Lecture 2: New data from old (recoding, reshaping, transforming)

Last time

We tried to make explicitly the role that computation plays in modern data analysis as part of a larger sales pitch for our approach to this class and the use of R as our main platform for understanding data

We then went on a (wobbly) detour back to the late 1600s and early 1700s to talk about the first test of hypothesis as well as other data use patterns that hold true today

Once collected and distributed, data find secondary uses that the “creators” can’t always anticipate

Our view of data (numerical, graphical, auditory, ...) is mediated by the available “technology” (actual tools as well as the accepted or dominant representations of the time)

We ended by introducing a data set that we’ll examine in more detail today...

Today

We are going to play with the idea of a data table a fair bit -- We will examine how we can reshape, aggregate and reformat data to provide us with alternate views of some phenomenon

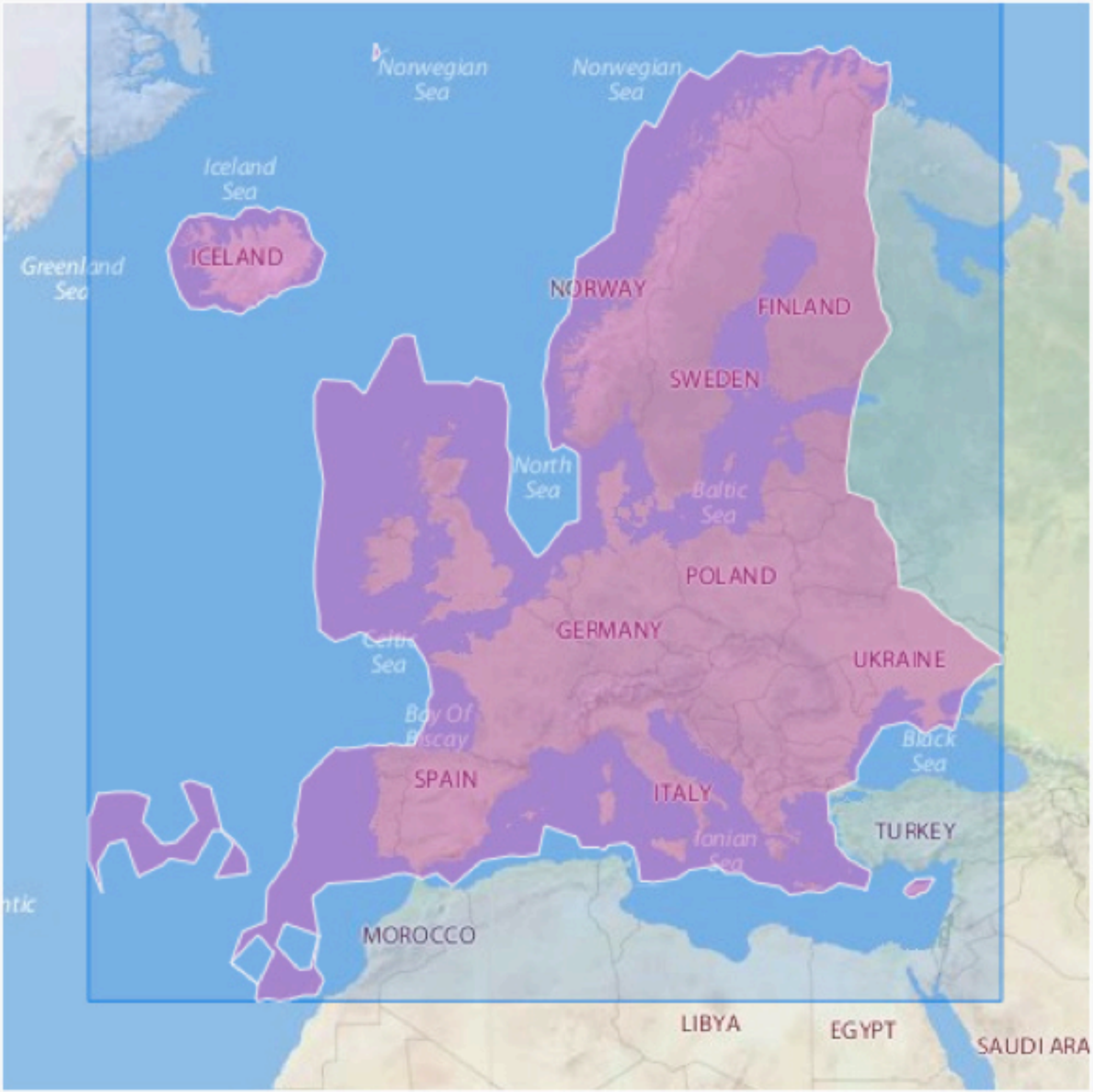
Along the way, we will spend some time talking about privacy and about how the computer represents time -- We'll also learn some basic graphical tools for visualizing simple data types

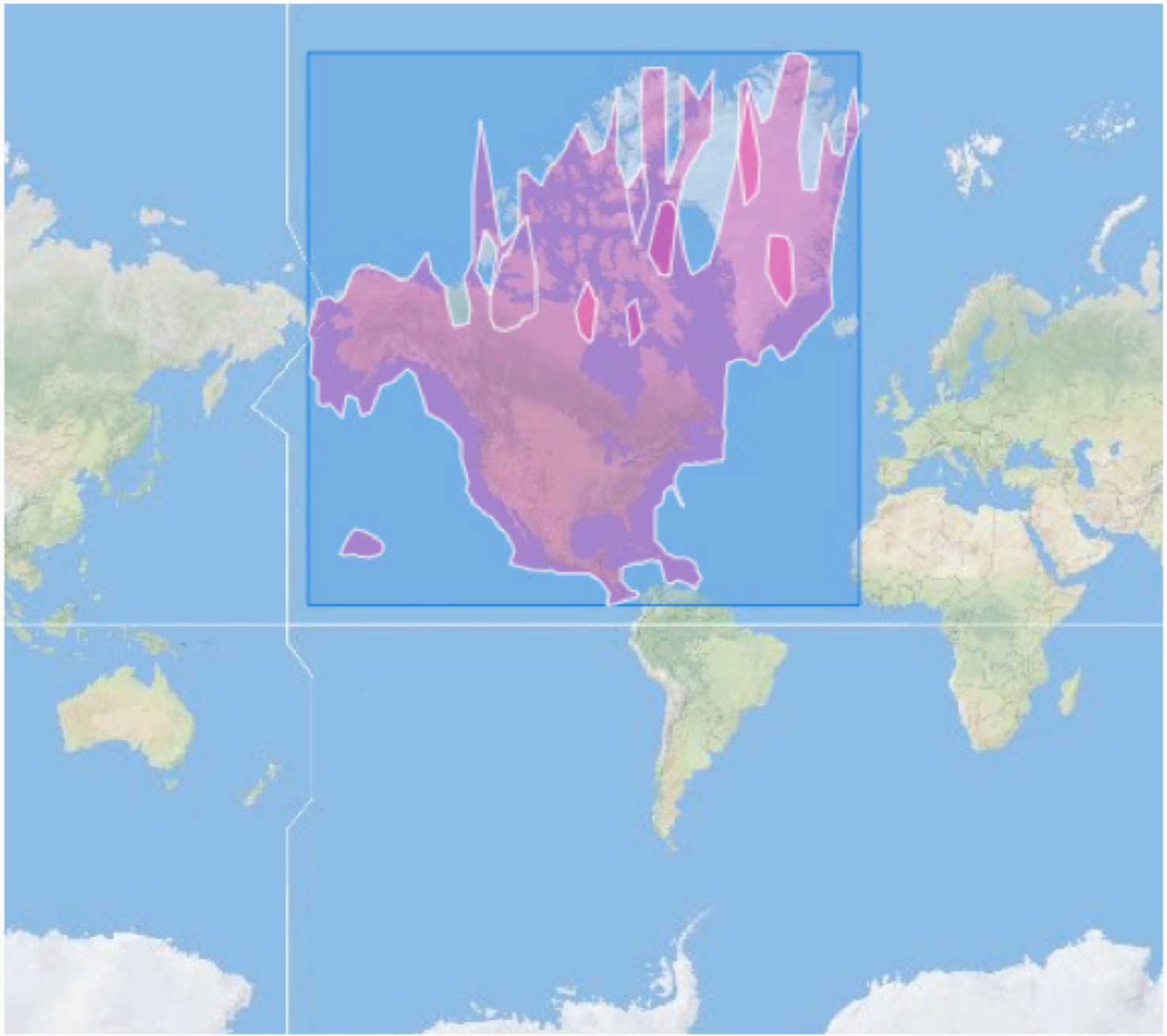
What's it add up to?

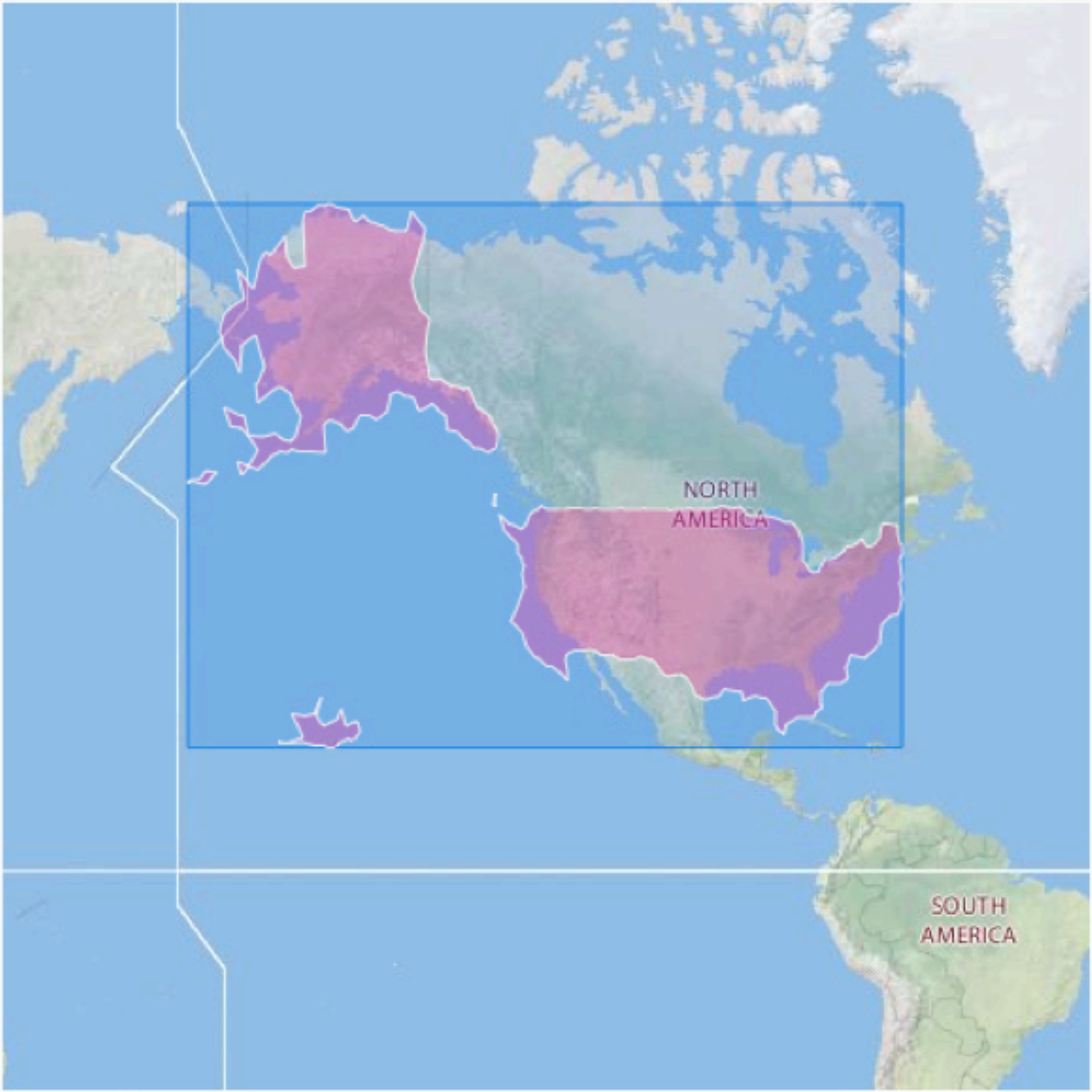
Last time I asked you to start to compile information about where you give off data -- At the end of the week I was hoping you'd think a little about what all that data might amount to if someone were to have access to it

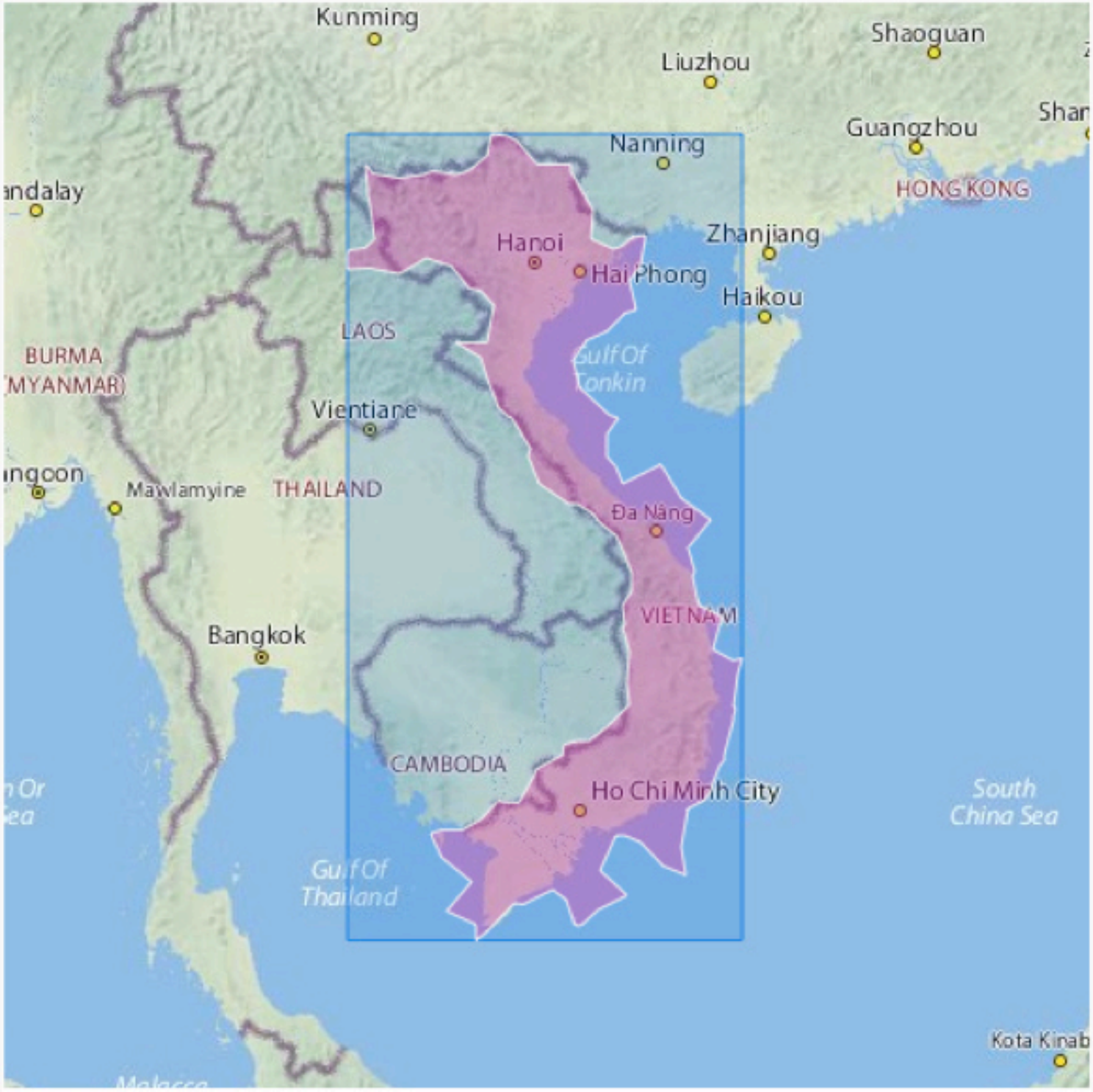
Under the heading of surprising results in this direction, Aaron Cope (formerly at Flickr) had a great idea to use the millions of geotagged images that have been uploaded

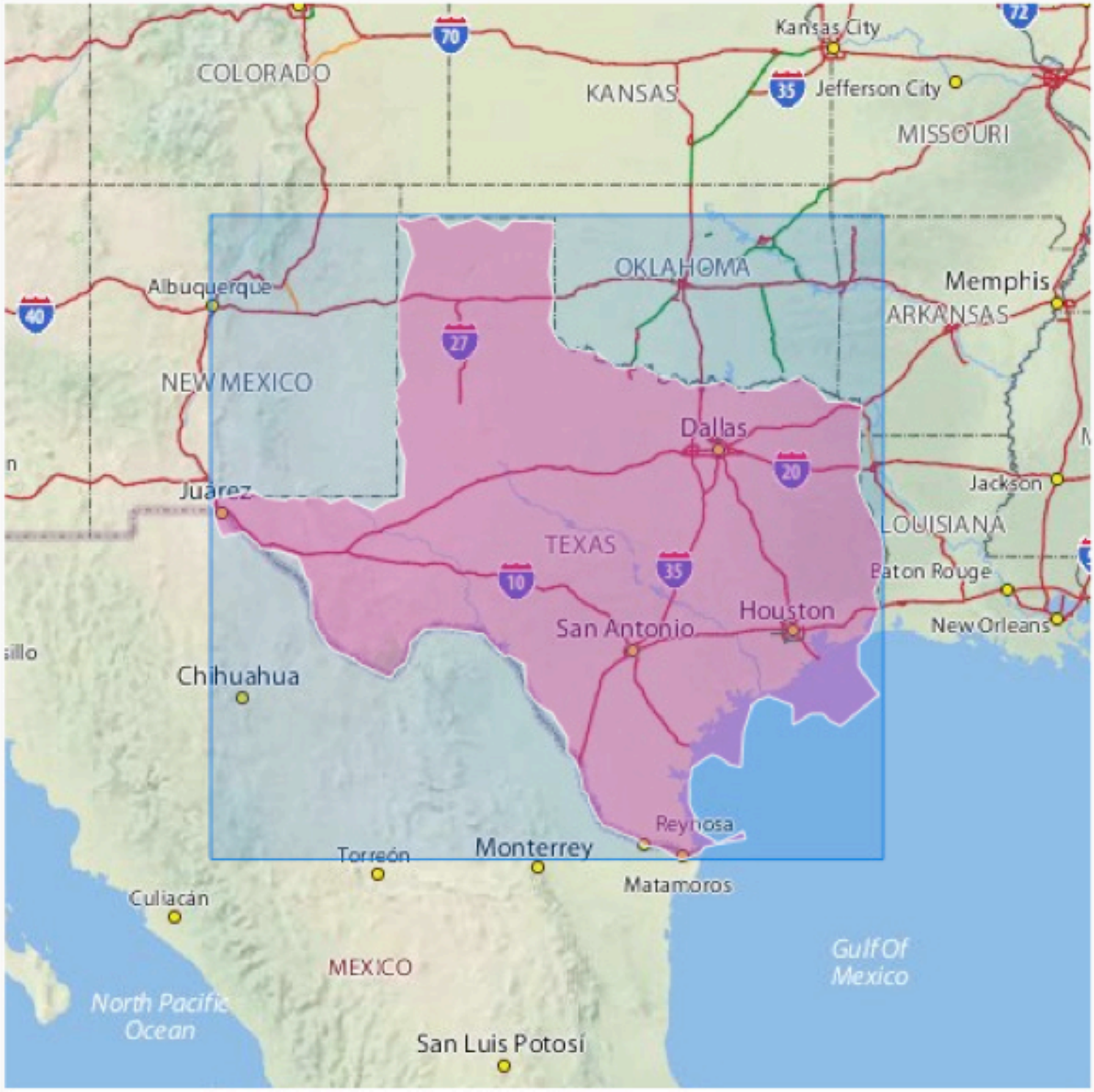
Often these images have an explicit geotag (think lat/lon) and users will add text tags that describe where they are (the United States, Texas, London) -- Aaron's idea was to use all the geotag points together with the explicit place names to build a map of those places...

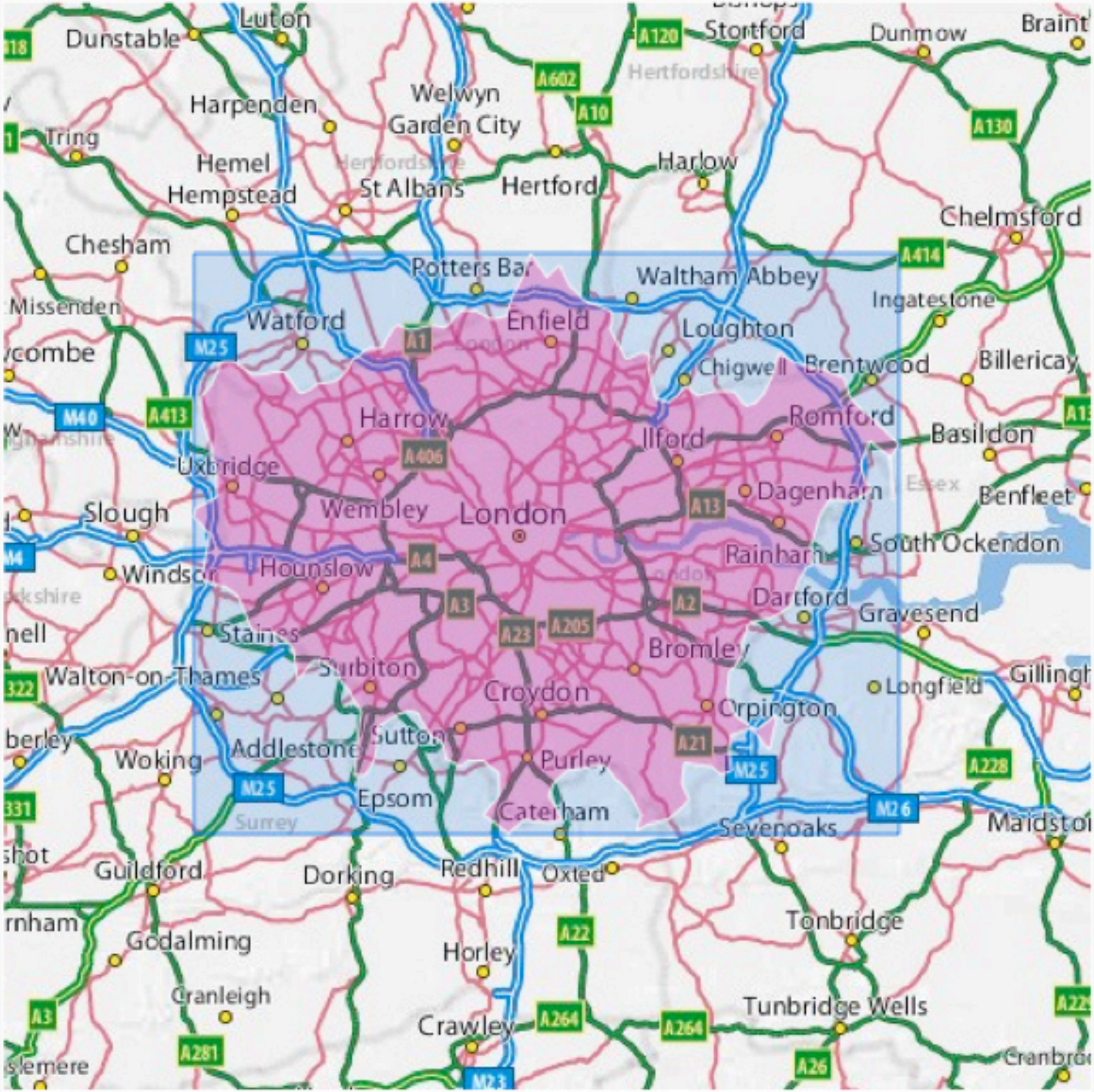

Norwegian Sea
Norwegian Sea
Iceland Sea
ICELAND
Greenland Sea
NORWAY
FINLAND
SWEDEN
North Sea
Baltic Sea
POLAND
GERMANY
Celtic Sea
UKRAINE
Bay Of Biscay
Black Sea
SPAIN
ITALY
TURKEY
Ionian Sea
MOROCCO
LIBYA
EGYPT
SAUDI ARA




NORTH
AMERICA
SOUTH
AMERICA


Kunming
Liuzhou
Shaoguan
Guangzhou
Nanning
HONG KONG
Zhanjiang
Hanoi
Hai Phong
Haikou
LAOS
BURMA
(MYANMAR)
Gulf Of
Tonkin
Vientiane
Mawlamyine
THAILAND
Đa Nẵng
VIETNAM
Bangkok
CAMBODIA
Ho Chi Minh City
South
China Sea
Gulf Of
Thailand
Kota Kinab


COLORADO
KANSAS
Kansas City
Jefferson City
MISSOURI
OKLAHOMA
Albuquerque
Memphis
ARKANSAS
NEW MEXICO
Dallas
Juárez
Jackson
LOUISIANA
TEXAS
Baton Rouge
Houston
San Antonio
New Orleans
Chihuahua
Reynosa
Monterrey
Torreón
Matamoros
Culiacán
MEXICO
Gulf Of Mexico
North Pacific Ocean
San Luis Potosí


Luton
Dunstable
Harpenden
Welwyn
Garden City
Stortford
Dunmow
Braint
Tring
Hemel
Hempstead
St Albans
Hertford
Harlow
Chelmsford
Chesham
Potters Bar
Waltham Abbey
Ingatestone
Missenden
Watford
Enfield
Loughton
Chigwell
Brentwood
Billericay
Harrow
Ilford
Romford
Basildon
Uxbridge
Dagenham
Benfleet
Slough
Wembley
London
Rainham
South Ockendon
Windsor
Hounslow
Dartford
Gravesend
Staines
Bromley
Walton-on-Thames
Surbiton
Croydon
Orpington
Longfield
Addlestone
Sutton
Purley
Woking
Epsom
Caterham
Sevenoaks
Guildford
Dorking
Redhill
Oxted
Tonbridge
Godalming
Horley
Cranleigh
Crawley
Tunbridge Wells
M25
A1
A2
A3
A4
A13
A23
A205
A406
A21
A22
A264
A26
A414
A120
A602
A10
A130
M40
A413
M4
M26
A228
A281
M23

One dangling thread

Last time we jumped from the christening records aggregated beginning in the sixteenth century to the proliferation of data that is recorded now when a child is born

Latanya Sweeny (a computer scientist from Carnegie Mellon University) presents a simple example of the “information explosion” that took place even in the span from 1925 to 1999

Field name
Child's first name
Child's middle name (sometimes or initial)
Child's last name
Day, month and year of birth
City and/or County of birth (sometimes hospital)
Father's name
Mother's name (including maiden name)
Place of birth (address and town/city)
Mother's age and address
Mother's birthplace (town/city, state, county)
Mother's occupation
Mother, number of previous children
Father's age and address
Father's birthplace (town/city, state, county)
Father's occupation

Figure 4 Typical Set of Birth Certificate Fields, post 1925

Figure 4 contains the typical list of fields available on a birth certificate from most states or counties in the United States, post 1925. This list was provided from the Commonwealth of Massachusetts [2].

Field#	Size	Field name	Field#	Size	Field name
1	1	File Status	31	3	Mother's State of Birth
2	50	Baby's First Name	32	7	Mother's Residence Address
3	50	Baby's Middle Name	33	2	Mother's Residence Direction
4	50	Baby's Last Name	34	20	Residence Street Address
5	1	Baby's Suffix Code	35	10	Residence Type
6	3	Baby's Suffix Text	36	2	Residence Extension
7	8	Baby's Date of Birth	37	10	Residence Apartment #
8	5	Baby's Time of Birth	38	20	Mother's Town of Residence
9	1	AM/PM Indicator	39	1	Mother's Residence in City Limits
10	1	Baby's Sex	40	14	Mother's County of Residence
11	3	Blood Type	41	3	Mother's State of Residence
12	1	Born Here?	42	10	Mother's Residence Zip Code
13	40	Place of Birth	43	38	Mother's Mailing Address
14	1	Facility Type	44	19	Mother's Mailing City
15	20	City of Birth	45	2	Mother's Mailing State
16	20	County of Birth	46	10	Mother's Mailing Zip Code
17	6	Certifier's Code	47	1	Mother Married?
18	30	Certifier's Name	48	50	Father's First Name
19	1	Certifier's Title	49	50	Father's Middle Name
20	30	Attendant's Name	50	50	Father's Last Name
21	1	Attendant's Title	51	1	Father's Suffix Code
22	23	Attendant's Address	52	9	Father's Suffix Text
23	19	Attendant's City	53	9	Father's Social Security Number
24	2	Attendant's State	54	8	Father's Date of Birth
25	10	Attendant's Zip Code	55	3	Father's State of Birth
26	50	Mother's First Name	56	14	Mother's Origin
27	50	Mother's Middle Name	57	14	Mother's Race
28	50	Mother's Last Name	58	2	Mother's Elementary Education
29	9	Mother's Social Security Number	59	2	Mother's College Education
30	8	Mother's Date of Birth	60	11	Mother's Occupation

Figure 5 Typical Set of Electronic Birth Certificate Fields in 1999 -*starting fields 1-60*

Field#	Size	Field name	Field#	Size	Field name
61	11	Mother's Industry	91	1	Alcohol Use During Pregnancy
62	14	Father's Origin	92	3	Number of Drinks/Week
63	14	Father's Race	93	3	Mother's Weight Gain
64	2	Father's Elementary Education	94	1	Release Info For SSN
65	2	Father's College Education	95	6	Operator Code
66	11	Father's Occupation	96	12	Hospital ID
67	11	Father's Industry	97	1	Sent to Romans
68	1	Plurality	98	1	Sent to APORS
69	1	Birth Order	99	16	Other Certifier Specify
70	2	Live Births Still Living	100	12	Temporary Audit Number
71	2	Live Births Now Dead	101	16	Other Facility Specify
72	4	Month/Year Last Live Birth	102	16	Other Attendant Specify
73	2	Number of Terminations	103	1	Mother's Race
74	4	Month/Year Last Termination	104	1	Father's Race
75	1	Baby's Weight Unit	105	2	Mother's Origin
76	5	Baby's Weight	106	2	Father's Origin
77	6	Date of Last Normal Menses	107	1	Attendant Same YN
78	1	Month Prenatal Care Began	108	1	Mailing Address Same YN
79	2	Total Number of Visits	109	1	Capture Father's Info YN
80	2	Apgar Score – 1 Minute	110	2	Mother's Age
81	2	Apgar Score – 5 Minute	111	2	Father's Age
82	2	Estimate of Gestation	112	12	Baby's Hospital Med. Rec.
83	6	Date of Blood Test	113	1	High Risk Pregnancy YN
84	22	Laboratory	114	1	Care Giver (For Chicago)
85	1	Mother Transferred In	115	1	Record Selected For Download
86	30	Facility Mother Transferred From	116	1	Downloaded
87	1	Baby Transferred Out	117	1	Printed
88	30	Facility Baby Transferred To	118	12	Form Number
89	1	Tobacco Use During Pregnancy			**MEDICAL RISK FACTORS**
90	3	Number of Cigarettes/Day	119	1	Anemia
			120	1	Cardiac Disease

Figure 6 Typical Set of Electronic Birth Certificate Fields in 1999 - *continued fields 61-120*

Field#	Size	Field name
121	1	Acute/Chronic Lung Disease
122	1	Diabetes
123	1	Genital Herpes
124	1	Hydramnios/Oligohydramnios
125	1	Hemoglobinopathy
126	1	Hypertension, Chronic
127	1	Hypertension, Preg. Assoc.
128	1	Eclampsia
129	1	Incompetent Cervix
130	1	Previous Infant 4000+ Grams
131	1	Previous Preterm or SGA Infant
132	1	Renal Disease
133	1	Rh Sensitization
134	1	Uterine Bleeding
135	1	No Medical Risk Factors
136	40	Other Medical Risk Factors
		OBSTETRIC PROCEDURES
137	1	Amniocentesis
138	1	Electronic Fetal Monitoring
139	1	Induction of Labor
140	1	Stimulation of Labor
141	1	Tocolysis
142	1	Ultrasound
143	1	No Obstetric Procedures
144	40	Other Obstetric Procedures
		COMPLICATIONS OF LABOR & DELIVERY
145	1	Febrile (>100 or 38C)
146	1	Meconium Moderate, Heavy
147	1	Premature Rupture (>12 Hrs)
148	1	Abruptio Placenta
149	1	Placenta Previa
150	1	Other Excessive Bleeding

Field#	Size	Field name
151	1	Seizures During Labor
152	1	Precipitous Labor (<3 Hrs)
153	1	Prolonged Labor (>20 Hrs)
154	1	Dysfunctional Labor
155	1	Breech/Malpresentation
156	1	Cephalopelvic Disproportion
157	1	Cord Prolapse
158	1	Anesthetic Complications
159	1	Fetal Distress
160	1	No Complications of L&D
161	40	Other Complications of L&D
		METHOD OF DELIVERY
162	1	Vaginal
163	1	Vaginal After Previous C-Section
164	1	Primary C-Section
165	1	Repeat C-Section
166	1	Forceps
167	1	Vacuum
		ABNORMAL CONDITIONS OF NEWBORN
168	1	Anemia
169	1	Birth Injury
170	1	Fetal Alcohol Syndrome
171	1	Hyaline Membrane Disease/RDS
172	1	Meconium Aspiration Syndrome
173	1	Assisted Ventilation <30
174	1	Assisted Ventilation >30
175	1	Seizures
176	1	No Abnormal Conditions of Newborn
177	40	Other Abnormal Condition of Newborn
		CONGENITAL ANOMALIES OF CHILD
178	1	Anencephalus
179	1	Spina Bifida/Meningocele
180	1	Hydrocephalus

Field#	Size	Field name
181	1	Microcephalus
182	40	Other CNS Anomalies
183	1	Heart Malformations
184	40	Other Circ./Resp. Anomalies
185	1	Rectal Atresia/Stenosis
186	1	Tracheo-Esophageal Fistula/Esophageal Atresia
187	1	Omphalocele/Gastroschisis
188	40	Other Gastrointestinal Ano.
189	1	Malformed Genitalia
190	1	Renal Agenesis
191	40	Other Urogenital Anomalies
192	1	Cleft Lip/Palate
193	1	Polydactyly/Syndactyly/Adactyly
194	1	Club Foot
195	1	Diaphragmatic Hernia
196	40	Other Musculoskeletal/Integumental Anomalies
197	1	Down's Syndrome
198	40	Other Chromosomal Anomalies
199	1	No Congenital Anomalies
200	40	Other Congenital Anomalies
		CODE STRIP
201	1	Record Complete YN
202	1	Record Type
203	4	Facility ID
204	4	City of Birth
205	3	County of Birth
206	2	Mother's State of Birth
207	2	Mother's State of Residence
208	4	Mother's Town of Residence
209	3	Mother's County of Residence
210	2	Father's State of Birth

Field#	Size	Field name
211	14	Certifier's License Number
212	6	Laboratory ID Number
213	4	Mother Xfer Code
214	3	Mother Xfer County Code
215	4	Baby Xfer Code
216	3	Baby Xfer County Code
217	4	Year of Birth
218	7	Certificate #
219	1	Unique Code
220	8	File Date
221	2	Community Area
222	4	Census Tract
223	2	Century of Last Live Birth
224	2	Century of Last Termination
225	2	Century of Last Menses
226	2	Century of Blood Test

Figure 8 Typical Set of Electronic Birth Certificate Fields in 1999 -*continued fields 181-226*

A dangling thread

We then looked at another source of data that registers births in the United States -- The Social Security Administration also collects information and has made it available in a relatively simple form

They have released a portion of their data for the years 1880 to 2009 -- In each case, they report the names given to babies born in that year together with a count broken down by gender

For privacy reasons, they only report name-gender pairs associated with five or more infants in a given year -- The data are exported in separate files, each a CSV file (Comma Separated Values)

Social Security

Social Security Numbers For Children

Social Security Numbers For Children

When you have a baby, one of the things that should be on your "to do" list is getting a Social Security number for your baby. The easiest time to do this is when you give information for your child's birth certificate. If you wait to apply for a number at a Social Security office, there may be delays while we verify your child's birth certificate.

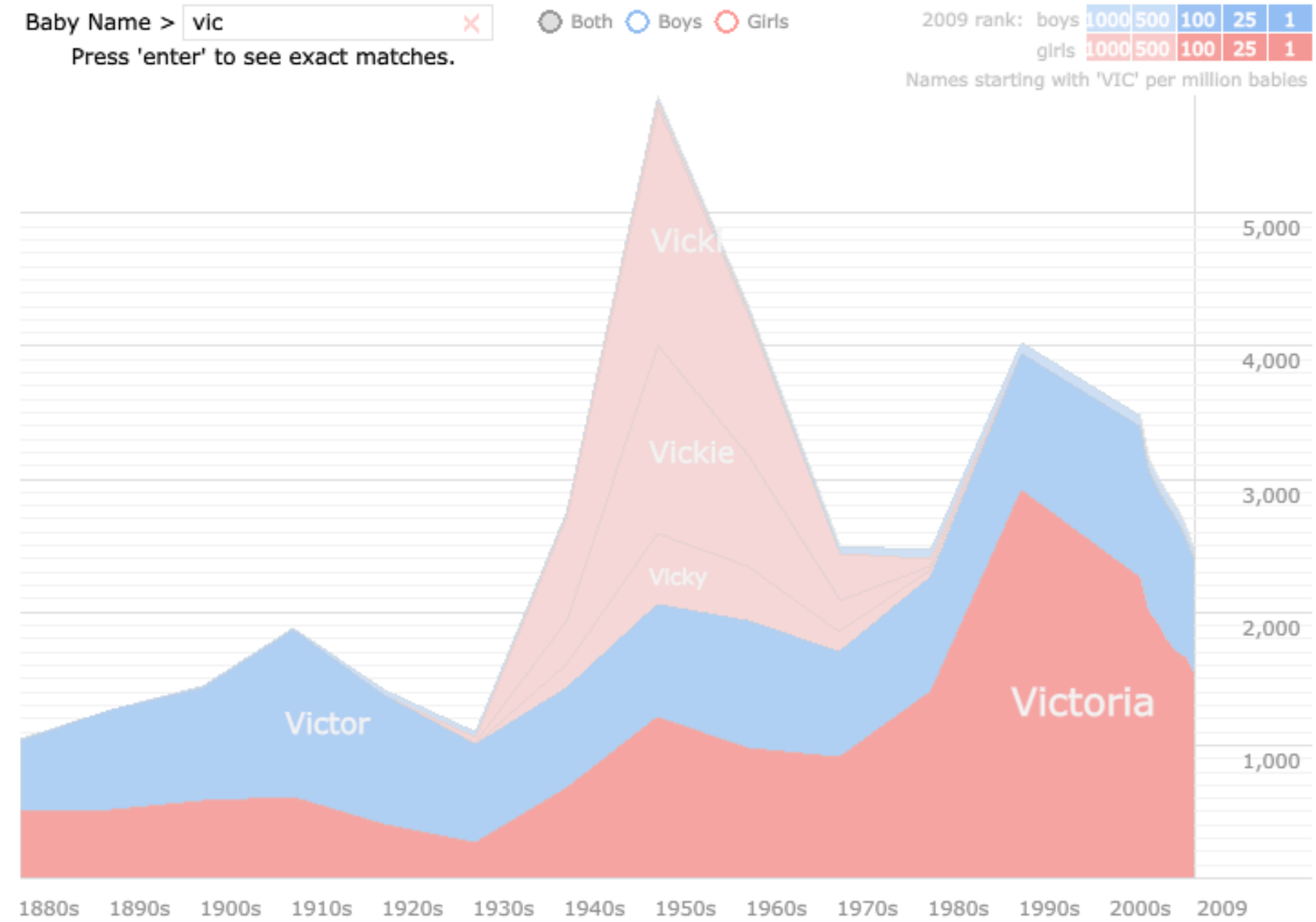
Why should I get a number for my child?

You need a Social Security number to claim your child as a dependent on your income tax return. Your child also may need a number if you plan to:

- Open a bank account for the child;
- Buy savings bonds for the child;
- Obtain medical coverage for the child; or
- Apply for government services for the child.

Must my child have a Social Security number?

No. Getting a Social Security number for your newborn is voluntary. But, it is a good idea to get a number when your child is born. You can apply for a Social Security number for your baby when you apply for your baby's birth

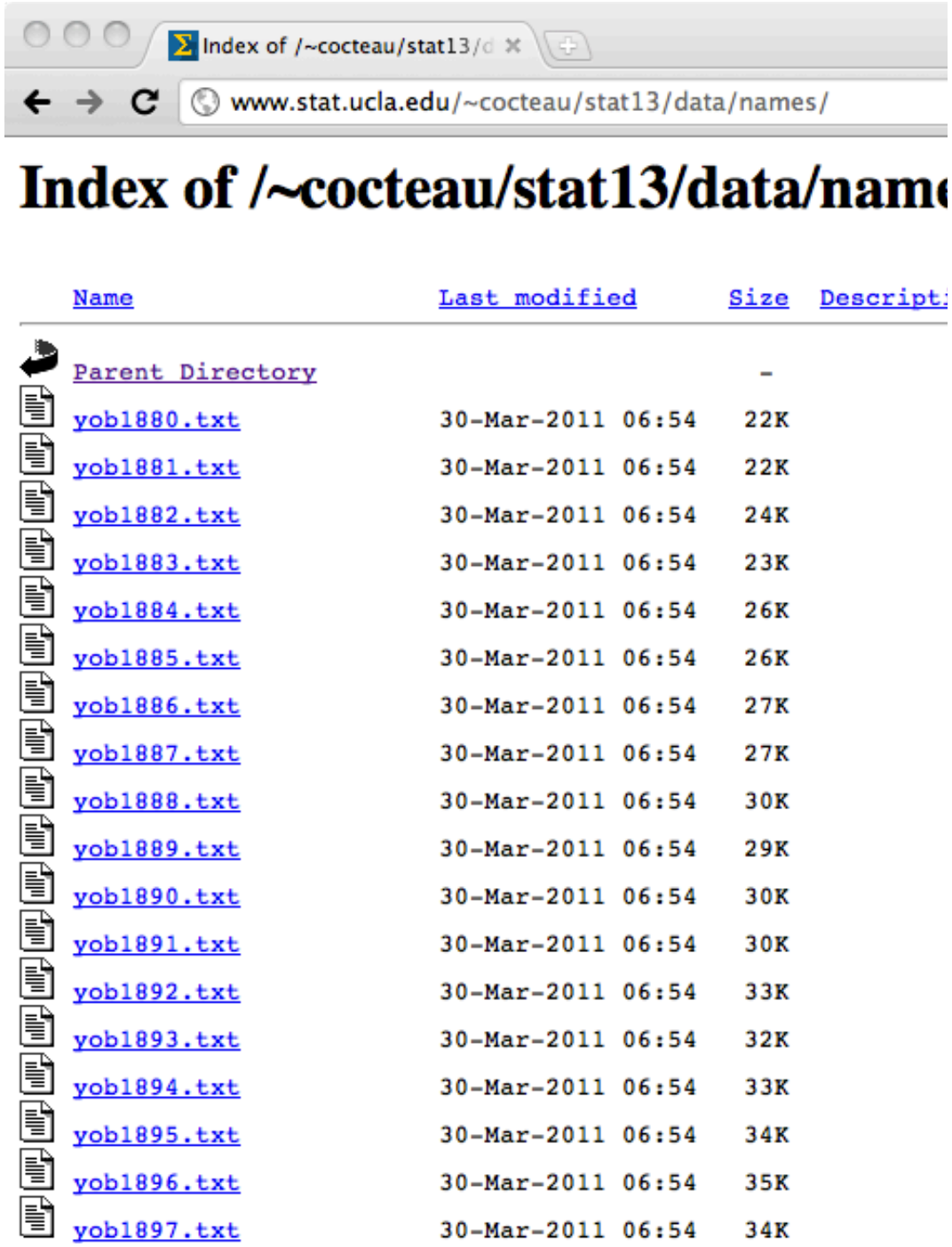

Baby Name > vic
Press 'enter' to see exact matches.
Both
Boys
Girls
2009 rank: boys 1000 500 100 25 1
girls 1000 500 100 25 1
Names starting with 'VIC' per million babies
Vicki
Vickie
Vicky
Victor
Victoria
5,000
4,000
3,000
2,000
1,000
1880s
1890s
1900s
1910s
1920s
1930s
1940s
1950s
1960s
1970s
1980s
1990s
2000s
2009

A dangling thread

I've put the data up on our course web site and you can have a look with your browser or you can read them directly into R

The data in each file is a table where the rows correspond to name-gender pairs together with a count for the year (encoded in the name of the file)

Each row consists of a series of fields (or attributes) separated by, well, commas



A dangling thread

You can read these data into R and make simple yearly comparisons -- For example, you might examine the most popular and least popular names each year

We might also consider the popularity of the most frequent name each year -- That is, are we seeing more unique names and fewer “heavy hitters” that are given to relatively large numbers of babies each year

2009's most frequent

```
Isabella,F,22067
Emma,F,17716
Olivia,F,17246
Sophia,F,16743
Ava,F,15730
Emily,F,15204
Madison,F,15097
Abigail,F,14232
Chloe,F,11785
Mia,F,11319
Elizabeth,F,10879
Addison,F,10567
Alexis,F,9839
Ella,F,9560
Samantha,F,9551
Natalie,F,9324
Grace,F,8194
Lily,F,8016
Alyssa,F,7900
Ashley,F,7741
Sarah,F,7652
Taylor,F,7517
Hannah,F,7482
Brianna,F,7281
Hailey,F,7262
```

```
Jacob,M,20858
Ethan,M,19664
Michael,M,18677
Alexander,M,18025
William,M,17696
Joshua,M,17418
Daniel,M,17336
Jayden,M,17082
Noah,M,17061
Anthony,M,16139
Christopher,M,16136
Aiden,M,15846
Matthew,M,15777
David,M,15236
Andrew,M,14675
Joseph,M,14674
Logan,M,14331
James,M,14022
Ryan,M,12986
Benjamin,M,12944
Elijah,M,12652
Gabriel,M,12648
Christian,M,12498
Nathan,M,11990
Jackson,M,11988
```

... and the least

```
Ziham,F,5
Zikia,F,5
Zimaya,F,5
Zimora,F,5
Zinaya,F,5
Zirah,F,5
Zoii,F,5
Zona,F,5
Zoriyah,F,5
Zowey,F,5
Zuheily,F,5
Zujeily,F,5
Zula,F,5
Zuleimy,F,5
Zuley,F,5
Zuliana,F,5
Zulmy,F,5
Zuriyah,F,5
Zyani,F,5
Zyien,F,5
Zykierra,F,5
Zynaria,F,5
Zynique,F,5
Zyrie,F,5
Zyriel,F,5
```

```
Zekhi,M,5
Zepplin,M,5
Zequan,M,5
Zereon,M,5
Zevion,M,5
Zhen,M,5
Zhyair,M,5
Zien,M,5
Zier,M,5
Zildjian,M,5
Zim,M,5
Zimir,M,5
Ziyun,M,5
Zlatan,M,5
Zoen,M,5
Zubayr,M,5
Zuhaib,M,5
Zykee,M,5
Zykell,M,5
Zylar,M,5
Zyquarius,M,5
Zyran,M,5
Zyreion,M,5
Zyrian,M,5
Zyvion,M,5
```

A dangling thread

Reading the data into R and manipulating things a little, we can look at not just the count but the frequency of name-gender pairs in each year

In 2009 the SSA says there were 2,095,910 boys born (well, SS cardholders born) and 2,001,968 girls -- The top name for a boy that year was Jacob with **a frequency** (count) of 20,858 or **a relative frequency** of 20858/2095910 = 1% while the opt name for a girl was Isabella with a relative frequency of 1.1%

Compare the relative frequencies in 1909 to 2009 on the right -- What do you notice? What other questions might you ask?

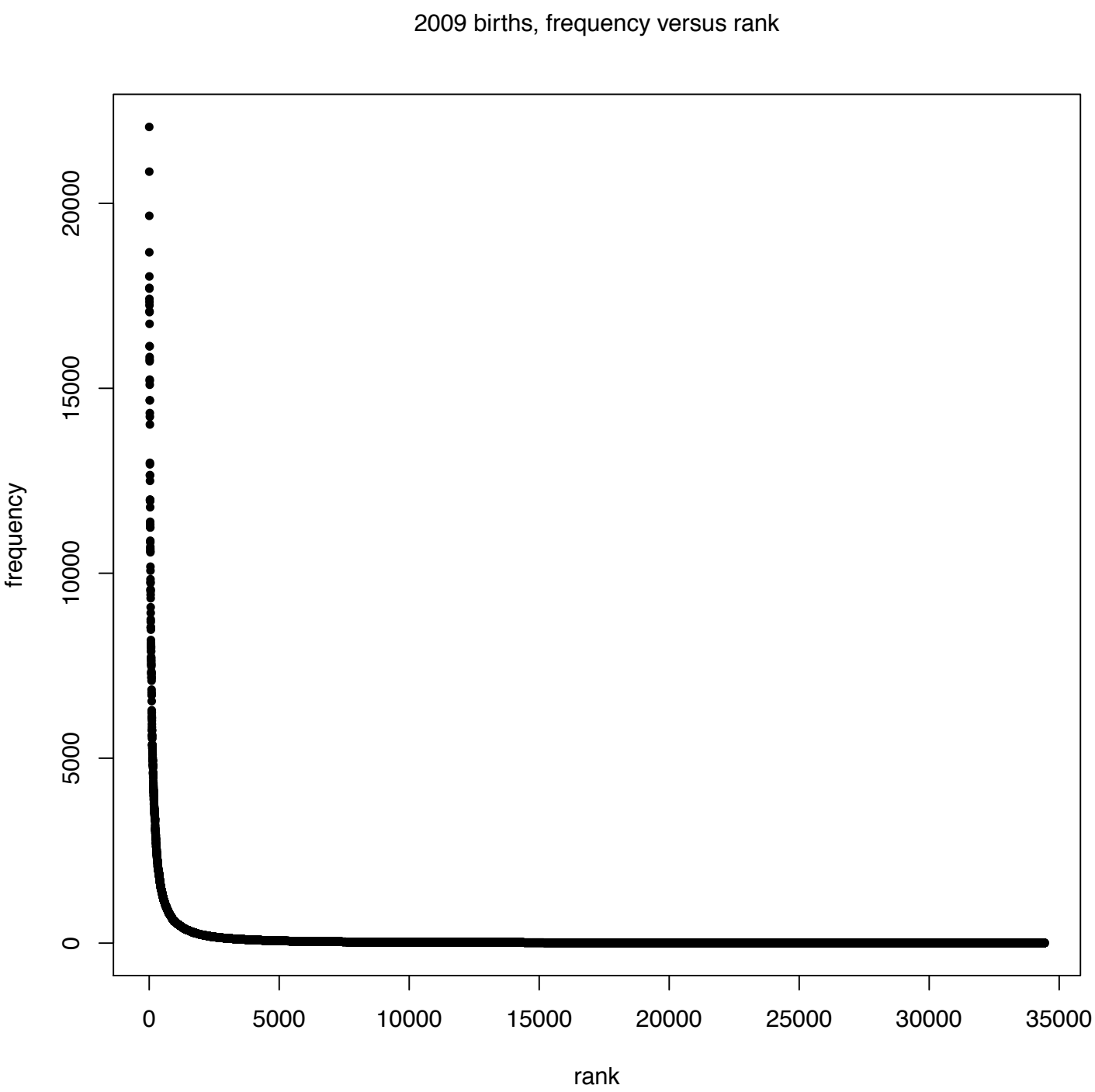
```
> head(boys1909)
        name gender  freq    relfreq
2548    John      M  9591 0.05849169
2549 William      M  7914 0.04826434
2550   James      M  7593 0.04630669
2551  George      M  4687 0.02858415
2552  Robert      M  4565 0.02784012
2553  Joseph      M  4348 0.02651672

> head(girls1909)
       name gender  freq    relfreq
1     Helen      F  9248 0.02820252
2  Margaret      F  7358 0.02243881
3      Ruth      F  6508 0.01984667
4   Dorothy      F  6250 0.01905988
5      Anna      F  5803 0.01769671
6 Elizabeth      F  5175 0.01578158

> head(boys2009)
       name gender  freq     relfreq
2     Jacob      M 20858 0.009951763
3     Ethan      M 19664 0.009382082
4   Michael      M 18677 0.008911165
5 Alexander      M 18025 0.008600083
7   William      M 17696 0.008443111
8    Joshua      M 17418 0.008310471

> head(girls2009)
       name gender  freq     relfreq
1  Isabella      F 22067 0.011022654
6      Emma      F 17716 0.008849292
10   Olivia      F 17246 0.008614523
13   Sophia      F 16743 0.008363271
18      Ava      F 15730 0.007857268
20    Emily      F 15204 0.007594527
```

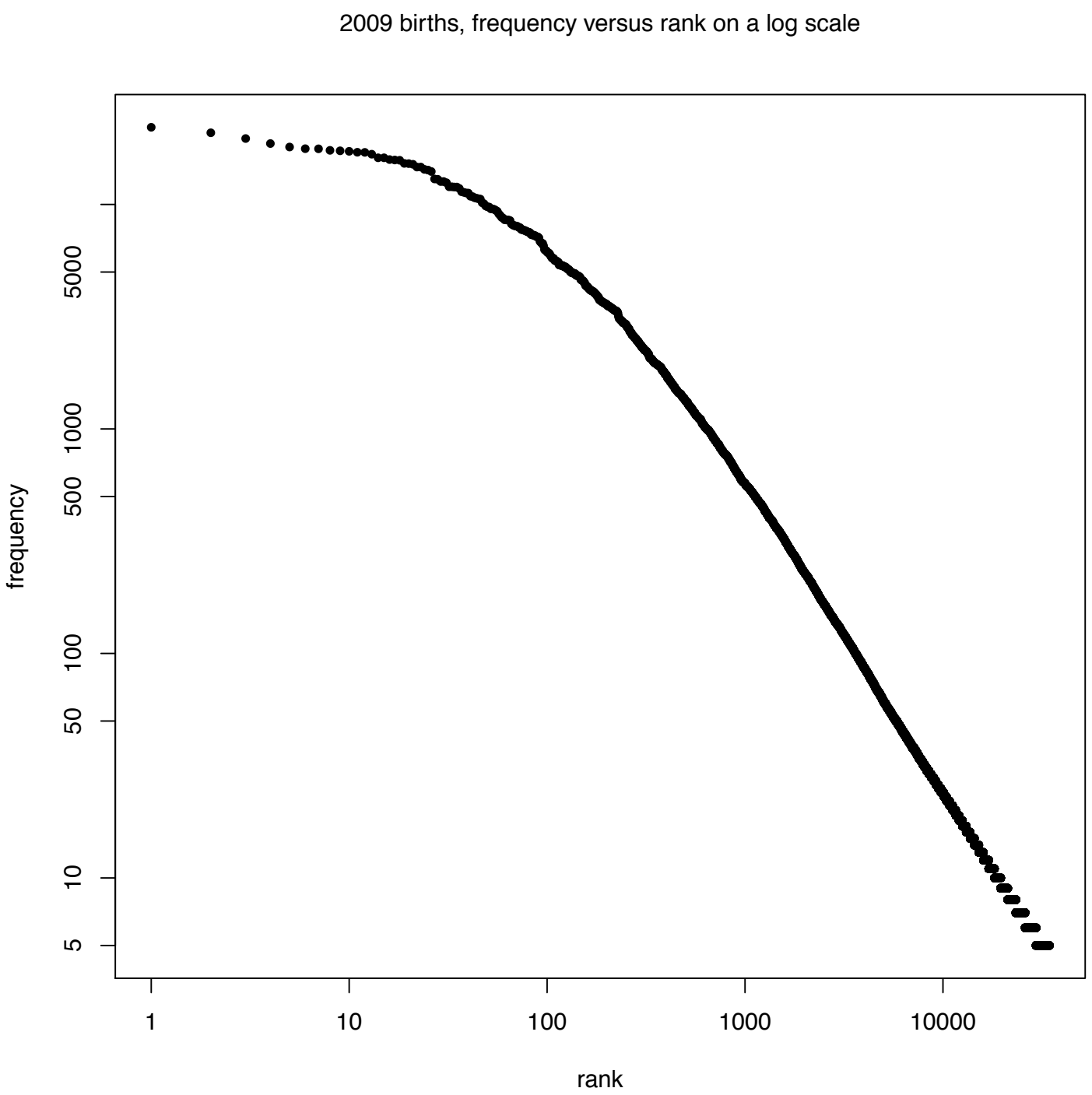
```
# The code below is the sort of thing you'll be able to do at the end of week 2
# in this class but we present it here so that you could see that it's easy to
# have a look at these data if you want

> births1959 <- read.csv(
                 url("http://www.stat.ucla.edu/~cocteau/stat13/data/names/yob1959.txt"),
                     col.names=c("name","gender","count"))

> girls1959 <- births1959[births1959$gender=="F",]
> boys1959 <- births1959[births1959$gender=="M",]

> girls1959$freq <- girls1959$count/2064606
> boys1959$freq <- boys1959$count/2152102

# totals from http://www.ssa.gov/oact/babynames/numberUSbirths.html
```

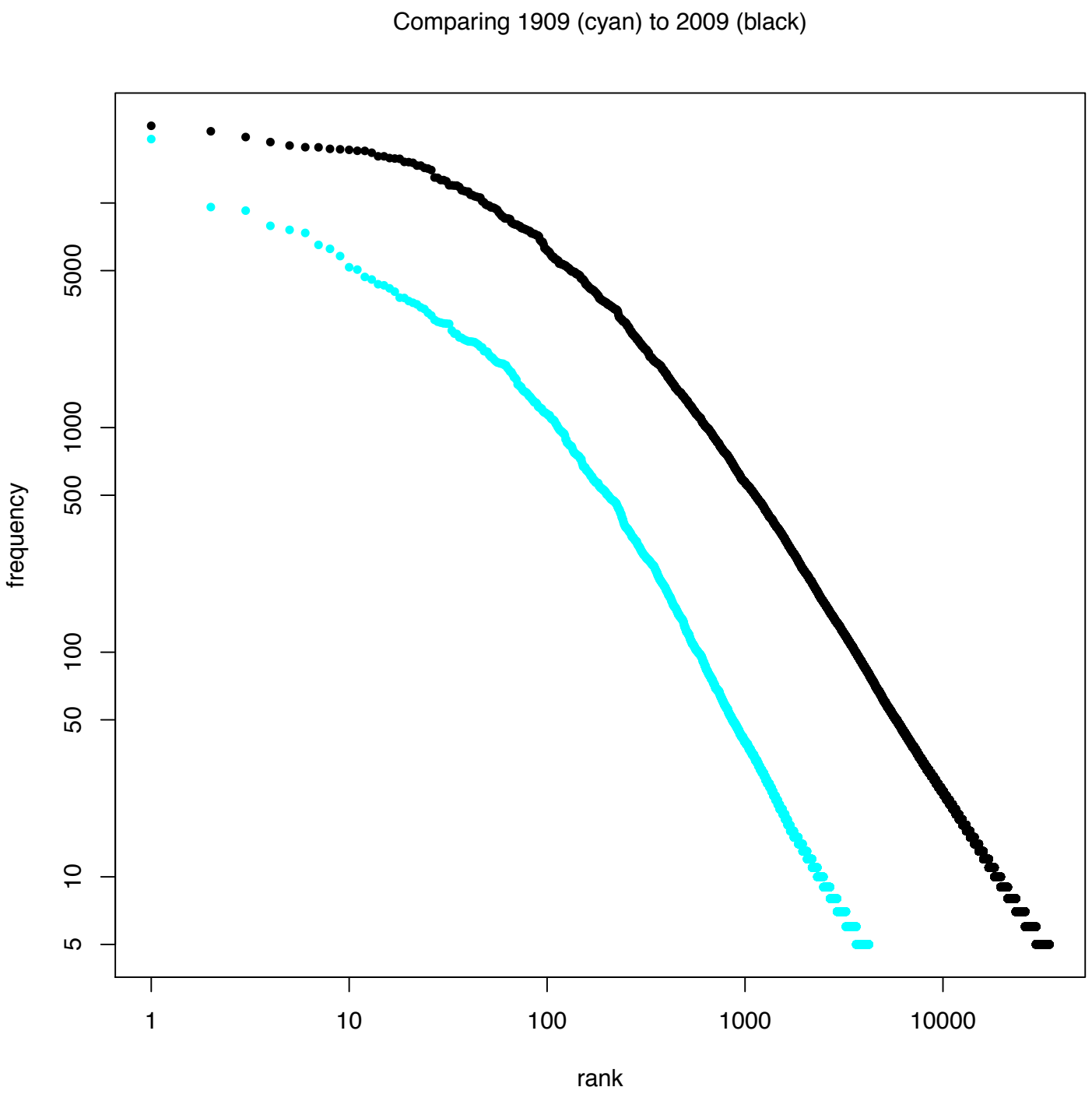
A dangling thread

The rank of a name is simply its position in the list of names sorted according to count -- The name Olivia is ranked 3rd among girls born in 2009 and William is 5th among boys in that year

We can compare the frequencies of names as a function of their rank to tell us something in general about the diversity of names being given to babies and how that is changing over time


2009 births, frequency versus rank
frequency
0
5000
10000
15000
20000
rank
0
5000
10000
15000
20000
25000
30000
35000


2009 births, frequency versus rank on a log scale
frequency
5000
1000
500
100
50
10
5
1
10
100
1000
10000
rank


Comparing 1909 (cyan) to 2009 (black)
frequency
5000
1000
500
100
50
10
5
1
10
100
1000
10000
rank

A comment

The data are distributed in “tabular” form, meaning that the CSV files are organized so that each row corresponds to a different name-gender pair -- Formally, each row consists of a name (a qualitative variable), a gender (again, qualitative) and its frequency (a discrete, quantitative variable)

Now, if you were interested in how your name varied over time, you could have to combine data from across different files and use the file name to encode the year

We could imagine another form for this data, one in which each name-gender pair appeared in rows and the columns were different years, with zeroes padded for names that didn’t occur in particular years (with an associated expansion of the data)

The registrar

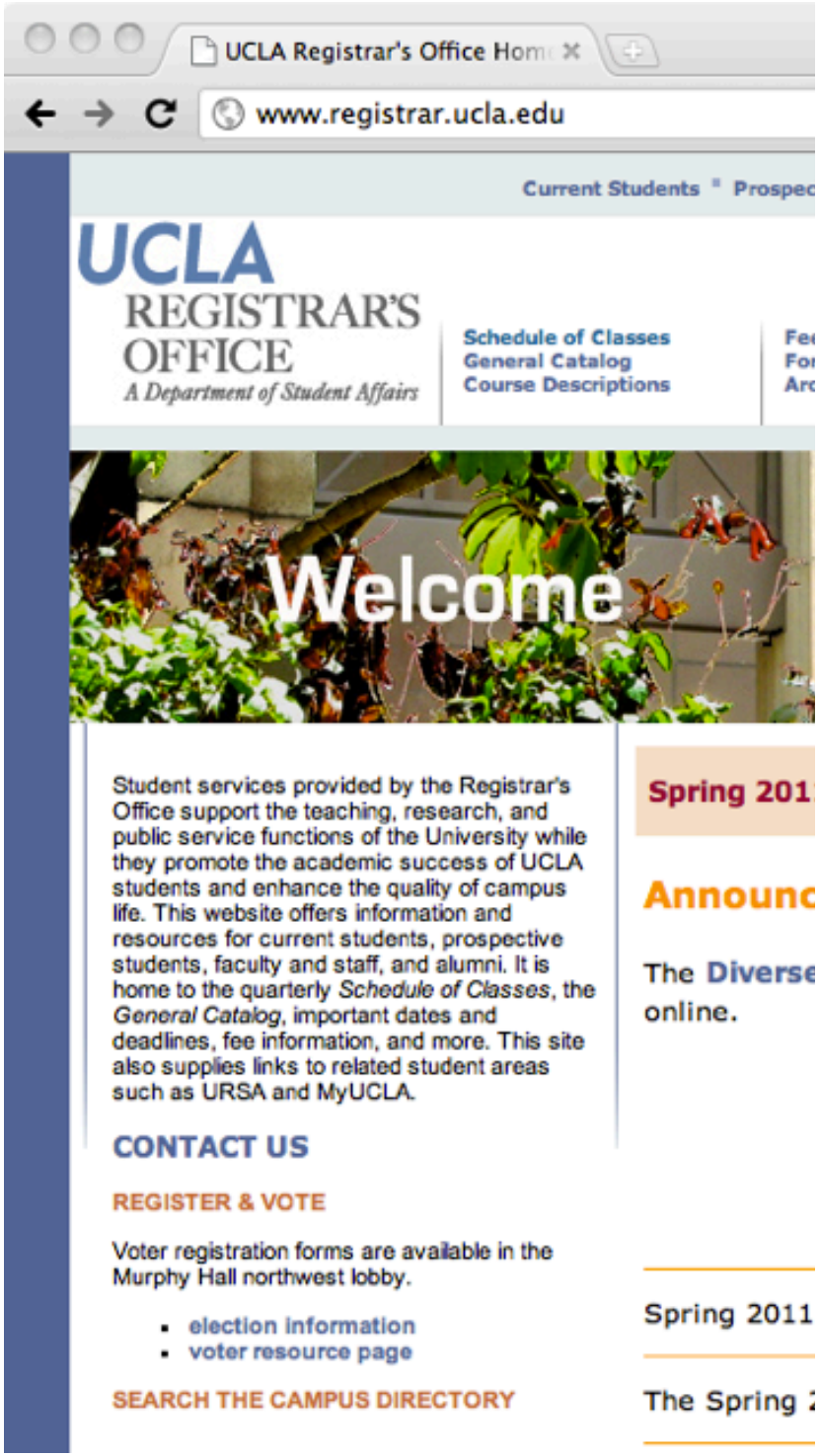
Let's return to the registrar data we introduced at the end of last lecture -- It will help us talk through certain kinds of plots and summaries

As we mentioned last time, the registrar maintains **a record of every class you take**; in addition to what class, it publishes a catalog of when classes meet and how many people were enrolled

On the next page, we present a few lines from a data file we will eventually consider in lab; it is was provided by the registrar (at a cost of $85) and contains the schedules for every student on campus last quarter*

In all, we have 162380 separate rows in this table, each corresponding to a different student and a single class with 31981 total students; What can we learn from these data? And, more importantly, how?

*Note that the identification number in this table is not your student ID, or even part of it, but a random number generated to replace your real ID



Anonymous or not?

There was some concern in the last lecture that the first column of the data set were your student ID's -- After a little reflection, this turned out not to be the case, how can we tell?

Consider, for example, the UID's from our class -- Here are the first three digits of the 170 or so students currently registered, what do you see? Then compare these to the first three digits of the ID's in the data the registrar provided, what do you see?

```
# prefix is an R data set that has all your student id's (don't worry,
# we won't make this public)

> prefix

 "203" "903" "703" "603" "903" "903" "903" "503" "103" "403" "403" "103"
 "903" "003" "803" "003" "203" "203" "503" "803" "403" "403" "703" "303"
 "703" "803" "603" "703" "903" "103" "903" "503" "403" "103" "903" "303"
 "603" "903" "103" "103" "503" "203" "103" "503" "603" "003" "403" "403"
 "803" "803" "503" "603" "803" "203" "103" "203" "803" "403" "303" "603"
 "703" "103" "203" "003" "003" "403" "903" "803" "603" "303" "603" "903"
 "003" "803" "103" "403" "903" "203" "603" "303" "603" "903" "003" "203"
 "003" "003" "803" "403" "003" "103" "903" "003" "503" "803" "703" "103"
 "403" "003" "503" "903" "003" "503" "003" "803" "203" "503" "703" "603"
 "503" "503" "103" "303" "803" "803" "203" "403" "303" "803" "403" "803"
 "303" "303" "603" "303" "603" "103" "203" "703" "203" "903" "703" "803"
 "003" "603" "203" "703" "303" "303" "003" "403" "103" "103" "603" "303"
 "403" "603" "003" "803" "303" "003" "903" "803" "503" "003" "703" "003"
 "403" "003" "403" "103" "003" "203" "003" "503" "703" "303" "403" "903"
 "703" "303" "603" "403" "203" "703" "403" "703" "803" "403" "303" "703"

> table(prefix)
003 103 203 303 403 503 603 703 803 903
 23  17  16  17  22  14  17  16  20  18
```

Anonymous or not?

The prefixes used on our registrar data set are

```
"816" "806" "821" "820" "807" "819" "822" "805"
"818" "106" "817" "815" "867" "835" "869" "837"
"836" "833" "834" "855" "868" "832" "808" "870"
"809" "872" "850" "871" "840" "866" "856" "842"
"857" "854" "853" "852"
```

Clearly, there's no overlap here -- So I think we can safely assume the registrar has given us an anonymous ID (one that is distinct for each student but is not the same as their UID)

Anonymous or not?

Ah but that's just the beginning -- Just because someone can't look up a person directly, does it mean the data cannot identify them uniquely?

For example, while releasing data about a person's gender, ZIP code and birth date would feel anonymous, in combination they can occur infrequently enough to uniquely identify individuals

In fact, Latanya Sweeny (mentioned before) found that 87% of the US population can be uniquely identified using gender, ZIP code and birth date! (Put another way, in the United States you have an 87% chance that your values of these attributes are unique to you)

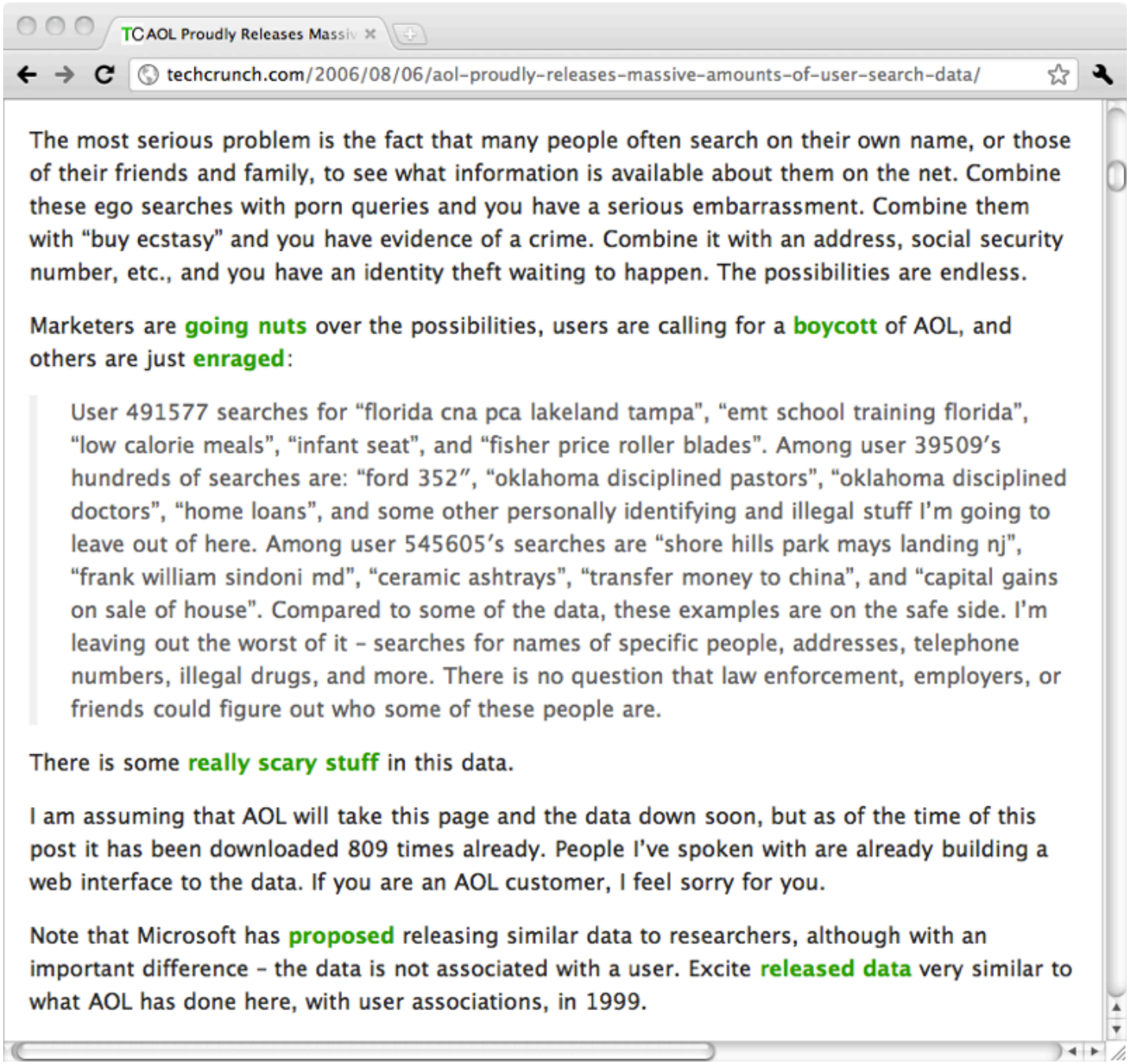
In some smaller contexts (say voting registration for a state), it might take even less to uniquely identify the vast majority of people -- This is an important lesson for the health sciences where data privacy is crucial

AOL’s data release

While on this subject, personally identifiable data can come in all forms -- A couple years ago AOL released 20 M search queries from 650,000 users and dutifully assigned each user a separate ID to distinguish them

Unfortunately, your search records say a lot about you!


AOL Proudly Releases Massiv
techcrunch.com/2006/08/06/aol-proudly-releases-massive-amounts-of-user-search-data/
Tech Gadgets Mobile Enterprise GreenTech CrunchBase TechCrunch TV Disrupt NYC More
TechCrunch
CREDANT
HEALTHCARE S
IS BECOMING INCRI
DOWNLOAD WHITEPAP
What's Hot: Android Apple Facebook Google Groupon Microsoft Twitter Zynga
AOL Proudly Releases Massive Amounts of Private Data
Michael Arrington
Aug 6, 2006
Like 7
Buzz 79
Tweet 2
0 Digg
200 Comments
Yet Another Update: AOL: "This was a screw up"
Further Update: Sometime after 7 pm the download link went down as well, but there is at least one mirror site. AOL is in damage control mode – the fact that they took the data down shows that someone there had the sense to realize how destructive this was, but it is also an admission of wrongdoing of sorts. Either way, the data is now out there for anyone that wants to use (or abuse) it.
Update: Sometime around 7 pm PST on Sunday, the AOL site referred to below was taken down. The direct link to the data is still live. A cached copy of the page is here.
AOL must have missed the uproar over the DOJ's demand for "anonymized" search data last year that caused all sorts of pain for Microsoft and Google. That's the only way to explain their release of data that includes 20 million web queries from 650,000 AOL
AOL



The most serious problem is the fact that many people often search on their own name, or those of their friends and family, to see what information is available about them on the net. Combine these ego searches with porn queries and you have a serious embarrassment. Combine them with "buy ecstasy" and you have evidence of a crime. Combine it with an address, social security number, etc., and you have an identity theft waiting to happen. The possibilities are endless.

Marketers are **going nuts** over the possibilities, users are calling for a **boycott** of AOL, and others are just **enraged**:

> User 491577 searches for "florida cna pca lakeland tampa", "emt school training florida", "low calorie meals", "infant seat", and "fisher price roller blades". Among user 39509's hundreds of searches are: "ford 352", "oklahoma disciplined pastors", "oklahoma disciplined doctors", "home loans", and some other personally identifying and illegal stuff I'm going to leave out of here. Among user 545605's searches are "shore hills park mays landing nj", "frank william sindoni md", "ceramic ashtrays", "transfer money to china", and "capital gains on sale of house". Compared to some of the data, these examples are on the safe side. I'm leaving out the worst of it – searches for names of specific people, addresses, telephone numbers, illegal drugs, and more. There is no question that law enforcement, employers, or friends could figure out who some of these people are.

There is some **really scary stuff** in this data.

I am assuming that AOL will take this page and the data down soon, but as of the time of this post it has been downloaded 809 times already. People I've spoken with are already building a web interface to the data. If you are an AOL customer, I feel sorry for you.

Note that Microsoft has **proposed** releasing similar data to researchers, although with an important difference – the data is not associated with a user. Excite **released data** very similar to what AOL has done here, with user associations, in 1999.

AOL's data release

Reporters at The New York Times were able to use data from these searches to actually find individuals and ask them what they thought about this kind of breach of privacy

Again, privacy is an under-discussed concept in a class of this kind but one that is essential if you find yourself in the medical profession or any research context in which you are dealing with human subjects data

A look at the data

OK back to the registrar! The data are a table where each row corresponds to a student enrolled in a class (we'll call this an enrollment event) -- There are 162,380 registration events for the Winter of 2011

The data we have include the (now infamous) ID, the subject and course number for the class, the building and room number it was held in, the start and end times for the class meeting, the days of the week it met and the level of the enrolled student (undergraduate, graduate and a handful of other designations)

	id	subject	course	building	room	start	end	dow	level
1	816640632	ANTHRO	0009	HAINES	00314	10:00:00	10:50:00	M	U
2	816640632	ANTHRO	0009	FOWLER	A00103B	11:00:00	12:15:00	TR	U
3	816640632	GEOG	0005	HAINES	00039	13:00:00	14:15:00	MW	U
4	816640632	ENGCOMP	0003	HUMANTS	A00046	09:30:00	10:45:00	TR	U
5	816640632	GEOG	0005	BUNCHE	A00170	11:00:00	12:50:00	M	U
6	816643648	MGMT	0403	GOLD	B00313	09:30:00	12:45:00	S	G
7	816643648	MGMT	0405	GOLD	B00313	14:00:00	17:15:00	S	G
8	816577472	COMM ST	0187	PUB AFF	01222	09:30:00	10:45:00	TR	U
9	816577472	COMM ST	0168	ROYCE	00362	17:00:00	19:50:00	M	U
10	816577472	COMM ST	0133	DODD	00175	10:00:00	10:50:00	MWF	U
12	806029941	EDUC	0491	KAUFMAN	00153	17:00:00	19:50:00	W	G
13	806029941	EDUC	0330D	FIELD		08:00:00	14:50:00	MTWRF	G
14	821748664	ANTHRO	0007	HAINES	00039	09:00:00	09:50:00	MWF	U
15	821748664	SPAN	0120	FOWLER	A00139	15:30:00	16:50:00	MW	U
16	821748664	SPAN	0120	HUMANTS	A00046	11:00:00	11:50:00	R	U
17	821748664	WOM STD	0107C M	HAINES	A00025	14:00:00	15:50:00	TR	U
18	821748664	ANTHRO	0007	HAINES	00350	12:00:00	12:50:00	R	U
19	820969784	ENGR	0180	BOELTER	02444	18:00:00	18:50:00	M	U
20	820969784	EL ENGR	0115AL	ENGR IV	18132	12:00:00	15:50:00	T	U
21	820969784	EL ENGR	0115A	ROLFE	01200	08:00:00	09:50:00	MW	U
22	820969784	EL ENGR	0115A	BOELTER	05280	09:00:00	09:50:00	F	U
23	820969784	STATS	0105	PAB	02434	15:00:00	15:50:00	R	U
24	820969784	STATS	0105	FRANZ	02258A	12:00:00	12:50:00	MWF	U
25	820969784	ENGR	0180	BOELTER	02444	16:00:00	17:50:00	MW	U
26	821030697	GEOG	0005	HAINES	00039	13:00:00	14:15:00	MW	U
27	821030697	COMM ST	0185	ROYCE	00362	17:00:00	18:50:00	T	U
28	821030697	SCAND	0180 C	PUB AFF	02214	14:00:00	15:50:00	M	U
29	821030697	SCAND	0180 C	PUB AFF	02214	14:00:00	15:15:00	W	U
30	821030697	GEOG	0005	BUNCHE	A00170	13:00:00	14:50:00	R	U

	id	subject	course	building	room	start	end	dow	level
1	816640632	ANTHRO	0009	HAINES	00314	10:00:00	10:50:00	M	U
2	816640632	ANTHRO	0009	FOWLER	A00103B	11:00:00	12:15:00	TR	U
3	816640632	GEOG	0005	HAINES	00039	13:00:00	14:15:00	MW	U
4	816640632	ENGCOMP	0003	HUMANTS	A00046	09:30:00	10:45:00	TR	U
5	816640632	GEOG	0005	BUNCHE	A00170	11:00:00	12:50:00	M	U
6	816643648	MGMT	0403	GOLD	B00313	09:30:00	12:45:00	S	G
7	816643648	MGMT	0405	GOLD	B00313	14:00:00	17:15:00	S	G
8	**816577472**	**COMM ST**	**0187**	**PUB AFF**	**01222**	**09:30:00**	**10:45:00**	**TR**	**U**
9	**816577472**	**COMM ST**	**0168**	**ROYCE**	**00362**	**17:00:00**	**19:50:00**	**M**	**U**
10	**816577472**	**COMM ST**	**0133**	**DODD**	**00175**	**10:00:00**	**10:50:00**	**MWF**	**U**
12	806029941	EDUC	0491	KAUFMAN	00153	17:00:00	19:50:00	W	G
13	806029941	EDUC	0330D	FIELD		08:00:00	14:50:00	MTWRF	G
14	821748664	ANTHRO	0007	HAINES	00039	09:00:00	09:50:00	MWF	U
15	821748664	SPAN	0120	FOWLER	A00139	15:30:00	16:50:00	MW	U
16	821748664	SPAN	0120	HUMANTS	A00046	11:00:00	11:50:00	R	U
17	821748664	WOM STD	0107C M	HAINES	A00025	14:00:00	15:50:00	TR	U
18	821748664	ANTHRO	0007	HAINES	00350	12:00:00	12:50:00	R	U
19	820969784	ENGR	0180	BOELTER	02444	18:00:00	18:50:00	M	U
20	820969784	EL ENGR	0115AL	ENGR IV	18132	12:00:00	15:50:00	T	U
21	820969784	EL ENGR	0115A	ROLFE	01200	08:00:00	09:50:00	MW	U
22	820969784	EL ENGR	0115A	BOELTER	05280	09:00:00	09:50:00	F	U
23	820969784	STATS	0105	PAB	02434	15:00:00	15:50:00	R	U
24	820969784	STATS	0105	FRANZ	02258A	12:00:00	12:50:00	MWF	U
25	820969784	ENGR	0180	BOELTER	02444	16:00:00	17:50:00	MW	U
26	821030697	GEOG	0005	HAINES	00039	13:00:00	14:15:00	MW	U
27	821030697	COMM ST	0185	ROYCE	00362	17:00:00	18:50:00	T	U
28	821030697	SCAND	0180 C	PUB AFF	02214	14:00:00	15:50:00	M	U
29	821030697	SCAND	0180 C	PUB AFF	02214	14:00:00	15:15:00	W	U
30	821030697	GEOG	0005	BUNCHE	A00170	13:00:00	14:50:00	R	U

	id	subject	course	building	room	start	end	dow	level
1	816640632	ANTHRO	0009	HAINES	00314	10:00:00	10:50:00	M	U
2	816640632	ANTHRO	0009	FOWLER	A00103B	11:00:00	12:15:00	TR	U
3	816640632	GEOG	0005	HAINES	00039	13:00:00	14:15:00	MW	U
4	816640632	ENGCOMP	0003	HUMANTS	A00046	09:30:00	10:45:00	TR	U
5	816640632	GEOG	0005	BUNCHE	A00170	11:00:00	12:50:00	M	U
6	816643648	MGMT	0403	GOLD	B00313	09:30:00	12:45:00	S	G
7	816643648	MGMT	0405	GOLD	B00313	14:00:00	17:15:00	S	G
8	816577472	COMM ST	0187	PUB AFF	01222	09:30:00	10:45:00	TR	U
9	816577472	COMM ST	0168	ROYCE	00362	17:00:00	19:50:00	M	U
10	816577472	COMM ST	0133	DODD	00175	10:00:00	10:50:00	MWF	U
12	806029941	EDUC	0491	KAUFMAN	00153	17:00:00	19:50:00	W	G
13	806029941	EDUC	0330D	FIELD		08:00:00	14:50:00	MTWRF	G
14	821748664	ANTHRO	0007	HAINES	00039	09:00:00	09:50:00	MWF	U
15	821748664	SPAN	0120	FOWLER	A00139	15:30:00	16:50:00	MW	U
16	821748664	SPAN	0120	HUMANTS	A00046	11:00:00	11:50:00	R	U
17	821748664	WOM STD	0107C M	HAINES	A00025	14:00:00	15:50:00	TR	U
18	821748664	ANTHRO	0007	HAINES	00350	12:00:00	12:50:00	R	U
19	**820969784**	**ENGR**	**0180**	**BOELTER**	**02444**	**18:00:00**	**18:50:00**	**M**	**U**
20	**820969784**	**EL ENGR**	**0115AL**	**ENGR IV**	**18132**	**12:00:00**	**15:50:00**	**T**	**U**
21	**820969784**	**EL ENGR**	**0115A**	**ROLFE**	**01200**	**08:00:00**	**09:50:00**	**MW**	**U**
22	**820969784**	**EL ENGR**	**0115A**	**BOELTER**	**05280**	**09:00:00**	**09:50:00**	**F**	**U**
23	**820969784**	**STATS**	**0105**	**PAB**	**02434**	**15:00:00**	**15:50:00**	**R**	**U**
24	**820969784**	**STATS**	**0105**	**FRANZ**	**02258A**	**12:00:00**	**12:50:00**	**MWF**	**U**
25	**820969784**	**ENGR**	**0180**	**BOELTER**	**02444**	**16:00:00**	**17:50:00**	**MW**	**U**
26	821030697	GEOG	0005	HAINES	00039	13:00:00	14:15:00	MW	U
27	821030697	COMM ST	0185	ROYCE	00362	17:00:00	18:50:00	T	U
28	821030697	SCAND	0180 C	PUB AFF	02214	14:00:00	15:50:00	M	U
29	821030697	SCAND	0180 C	PUB AFF	02214	14:00:00	15:15:00	W	U
30	821030697	GEOG	0005	BUNCHE	A00170	13:00:00	14:50:00	R	U

Types of data

A variable is a characteristic of a person or thing that can be assigned a number or a category -- Your book distinguishes between two kinds of variables

Qualitative data “arise when individuals may fall into separate” categories which may not have a numerical relationship (gender and name in the SSA data and course and building for our registrar data) -- Qualitative data may be ordinal in the sense that there is a natural order to the categories

Quantitative data, on the other hand, are numerical, “arising from counts or measurements” -- These data can, in turn, be loosely described as being continuous (able to take any number) or discrete (integers, say, or numerical values with just a small number of unique entries)

We go through the effort of recording these differences because they often **inform the kinds of summaries or displays that are appropriate** for a variable

Frequency graphs

A frequency distribution is a display of the frequency or count of all the values in a sample; it is often tabular or graphical

For categorical variables or discrete variables with a small number of values, this idea makes sense -- For continuous variables we might consider grouping the data in some way

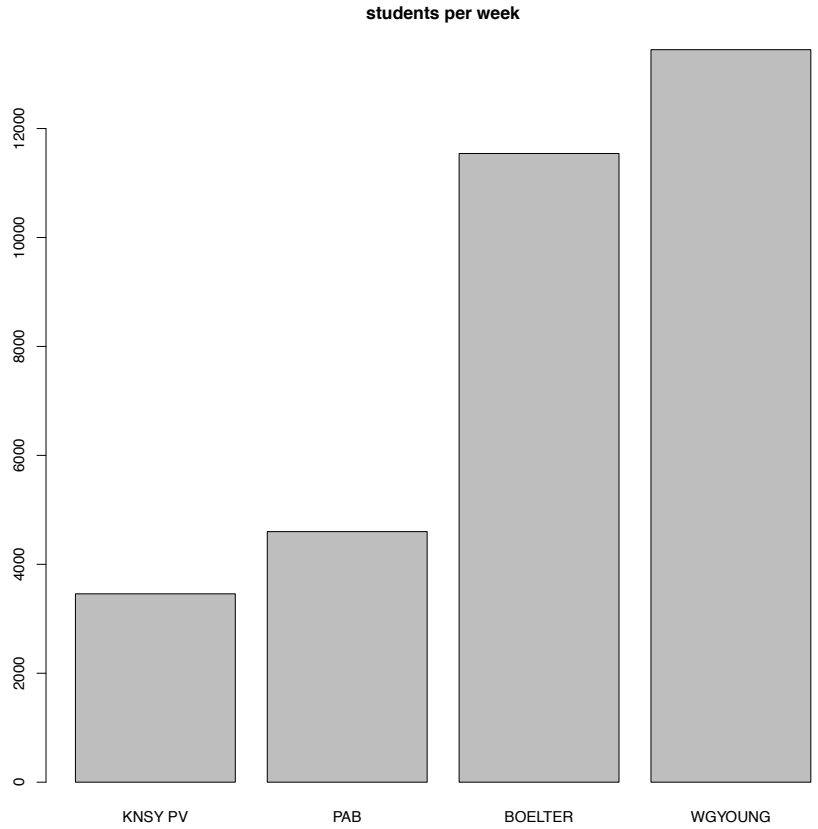
Here is a tabular display exhibiting the frequency of the number of students each week who make use of a selection of buildings

KNSY	PV	PAB	BOELTER	WGYOUNG
3458		4600	11542	13448

(Note that we are simply counting the number of times PAB or BOELTER appear in our data so if a student has two classes in BOELTER, they contribute twice to its count)

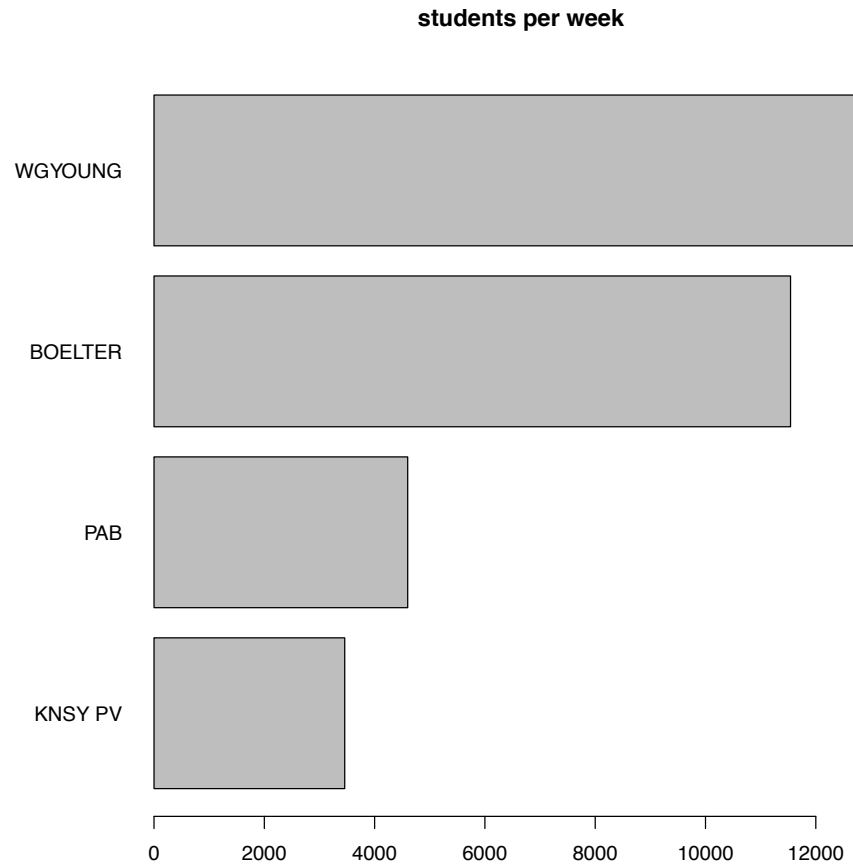
Graphical displays

A **barplot** can be formed to make comparisons easier



Graphical displays

Some have argued that comparisons are better made when the bars run horizontally*



* Cleveland, W. S. (1993), Visualizing Data, Hobart Press

Alternate views

To turn the data from a registrar-centric unit of observation (an enrollment event) into something we might care about, we have to reshape or reformat or process the data

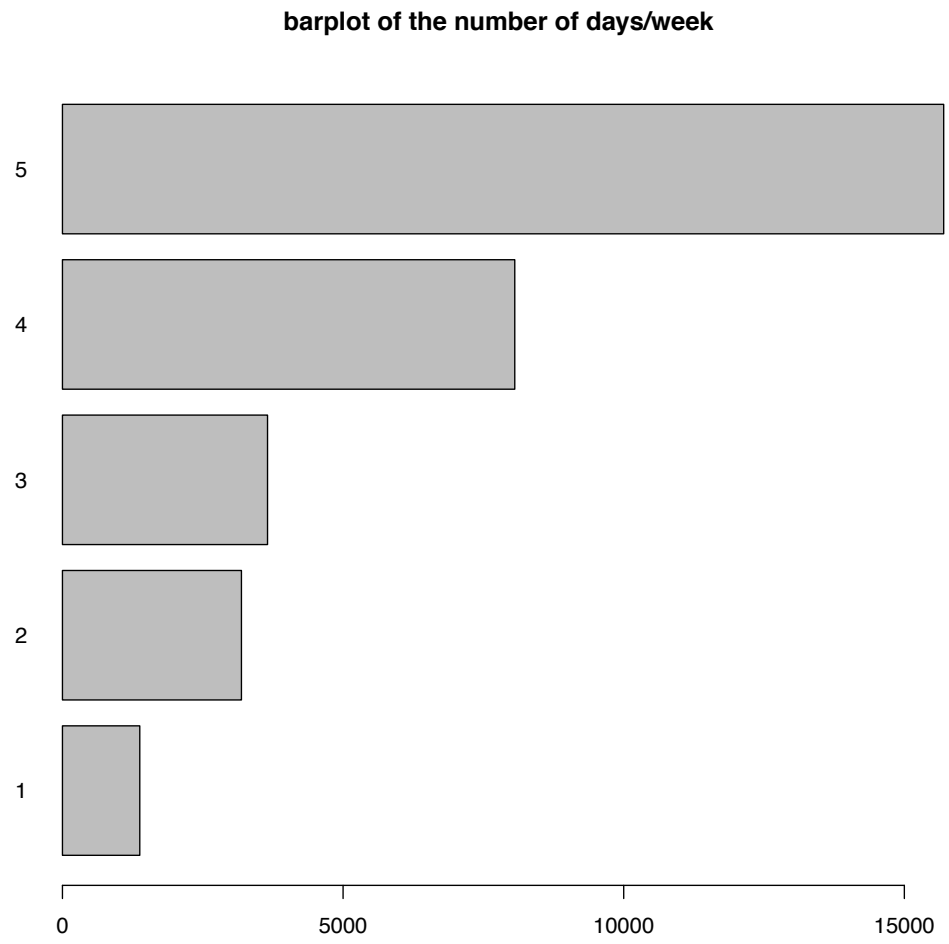
Here, for example, is a table where the basic unit of observation is now a student; for each student, we record their level (graduate or undergraduate), the number of classes they've enrolled in, the number of hours spent in class, the number of days per week they have to be on campus and the time of their earliest class

The new data set has 31,829 rows, each corresponding to a different student...

	id	level	numclasses	numdays	minhour
1	805307175	G	2	1	9
2	805310261	G	3	2	13
3	805314438	G	2	2	18
4	805315160	G	2	2	8
5	805315434	G	1	1	18
6	805355311	G	1	1	12
7	805358888	G	3	4	9
8	805359669	G	5	4	10
9	805364118	G	2	2	9
10	805366583	G	2	2	19
11	805367604	G	2	3	14
12	805370786	G	1	1	13
13	805370964	G	1	1	16
14	805371468	G	2	3	15
15	805371514	G	2	2	13
16	805371734	G	2	3	9
17	805372600	G	2	4	10
18	805372716	G	2	3	13
19	805374592	U	2	2	12
20	805374823	U	4	3	9
21	805376428	G	1	1	12
22	805383993	G	3	4	9
23	805384347	G	2	1	9
24	805384743	G	5	5	8
25	805384808	U	2	4	9
26	805387316	U	2	5	12
27	805413252	G	2	1	17
28	805417569	G	5	3	8
29	805419845	G	2	1	10

Number of days per week

With these new data, we can tabulate the number of days per week students have to be on campus -- This is probably not that surprising to you, but about half of you have to be here 5 days/week



Tables

Here is a two-by-two table (also referred to as a contingency table) that breaks down this total by level (graduate or undergraduate) -- What do you see?

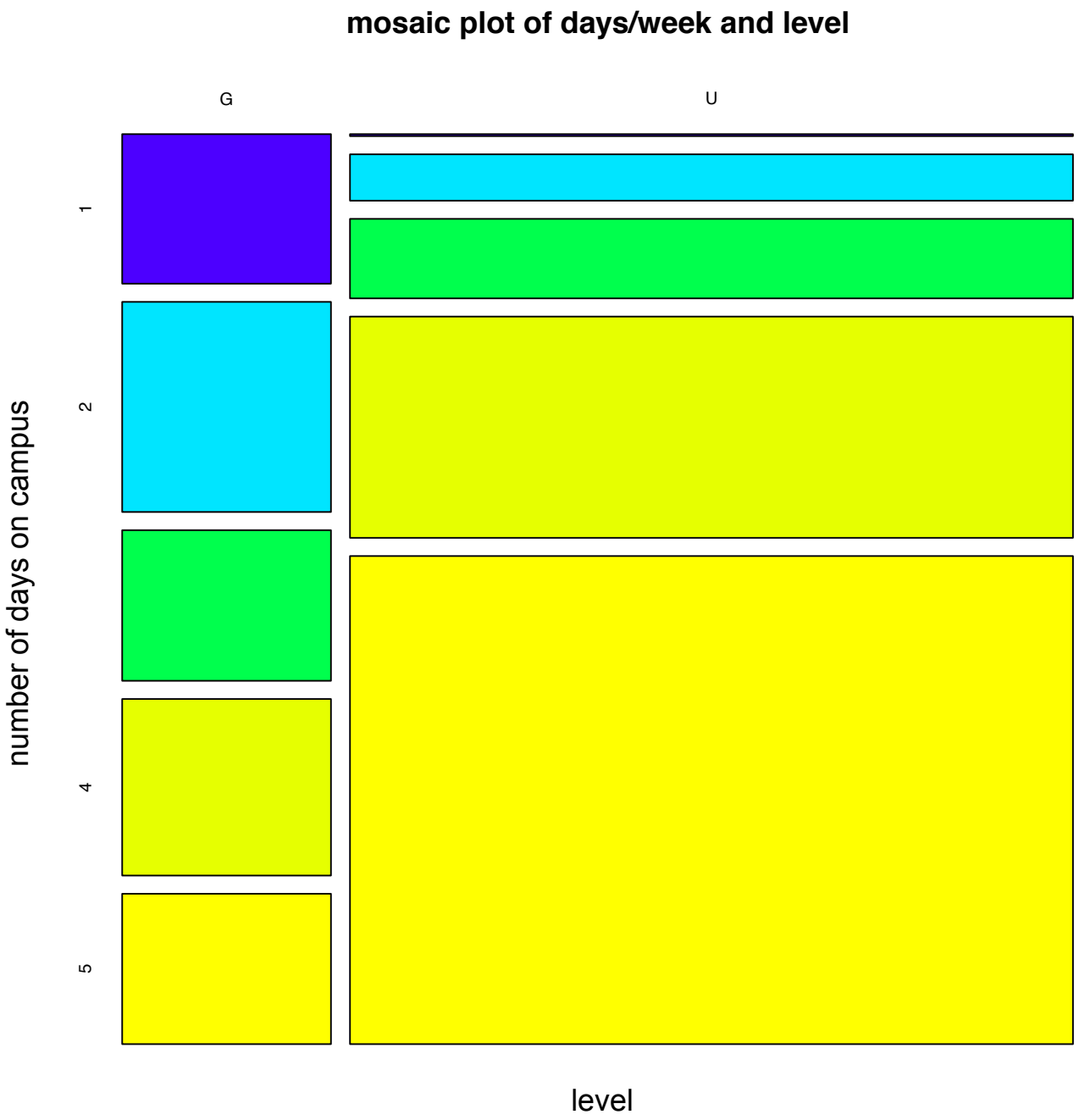
	num days/week				
level	1	2	3	4	5
G	1274	1790	1283	1504	1281
U	59	1367	2346	6527	14398

Mosaic plots

These displays represent the counts in a contingency table by tiles whose size (area) is proportional to the cell count or frequency

It is also possible to extend these displays to tabulations with more than two variables -- How might this work?

* Hartigan, J.A., and Kleiner, B. (1984) A mosaic of television ratings. *The American Statistician*, 38, 32-35


mosaic plot of days/week and level
G
U
1
2
4
5
number of days on campus
level

Mosaic plots

As a kind of multidimensional barplot, the mosaic plot makes the height and width of the boxes proportional to the counts in the corresponding cell of the table -- It does this in two passes

In our case, the width of the boxes are chosen according to the total number of undergraduate and graduate students in our data set -- About 20% are graduates

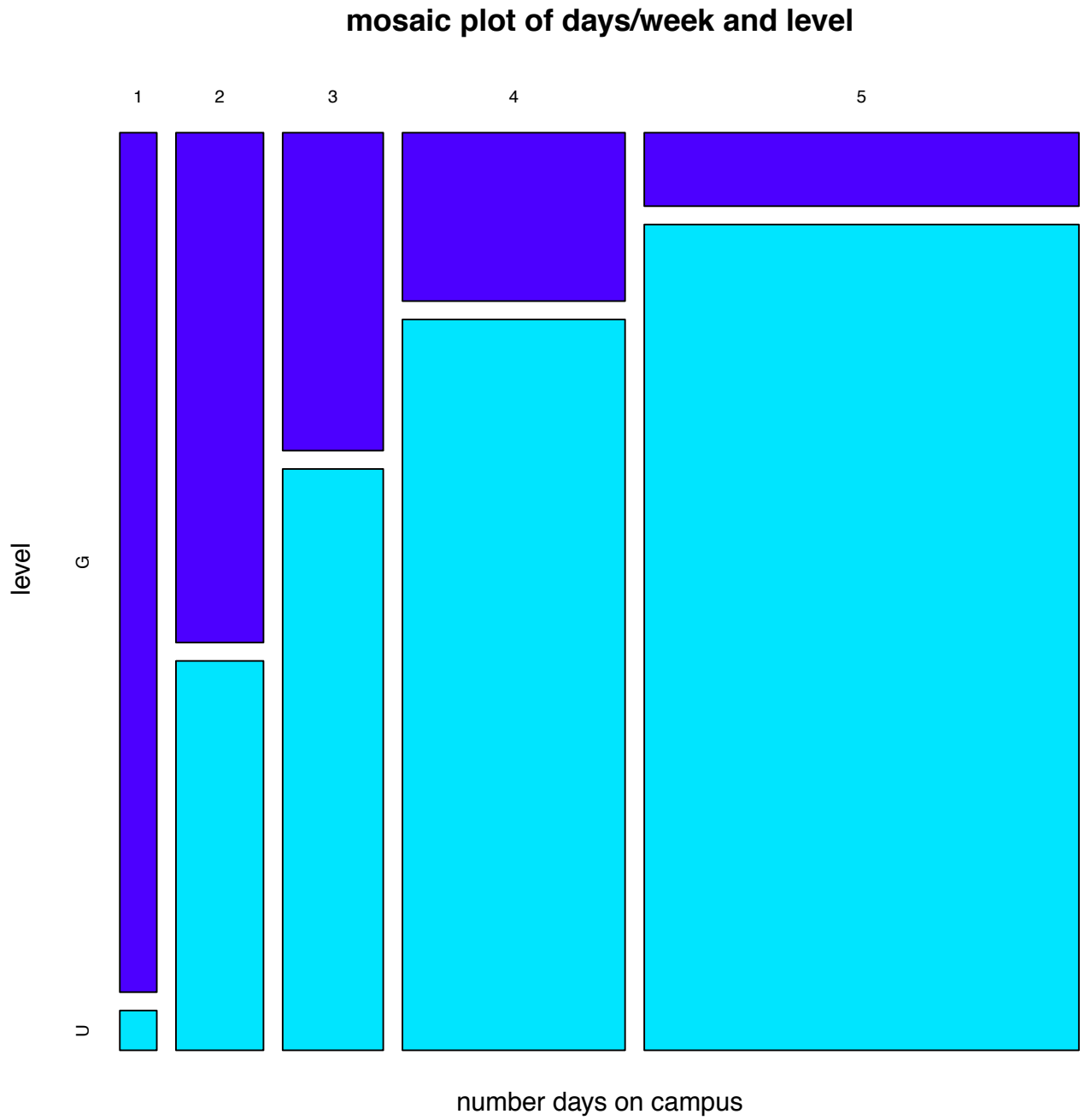
	num days/week					
level	1	2	3	4	5	
G	1274	1790	1283	1504	1281	7132
U	59	1367	2346	6527	14398	24697
	1333	3157	3629	8031	15679	

Then, within each category (U or G), we adjust the height of each box so that it is proportional to the count of student (U or G) who have to be on campus 1, 2, ... days a week

Mosaic plots

Given this two-pass construction idea, there are actually two mosaic plots that can be formed from a table -- On the next page we use the “transpose” of our original table

	level	
num days/week	G	U
1	1274	59
2	1790	1367
3	1283	2346
4	1504	6527
5	1281	14398


mosaic plot of days/week and level
1
2
3
4
5
level
G
U
number days on campus

New data from old

The start and end times can be used to derive a new variable, the total minutes, for each registration event -- At a practical level, the H:M:S format means we need to convert the times to minutes since midnight, say, and then subtract

We can then aggregate data across students to give us the number of minutes each student on campus spent in class last quarter -- On the next page we have a summary of these numbers known as a histogram

Time

One trailing idea we should return to over the course of the quarter is the representation of time on the computer -- the H:M:S descriptor is fine but it's hard to work with when we start to involve dates and times

What we start to want is an origin against which we can track the passage of time -- So for our example from the registrar, we might imagine Sunday night at midnight representing our 0 and then count the number of seconds past that

This would allow us to make a time grid on which we can place classes and so on -- It turns out that this strategy is followed in modern computer systems

Seconds since the UNIX epoch (the number of seconds since January 1, 1970) is a common count of time (today at 2:31 UNIX time was 1301520761

	id	subject	course	building	room	start	end	dow	level	tottime
1	816640632	ANTHRO	0009	HAINES	00314	10:00:00	10:50:00	M	U	50
2	816640632	ANTHRO	0009	FOWLER	A00103B	11:00:00	12:15:00	TR	U	75
3	816640632	GEOG	0005	HAINES	00039	13:00:00	14:15:00	MW	U	75
4	816640632	ENGCOMP	0003	HUMANTS	A00046	09:30:00	10:45:00	TR	U	75
5	816640632	GEOG	0005	BUNCHE	A00170	11:00:00	12:50:00	M	U	110
6	816643648	MGMT	0403	GOLD	B00313	09:30:00	12:45:00	S	G	195
7	816643648	MGMT	0405	GOLD	B00313	14:00:00	17:15:00	S	G	195
8	816577472	COMM ST	0187	PUB AFF	01222	09:30:00	10:45:00	TR	U	75
9	816577472	COMM ST	0168	ROYCE	00362	17:00:00	19:50:00	M	U	170
10	816577472	COMM ST	0133	DODD	00175	10:00:00	10:50:00	MWF	U	50
12	806029941	EDUC	0491	KAUFMAN	00153	17:00:00	19:50:00	W	G	170
13	806029941	EDUC	0330D	FIELD		08:00:00	14:50:00	MTWRF	G	410
14	821748664	ANTHRO	0007	HAINES	00039	09:00:00	09:50:00	MWF	U	50
15	821748664	SPAN	0120	FOWLER	A00139	15:30:00	16:50:00	MW	U	80
16	821748664	SPAN	0120	HUMANTS	A00046	11:00:00	11:50:00	R	U	50
17	821748664	WOM STD	0107C M	HAINES	A00025	14:00:00	15:50:00	TR	U	110
18	821748664	ANTHRO	0007	HAINES	00350	12:00:00	12:50:00	R	U	50
19	820969784	ENGR	0180	BOELTER	02444	18:00:00	18:50:00	M	U	50
20	820969784	EL ENGR	0115AL	ENGR IV	18132	12:00:00	15:50:00	T	U	230
21	820969784	EL ENGR	0115A	ROLFE	01200	08:00:00	09:50:00	MW	U	110
22	820969784	EL ENGR	0115A	BOELTER	05280	09:00:00	09:50:00	F	U	50
23	820969784	STATS	0105	PAB	02434	15:00:00	15:50:00	R	U	50
24	820969784	STATS	0105	FRANZ	02258A	12:00:00	12:50:00	MWF	U	50
25	820969784	ENGR	0180	BOELTER	02444	16:00:00	17:50:00	MW	U	110
26	821030697	GEOG	0005	HAINES	00039	13:00:00	14:15:00	MW	U	75
27	821030697	COMM ST	0185	ROYCE	00362	17:00:00	18:50:00	T	U	110
28	821030697	SCAND	0180 C	PUB AFF	02214	14:00:00	15:50:00	M	U	110
29	821030697	SCAND	0180 C	PUB AFF	02214	14:00:00	15:15:00	W	U	75
30	821030697	GEOG	0005	BUNCHE	A00170	13:00:00	14:50:00	R	U	110

	id	level	numclasses	numdays	minhour	tottime
1	805307175	G	2	1	9	280
2	805310261	G	3	2	13	600
3	805314438	G	2	2	18	365
4	805315160	G	2	2	8	340
5	805315434	G	1	1	18	230
6	805355311	G	1	1	12	50
7	805358888	G	3	4	9	430
8	805359669	G	5	4	10	600
9	805364118	G	2	2	9	340
10	805366583	G	2	2	19	340
11	805367604	G	2	3	14	245
12	805370786	G	1	1	13	170
13	805370964	G	1	1	16	170
14	805371468	G	2	3	15	125
15	805371514	G	2	2	13	400
16	805371734	G	2	3	9	295
17	805372600	G	2	4	10	220
18	805372716	G	2	3	13	220
19	805374592	U	2	2	12	225
20	805374823	U	4	3	9	410
21	805376428	G	1	1	12	50
22	805383993	G	3	4	9	450
23	805384347	G	2	1	9	160
24	805384743	G	5	5	8	1210
25	805384808	U	2	4	9	235
26	805387316	U	2	5	12	250
27	805413252	G	2	1	17	580
28	805417569	G	5	3	8	1000
29	805419845	G	2	1	10	130

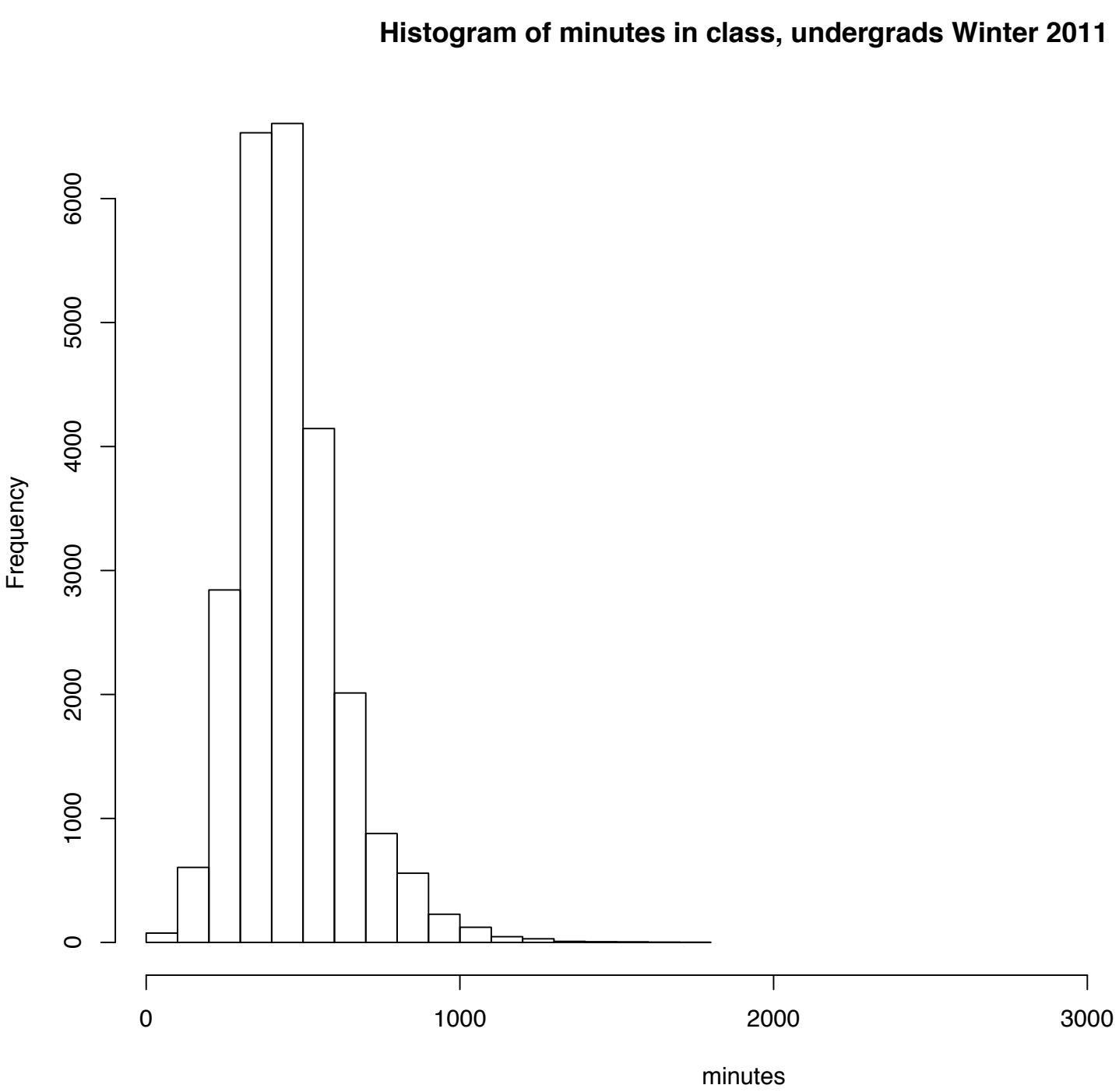
Histograms

A histogram groups or bins the data and, like a barplot, presents the number of data points that fall into each group

This display involves a “tuning parameter” -- That is, we are free to choose how many bins we want to make the display (of course statisticians have thought for decades about how to choose this parameter in a more automatic fashion)

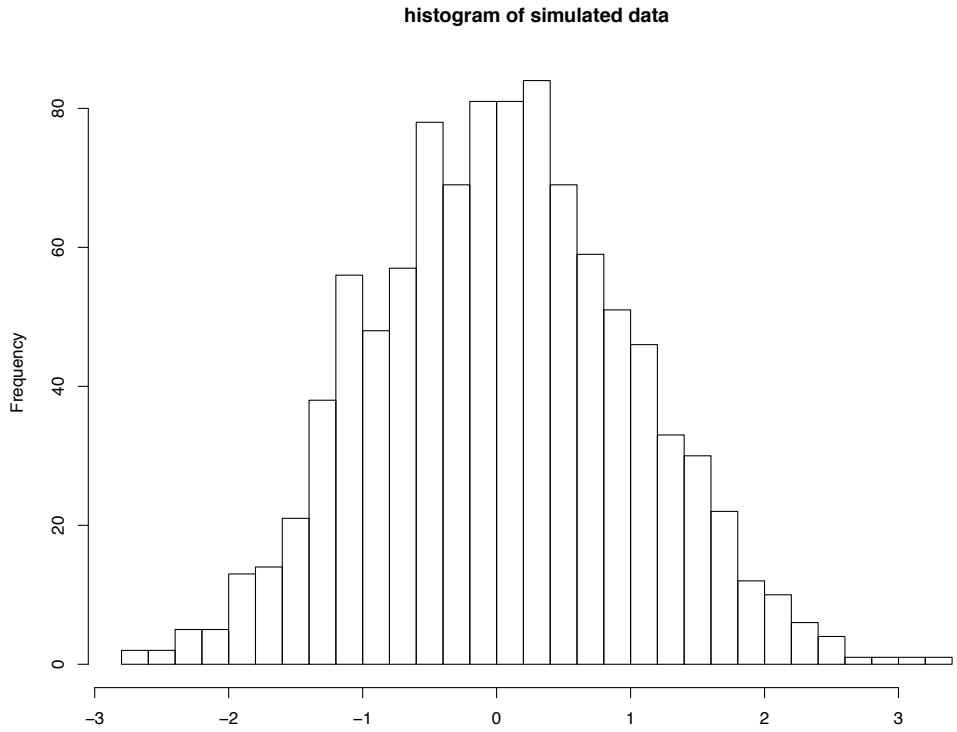
In situations like this, it is always good to vary the number of bins and examine the plot for any structure that emerges; in so doing, we want to get a sense of the “shape” of the data

What do we see?

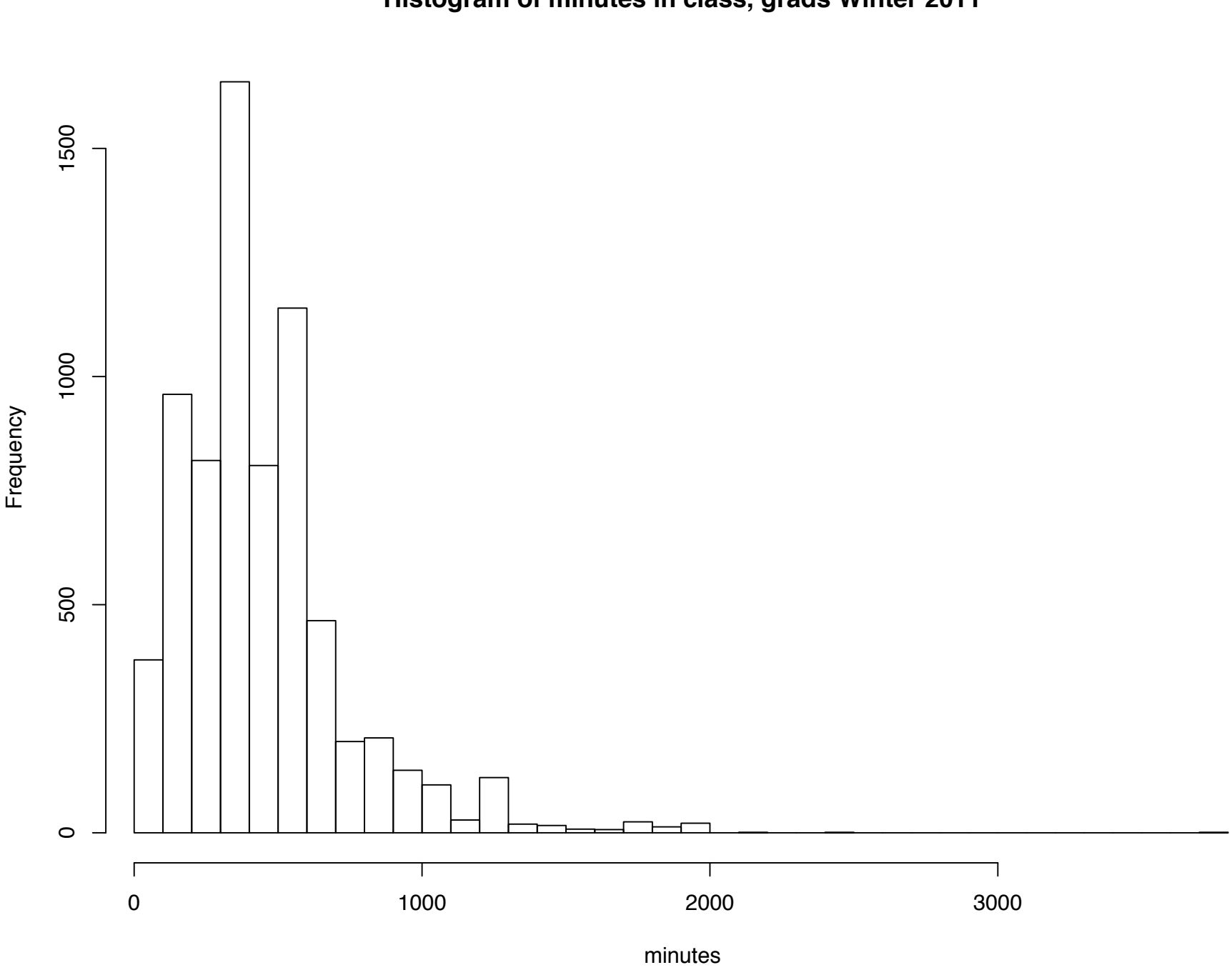

Histogram of minutes in class, undergrads Winter 2011
Frequency
0
1000
2000
3000
4000
5000
6000
0
1000
2000
3000
minutes

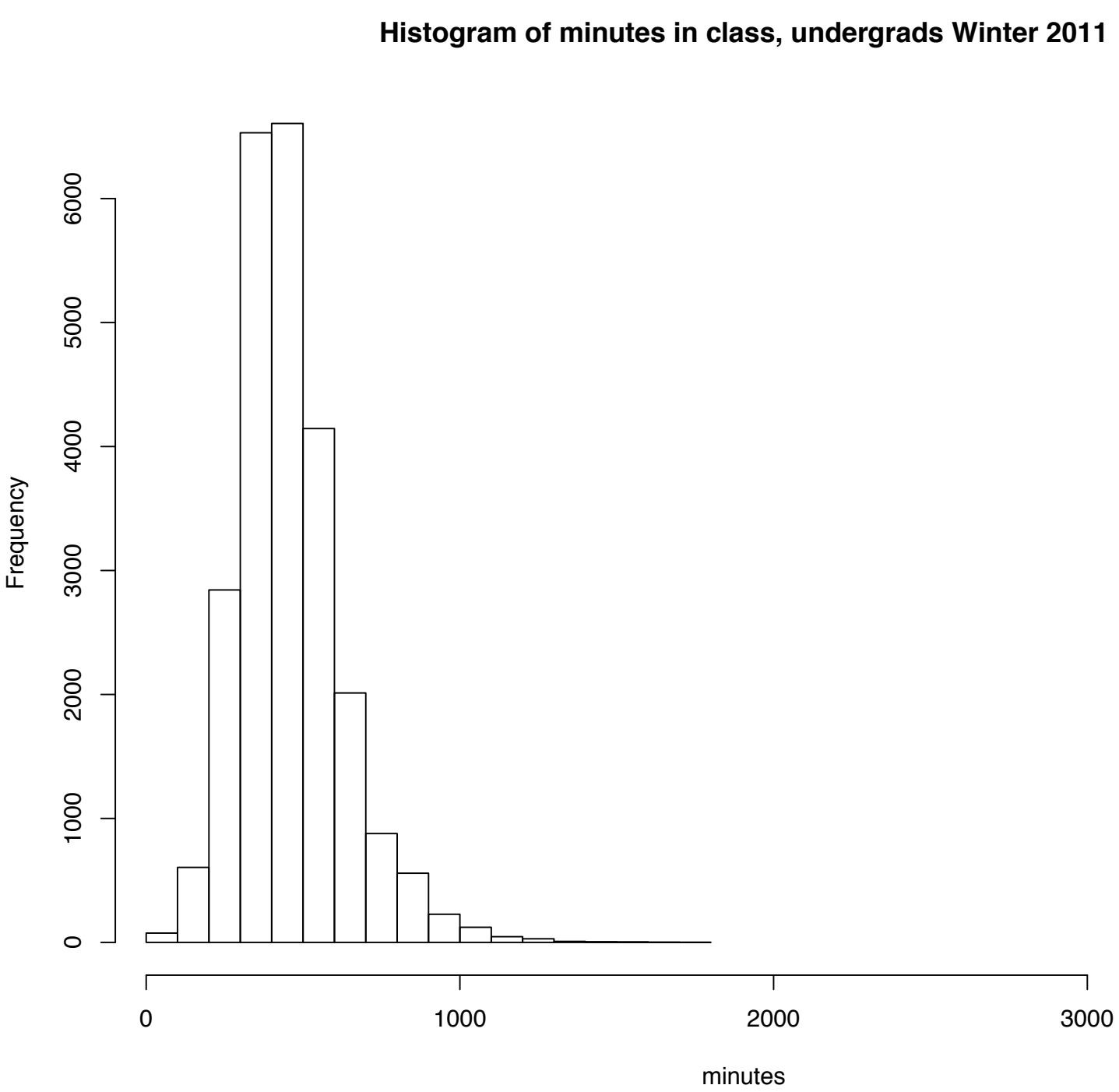
A taxonomy of shapes

Over the course of your lab and future lectures, we'll start to build up a vocabulary describing the shapes of distributions -- We refer to a distribution as **symmetric** if you can fold the histogram in half and have roughly the same shape on either side



In the case of our total time spent in class, the distribution is said to be **skewed to the right**, because there are too many large values than we might expect -- When the distribution is pushed in the other direction we say that it is **skewed to the left**


Histogram of minutes in class, grads Winter 2011
Frequency
0
500
1000
1500
0
1000
2000
3000
minutes


Histogram of minutes in class, undergrads Winter 2011
Frequency
0
1000
2000
3000
4000
5000
6000
0
1000
2000
3000
minutes

Comparing distributions

Comparing the hours in class between graduate and undergraduate students, what do you notice? How could we make these comparisons easier?

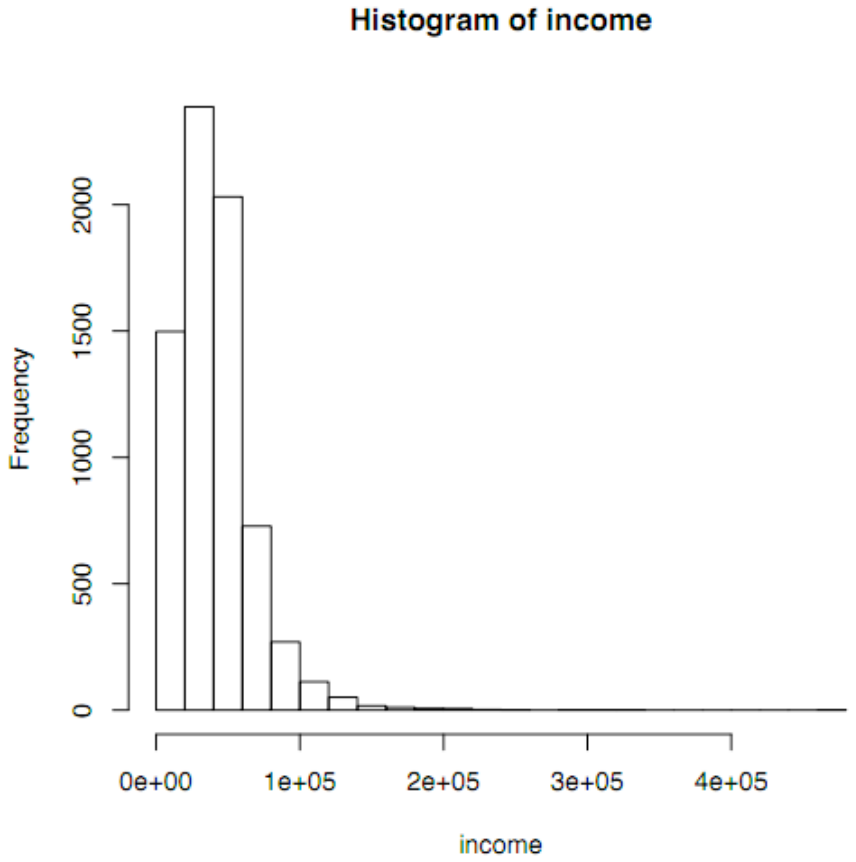
In lab you will look at a smoothed histogram, a density plot, that allows you to overlay two histograms more easily (we can do it here but there's a lot of occlusion)

Taxonomy of shapes: Multiple modes

For the 1983 Family Expenditure Survey in the UK, a sample of n=7,125 households reported their annual income*

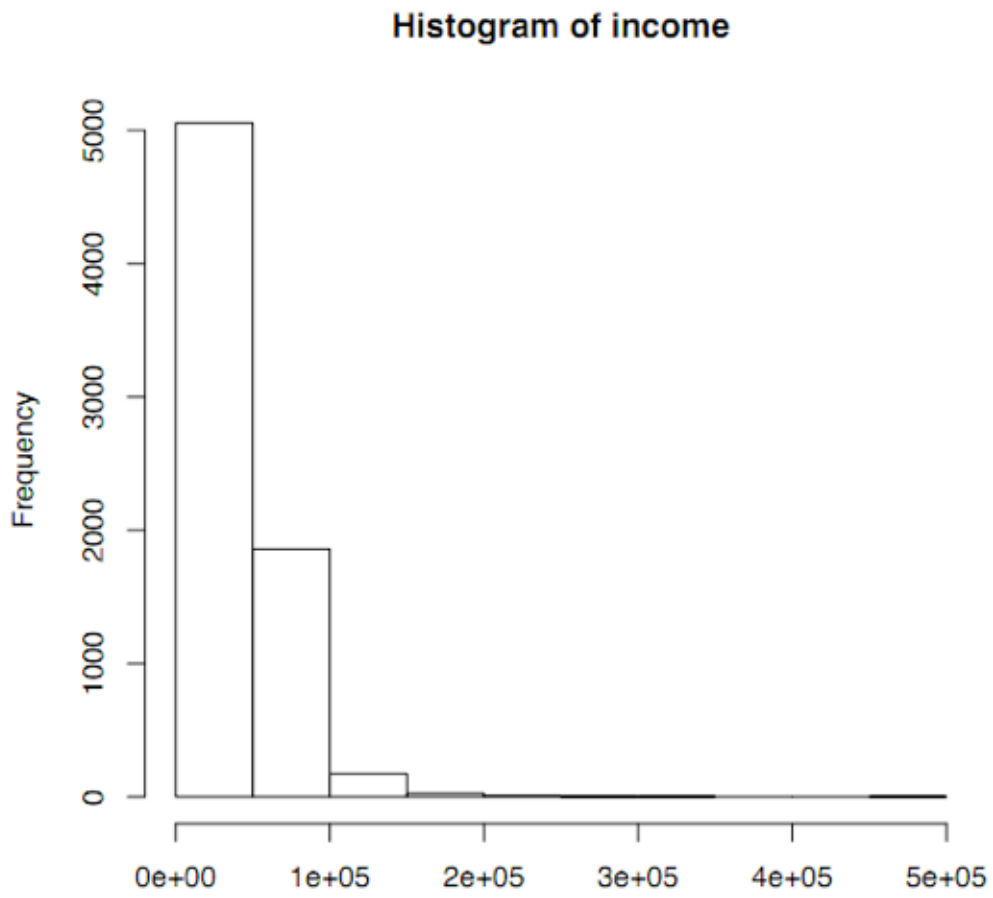
The minimum of these data is 529 (~$850), the maximum 472,821 (~$760K)

The mean of these data is 41,262 (roughly $66K) and the median is 37,520 ($60K)

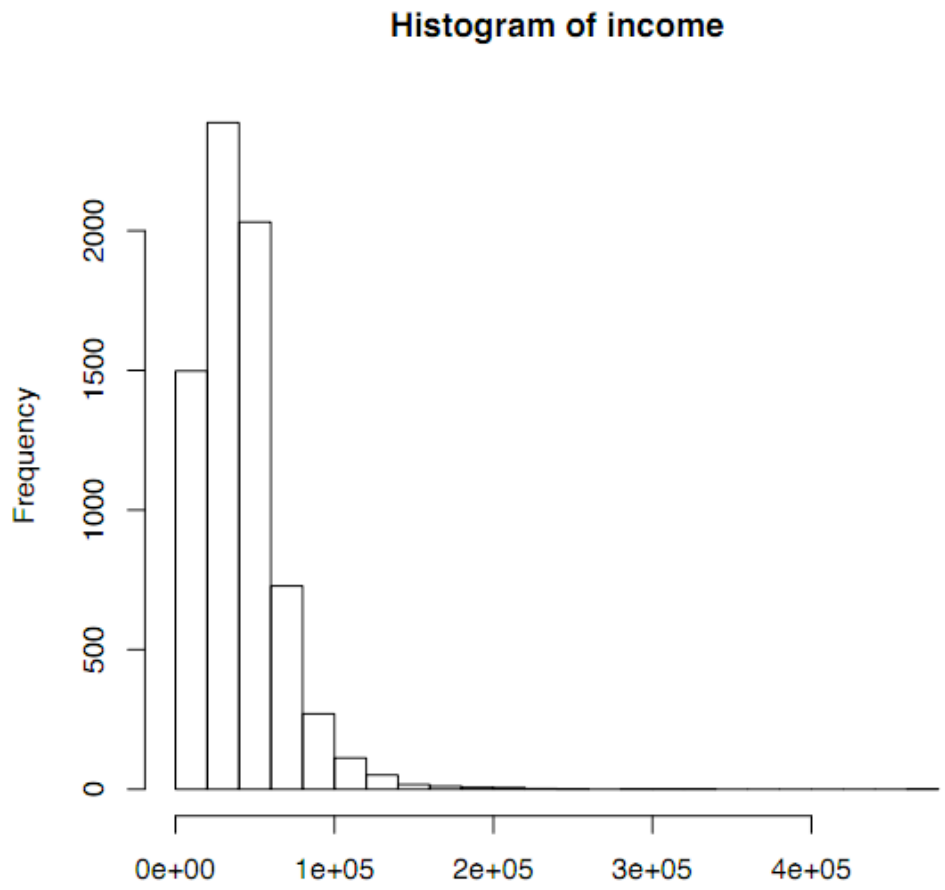


* Yeah, yeah, cheesy example, sorry.

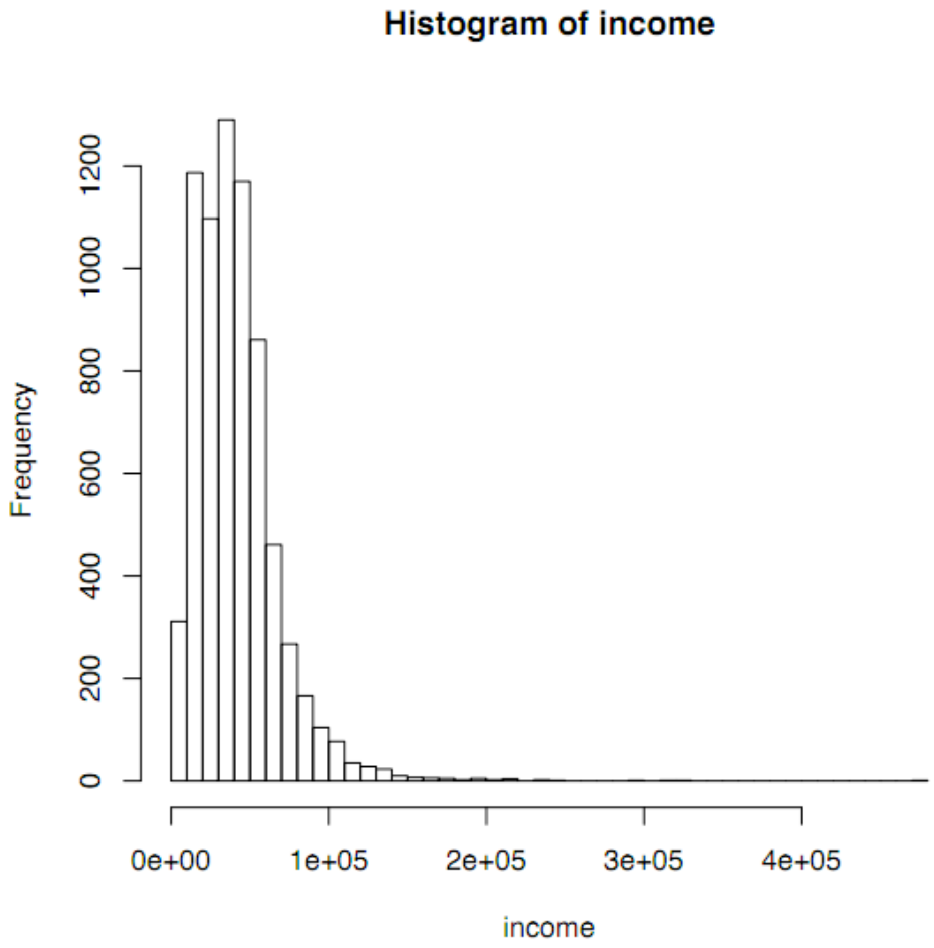
A comment on varying bin size... we see skew



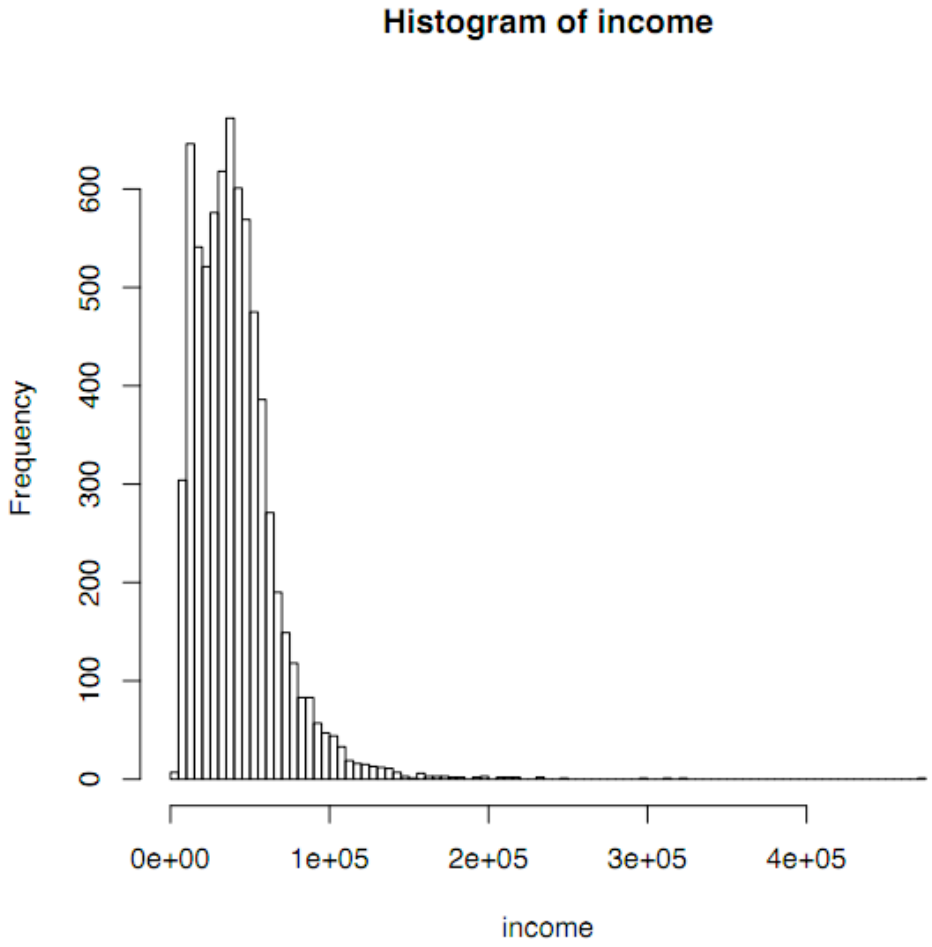
... and more skew



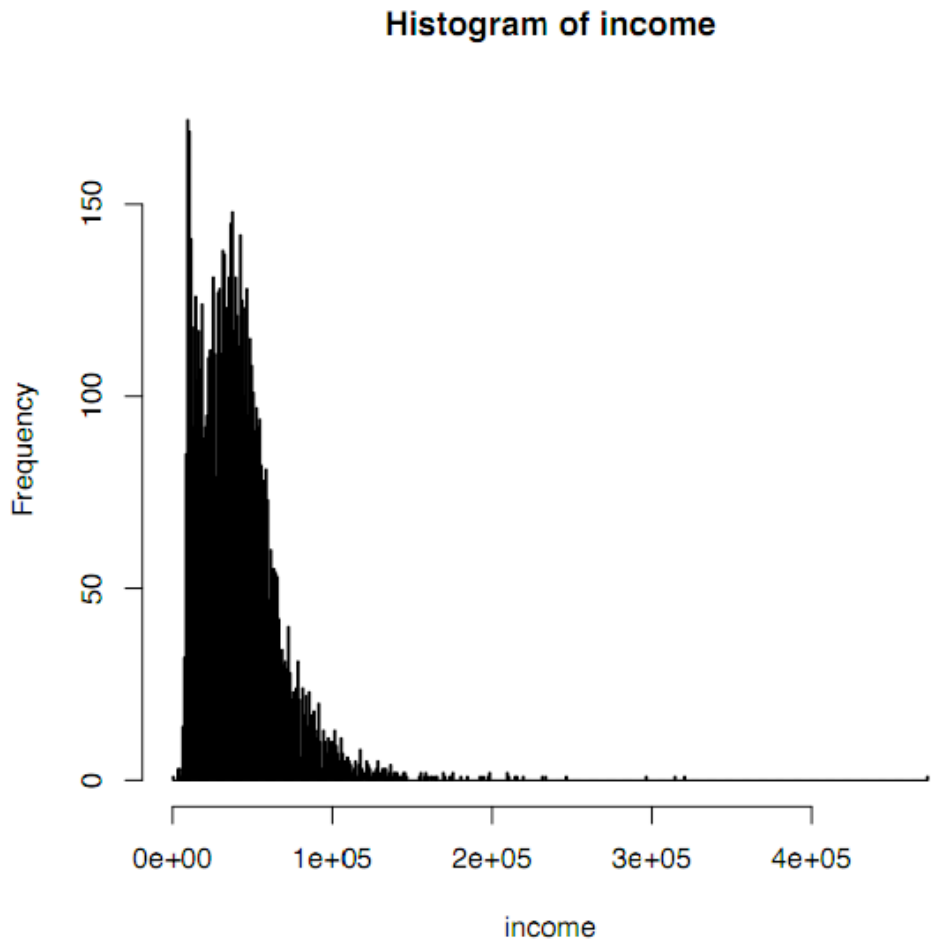
... but something starts to emerge at the left



... and more


Histogram of income
Frequency
0
100
200
300
400
500
600
0e+00
1e+05
2e+05
3e+05
4e+05
income

... and it's quite visible now; what might this be?



Taxonomy of shapes

We say that the income distribution has **two modes or peaks**

In categorical settings, the mode is defined to be the value of the variable receiving the most counts -- By analogy, we call peaks or regions of high concentration in continuous data, modes

Notice however, that the question of the number of modes is **a bit slippery** for continuous data; which bin width do you use?

Alternate views

So far we have had a record view and a student view -- We can also form a building view and examine the times of the day each building is occupied

As a start, at 3:00 on Wednesdays last quarter he is the distribution of people across campus (at least in terms of scheduled events)

WGYOUNG	BOELTER	MS	PUB AFF	HAINES	BUNCHE
800	752	741	674	605	352
PAB	HUMANTS	LAKRETZ	ROYCE	DODD	FOWLER
342	321	284	272	268	247
FRANZ	MOORE	KNSY PV	DE NEVE	MELNITZ	ROLFE
229	219	187	177	171	146
GEOLOGY	KAUFMAN	HLTHSCI	CORNELL	GOLD	KNUDSEN
121	113	104	101	84	84
ENTRPNR	PUB HLT	MACGOWN	NEUROSC	MCGWN E	BROAD
77	76	63	54	44	43
COVEL	SLICHTR	GSEIS	SMB	ENGR IV	POWELL
35	24	21	20	19	19
BOTANY	DENT	MOL SCI	FACTOR		AU
17	17	12	5	1	0
BIO SCI	BMC	BOYER	CAMPBEL	COLLINS	ENGR I
0	0	0	0	0	0
ENGR V	FERNALD	FIELD	GONDA	KORN	LAW
0	0	0	0	0	0
LS	LUVALLE	MACDNLD	OFF CAM	PERLOFF	PVUB
0	0	0	0	0	0
SAC	SCULPT	SEMEL	SPROUL	STRTHMR	TBA
0	0	0	0	0	0
UES					
0					

Privacy

Returning to a moment to our privacy ideas, I asked in the last lecture if you thought many students had your exact same schedule last term -- The general feeling seemed to be, um, “lots”

As part of the “student” view of these data, we can include a character string that represents all the classes a student -- We simply paste together the sorted list of their classes

We can then see how many of these strings are shared by different people --

"NURSING 0218C NURSING 0418C"
"INF STD 0425 INF STD 0464 INF STD 0498"
"MGMT 0297B MGMT 0420"
"HLT SER 0205 HLT SER 0289"
"EDUC 0458B"
"M PHARM 0251"
"BIOSTAT 0201 HLT SER 0225B HLT SER 0266B"
"BIOSTAT 0288 BIOSTAT 0411 EPIDEM 0291 EPIDEM 0292 EPIDEM 0412"
"ENGL 0244 ENGL 0247"
"MGMT 0295D MGMT 0297B"
"GERMAN 0209C GERMAN 0260"
"FILM TV 0434"
"COM LIT 0375"
"ART HIS 0113C ART HIS 0375"
"INF STD 0289 INF STD 0438B"
"HIN-URD 0005 LING 0276"
"GERMAN 0005 GREEK 0202A"
"BIOSTAT 0208 M STATS 0200B"
"HIST 0122E M HIST 0179A"
"ANTHRO 0133P ANTHRO 0149C MATH 0071SL PSYCH 0098T M"
"E&S SCI 0245B"
"CHICANO 0191 SOCIOL 0236A SOCIOL 0261"
"PSYCH 0296A PSYCH 0410E"
"SOC WLF 0201B SOC WLF 0201C SOC WLF 0230B SOC WLF 0240B SOC WLF 0401B"
"SOCIOL 0175 M SOCIOL 0180A"
"MATH 0002 STATS 0010"
"EDUC 0296G EDUC 0296H"
"NURSING 0174 NURSING 0225A NURSING 0230B NURSING 0260 NURSING 0266"
"CH ENGR 0298C CH ENGR 0299"
"HIST 0201B"
"MGMT 0215 MGMT 0281B MGMT 0406

Top course schedule “signatures”

MGMT 0403 MGMT 0405 (256)
MGMT 0410 MGMT 0411B MGMT 0421B MGMT 0430 (235)
EDUC 0330D EDUC 0491 (128)
SOC WLF 0201B SOC WLF 0201C SOC WLF 0230B SOC WLF 0240B SOC WLF 0401B (089)
NURSING 0239C NURSING 0264 NURSING 0439C (074)
EDUC 0288 (060)
NURSING 0174 NURSING 0225A NURSING 0230B NURSING 0260 NURSING 0266 (058)
BIOL CH 0254C BIOL CH 0254D (055)
NURSING 0115 NURSING 0160 NURSING 0174 (054)
ARCH&UD 0289 ARCH&UD 0401 (049)
NURSING 0171 NURSING 0267 NURSING 0270 (046)
M PHARM 0251 (039)
CHEM 0020B CHEM 0020L MATH 0032B PHYSICS 0001A (038)
MGMT 0298D MGMT 0420 (038)
CHEM 0020B CHEM 0020L MATH 0032A PHYSICS 0001A (036)
EDUC 0296G EDUC 0296H EDUC 0411 (035)
MGMT 0232F MGMT 0420 (035)
MGMT 0295D MGMT 0410 MGMT 0411B MGMT 0421B MGMT 0430 (034)
MGMT 0237A MGMT 0237B MGMT 0237C MGMT 0237E (033)
CHEM 0204 (031)
EDUC 0452B EDUC 0454A (029)
NURSING 0211 NURSING 0224 NURSING 0429A (028)
ECON 0201B ECON 0202B ECON 0203B (026)
EDUC 0499B (025)
PUB PLC 0202 PUB PLC 0204 PUB PLC 0208 (025)

Anonymity

As it turns out, 76% of all students have unique course lists -- If you restrict that just to undergraduates, 87% of you had a unique course load last quarter

This illustrates the idea we were trying to motivate with the social security system and its data releases -- Adding variables (adding more classes) serves to make you unique, we'll talk about this a bit later in the quarter when we discuss the so-called "curse of dimensionality"