上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

EE448
Big Data Mining

Weinan Zhang
Shanghai Jiao Tong University
http://wnzhang.net

Spring Semester 2019

http://wnzhang.net/teaching/ee448/index.html

Self Introduction – Weinan Zhang

- Position
  - Assistant Professor at CS Dept. of SJTU 2016-now
  - Apex Data and Knowledge Management Lab
  - John Hopcroft Research Center for Computer Science
  - Research on machine learning and data mining topics

- Education
  - Ph.D. on Computer Science from University College London (UCL), United Kingdom, 2012-2016
  - B.Eng. on Computer Science from ACM Class 07 of Shanghai Jiao Tong University, China, 2007-2011

Course Administration

- No official text book for this course, some recommended books are

  - Jiawei Han, Micheline Kamber, Jian Pei. “Data Mining: Concepts and Techniques, 3rd Edition”. Morgan Kaufmann Series, 2011.
  - 范明，孟小峰 译《数据挖掘 概念与技术》机械工业出版社，2012.
  - Bing Liu. “Web Data Mining, 2nd Edition”. Springer, 2011.
  - 俞勇等 译《Web数据挖掘》清华大学出版社，2012.

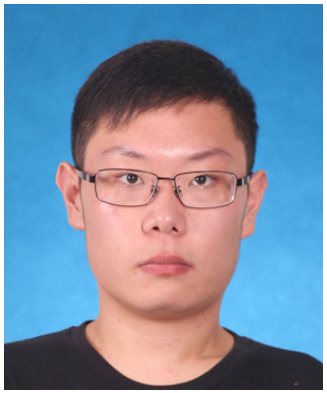
  - 李航《统计学习方法》清华大学出版社，2012.
  - 周志华《机器学习》清华大学出版社，2016.
  - Tom Mitchell. “Machine Learning”. McGraw-Hill, 1997

Course Administration

- A hands-on big data mining course
  - No assignment, no final exam
  - Two course works (80%)
    - Text Classification (40%)
    - Recommendation (40%)
  - Poster session (10%)
  - Attending (10%)
    - Could be evaluated by quiz

Teaching Assistants

- Haiwen Wang (王海文)
- Email: wanghaiwencn[A.T.]foxmail.com
- 2018 Ph.D student at IIoT
- Research on data mining, graph deep learning



- Zhaorun Han (撖朝润)
- Email: hanzhaorun [A.T.] sjtu.edu.cn
- 2018 M.S. student at IIoT
- Research on big data analysis

TA Administration

- Join the mail list
  - Please send your
    - Chinese name
    - Student number
    - Email address

    to wanghaiwencn [A.T.] foxmail.com

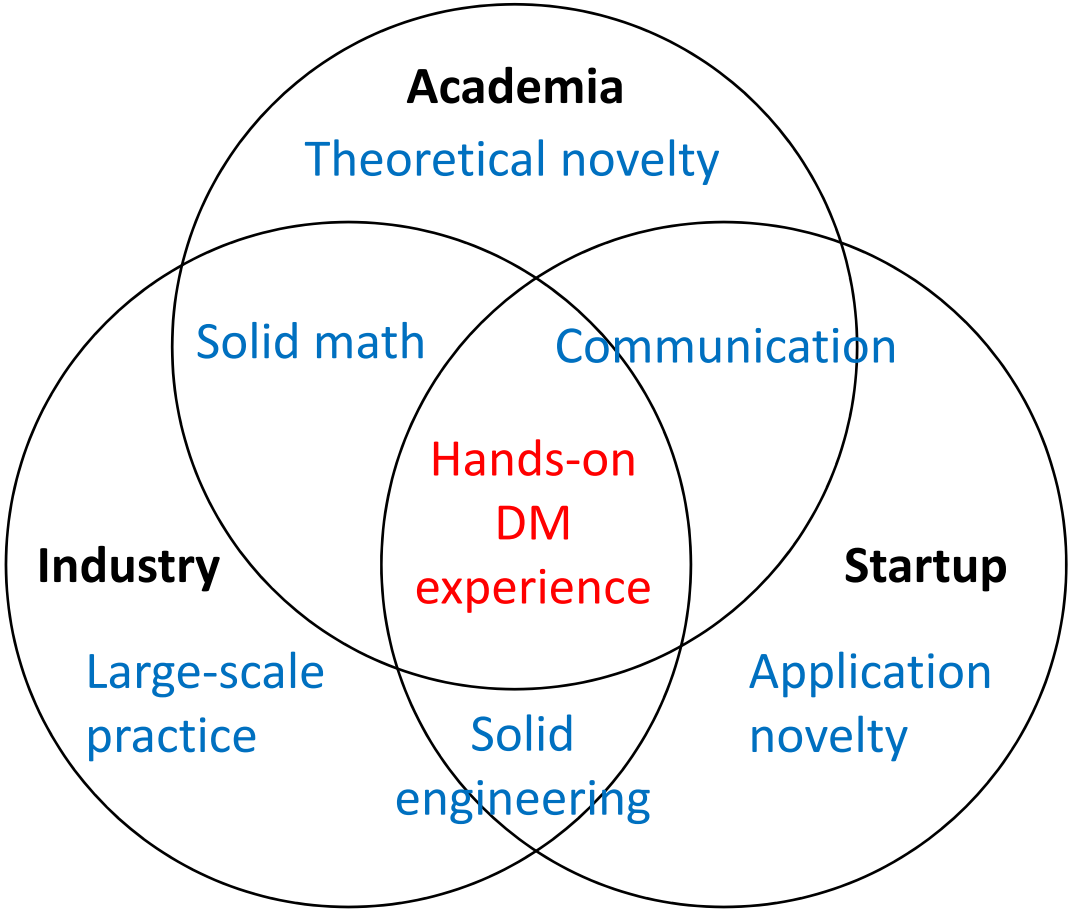
    with email title “Check in EE448 2019”

- Office hour
  - Every Monday 7-8pm, 1-432 SEIEE Building
  - TAs will be there for QA

Goals of This Course

- Know about the big picture of data science
- Get familiar with popular data mining methodologies
  - Data representations
  - Problem formulation
  - Machine learning & data mining algorithms
  - Experimental methodologies
- Get some first-hand DM developing experiences
- Present your own DM solutions to real-world problems

Why we focus on hands-on DM



- Get familiar with various data mining applications.
- Play with the data and get your hands dirty!

Course Landscape

1. Data Mining Intro
2. Fundamentals of Data
3. Basic DM Algorithms
4. Supervised Learning 1
5. Supervised Learning 2
6. Supervised Learning 3
7. Unsupervised Learning
8. Text Mining
9. Search Engines
10. Ranking Information Items
11. Recommender Systems
12. Computational Ads
13. Behavioral Targeting
14. Knowledge Graphs
15. Social Networks
16. Poster Session

EE448, Big Data Mining, Lecture 1

Introduction to Big Data Mining

Weinan Zhang

Shanghai Jiao Tong University

http://wnzhang.net

http://wnzhang.net/teaching/ee448/index.html

Content of This Lecture

- An example as an intro of data mining
- Concepts of data mining
- Real-world examples of data mining

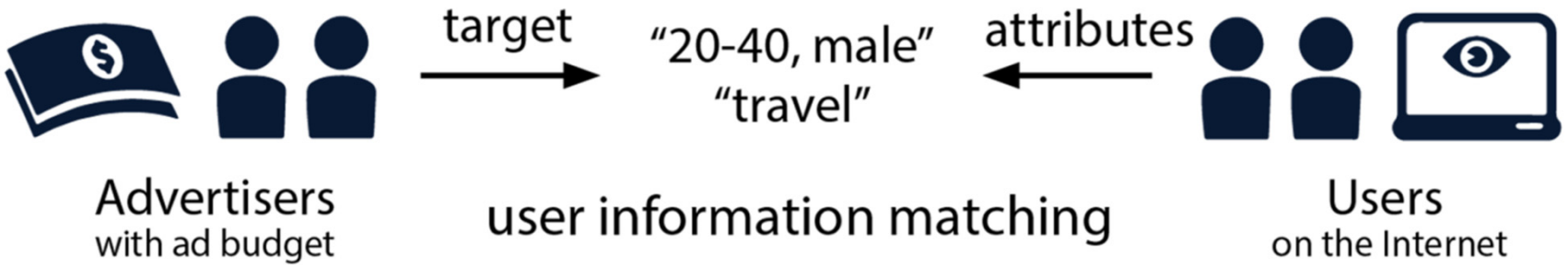
Display Advertising

- A display ad example

How likely the user is going to click the ad?

http://news.ifeng.com

Display Advertising



- Advertiser targets a segment of users
  - E.g. by age, gender, occupation, interest tags etc.
- Intermediary matches users and ads by user information

Internet Advertising Frontier:
Real-Time Bidding (RTB) based Display Advertising

What is Real-Time Bidding?

- Every online **ad view** can be evaluated, bought, and sold, all **individually**, and all **instantaneously**.

- Instead of buying keywords or a bundle of ad views, advertisers are now **buying users** directly.

Wang J, Zhang W, Yuan S. Display Advertising with Real-Time Bidding (RTB) and Behavioural Targeting. Foundations and Trends® in Information Retrieval, 2017, 11(4-5): 297-435.

An RTB Example

- Weinan regularly reads articles on emarketer.com

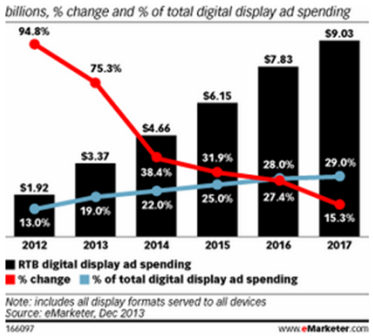
eMarketer Research Topics Products Why eMarketer Customer Stories Articles

Advertisers Continue Rapid Adoption of Programmatic Buying

By 2017, advertisers will spend more than $9 billion on RTB

Nov 26, 2013 Share Print Email

Advertisers are spending more than expected on real-time bidding, which is expected to account for a significant share of all display ad spending in the US



continues its rapid transition from infancy to a well-established display purchase method in just a few years.

eMarketer projects RTB digital display ad spending in the US will account for 29.0% of total US digital display ad spending by 2017, or $9.03 billion. In 2013, it will account for 19.0%, or $3.37 billion. These estimates are revised slightly upward from our previous forecast in August

Latest from eMarketer

Latest Articles | Latest Webinars

Hispanic Gen Xers Lead in Daily Tablet Usage

Chrysler's Multichannel Approach to Online Video Gets Greater Recall

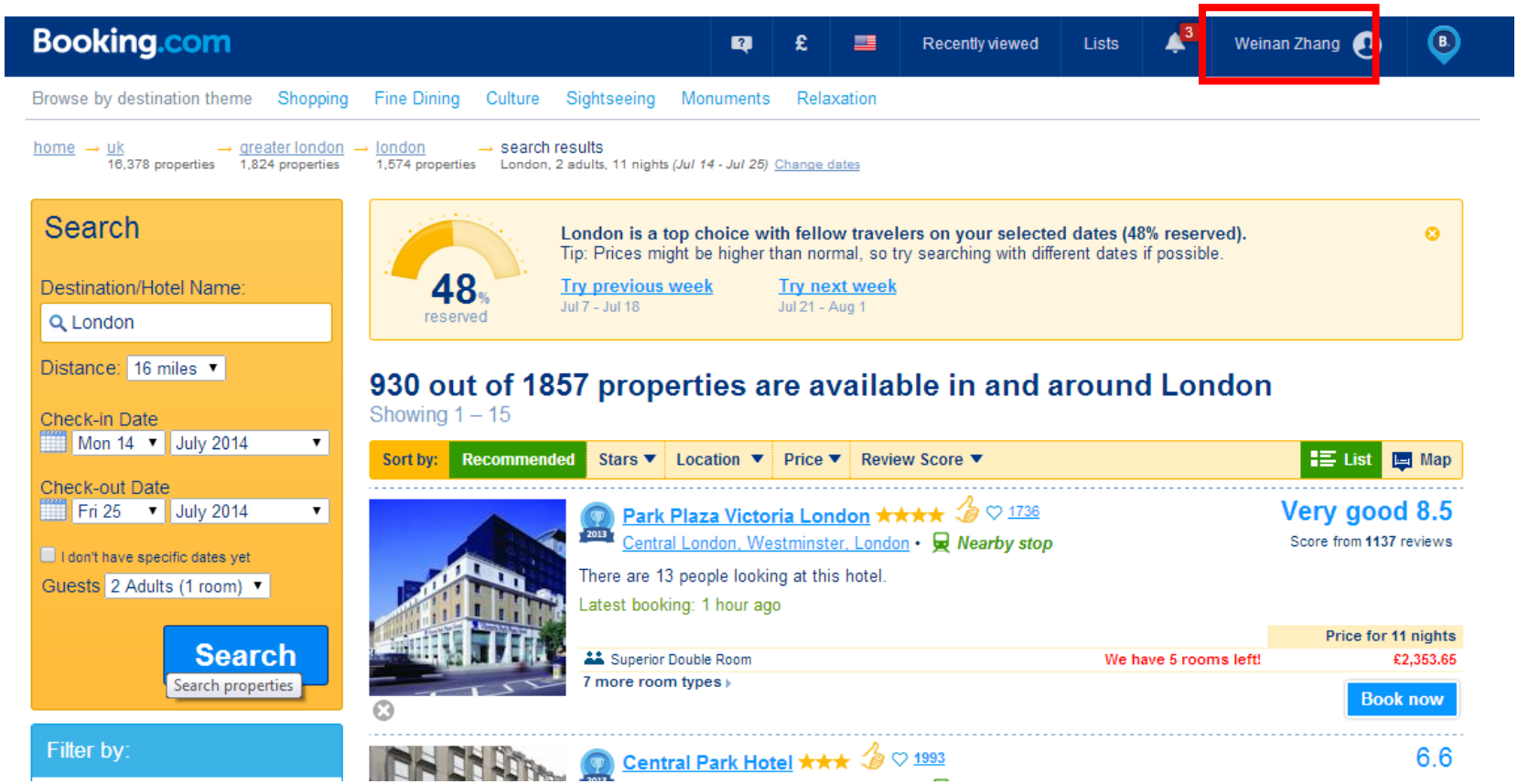
Android Rules UK Smartphone Sales

More Articles » eMarketer Daily Newsletter »

Contact Sign-Up
Contact Sign-Up
Contact Sign-Up
Contact Sales

An RTB Example

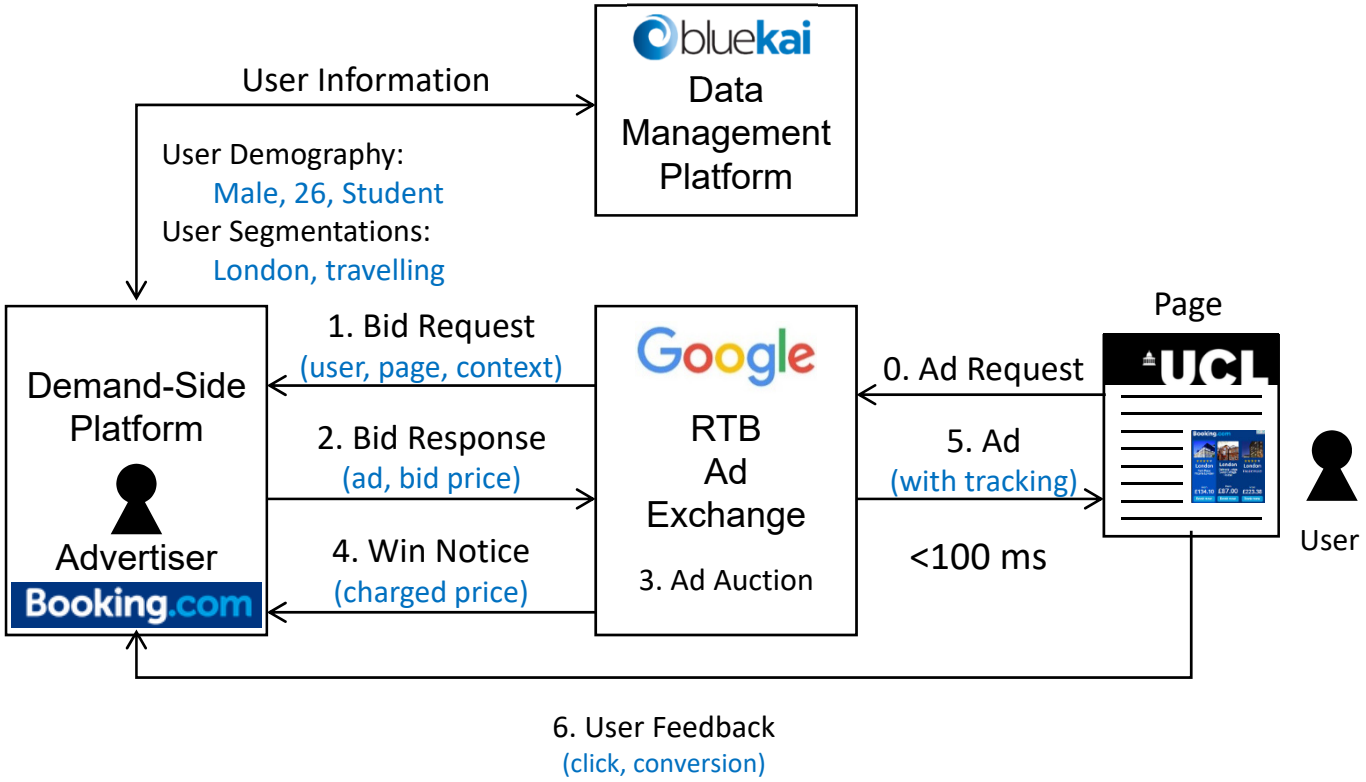
- Weinan recently checked the London hotels on booking.com



An RTB Example

- The day after, he found relevant ads on facebook.com

Players Interaction in RTB



- A demand-side platform buys ads via real-time bidding (RTB) 10 billion per day
- A data management platform analyzes and maintains the information billions of Internet users

Data Technology as a Service

A data service example in RTB ads

- The data service (or DaaS) is a cousin of software as a service (SaaS)
  - takes the input of high-quality data request based on raw data
  - returns the requested high-quality data for higher-level (intelligent) applications

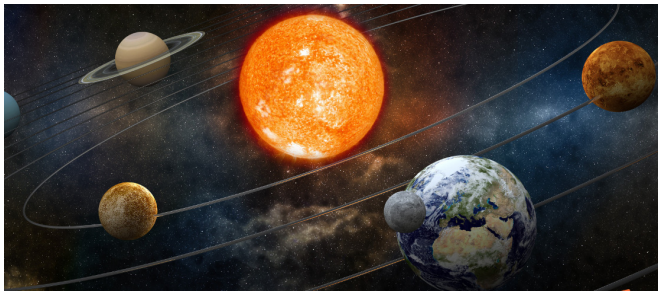
Content of This Lecture

- An example as an intro of data mining
- Concepts of data mining
- Real-world examples of data mining

The Underlying Data Science

- Data science is the subject concerned with the methodology of discovering the underlying principles and patterns from massive amount of data.

- Physics
  - Goal: discover the underlying principle of the world



  - Solution: build the model of the world

$$F = G\frac{m_1 m_2}{r^2}$$

Example: Newton's gravity law

- Data Science
  - Goal: discover the underlying principle of the data

  - Solution: build the model of the data

$$p(x) = \frac{e^{f(x)}}{\sum_{x'} e^{f(x')}}$$

Example: Energy-based distribution

- In fact, data science could be a more general concept for natural science.

Evolution of Sciences

- Before 1600, **empirical science**
- 1600-1950s, **theoretical science**
  - Each discipline has grown a *theoretical* component. Theoretical models often motivate experiments and generalize our understanding.
- 1950s-1990s, **computational science**
  - Over the last 50 years, most disciplines have grown a third, *computational* branch (e.g. empirical, theoretical, and computational ecology, or physics, or linguistics.)
  - Computational Science traditionally meant simulation. It grew out of our inability to find closed-form solutions for complex mathematical models.
- 1990-now, **data science**
  - The flood of data from new scientific instruments and simulations
  - The ability to economically store and manage petabytes of data online
  - The Internet and computing Grid that makes all these archives universally accessible
  - Scientific info. management, acquisition, organization, query, and visualization tasks scale almost linearly with data volumes. Data mining is a major new challenge!

Jim Gray and Alex Szalay, *The World Wide Telescope: An Archetype for Online Science*, Comm. ACM, 45(11): 50-54, Nov. 2002

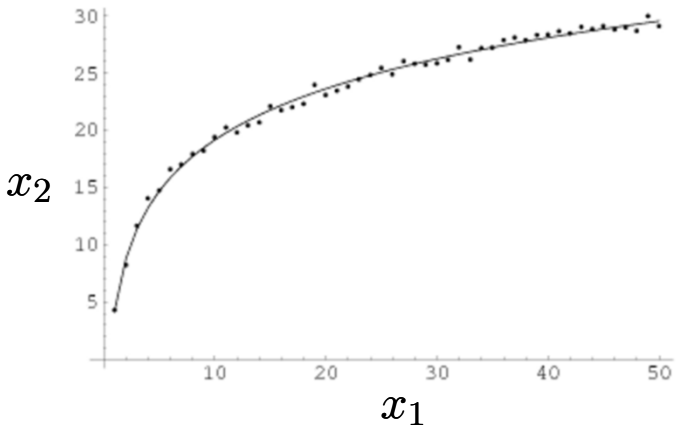
Slide credit: Jiawei Han

Data Science

- A deterministic view
  - For a high-dimensional data $\mathbf{x}$
  - Find the underlying function

$$\mathbf{x}_i = f(\mathbf{x}_{\neq i})$$

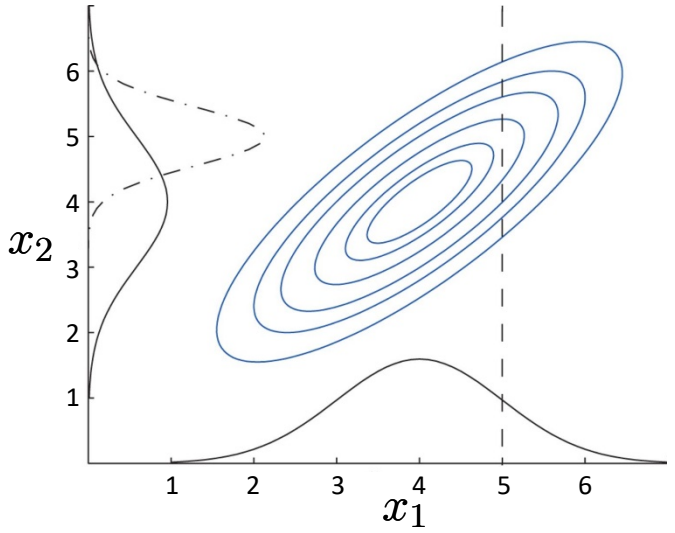
for a certain target dimension data $\mathbf{x}_i$



- A probabilistic view
  - For a high-dimensional data $\mathbf{x}$
  - Find joint data distribution $p(\mathbf{x})$
  - Then the conditional distribution

$$p(\mathbf{x}_i|\mathbf{x}_{\neq i})$$

for a certain target dimension data $\mathbf{x}_i$



An Example in User Behavior Modeling

Interest	Gender	Age	BBC Sports	PubMed	Bloomberg Business	Spotify
Finance	Male	29	Yes	No	Yes	No
Sports	Male	21	Yes	No	No	Yes
Medicine	Female	32	No	Yes	No	No
Music	Female	25	No	No	No	Yes
Medicine	Male	40	Yes	Yes	Yes	No

Expensive data

Cheap data

- A 7-field record data
  - 3 fields of data that are expensive to obtain
    - Interest, gender, age collected by user registration information or questionnaires
  - 4 fields of data that are easy or cheap to obtain
    - Raw data of whether the user has visited a particular website during the last two weeks, as recorded by the website log

An Example in User Behavior Modeling

Interest	Gender	Age	BBC Sports	PubMed	Bloomberg Business	Spotify
Finance	Male	29	Yes	No	Yes	No
Sports	Male	21	Yes	No	No	Yes
Medicine	Female	32	No	Yes	No	No
Music	Female	25	No	No	No	Yes
Medicine	Male	40	Yes	Yes	Yes	No

Expensive data

Cheap data

- **Deterministic view**: fit a function

  Age = f(Browsing=BBC Sports, Bloomberg Business)

- **Probabilistic view**: fit a joint data distribution

  p(Interest=Finance, Gender=Male, Age=29, Browsing=BBC Sports, Bloomberg Business)

- Then build the conditional data distribution

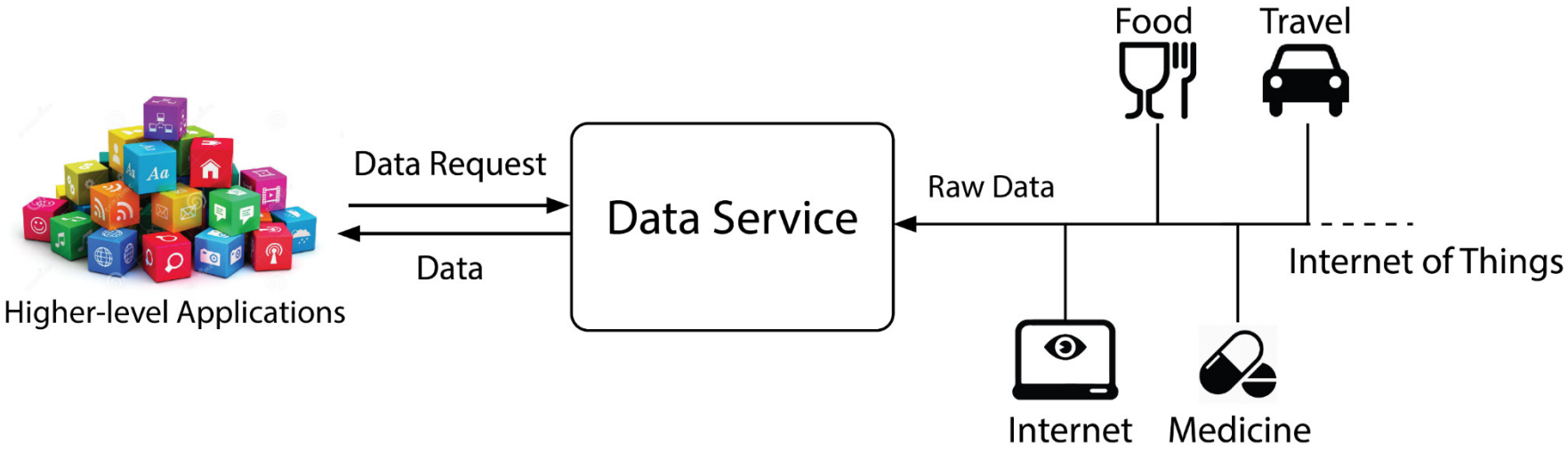
  p(Interest=Finance | Browsing=BBC Sports, Bloomberg Business)

  p(Gender=Male | Browsing=BBC Sports, Bloomberg Business)

Data Technology as a Service

- The data service is just like a data processing factory that
  - collects raw and cheap data
  - supports the higher-level (intelligent) applications with quality data

Data Technology Everywhere



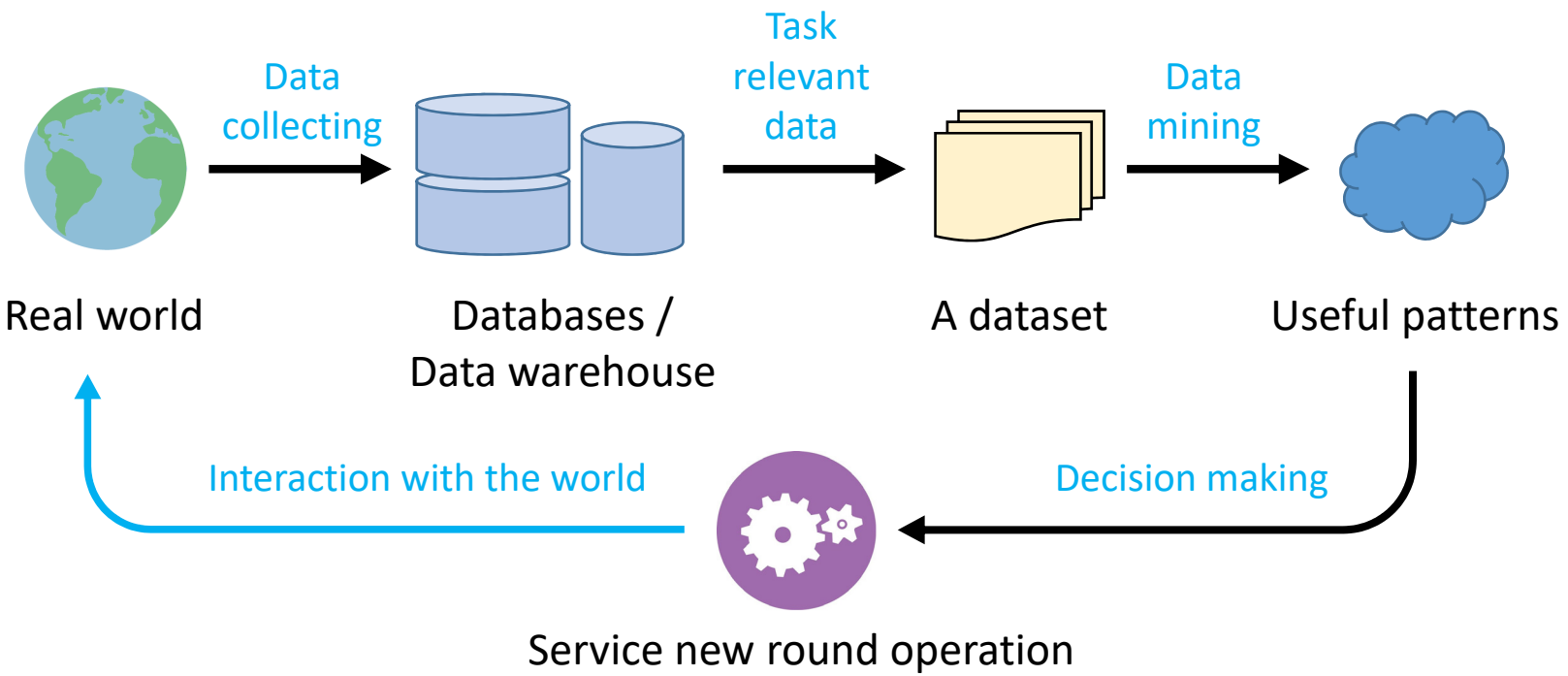
- The data itself is not valuable without the data service!
- How to perform proper and effective mining for the principles, patterns and knowledge from massive amount of data is what we focus in this course.

What is Data Mining?

- Data mining is about the extraction of non-trivial, implicit, previously unknown and potentially useful principles, patterns or knowledge from massive amount of data.

- Data Science is the subject concerned with the scientific methodology to properly, effectively and efficiently perform data mining
  - an interdisciplinary field about scientific methods, processes, and systems

A Typical Data Mining Process



- Data mining plays a key role of enabling and improving the various data services in the world
- Note that the (improved) data services would then change the world data, which would in turn change the data to mine

A Multi-Dimensional View of Data Mining

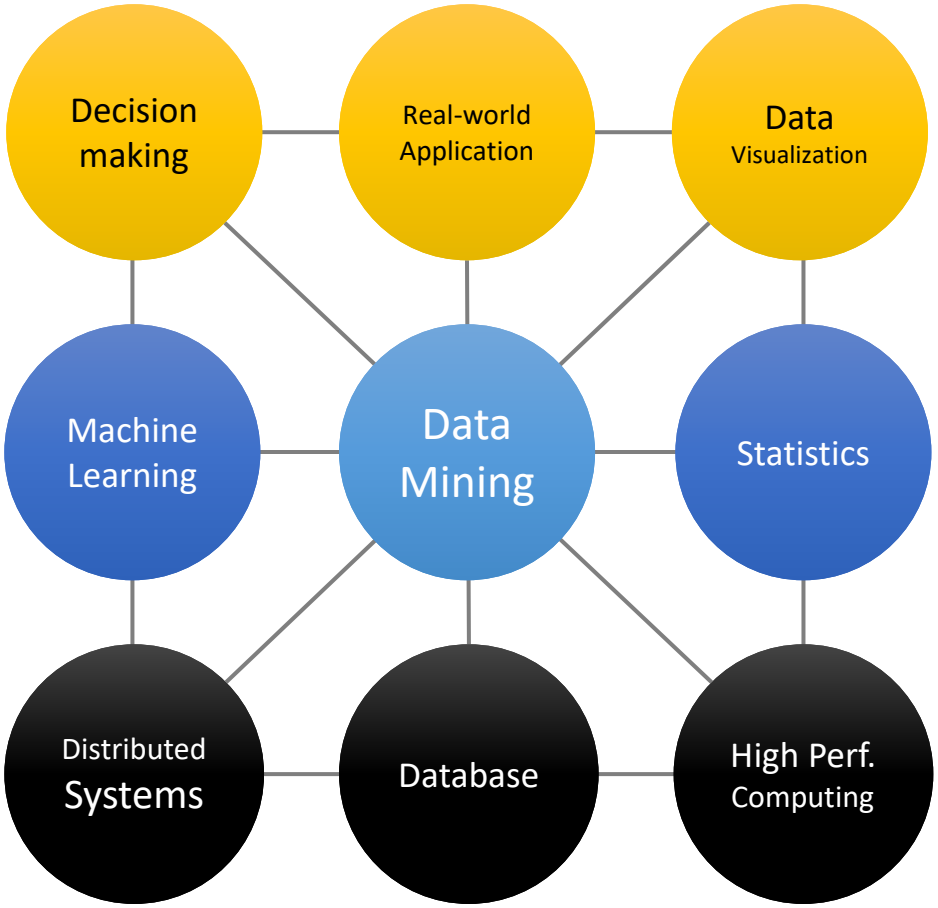
- Data to be mined
  - Database data (extended-relational, object-oriented, heterogeneous, legacy), data warehouse, transactional data, stream, spatiotemporal, time-series, sequence, text and web, multi-media, graphs & social and information networks
- Knowledge to be mined (or data mining functions)
  - Characterization, discrimination, association, classification, clustering, trend/deviation, outlier analysis, etc.
  - Descriptive vs. predictive data mining
  - Multiple/integrated functions and mining at multiple levels
- Techniques utilized
  - Data warehouse, **machine learning**, statistics, pattern recognition, visualization, distributed computing, high-performance, etc.
- Applications adapted
  - Retail, telecommunication, banking, fraud analysis, bio-data mining, stock market analysis, text mining, Web mining, etc.

More application examples will be provided.

Slide credit: Jiawei Han

Data Mining Techniques

- Application level
  - Intelligent systems & applications with further feedbacks
- Methodology level
  - Machine learning & statistics techniques based on large amount of formatted data
- System level
  - Scalable systems & architectures for hosting, retrieving and computing big data



Data Mining and Machine Learning

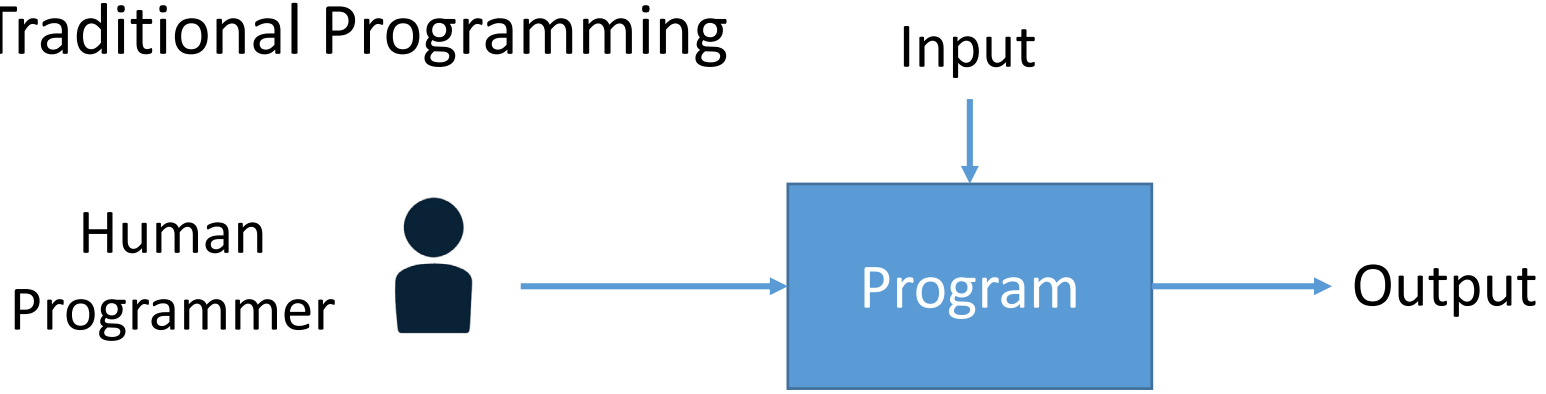
- What is the difference between data mining and machine learning?

- Data mining is about the extraction of non-trivial, implicit, previously unknown and potentially useful principles, patterns or knowledge from massive amount of data.

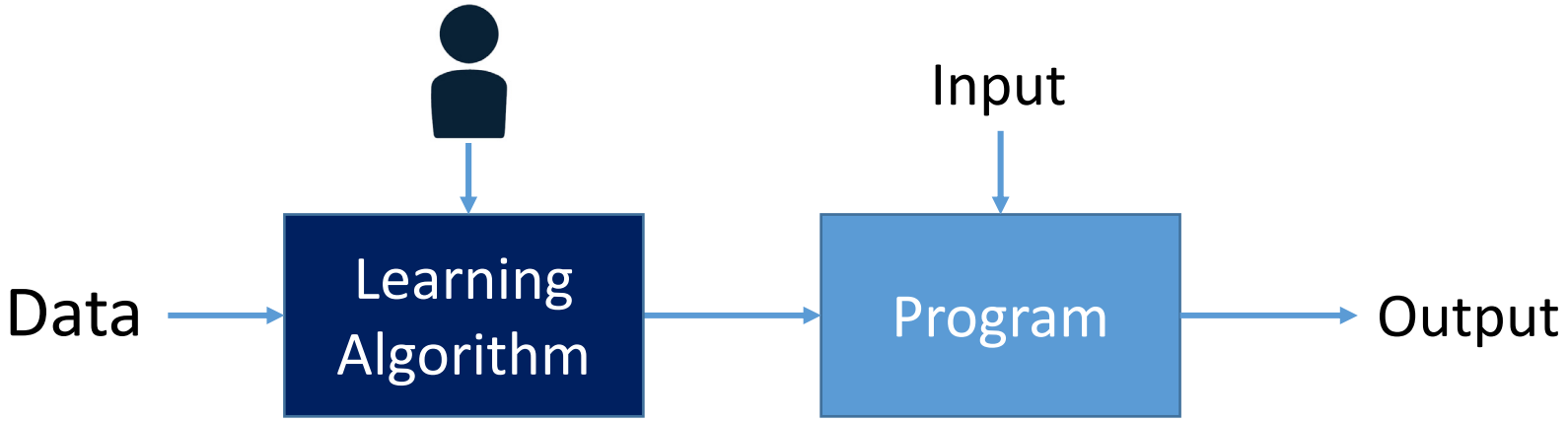
- Machine learning is the study of algorithms that improves a particular quantitative performance at some task based on data with non-explicit programming.

Programming vs. Machine Learning

- Traditional Programming



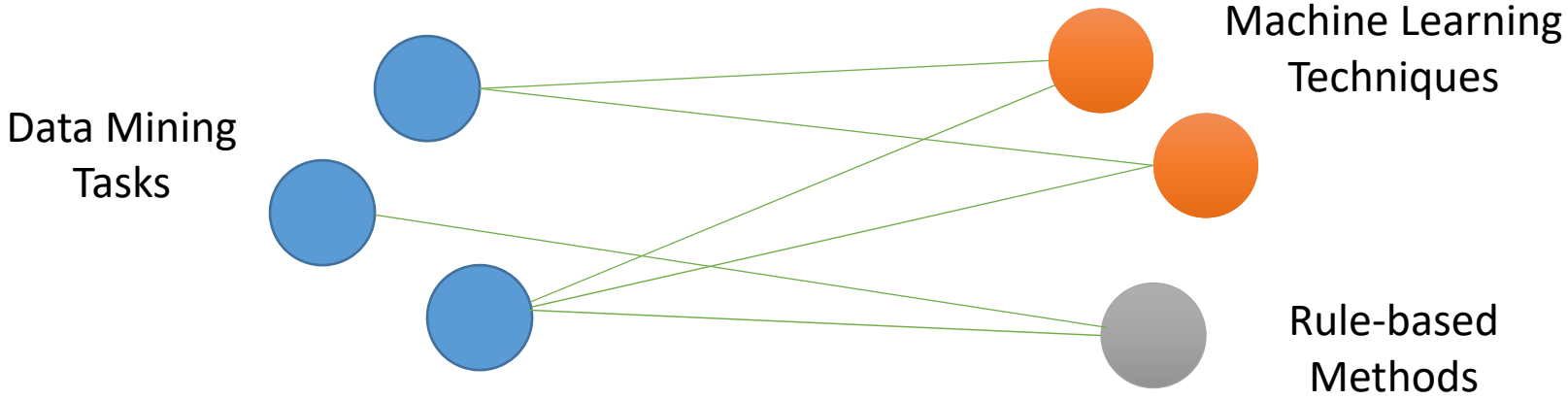
- Machine Learning



Slide credit: Feifei Li

Data Mining and Machine Learning

- What is the difference between data mining and machine learning?
  - They are solving similar tasks with different focuses
    - Data mining focuses on solving the problems
    - Solving a DM problem could involve different methods including ML
    - Machine learning focuses on modeling based on the data
    - An ML model could be applied to various DM tasks



A Brief History of Data Mining Society

- 1989 IJCAI Workshop on Knowledge Discovery in Databases
  - Knowledge Discovery in Databases (G. Piatetsky-Shapiro and W. Frawley, 1991)
- 1991-1994 Workshops on Knowledge Discovery in Databases
  - Advances in Knowledge Discovery and Data Mining (U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 1996)
- 1995-1998 International Conferences on Knowledge Discovery in Databases and Data Mining (KDD'95-98)
  - Journal of Data Mining and Knowledge Discovery (1997)
- ACM SIGKDD conferences since 1998 and SIGKDD Explorations
- More conferences on data mining
  - PAKDD (1997), PKDD (1997), SIAM-Data Mining (2001), (IEEE) ICDM (2001), etc.
- ACM Transactions on KDD starting in 2007

Slide credit: Jiawei Han

Conferences and Journals on Data Mining

- KDD Conferences
  - ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD)
  - SIAM Data Mining Conf. (SDM)
  - (IEEE) Int. Conf. on Data Mining (ICDM)
  - Int. Conf. on Web Search and Data Mining (WSDM)
  - European Conf. on Machine Learning and Principles and practices of Knowledge Discovery and Data Mining (ECML-PKDD)
  - Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD)
- Other related conferences
  - DB conferences: ACM SIGMOD, VLDB, ICDE, EDBT, ICDT, …
  - Web and IR conferences: WWW, SIGIR, CIKM
  - ML conferences: ICML, NIPS
  - PR conferences: CVPR
- Journals
  - IEEE Trans. On Knowledge and Data Eng. (TKDE)
  - KDD Explorations
  - ACM Trans. on KDD (TKDD)

Slide credit: Jiawei Han

Content of This Lecture

- An example as an intro of data mining
- Concepts of data mining
- Real-world examples of data mining

DM Use Case 1: Frequent Item Set Mining

WRAPPING PAPER	0.99
INSTANT COFFEE GOLD	1.99
INSTANT COFFEE GOLD	1.99
ORANGE JUICE 1.5L	0.79
ORANGE JUICE 1.5L	0.79
RICE CRACKERS SALT	0.29
RICE CRACKERS SALT	0.29
PLAIN MARGARINE	0.44
GARDEN GLOVES	1.49
FREE RANGE EGGS	1.05
ASSORTED MUESLI	1.49
COOKIES	1.05
MACARONI	0.42
BUTTERMILK DESSERT	0.29
BUTTERMILK DESSERT	0.29
BUTTERMILK DESSERT	0.29
BUTTERMILK DESSERT	0.29
TOTAL	14.23
CASH	20.00
CHANGE	5.77

THANK YOU AND GOODBYE

Some intuitive patterns:

$$\{\text{milk, bread, butter}\}$$
$$\{\text{onion, potatoes, beef}\}$$

Some non-intuitive ones:

$$\{\text{diaper, beer}\}$$

Agrawal, R.; Imieliński, T.; Swami, A. (1993). "Mining association rules between sets of items in large databases". *ACM SIGMOD 1993*

DM Use Case 1: Association Rule Mining

WRAPPING PAPER	0.99
INSTANT COFFEE GOLD	1.99
INSTANT COFFEE GOLD	1.99
ORANGE JUICE 1.5L	0.79
ORANGE JUICE 1.5L	0.79
RICE CRACKERS SALT	0.29
RICE CRACKERS SALT	0.29
PLAIN MARGARINE	0.44
GARDEN GLOVES	1.49
FREE RANGE EGGS	1.05
ASSORTED MUESLI	1.49
COOKIES	1.05
MACARONI	0.42
BUTTERMILK DESSERT	0.29
BUTTERMILK DESSERT	0.29
BUTTERMILK DESSERT	0.29
BUTTERMILK DESSERT	0.29
TOTAL	14.23
CASH	20.00
CHANGE	5.77

THANK YOU AND GOODBYE

Some intuitive patterns:

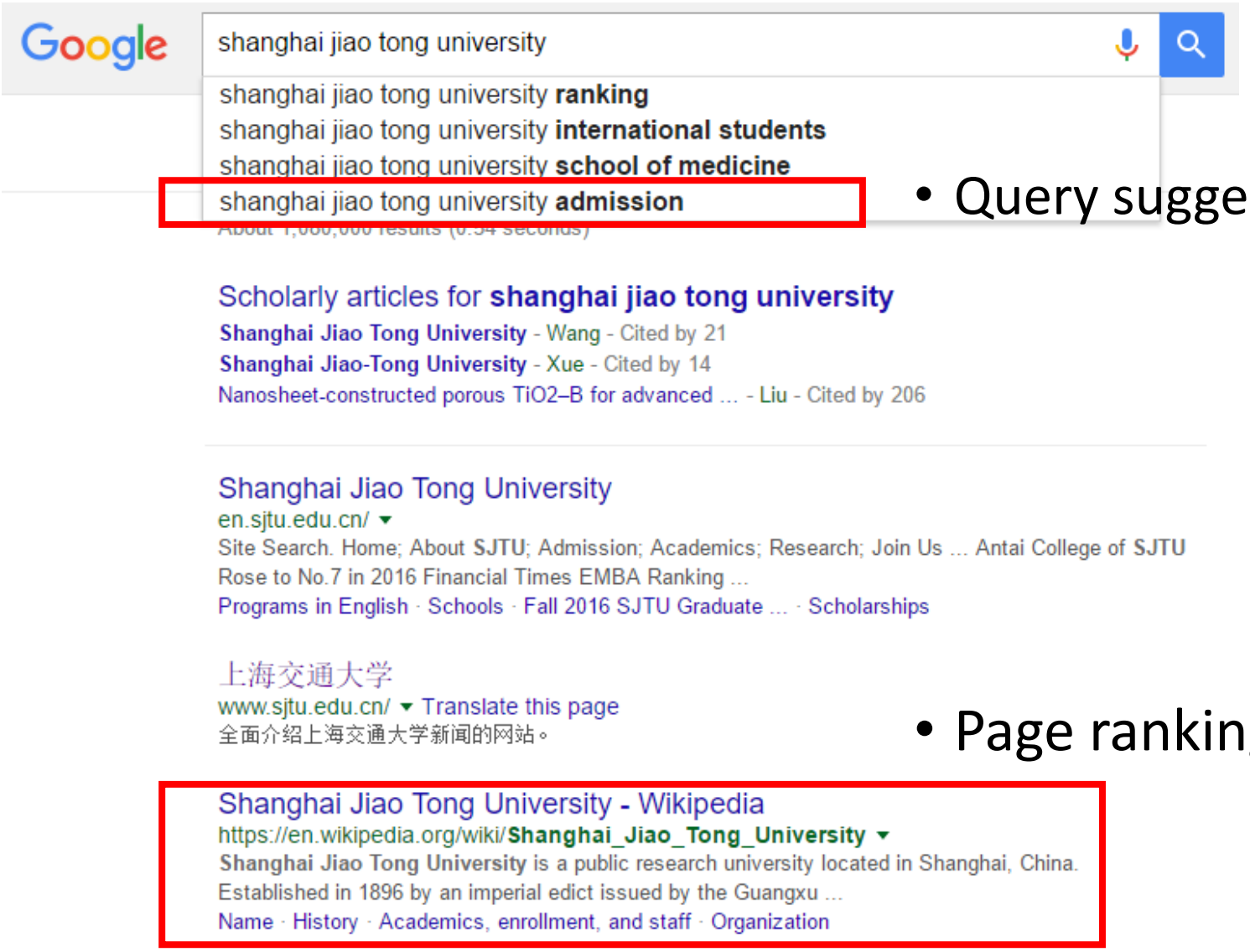
$$\{\text{milk, bread}\} \Rightarrow \{\text{butter}\}$$
$$\{\text{onion, potatoes}\} \Rightarrow \{\text{burger}\}$$

Some non-intuitive ones:

$$\{\text{diaper}\} \Rightarrow \{\text{beer}\}$$

Agrawal, R.; Imieliński, T.; Swami, A. (1993). "Mining association rules between sets of items in large databases". *ACM SIGMOD 1993*

DM Use Case 2: Web Search



- Query suggestion
- Page ranking

DM Use Case 3: News Recommendation

美国大选 + 关注

"特朗普时代"的中美新局

周浩：美国在经济上强势可能带来外交上相对弱势，美国可能给予中国在亚太更大的主动权，以为经济发展蓄势。

收藏

更新于2016年11月16日 07:07 德国商业银行首席中国经济师 周浩 为英国《金融时报》中文网撰稿

特朗普强势当选美国总统，给全市场留下了一个费解的难题：到底这位特立独行的美国白人会给世界带来怎样的变化，而未来世界格局中，中美两大经济体又将会以怎样的方式来进行互动。

到目前为止，我们只能通过特朗普在竞选过程中的讲话，部分了解未来美国政策的走向。比如说，特朗普反对TPP，认为目前的全球化策略并没有能够解决美国企业的困境，并表示要对中国商品征收45%的关税，同时要在美国和墨西哥边境建造"长城"来防止非法移民。特朗普也反对美国目前的世界警察角色，认为这给美国普通家庭带来了负担和悲痛，这意味着美国在全球战略布局中将更多采取收缩策略。此外，特朗普认为美国的能源政策和医疗保险制度是个灾难，认为政府插手太多，造成了巨大的浪费。

- Predict whether a user will like a news given its reading context

您可能感兴趣的文章

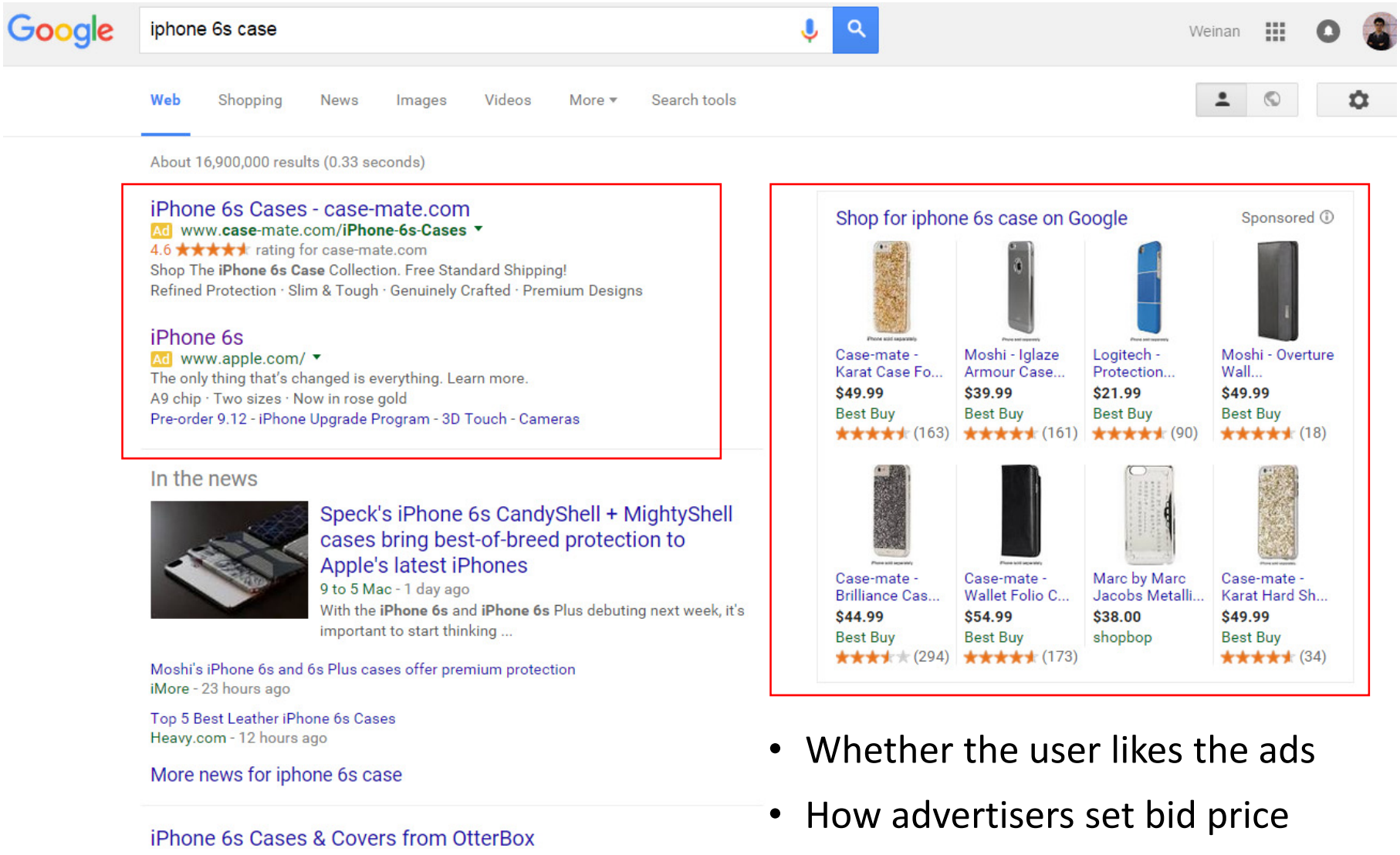
焦虑与希望——选后华盛顿侧记

这是特朗普的1966年

特朗普能被政治精英驯服吗？

从特朗普胜选看美国政治

DM Use Case 4: Sponsored Search



- Whether the user likes the ads
- How advertisers set bid price

DM Use Case 5: Displayed Advertising

- Whether the user likes the ads
- How advertisers set bid price

https://github.com/wnzhang/rtb-papers

DM Use Case 6: Information Extraction

Kinect - Fastest Selling Electronic Product in History

Posted on: 3/10/2011 1:09:45 PM by **David Lewis**

Microsoft's Kinect sensor system has been officially recognised as the fastest selling electrical device in history.

Manufactured to give wireless interactivity with the company's Xbox game platform, the device has sold eight million units in its first two months, outstripping the sales of Apple's iPhone and iPad when they were launched.

The news comes as a welcome relief for Microsoft who have been trailing Apple in the technology stakes over the last few years with the Apple brand being seen as more cool and sexy than Microsoft.

The figures, which have been verified by the Guinness Book of World Records, represent sales of the camera add-on which uses infrared technology to track the movement of the participant and translate their movements to action in the game.

For some time Microsoft's Xbox was at a disadvantage to Nintendo's Wii system because of the lack of a motion detector but the Kinect addresses the issue well. Microsoft were keen on using a different technological base for their system to avoid being accused of copyright infringement and so the solution was built around infrared technology.

Microsoft says that sales of the Kinect reflect the popularity of the games platform in comparison with the Wii and hope that the availability of Kinect will also boost sales of the Xbox itself.

It notes that sales of games for the Xbox have also rocketed since the device became available with total sales now exceeding ten million.

In January Microsoft reported profits of $6.63bn (£4.1bn) for the last three months of 2010, down from $6.66bn a year earlier despite the excellent sales performance of Kinect.

Posted: 3/10/2011 1:09:45 PM by **David Lewis** | with 0 comments

Webpage

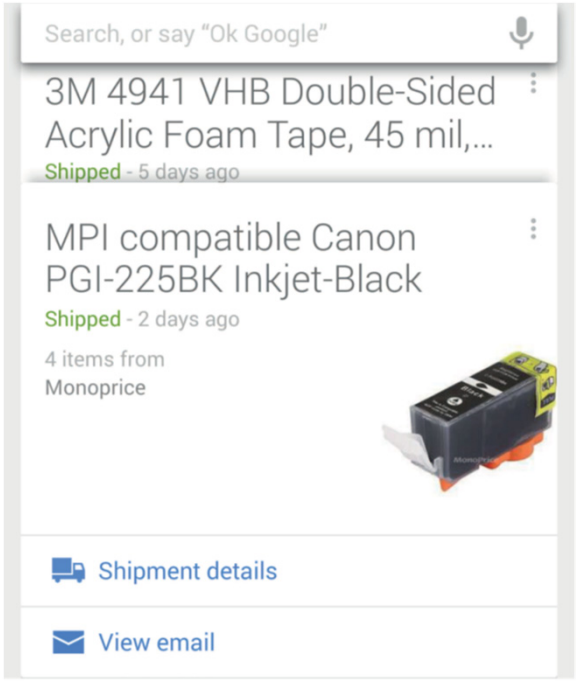
→

Kinect
Electronic Product
Microsoft's Xbox
Games
Xbox Game Platform
•
•
•

Keywords

DM Use Case 7: Information Extraction

- Structural information extraction and illustration



Gmail

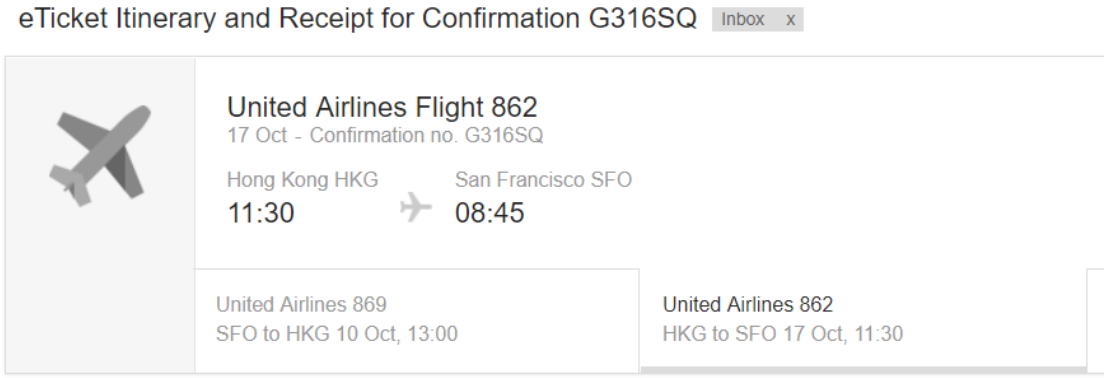
Google Now

Zhang, Weinan, et al. "Annotating needles in the haystack without looking: Product information extraction from emails." *KDD* 2015.

DM Use Case 7: Information Extraction

- Structural information extraction and illustration

Google Now

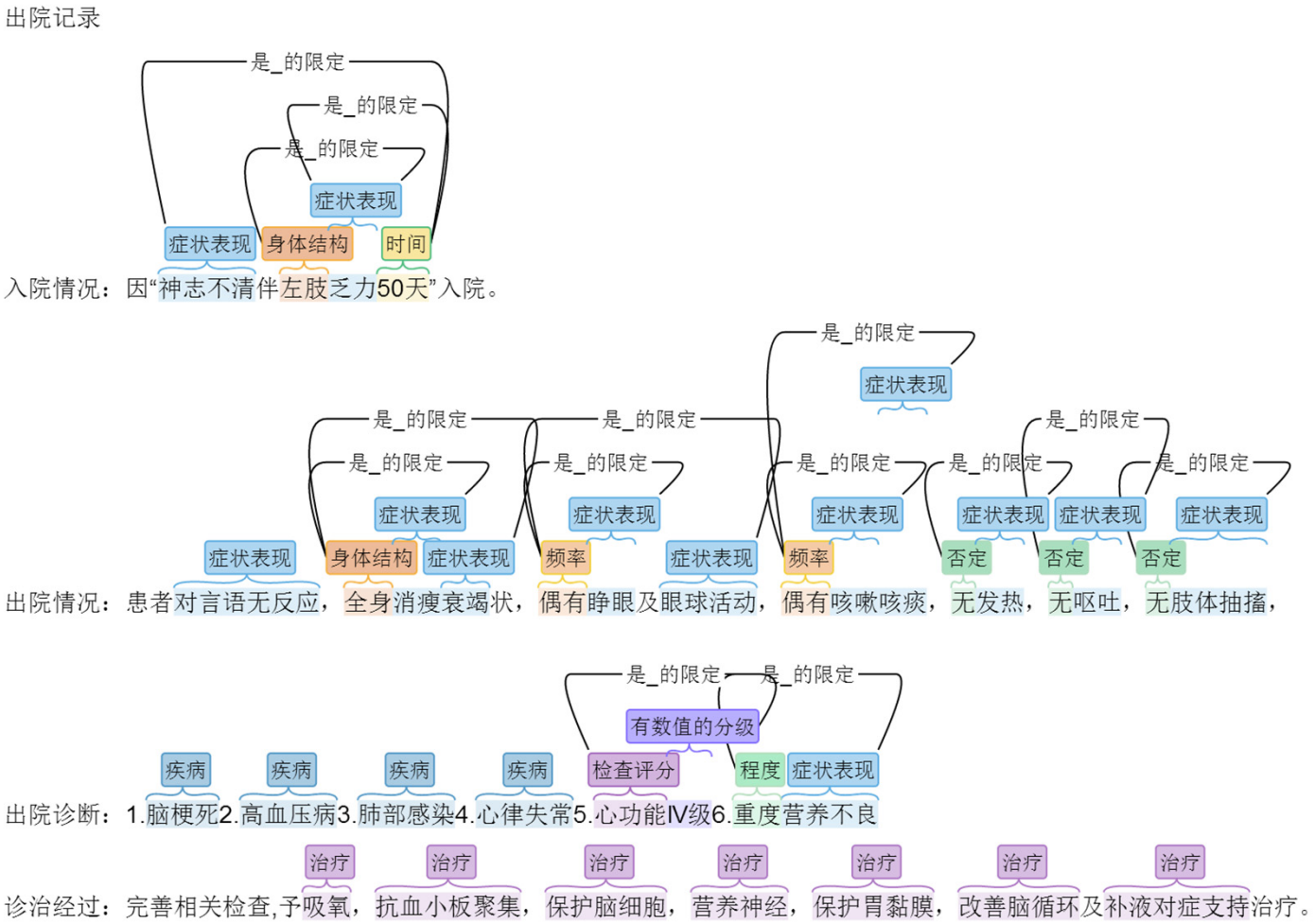


Gmail

Zhang, Weinan, et al. "Annotating needles in the haystack without looking: Product information extraction from emails." *KDD* 2015.

DM Use Case 7: Information Extraction

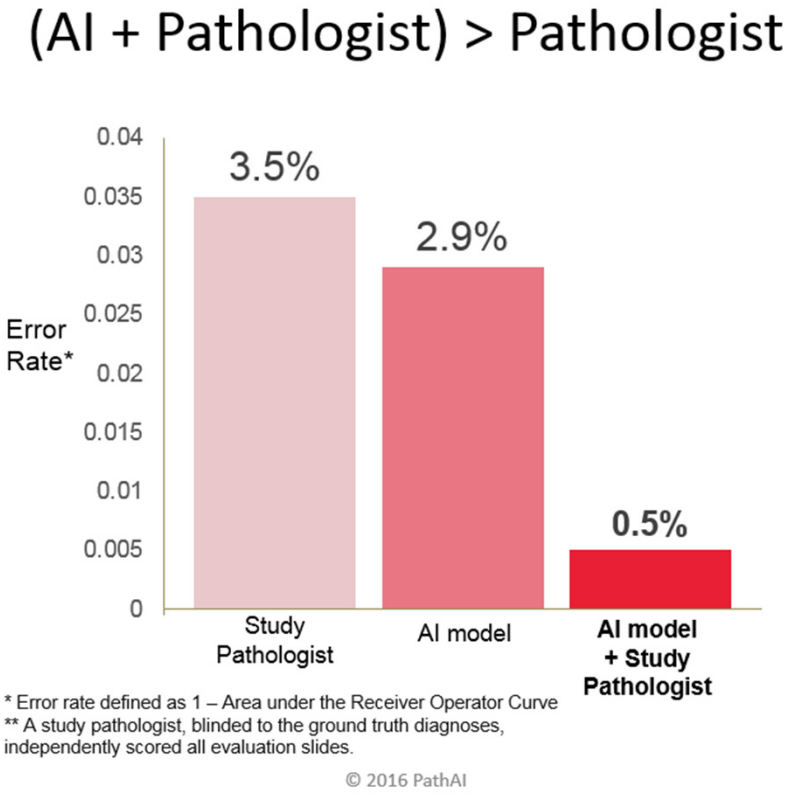
- Synyi.com medical structural information extraction



Wang, Zhenghui et al. Label-aware Double Transfer Learning for Cross-Specialty Medical Named Entity Recognition. NAACL 2018.

DM Use Case 8: Medical Image Analysis

- Breast Cancer Diagnoses



Wang, Dayong, et al. "Deep learning for identifying metastatic breast cancer." arXiv preprint arXiv:1606.05718 (2016).
https://blogs.nvidia.com/blog/2016/09/19/deep-learning-breast-cancer-diagnosis/

DM Use Case 8: Clinic Medicine Data Mining

- Predict the patient's health (e.g. diabetes) after 3 years given the current internal secretion test results

Clinic tests

Factors Associated With Patients' Adherence To Anti-Diabetic Medications

I. Baseline characteristics:

Age: Gender: Male/Female Education: Occupation: Nationality: Marital status:

II. Profile of Diabetes Mellitus

1. Duration of diabetes mellitus 2. Age at onset:: 3. Family history of diabetes: Yes/No

III. Patient adherence to drug therapy

1. Do you take the anti-diabetic drugs as advised by your doctor? Yes/No If No, please tick the options [✓]

Items	Yes	No	Items	Yes	No
Lack of finance			Side effects		
Feeling drug is not effective			Feeling the dose given is high		
Interferes with my meal plan			Complexity of drug regimen		
Taking them since many years			Multiple medications		
I forget			Poor family support		

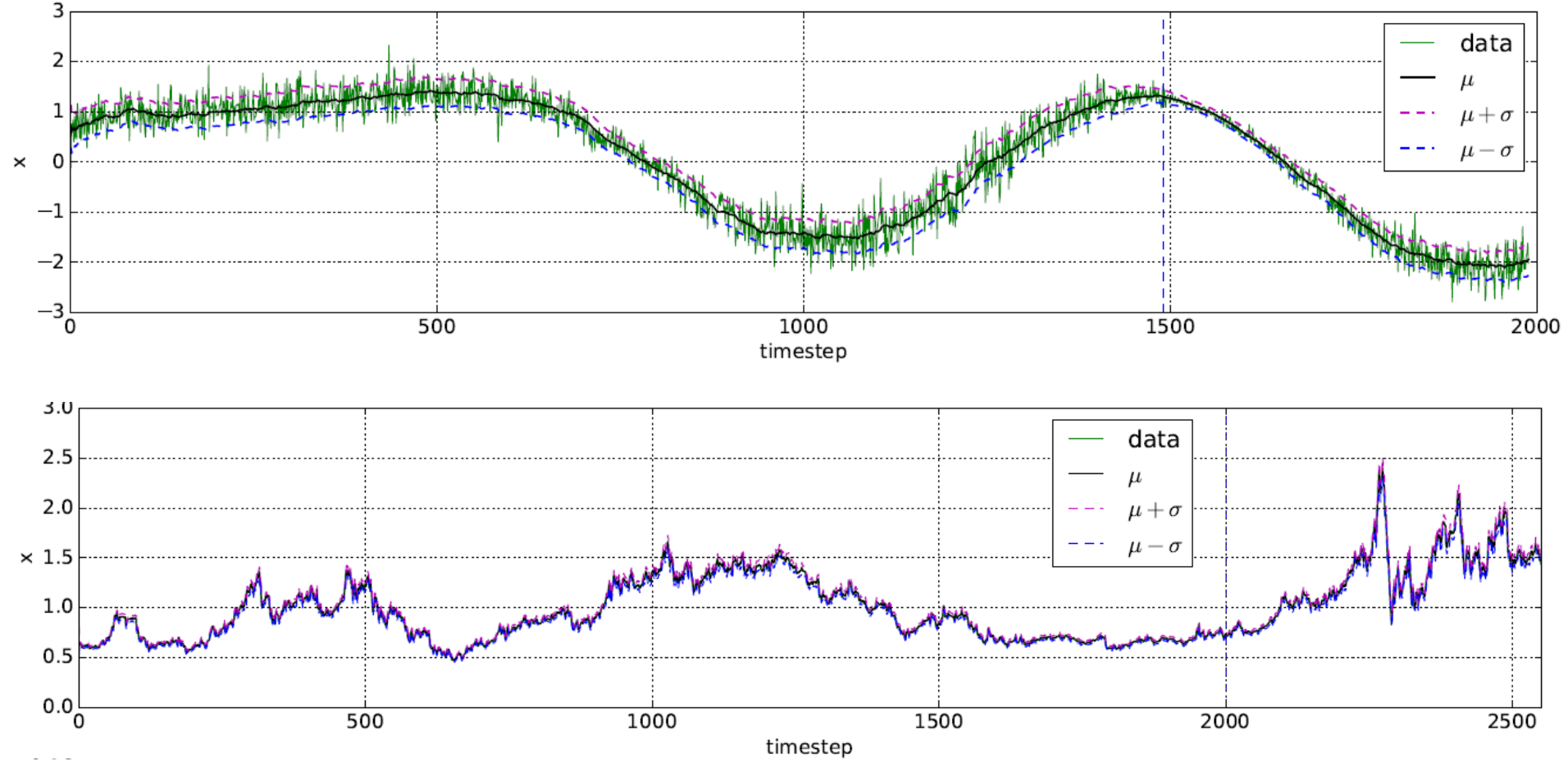
Items	Yes	No
Do you regularly monitor your blood glucose?		
Do you make your own modification in the dose of drugs prescribed?		
Do you make your own modification in the timing of anti-diabetic drugs?		
Do you have good knowledge about anti-diabetic medications prescribed to you?		
Do you know the importance of anti-diabetic medication		
Did your physician give information on diabetes		
Did your physician give information on anti-diabetic medications		
Were you involved in treatment decisions		
Do you feel comfortable to ask questions to your doctor		

Questionnaires

- Explainable patterns are always desirable for clinic medicine to provide informative guidance to doctors

DM Use Case 9: Financial Data Prediction

- Predict the trend and volatility of financial time series data



DM Use Case 10: Social Networks

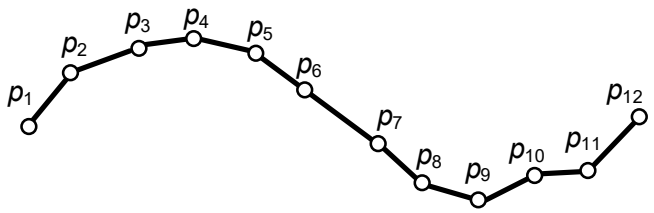
- Community detection / node classification
- Information diffusion modeling
- Friends/Tweets/Job Candidates suggestion

DM Use Case 11: Spatio-Temporal DM

- A spatio-temporal trajectory

$$p_1 \rightarrow p_2 \rightarrow \cdots \rightarrow p_n$$

$$p_i = (x, y, t)$$



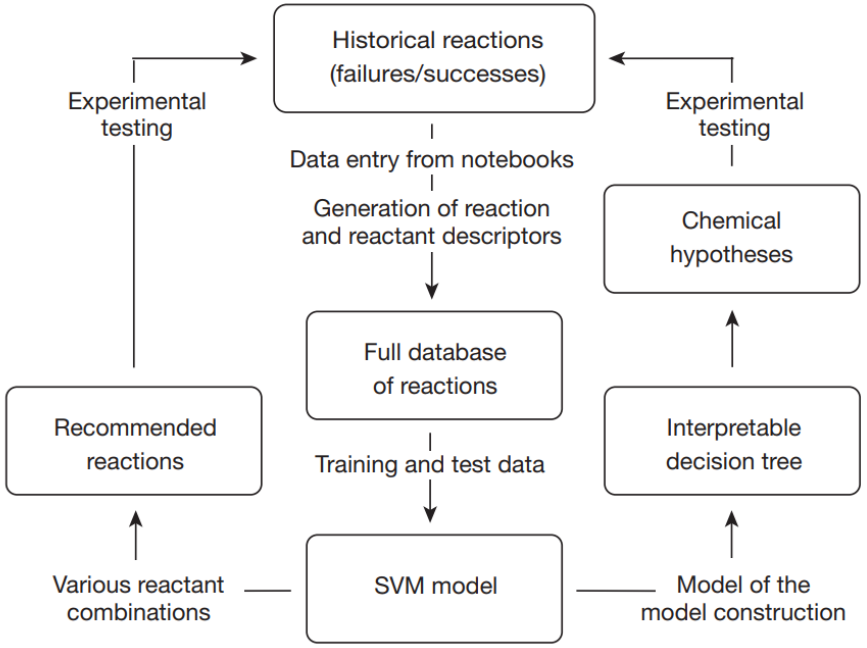
- Behavior modeling of humans and vehicles in the cities
- Prediction of human / vehicles / environment in a certain spatio-temporal point
- Optimization including car route scheduling, lane design, factory relocation

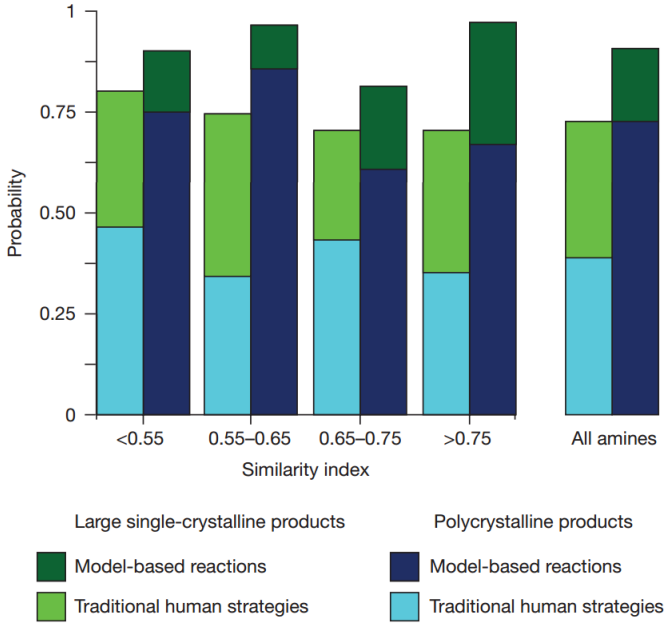
https://www.microsoft.com/en-us/research/project/trajectory-data-mining/

Slide credit: Yu Zheng

DM Use Case 12: New Material Discovery

- Driven by Materials Genome Initiative
- Mine the underlying patterns between the experiment conditions and the properties of the resulted material

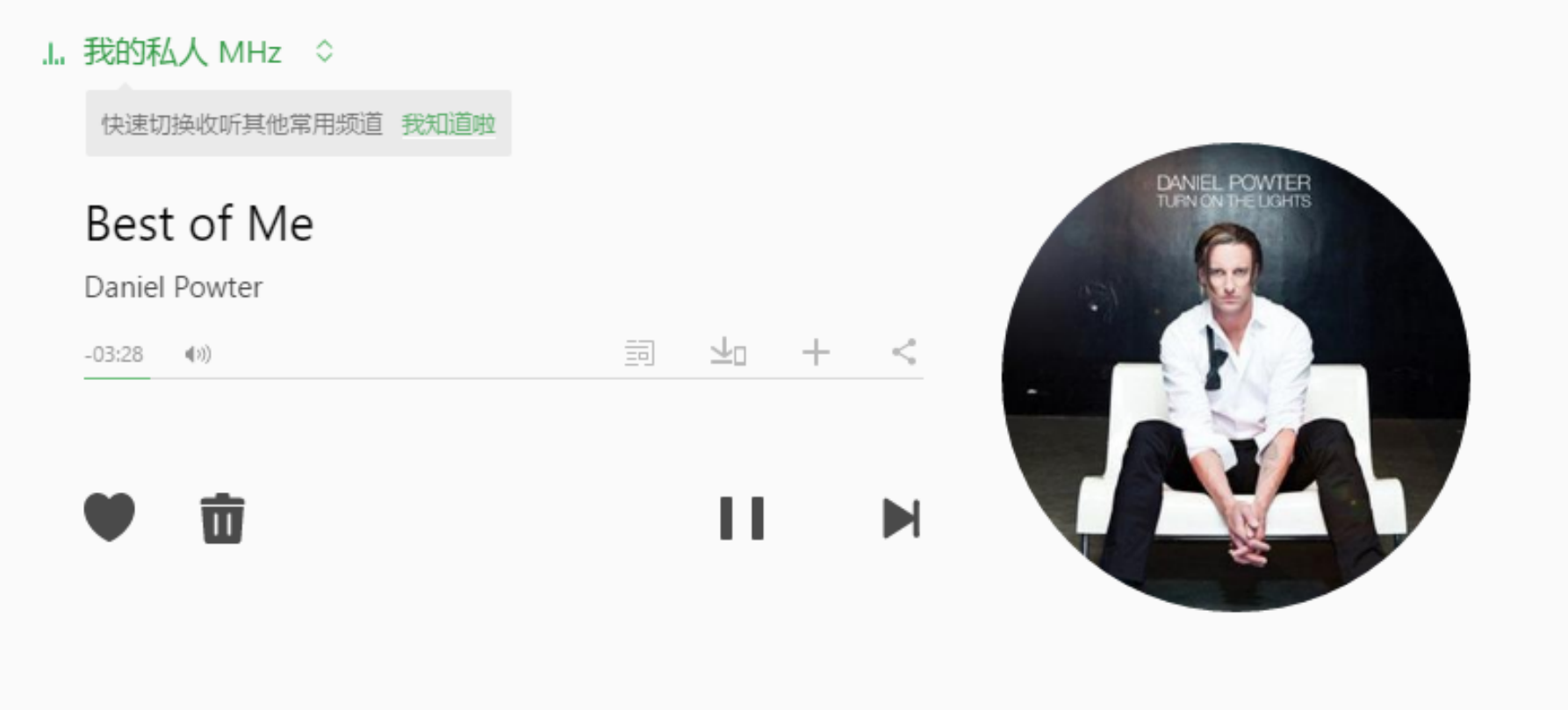




Raccuglia, Paul, et al. "Machine-learning-assisted materials discovery using failed experiments." *Nature* 533.7601 (2016): 73-76.

DM Use Case 13: Interactive Recommendation

- Douban.fm music recommend and feedback
  - The machine needs to make decisions, not just prediction



Summary of This Lecture

- An example as an intro of data mining
- Concepts of data mining
- Real-world examples of data mining

- Data mining is about the extraction of non-trivial, implicit, previously unknown and potentially useful principles, patterns or knowledge from massive amount of data.