ORACLE

AI Productization: Models and Beyond

Dr. Ying Xu

Principal Member Technical Staff

Oracle Digital Assistant, OCI Language, Oracle

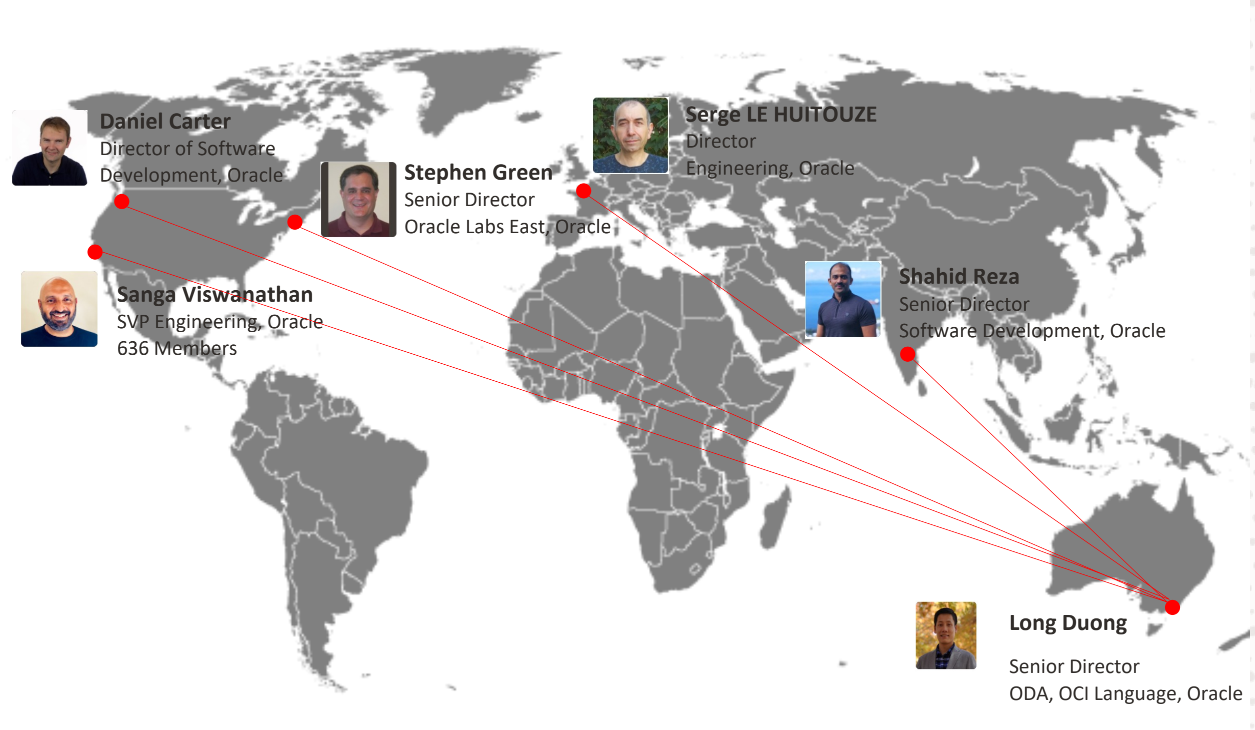
May 23, 2022

Our Team

Long Duong
Senior Director
ODA, OCI Language, Oracle

Mark Johnson
Chief AI Scientist
ODA, OCI Language, Oracle

Ahmed Abobakr – AI Engineer
Bhagya Hettige – AI Engineer
budhaditya Saha – AI Scientist
Chang Xu – AI Engineer
Cong Duy Vu Hoang – Consulting Member of Techni...
Dalu Guo – AI Engineer
Duy Vu – Data Scientist - Conversation...
Fahimeh Sadat Saleh – AI Engineer
Gioacchino Tangari – AI Engineer
Mohammad Najafi – AI Engineer
Nitika Mathur – AI Scientist
Omid Mohamad Nezami – AI Engineer
Pallavi D – Non Billable Contractor
Paria Jamshid Lou – AI Engineer
Pavan Jahagirdar – Non Billable Contractor
Philip Arthur – AI Scientist
Poorya Zaremoodi – Computer Scientist - Artificial...
Sagar Vetkar – Non Billable Contractor
Shivashankar Subramanian – AI Scientist
Shorya Shrivastava – Non Billable Contractor
Steve Siu – AI Engineer
Syed Najam Abbas Zaidi – AI Engineer
Thanh Vu – Computer Scientist - Artificial...
Thomas Pham – Senior Member Technical Staff
UMANGA BISTA – AI Engineer
Vlad Blinov – Data Scientist
Peter Zhong – Data Scientist
Snow Situ – AI Engineer
Don Dharmasiri – Computer Scientist - Artificial...
Ying Xu – Computer Scientist
Davis Hong – Senior Member Technical Staff

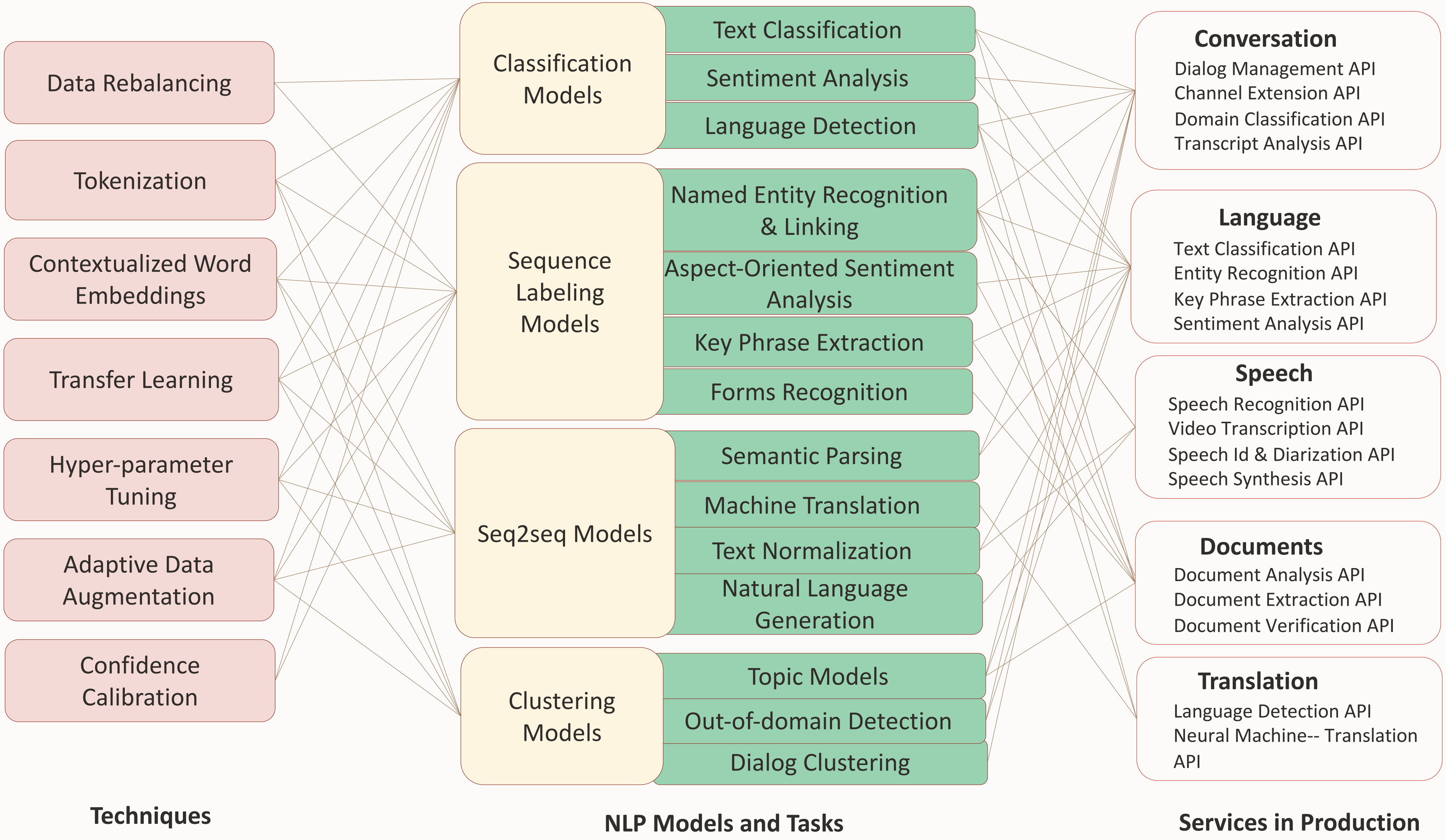


What is AI Productization?

What is AI Productization?

- Label Smoothing
- Contextualized Word Embeddings
- Data Rebalancing
- Knowledge Distillation
- Regularization
- Multitask Learning
- Prompt based Generation
- Transfer Learning
- Adaptive Data Augmentation
- Confidence Calibration
- Hyper-parameter Tuning
- Learning Rate Decay
- Loss Function Design
- Zero/few Short Learning
- Quantization
- Dropout
- Temperature Scaling
- Transformers
- Tokenization
- …

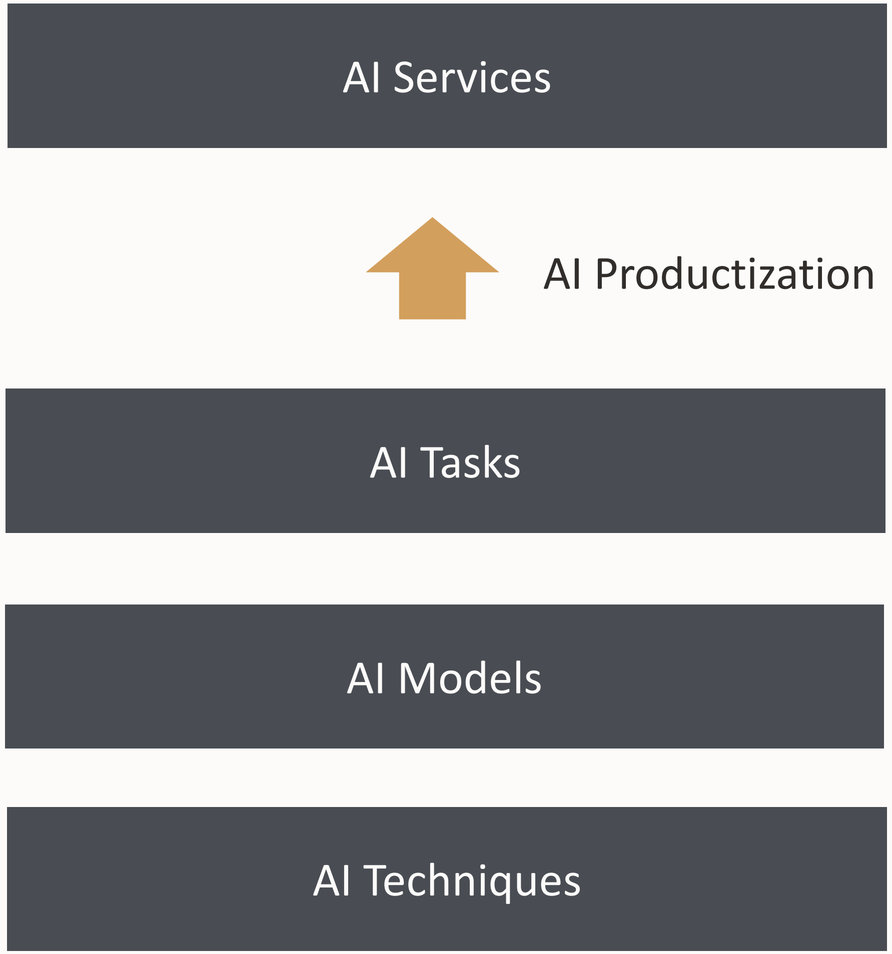
What is AI Productization?



What is AI Productization?

DEMO

What is AI Productization?



Overview

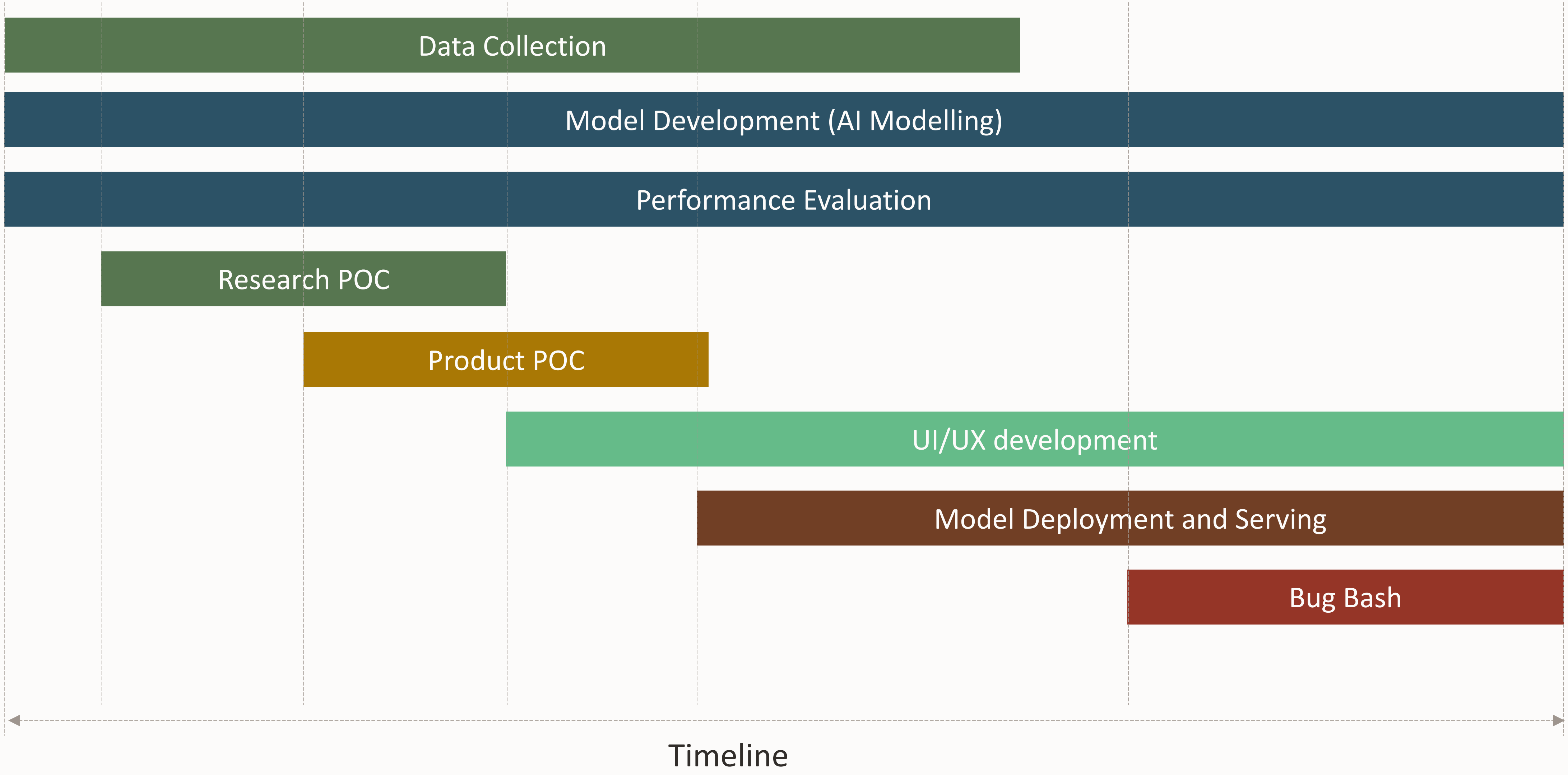
What is AI Productization?

- NLP Techniques
- Models and Tasks
- Services in Production

Beyond AI Modellings

- Running an AI project
- Discussion: Data Collection
- Discussion: Deterministic and Stable Training
- Discussion: Performance - Latency Trade-off

Beyond AI Modellings: Running an AI Project



Beyond AI Modellings: Data Collection

Public Data | Customer Data | From 3rd Party

Data Collection

Model Development (AI Modelling)

Public Data

- License checking → Legal Approve
- This determines whether the dataset can be used for different purposes, e.g. benchmarking/evaluation, hyperparameter tuning, production, etc.

Customer Data

- Highly confidential
- It may involve PII/PHI application → PII /PHI removal

Beyond AI Modellings: Data Collection

Data Collection From 3rd Party

- Data Collection Guidelines
  - Definition of the Task; Example Annotations; Borderline Use Cases; Annotation Rules;
  - Evaluation Metrics; Estimated Target Performance

- Quality Control
  - Pilot Run and Production Batches
  - Annotation Auditing
  - Data Quality Evaluation
  - Feedback List Curation
  - Annotator Performance Evaluation

Beyond AI Modellings: Running an AI Project

Model Development (AI Modelling)		
Literature Review	Model Research and Development	Hyper-parameter Tuning

For **Literature Review**, we collect recent research papers that are relevant to the target task, compile them into a list, share with the team, and call for discussion.

For **Model Research and Development**, we aim at

- recover the performance reported in the paper; then
- customize the model for our specific tasks; and
- if the model achieves promising performance, we will incorporate it into our internal ML/DL framework: OPALS

Beyond AI Modellings: Running an AI Project

Model Development (AI Modelling)		
Literature Review	Model Research and Development	Hyper-parameter Tuning

Hyper-parameter tuning is used to select the pretrained contextualized embeddings and plausible hyper-parameter combinations.

- In real application, hyperparameters could be tuned based on large number of evaluation sets, e.g. thousands of evaluation sets. → Multi-objective optimization

- Hyperparameter tuning with customization:
  - Imposes training and runtime resource constraints, e.g. time, memory, etc
  - Prefers hyper-parameters that provide training stability
  - Ignores unstable data points

- Auto-ML style functionality automatically suggests hyperparameters according to data shapes.

Beyond AI Modellings: Running an AI Project

Model Development (AI Modelling)

Performance Evaluation

Competitor Benchmarking	Customer Regression	Evaluation Tool

Competitor Profiling

- Competitor Profiling involves service feature and function comparison.
- For example, with text classification, some competitors may provide Out-of-Domain Detection, some may provide multi-label classification, some may provide support for more languages, etc.

Competitor Benchmarking

- For AI models, we care about
  - Model performance based on metrics such as accuracy, F1, AUC, etc. depending on the task
  - Training/Inference latency
  - How easy it is for developers/users to use

Beyond AI Modellings: Running an AI Project

Model Development (AI Modelling)		
Performance Evaluation		
Competitor Benchmarking	Customer Regression	Evaluation Tool

Customer Regression

- Performance maintenance over releases
  - Macro regression: drop of performance in percentage, e.g. on accuracy
  - Micro regression: drop of performance on specific test examples

- Model determinism and stable training
  - Model determinism: deterministic training across N training runs
  - Stable training: across N training runs, different computing shapes, dependency package upgrades, model updates, etc

- Customer regression minimization and Hyperparameter tuning strategies
  - Penalizes regression more than reward improvements

Beyond AI Modellings: Running an AI Project

Evaluation Tool

- Different Test Types that are representative of different model features, e.g. class imbalance, out-of-domain (OOD) detection, multi-label classification, overfitting issue, etc.
- Different Test Types that are reflective of different *behavioral tests*, e.g. robustness, temporal, negation, etc.
- Delta columns that show the performance regressions between model releases.
- Status columns that indicate whether the test on a dataset is a PASS or FAIL: acceptance rate.
- Each individual regressed example should be tracible for manual investigation.
- Results from internal model explainability tools to help manual investigation.

Behavioral Testing

Negation	***MFT:*** Negated negative should be positive or neutral	18.8	54.2	29.4	13.2	2.6	The food is not poor. pos or neutral It isn't a lousy customer service. pos or neutral
	MFT: Negated neutral should still be neutral	40.4	39.6	74.2	98.4	95.4	This aircraft is not private. neutral This is not an international flight. neutral
	MFT: Negation of negative at the end, should be pos. or neut.	100.0	90.4	100.0	84.8	7.2	I thought the plane would be awful, but it wasn't. pos or neutral I thought I would dislike that plane, but I didn't. pos or neutral
	MFT: Negated positive with neutral content in the middle	98.4	100.0	100.0	74.0	30.2	I wouldn't say, given it's a Tuesday, that this pilot was great. neg I don't think, given my history with airplanes, that this is an amazing staff. neg

NER	***INV:*** Switching locations should not change predictions	7.0	20.8	14.8	7.6	6.4	@JetBlue I want you guys to be the first to fly to # Cuba → Canada... INV @VirginAmerica I miss the #nerdbird in San Jose → Denver INV
	INV: Switching person names should not change predictions	2.4	15.1	9.1	6.6	2.4	...Airport agents were horrendous. Sharon → Erin was your saviour INV @united 8602947, Jon → Sean at http://t.co/58tuTgli0D, thanks. INV

Robust.	***INV:*** Add randomly generated URLs and handles to tweets	9.6	13.4	24.8	11.4	7.4	@JetBlue that selfie was extreme. @pi9QDK INV @united stuck because staff took a break? Not happy 1K.... https://t.co/PWK1jb INV
	INV: Swap one character with its neighbor (typo)	5.6	10.2	10.4	5.2	3.8	@JetBlue → @JeBtlue I cri INV @SouthwestAir no thanks → thakns INV

Examples from "**Beyond Accuracy: Behavioral Testing of NLP Models with CheckList, ACL 2020**"

Beyond AI Modellings: Model Prototyping and Service Demo

Model Development (AI Modelling)

Research POC

Product POC

Research POC (proof-of-concept)

- To prove that an AI service is theoretically and empirically possible
- It involves public data collection; model prototyping; performance evaluation; and comparison with SOTA performance in research papers on public benchmarks.

Product POC

- To provide a service demo that shows the main functions and features of an AI service
- It involves use case or feature planning, UI/UX development, competitor profiling, and <u>potential customer identification</u>
- In this stage, we can start designing input and output APIs, e.g. what is expected from the end user, and what will be responded to them.
- DEMO

Beyond AI Modellings: Running an AI Project

Model Research and Development (AI Modelling)

Model Deployment and Serving

Bug Bash

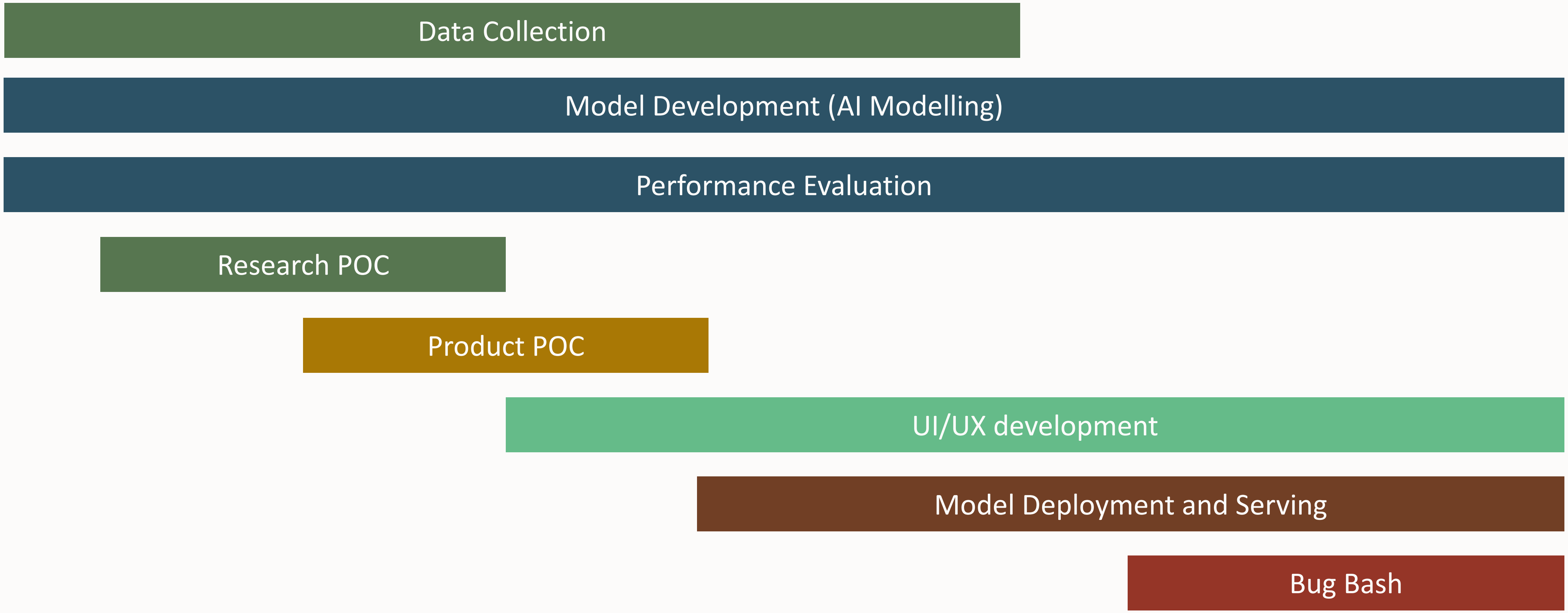
Model Deployment and Serving

- Running machine learning model in Docker
- Deploying model based on different DL libraries, e.g. TensorFlow, PyTorch, etc
- Loading test: to test how the model is coping with N requests per second (RPS)
- Resource sharing: embedding as a service, caching mechanism, language detection, etc

Bug Bash

- Internal or cross-team bug bash
- Pre-production bug bash helps identify bugs that could appear at all corners
  - unexpected model output, e.g. a model could give weird output for gibberish inputs
  - latency issues, e.g. a response takes longer than expected to complete
  - UI/UX bugs, e.g. not responsive to a button

Beyond AI Modellings: Running an AI Project



Innovation is encouraged in every stage of this pipeline.

- Our team filed **58 patents** in the last 3 years
- We encourage team members to attend **AI/NLP/CV conferences** and **forums**
- We encourage literature reviews and **blog style discussions**

Discussion: Data collection

Text Classification

We want to build a text classifier that classifies text documents to1000 pre-defined classes identified from prioritized domains. We don’t have training examples for building this model; and would like to collect data from 3rd party data collection companies.

- There are hierarchical relations between the 1000 classes.
- One document can be assigned to more than 1 label.
- We can collaborate with multiple 3rd party companies.

How would you run the data collection project?

How can you guarantee the high quality of the collected data?

How would you cope with subjectiveness of the annotation?

PollEv.com/jeyhanlau569

Discussion: Data collection

Text Classification

We want to build a text classifier that classifies text documents to1000 pre-defined classes identified from prioritized domains. We don't have training examples for building this model; and would like to collect data from 3rd party data collection companies.

- There are hierarchical relations between the 1000 classes.
- One document can be assigned to more than 1 class labels.
- We can collaborate with multiple 3rd party companies.

How would you run the data collection project?

How can you guarantee the high quality of the collected data?

How would you cope with subjectiveness of the annotation?

As a follow-up, we also want to collect data in Arabic.

How would you deal with it?

PollEv.com/jeyhanlau569

Discussion: Deterministic and stable training

Regression Analysis

We have a test dataset with 20 test examples from a customer that are considered very important (It could be that our customer uses these 20 examples as demo cases to show to their customers). However, 5 of the previously PASSED test examples FAILED in the recent model update release.

- The model update is the only source of change; nothing in the pipeline around the model has changed; no data change.
- The model update seems not relevant to the 20 test examples. For example, the model update should be only affecting datasets with 10K training examples; while this customer only provided 5K training examples.
- The model is based on adapting a pretrained language model. It involves updating the embedding layers, encoder layers, and the task-specific layers.
- The model is trained on both original data and augmented data. A number of augmentation techniques are applied.
- The model is build based on TensorFlow.

How would you investigate these failed cases and provide an explanation to the customer?

PollEv.com/jeyhanlau569

Discussion: Performance - Latency trade-off

Model Inference Latency

We want to release a service; and the project is in the deployment and serving stage. However, we found that the model is responding super slow when it is served for deployment; and the PM is expecting the model to give response in 1/5 of the time.

- The service takes plain text documents as input; and output analytic information about the input documents.
- The current model is based on adapting a pretrained language model. Training the model involves updating the embedding layers, encoder layers, and the task-specific layers.
- The current model is implemented in TensorFlow; and served with TF-Serving.

How would you investigate this problem and provide solution?

PollEv.com/jeyhanlau569

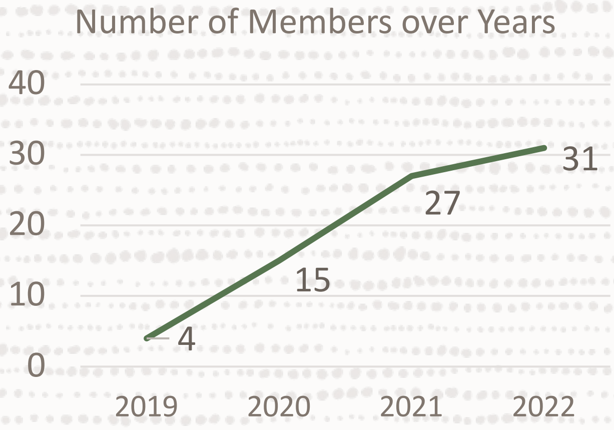
We are hiring!

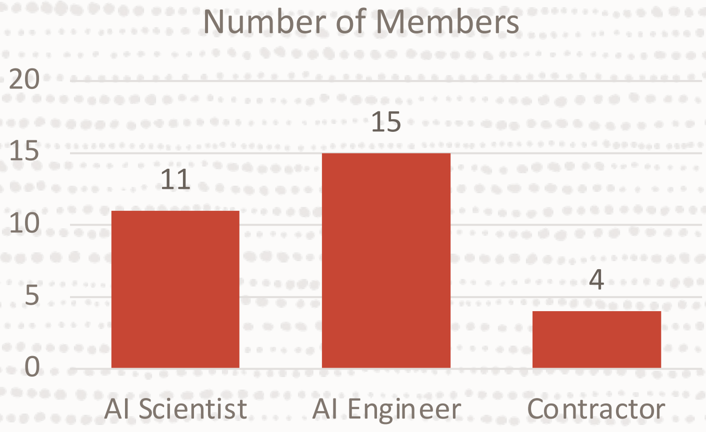
1. Our open positions
    - **AI/NLP/CV Scientist**
    - **AI/NLP/CV Engineer**
    - **AI/NLP/CV Interns** (coming soon)
    - **Language Engineer** (coming soon)

2. Our daily work
    - **Start-up style**
    - **Challenging problems and tasks**
    - **Research-backed product development**

3. OCI Values
    - PUT CUSTOMERS FIRST
    - ACT NOW, ITERATE
    - NAIL THE BASICS
    - EXPECT AND EMBRACE CHANGE
    - TAKE RISKS, REMAIN CALM
    - INNOVATE TOGETHER
    - OWN WITHOUT EGO
    - EARN TRUST, GIVE TRUST
    - TAKE PRIDE IN YOUR WORK
    - CHALLENGE IDEAS, CHAMPION EXECUTION

About Our Team





94%

ODA Revenue YoY Growth
from
FY21 to FY22

ORACLE Digital Assistant

Language AI Services

~80M

Requests to OCI Language services
in FY22

Thank you!

ORACLE

Our mission is to help people see data in new ways, discover insights, unlock endless possibilities.