

Machine Learning for IT company (Matrix completion & Active Learning)

Makoto Yamada

myamada@i.Kyoto-u.ac.jp



Types of Data

- Click data
 - News (Yahoo, Google)
 - E-commerce (Yahoo Auction, eBay, Alibaba, Amazon)
 - Ad recommendation (Google, Criteo, Cyber Agent)
 - Video sharing (Youtube, ニコニコ, SnapChat)
 - Image sharing (Facebook, Flickr, Picasa)
- Text data (web page, item description, etc.)



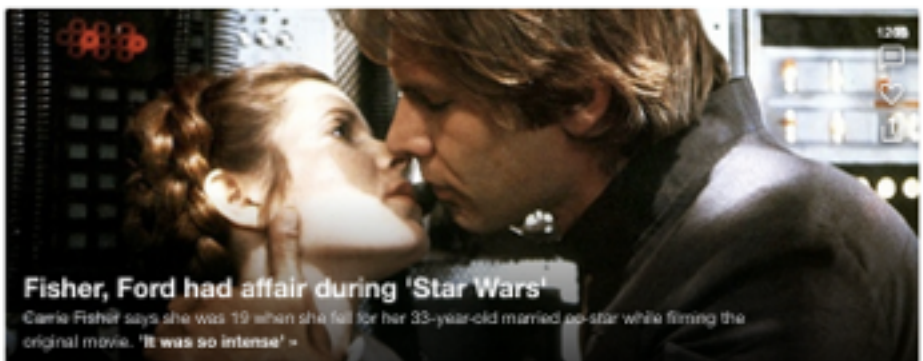
GBDT

- Mail
- News
- Finance
- Sports
- Fantasy Football
- Politics
- Celebrity
- View
- TV
- Movies
- Style
- Beauty
- Shopping
- Tech
- More on Yahoo

2016 Big Data Trends

Top 8 Trends in Big Data for 2016. Get the Whitepaper!

Logistic Regression Factorization Machine



- Trending Now**
1. Beastie Boys
 2. Kim Kardashian
 3. Tiffany Trump
 4. Blake Lively
 5. Dolly Parton
 6. Luxury SUV Deals
 7. Amber Heard
 8. Rheumatoid Arthritis...
 9. Kareena Kapoor
 10. 2016 Cars

Oxford Dictionaries' word of the year

New Balance shoes caught in controversy

Report: 3 NSA teams boycotting Trump hotels

'Catch' star Jake Harris has cracked skull

YAHOO! MAIL

One inbox. Every email.

Yahoo, Gmail, Hotmail, and AOL together at last.

Get the mail app

Science

Stephen Hawking Puts An Expiry Date On Humanity

The physicist says that humanity has a limited time to make progress before it's too late. "Earth is more a wasteland than a home."

By @STimes

Sponsored

The Simple Tools That Made Me Millions

This rare technique is making Japanese investors a fortune in 2016...

France Deep Dive

Politics

Megyn Kelly Believes President Donald Trump Could Be 'Dangerous'

Megyn Kelly is speaking up about her controversial relationship with Donald...

Factorization Machine, GBDT Etc.

IBIS 2016|



- ibis 2016
- ibis 2016 **ripley**
- ibis 2016 **conference**
- ibis 2016 **las vegas**
- ripley** ibis 2016

Query auto completion
 GBDT (Decision Tree)
 CRF

IBIS 2016. The premier BusinessObjects training and information exchange of the year hosted by InfoSol.

IBIS Academy 2016 - HOME

www.ibisacademy.com v

IBIS Academy Attendees; IBIS Conference 2016; History of the IBIS Academy; Dates to Remember. Dates to Remember. 23 Oct Registration Opens: Partners.

Web ranking
 GBDT (Decision Tree)

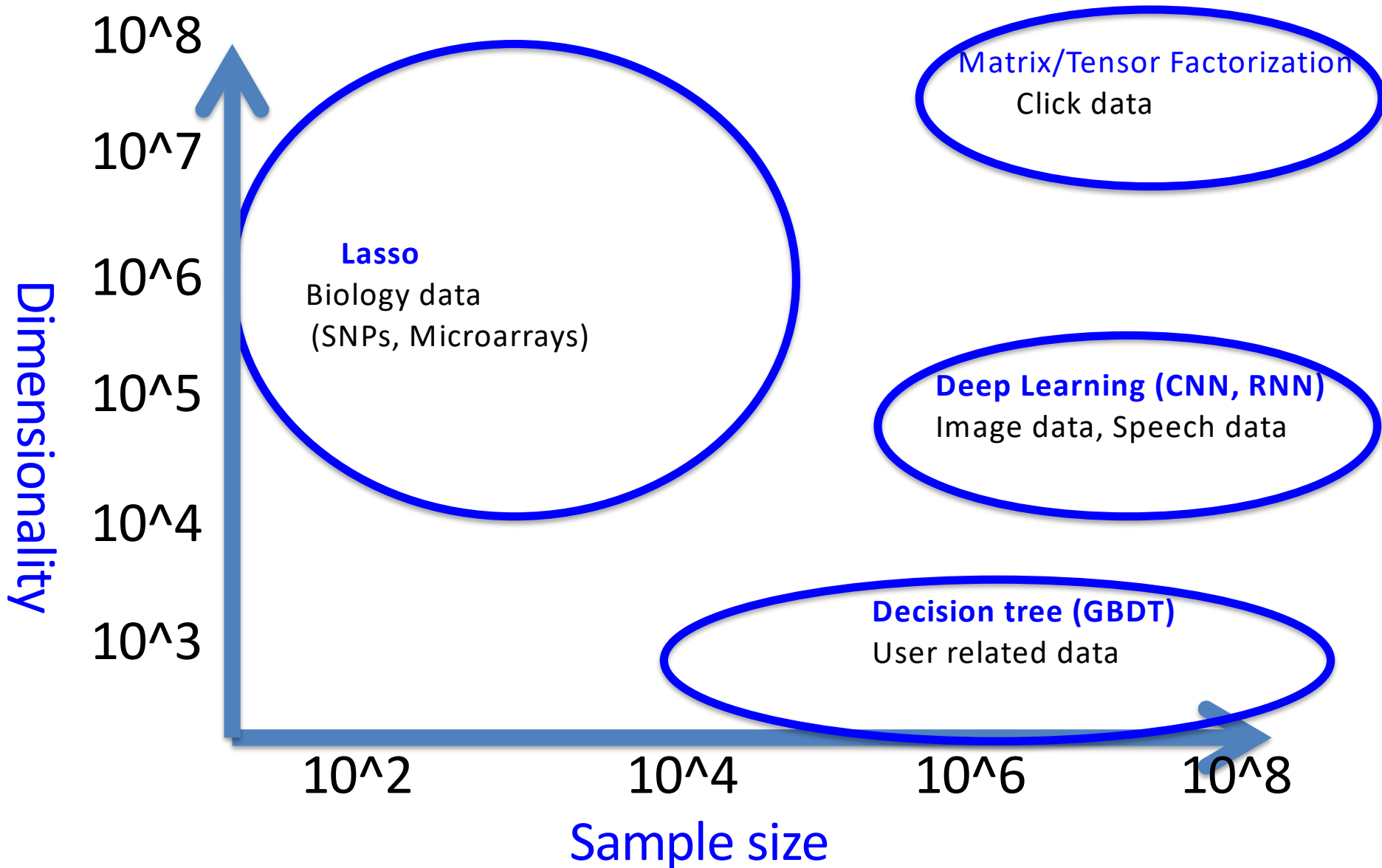
Ibis Demo Tour 2016 | Buy/Try

www.ibiscycles.com/buytry/ibis_demo_tour v

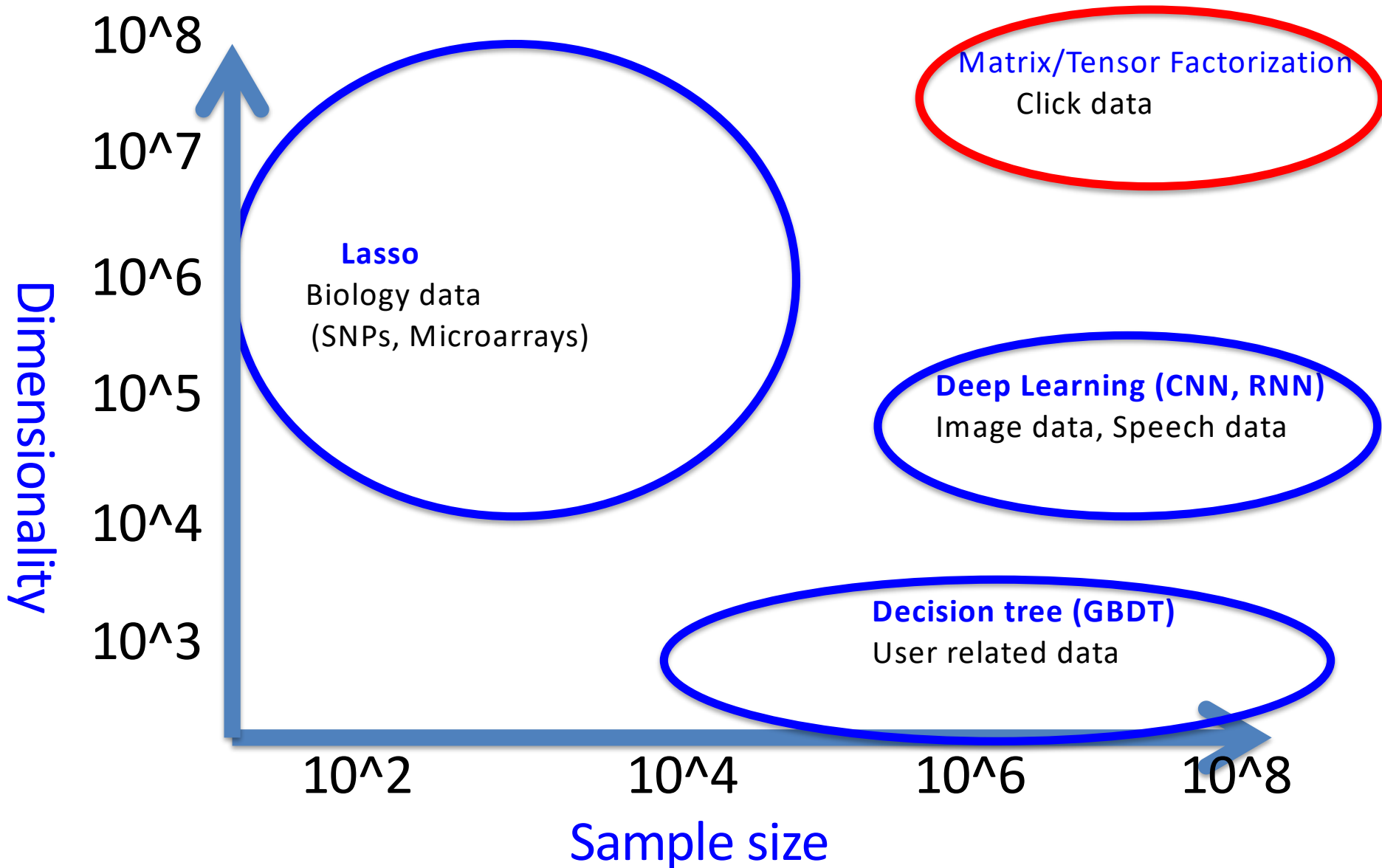
Ibis Demo Tour 2016 | Buy/Try. Ibis Cycles Inc. designs, develops, sources, distributes, and markets the best bicycles and cycling related products in the world.

IBIS 2016 - Image Results

Data & Related methods

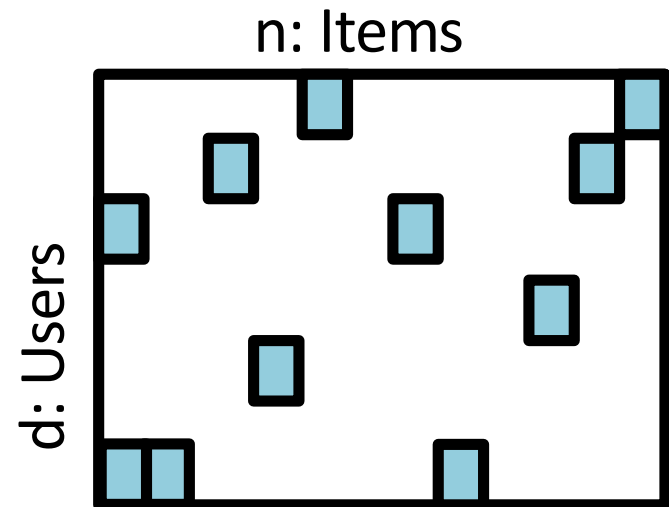


Data & Related methods



High-dimensional Sparse data

- Dimensionality and sample size are both large!
 - Text data
 - Click data
 - Link data



Text classification, Sentiment analysis: Logistic regression

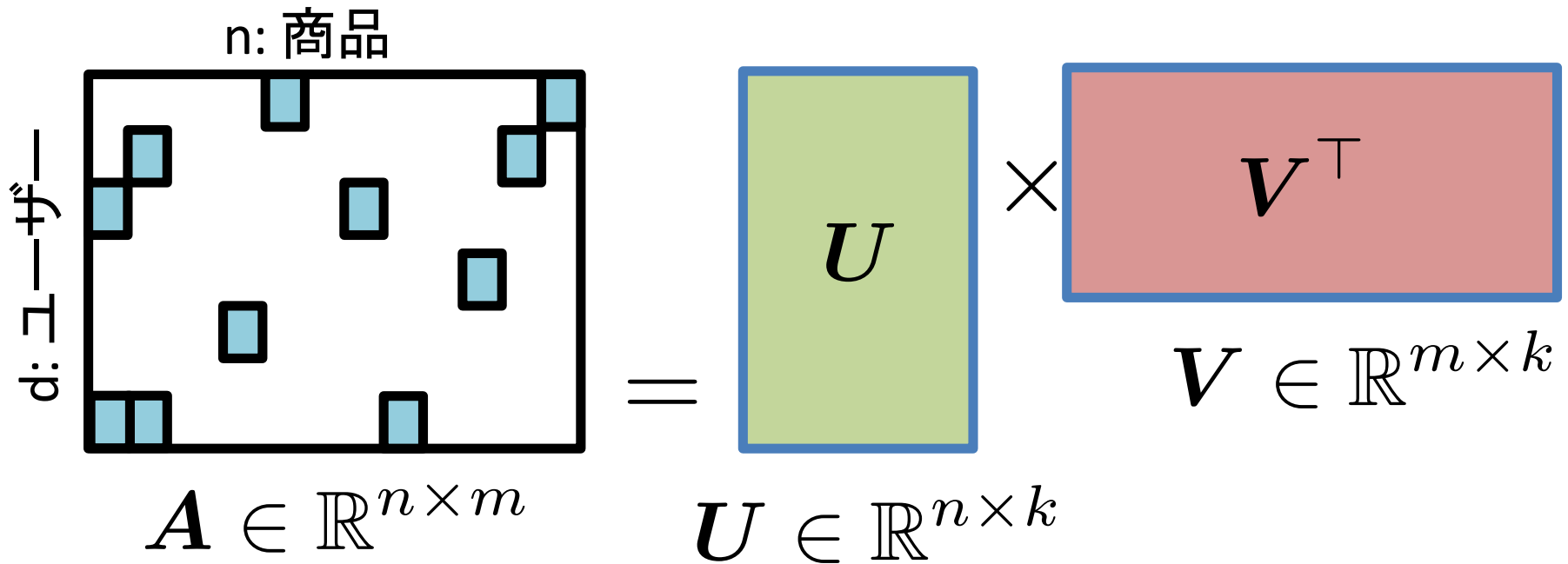
Ad recommendation, item recommendation: Matrix completion

Example of data format

- Movielens data (1M)
- 1::1193::5::978300760
- 1::661::3::978302109
- 1::914::3::978301968
- 1::3408::4::978300275
- 1::2355::5::978824291
- 1::1197::3::978302268
- 1::1287::5::978302039
- 1::2804::5::978300719
- 1::594::4::978302268
- 1::919::4::978301368
- 1::595::5::978824268

Collaborative filtering

- Recommending items using user click information (Amazon, Netflix, etc.)



A is a sparse matrix

Singular Value Decomposition

- Fill un-observed element by 0 and do SVD.

$$A = U \Sigma V^T$$

- Use low-rankness to estimate un-observed elements

$$\hat{A} = U_k \Sigma_k V_k^T$$

- This approach makes un-observed elements as 0. However, those elements are simply not observed (not zero!)

Alternating Least Squares (ALS)

- Fitting only observed entries.

$$\min_{\mathbf{U}, \mathbf{V}} \sum_{(i,j) \in \Omega} (a_{ij} - \mathbf{u}_i^\top \mathbf{v}_j)^2 + \lambda_1 \|\mathbf{U}\|_F^2 + \lambda_2 \|\mathbf{V}\|_F^2$$

- \mathbf{U} and \mathbf{V} can be alternately optimized.

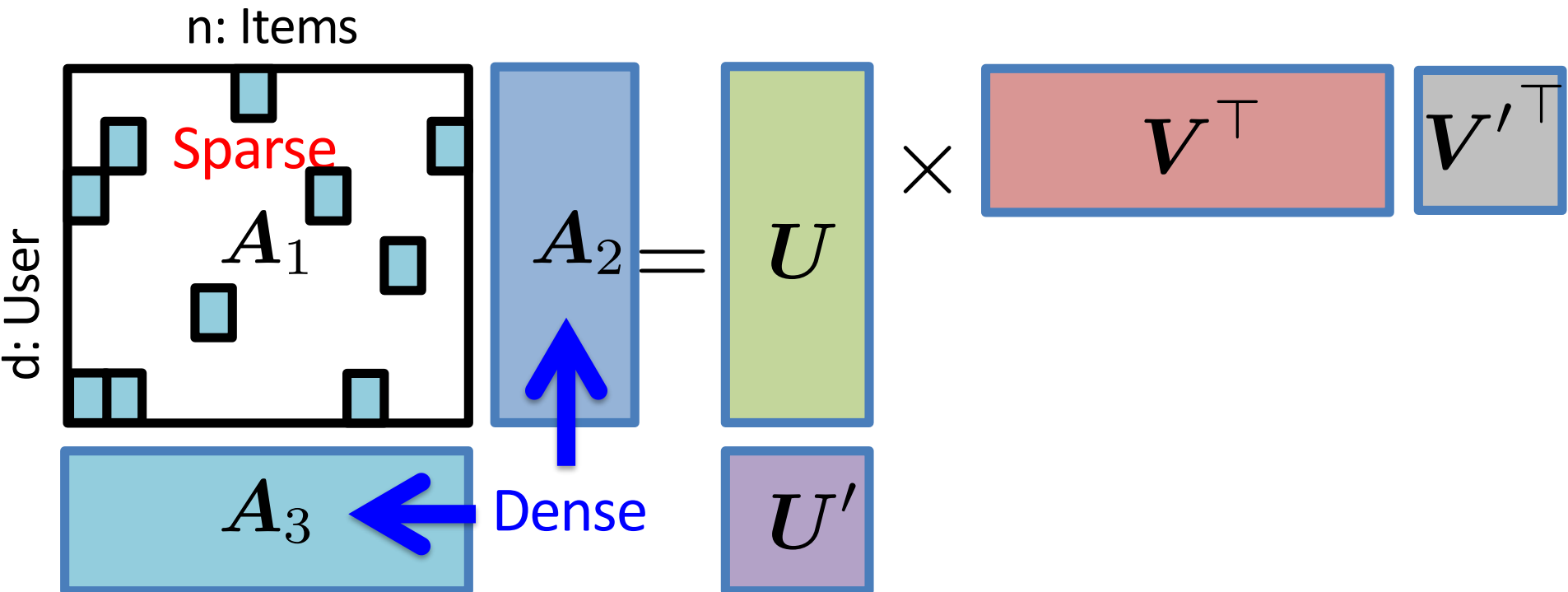
$$\mathbf{u}_i = \left(\sum_{(i,j) \in \Omega} \mathbf{v}_j \mathbf{v}_j^\top + \lambda_1 \mathbf{I} \right)^{-1} \sum_{(i,j) \in \Omega} a_{ij} \mathbf{v}_j$$

$$\mathbf{v}_j = \left(\sum_{(i,j) \in \Omega} \mathbf{u}_i \mathbf{u}_i^\top + \lambda_2 \mathbf{I} \right)^{-1} \sum_{(i,j) \in \Omega} a_{ij} \mathbf{u}_i$$

Advanced topic: Cold start problems

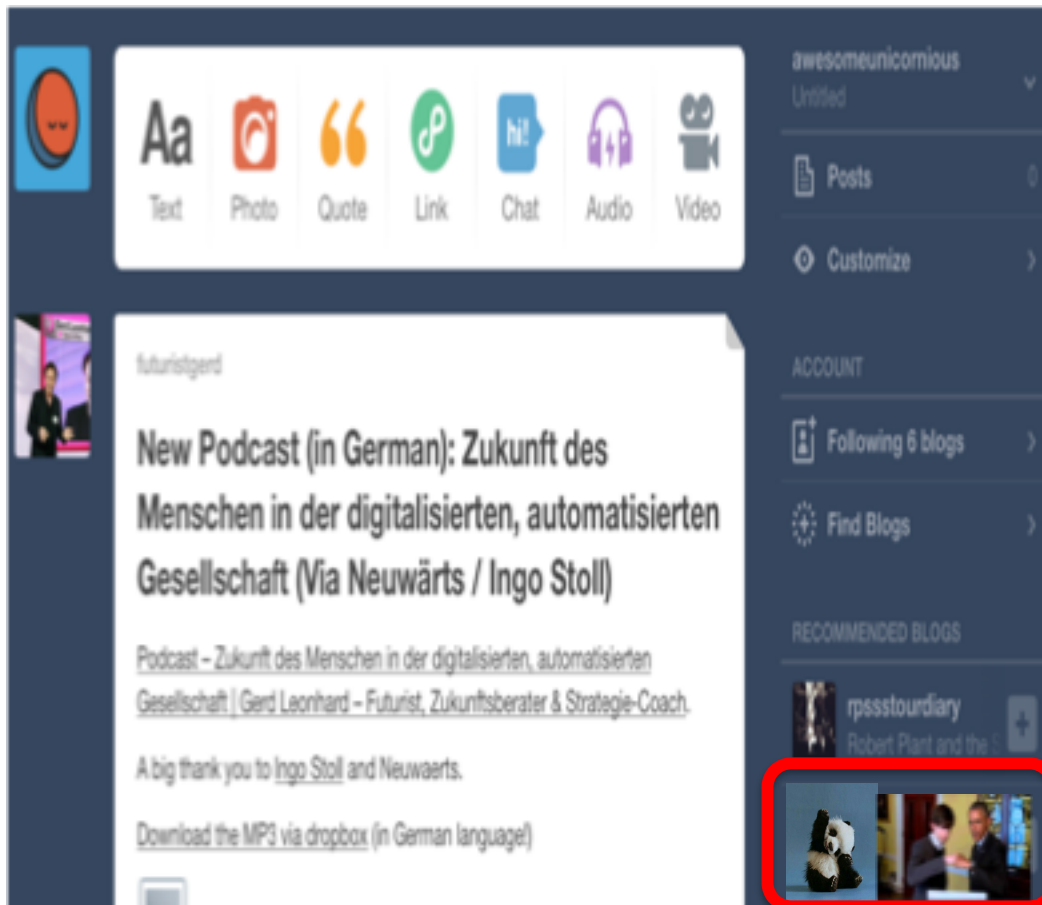
Singh, A. & Gordon, G, KDD 2008

- **Cold start:** Matrix A is very sparse. Some row (user) or column (item) can be completely missing.



Tumblr Blog recommendation

- Which blog we should recommend to users?



The image shows a screenshot of a Tumblr blog post. The post is by the user 'futuristgerd' and is titled 'New Podcast (in German): Zukunft des Menschen in der digitalisierten, automatisierten Gesellschaft (Via Neuwärts / Ingo Stoll)'. The post content includes a description of the podcast and a thank you message to Ingo Stoll and Neuwärts. Below the post, there is a 'RECOMMENDED BLOGS' section. A red box highlights two recommended blog thumbnails: one featuring a panda and another featuring a man in a suit (likely Barack Obama) shaking hands with another man.

awesomeunicornious
United

Text Photo Quote Link Chat Audio Video

futuristgerd

New Podcast (in German): Zukunft des Menschen in der digitalisierten, automatisierten Gesellschaft (Via Neuwärts / Ingo Stoll)

Podcast – Zukunft des Menschen in der digitalisierten, automatisierten Gesellschaft | Gard Leonhard – Futurist, Zukunftsberater & Strategie-Coach.

A big thank you to [Ingo Stoll](#) and [Neuwärts](#).

[Download the MP3 via dropbox \(in German language!\)](#)

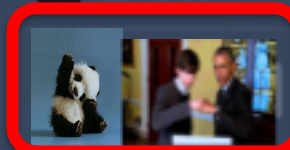
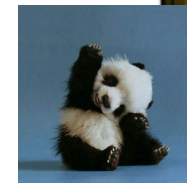
ACCOUNT

Following 6 blogs

Find Blogs

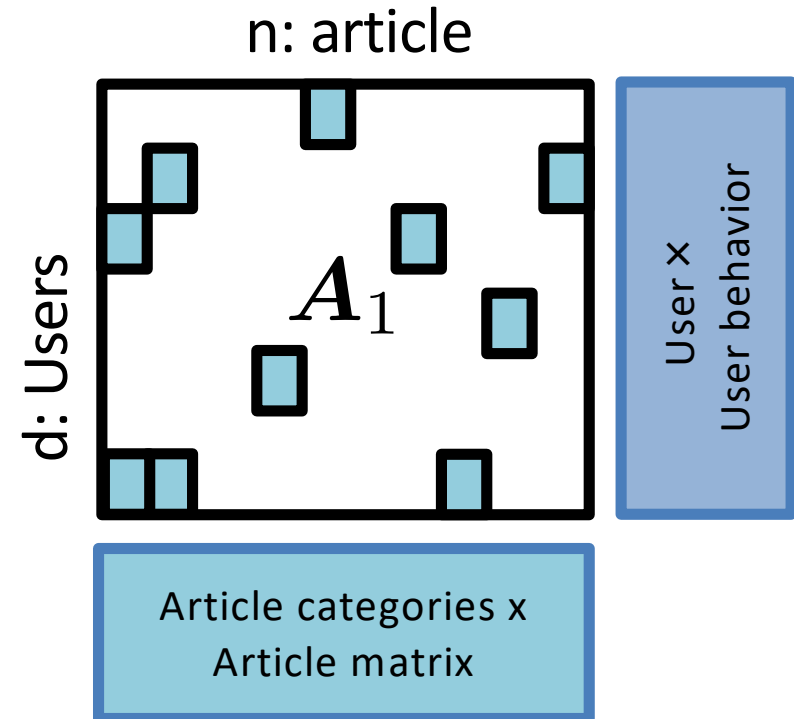
RECOMMENDED BLOGS

rpsstourdiary
Robert Plant and the



News recommendation

- News recommendation
 - #of users $\sim 180,000$
 - #of articles ~ 750
 - #of categories 34
 - #of Rating(1.4 million)



| Method | News-Cold-Start | News-No-Cold-Start |
|------------|-----------------------|-----------------------|
| CMF-Hazans | 0.27408 ± 0.00016 | 0.21559 ± 0.00143 |
| SMF | 0.29051 ± 0.00074 | 0.21488 ± 0.00076 |

Factorization Machine

Rendle, ICDM 2010

- Generalized version of matrix completion
 - Matrix completion + User bias + Item bias
 - We can easily add user information
- Idea (super simple)
 - Solve matrix completion problems by regression
 - (i,j)-th rating input and output can be written as

$$\mathbf{x}_i = \left[\overbrace{0 \cdots 0}^{|U|} \underbrace{1}_{k\text{-th user}} \overbrace{0 \cdots 0}^{|I|} \underbrace{1}_{k'\text{-th item}} \overbrace{0 \cdots 0}^{|I|} \right]^T \in \mathbb{R}^d,$$

$$y_i = [\mathbf{A}]_{k,k'}.$$

Factorization Machine

Rendle, ICDM 2010

- Regression model

$$f(\mathbf{x}; \mathbf{w}, \mathbf{G}) = w_0 + \mathbf{w}_0^\top \mathbf{x} + \sum_{\ell=1}^d \sum_{\ell'=\ell+1}^d \mathbf{g}_\ell^\top \mathbf{g}_{\ell'} x_\ell x_{\ell'},$$

- FM is equivalent to matrix completion

$$\hat{\mathbf{A}}_{k,k'} = w_0 + [\mathbf{w}_0]_k + [\mathbf{w}_0]_{|U|+k'} + \mathbf{g}_k^\top \mathbf{g}_{|U|+k'},$$

- We can also handle the cold start problems by simply concatenating user and item information.

Factorization Machine

- Optimization problem:

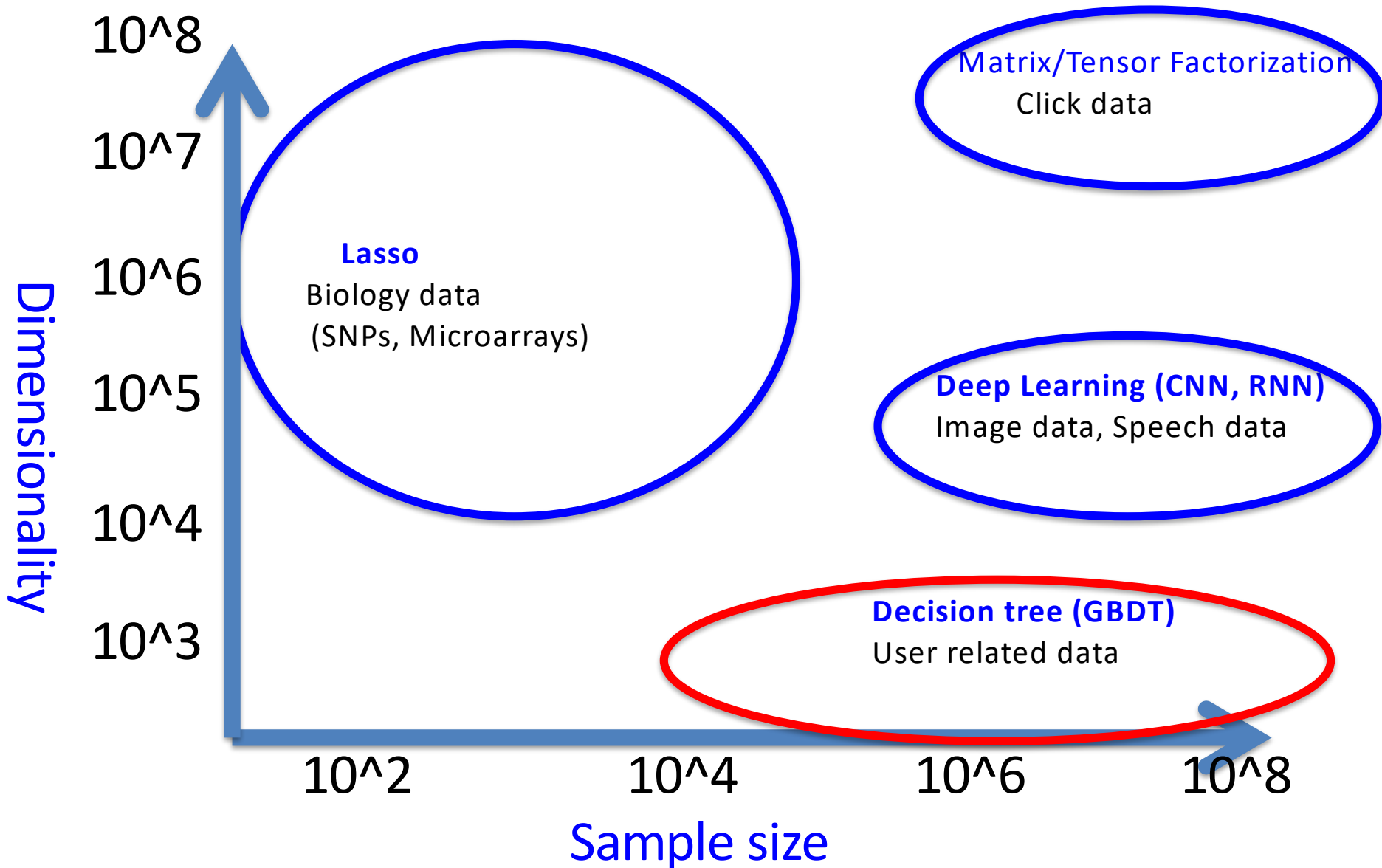
$$\min_{w_0, \mathbf{w}, \mathbf{G}} \sum_{i=1}^n (y_i - f(\mathbf{x}_i; w_0, \mathbf{w}_0, \mathbf{G}))^2 + \lambda_1 \|w_0\|_2^2 + \lambda_2 \|\mathbf{w}_0\|_2^2 + \lambda_3 \|\mathbf{G}\|_F^2,$$

- Alternating Least Squares, SGD, Markov Chain Monte Carlo (MCMC).
- Convex optimization version (Blondel, ECML 2015, Yamada KDD 2017)

Factorization Machine Usage

- URL: <http://www.libfm.org/>
- Movielens data:
<https://grouplens.org/datasets/movielens/>
- `./triple_format_to_libfm.pl -in ml-1m/ratings.dat -target 2 -delete_column 3 -separator "::"`
- `./libFM -task r -train ratings.dat.libfm -test ratings.dat.libfm -dim '1,1,8'`

Data & Related methods



Low dimensionality & large sample

- The number of samples is larger than that of dimension ($n \gg d$)
 - Images
 - Speech
 - User related data

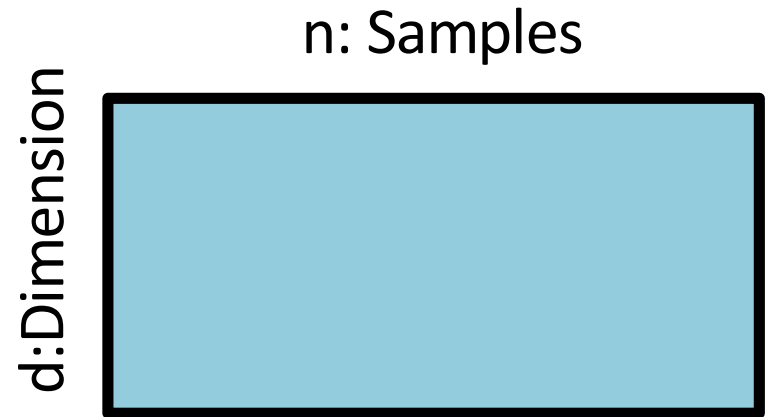


Image and speech recognition: [Deep Learning](#)

Spam detection, Web ranking: GBDT (xgboost),
Logistic regression

Yahoo Auction

The screenshot displays the Yahoo Auction homepage with the following elements:

- Navigation Bar:** Includes the site URL 'auctions.yahoo.co.jp', search bar, and user account options.
- Top Promotions:** Three banners at the top: 'かんたん決済が無料に! (外部)', 'ブックオフと検索してお得なクーポン', and '家の中の不用品を探そう! (外部)'.
- Category List (Left):** A vertical menu with categories such as 'コンピュータ', '家電、AV | カメラ', '音楽 | CD', '本、雑誌 | 漫画', '映画、ビデオ | DVD', 'おもちゃ | ゲーム', 'ホビー、カルチャー', 'アンティーク、コレクション', 'スポーツ、レジャー', '自動車 | オートバイ', 'ファッション | ブランド別', 'アクセサリ、時計', 'ビューティ、ヘルスケア', '食品 | 飲料', '住まい、インテリア | DIY', 'ペット、生き物', '事務、店舗用品', '花、園芸 | 農業', 'チケット、金券 | 宿泊予約', 'ベビー用品', and 'おもちゃ | ゲーム'.
- Main Content Area:**
 - 100万円山分け!** and **1000ポイントプレゼント!** promotional banners.
 - マイ・オークション:** A section with tabs for 'ウォッチリスト', '入札中', '落札分', '値下げ交渉', '出品中', and '出品終了分'.
 - チェックした商品の関連商品:** A carousel of related products including smartphones and tablets.

| 商品名 | 現在価格 | 入札条件 |
|-------------------------------------|----------|------------|
| 超美品 AU/LG isai LGL 22 4G LTE イ... | 21,681 円 | 入札0/残り1日 |
| ●docomo SHARP AQ UOS SH-06F ホワイ... | 31,000 円 | 入札24/残り2日 |
| au/ISW11HT htc EV Qi/中西/BDOA/mbm... | 2,000 円 | 入札0/残り20時間 |
 - Chrysler 300:** A featured advertisement for a white Chrysler 300 with a '7日間モニターキャンペーン'.
 - 日帰りバスツアー特集:** A banner for a 'Yahoo! Travel' special offer.
 - リユースをはじめよう:** A section encouraging users to sell or buy items, with buttons for '出品する' and '買取を申し込む'.
 - ヤフオク!へようこそ:** A welcome message stating that 8,774 items were auctioned on August 18th and 2,189,348 items were newly listed.
- Footer:** Search bar and 'Highlight All' / 'Match Case' options.

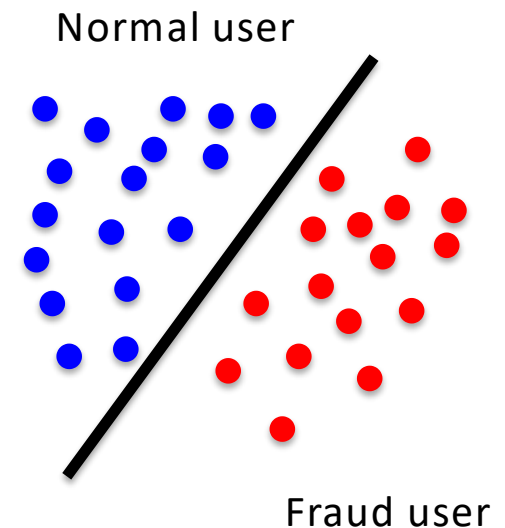
Auction Fraud Detection

- Example of frauds
 - Selling fake items
 - Do not send items
 - Do a big frauds after gathering trust scores.
 - Etc.
- Detecting fraud is a very important to make users happy!
- Challenge
 - The fraud types changes over season
 - Active learning, transfer learning, etc.
 - Big data (the number of samples can be hundred million of items)

Formulation

- Classification problem (Fraud or non-Fraud)
 - Build a classifier using a labeled data

| | Gender | Age | ... | Location | Label |
|--------|--------|-----|-----|----------|-------|
| User 1 | Male | 25 | ... | Tokyo | +1 |
| User 2 | Female | 20 | ... | Kyoto | -1 |
| ... | | | ... | | |
| User n | Male | 36 | ... | Tokyo | -1 |



Supervised Learning (review)

- Input and output: $\mathbf{x} \in \mathbb{R}^d$, $y \in \mathbb{R}$
- Training samples: $\{(\mathbf{x}_i, y_i)\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} p(\mathbf{x}, y)$
- Goal: Training classifier from the training samples.

- Model (Linear model)

$$f(\mathbf{x}; \mathbf{w}) = w_1 x_1 + w_2 x_2 + \dots + w_d x_d = \mathbf{w}^\top \mathbf{x}$$

- Model parameter

$$\mathbf{w} = [w_1, w_2, \dots, w_d]^\top \in \mathbb{R}^d$$

Prediction

- How to prediction (User probability)

$$p(y = +1|\mathbf{x}) = \frac{1}{1 + \exp(-\mathbf{w}^\top \mathbf{x})}$$

$$p(y = -1|\mathbf{x}) = \frac{\exp(-\mathbf{w}^\top \mathbf{x})}{1 + \exp(-\mathbf{w}^\top \mathbf{x})}$$

- The probability should be sum to 1.

$$p(y = +1|\mathbf{x}) + p(y = -1|\mathbf{x}) = \frac{1 + \exp(-\mathbf{w}^\top \mathbf{x})}{1 + \exp(-\mathbf{w}^\top \mathbf{x})} = 1$$

Parameter training (Review)

- We train model parameter \mathbf{w} from data
- How to estimate?
 - The positive class probability is high if the data is normal, and the negative class probability is high if the data is fraud.

- Likelihood function: $L(\mathbf{w}) = \prod_{i=1}^n p(y_i | \mathbf{x}_i; \mathbf{w})$

- Log likelihood function:

$$L(\mathbf{w}) = \log \prod_{i=1}^n p(y_i | \mathbf{x}_i; \mathbf{w}) = \sum_{i=1}^n \log p(y_i | \mathbf{x}_i; \mathbf{w})$$

Parameter training

- Optimization

$$\max_{\mathbf{w}} L(\mathbf{w}) \rightarrow \min_{\mathbf{w}} - \sum_{i=1}^n \log p(y_i | \mathbf{x}_i; \mathbf{w})$$

- We can use a gradient descent.

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta \nabla_{\mathbf{w}} (-L(\mathbf{w}))$$

Fraud Detection

- How to find fraud users?

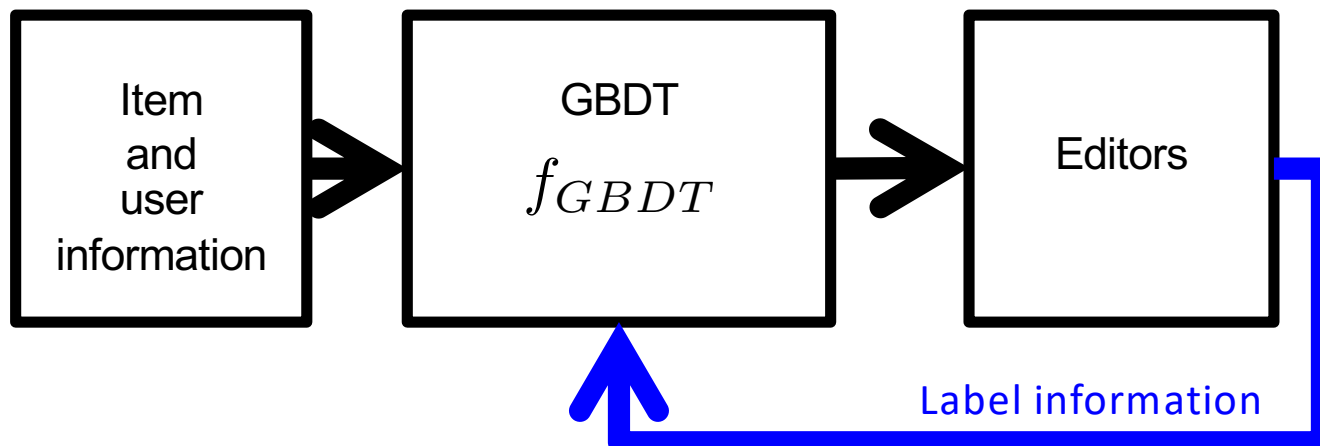
The screenshot shows the Yahoo! Auctions Japan homepage. The browser address bar displays 'auctions.yahoo.co.jp'. The page features a navigation menu on the left with categories like 'コンピュータ' (Computers), '家電・AV・カメラ' (Home Appliances), and 'ファッション' (Fashion). The main content area includes several promotional banners and product listings. One prominent banner advertises a '100万円山分け!' (100,000 Yen Split!) promotion. Below this, there are sections for 'マイ・オークション' (My Auctions) and 'チェックした商品の関連商品' (Related products for items you've checked). The bottom right corner contains a 'ヤフオク!へようこそ' (Welcome to Yahoo! Auctions!) message, mentioning that 8,744 people participated in an auction on August 18th, with 2,189,348 items sold.

Practical issues

- Various types of frauds
 - Selling fake items
 - Do not send items
 - Do a big frauds after gathering trust scores.
 - Etc.
- Detecting fraud is a very important to make users happy!
- Challenge
 - The fraud types changes over season
 - Active learning, transfer learning, etc.
 - Big data (**the number of samples can be hundred million of items**)
- Can we automatically erase user account?

One solution: Using Active learning

- **Key idea:** We ask human editor to judge fraud or not → Feedback the result to machine learning model (Active learning)
- **Use supervised learning** (GBDT, xgboost)
 - Semi-supervised and unsupervised method tends not to work for real problems.
- **Feature engineering is super important!**



Results

- Detection results (rule based approach is a baseline)

