AI: The LinkedIn Way

Deepak Agarwal
VP of LinkedIn AI

September 19, 2018
Let people know you’re open

On

We are all seeking new professional opportunities

To Advance our Careers
LinkedIn operates the largest professional network on the Internet

Tell your Story

- 575M+ members
- 35K+ skills
- 25K+ titles
- 3.5B
- 1.5B
- 10M+ orgs
- 148 industries
- certificates, degrees, and more...
- 200+ countries
- states, cities, postal codes, ...
- roles, occupations
- speciality
- tools, products, technologies, ...
- 4B
Artificial Intelligence at LinkedIn

Every professional interaction with LinkedIn is personal, relevant and helps in providing opportunity.

Automatically and optimally deliver the right information to the right user at the right time through the right channel.

Accelerator of LinkedIn value propositions: engagement on the platform, customer ROI.
People

- Computer Scientists (optimization, NLP, computer vision)
- Infrastructure and System Engineers
- Statisticians
Principal Data Scientist
Microsoft - Bangalore, IN
Posted 3 days ago - 1,142 views

Jobs
- 147 applicants
- Full-time

Company
- 10,001 employees
- Computer Software

Connections
- 32 connections
- 100 company alumni

32 connections can refer you
Get referred to increase your chances of landing an interview.

Job description
We are currently building the next generation of cloud services for partner monetization, user acquisition, engagement & membership platform. These services have a huge global footprint of over 240 markets and process millions of transactions daily, with loads growing linearly as Microsoft moves to a "cloud first", "mobile first" strategy. The platform powers all of Microsoft's key services - Windows App Store, Windows Phone, XBOX, Bing Ads, Office 365, Microsoft Azure to name just a few. This endeavor offers big opportunities for data science and machine learning.

Seniority Level
Not Applicable

Industry
Computer, Hardware, Computer Software, Information Technology and Services

Learning
Courses related to Deepak's skills
- Data Science Foundations: Python Scientific Stack
  Viewers: 6,033
- Pandas for Data Science
  Viewers: 10,197
- Building and Deploying Deep Learning Applications with TensorFlow
  Viewers: 2,636

See my courses
See more courses
AI/ML process

Prod & Business understanding, Analytics, Data science

Product Metrics
- e.g., Active communities

Relevance Metrics
- e.g., consumption, production and diversity

ML Optimization Framework
- ML platform, automation, monitoring,...

A/B Testing Framework
- Scientific rigor, statistical methodology

Goals

User Inputs
- Needs
- Problems
- Time
- Money

Product Outputs
- Outcome
- Solution
- Results
- Value

PRODUCT
App 1: Member engagement

AI Platform
Contributors on LinkedIn

- John contributes by sharing

Messaging, commenting, publishing are other ways to contribute on LinkedIn.
Receiving Response to Contributions

John Smith

"The bot gets better and better every day. It's like a tougher version of us"

AI program beats humans in poker game - BBC News
Response to contributions leads to repeat contributions

(WSDM' 17; Guangde, Chen and Agarwal)
Contributions and Daily Active Uniques

- 1 unit increase in contributors leads to 0.4-0.7 increase in Daily Active Users- Network A/B testing (Eckles et al, Journal of Causal Inference, 2016)

- **Summary:** Increase in frequent contributors improves daily active uniques and engagement
Standard approach: Local optimization

PYMK

Feed

Notifications
Optimize Eco-system

Notifications -> Emails -> Feed

Right Distribution through right channel

Builder the right Network

Contributor -> PYMK, Follows -> Responder
Problem Formulation

- Create the right network that is value based, not volume based
  - Incorporate future downstream value in recommending new edges

- Optimize distribution of contributions to network across channels to encourage repeat contributions

- Guard against “high concentration” when distributing contributions
  - Better to have more contributors receiving some feedback instead of a few receiving a lot
Optimizing right network (PYMK)

Build connections that are likely to drive conversations:

\[ \text{score} = P(\text{connection}) + \alpha P(\text{conversation}) \]

- \textit{connection}: same school? same company? ...
- \textit{conversation}: messaging(private) or comment/like/share (public)
\[ P(\text{click}) + \alpha P(\text{viral}) E(\#DSI) + \beta P(\text{resp}) \exp(-\gamma E(\#\text{creator\ response})) \]

- \( DS\text{I} \) = downstream impact
- \( \text{viral} \) = share, like, comment
- \( \text{resp} \) = response to the creator via comment, like, re-share, messaging
- \( \gamma \) = temperature
- \( \alpha, \beta \) are weights

\[ pVisit = \begin{cases} 
\text{small} : & \text{notify viewer} \\
\text{big} : & \text{wait to show on feed}
\end{cases} \]
Optimize **distribution**

- **Alice**
  - pVisit = 0.05
  - DSI = 100
  - Bob's network

- **Bob**
  - pVisit = 0.05
  - DSI = 80
  - No notifications sent

- **Caitlin**
  - pVisit = 0.05
  - DSI = 1
  - Caitlin's network

- **Dave**
  - pVisit = 0.5
  - DSI = 1
  - Dave's network

- **Ed**
  - pVisit = 0.5
  - DSI = 10
  - Ed's network

- **Francis**
  - pVisit = 0.6
  - DSI = 100
  - Francis's network

---

**Notifications sent**

- Alice
- Bob

**No notifications sent**

- Caitlin
- Dave
- Ed
- Francis
What Kind of Model?

A combination of deep learning and trees models and real time “corrections” at fine granularities
Features in ML model: Knowledge Graph, Learned Representation

Information extraction
Human curation
Deep Learning
Challenges

- Jointly train a combination of different classes of models
- Automatically determine features, model structure & hyperparameters
Challenges

Manage data/modeling pipelines with complex dependencies

Automatically detect issues and gracefully degrade model performance

![Diagram showing data processing pipelines](image)

1. Data-Driven Anomaly Detection
2. Auto-tune based on labels
Challenges: 1) deploy large models to different places in production 2) manage complex model/data deployment processes
App 1: Member engagement

AI Platform
Pro-ML: Productive Machine Learning

We make the end-to-end process of running and iterating on large ML workflows easy, robust and almost automated.
Pro-ML Key Components

- Feature Marketplace
- Pro-ML Jupyter Notebook
- Model Creation Service (Photon-connect)
- ML Algorithms
  - Large-scale deep learning Trees (XGBoost)
  - Personalization (GLMix)
- Model Repo
- Model Deployer & Release Tool
- Debug & Explain Tool
- Production
- Health Metric Monitoring
Frame: Feature Access Abstraction Layer

Replaces complex feature preparation workflows with a simple, config-driven feature specification that makes features available for training and scoring in offline and online environments with automated consistency monitoring.

targetFeatures: 
{
  key: targetId
  featureList: [
    waterloo_job_companyDesc
    waterloo_job_companySize
    ...
  ]
}

sourceFeatures: 
{
  key: sourceId
  featureList: [
    jfu_member_geoCountry
    jfu_member_geoRegion
    ...
  ]
}

Quasar: Unified Inference Engine @ LinkedIn

Provides a DSL to define the scoring and ranking computation logic of a model, and a high-performance execution engine to run the model.
Democratize ML at LinkedIn

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

**Curriculum:**
- AI Boot Camp: 5 full days (aimed at ICs)
- One-day course: Aimed at managers

Continuing education via LinkedIn Learning
Conversational AI: seq2seq model

Coming the future:
Question and Answering
Help center chatbot
More Methodological Research

Richer class of non-parametric models for personalization

Transfer learning

Active learning to collect human labels more effectively

Large scale & real-time Graph computations

Fairness, Transparency and Privacy

Efficient network A/B testing

......
Summary

- ML/AI is “oxygen” of LinkedIn product
- Converting ill-posed business/product problem to well posed ML problem requires effort, rigor and collaboration
- Scaling ML process for large workflows challenging both from infra and organizational perspective
Appendix
Quasar Components

Model Development

model.quasar

MODEL ID "..."
import com.linkedin.quasar
...
DOCUMENT FEATURE ...
...

Model Execution

Quasar DSL Parser

Quasar Execution Engine

Quasar Executor (per thread)

Operators

Doc 1 Features

Doc 2 Features

Doc 3 Features

Feature Transformers

consume finishUp

request parameters
model parameters

Doc 1

Features

Doc 2

Features

Doc 3

Features

QuasarModel

share

consume

finishUp

(request parameters)

(model parameters)
Trains a DAG of different classes of ML models specified using Quasar DSL with features specified using Frame config
Applied Research

1. Focused on LinkedIn Ecosystem - create member delight
2. Accelerator of Member Value
3. Centrally Managed but Deeply Aligned within Product Development
Knowledge Marketplace

Mission
Create a knowledge marketplace where any member and enterprise customer can ask a question in natural language and we provide an answer almost instantaneously, eliciting more information if necessary.

Hey, we found you’re skilled in {#skills}id, {/skills} technologies and hence you seem to be a good match for a role at {company}

Enrich conversations

Surface premium value

Data access
# ML Infrastructure Ecosystem

<table>
<thead>
<tr>
<th>Exploration</th>
<th>Data Prep + Tracking</th>
<th>Creation</th>
<th>Delivery</th>
<th>Inference</th>
<th>Experimentation</th>
<th>Maintenance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jupyter Notebooks</td>
<td>Tracking Data Collection (Kafka)</td>
<td>Feature Data Joins (Frame)</td>
<td>Nearline Feature Processing</td>
<td>Entity Feature Store</td>
<td>A/B Experimentation (TREX)</td>
<td>Monitoring</td>
</tr>
<tr>
<td>Spark</td>
<td>Label Collection</td>
<td>Training (Photon ML on Spark)</td>
<td>Feature Sharing (Frame)</td>
<td>Dynamic Feature Store (Pinot)</td>
<td>Analytics (Raptor)</td>
<td>Anomaly Detection (ThirdEye)</td>
</tr>
<tr>
<td>Query Explain / Debug (SLoQ)</td>
<td>Xgboost</td>
<td>Grid to Prod K/V push (Venice)</td>
<td>Inference Engine (Quasar)</td>
<td></td>
<td></td>
<td>Root Cause Analysis</td>
</tr>
<tr>
<td>Feed-admin tool (Debug)</td>
<td>Tensorflow</td>
<td>Grid to Prod Index push</td>
<td>L3 Blending Layer (ReMix)</td>
<td></td>
<td></td>
<td>Process Automation</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>L1/L2 Indices: Graph, Search, Feed</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
LinkedIn operates the largest professional network on the Internet

Tell your Story

Satya Nadella • 2nd
CEO at Microsoft
Greater Seattle Area

View in Recruiter • More...

Current CEO at Microsoft, where our mission is to empower every person and every organization on the planet to achieve more. Joined Microsoft in 1992, and have held a variety of roles, most recently executive vice president of Microsoft’s Cloud and Enterprise...

Show more ▼

Highlights
3 Mutual Connections
You and Satya both know Michael Narayan, Kevin Scott, and 1 other

Satya’s Articles & Activity
4,864,568 followers

My views on U.S. immigration policy
Satya shared this
It was a pleasure to sit down with Microsoft CEO Satya... Satya liked
I always appreciate the opportunity to learn from... Satya shared this

See all articles
See all activity

Experience
CEO Microsoft
Feb 2016 – Present • 4 yrs 5 mos
Greater Seattle Area

560M+ members

3.5B
35K+ skills

4B
speciality
tools, products, technologies, ...

25K+ titles

500+ connections

2.5B
titles

200+
countries

1.5B

states, cities, postal codes, ...

10M+ orgs

148 industries
certificates, degrees, and more ...
LinkedIn operates the largest professional network on the Internet.
LinkedIn operates the largest professional network on the Internet

Explore your Opportunities

300M+ jobs  8K+ courses

?B course learning

?B job applies

?B views

130K+ weekly articles

560M+ members
Network embedding

- Use the embedding:
  - connect with other data, e.g., natural language
  - ML model features
Challenge: Complex Model Deployment Process

**Online Data**
- UI
- Scoring & Ranking
- Parameter Store
- User Features
- Item Index
- Live Index Updater

**Offline Data**
- Offline Index Builder
- Model Training
- ETL
- Offline Data

**Database Connections**
- User DB
- Item DB

**Feature Pipelines**
- User
- Item

**Weekly Processes**
- Offline Index Builder

**Data Flows**
- Recommendations
- Streaming tracking
- Click, impression
Pro-ML Key Components

Feature Marketplace

Pro-ML Jupyter Notebook

ML Algorithms
Large-scale deep learning
Trees (XGBoost)
Personalization (GLMix)

Photon-Connect

Model Repo

Model Deployer and CRT

Triton (Debug and Explain)

Frame

Quasar

Production

Health Metric Monitoring
Provide

- A variety of common-used features
- A mechanism to share features across applications
- A feature data access abstraction layer

Enable leverage of features and simplified configuration of complex feature access patterns
Provide

- A model repo to record all ML models
- A deployer to deploy ML models and the associated data to different destinations in the production environment
- A central release tool to manage the deployment process

Enable ML models and the associated data to be first-class citizens in LinkedIn deployment tooling
Provide
- Monitoring of health metrics of ML models, feature data and workflows
- Mechanisms for issue detection and model graceful degradation
- Tools to debug and explain ML models

Enable automatic detection of issues and reduce developers’ time on issue investigation
LinkedIn operates the largest professional network on the Internet

Explore your Opportunities

- 300M+ jobs
- 10K+ courses in 5 languages
- 9M weekly interactions
- 18M monthly job applies
- 581M weekly views
- 600M+ members
- 581M weekly views
- 30B
- feed updates (130K+ weekly articles, ...)

Jobs you may be interested in

- Principal Software Engineering Manager
  Microsoft
  Mountain View, CA, US

- Director Of Engineering
  Gliffy
  San Francisco Bay Area

- Senior Director of Engineering, Search
  Move, Inc
  Santa Clara, CA, US

- Director of Engineering (Platform)