

Big Data

*Lawrence Snyder
University of Washington, Seattle*

2 Stories ... 1 Written By Computer

- “Things looked bleak for the Angels when they **M** trailed by two runs in the ninth inning, but Los Angeles recovered thanks to a key single from Vladimir Guerrero to pull out a 7-6 victory over the Boston Red Sox at Fenway Park on Sunday.”
- “The University of Michigan baseball team used a **H** four-run fifth inning to salvage the final game in its three-game weekend series with Iowa winning 7-5 on Saturday afternoon (April 24) at the Wilpon Baseball Complex, home of historic Ray Fisher Stadium.”

2 Stories ... 1 Written By Computer

- Scan many thousands of sports stories ... and detect key phrases
- <situation description> for the <slang name> when they trailed by <number> runs in the <point in game>, but <short name> recovered thanks to a key <play> from <player> to pull out a <score> victory over the <opponent formal name> at <location> on <when>.
- “Things looked bleak for the Angels when they trailed by two runs in the ninth inning, but Los Angeles recovered thanks to a key single from Vladimir Guerrero to pull out a 7-6 victory over the Boston Red Sox at Fenway Park on Sunday.”

Big Data

- “Big data” refers to the analysis of any large corpus of data to extract information, usually by means of statistical techniques
- We’ve already seen big data in action
 - In Target’s analysis of purchasing behavior to determine if a woman is pregnant
 - In Google’s Page Rank, their estimate of how significant a page is
- Today we discuss big data ... but there is a limit to how big it can get in a lecture!



Where is Big Data Found?

- Data is everywhere –
 - All Facebook users constitute a data archive that Facebook analyzes continually
 - Google has crawled the WWW for years ... their data is truly big!
 - Census bureau
 - UW's student database
 - Every company's consumer records
 - Governmental DBs like licensing, tax revenues ,etc.
 - ...

More Than No.s

- Data that is regularly gathered for some purpose typically contains a lot more information than the recorded numbers

reveal ...
and it's
often
easy to
get

 8  2 

Vital Statistics Data Available Online

This page is a portal to the online data dissemination activities of the [Division of Vital Statistics](#), including both interactive online data access tools and downloadable public use data files.

On this Page

- [Downloadable Data Files](#)
- [Data Access Tools](#)




Downloadable Data Files

Public use [Birth](#), [Period Linked Birth - Infant Death](#), [Birth Cohort Linked Birth - Infant Death](#), [Mortality Multiple Cause](#), and [Fetal Death](#) data files are available for independent research and analyses.

- [Vital Statistics Data Release Policy](#)
- [Data Users Agreement](#)

Birth Data Files

User's Guide (.pdf files)

2012 (1.4 MB) 
2011 (1 MB) 
2010 Addendum (210 KB) 

U.S. Data (.zip files)*

2012 (218 MB) 
2011 (215 MB) 
2010 (209 MB) 

U.S. Territories Data (.zip files)

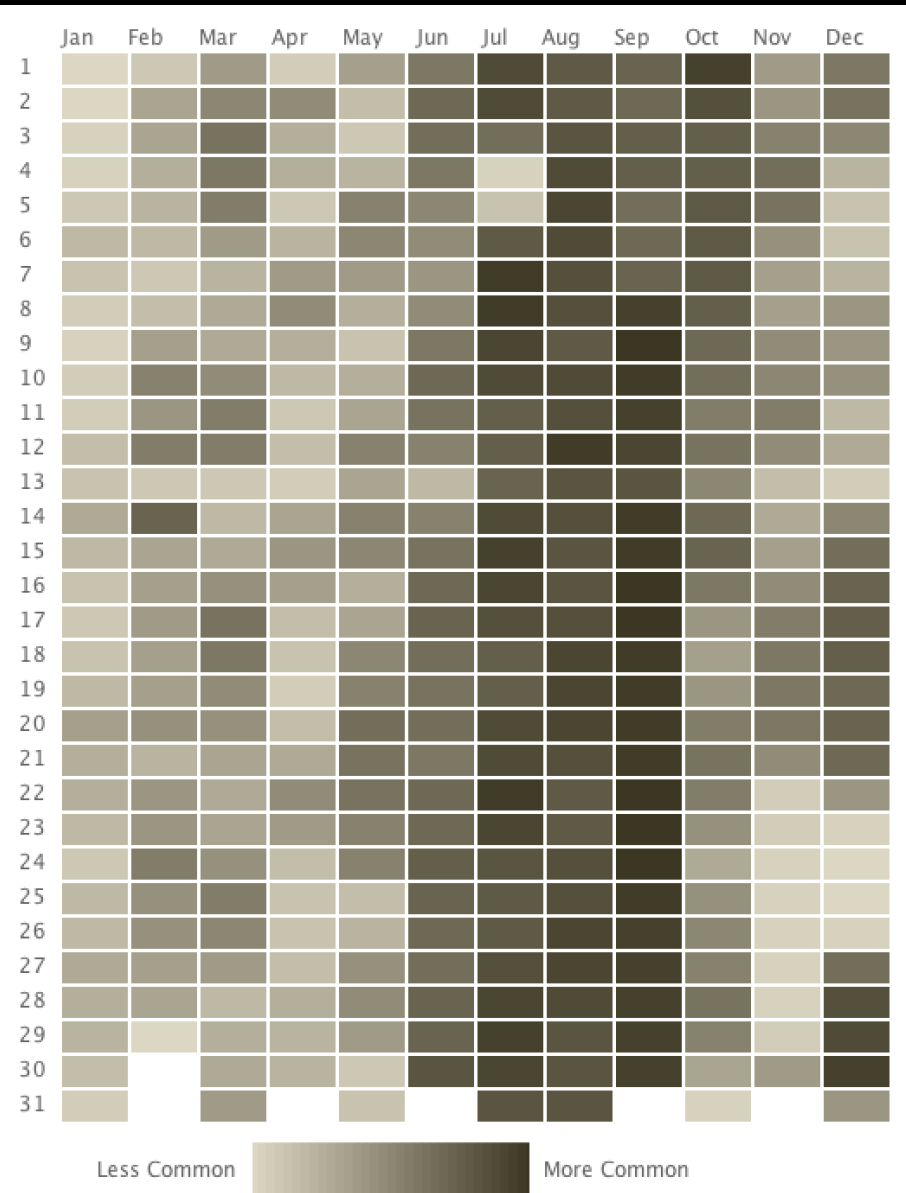
2012 (2.8 MB) 
2011 (1.7 MB) 
2010 (1.7 MB) 

More Than No.s

- The data used for our “heat map” of birthdays came from birth certificate records

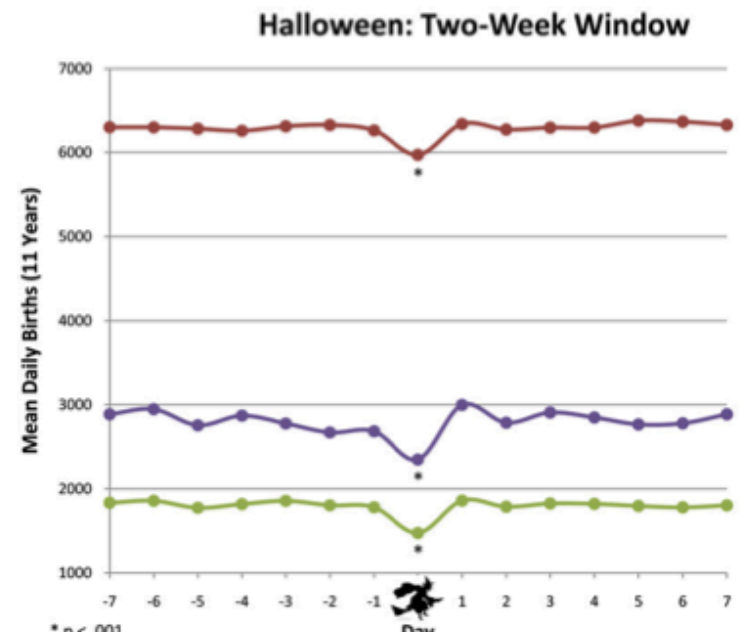
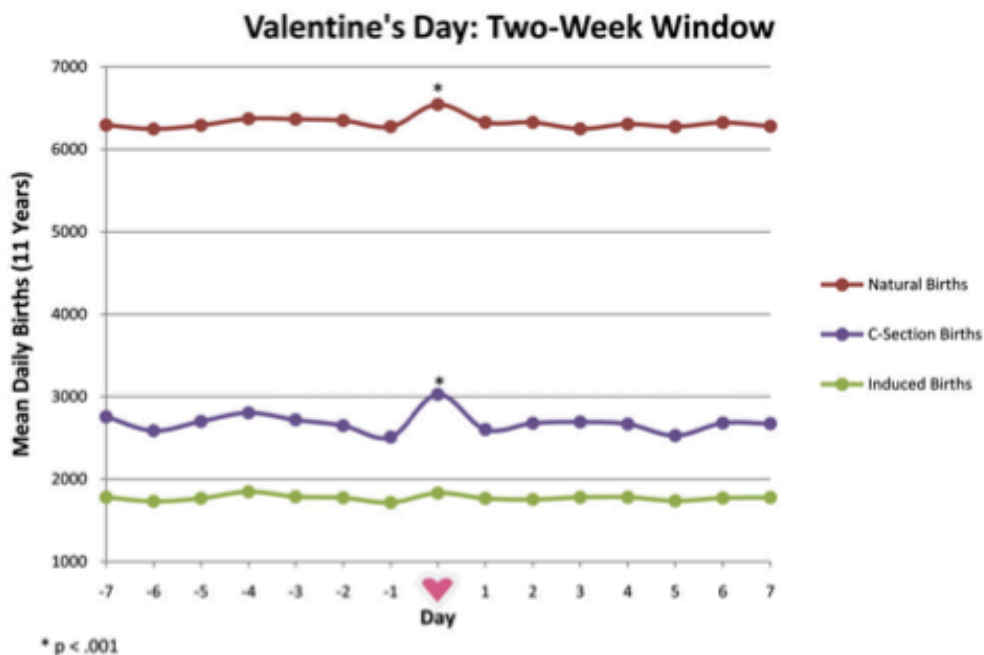
CDC birth data for the years 1969-1988

Processing: add by day
sort descending, plot



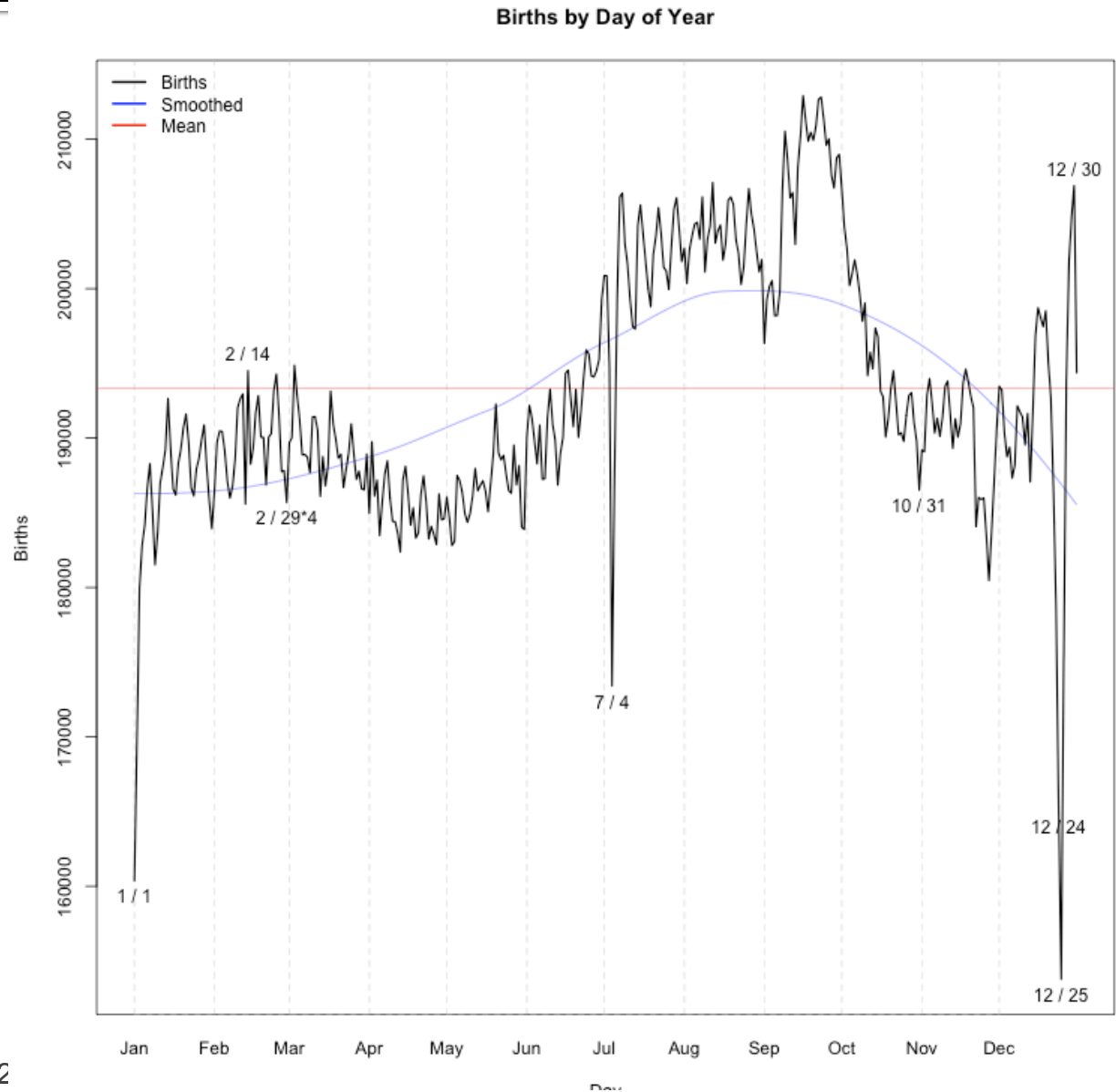
More Than No.s

- Preference for birthdays ...
 - People love Valentines Day, and hate Halloween
 - Compute average for each day around date, plot



More Than No.s

- Plot raw data
- <Smooth>



Suppose You Have Lots of Text

- One thing you can do is figure out how often certain letters occur ... good for “wink-comm”



Frequency Of Longer Sequences

- Counting the frequency of letters is more technically called “computing a 1-gram”
- More generally, an n -gram is counting up the frequency of sequences (of pretty much anything digital) of length n
- The 2-grams of this DNA: CGTTGACAACGT are: CG, GT, TT, TG, GA, AC, CA, AA, AC, CG, GT ... so CG & GT occur twice, others just once
- The 2-grams of words in “To be or not to be” are: to-be, be-or, or-not, not-to, to-be
- Etc.

Google's List

- Google has a lot of text, and has compiled the n -grams for tokens (i.e. words, non-blank letter sequences followed by punctuation or blank)
- Number of tokens: 1,024,908,267,229
- Number of sentences: 95,119,665,584
- Number of unigrams: 13,588,391
- Number of bigrams: 314,843,401
- Number of trigrams: 977,069,902
- Number of fourgrams: 1,313,818,354
- Number of fivegrams: 1,176,470,663

What Are n -grams Good For

- Spelling correction software: Using an n -gram of letters, what's wrong with "thniking"?
- Optical Character Recognition ... if you have figured out "to be or not to <smudge>" you might use word 2-grams starting with "to"
- Google will just show you cool plots ...

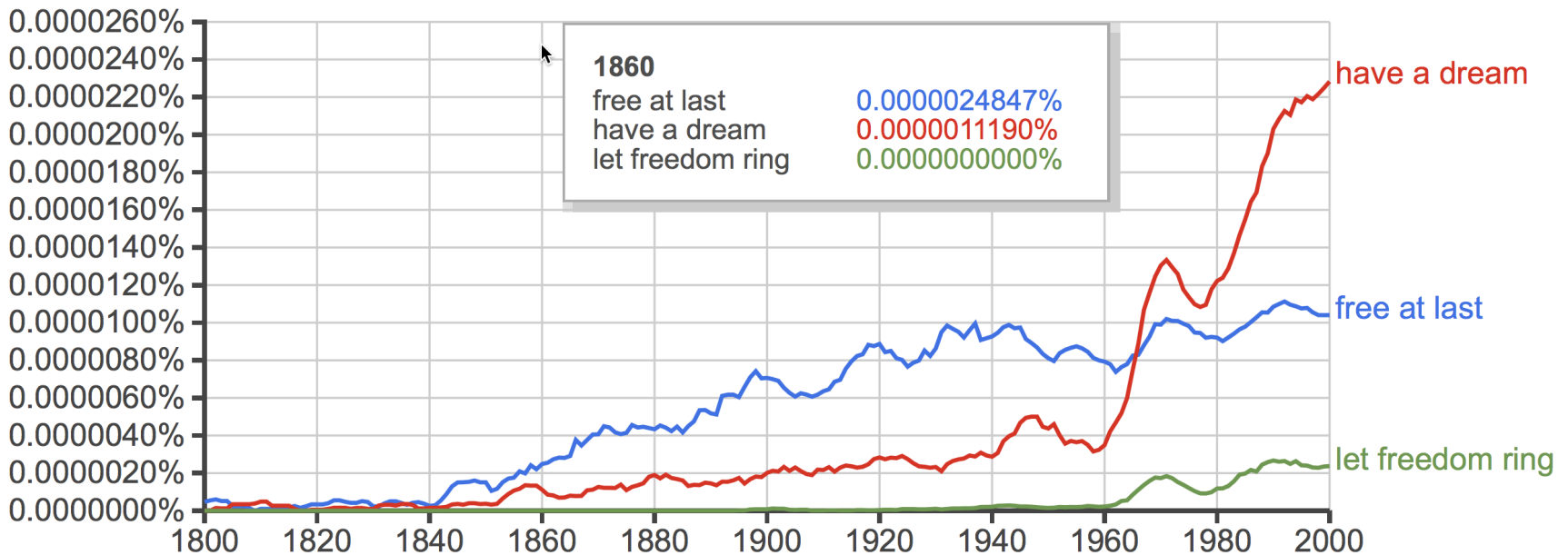
Google's n-gram Viewer

Google books Ngram Viewer

Graph these comma-separated phrases: case-insensitive

between and from the corpus with smoothing of .

[Search lots of books](#)

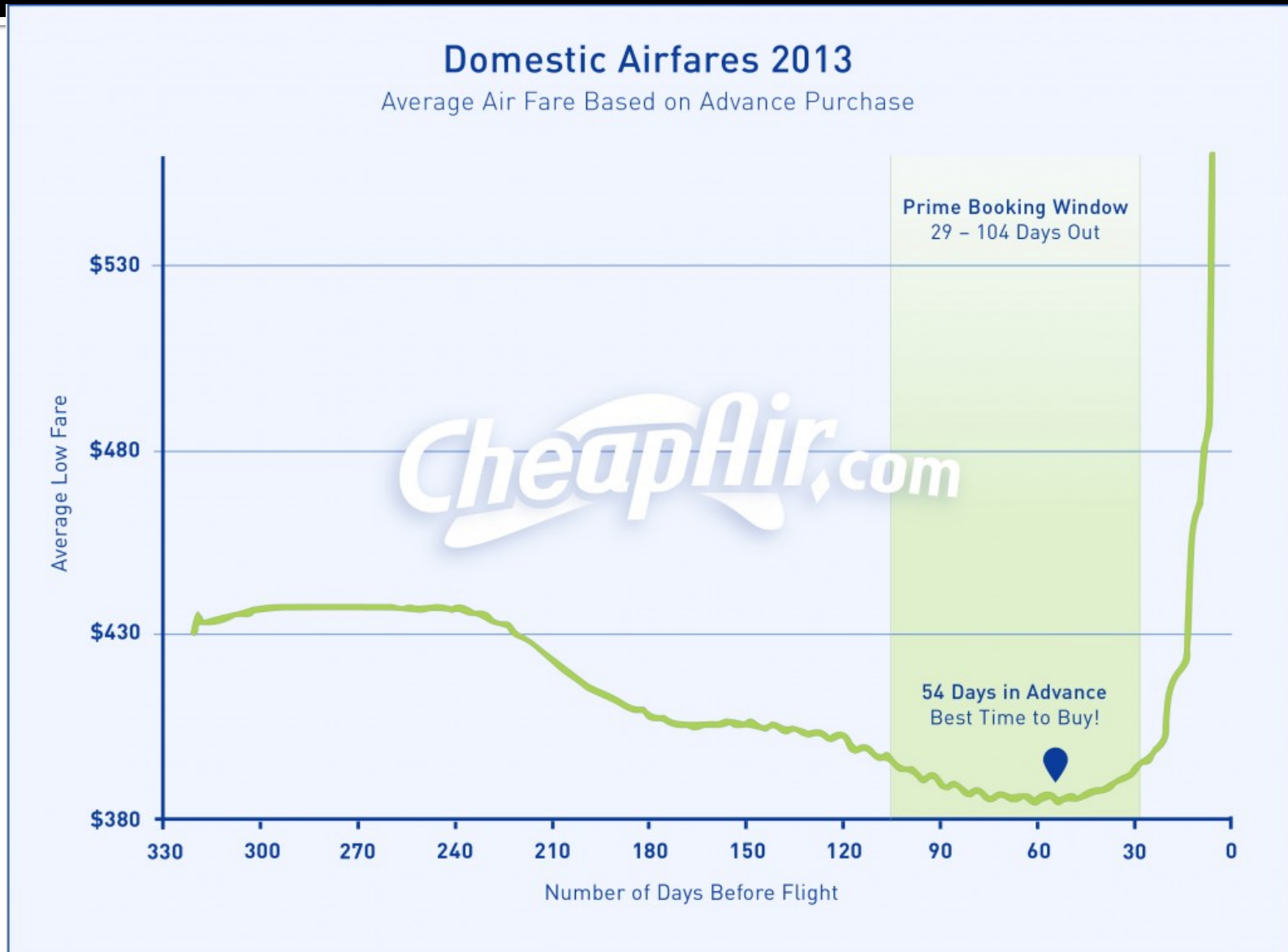


(click on line/label for focus)

Analyze Airline Prices

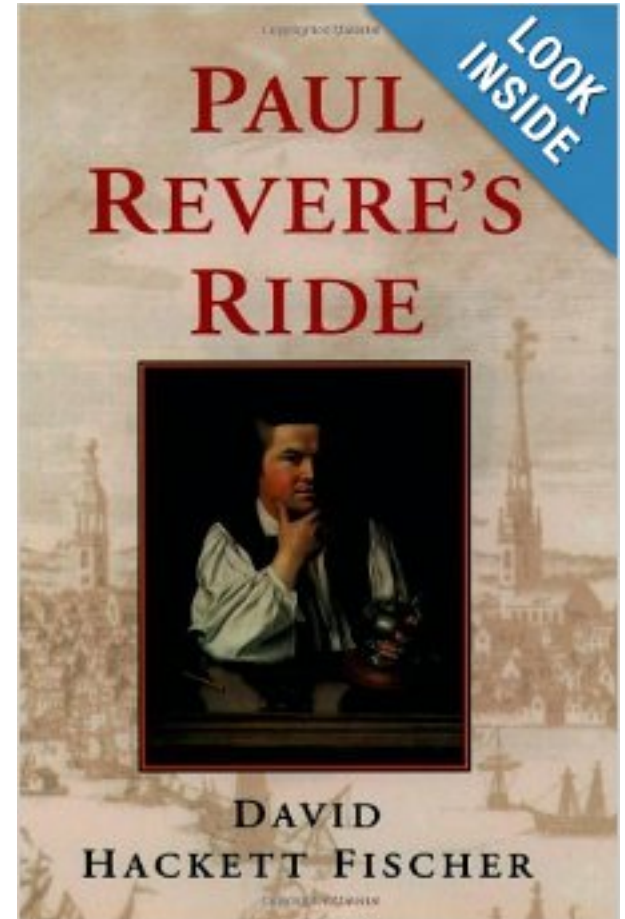
- CheapOair analyzed ticket price for 4 million airline trips in 2013 from 320 days before flight
- Fifty-four days before takeoff is, on average, when domestic airline tickets are at their **absolute lowest price**.
- Prime booking window: 104 - 29 days before your trip ... usually within \$10 of best price

Information You Can Use



But There's Always Been Data

- Revolutionary War Boston Club membership as relayed in Fischer's Book
- Find it in the appendix ... type it in
- Analysis by Kieran Healy <http://kieranhealy.org/blog/archives/2013/06/09/using-metadata-to-find-paul-revere/>



The Table Associations ...

- A 254 x 7 table: colonist x organization

Q Search this file...

		StAndrewsLodge	LoyalNine	NorthCaucus	LongRoomClub	TeaParty	BostonCommittee	LondonEnemies
1								
2	Adams.John	0	0	1	1	0	0	0
3	Adams.Samuel	0	0	1	1	0	1	1
4	Allen.Dr	0	0	1	0	0	0	0
5	Appleton.Nathaniel	0	0	1	0	0	1	0
6	Ash.Gilbert	1	0	0	0	0	0	0
7	Austin.Benjamin	0	0	0	0	0	0	1
8	Austin.Samuel	0	0	0	0	0	0	1
9	Avery.John	0	1	0	0	0	0	1
10	Baldwin.Cyrus	0	0	0	0	0	0	1
11	Ballard.John	0	0	1	0	0	0	0

- Organize such data with spreadsheet software

The Table Associations ...

- A 254 x 7 table: colonist x organization

Search this file...

		StAndrewsLodge	LoyalNine	NorthCaucus	LongRoomClub	TeaParty	BostonCommittee	LondonEnemies
1								
2	Adams.John	0	0	1	1	0	0	0
3	Adams.Samuel	0	0				1	1
4	Allen.Dr	0	0				0	0
5	Appleton.Nathaniel	0	0				1	0
6	Ash.Gilbert	1	0	0	0	0	0	0
7	Austin.Benjamin	0	0	0	0	0	0	1
8	Austin.Samuel	0	0	0	0	0	0	1
9	Avery.John	0	1	0	0	0	0	1
10	Baldwin.Cyrus	0	0	0	0	0	0	1
11	Ballard.John	0	0	1	0	0	0	0

Metadata

The Table of Colonists ...

- Now, transpose the table – make columns rows, and rows columns
- A 7×254 table: organization x colonist

	Adams.John	Adams.Samuel	Allen.Dr	Appleton.Nathaniel	Ash.Gilbert
StAndrewsLodge	0	0	0	0	1
LoyalNine	0	0	0	0	0
NorthCaucus	1	1	1	1	0
LongRoomClub	1	1	0	0	0
TeaParty	0	0	0	0	0
BostonCommittee	0	1	0	1	0
LondonEnemies	0	1	0	0	0

Multiply The Two Matrices $A(A^T)$

- It produces a 254×254 table that shows for any pair of people (one in row and one in column) how many associations they have in common!

	Adams.John	Adams.Samuel	Allen.Dr	Appleton.Nathaniel	Ash.Gilbert
Adams.John	-	2	1	1	0
Adams.Samuel	2	-	1	2	0
Allen.Dr	1	1	-	1	0
Appleton.Nathaniel	1	2	1	-	0
Ash.Gilbert	0	0	0	0	-
Austin.Benjamin	0	1	0	0	0

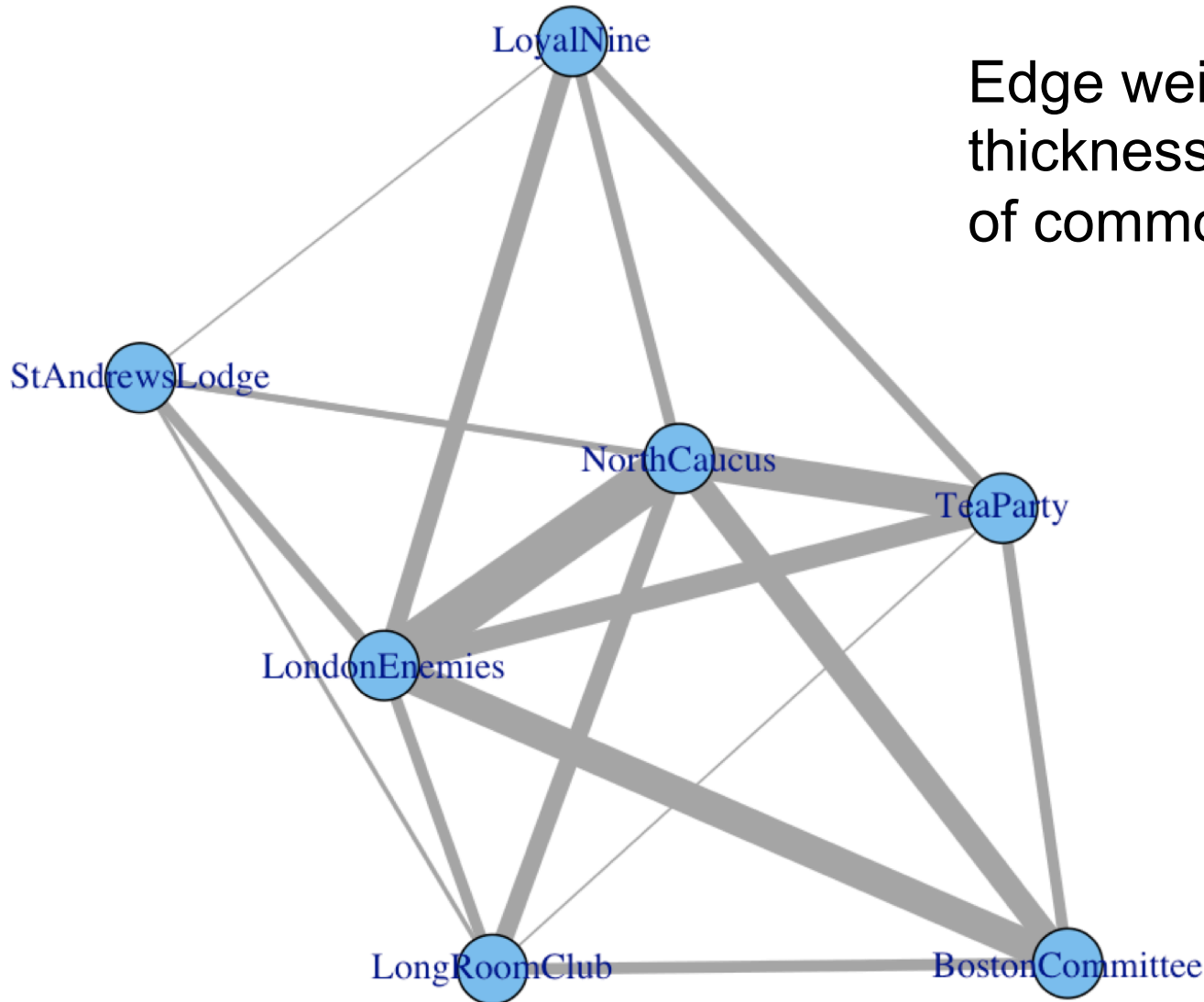
- The non-zero entries indicate pairs that might be collaborators

Multiply In Other Order (A^T)A

- Produces an organization x organization table saying how many members each pair (row, column) have in common

	StAndrewsLo	LoyalNine	NorthCaucus	LongRoomCl	TeaParty	BostonComm	LondonEnem
StAndrewsLo	-	1	3	2	3	0	5
LoyalNine	1	-	5	0	5	0	8
NorthCaucus	3	5	-	8	15	11	20
LongRoomCl	2	0	8	-	1	5	5
TeaParty	3	5	15	1	-	0	1
BostonComm	0	0	11	5	0	-	0
LondonEnem	5	8	20	5	1	0	-

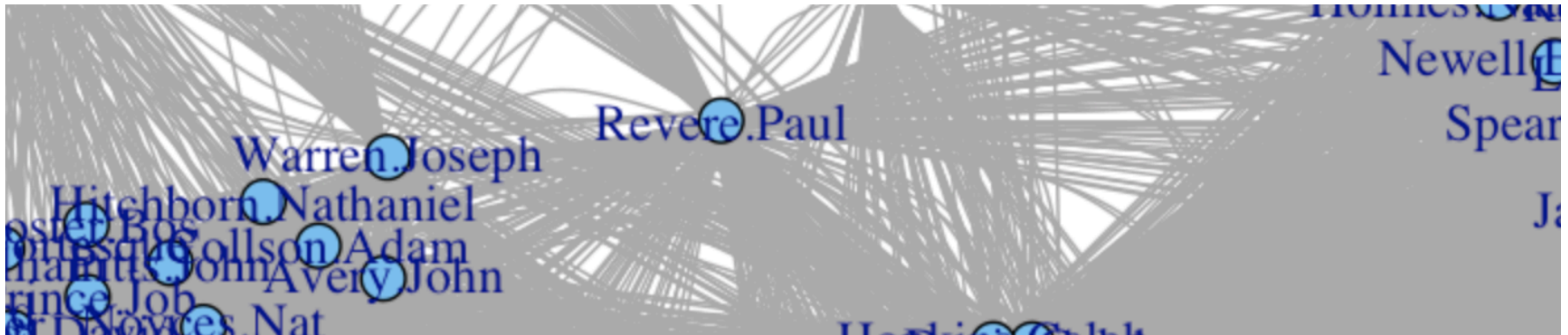
Diagram 7x7 Table for "A Visual"



Edge weight (line thickness) is number of common members

Zoom To See Who's Best Connected

- A person of suspicion!



- Used membership metadata, performed normal analysis on it, identified key player

Betweenness Centrality

- How likely is that in the graph of who's connected to whom, a shortest path goes through a specific person – measure of connectedness

Revere.Paul	Urann.Thomas	Warren.Joseph	Peck.Samuel
3839	2185	1817	1150
Barber.Nathaniel	Cooper.William	Hoffins.John	Bass.Henry
931	931	931	852
Chase.Thomas	Davis.Caleb		
852	852		

- Paul Revere is on 3839 shortest paths in the graph

Summary

- Data collections are everywhere
- Analyzing them can discover amazing facts
- Forms of analysis
 - Many techniques reveal interesting results with very primitive tools
 - We saw sorting, plotting, averaging, matrix product, centrality measures
 - Statistical software already exists
 - Mostly, the information can be “anonymized”

Summary

- Data collections are everywhere
- Analyzing them can discover amazing facts
- Forms of analysis
 - Many techniques reveal interesting results with very primitive tools
 - We saw sorting, plotting, averaging, matrix product, centrality measures
 - Statistical software already exists
 - Mostly, the information can be “anonymized”

How can big data be useful to you?