Exploring Data

CSESS Seminar
(7 Oct, 2021)

Prepared by Raymond Wong
Presented by Raymond Wong
raywong@cse

- Do you think that you could earn "US$1 billion" by doing the gambling on the HK horse racing?

Let me give some information to you.

If you are the single winner for the Triple Trio, you could obtain ~US$13 million

There are more than 10 million combinations

What is the total number of times we need to be a single winner of Triple Trio?

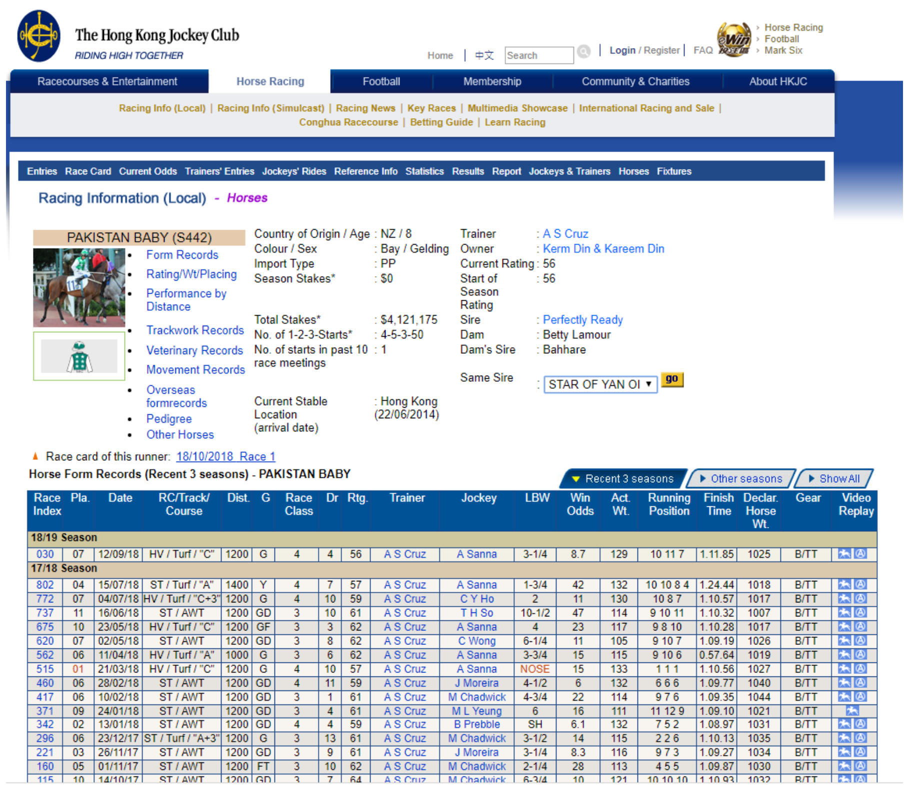
= 1,000/13 = 76.92 = ~77

- Do you think that you could earn "US$1 billion" by doing the gambling on the HK horse racing?

Someone did that!

Bill Benter used the data analysis technology to do that!

- Bill Benter collected a lot of statistics about the “horse performance” in the past.


The Hong Kong Jockey Club
RIDING HIGH TOGETHER
Home | 中文 | Search | Login / Register | FAQ
Horse Racing
Football
Mark Six

Racecourses & Entertainment | Horse Racing | Football | Membership | Community & Charities | About HKJC

Racing Info (Local) | Racing Info (Simulcast) | Racing News | Key Races | Multimedia Showcase | International Racing and Sale | Conghua Racecourse | Betting Guide | Learn Racing

Entries Race Card Current Odds Trainers' Entries Jockeys' Rides Reference Info Statistics Results Report Jockeys & Trainers Horses Fixtures

Racing Information (Local) - Horses

PAKISTAN BABY (S442)

- Form Records
- Rating/Wt/Placing
- Performance by Distance
- Trackwork Records
- Veterinary Records
- Movement Records
- Overseas formrecords
- Pedigree
- Other Horses

Country of Origin / Age : NZ / 8
Colour / Sex : Bay / Gelding
Import Type : PP
Season Stakes* : $0
Total Stakes* : $4,121,175
No. of 1-2-3-Starts* : 4-5-3-50
No. of starts in past 10 race meetings : 1
Current Stable Location (arrival date) : Hong Kong (22/06/2014)

Trainer : A S Cruz
Owner : Kerm Din & Kareem Din
Current Rating : 56
Start of Season Rating : 56
Sire : Perfectly Ready
Dam : Betty Lamour
Dam's Sire : Bahhare
Same Sire : STAR OF YAN OI go

Race card of this runner: 18/10/2018 Race 1

Horse Form Records (Recent 3 seasons) - PAKISTAN BABY

Recent 3 seasons | Other seasons | Show All

Race Index	Pla.	Date	RC/Track/ Course	Dist.	G	Race Class	Dr	Rtg.	Trainer	Jockey	LBW	Win Odds	Act. Wt.	Running Position	Finish Time	Declar. Horse Wt.	Gear	Video Replay
18/19 Season																		
030	07	12/09/18	HV / Turf / "C"	1200	G	4	4	56	A S Cruz	A Sanna	3-1/4	8.7	129	10 11 7	1.11.85	1025	B/TT	
17/18 Season																		
802	04	15/07/18	ST / Turf / "A"	1400	Y	4	7	57	A S Cruz	A Sanna	1-3/4	42	132	10 10 8 4	1.24.44	1018	B/TT	
772	07	04/07/18	HV / Turf / "C+3"	1200	G	4	10	59	A S Cruz	C Y Ho	2	11	130	10 8 7	1.10.57	1017	B/TT	
737	11	16/06/18	ST / AWT	1200	GD	3	10	61	A S Cruz	T H So	10-1/2	47	114	9 10 11	1.10.32	1007	B/TT	
675	10	23/05/18	HV / Turf / "C"	1200	GF	3	3	62	A S Cruz	A Sanna	4	23	117	9 8 10	1.10.28	1017	B/TT	
620	07	02/05/18	ST / AWT	1200	GD	3	8	62	A S Cruz	C Wong	6-1/4	11	105	9 10 7	1.09.19	1026	B/TT	
562	06	11/04/18	HV / Turf / "A"	1000	G	3	6	62	A S Cruz	A Sanna	3-3/4	15	115	9 10 6	0.57.64	1019	B/TT	
515	01	21/03/18	HV / Turf / "C"	1200	G	4	10	57	A S Cruz	A Sanna	NOSE	15	133	1 1 1	1.10.56	1027	B/TT	
460	06	28/02/18	ST / AWT	1200	GD	4	11	59	A S Cruz	J Moreira	4-1/2	6	132	6 6 6	1.09.77	1040	B/TT	
417	06	10/02/18	ST / AWT	1200	GD	3	1	61	A S Cruz	M Chadwick	4-3/4	22	114	9 7 6	1.09.35	1044	B/TT	
371	09	24/01/18	ST / AWT	1200	GD	3	4	61	A S Cruz	M L Yeung	6	16	111	11 12 9	1.09.10	1021	B/TT	
342	02	13/01/18	ST / AWT	1200	GD	4	4	59	A S Cruz	B Prebble	SH	6.1	132	7 5 2	1.08.97	1031	B/TT	
296	06	23/12/17	ST / Turf / "A+3"	1200	G	3	13	61	A S Cruz	M Chadwick	3-1/2	14	115	2 2 6	1.10.13	1035	B/TT	
221	03	26/11/17	ST / AWT	1200	GD	3	9	61	A S Cruz	J Moreira	3-1/4	8.3	116	9 7 3	1.09.27	1034	B/TT	
160	05	01/11/17	ST / AWT	1200	FT	3	10	62	A S Cruz	M Chadwick	2-1/4	28	113	4 5 5	1.09.87	1030	B/TT	
115	10	14/10/17	ST / AWT	1200	GD	3	7	64	A S Cruz	M Chadwick	6-3/4	10	121	10 10 10	1.10.93	1032	B/TT	

- He used some data analysis technology to predict the “next” horse racing result.

data analysis = US$1 billion!

- We have just discussed one successful story of using data analysis.
- Let us discuss one more successful story of using data analysis.

AlphaGo

What is it?

a computer program with data analysis developed by Google DeepMind.

Who is he?

Ke Jie (a 9-dan professional)
World GO number ONE in 2017

AlphaGo

They competed each other in March 2017.
Finally, who won the game?

AlphaGo

- AlphaGo collected a lot of statistics about the “GO player performance” in the past.
- It used some data analysis technology to predict the “next” move for each round in the game.

data analysis “wins” human!

- Maybe, you are interested in why AlphaGo may win the game.
- Let us describe one basic game related to this.
- The details of AlphaGo may be known by you when you study some courses about data analysis.

Gambling

- Suppose that Raymond has enough money and enough time for gambling.
- Consider that Raymond wants to do a gambling.
- The gambling game has only two possible outcomes, namely “large” (with probability = 0.5) and “small” (with probability = 0.5).
- Raymond could play the gambling game multiple times by always guessing that the outcome is “large”.
- Is it always true that Raymond must earn money at the end with a “smart” strategy by playing a number of times of the gambling game?

Yes.

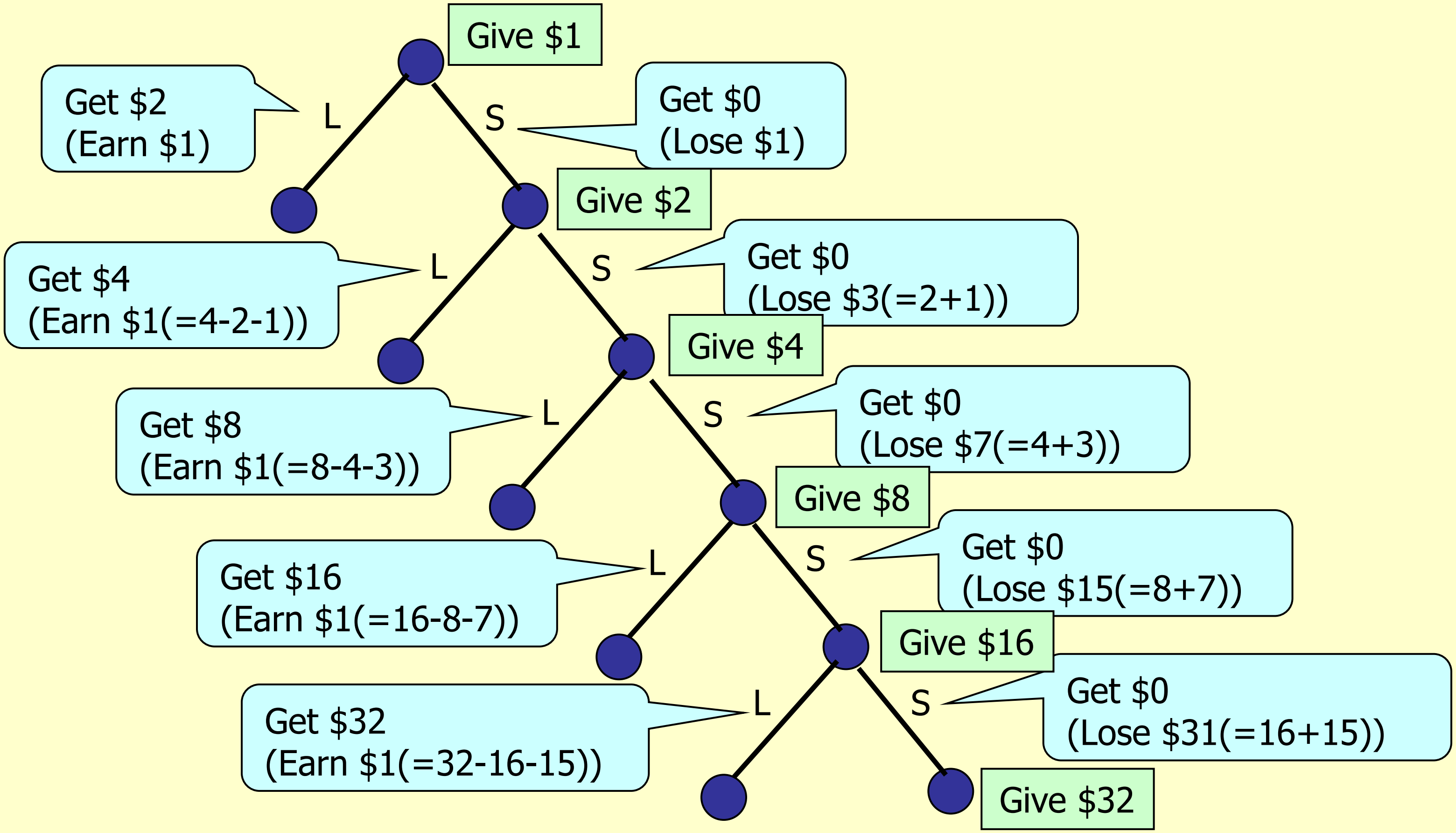
Gambling

- Suppose that Raymond has enough money and enough time for gambling.
- Consider that Raymond wants to do a gambling.
- The gambling game has only two possible outcomes, namely "large" (with probability = 0.3) and "small" (with some probabilities).
- Raymond could play the gambling game multiple times by always guessing that the outcome is "large".
- Is it always true that Raymond must earn money at the end with a "smart" strategy by playing a number of times of the gambling game?

Yes.

Gambling

- Consider the following.
  - For each round,
    - I should give $1 to play the game
    - If I lose the game,
      I will get $0 (i.e., will lose $1).
    - If I win the game,
      I will get $2 (i.e., will earn $1).

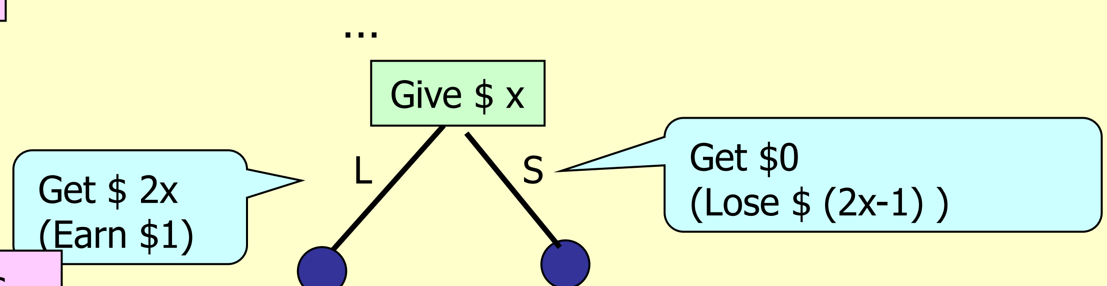


Finally, Raymond must earn $1 at the end.

However, Raymond needs to spend a lot of time and a lot of money.

This is true when P(L) = 0.5.

This is true when P(L) = 0.3 or other values.



Gambling

- What is the "expected" number of rounds that Raymond could earn \$1 when $P(L) = 0.5$?

The expected number of rounds $= 1/0.5$
$= 2$

Gambling

- When $P(L) = 0.5$, is it always true that Raymond could earn \$1 after playing 2 rounds?

  No

- When $P(L) = 0.5$, is it always true that Raymond could earn \$1 after playing 1000 rounds?

  No

Gambling

- What is the “expected” number of rounds that Raymond could earn \$1 when $P(L) = 0.3$?

The expected number of rounds = 1/0.3
=3.33

Gambling

- Do you think that the casino must lose?

No. In some casinos there are the following rules.
1. For each game, the player has a “minimum” amount of the money (e.g., \$100) for playing. This means that the player has to bring more money to the casino.

2. For each game, the player has a maximum amount of the money (e.g., \$1000) for playing. This means that the player could not play the game with a large amount of money.

Gambling

- Caution
  - Each player may need to spend a lot of time (maybe, more than 1 day)
  - Each player may need to spend a lot of money (more than what the player has).

Is it valuable to play a time-consuming game to win $1?

- The probabilities (0.5 or 0.3) could be learnt from the past data using the data analysis technology
- No matter how "accurate" the probabilities could be learnt from the past data, Raymond must always win the game!

- This concept is NOT restricted to the gambling.
- This could also be applied to the stock market for investment where “large” could be replaced by “up” and “small” could be replaced by “down”.

- Now, we know two successful stories of using data analysis.
- Next, let us see a case study of how to analyze data analysis.

Singapore Taxis

- In Singapore, each taxi is equipped with a sensor.
- We know the path/trajectory of each taxi.
- We can collect the location of each taxi every second.
- There are a lot of points generated.

Singapore Taxis

- During rainy days, hard to find taxi in Singapore

Why?

- **EXPLANATION 1:** Taxis are slow to avoid accident
- **DATA:** GPS location of taxis are often on roadside; few cars are on the road

- **EXPLANATION 2:** Higher customer need
- **DATA:** Taxi income drops significantly

- **FACT:** A Singapore law says that higher penalty is imposed for car accidents that happen in rain

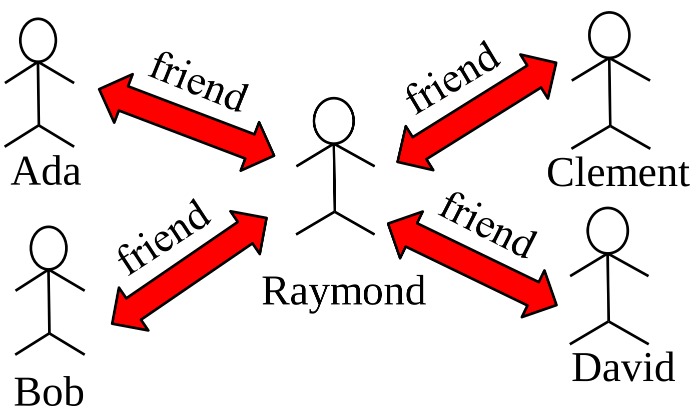
- Let us see one more successful story of using social networks.

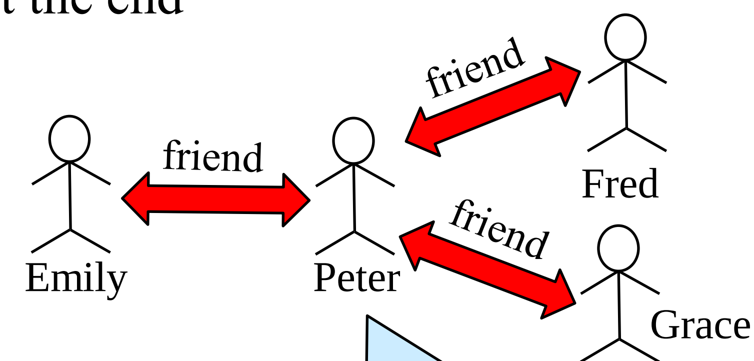
- In 2007, nobody knew the top-secret of "The Wizarding World of Harry Potter" (i.e., building a theme park in Orlando).
- The Marketing Manager of this project did not spend money on TV or other media for advertisement.

- 7 top fans of "Harry Potter" were invited to participate in a top-secret Webcast held at midnight on May 31, 2007.
- Finally, 350,000,000 knew this secret!
- 7 = 350,000,000!!

Finding "Good" People for Marketing

Objective: to find a **limited** number of people for marketing in order to "influence" as many people as possible at the end

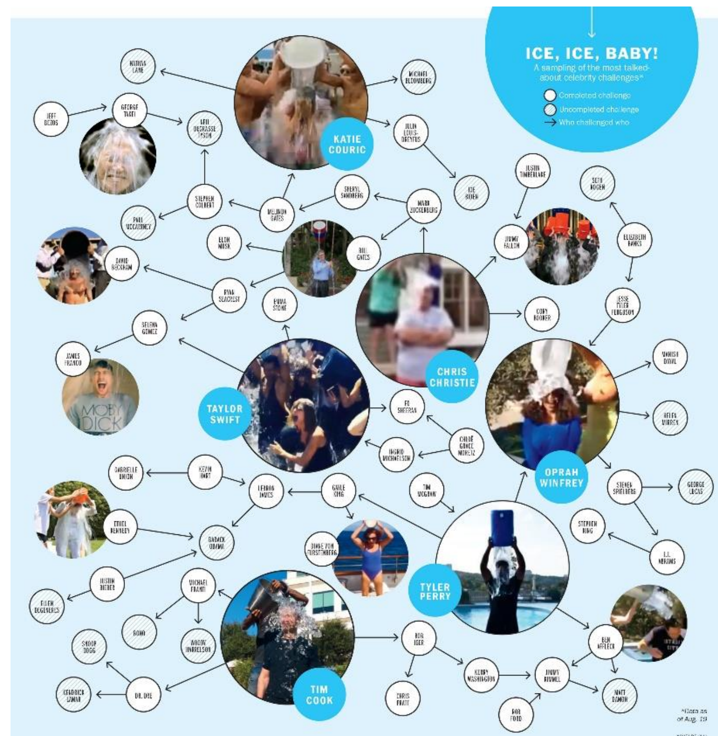




Which two people should I choose for marketing?

Other Applications

- Ice Bucket Challenge



- Next, let us see a list of data analysis topics.

Major Topics

1. Association
2. Clustering
3. Classification
4. Data Warehouse
5. Web Databases

1. Association

Customer	Apple	Orange	Milk
Raymond	Apple	Orange	
Ada		Orange	Milk
Grace	Apple	Orange	
…	…	…	…

We are interested in the items/itemsets with frequency >= 2

Items/Itemsets	Frequency
Apple	2
Orange	3
Milk	1
{Apple, Orange}	2
{Orange, Milk}	1

Frequent Pattern (or Frequent Item)

Frequent Pattern (or Frequent Item)

Frequent Pattern (or Frequent Itemset)

1. Association

Customer	Apple	Orange	Milk
Raymond	Apple	Orange	
Ada		Orange	Milk
Grace	Apple	Orange	
…	…	…	

We are interested in the items/itemsets with frequency >= 2

Items/Itemsets	Frequency
Apple	2
Orange	3
Milk	1
{Apple, Orange}	2

Association Rule:

1. Apple → Orange
(100% customers who buy apple will probably buy orange.)

2. Orange → Apple
(67% customer who buy orange will probably buy apple.)

Problem: to find all frequent patterns and association rules

1. Association

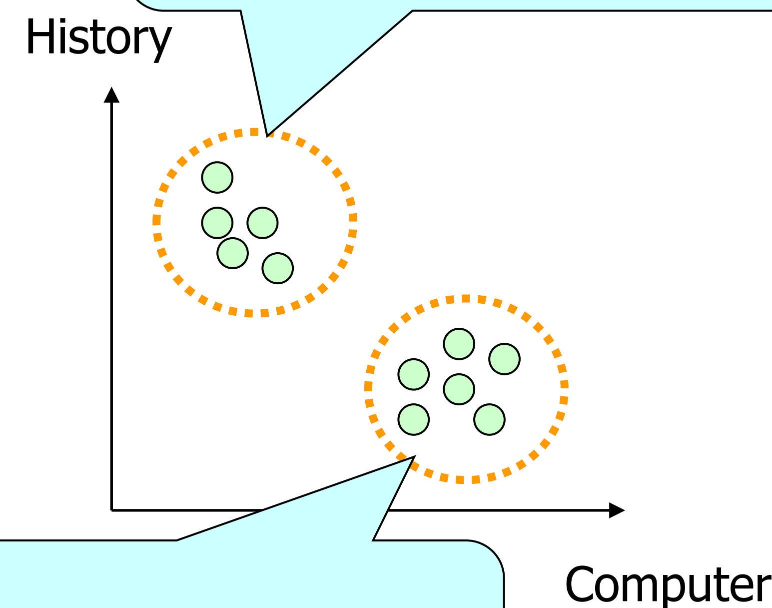
- Applications of Association Rule Mining
  - Supermarket
  - Web Mining
  - Medical analysis
  - Bioinformatics
  - Network analysis
    (e.g., Denial-of-service (DoS))
  - Programming Pattern Finding

Major Topics

1. Association
2. Clustering
3. Classification
4. Data Warehouse
5. Web Databases

2. Clustering

	Computer	History
Raymond	100	40
Louis	90	45
Wyman	20	95
…	…	…



Problem: to find all clusters

2. Clustering

- Clustering for Understanding
  - Applications
    - Biology
      - Group different species
    - Psychology and Medicine
      - Group medicine
    - Business
      - Group different customers for marketing
    - Network
      - Group different types of traffic patterns
    - Software
      - Group different programs for data analysis

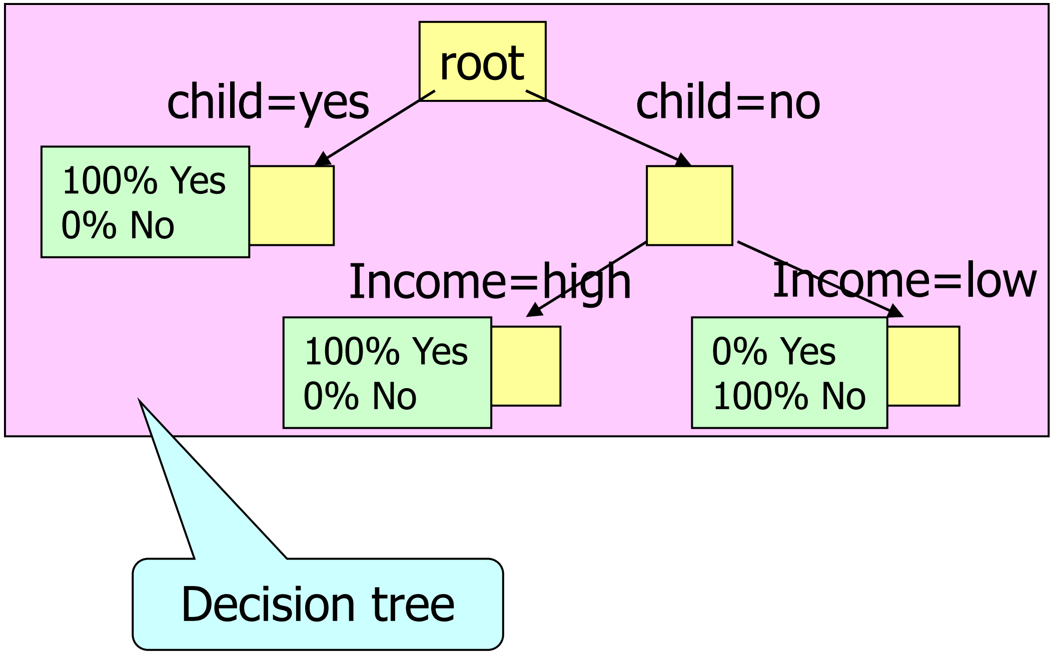
Major Topics

1. Association
2. Clustering
3. Classification
4. Data Warehouse
5. Web Databases

3. Classification

Suppose there is a person.

Race	Income	Child	Insurance
white	high	no	?



Applications

- Insurance
  - According to the attributes of customers,
    - Determine which customers will buy an insurance policy
- Marketing
  - According to the attributes of customers,
    - Determine which customers will buy a product such as computers
- Bank Loan
  - According to the attributes of customers,
    - Determine which customers are “risky” customers or “safe” customers

Applications

- Network
  - According to the traffic patterns,
    - Determine whether the patterns are related to some “security attacks”
- Software
  - According to the experience of programmers,
    - Determine which programmers can fix some certain bugs

Applications

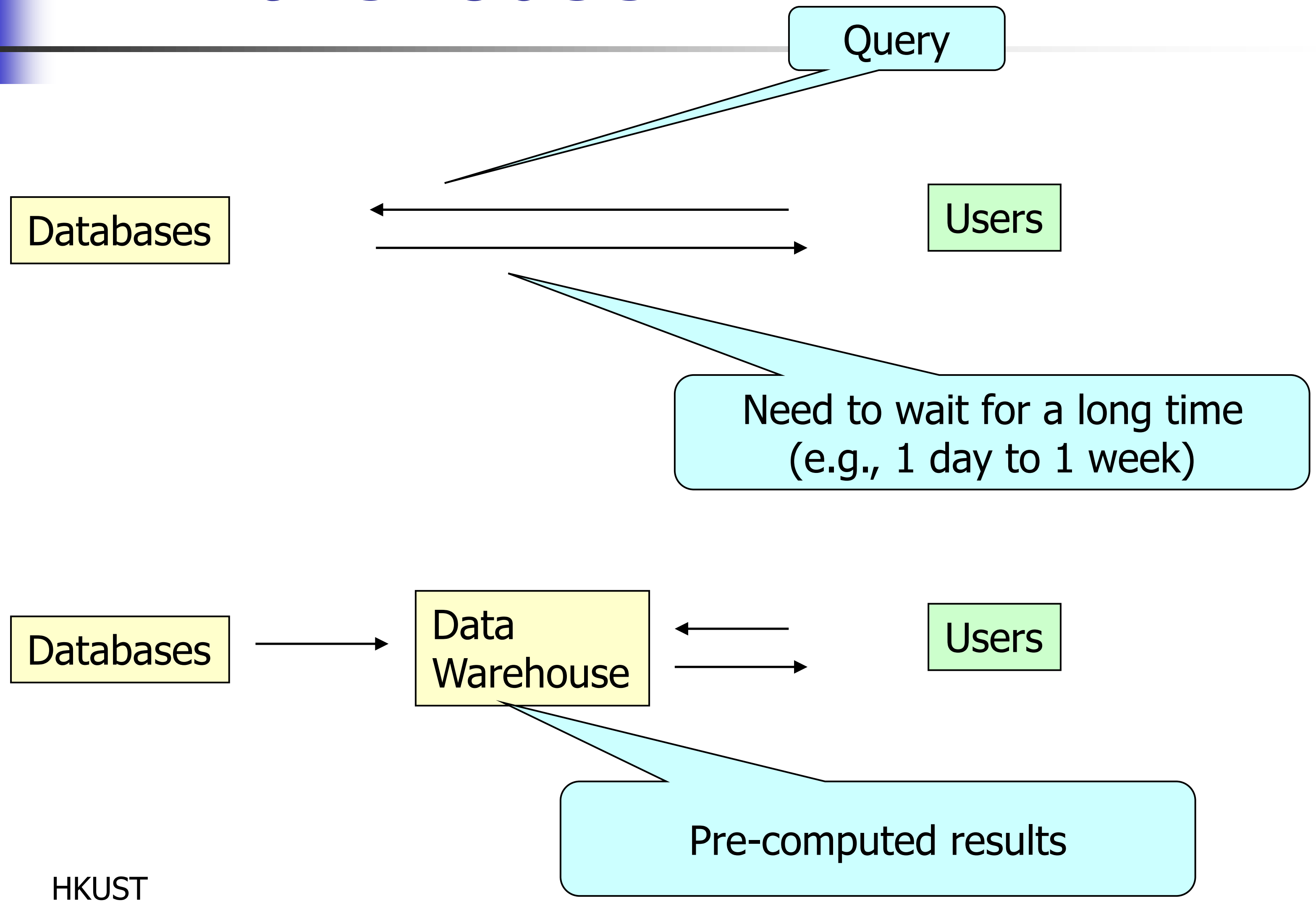
- We could have a more “general” application.
- It is NOT just to determine whether something is related to “yes” or “no”.
- E.g., Automatic Image Caption Generation


A person riding a motorcycle on a dirt road.
Two dogs play in the grass.
A skateboarder does a trick on a ramp.
A dog is jumping to catch a frisbee.
A group of young people playing a game of frisbee.
Two hockey players are fighting over the puck.
A little girl in a pink hat is blowing bubbles.
A refrigerator filled with lots of food and drinks.
A herd of elephants walking across a dry grass field.
A close up of a cat laying on a couch.
A red motorcycle parked on the side of the road.
A yellow school bus parked in a parking lot.
Describes without errors
Describes with minor errors
Somewhat related to the image
Unrelated to the image

Major Topics

1. Association
2. Clustering
3. Classification
4. Data Warehouse
5. Web Databases

4. Warehouse


Query
Databases
Users
Need to wait for a long time
(e.g., 1 day to 1 week)
Databases
Data
Warehouse
Users
Pre-computed results

Advantages

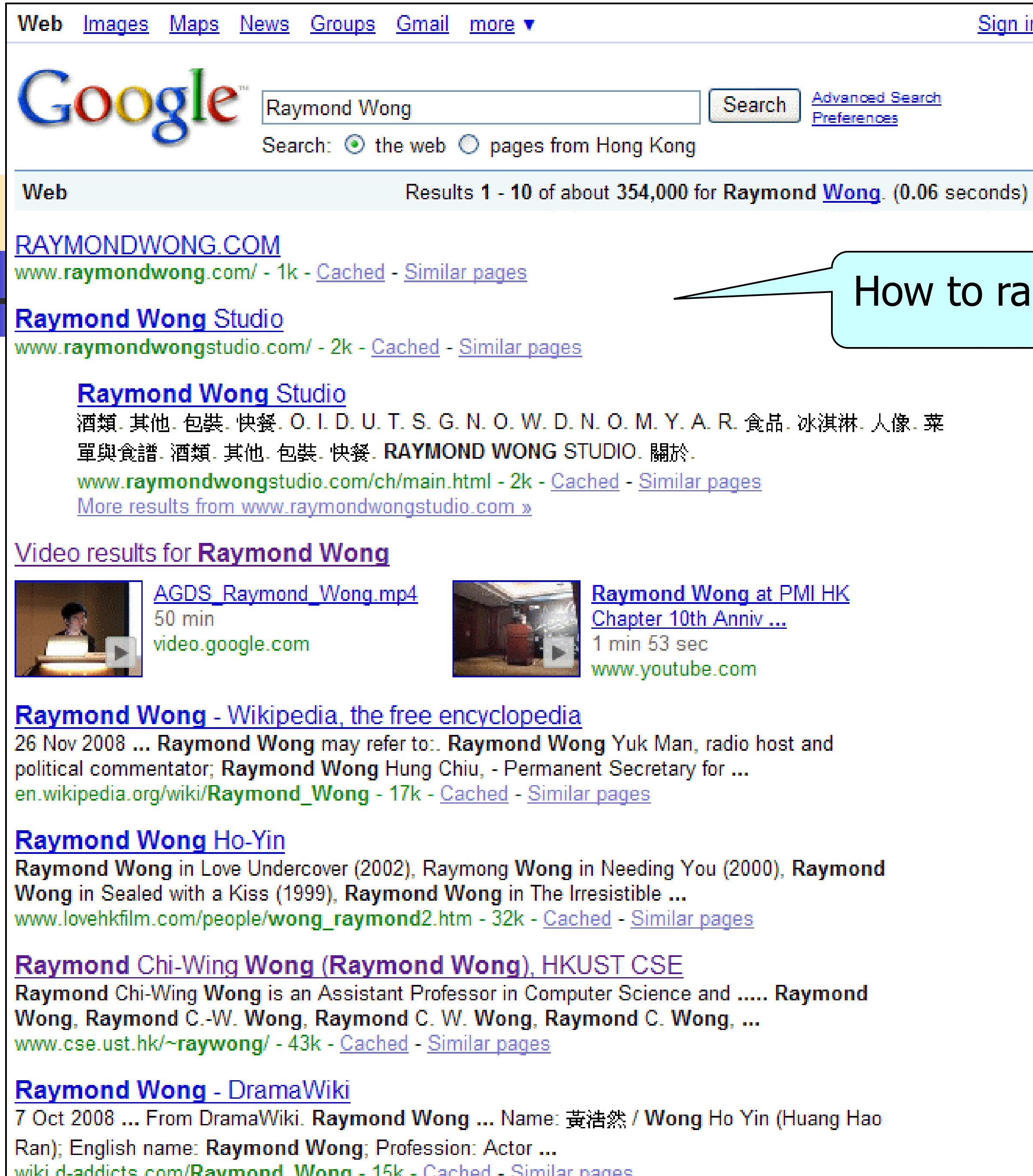
- Fast Query Response

Major Topics

1. Association
2. Clustering
3. Classification
4. Data Warehouse
5. Web Databases

5. Web Databases

Raymond Wong



How to rank the webpages?

- We have illustrated a list of major topics in data analysis
- Next, let me illustrate 2 recent research papers from me.

Which apartment should Raymond buy?

Suppose that user Raymond wants to buy an apartment

If the value is larger, then it is better to a user.
One example is the apartment size.

There are 2 popular queries for this problem.

- Top-k queries
- Skyline queries

In this talk, we will talk about a new type of queries.

- k-regret queries

D

Apartment	X_1	X_2
p_1	0	1
p_2	0.2	1
p_3	0.6	0.9
p_4	0.9	0.6
p_5	1	0.2
p_6	1	0
...	...	...

Suppose that user Raymond wants to buy an apartment

Top-k queries

D

Apartment	**X_1**	**X_2**
p_1	0	1
p_2	0.2	1
p_3	0.6	0.9
p_4	0.9	0.6
p_5	1	0.2
p_6	1	0
...	...	...

Suppose that user Raymond wants to buy an apartment

- Assume that Raymond has a "known" utility function.
- Utility function f

  $f(p) = 0.3\ X_1 + 0.7\ X_2$
- Utility vector u = (0.3, 0.7)
- Suppose that we want to find the top-1 apartment.
- **Output**

  Maximum utility point of D = p_3

D

Apartment	X_1	X_2	Utility
p_1	0	1	0.7
p_2	0.2	1	0.76
p_3	0.6	0.9	0.81
p_4	0.9	0.6	0.69
p_5	1	0.2	0.44
p_6	1	0	0.3
…	…	…	…

Advantage: The output size is "fixed"

Disadvantage: We need to know the "exact" utility function of Raymond

Suppose that user Raymond wants to buy an apartment

- My previous work
  - k-Hit Query: Top-k Query with Probabilistic Utility Function (SIGMOD 2015)

Which apartment should Raymond buy?

Suppose that user Raymond wants to buy an apartment

If the value is larger, then it is better to a user.
One example is the apartment size.

There are 2 popular queries for this problem.

- Top-k queries
- Skyline queries

In this talk, we will talk about a new type of queries.

- k-regret queries

D

Apartment	X_1	X_2
p_1	0	1
p_2	0.2	1
p_3	0.6	0.9
p_4	0.9	0.6
p_5	1	0.2
p_6	1	0
...	...	...

Suppose that user Raymond wants to buy an apartment

Skyline queries

D

Apartment	**X_1**	**X_2**
p_1	0	1
p_2	0.2	1
p_3	0.6	0.9
p_4	0.9	0.6
p_5	1	0.2
p_6	1	0
...	...	...

Skyline queries

Suppose that user Raymond wants to buy an apartment

- There is no assumption that we know the "exact" utility function of Raymond
- There is a concept called "dominance"

D

Apartment	X_1	X_2
p_1	[illegible]	[illegible]
p_2	0.2	1
p_3	0.6	0.9
p_4	0.9	0.6
p_5	1	0.2
p_6	1	0
...	...	...

p_2 dominates p_1 because
(1) the X_1 value of p_2 is better than that of p_1.
(2) the X_2 value of p_2 is equal to that of p_1.

Skyline queries

Suppose that user Raymond wants to buy an apartment

- There is no assumption that we know the "exact" utility function of Raymond
- There is a concept called "dominance"

D

Apartment	X_1	X_2
~~p_1~~	[illegible]	[illegible]
p_2	0.2	1
p_3	0.6	0.9
p_4	0.9	0.6
p_5	1	0.2
~~p_6~~	~~1~~	[illegible]
...	...	...

p_5 dominates p_6 because
(1) the X_1 value of p_5 is equal to that of p_6.
(2) the X_2 value of p_5 is better than that of p_6.

Skyline queries

Suppose that user Raymond wants to buy an apartment

- There is no assumption that we know the "exact" utility function of Raymond
- There is a concept called "dominance"
- Apartments are called **skyline apartments** if they are not dominated by any other apartments
- **Output**
  Skyline apartments = {p_2, p_3, p_4, p_5}

D

Apartment	X_1	X_2
p_1	[illegible]	[illegible]
p_2	0.2	1
p_3	0.6	0.9
p_4	0.9	0.6
p_5	1	0.2
p_6	1	[illegible]
...	...	...

Advantage: There is no need to specify the utility function of Raymond

Disadvantage: The output size is uncontrollable.

Skyline queries

Suppose that user Raymond wants to buy an apartment

- My previous work
  - Skyline Queries and Pareto Optimality (Encyclopedia of Database Systems, 2016)
  - Finding Competitive Price (SIGSPATIAL GIS 2013)
  - Finding Top-k Preferable Products (TKDE 2012)
  - Finding Top-k Profitable Products (ICDE 2011)
  - Creating Competitive Products (VLDB 2009)
  - Online Skyline Analysis with Dynamic Preferences on Nominal Attributes (TKDE 2009)
  - Finding the Influence Set through Skylines (EDBT 2009)
  - Efficient Skyline Querying with Variable User Preferences on Nominal Attributes (VLDB 2008)
  - Mining Favorable Facets (SIGKDD 2007)

Which apartment should Raymond buy?

Suppose that user Raymond wants to buy an apartment

If the value is larger, then it is better to a user.
One example is the apartment size.

There are 2 popular queries for this problem.

- Top-k queries
- Skyline queries

In this talk, we will talk about a new type of queries.

- k-regret queries

D

Apartment	X_1	X_2
p_1	0	1
p_2	0.2	1
p_3	0.6	0.9
p_4	0.9	0.6
p_5	1	0.2
p_6	1	0
...	...	...

Suppose that user Raymond wants to buy an apartment

D

Apartment	X_1	X_2
p_1	0	1
p_2	0.2	1
p_3	0.6	0.9
p_4	0.9	0.6
p_5	1	0.2
p_6	1	0
…	…	…

k-regret queries

k-regret queries

Suppose that user Raymond wants to buy an apartment

- It has **both** the advantage of the top-k queries and the advantage of the skyline queries.

The output size is specified by parameter k (e.g., 2)

Advantage: The output size is “fixed”

Advantage: There is no need to specify the utility function of Raymond

D

Apartment	X_1	X_2
p_1	0	1
p_2	0.2	1
p_3	0.6	0.9
p_4	0.9	0.6
p_5	1	0.2
p_6	1	0
…	…	…

Suppose that user Raymond wants to buy an apartment

- My previous work
  - Interactive Search for One of the Top-k (SIGMOD 2021)
  - Being Happy with the Least: Achieving α-happiness with Minimum Number of Tuples (ICDE 2020)
  - Strongly Truthful Interactive Regret Minimization (SIGMOD 2019)
  - FindYourFavorite: An Interactive System for Finding the User's Favorite Tuple in the Database (SIGMOD 2019 (demo paper))
  - Finding Average Regret Ratio Minimizing Set in Database (ICDE 2019)
  - Efficient k-Regret Query Algorithm with Restriction-free Bound for any Dimensionality (SIGMOD 2018)
  - k-Regret Minimizing Set: Efficient Algorithms and Hardness (ICDT 2017)
  - Minimizing Average Regret Ratio in Database (SIGMOD 2016 (Undergraduate Research Competition))
  - Geometry Approach for k-Regret Query (ICDE 2014)

k-regret queries

Suppose that user Raymond wants to buy an apartment

- Consider that Raymond has a utility function with the utility vector (0.3, 0.7).
  But, we do not know this utility function.
- Suppose that the whole dataset is seen by user Raymond.
  - We could find his favorite apartment p_3 (Raymond's maximum utility point) 0.81

D

Apartment	X_1	X_2	Utility
p_1	0	1	0.7
p_2	0.2	1	0.76
p_3	0.6	0.9	0.81
p_4	0.9	0.6	0.69
p_5	1	0.2	0.44
p_6	1	0	0.3
…	…	…	…

If Raymond's maximum utility in S is equal to Raymond's maximum utility in D, then Raymond's regret ratio is equal to 0

Thus, we would like to have a smaller value for Raymond's regret ratio.

k-regret queries

Suppose that user Raymond w

- Consider that Raymond has a utility function w utility vector (0.3, 0.7).
  But, we do not know this utility function.
- Suppose that the whole dataset is seen by user Raymond.
  - We could find his favorite apartment p_3 (Raymond's maximum utility point) 0.81
- Consider a set S = $\{p_2, p_4\}$ (which could be an output of this query).
- Suppose that set S is seen by user Raymond.
  - We could find his favorite apartment p_2 (Raymond's maximum utility point) 0.76

D

There is a difference between these 2 utility values

Raymond's **Regret Ratio**

Raymond's maximum utility in S

$$= 1 - \frac{0.76}{0.81}$$

Raymond's maximum utility in D

$= 0.06173$

k-regret queries

Suppose that user Raymond wants to buy an apartment

- Consider that Raymond has a utility function with the utility vector (0.3, 0.7).
  But, we do not know this utility function.

D

There is a difference between these 2 utility values

Raymond's **Regret Ratio**

Raymond's maximum utility in S

$$= 1 - \frac{0.76}{0.81}$$

Raymond's maximum utility in D

$$= 0.06173$$

Problem (k-regret): Given a set D, we want to find a set S of k points such that the mrr of S is minimized.

Advantage: The output size is "fixed"

Advantage: There is no need to specify the utility function of Raymond

k-regret queries

Suppose that user

- Consider that Raymond has a utility function with the utility vector (0.3, 0.7).
  But, we do not know this utility function.
  - Raymond's Regret Ratio = 0.06173
- There are many other users (e.g., Mary and Peter).
- Each of them has different utility functions.
- E.g., Mary's Regret Ratio = 0.05120
  E.g., Peter's Regret Ratio = 0
- **Maximum Regret Ratio (mrr)** = the maximum of all regret ratios (among all users) (e.g., 0.06173)

D

Ap

There is a difference between these 2 utility values

Raymond's **Regret Ratio**

Raymond's maximum utility in S

Each user's preference is represented in form of a utility vector (w_1, w_2)

Consider all possible vectors.

k-regret queries

Suppose that user Raymond wants to buy an apartment

- Next, let us give some details.

Problem (k-regret): Given a set D, we want to find a set S of k points such that the mrr of S is minimized.

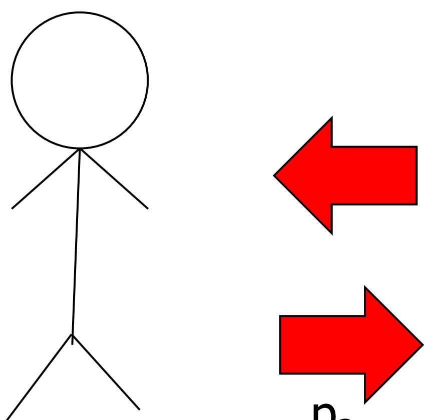
k-regret queries

Suppose that user Raymond wants to buy an apartment

- Consider that Raymond has a utility function with the utility vector (0.3, 0.7).
  But, we do not know this utility function.

D

Apartment	**X_1**	**X_2**
p_1	0	1
p_2	0.2	1
p_3	0.6	0.9
p_4	0.9	0.6
p_5	1	0.2
p_6	1	0
...	...	...



p_3

Raymond

Which one is better?

1. p_2
2. p_3

After this round,
we understand Raymond's preference better.

Problem (k-regret): Given a set D, we want to find a set S of k points such that the mrr of S is minimized.

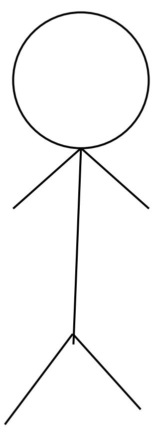
k-regret queries

Suppose that user Raymond wants to buy an apartment

- Consider that Raymond has a utility function with the utility vector (0.3, 0.7).
  But, we do not know this utility function.

D

Apartment	X_1	X_2
p_1	0	1
p_2	0.2	1
p_3	0.6	0.9
p_4	0.9	0.6



Raymond

Which one is better?

1. p_4
2. p_5

p_4

After this round,
we understand Raymond's preference much better.

With more rounds/questions, we could know Raymond's preference more.

Problem 1 (Interactive Regret): Given a set D, we want to ask a number of questions to Raymond and return an apartment such that Raymond's regret ratio is at most ε.

Problem 2 (Maximum Utility Point Determination): Given a set D, we want to ask a number of questions to Raymond and return an apartment such that this apartment is Raymond's maximum utility point in D.

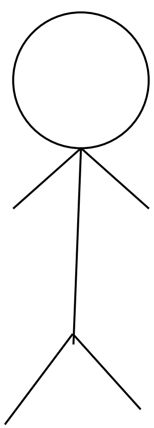
k-regret queries

Suppose that use

- Consider that Raymond
(0.3, 0.7).
But, we do not know th

D

Apartment	X_1	X_2
p_1	0	1
p_2	0.2	1
p_3	0.6	0.9
p_4	0.9	0.6



Raymond

Which one is better?

1. p_4
2. p_5

p_4

After this round,
we understand Raymond's preference much better.

With more rounds/questions, we could know Raymond's preference more.

k-regret queries

Suppose that user Raymond wants to buy an apartment

- My previous work
  - Interactive Search for One of the Top-k (SIGMOD 2021)
  - Being Happy with the Least: Achieving α-happiness with Minimum Number of Tuples (ICDE 2020)
  - Strongly Truthful Interactive Regret Minimization (SIGMOD 2019)
  - FindYourFavorite: An Interactive System for Finding the User's Favorite Tuple in the Database (SIGMOD 2019 (demo paper))
  - Finding Average Regret Ratio Minimizing Set in Database (ICDE 2019)
  - Efficient k-Regret Query Algorithm with Restriction-free Bound for any Dimensionality (SIGMOD 2018)
  - k-Regret Minimizing Set: Efficient Algorithms and Hardness (ICDT 2017)
  - Minimizing Average Regret Ratio in Database (SIGMOD 2016 (Undergraduate Research Competition))
  - Geometry Approach for k-Regret Query (ICDE 2014)

Demo System

- We developed a demo system on a car database with the following attributes
  - Price
  - Year
  - Power
  - Used km

Find Your Favorite!

This is a demostration system for finding your favorite car in a used car database.
Enter your acceptable range for each attribute (leave blank to use the default).
You will be presented two cars each time and you need to choose the one you favor more.
Click the "Start" button to find your favorite car in the database!

Attribute	Smallest Acceptable Range	Greatest Acceptable Range
Price (USD)	1000	50000
Year	2001	2017
Power (PS)	50	400
Used KM	10000	150000
Max No. of Cars	1000	**Mode** ◉ Simplex ○ Random

Start

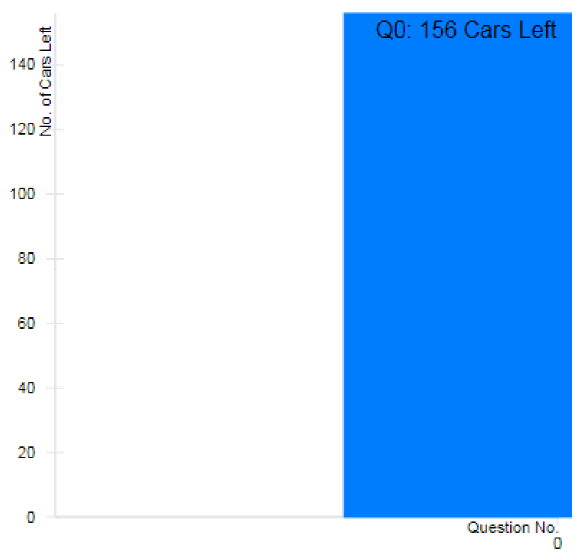
Your Choice

Q1: Choose the Car You Favor More among the Following Options

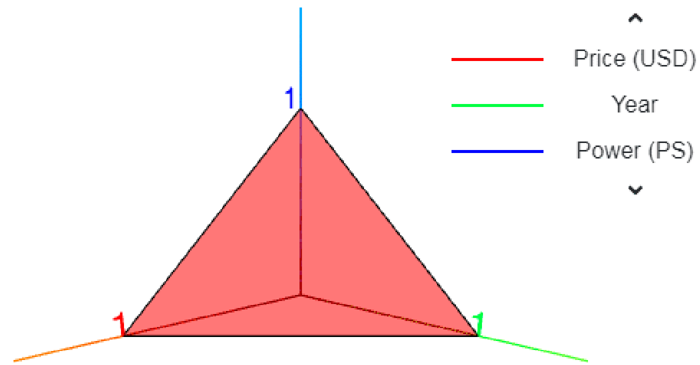
Option	Price (USD)	Year	Power (PS)	Used KM		
1	10500	2015	110	10000	Choose	Stop
2	8000	2011	156	30000	Choose	

Visualization

Cars Left vs. Questions Asked



Preference Space Visualization



Statistics

No. of Cars Pruned: 0

Step	Price (USD)	Year	Power (PS)	Used KM

No. of Cars Left: 156

Price (USD)	Year	Power (PS)	Used KM
14500	2014	125	30000
3699	2002	231	150000
2500	2002	193	150000
18000	2007	218	20000
39600	2014	306	30000
14999	2007	218	50000
7999	2013	80	40000

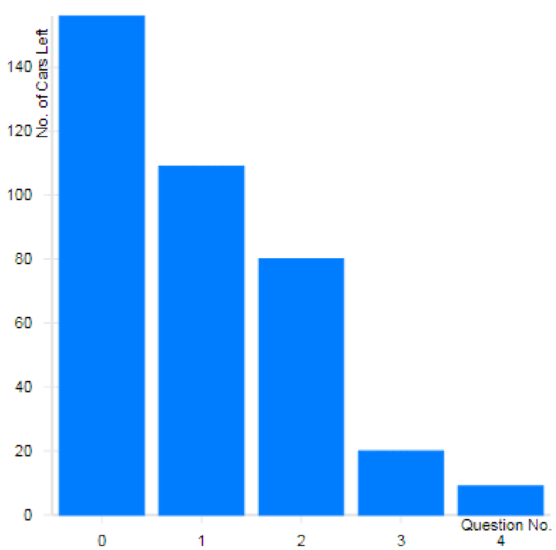
Your Choice

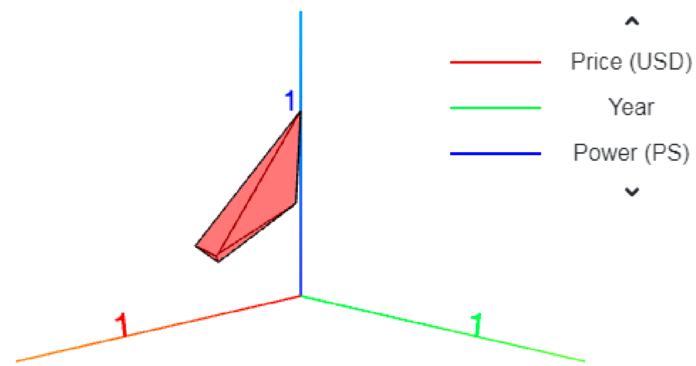
Q1: Choose the Car You Favor More among the Following Options

Option	Price (USD)	Year	Power (PS)	Used KM		
1	10500	2015	110	10000	Choose	Stop
2	8000	2011	156	30000	Choose	

Raymond chooses this option.
Then, he is asked for several questions and keep choosing options.

Visualization





Statistics

No. of Cars Pruned: 147

Step	Price (USD)	Year	Power (PS)	Used KM
4	3699	2002	231	150000
4	7300	2009	235	100000
4	6799	2002	286	150000
4	2450	2001	231	150000
4	8500	2003	300	150000
4	2990	2002	116	30000
4	1250	2001	129	150000

No. of Cars Left: 9

Price (USD)	Year	Power (PS)	Used KM
30900	2011	354	60000
12999	2002	306	90000
4200	2003	276	150000
17000	2008	344	125000
23000	2008	349	40000
8790	2010	299	100000
10900	2003	334	150000

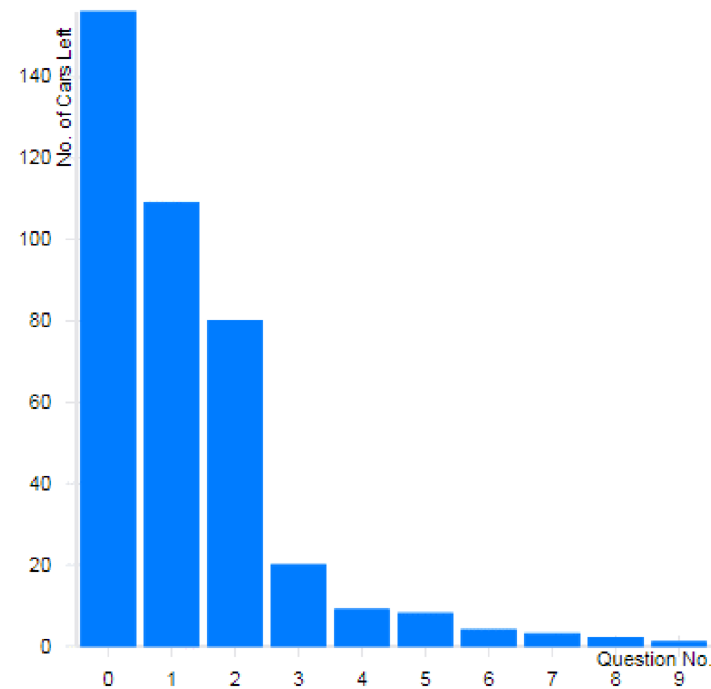
- Raymond keeps choosing one of the two choices.
- Finally, he obtains the following answer.

Total No. of Questions Asked is: 9.

Your Favourite Car is:

Price (USD)	Year	Power (PS)	Used KM
4200	2003	276	150000

Cars Left vs. Questions Asked



Return to Welcome

Conclusion

- We have illustrated a lot of applications how using data "smartly" could improve our life.
- We have illustrated some common topics in data analysis
- We have illustrated some of our recent papers.

Q&A