



Democratizing AI

Back-fitting end-to-end machine
learning at LinkedIn *at scale!*



Bo Long



Joel Young



Today's agenda

1:30

Introduction: LI Products & AI Overview

1:40

The AI Problem Space at LinkedIn

1:45

Deep Dive: Scaling Our Technology: Pro-ML

1:55

Scaling Our People: AI Academy

2:00

Conclusion/Q&A

People You May Know

This screenshot shows the LinkedIn mobile application's main feed. At the top, there's a search bar with placeholder text "People, jobs, posts and more...". Below it, a notification badge indicates "670 Connections". A banner highlights "5 of your batchmates from Indian Institute of Management (2010-2012)". Below the banner, a grid of user profiles for Suresh, Kiran, Venk..., Zishan, and Nikita, each with a "CONNECT" button. The main feed displays a post from Tomer Cohen with the caption "Congrats Lior Ron on the acquisition! Drinks are on you next time we meet." Below the post is a large image of a human brain with the text "Artificial intelligence is more artificial than intelligent". At the bottom of the screen are navigation tabs for Home, Jobs, Network, and Profile.

AI for LinkedIn products

Feed

This screenshot shows the LinkedIn mobile application's "Feed" section. It features a post from Tomer Cohen with the caption "Congrats Lior Ron on the acquisition! Drinks are on you next time we meet." Below the post is a large image of a human brain with the text "Artificial intelligence is more artificial than intelligent". At the bottom of the screen are navigation tabs for Home, My Network, Messaging, Notifications, and Jobs.

Jobs

This screenshot shows a job listing for a "Principal Data Scientist" position at Microsoft in Bangalore, posted 3 days ago with 1,142 views. The listing includes a Microsoft logo, a "Save" button, and an "Apply" button. Below the listing, it says "32 connections can refer you" and "Ask for a referral". The job description section discusses building the next generation of cloud services for partner monetization, user acquisition, engagement & membership platform. It mentions a global footprint of over 240 markets and process millions of transactions daily. The industry is listed as Computer Hardware, Computer Software, Information Technology and Services. The seniority level is "Not Applicable".

Learning

Courses related to Deepak's skills

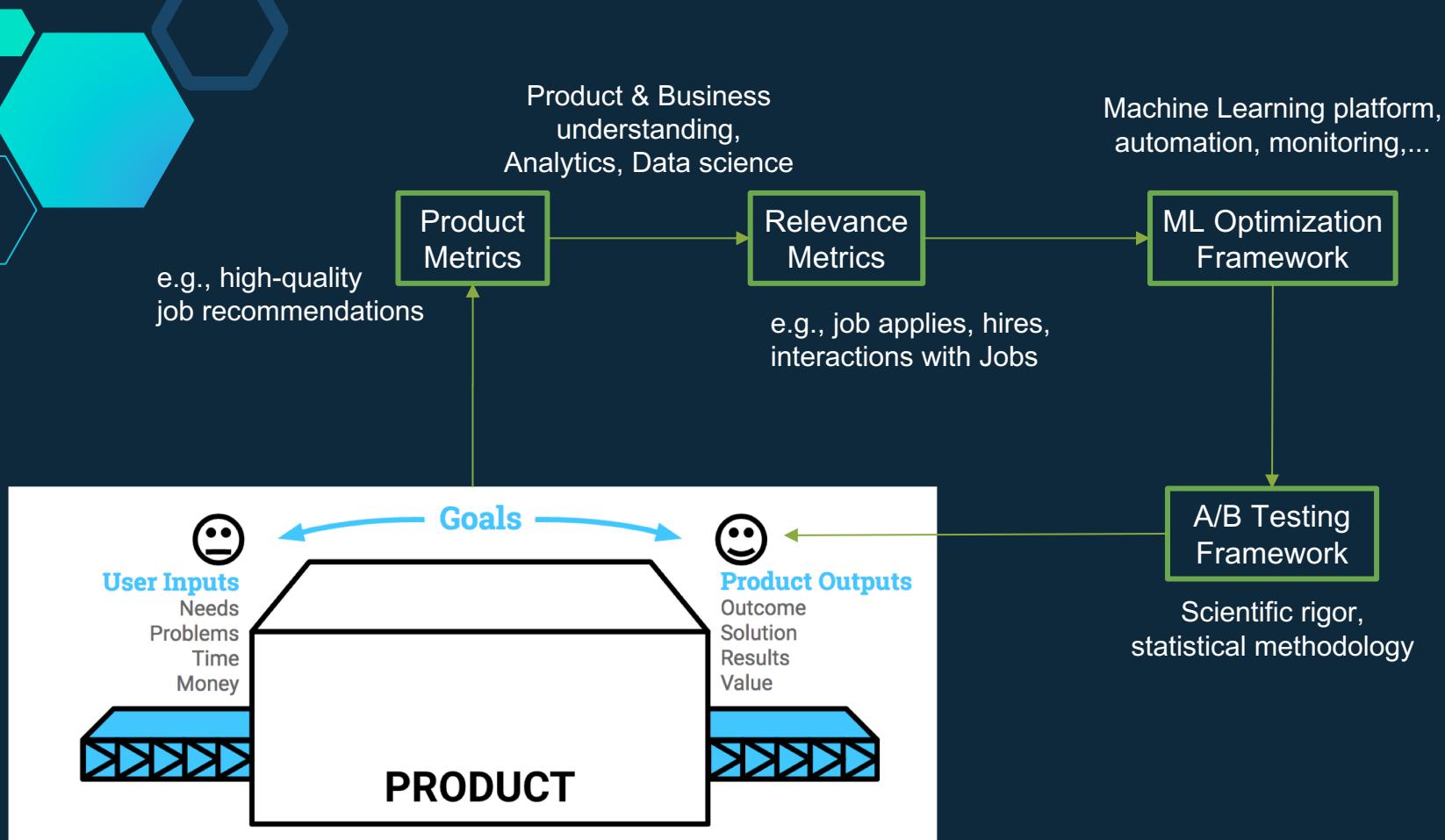
This screenshot shows a list of three LinkedIn Learning courses: "Data Science Foundations: Python Scientific Stack" (Viewers: 6,033), "Pandas for Data Science" (Viewers: 10,197), and "Building and Deploying Deep Learning Applications with TensorFlow" (Viewers: 2,636). Each course entry includes a play button icon and a brief description. At the bottom right, there is a "See more courses" link.

Recruiter Search

This screenshot shows the LinkedIn Recruiter mobile application. It displays a search results page for "Project Manager" in "Greater Chicago Area". The results show 9K total candidates, 463 have company connections, 201 are present in your talent board, and 27 have applied. The interface includes sections for "Updates", "Your search results", "Project activity", and "Upcoming events".

This screenshot shows the LinkedIn Sales Navigator web interface. It displays a pipeline of deals across various accounts. Key columns include Account, Deal Name, Buyer circle, Amount, Close date, Stage, Probability, and Next steps. A sidebar on the right provides notes for specific accounts, such as "Met with the legal team to finalize the paperwork." and "Jacob seems really interested in this role...".

Sales





Opportunities for AI at LinkedIn

Machine Learning Contributes Everywhere!



Opportunities for AI at LinkedIn

Many Technologies!



Opportunities for AI at LinkedIn

All Mission Critical!



Opportunities for AI at LinkedIn

Machine Learning Is Becoming Democratized!



Opportunities for AI at LinkedIn

Too Much Friction!



Opportunities for AI at LinkedIn

State of the Art Advancing Quickly!



Opportunities for AI at LinkedIn

Which Stack Will Win?



Productive Machine Learning

Pro-ML



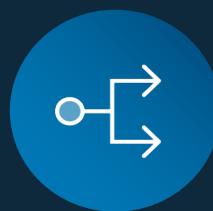
Goal of Pro-ML

Make the end-to-end ML experience
easy, fast, robust, and automatic



Model Creation

Explore a large space
of different variations
of a model



Deployment

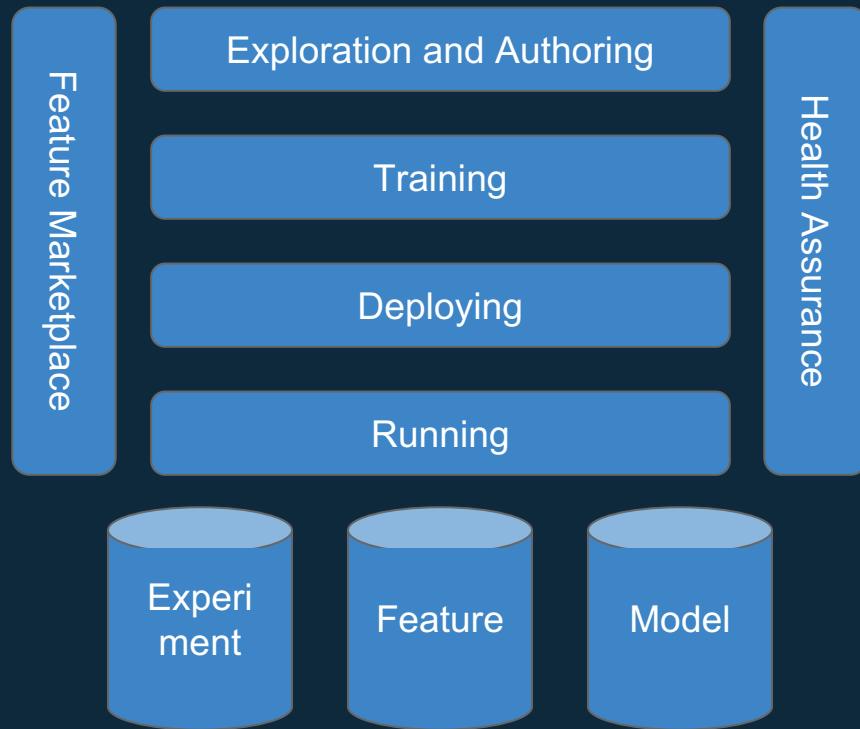
Deploy different
components of the model
and features in different
places in production



Maintenance

Continuously monitor
model health, data
quality and detect
anomalies

Pro-ML Layers





Organizationally

Team of Teams

Finding a more
global optima!

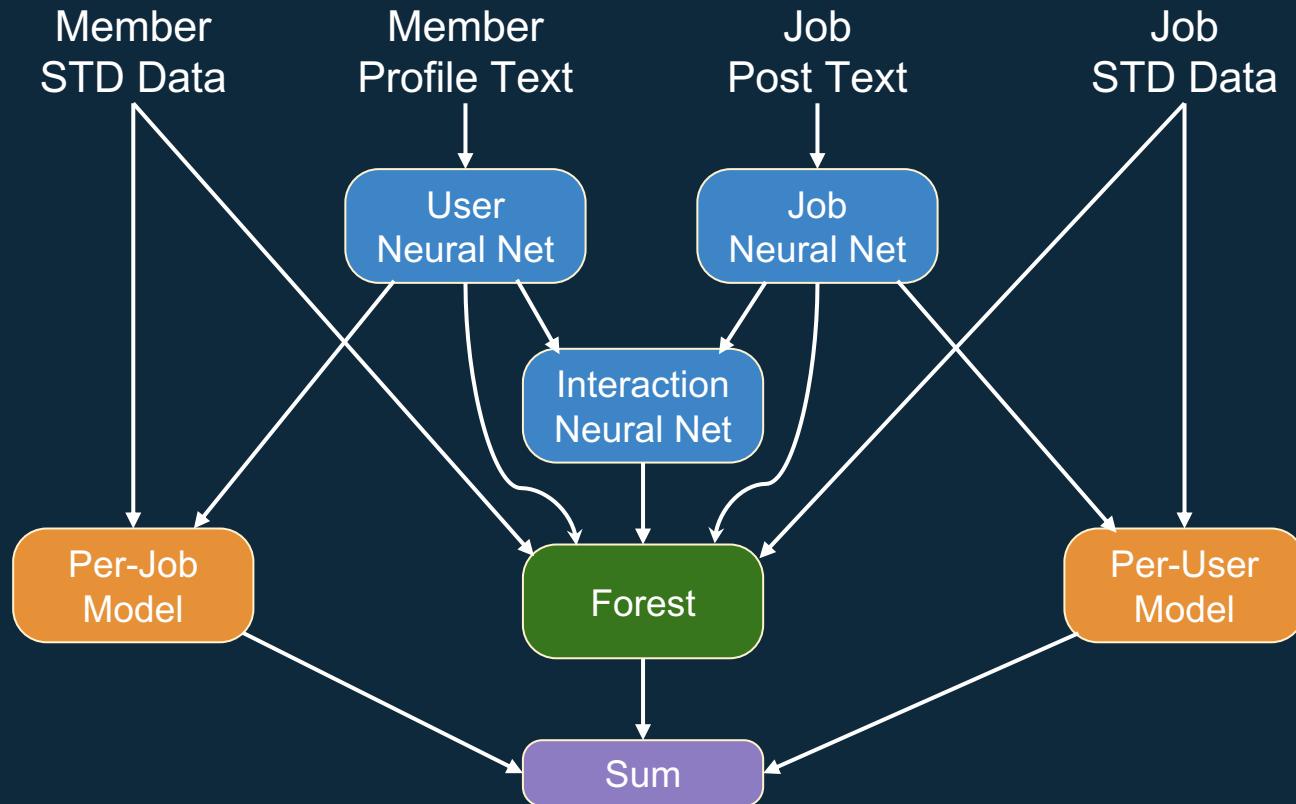
Team of Teams





What is a Model?





Exploring and Authoring - Jupyter

The features that we will look at are title embedding features. Using an advanced neural network, we learn a vector representation (called an **embedding vector**) for each title from a large amount of data, so that similar titles are close to one another in the vector space. The following figure shows a few examples. For instance, "Recruiter" and "Human Resources Consultant" are very close to each other, although they have not words in common. Text-based features would not be able to capture this relationship.

The figure is a 2D scatter plot illustrating the results of a neural network's title embedding. The horizontal and vertical axes are represented by a light gray grid. Blue square markers are placed on the grid, each representing the vector embedding for a specific job title. The titles are labeled as follows:

- Top-left cluster: Chef, Sous Chef, Chef De Partie
- Middle-left cluster: Cook
- Bottom-left cluster: Legal Specialist, Legal Counsel, Partner, Senior Advisor, Committee Member
- Bottom-right cluster: Recruitment Specialist, Recruiter, Human Resources Consultant

The labels are positioned near their respective blue square markers. The plot demonstrates that the neural network has learned to place similar job titles close together in the vector space, even if they do not share common words.

Exploring and Authoring - Quasar



```
75      [ "", "cosine", {"type": "SPARSE", "indexMap": "DYNAMIC"});  
76 Vector interaction_titleNwEmbeddingV1 = SimilarityFeatureProducer(  
77     [member_primaryTitleNwEmbeddingV1],  
78     [job_jobTitleNwEmbeddingV1],  
79     [ "", "cosine", {"type": "SPARSE", "indexMap": "DYNAMIC"});  
80 Vector interaction_titleNwEmbeddingV3 = SimilarityFeatureProducer(  
81     [member_primaryTitleNwEmbeddingV3],  
82     [job_jobTitleNwEmbeddingV3],  
83     [ "", "cosine", {"type": "SPARSE", "indexMap": "DYNAMIC"});  
84  
85 // Use a set of trees to compute the final score  
86 double score = DecisionTreesScorer(treeModel);  
87  
88 list = ORDER DOCUMENTS BY score WITH DESC;  
89  
90 RETURN list;  
91  
92 TRAINING CONFIG sentinel SCORING BY score REQUIRED WEIGHT PAIRS ([  
93
```

Exploring and Authoring - Plotting



```
In [*]: %matplotlib inline

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

fig, ax = plt.subplots(figsize=(8,7))

for key, grp in featureStats.groupby(['term']):
    ax = grp.plot(ax=ax, kind='line', x='bucket', y='dismissRate', label=key)

ax.set_xlabel('Feature value (title similarity score)')
ax.set_ylabel('Dismiss rate')
plt.legend(loc='best')
plt.show()
```

In the above figure, each curve correspond to one feature (each of which corresponds to a way of

Exploring and Authoring - Spark



```
val baseline = trainingOutput("baseline").select($"Metric", $"value".as("Baseline"))
val newModel = trainingOutput("newModel").select($"Metric", $"value".as("NewModel"))
val modelComparison = baseline.join(newModel, "Metric").
    select($"Metric", $"Baseline", $"NewModel", lift($"Baseline", $"NewModel"))

modelComparison: org.apache.spark.sql.DataFrame = [Metric: string, Baseline: double ... 2 more fields]
```

In [12]: %matplotlib inline
modelComparison

Type:

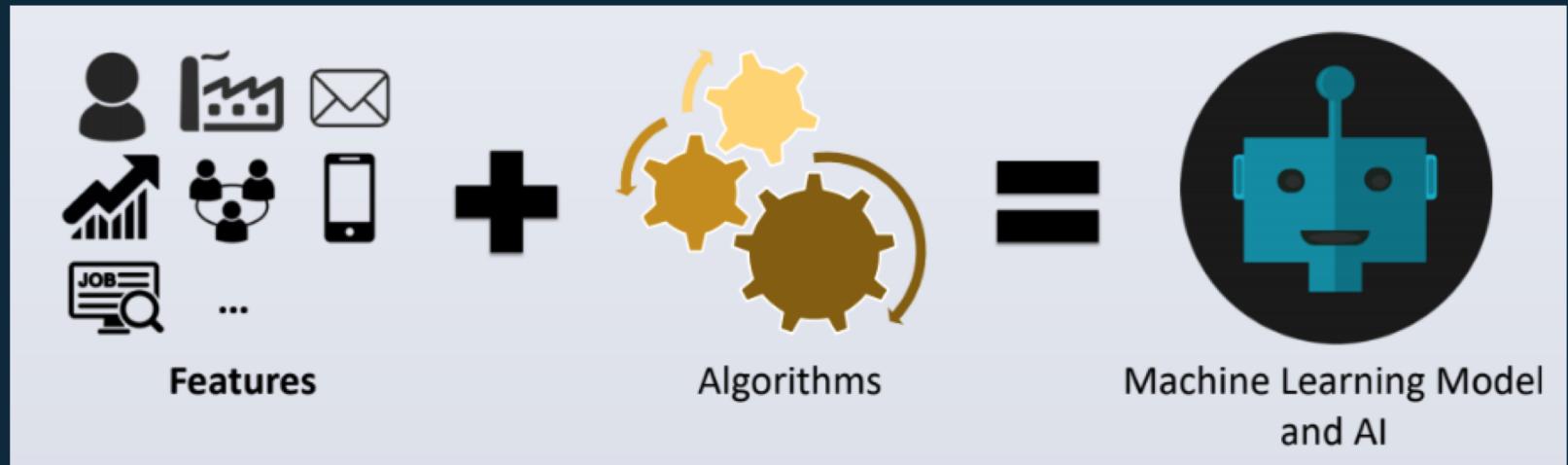
Metric	Baseline	NewModel	Lift
ACCURACY	0.612147	0.624582	0.020314
AUROC	0.660319	0.673524	0.019997



How to Train the Model



Features and Algorithms



Feature Marketplace

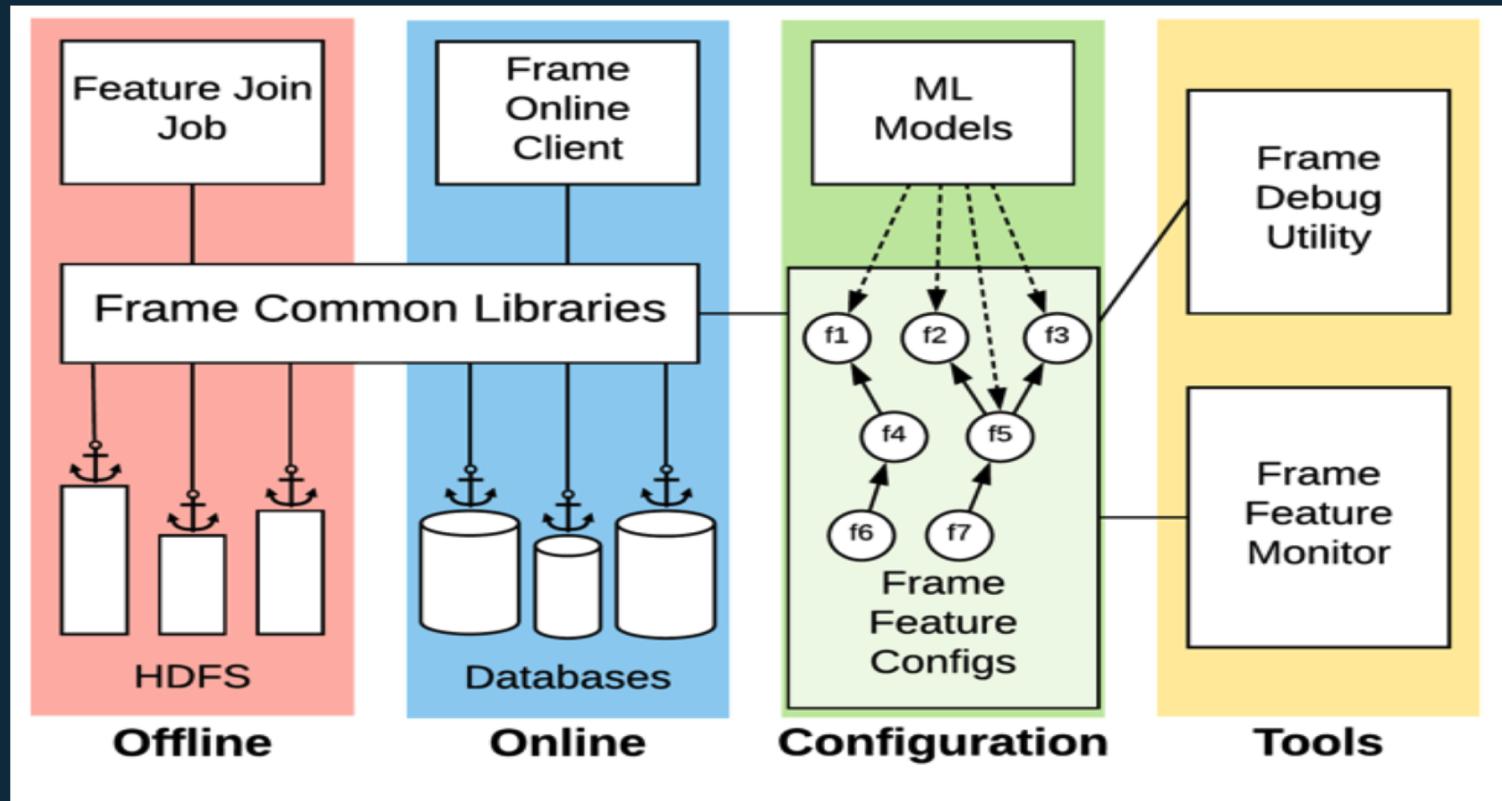


MODOOP FEATURES Search features...

13413 results found in 25ms
Clear all filters

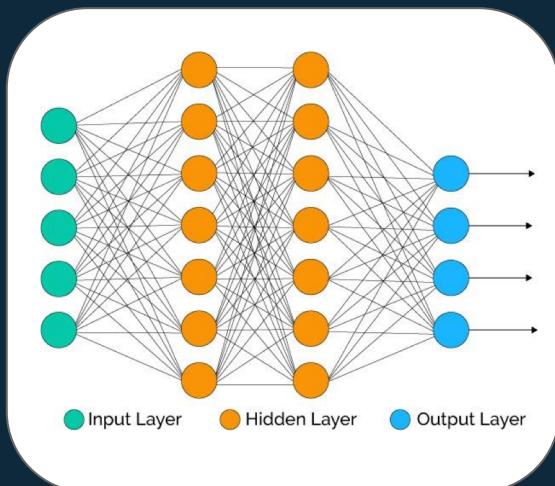
Feature ID	Feature Name	Data Type	Feature Type	Description	Aggregation Length
1	imported_contacts	long	numerical	total imported contacts	
2	imported_contacts_107d	long	numerical	total imported contacts within last 7 days	
3	imported_contacts_130d	long	numerical	total imported contacts within last 30 days	
4	is_uploaded_abook_07d	long	binary	has uploaded abook within last 7 days	
5	is_uploaded_abook_130d	long	binary	has uploaded abook within last 30 days	
6	is_uploaded_abook_190d	long	binary	has uploaded abook within last 90 days	
7	conn_uploaded_abook	long	numerical	total abook uploaded from connections	all,last 7 days
8	email_appeared	long	numerical	times of profile appeared in others abook email	all,last 7 days

Feature Access Consistency

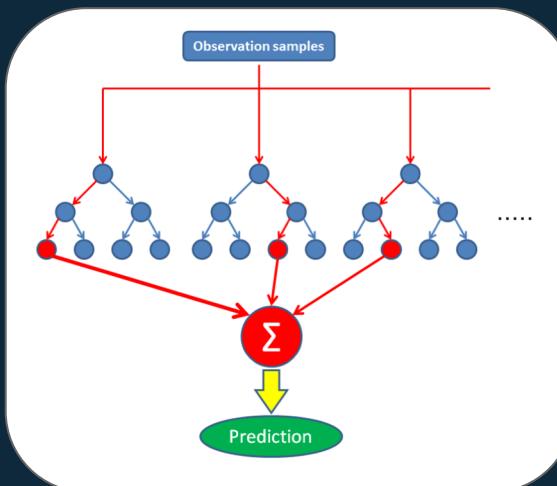


Algorithms

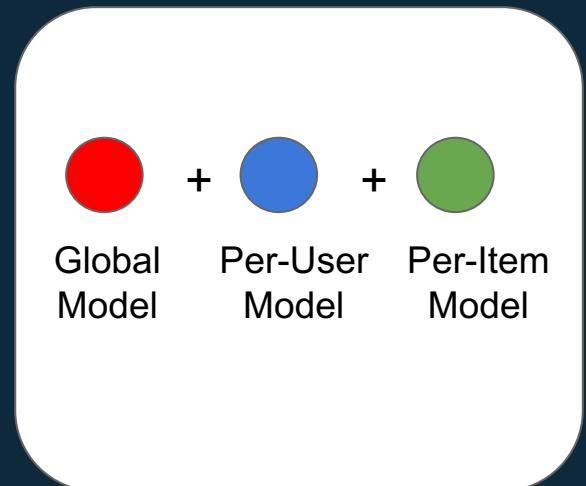
Deep Learning



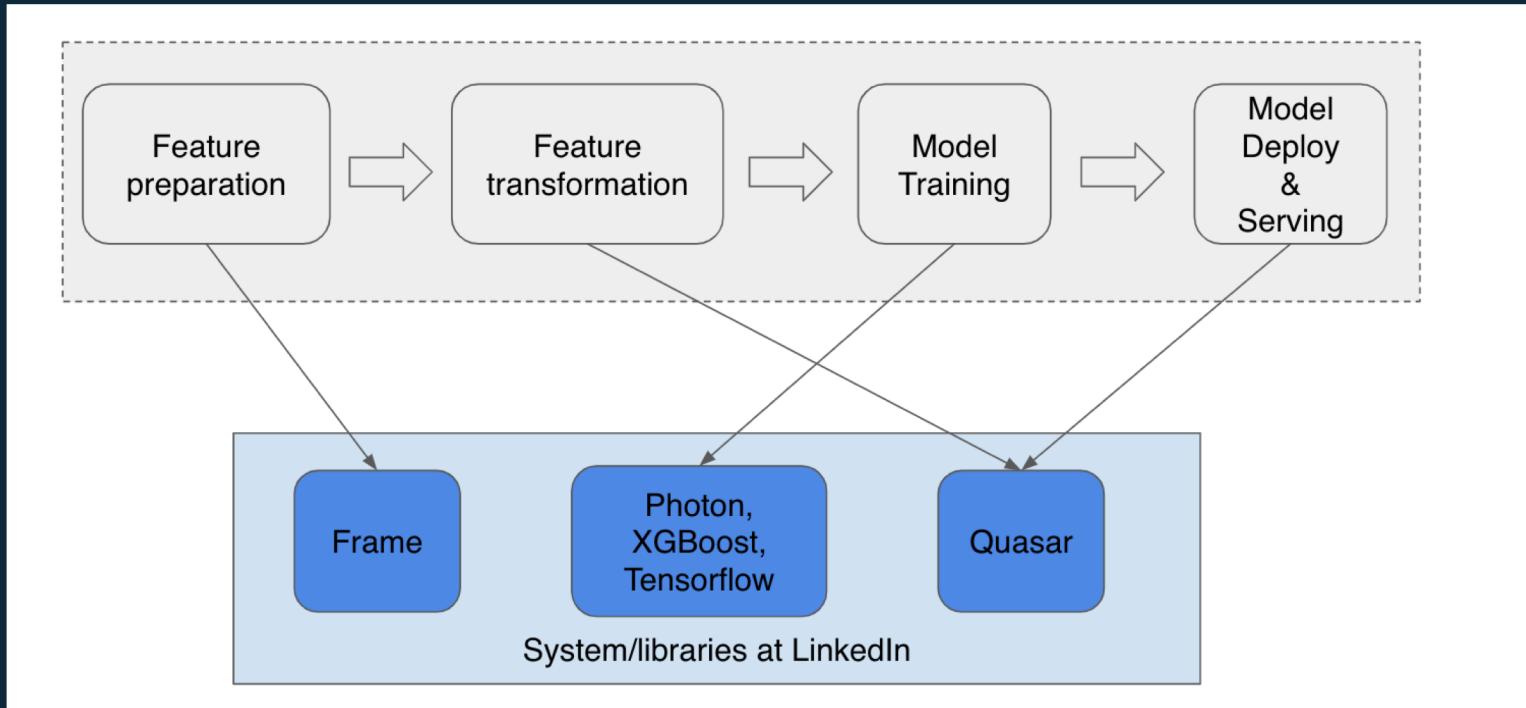
Trees

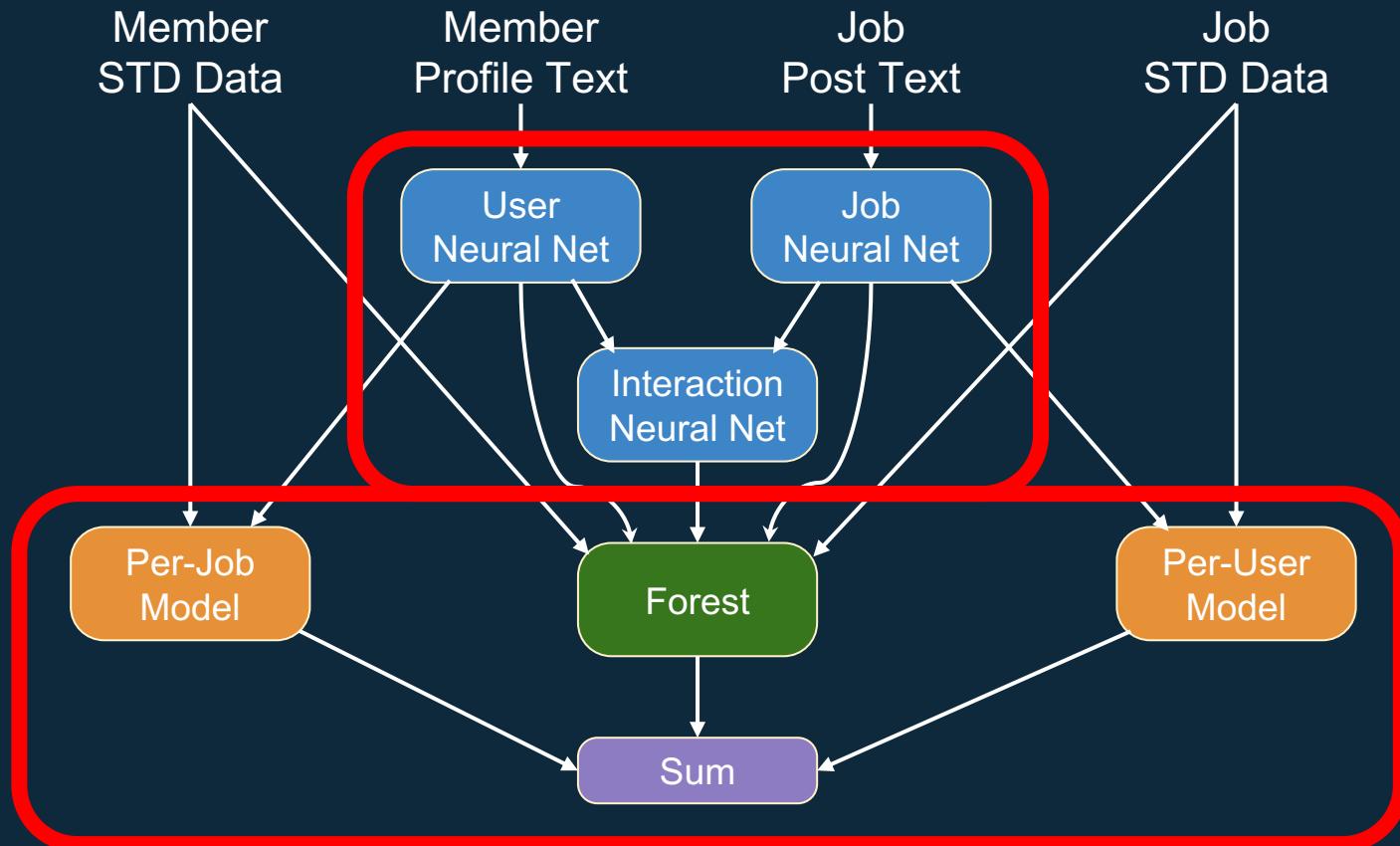


General Linear
Mixed Models



Training Engine: Photon-Connect







The Problem is Huge

$$\begin{array}{c} \text{User icon} \\ 500M \end{array} \times \begin{array}{c} \text{Briefcase icon} \\ 10M \end{array} = 5,000,000,000,000$$



How We Learn From Data



Global
Component

Macro patterns



Member
Component

Member specific micro patterns



Job
Component

Job specific micro patterns



It's Looking Better!



1

x



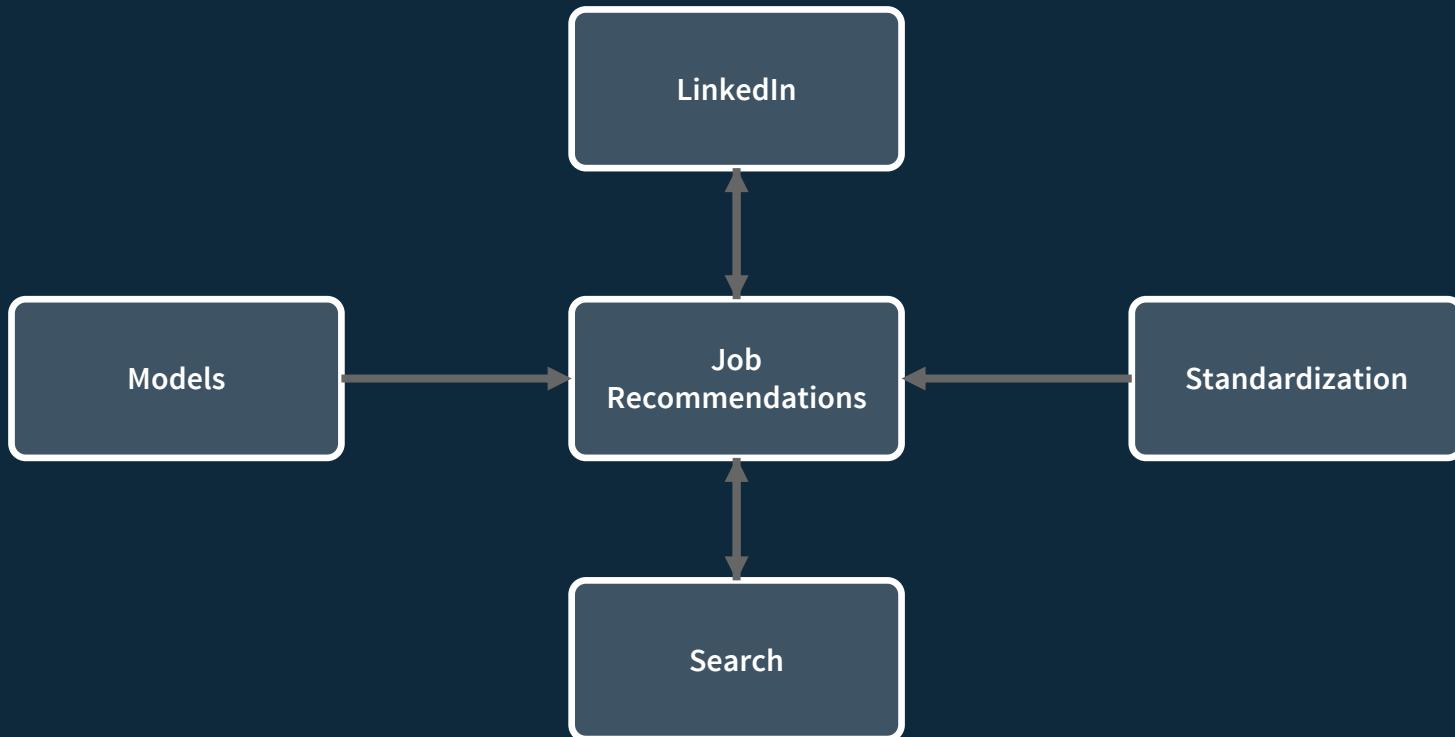
10M

=

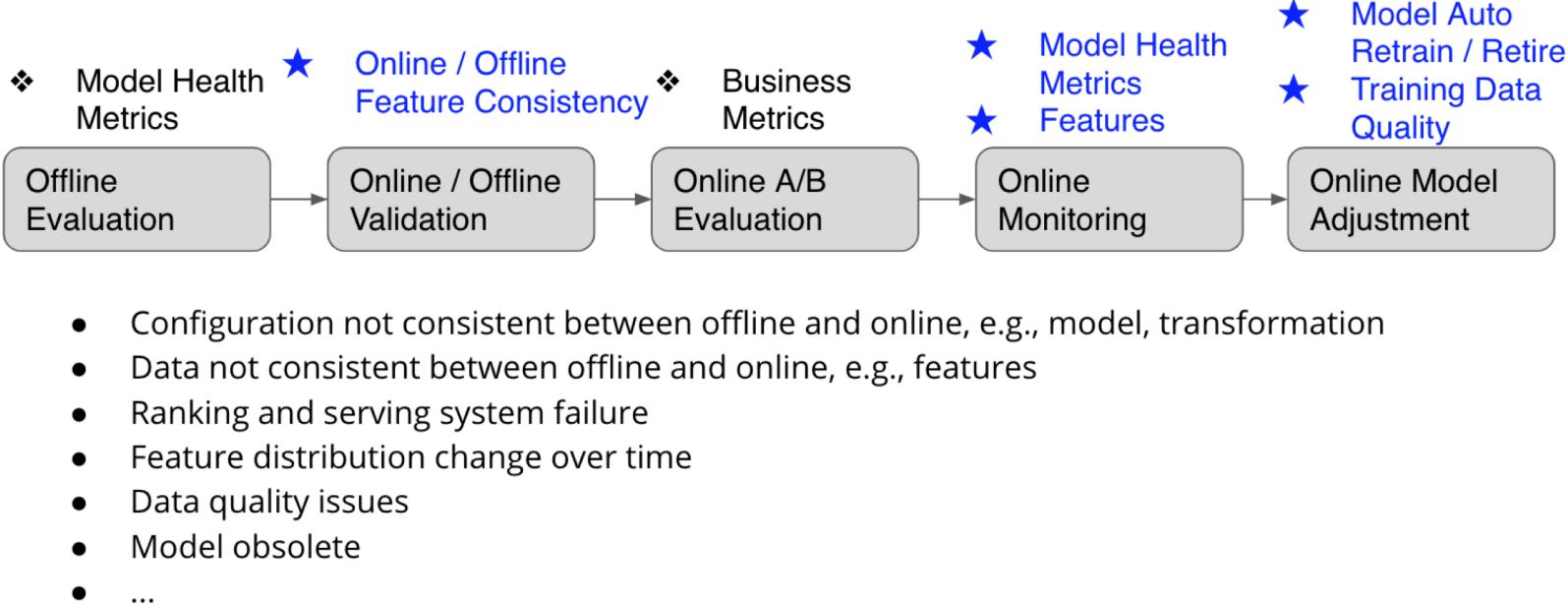
1	0	,	0	0	0	,	0	0	0
---	---	---	---	---	---	---	---	---	---



How We Make it Happen



Health Assurance Model Development Cycle



Anomaly Detection

Anomaly Overview

Anomaly ID: guest_invite from business_intraday_metrics_hourly_additive
Metric: intraday_hourly_guest_invites
Dimensions: Time & Duration: [2018, 1 pm] - [2018, 5 pm] (4 days)

Duration	Current Avg	Baseline Avg	Change
4 days	A REALLY BIG NUMBER!	A BIG NUMBER!!!	PCT%

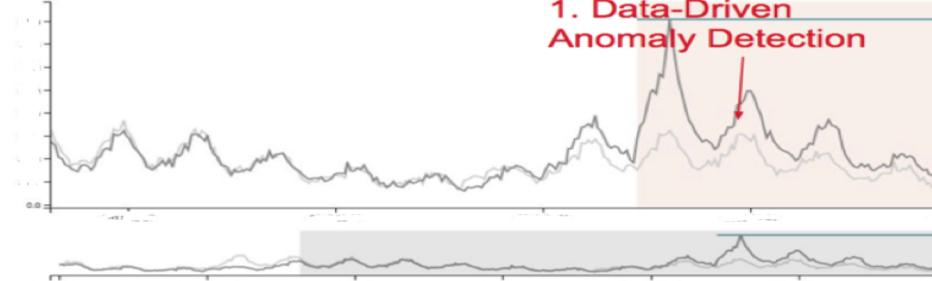
Is this an anomaly? ?

Yes (True Anomaly) Yes (But New Trend) No (False Alarm) To Be Determined

2. Auto-tune based on labels

Display Window: [2018, 1 pm] - [2018, 5 pm]
Granularity: 1_HOURS
Compare by: absolute

1. Data-Driven Anomaly Detection



3. Diagnosis on Metric, Dimension and Event

Metrics to Investigate: business_intraday_metrics_hourly_addit...
Filter by: Add a filter (Type to search)
+ add to chart

Metrics **Dimensions** **Events**

Card | Table



AI Academy Accelerated Democratization

Using LinkedIn tools and infra, train all engineers to use basic supervised ML in their day-to-day job

Partial Curriculum:

- AI 100 – Managing AI-driven Products
- AI 200 – AI For Software Engineers
- AI 300 – AI Engineering in Depth

Continuing education via LinkedIn Learning



LinkedIn
AI Academy



Adoption

Success Traps!

Support Trap

Democratizing AI

Back-fitting end-to-end machine
learning at LinkedIn *at scale!*



Productive Machine Learning

<https://engineering.linkedin.com/blog>