

DATA MINING

LECTURE 2

Data Preprocessing
Exploratory Analysis
Post-processing



What is Data Mining?

- Data mining is the use of **efficient** techniques for the analysis of **very large** collections of data and the extraction of **useful** and possibly **unexpected** patterns in data.
- “Data mining is the analysis of (often large) observational data sets to find **unsuspected relationships** and to **summarize** the data in novel ways that are both **understandable and useful** to the data analyst” (Hand, Mannila, Smyth)
- “Data mining is the discovery of **models** for data” (Rajaraman, Ullman)
 - We can have the following types of models
 - Models that **explain** the data (e.g., a single function)
 - Models that **predict** the future data instances.
 - Models that **summarize** the data
 - Models that ~~extract~~ **extract** the most prominent **features** of the data.

Why do we need data mining?

- Really **huge** amounts of **complex** data generated from multiple sources and **interconnected** in different ways
 - **Scientific** data from different disciplines
 - Weather, astronomy, physics, biological microarrays, genomics
 - Huge **text** collections
 - The Web, scientific articles, news, tweets, facebook postings.
 - **Transaction** data
 - Retail store records, credit card records
 - **Behavioral** data
 - Mobile phone data, query logs, browsing behavior, ad clicks
 - **Networked** data
 - The Web, Social Networks, IM networks, email network, biological networks.
 - All these types of data can be **combined** in many ways
 - Facebook has a network, text, images, user behavior, ad transactions.
- We need to **analyze** this data to **extract knowledge**
 - Knowledge can be used for **commercial** or **scientific** purposes.
 - Our solutions should **scale** to the size of the data

The data analysis pipeline

- Mining is not the only step in the analysis process



- **Preprocessing**: real data is noisy, incomplete and inconsistent. **Data cleaning** is required to make sense of the data
 - Techniques: Sampling, Dimensionality Reduction, Feature selection.
 - A dirty work, but it is often the most important step for the analysis.
- **Post-Processing**: Make the data actionable and useful to the user
 - Statistical analysis of importance
 - Visualization.
- Pre- and Post-processing are often data mining tasks as well

Data Quality

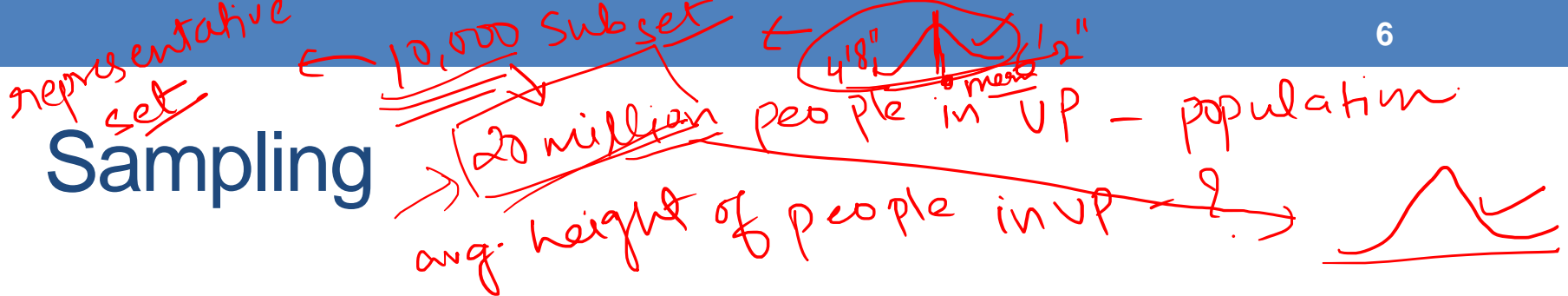
- Examples of data quality problems:
 - Noise and outliers
 - Missing values
 - Duplicate data

A mistake or a millionaire?

Missing values

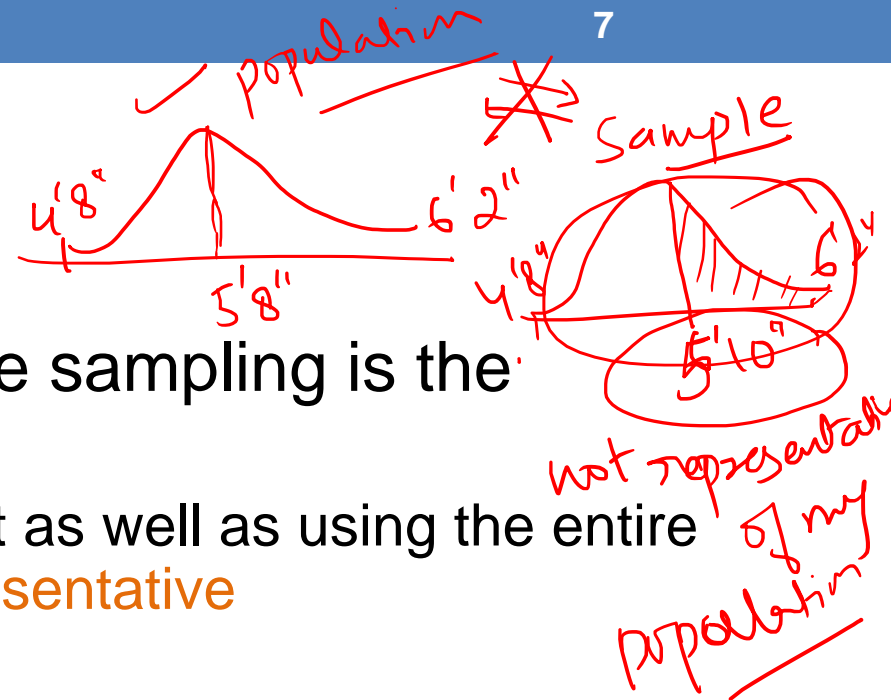
Inconsistent duplicate entries

<i>Tid</i>	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	10000K	Yes
6	No	NULL	60K	No
7	Yes	Divorced	220K	NULL
8	No	Single	85K	Yes
9	No	Married	90K	No
9	No	Single	90K	No



- Sampling is the main technique employed for data selection.
 - It is often used for both the preliminary investigation of the data and the final data analysis.
- Statisticians sample because obtaining the entire set of data of interest is too expensive or time consuming.
 - Example: What is the average height of a person in Ioannina? *UP-*
 - We cannot measure the height of everybody
- Sampling is used in data mining because processing the entire set of data of interest is too expensive or time consuming.
 - Example: We have **1M** documents. What fraction has at least 100 words in common?
 - Computing number of common words for all pairs requires **10¹²** comparisons
 - Example: What fraction of tweets in a year contain the word "Greece"?
 - **300M** tweets per day, if **100** characters on average, **86.5TB** to store all tweets

Sampling ...



- The key principle for effective sampling is the following:
 - using a sample will work almost as well as using the entire data sets, if the sample is **representative**
 - A sample is representative if it has approximately the same property (of interest) as the original set of data
 - Otherwise we say that the sample introduces some **bias**
 - What happens if we take a sample from the university campus to compute the average height of a person at loannina?

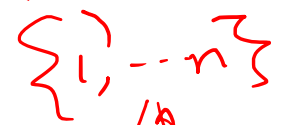
JK.

Types of Sampling

sample set



population set



~~it is removed~~

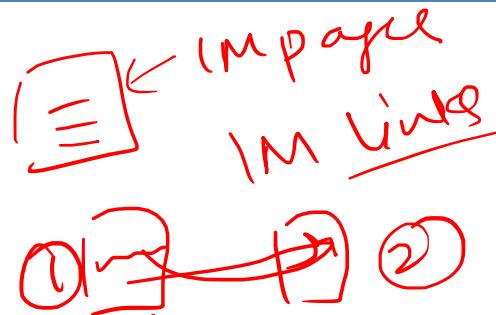
- Simple Random Sampling
 - There is an equal probability of selecting any particular item
- Sampling **without replacement**
 - As each item is selected, it is removed from the population
- Sampling **with replacement**
 - Objects are not removed from the population as they are selected for the sample.
 - In sampling with replacement, the same object can be picked up more than once. This makes analytical computation of probabilities easier
 - E.g., we have 100 people, 51 are women $P(W) = 0.51$, 49 men $P(M) = 0.49$. If I pick two persons what is the probability $P(W, W)$ that both are women?
 - Sampling ~~with replacement~~: $P(W, W) = 0.51^2$
 - Sampling without replacement: $P(W, W) = \frac{51}{100} * \frac{50}{99}$

$$P(W) * P(W)$$

50 99

Types of Sampling

10K 1M



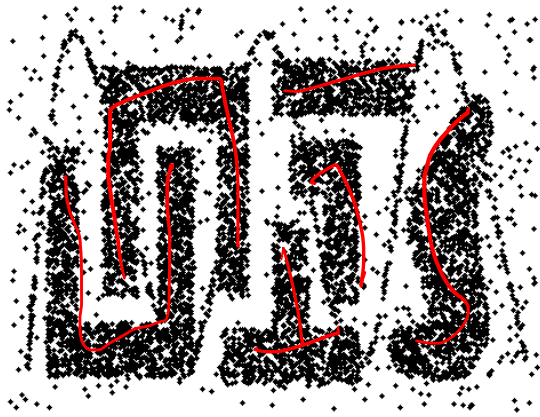
• Stratified sampling

- Split the data into several **groups**; then draw random samples from each group.
 - Ensures that both groups are represented.
- **Example 1.** I want to understand the differences between legitimate and fraudulent credit card transactions. **0.1%** of transactions are fraudulent. What happens if I select **1000** transactions at random?
 - I get **1** fraudulent transaction (in expectation). Not enough to draw any conclusions. Solution: sample **1000** legitimate and **1000** fraudulent transactions

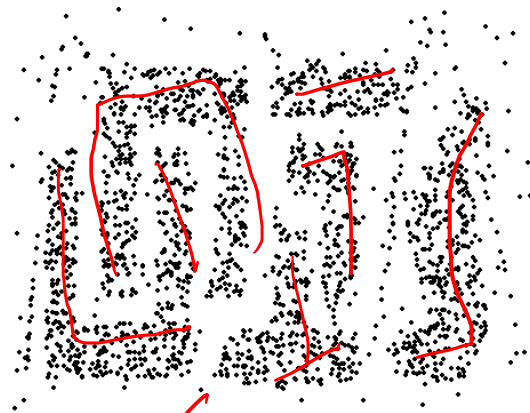
Probability Reminder: If an event has probability p of happening and I do N trials, the expected number of times the event occurs is pN

- **Example 2.** I want to answer the question: Do web pages that are linked have on average more words in common than those that are not? I have **1M** pages, and **1M** links, what happens if I select **10K** pairs of pages at random?
 - Most likely I will not get any links. Solution: sample **10K** random pairs, and **10K** links

Sample Size



8000 points



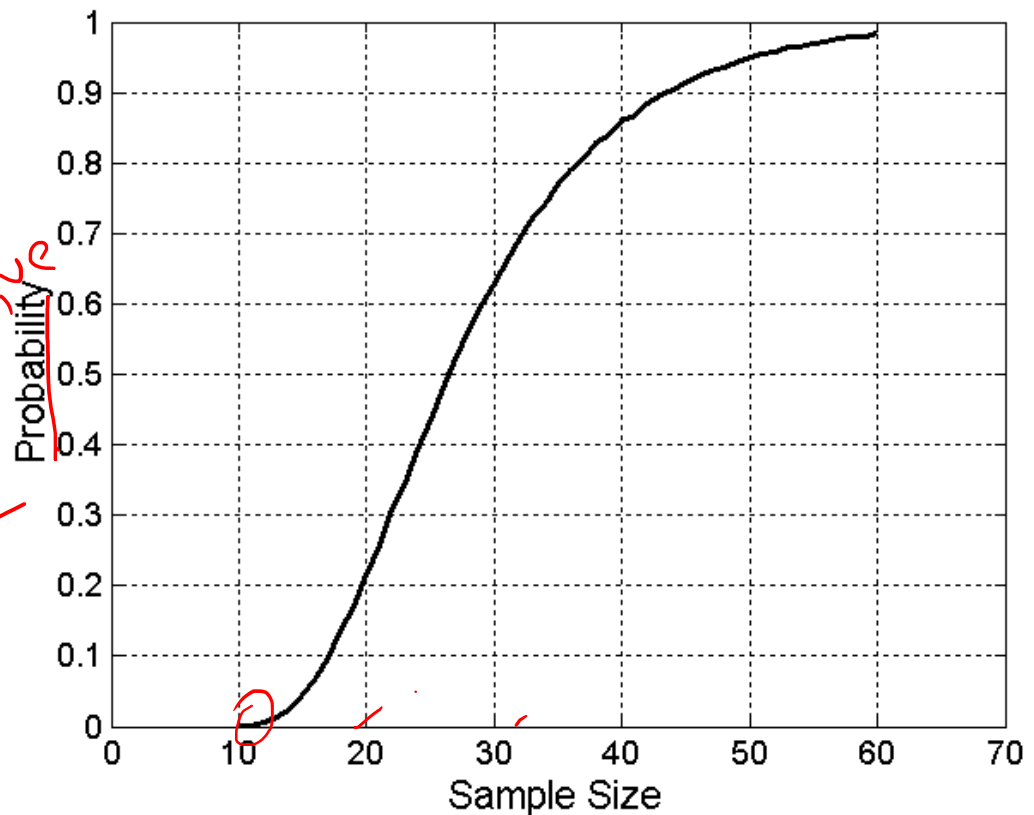
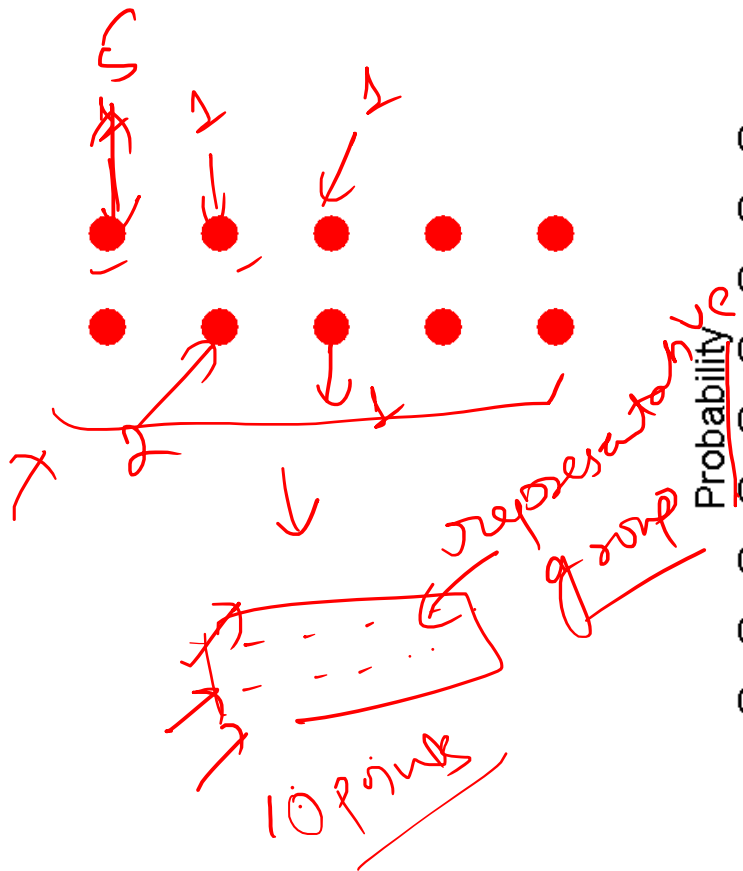
2000 Points



500 Points

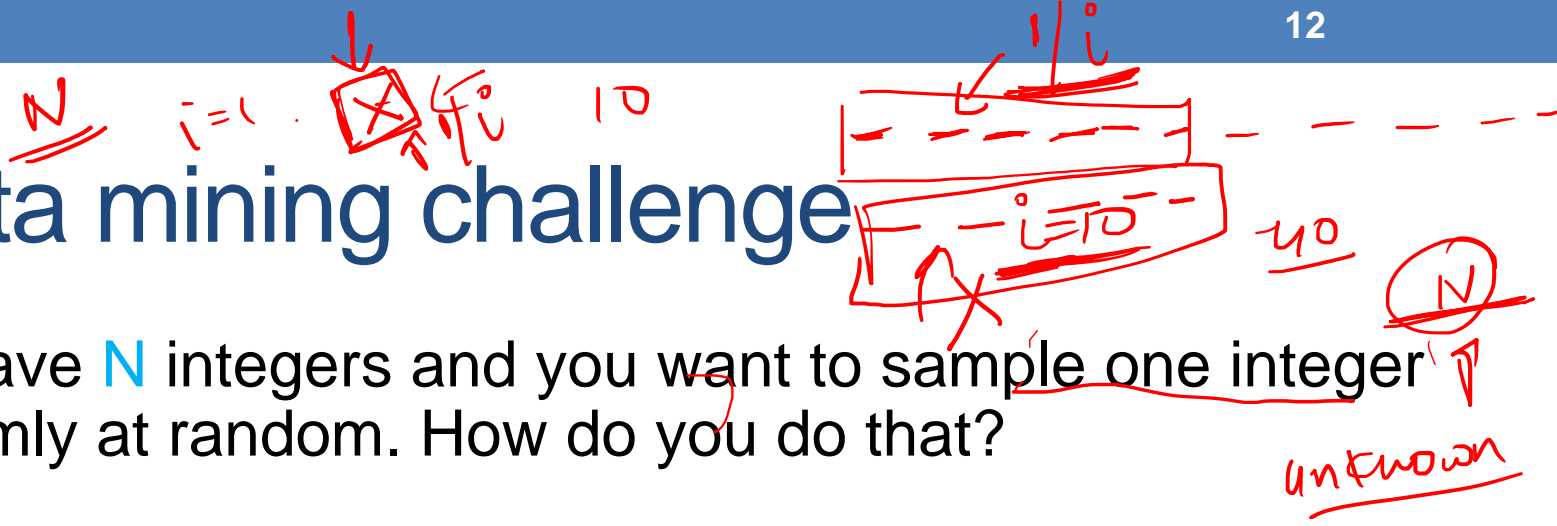
Sample Size

- What sample size is necessary to get at least one object from each of 10 groups.



A data mining challenge

- You have N integers and you want to sample one integer uniformly at random. How do you do that?
- The integers are coming in a stream: you do not know the size of the stream in advance, and there is not enough memory to store the stream in memory. You can only keep a constant amount of integers in memory.
- How do you sample?
 - Hint: if the stream ends after reading n integers the last integer in the stream should have probability $1/n$ to be selected.
- Reservoir Sampling:**
 - Standard interview question for many companies



$1/i$

$N=50$

$1/n$

Reservoir sampling

- Algorithm: With probability $1/n$ select the n -th item of the stream and replace the previous choice.

- Claim: Every item has probability $1/N$ to be selected after N items have been read.

Proof

- What is the probability of the n -th item to be selected?

- $\frac{1}{n}$

- What is the probability of the n -th items to survive for $N-n$ rounds?

- $\left(1 - \frac{1}{n+1}\right) \left(1 - \frac{1}{n+2}\right) \cdots \left(1 - \frac{1}{N}\right)$

$$P(S = a_i) = \frac{1}{i} \left(1 - \frac{1}{i+1}\right) \left(1 - \frac{1}{i+2}\right) \cdots \left(1 - \frac{1}{n}\right)$$

$$= \frac{1}{i} \left(\frac{i+1-1}{i+1}\right) \left(\frac{i+2-1}{i+2}\right) \cdots \left(\frac{n-1}{n}\right)$$

$$= \frac{1}{i} \left(\frac{i}{i+1}\right) \left(\frac{i+1}{i+2}\right) \cdots \left(\frac{n-1}{n}\right)$$

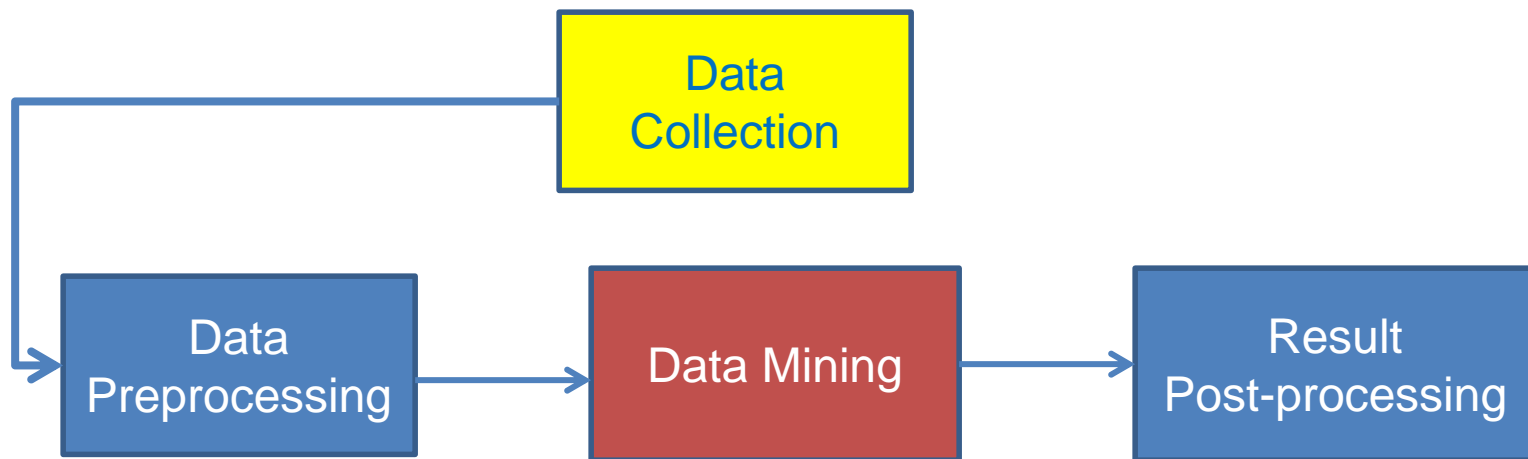
$= 1/n$
 prob. of selecting one
 no. of N nos.
 (should be $1/n$)

A (detailed) data preprocessing example

- Suppose we want to mine the comments/reviews of people on [Yelp](#) and [Foursquare](#).



Data Collection



- Today there is an abundance of data online
 - Facebook, Twitter, Wikipedia, Web, etc...
- We can extract interesting information from this data, but first we need to collect it
 - Customized crawlers, use of public APIs
 - Additional cleaning/processing to parse out the useful parts
 - Respect of crawling etiquette

Mining Task

- Collect all reviews for the top-10 most reviewed restaurants in NY in Yelp
 - (thanks to Hady Law)
- Find few terms that best describe the restaurants.
- Algorithm?

Example data

- I heard so many good things about this place so I was pretty juiced to try it. I'm from Cali and I heard Shake Shack is comparable to IN-N-OUT and I gotta say, Shake Shack wins hands down. Surprisingly, the line was short and we waited about 10 MIN. to order. I ordered a regular cheeseburger, fries and a black/white shake. So yummerz. I love the location too! It's in the middle of the city and the view is breathtaking. Definitely one of my favorite places to eat in NYC.
- I'm from California and I must say, Shake Shack is better than IN-N-OUT, all day, err'day.
- Would I pay \$15+ for a burger here? No. But for the price point they are asking for, this is a definite bang for your buck (though for some, the opportunity cost of waiting in line might outweigh the cost savings) Thankfully, I came in before the lunch swarm descended and I ordered a shake shack (the special burger with the patty + fried cheese & portabella topping) and a coffee milk shake. The beef patty was very juicy and snugly packed within a soft potato roll. On the downside, I could do without the fried portabella-thingy, as the crispy taste conflicted with the juicy, tender burger. How does shake shack compare with in-and-out or 5-guys? I say a very close tie, and I think it comes down to personal affiliations. On the shake side, true to its name, the shake was well churned and very thick and luscious. The coffee flavor added a tangy taste and complemented the vanilla shake well. Situated in an open space in NYC, the open air sitting allows you to munch on your burger while watching people zoom by around the city. It's an oddly calming experience, or perhaps it was the food coma I was slowly falling into. Great place with food at a great price.

First cut

- Do simple processing to “normalize” the data (remove punctuation, make into lower case, clear white spaces, other?)
- Break into words, keep the most popular words

the 27514
and 14508
i 13088
a 12152
to 10672
of 8702
ramen 8518
was 8274
is 6835
it 6802
in 6402
for 6145
but 5254
that 4540
you 4366
with 4181
pork 4115
my 3841
this 3487
wait 3184
not 3016
we 2984
at 2980
on 2922

the 16710
and 9139
a 8583
i 8415
to 7003
in 5363
it 4606
of 4365
is 4340
burger 432
was 4070
for 3441
but 3284
shack 3278
shake 3172
that 3005
you 2985
my 2514
line 2389
this 2242
fries 2240
on 2204
are 2142
with 2095

the 16010
and 9504
i 7966
to 6524
a 6370
it 5169
of 5159
is 4519
sauce 4020
in 3951
this 3519
was 3453
for 3327
you 3220
that 2769
but 2590
food 2497
on 2350
my 2311
cart 2236
chicken 2220
with 2195
rice 2049
so 1825

the 14241
and 8237
a 8182
i 7001
to 6727
of 4874
you 4515
it 4308
is 4016
was 3791
pastrami 3748
in 3508
for 3424
sandwich 2928
that 2728
but 2715
on 2247
this 2099
my 2064
with 2040
not 1655
your 1622
so 1610
have 1585

First cut

- Do simple processing to “normalize” the data (remove punctuation, make into lower case, clear white spaces, other?)
- Break into words, keep the most popular words

the 27514
and 14508
i 13088
a 12152
to 10672
of 8702
ramen 8518
was 8274
is 6835
it 6802
in 6402
for 6145
but 5254
that 4540
you 4366
with 4181
pork 4115
my 3841
this 3487
wait 3184
not 3016
we 2984
at 2980
on 2922

the 16710
and 9139
a 8583
i 8415
to 7003
in 5363
it 4606
of 4365
is 4340
burger 432
was 4070
for 3441
but 3284
shack 3278
shake 3172
that 3005
you 2985
my 2514
line 2389
this 2242
fries 2240
on 2204
are 2142
with 2095

the 16010
and 9504
i 7966
to 6524
a 6370
it 5169
of 5159
is 4519
sauce 4020
in 3951
this 3519
was 3453
for 3327
you 3220
that 2769
but 2590
food 2497

cart 2236
chicken 2220
with 2195
rice 2049
so 1825

the 14241
and 8237
a 8182
i 7001
to 6727
of 4874
you 4515
it 4308
is 4016
was 3791
pastrami 3748
in 3508
for 3424
sandwich 2928
that 2728
but 2715
on 2247

not 1655
your 1622
so 1610
have 1585

Most frequent words are **stop words**

Second cut

- Remove stop words
 - Stop-word lists can be found online.

a, about, above, after, again, against, all, am, an, and, any, are, aren't, as, at, be, because, been, before, being, below, between, both, but, by, can't, cannot, could, couldn't, did, didn't, do, does, doesn't, doing, don't, down, during, each, few, for, from, further, had, hadn't, has, hasn't, have, haven't, having, he, he'd, he'll, he's, her, here, here's, hers, herself, him, himself, his, how, how's, i, i'd, i'll, i'm, i've, if, in, into, is, isn't, it, it's, its, itself, let's, me, more, most, mustn't, my, myself, no, nor, not, of, off, on, once, only, or, other, ought, our, ours, ourselves, out, over, own, same, shan't, she, she'd, she'll, she's, should, shouldn't, so, some, such, than, that, that's, the, their, theirs, them, themselves, then, there, there's, these, they, they'd, they'll, they're, they've, this, those, through, to, too, under, until, up, very, was, wasn't, we, we'd, we'll, we're, we've, were, weren't, what, what's, when, when's, where, where's, which, while, who, who's, whom, why, why's, with, won't, would, wouldn't, you, you'd, you'll, you're, you've, your, yours, yourself, yourselves,

Second cut

- Remove stop words
 - Stop-word lists can be found online.

ramen 8572
pork 4152
wait 3195
good 2867
place 2361
noodles 2279
ippudo 2261
buns 2251
broth 2041
like 1902
just 1896
get 1641
time 1613
one 1460
really 1437
go 1366
food 1296
bowl 1272
can 1256
great 1172
best 1167

burger 4340
shack 3291
shake 3221
line 2397
fries 2260
good 1920
burgers 1643
wait 1508
just 1412
cheese 1307
like 1204
food 1175
get 1162
place 1159
one 1118
long 1013
go 995
time 951
park 887
can 860
best 849

sauce 4023
food 2507
cart 2239
chicken 2238
rice 2052
hot 1835
white 1782
line 1755
good 1629
lamb 1422
halal 1343
just 1338
get 1332
one 1222
like 1096
place 1052
go 965
can 878
night 832
time 794
long 792
people 790

pastrami 3782
sandwich 2934
place 1480
good 1341
get 1251
katz's 1223
just 1214
like 1207
meat 1168
one 1071
deli 984
best 965
go 961
ticket 955
food 896
sandwiches 813
can 812
beef 768
order 720
pickles 699
time 662

Second cut

- Remove stop words
 - Stop-word lists can be found online.

ramen 8572
 pork 4152
 wait 3195
 good 2867
 place 2361
 noodles 2279
 ippudo 2261
 buns 2251
 broth 2041
 like 1902
 just 1896
 get 1641
 time 1613
 one 1460
 really 1437
 go 1366
 food 1296
 bowl 1272
 can 1256
 great 1172
 best 1167

burger 4340
 shack 3291
 shake 3221
 line 2397
 fries 2260
 good 1920
 burgers 1643
 wait 1508
 just 1412
 cheese 1307
 like 1204
 food 1175
 get 1162

sauce 4023
 food 2507
 cart 2239
 chicken 2238
 rice 2052
 hot 1835
 white 1782
 line 1755
 good 1629
 lamb 1422
 halal 1343
 just 1338
 get 1332

pastrami 3782
 sandwich 2934
 place 1480
 good 1341
 get 1251
 katz's 1223
 just 1214
 like 1207
 meat 1168
 one 1071
 deli 984
 best 965
 go 961

Commonly used words in reviews, not so interesting

long 1015
 go 995
 time 951
 park 887
 can 860
 best 849

place 1052
 go 965
 can 878
 night 832
 time 794
 long 792
 people 790

sandwiches 813
 can 812
 beef 768
 order 720
 pickles 699
 time 662

IDF

- Important words are the ones that are unique to the document (differentiating) compared to the rest of the collection
 - All reviews use the word “like”. This is not interesting
 - We want the words that characterize the specific restaurant
- **Document Frequency** $DF(w)$: fraction of documents that contain word w .

$$DF(w) = \frac{D(w)}{D} \quad \begin{array}{l} D(w): \text{num of docs that contain word } w \\ D: \text{total number of documents} \end{array}$$

- **Inverse Document Frequency** $IDF(w)$:

$$IDF(w) = \log\left(\frac{1}{DF(w)}\right)$$

- Maximum when unique to one document : $IDF(w) = \log(D)$
- Minimum when the word is common to all documents: $IDF(w) = 0$

TF-IDF

- The words that are best for describing a document are the ones that are **important for the document**, but also **unique to the document**.
- **TF(w,d)**: term frequency of word w in document d
 - Number of times that the word appears in the document
 - Natural measure of **importance** of the word for the document
- **IDF(w)**: inverse document frequency
 - Natural measure of the **uniqueness** of the word w
- **TF-IDF(w,d)** = $TF(w,d) \times IDF(w)$

Third cut

- Ordered by TF-IDF

ramen 3057.4176194	fries 806.08537330	lamb 985.655290756243	pastrami 1931.94250908298 6
akamaru 2353.24196	custard 729.607519	halal 686.038812717726	katz's 1120.62356508209 4
noodles 1579.68242	shakes 628.4738038	53rd 375.685771863491	rye 1004.28925735888 2
broth 1414.7133955	shroom 515.7790608	gyro 305.809092298788	corned 906.113544700399 2
miso 1252.60629058	burger 457.2646379	pita 304.984759446376	pickles 640.487221580035 4
hirata 709.1962086	crinkle 398.347221	cart 235.902194557873	reuben 515.779060830666 1
hakata 591.7643688	burgers 366.624854	platter 139.45990308004	matzo 430.583412389887 1
shiromaru 587.1591	madison 350.939350	chicken/lamb 135.852520	sally 428.110484707471 2
noodle 581.8446147	shackburger 292.42	carts 120.274374158359	harry 226.323810772916 4
tonkotsu 529.59457	'shroom 287.823136	hilton 84.2987473324223	mustard 216.079238853014 6
ippudo 504.5275695	portobello 239.806	lamb/chicken 82.8930633	cutter 209.535243462458 1
buns 502.296134008	custards 211.83782	yogurt 70.0078652365545	carnegie 198.655512713779 3
ippudo's 453.60926	concrete 195.16992	52nd 67.5963923222322	katz 194.387844446609 7
modern 394.8391629	bun 186.9621782983	6th 60.7930175345658 9	knish 184.206807439524 1
egg 367.3680056967	milkshakes 174.996	4am 55.4517744447956 5	sandwiches 181.415707218 8
shoyu 352.29551922	concretes 165.7861	yellow 54.4470265206673	brisket 131.945865389878 4
chashu 347.6903490	portabello 163.483	tzatziki 52.95945713886	fries 131.613054313392 7
karaka 336.1774235	shack's 159.334353	lettuce 51.323016802268	salami 127.621117258549 3
kakuni 276.3102111	patty 152.22603588	sammy's 50.656872045869	knishes 124.339595021678 1
ramens 262.4947006	ss 149.66803104461	sw 50.5668577816893 3	delicatessen 117.488967607 2
bun 236.5122638036	patties 148.068287	platters 49.90659700031	deli's 117.431839742696 1
wasabi 232.3667512	cam 105.9496067806	falafel 49.479699521204	carver 115.129254649702 1
dama 221.048168927	milkshake 103.9720	sober 49.2211422635451	brown's 109.441778045519 2
brulee 201.1797390	lamps 99.011158998	moma 48.1589121730374	matzoh 108.22149937072 1

Third cut

- TF-IDF takes care of stop words as well
- We do not need to remove the stopwords since they will get $IDF(w) = 0$

Decisions, decisions...

- When mining real data you often need to make some
 - What data should we collect? How much? For how long?
 - Should we throw out some data that does not seem to be useful?

An actual review

AAAAAAAAAAAAA

AAAAAAAAAAAAAAAAAAAAAAAAAAAAA AAAAAAAAAAAAAAAAAAAAAAAAAAAAA AAA

- Too frequent data (stop words), too infrequent (errors?), erroneous data, missing data, outliers
 - How should we weight the different pieces of data?
- Most decisions are application dependent. Some information may be lost but we can usually live with it (most of the times)
- We should make our decisions clear since they affect our findings.
- Dealing with real data is hard...

Exploratory analysis of data

- **Summary statistics**: numbers that summarize properties of the data
 - Summarized properties include **frequency**, **location** and **spread**
 - Examples: location - mean
 spread - standard deviation
 - Most summary statistics can be calculated in a single pass through the data

Frequency and Mode

- The **frequency** of an attribute value is the percentage of time the value occurs in the data set
 - For example, given the attribute 'gender' and a representative population of people, the gender 'female' occurs about 50% of the time.
- The **mode** of a an attribute is the most frequent attribute value
- The notions of frequency and mode are typically used with categorical data

Percentiles

- For continuous data, the notion of a **percentile** is more useful.

Given an ordinal or continuous attribute x and a number p between 0 and 100, the p^{th} percentile is a value x_p of x such that $p\%$ of the observed values of x are less than x_p .

- For instance, the 50th percentile is the value $x_{50\%}$ such that 50% of all values of x are less than $x_{50\%}$.

Measures of Location: Mean and Median

- The **mean** is the most common measure of the location of a set of points.
- However, the mean is very sensitive to outliers.
- Thus, the **median** or a trimmed mean is also commonly used.

$$\text{mean}(x) = \bar{x} = \frac{1}{m} \sum_{i=1}^m x_i$$

$$\text{median}(x) = \begin{cases} x_{(r+1)} & \text{if } m \text{ is odd, i.e., } m = 2r + 1 \\ \frac{1}{2}(x_{(r)} + x_{(r+1)}) & \text{if } m \text{ is even, i.e., } m = 2r \end{cases}$$

Example

<i>Tid</i>	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	10000K	Yes
6	No	NULL	60K	No
7	Yes	Divorced	220K	NULL
8	No	Single	85K	Yes
9	No	Married	90K	No
10	No	Single	90K	No

Mean: 1090K

Trimmed mean (remove min, max): 105K

Median: $(90+100)/2 = 95K$

Measures of Spread: Range and Variance

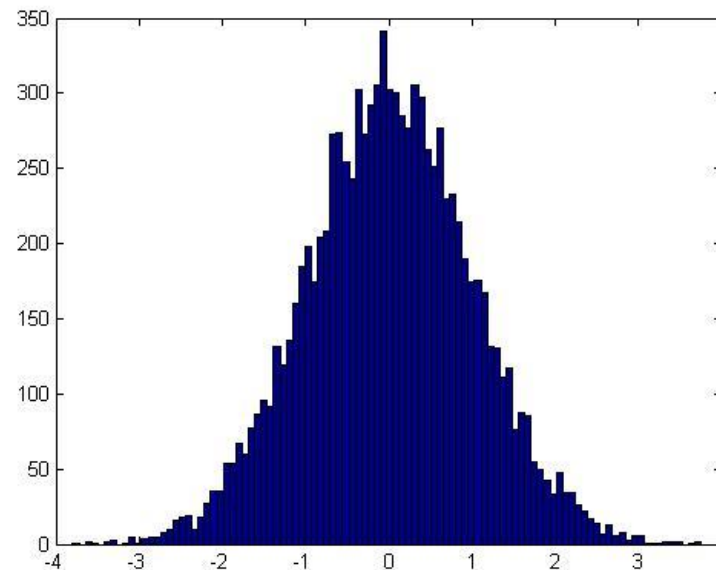
- **Range** is the difference between the max and min
- The **variance** or **standard deviation** is the most common measure of the spread of a set of points.

$$var(x) = \frac{1}{m} \sum_{i=1}^m (x - \bar{x})^2$$

$$\sigma(x) = \sqrt{var(x)}$$

Normal Distribution

- $$\phi(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

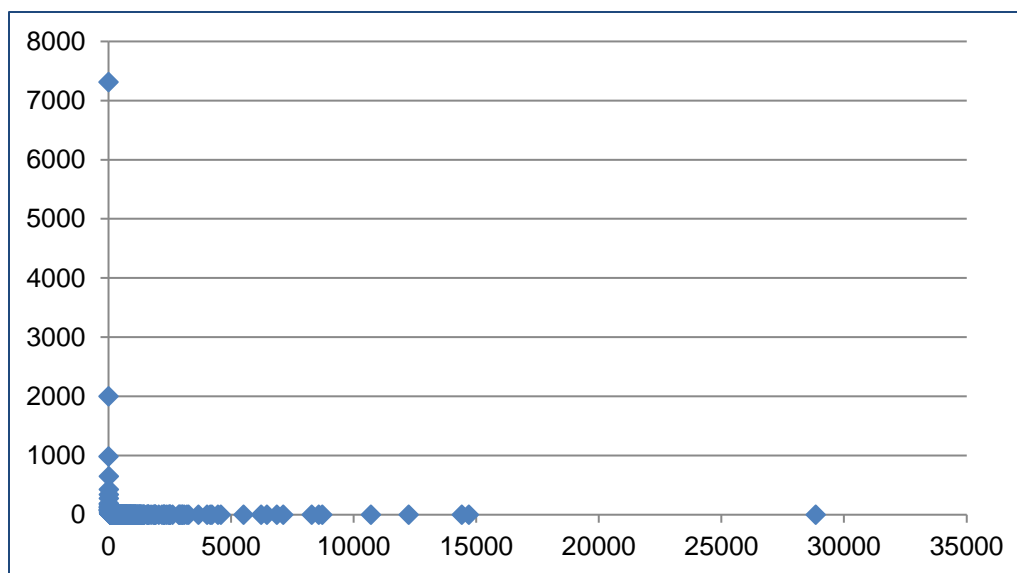


This is a value histogram

- An important distribution that characterizes many quantities and has a central role in probabilities and statistics.
 - Appears also in the central limit theorem
- Fully characterized by the mean μ and standard deviation σ

Not everything is normally distributed

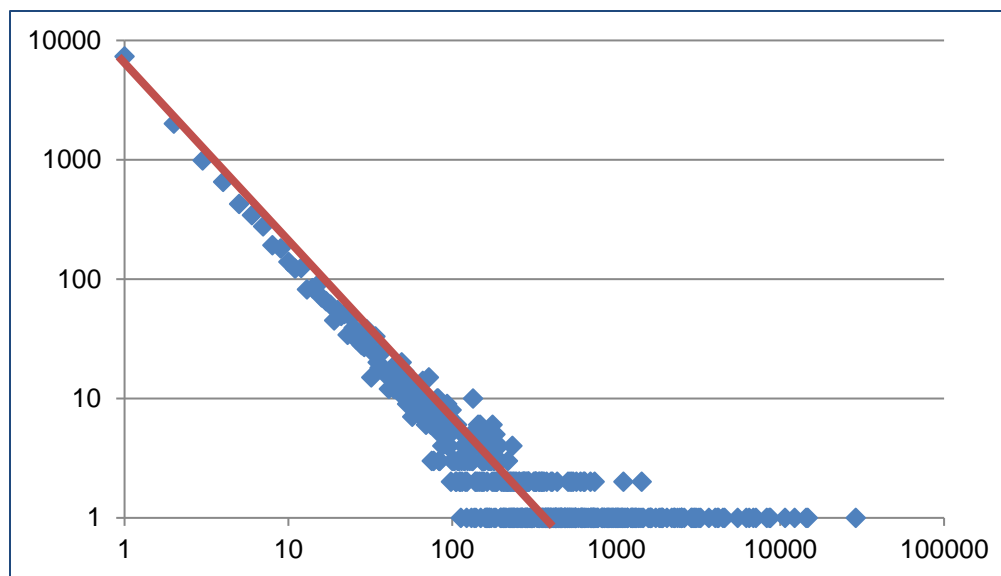
- Plot of number of words with x number of occurrences



- If this was a normal distribution we would not have a frequency as large as 28K

Power-law distribution

- We can understand the distribution of words if we take the **log-log** plot

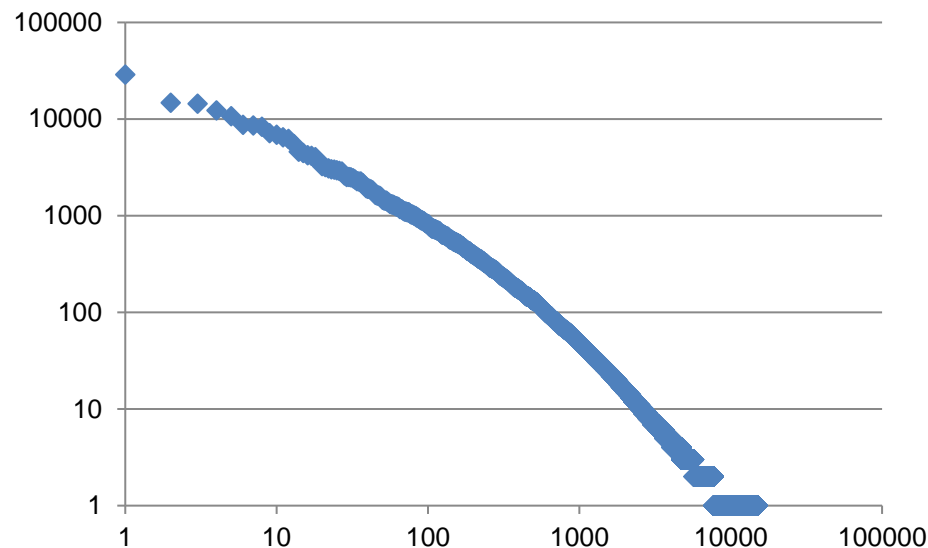


- Linear relationship in the log-log space

$$p(x = k) = k^{-a}$$

Zipf's law

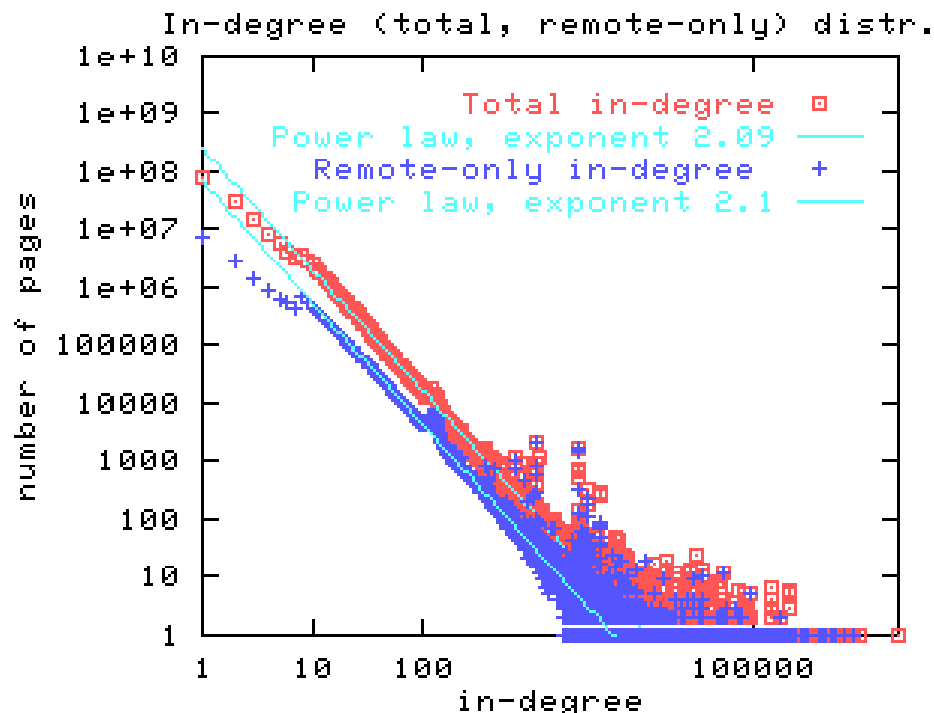
- Power laws can be detected by a linear relationship in the log-log space for the **rank-frequency** plot



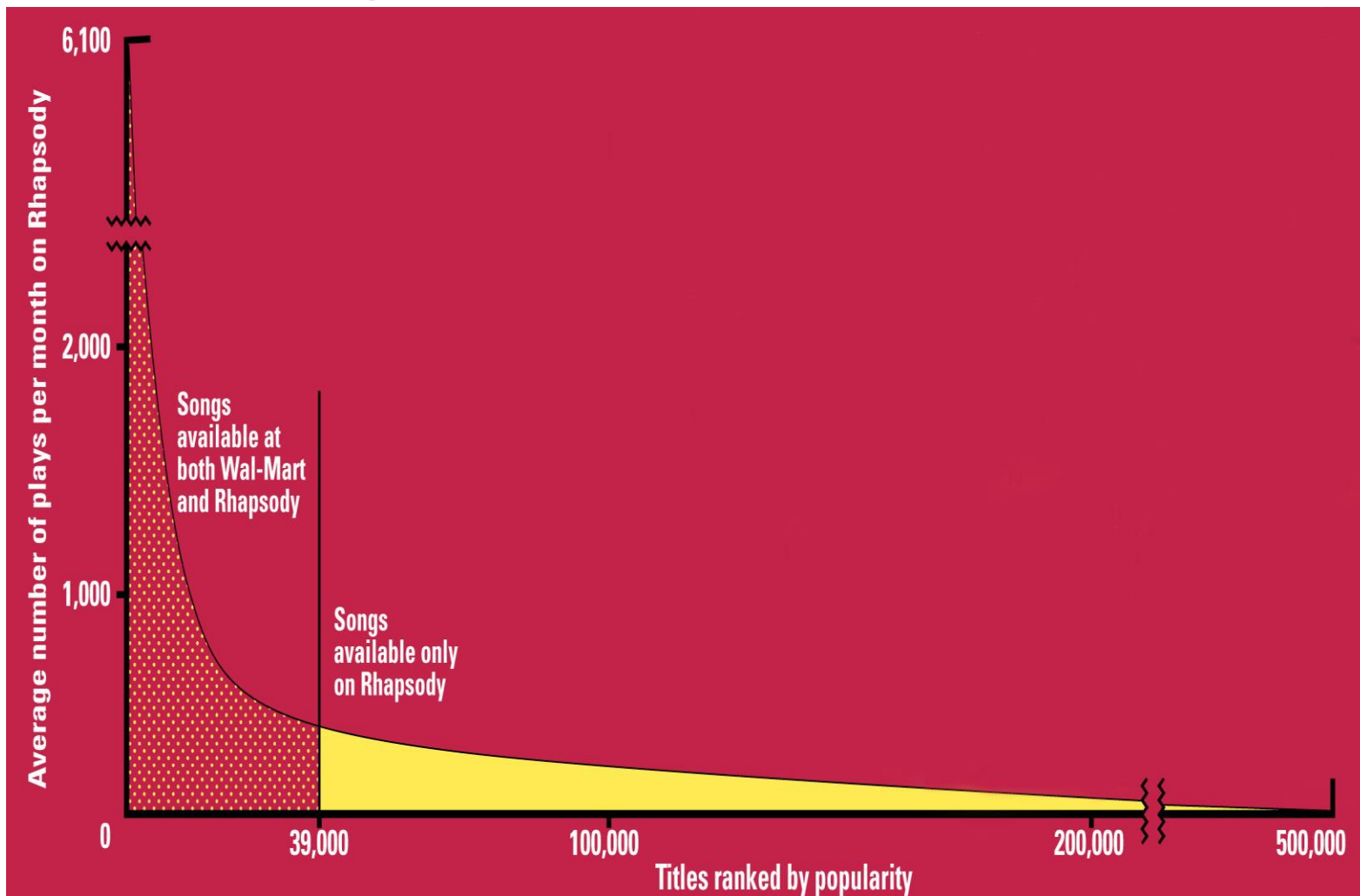
- $f(r)$: Frequency of the r -th most frequent word
$$f(r) = r^{-\beta}$$

Power-laws are everywhere

- Incoming and outgoing links of web pages, number of friends in social networks, number of occurrences of words, file sizes, city sizes, income distribution, popularity of products and movies
 - Signature of human activity?
 - A mechanism that explains everything?
 - Rich get richer process



The Long Tail



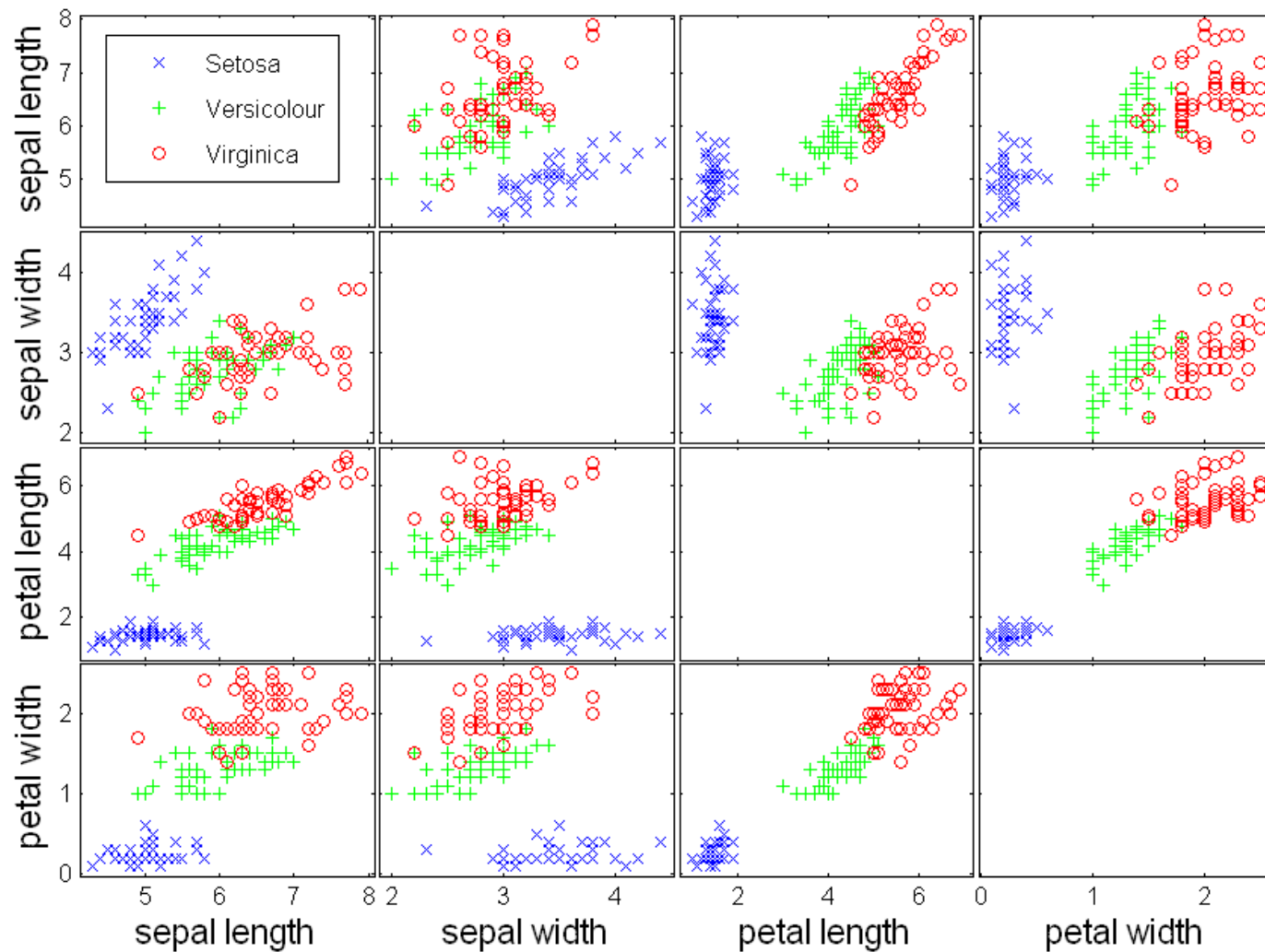
Source: Chris Anderson (2004)

Sources: Erik Brynjolfsson and Jeffrey Hu, MIT, and Michael Smith, Carnegie Mellon; Barnes & Noble; Netflix; RealNetworks

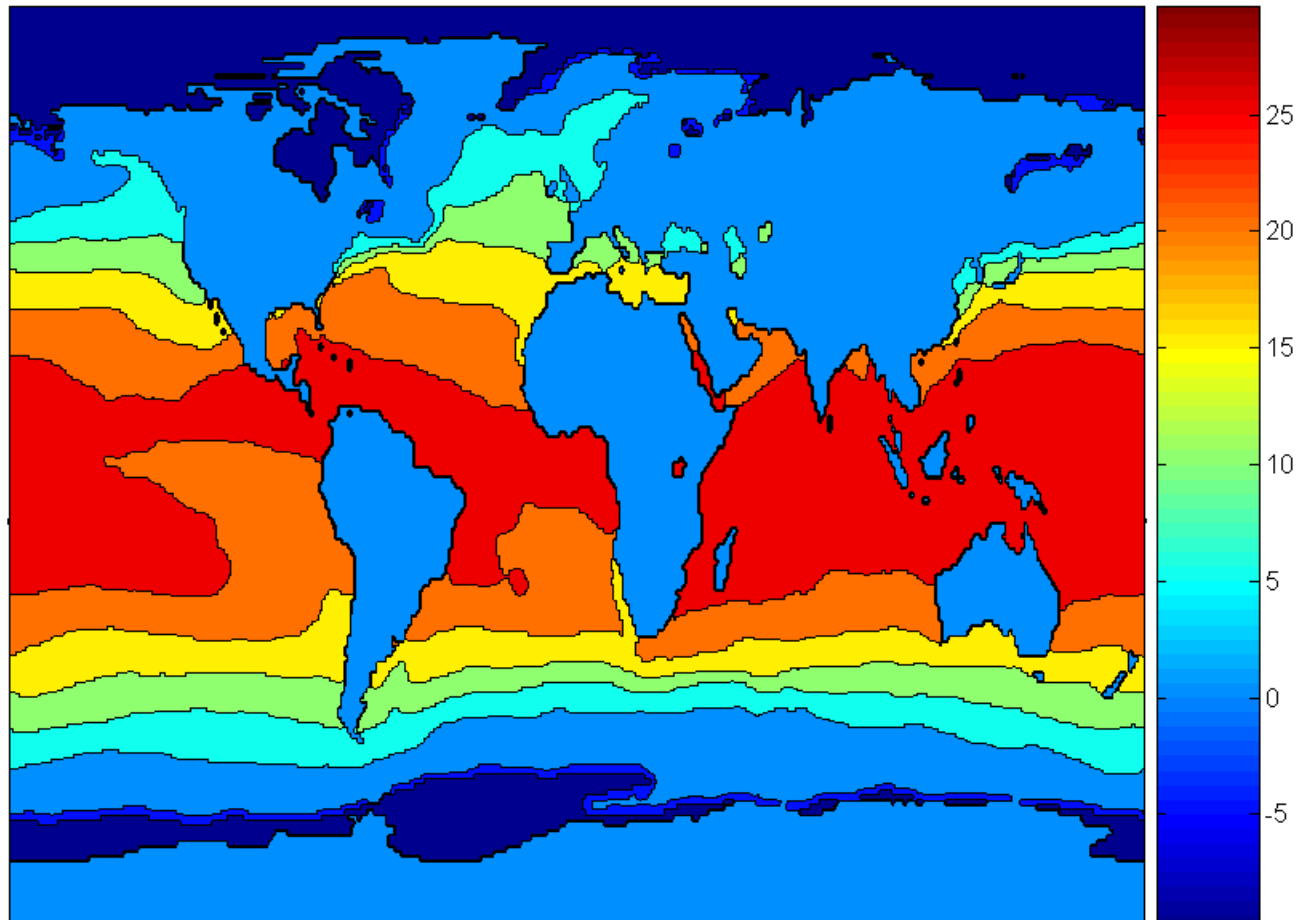
Post-processing

- Visualization
 - The human eye is a powerful analytical tool
 - If we visualize the data properly, we can discover patterns
 - Visualization is the way to present the data so that patterns can be seen
 - E.g., histograms and plots are a form of visualization
 - There are multiple techniques (a field on its own)

Scatter Plot Array of Iris Attributes



Contour Plot Example: SST Dec, 1998



Celsius

Meaningfulness of Answers

- A big data-mining risk is that you will “discover” patterns that are meaningless.
- Statisticians call it **Bonferroni's principle**: (roughly) if you look in more places for interesting patterns than your amount of data will support, you are bound to find crap.
- The **Rhine Paradox**: a great example of how not to conduct scientific research.

Rhine Paradox – (1)

- Joseph Rhine was a parapsychologist in the 1950's who hypothesized that some people had Extra-Sensory Perception.
- He devised (something like) an experiment where subjects were asked to guess 10 hidden cards – red or blue.
- He discovered that almost 1 in 1000 had ESP – they were able to get all 10 right!

Rhine Paradox – (2)

- He told these people they had ESP and called them in for another test of the same type.
- Alas, he discovered that almost all of them had lost their ESP.
- What did he conclude?
 - Answer on next slide.

Rhine Paradox – (3)

- He concluded that you shouldn't tell people they have ESP; it causes them to lose it.

