https://unsplash.com/photos/gm3bxHin8VA

Aalto University
School of Business

MySQL for Data Analytics

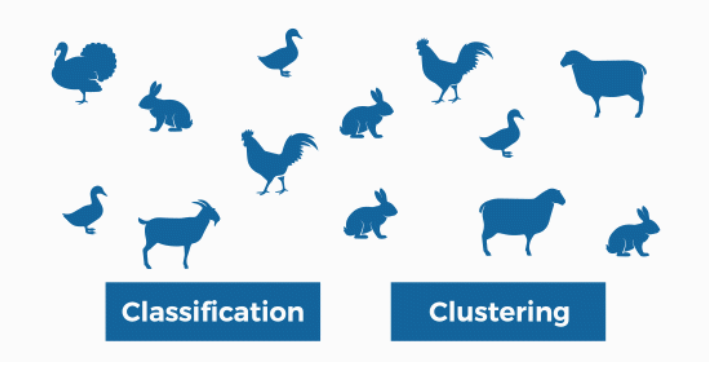
Lecturer: Yong Liu
Contact me at: Yong.liu@aalto.fi

Content

- **Association analysis**
- **Connecting database to R**
- **Commands for Root Users**
- **Text mining (Optional)**
- **Closure of the course**

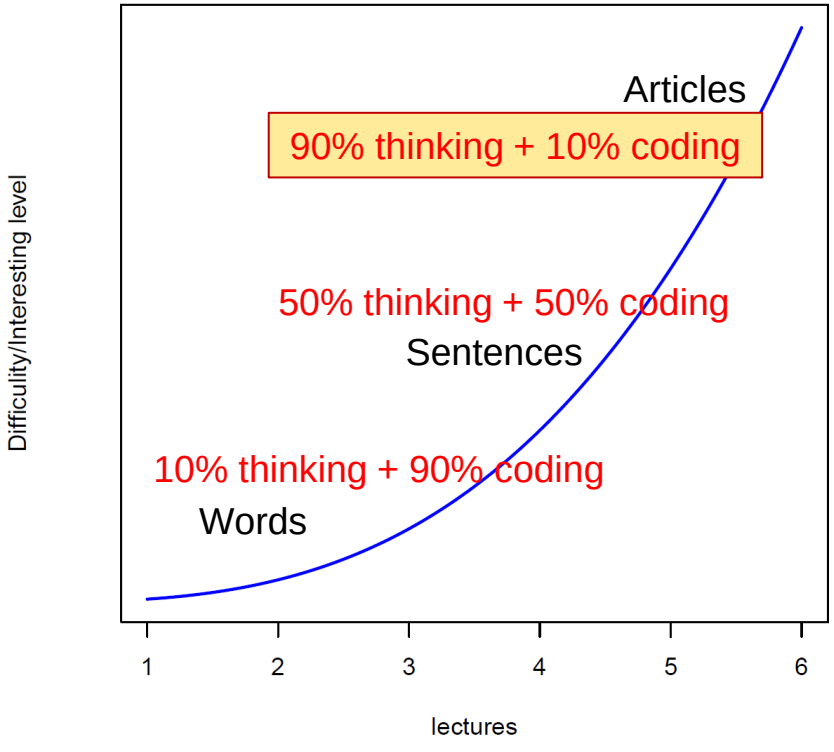
Principles Behind Common Analytic Tasks

- **Prediction** – e.g. daily revenue of next 7 days
- **Optimization** – e.g. logistic planning
- **Recommendation** – e.g. **association analysis**
- **Classification**
- **Clustering**



https://blog.bismart.com/en/classification-vs.-clustering-a-practical-explanation

Allocating your resources properly



- The first few lectures demand more memorization work, and very limited critical thinking.
- The last few lectures demand more critical thinking, in addition to some memorization work.

Advices:

- **Keep pace with the lectures**. Otherwise, you will find the course tremendously difficult in the last few lectures, if you do not get familiar with basic MySQL vocabulary.

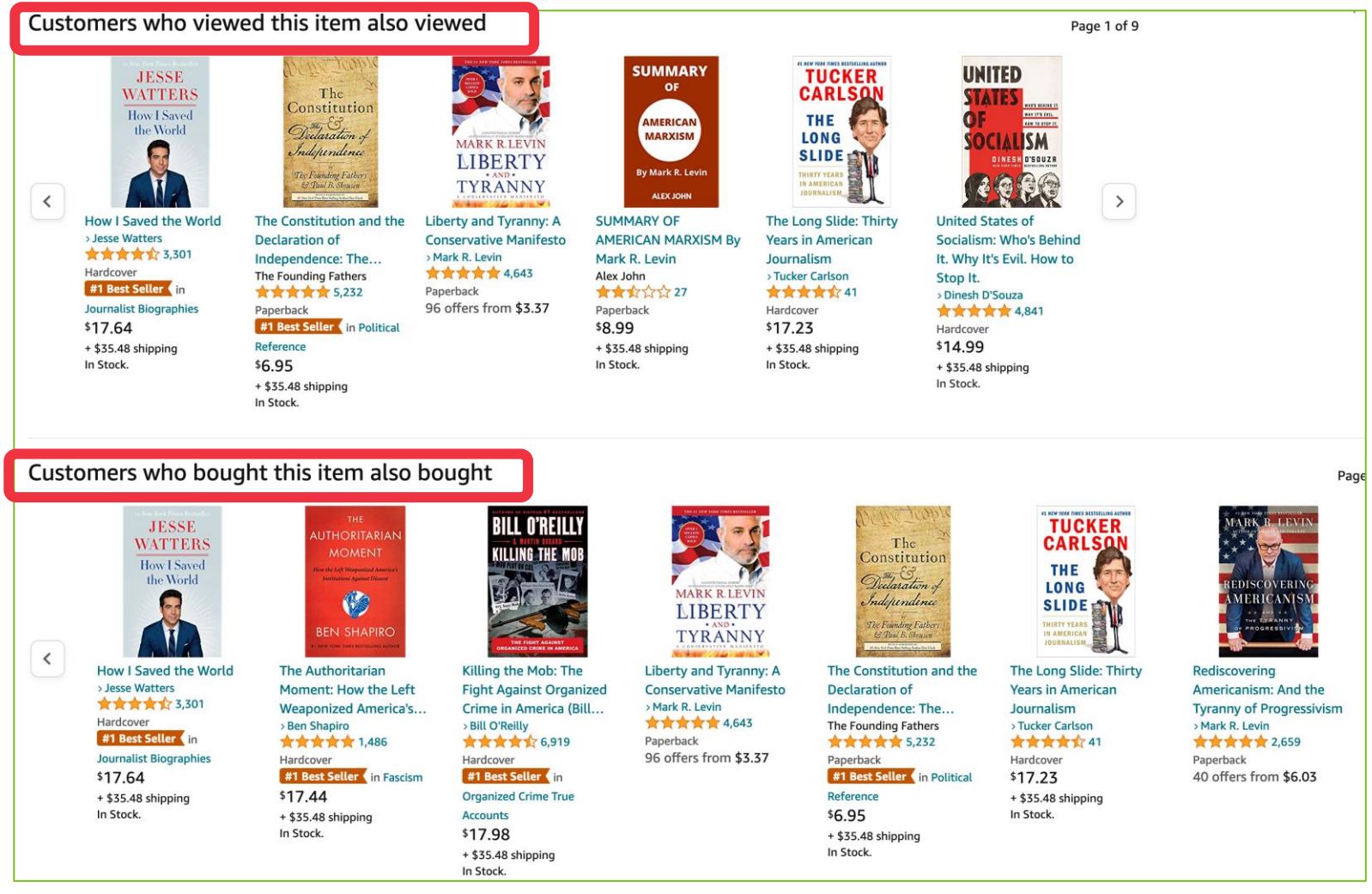
MySQL as a language: Words → Sentences → Articles

Association Analysis

- **Association analysis (AA)** discovers the probability of the co-occurrence of items in a collection.
- **Association rules**: the relationships between co-occurring items.
- Applications of AA: Market-basket analysis & network analysis

https://docs.oracle.com/cd/E18283_01/datamine.112/e16808/market_basket.htm

Association Analysis – Example I



Conversion Rate

"As all other aspects remain the same except for the **increase of the conversion rate by 1%**, this single digit increase has the overall effect of **increasing the marketing ROI from 25% to 88%**."

	BEFORE	AFTER
Visitors	100,000	100,000
Cost per visit	$0.40	$0.40
Cost of all visits visitors * cost per visit	$40,000	$40,000
Conversion rate	2%	3%
Total conversions conversion rate * visitors	2000	3000
Revenue per conversion revenue per conversion * conversions	$50	$50
Total Revenue	$100,000	$150,000
Non-marketing profit	50%	50%
Non-marketing costs (1 - non-marketing profit margin %) * total revenue	$50,000	$75,000
Marketing costs visitors * cost per visit	$40,000	$40,000
Total profit total revenue – (non-marketing costs + cost of all visits)	$10,000	$35,000
Total Marketing ROI total profit / marketing costs	25%	88%

https://www.trm-marketing.com/blog/post/what-is-a-conversion-rate-and-what-is-the-value-of-increasing-it

Market Basket Example

Example II

? Where should detergents be placed in the Store to maximize their sales?

? Are window cleaning products purchased when detergents and orange juice are bought together?

? Is soda typically purchased with bananas? Does the brand of soda make a difference?

? How are the demographics of the neighborhood affecting what customers are buying?

Image source: deepclimate.org

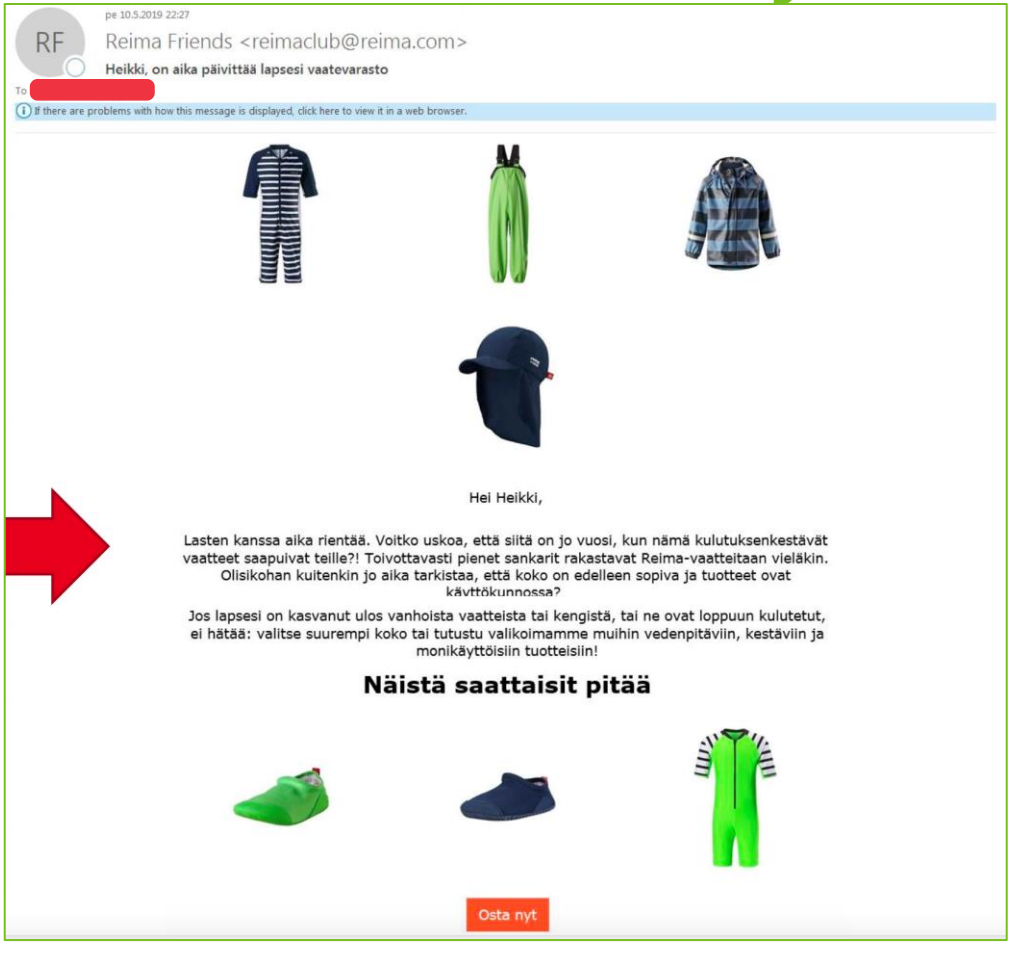
https://www.analyticsvidhya.com/blog/2014/08/effective-cross-selling-market-basket-analysis/

Market-basket analysis

- Valuable for direct marketing, sales promotions, and for discovering business trends. Market-basket analysis can also be used effectively for store layout, catalog design, and cross-sell.
- In e-commerce, association rules may be used for Web page personalization.

Example: An association model might find that a user who visits pages A and B is 70% likely to also visit page C in the same session. Based on this rule, a dynamic link could be created for users who are likely to be interested in page C.

Association Analysis – Example III



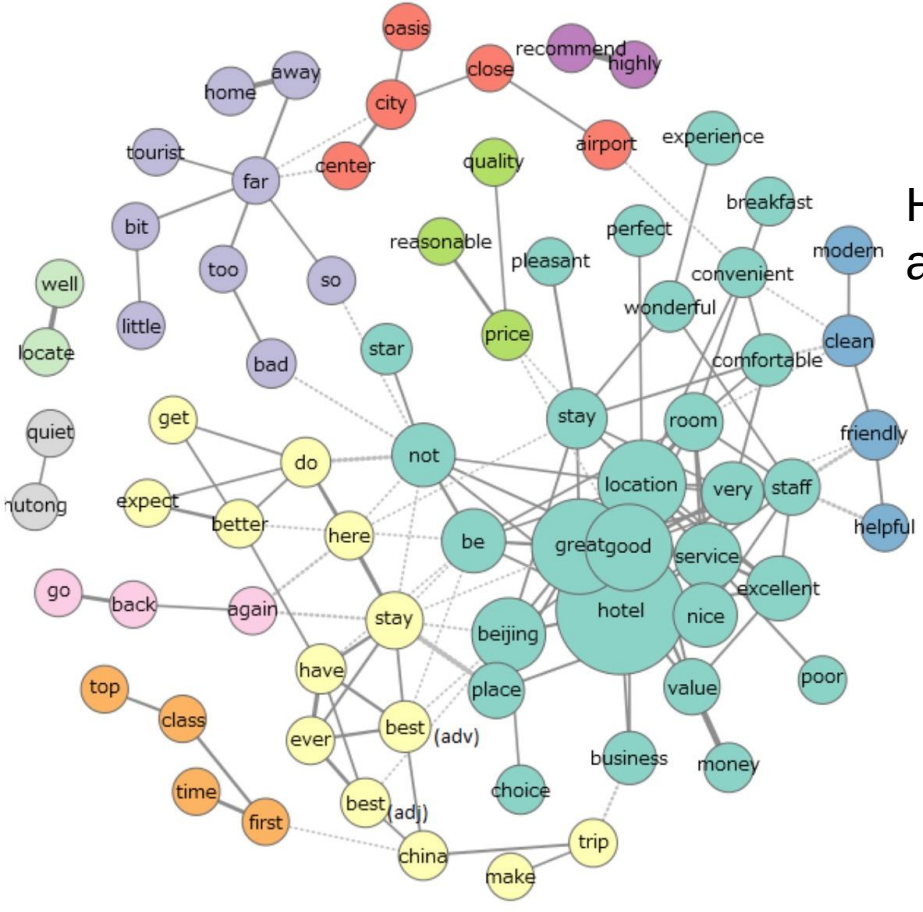
Translation of the Finnish words:

With children the time is rushing. Can you believe it's been a year since these wear-resistant clothes arrived for you ?! Hopefully the little heroes will settle their Reima clothes even more. However, it would be time to check that the size is still suitable and the products are in good condition.

If your child has grown out of old clothes or shoes, or they are worn out, no worries: choose a larger size or check out our range of other waterproof, durable and multi-purpose products!

These you might like.

Association Analysis – Example IV



Words association

How do consumers evaluate different attributes of your products or services?

Association rules

$$Support = \frac{frq(X,Y)}{N}$$

$$Rule: X \Rightarrow Y \quad Confidence = \frac{frq(X,Y)}{frq(X)}$$

$$Lift = \frac{Support}{Supp(X) \times Supp(Y)}$$

Rule	Support	Confidence	Lift
$A \Rightarrow D$	2/5	2/3	10/9
$C \Rightarrow A$	2/5	2/4	5/6
$A \Rightarrow C$	2/5	2/3	5/6
$B \& C \Rightarrow D$	1/5	1/3	5/9

An example of Association Rules

1. Assume there are 100 customers.
2. 10 of them bought milk, 8 bought butter and 6 bought both of them.
3. bought milk => bought butter.
4. support = P(Milk & Butter) = 6/100 = 0.06.
5. confidence = support/P(Butter) = 0.06/0.08 = 0.75.
6. lift = confidence/P(Milk) = 0.75/0.10 = 7.5.

Please note the rule $A \Rightarrow D$ differs from the rule $D \Rightarrow A$

https://ghumare64.medium.com/market-basket-analysis-using-association-rule-mining-2f87e4064ee7
https://www.saedsayad.com/association_rules.htm

Core Function of Association Analysis

id	uid
1	Product a
1	Product b
1	Product c
1	Product d
2	Product a
2	Product b
2	Product c
2	Product e
3	Product d
3	Product c
3	Product b
3	Product e

Raw data

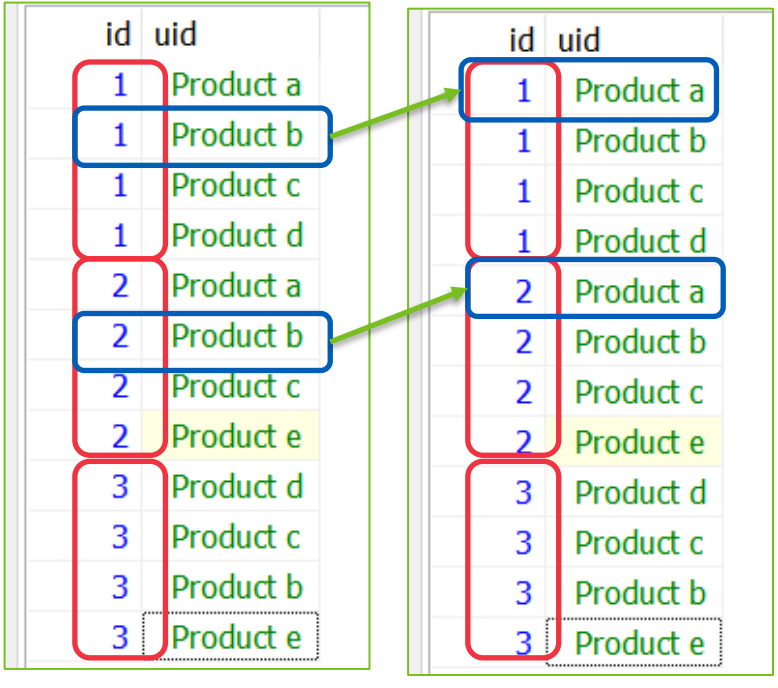
```
SELECT   a.uid as Product1, b.uid as Product2,
         COUNT(*) as Frequency
FROM EXAMPLE1 as a JOIN EXAMPLE1 as b
         ON a.id = b.id AND a.uid > b.uid
GROUP BY a.uid, b.uid
```

example1 (3×10)

Product1	Product2	Frequency
Product b	Product a	2
Product c	Product a	2
Product d	Product a	1
Product c	Product b	3
Product d	Product b	2
Product d	Product c	2
Product e	Product a	1
Product e	Product b	2
Product e	Product c	2
Product e	Product d	1

Co-occurrence frequency

```
SELECT   *
FROM EXAMPLE1 as a JOIN EXAMPLE1 as b
ON a.id = b.id AND a.uid > b.uid
```



id	uid	id	uid
1	Product b	1	Product a
1	Product c	1	Product a
1	Product d	1	Product a
1	Product c	1	Product b
1	Product d	1	Product b
1	Product d	1	Product c
2	Product b	2	Product a
2	Product c	2	Product a
2	Product e	2	Product a
2	Product c	2	Product b
2	Product e	2	Product b
2	Product e	2	Product c
3	Product c	3	Product b
3	Product d	3	Product b
3	Product e	3	Product b
3	Product d	3	Product c
3	Product e	3	Product c
3	Product e	3	Product d

id	uid	id	uid
1	Product b	1	Product a
1	Product c	1	Product a
1	Product d	1	Product a
1	Product c	1	Product b
1	Product d	1	Product b
1	Product d	1	Product c
2	Product b	2	Product a
2	Product c	2	Product a
2	Product e	2	Product a
2	Product c	2	Product b
2	Product e	2	Product b
2	Product e	2	Product c
3	Product c	3	Product b
3	Product d	3	Product b
3	Product e	3	Product b
3	Product d	3	Product c
3	Product e	3	Product c
3	Product e	3	Product d

```
SELECT   a.uid as Person1, b.uid as Person2,
                COUNT(*) as Frequency
FROM EXAMPLE1 as a JOIN EXAMPLE1 as b
                ON a.id = b.id AND a.uid > b.uid
GROUP BY a.uid, b.uid
```

example1 (3×10)

Product1	Product2	Frequency
Product b	Product a	2
Product c	Product a	2
Product d	Product a	1
Product c	Product b	3
Product d	Product b	2
Product d	Product c	2
Product e	Product a	1
Product e	Product b	2
Product e	Product c	2
Product e	Product d	1

Compute the confidence level

```
SELECT tb1.Product1, tb1.Product2, tb1.Frequency,
tb1.Frequency/tb2.overall_frequency from

(SELECT   a.uid as Product1, b.uid as Product2,
COUNT(*) as Frequency
FROM EXAMPLE1 as a JOIN EXAMPLE1 as b
ON a.id = b.id AND a.uid > b.uid
GROUP BY a.uid, b.uid)  AS tb1

Join

(SELECT uid AS focused_product, COUNT(*) AS
overall_frequency
FROM EXAMPLE1 GROUP BY uid) AS tb2

ON tb1.Product2 = tb2.focused_product;
```

Product1	Product2	Frequency	tb1.Frequency/tb2.overall_frequency
Product b	Product a	2	1.0000
Product c	Product a	2	1.0000
Product d	Product a	1	0.5000
Product c	Product b	3	1.0000
Product d	Product b	2	0.6667
Product d	Product c	2	0.6667
Product e	Product a	1	0.5000
Product e	Product b	2	0.6667
Product e	Product c	2	0.6667
Product e	Product d	1	0.5000

$$Rule: X \Rightarrow Y$$

$$Support = \frac{frq(X,Y)}{N}$$

$$Confidence = \frac{frq(X,Y)}{frq(X)}$$

$$Lift = \frac{Support}{Supp(X) \times Supp(Y)}$$

Apply the code to real-life data

Core Function of Association Analysis

id	uid
1	Product a
1	Product b
1	Product c
1	Product d
2	Product a
2	Product b
2	Product c
2	Product e
3	Product d
3	Product c
3	Product b
3	Product e

```
SELECT  a.uid as Person1, b.uid as Person2,
        COUNT(*) as Frequency
FROM EXAMPLE1 as a JOIN EXAMPLE1 as b
        ON a.id = b.id AND a.uid > b.uid
GROUP BY a.uid, b.uid
```

example1 (3×10)

Product1	Product2	Frequency
Product b	Product a	2
Product c	Product a	2
Product d	Product a	1
Product c	Product b	3
Product d	Product b	2
Product d	Product c	2
Product e	Product a	1
Product e	Product b	2
Product e	Product c	2
Product e	Product d	1

Raw data → Co-occurrence frequency

Can you revise the code to generate results for a hotel recommendation system?

"Sample solution"

id	hotel_id	user_id	username	overall_rating	review_date	checkin_year
646,693	199,923	(NULL)	(NULL)	5	2002-08-11	2,002
646,694	228,670	(NULL)	(NULL)	4	2002-08-11	2,002
809,978	228,670	(NULL)	(NULL)	3	2002-11-23	2,002
833,841	249,056	(NULL)	(NULL)	5	2003-01-05	2,003
841,271	206,760	(NULL)	(NULL)	5	2003-01-20	2,003
853,633	206,761	E4ED74B155D08686D9C032A5286D...	TC1968	4	2003-02-05	2,003
860,530	267,464	(NULL)	(NULL)	5	2003-02-23	2,003
908,091	267,464	(NULL)	(NULL)	5	2003-03-04	2,003
987,116	263,808	(NULL)	(NULL)	4	2003-05-02	2,003
1,070,394	228,682	CC9AB0C275A231756D0D1C0E443...	Helga88	4	2003-06-04	2,003
1,108,965	232,283	C164F53CD27D809BC7210E34703B...	Maura C	5	2003-06-25	2,003
1,137,703	228,670	(NULL)	(NULL)	4	2003-07-14	2,003
1,138,665	596,136	8571D7E8520AA15C3EF73567142A...	Nils S H	4	2003-07-15	2,003
1,153,472	263,808	(NULL)	(NULL)	4	2003-07-29	2,003
1,216,222	206,760	307FFC64E94BB40C7FB0E6674526F...	european1	4	2003-08-05	2,003
1,230,492	228,673	(NULL)	(NULL)	3	2003-08-16	2,003

Table: tripadvisor_data_for_handson_assignment_ONLY

Submit your answer to Presemo:
https://presemo.aalto.fi/drm/

```
SELECT  a.hotel_id as hotel1, b.hotel_id as hotel2,
COUNT(*) as Frequency
FROM tripadvisor_review_sample_without_reviewtext as a
JOIN tripadvisor_review_sample_without_reviewtext as b
ON a.username = b.username AND a.hotel_id > b.hotel_id
WHERE a.hotel_id IS NOT null
GROUP BY a.hotel_id, b.hotel_id
ORDER BY Frequency desc
```

Whether the above code will yield correct results and why?

hotel1	hotel2	Frequency
1,389,502	99,468	15
108,141	99,468	15
99,468	83,660	15
645,215	99,468	10
223,712	99,468	10
1,175,533	99,468	10
99,468	88,440	10
1,218,737	99,468	10
99,476	99,468	10
285,427	99,468	10
218,777	99,468	10
224,843	99,468	10
99,468	88,442	10
99,468	80,593	10
99,468	87,571	10
1,389,502	83,660	9
108,141	83,660	9
1,389,502	108,141	9

id	hotel_id	user_id	username	overall_rating	review_date	checkin_year
646,693	199,923	(NULL)	(NULL)	5	2002-08-11	2,002
646,694	228,670	(NULL)	(NULL)	4	2002-08-11	2,002
809,978	228,670	(NULL)	(NULL)	3	2002-11-23	2,002
833,841	249,056	(NULL)	(NULL)	5	2003-01-05	2,003
841,271	206,760	(NULL)	(NULL)	5	2003-01-20	2,003
853,633	206,761	E4ED74B155D08686D9C032A5286D...	TC1968	4	2003-02-05	2,003
860,530	267,464	(NULL)	(NULL)	5	2003-02-23	2,003
908,091	267,464	(NULL)	(NULL)	5	2003-03-04	2,003
987,116	263,808	(NULL)	(NULL)	4	2003-05-02	2,003
1,070,394	228,682	CC9AB0C275A231756D0D1C0E443...	Helga88	4	2003-06-04	2,003
1,108,965	232,283	C164F53CD27D809BC7210E34703B...	Maura C	5	2003-06-25	2,003
1,137,703	228,670	(NULL)	(NULL)	4	2003-07-14	2,003
1,138,665	596,136	8571D7E8520AA15C3EF73567142A...	Nils S H	4	2003-07-15	2,003
1,153,472	263,808	(NULL)	(NULL)	4	2003-07-29	2,003
1,216,222	206,760	307FFC64E94BB40C7FB0E6674526F...	european1	4	2003-08-05	2,003
1,230,492	228,673	(NULL)	(NULL)	3	2003-08-16	2,003

Uniqueness of the data [?]:
A user may visit the same hotel multiple times

id	uid
1	Product a
1	Product b
1	Product c
1	Product d
2	Product a
2	Product b
2	Product c
2	Product e
3	Product b
3	Product c
3	Product d
3	Product e

Table: EXAMPLE1

```
SELECT   a.uid as Person1, b.uid as Person2,  COUNT(*) as Frequency
FROM EXAMPLE1 as a JOIN EXAMPLE1 as b
         ON a.id = b.id AND a.uid > b.uid
GROUP BY a.uid, b.uid;
```

Code

Person1	Person2	Frequency
Product b	Product a	2
Product c	Product a	2
Product d	Product a	1
Product e	Product a	1
Product c	Product b	3
Product d	Product b	2
Product e	Product b	2
Product d	Product c	2
Product e	Product c	2
Product e	Product d	1

Output

id	uid
1	Product a
1	Product a
1	Product b
1	Product c
1	Product d
2	Product a
2	Product b
2	Product c
2	Product e
3	Product b
3	Product c
3	Product d
3	Product e

Table: EXAMPLE2

```
SELECT   a.uid as Person1, b.uid as Person2,  COUNT(*) as Frequency
FROM EXAMPLE2 as a JOIN EXAMPLE2 as b
         ON a.id = b.id AND a.uid > b.uid
GROUP BY a.uid, b.uid
```

Code

Person1	Person2	Frequency
Product b	Product a	3
Product c	Product a	3
Product d	Product a	2
Product e	Product a	1
Product c	Product b	3
Product d	Product b	2
Product e	Product b	2
Product d	Product c	2
Product e	Product c	2
Product e	Product d	1

Output

id	hotel_id	user_id	username	overall_rating	review_date	checkin_year
646,693	199,923	(NULL)	(NULL)	5	2002-08-11	2,002
646,694	228,670	(NULL)	(NULL)	4	2002-08-11	2,002
809,978	228,670	(NULL)	(NULL)	3	2002-11-23	2,002
833,841	249,056	(NULL)	(NULL)	5	2003-01-05	2,003
841,271	206,760	(NULL)	(NULL)	5	2003-01-20	2,003
853,633	206,761	E4ED74B155D08686D9C032A5286D...	TC1968	4	2003-02-05	2,003
860,530	267,464	(NULL)	(NULL)	5	2003-02-23	2,003
908,091	267,464	(NULL)	(NULL)	5	2003-03-04	2,003
987,116	263,808	(NULL)	(NULL)	4	2003-05-02	2,003
1,070,394	228,682	CC9AB0C275A231756D0D1C0E443...	Helga88	4	2003-06-04	2,003
1,108,965	232,283	C164F53CD27D809BC7210E34703B...	Maura C	5	2003-06-25	2,003
1,137,703	228,670	(NULL)	(NULL)	4	2003-07-14	2,003
1,138,665	596,136	8571D7E8520AA15C3EF73567142A...	Nils S H	4	2003-07-15	2,003
1,153,472	263,808	(NULL)	(NULL)	4	2003-07-29	2,003
1,216,222	206,760	307FFC64E94BB40C7FB0E6674526F...	european1	4	2003-08-05	2,003
1,230,492	228,673	(NULL)	(NULL)	3	2003-08-16	2,003

Uniqueness of the data [?]:
A user may visit the same hotel multiple times

What would be the solution to make the data ready for the association analysis?

Reflections

1. Understanding the nature of your data is very important before any analysis.
2. Code that generates no error message does not necessarily generate the right results.
3. A good understanding of your data – albeit take time and yields no direct output - is very important!

Association of more than two products

id	uid
1	Product a
1	Product b
1	Product c
1	Product d
2	Product a
2	Product b
2	Product c
2	Product e
3	Product d
3	Product c
3	Product b
3	Product e

Raw data

```
SELECT   a.uid as Person1, b.uid as Person2,
                COUNT(*) as Frequency
FROM EXAMPLE1 as a
JOIN EXAMPLE1 as b
                ON a.id = b.id AND a.uid > b.uid
GROUP BY a.uid, b.uid
```

example1 (3×10)

Product1	Product2	Frequency
Product b	Product a	2
Product c	Product a	2
Product d	Product a	1
Product c	Product b	3
Product d	Product b	2
Product d	Product c	2
Product e	Product a	1
Product e	Product b	2
Product e	Product c	2
Product e	Product d	1

Co-occurrence frequency of **two** products

Solution for three-products co-occurrence frequency?
Submit your answer to Presemo

Product1	Product2	Product3	Frequency
Product c	Product b	Product a	2
Product d	Product b	Product a	1
Product d	Product c	Product a	1
Product d	Product c	Product b	2
Product e	Product b	Product a	1
Product e	Product c	Product a	1
Product e	Product c	Product b	2
Product e	Product d	Product c	1
Product e	Product d	Product b	1

Co-occurrence frequency of **three** products

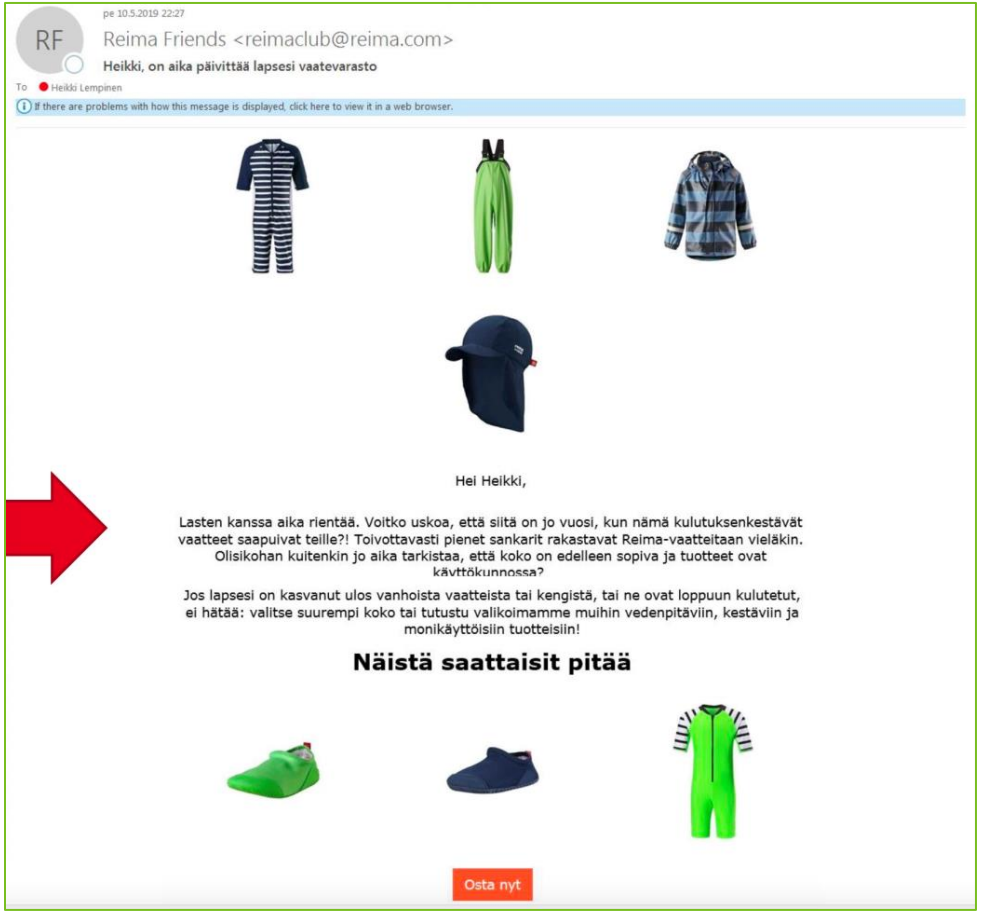
Solution for three-products co-occurrence frequency

```
SELECT   a.uid as Product1, b.uid as Product2, c.uid as Product3, COUNT(*) as Frequency
FROM EXAMPLE1 a
JOIN EXAMPLE1 b ON a.id = b.id AND a.uid > b.uid
JOIN EXAMPLE1 c ON a.id = c.id AND b.uid > c.uid
GROUP BY a.uid, b.uid, c.uid;
```

Product1	Product2	Product3	Frequency
Product c	Product b	Product a	2
Product d	Product b	Product a	1
Product d	Product c	Product a	1
Product d	Product c	Product b	2
Product e	Product b	Product a	1
Product e	Product c	Product a	1
Product e	Product c	Product b	2
Product e	Product d	Product c	1
Product e	Product d	Product b	1

Co-occurrence frequency
of **three** products

Can you generate the solution?



Data preparation:

1. What kinds of data you would need?
2. Any necessary manipulation to the data before analysis?

Analysis → Prediction:

1. What customers bought before predict what they will buy now?!
2. Does the sequence of purchase matter and how?
3. How to code?

review (5×56,709)

id	user_id	hotel_id	review_date	checkin_year
853,633	E4ED74B155D08686D9C032A5286D...	206,761	2003-02-05	2,003
1,070,394	CC9AB0C275A231756D0D1C0E443...	228,682	2003-06-04	2,003
1,108,965	C164F53CD27D809BC7210E34703B...	232,283	2003-06-25	2,003
1,138,665	8571D7E8520AA15C3EF73567142A...	596,136	2003-07-15	2,003
1,216,222	307FFC64E94BB40C7FB0E6674526F...	206,760	2003-08-05	2,003
1,383,626	DD49C7A0F9B575384874F75C9B21...	262,286	2003-10-09	2,003
1,496,549	BAA10872C4E4380C336DF0F9EF52...	199,923	2003-11-22	2,003
1,541,745	8854F4A62CB84DC2DB48465E2907...	202,626	2004-01-06	2,004
1,644,481	2B40B65CFF95534211587D0C3DC0...	206,763	2004-02-09	2,004
1,715,016	023696A80B035229F9873306B3D8...	293,333	2004-03-08	2,004
1,754,696	07A70590205F70B4329E58749BFAB...	199,923	2004-03-29	2,004
1,777,602	E2C4BE6FA54491536BB6DAB1AB10...	228,677	2004-04-12	2,004
1,777,683	E2C4BE6FA54491536BB6DAB1AB10...	228,682	2004-04-12	2,004
1,785,803	DCB34C13526F81DF5975E7E5CB16...	228,682	2004-04-13	2,004
1,869,142	2A8F276D2471783E1053D68A370C...	238,453	2004-05-03	2,004
2,078,376	454C2DAAA0BD0F6526D90EF10986...	206,765	2004-05-18	2,004
2,078,379	454C2DAAA0BD0F6526D90EF10986...	237,708	2004-05-18	2,004
2,128,683	3C9DF04638A7A8228333A3355273...	267,464	2004-05-28	2,004
2,268,991	6325AD565B1B895111151F30A963...	199,923	2004-07-05	2,004
2,498,342	EC710E4F8F1F7B9365224C9C66CC7...	228,674	2004-08-26	2,004

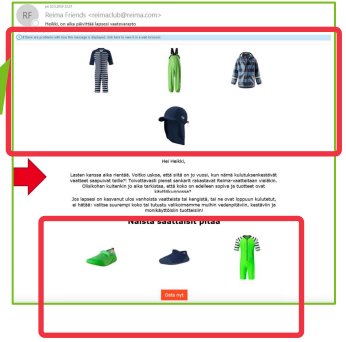
Raw data

review (2×15)

checkin_year	COUNT(*)
2,002	3
2,003	19
2,004	75
2,005	117
2,006	313
2,007	534
2,008	710
2,009	1,314
2,010	1,997
2,011	3,696
2,012	6,250
2,013	8,982
2,014	12,813
2,015	18,091
2,016	2,203

Manipulation I

Predictors



Outcomes

Data Preparation/Manipulation:

If time variable matters now, how to add that factor/variable into the association analysis?

Other considerations:

1. Association analysis AND/OR Prediction?
2. Association based on this year's sale data?
3. Association based on last year's sale data?
4. Association based on an integration of this year and last year's sale data?
5. An integration of association and prediction analysis?

review (5×56,709)

id	user_id	hotel_id	review_date	new_product_id
853,633	E4ED74B155D08686D9C032A5286D...	206,761	2003-02-05	2,003
1,070,394	CC9AB0C275A231756D0D1C0E443...	228,682	2003-06-04	2,003
1,108,965	C164F53CD27D809BC7210E34703B...	232,283	2003-06-25	2,003
1,138,665	8571D7E8520AA15C3EF73567142A...	596,136	2003-07-15	2,003
1,216,222	307FFC64E94BB40C7FB0E6674526F...	206,760	2003-08-05	2,003
1,383,626	DD49C7A0F9B575384874F75C9B21...	262,286	2003-10-09	2,003
1,496,549	BAA10872C4E4380C336DF0F9EF52...	199,923	2003-11-22	2,003
1,541,745	8854F4A62CB84DC2DB48465E2907...	202,626	2004-01-06	2,004
1,644,481	2B40B65CFF95534211587D0C3DC0...	206,763	2004-02-09	2,004
1,715,016	023696A80B035229F9873306B3D8...	293,333	2004-03-08	2,004
1,754,696	07A70590205F70B4329E58749BFAB...	199,923	2004-03-29	2,004
1,777,602	E2C4BE6FA54491536BB6DAB1AB10...	228,677	2004-04-12	2,004
1,777,683	E2C4BE6FA54491536BB6DAB1AB10...	228,682	2004-04-12	2,004
1,785,803	DCB34C13526F81DF5975E7E5CB16...	228,682	2004-04-13	2,004
1,869,142	2A8F276D2471783E1053D68A370C...	238,453	2004-05-03	2,004
2,078,376	454C2DAAA0BD0F6526D90EF10986...	206,765	2004-05-18	2,004
2,078,379	454C2DAAA0BD0F6526D90EF10986...	237,708	2004-05-18	2,004
2,128,683	3C9DF04638A7A8228333A3355273...	267,464	2004-05-28	2,004
2,268,991	6325AD565B1B895111151F30A963...	199,923	2004-07-05	2,004
2,498,342	EC710E4F8F1F7B9365224C9C66CC7...	228,674	2004-08-26	2,004

Raw data

review (2×15)

checkin_year	COUNT(*)
2,002	3
2,003	19
2,004	75
2,005	117
2,006	313
2,007	534
2,008	710
2,009	1,314
2,010	1,997
2,011	3,696
2,012	6,250
2,013	8,982
2,014	12,813
2,015	18,091
2,016	2,203

Manipulation I

Question: Are we ready to run the code for association analysis?

Data Preparation/Manipulation:
If time variable matters now, how to add that fact/variable into the association analysis?

```
(SELECT id, user_id, hotel_id, review_date, checkin_year,
CONCAT(hotel_id, '_','Predictor') AS new_product_id
FROM review
WHERE user_id IS NOT NULL AND checkin_year < 2015
LIMIT 15)
union
(SELECT id, user_id, hotel_id, review_date, checkin_year,
CONCAT(hotel_id, '_','Outcome') AS new_product_id
FROM review
WHERE user_id IS NOT NULL AND checkin_year >=
2015 LIMIT 10)
```

Result #1 (6×15)

id	user_id	hotel_id	review_date	checkin_year	new_product_id
314,109,325	000A1DFEFDAE9CAB77EB...	228,682	2015-09-27	2,015	228682_Outcome
275,334,075	000ECDFA5D2DDBDF7479...	228,686	2015-05-28	2,015	228686_Outcome
275,341,210	000ECDFA5D2DDBDF7479...	228,673	2015-05-28	2,015	228673_Outcome
147,423,657	0004FF72BDDF25A752B70...	267,464	2012-12-13	2,012	267464_Predictor
312,100,406	0008D1BACA977284D9DD...	232,150	2015-09-20	2,015	232150_Outcome
225,432,633	00017FF6229848D7515CB...	232,143	2014-08-30	2,014	232143_Predictor
227,105,065	00017FF6229848D7515CB...	1,840,189	2014-09-05	2,014	1840189_Predictor
294,869,698	00044C8B32E2D18B687E6...	2,151,632	2015-08-01	2,015	2151632_Outcome
106,676,859	000491C9528EBE3302CA4...	232,307	2011-05-03	2,011	232307_Predictor
256,942,638	000859B4AFE2A75400EDD...	281,329	2015-02-28	2,015	281329_Outcome
293,192,416	0002576AD91EE9065846B...	578,920	2015-07-27	2,015	578920_Outcome
335,151,566	0003911E45A14147EF812...	578,920	2015-12-26	2,015	578920_Outcome
336,762,288	0003976319D8996F54F4A...	293,333	2016-01-02	2,016	293333_Outcome

review (5×56,709)

id	user_id	hotel_id	review_date	new_product_id
853,633	E4ED74B155D08686D9C032A5286D...	206,761	2003-02-05	2,003
1,070,394	CC9AB0C275A231756D0D1C0E443...	228,682	2003-06-04	2,003
1,108,965	C164F53CD27D809BC7210E34703B...	232,283	2003-06-25	2,003
1,138,665	8571D7E8520AA15C3EF73567142A...	596,136	2003-07-15	2,003
1,216,222	307FFC64E94BB40C7FB0E6674526F...	206,760	2003-08-05	2,003
1,383,626	DD49C7A0F9B575384874F75C9B21...	262,286	2003-10-09	2,003
1,496,549	BAA10872C4E4380C336DF0F9EF52...	199,923	2003-11-22	2,003
1,541,745	8854F4A62CB84DC2DB48465E2907...	202,626	2004-01-06	2,004
1,644,481	2B40B65CFF95534211587D0C3DC0...	206,763	2004-02-09	2,004
1,715,016	023696A80B035229F9873306B3D8...	293,333	2004-03-08	2,004
1,754,696	07A70590205F70B4329E58749BFAB...	199,923	2004-03-29	2,004
1,777,602	E2C4BE6FA54491536BB6DAB1AB10...	228,677	2004-04-12	2,004
1,777,683	E2C4BE6FA54491536BB6DAB1AB10...	228,682	2004-04-12	2,004
1,785,803	DCB34C13526F81DF5975E7E5CB16...	228,682	2004-04-13	2,004
1,869,142	2A8F276D2471783E1053D68A370C...	238,453	2004-05-03	2,004
2,078,376	454C2DAAA0BD0F6526D90EF10986...	206,765	2004-05-18	2,004
2,078,379	454C2DAAA0BD0F6526D90EF10986...	237,708	2004-05-18	2,004
2,128,683	3C9DF04638A7A8228333A3355273...	267,464	2004-05-28	2,004
2,268,991	6325AD565B1B895111151F30A963...	199,923	2004-07-05	2,004
2,498,342	EC710E4F8F1F7B9365224C9C66CC7...	228,674	2004-08-26	2,004

Raw data

review (2×15)

checkin_year	COUNT(*)
2,002	3
2,003	19
2,004	75
2,005	117
2,006	313
2,007	534
2,008	710
2,009	1,314
2,010	1,997
2,011	3,696
2,012	6,250
2,013	8,982
2,014	12,813
2,015	18,091
2,016	2,203

Manipulation I

Question: Are we ready to run the code for association analysis?

Answer: NO

- There can be multiple values for the same products for the same users in the category of predictor or outcome variable. For instance, a user repeatedly visited the same hotels in the past.

Data Preparation/Manipulation:
If time variable matters now, how to add that fact/variable into the association analysis?

```
(SELECT id, user_id, hotel_id, review_date, checkin_year,
CONCAT(hotel_id, '_','Predictor') AS new_product_id
FROM review
WHERE user_id IS NOT NULL AND checkin_year < 2015
LIMIT 15)
union
(SELECT id, user_id, hotel_id, review_date, checkin_year,
CONCAT(hotel_id, '_','Outcome') AS new_product_id
FROM review
WHERE user_id IS NOT NULL AND checkin_year >=
2015 LIMIT 10)
```

Result #1 (6×15)

id	user_id	hotel_id	review_date	checkin_year	new_product_id
314,109,325	000A1DFEFDAE9CAB77EB...	228,682	2015-09-27	2,015	228682_Outcome
275,334,075	000ECDFA5D2DDBDF7479...	228,686	2015-05-28	2,015	228686_Outcome
275,341,210	000ECDFA5D2DDBDF7479...	228,673	2015-05-28	2,015	228673_Outcome
147,423,657	0004FF72BDDF25A752B70...	267,464	2012-12-13	2,012	267464_Predictor
312,100,406	0008D1BACA977284D9DD...	232,150	2015-09-20	2,015	232150_Outcome
225,432,633	00017FF6229848D7515CB...	232,143	2014-08-30	2,014	232143_Predictor
227,105,065	00017FF6229848D7515CB...	1,840,189	2014-09-05	2,014	1840189_Predictor
294,869,698	00044C8B32E2D18B687E6...	2,151,632	2015-08-01	2,015	2151632_Outcome
106,676,859	000491C9528EBE3302CA4...	232,307	2011-05-03	2,011	232307_Predictor
256,942,638	000859B4AFE2A75400EDD...	281,329	2015-02-28	2,015	281329_Outcome
293,192,416	0002576AD91EE9065846B...	578,920	2015-07-27	2,015	578920_Outcome
335,151,566	0003911E45A14147EF812...	578,920	2015-12-26	2,015	578920_Outcome
336,762,288	0003976319D8996F54F4A...	293,333	2016-01-02	2,016	293333_Outcome

Connect R to MySQL: Building Connection

```
#install R package to obtain the relevant functions
install.packages("RMySQL")
#activate the package and relevant functions
library(RMySQL)
#Connect R to MySQL (template):
mydb = dbConnect(MySQL(), user='user', password='password', dbname='database_name', host='host / database url')
#An example of the connection
mydb = dbConnect(MySQL(), user='555555', password='P555555', dbname='D555555', host='johnson.org.aalto.fi')
```

Connect R to MySQL: Fetch Data

```
#Show the list of the tables in our database connection.
dbListTables(mydb)
#Retrieve data from the database.
rs = dbSendQuery(mydb, "select * from payments")
#Save the retrieved data to a R dataframe.
df = fetch(rs, n=-1)
#This function saves retrieved MySQL data as a data frame object. The n in the function specifies the number of records to retrieve, using n=-1 retrieves all pending records.
View(df) #This function helps you to see the data.
```

Connect R to MySQL: Descriptive Statistics

```
View(attitude)  #plot an embedded dataset termed as attitude
plot(attitude)  #plot an embedded dataset termed as attitude
install.packages("skimr")
#install a package to get function for descriptive statistics
library(skimr) #activate the package
skim(attitude)
skim(df) #activate the package
cor(attitude) #correlation analysis
```

Other functions for root users

Task: create a new database and a new user account (including both ID and password), and grant permission to the user to use the database

```
create database newdatabase;
- create USER 'user' identified BY 'password';
- create USER 'newtest5' identified BY 'newtest5' password expire;
grant all ON newdatabase.* TO 'newtest5';
```

Other functions for root users

Importing large file: MySQL only allows to import relatively small size file in default. For importing large file, you need to change settings.

```
ERROR 1153 (08S01) at line 96: Got a packet bigger than 'max_allowed_packet' bytes
```

Set global max_allowed_packet = 2*1024*1024*10 ;

Set global wait_timeout=1000;

Set global interactive_timeout=1000;

Text mining using MySQL

- Your company received a huge amount of customer reviews on the products of the company from e.g., product review website or Facebook.
- You are supposed to compute associations between words, customers emotion expressed in the review, and how an emotion is associated with a particular product attribute.
- You have found and download the list of positive and negative emotion words.

Demonstration via MySQL as an analytic project

1. Code is a treasure!
2. The code can be copied and reused to do association or emotion analysis on any textual data, if imported into MySQL.

Summary of the MySQL skills (1)

- **1. Skills of Managing MySQL data file**
  - **Nature of .sql file (a summary MySQL commands).**
  - **Import .sql file**
  - **Import .csv file**
  - **Export .sql file (drop versus not drop)**
  - **Export to be .csv file (ctrl + End)**

Summary of the MySQL skills (2)

- **2. Skills of managing MySQL account**
  - **Remote server account**
  - **Local user account**

Summary of the MySQL skills (3)

- **3. Skills of operating MySQL database**
  - **Creating a new database**
    - **Create database [if not exists] *DatabaseName;***
  - **Removing an existing database**
    - **Drop database [if exists] *DatabaseName;***
  - **Activate a database (default database)**
    - **Use *DatabaseName***
  - **Show the names of the all the databases**
    - **Show databases**

Summary of the MySQL skills (4)

- **4. Skills of operating MySQL tables (2)**
  - **Show columns information of a table**
    - **Show columns from *TableName;***
    - **Describe *TableName*;**
    - **Desc *TableName*;**

COLUMNS (6×8)

Field	Type	Null	Key	Default	Extra
employeeNumber	int(11)	NO	PRI	(NULL)	
lastName	varchar(50)	NO		(NULL)	
firstName	varchar(50)	NO		(NULL)	
extension	varchar(10)	NO		(NULL)	
email	varchar(100)	NO		(NULL)	
officeCode	varchar(10)	NO	MUL	(NULL)	
reportsTo	int(11)	YES	MUL	(NULL)	
jobTitle	varchar(50)	NO		(NULL)	

Summary of the MySQL skills (5)

- **5. Skills of operating MySQL tables (3)**
  - **Copy the structure and indexes, but not the data:**
    - **create table new_table like old_table;**
  - **Copy the structure, indexed and the data**
    - **Create table new_table like old_table;**
    - **Insert new_table select * from old_table;**
  - **Copy the data and the structure, but not the indexes:**
    - **create table new_table as select * from old_table;**

Summary of the MySQL skills (6)

- **6. Skills of adding comment to MySQL query**
  - **From a “#” character to the end of the line.**
  - **From a “-- ” sequence to the end of the line.**
  - **From a /* sequence to the following */ sequence.**

Summary of the MySQL skills (7)

- **7. Skills of creating a table**

```
create table TableName (Variable1    datatype [constraint],
                        Variable2    datatype [constraint],
                        Variable3    datatype [constraint],
                        ……
                        );
```

Summary of the MySQL skills (8)

- **8. Skills of operating datatype**
  - **Numeric Types (integer, decimal and float)**
  - **Date and Time Types (year, date, datatime)**
  - **String Types (Char, Varchar)**

Summary of the MySQL skills (9)

- **9. Skills of operating key in a table**
  - **Primary key, unique key and Foreign key**
  - **Adding key**
  - **Removing key**
  - **Set not null and auto_increment function**
  - **Entity-relationship diagram (ERD)**

Summary of the MySQL skills (10)

- **Skills of using select commands**
  - **Select for calculation and other function**
    - **Select 5+5; Select curtime(), curdate();**
  - **Select reserved word using `**
  - **Select all columns and rows using ***
  - **Select … where…**
    - **=, <, <=, >, >=, !=**
    - **And / or / not**
    - **()**
    - **Between… and …**
    - **Not between … and …**
    - **+ ; - ; * ; / ;**

Summary of the MySQL skills (11)

- **Skills of using select commands**
  - **Select … limit x, y**
  - **Select … order by desc | asc**

Summary of the MySQL skills (12)

- **Select…like [binary]…**
- **Select… IN…**
- **Select … REGEXP…**
- **Select…Distinct…**
- **Select … LEFT(*str*,*len*) …**
- **Select … LENGTH(*str*) …**

Summary of the MySQL skills (13)

- Select … TRIM …
- Select … SUBSTRING(str, pos, len) …
- SUBSTRING_INDEX(*str*,*delim*,*count*)
- Select… REPLACE(*str*,*from_str*,*to_str*) …
- Select…Group by…
- Count() + group by

Summary of the MySQL skills (14)

- **Count(Distinct)+ Group by**
- **group_concat()+ Group by**
- **Select…Group by + having**
- **As [alias]**
- **= '' != '' is null is not null**

Summary of the MySQL skills (15)

- **DATE(*expr*)**
- **STR_TO_DATE()**
- **DAYNAME(date)**
- **DAYOFMONTH(date)**
- **DATE_ADD(date, INTERVAL expr unit)**
- **DATEDIFF(expr1,expr2)**

Summary of the MySQL skills (15)

- **Alter Table table_name Add column_name datatype**
- **Alter Table table_name Drop column_name**

- **Delete from table_name [where conditions]**

Summary of the MySQL skills (16)

Sub-Queries

- **The result of a select command represents one column (or a list of values). E.g.:**

```
Select attributes
        from table_1
        Where attributes IN| NOT IN
        (Select ONE_column
                from table_2
                Where attributes )
```

Summary of the MySQL skills (17)

- **Update** table_name

  Set column_name1 = value|expression,

  column_name2 = value|expression,

  …

  column_nameN = value|expression

  Where conditions;

Summary of the MySQL skills (I)

- **If**
- **Case when**
- **Join**
- **Table and view**
- **Association analysis**

Bonus of sending your feedback

- **Method:** Webropol-survey, link of which will be sent to your email address.
- **Bonus**: One additional points (the full mark is 100 points)

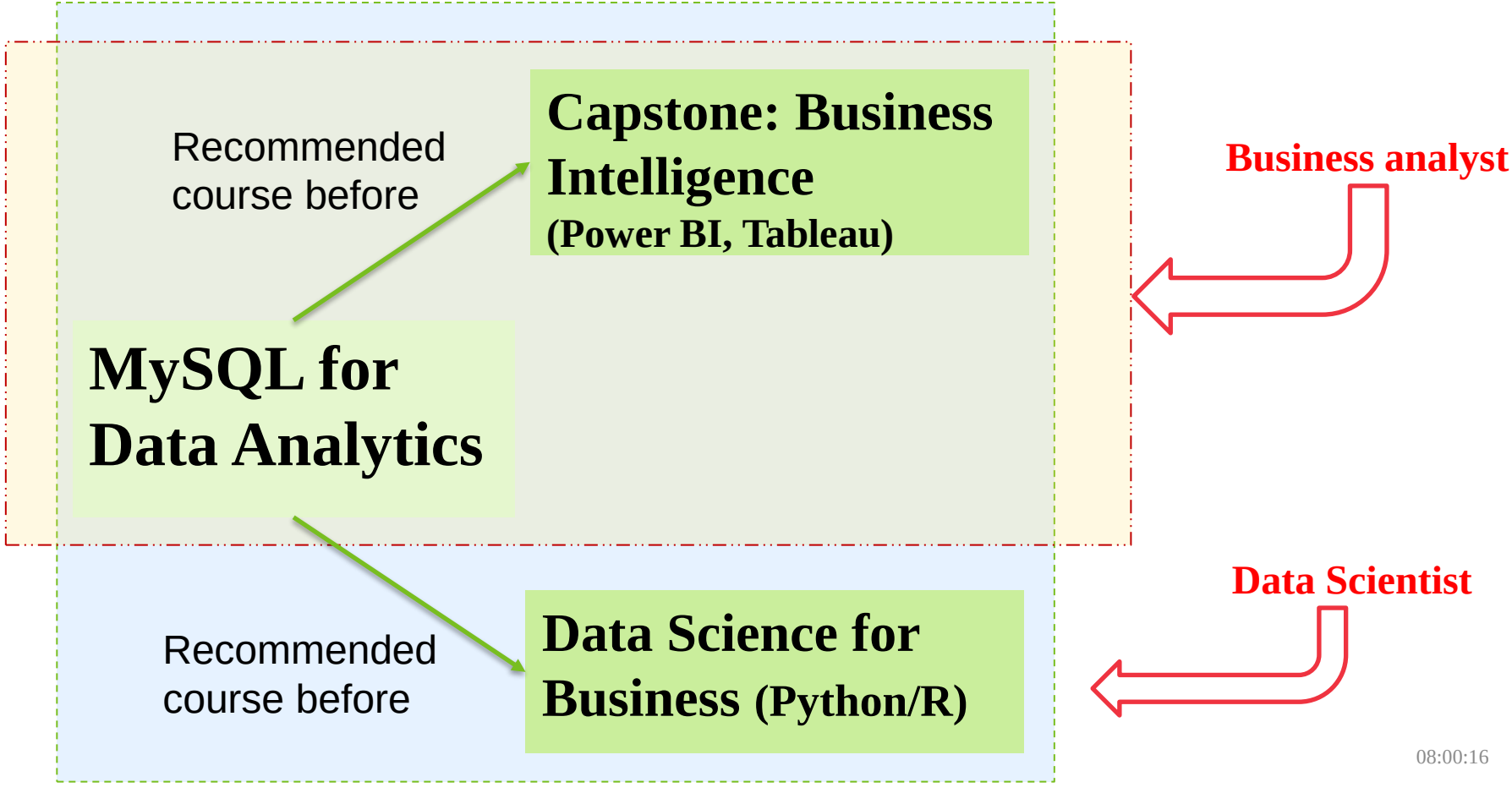
Reflection

- **Programming???**

- **R? Python?**
- **Stata Matlab?**
- **SPSS is not recommended – you cannot remember what you have done to your data**

http://stanfordphd.com/Statistical_Software.html

Finally

- **Statistics Explained - A Guide for Social Science Students**
- **R, Stata, Matlab, Python?**
- **Econometrics**

Connection to Other Courses/Jobs