SPL Data and Frequency Pattern Mapping

George Legrady, legrady@mat.ucsb.edu
Ambika Yadav, ambikayadav@gmail.com

Data Mining & Knowledge Discovery in Databases

- **Data Mining**: Computational process for discovering patterns in large datasets (*Algorithms for extracting patterns*)
- **KDD**: the process of discovering useful knowledge from a collection of data (*high-level emphasis on knowledge discovery*)
- **KDD**: An iterative process [p42, Fayyad]

Data Mining Tasks

The most time intensive effort in data mining is **understanding your data** before any analysis or visualization can take place

The next step is *Discovery*: Detecting something new or relevant. Some common tasks:

- **Anomaly detection** : Unusual records (Outlier, deviation detection) that may be interesting
- **Classification:** Task of identifying which set of categories a data belongs to (Spam OR not spam)
- **Clustering**: The task of discovering groups and structures in the data that are in some way or another "similar"
- **Regression analysis**: A statistical process for estimating the relationships among variables
- **Association rule learning:** (Dependency modeling) Searches for relationships between variables. Using association rule learning, a supermarket can determine which products are frequently bought together

Data in XML

```
<transaction>
        <itemNumber>1946514</itemNumber>
        <bibNumber>2068458</bibNumber>
        <ckodate>2006-10-03</ckodate>
        <ckotime>13:44:00</ckotime>
        <ckidate>2007-02-19</ckidate>
        <ckitime>13:42:00</ckitime>
        <collcode>nanf</collcode>
        <itemtype>acbk</itemtype>
        <barcode>0010041282327</barcode>
        <title>Calm energy how people regulate mood with food...</title>
        <callNumber>152.4 T338C 2001</callNumber>
        <deweyClass>152.4</deweyClass>
        <subjects>
                <subject>Nutrition Psychological aspects</subject>
                <subject>Mood Psychology</subject>
                <subject>Mental health Nutritional aspects</subject>
                <subject>Exercise Psychological aspects</subject>
        </subjects>
</transaction>
```

The Seattle Public Library Data & its Metadata

The metadata for each item in the collection is *multivariate*

Ordinal (*In a numeric sequence*)

- **ItemNumber**: Assigned when object enters system
- **Dewey Classification** (Dewey numeric)

Interval Scale (*Time-Stamp*)

- **Check-out/check-in hour, day, month, year**

Categorical (*Not necessarily numerically orderable*)

- **BibNumber**: Each title has a specific number, copies of titles all have same number
- **Barcode**: Each item has a unique number on RFID sticker
- **CallNumber**: by which to locate items on shelves - Ordinal if Dewey, otherwise categorical

Semantic (*Text-based*)

- **Title:** Each Item has a title
- **ItemType**: books, cds, dvds, music sheets, etc.
- **Collection Code**: The physical home of the item and other info
- **Subjects**: Keywords (arbitrary labeling)

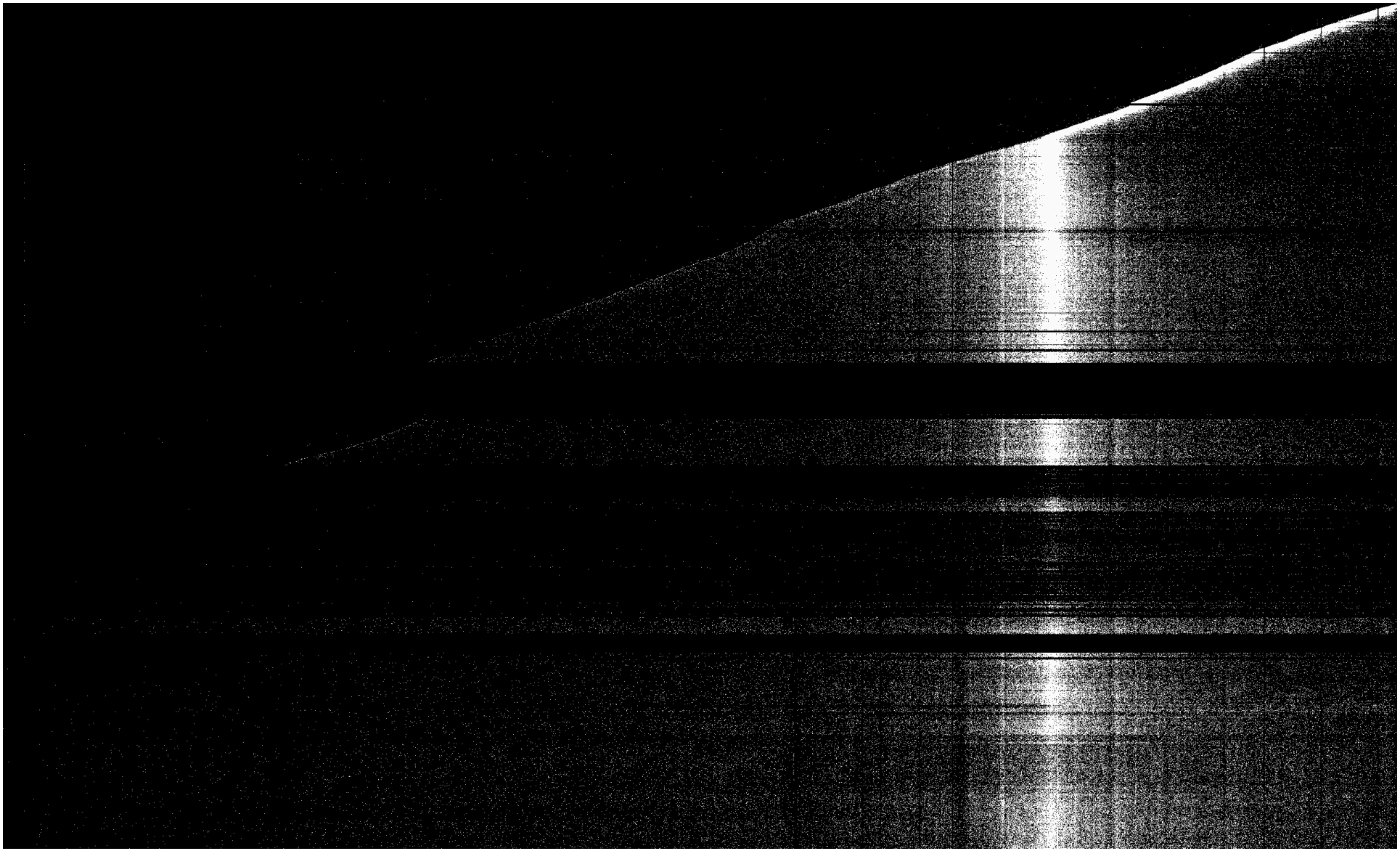
MySQL: Select itemNumber, cout, collcode, itemtype, barcode, title, callNumber, deweyClass, subj from inraw where year(cout) = 2007 and month(cout) >= 1 and month(cout) <= 4 limit 50;

itemNumber	cout	collcode	itemtype	barcode	title	callNumber	deweyClass	subj
2124890	2007-01-02 09:20:00	nacd	accd	0010053294418	tigers have spoken	CD 782.421642 C2665T	782.421642	Rock music 2001 2010^Country music 2001 2010
1963406	2007-01-02 09:12:00	cafic	acbk	0010046075908	Heart on the line	FIC ARNOLD	NULL	Love stories^Women television producers and directors New York State N...
1348064	2007-01-02 09:03:00	nalpfic	acbk	0010047150288	Too hot to handle	FIC LOWELL	NULL	Love stories^Large type books^Ranch life Fiction
264450	2007-01-02 09:13:00	canf	acbk	0010027180271	Governing public schools new times new requirements	371.2 DANZBER 1992	371.2	School boards Rating of
436074	2007-01-02 09:13:00	cs9	acbk	0010002326972	great deception the inside story of how the Kremlin took over Cuba	972.91 M741G	972.91	Cuba History 1959^Communism Cuba^Soviet Union Foreign relations Cu...
660741	2007-01-02 09:13:00	canf	acbk	0010041599258	Logic or The right use of reason in the inquiry after truth with a va...	160 L793W 1996	160	Locke John 1632 1704^Logic Early works to 1800^Conduct of life Early w...
2225714	2007-01-02 09:12:00	cafic	acbk	0010050230746	Something about Emmaline	FIC BOYLE2005	NULL	Love stories^Historical fiction^Regency fiction^London England Fiction
860548	2007-01-02 09:14:00	canf	acbk	0010048238389	Understanding installation art from Duchamp to Holzer	709.04 R7277U 2003	709.04	Installations Art^Art Modern 20th century
15316	2007-01-02 09:12:00	cs9	acbk	0000103502175	Frank B Kellogg a biography	B K292 B	NULL	Kellogg Frank B Frank Billings 1856 1937^Statesmen United States Biography
418329	2007-01-02 09:13:00	cs9	acbk	0010002324977	German raider Atlantis	940.953 F8517G	940.953	World War 1939 1945 Naval operations German^Atlantis Ship
745716	2007-01-02 09:13:00	cafic	acbk	0010033394031	Zabelle	FIC KRICORI1998	NULL	Armenian American women Massachusetts Boston Fiction
1529347	2007-01-02 09:14:00	canf	acbk	0000102981248	Reservation narrow gauge Omak Creek railroad Bow Arrow Short Li...	385.52097 L587R	385.52097	Biles Coleman Lumber Company History^Logging railroads Washington St...
1558007	2007-01-02 09:13:00	canf	acbk	0010004731641	place no one knew Glen Canyon on the Colorado	779 P833P2a	779	Glen Canyon Utah and Ariz^Glen Canyon Utah and Ariz Pictorial works
1681487	2007-01-02 09:12:00	canf	acbk	0010002691342	Buddhism in translations passages selected from the Buddhist sacr...	294 W252B	294	Buddhism Sacred books
1688675	2007-01-02 09:14:00	canf	acbk	0010048100100	death in Washington Walter G Krivitsky and the Stalin terror	327.1247 K4592D 2003	327.1247	Krivitsky W G Walter G 1899 1941^Espionage American^Espionage Soviet...
2424499	2007-01-02 09:13:00	canf	acbk	0010054269096	101 diseases you dont want to get	614.5 P8715o 2005	614.5	Epidemiology Popular works^Diseases Popular works
27939	2007-01-02 09:09:00	nafic	acbk	0010040162363	Twilight in Texas	FIC THOMAS	NULL	Love stories^Texas Fiction^Texas Rangers Fiction
1534336	2007-01-02 09:00:00	nanf	acbk	0010041237933	Boundless healing meditation exercises to enlighten the mind and...	294.34435 T387B 2000	294.34435	Mind and body^Meditation Buddhism^Healing Religious aspects Buddhism
2205408	2007-01-02 09:14:00	canf	acbk	0010050859288	fragrance of faith the enlightened heart of Islam	297 R1294F 2004	297	Islam
2631813	2007-01-02 09:14:00	nanf	acbk	0010053789672	How to reduce workplace conflict and stress how leaders and their...	658.1053 M327H 2005	658.1053	Teams in the workplace^Interpersonal conflict^Mediation^Negotiation^O...
702785	2007-01-02 09:13:00	canf	acbk	0010034758416	home of the blizzard a true story of Antarctic survival	919.8904 MAWSON 1998	919.8904	Antarctica Discovery and exploration^Australasian Antarctic Expedition 1...
1168659	2007-01-02 08:58:00	nab	acbk	0010045761730	Lives of mothers daughters growing up with Alice Munro	B M9265M 2001	NULL	Novelists Canadian 20th century Family relationships^Novelists Canadian...
2498965	2007-01-02 09:14:00	canf	acbk	0010054107684	Foxes in the henhouse how the Republicans stole the South and th...	324.70973 R29996J 2006	324.70973	Democratic Party U S^Republican Party U S 1854^Politics Practical United...
1076021	2007-01-02 09:13:00	canf	acbk	0010045523353	False intimacy understanding the struggle of sexual addiction	241.66 Sch196F 1997	241.66	Sex addicts Rehabilitation^Intimacy Psychology Religious aspects Christia...
1703576	2007-01-02 09:12:00	canf	acbk	0010025511717	Lakota recollections of the Custer fight new sources of Indian milit...	973.82 LAKOTA 1991	973.82	Dakota Indians Wars 1876^Little Bighorn Battle of the Mont 1876 Person...
2033121	2007-01-02 09:13:00	cafic	acbk	0010020070768	Furors die a novel	FIC HOFFMAN	NULL	NULL
2633117	2007-01-02 09:06:00	nanf	acbk	0010042818186	devils dictionary of business monkey business high finance and lo...	330.0207 V895D 2005	330.0207	Finance Dictionaries^Business Dictionaries
295289	2007-01-02 09:13:00	canf	acbk	0010041330571	Running with the Buffaloes a season inside with Mark Wetmore Ad...	796.42809 C7193L 2000	796.42809	Colorado Buffaloes Cross country team^Cross country running Colorado...
2684484	2007-01-02 09:13:00	canew	acbk	0010054076608	opened grave Sherlock Holmes investigates his ultimate case	FIC JAMES2006	NULL	Mystery fiction^Holmes Sherlock Fictitious character Fiction^Missing pers...
756568	2007-01-02 09:14:00	canf	acbk	0010028375219	Transforming vision writers on art	810.80357 TRANSFO 1994	810.80357	Art^American poetry 20th century^Art Poetry
1437750	2007-01-02 09:13:00	cs9	acbk	0000102286903	De Shazer the Doolittle raider who turned missionary a true and th...	B D459W	NULL	Missions Japan^De Shazer Jacob 1912
67201	2007-01-02 09:45:00	nchol	jcbk	0010045871109	Jothams journey a storybook for Advent	J YTREEID	NULL	Family Prayer books and devotions English^Advent Prayer books and dev...
1336766	2007-01-02 09:44:00	cs9o	acbk	0010001512887	Burri	B B942B	NULL	Burri Alberto 1915
2582486	2007-01-02 09:44:00	ncnew	jcbk	0010050556819	Lady in the water a bedtime story	E SHYAMAL	NULL	Imaginary creatures Fiction
1978356	2007-01-02 09:45:00	cs6ro	arbk	0010019836690	Bullard Arms	338.76834 J241B	338.76834	Bullard James Herbert 1842 1914^Bullard Repeating Arms Company Hist...
1404338	2007-01-02 09:44:00	canf	acbk	0010038257472	Perfect bones a six point plan to promote healthy bones	616.716 L5785P 2000	616.716	Osteoporosis Popular works^Women Nutrition
2347671	2007-01-02 09:44:00	nanf	acbk	0010050592780	Satisfaction the science of finding true fulfillment	155.9 B4581S 2005	155.9	Satisfaction
468341	2007-01-02 09:41:00	nafic	acbk	0010022410574	Fly fishing tales literary bait by angling authors	FIC FLY FIS1994	NULL	Fishing stories American^Fly fishing Fiction
1558499	2007-01-02 10:02:00	nanf	acbk	0010037361580	Chariots of the gods unsolved mysteries of the past	001.94 DANIKEN 1999	001.94	Archaeology^Life on other planets^Civilization Ancient Extraterrestrial in...
2064174	2007-01-02 09:41:00	canf	acbk	0010036046778	way of agape	241.4 MISSLER 1999	241.4	Agape^Love Religious aspects Christianity
18430	2007-01-02 09:41:00	canf	acbk	0010026014141	Future of medicine toward a science of prevention based on ancie...	613 DUGLISS 1993	613	Medical care United States^Medicine Preventive United States
614085	2007-01-02 09:41:00	canf	acbk	0010045701462	Tanker operations a handbook for the person in charge PIC	623.88245 H8627T 2001	623.88245	Tankers Handbooks manuals etc
1591175	2007-01-02 09:39:00	caesl	bcbk	0010034398809	Angliiskii iazyk prosto o slozhnom prakticheskii kurs	RUSSIAN 428.24917 L576A	428.24917	English language Textbooks for foreign speakers Russian
2670845	2007-01-02 08:48:00	canf	acbk	0010051680022	Gorgeous disaster the tragic story of Debra LaFave	364.153 L131L 2006	364.153	Sexual abuse victims United States Case studies^Female sex offenders U...
1064531	2007-01-02 10:16:00	nacd	accd	0010040958539	Spirituals in concert	CD784.73 B322S	784.73	Spirituals Songs^Songs High voice with instrumental ensemble^Songs Hi...
965750	2007-01-02 09:38:00	capf	acbk	0010025431536	Kathy and Mo show parallel lives	812.54 GAFFNEY 1992	812.54	Women Drama
1970799	2007-01-02 09:39:00	nafic	acbk	0010046761242	revelation	FIC LITTLE	NULL	Horror fiction
1006825	2007-01-02 10:17:00	cacd	accd	0010046200415	peaceful Christmas	CD 782.21723 P3133	782.21723	Christmas music^New Age music
1831232	2007-01-02 10:17:00	canf	acbk	0010048571573	Tauntons family home idea book	728.37 St546T 2003	728.37	Architecture Domestic United States^Interior architecture United States^R...

Typical daily Dewey activity
20 most active

hour	789	613	784	917	746	973	796	398	616	914	791	306	332	305	158	792	811	658	635	895
0	0	0	0	8	0	4	12	1	0	0	0	1	0	0	1	0	5	0	0	4
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0
7	1	2	0	5	0	1	1	0	0	1	0	0	0	2	1	0	1	0	0	2
8	15	58	57	109	38	90	39	73	25	50	41	34	20	31	13	33	33	14	24	42
9	27	71	48	134	40	87	62	73	41	70	78	36	35	42	12	32	52	14	28	45
10	1192	942	812	777	619	632	582	515	465	472	440	383	466	394	429	362	420	370	402	396
11	1560	1149	1125	969	880	1006	786	966	702	701	567	610	453	556	542	526	628	488	532	560
12	2185	1683	1878	1177	1025	1108	1156	889	881	935	782	706	779	700	715	625	788	680	647	960
13	3266	2334	2085	1769	1717	1615	1462	1324	1181	1275	972	1117	1175	1025	1025	997	968	967	986	987
14	3336	2698	2227	1811	1890	1744	1685	1487	1421	1271	1144	1146	1059	1044	1062	1041	990	962	1105	1140
15	3995	2810	2426	1870	1850	1813	1703	1417	1444	1360	1388	1142	1187	1172	1073	1143	1139	1074	1139	1044
16	3959	2855	2599	2068	1857	1867	1904	1798	1520	1449	1411	1390	1250	1306	1321	1265	1101	1126	1110	1119
17	4322	2870	2547	2203	1992	1817	1904	1374	1438	1416	1322	1402	1366	1317	1346	1359	1152	1248	1045	798
18	2398	1574	1381	1012	1024	1034	1059	630	756	615	719	682	678	680	677	685	604	562	532	280
19	2233	1557	1284	1050	845	942	1027	643	612	524	595	713	706	594	623	645	562	651	497	341
20	48	18	17	35	9	15	25	18	9	6	15	4	3	3	4	6	11	17	22	4
21	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	0	2	2	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0
23	6	1	2	5	1	0	6	0	0	0	4	2	2	0	2	7	1	1	0	0

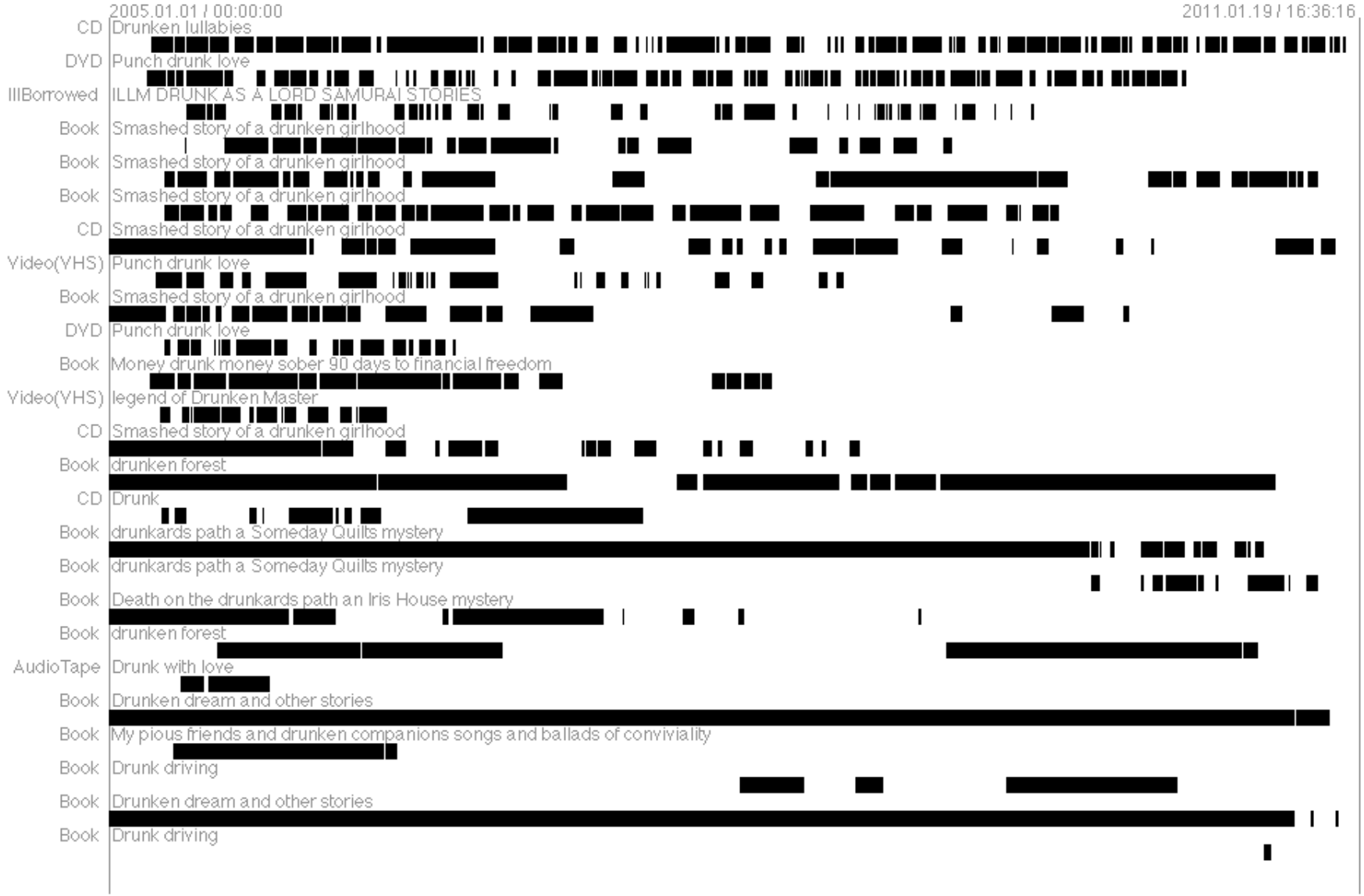
ItemNumber (*scalar numeric code maps acquisition history*)



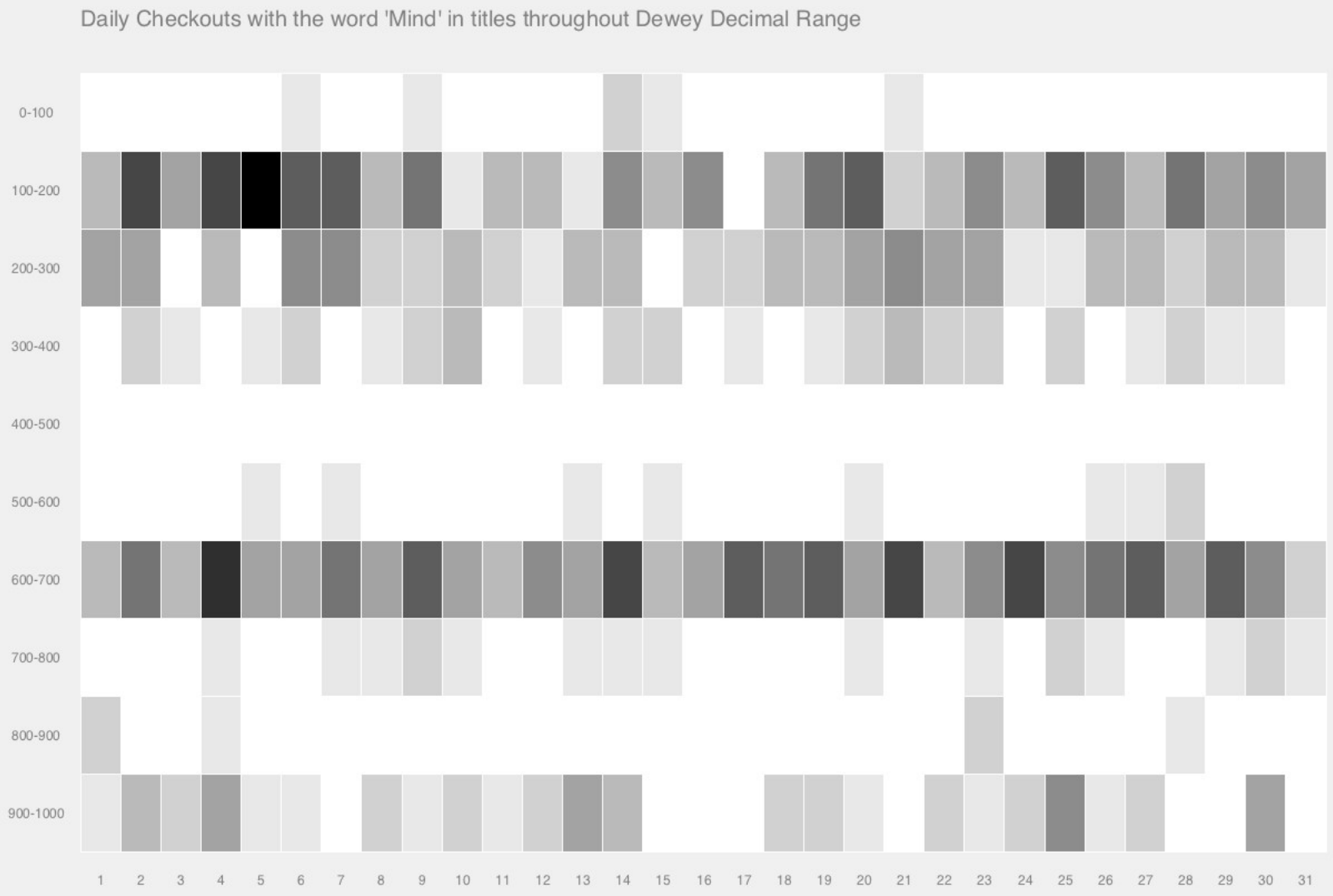
MySQL

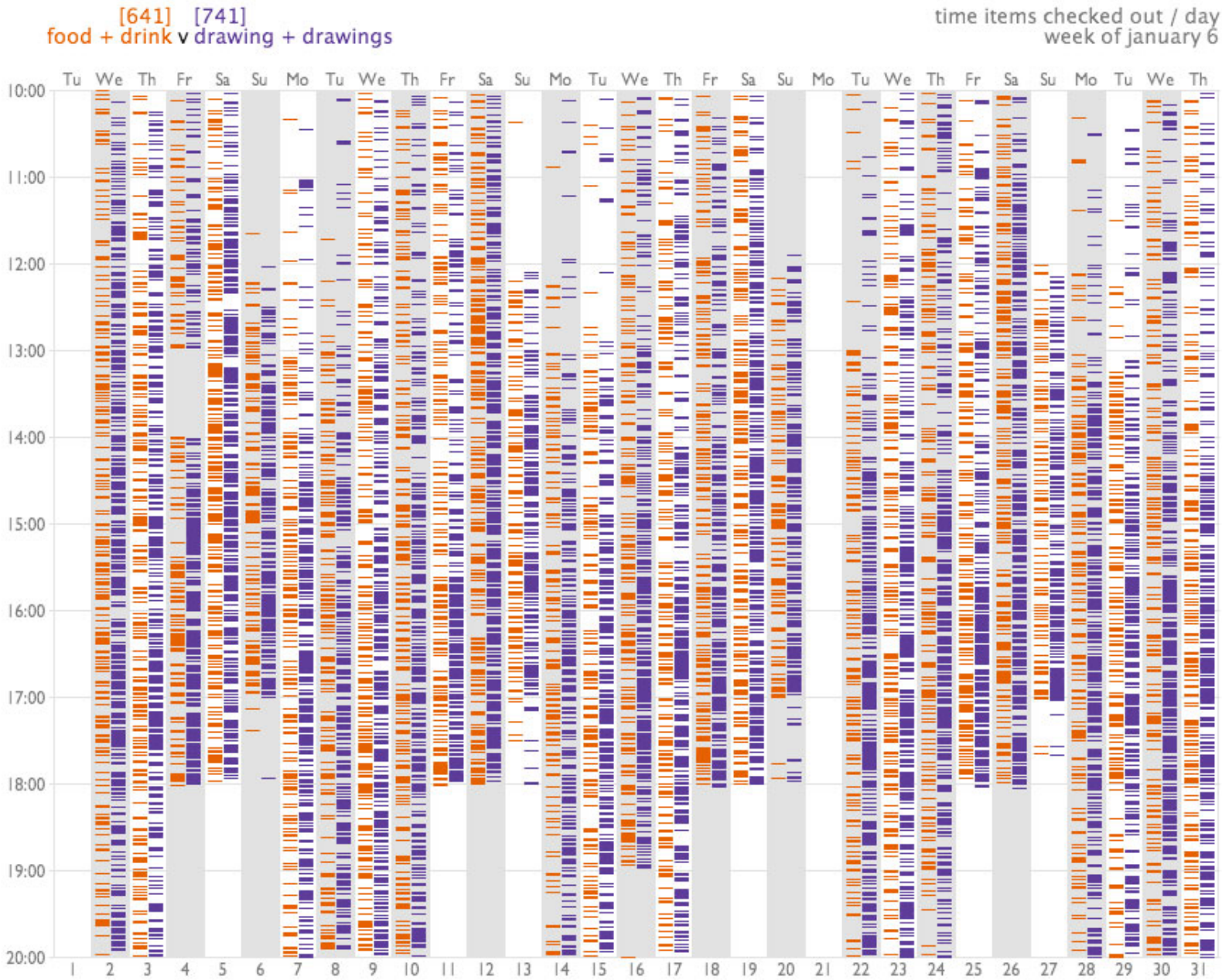
- **MySQL:** Open-source relational database (Structured Query Language)
- **http://dev.mysql.com/doc/refman/5.6/en/index.html**
- **Industry standard**
- MySQL exercises to develop skills in retrieving meaningful information

title like '%drunk%' group by collcode;

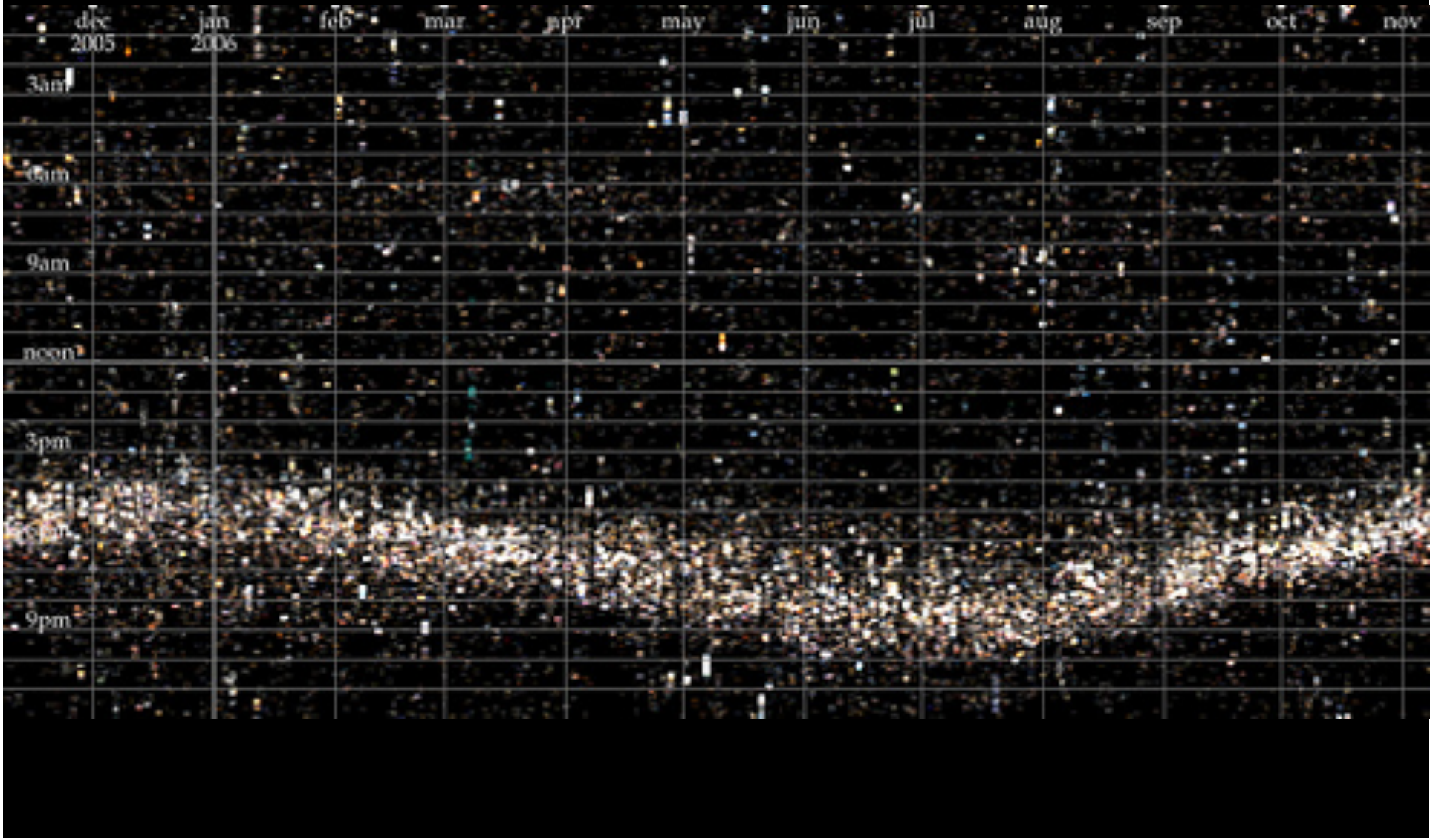


2D Visualization / Mapping

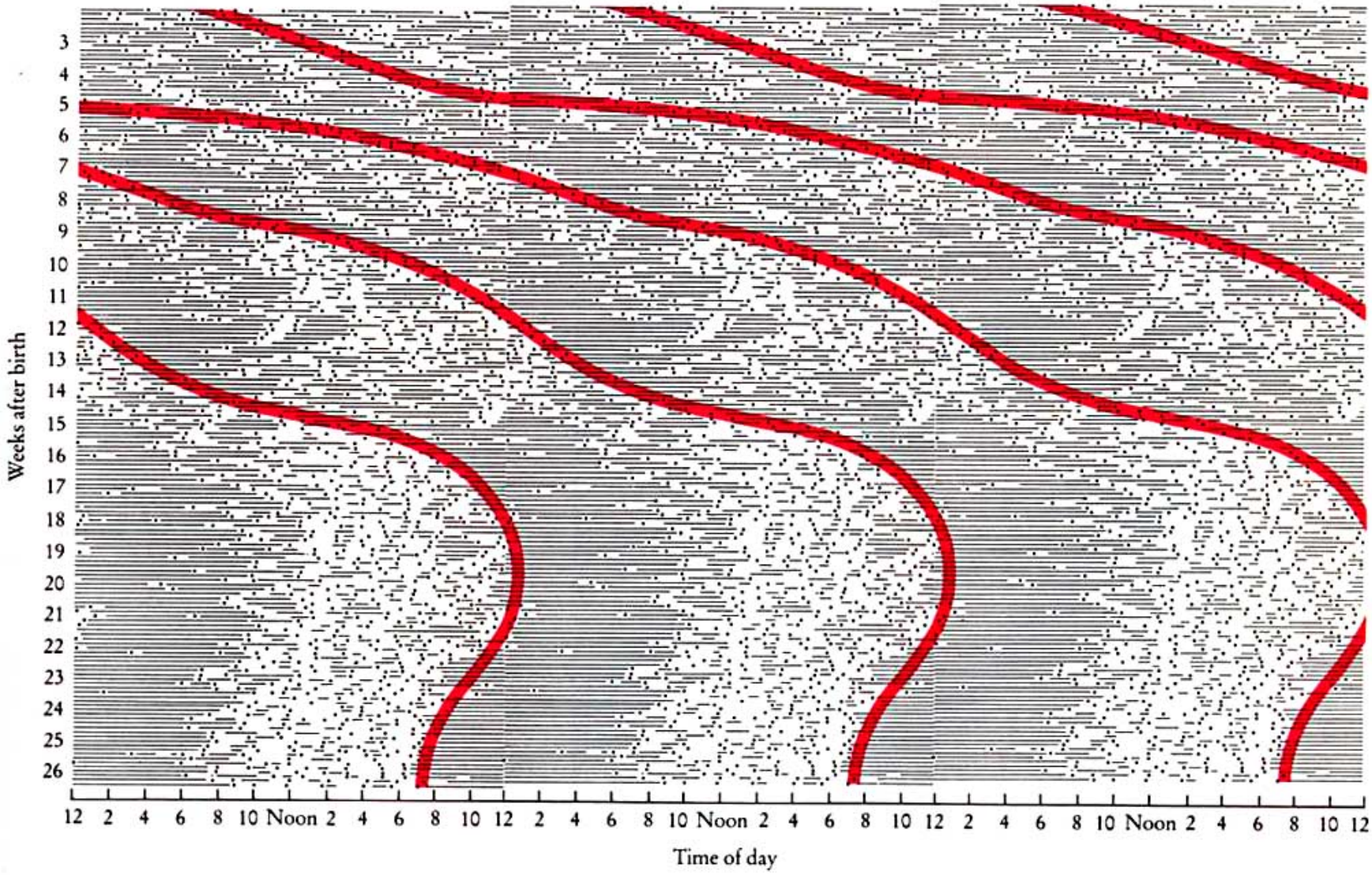



[641] [741]
food + drink v drawing + drawings
time items checked out / day
week of january 6
Tu We Th Fr Sa Su Mo Tu We Th Fr Sa Su Mo Tu We Th Fr Sa Su Mo Tu We Th Fr Sa Su Mo Tu We Th
10:00
11:00
12:00
13:00
14:00
15:00
16:00
17:00
18:00
19:00
20:00
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31

Linear Time Graphs

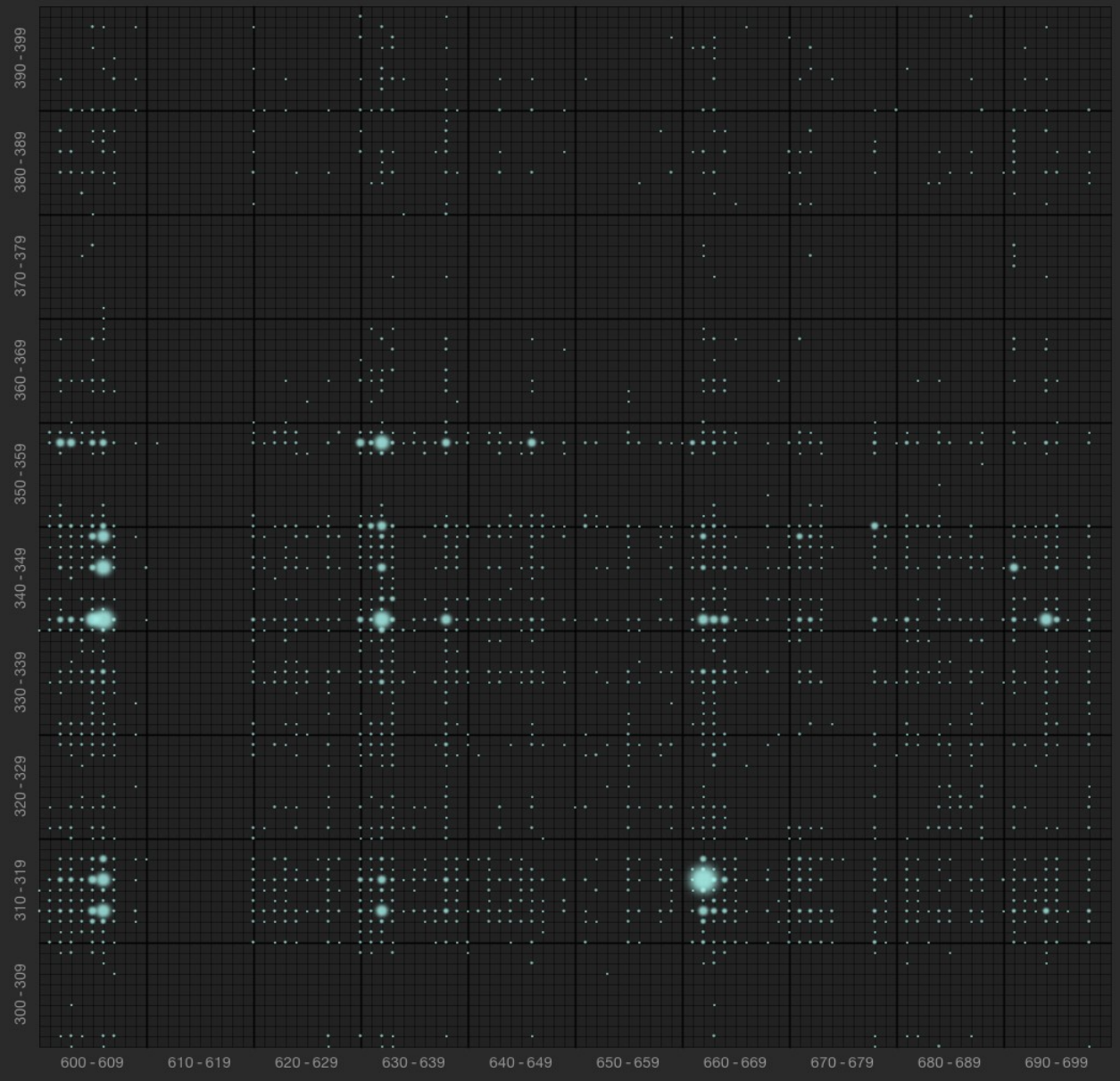


Bio-Rhythm: Frequency Map reveals pattern



Frequency Pattern Mining & Knowledge Discovery

- **FPM**: Process of identifying, interesting patterns to create knowledge about the data
- **KDD**: Find relationships among items in a database
- **Sequential pattern mining**, a related problem where an order is present in the transactions
- **FP Tree Algorithm** − what items are checked-out together



Frequently Borrowed Together

use the radio buttons to zoom

● ●	000 - 099	Computer science, information & general works
● ●	100 - 199	Philosophy and psychology
● ●	200 - 299	Religion
◉ ●	300 - 399	Social sciences
● ●	400 - 499	Language
● ●	500 - 599	Science
● ◉	600 - 699	Technology
● ●	700 - 799	Arts & recreation
● ●	800 - 899	Literature
● ●	900 - 999	History & geography

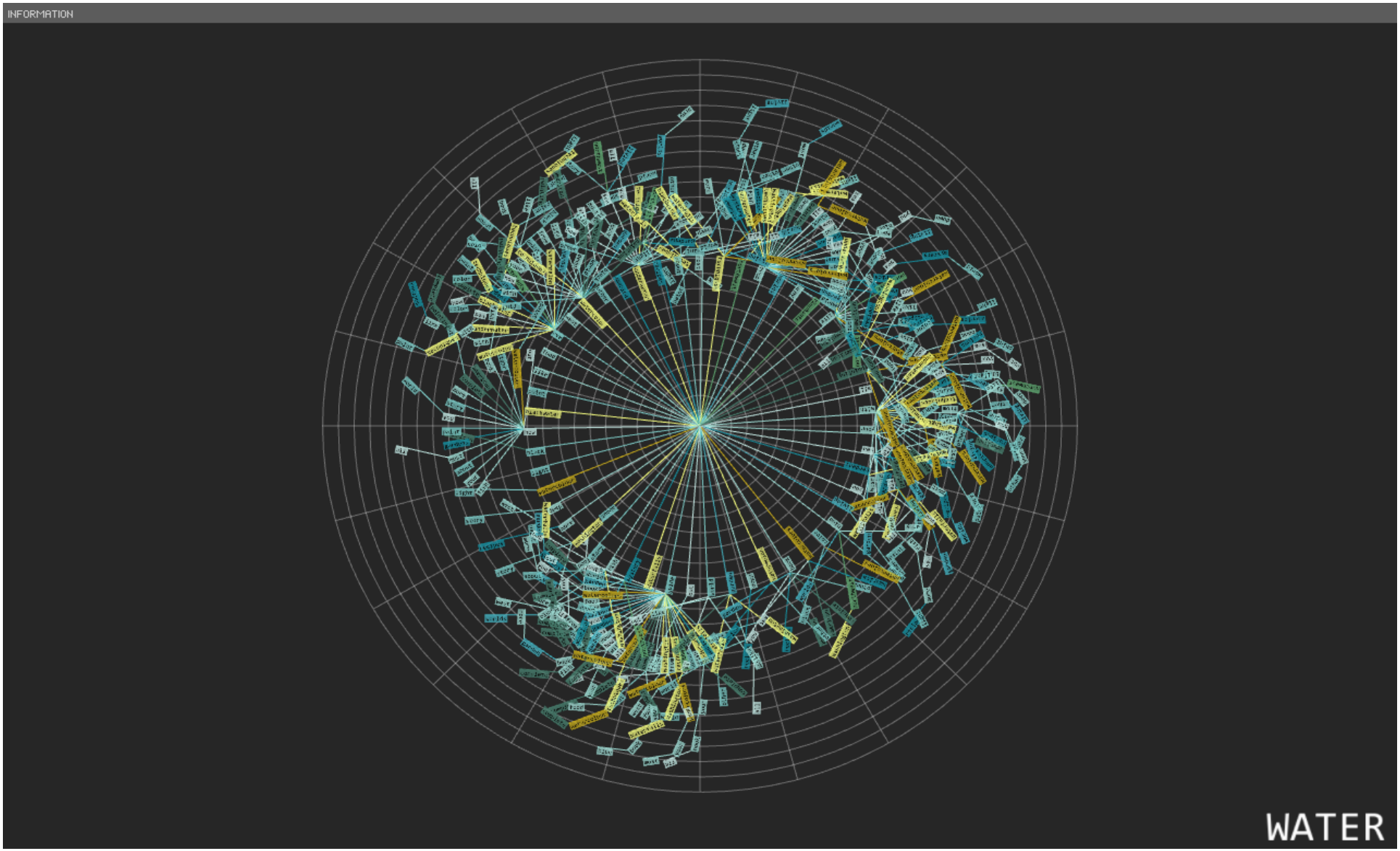
300 - 309	Social sciences, sociology & anthropology
310 - 319	Statistics
320 - 329	Political science
330 - 339	Economics
340 - 349	Law
350 - 359	Public administration & military science
360 - 369	Social problems & social services
370 - 379	Education
380 - 389	Commerce, communications, & transportation
390 - 399	Customs, etiquette, & folklore

600 - 609	Technology
610 - 619	Medicine & health
620 - 629	Engineering
630 - 639	Agriculture
640 - 649	Home & family management
650 - 659	Management & public relations
660 - 669	Chemical engineering
670 - 679	Manufacturing
680 - 689	Manufacture for specific uses
690 - 699	Construction of buildings

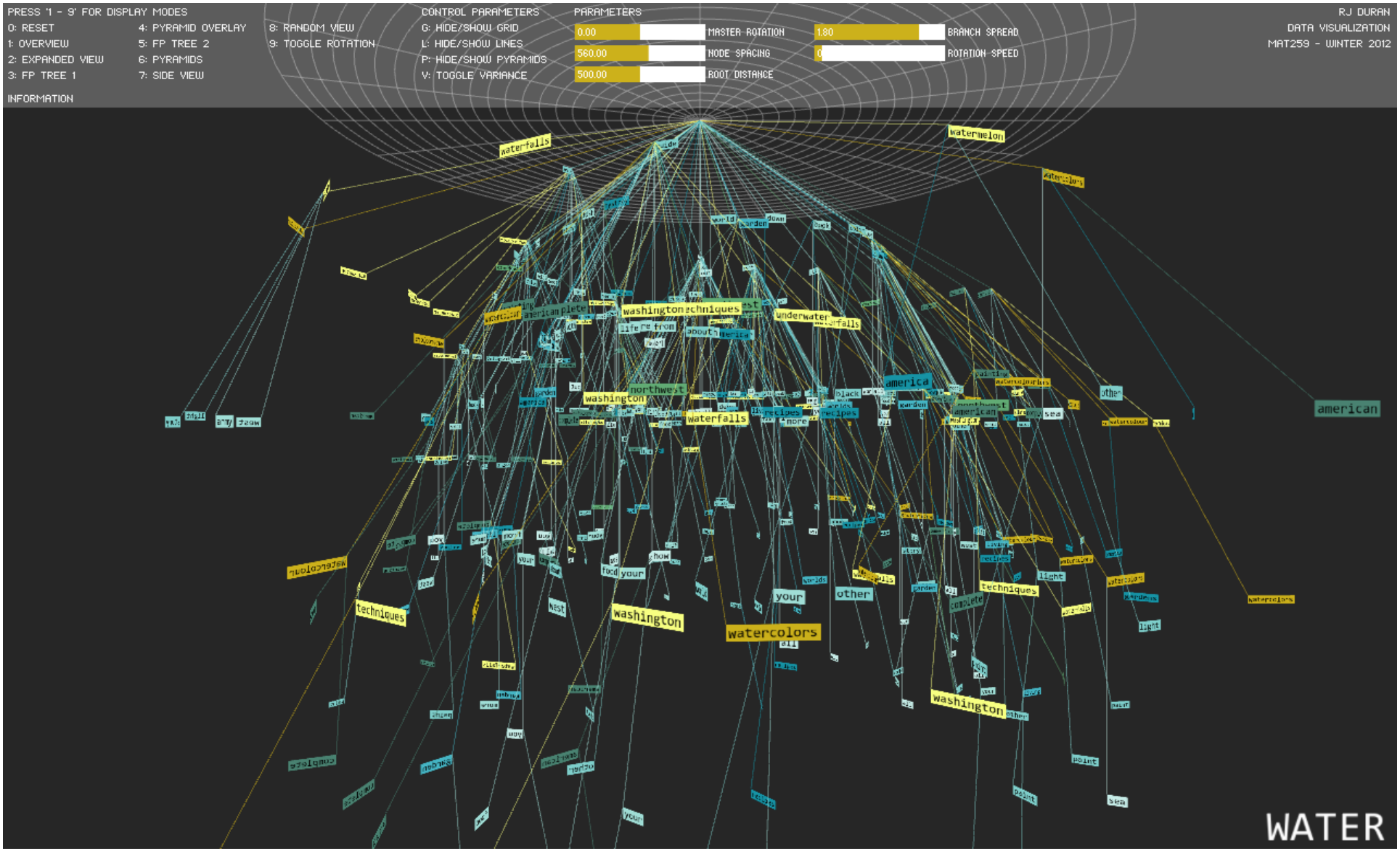
Associative Relationships

- Association rule learning (Dependency modeling)
- Searches for relationships between variables. For example, a supermarket might gather data on customer purchasing habits, which products are frequently bought together.
- FP-Growth Algorithm: Frequency-Pattern uses recursive-built tree structure to show paired occurrences

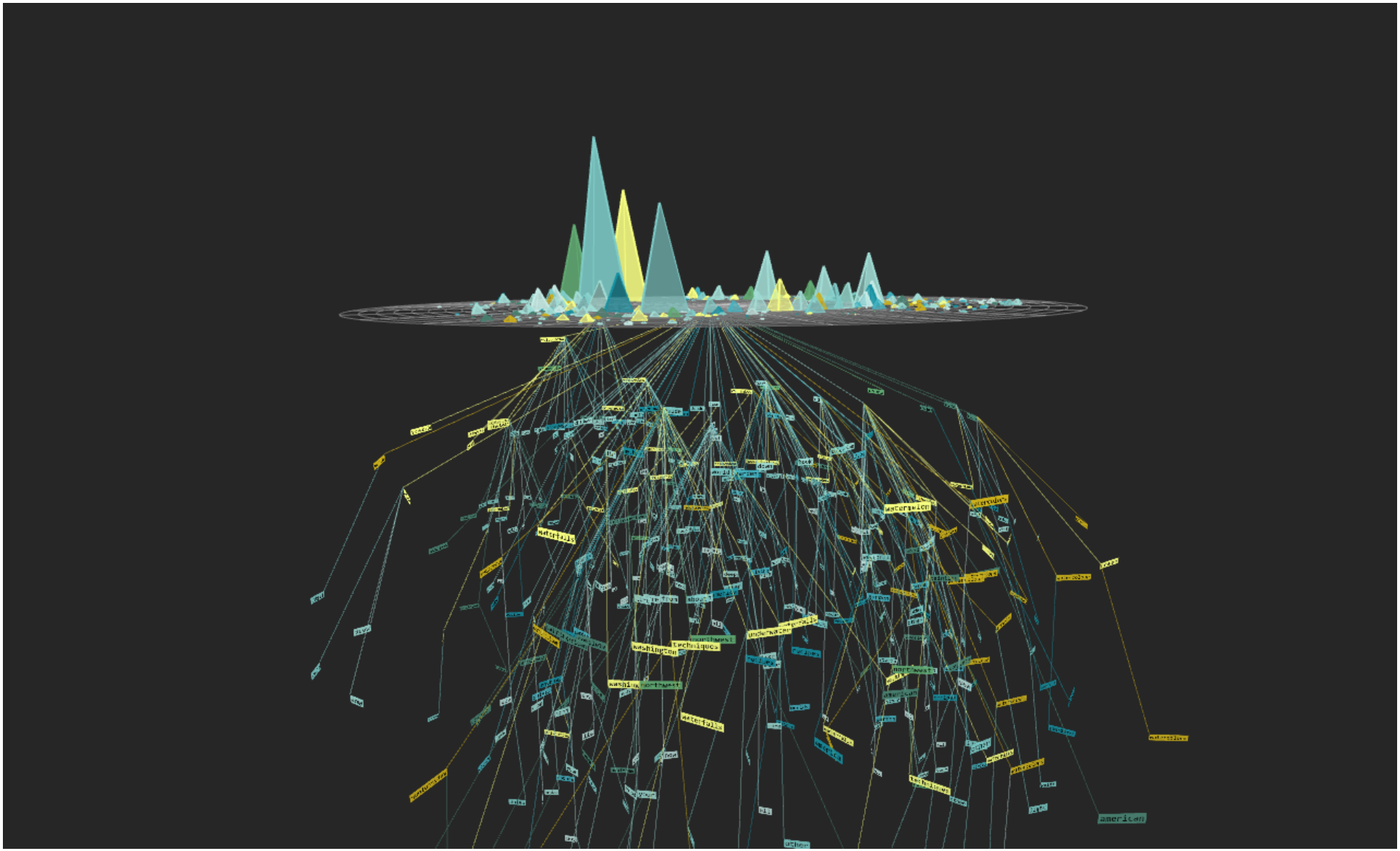
Associative (FPTree Algorithm)



Associative (FPTree Algorithm)



Associative (FPTree Algorithm)

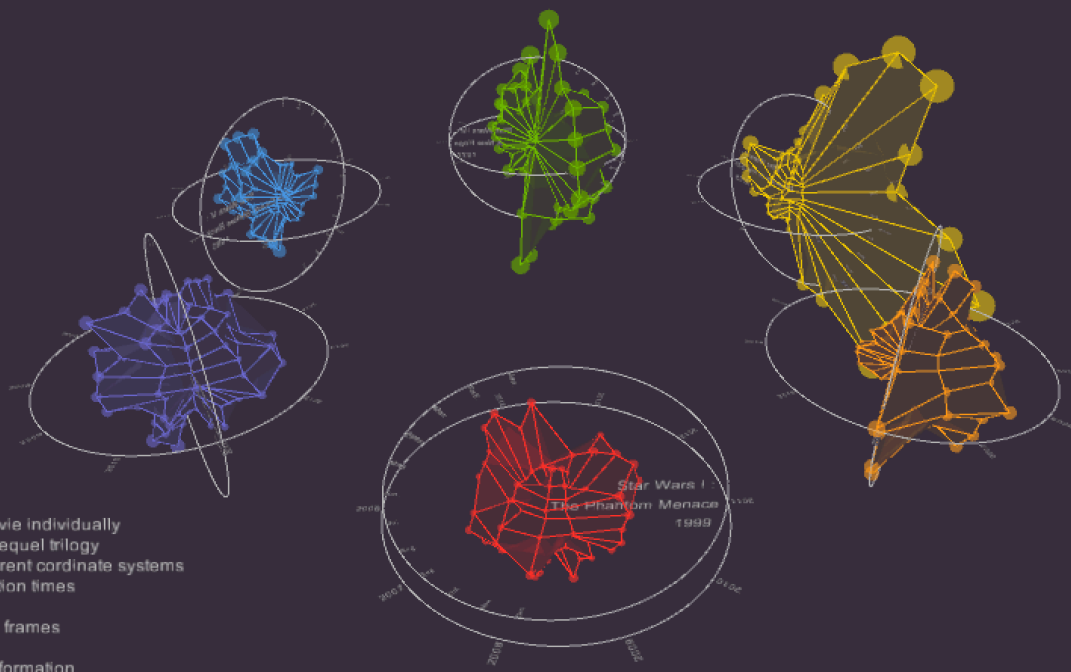


STAR WARS NEBULA

Using Data of Checkout Times and Duration Times of the Former 6 in This Movie Series in Seattle Public Library

Year	Month
2006	January
2007	February
2008	March
2009	April
2010	May
2011	June
2012	July
2013	August
2014	September
2015	October
	November
	December

Reset
Years
Months
All



Junxiang Yao
MAT259 PROJ 2
3D Interaction & Change Over Time

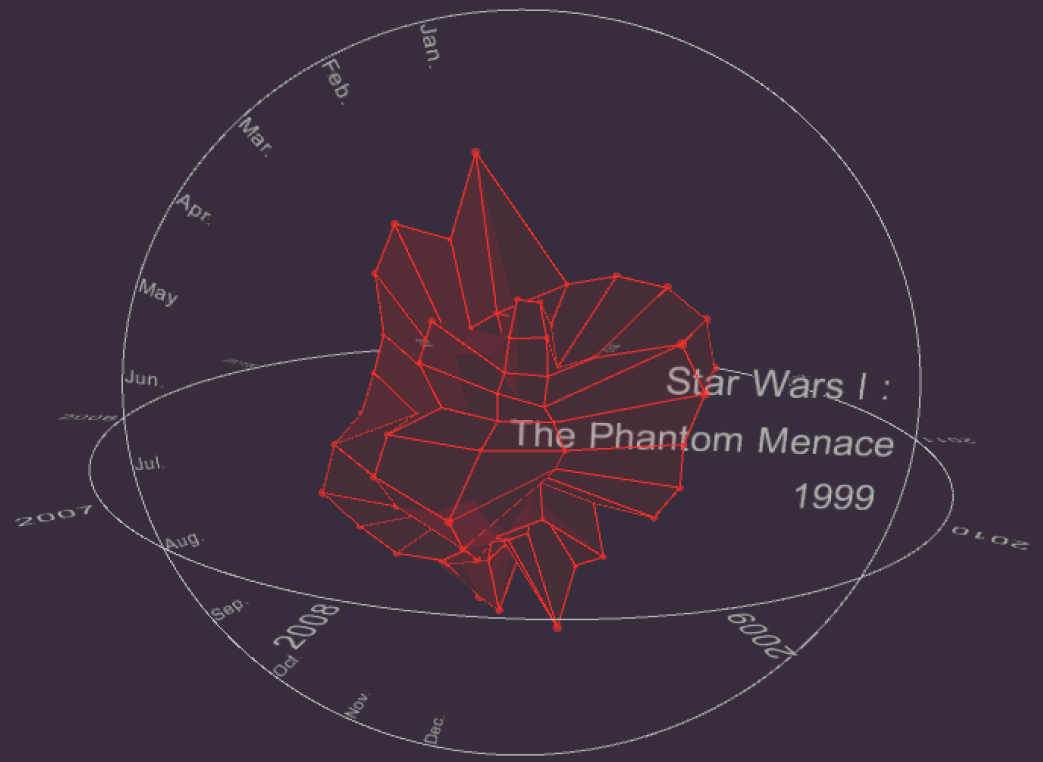
Keyboard Control

Press 1 / 2 / 3 / 4 / 5 / 6 to check movie individually
Press 7 / 8 to check the original / prequel trilogy
Press 9 / 0 to check all in one / different cordinate systems
Press D to show / hide dots of duration times
Press S to show / hide the solids
Press N to show / hide the cromatic frames
Press F to show / hide grey frames
Press T to show / hide the verbal information
Press L to show / hide the cordinate axis and lables

The radius of the circular cordinate axis is 200 checkout times.

Duration time scaling:

40 80 120 160 200 240


Jan.
Feb.
Mar.
Apr.
May
Jun.
Jul.
Aug.
Sep.
Oct.
Nov.
Dec.
2007
2008
2009
2010
Star Wars I :
The Phantom Menace
1999

STAR WARS NEBULA

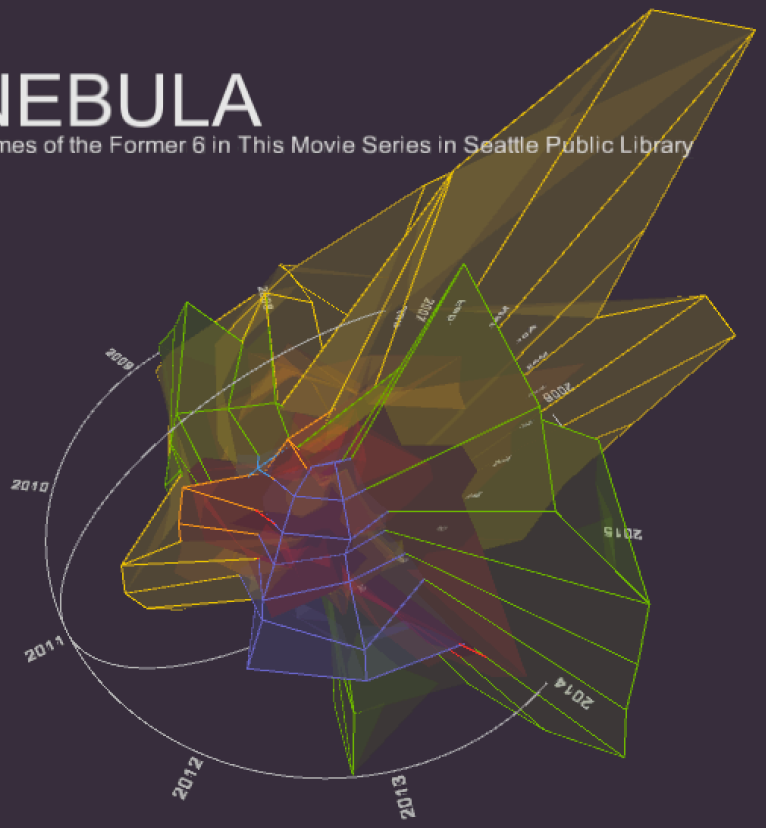
Using Data of Checkout Times and Duration Times of the Former 6 in This Movie Series in Seattle Public Library

Year	Month
2006	January
2007	February
2008	March
2009	April
2010	May
2011	June
2012	July
2013	August
2014	September
2015	October
	November
	December

Reset

Years
Months
All

- Star Wars IV (1977)
- Star Wars V (1980)
- Star Wars VI (1983)
- Star Wars I (1999)
- Star Wars II (2002)
- Star Wars III (2005)



Keyboard Control

Press 1 / 2 / 3 / 4 / 5 / 6 to check movie individually
Press 7 / 8 to check the original / prequel trilogy
Press 9 / 0 to check all in one / different cordinate systems
Press D to show / hide dots of duration times
Press S to show / hide the solids
Press N to show / hide the cromatic frames
Press T to show / hide the verbal information
Press L to show / hide the cordinate axis and lables

Junxiang Yao
MAT259 PROJ 2
3D Interaction & Change Over Time

The radius of the circular cordinate axis is 200 checkout times.

Duration time scaling:

40 80 120 160 200 240

STAR WARS NEBULA

Using Data of Checkout Times and Duration Times of the Former 6 in This Movie Series in Seattle Public Library

Year	Month
2006	January
2007	February
2008	March
2009	April
2010	May
2011	June
2012	July
2013	August
2014	September
2015	October
	November
	December

Reset

Years
Months
All

- Star Wars IV (1977)
- Star Wars V (1980)
- Star Wars VI (1983)
- Star Wars I (1999)
- Star Wars II (2002)
- Star Wars III (2005)

Keyboard Control

Press 1 / 2 / 3 / 4 / 5 / 6 to check movie individually
Press 7 / 8 to check the original / prequel trilogy
Press 9 / 0 to check all in one / different cordinate systems
Press D to show / hide dots of duration times
Press S to show / hide the solids
Press N to show / hide the cromatic frames
Press F to show / hide grey frames
Press T to show / hide the verbal information
Press L to show / hide the cordinate axis and lables

Junxiang Yao
MAT259 PROJ 2
3D Interaction & Change Over Time

The radius of the circular cordinate axis is 200 checkout times.

Duration time scaling:

40 80 120 160 200 240

Data Processing Functions

- Validation: Ensuring that data is "clean, correct and useful”
- Sorting: Arranging items in some sequence and/or in different sets
- Summarization: Reducing detail data to its main points
- Aggregation: Combining multiple pieces of data (possibly from various sources)
- Analysis: Collection, organization, analysis, interpretation and presentation of data
- Reporting: List detail or summary data or computed information