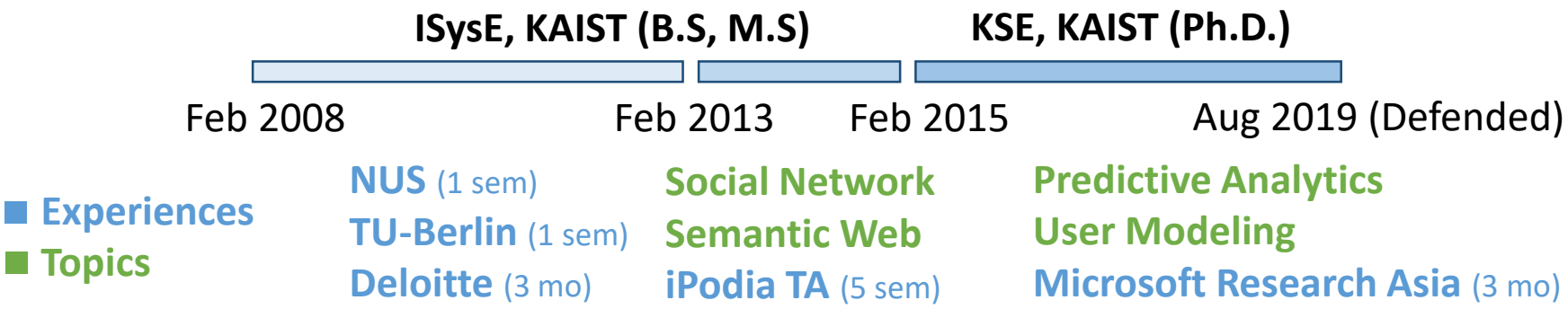
# What I have done so far
# &
# What I am interested in

**Sundong Kim, KAIST**

http://seondong.github.io

July 12th, 2019

# About Me




- **Ph.D. Candidate @ Data Mining Lab, KAIST (Prof. Jae-Gil Lee)**
- Open-minded, Opinionative, Optimist, Optimizer
- **Career goal:**
  - Willing to work **in a team** with **strong engineering backgrounds**
  - Willing to work **in a project** which contributes to **further customer satisfaction**
  - Willing to work as an Applied Scientist / Data Scientist / Research Scientist where we can **develop things by getting constant feedback** from others

# Have been doing / Would like to do

- **What I have been doing recently:**
  - **Predictive Analytics** (in Longitudinal Setup, Imbalanced Dataset)
  - **User Modeling** (with Deep Learning, NLP techniques)
  - **Communication** (Research collaboration & led small groups)
  - **Python Programming** (numpy, pandas, sklearn, tensorflow2.0)

- **What I am interested in:**
  - **AutoML in Relational Data**: To be **free** from engineering efforts
  - **Competitive Data Science:** To develop **light and effective** model
  - **Decision Making:** To learn from the causal effect by **A/B test**
  - **Facilitating Teamwork:** To make **our group better** than before

# R1) Revisit Prediction (My PhD Thesis)

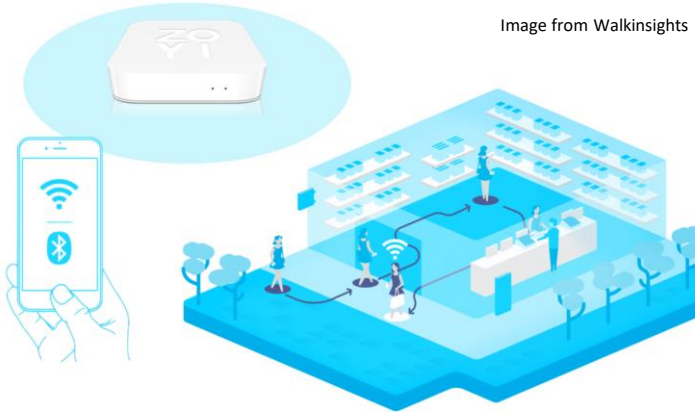


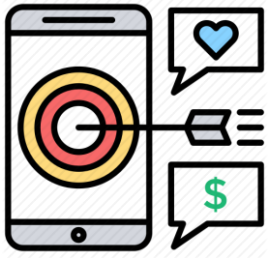
Image from Walkinsights

⇨ **Revisit in 60 days**

**"Revisit Prediction for Targeted Marketing"**

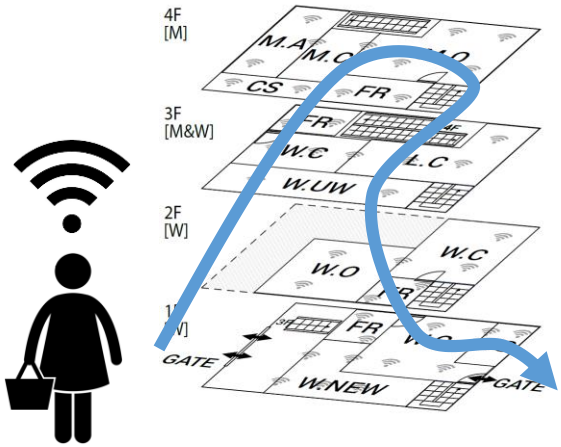
**Potential to revisit?**

- High → Invite to premium brands (Upselling)
- Low → Discount or Mass sale



# For R1) In-Store Sensors Data in Offline Stores

- 7 Flagship stores
- 110K-2M visits/store
- 220-990 days collected
- Average number of areas = 6.56




| Shop ID | A_GN | A_MD | E_GN | E_SC | L_GA | L_MD | O_MD |
|---|---|---|---|---|---|---|---|
| Location | Seoul, Korea | | | | | | |
| Length (days) | 222 | 220 | 300 | 373 | 990 | 747 | 698 |
| # sensors | 16 | 27 | 40 | 22 | 14 | 11 | 27 |
| Data size | 15GB | 77GB | 148GB | 99GB | 164GB | 242GB | 567GB |
| # visits > 60s | 0.11M | 0.33M | 0.18M | 0.27M | 1.06M | 1.72M | 2.01M |
| Revisit rate | 11.73% | 31.99% | 21.18% | 36.55% | 21.22% | 32.98% | 48.73% |

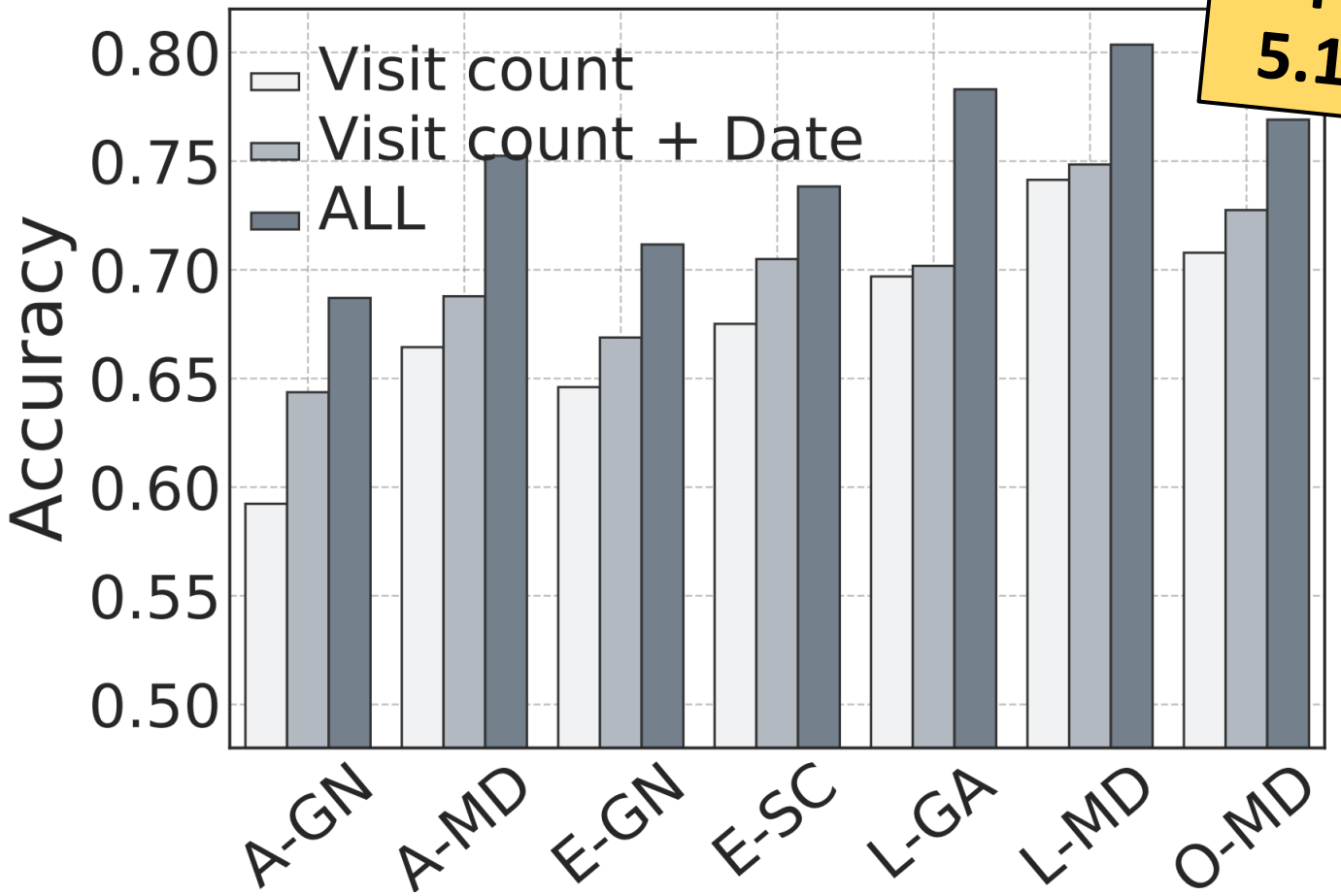
# R1-1) Feature Engineering Model

- Overall statistics
- Travel distance/speed/acceleration
- Area preference
- Entrance and exit pattern
- Heuristics
- Statistics of each area
- Store accessibility
- Group movement
- Time of visit
- Upcoming events

Motion pattern

Temporal Information

# R1-1) Effectiveness of Our Features

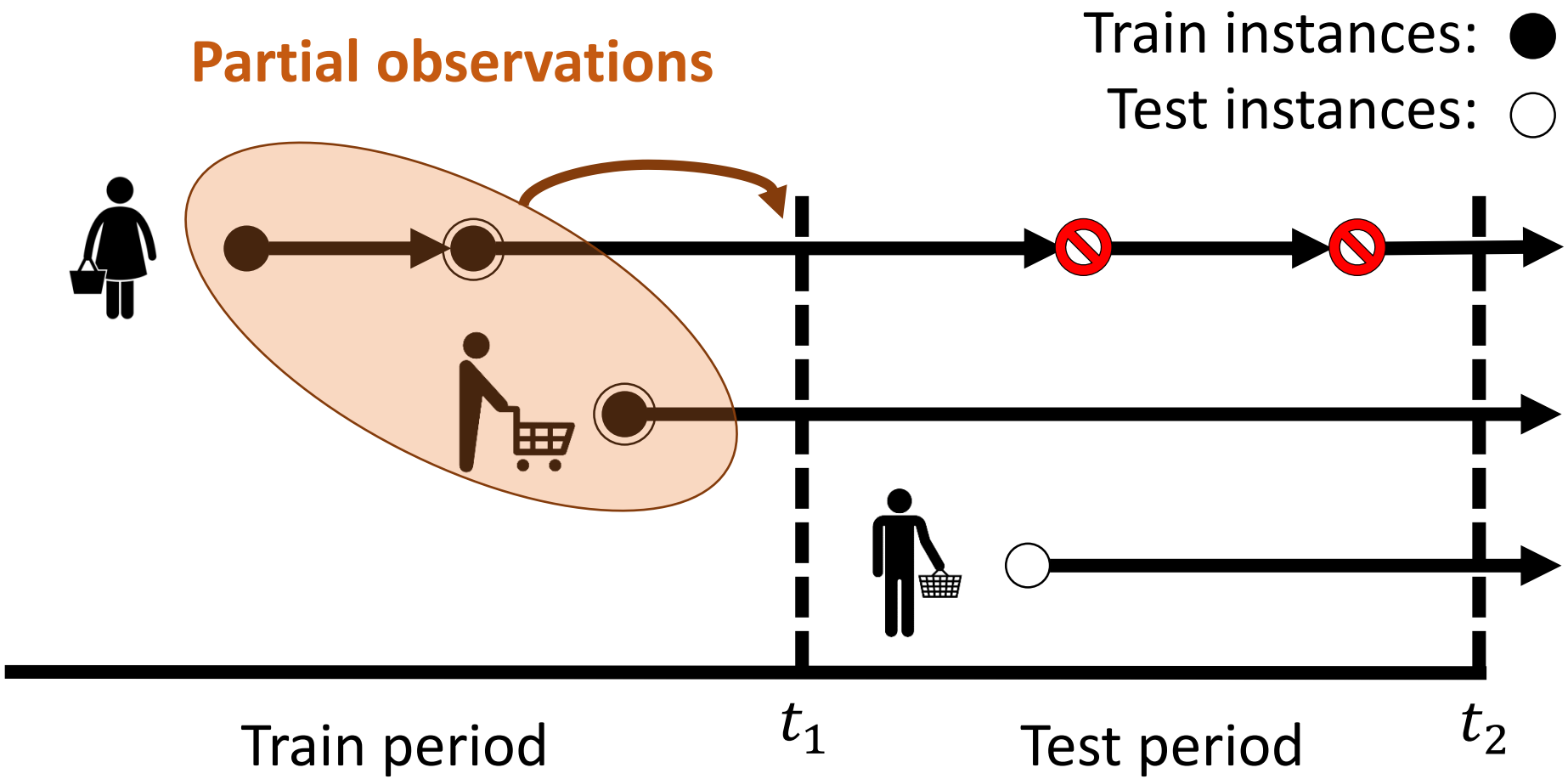



**Improved by 5.1-12.4%.**

***XGBoost***
***LightGBM***

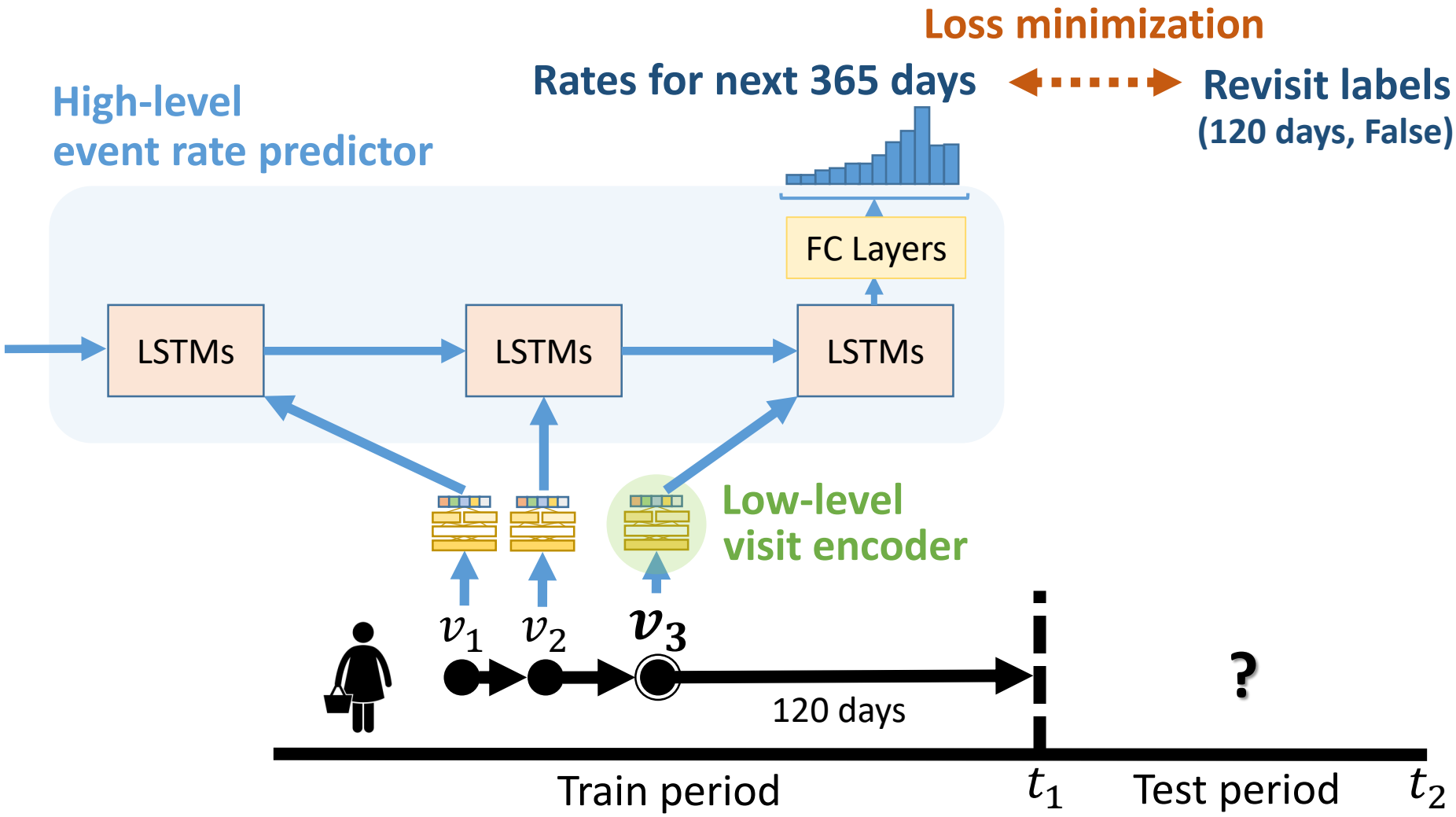
[ICDM'18, KAIS Journal]

# R1-2) How to Use Partial Observations

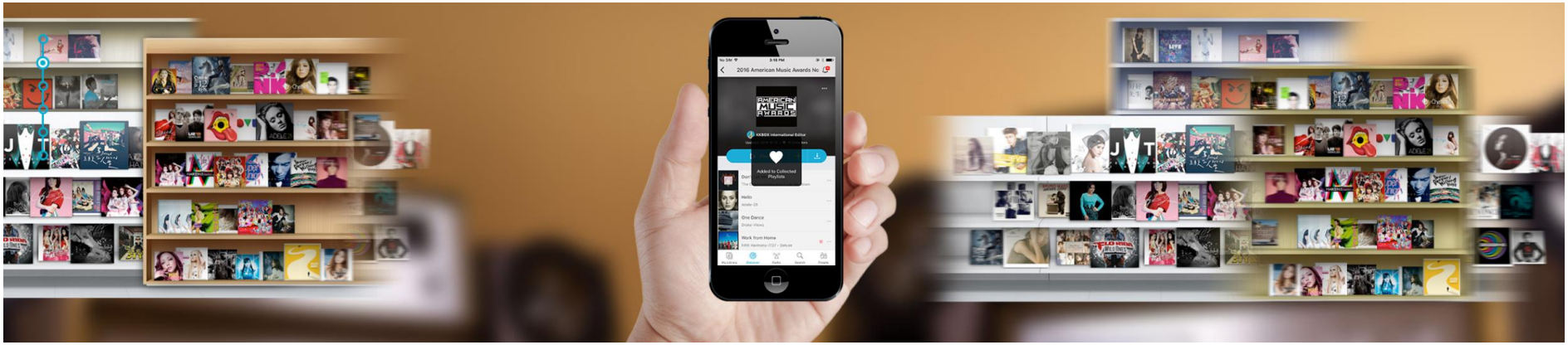


In preparation

# R1-2) Solution: Deep Survival Analysis



# R2) Churn Prediction

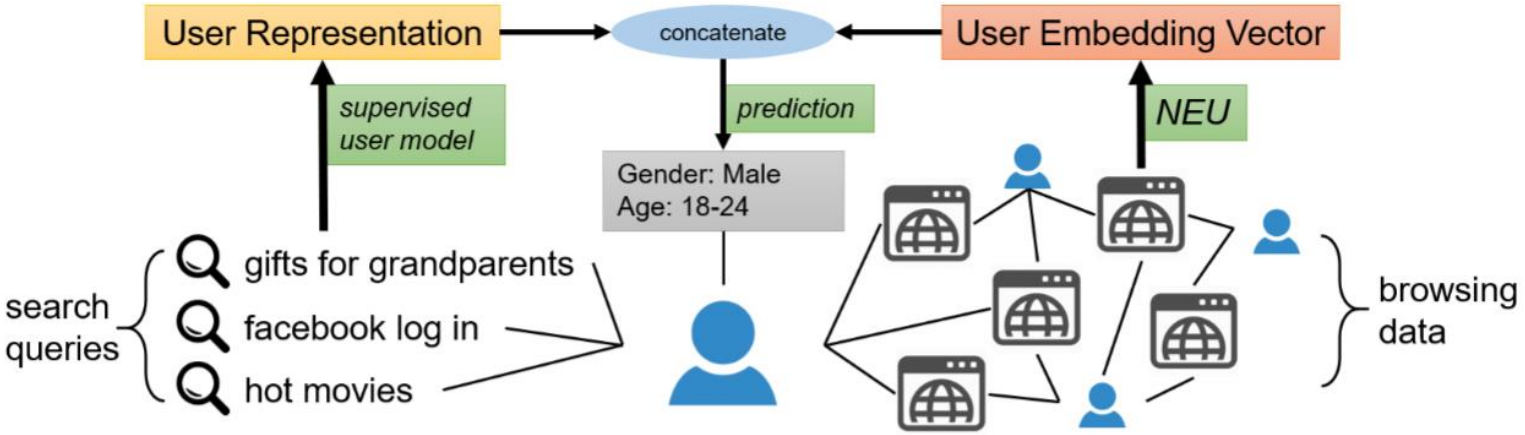


- Participated to improve predictive analytics skills using public data
- Played with online music subscription data
- Formed a team through online
- Achieved Top-10 position (over 575 teams, unofficial)

**Learned: Temporal Feature Engineering, Model Ensemble, Apache Spark, Competitive Data Science**

With M. An, in MSRA, Submitted

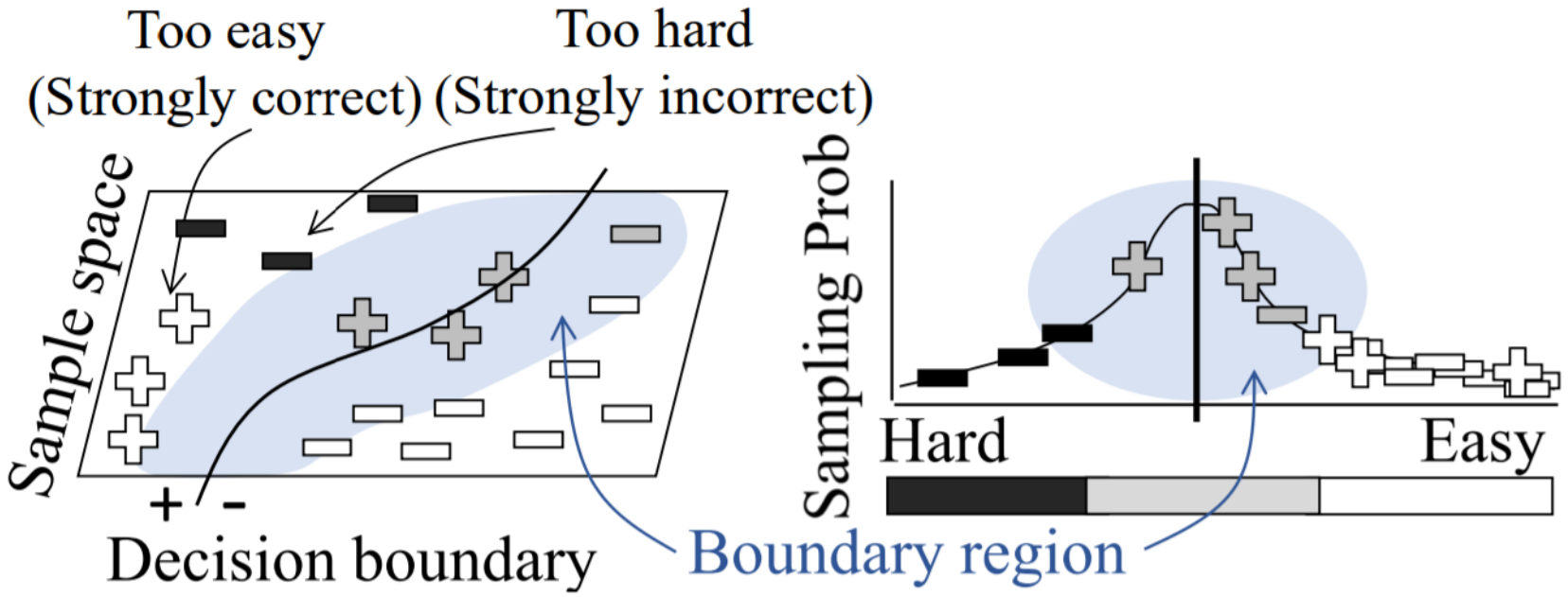
# R3) User Embedding for Profile Prediction

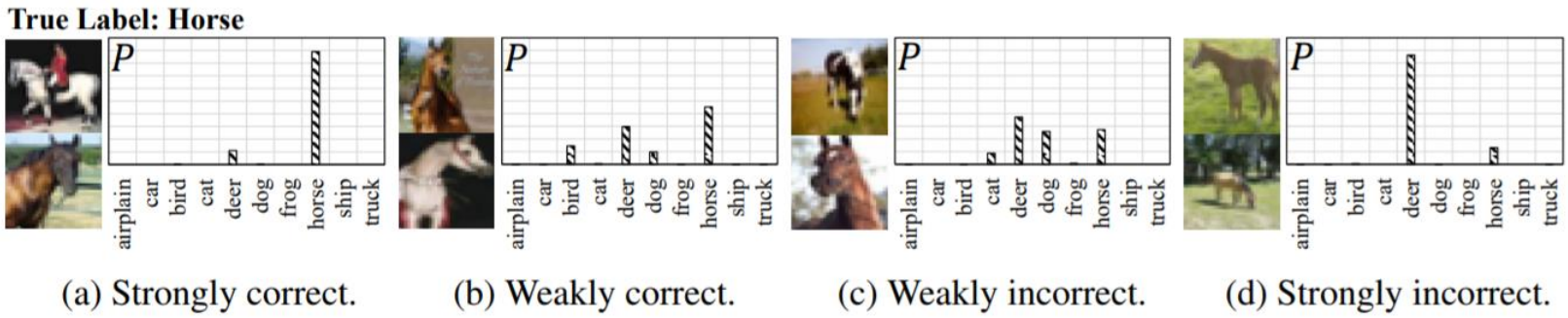




With H. Song, Submitted

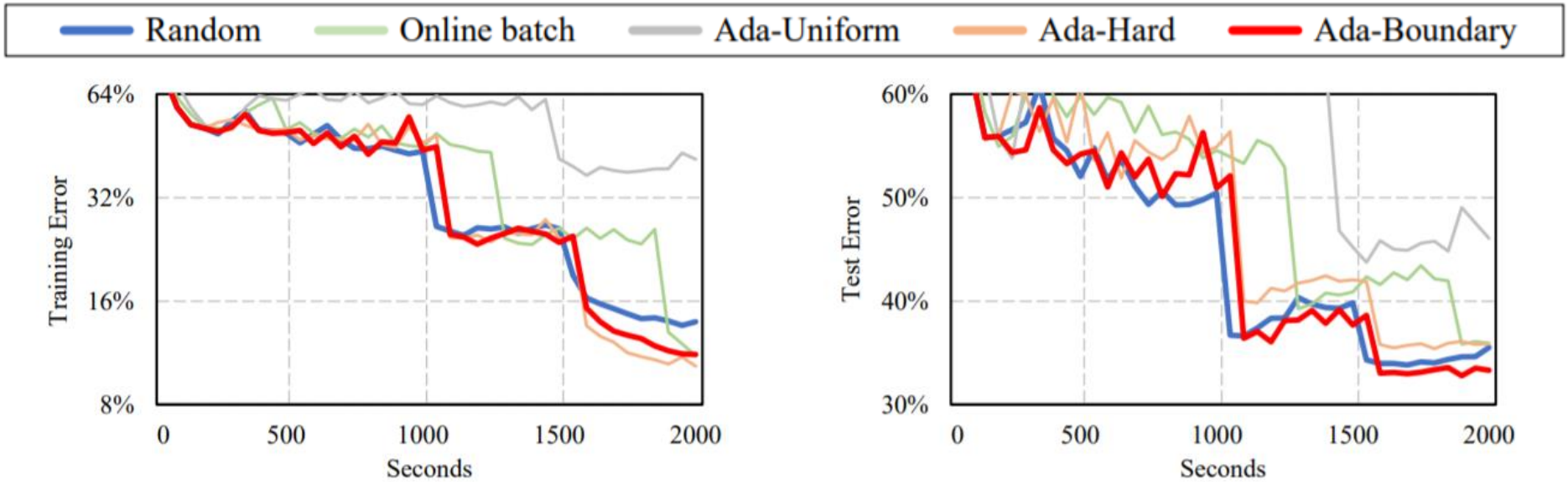
# R4) Batch Selection for Faster DNN Training




(a) Strongly correct. (b) Weakly correct. (c) Weakly incorrect. (d) Strongly incorrect.

# R4) Batch Selection for Faster DNN Training



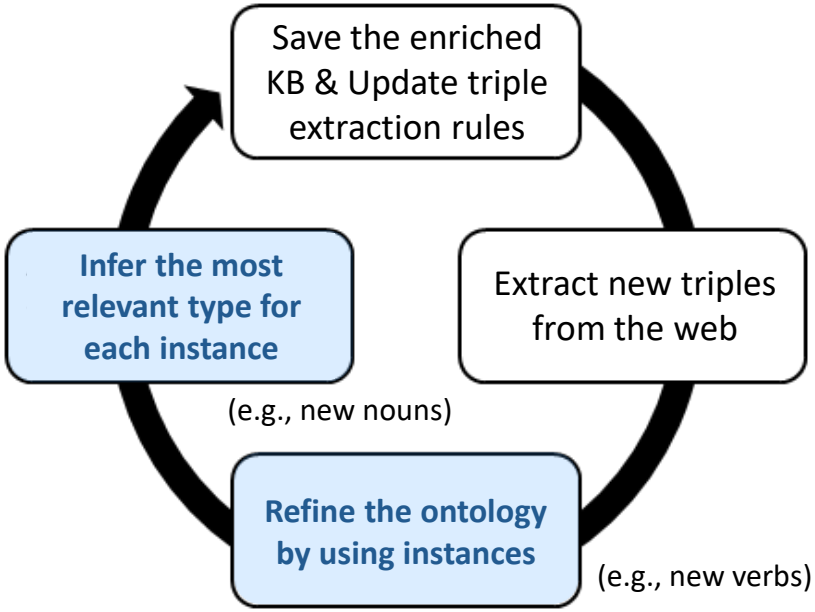

CIFAR-100 Training Error.



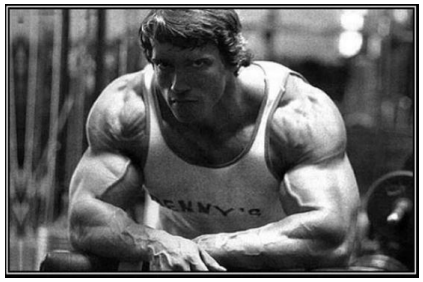

CIFAR-100 Test Error.

KIPS'14, Deployed in WiseKB

# R5) Knowledge Base (KB) Bootstrapping

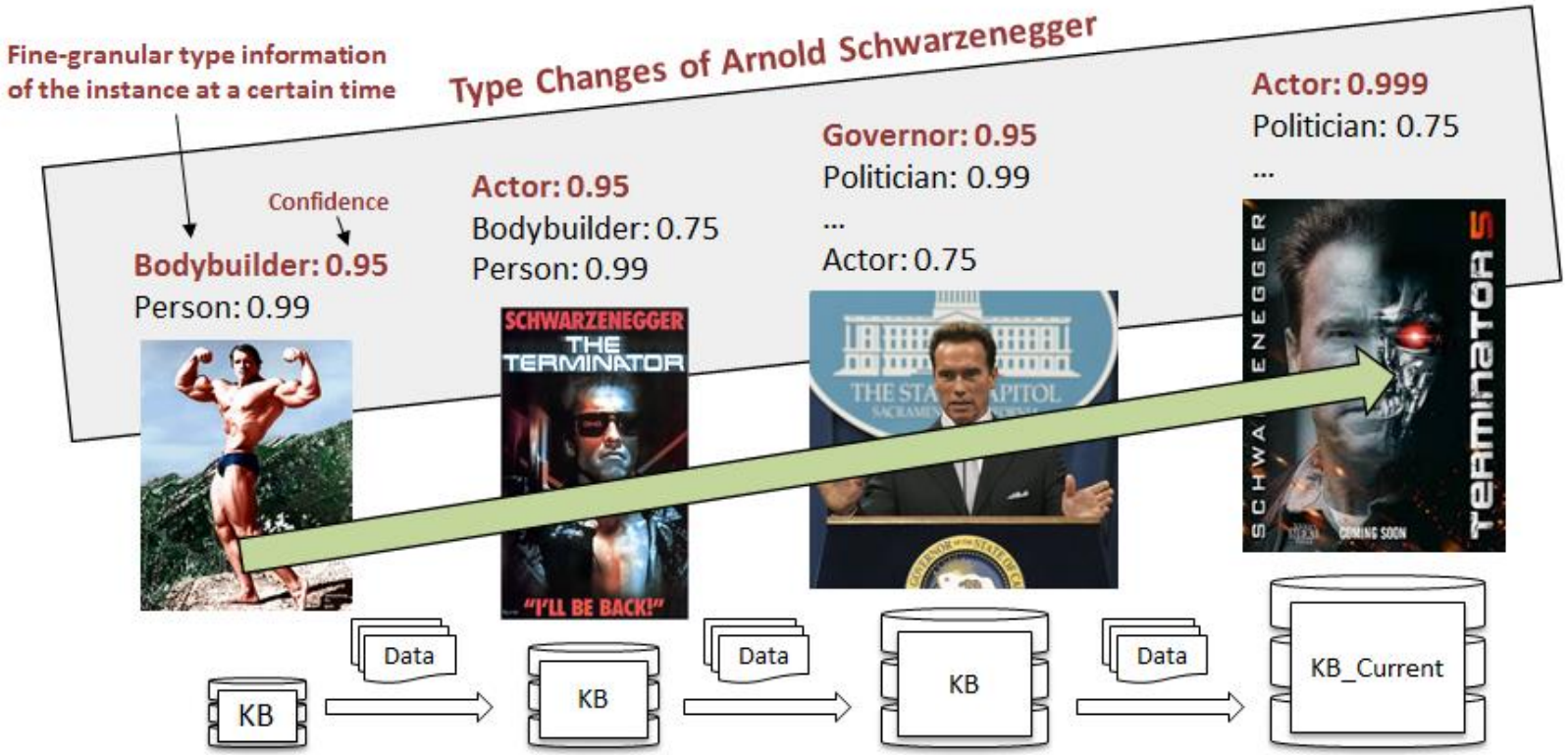

Bodybuilder



Actor

Governor

# R5) Knowledge Base (KB) Bootstrapping




**Learned: Java, Semantic Web, SPARQL, Jena, REST API**

**Joined the project before deep learning is widely applied.**

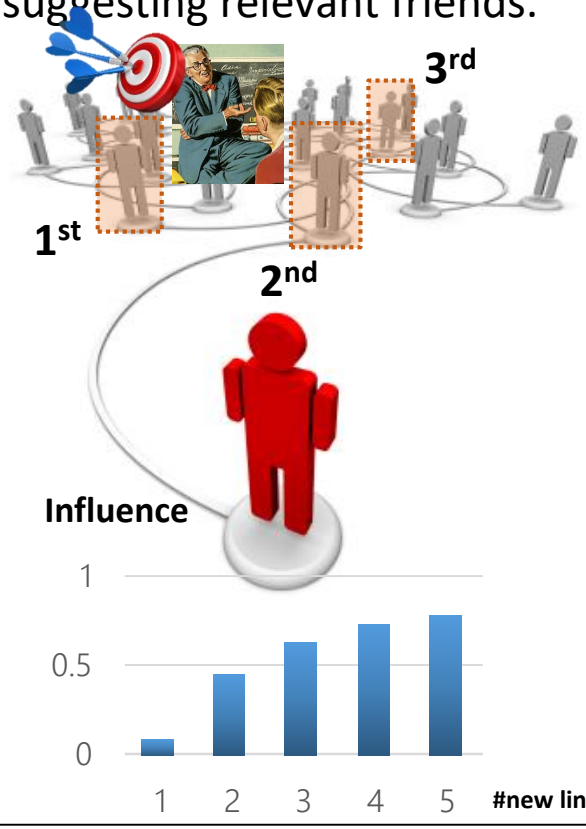
# R6) Friend Recommendation With a Target

One of the most famous professors in data mining

**Asymmetric relationship**

**Application**

SNS administrator can help agents to get more attention by suggesting relevant friends.




**Field:** Network Science, Influence Maximization

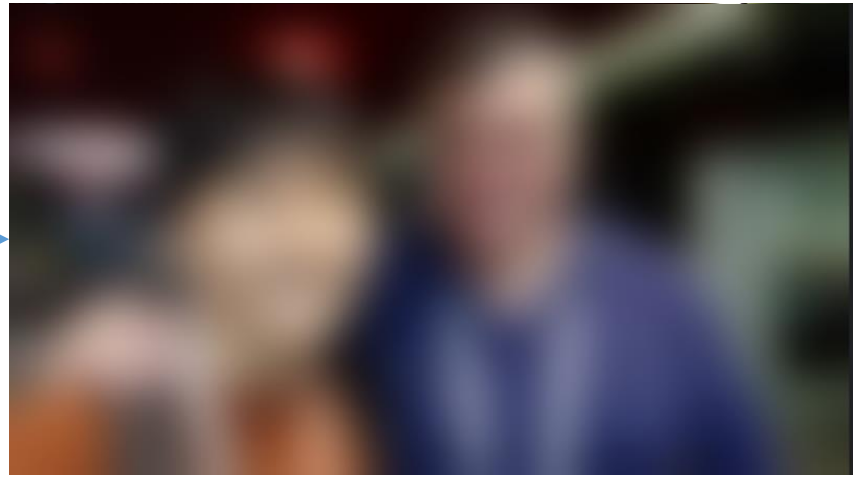
**Proposed Concept:** Influence & Reluctance

**Known Measure:** Katz Centrality

**Solution:** Monte-Carlo Simulation

# R6) Friend Recommendation With a Target

**After few years ...**
**(Read hundreds of papers,**
**Get to know dozens of peoples,**
**Attend several conferences)**



**Original intention (2015):** Maximize my influence to him through new friends → I can reach out him without worrying about getting rejected.

**Actual result (2018):** His influence is maximized through my new friends → As time goes, my interest in him has grown more and more.