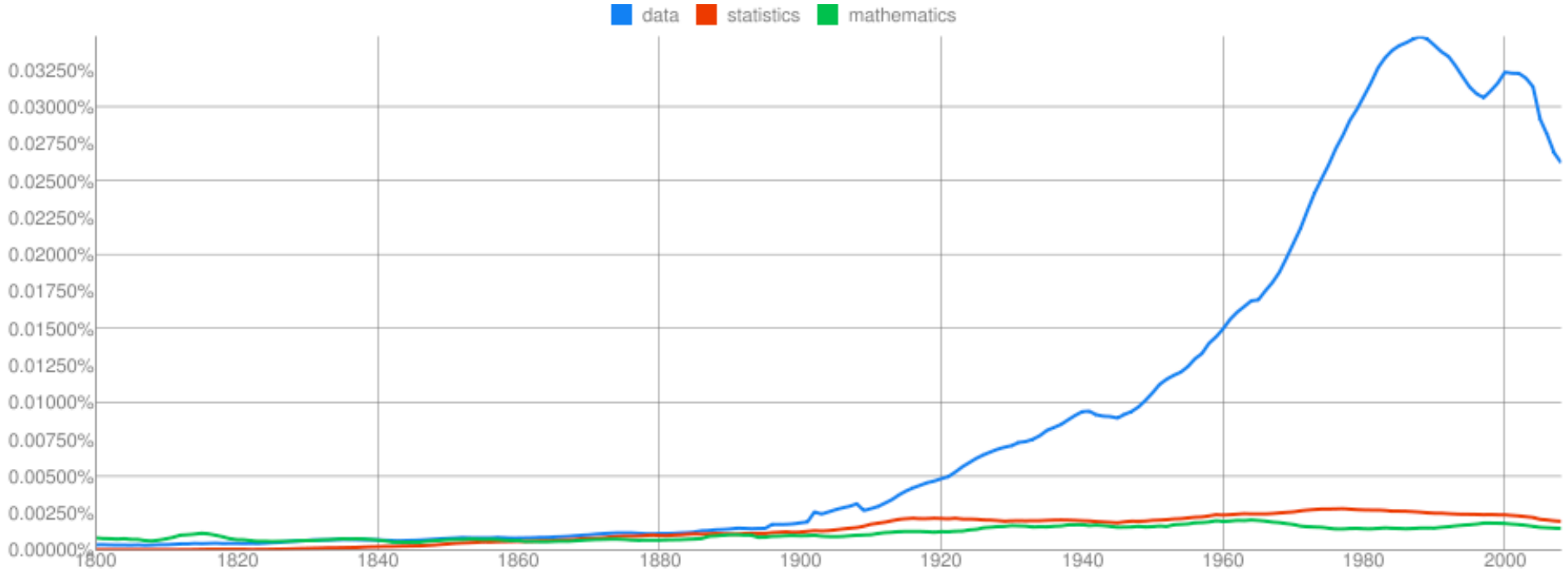


Lecture 16: Words as data

Pivot: A step back

So far in lab you've looked at a couple ways to derive confidence intervals -- One involving the quantiles of the bootstrap distribution, one involving the +/- two standard error approach, and then the classical t-statistic technique involving quantiles from the t-distribution

We are going to take a step back briefly and recall the small miracle that is the t-distribution -- Considering the simple case of the sample mean and its relationship to the population mean will help us understand regression technology a little better

Pivot: A step back

Recall that Gosset worked out the **sampling distribution of a standardized statistic, the t-statistic,** under the assumption that the data we've observed come form a population that is well described by a normal distribution

Under that assumption, he demonstrated that the quantity

$$\frac{\bar{x} - \mu}{s/\sqrt{n}}$$

has a sampling distribution that does not depend on anything but the number of observations that go into the sample mean and sample standard deviation (where μ is the unknown mean of the population)

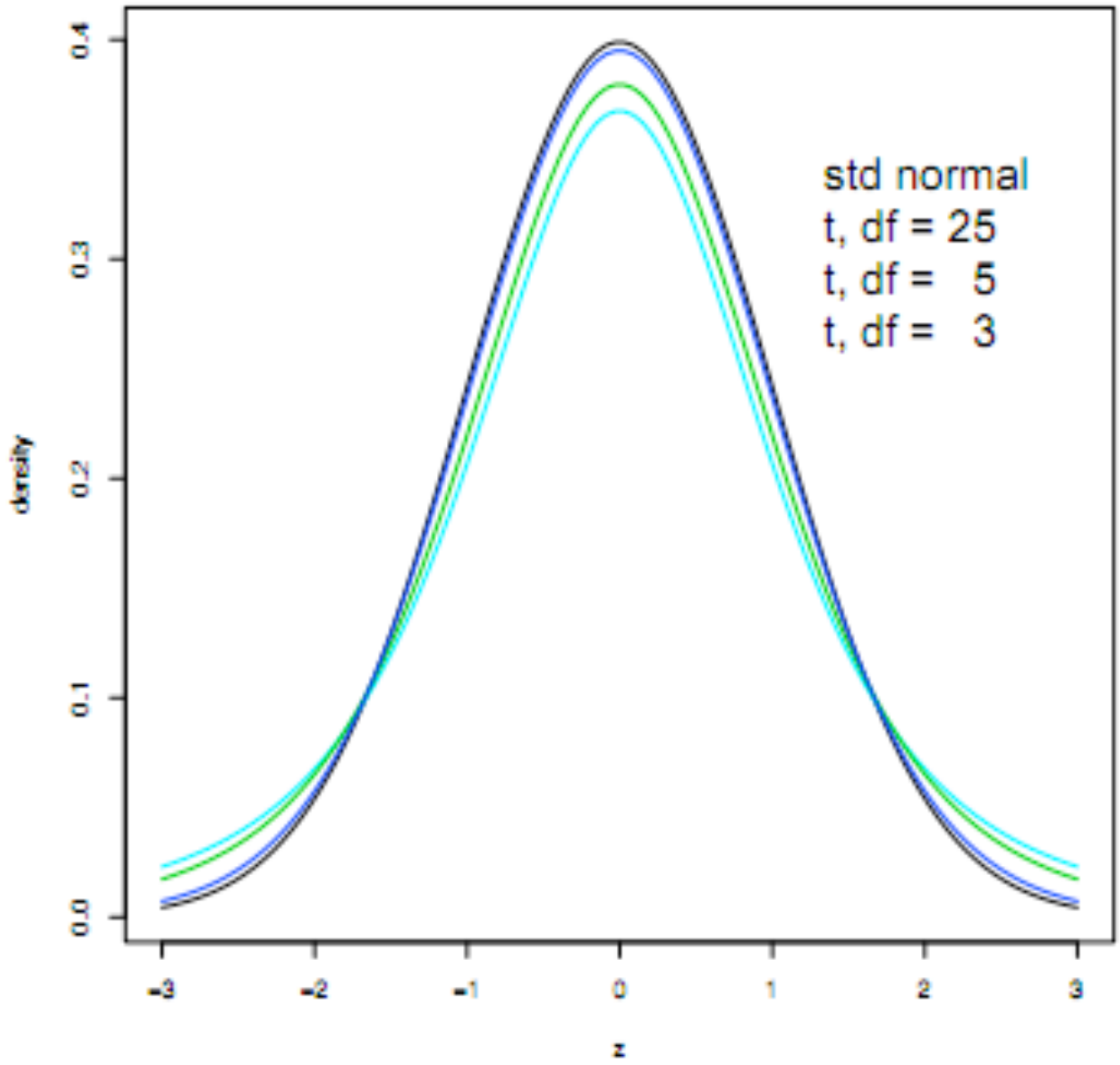
Pivot: A step back

In technical parlance, we refer to

$$\frac{\bar{x} - \mu}{s/\sqrt{n}}$$

as a pivotal quantity -- Its distribution can be written down explicitly (although we won't) and it depends only on sample size

We used this fact to help us form confidence intervals...


std normal
t, df = 25
t, df = 5
t, df = 3
density
z
0.0
0.1
0.2
0.3
0.4
-3
-2
-1
0
1
2
3

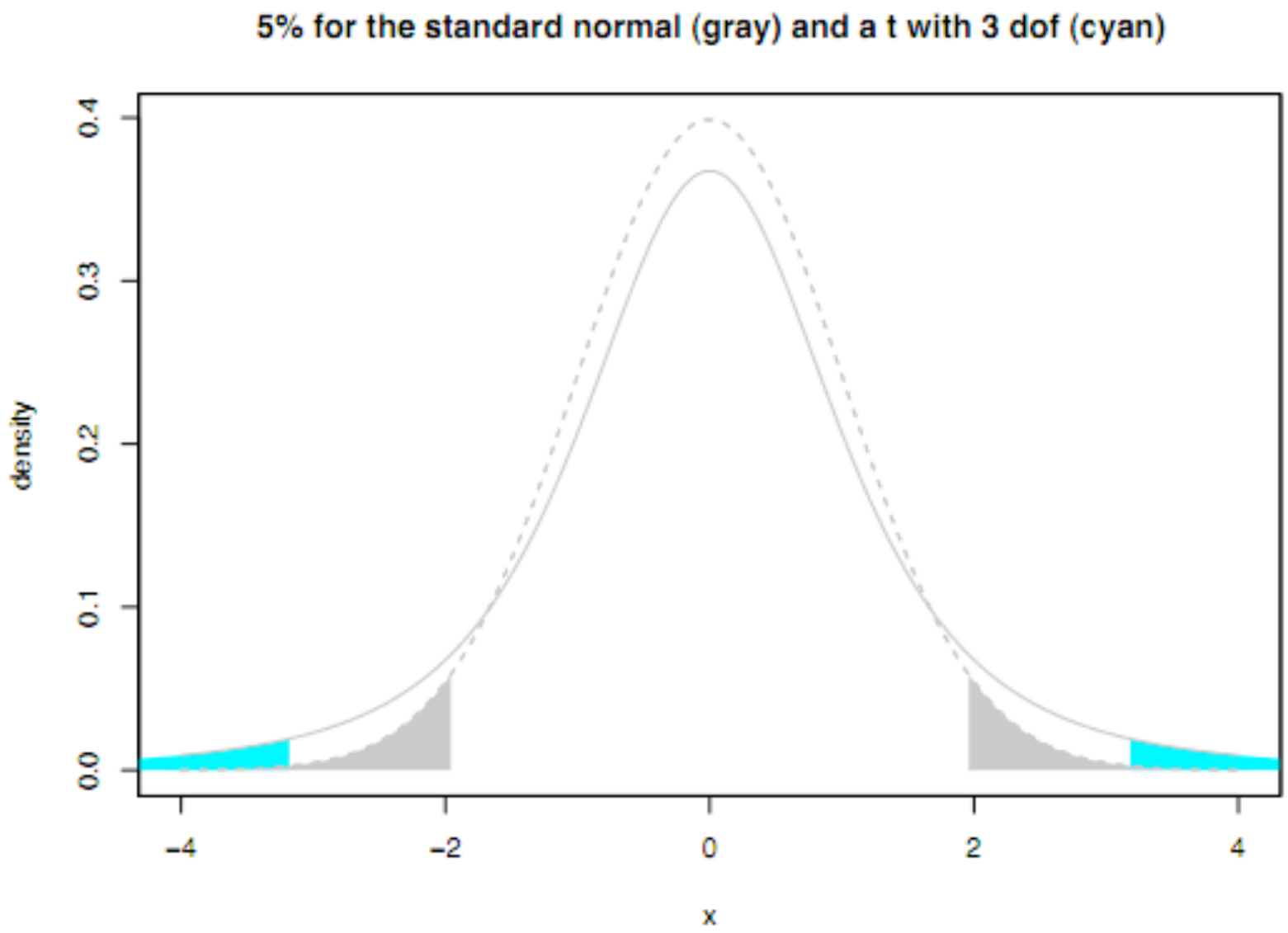
Pivot

If we know that

$$\frac{\bar{x} - \mu}{s/\sqrt{n}}$$

has a specific sampling distribution, we can find upper and lower points that define where we typically expect to see our observations (again, “typical” referring to lots of repeated experiments)

In this case, the points are the quantiles of the Student’s t-distribution...


5% for the standard normal (gray) and a t with 3 dof (cyan)
density
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4
x

Pivot

If we let q_t define the upper and lower points which contain 95% of our statistics (95% referring to 95 out of 100 times we repeated our experiment), then

$$\text{Prob}\left(-q_t \leq \frac{\bar{x}-\mu}{s/\sqrt{n}} \leq q_t\right) = 0.95$$

which we can simplify to give

$$\text{Prob}\left(-q_t\ s/\sqrt{n} \leq \bar{x}-\mu \leq q_t\ s/\sqrt{n}\right) = 0.95$$

or

$$\text{Prob}\left(\bar{x}-q_t\ s/\sqrt{n} \leq \quad \mu \leq \bar{x}+q_t\ s/\sqrt{n}\right) = 0.95$$

which is the basic line of reasoni g we followed for the t-based confidence intervals

A small, antique example

To illustrate the use of the t-distribution, we are going to go **right back to the source**; the reason is that it will provide us with a bridge back to our computational approach and a look at a more recent set of problems

The data we will consider were originally collected by Charles Darwin; his experiment involved 15 pairs of plants (*Zea Mays*, a corn plant) **descended from the same parents, having exactly the same age and having been subjected "first to last to the same conditions"**

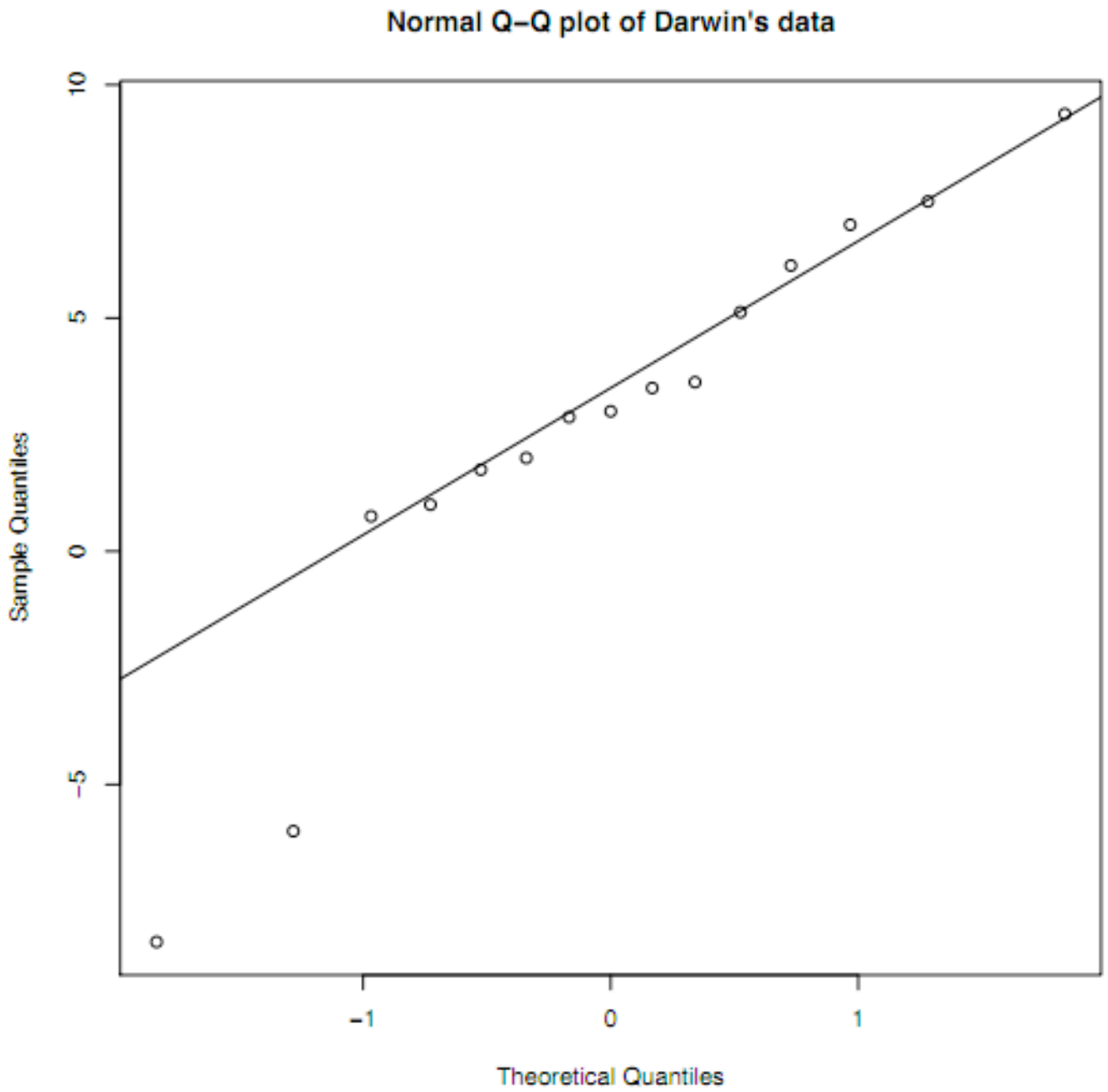
One individual from each pair was selected at random and cross-fertilized and the other self-fertilized; the **heights of the offspring** were then measured to the nearest eighth of an inch -- the results are on the next page

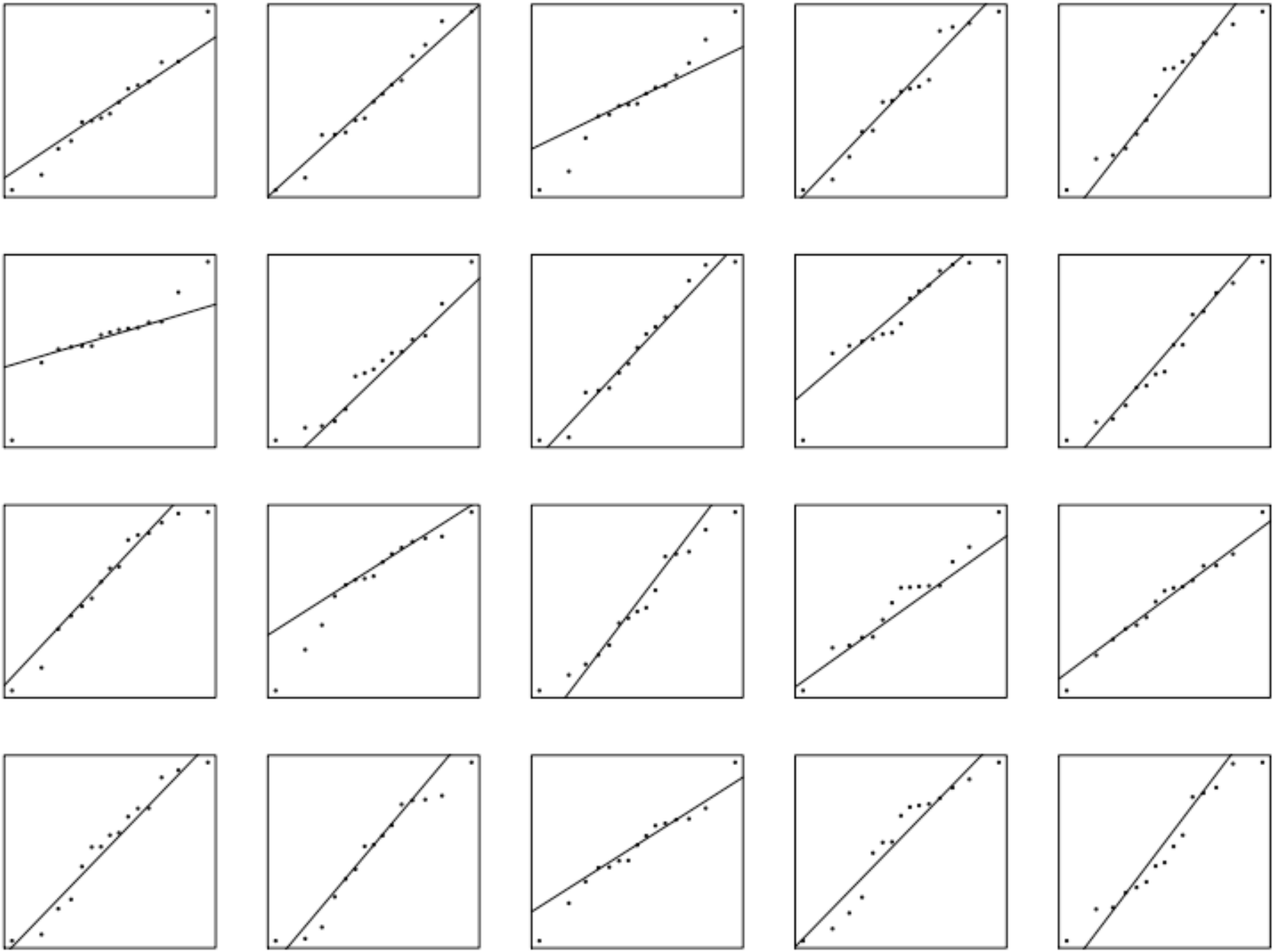
Pot	Crossed	Self-Fertilized	Difference
I	23.500	17.375	6.125
I	12.000	20.375	-8.375
I	21.000	20.000	1.000
II	22.000	20.000	2.000
II	19.124	18.375	0.749
II	21.500	18.625	2.875
III	22.125	18.625	3.500
III	20.375	15.250	5.125
III	18.250	16.500	1.750
III	21.625	18.000	3.625
III	23.250	16.250	7.000
IV	21.000	18.000	3.000
IV	22.125	12.750	9.375
IV	23.000	15.500	7.500
IV	12.000	18.000	-6.000

A small, antique example

Fisher analyzed these data by **first taking differences**; under the assumption that plant heights in the two groups (self- and cross-fertilized) were each normally distributed, so that their difference, in turn, **had a normal distribution**

On the next page we present a normal Q-Q plot for the 15 differences; what do you notice about these values? What do you think about the normality assumption?


Normal Q–Q plot of Darwin's data
Sample Quantiles
Theoretical Quantiles
10
5
0
−5
−1
0
1



A small, antique example

Fisher computed the sample mean difference to be 2.62 inches (the offspring of cross-fertilized plants being taller than the self-fertilized plants by 2.62 inches) and found that the sample standard deviation is s = 4.72

With a sample size of n=15, the multiplier for the t-distribution for a 95% confidence interval is

```
> qt(0.975,df=14)
[1] 2.144787
```

This gives a 95% confidence interval of $\bar{x} \pm 2.14\, s/\sqrt{n}$ or [0.004,5.229];

We've exhibited the interval with so many significant digits to make a point -- if the confidence interval maps out a range of plausible values for the true difference in heights between cross- and self-fertilized corn plants, then what does this interval suggest?

A small, antique example

The fact that the t-distribution depends only on sample size also provides us with a way to test hypotheses; here is a simple example

In the case of differences, there is a point we consider special, zero; what we have done in the previous slide is to, essentially, **interpret the size of a difference** (the size of an "effect") **in terms of its standard error** (or an estimate of the standard error)

That is, while 2.62 inches may seem like a lot for a small plant, we are judging it's size relative to its standard error $4.72/\sqrt{15} = 1.21$; in other words, the size of the **effect of cross- versus self-fertilization is 2.62/1.21 or 2.15 standard errors from 0**

We could ask, under the null hypothesis that the population mean is zero (that there is really no difference in heights between the offspring of cross- and self-fertilized plants), how likely are we to see an effect this large or larger?

A small, antique example

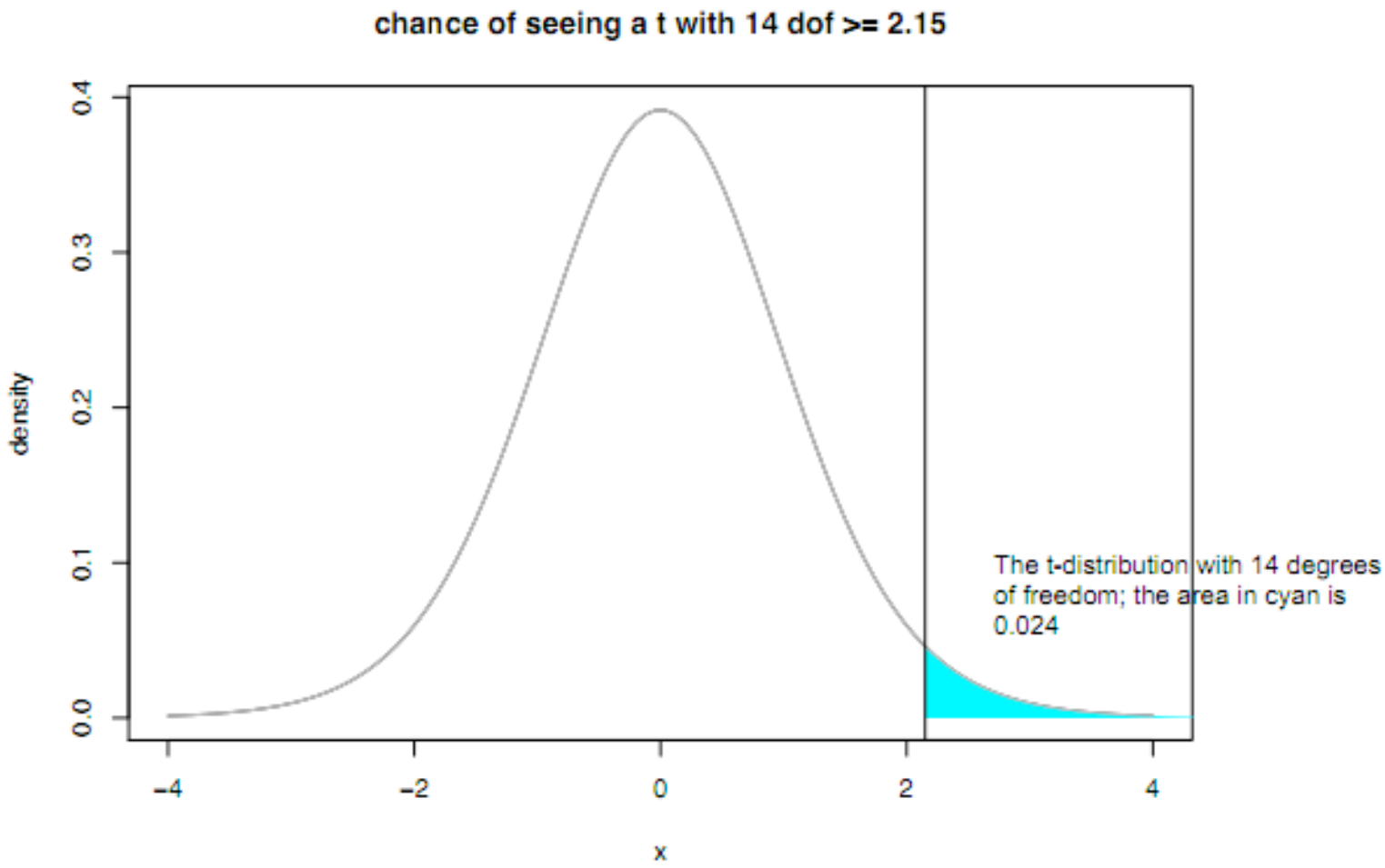
Remember that the quantity

$$\frac{\bar{x} - \mu}{s/\sqrt{n}}$$

depends only on sample size; so if we hypothesize $\mu = 0$ (in this case that the average difference in plant heights is zero), then the t-statistic

$$\frac{\bar{x}}{s/\sqrt{n}} = 2.15$$

should have a t-distribution with n-1 degrees of freedom; and so we can ask, what probability does the t-distribution assign to this event?


chance of seeing a t with 14 dof >= 2.15
density
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4
x
The t-distribution with 14 degrees of freedom; the area in cyan is 0.024

Some comments

In this case, we are using the t-distribution as our reference distribution for the test statistic for the null hypothesis that $\mu = 0$

$$\frac{\bar{x}}{s/\sqrt{n}}$$

It is known as a t-test (or a paired t-test since we're working with differences of matched pairs)

Testing and confidence intervals

Recall from last lecture that hypothesis tests and confidence intervals are looking for consistency between samples and population parameters, but they are coming at it from slightly different perspectives

Confidence intervals: Fix the (sample) statistic and ask what values of the population parameter are consistent with the fixed statistic

Hypothesis tests: Fix the population parameter value and ask what (sample) statistics are consistent with that fixed value

It turns out there is a one-to-one correspondence between tests and confidence intervals -- This dance with the t-distribution makes that clear

Testing and confidence intervals

Let [lo,hi] be a 95% confidence interval, say, for a population parameter θ -- Then for any θ_0 we can test the null hypothesis that $H_0 : \theta = \theta_0$, rejecting the null if θ_0 is not contained in [lo,hi]

The resulting test has significance level 0.05 -- In general, any $100(1-\alpha)$ percent confidence interval is equivalent to a test with significance level α

The logic works in reverse if we start with a hypothesis test and consider the set of values θ_0 for which we would fail to reject the null hypothesis that $H_0 : \theta = \theta_0$ -- These values form a confidence interval for the population parameter

Some comments

The last example we are using the **t-distribution as our reference distribution** for the test -- it is known as **a t-test** (or a paired t-test since we're working with differences of matched pairs)

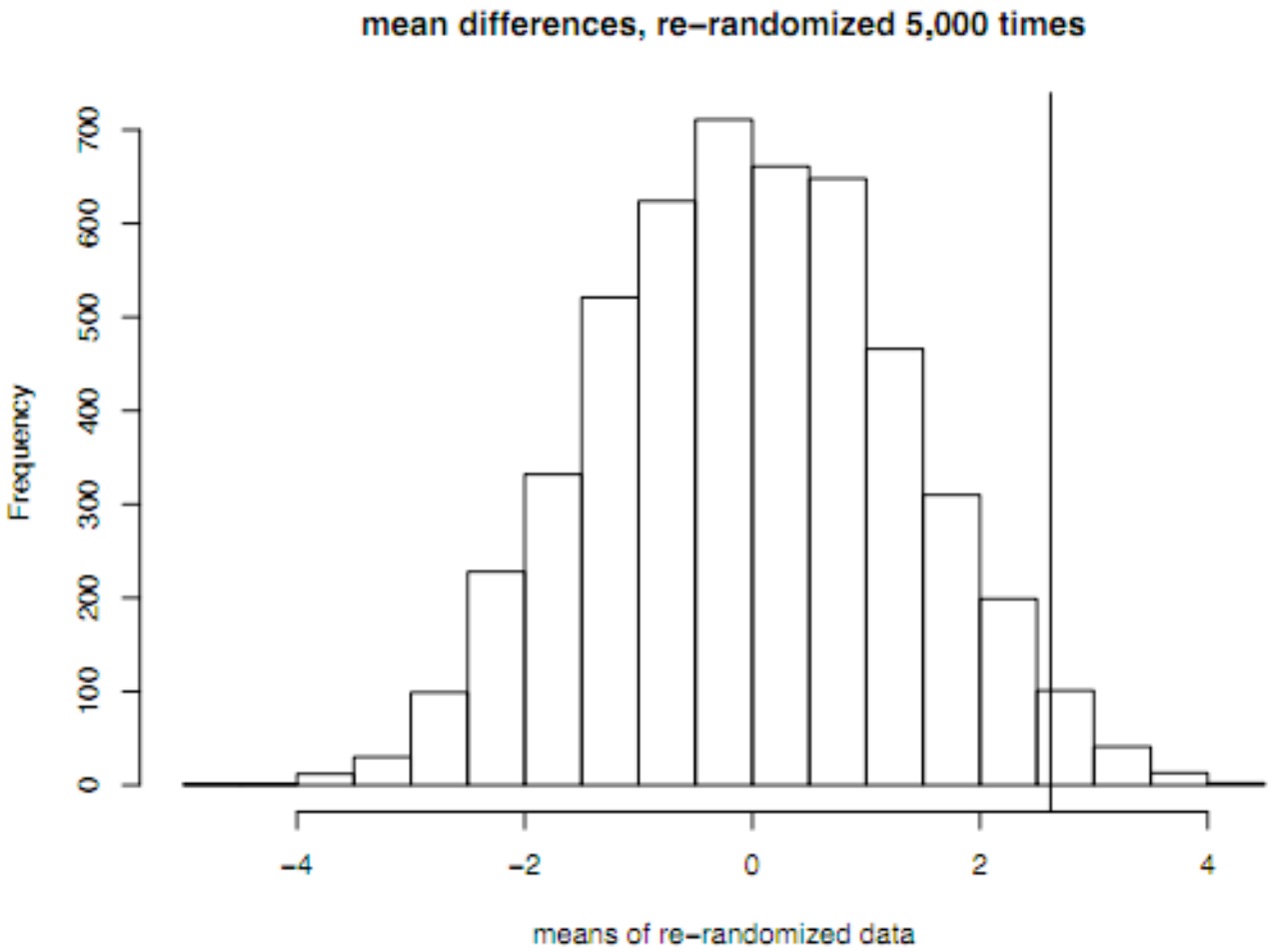
While this approach is classical, this is not how we would have approached the testing problem; instead, **we would have considered some form of re-randomization to generate a null distribution**

So... can we?

Some comments

Recall that the choice between cross- and self-fertilization was made at random; Fisher proposed re-randomizing each pair (so for each of the 15 rows in his original data table, we swap the two values on the toss of a coin) to come up with a null distribution for the sample mean

On the next slide we present the distribution of 5,000 re-randomized trials; the distribution looks fairly normal and we can (by now easily) compute a P-value...


mean differences, re−randomized 5,000 times
Frequency
0
100
200
300
400
500
600
700
−4
−2
0
2
4
means of re−randomized data

Some comments

The P-value we compute from 5,000 re-randomizations is 0.026, **in pretty good agreement with what we computed from the t-test**!

This is not surprising in that, along with the bootstrap and the other computational approaches we've been taking, **when the classical assumptions are met** (large sample sizes, or small sample sizes but normal populations) **the two approaches tend to agree**

The advantage of the computational procedure is that the same general principles ("analyze as you randomized") **apply to a variety of statistics and a range of applications**; the work we've presented on the previous 20 slides is all about one estimate, the sample mean

Back to the river

Last time, we performed a simple regression analysis to assess the relationship between Mercury levels and fish length for 98 fish taken from the Waccamaw river in North Carolina

We had a look at the R commands to perform the fit and even assessed the sampling variability using the bootstrap; we saw that the computationally intensive approach gave results that were close to what the classical formula told us (based on the t-distribution)

There are two dangling ideas I wanted to follow up on before we branched out into a different topic...

Prelude

In Lab this week, you will fit a linear model, and possibly a decision tree; we specify statistical models in R using its formula language -- Below we given an example for the linear model relating Mercury levels and length

```
> names(waccamaw)
[1] "river"   "station" "length"  "weight"  "mercury"

> fit = lm(mercury~length,data=waccamaw)
> summary(fit)

Call:
lm(formula = mercury ~ length, data = waccamaw)

Residuals:
     Min       1Q   Median       3Q      Max
-1.27784 -0.32696 -0.08177  0.31462  1.88604

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.450166   0.284687  -5.094 1.75e-06 ***
length       0.067510   0.006893   9.794 4.12e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5895 on 96 degrees of freedom
Multiple R-squared: 0.4998,   Adjusted R-squared: 0.4946
F-statistic: 95.92 on 1 and 96 DF,  p-value: 4.118e-16
```

Comparing the classics

The R output refers to t-statistics and tests of hypothesis for each of the coefficients in the regression equation (we even see multiple *'s to denote significance)

There are **close connections between the analytical approach** to the sampling distribution we mapped out for $\bar{x}$ and that for $\widehat{\beta}_0, \widehat{\beta}_1$; in both cases, we end up with a t-distribution when the data (or for least squares, when the errors) are normal or an approximate t for large samples

Comparing the classics

Sample mean $\bar{x}$

Approximately normal sampling distribution for large n

$\frac{\bar{x} - \mu}{\widehat{SE}}$ has a t-distribution with n-1 degrees of freedom when the data are normal

95% confidence intervals are of the form $\bar{x} \pm t^* \widehat{SE}$

$\frac{\bar{x}}{\widehat{SE}}$ can be used to test the null hypothesis that $\mu = 0$

Least squares estimates $\hat{\beta}_0, \hat{\beta}_1$

Approximately normal sampling distributions for large n

$\frac{\hat{\beta}_1 - \beta_1}{\widehat{SE}}$ has a t-distribution with n-2 degrees of freedom when the errors are normal

95% confidence intervals are of the form $\hat{\beta}_1 \pm t^* \widehat{SE}$

$\frac{\hat{\beta}_1}{\widehat{SE}}$ can be used to test the null hypothesis that $\beta_1 = 0$

The standard error, classically

There are three parameters in our population model -- The two regression coefficients (slope and intercept) $\beta_1,\ \beta_0$ and the error standard deviation σ

We estimate the regression coefficients by least squares giving us $\widehat{\beta}_1$ and $\widehat{\beta}_0$ and we can assess sampling variability using the bootstrap

The classical tools, however, follow what we did for the mean -- The estimate of the error standard deviation is simply

$$\widehat{\sigma} = \sqrt{\frac{\sum_i (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2}{n-2}}$$

and the classical estimate of the standard error of $\widehat{\beta}_1$, say, is

$$\frac{\widehat{\sigma}}{\sqrt{\sum_i (x_i - \bar{x})^2}}$$

The classics

Using the chart on the previous page, you see the link between testing and confidence intervals for the sample mean and testing and confidence intervals for the regression slope and intercept estimates

The bootstrap confidence intervals are providing a generalization of the classics when we're not sure the assumptions Gosset required hold -- And the confidence intervals can be inverted to provide us with an example of a bootstrap hypothesis test

But before the bootstrap, we were testing hypotheses in a different way...

A test

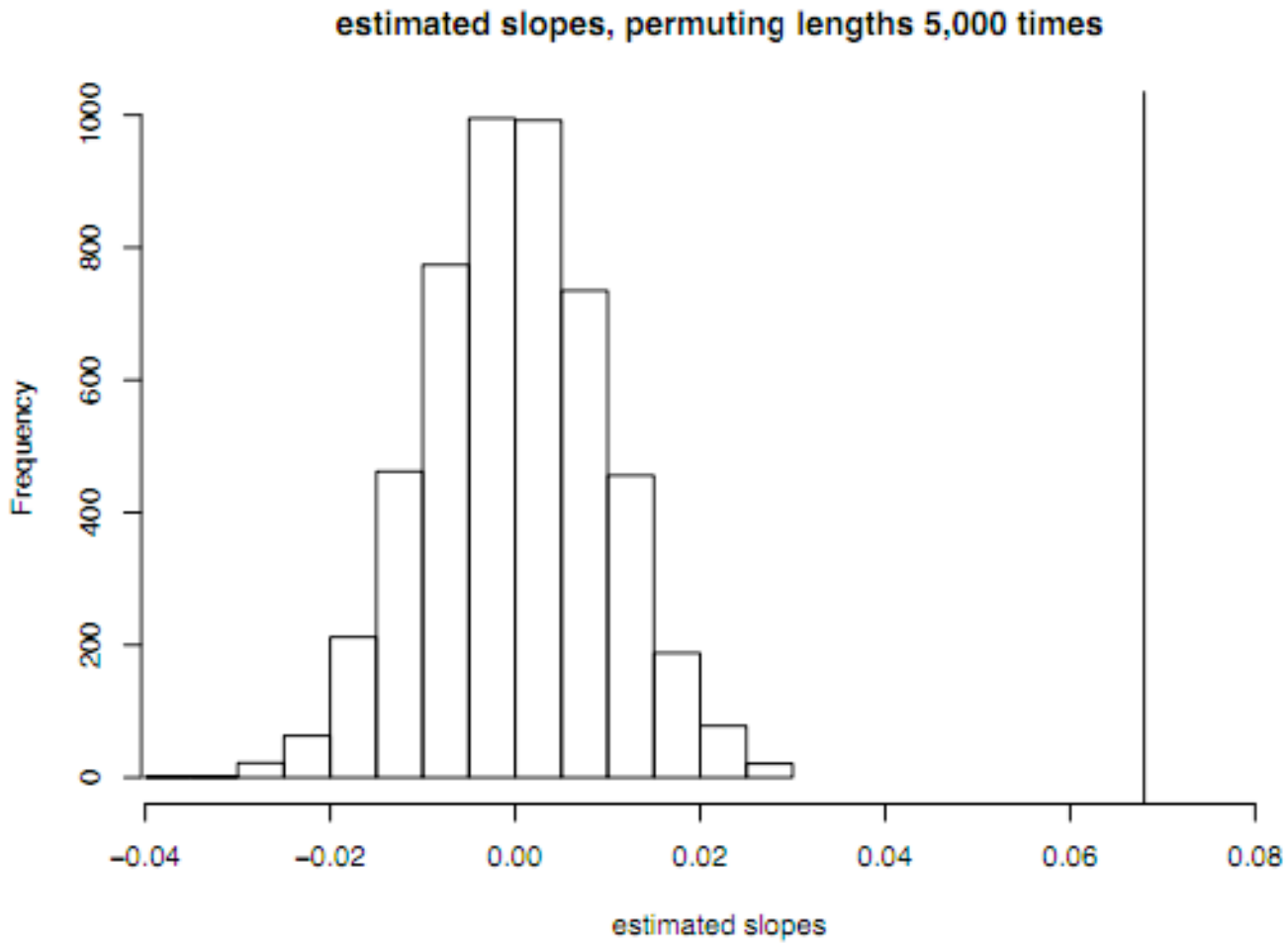
We could, of course, conduct a test along the lines of our earlier work on testing -- In particular, suppose we want to test the hypothesis that $H_0 : \beta_1 = 0$, or that the population parameter in a linear relationship relating Mercury levels and fish length is zero

Thinking back to the shuffling of labels that we started with at the beginning of the quarter, what special symmetry does this hypothesis suggest? What might we shuffle here?

A test

Recall that our data are a random sample of fish from the Waccamaw river; to say that $\beta_1 = 0$ is to suggest that **the distribution of Mercury levels in a fish is the same no matter what the length of the fish**

Therefore, under the null hypothesis, **we can randomly permute Mercury levels among the fish** to come up with a reference distribution for the test; for each permutation, we recompute the least squares fit and repeating this 5,000 times we get something like...

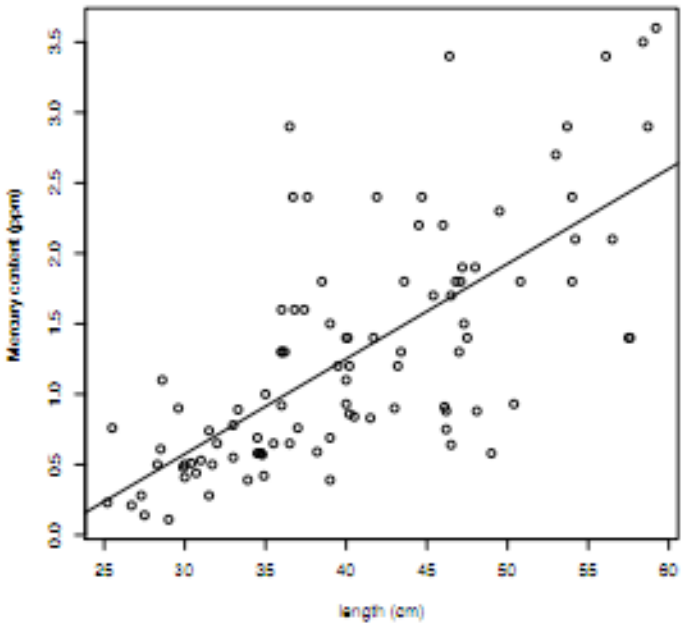

estimated slopes, permuting lengths 5,000 times
Frequency
0
200
400
600
800
1000
−0.04
−0.02
0.00
0.02
0.04
0.06
0.08
estimated slopes

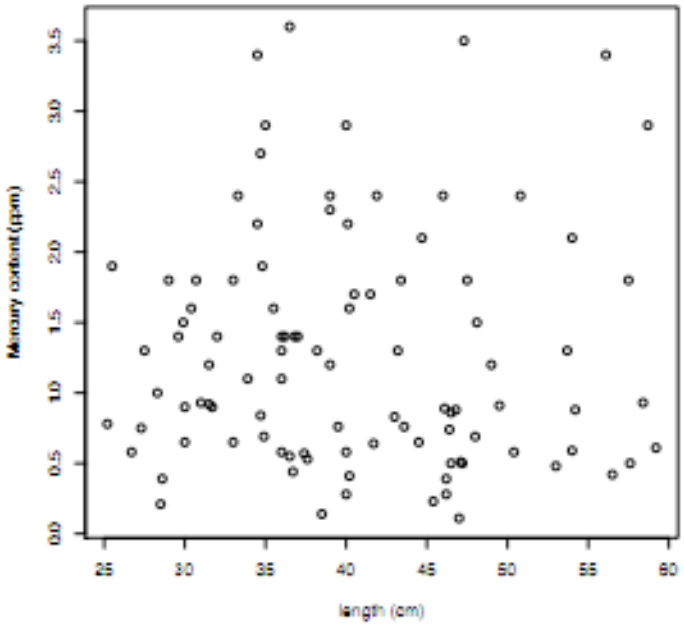
Hypothesis tests

This is a randomization test of the null hypothesis that the coefficient in a regression of Mercury content on fish length is zero, $\beta_1 = 0$

At the right we have a plot that might make the whole thing a little clearer -- In the top panel we have the relationship that we observed, the real data; in the lower panel, we have plotted one of the **relabeled data sets**

There is, by design, no relationship in the second between fish length and Mercury (r=0.006)





Testing

I present this alternative test (alternative to working with the bootstrap confidence interval) because it demonstrates **the richness of the tools you are learning** -- There are several different ways to think about the same problem and, for the most part, the answers should agree

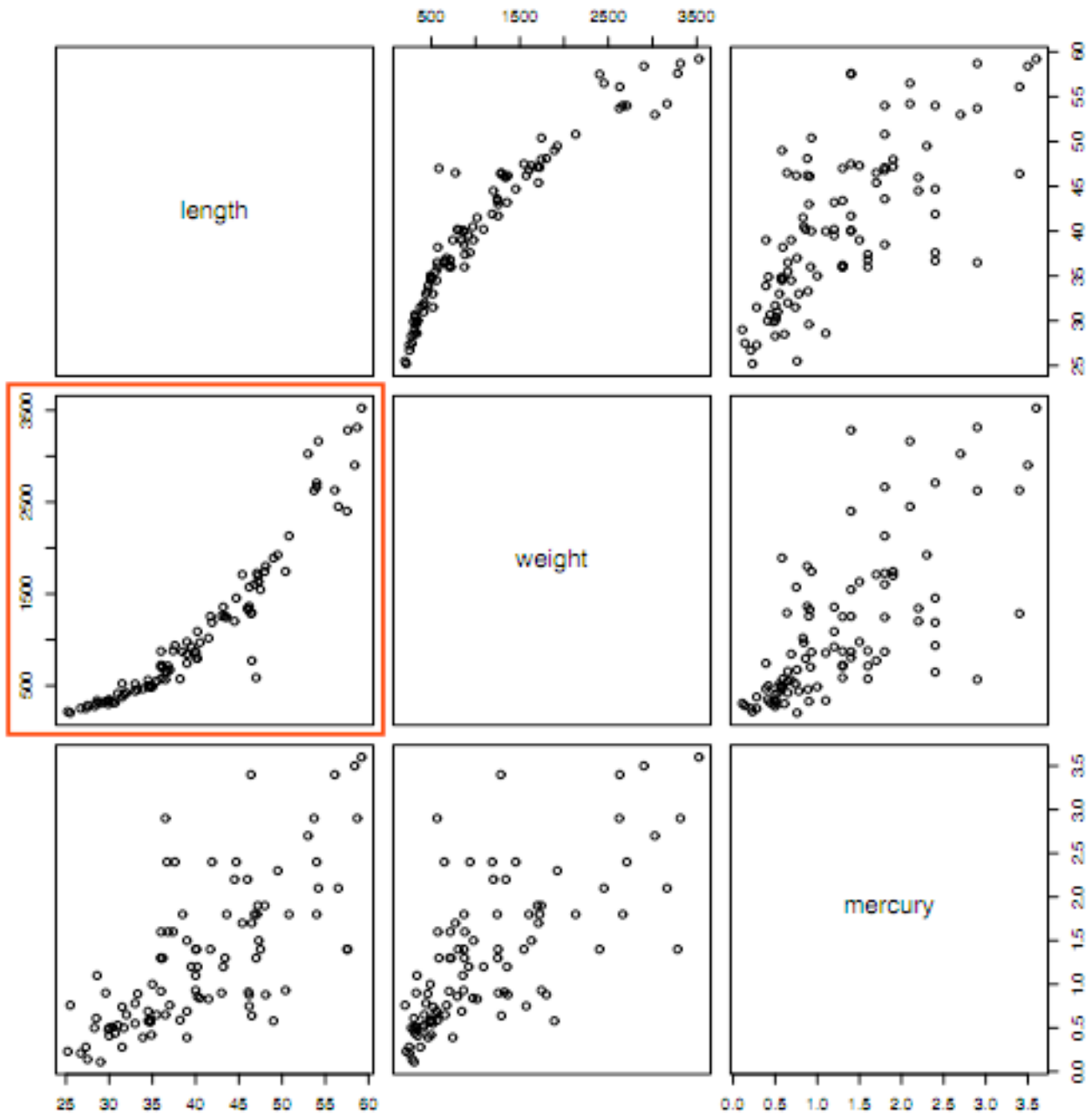
In general, if you are forming a test and there is something natural to permute (as in the case of randomized trials or this regression example), then you're better off making use of that structure for a test

In the classical case, you're reasoning from the t-distribution consistently, and so you don't have as many choices to make -- You are also somewhat limited in the problems you can address (think about the median example from your lab)

Back to the river

We’re going to next consider a second pair of variables -- It will introduce a small complication that will provide us with a richer view of regression analysis

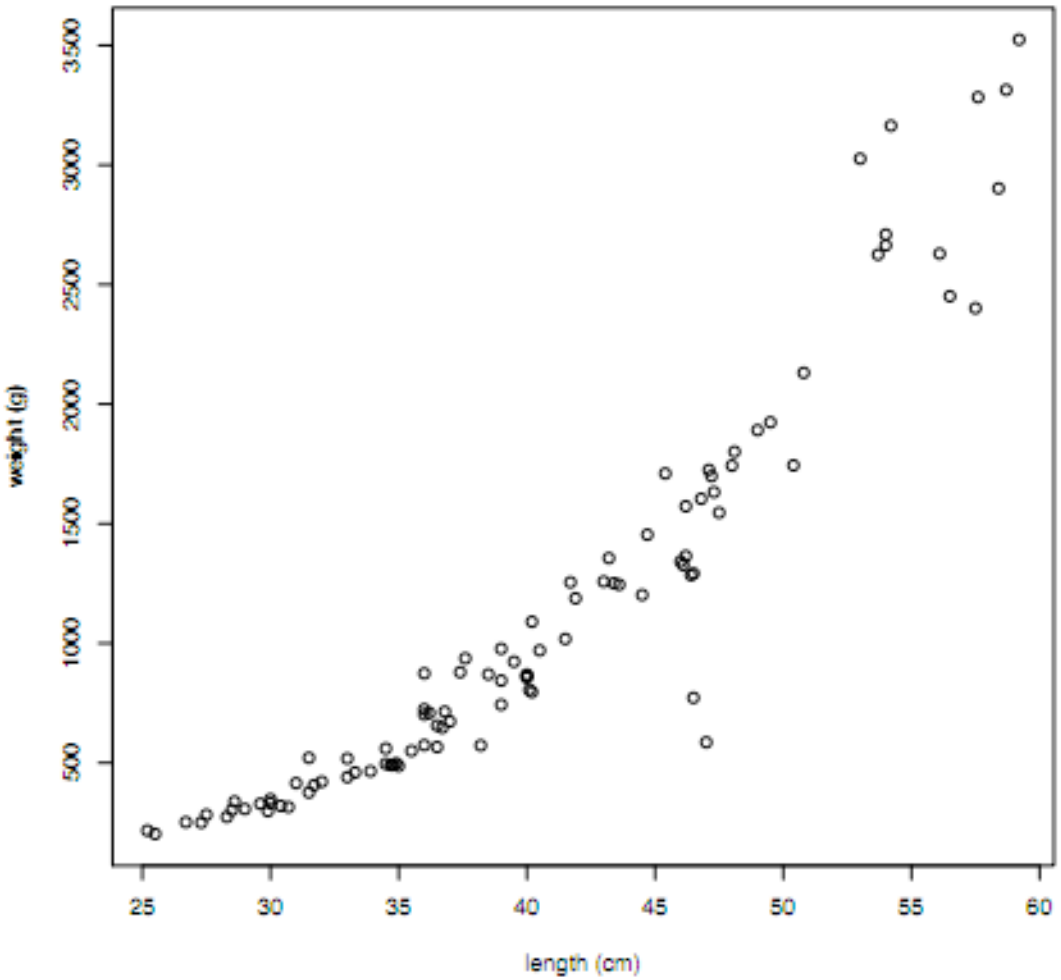
Let’s have a look at the relationship between fish length and fish weight -- Recall our scatterplot matrix...

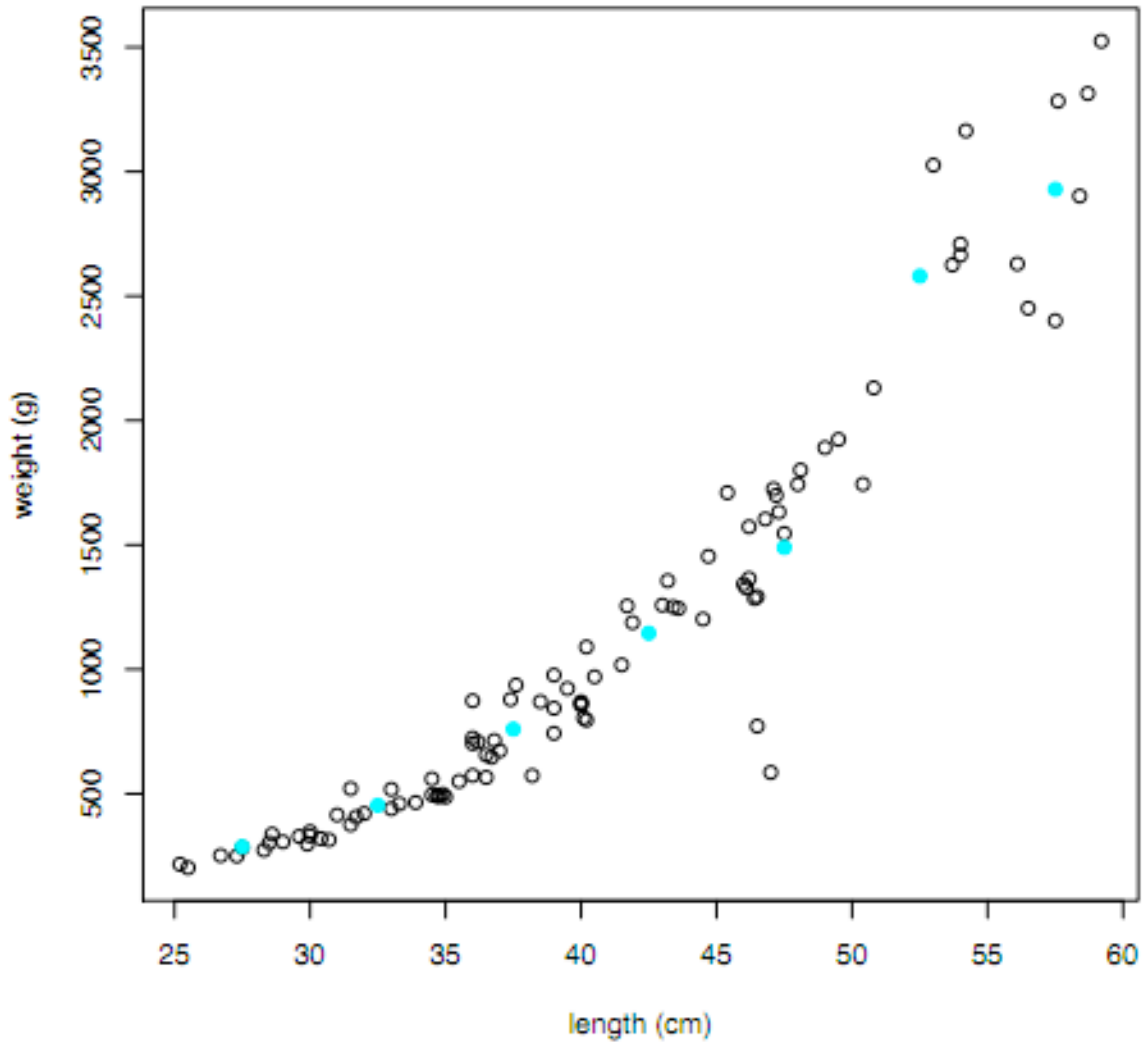

length
weight
mercury
500
1500
2500
3500
25
30
35
40
45
50
55
60
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5

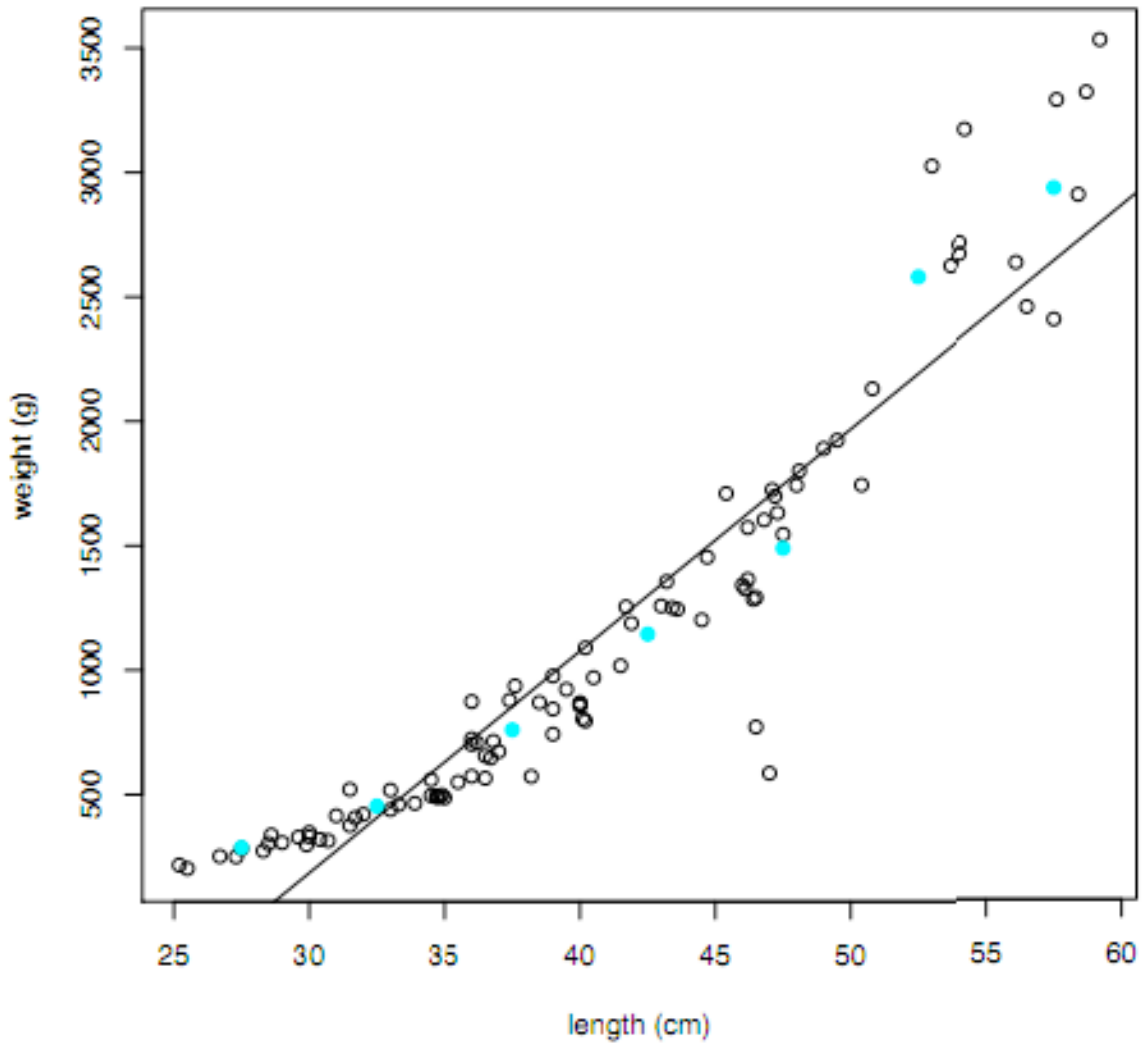
Weight

Either recalling Galton's mental image of data or our somewhat more practical application of least squares, **what do you think of this relationship?**

How would we model it mathematically? Will a line work?




weight (g)
500
1000
1500
2000
2500
3000
3500
25
30
35
40
45
50
55
60
length (cm)


weight (g)
500
1000
1500
2000
2500
3000
3500
25
30
35
40
45
50
55
60
length (cm)

Assessing the fit

When examining a regression model, there are several things to consider

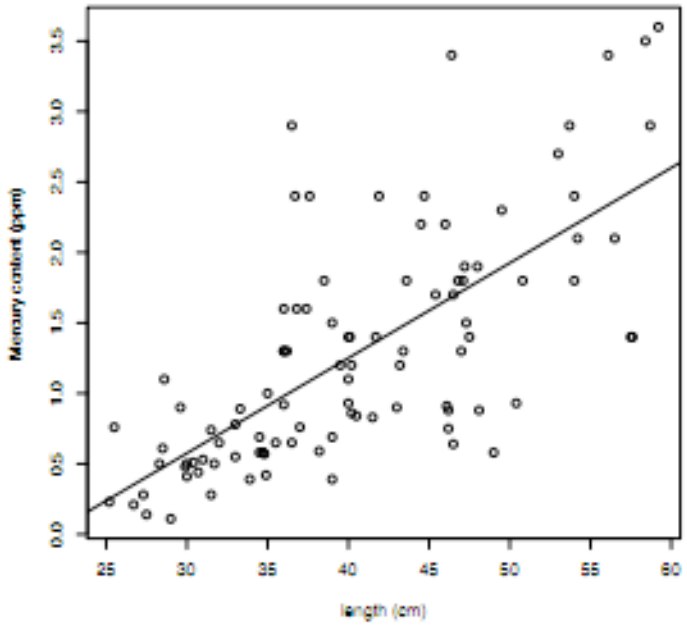
1. Is the relationship between outcomes and predictors linear?
2. Is the error variance constant?
3. Do the errors look roughly normally distributed?

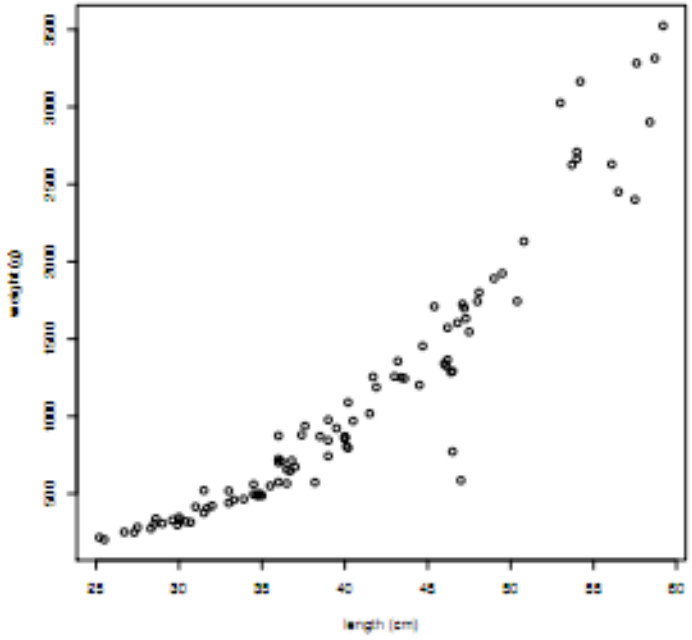
Regression analysis

For a simple linear regression with just one predictor variable, we can make scatterplots to assess the relationship between inputs and outputs easily

Assuming a linear relationship, then the errors from our least squares procedure should look like a sample from the normal distribution

This suggests other plots...

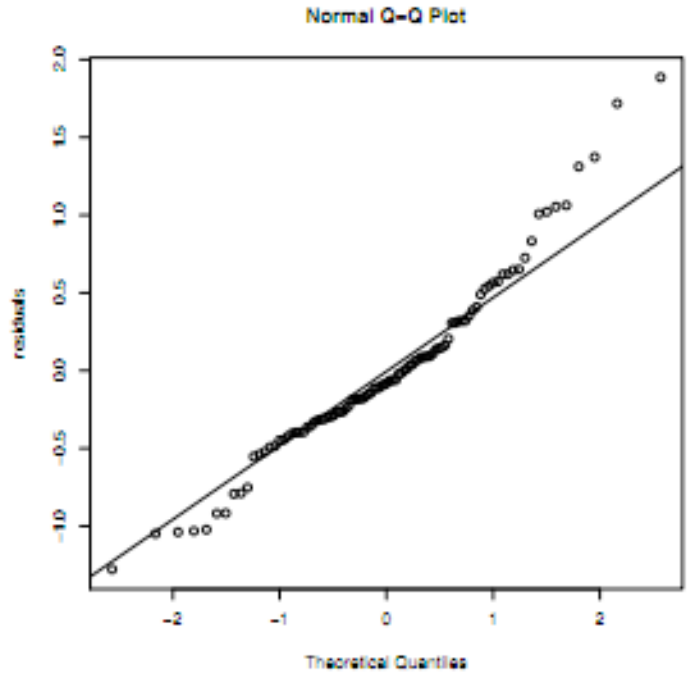


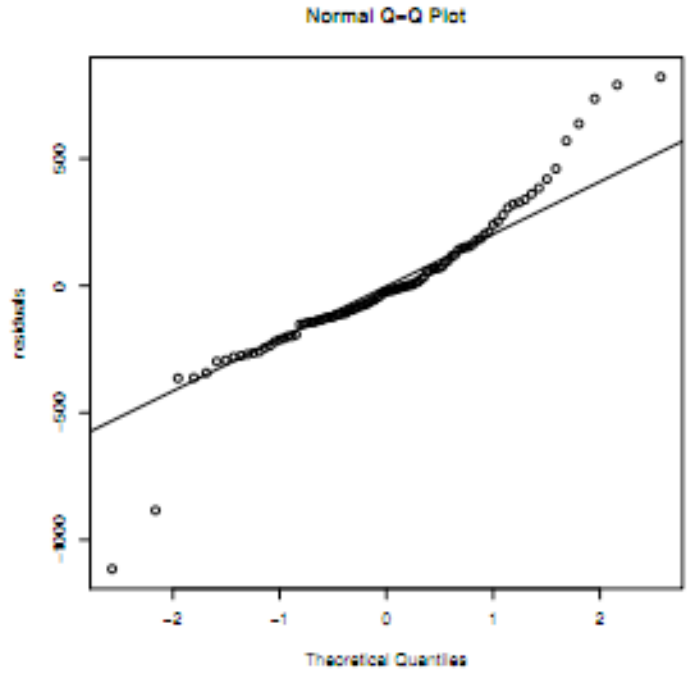


Residual analysis

Again, we want to inspect the residuals for “bends” which would indicate departures from normality

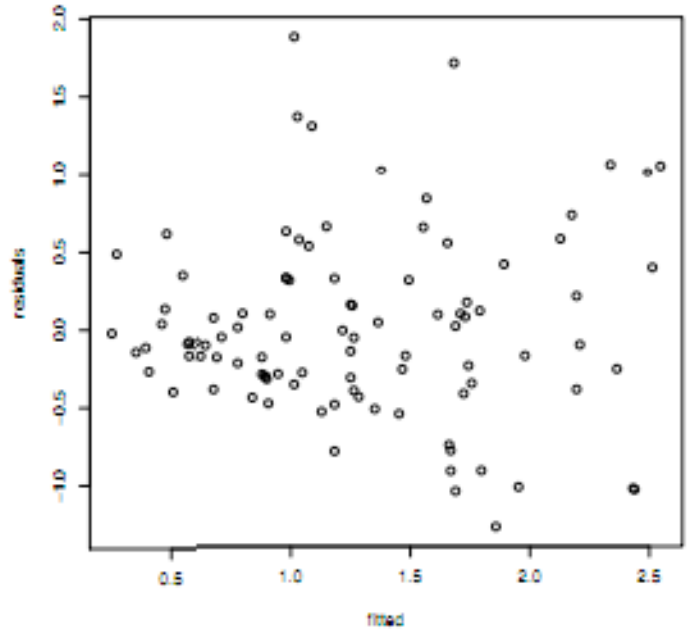
We should also examine the plots for large (positive or negative) values that might indicate outlying points





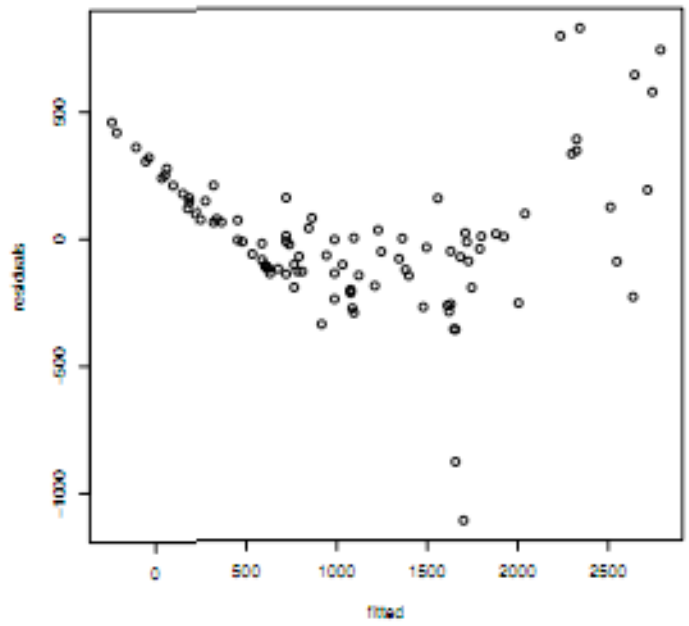
Residual analysis

Another informative plot compares the residuals to the fitted values (the points on the line); ideally **we should see no pattern here** - remember our model assumes the errors are independent normal observations



For the Mercury regression we see a slight indication of changing variability -- **for short fish we see less variation in the residuals than the long fish**

For the weight regression, we see that the fitted model **consistently overestimates (positive errors) the weights of small and large fish**, giving the plot a U-shape and suggesting a problem with the model



Polynomials

The relationship between weight and length is not linear and our basic inferential model breaks down when this assumption is violated (we no longer have just random errors from our model, but also considerable bias from the structural components we've left out)

Often, we consider fitting low-degree polynomials instead of just a line; on the next few slides, we go from a linear fit to a cubic -- the R command `poly()` returns a polynomial with the indicated degree

Polynomials

In terms of our model, we move from

$$(\text{weight}) = \beta_0 + \beta_1(\text{length}) + (\text{error})$$

to a model with one extra term

$$(\text{weight}) = \beta_0 + \beta_1(\text{length}) + \beta_2(\text{length}^2) + (\text{error})$$

We then choose estimates $\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2$ for the coefficients $\beta_0, \beta_1, \beta_2$ to minimize the sum of squared errors in the same way we did before

$$\sum_{i=1}^{n} \left[y_i - (\beta_0 + \beta_1 x_i + \beta_2 x_i^2)\right]^2$$

where we now denote the pairs of (length,weight) for each fish as $(x_i, y_i), i = 1, ..., 98$

Fitting a line

Below we provide the code to fit a line using the poly() function; this allows us to go from degree 1 to 2 to 3 easily; what do you see?

```
> fit = lm(weight~poly(length,1),data=waccamaw)
> summary(fit)

Call:
lm(formula = weight ~ poly(length, 1), data = waccamaw)

Residuals:
     Min       1Q   Median       3Q      Max
-1114.82  -141.65   -22.94   136.13   821.17

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      1111.22      28.86   38.50   <2e-16 ***
poly(length, 1)  7625.26     285.70   26.69   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 285.7 on 96 degrees of freedom
Multiple R-squared: 0.8812,   Adjusted R-squared:  0.88
F-statistic: 712.3 on 1 and 96 DF,  p-value: < 2.2e-16
```

Fitting a quadratic

Below we provide the code to fit a quadratic (including both length and $length^2$ in the model) relationship between weight and length; what do you see?

```
> names(waccamaw)
[1] "river"   "station" "length"  "weight"  "mercury"

> fit = lm(weight-poly(length,2),data=waccamaw)

> summary(fit)

Call:
lm(formula = weight - poly(length, 2), data = waccamaw)

Residuals:
     Min       1Q   Median       3Q      Max
-992.013  -49.733    3.498   87.098  684.114

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       1111.22      21.27   52.24  < 2e-16 ***
poly(length, 2)1  7625.26     210.58   36.21  < 2e-16 ***
poly(length, 2)2  1903.54     210.58    9.04 1.87e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 210.6 on 95 degrees of freedom
Multiple R-squared: 0.9362,   Adjusted R-squared: 0.9348
F-statistic: 696.5 on 2 and 95 DF,  p-value: < 2.2e-16
```

Fitting a cubic

Below we provide the code to fit a cubic (including both length, $length^2$ and $length^3$ in the model) relationship between weight and length; w you see?

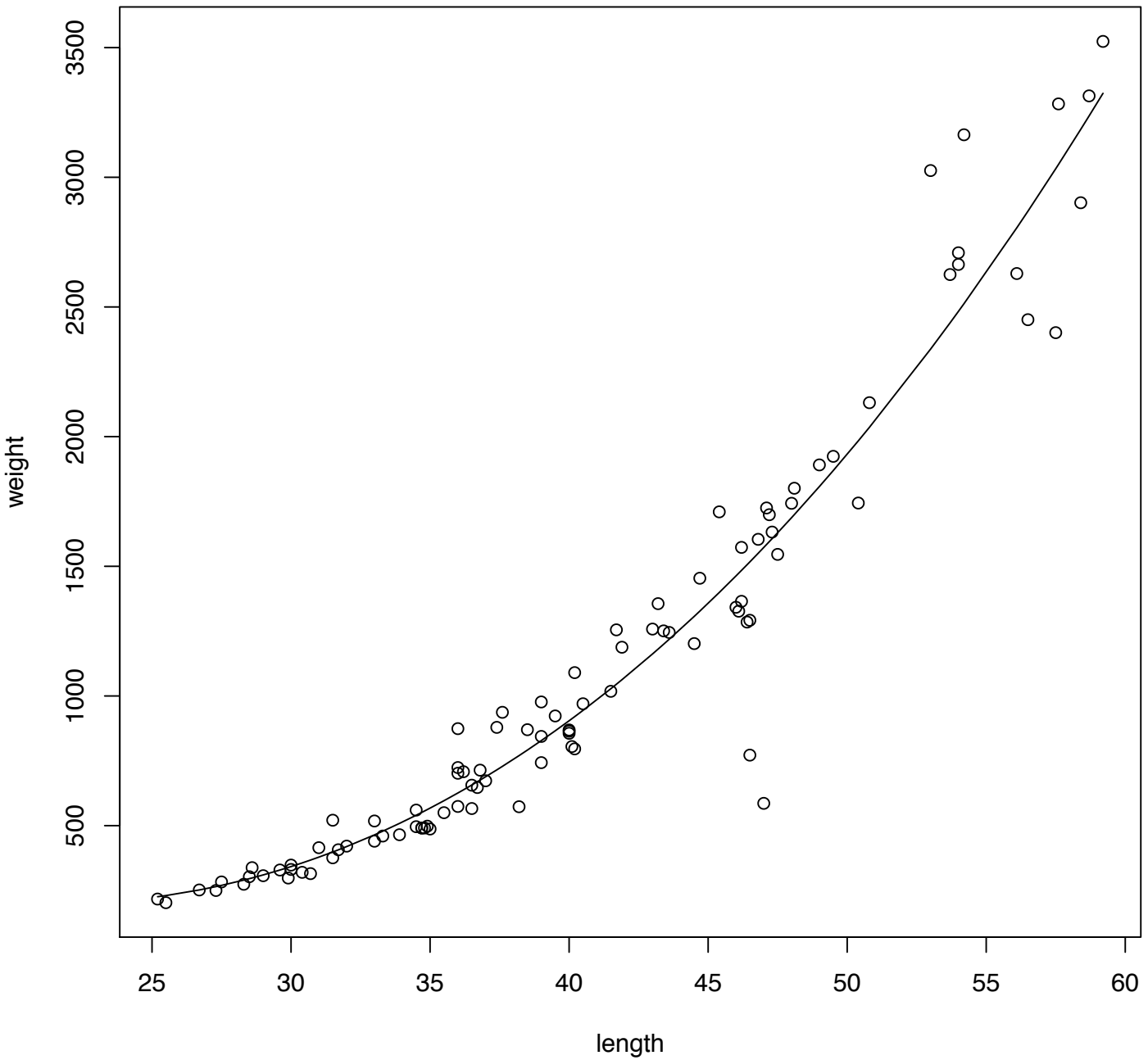
```
> fit = lm(weight-poly(length,3),data=waccamaw)
> summary(fit)

Call:
lm(formula = weight - poly(length, 3), data = waccamaw)

Residuals:
      Min        1Q    Median        3Q       Max
-986.6574  -52.1110    0.3027   87.4199  688.4783

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        1111.22      21.38  51.978  < 2e-16 ***
poly(length, 3)1   7625.26     211.64  36.030  < 2e-16 ***
poly(length, 3)2   1903.54     211.64   8.994 2.53e-14 ***
poly(length, 3)3     48.03     211.64   0.227    0.821
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 211.6 on 94 degrees of freedom
Multiple R-squared: 0.9362,   Adjusted R-squared: 0.9342
F-statistic: 459.7 on 3 and 94 DF,  p-value: < 2.2e-16
```

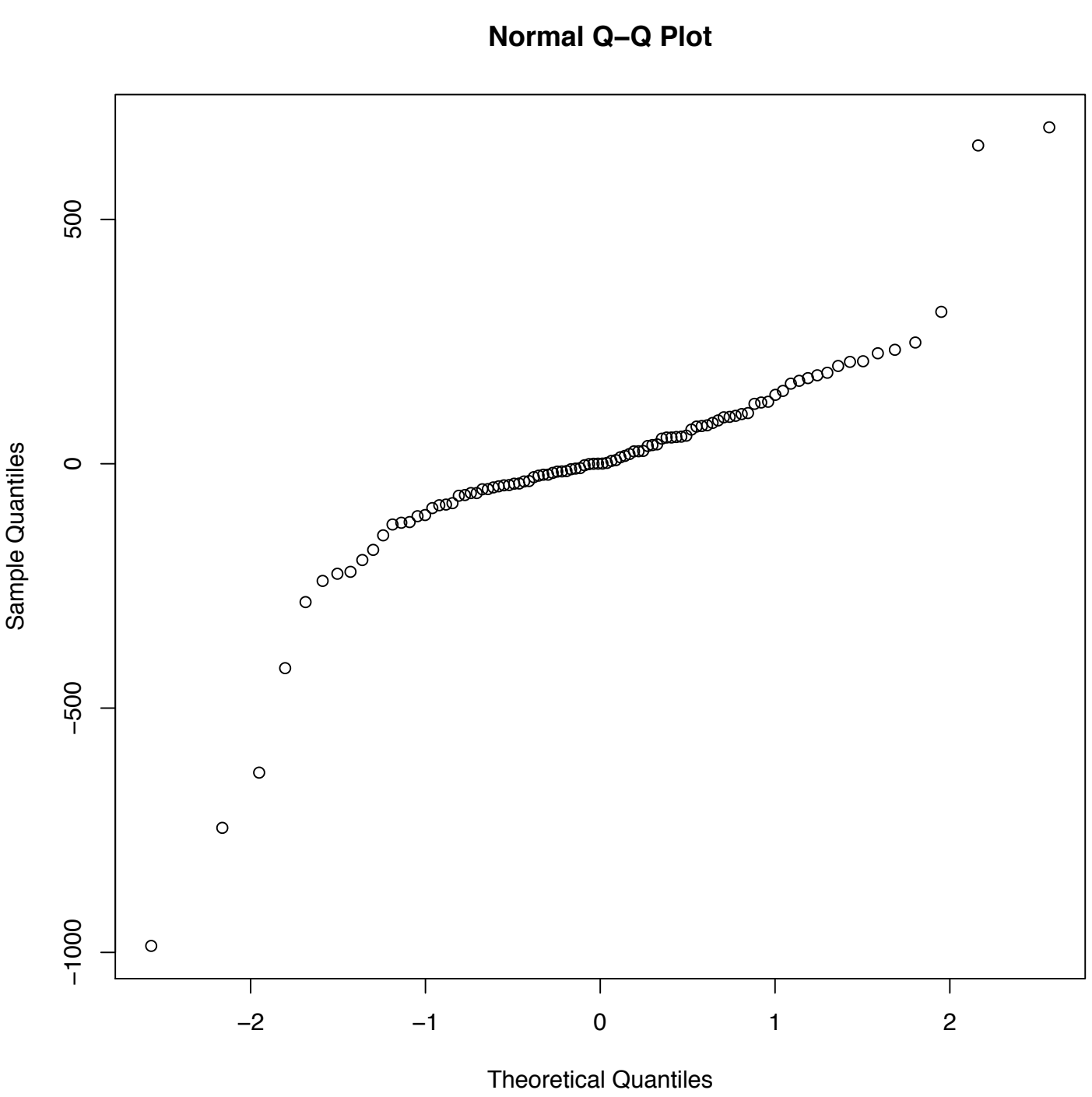
Fitting a cubic

In this case, the data seem to suggest a quadratic function is enough -- The extra cubic terms isn't "statistically significant", meaning its estimated value is a modest 0.27 standard errors away from zero (close enough to be judged unimportant)

By contrast, the linear and quadratic terms are 36 and 9 standard errors from zero, making them highly unlikely to be the result of chance

Here's the data and the fit, and then a Q-Q plot of the residuals...


weight
500
1000
1500
2000
2500
3000
3500
25
30
35
40
45
50
55
60
length


Normal Q–Q Plot
Sample Quantiles
Theoretical Quantiles
500
0
−500
−1000
−2
−1
0
1
2

Where to?

There are numerous open questions

1. How do you know you've included the most informative variables?
2. How do you decide on their functional form?
3. What do you do about missing data?
4. How do you handle changing variances?
5. What if my data are not normally distributed but are discrete counts or even binary?

A new data set

For the last couple of summers, I have done work related to climate change and the differential impacts that the accompanying extreme climate events have globally

Broadly, those least responsible for climate change are most impacted -- Several organizations have tried to quantify this statement, creating indices of various kinds to order countries according to their vulnerability to climatic events

Call and response

To get started on this topic, I sent out a handful of emails to try to acquire some of the base vulnerability data from researchers publishing indices -- Here's a typical note, with identifying references stripped

This is "the call"...

hi

i'm a professor in the statistics department at ucla... i'm writing to see if i might be able to get a copy of the data you assembled for your XXXXXX article on vulnerability to extreme climate events. i was specifically looking for the data referred to for your first regression in your XXXXXX section.

is that possible?

thanks!

M.

Mark Hansen | www.stat.ucla.edu/~cocteau

Call and response

In a graduate class I teach, we discussed how much of computational science is not particularly reproducible -- Two of the emails I received tell the story nicely...

Mark,

Sure. It has been a little while, and I am not 100% sure which files I used, but I am 99% sure that it was the sheet "late with population" in the attached excel spreadsheet, which ought to match what is in the STATA file. If this doesn't seem right, let me know, and I can spend a few more minutes hunting for the right data.

Cheers,

XXXXX

Dear Mark

This will take some digging through old files and backup disks as this was some time ago and my filing system is probably not what it should be. However, I'll try and find some time to do the necessary excavation. The data we used were all publicly available, and you should be able to reproduce the analysis based on the description of the methodology in the paper, if that's your interest. In any case feel free to give me a nudge in a week or so. When do you need to have the data, given your teaching schedule?

It would be good to have a professional statistician cast an eye over this, and the data we used were up to 2000, and could do with being updated. In any case I'm sure you'll find much to criticise! As someone who finds themselves having to deal with vulnerability indices I'm very sceptical of them, even the ones I've produced myself.

All the best

XXXXXX

Call and response

What's beautiful here is that you find two very different ways of working -- In one, the researcher is able to produce data almost immediately that (while not exactly right) got us really close to the published results

In the other case, there's a certain amount of hunting that has to take place -- I don't mean to fault this researcher in any way, I simply wanted to make the case that we should strive to be more like the first researcher (and better)

It's also worth noting the reception that statisticians get if we're not careful -- Writing to researchers outside of statistics often elicits a kind of fear that we're going to check up on them or criticize their work in some way

My (unwanted and highly biased) advice to you is to be the kind of statistician that tries to help!

PNAS PNAS PNAS

Estimating least-developed countries' vulnerability to climate-related extreme events over the next 50 years

Anthony G. Patt[a,1], Mark Tadross[b], Patrick Nussbaumer[c], Kwabena Asante[d], Marc Metzger[e,f], Jose Rafael[g], Anne Goujon[a,h], and Geoff Brundrit[i]

[a]International Institute for Applied Systems Analysis, 2361 Laxenburg, Austria; [b]Climate Systems Analysis Group, University of Cape Town, Rondebosch 7701, South Africa; [c]Institute of Environmental Science and Technology, Autonomous University of Barcelona, 08193 Bellaterra, Spain; [d]Climatus LLC, Mountain View, CA 94041; [e]Centre for the Study of Environmental Change and Sustainability, University of Edinburgh, EH8 9XP, Scotland; [f]Alterra, Wageningen University and Research Centre, 6700 AA Wageningen, The Netherlands; [g]Department of Geography, University of Eduardo Mondlane, Maputo, Mozambique; [h]Vienna Institute of Demography, Austrian Academy of Sciences, 1040 Vienna, Austria; and [i]Department of Oceanography, University of Cape Town, Rondebosch 7701, South Africa

Edited by Stephen H. Schneider, Stanford University, Stanford, CA, and approved December 4, 2009 (received for review September 10, 2009)

When will least developed countries be most vulnerable to climate change, given the influence of projected socio-economic development? The question is important, not least because current levels of international assistance to support adaptation lag more than an order of magnitude below what analysts estimate to be needed, and scaling up support could take many years. In this paper, we examine this question using an empirically derived model of human losses to climate-related extreme events, as an indicator of vulnerability and the need for adaptation assistance. We develop a set of 50-year scenarios for these losses in one country, Mozambique, using high-resolution climate projections, and then extend the results to a sample of 23 least-developed countries. Our approach takes into account both potential changes in countries' exposure to climatic extreme events, and socio-economic development trends that influence countries' own adaptive capacities. Our results suggest that the effects of socio-economic development trends may

sensitivity to those stressors, which in turn is determined by a complex set of social, economic, and institutional factors collectively described as determining its adaptive capacity (5, 6). As the UNFCCC secretariat suggested in its needs assessment, "one of the key limitations in estimating the costs of adaptation is the uncertainty about adaptive capacity. Adaptive capacity is essentially the ability to adapt to stresses such as climate change. It does not predict what adaptations will happen, but gives an indication of differing capacities of societies to adapt *on their own* to climate change or other stresses" (1, p. 97).

Human losses to extreme weather events can serve as a reliable indicator for this vulnerability, and with it the need for financial assistance, for two reasons. First, measures to reduce vulnerability to extreme weather events account for a particularly large share of estimated adaptation financial needs (1). Second, in the context of efforts to achieve a wide range of development goals, it is only

Vulnerability

The underlying question here is interesting and relevant (they usually are, for what it's worth) -- Here we are interested in understanding how climate change (and the accompanying increase in extreme weather events) will affect different parts of the world

Specifically, the researchers produce a model that relates variables capturing some notion of vulnerability to the impacts that weather-related natural disasters have had, country by country

Estimating least-developed to climate-related extreme 50 years

Anthony G. Patt[a,1], Mark Tadross[b], Patrick Nussbaumer[c], Kwa
Anne Goujon[a,h], and Geoff Brundrit[i]

[a]International Institute for Applied Systems Analysis, 2361 Laxenburg, Austria; [b]
South Africa; [c]Institute of Environmental Science and Technology, Autonomous
View, CA 94041; [d]Centre for the Study of Environmental Change and Sustainab
University and Research Centre, 6700 AA Wageningen, The Netherlands; [g]Depa
Mozambique; [h]Vienna Institute of Demography, Austrian Academy of Sciences
Cape Town, Rondebosch 7701, South Africa

Edited by Stephen H. Schneider, Stanford University, Stanford, CA, and approv

When will least developed countries be most vulnerable to climate change, given the influence of projected socio-economic development? The question is important, not least because current levels of international assistance to support adaptation lag more than an order of magnitude below what analysts estimate to be needed, and scaling up support could take many years. In this paper, we examine this question using an empirically derived model of human losses to climate-related extreme events, as an indicator of vulnerability and the need for adaptation assistance. We develop a set of 50-year scenarios for these losses in one country, Mozambique, using high-resolution climate projections, and then extend the results to a sample of 23 least-developed countries. Our approach takes into account both potential changes in countries' exposure to climatic extreme events, and socio-economic development trends that influence countries' own adaptive capacities. Our results suggest that the effects of socio-economic development trends may begin to offset rising climate exposure in the second quarter of the century, and that it is in the period between now and then that vulnerability will rise most quickly. This implies an urgency to the need for international assistance to finance adaptation.

vulnerability | adaptive capacity | development | natural disasters | natural hazards

Results

The first stage of our analysis was to estimate statistical models of losses from climate-related disasters, based on a set of climatic and socio-economic variables that will likely change over time, which appear in Table 1. The dependent variables are logged values of the number of people per million of national population killed or affected, respectively, by droughts, floods, or storms over the period 1990–2007. The variable number of disasters is the logged value of numbers reported by each country over the same period, and accounts for climate exposure; estimated coefficient values greater than 1 in both models indicate that average losses per disaster are higher in more disaster-prone countries. We expected that larger countries are likely to experience disasters over a smaller proportion of their territory or population, and also benefit from potential economies of scale in their disaster management infrastructure, both resulting in lower average per capita losses; the negative coefficient estimates for the variable national population in both models are consistent with this expectation. The variable HDI represents the Human Development Index, a United Nations (UN) indicator comprised of per capita income, average education and literacy rates, and average life expectancy at birth. Recent studies of disaster losses —not limited to climate-related events—have shown that countries with medium HDI values experience the highest average losses, whereas countries with high HDI values experience the lowest (14, 15). We therefore included the logged HDI values in quadratic form. Negative coefficient estimates for both HDI and HDI^2 in both models are thus consistent with these expectations, given that logged HDI values are always negative, and the square of the logged values are in turn positive. Finally, we considered several additional socio-economic variables not directly captured by HDI, and found only two that improved model fit. For the model of the number of people killed, the positive coefficient estimate for female fertility indicates that countries with higher birth rates experience greater average numbers of deaths. We do not take this to mean that there is a direct connection between fertility and natural hazard deaths, but rather that higher birth rates are associated with lower female empowerment, and lower female empowerment is associated with higher disaster vulnerability, as has been shown previously (16, 17). For the model of the number of people affected, the negative coefficient estimate for the proportion urban population is consistent with urban residents being less likely to require post-disaster assistance than rural residents, also observed previously (18, 19). Both models yield an R^2 statistic slightly greater than 0.5, indicating that variance in the independent variables explains just over half of the variance in the numbers killed and affected. This is consistent with results from past analyses based on similar data and methods (8–10).

Vulnerability

In the end, a great deal of attention is paid to a regression table (below), the form of which we should be fairly familiar with

In each row they present the regression of the logarithm of the number of people killed by weather-related natural disasters from 1990 to 2007 as a function of several predictors, one of which is slightly special...

Table 1. Ordinary least-squares regression results

Independent variables	Killed	Affected
Number of disasters	1.36* (0.15)	1.88* (0.19)
National population	−0.56* (0.09)	−0.79* (0.11)
HDI	−5.97* (1.95)	−13.55* (2.16)
HDI^2	−6.26* (1.52)	−9.82* (1.86)
Female fertility	1.45* (0.43)	
Proportion urban population		−0.41 (0.37)
Constant	−3.86* (0.49)	5.33* (1.71)
Number of observations	150	154
R^2	0.52	0.55

The dependent variable in the Killed model is the logged value of the number of people reported by CRED as killed by the three types of disasters considered (droughts, floods, and storms) divided by population. The dependent variable in the *Affected* model is the same for the number of people reported affected, but not killed, by the same disasters. All independent variables are logged values. Because HDI occupies the range of 0–1, all logged HDI values were negative, whereas the squares of these values were positive. *Values significant (*two-tailed student's t* test) at the 99% confidence level. Values in parentheses are SEs.

HDI

The HDI or Human Development Index, a United Nations (UN) is an “indicator comprised of per capita income, average education and literacy rates, and average life expectancy at birth”

From the table on the previous page, we see that the variable and its square are both included in the final model and are given the following interpretation

“Of particular importance to rapidly developing countries is the observed nonlinear relationship between HDI and disaster losses. Fig. 1 illustrates the magnitude of this effect in both models, compared with the background variance, and taking into account the effects of the other variables. The estimated regression curve in Fig. 1A suggests that the risk of being affected by a climate disaster is highest in countries with HDI values of ~0.5, whereas the curve in Fig. 1B suggests that the highest risk level is for countries with HDI values somewhat higher, ~0.6. This suggests that for countries with HDI values of less than 0.5, the transition to higher levels of development could potentially, in the absence of targeted intervention, exacerbate vulnerability.”

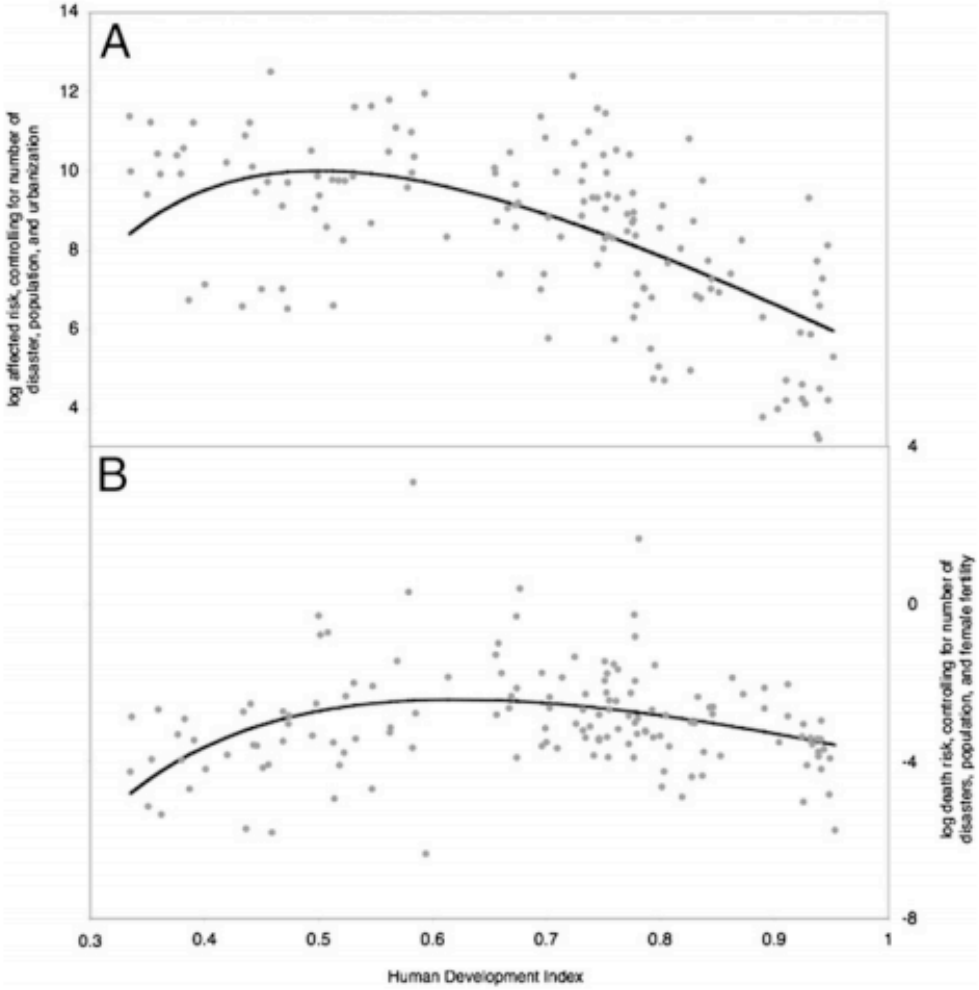


Fig. 1. Relationship between risk and HDI for (*A*) the number of people affected, i.e., needing emergency or recovery assistance, by a flood, drought, or cyclone, per million of population, and (*B*) the number of people killed. Each dot represents a country in the CRED database during the period 1990–2007, with its position on the vertical scale being the logarithm of the annual value per million population, after subtracting the predicted influence of other risk factors. Regression line in each figure shows predicted values including the influence of HDI.

Looking at the data

The data we were given consist of measurements associated with 144 different countries -- For each we have the following variables

`country_name` the name of the country

`ln_events` the natural logarithm of the number of droughts, floods and storms occurring in the country from 1990-2007

`ln_pop` the natural logarithm of the country's population

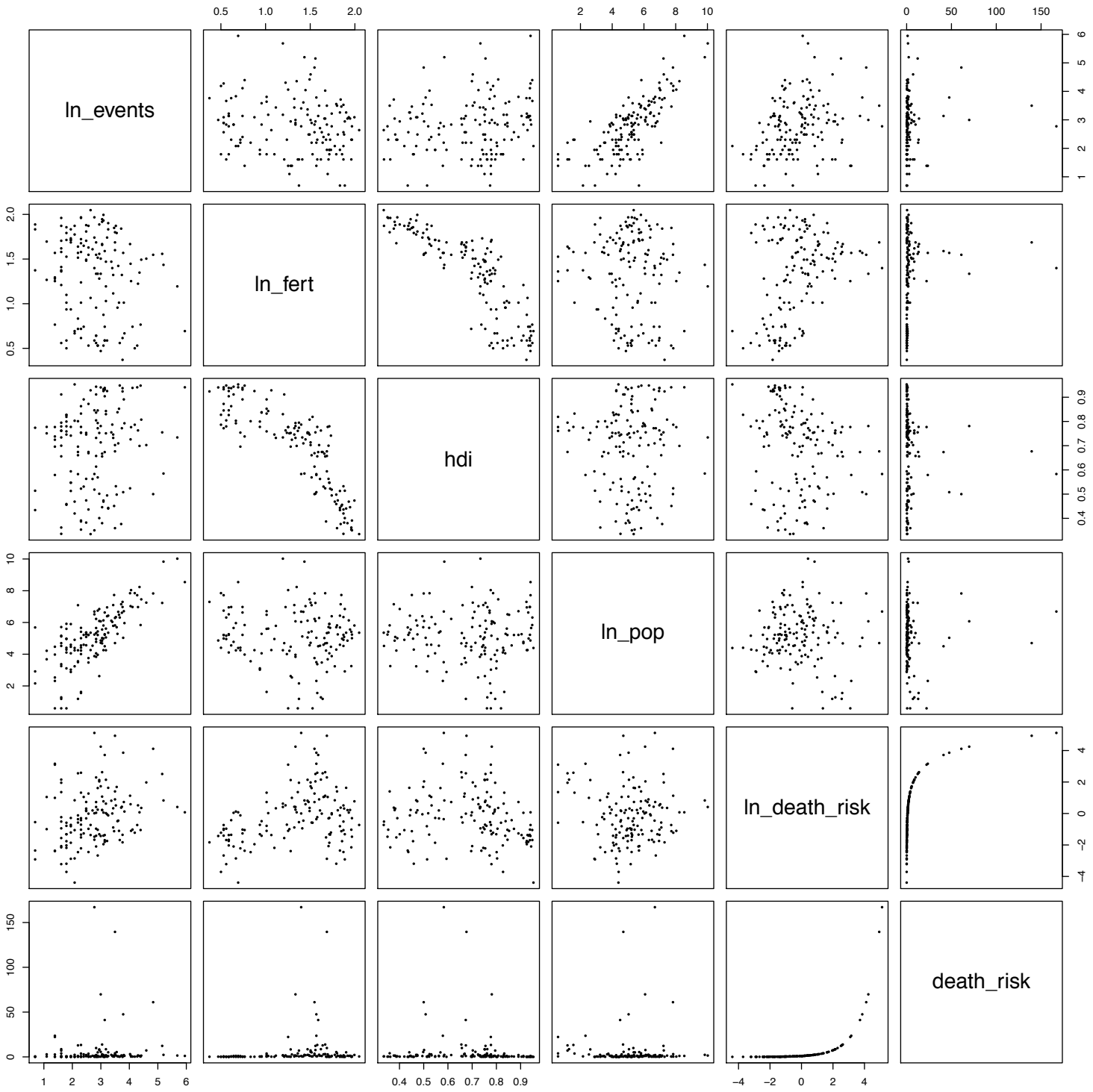
`ln_fert` the natural logarithm of an estimate of the country's female fertility

`hdi` the Human Development Index for the country

`death_risk` the proportion of people out of 1M in population killed in droughts, floods and storms

There are four predictor variables (if you count HDI and its square as one) which, while not big by any stretch of the imagination, is complex enough to keep us from "seeing" the whole data set

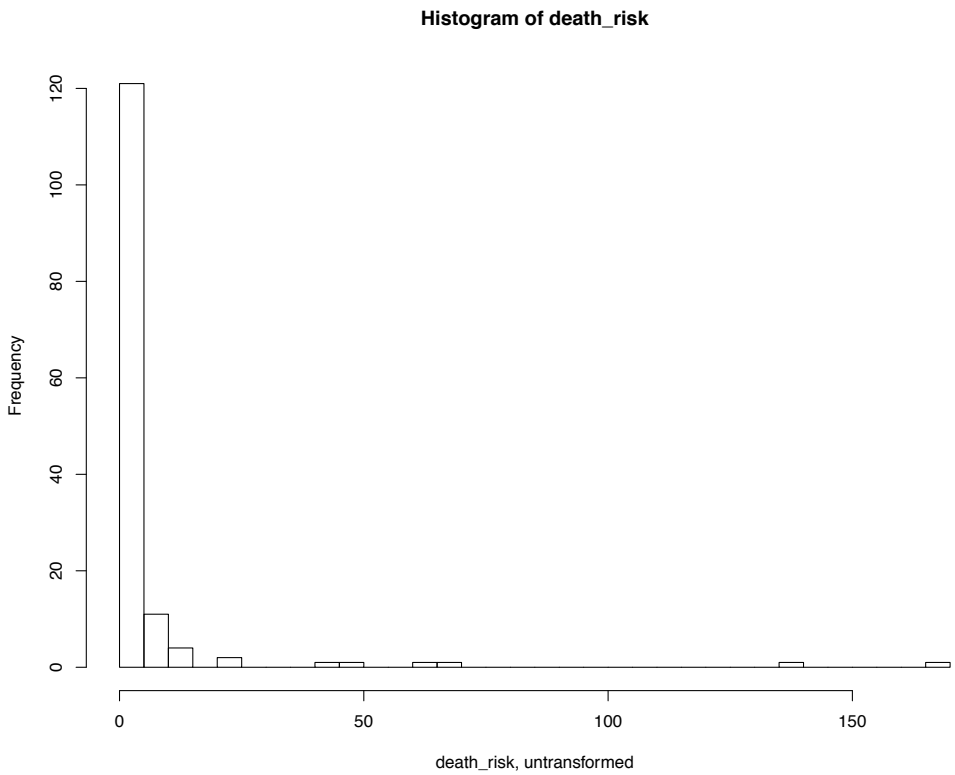
Instead, we might opt for partial views...

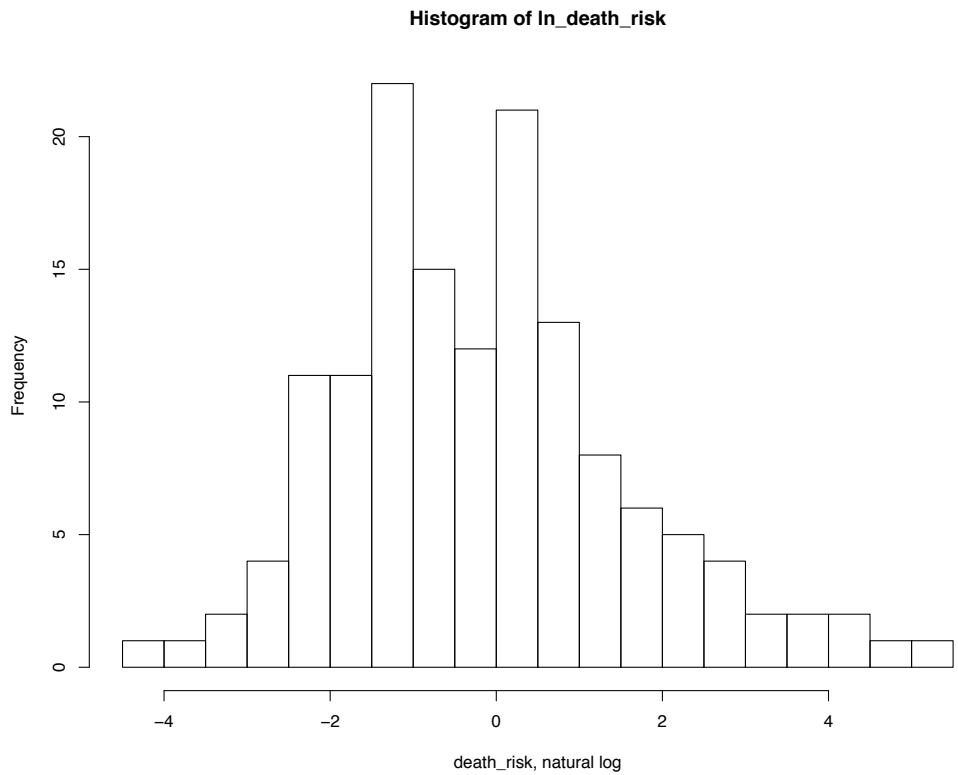

ln_events
ln_fert
hdi
ln_pop
ln_death_risk
death_risk

A first model

Rather than repeat the author's final analysis, we might usefully question the rationale for taking each step -- Let's start with the response variable and the decision to take a logarithm

If we look at the response itself, we see that it's quite skewed...


Histogram of death_risk
Frequency
0
20
40
60
80
100
120
0
50
100
150
death_risk, untransformed


Histogram of ln_death_risk
Frequency
0
5
10
15
20
−4
−2
0
2
4
death_risk, natural log

An extended model

WIth these data, we move from simple regression (one predictor) to multiple regression (um, multiple predictors) -- The model now transitions from

$$(\text{mercury}) \quad = \quad \beta_0 + \beta_1(\text{length}) + (\text{error})$$

to the slightly more elaborate

$$\begin{aligned}(\text{vuln}) \quad = \quad & \beta_0 + \beta_1(\text{num disasters}) + \beta_2(\text{population}) + \beta_3(\text{hdi}) + \\ & \beta_4(\text{hdi}^2) + \beta_5(\text{fertility}) + \beta_6(\text{urban population}) + (\text{error})\end{aligned}$$

```
> fit <- lm(ln_death_risk~ln_urb+ln_pop+ln_fert+poly(hdi,2),data=vul)
> summary(fit)

Call:
lm(formula = ln_death_risk ~ ln_urb + ln_pop + ln_fert + poly(hdi,
    2), data = vul)

Residuals:
    Min      1Q  Median      3Q     Max
-3.7736 -0.9064  0.0217  0.7121  4.7937

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -0.43483    1.41072  -0.308  0.75837
ln_urb        -0.83603    0.33099  -2.526  0.01267 *
ln_pop         0.09735    0.07728   1.260  0.20990
ln_fert        2.19853    0.54191   4.057 8.28e-05 ***
poly(hdi, 2)1 10.61737    3.42476   3.100  0.00234 **
poly(hdi, 2)2 -5.38801    1.62259  -3.321  0.00115 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.514 on 138 degrees of freedom
Multiple R-squared: 0.278, Adjusted R-squared: 0.2519
```

Read

She explained to me that a suitably programmed computer can read a novel in a few minutes and record the list of all the words contained in the text, in order of frequency. "That way I can have an already completed reading at hand," Lotaria says, "with an incalculable saving of time. What is the reading of a text, in fact, except the recording of certain thematic recurrences, certain insistences of forms and meanings?

"Words that appear eighteen times: boys, cap, come, dead, eat, enough, evening, French, go, handsome, new, passes, period, potatoes, those, until…

"Don't you already have a clear idea what it's about?" Lotario says. "There's no question: it's a war novel, all action, brisk writing, with a certain underlying violence.

Italo Calvino
"If on a winter's night a traveler"


market get home
cost world address debt
new must
percent long-term
chamber financial century clean
way challenges ask values programs
united businesses work meet global
family pay made
put economic buy need back see
many states
nation college fact help even
ensure congress day recession
soon send
spend time tonight one tax
afford come lead end
energy money president
school ever may
million security days know first
goal force high credit
business already let years reform
necessary
provide take government bad
issue economy now invest
bring system begin care
difficult like prosperity
families country american
confidence
commitment children
restore cut schools
single future every crisis health right
last opportunity administration keep still
effort make people yet words responsibility
save
banks jobs longer next helping
investment
just americans times
america history action largest
education
asked tell
lending budget job power
another also
deficit called
recovery
year
never loans support
finally

The Words That Were Used

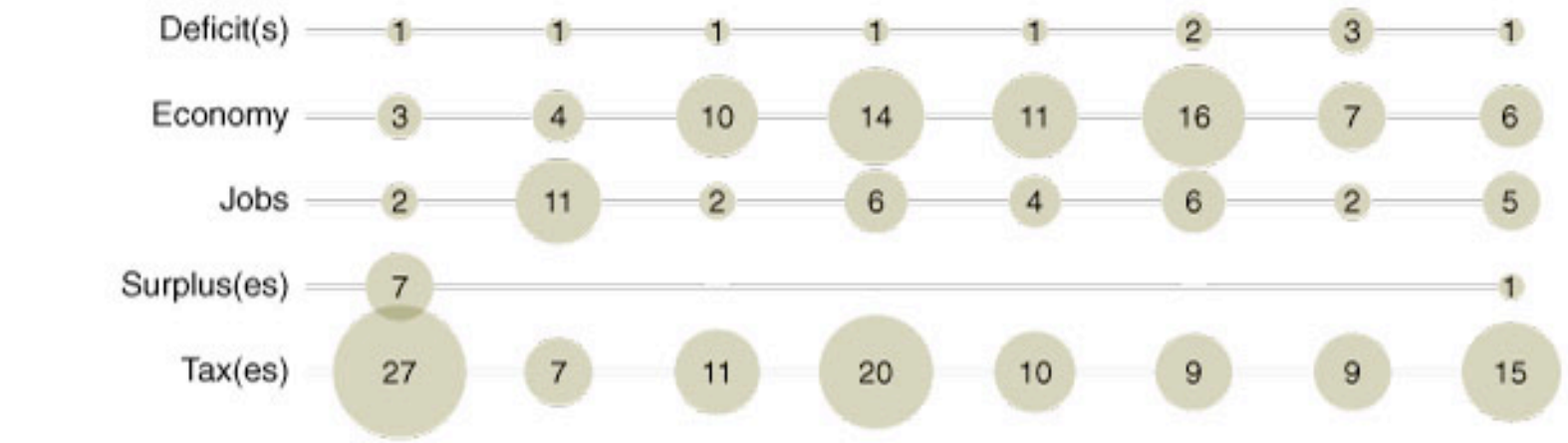
Number of times President Bush used the following words or phrases in State of the Union addresses.

2001	2002	2003	2004	2005	2006	2007	2008
FEB. 27*	JAN. 29	JAN. 28	JAN. 20	FEB. 2	JAN. 31	JAN. 23	JAN. 28

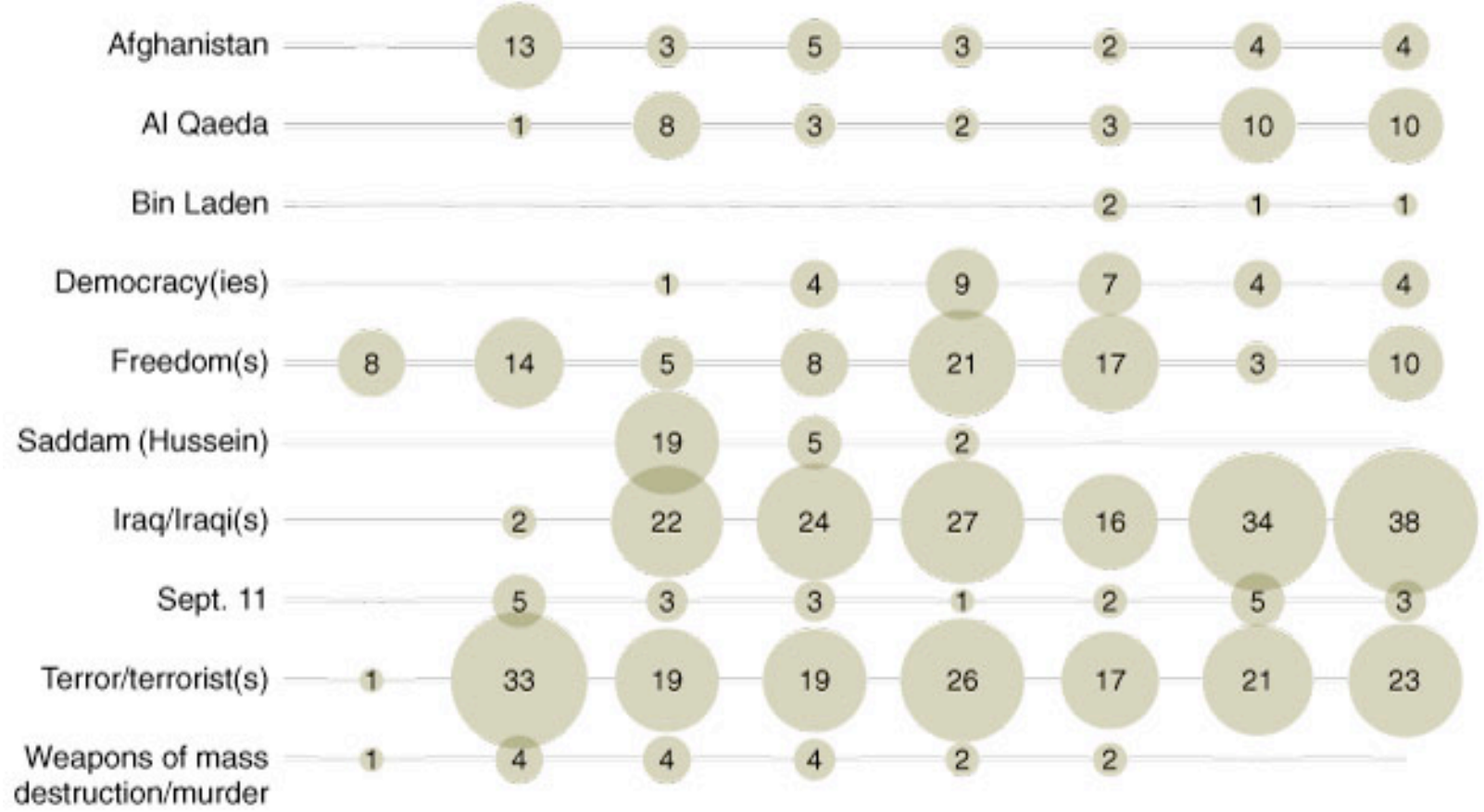
DOMESTIC AFFAIRS

	2001	2002	2003	2004	2005	2006	2007	2008
Earmark(s)						1	5	5
Energy	13	3	8	2	4	8	3	5
Health care	9	1	6	9	3	2	4	2
(Health) insurance	2		1	6	1	3	14	1
Medicare	12	1	5	8		2	2	1
Oil		1				3	9	3
Retirement(s)	5	3	1	1	12	3		
Social Security	15	2	2	2	18	3	2	1

TAXES AND THE ECONOMY



TERRORISM AND FOREIGN AFFAIRS



HINDSIGHT

IS ALWAYS 20/20

R. LUKE DUBOIS

DATA FROM THE AMERICAN PRESIDENCY PROJECT

UNIVERSITY OF CALIFORNIA, SANTA BARBARA

PRODUCED BY DANA KARWAS FOR BITFORMS GALLERY, NYC

GRAPHIC DESIGN BY ROGERBOVA.COM, NYC

LETTERPRESS BY COEUR NOIR, BROOKLYN, NY

SOFTWARE DEVELOPED USING MAX/MSP/JITTER, CYCLING'74, SAN FRANCISCO, CA

FOR MY FATHER, ROGER

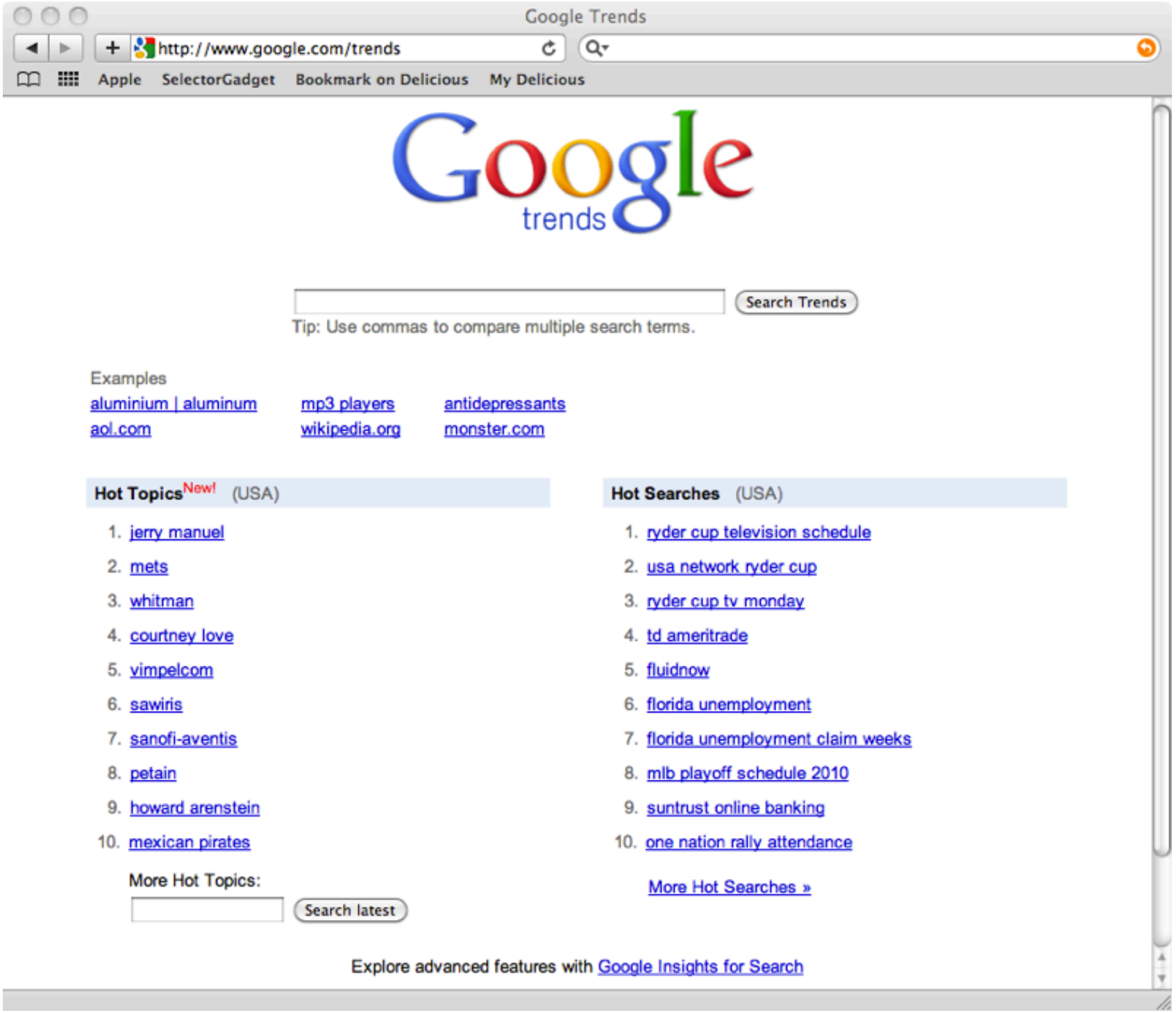
R. LUKE DUBOIS / 1975-

R. Luke DuBois
Detail: "Lyndon Baines Johnson / 1963-1969,"
from *Hindsight is Always 20/20*, 2008

R. Luke DuBois
Detail: "Ronald Wilson Reagan / 1981-1989,"
from *Hindsight is Always 20/20*, 2008

R. Luke DuBois
Detail: "William Jefferson Clinton / 1993-2001,"
from *Hindsight is Always 20/20*, 2008

R. Luke DuBois
Detail: "George Walker Bush / 2001-Present,"
from *Hindsight is Always 20/20*, 2008


Google Trends
http://www.google.com/trends
Apple
SelectorGadget
Bookmark on Delicious
My Delicious
Google
trends
Search Trends
Tip: Use commas to compare multiple search terms.
Examples
aluminium | aluminum
aol.com
mp3 players
wikipedia.org
antidepressants
monster.com
Hot Topics New! (USA)
1. jerry manuel
2. mets
3. whitman
4. courtney love
5. vimpelcom
6. sawiris
7. sanofi-aventis
8. petain
9. howard arenstein
10. mexican pirates
More Hot Topics:
Search latest
Hot Searches (USA)
1. ryder cup television schedule
2. usa network ryder cup
3. ryder cup tv monday
4. td ameritrade
5. fluidnow
6. florida unemployment
7. florida unemployment claim weeks
8. mlb playoff schedule 2010
9. suntrust online banking
10. one nation rally attendance
More Hot Searches »
Explore advanced features with Google Insights for Search

Google Trends: sneezing, coughing, runny nose

http://www.google.com/trends?q=sneezing,+c

Google

Apple SelectorGadget Bookmark on Delicious My Delicious

Google trends

sneezing, coughing, runny nose

Search Trends

Tip: Use commas to compare multiple search terms.

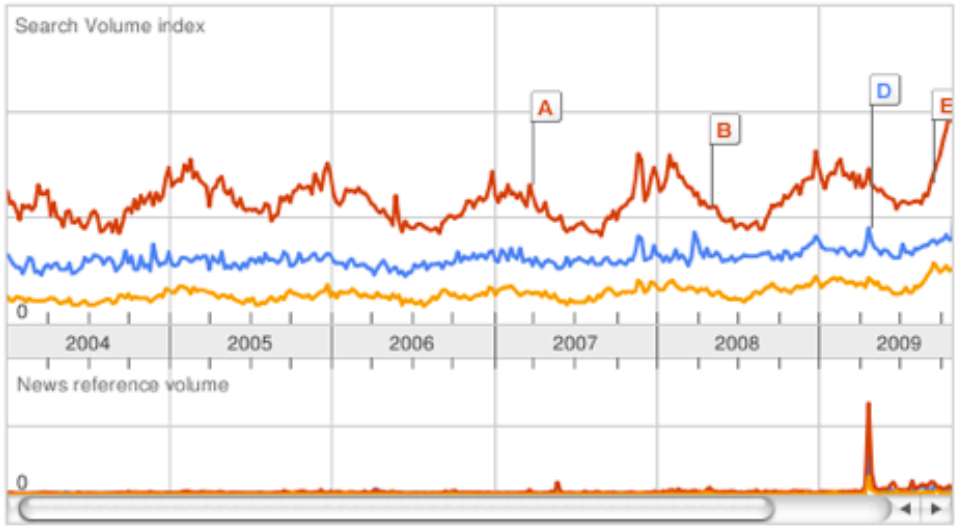
Sign in to see and export additional Trends data.

Searches Websites

All regions

All years

● sneezing ● coughing ● runny nose



Rank by sneezing

A Girl, 16, Kicked Off Plane for Coughing
ABC News - Mar 30 2007

B Mauresmo eyes Rome return after coughing problem
Reuters India - May 6 2008

C • Wash your hands with soap and water frequently, especially after coughing/sneezing.
TODAYonline - Apr 28 2009

D • Wash your hands with soap and water frequently, especially after coughing/sneezing.
TODAYonline - Apr 28 2009

E Doctors diagnose coughing problem, pull fragment of Wendy's plastic utensil from NC man's lung
Chicago Tribune - Sep 17 2009

F Achoo! Girl Can't Stop Sneezing
ABC News - Nov 11 2009

More news results »

Regions

1. United States
2. Canada
3. New Zealand
4. Australia
5. United Kingdom
6. India

Cities

1. Waterloo, Canada
2. Tampa, FL, USA
3. Philadelphia, PA, USA
4. Orlando, FL, USA
5. Minneapolis, MN, USA

Languages

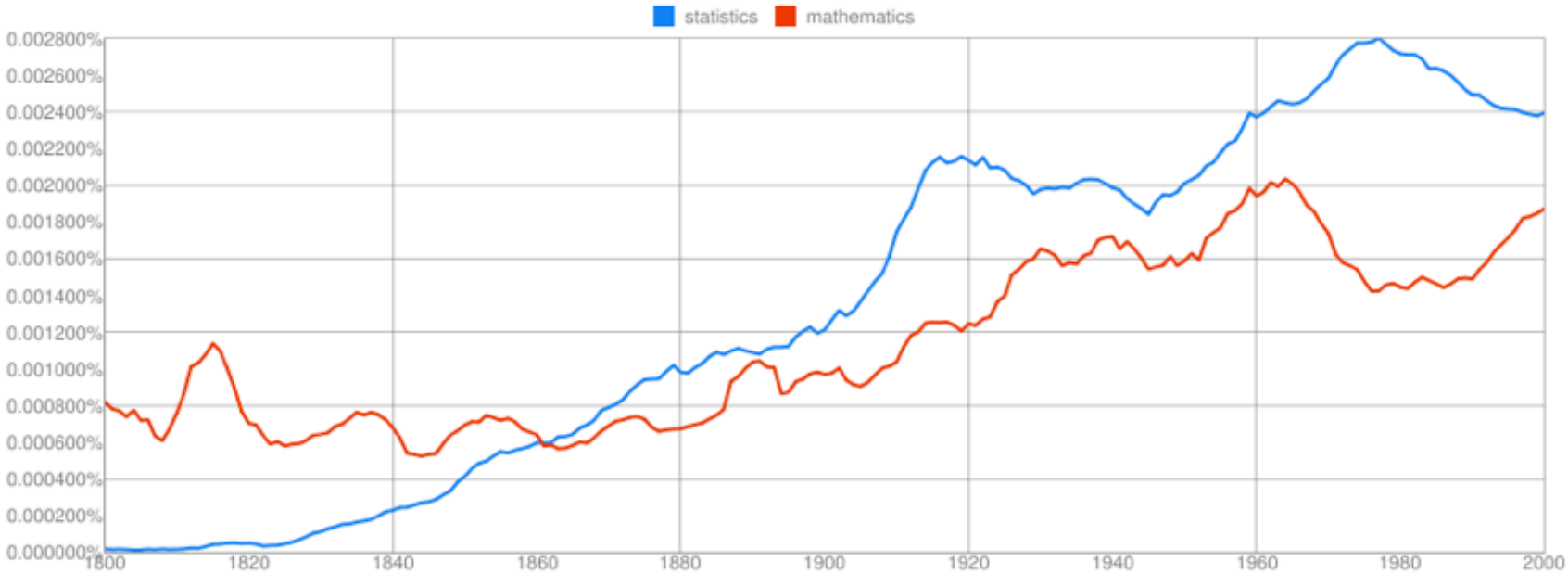
1. English
2. Dutch
3. German

Google labs **Books Ngram Viewer**

Graph these **case-sensitive** comma-separated phrases: statistics,mathematics

between 1800 and 2000 from the corpus English with smoothing of 3.

Search lots of books



Search in Google Books:

1800 - 1829	1830 - 1958	1959 - 1970	1971 - 1985	1986 - 2000	mathematics (English)
1800 - 1880	1881 - 1970	1971 - 1980	1981 - 1990	1991 - 2000	statistics (English)

Run your own experiment! Raw data is available for download here.

Google Research Blog

All Our N-gram are Belong to You

Thursday, August 03, 2006 at 8/03/2006 11:26:00 AM

Posted by Alex Franz and Thorsten Brants, Google Machine Translation Team

Here at Google Research we have been using word n-gram models for a variety of R&D projects, such as statistical machine translation, speech recognition, spelling correction, entity detection, information extraction, and others. While such models have usually been estimated from training corpora containing at most a few billion words, we have been harnessing the vast power of Google's datacenters and distributed processing infrastructure to process larger and larger training corpora. We found that there's no data like more data, and scaled up the size of our data by one order of magnitude, and then another, and then one more - resulting in a training corpus of *one trillion words* from public Web pages.

We believe that the entire research community can benefit from access to such massive amounts of data. It will advance the state of the art, it will focus research in the promising direction of large-scale, data-driven approaches, and it will allow all research groups, no matter how large or small their computing resources, to play together. That's why we decided to share this enormous dataset with everyone. We processed 1,024,908,267,229 words of running text and are publishing the counts for all 1,176,470,663 five-word sequences that appear at least 40 times. There are 13,588,391 unique words, after discarding words that appear less than 200 times.

Watch for an announcement at the Linguistics Data Consortium (LDC), who will be distributing it soon, and then order your set of 6 DVDs. And let us hear from you - we're excited to hear what you will do with the data, and we're always interested in feedback about this dataset, or other potential datasets that might be useful for the research community.

Update (22 Sept. 2006): The LDC now has the data available in their catalog. The counts are as follows:

```
File sizes: approx. 24 GB compressed (gzip'ed) text files

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
```

Published Online 16 December 2010

Science DOI: 10.1126/science.1199644

< *Science* Express Index

RESEARCH ARTICLE

Quantitative Analysis of Culture Using Millions of Digitized Books

Jean-Baptiste Michel[1,2,3,4,*†], Yuan Kui Shen[5], Aviva P. Aiden[6], Adrian Veres[7], Matthew K. Gray[8], The Google Books Team[8], Joseph P. Pickett[9], Dale Hoiberg[10], Dan Clancy[8], Peter Norvig[8], Jon Orwant[8], Steven Pinker[4], Martin A. Nowak[1,11,12] and Erez Lieberman Aiden[1,12,13,14,15,16,*†]

± Author Affiliations

†To whom correspondence should be addressed. E-mail: jb.michel@gmail.com (J.B.M.); erez@erez.com (E.A.).

↵* These authors contributed equally to this work.

ABSTRACT

We constructed a corpus of digitized texts containing about 4% of all books ever printed. Analysis of this corpus enables us to investigate cultural trends quantitatively. We survey the vast terrain of "culturomics", focusing on linguistic and cultural phenomena that were reflected in the English language between 1800 and 2000. We show how this approach can provide insights about fields as diverse as lexicography, the evolution of grammar, collective memory, the adoption of technology, the pursuit of fame, censorship, and historical epidemiology. "Culturomics" extends the boundaries of rigorous quantitative inquiry to a wide array of new phenomena spanning the social sciences and the humanities.

Google labs Books Ngram Viewer

Here are the datasets backing the Google Books Ngram Viewer. These datasets were generated in July 2009; we will update these datasets as our book scanning continues, and the updated versions will have distinct and persistent version identifiers (20090715 for the current set).

Each of the numbered links below will directly download a fragment of the given corpus. For instance, the first ten links below collectively comprise the 1-gram (i.e., individual words) counts for English, as collected from Google's scanned books around July 15, 2009. In addition, for each corpus we provide the file **total counts**, which records the total number of 1-grams contained in the books that make up the corpus. This file is useful to compute the relative frequencies of n-grams.

Details on the corpus construction can be found in the Science article written by Jean-Baptiste Michel et al. but are abbreviated here. Of note, we report only the n-grams that appeared over 40 times in the whole corpus. Therefore, the sum of the 1-gram occurences in any given corpus is smaller than the number given in the total counts file.

File format: Each of the numbered files below is zipped *tab*-separated data. (Yes, we know the files have .csv extensions.) Each line has the following format:

```
ngram TAB year TAB match_count TAB page_count TAB volume_count NEWLINE
```

As an example, here are the 30,000,000th and 30,000,001st lines from file 0 of the English 1-grams (googlebooks-eng-all-1gram-20090715-0.csv.zip):

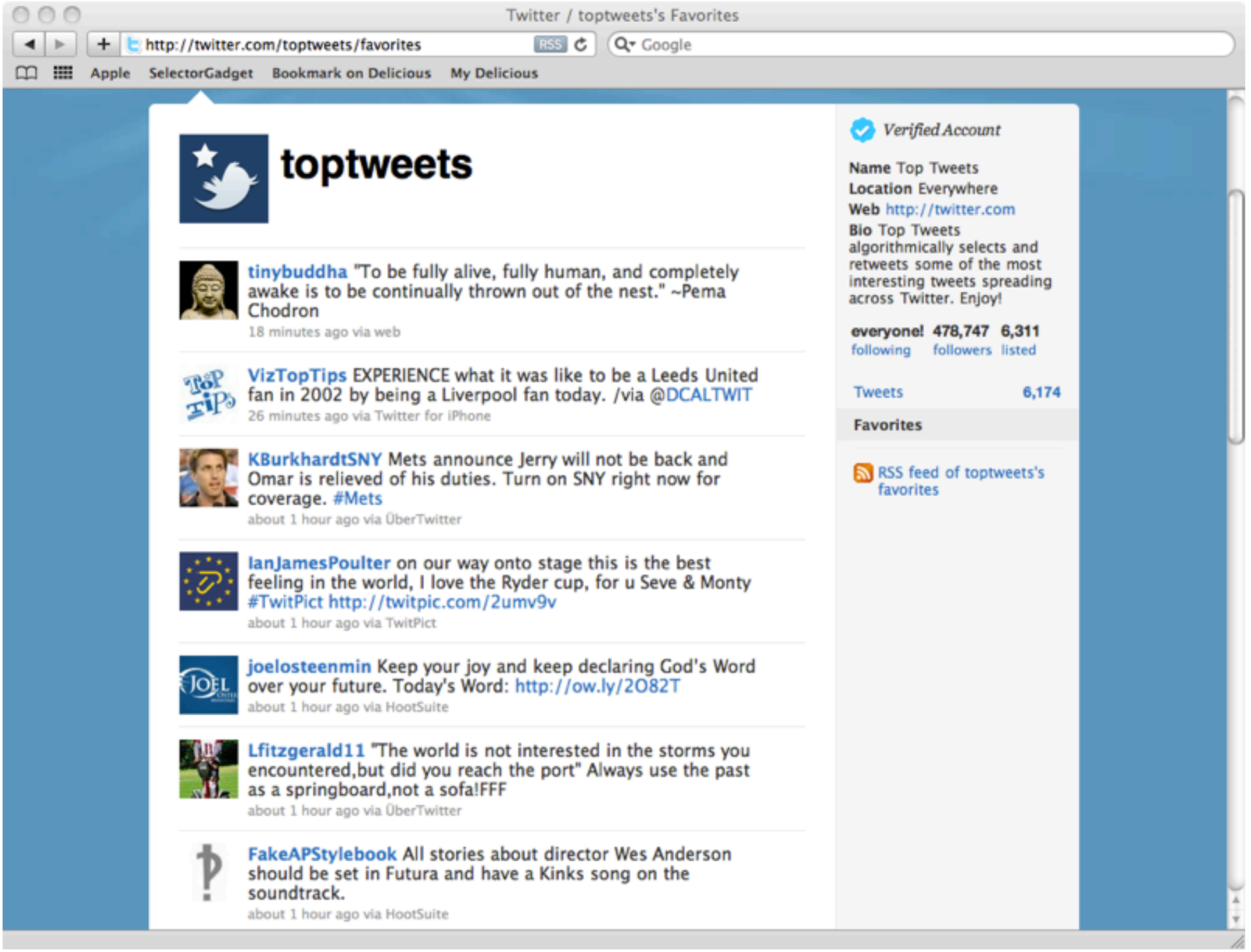
```
circumvallate   1978   313    215   85
circumvallate   1979   183    147   77
```

The first line tells us that in 1978, the word "circumvallate" (which means "surround with a rampart or other fortification", in case you were wondering) occurred 313 times overall, on 215 distinct pages and in 85 distinct books from our sample.

The format of the total counts file is identical, except that the `ngram` field is absent: there is only one triplet of values (match_count, page_count, volume_count) per year.

Here's the 9,000,000th line from file 0 of the English 5-grams (googlebooks-eng-all-5gram-20090715-0.csv.zip):

```
analysis is often described as  1991  1   1   1
```


Twitter / toptweets's Favorites
http://twitter.com/toptweets/favorites
RSS
Google
Apple
SelectorGadget
Bookmark on Delicious
My Delicious
toptweets
tinybuddha "To be fully alive, fully human, and completely awake is to be continually thrown out of the nest." ~Pema Chodron
18 minutes ago via web
VizTopTips EXPERIENCE what it was like to be a Leeds United fan in 2002 by being a Liverpool fan today. /via @DCALTWIT
26 minutes ago via Twitter for iPhone
KBurkhardtSNY Mets announce Jerry will not be back and Omar is relieved of his duties. Turn on SNY right now for coverage. #Mets
about 1 hour ago via ÜberTwitter
IanJamesPoulter on our way onto stage this is the best feeling in the world, I love the Ryder cup, for u Seve & Monty #TwitPict http://twitpic.com/2umv9v
about 1 hour ago via TwitPict
joelosteenmin Keep your joy and keep declaring God's Word over your future. Today's Word: http://ow.ly/2O82T
about 1 hour ago via HootSuite
Lfitzgerald11 "The world is not interested in the storms you encountered,but did you reach the port" Always use the past as a springboard,not a sofa!FFF
about 1 hour ago via ÜberTwitter
FakeAPStylebook All stories about director Wes Anderson should be set in Futura and have a Kinks song on the soundtrack.
about 1 hour ago via HootSuite
Verified Account
Name Top Tweets
Location Everywhere
Web http://twitter.com
Bio Top Tweets algorithmically selects and retweets some of the most interesting tweets spreading across Twitter. Enjoy!
everyone! 478,747 6,311
following followers listed
Tweets 6,174
Favorites
RSS feed of toptweets's favorites

John Tukey

From Wikipedia, the free encyclopedia

John Wilder Tukey (June 16, 1915 – July 26, 2000) was an American statistician.

Contents [hide]

Biography [edit]

Tukey was born in New Bedford, Massachusetts in 1915, and obtained a B.A. in 1936 and M.Sc. in 1937, in chemistry, from Brown University, before moving to Princeton University where he received a Ph.D. in mathematics.

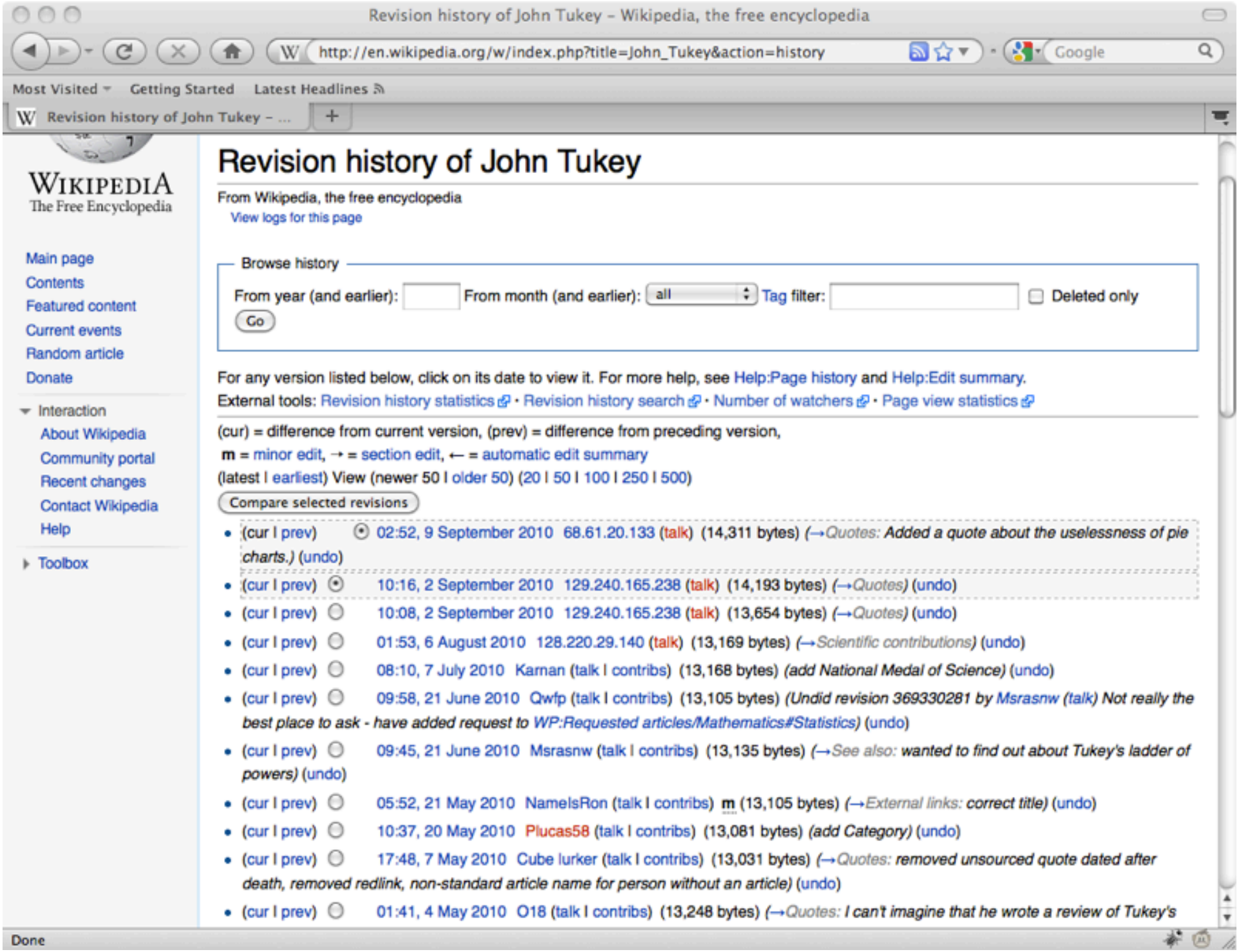
During World War II, Tukey worked at the Fire Control Research Office and collaborated with Samuel Wilks and William Cochran. After the war, he returned to Princeton, dividing his time between the university and AT&T Bell Laboratories.

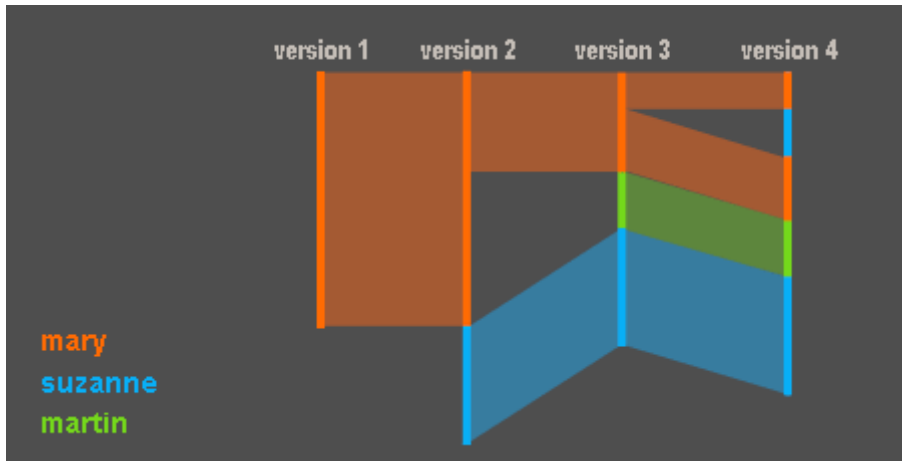
Among many contributions to civil society, Tukey served on a committee of the American Statistical Association that produced a report challenging the conclusions of the Kinsey Report, *Statistical Problems of the Kinsey Report on Sexual Behavior in the Human Male*.

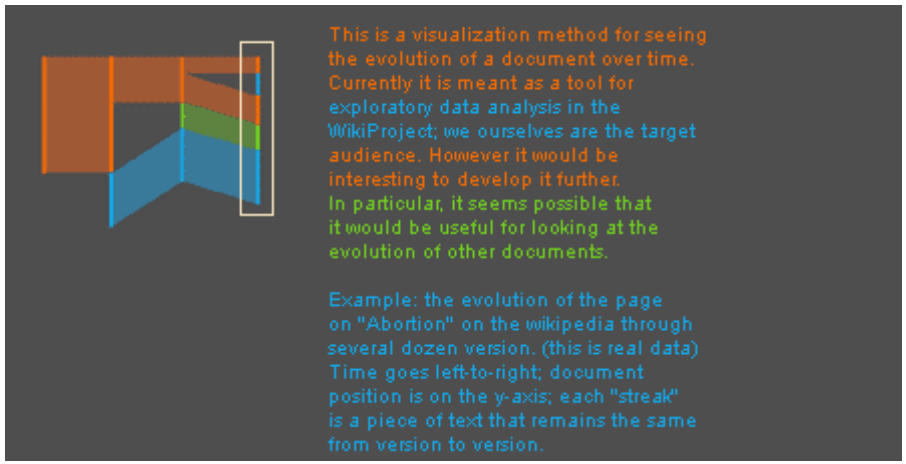
John Tukey

John Wilder Tukey

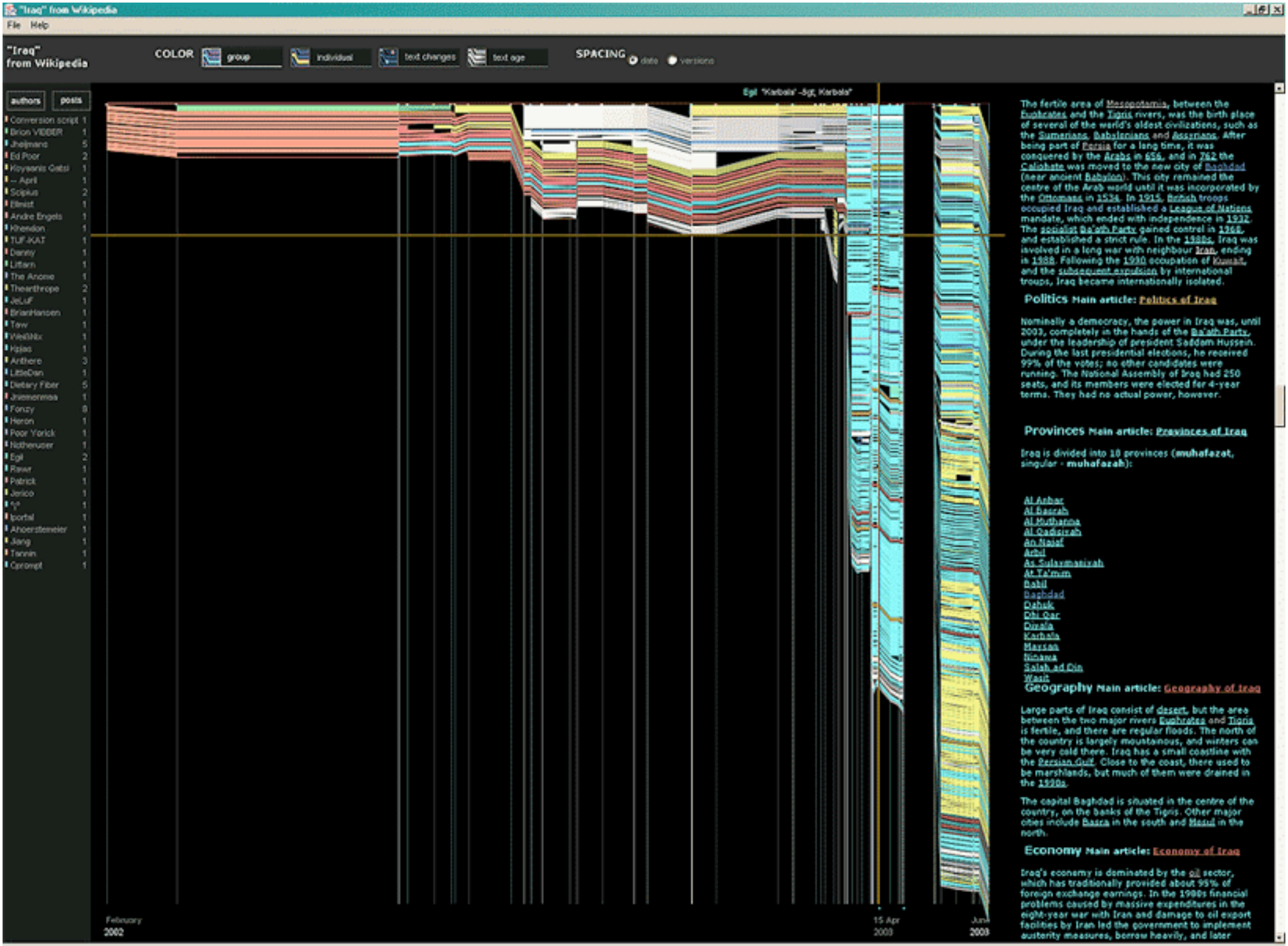
Born	June 16, 1915 New Bedford, Massachusetts, USA
Died	July 26, 2000 (aged 85) New Brunswick, New Jersey
Residence	United States
Nationality	American
Fields	Mathematician
Institutions	Bell Labs

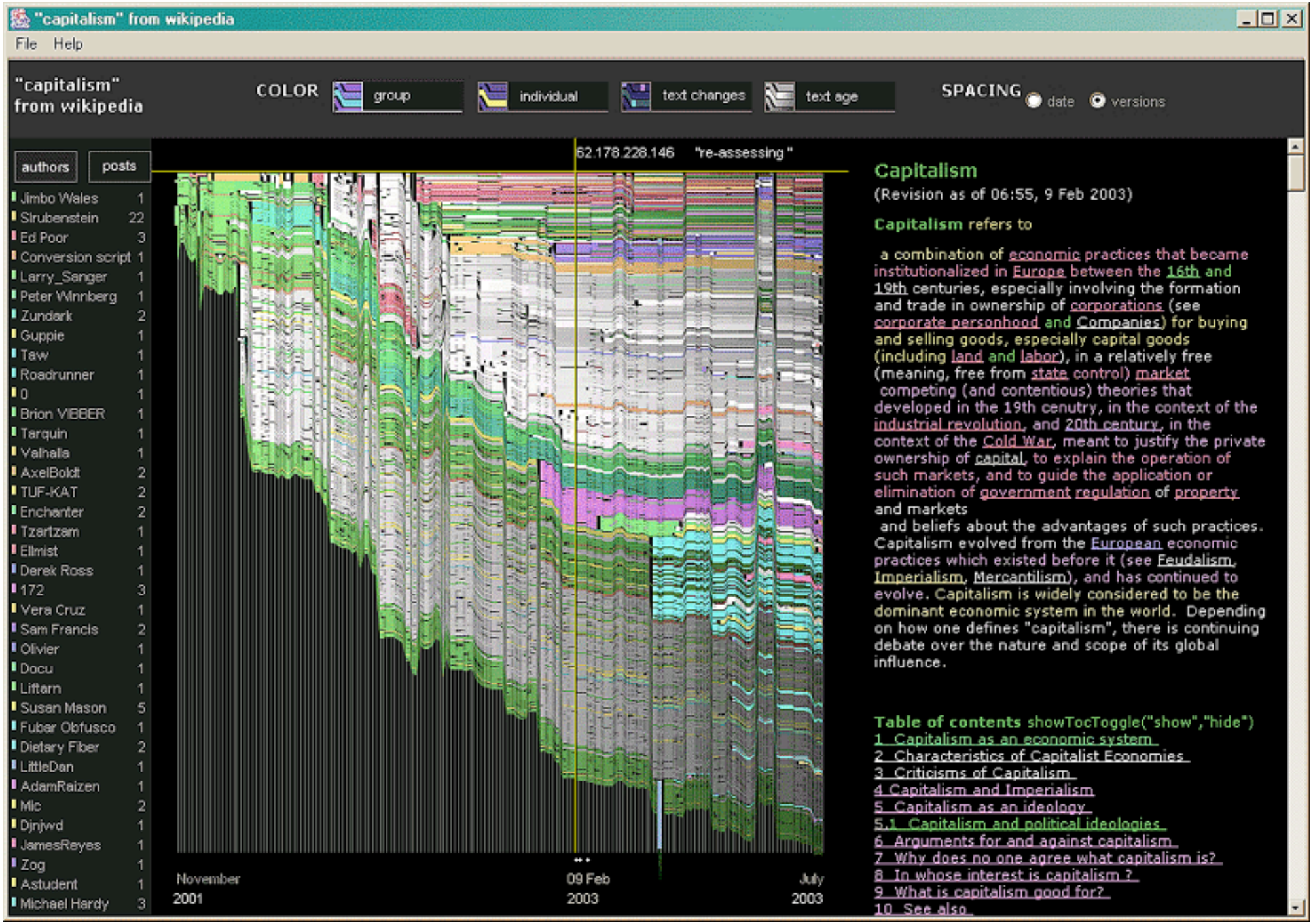

Revision history of John Tukey - Wikipedia, the free encyclopedia
http://en.wikipedia.org/w/index.php?title=John_Tukey&action=history
Google
Most Visited
Getting Started
Latest Headlines
Revision history of John Tukey - ...
WIKIPEDIA
The Free Encyclopedia
Main page
Contents
Featured content
Current events
Random article
Donate
Interaction
About Wikipedia
Community portal
Recent changes
Contact Wikipedia
Help
Toolbox
Revision history of John Tukey
From Wikipedia, the free encyclopedia
View logs for this page
Browse history
From year (and earlier):
From month (and earlier):
all
Tag filter:
Deleted only
Go
For any version listed below, click on its date to view it. For more help, see Help:Page history and Help:Edit summary.
External tools: Revision history statistics · Revision history search · Number of watchers · Page view statistics
(cur) = difference from current version, (prev) = difference from preceding version,
m = minor edit, → = section edit, ← = automatic edit summary
(latest | earliest) View (newer 50 | older 50) (20 | 50 | 100 | 250 | 500)
Compare selected revisions
(cur | prev) 02:52, 9 September 2010 68.61.20.133 (talk) (14,311 bytes) (→Quotes: Added a quote about the uselessness of pie charts.) (undo)
(cur | prev) 10:16, 2 September 2010 129.240.165.238 (talk) (14,193 bytes) (→Quotes) (undo)
(cur | prev) 10:08, 2 September 2010 129.240.165.238 (talk) (13,654 bytes) (→Quotes) (undo)
(cur | prev) 01:53, 6 August 2010 128.220.29.140 (talk) (13,169 bytes) (→Scientific contributions) (undo)
(cur | prev) 08:10, 7 July 2010 Karnan (talk | contribs) (13,168 bytes) (add National Medal of Science) (undo)
(cur | prev) 09:58, 21 June 2010 Qwfp (talk | contribs) (13,105 bytes) (Undid revision 369330281 by Msrasnw (talk) Not really the best place to ask - have added request to WP:Requested articles/Mathematics#Statistics) (undo)
(cur | prev) 09:45, 21 June 2010 Msrasnw (talk | contribs) (13,135 bytes) (→See also: wanted to find out about Tukey's ladder of powers) (undo)
(cur | prev) 05:52, 21 May 2010 NamelsRon (talk | contribs) m (13,105 bytes) (→External links: correct title) (undo)
(cur | prev) 10:37, 20 May 2010 Plucas58 (talk | contribs) (13,081 bytes) (add Category) (undo)
(cur | prev) 17:48, 7 May 2010 Cube lurker (talk | contribs) (13,031 bytes) (→Quotes: removed unsourced quote dated after death, removed redlink, non-standard article name for person without an article) (undo)
(cur | prev) 01:41, 4 May 2010 O18 (talk | contribs) (13,248 bytes) (→Quotes: I can't imagine that he wrote a review of Tukey's
Done



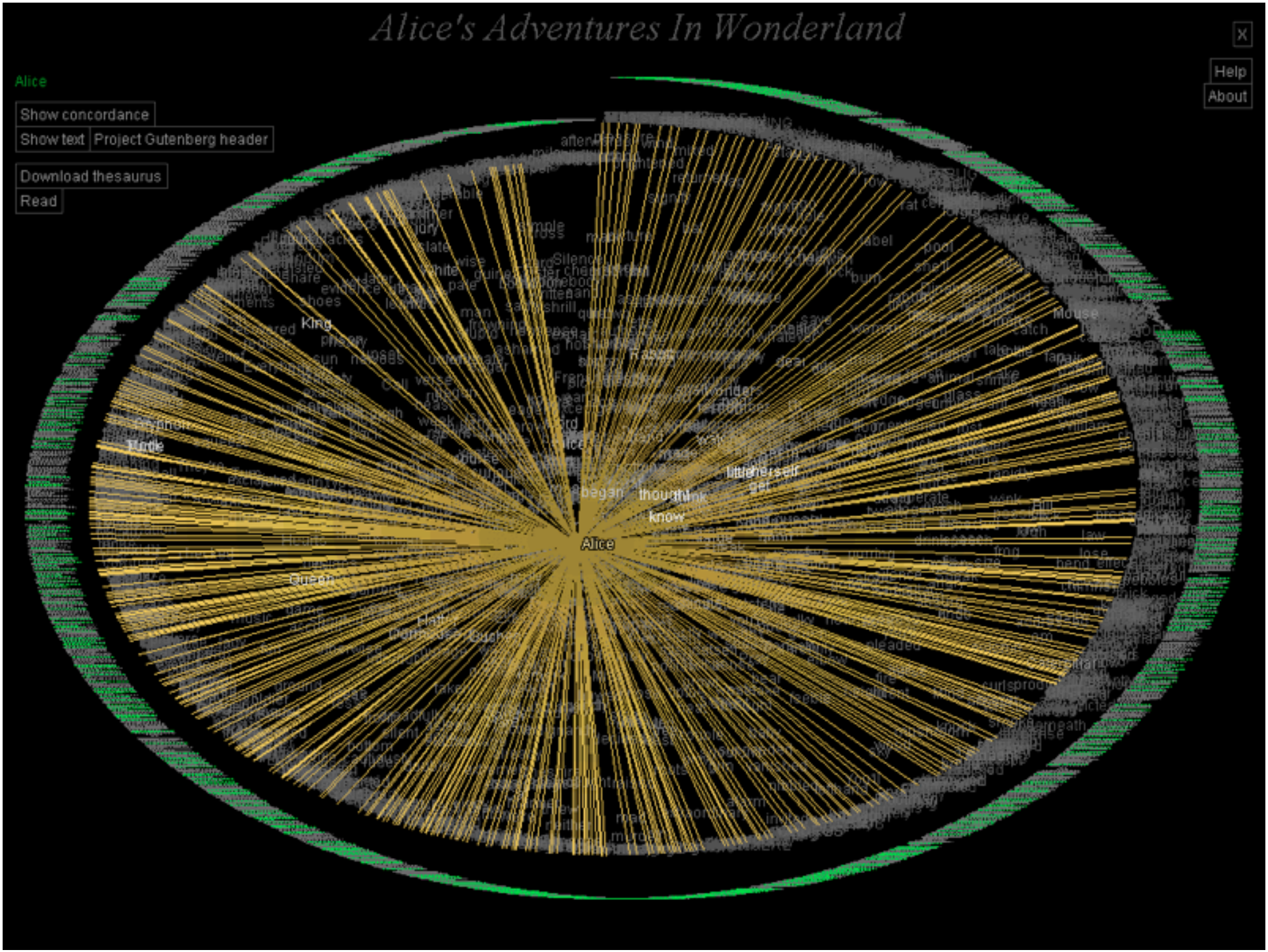


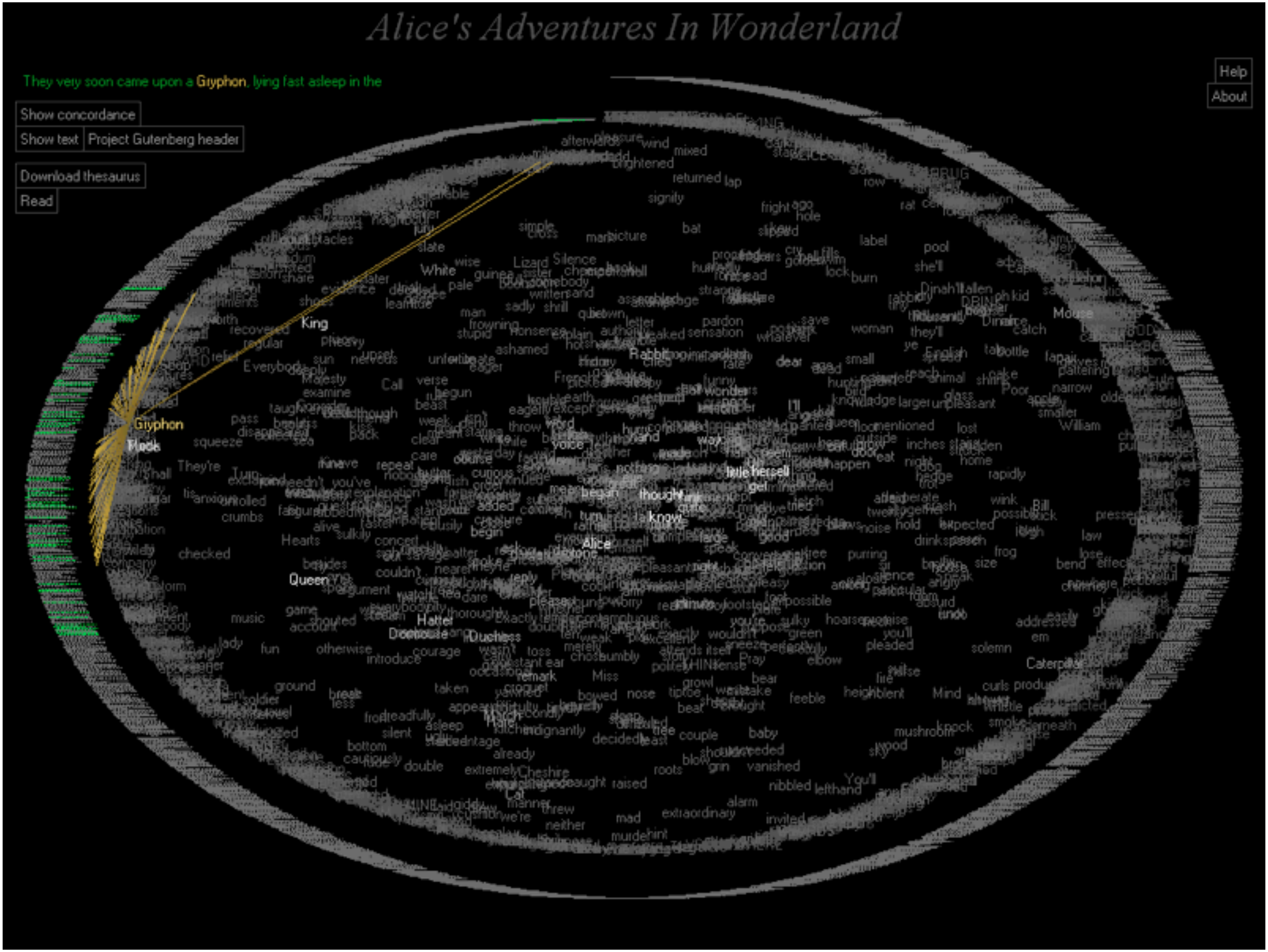
http://www.research.ibm.com/visual/projects/history_flow/


"Iraq" from Wikipedia
File Help
"Iraq" from Wikipedia
COLOR
group
individual
text changes
text age
SPACING
date
version
authors
posts
Egil "Karbala' -Sgt, Karbala"
February 2002
15 Apr 2003
Politics Main article: Politics of Iraq
Provinces Main article: Provinces of Iraq
Al Anbar
Al Basrah
Al Muthanna
Al Qadisiyah
An Najaf
Arbil
As Sulaymaniyah
At Ta'mim
Babil
Baghdad
Dahuk
Dhi Qar
Diyala
Karbala
Maysan
Ninawa
Salah ad Din
Wasit
Geography Main article: Geography of Iraq
Economy Main article: Economy of Iraq

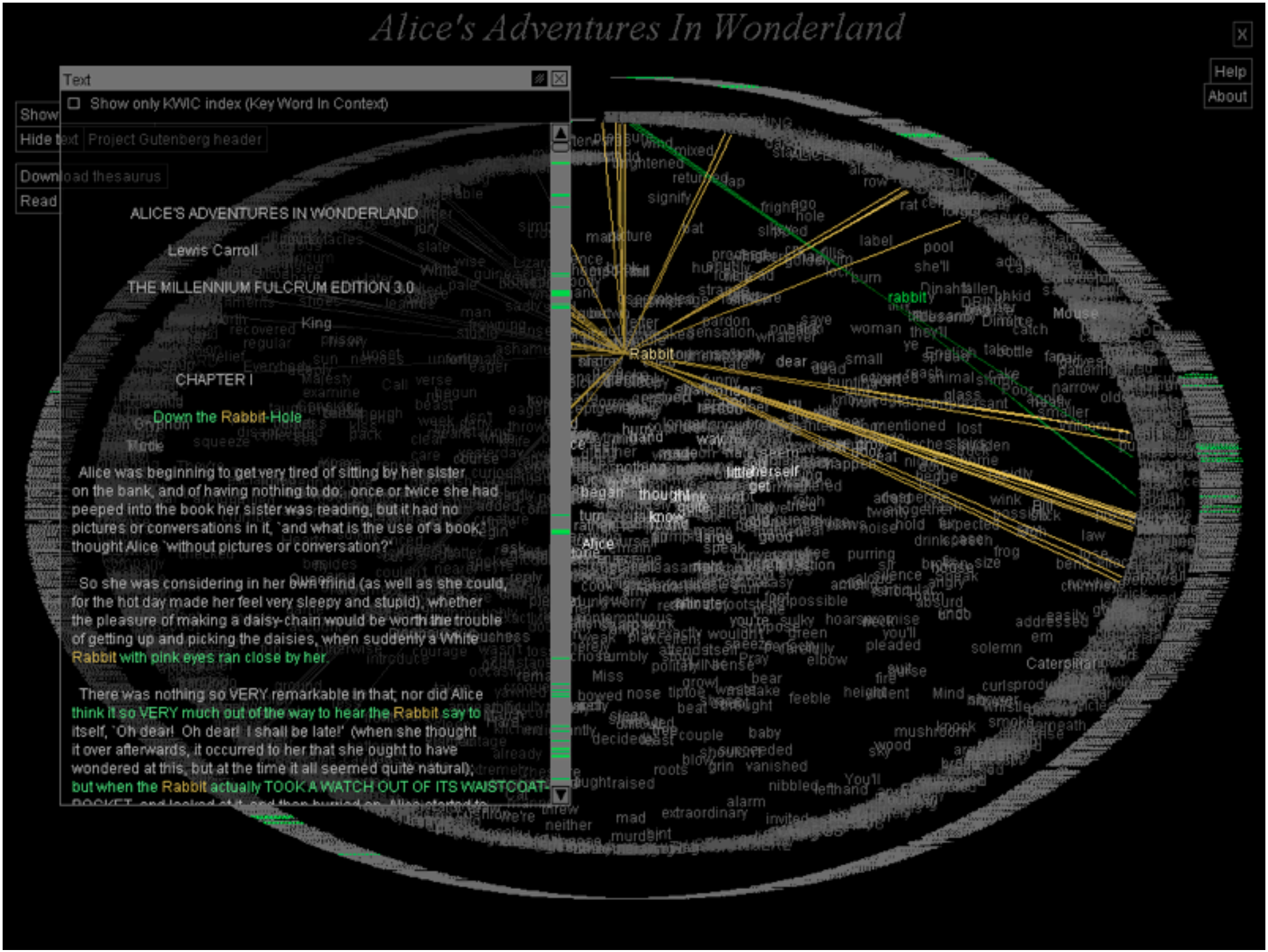

"capitalism" from wikipedia
File Help
"capitalism" from wikipedia
COLOR
group
individual
text changes
text age
SPACING
date
versions
authors
posts
Jimbo Wales 1
Slrubenstein 22
Ed Poor 3
Conversion script 1
Larry_Sanger 1
Peter Winnberg 1
Zundark 2
Guppie 1
Tew 1
Roadrunner 1
0 1
Brion VIBBER 1
Tarquin 1
Valhalla 1
AxelBoldt 2
TUF-KAT 2
Enchanter 2
Tzartzam 1
Ellmist 1
Derek Ross 1
172 3
Vera Cruz 1
Sam Francis 2
Olivier 1
Docu 1
Liftarn 1
Susan Mason 5
Fubar Obfusco 1
Dietary Fiber 2
LittleDan 1
AdamRaizen 1
Mic 2
Djnjwd 1
JamesReyes 1
Zog 1
Astudent 1
Michael Hardy 3
62.178.226.146 "re-assessing"
November 2001
09 Feb 2003
July 2003
Capitalism
(Revision as of 06:55, 9 Feb 2003)
Capitalism refers to
a combination of economic practices that became institutionalized in Europe between the 16th and 19th centuries, especially involving the formation and trade in ownership of corporations (see corporate personhood and Companies) for buying and selling goods, especially capital goods (including land and labor), in a relatively free (meaning, free from state control) market
competing (and contentious) theories that developed in the 19th cenutry, in the context of the industrial revolution, and 20th century, in the context of the Cold War, meant to justify the private ownership of capital, to explain the operation of such markets, and to guide the application or elimination of government regulation of property and markets
and beliefs about the advantages of such practices. Capitalism evolved from the European economic practices which existed before it (see Feudalism, Imperialism, Mercantilism), and has continued to evolve. Capitalism is widely considered to be the dominant economic system in the world. Depending on how one defines "capitalism", there is continuing debate over the nature and scope of its global influence.
Table of contents showTocToggle("show","hide")
1 Capitalism as an economic system
2 Characteristics of Capitalist Economies
3 Criticisms of Capitalism
4 Capitalism and Imperialism
5 Capitalism as an ideology
5.1 Capitalism and political ideologies
6 Arguments for and against capitalism
7 Why does no one agree what capitalism is?
8 In whose interest is capitalism ?
9 What is capitalism good for?
10 See also

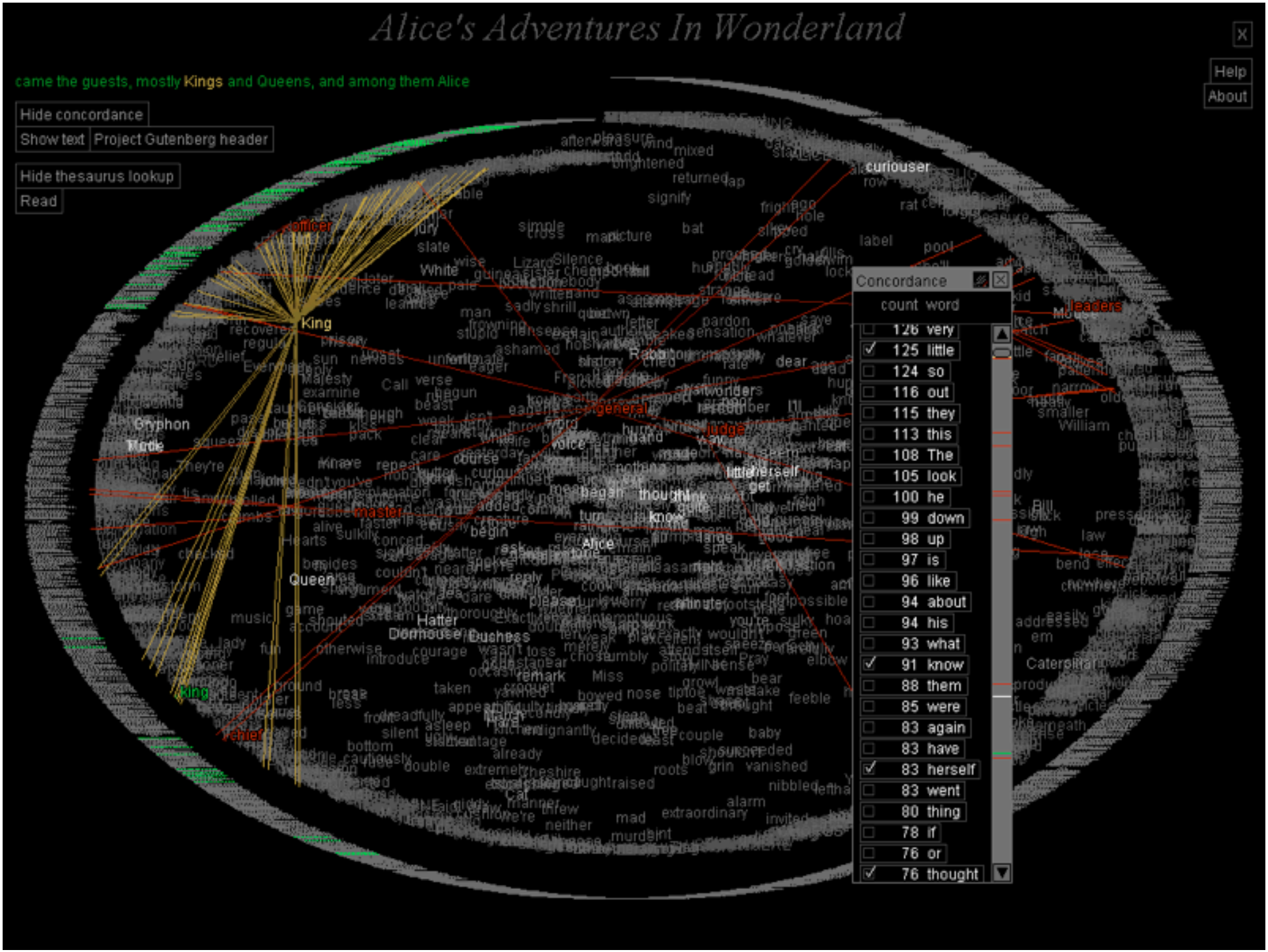

Alice's Adventures In Wonderland
X
Help
About
Show concordance
Show text
Project Gutenberg header
Download thesaurus
Read
King
Rabbit
Mouse
Gryphon
Turtle
Alice
Queen
Hatter
Dormouse
Duchess
Caterpillar
Cat
White
Hearts
William
Dinah
Miss
Majesty
remark
thought
herself
little
begin
know
http://www.textarc.org

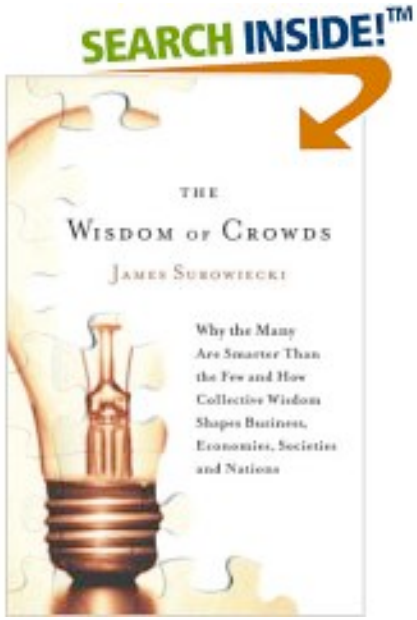

Alice's Adventures In Wonderland
X
Help
About
Alice
Show concordance
Show text
Project Gutenberg header
Download thesaurus
Read
King
Turtle
Queen
Alice
began
thought
know
herself
Mouse



http://www.textarc.org


Alice's Adventures In Wonderland
Text
Show only KWIC index (Key Word In Context)
Show
Hide text
Project Gutenberg header
Download thesaurus
Read
ALICE'S ADVENTURES IN WONDERLAND
Lewis Carroll
THE MILLENNIUM FULCRUM EDITION 3.0
CHAPTER I
Down the Rabbit-Hole
Alice was beginning to get very tired of sitting by her sister on the bank, and of having nothing to do: once or twice she had peeped into the book her sister was reading, but it had no pictures or conversations in it, `and what is the use of a book,' thought Alice `without pictures or conversation?'
So she was considering in her own mind (as well as she could, for the hot day made her feel very sleepy and stupid), whether the pleasure of making a daisy-chain would be worth the trouble of getting up and picking the daisies, when suddenly a White Rabbit with pink eyes ran close by her.
There was nothing so VERY remarkable in that; nor did Alice think it so VERY much out of the way to hear the Rabbit say to itself, `Oh dear! Oh dear! I shall be late!' (when she thought it over afterwards, it occurred to her that she ought to have wondered at this, but at the time it all seemed quite natural); but when the Rabbit actually TOOK A WATCH OUT OF ITS WAISTCOAT-
Rabbit
rabbit
Alice
Help
About
X

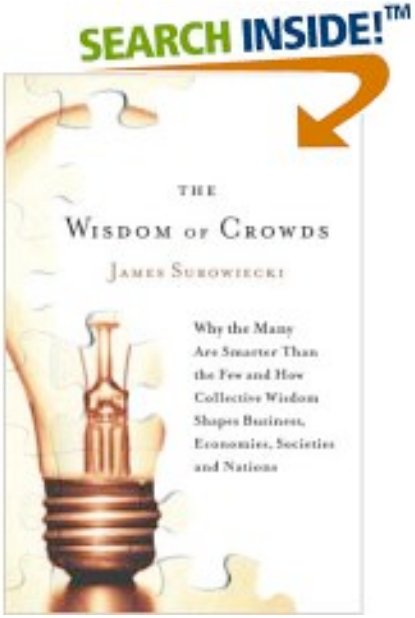

Alice's Adventures In Wonderland
came the guests, mostly Kings and Queens, and among them Alice
Hide concordance
Show text
Project Gutenberg header
Hide thesaurus lookup
Read
Help
About
Concordance
count word
126 very
125 little
124 so
116 out
115 they
113 this
108 The
105 look
100 he
99 down
98 up
97 is
96 like
94 about
94 his
93 what
91 know
88 them
85 were
83 again
83 have
83 herself
83 went
80 thing
78 if
76 or
76 thought
King
Queen
Alice
Hatter
Dormouse
Duchess
Gryphon
Turtle
officer
curiouser
Caterpillar
Rabbit
thought
know
begin
remark
judge
general
master
leaders



Concordance (learn more)

These are the 100 most frequently used words in this book.

american answer best better between business cars case collective come companies course crowd day decisions different does down else end even everyone experiment fact few first found game get go going good group idea important individual information instance instead intelligence investors kind know least less likely line makes making market may means members might money new often others own part people percent person place point possible price problem question rather result right say scientists see seems sense should since small solution something stock study system take team things think though time traffic two want whether whole wisdom work world years



Text Stats

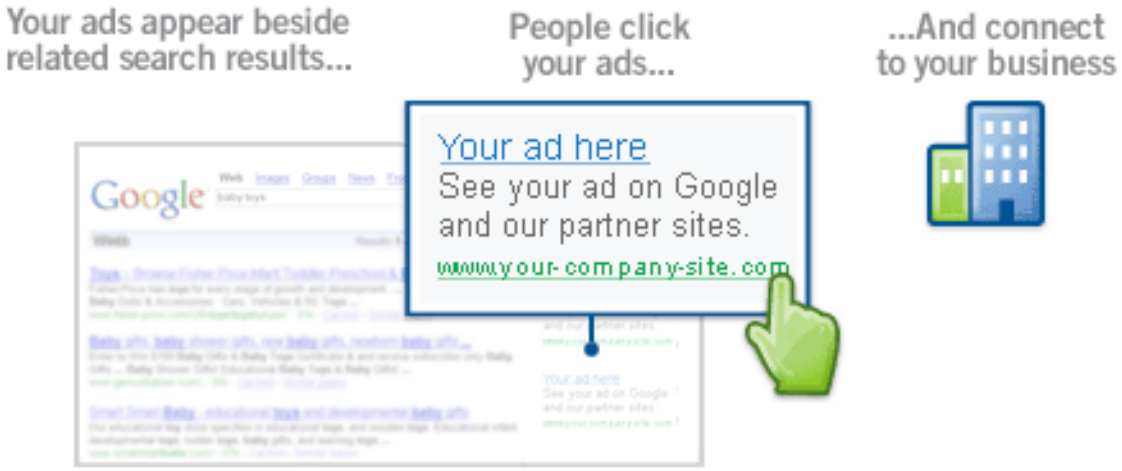
These statistics are computed from the text of this book. (learn more)

Readability (learn more)		Compared with books in All Categories	
Fog Index:	14.9	67% are easier	33% are harder
Flesch Index:	47.5	56% are easier	44% are harder
Flesch-Kincaid Index:	12.3	69% are easier	31% are harder

Complexity (learn more)			
Complex Words:	15%	49% have fewer	51% have more
Syllables per Word:	1.6	48% have fewer	52% have more
Words per Sentence:	22.6	82% have fewer	18% have more

Number of			
Characters:	570,231	68% have fewer	32% have more
Words:	94,421	70% have fewer	30% have more
Sentences:	4,172	56% have fewer	44% have more

Fun stats	
Words per Dollar:	5,733
Words per Ounce:	5,901



Learn about AdWords

How it works

[Why it works](#)

[Costs and payment](#)

[For local businesses](#)

[Success stories](#)

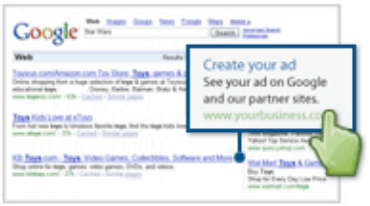
You create your ads
You create ads and choose keywords, which are words or phrases related to your business.
[Get keyword ideas](#)

Your ads appear on Google
When people search on Google using one of your keywords, your ad may appear next to the search results. Now you're advertising to an audience that's already interested in you.

You attract customers
People can simply click your ad to make a purchase or learn more about you. You don't even need a webpage to get started - Google will help you create one for free. It's that easy!

[Sign up now](#) | [Next topic »](#)

Keywords are what people search for on Google.



Your *ad* appears beside relevant search results.

Text as data: Why we teach it

From early problems in authorship attribution (how many of you have read Wallace and Mosteller's the Federalist Papers?) to more recent work in large-scale text mining, there's plenty of interesting problems and data sources to analyze

Artifacts from computer-mediated communication (web logs, bulletin boards, chat transcripts, email) all provide complex and socially interesting data for students to work with

These sources can be immediately compelling; in some sense they are closer to home, are more recognizable than certain scientific data sources, and can kick off important discussions about privacy and computer technologies (OK, I might be the only one interested in that)

Bursty and Hierarchical Structure in Streams *

Jon Kleinberg †

Abstract

A fundamental problem in text data mining is to extract meaningful structure from document streams that arrive continuously over time. E-mail and news articles are two natural examples of such streams, each characterized by topics that appear, grow in intensity for a period of time, and then fade away. The published literature in a particular research field can be seen to exhibit similar phenomena over a much longer time scale. Underlying much of the text mining work in this area is the following intuitive premise — that the appearance of a topic in a document stream is signaled by a "burst of activity," with certain features rising sharply in frequency as the topic emerges.

The goal of the present work is to develop a formal approach for modeling such "bursts," in such a way that they can be robustly and efficiently identified, and can provide an organizational framework for analyzing the underlying content. The approach is based on modeling the stream using an infinite-state automaton, in which bursts appear naturally as state transitions; it can be viewed as drawing an analogy with models from queueing theory for bursty network traffic. The resulting algorithms are highly efficient, and yield a nested representation of the set of bursts that imposes a hierarchical structure on the overall stream. Experiments with e-mail and research paper archives suggest that the resulting structures have a natural meaning in terms of the content that gave rise to them.

*This work appears in the Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2002.

†Department of Computer Science, Cornell University, Ithaca NY 14853. Email: kleinber@cs.cornell.edu.
Supported in part by a David and Lucile Packard Foundation Fellowship, an ONR Young Investigator Award, NSF ITR/IM Grant IIS-0081334, and NSF Faculty Early Career Development Award CCR-9701399.

Tracing information flow on a global scale using Internet chain-letter data

David Liben-Nowell*† and Jon Kleinberg†‡

*Department of Computer Science, Carleton College, Northfield, MN 55057; and ‡Department of Computer Science, Cornell University, Ithaca, NY 14853

Edited by Ronald L. Graham, University of California at San Diego, La Jolla, CA, and approved January 25, 2008 (received for review September 6, 2007)

Although information, news, and opinions continuously circulate in the worldwide social network, the actual mechanics of how any single piece of information spreads on a global scale have been a mystery. Here, we trace such information-spreading processes at a person-by-person level using methods to reconstruct the propagation of massively circulated Internet chain letters. We find that rather than fanning out widely, reaching many people in very few steps according to "small-world" principles, the progress of these chain letters proceeds in a narrow but very deep tree-like pattern, continuing for several hundred steps. This suggests a new and more complex picture for the spread of information through a social network. We describe a probabilistic model based on network clustering and asynchronous response times that produces trees with this characteristic structure on social-network data.

social networks | algorithms | epidemics | diffusion in networks

The dissemination of information is a ubiquitous process in human social networks. It plays a fundamental role in settings that include the spread of technological innovations (1, 2), word-of-mouth effects in marketing (3–5), the spread of news and opinion (6–8), collective problem-solving (9, 10), and sampling methods for hidden populations (11, 12). The basic models for studying such phenomena posit that information will diffuse from person to person in the style of an epidemic (13–16), expanding widely in a short number of steps according to "small-world" principles (17, 18). However, despite recent studies in online domains (5–8), it has been difficult to obtain detailed traces of the dissemination of a single piece of news or information on a global scale to assess the predictions of these models. As such, it has remained an open question whether the spreading of information truly proceeds with a rapid, epidemic-style fan-out or whether it follows a potentially more complex structure. The difference between these possibilities has consequences not only for the models that are used to capture their essential properties but also potentially for the "life cycle" of a piece of information as it spreads through the global social network.

Here, we trace these types of large-scale information-spreading processes at a person-by-person level using methods to reconstruct the propagation of massively circulated Internet chain letters, and from these observations we propose a new set of principles for how such processes work. We focus in particular on two such chain letters, which exhibit tree-like patterns of dissemination that are quite similar to each other but are initially in conflict with the intuitive picture of how information spreads in these settings. Rather than expanding to many individuals in a few steps, the trees are very narrow and continue reaching people several hundred levels deep. We describe a mathematical model that produces trees with this characteristic structure, grounded fundamentally in the observations that social networks are highly clustered and that information can take widely varying amounts of time to traverse different edges in the network. The simple structure of the model, and the fact that it is based on earlier empirical studies of human response times (19–21), thus suggests a possible basis for this narrow and deeply reaching style of information transmission in the local dynamics of communication within highly clustered social networks.

Reconstructing the Spread of Internet Chain Letters

To reconstruct instances in which specific pieces of information spread through large, globally distributed populations, we analyzed the dissemination of petitions that circulated widely in chain-letter form on the Internet over the past several years. The petitions instruct each recipient to append his or her name to a copy of the letter and then forward it to friends. Each copy will thus contain a list of people, representing a particular sequence of forwardings of the message; and hence different copies will contain different but overlapping lists of people, reflecting the paths they followed to their respective current recipients. This forwarding process is a readily recognizable mechanism by which jokes and news clippings can also achieve wide circulation through the global e-mail network; the explicit lists of names in the petition format, however, make it much easier to trace the propagation of the messages. The main chain letter that we analyze is based on a widely circulated petition from 2002–2003 claiming to organize opposition to the impending war in Iraq. We obtained copies via Internet searches of mailing-list archives in which they were publicly posted; these searches resulted in 637 copies with distinct chains of recipients, representing nearly 20,000 distinct signatories in aggregate. [See supporting information (SI) *Appendix* for the specifics of the data-collection process.]

We performed a similar analysis for a second chain letter, a petition that began circulating in 1998, purporting to organize political support for continued United States governmental funding of National Public Radio (NPR) and the Public Broadcasting System (PBS). Through similar means to those used for the Iraq petition, we acquired 316 distinct copies of the NPR petition, comprising a total of 13,052 people. The dissemination of the two chain letters exhibited qualitatively very similar structures, and for purposes of the discussion here, we focus on the analysis of the chain letter associated with the Iraq petition. Although both petitions in fact had their origins in hoaxes and naive misunderstandings, as a large fraction of the most widespread Internet chain letters do (22, 23), this fact is immaterial to our purposes, especially because almost all signatories to each appeared to believe them to be authentic; hence, we are studying genuine instances of the dissemination of individual pieces of information along links in the global social network.

People may in general receive a copy of the chain letter multiple times, but if each appends his or her name to just one copy, then the full propagation of the letter can be represented as a tree

Author contributions: D.L.-N. and J.K. designed research, performed research, contributed new reagents/analytic tools, analyzed data, and wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

†To whom correspondence may be addressed. E-mail: dlibenno@carleton.edu or kleinber@cs.cornell.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0708471105/DC1.