Data Mining: KDD Process, Frequent Itemsets, Association Rules, Frequent-Pattern Trees

Text:

Chapter 26

Other references:

Data Mining: Concepts and Techniques, by Han and Kimber, Second Edition (Chapters 1 & 5)

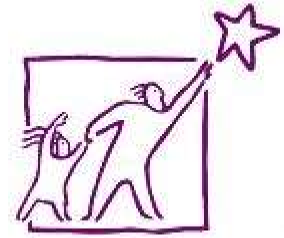
Databases: the continuing saga

When last we left databases…

- We had decided they were great things
- We knew how to conceptually model them in ER diagrams
- We knew how to logically model them in the relational model
- We knew how to normalize our database relations
- We could write queries in different languages
- We'd processed things so people could analyze them

Now: What do we do with all that data?

Learning Goals



- Define the term *knowledge discovery*.
- Explain the general steps involved in the *knowledge discovery in databases* (KDD) process
- Comment on the benefits and challenges that data mining has when dealing with imperfect data quality, especially in large datasets (e.g., data mining can point out anomalies (outliers), optimize the use of human time, detect patterns in data (including patterns that are there just by chance).
- Explain the value of finding frequent itemsets and association rules. Provide some real-world examples of their use (e.g., retailing, biology).
- Explain the purpose of association rules.
- Apply the Apriori Algorithm and compute frequent itemsets and association rules (by hand, for a small dataset).

A Definition of Data Mining

Data mining is the exploration and analysis of large quantities of data in order to discover valid, novel, potentially useful, and ultimately understandable patterns in data.

Valid:	The patterns hold in general.
Novel:	We did not know the pattern beforehand.
Useful:	We can devise actions from the patterns (business intelligence).
Under-standable:	We can interpret and comprehend the patterns.

What about exceptions to patterns? (Outliers or anomalies)

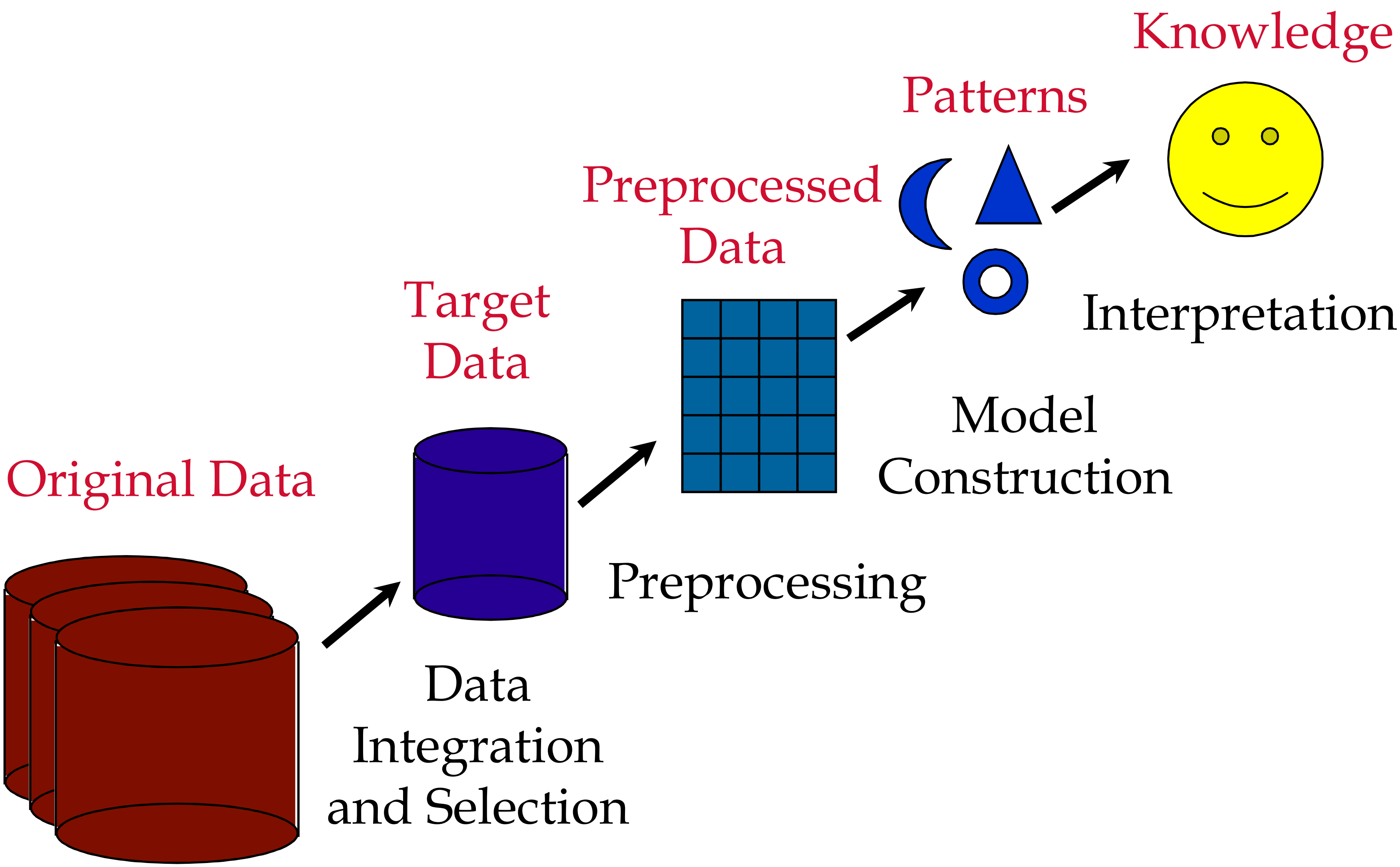
Characteristics of Data Mining

- Key Characteristics
  - Large, multidimensional datasets
  - Efficient algorithms to “discover” knowledge
- What’s the connection with database systems?
  - Managing the data
    - Extract, Transform, and Load
    - There may be many distributed, heterogeneous sources.
  - Large numbers of records; many dimensions
  - Heavy emphasis on I/Os
  - Query evaluation and optimization considerations

Characteristics of Data Mining (cont.)

- Where does the data come from?
  - Databases, flat files, Excel, numerous applications, …
  - Data warehouses: Large, read-mostly, summarized data, that's often downloaded from OLTP (transactional) systems
  - Sensor networks; cameras; satellites (already 15-20 years ago, we were generating 250,000 CDs of data per day)
  - Web logs
  - Transaction data (e.g., supermarkets)

The KDD Process



Market Analysis and Management

- Where are the data sources for market analysis?
  - Credit card transactions, loyalty cards, discount coupons, customer complaints, demographics, surveys
- Target marketing
  - Find clusters of “model” customers who share the same characteristics: interest, income level, spending habits, etc.
    - Higher income groups
    - Lower income groups
    - BMW owners
    - New baby in the household
- Determine customer purchasing patterns over time
  - Conversion of single to a joint bank account: marriage, etc.

Market Analysis and Management (cont.)

- Cross-market analysis
  - Associations or correlations between product sales
    - Amazon.com: People who bought X also bought Y.
  - Predictions based on association information (e.g., Recommender Systems)
    - Suppose you watch some movies.
    - You rank/rate those movies (e.g., You liked “Titanic” and “Slumdog Millionaire”; but, you didn’t like “Apollo 13” and “Terminator 2”).
    - People who rated movies similarly to the way you did, share a common profile or “cluster”.
      - So, let’s find other points (people) in this cluster, and then find out what other movies they liked (and that you haven’t seen yet).
      - Chances are that you’ll like their choices, too; but, if not, we can further refine the cluster, and establish a narrower profile. Iterate.
      - Compare this to the old way of “film critics” deciding which are the “best” movies for you to see.

Market Analysis and Management (cont.)

- Customer Profiling
  - Data mining can tell you what types of customers buy what products (e.g., via clustering or classification).
- Identifying Customer Requirements
  - Identify the best products for different customers.
  - Try to predict the factors that will attract new customers, or retain current customers.
- Summary Information
  - Statistical summaries (e.g., central tendency and variation)

Fraud Detection and Management

- Applications
  - Widely used in health care, retail, credit card services, telecommunications (phone card fraud), etc.
- Approach
  - Use historical data to build models of fraudulent behavior and use data mining to help identify similar instances
- Examples
  - Auto insurance: Detect a group of people who stage accidents to collect on insurance
  - Money laundering: Detect suspicious money transactions (US Treasury's Financial Crimes Enforcement Network—FINCEN)
  - Medical insurance: Detect professional "patients" and rings of doctors and rings of references
  - Telephone calls: Detect patterns that deviate from an expected norm.
  - Retail shrink: Analysts estimate that 38% of retail shrink is due to dishonest employees.

Data Mining Techniques

- Supervised Learning
  - Classification and regression
    - Classify into pre-defined groups, based on previous information
  - Some examples
    - Handwritten postal codes on envelopes
    - Insurance or medical expert systems (decision trees)
      - If (16 ≤ age ≤ 25 & car_type = (“sports car” or “Hummer”) then:
        - risk_status = HIGH
      - If (symptomA = … and symptomB = …) then …

Support question

What is the support of Sushi → Bread (express as a fraction)?

(Reminder: a rule X→Y holds with support *sup* if sup% of transactions contain *X* **AND** *Y*.)

A. 3/7

B. 3/4

C. 4/7

D. None of the above

T1	Sushi, Chicken, Milk
T2	Sushi, Bread
T3	Bread, Vegetables
T4	Sushi, Chicken, Bread
T5	Sushi, Chicken, Ramen, Bread, Milk
T6	Chicken, Ramen, Milk
T7	Chicken, Milk, Ramen

Confidence

Informally: confidence measures which items suggest the others will be there, too.

Formally: A rule X→Y holds with *confidence* conf% if conf% of transactions that contain X also contain Y

T1	Sushi, Chicken, Milk
T2	Sushi, Bread
T3	Bread, Vegetables
T4	Sushi, Chicken, Bread
T5	Sushi, Chicken, Ramen, Bread, Milk
T6	Chicken, Ramen, Milk
T7	Chicken, Milk, Ramen

Ramen → Milk, Chicken [conf = 3/3 = 100%]

Ramen, Chicken → Milk [conf = 3/3 = 100%]

Confidence question

What is the confidence of Sushi → Chicken (express as a fraction)?

(Reminder: A rule X→Y holds with *confidence* conf% if conf% of transactions that contain X also contain Y)

A. 3/7

B. 3/4

C. 3/5

D. None of the above

T1	Sushi, Chicken, Milk
T2	Sushi, Bread
T3	Bread, Vegetables
T4	Sushi, Chicken, Bread
T5	Sushi, Chicken, Ramen, Bread, Milk
T6	Chicken, Ramen, Milk
T7	Chicken, Milk, Ramen

So when is a rule valid?

A rule is valid if its support is above a given threshold (minimum support) and its confidence is over another given threshold (minimum confidence).

A frequent itemset is a set of items that has at least minimum support

T1	Sushi, Chicken, Milk
T2	Sushi, Bread
T3	Bread, Vegetables
T4	Sushi, Chicken, Bread
T5	Sushi, Chicken, Ramen, Bread, Milk
T6	Chicken, Ramen, Milk
T7	Chicken, Milk, Ramen

In this example, {chicken, milk, ramen} is a frequent itemset if the minimum support is less than 3/7.

The Apriori algorithm key idea

Calculating association rules on terabytes of data can be sloooowww. The slowest part is *counting the support*.

The Apriori algorithm speeds things up based on the observation that each subset of a frequent itemset must *also* be a frequent itemset

For example, since rice only appears one time, it can't appear 2 or times with anything else.

Transaction	Items
T1	apple, dates, rice, corn
T2	corn, dates, tuna
T3	apple, corn, dates, tuna
T4	corn, tuna

Minimum support = 50%

Apriori exercise, part the first

Start by finding the support of all itemsets of size 1

Support:

{apple} = 2/4
{corn} = 4/4
{dates} = 3/4
{rice} = 1/4
{tuna} = 3/4

Transaction	Items
T1	apple, dates, rice, corn
T2	corn, dates, tuna
T3	apple, corn, dates, tuna
T4	corn, tuna

Minimum support = 50%

Apriori round 2:
Find all frequent itemsets of size 2

All possible itemsets of size 2:

{apple, corn}
{apple, dates}
{apple, rice}
{apple, tuna}
{corn, dates}
{corn, rice}
{corn, tuna}
{dates, rice}
{dates, tuna}
{rice, tuna}

Transaction	Items
T1	apple, dates, rice, corn
T2	corn, dates, tuna
T3	apple, corn, dates, tuna
T4	corn, tuna

Minimum support = 50%

Because {rice} only occurs once, anything including {rice} can’t occur 2 or more times, so we can ignore itemsets including {rice}.

Apriori round 2:
Find all frequent itemsets of size 2

All possible itemsets of size 2:

{apple, corn}
{apple, dates}
~~{apple, rice}~~
{apple, tuna}
{corn, dates}
~~{corn, rice}~~
{corn, tuna}
~~{dates, rice}~~
{dates, tuna}
~~{rice, tuna}~~

Transaction	Items
T1	apple, dates, rice, corn
T2	corn, dates, tuna
T3	apple, corn, dates, tuna
T4	corn, tuna

Minimum support = 50%

Because {rice} is not frequent, anything including {rice} is not frequent, so we can ignore itemsets including {rice}.

Apriori round 2:
Find all frequent itemsets of size 2

All possible itemsets of size 2:

{apple, corn}
{apple, dates}
~~{apple, rice}~~
{apple, tuna}
{corn, dates}
~~{corn, rice}~~
{corn, tuna}
~~{dates, rice}~~
{dates, tuna}
~~{rice, tuna}~~

Transaction	Items
T1	apple, dates, rice, corn
T2	corn, dates, tuna
T3	apple, corn, dates, tuna
T4	corn, tuna

Minimum support = 50%

Group exercise: count support for remaining itemsets.

Apriori round 2:
Find all frequent itemsets of size 2

Support for all possible itemsets of size 2:

{apple, corn} = 2/4
{apple, dates} = 2/4
~~{apple, rice}~~
{apple, tuna} = 1/4
{corn, dates} = 3/4
~~{corn, rice}~~
{corn, tuna} = 3/4
~~{dates, rice}~~
{dates, tuna} = 2/4
~~{rice, tuna}~~

Transaction	Items
T1	apple, dates, rice, corn
T2	corn, dates, tuna
T3	apple, corn, dates, tuna
T4	corn, tuna

Minimum support = 50%

Group exercise: what are the frequent itemsets of size 2?

Apriori round 2:
Find all frequent itemsets of size 2

All frequent itemsets of size 2:

{apple, corn}
{apple, dates}
{corn, dates}
{corn, tuna}
{dates, tuna}

Transaction	Items
T1	apple, dates, rice, corn
T2	corn, dates, tuna
T3	apple, corn, dates, tuna
T4	corn, tuna

Minimum support = 50%

Apriori round 3: Find all frequent itemsets of size 3

Given frequent itemsets of size 2

{apple, corn}
{apple, dates}
{corn, dates}
{corn, tuna}
{dates, tuna}

Transaction	Items
T1	apple, dates, rice, corn
T2	corn, dates, tuna
T3	apple, corn, dates, tuna
T4	corn, tuna

Minimum support = 50%

Without counting support, what are all the possible frequent itemsets of size 3?

(remember: in order for an itemset to be possibly frequent, all subsets of it must be frequent, e.g., {apple, corn, rice} is not a possible frequent itemset because {rice} is not a frequent itemset)

Apriori round 3: Find all frequent itemsets of size 3

Great! Now count support for the remaining itemsets

(Group exercise):

{apple, corn, dates}
{corn, dates, tuna}

Transaction	Items
T1	apple, dates, rice, corn
T2	corn, dates, tuna
T3	apple, corn, dates, tuna
T4	corn, tuna

Minimum support = 50%

Apriori round 3: Find all frequent itemsets of size 3

Great! Now count support for the remaining itemsets

(Group exercise):

{apple, corn, dates} = 2/4
{corn, dates, tuna} = 2/4

Since 2/4 = 50%, both are frequent

Transaction	Items
T1	apple, dates, rice, corn
T2	corn, dates, tuna
T3	apple, corn, dates, tuna
T4	corn, tuna

Minimum support = 50%

Apriori example: done!

The whole list of frequent itemsets for this example is:

{apple}
{corn}
{dates}
{tuna}
{apple, corn}
{apple, dates}
{corn, dates}
{corn, tuna}
{dates, tuna}
{apple, corn, dates}
{corn, dates, tuna}

Transaction	Items
T1	apple, dates, rice, corn
T2	corn, dates, tuna
T3	apple, corn, dates, tuna
T4	corn, tuna

Minimum support = 50%

Apriori example: done!

Frequent itemsets

{apple}
{corn}
{dates}
{tuna}
{apple, corn}
{apple, dates}
{corn, dates}
{corn, tuna}
{dates, tuna}
{apple, corn, dates}
{corn, dates, tuna}

Itemsets we counted support for:

{apple}
{corn}
{dates}
{rice}
{tuna}
{apple, corn}
{apple, dates}
{apple, tuna}
{corn, dates}
{corn, tuna}
{dates, tuna}
{apple, corn, dates}
{corn, dates, tuna}

All possible itemsets:

{apple}
{corn}
{dates}
{rice}
{tuna}
{apple, corn}
{apple, dates}
{apple, rice}
{apple, tuna}
{corn, dates}
{corn, rice}
{corn, tuna}
{dates, rice}
{dates, tuna}
{rice, tuna}
{apple, corn, dates}
{apple, corn, rice}
{apple, corn, tuna}
{corn, dates, rice}
{corn, dates, tuna}
{dates, rice, tuna}
{apple, corn, dates, rice}
{apple, corn, dates, tuna}
{corn, dates, rice, tuna}
{apple, corn, dates, rice, tuna}

Apriori algorithm formalized

1. Find the frequent itemsets of size 1; call this F_1
2. For k=1 until there are no more frequent itemsets
   1. Form candidate itemsets of size k+1: C_{k+1} is the set of itemsets of size k+1 where all subsets of C_{k+1} are frequent itemsets
   2. Count support of items in C_{k+1}
   3. F_{k+1} = itemsets in C_{k+1} that are frequent itemsets
3. Answer is the union of all F_k

Transaction	Items
T1	apple, dates, rice, corn
T2	corn, dates, tuna
T3	apple, corn, dates, tuna
T4	corn, tuna

Minimum support = 75%

F_1 = {{corn}, {dates}, {tuna}}

Apriori algorithm formalized

1. Find the frequent itemsets of size 1; call this F_1
2. For k=1 until there are no more frequent itemsets
   1. Form candidate itemsets of size k+1: C_{k+1} is the set of itemsets of size k+1 where all subsets of C_{k+1} are frequent itemsets
   2. Count support of items in C_{k+1}
   3. F_{k+1} = itemsets in C_{k+1} that are frequent itemsets
3. Answer is the union of all F_k

Reminder: F_1 = {{corn},
{dates}
{tuna}}

Transaction	Items
T1	apple, dates, rice, corn
T2	corn, dates, tuna
T3	apple, corn, dates, tuna
T4	corn, tuna

Minimum support = 75%

{C_2 = {{corn, dates},
{corn, tuna},
{dates, tuna}}

Apriori algorithm formalized

1. Find the frequent itemsets of size 1; call this F_1
2. For k=1 until there are no more frequent itemsets
   1. Form candidate itemsets of size k+1: C_{k+1} is the set of itemsets of size k+1 where all subsets of C_{k+1} are frequent itemsets
   2. Count support of items in C_{k+1}
   3. F_{k+1} = itemsets in C_{k+1} that are frequent itemsets
3. Answer is the union of all F_k

Transaction	Items
T1	apple, dates, rice, corn
T2	corn, dates, tuna
T3	apple, corn, dates, tuna
T4	corn, tuna

Minimum support = 75%

Support: {{corn, dates} =3/4
{corn, tuna} =3/4
{dates, tuna}=2/4

Apriori algorithm formalized

1. Find the frequent itemsets of size 1; call this F_1
2. For k=1 until there are no more frequent itemsets
   1. Form candidate itemsets of size k+1: C_{k+1} is the set of itemsets of size k+1 where all subsets of C_{k+1} are frequent itemsets
   2. Count support of items in C_{k+1}
   3. F_{k+1} = itemsets in C_{k+1} that are frequent itemsets
3. Answer is the union of all F_k

Transaction	Items
T1	apple, dates, rice, corn
T2	corn, dates, tuna
T3	apple, corn, dates, tuna
T4	corn, tuna

Minimum support = 75%

F_2 = {{corn, dates}, = 3/4
{corn, tuna} = 3/4
~~{dates, tuna}} = 2/4~~

Apriori algorithm formalized

1. Find the frequent itemsets of size 1; call this F_1
2. For k=1 until there are no more frequent itemsets
   1. Form candidate itemsets of size k+1: C_{k+1} is the set of itemsets of size k+1 where all subsets of C_{k+1} are frequent itemsets
   2. Count support of items in C_{k+1}
   3. F_{k+1} = itemsets in C_{k+1} that are frequent itemsets
3. Answer is the union of all F_k

Transaction	Items
T1	apple, dates, rice, corn
T2	corn, dates, tuna
T3	apple, corn, dates, tuna
T4	corn, tuna

Answer = F1 ∪ F2 ∪ F_3 =
{{corn}, {dates}, {tuna},
{corn, dates},{corn, tuna}}

Great! Your turn!

Use the Apriori algorithm to find frequent itemsets with a support threshold of 3/7 transactions. Write down all steps!

Algorithm reminder:

1. Find the frequent itemsets of size 1; call this F_1
2. For k=1 until there are no more
   1. Form candidate itemsets of size k+1: C_{k+1} is the set of itemsets of size k+1 where all subsets of C_{k+1} are frequent itemsets
   2. Count support of items in C_{k+1}
   3. F_{k+1} = itemsets in C_{k+1} that are frequent itemsets
3. Answer is the union of all F_k

Transaction	Items
T1	cake, jam, rolls, tea
T2	cake, jam, tea
T3	cake, jam
T4	jam, rolls, tea
T5	jam, rolls
T6	rolls, tea
T7	jam, tea

Walking through the example

- Minimum support = 3/7
- Support for C_1:
  {cake} = 3
  {jam} = 6
  {rolls} = 4
  {tea} = 5
- F_1={{cake},{jam},{rolls},{tea}}

Transaction	Items
T1	cake, jam, rolls, tea
T2	cake, jam, tea
T3	cake, jam
T4	jam, rolls, tea
T5	jam, rolls
T6	rolls, tea
T7	jam, tea

Walking through the example

- Minimum support = 3/7
- Support for C_2:
  {cake, jam} = 3
  {cake, rolls} = 1
  {cake, tea} = 2
  {jam, rolls} = 3
  {jam, tea} = 4
  {rolls, tea} = 3
- F_2={{cake, jam}, {jam,rolls}, {jam, tea}, {rolls, tea}}
- Support for C_3: {jam, rolls, tea} = 2
- F_3 = {}

Transaction	Items
T1	cake, jam, rolls, tea
T2	cake, jam, tea
T3	cake, jam
T4	jam, rolls, tea
T5	jam, rolls
T6	rolls, tea
T7	jam, tea

Reminder: F_1={{cake},{jam},{rolls},{tea}}

It shook up the research world

It has over 20,000 citations! Why?

- It’s something people really needed
- It scales really well
- It’s easy to understand
- Lots to extend

Research? See ACM SIGKDD, etc. Conferences and Proceedings

- Association for Computing Machinery's Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD)
  - Related organizations: VLDB (Very Large Data Bases) and ACM SIGMOD (Special Interest Group on the Management of Data)
- SIGKDD has participants from 3 major disciplines:
  - Artificial Intelligence (especially Machine Learning)
  - Statistics
  - Databases
- All 3 disciplines deal heavily with algorithms.
  - AI has an emphasis on supervised and unsupervised learning.
  - Statistics has an emphasis on exploratory data analysis, probabilities, inference, and validation.
  - DB has an emphasis on managing large volumes of disk-resident data, especially with respect to I/Os and scalability.

But there's more to data mining than just algorithms

Let's take a bit to think about how data mining is impacting the world around us

An example data mining quote from the NY Times:

"We have the capacity to send every customer an ad booklet, specifically designed for them, that says, 'Here's everything you bought last week and a coupon for it,' " one Target executive told me. 'We do that for grocery products all the time.' But for pregnant women, Target's goal was selling them baby items they didn't even know they needed yet."

https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html

On the other hand...

Grocery store loyalty cards help trace hepatitis A outbreak:

http://o.canada.com/health-2/grocery-store-loyalty-cards-help-b-c-disease-detectives-trace-hepatitis-a-outbreak

And then there’s Facebook

Full page apology in many US and British papers

“You may have heard about a quiz app built by a university researcher that leaked Facebook data of millions of people in 2014. This was a breach of trust, and I’m sorry we didn’t do more at the time. We’re now taking steps to make sure this doesn’t happen again.

We’ve already stopped apps like this from getting so much information. Now we’re limiting the data apps get when you sign in using Facebook.

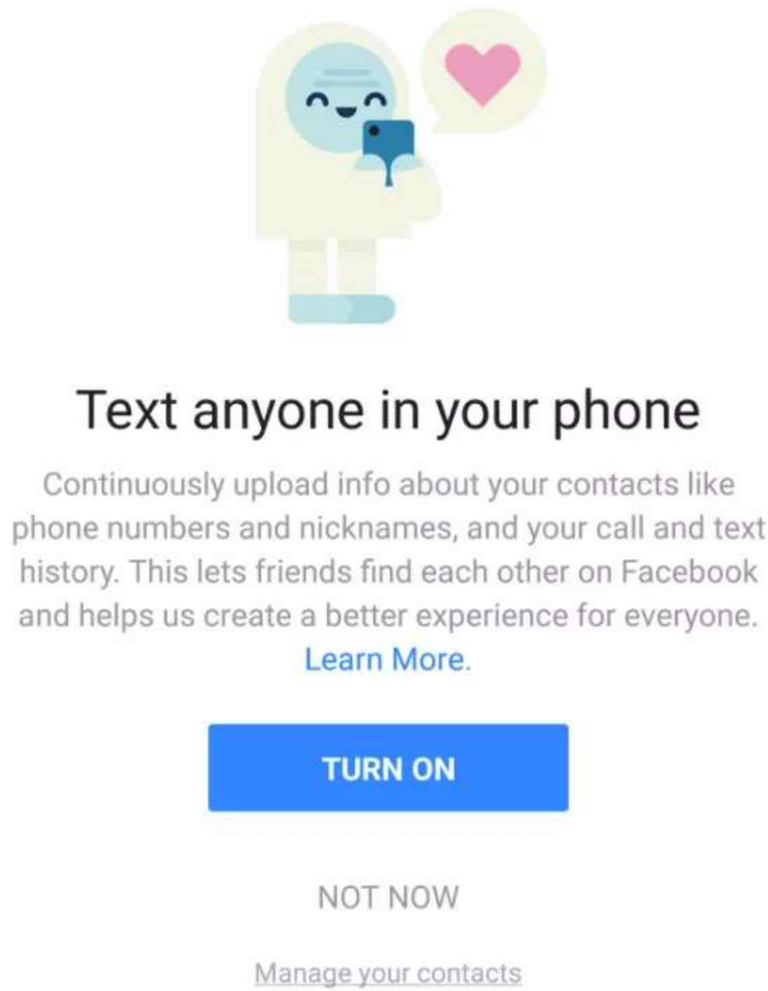
We’re also investigating every single app that had access to large amounts of data before we fixed this. We expect there are others. And when we find them, we will ban them and tell everyone affected.

Finally, we’ll remind you which apps you’ve given access to your information — so you can shut off the ones you don’t want anymore.

Thank you for believing in this community. I promise to do better for you.

Mark Zuckerberg”

Okay, they're sorry. Obviously won't happen again, right?



What actually happened?

"*Ars Technica* reports that Facebook has been requesting access to contacts, SMS data, and call history on Android devices to improve its friend recommendation algorithm and distinguish between business contacts and your true personal friendships."

"Several Twitter users have reported finding months or years of call history data in their downloadable Facebook data file"

"Facebook has responded to the findings, but the company appears to suggest it's normal for apps to access your phone call history when you upload contacts to social apps."

https://www.theverge.com/2018/3/25/17160944/facebook-call-history-sms-data-collection-android

And there's this

An executive order signed by President Trump in January 2017 states:

> "Agencies shall, to the extent consistent with applicable law, ensure that their privacy policies exclude persons who are not United States citizens or lawful permanent residents from the protections of the Privacy Act regarding personally identifiable information."

But even if you trust Facebook with this data, there are still concerns

“**The Trump administration has said it wants to start collecting the social media history of nearly everyone seeking a visa to enter the US.**

The proposal, which comes from the state department, would require most visa applicants to give details of their Facebook and Twitter accounts.

They would have to disclose all social media identities used in the past five years.

About 14.7 million people a year would be affected by the proposals.”

http://www.bbc.com/news/world-us-canada-43601557

How to study?

- Do the practice exercises without looking at the answer key
- Read over your notes
- Look over the book