BUILDING AN AI PLATFORM
KEY PRINCIPLES AND LESSONS LEARNED

Moty Fania
Principle Engineer (CTO), Advanced Analytics

NY, Sep 2019
Legal notices

This presentation is for informational purposes only. INTEL MAKES NO WARRANTIES, EXPRESS OR IMPLIED, IN THIS SUMMARY.

For more complete information about performance and benchmark results, visit www.intel.com/benchmarks

Intel and the Intel logo are trademarks of Intel Corporation in the U.S. and/or other countries.

* Other names and brands may be claimed as the property of others.

Copyright © 2019, Intel Corporation. All rights reserved.
AI in a Nutshell

Artificial intelligence is about **replacing human decision making**

- These are not repetitive tasks, but rather **judgment-based** decisions
- Measured by the **quality of decisions** & actions taken to achieve a desired objective
- **Machine learning** as the key technology to use a variety of inputs, recognize patterns, predict future outcomes and make decisions.

**Narrow AI**

Also referred to as “weak AI.” This is AI that works within a very limited context, and can’t take on tasks beyond its field. It is the only form of Artificial Intelligence achieved so far.
### About Us – Advanced Analytics @ Intel

<table>
<thead>
<tr>
<th>AI IN PRODUCTS</th>
<th>HW VALIDATION</th>
<th>PRODUCT DEV</th>
<th>INDUSTRIAL AI</th>
<th>SALES</th>
<th>MOBILEYE</th>
<th>HEALTH</th>
</tr>
</thead>
<tbody>
<tr>
<td>Improve products power/ performance</td>
<td>Cut product time to market</td>
<td>Reduce test cost, improve quality, increase capacity</td>
<td>Yield and quality improvement</td>
<td>Increase sales revenue</td>
<td>Differentiated data services</td>
<td>Improve clinical trials outcome</td>
</tr>
</tbody>
</table>

**MISSION:** TRANSFORM CRITICAL WORK USING AI
Why AI Platforms?

- Developing good ML models is not sufficient by itself
- It’s equally important to effectively deploy these models to production and integrate them into the business process in a sustainable way
- We build designated AI platforms to:
  - Enable easy way to rapidly deploy models to production
  - Bring relevant data to the algorithm or vice versa (stream, batch)
  - Enable a closed feedback loop (at scale and in a timely manner)
  - Sustain over time – manageability, re-train models, logs, indicators etc.

AI Platforms enable a self-sustained, on-going AI Service in production with built-in integration points for different kinds of models
The Characteristics of Our AI Platforms

- Deep integration to biz processes
- Continuous delivery-deployment
- On-Prem, Cloud, Hybrid
- Closed feedback loop & Actuation

<table>
<thead>
<tr>
<th>Category</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PRODUCT DEV</strong></td>
<td>- Dedicated platform to fully automate Unit flow from assembly to end of test</td>
</tr>
<tr>
<td></td>
<td>- Unique test route and actuation at unit level</td>
</tr>
<tr>
<td></td>
<td>- Mostly structured data - Models are Python based</td>
</tr>
<tr>
<td><strong>HW VALIDATION</strong></td>
<td>- Collects millions of Test logs (stream)</td>
</tr>
<tr>
<td></td>
<td>- Need to run distributed algorithms on Big data (Volume, Variance, Velocity)</td>
</tr>
<tr>
<td></td>
<td>- Simple interface to Biz (REST API)</td>
</tr>
<tr>
<td><strong>SALES AI</strong></td>
<td>- Specialized data stores – Knowledge Graph, Search, Big data object store</td>
</tr>
<tr>
<td></td>
<td>- Crawling (Web and social media )</td>
</tr>
<tr>
<td></td>
<td>- Hybrid platform (On-Prem + Public cloud)</td>
</tr>
<tr>
<td><strong>INDUSTRIAL AI</strong></td>
<td>- Edge, Fog, Cloud solution</td>
</tr>
<tr>
<td></td>
<td>- Can be deployed anywhere – On any premise / any cloud</td>
</tr>
<tr>
<td></td>
<td>- Time-series modeling + Real-time actuation, Visual Inspection</td>
</tr>
<tr>
<td><strong>HEALTH</strong></td>
<td>- Hosted in the Public cloud, Customers are paying for service,</td>
</tr>
<tr>
<td></td>
<td>- Serverless implementation to optimize cost</td>
</tr>
<tr>
<td></td>
<td>- Edge (Sensors, Mobile) + cloud integration (Big Data)</td>
</tr>
</tbody>
</table>
Sales AI

**SENSE**
Real time scan of customer information

**REASON**
Detect intent to buy and potential opportunities

**INTERACT**
Automate a personalized best action

**LEARN**
Automatically learn based on customer response

Our sales AI program led to a cumulative sales revenue growth > $450M in the past five years
Sales AI Platform – What is required?

- Online streaming system
- Agile Micro-services, architecture
- Real time DL inference service
- Specialized data stores – Knowledge Graph, Search, Big data object store
- NLU and auto labeling
Online streaming system

Twitter Service

Handles Queue

Seed Queue

Crawl Service

MESSAGE BUS

DATA STORES

Reasoning Agents

* Intel's Sales AI uses APIs provided by Twitter to collect data only from accounts the user has designated as public and adheres to individual websites’ terms of use. The only data collected is what is permitted by the site owner under their terms of use agreement.
Micro-services architecture
DEEP LEARNING INFERENCE SERVICE
Archelon - Deep Learning Inference System

- AI may involve massive usage of Deep Learning models
  - NLP / NLU
  - Recommendation Systems
  - Visual inspection / understanding
- A production grade, managed, inference system that serves DL models and enables online, closed feedback loop
- Continuous delivery of models to production

- Production Inference service for DL models
- Smart in-memory cache for data batching, sequencing
- Fast, scalable APIs for data ingestion & real time responses
- Full Scalability
The Latency Challenge

Capture

$\pm 100 \text{ f/sec}$

Process

ML Reasoning

$5 \times$ CNNs

$2 \times$ LSTM

Act

$< 1 \text{ second}$

* Other names and brands may be claimed as the property of others.
Asynchronous Inference Unit (AIU)

- Always ON
- Continuous resource utilization
- Stateless
- Dynamic Batching
- Logical grouping of Models – “Units”
- Easily scalable

* Other names and brands may be claimed as the property of others.
AIU Allocation

AIUs enable flexibility and Separation of concerns

Can be allocated as required

- per model
- per data source
- Latency / throughput
- Inference HW

Example: Dedicate AIU per model

Data Store

* Other names and brands may be claimed as the property of others.
Tensor Flow Serving

A flexible, high-performance serving system for machine learning models, designed for production environments

We have added the following capabilities:

- Simple APIs to deploy new models
- Generic client to query any model (sync / Async)
- Optimized Docker image for CPU & GPU
- Implementation within Kubernetes with performance optimizations & scaling
- Full automation of deployment

* Other names and brands may be claimed as the property of others.
DATA STORE
Flexible Data Store(s)

SQL

SEARCH (ENGINE)

BIG DATA STORE

KNOWLEDGE GRAPH
Why Knowledge Graphs?

- Graphs are a natural way to represent entities and their relationships
- **Structured** and formal representation of knowledge **over time**
- Enable **reasoning** to infer new knowledge
- Graphs can be managed **efficiently**

**Semantic descriptions of entities and their relationships**
Mastodon

A generic service for building and sustaining graphs at scale

- Microservice architecture
- Asynchronous stream processing using Kafka as the message bus
- Easily deployable with Docker, K8S and Helm
- Configuration driven - configurable schema
- Decoupled from specific Graph technology
Graph Extractors and Transformers

- Always ON (Docker containers)
- Stateless
- Configuration driven behavior
- Each handles a certain graph entity
- May produce work for others!
- Implemented with Kafka streams

**Transformer** – Transforms raw string data from message bus to specific graph semantics
**Extractor** – Implements the logic for a specific entity extraction from an external source

*Other names and brands may be claimed as the property of others.*
High level architecture

* Other names and brands may be claimed as the property of others.
A Sample Flow

Graph Builder Queue

Company Queue

Company Extract Queue

Product Queue

DATA LOADER

CRM
Companies
Data

“COMPANY” TRANSFORMER

“COMPANY” EXTRACTOR

“PRODUCT” EXTRACTOR

* Other names and brands may be claimed as the property of others.
Natural Language Understanding

- Named-entity recognition (NER) ➔ We’re excited to report that Amazon buys Ring to get into home security business.
- Text-classification ➔ M&A
- Relations extraction ➔ Amazon To Acquire Ring
- Intent classification ➔ Amazon Enters Industry Home Security
- Sentiment analysis ➔ Positive

We’re excited to report that Amazon buys Ring to get into home security business
Osprey - a weak-supervision labeling system

- Many (text-related) business problems require supervised methods but require expensive labeled dataset creation and maintenance
- Snorkel is an open-source framework capable of training models for supervised problems without manual-labeling
- Osprey is a weak-supervision system which we implemented on top of Snorkel to support highly-imbalanced data
- With Osprey, the programmatic labeling is decoupled into a code layer and a configuration one.
- Osprey and Snorkel offer better performance over prior traditionally supervised and rules-based

Putting it all together

**Archelon**
- NLU / Information extraction
- Deep Learning Inference

**RAY**
- Computation Engine
- Action service - Points
- Action service - Training
- Action service - Sales Assists
- Action service Matchmaking

**Message Bus**
- Crawler
- Graph Builder
- Relations analytics

**Databases**

---

*Intel’s Sales AI uses APIs provided by Twitter to collect data only from accounts the user has designated as public and adheres to individual websites' terms of use. The only data collected is what is permitted by the site owner under their terms of use agreed to.
Sales Assists - User Interface

Sales Assist provides relevant customer activity to the account manager with details that provide additional information.
With the “Details” for a “Viewed Intel product” assist, the account manager can see the specific pages the customer visited on intel.com and Sales Assist highlights that the customer has not previously purchased the product.
An Autonomous sales Example

SENSE DIRECT INTENT

REASONING

RECOMMENDER SYSTEM

Available Points Offers

490

Customer DNA

Single Customer View

INTERACT

Communication in 15 different languages

INTERACT

Sent to ~3% of contacts

LEARN

A/B Testing

- Open rate
- Recommended SKU's purchased in relevant period

~150K Contacts
THANK YOU!
Please rate this session ;-)
Sharing Intel IT Best Practices With the World

Learn more about Intel IT's initiatives at: www.intel.com/IT
RAY – A Scalable computation engine

• Ray is a flexible, high-performance, distributed execution framework for Emerging AI Applications (UC Berkeley)
• A scalable system for parallel and distributed Python.
• Shortens path to production (scalability, throughput, tolerance)
• Dev & DS friendly – Powerful yet easy actor system
• Superior performance to Spark in certain use cases

<table>
<thead>
<tr>
<th>Input Value</th>
<th>Without Ray(Single Core)</th>
<th>With Ray(Standalone/Loc al with 4 CPU cores)</th>
<th>With Ray (Using a cluster of 3 nodes with 28 CPU cores per nodes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>200</td>
<td>26.05 sec</td>
<td>8.32 sec</td>
<td>1.31 sec</td>
</tr>
<tr>
<td>300</td>
<td>134.77 sec</td>
<td>40.42 sec</td>
<td>7.72 sec</td>
</tr>
<tr>
<td>400</td>
<td>435.57 sec</td>
<td>128.89 sec</td>
<td>23.19 sec</td>
</tr>
</tbody>
</table>

Ray on standalone mode is on an average 3x faster and on cluster mode is on an average 18x faster than the scenarios performed without Ray.