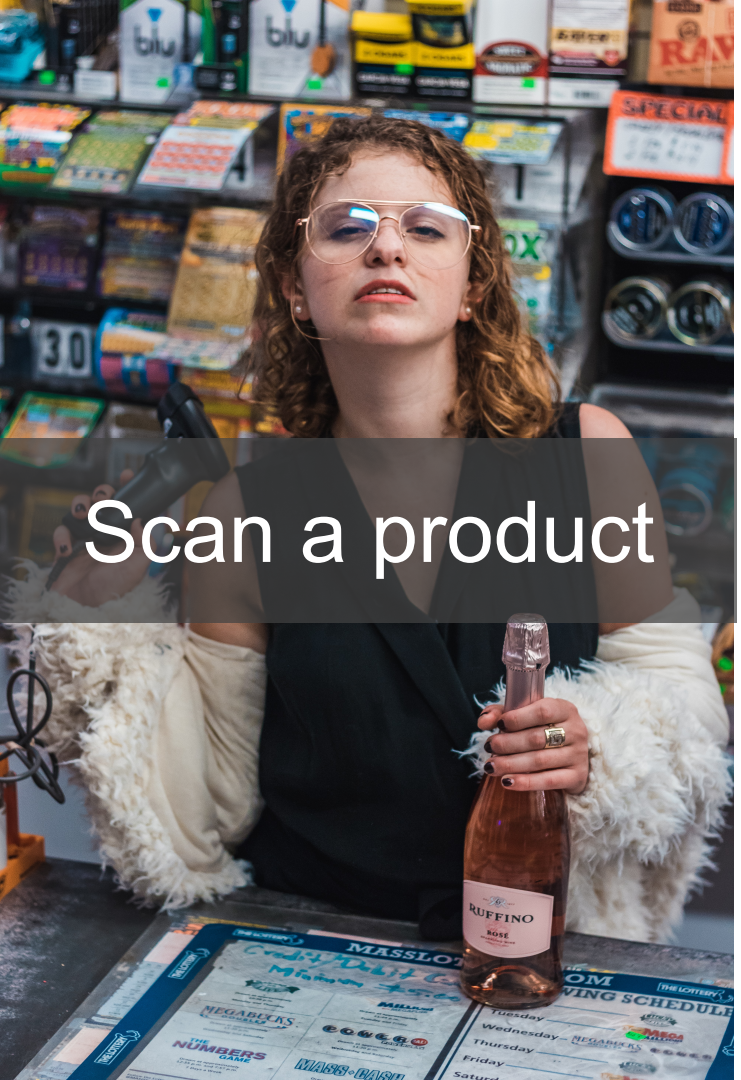
Use Cases



- **Use the ICA App and your smartphone to scan a dish or upload an image from the web**
- **Get a shortlist of similar recipes from ICAs recipe bank.**
- **Get nutritional facts and ingredients.**
- **Save the recipe to your favourites**
- **Send the ingredients to a purchase list or to your e-commerce basket**

Our approach

- Augment publicly available food dataset with ICA's food. APIs to get nutritional facts about the food
- Use existing data pipeline for downstream activities.
- Baseline can be generated within 6 - 8 weeks for effort



- **Use the ICA App and a smartphone to scan a product to get nutritional facts and other product related content.**

- **Customer uses the ICA App and a smartphone to scan an isle to find a product. Can also be used by personnel picking e-commerce orders in store.**

Our approach

- Curate product item dataset and its nutritional facts. Information can be scrapped from available product catalogue or from internet.
- Baseline can be generated within 6 - 8 weeks for effort

Footfall analysis

- **Capture <u>anonymous</u> customer data via face recognition to predict:**
  - Age
  - Gender
  - Mood
  - Returning customer to the store

- **Analyse in-store customer behaviour**
  - Items put in basket
  - Time spent looking for items
  - Time spent per aisle
  - Movement patterns

Our approach

- Detect face and associated features to derive various demographic information
- Look for historical information about user purchase or purchase in that demography
- Create offers on fly
- Baseline in 6 – 8 weeks

Just walk out

- **Assisted shopping as Amazon Go in Seattle.**
- **Customer is identified by scanning a barcode with the ICA app when entering the store.**
- **Cameras scan what customers put in their baskets.**
- **Check-out is fully automated.**

Our approach

- Curate product item dataset. Information can be scrapped from available product catalogue or from internet.
- Baseline can be generated within 6 - 8 weeks for effort.

Translation

- **Automated translation of recipes to multiple languages**
- **Automated translation of ICA.se and ICA's e-commerce sites to multiple languages**
- **Automated translation of product manuals etc**

Our approach

- Create and collect parallel sentences in the respective language of interest.
- Use existing big data pipeline to achieve all downstream activities like training, regression, deployment.
- Baseline can be generated within 6 - 8 weeks for other interface to consume the predictions

Recipe to basket

- **Ingredient matching based on actual behaviour of customers.**
- **Optimize shopping basket based on total purchase and customer preferences.**
- **Suggest recipes based on what the customer already has put in the basket.**
- **Upselling & recommendation engine based on customer preferences**

Our approach

- Curate product item dataset. Information can be scrapped from available product catalogue or from internet.
- Baseline can be generated within 6 - 8 weeks for effort.

Master data mgmt

- **Improve efficiency of master data management by accurately tagging of products and their attributes**
- **Improve SEO and searchability on site with better product tagging**
- **Associated content can be created based on detected tags**

Our approach

- Curate product item dataset. Information can be scrapped from available product catalogue or from internet.
- Run NLP activities on existing product information
- Baseline can be generated within 8 - 12 weeks for effort.

References

Production & Pilot

http://developers.anuvaad.org

- Anuvaad means Translation.
- Big data pipeline
- https://github.com/project-anuvaad
- Please ask for username and password

Facial detection APIs

- <endpoint>/api/v1/face/verify
- End configuration on request

OCR Android application for handwriting detection and OMR

- Please look at the video demo
- APK on request

Questions?

Kumar Deepak
+91 9880216150
kumar.deepak@tarento.com