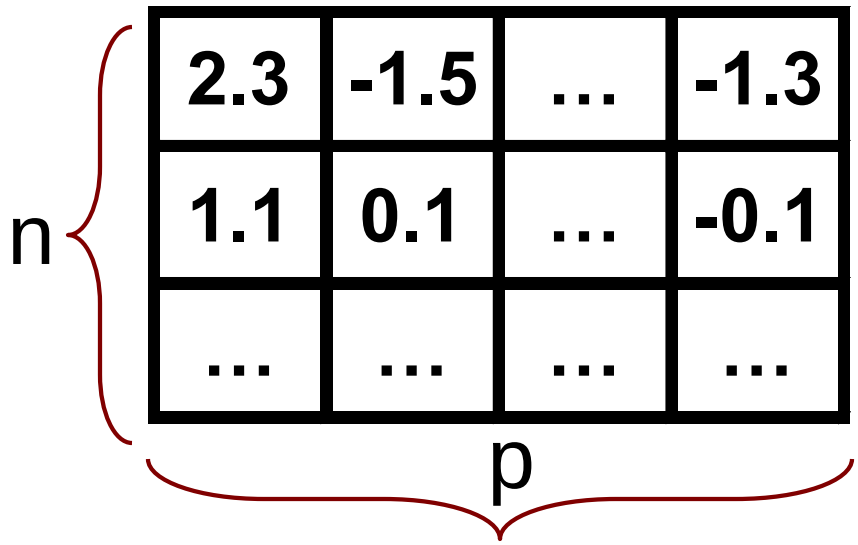
Topic 1:
Introduction to Data Mining

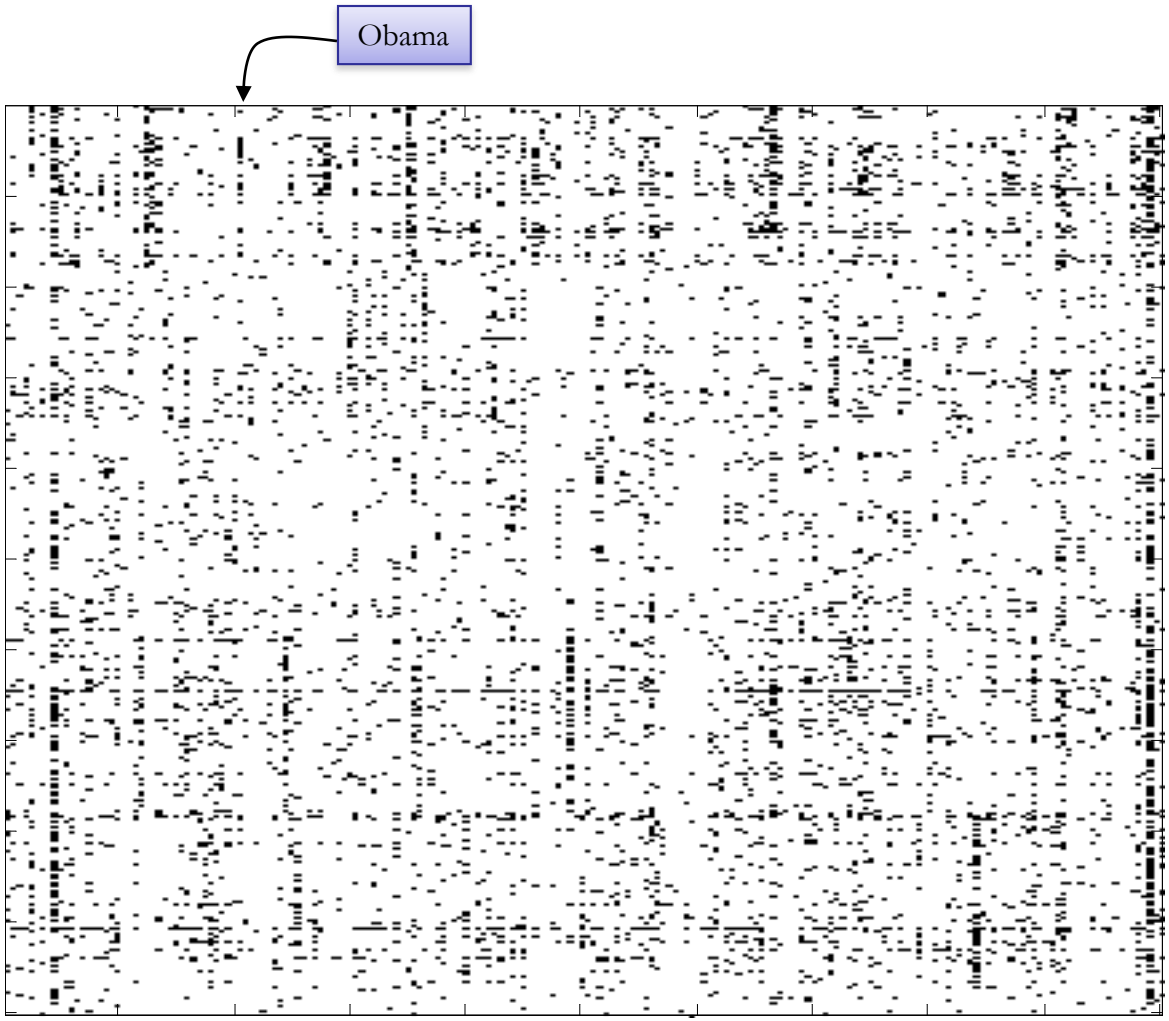
Credits: Chris Volinsky

Types of Data: Flat File Data



- Rows = objects
- Columns = measurements on objects
- Both n and p can be very large in data mining

Types of Data: Text Data



Types of Data: Relational Data

128.200.39.17, -, 3/22/00, 20:55:07, W3SVC, SRVR1, 128.200.39.181, 0, 258, 111, 404, 3, GET, /spt/images/bk1.jpg, -,
128.200.39.17, -, 3/22/00, 20:55:36, W3SVC, SRVR1, 128.200.39.181, 1061, 382, 414, 200, 0, POST, /spt/main.html, -,
128.200.39.17, -, 3/22/00, 20:55:36, W3SVC, SRVR1, 128.200.39.181, 0, 258, 111, 404, 3, GET, /spt/images/bk1.jpg, -,
128.195.36.195, -, 3/22/00, 10:35:11, W3SVC, SRVR1, 128.200.39.181, 781, 363, 875, 200, 0, GET, /top.html, -,
128.195.36.195, -, 3/22/00, 10:35:16, W3SVC, SRVR1, 128.200.39.181, 5288, 524, 414, 200, 0, POST, /spt/main.html, -,
128.195.36.195, -, 3/22/00, 10:35:17, W3SVC, SRVR1, 128.200.39.181, 30, 280, 111, 404, 3, GET, /spt/images/bk1.jpg, -,
…,

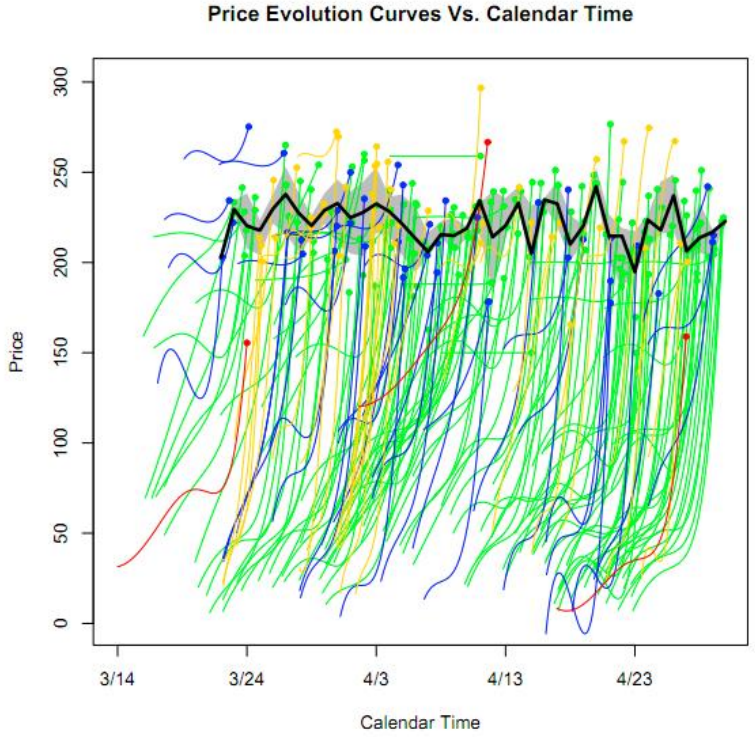
128.195.36.195, Doe, John, 12 Main St, 973-462-3421, Madison, NJ, **07932**
114.12.12.25,Trank, Jill, 11 Elm St, 998-555-5675, Chester, NJ, 07911
…

07911, Chester, NJ, 07954, 34000, , 40.65, -74.12
07932, Madison, NJ, 56000, 40.642, -74.132
…

- Most large data sets are stored in relational data sets
- Special data query language: SQL

Types of Data: Time Series Data

Types of Data: Time Series Data



Jank, Shmueli, et al (2005)

Fig. 10. Rug Plot displaying the price evolution (y-axis) of 217 online auctions over calendar time (x-axis) during a 3-month period. The colored lines show the price path of each auction with color indicating auction length (yellow = 3-day; blue=5-day; green = 7-day; red = 10-day). The dot at the end of each line indicates the final price of the auction. The black line represents the average of the daily closing price , and the gray band is the inter-quartile range.

Types of Data: Transactional Data

Date stamped events (logs):

128.195.36.195, -, 3/22/00, 10:35:11, W3SVC, SRVR1, 128.200.39.181, 781, 363, 875, 200, 0, GET, /top.html, -,
128.195.36.195, -, 3/22/00, 10:35:16, W3SVC, SRVR1, 128.200.39.181, 5288, 524, 414, 200, 0, POST, /spt/main.html, -,
128.195.36.195, -, 3/22/00, 10:35:17, W3SVC, SRVR1, 128.200.39.181, 30, 280, 111, 404, 3, GET, /spt/images/bk1.jpg, -,
128.195.36.101, -, 3/22/00, 16:18:50, W3SVC, SRVR1, 128.200.39.181, 60, 425, 72, 304, 0, GET, /top.html, -,
128.195.36.101, -, 3/22/00, 16:18:58, W3SVC, SRVR1, 128.200.39.181, 8322, 527, 414, 200, 0, POST, /spt/main.html, -,
128.195.36.101, -, 3/22/00, 16:18:59, W3SVC, SRVR1, 128.200.39.181, 0, 280, 111, 404, 3, GET, /spt/images/bk1.jpg, -,
128.200.39.17, -, 3/22/00, 20:54:37, W3SVC, SRVR1, 128.200.39.181, 140, 199, 875, 200, 0, GET, /top.html, -,
128.200.39.17, -, 3/22/00, 20:54:55, W3SVC, SRVR1, 128.200.39.181, 17766, 365, 414, 200, 0, POST, /spt/main.html, -,
128.200.39.17, -, 3/22/00, 20:54:55, W3SVC, SRVR1, 128.200.39.181, 0, 258, 111, 404, 3, GET, /spt/images/bk1.jpg, -,
128.200.39.17, -, 3/22/00, 20:55:07, W3SVC, SRVR1, 128.200.39.181, 0, 258, 111, 404, 3, GET, /spt/images/bk1.jpg, -,
128.200.39.17, -, 3/22/00, 20:55:36, W3SVC, SRVR1, 128.200.39.181, 1061, 382, 414, 200, 0, POST, /spt/main.html, -,
128.200.39.17, -, 3/22/00, 20:55:36, W3SVC, SRVR1, 128.200.39.181, 0, 258, 111, 404, 3, GET, /spt/images/bk1.jpg, -,
128.200.39.17, -, 3/22/00, 20:55:39, W3SVC, SRVR1, 128.200.39.181, 0, 258, 111, 404, 3, GET, /spt/images/bk1.jpg, -,
128.200.39.17, -, 3/22/00, 20:56:03, W3SVC, SRVR1, 128.200.39.181, 1081, 382, 414, 200, 0, POST, /spt/main.html, -,
128.200.39.17, -, 3/22/00, 20:56:04, W3SVC, SRVR1, 128.200.39.181, 0, 258, 111, 404, 3, GET, /spt/images/bk1.jpg, -,
128.200.39.17, -, 3/22/00, 20:56:33, W3SVC, SRVR1, 128.200.39.181, 0, 262, 72, 304, 0, GET, /top.html, -,
128.200.39.17, -, 3/22/00, 20:56:52, W3SVC, SRVR1, 128.200.39.181, 19598, 382, 414, 200, 0, POST, /spt/main.html, -,

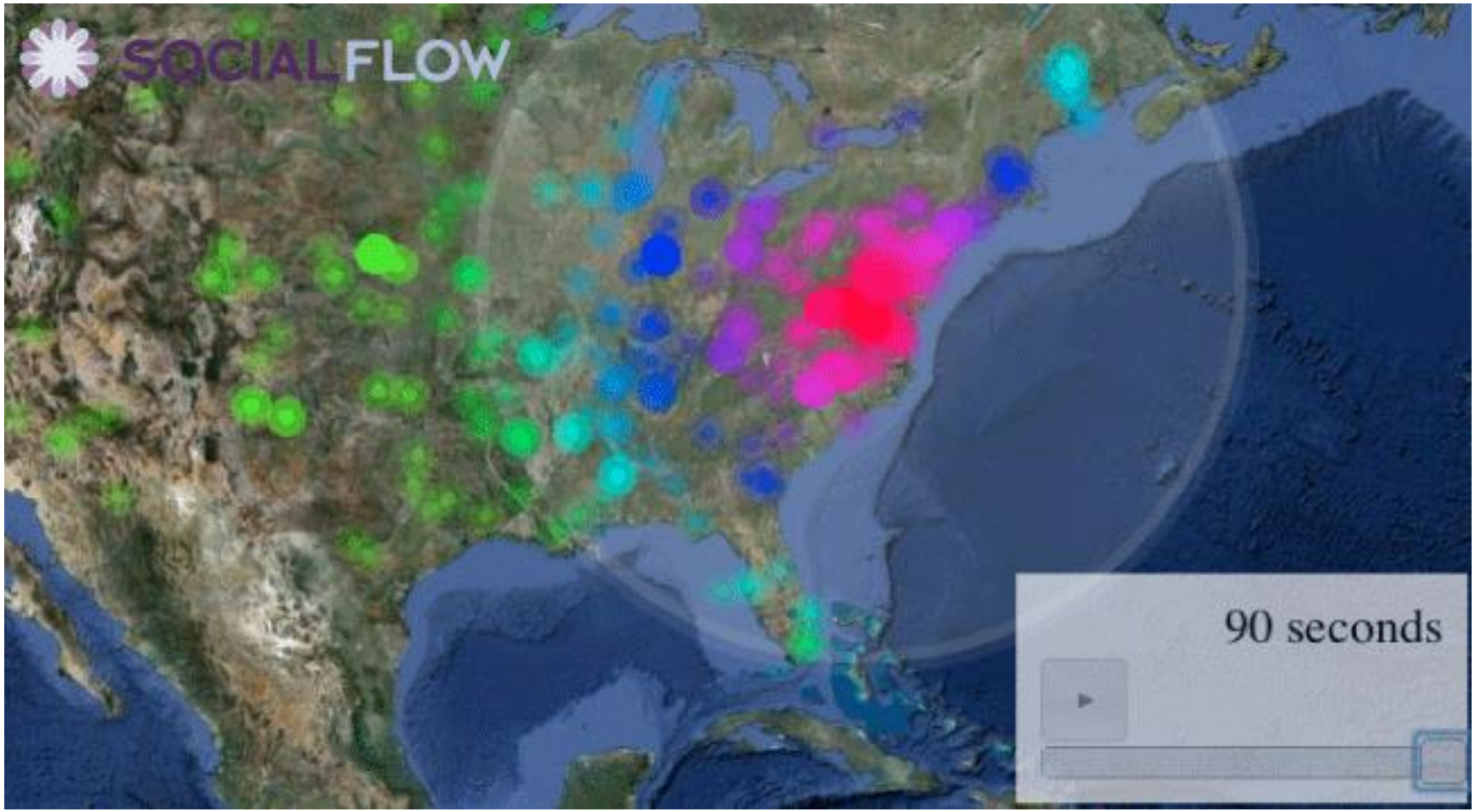
Can be represented as a time series:

User 1	2	3	2	2	3	3	3	1	1	1	3	1	3	3	3	3
User 2	3	3	3	1	1	1										
User 3	7	7	7	7	7	7	7	7								
User 4	1	5	1	1	1	5	1	5	1	1	1	1	1	1		
User 5	5	1	1	5												
…		…														

Types of Data: Spatio-Temporal Data

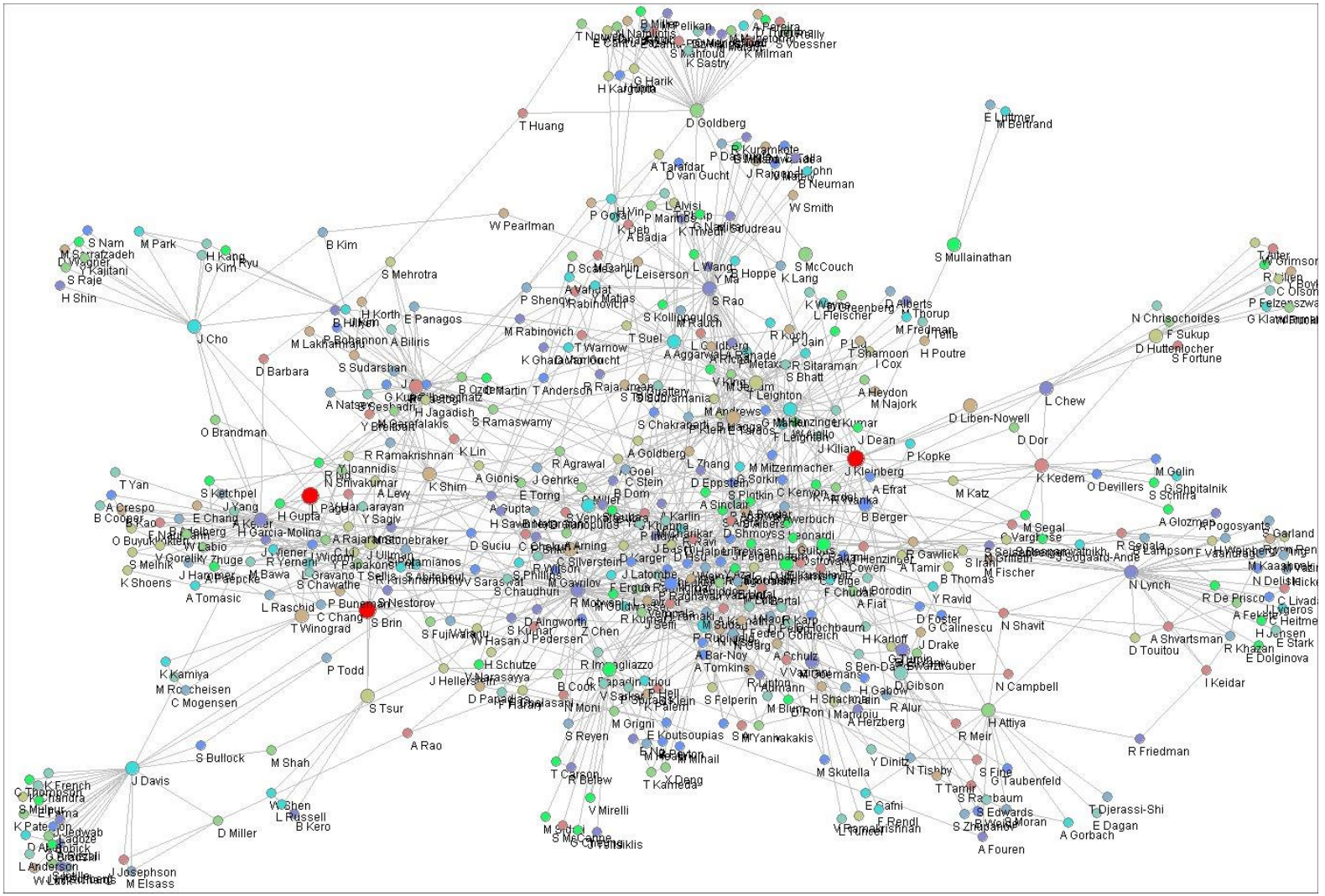
@b_mc817
Glendaaaaa

Omg earthquake!!!



Types of Data: Image Data

Types of Data: Network Data



Algorithms for estimating relative importance in networks
S. White and P. Smyth, *ACM SIGKDD*, 2003.