ISA 401: Business Intelligence & Data Visualization

25: A Short Introduction to Exploratory Data Mining

Fadel M. Megahed, PhD

Raymond E. Glos Professor in Business
Farmer School of Business
Miami University

@FadelMegahed
fmegahed
fmegahed@miamioh.edu
? Automated Scheduler for Office Hours

Fall 2025

A Recap of What we Learned Last Week

- Define a “business report” & its main functions
- Understand the importance of the right KPIs
- Automate traditional business reports
- Dashboards as real-time business reporting tools

Course Objectives Covered so Far

[Y]ou will be re-introduced to **how data should be explored** ... Instead, the focus is on understanding the underlying methodology and mindset of **how data should be approached, handled, explored, and incorporated back into the domain of interest.** ... You are expected to:

✓ **Be capable of extracting, transforming and loading (ETL) data using multiple platforms (e.g. R & Tableau).**

✓ **Write basic R scripts to preprocess and clean the data.**

✓ **Explore the data using visualization approaches that are based on sound human factors (i.e. account for human cognition and perception of data).**

⊗ **Understand how data mining and other analytical tools can capitalize on the insights generated from the data viz process.**

✓ **Create interactive dashboards that can be used for business decision making, reporting and/or performance management.**

⊗ **Be able to apply the skills from this class in your future career.**

Learning Objectives for Today's Class

- Describe the goals & functions of data mining
- Understand the statistical limits on data mining
- Describe the data mining process
- What is “frequent itemsets” & the application of this concept
- Explain how and why “association rules” are constructed
- Use R to populate both concepts

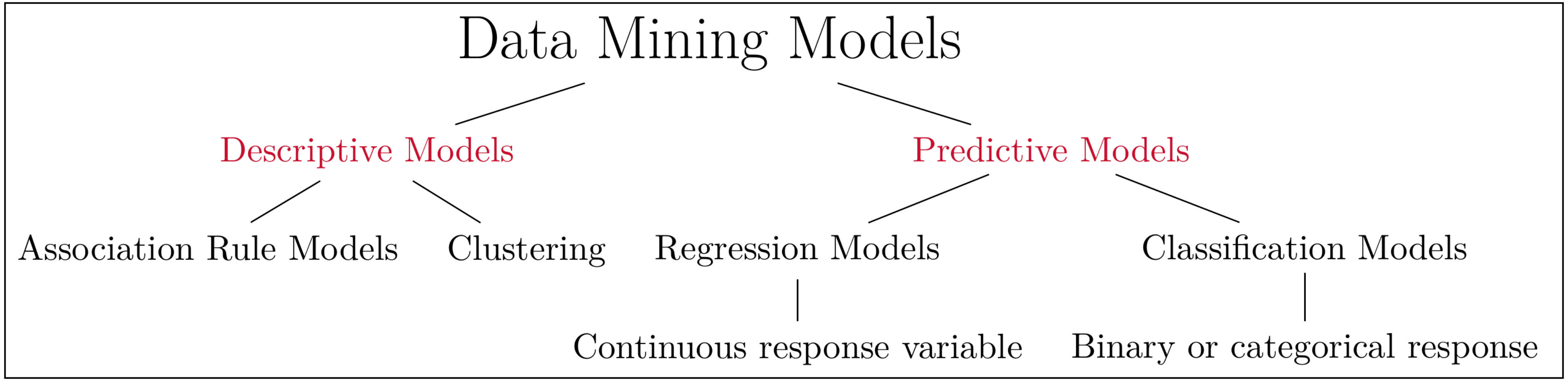
An Overview of Data Mining

What is Data Mining?

- The most common definition of data mining is the discovery of models from data.
- Discovery of **patterns and models that are:**
  - **Valid:** hold on new data with some certainty
  - **Useful:** should be possible to act on the item
  - **Unexpected:** non-obvious to the system
  - **Understandable:** humans should be able to interpret the pattern
- Subsidiary Issues:
  - **Data cleansing:** detection of bogus data
  - **Data visualization:** something better than MBs of output
  - **Warehousing** of data (for retrieval)

Source: The slide is adapted from Jure Leskovic, Stanford CS246, Lecture Notes, see http://cs246.stanford.edu

A Simplistic View of Data Mining Models



A simplistic summary of data mining models. Note that, in ISA 401, we will only briefly cover descriptive/exploratory data mining models

Data Mining is Hard

Data mining is hard since it has the following issues:

- Scalability
- Dimensionality
- Complex and Heterogeneous Data
- Data Quality
- Data Ownership and Distribution
- Privacy Preservation

Note that I have intentionally not included fitting/training a model since this is relatively easy if you understand the data, engineered/captured the important predictors, and have the data in the "correct" shape/quality.

Association Rules

Data

Top 5 Rules

Scatter Plot of all Rules

Graph-based Plot of Top 5 Rules

```
## transactions as itemMatrix in sparse format with
##  9835 rows (elements/itemsets/transactions) and
##  169 columns (items) and a density of 0.02609146
##
## most frequent items:
##       whole milk other vegetables       rolls/buns             soda
##             2513             1903             1809             1715
##           yogurt          (Other)
##             1372            34055
##
## element (itemset/transaction) length distribution:
## sizes
##    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16
## 2159 1643 1299 1005  855  645  545  438  350  246  182  117   78   77   55   46
##   17   18   19   20   21   22   23   24   26   27   28   29   32
##   29   14   14    9   11    4    6    1    1    1    1    3    1
##
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   3.000   4.409   6.000  32.000
##
```

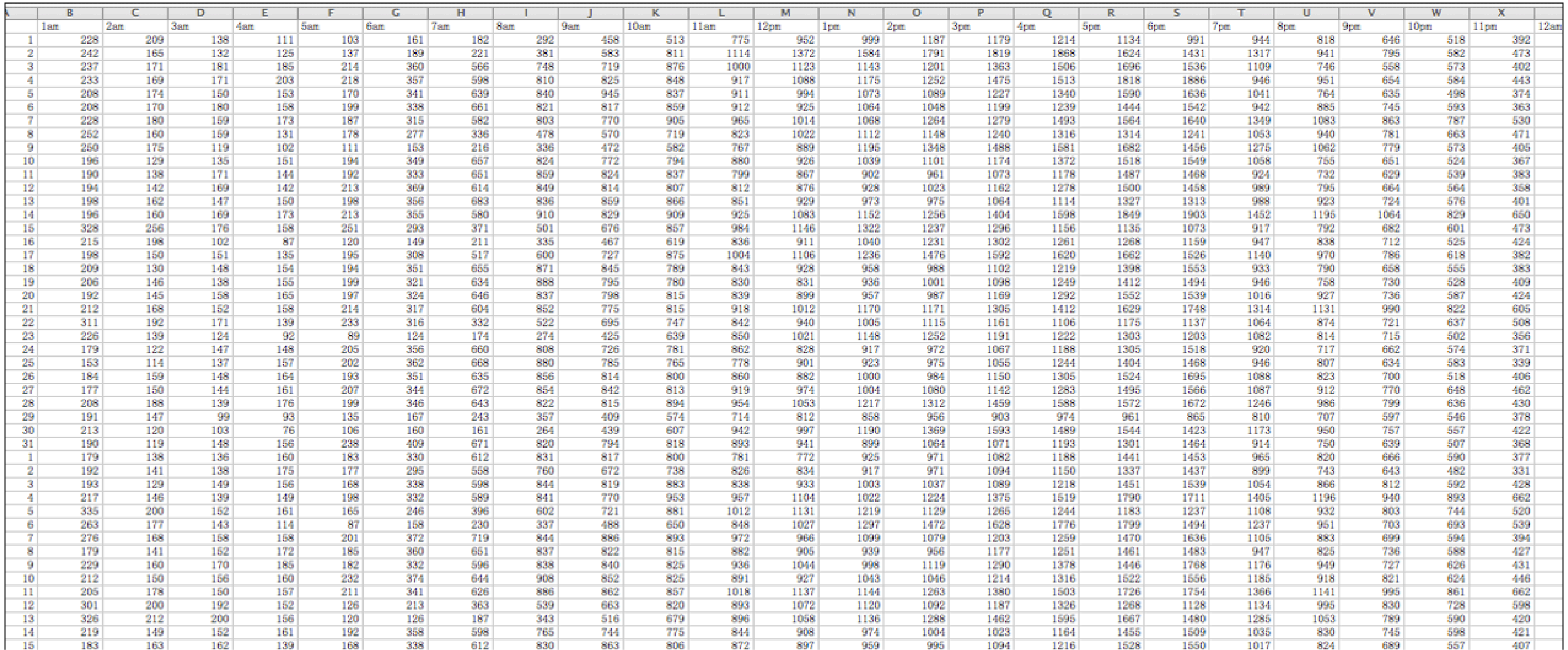
Clustering of Traffic Volume on I-85

04:00

Data

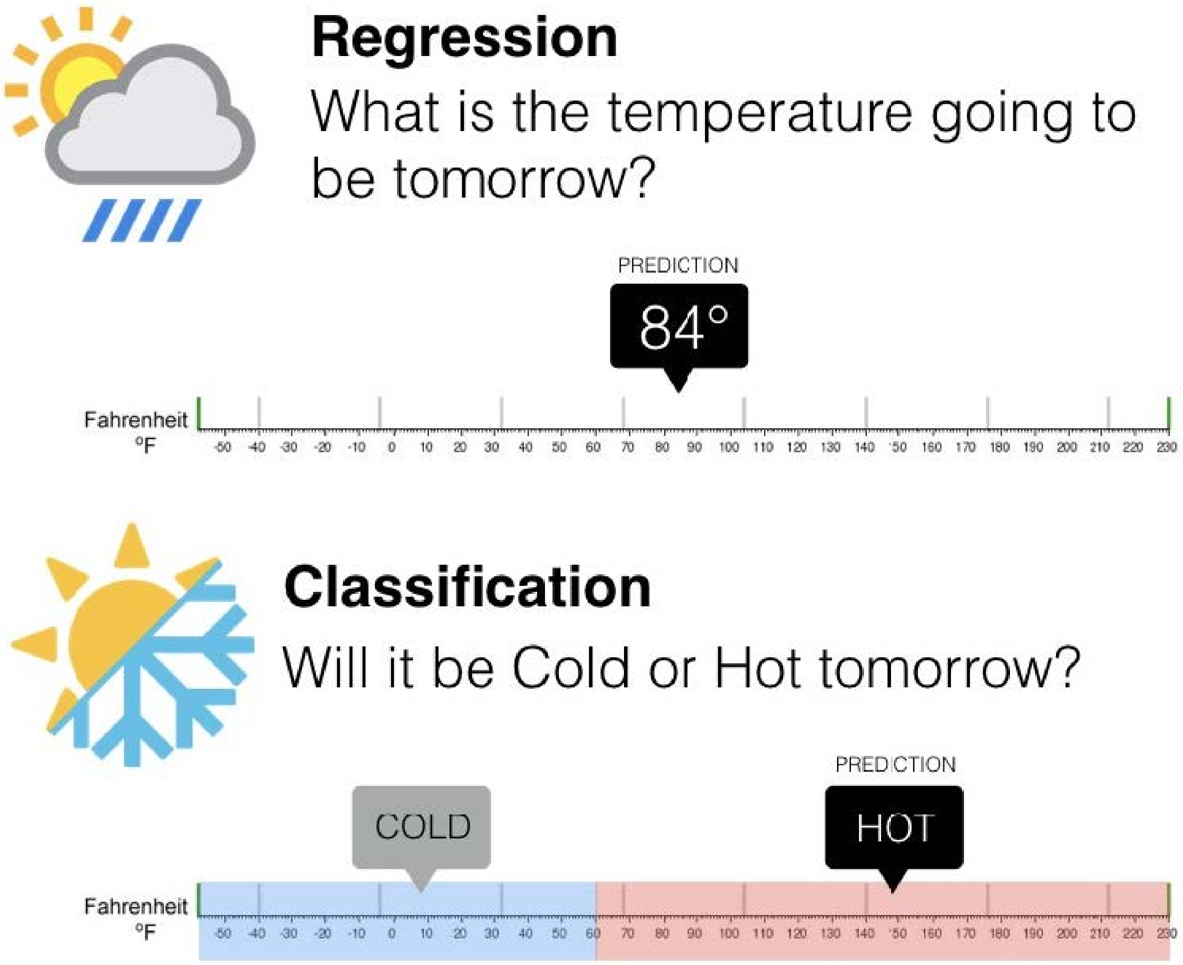
Calendar Plot of Clustered Data

Insights from Chart?



A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	
	1am	2am	3am	4am	5am	6am	7am	8am	9am	10am	11am	12pm	1pm	2pm	3pm	4pm	5pm	6pm	7pm	8pm	9pm	10pm	11pm	12am
1	228	209	138	111	103	161	182	292	458	513	775	952	999	1187	1179	1214	1134	991	944	818	646	518	392	
2	242	165	132	125	137	189	221	381	583	811	1114	1372	1584	1791	1819	1868	1624	1431	1317	941	795	582	473	
3	237	171	181	185	214	360	566	748	719	876	1000	1123	1143	1201	1363	1506	1696	1536	1109	746	558	573	402	
4	233	169	171	203	218	357	598	810	825	848	917	1088	1175	1252	1475	1513	1818	1886	946	951	654	584	443	
5	208	174	150	153	170	341	639	840	945	837	911	994	1073	1089	1227	1340	1590	1636	1041	764	635	498	374	
6	208	170	180	158	199	338	661	821	817	859	912	925	1064	1048	1199	1239	1444	1542	942	885	745	593	363	
7	228	180	159	173	187	315	582	803	770	905	965	1014	1068	1264	1279	1493	1564	1640	1349	1083	863	787	530	
8	252	160	159	131	178	277	336	478	570	719	823	1022	1112	1148	1240	1316	1314	1241	1053	940	781	663	471	
9	250	175	119	102	111	153	216	336	472	582	767	889	1195	1348	1488	1581	1682	1456	1275	1062	779	573	405	
10	196	129	135	151	194	349	657	824	772	794	880	926	1039	1101	1174	1372	1518	1549	1058	755	651	524	367	
11	190	138	171	144	192	333	651	859	824	837	799	867	902	961	1073	1178	1487	1468	924	732	629	539	383	
12	194	142	169	142	213	369	614	849	814	807	812	876	928	1023	1162	1278	1500	1458	989	795	664	564	358	
13	198	162	147	150	198	356	683	836	859	866	851	929	973	975	1064	1114	1327	1313	988	923	724	576	401	
14	196	160	169	173	213	355	580	910	829	909	925	1083	1152	1256	1404	1598	1849	1903	1452	1195	1064	829	650	
15	328	256	176	158	251	293	371	501	676	857	984	1146	1322	1237	1296	1156	1135	1073	917	792	682	601	473	
16	215	198	102	87	120	149	211	335	467	619	836	911	1040	1231	1302	1261	1268	1159	947	838	712	525	424	
17	198	150	151	135	195	308	517	600	727	875	1004	1106	1236	1476	1592	1620	1662	1526	1140	970	786	618	382	
18	209	130	148	154	194	351	655	871	845	789	843	928	958	988	1102	1219	1398	1553	933	790	658	555	383	
19	206	146	138	155	199	321	634	888	795	780	830	831	936	1001	1098	1249	1412	1494	946	758	730	528	409	
20	192	145	158	165	197	324	646	837	798	815	839	899	957	987	1169	1292	1552	1539	1016	927	736	587	424	
21	212	168	152	158	214	317	604	852	775	815	918	1012	1170	1171	1305	1412	1629	1748	1314	1131	990	822	605	
22	311	192	171	139	233	316	332	522	695	747	842	940	1005	1115	1161	1106	1175	1137	1064	874	721	637	508	
23	226	139	124	92	89	124	174	274	425	639	850	1021	1148	1252	1191	1222	1303	1203	1082	814	715	502	356	
24	179	122	147	148	205	356	660	808	726	781	862	828	917	972	1067	1188	1305	1518	920	717	662	574	371	
25	153	114	137	157	202	362	668	880	785	765	778	901	923	975	1055	1244	1404	1468	946	807	634	583	339	
26	184	159	148	164	193	351	635	856	814	800	860	882	1000	984	1150	1305	1524	1695	1088	823	700	518	406	
27	177	150	144	161	207	344	672	854	842	813	919	974	1004	1080	1142	1283	1495	1566	1087	912	770	648	462	
28	208	188	139	176	199	346	643	822	815	894	954	1053	1217	1312	1459	1588	1572	1672	1246	986	799	636	430	
29	191	147	99	93	135	167	243	357	409	574	714	812	858	956	903	974	961	865	810	707	597	546	378	
30	213	120	103	76	106	160	161	264	439	607	942	997	1190	1369	1593	1489	1544	1423	1173	950	757	557	422	
31	190	119	148	156	238	409	671	820	794	818	893	941	899	1064	1071	1193	1301	1464	914	750	639	507	368	
1	179	138	136	160	183	330	612	831	817	800	781	772	925	971	1082	1188	1441	1453	965	820	666	590	377	
2	192	141	138	175	177	295	558	760	672	738	826	834	917	971	1094	1150	1337	1437	899	743	643	482	331	
3	193	129	149	156	168	338	598	844	819	883	838	933	1003	1037	1089	1218	1451	1539	1054	866	812	592	428	
4	217	146	139	149	198	332	589	841	770	953	957	1104	1022	1224	1375	1519	1790	1711	1405	1196	940	893	662	
5	335	200	152	161	165	246	396	602	721	881	1012	1131	1219	1129	1265	1244	1183	1237	1108	932	803	744	520	
6	263	177	143	114	87	158	230	337	488	650	848	1027	1297	1472	1628	1776	1799	1494	1237	951	703	693	539	
7	276	168	158	158	201	372	719	844	886	893	972	966	1099	1079	1203	1259	1470	1636	1105	883	699	594	394	
8	179	141	152	172	185	360	651	837	822	815	882	905	939	956	1177	1251	1461	1483	947	825	736	588	427	
9	229	160	170	185	182	332	596	838	840	825	936	1044	998	1119	1290	1378	1446	1768	1176	949	727	626	431	
10	212	150	156	160	232	374	644	908	852	825	891	927	1043	1046	1214	1316	1522	1556	1185	918	821	624	446	
11	205	178	150	157	211	341	626	886	862	857	1018	1137	1144	1263	1380	1503	1726	1754	1366	1141	995	861	662	
12	301	200	192	152	126	213	363	539	663	820	893	1072	1120	1092	1187	1326	1268	1128	1134	995	830	728	598	
13	326	212	200	156	120	126	187	343	516	679	896	1058	1136	1288	1462	1595	1667	1480	1285	1053	789	590	420	
14	219	149	152	161	192	358	598	765	744	775	844	908	974	1004	1023	1164	1455	1509	1035	830	745	598	421	
15	183	163	162	139	168	338	612	830	863	806	872	897	959	995	1094	1216	1528	1550	1017	824	689	557	407	

Regression vs Classification



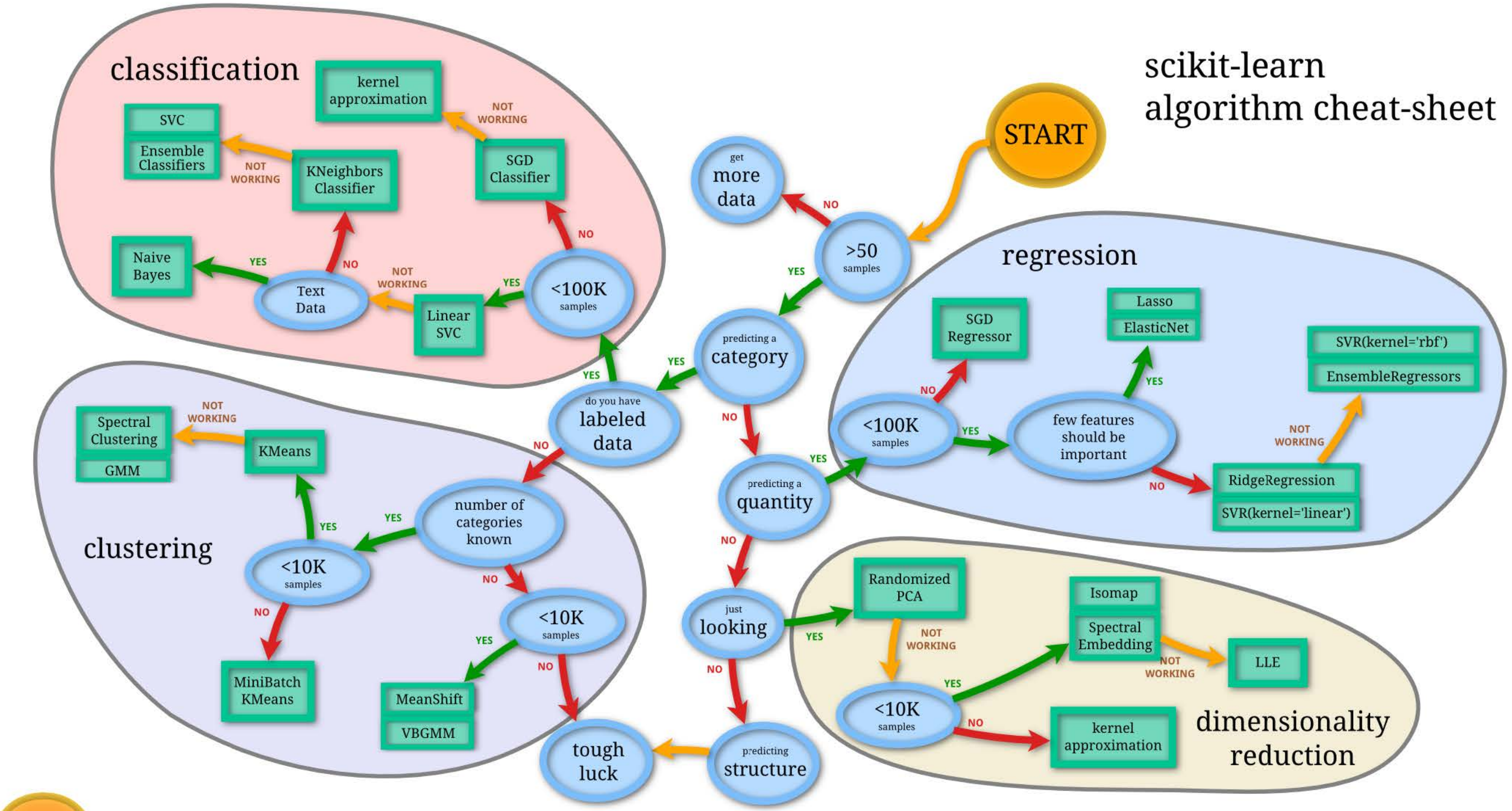
Regression

What is the temperature going to be tomorrow?

Classification

Will it be Cold or Hot tomorrow?

An Overview of Common Data Mining Models

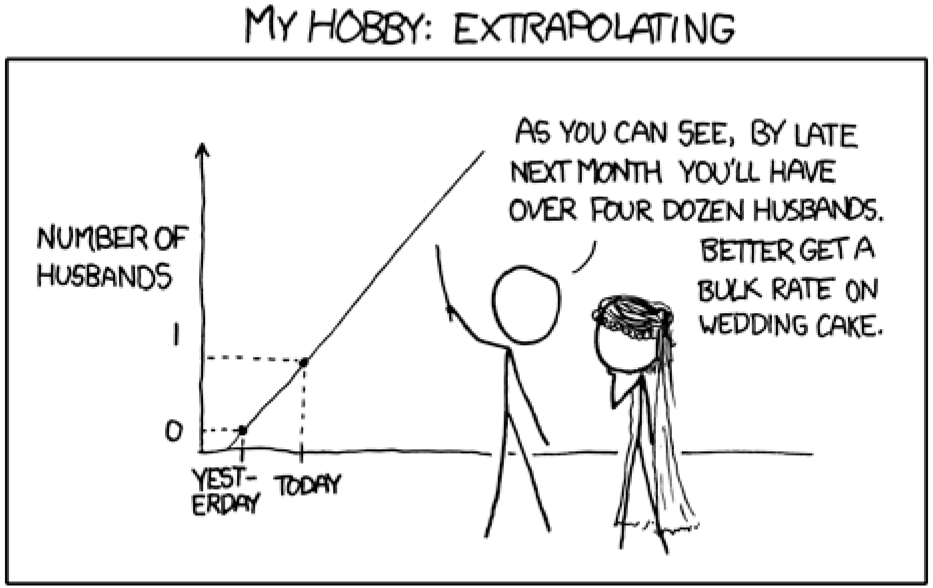


Back

Limits on Data Mining

Meaningfulness of Answers from DM Models

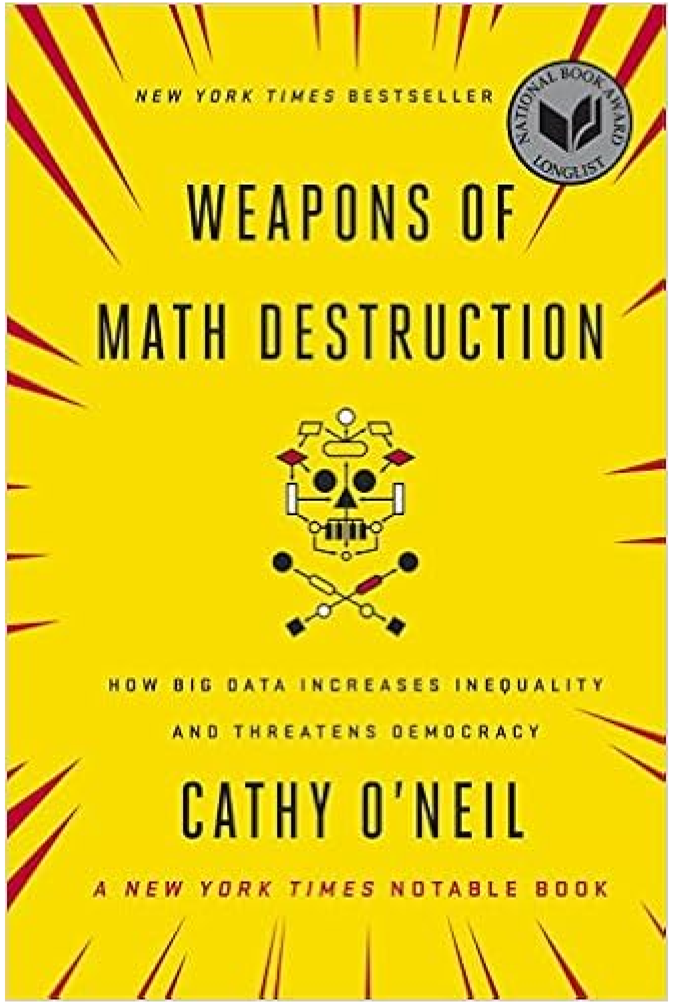
- **A big risk when data mining is that you will discover patterns that are meaningless.**
- **Bonferroni's Principle:** (roughly) if you look in more places for interesting patterns than your amount of data will support, you are bound to find.

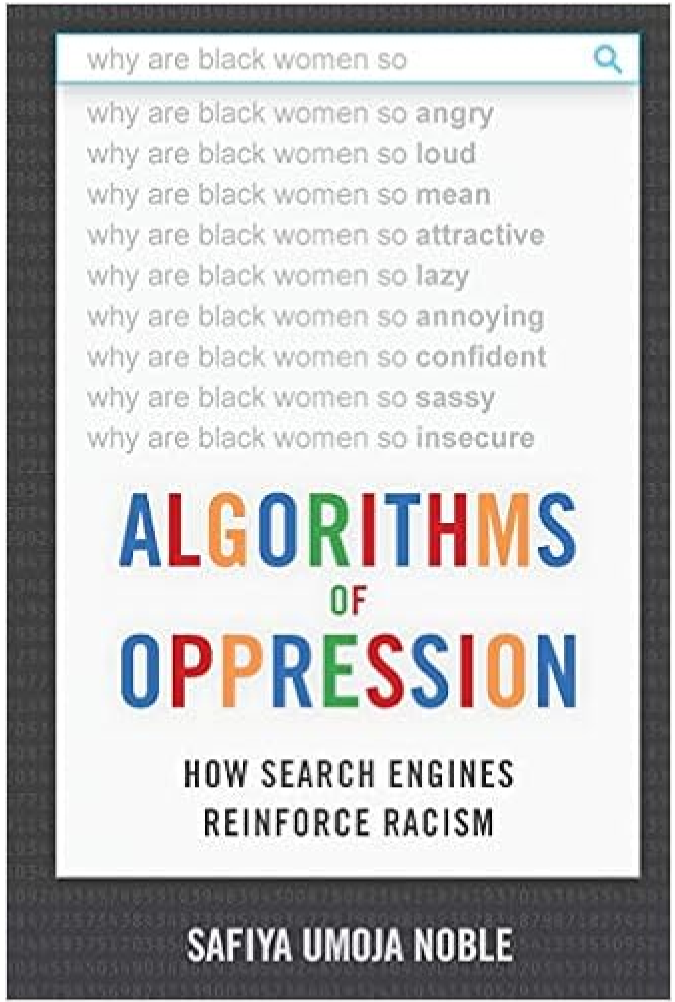

MY HOBBY: EXTRAPOLATING
NUMBER OF HUSBANDS
1
0
YEST- ERDAY
TODAY
AS YOU CAN SEE, BY LATE NEXT MONTH YOU'LL HAVE OVER FOUR DOZEN HUSBANDS. BETTER GET A BULK RATE ON WEDDING CAKE.

Rhines Paradox: An Example of Overzealous DM?

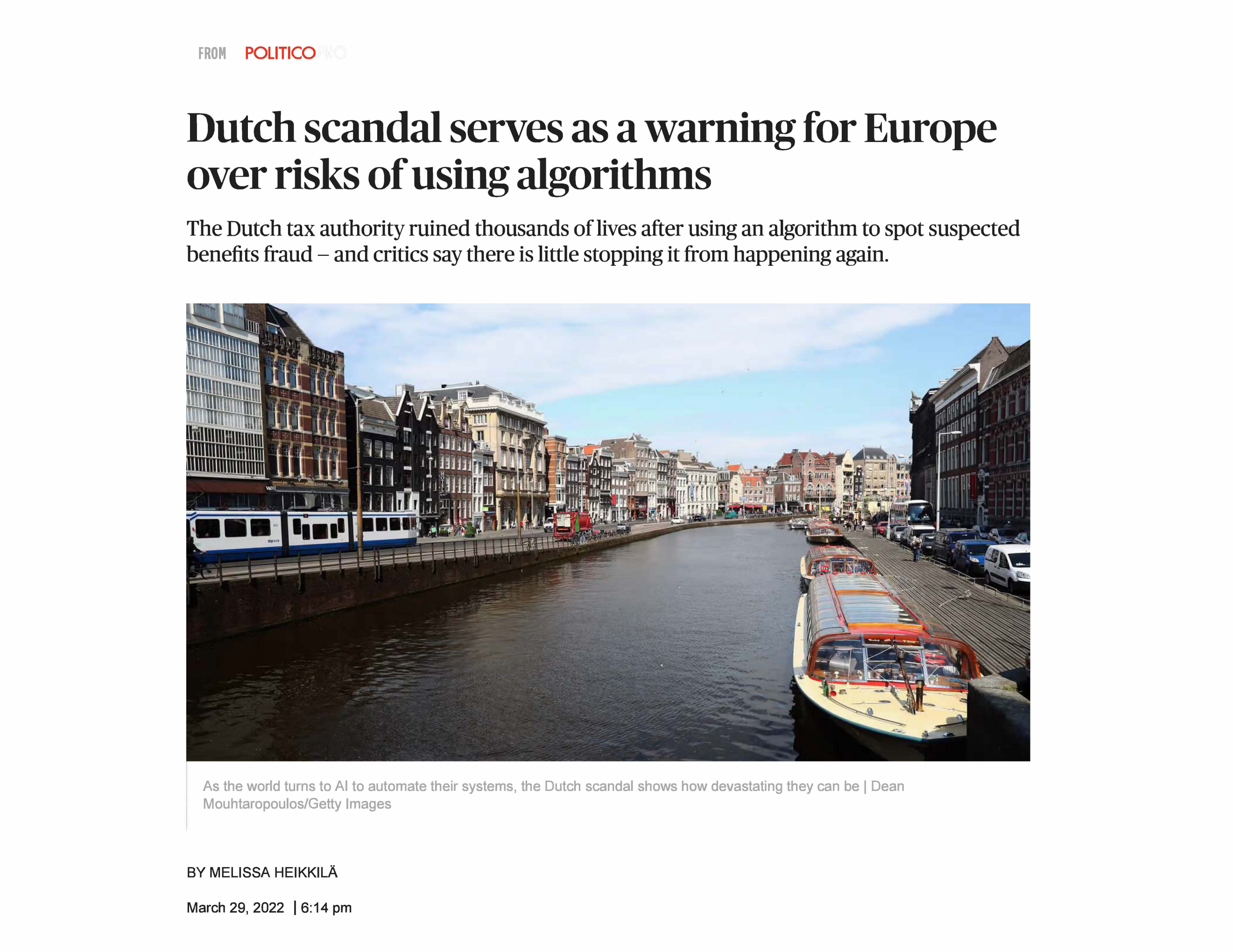
- Joseph Rhine was a parapsychologist in the 1950s who hypothesized that some people had **Extra-Sensory Perception**.
- He devised an experiment where subjects were asked to guess 10 hidden cards red or blue.
- He discovered that almost 1 in 1000 had ESP they were able to get all 10 right!
- He told these people they had ESP and called them in for another test of the same type.
- Alas, he discovered that almost all of them had lost their ESP.
- **What did he conclude?**
  - He concluded that you should not tell people they have ESP; it causes them to lose it.
  - **Why is this an incorrect conclusion?**

Ethical Issues with Data Mining


NEW YORK TIMES BESTSELLER
NATIONAL BOOK AWARD LONGLIST
WEAPONS OF MATH DESTRUCTION
HOW BIG DATA INCREASES INEQUALITY AND THREATENS DEMOCRACY
CATHY O'NEIL
A NEW YORK TIMES NOTABLE BOOK


why are black women so
why are black women so angry
why are black women so loud
why are black women so mean
why are black women so attractive
why are black women so lazy
why are black women so annoying
why are black women so confident
why are black women so sassy
why are black women so insecure
ALGORITHMS OF OPPRESSION
HOW SEARCH ENGINES REINFORCE RACISM
SAFIYA UMOJA NOBLE

In the News: AI Implementation Scandals



FROM POLITICO

Dutch scandal serves as a warning for Europe over risks of using algorithms

The Dutch tax authority ruined thousands of lives after using an algorithm to spot suspected benefits fraud – and critics say there is little stopping it from happening again.

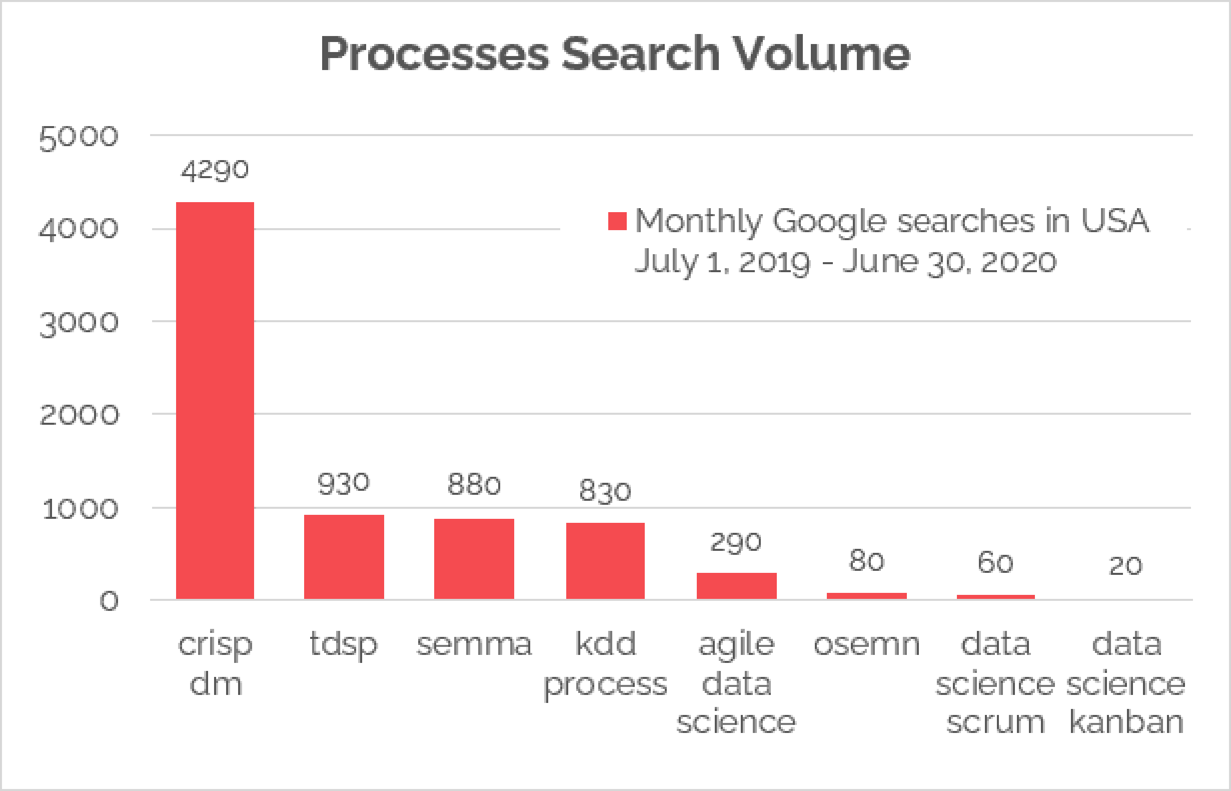
As the world turns to AI to automate their systems, the Dutch scandal shows how devastating they can be | Dean Mouhtaropoulos/Getty Images

BY MELISSA HEIKKILÄ

March 29, 2022 | 6:14 pm

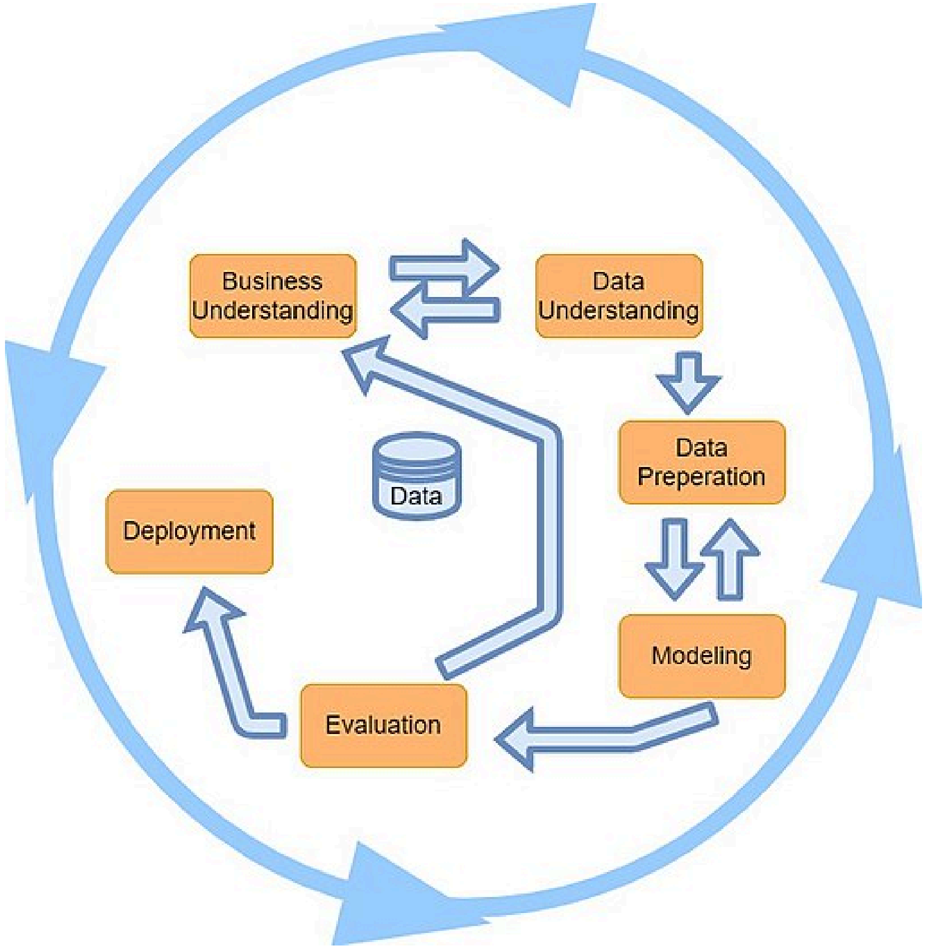
The Data Mining Process

Frameworks for Data Mining Projects



The CRISP-DM Process

- **You are expected to read the original CRISP-DM paper**
- Each step has several substeps
- **Most of the project time is typically spent in steps 1-3**



Frequent Itemsets, Market Basket Analysis and Association Rule Mining

Association Rule Discovery

Supermarket shelf management – Market-basket model:

- **Goal:** Identify items that are bought together by sufficiently many customers
- **Approach:** Process the sales data collected with barcode scanners to find dependencies among items
- **A classic rule:**
  - If someone buys diaper and milk, then he/she is likely to buy beer
  - Don’t be surprised if you find six-packs next to diapers!

Source: J. Leskovec, A. Rajaraman, J. Ullman: Mining of Massive Datasets, http://www.mmds.org

The Market-Basket Model

- A large set of **items**
  - e.g., things sold in a supermarket
- A large set of **baskets**
- Each basket is a **small subset of items**
  - e.g., the things one customer buys on one day
- Want to discover **association rules**
  - People who bought {x,y,z} tend to buy {v,w}
  - Amazon!

Input:

Basket #	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Output: Discovered Rules

{Milk} --> {Coke}

{Diaper, Milk} --> {Beer}

Source: J. Leskovec, A. Rajaraman, J. Ullman: Mining of Massive Datasets, http://www.mmds.org

Definitions: Support & Support Threshold

- **Simplest question:** Find sets of items that appear together “frequently” in baskets
- **Support for itemset I:** Number of baskets containing all items in I
  - Often expressed as a fraction of the total number of baskets
- Given a **support threshold** s, then sets of items that appear in at least s baskets are called frequent itemsets

Input:

Basket #	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Support of {Beer, Bread}: = 2

Source: J. Leskovec, A. Rajaraman, J. Ullman: Mining of Massive Datasets, http://www.mmds.org

Non-graded Activity: Frequent Itemsets

04:00

Activity   Your Solution

Items = {Milk, Coke, Pepsi, Beer, Juice}

With a support threshold of 3 baskets, find all frequent itemsets based on these 8 baskets:

- $B_1 = \{\text{Milk}, \text{Coke}, \text{Beer}\}$ $B_2 = \{\text{Milk}, \text{Pepsi}, \text{Juice}\}$
- $B_3 = \{\text{Milk}, \text{Beer}\}$ $B_4 = \{\text{Coke}, \text{Juice}\}$
- $B_5 = \{\text{Milk}, \text{Pepsi}, \text{Beer}\}$ $B_6 = \{\text{Milk}, \text{Coke}, \text{Beer}, \text{Juice}\}$
- $B_6 = \{\text{Coke}, \text{Beer}, \text{Juice}\}$ $B_8 = \{\text{Coke}, \text{Beer}\}$

Association Rules

- **Association Rules:** If-then rules about the contents of baskets
- {i$_1$, i$_2$,...,i$_k$} → j means: "if a basket contains all of $i_1, \dots, i_k$ then it is likely to contain j"
- **In practice there are many rules, want to find significant/interesting ones!**
- **Confidence** of this association rule is the probability of j given $I =$ {i$_1$,...,i$_k$}

$$conf(I \rightarrow j) = P(j \mid I) = \frac{support(I \cap j)}{support(I)}$$

- **Not all high-confidence rules are interesting**
  - The rule **X** → **milk** may have high confidence for many itemsets **X**, because **milk** is just purchased very often (independent of **X**) and the confidence will be high
- **Lift** of an association rule $I \rightarrow J$ is the ratio between its confidence and the fraction of baskets containing j: $lift(I \rightarrow j) = \frac{conf(I \rightarrow j)}{Pr(j)}$

Adaped from J. Leskovec, A. Rajaraman, J. Ullman: Mining of Massive Datasets, http://www.mmds.org

Non-Graded Activity: Confidence and Lift

04:00

Activity | Your Solution

- $B_1 = \{$Milk, Coke, Beer$\}$
- $B_2 = \{$Milk, Pepsi, Juice$\}$
- $B_3 = \{$Milk, Beer$\}$
- $B_4 = \{$Coke, Juice$\}$
- $B_5 = \{$Milk, Pepsi, Beer$\}$
- $B_6 = \{$Milk, Coke, Beer, Juice$\}$
- $B_6 = \{$Coke, Beer, Juice$\}$
- $B_8 = \{$Coke, Beer$\}$

For the association rule: $\{$Milk, Beer$\} \rightarrow$ Coke, compute both its confidence and lift.

Finding Association Rules

- **Problem: Find all association rules with support ≥ s and confidence ≥ c**
  - **Note:** Support of an association rule is the support of the set of items on the left side
- **Hard part: Finding the frequent itemsets!**
  - If $\{i_1, i_2,\ldots,i_k\} \rightarrow j$ has high support and confidence, then:
  - both $\{i_1, i_2,\ldots ,i_k\}$ and both $\{i_1, i_2,\ldots,i_k, j\}$ will be "frequent"

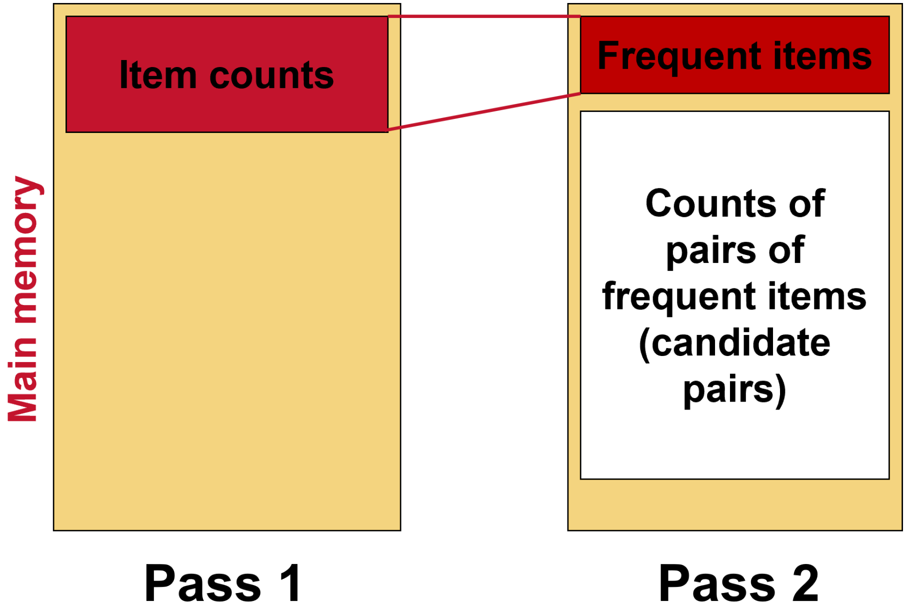
Naïve Approach to Counting Frequent Itemsets

- Naïve approach to finding frequent pairs
- **Read file once, counting in main memorythe occurrences of each pair:**
    - From each basket of n items, generate its $\frac{n(n-1)}{2}$ pairs by two nested loops
- Fails if (#items)^2 exceeds main memory
    - Remember: #items can be 100K (Wal-Mart) or 10B (Web pages)
    - Suppose 10^5 items, counts are 4-byte integers
    - Number of pairs of items: $\frac{10^5(10^5-1)}{2} = 5 * 10^9$
    - Therefore, $2 * 10^{10}$ (20 gigabytes) of memory needed

Source: J. Leskovec, A. Rajaraman, J. Ullman: Mining of Massive Datasets, http://www.mmds.org

A-Priori Algorithm

- **Pass 1:** Read baskets and count in main memory the occurrences of each **individual item**
    - Requires only memory proportional to #items
- **Items that appear $\geq s$ times are the frequent items**
- **Pass 2:** Read baskets again and count in main memory **only those pairs where both elements are frequent (from Pass 1)**



Source: J. Leskovec, A. Rajaraman, J. Ullman: Mining of Massive Datasets, http://www.mmds.org

Using R to Mine Association Rules

In class, we will go through this R code, explaining: (a) what each function is doing, and (b) the outputs from each step.

```r
if(require(pacman)==FALSE) install.packages('pacman')
pacman::p_load(arules, tidyverse)

data('Groceries') # note its class

summary(Groceries)

itemFrequency(Groceries) # returns frequency in alphabetic order
itemFrequency(Groceries) %>% sort(decreasing = T)

itemFrequencyPlot(Groceries, support = 0.1)
itemFrequencyPlot(Groceries, topN = 20)

# mine association rules with a certain min support and confidence
grocery_rules = apriori(
  Groceries, parameter = list(
    support = 0.01, confidence = 0.5, minlen = 2, maxlen = 5)  )

summary(grocery_rules)
inspect(grocery_rules)

sort(grocery_rules, by ='lift', decreasing = T)[1:3] %>% inspect()
```

Recap

Summary of Main Points

- Describe the goals & functions of data mining
- Understand the statistical limits on data mining
- Describe the data mining process
- What is “frequent itemsets” & the application of this concept
- Explain how and why “association rules” are constructed
- Use R to populate both concepts