Experimental Design for Collection-based Comparative Evaluation

Evangelos Kanoulas

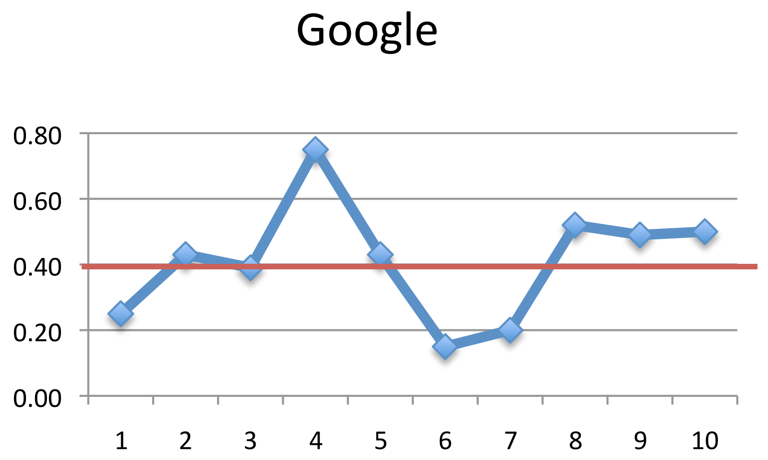
UNIVERSITY OF AMSTERDAM

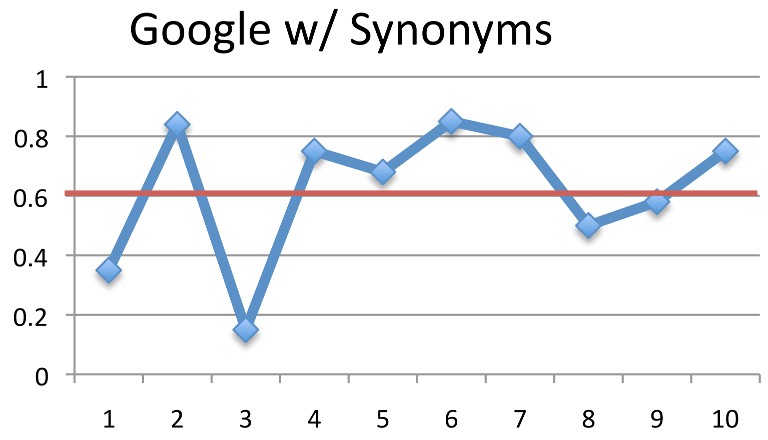
Comparing Retrieval Systems

- Hypothesis: Synonyms will improve search engine effectiveness

- Google:
  - Mean AP= 0.41

- Google w/ Synonyms
  - Mean AP= 0.63



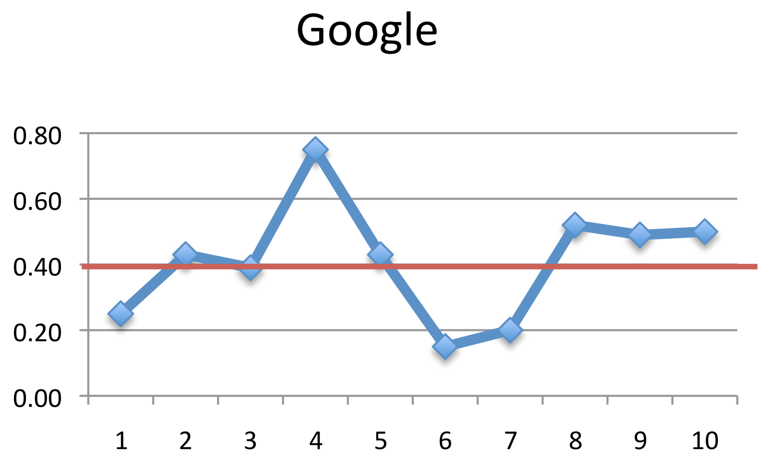


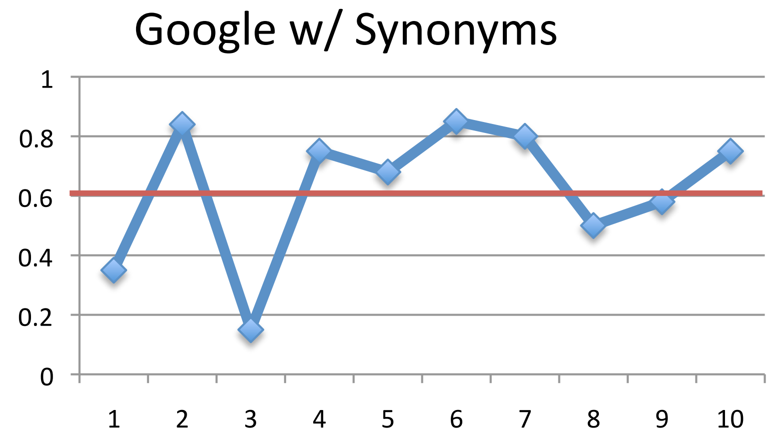
So what?

- Do these results support my hypothesis?

- Is it possible that my results are just random?

→ statistical significance testing





Statistical Significance Testing

- Two hypotheses, e.g.
  - H_0: B-A = 0
  - H_a: B-A ≠ 0 or B-A>0

- We want to prove the null hypothesis wrong

Statistical Significance Testing

- Obtain system performance measurements over a sample of queries
- Compute a test statistic t from those measurements
  - with known distribution under H_0
- Compute the p-value, i.e.
  - the probability of observing the test statistic t …
  - … under a distribution obtained by assuming H_0 is true
    - If the p-value is low, conclude H_0 is false

Commonly used tests

- Non-parametric
  - Sign test/binomial test
  - Wilcoxon signed rank test

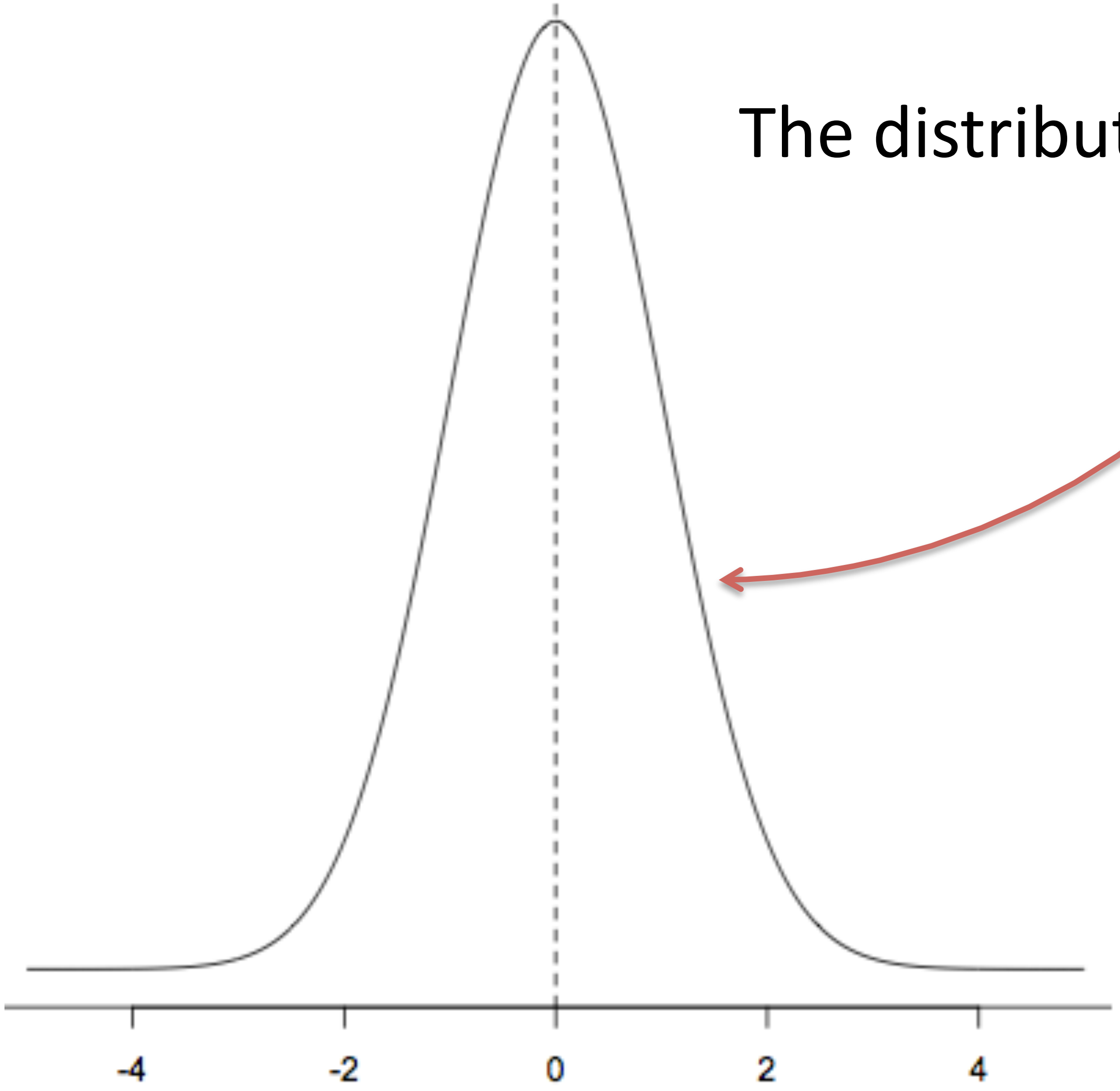
- Parametric
  - Student's t-test

- Distribution-free
  - Randomization test
  - Bootstrap test

Student’s t-test

- Statistic: $t = \frac{\overline{B - A}}{\frac{\sigma_{B-A}}{\sqrt{N}}}$
  - (Assumption)
    - mean measures follow normal distribution

Student’s t-test



The distribution of t under H_0

Student’s t-test

Query	A	B	B-A
1	.25	.35	+.10
2	.43	.84	+.41
3	.39	.15	-.24
4	.75	.75	0
5	.43	.68	+.25
6	.15	.85	+.70
7	.20	.80	+.60
8	.52	.50	-.02
9	.49	.58	+.09
10	.50	.75	+.25

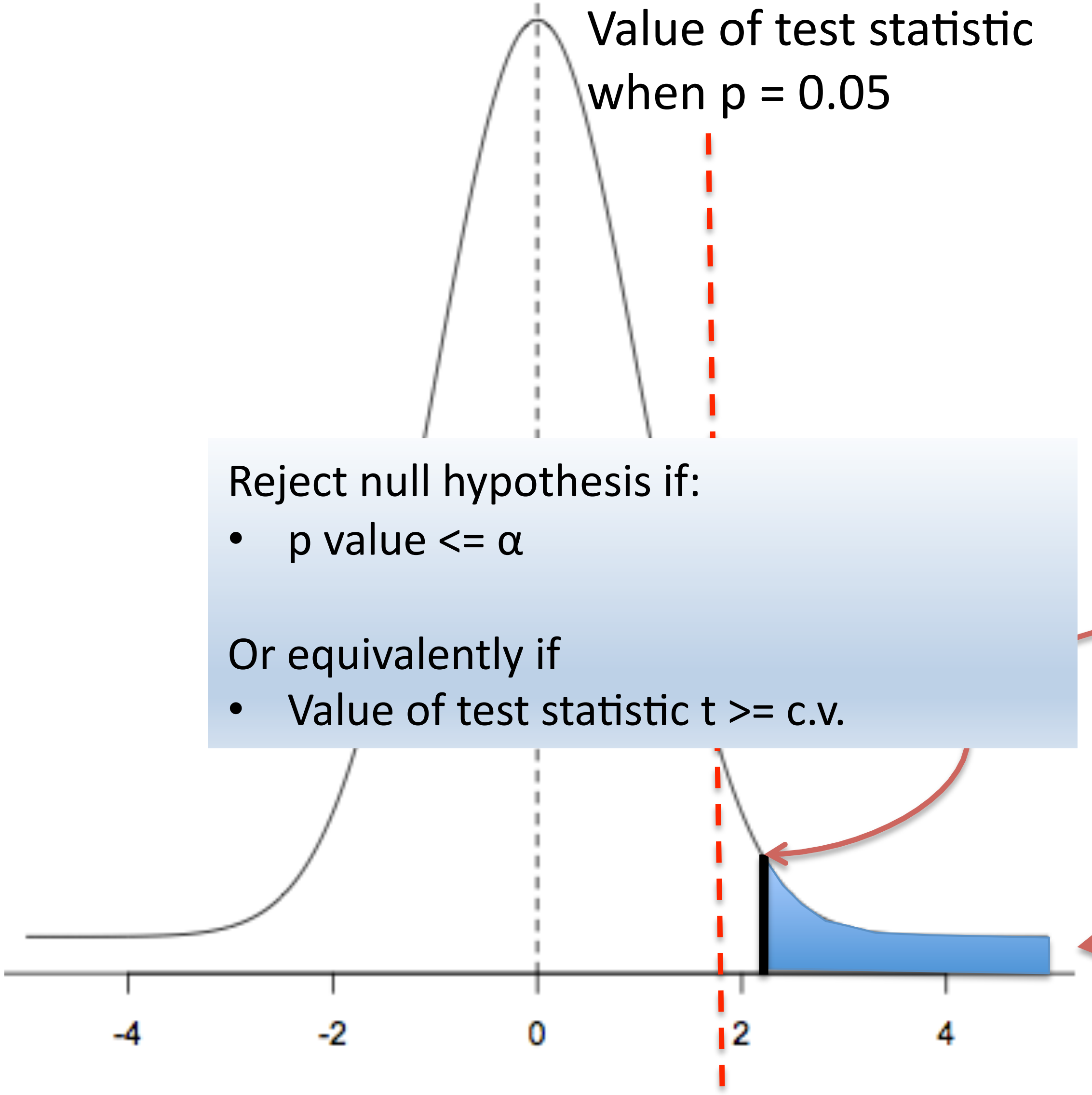
$$\hat{\mu} = \overline{B - A} = 0.214$$

$$\hat{\sigma}_{B-A} = 0.291$$

$$t = \frac{\hat{\mu}}{\hat{\sigma}_{B-A}}\sqrt{n} = 2.33$$

Student’s t-test

Critical value (c.v.):
Value of test statistic
when p = 0.05



Reject null hypothesis if:

- p value <= α

Or equivalently if

- Value of test statistic t >= c.v.

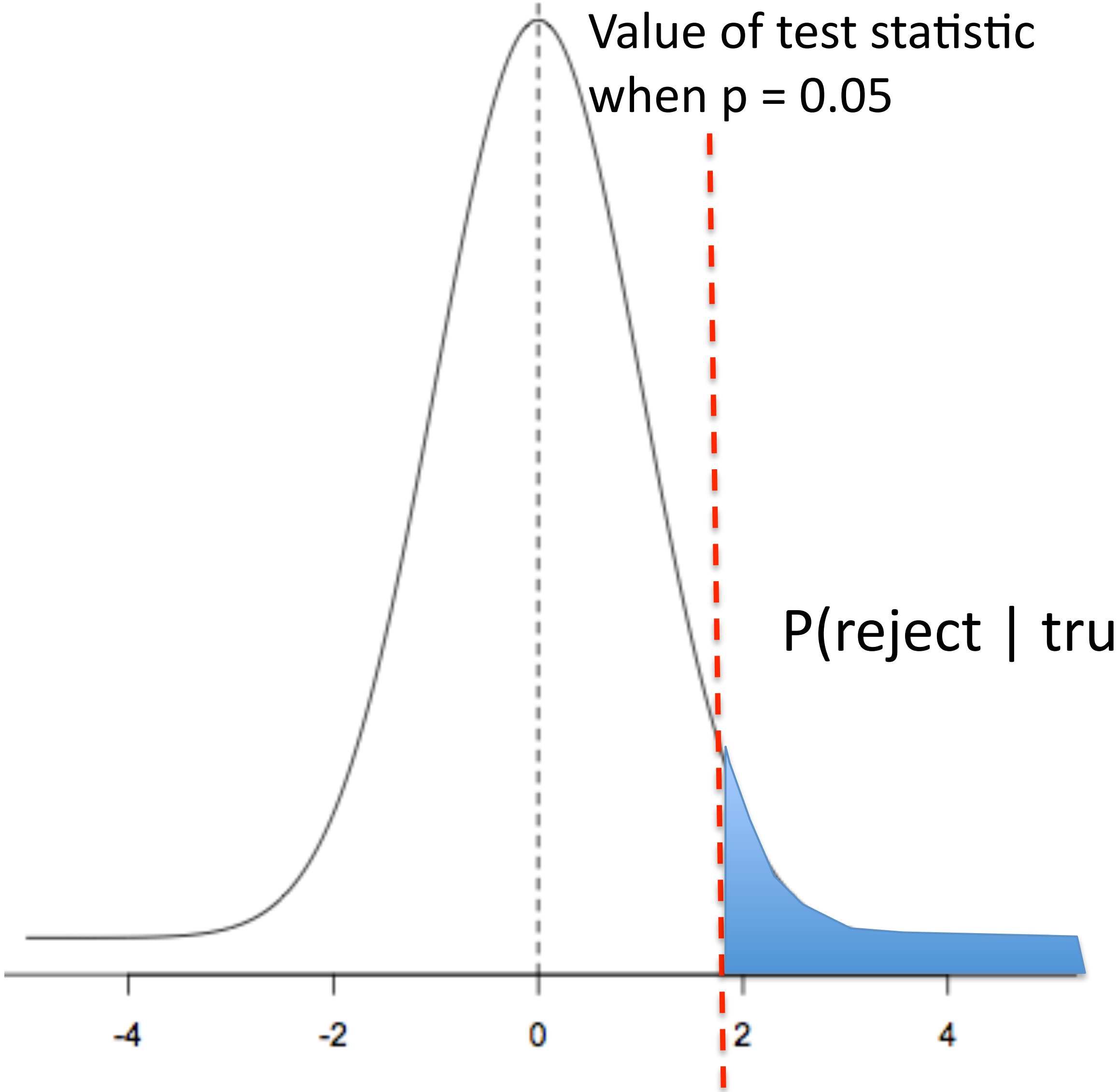
$$\hat{\mu} = \overline{B - A} = 0.214$$

$$\hat{\sigma}_{B-A} = 0.291$$

$$t = \frac{\hat{\mu}}{\hat{\sigma}_{B-A}} \sqrt{n} = 2.33$$

$$p - value = 0.02$$

Student's t-test

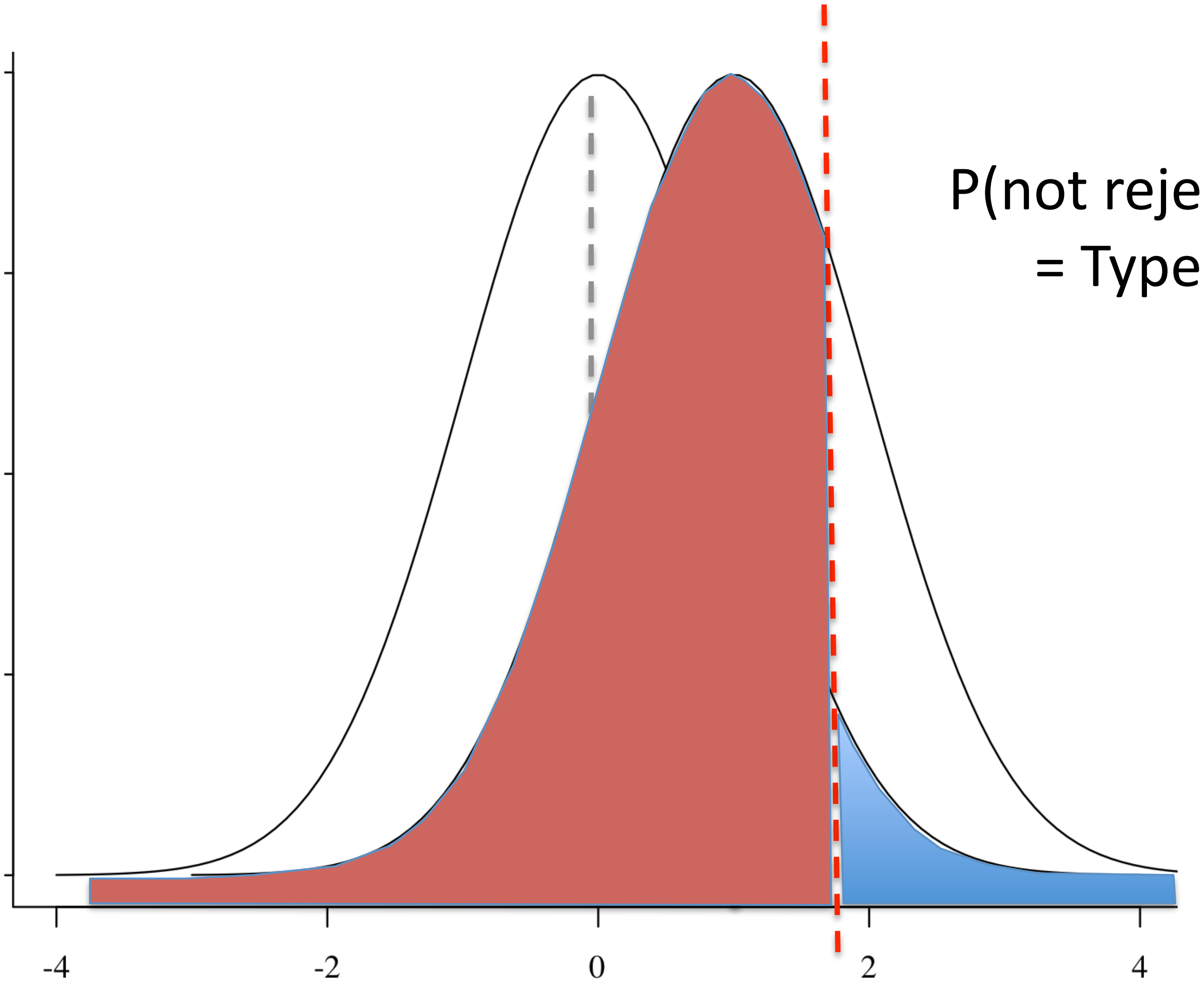


Critical value (c.v.):
Value of test statistic
when $p = 0.05$

P(reject | true) = Type I error = α

Student's t-test

Critical value (c.v.):
Value of test statistic
when p = 0.05



P(not reject | false)
= Type II error = β

Errors in Inference

- A significance test is basically a classifier

H_0	true	false
not rejected	accuracy: 1-α	Type II error: **β**
rejected	Type I error: **α**	power: 1-β

- We can’t actually know whether H_0 is true or not
  - If we could, we wouldn’t need the test

- But we set up the test to control the expected Type I (significance) and Type II (power) error rates

Expected Type I Error Rate

- Test parameter α is used to decide whether to reject H_0 or not—if $p < \alpha$, then reject H_0

- Choosing α is equivalent to stating an expected Type I error rate
  - e.g. if $p < 0.05$ is considered significant, we are saying that we expect that we will incorrectly reject H_0 5% of the time

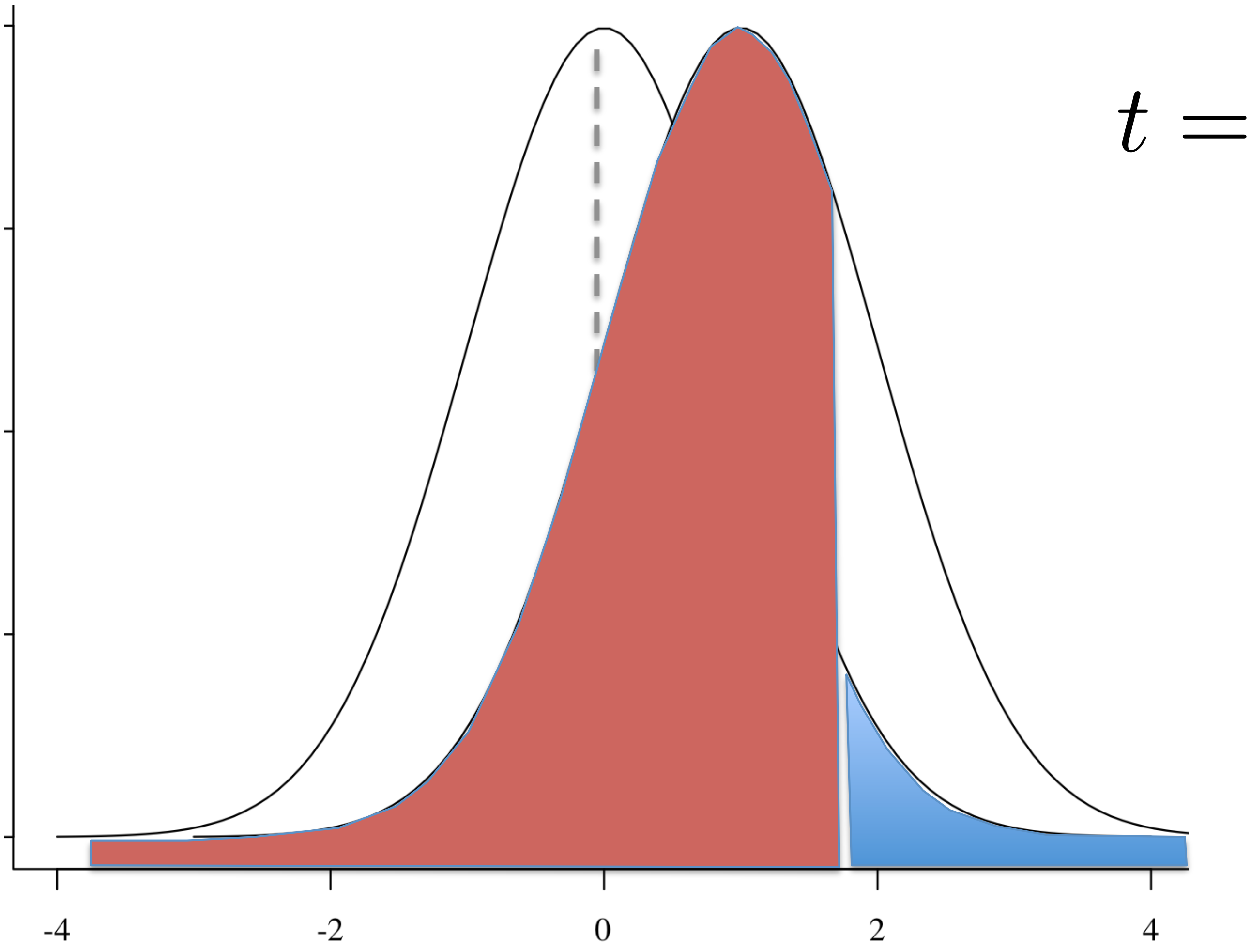
- Why?
  - Because when H_0 is true, every p-value is equally likely to be observed
  - 5% of the time we will observe a p-value less than 0.05… and therefore there is a 5% Type I error rate

Expected Type II Error Rate

- What about Type II errors?
  - False negatives are bad: if we can't reject H_0 when it's false, we may miss out on interesting results

- What is the distribution of p-values when H_0 is false?
  - Problem: there is only one way H_0 can be true, but there are many ways it can be false

Student’s t-test



Effect Size

$$t = \frac{\boxed{\frac{\overline{B-A}}{\sigma_{B-A}}}}{\sqrt{N}}$$

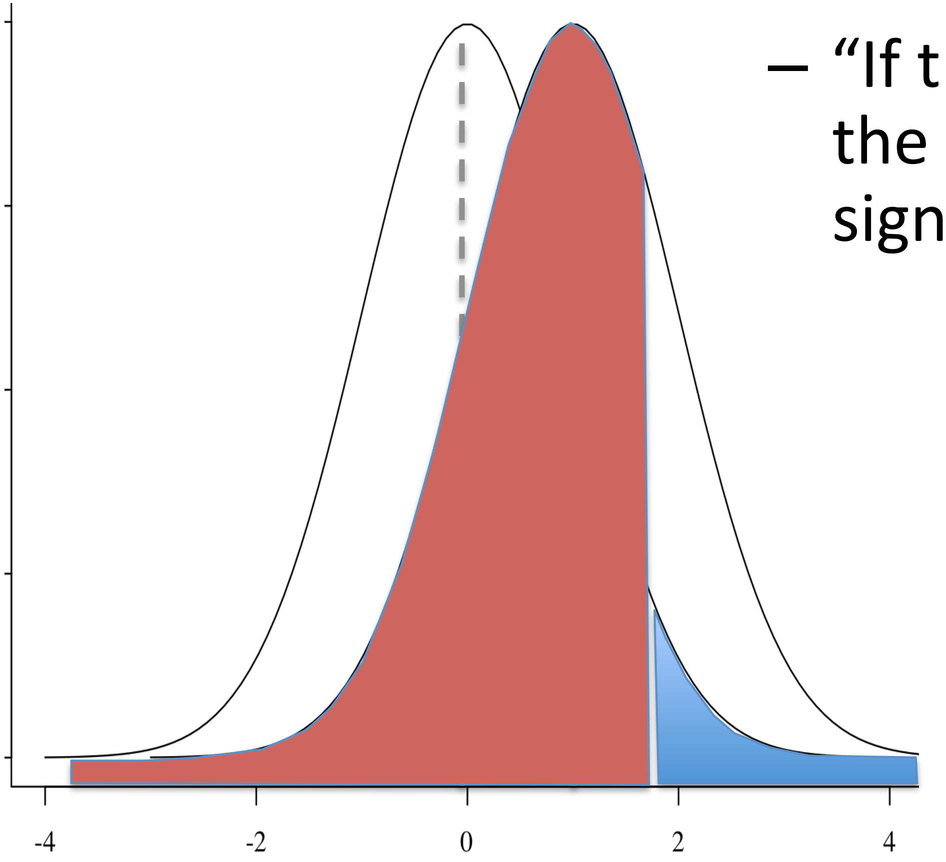
Effect Size

- A measure of the magnitude of the difference between two systems
  - Effect size is dimensionless; intuitively similar to % change in performance
  - Bigger population effect size => more likely to find a significant difference in a sample

Power and Effect Size

- Before testing, we can say "I want to be able to detect an effect size of h with probability β"
  - "If there is at least a 5% difference, the test should say the difference is significant with 80% probability"

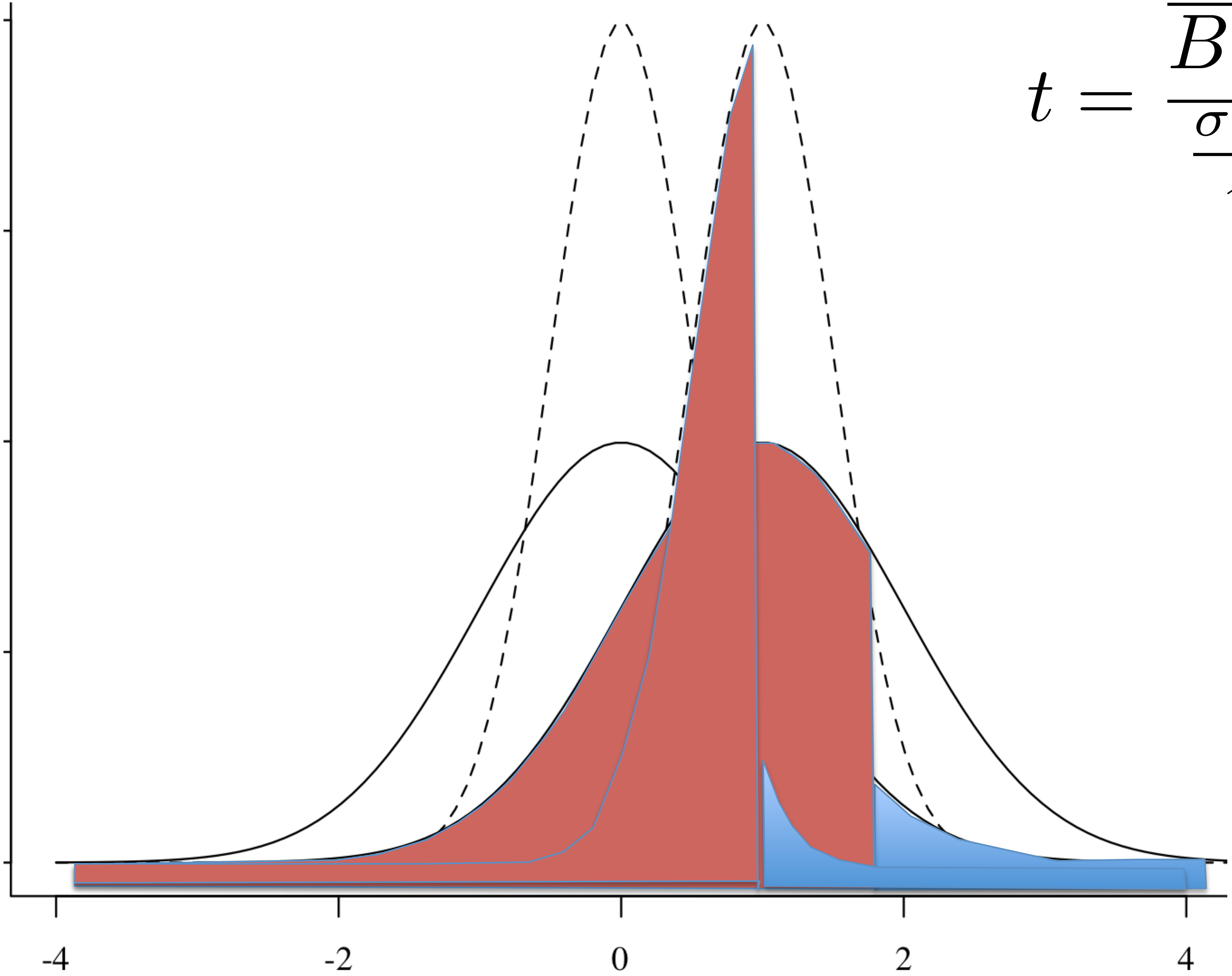
⇒ $h = 0.05$, $\beta = 0.8$



Sample Size

- Once we have chosen α, β, h, we can determine the sample size needed to make the error rates come out as desired
  - n = f(α, β, h)
  - Usually involves a linear search
  - There are software tools to do this
- Basically:
  - Sample size n increases with β if other parameters held constant
  - If you want more power, you need more queries

Sample Size



$$t = \frac{\overline{B - A}}{\frac{\sigma_{B-A}}{\sqrt{N}}}$$

Power Analysis

- Statistical significance testing:
  1. sample size
  2. effect size = diff of means / st. dev.
  3. significance level = P(Type I error) = probability of finding an effect that is not there
  4. power = 1 - P(Type II error) = probability of finding an effect that is there

- Given any three, we can determine the fourth
  - Easier under normality assumption

So far… statistics 101…

- Two sides of the same coin:
  - Statistical significance => results generalize from a sample of queries to the population
  - Power analysis => number of queries necessary to stat. detect a given difference

Why care about significance testing?

- Sources of variance specific to IR:
  - Properties of queries
  - Properties of document corpus
  - Properties of effectiveness measures
  - Assessor error and disagreement
  - Missing relevance judgments
  - Total number of relevant documents
  - …

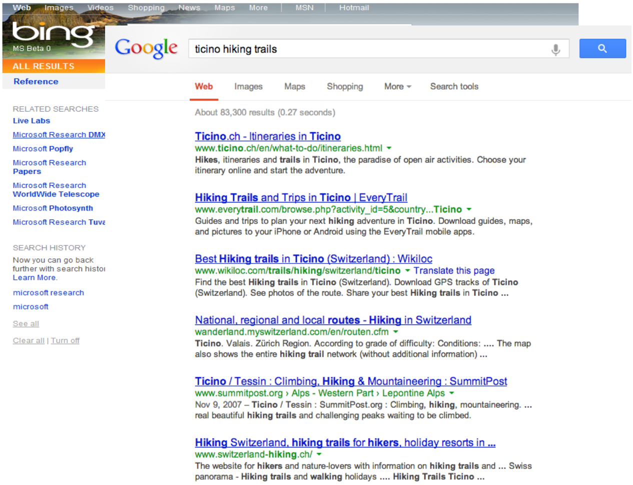
- Only variance due to queries included in standard statistical testing

=> Wrong conclusions!

Collection-based Experiment

- uses of alternative dispute resolution
- job search vancouver washington
- poem of arrival of columbus

Results



Web Images Videos Shopping News Maps More | MSN | Hotmail
bing
MS Beta 0
ALL RESULTS
Reference
Google
Web Images Maps Shopping More Search tools

ticino hiking trails

RELATED SEARCHES

Live Labs

Microsoft Research DMX

Microsoft Popfly

Microsoft Research Papers

Microsoft Research WorldWide Telescope

Microsoft Photosynth

Microsoft Research Tuva

SEARCH HISTORY

Now you can go back further with search histo Learn More.

microsoft research

microsoft

See all

Clear all | Turn off

About 83,300 results (0.27 seconds)

Ticino.ch - Itineraries in **Ticino**
www.**ticino**.ch/en/what-to-do/itineraries.html
Hikes, itineraries and **trails** in **Ticino**, the paradise of open air activities. Choose your itinerary online and start the adventure.

Hiking Trails and Trips in **Ticino** | EveryTrail
www.every**trail**.com/browse.php?activity_id=5&country...**Ticino**
Guides and trips to plan your next **hiking** adventure in **Ticino**. Download guides, maps, and pictures to your iPhone or Android using the EveryTrail mobile apps.

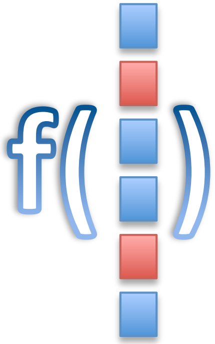
Best **Hiking trails** in **Ticino** (Switzerland) : Wikiloc
www.wikiloc.com/**trails/hiking**/switzerland/**ticino** Translate this page
Find the best **Hiking trails** in **Ticino** (Switzerland). Download GPS tracks of **Ticino** (Switzerland). See photos of the route. Share your best **Hiking trails** in **Ticino ...**

National, regional and local **routes** - **Hiking** in Switzerland
wanderland.myswitzerland.com/en/routen.cfm
Ticino. Valais. Zürich Region. According to grade of difficulty: Conditions: **....** The map also shows the entire **hiking trail** network (without additional information) **...**

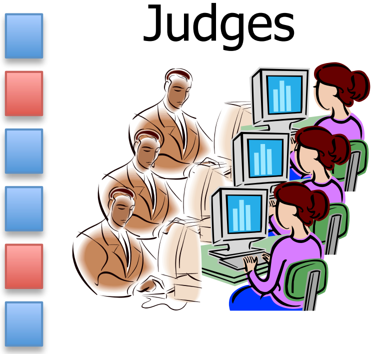
Ticino / Tessin : Climbing, **Hiking** & Mountaineering : SummitPost
www.summitpost.org › Alps - Western Part › Lepontine Alps
Nov 9, 2007 – **Ticino** / Tessin : SummitPost.org : Climbing, **hiking**, mountaineering. **...** real beautiful **hiking trails** and challenging peaks waiting to be climbed.

Hiking Switzerland, **hiking trails** for **hikers**, holiday resorts in **...**
www.switzerland-**hiking**.ch/
The website for **hikers** and nature-lovers with information on **hiking trails** and **...** Swiss panorama - **Hiking trails** and **walking** holidays **....** **Hiking Trails Ticino ...**

Search Algorithms



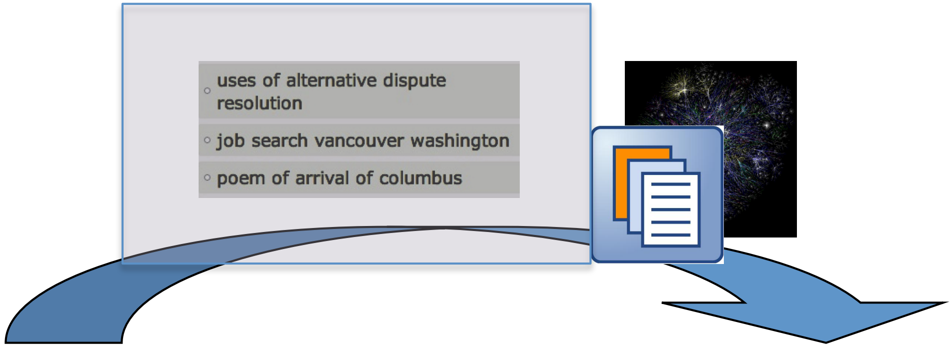
Judges



QUERY: job search vancouver washington

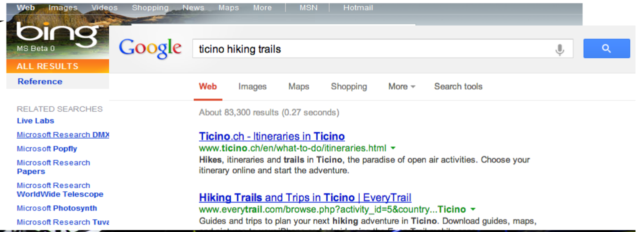
DESCRIPTION: I would like to find web page that aggregate job opportunities in the IT industry in Vancouver, Washington.

Variance due to Queries



- uses of alternative dispute resolution
- job search vancouver washington
- poem of arrival of columbus

Results



Web Images Videos Shopping News Maps More | MSN | Hotmail
bing
MS Beta 0
ALL RESULTS
Reference

RELATED SEARCHES
Live Labs
Microsoft Research DMX
Microsoft Popfly
Microsoft Research Papers
Microsoft Research WorldWide Telescope
Microsoft Photosynth
Microsoft Research Tuva

Google
Web Images Maps Shopping More Search tools

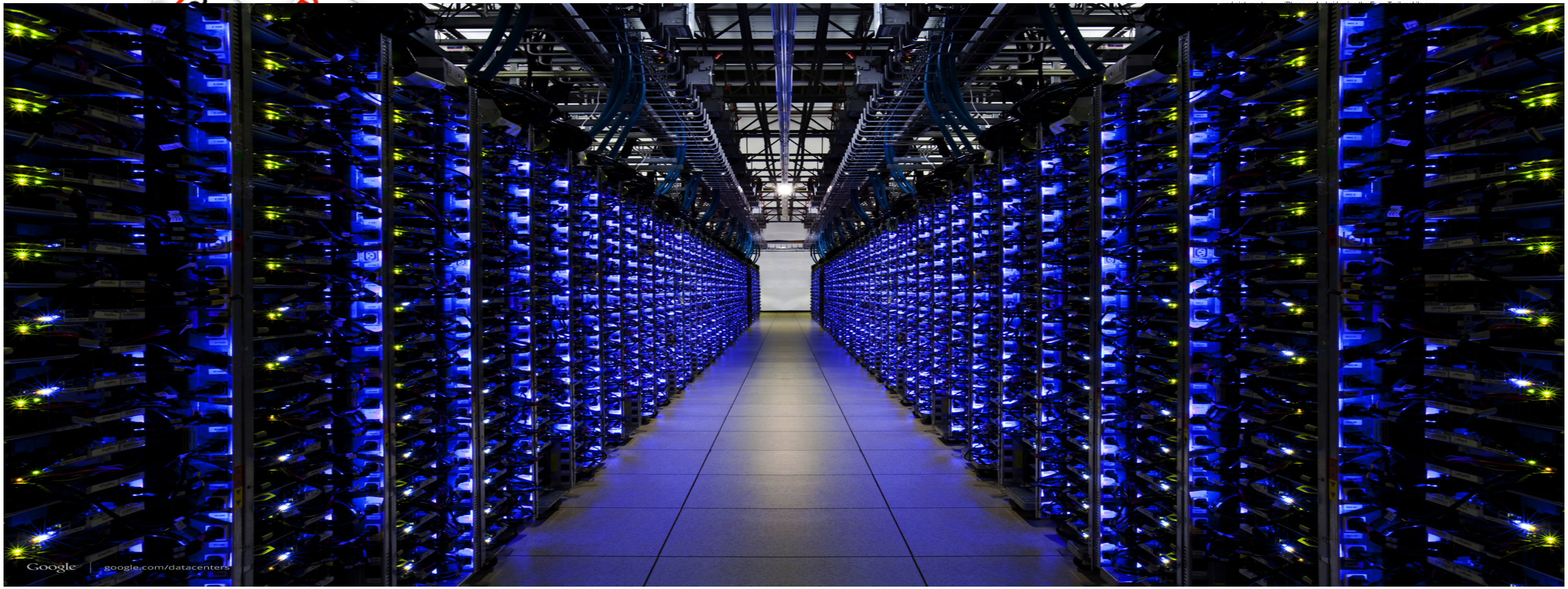
ticino hiking trails

About 83,300 results (0.27 seconds)

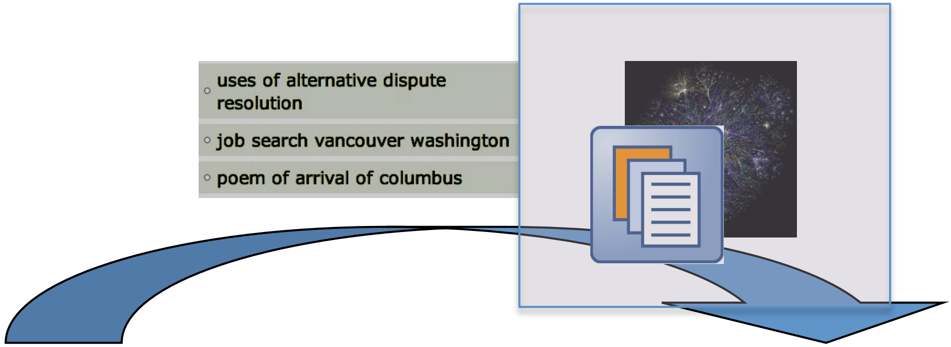
Ticino.ch - Itineraries in Ticino
www.ticino.ch/en/what-to-do/itineraries.html
Hikes, itineraries and trails in Ticino, the paradise of open air activities. Choose your itinerary online and start the adventure.

Hiking Trails and Trips in Ticino | EveryTrail
www.everytrail.com/browse.php?activity_id=5&country...Ticino
Guides and trips to plan your next hiking adventure in Ticino. Download guides, maps,

Search Algorithms

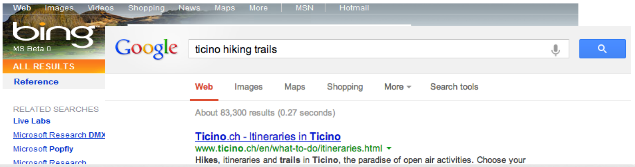


Variance due to Document Collection



- uses of alternative dispute resolution
- job search vancouver washington
- poem of arrival of columbus

Results

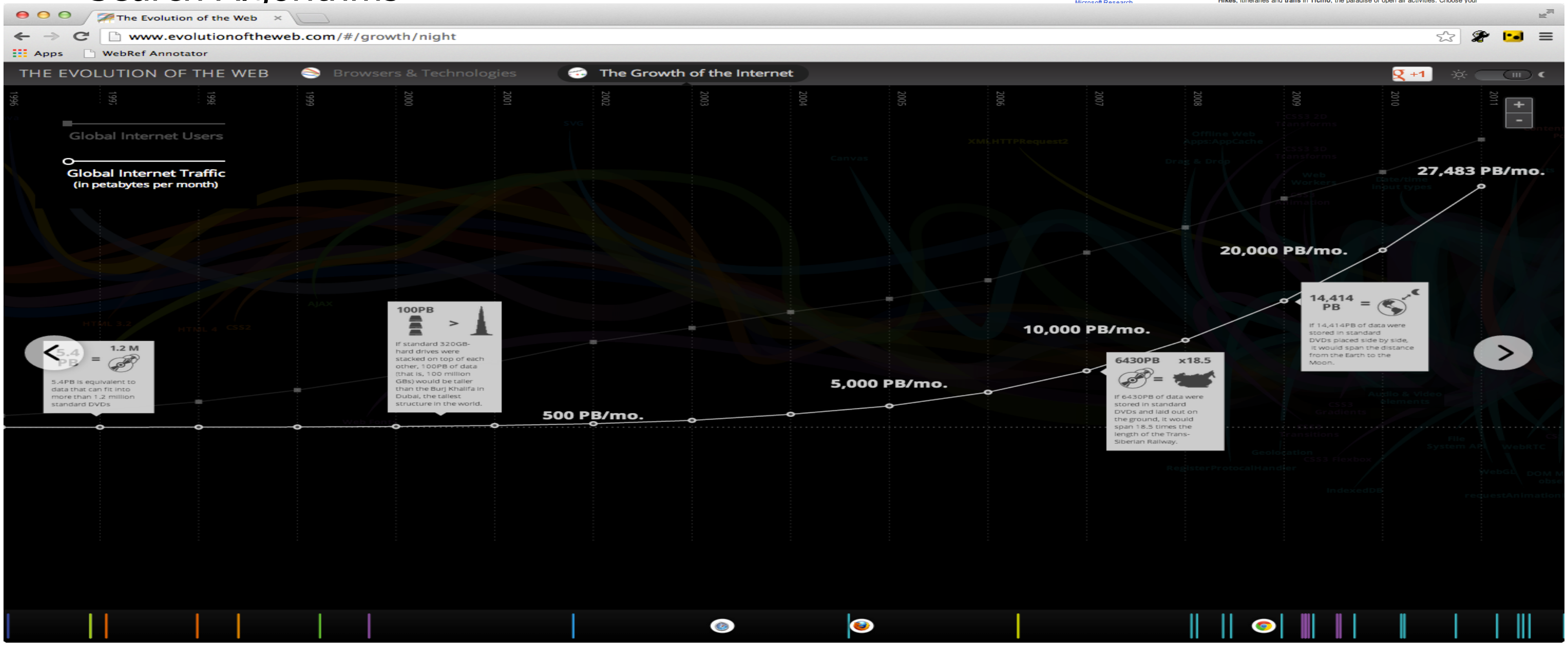


ticino hiking trails

About 83,300 results (0.27 seconds)

Ticino.ch - Itineraries in Ticino
www.ticino.ch/en/what-to-do/itineraries.html
Hikes, itineraries and trails in Ticino, the paradise of open air activities. Choose your

Search Algorithms



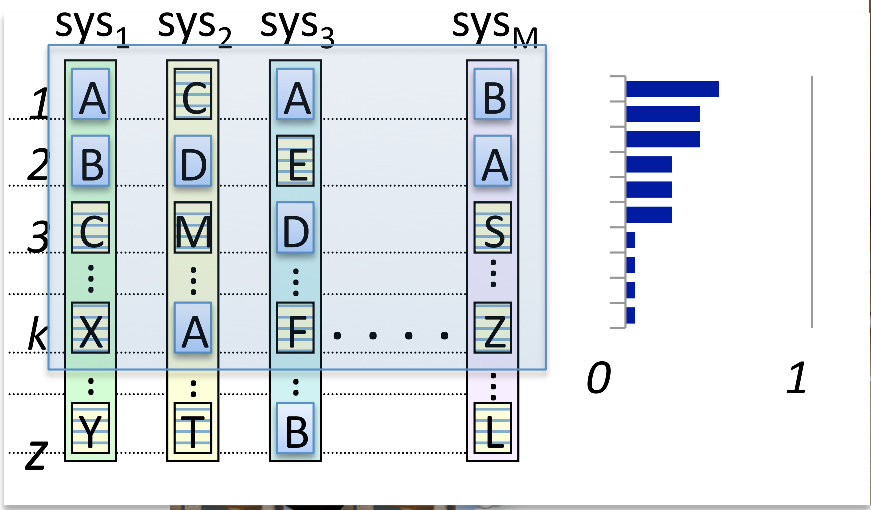
5.4PB is equivalent to data that can fit into more than 1.2 million standard DVDs

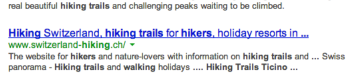
If standard 320GB-hard drives were stacked on top of each other, 100PB of data (that is, 100 million GBs) would be taller than the Burj Khalifa in Dubai, the tallest structure in the world.

If 6430PB of data were stored in standard DVDs and laid out on the ground, it would span 18.5 times the length of the Trans-Siberian Railway.

If 14,414PB of data were stored in standard DVDs placed side by side, it would span the distance from the Earth to the Moon.

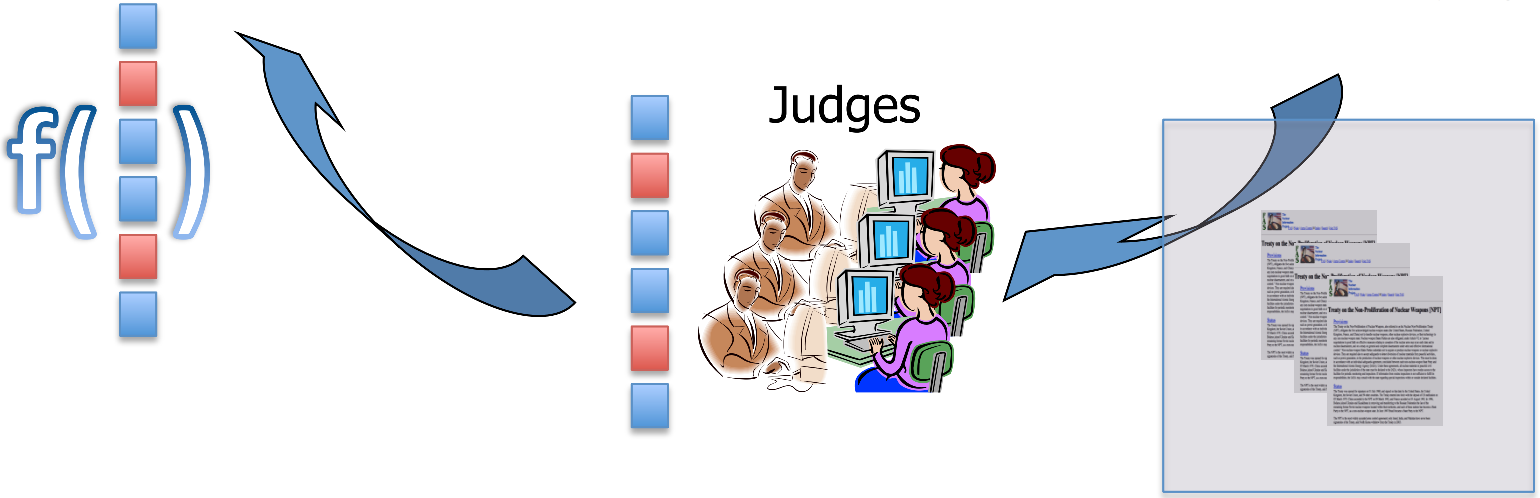
Variance due to missing judgments





real beautiful **hiking trails** and challenging peaks waiting to be climbed.

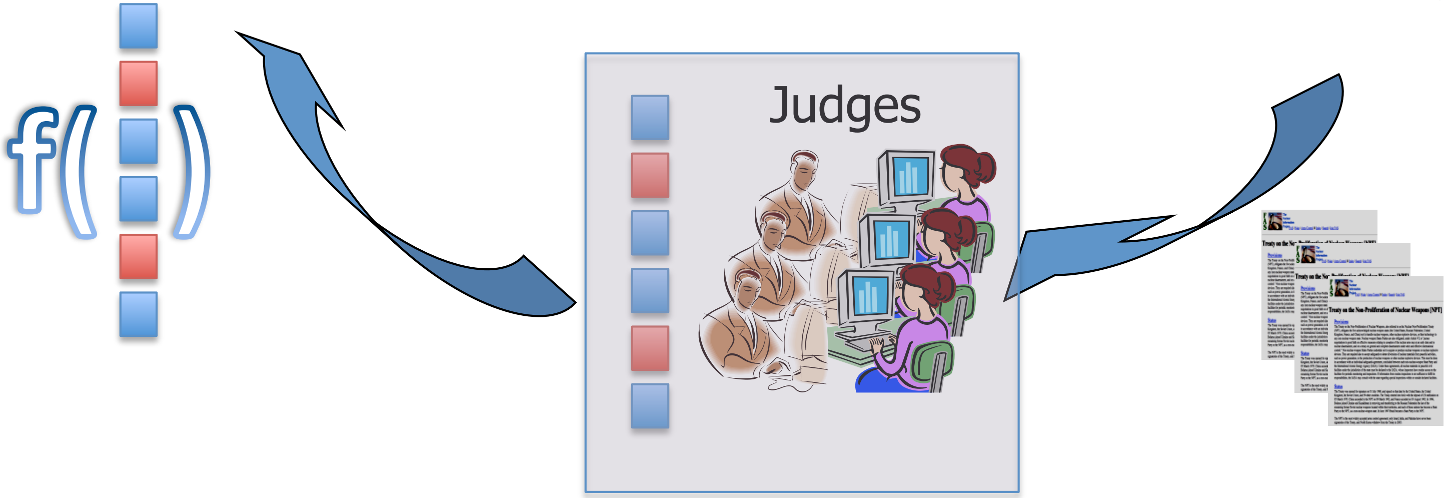
Hiking Switzerland, hiking trails for hikers, holiday resorts in ...
www.switzerland-hiking.ch/
The website for **hikers** and nature-lovers with information on **hiking trails** and ... Swiss panorama - **Hiking trails** and **walking** holidays .... **Hiking Trails** Ticino ...



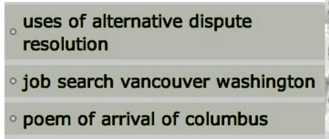
Ticino. Valais. Zürich Region. According to grade of difficulty: Conditions: **....** The map also shows the entire **hiking trail** network (without additional information) **...**

Ticino / Tessin : Climbing, **Hiking** & Mountaineering : SummitPost
www.summitpost.org › Alps - Western Part › Lepontine Alps
Nov 9, 2007 – **Ticino** / Tessin : SummitPost.org : Climbing, **hiking**, mountaineering. **...** real beautiful **hiking trails** and challenging peaks waiting to be climbed.

Hiking Switzerland, **hiking trails** for **hikers**, holiday resorts in **...**
www.switzerland-**hiking**.ch/
The website for **hikers** and nature-lovers with information on **hiking trails** and **...** Swiss panorama - **Hiking trails** and **walking** holidays **....** **Hiking Trails Ticino ...**

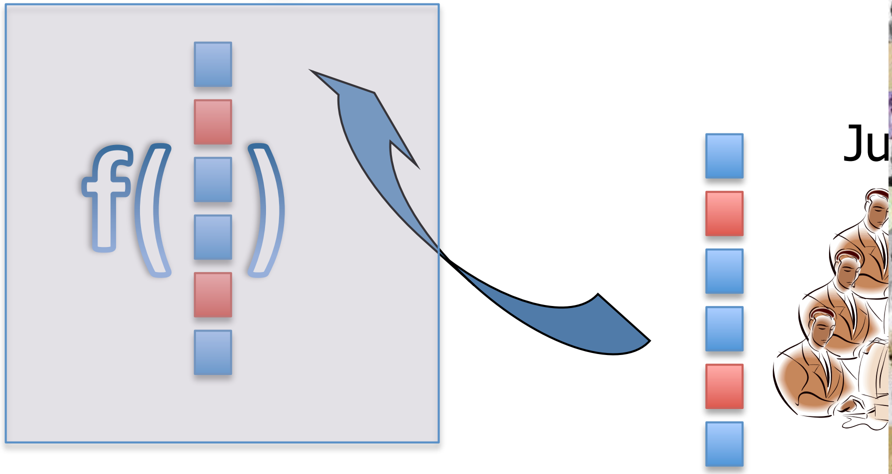


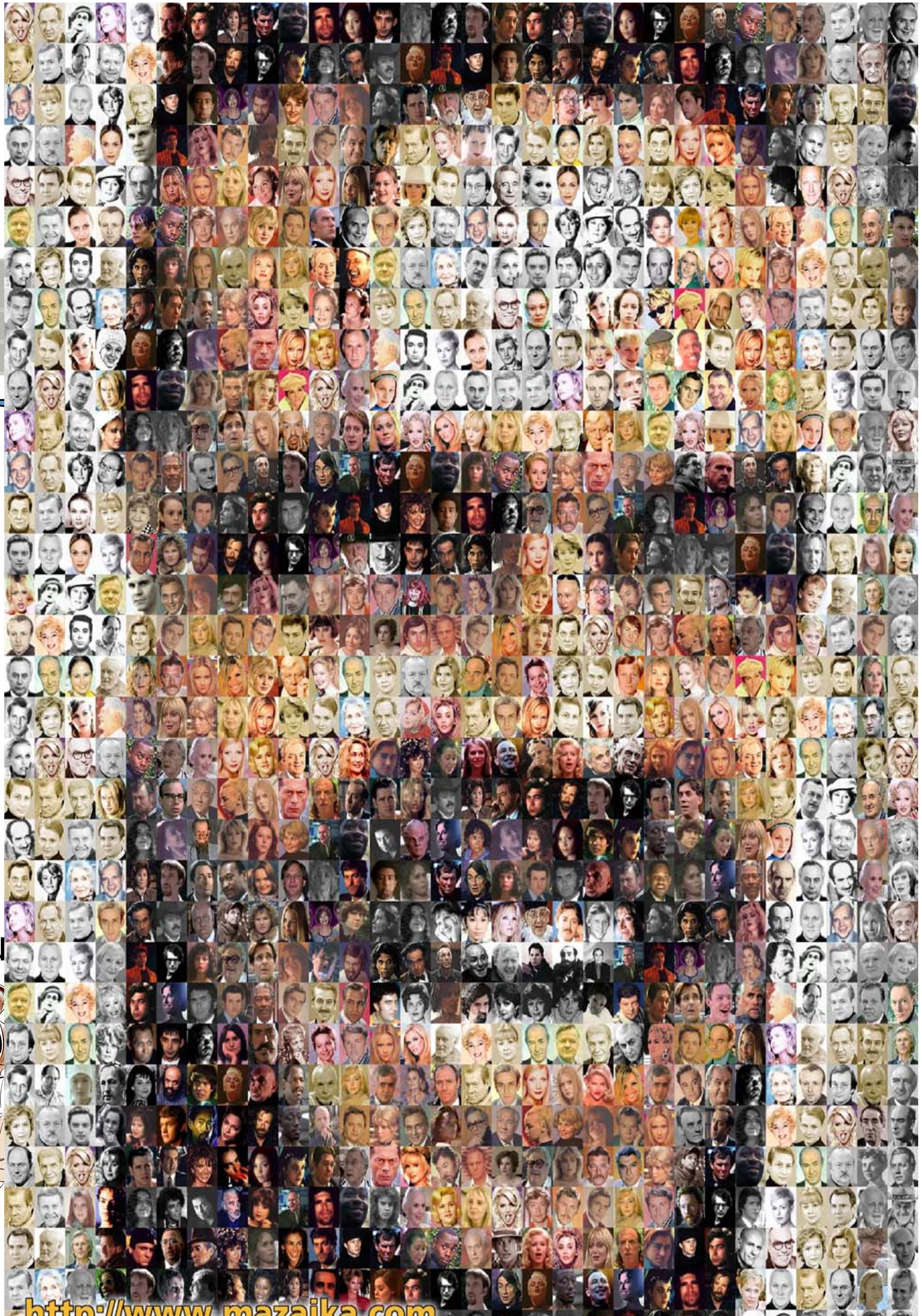
Variance due to Measure Parameters



- uses of alternative dispute resolution
- job search vancouver washington
- poem of arrival of columbus

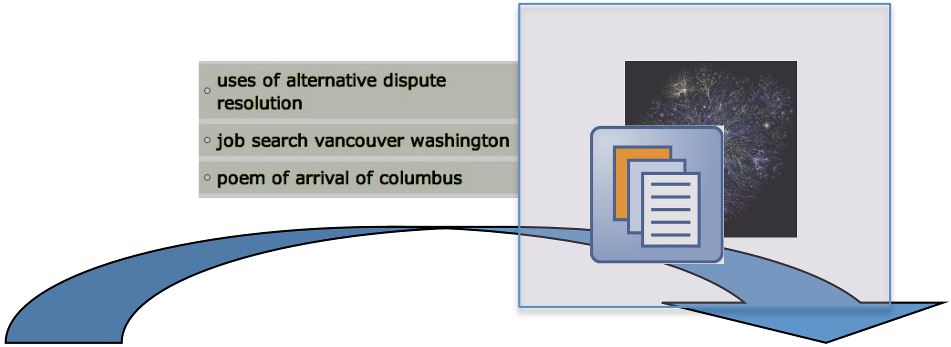
Search Algorithms





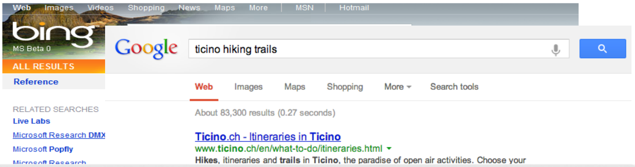
http://www.mazaika.com

Variance due to Document Collection



- uses of alternative dispute resolution
- job search vancouver washington
- poem of arrival of columbus

Results

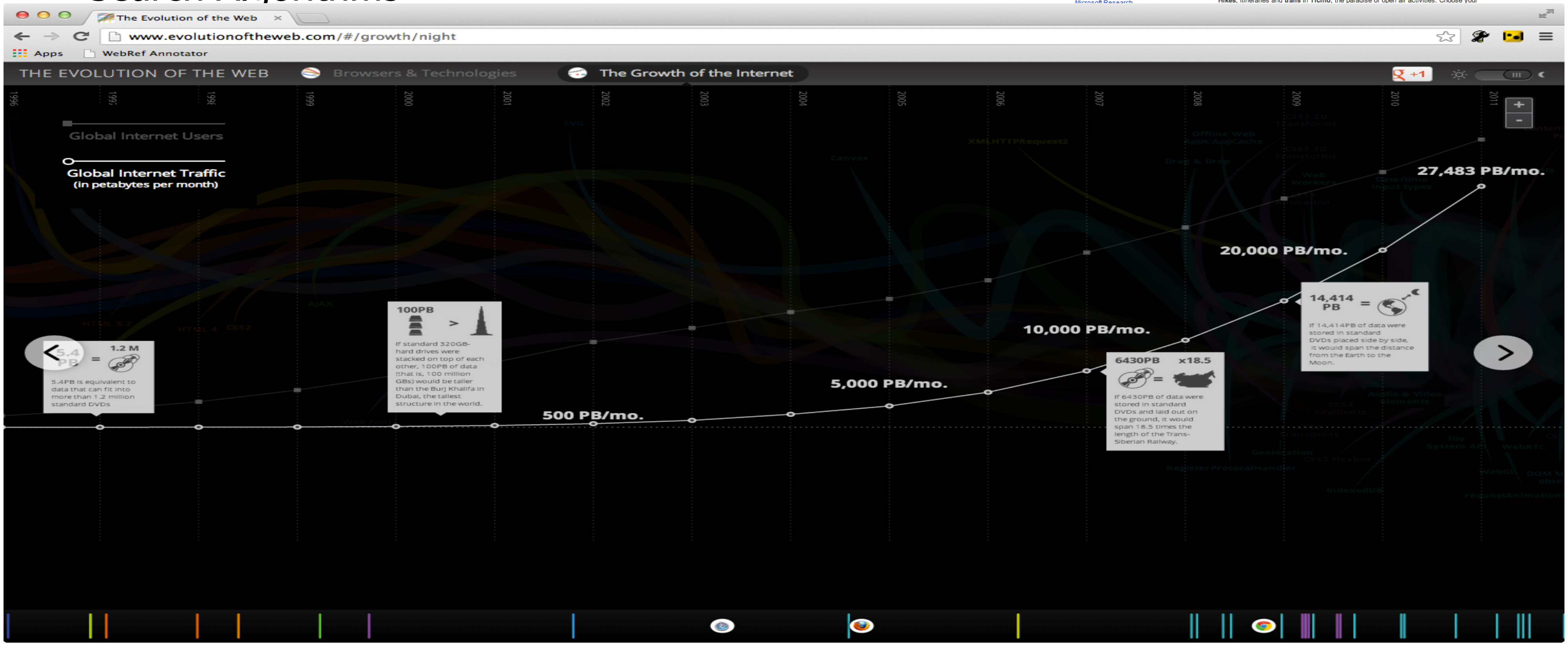


ticino hiking trails

About 83,300 results (0.27 seconds)

Ticino.ch - Itineraries in Ticino
www.ticino.ch/en/what-to-do/itineraries.html
Hikes, itineraries and trails in Ticino, the paradise of open air activities. Choose your

Search Algorithms



www.evolutionoftheweb.com/#/growth/night

5.4PB is equivalent to data that can fit into more than 1.2 million standard DVDs

If standard 320GB-hard drives were stacked on top of each other, 100PB of data (that is, 100 million GBs) would be taller than the Burj Khalifa in Dubai, the tallest structure in the world.

If 6430PB of data were stored in standard DVDs and laid out on the ground, it would span 18.5 times the length of the Trans-Siberian Railway.

If 14,414PB of data were stored in standard DVDs placed side by side, it would span the distance from the Earth to the Moon.

Variance due to Document Collection

- The document collection is not absolute
  - may think of it as a sample
    - from some large/infinite universe of possible items

- Each query measurement is an estimate
  - of a population measure
    - one query, population of documents

- Quality of estimate varies between topics
  - therefore a mean is misleading
    - and so is a t-test

Motivation Hypothesis

A test document collection

should be thought of as

a sample from some hypothetical universe

of possible documents

Statistical significance

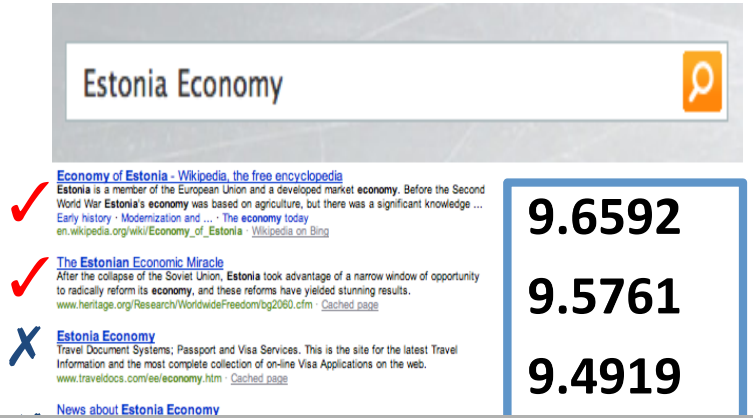
- Traditional significance testing:
  - consider the *queries* as a sample from some universe
    - what does this sample tell us about the population?

A test document collection should be thought of as a sample from some hypothetical universe

- Now we have two simultaneous sampling processes
  - need to revise the question
    - what does this (sample x sample) tell us about the (population x population)?

Simulation of multiple collections

- How can you compute a measure over multiple collections?
  - Consider multiple collections
  - Simulate multiple collections



Estonia Economy

✓ Economy of Estonia - Wikipedia, the free encyclopedia
Estonia is a member of the European Union and a developed market economy. Before the Second World War Estonia's economy was based on agriculture, but there was a significant knowledge ...
Early history · Modernization and ... · The economy today
en.wikipedia.org/wiki/Economy_of_Estonia · Wikipedia on Bing

✓ The Estonian Economic Miracle
After the collapse of the Soviet Union, Estonia took advantage of a narrow window of opportunity to radically reform its economy, and these reforms have yielded stunning results.
www.heritage.org/Research/WorldwideFreedom/bg2060.cfm · Cached page

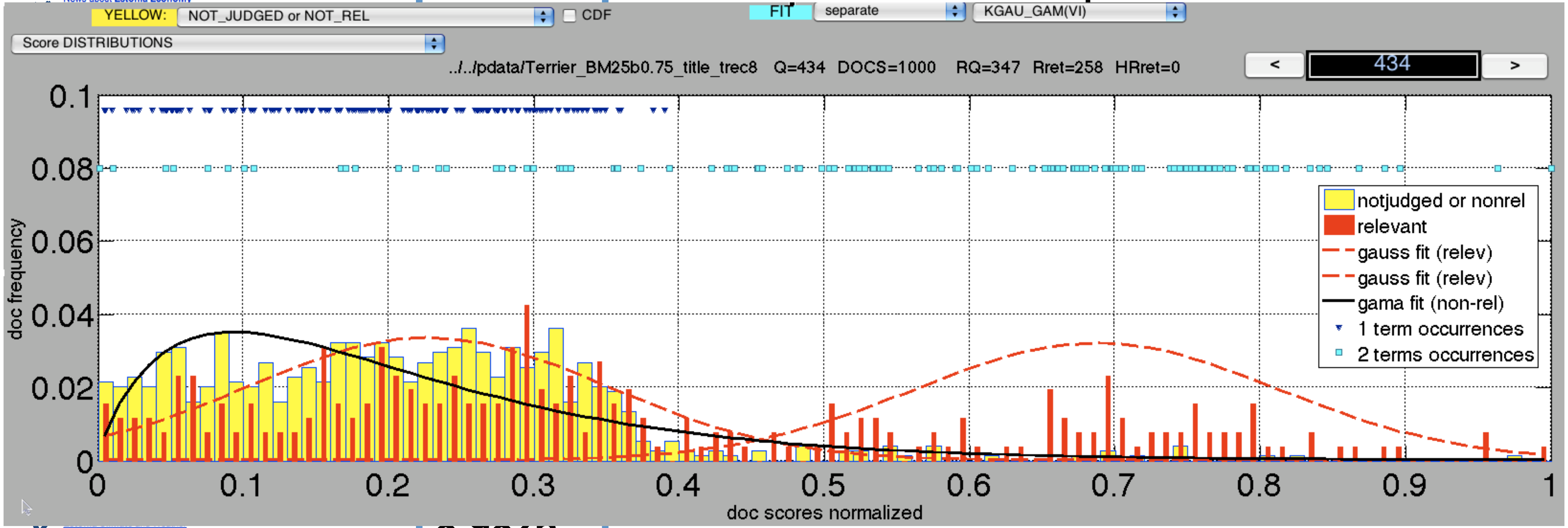
✗ Estonia Economy
Travel Document Systems; Passport and Visa Services. This is the site for the latest Travel Information and the most complete collection of on-line Visa Applications on the web.
www.traveldocs.com/ee/economy.htm · Cached page

News about Estonia Economy

9.6592

9.5761

9.4919



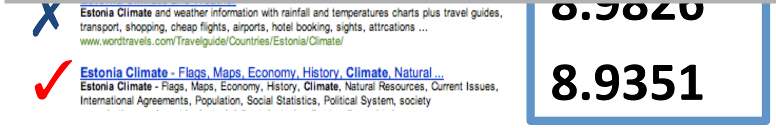
YELLOW: NOT_JUDGED or NOT_REL ☐ CDF

FIT separate KGAU_GAM(VI)

Score DISTRIBUTIONS

../../pdata/Terrier_BM25b0.75_title_trec8 Q=434 DOCS=1000 RQ=347 Rret=258 HRret=0

< 434 >



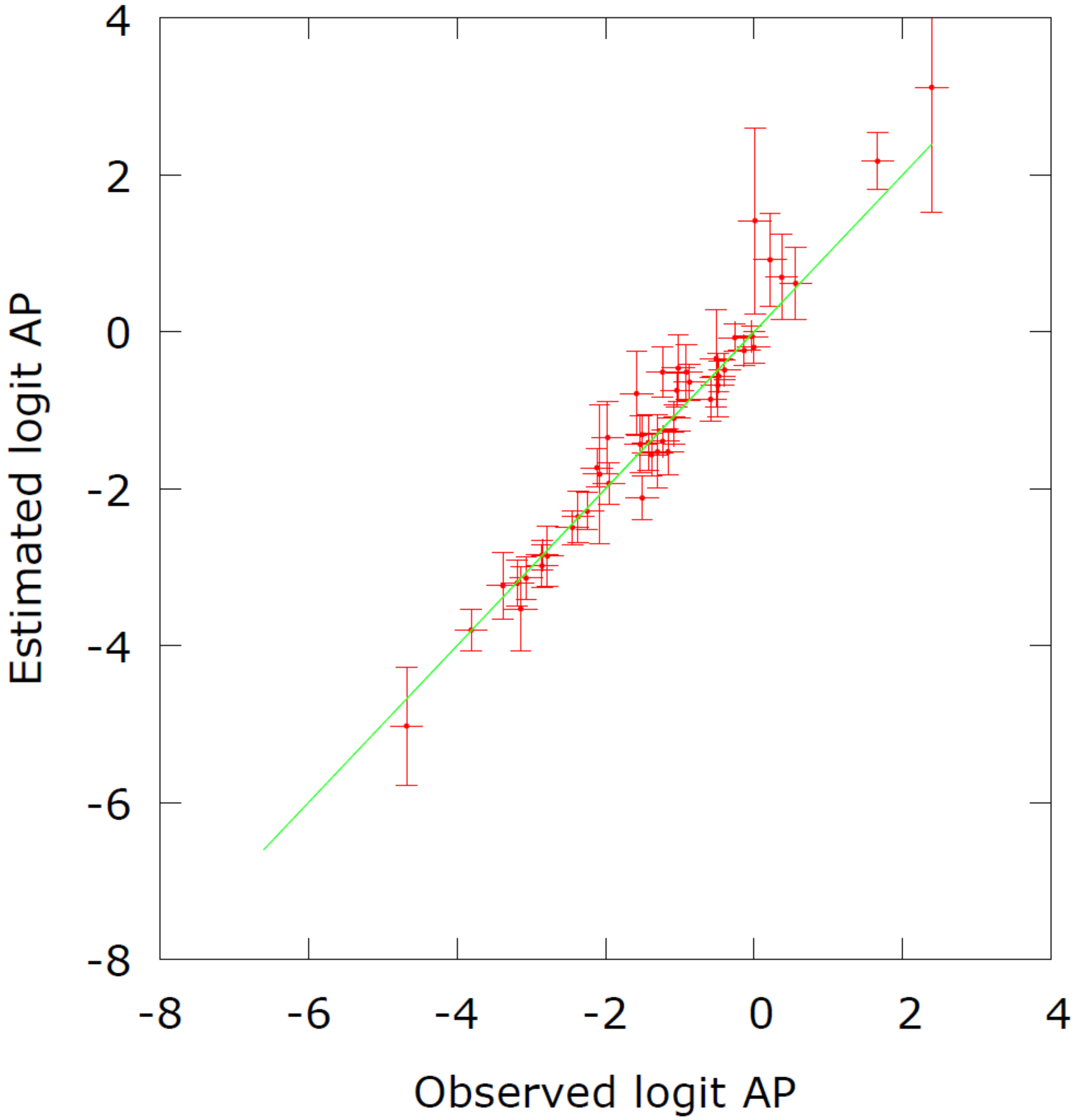
✗ Estonia Climate and weather information with rainfall and temperatures charts plus travel guides, transport, shopping, cheap flights, airports, hotel booking, sights, attrcations ...
www.wordtravels.com/Travelguide/Countries/Estonia/Climate/

✓ Estonia Climate - Flags, Maps, Economy, History, Climate, Natural ...
Estonia Climate - Flags, Maps, Economy, History, Climate, Natural Resources, Current Issues, International Agreements, Population, Social Statistics, Political System, society

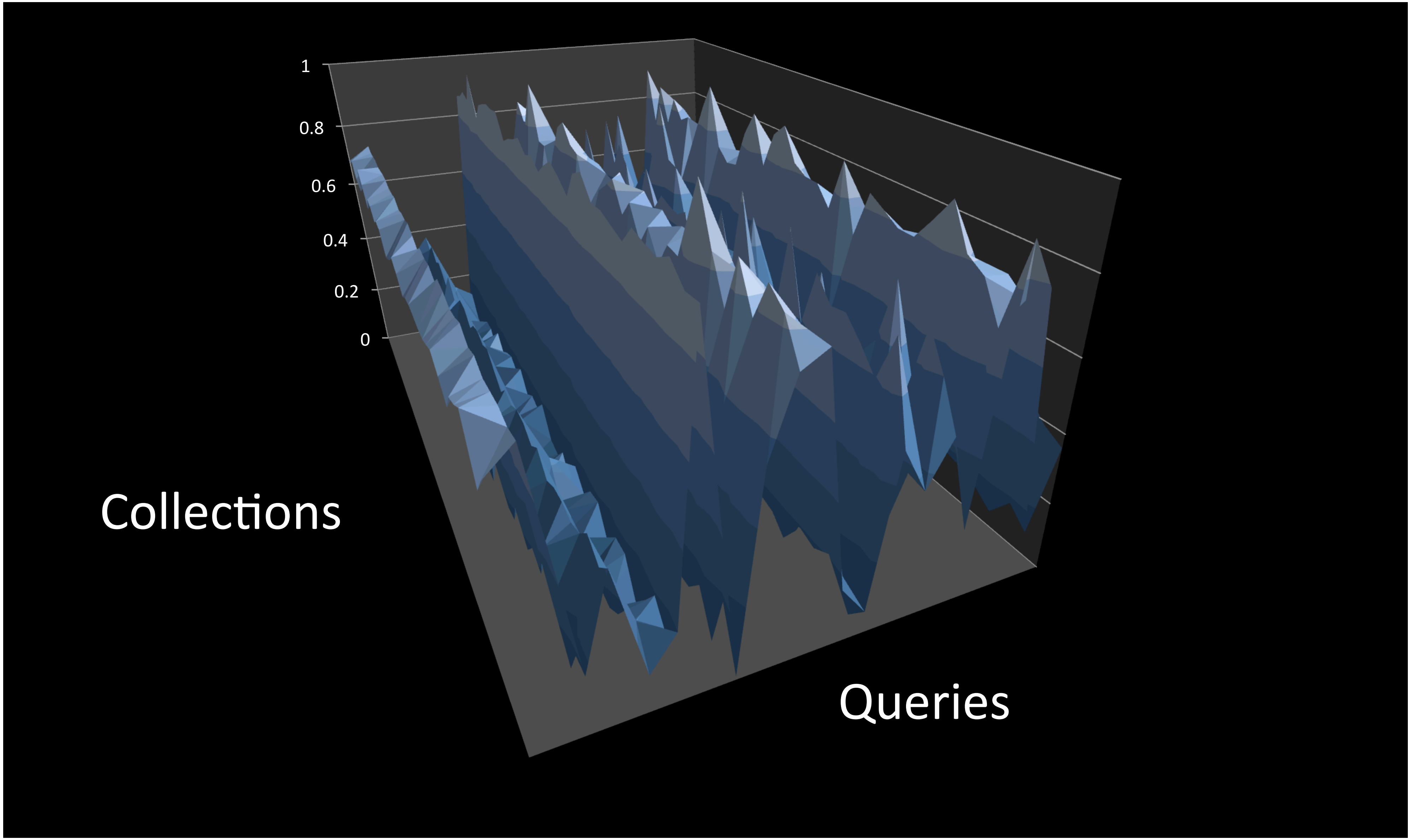
8.9826

8.9351

Single Query Measurements



Single System Measurements

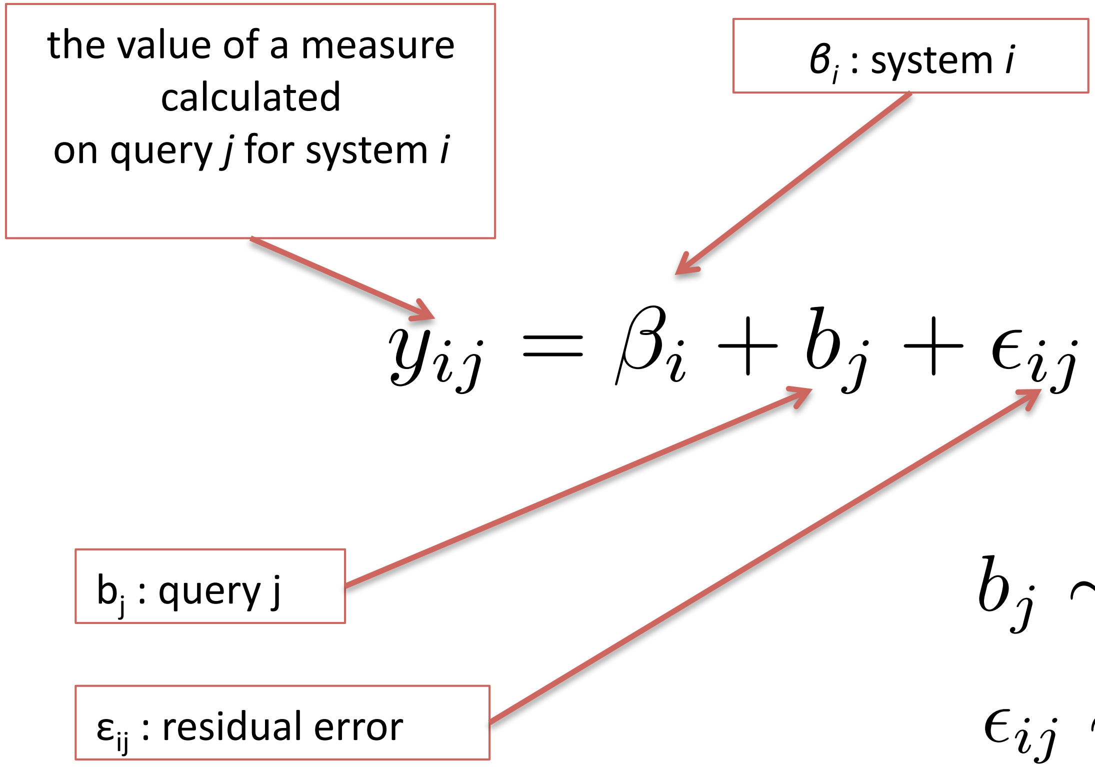


The Linear Model

- The t-test is based on a linear regression model

The Linear Model

- The t-test is based on a linear regression model



the value of a measure calculated on query *j* for system *i*

β

$$y_{ij} = \beta_i + b_j + \epsilon_{ij}$$

β_i : system *i*

b_j : query j

ε_{ij} : residual error

$$b_j \sim \mathcal{N}(0, \sigma_1^2)$$

$$\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$$

The Linear Model

- In the statistical programming environment **R**

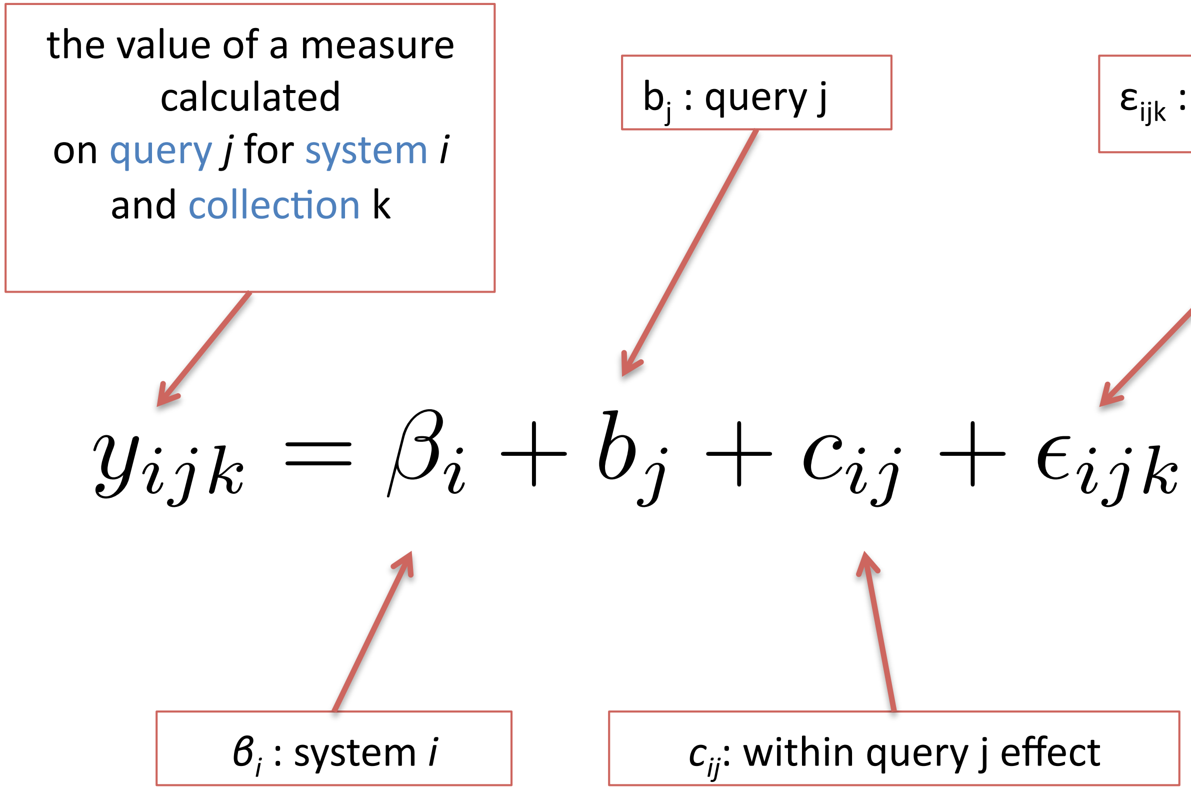
```
      lme(effectiveness ~ system, data=data, random=~1|
        query)
```

… equivalent to …

```
      t.test(effectiveness ~ system, data=data,
paired=TRUE)
```

Mixed Effects Models

- Two sources of variance
  - Query effect
  - Collection effect (within query variance)



the value of a measure calculated on query *j* for system *i* and collection k

b_j : query j

ε_{ijk} : residual error

$$y_{ijk} = \beta_i + b_j + c_{ij} + \epsilon_{ijk}$$

β_i : system *i*

c_{ij}: within query j effect

$$b_j \sim \mathcal{N}(0, \sigma_1^2)$$

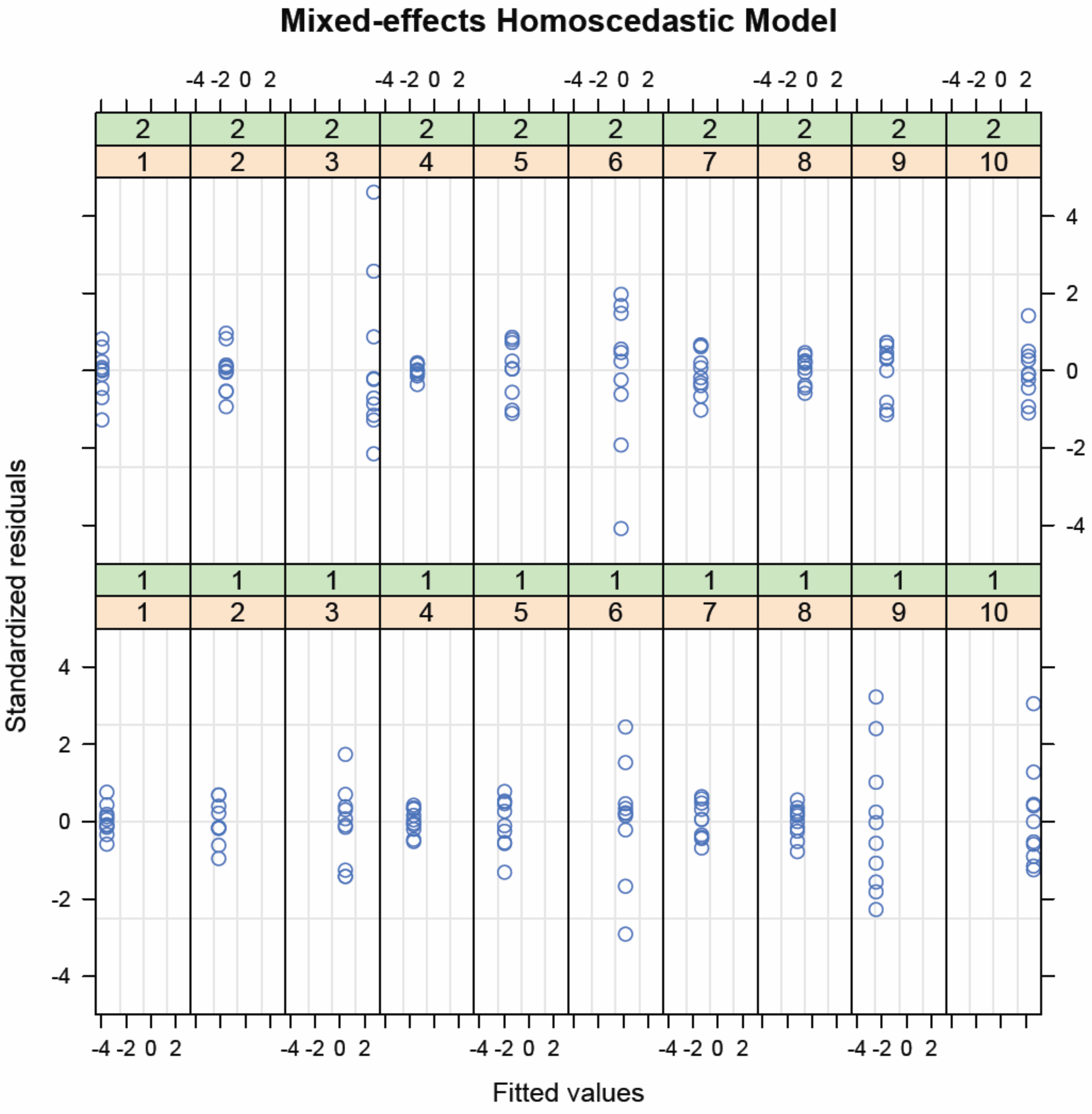
$$c_{ij} \sim \mathcal{N}(0, \sigma_2^2)$$

$$\epsilon_{ijk} \sim \mathcal{N}(0, \sigma^2)$$

Mixed Effects Models

- In the statistical programming environment **R**

```
lme1 <- lme(effectiveness~system, data=df, random=~1|query/system)

summary(lme1)

Random effects:
  Formula: ~1 | query
          (Intercept)
StdDev:   1.539644

  Formula: ~1 | system %in% query
          (Intercept)   Residual
StdDev:       0.6191864 0.6386645

Fixed effects: y ~ system
               Value     Std.Error      DF      t-value        p-value
(Intercept)        -1.3445  0.2438470  846          -5.514077      0.0000
system2          0.0999  0.1343512  46             0.744112     0.4606
```


Mixed-effects Homoscedastic Model
Standardized residuals
Fitted values
-4 -2 0 2
4
2
0
-2
-4
2
1
2
3
4
5
6
7
8
9
10

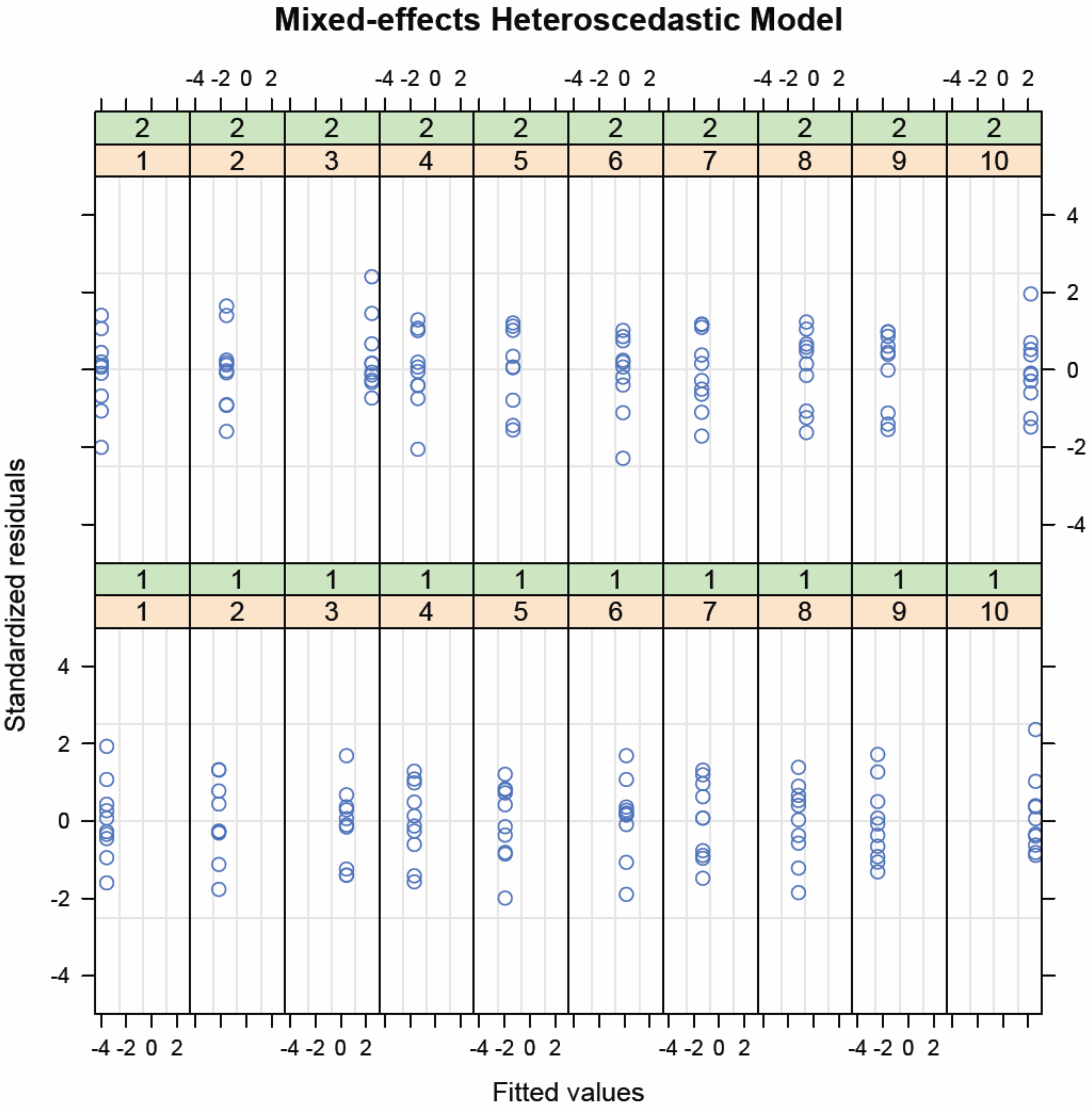
Mixed Effects Models

- Two sources of variance
  - Query effect
  - Collection effect (within query variance)

$$y_{ijk} = \beta_i + b_j + c_{ij} + \epsilon_{ijk}$$

$$b_j \sim \mathcal{N}(0, \sigma_1^2) \qquad \boxed{c_{ij} \sim \mathcal{N}(0, \sigma_{ij}^2)} \qquad \epsilon_{ijk} \sim \mathcal{N}(0, \sigma^2)$$

Heteroscedastic Model

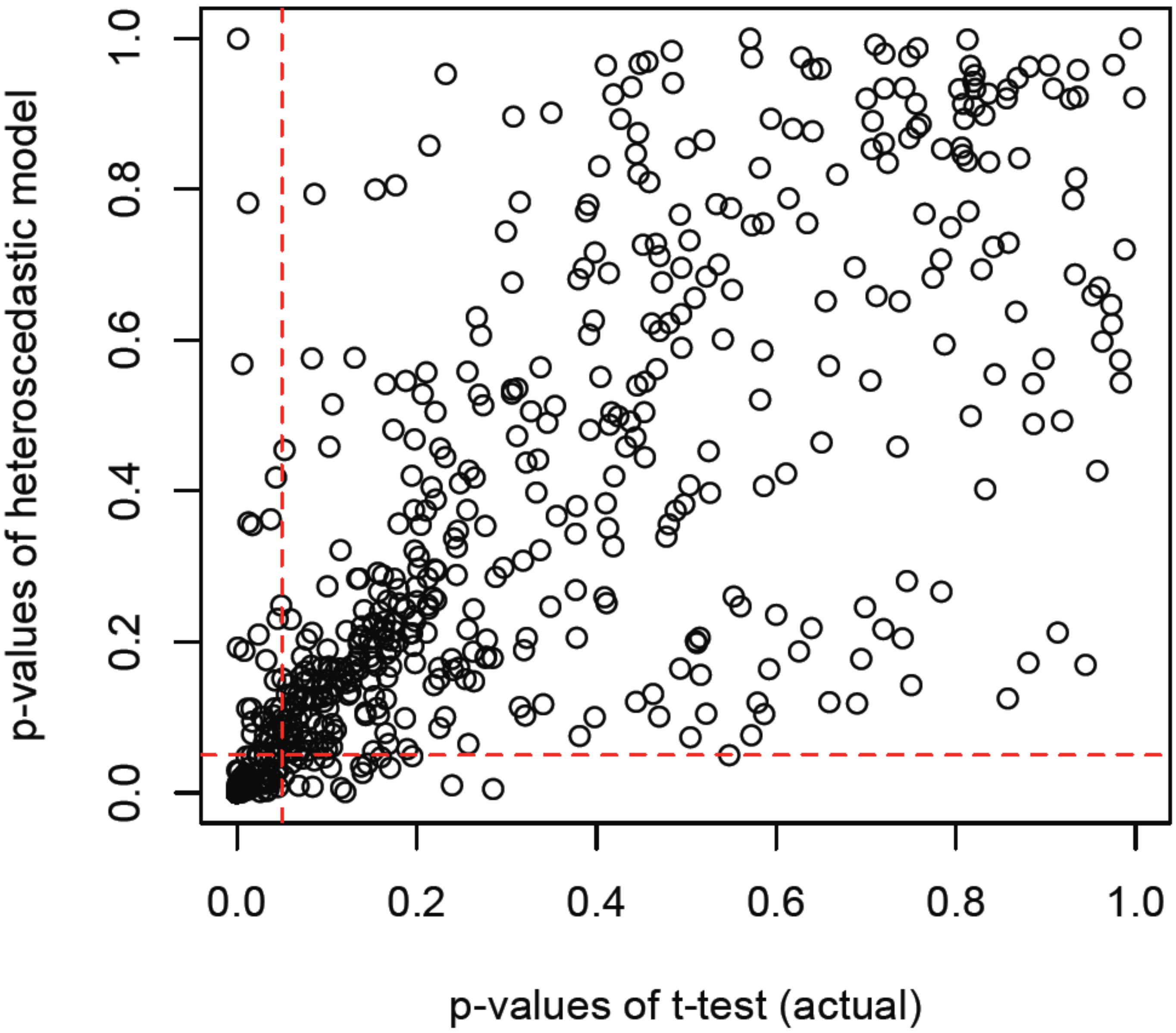

Mixed-effects Heteroscedastic Model
-4 -2 0 2
2 2 2 2 2 2 2 2 2 2
1 2 3 4 5 6 7 8 9 10
1 1 1 1 1 1 1 1 1 1
1 2 3 4 5 6 7 8 9 10
4
2
0
-2
-4
Standardized residuals
Fitted values

Mixed Effects Models

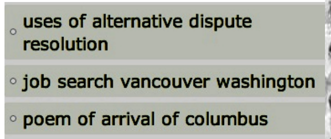
- In the statistical programming environment **R**

```
lme2 <- lme(effectiveness~system, data=df, random=~1|query/system,
     weights=varIdent(form=~1|query*system))
```

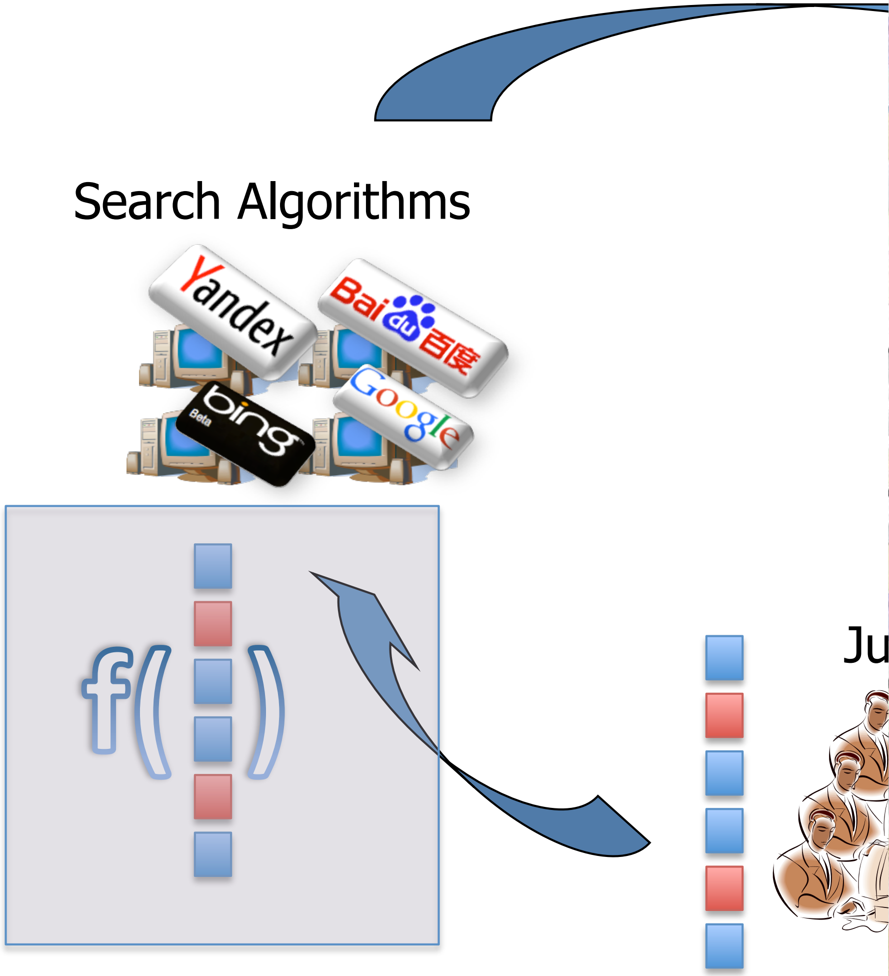
```
Random effects:
  Formula: ~1 | query            |         Formula: ~1 | system %in% query
         (Intercept)             |                     (Intercept)   Residual
StdDev:  1.447164                |         StdDev:     0.4537618     0.186183


Variance function:
  Structure: Different standard deviations per stratum
  Formula: ~1 | query* system
  Parameter estimates:
         1*1       1*2           2*1           2*2          ...
         1.0000000 1.6108387     1.3969085     1.5405710    ...

Fixed effects: y ~ system
                    Value     Std.Error          DF   t-value      p-value
  (Intercept)           -1.4385  0.22266286      846  -6.460817     0.0000
  system2           0.1834    0.09844907    46   1.863342      0.0688
```


p-values of heteroscedastic model
p-values of t-test (actual)
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0

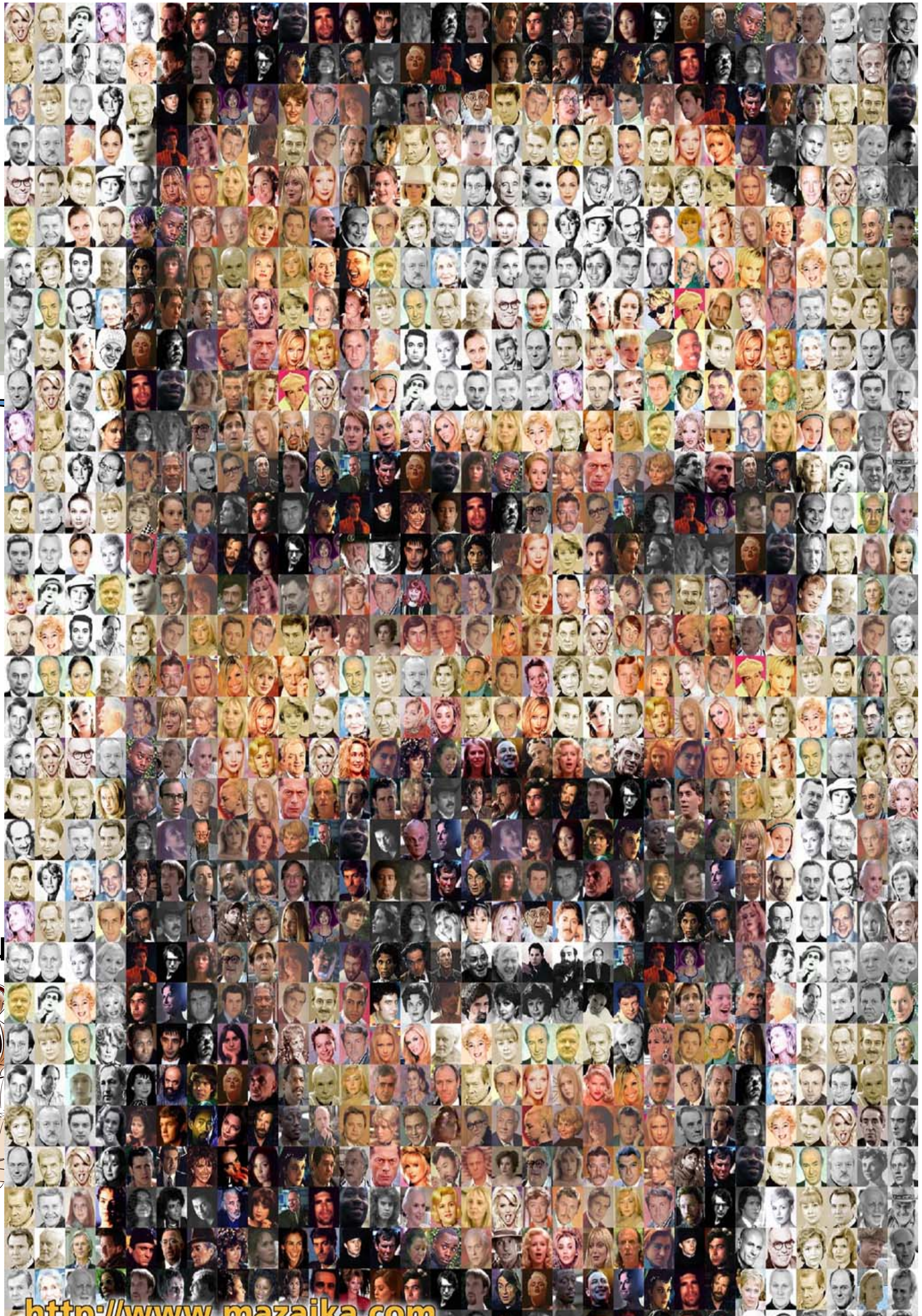
Variance due to Measure Parameters



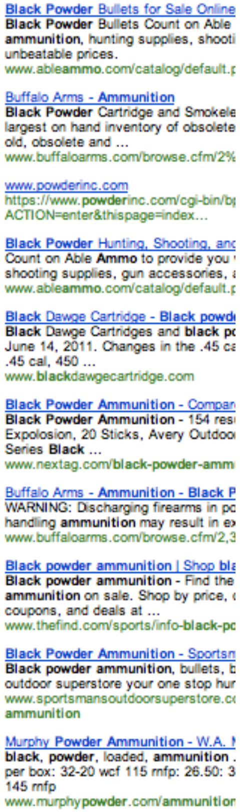
- uses of alternative dispute resolution
- job search vancouver washington
- poem of arrival of columbus



Search Algorithms



RBP – User Model



1

Black Powder Bullets for Sale Online
Black Powder Bullets Count on Able
ammunition, hunting supplies, shooti
unbeatable prices.
www.ableammo.com/catalog/default.p

2

Buffalo Arms - Ammunition
Black Powder Cartridge and Smokele
largest on hand inventory of obsolete
old, obsolete and ...
www.buffaloarms.com/browse.cfm/2%

3

www.powderinc.com
https://www.powderinc.com/cgi-bin/bp
ACTION=enter&thispage=index...

4

Black Powder Hunting, Shooting, anc
Count on Able Ammo to provide you
shooting supplies, gun accessories, a
www.ableammo.com/catalog/default.p

5

Black Dawge Cartridge - Black powde
Black Dawge Cartridges and black po
June 14, 2011. Changes in the .45 ca
.45 cal, 450 ...
www.blackdawgecartridge.com

6

Black Powder Ammunition - Compar
Black Powder Ammunition - 154 resu
Expolosion, 20 Sticks, Avery Outdoo
Series Black ...
www.nextag.com/black-powder-amm

7

Buffalo Arms - Ammunition - Black P
WARNING: Discharging firearms in po
handling ammunition may result in ex
www.buffaloarms.com/browse.cfm/2,3

8

Black powder ammunition | Shop bla
Black powder ammunition - Find the
ammunition on sale. Shop by price, c
coupons, and deals at ...
www.thefind.com/sports/info-black-po

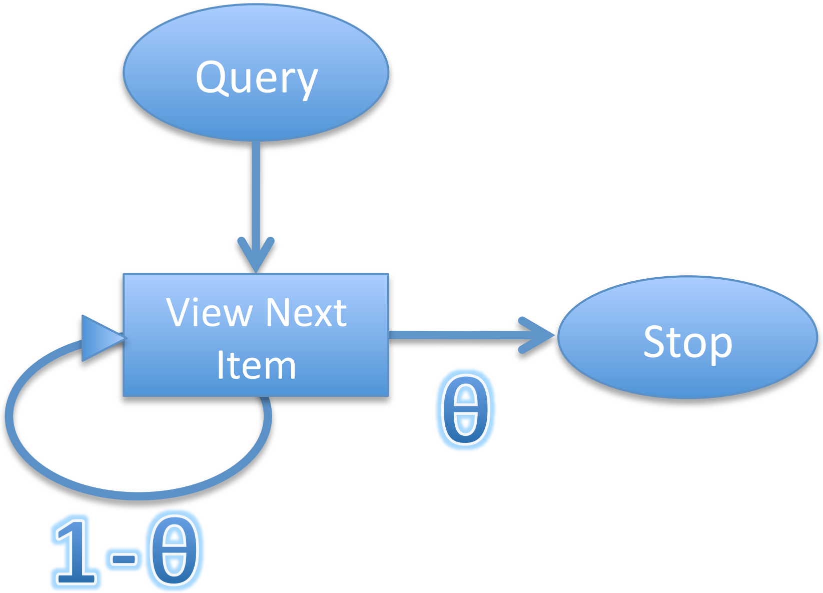
9

Black Powder Ammunition - Sportsm
Black powder ammunition, bullets, b
outdoor superstore your one stop hur
www.sportsmansoutdoorsuperstore.co
ammunition

10

Murphy Powder Ammunition - W.A. M
black, powder, loaded, ammunition .
per box: 32-20 wcf 115 mfp: 26.50: 3
145 mfp
www.murphypowder.com/ammunition

…



Query

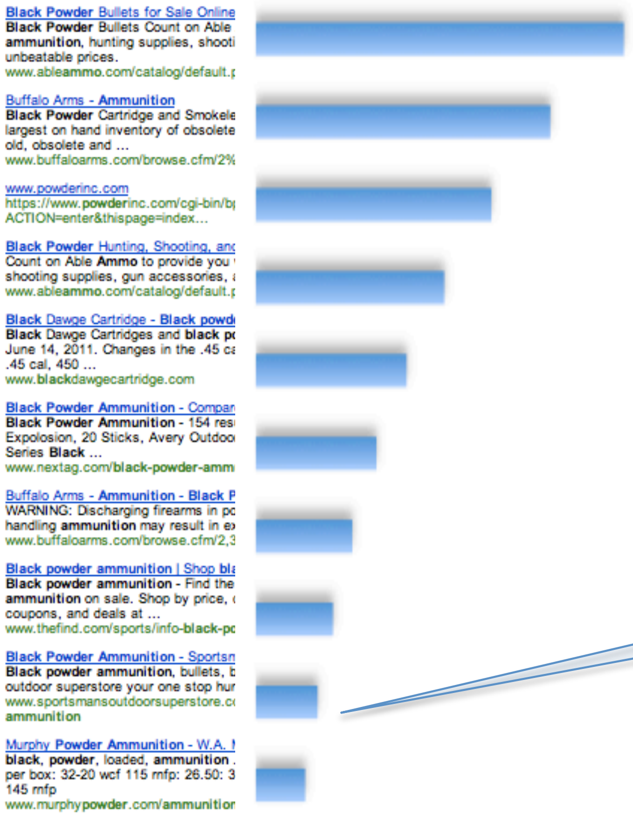
View Next Item

Stop

θ

1-θ

RBP – The Measure



1
Black Powder Bullets for Sale Online
Black Powder Bullets Count on Able
ammunition, hunting supplies, shooti
unbeatable prices.
www.ableammo.com/catalog/default.p

2
Buffalo Arms - Ammunition
Black Powder Cartridge and Smokele
largest on hand inventory of obsolete
old, obsolete and ...
www.buffaloarms.com/browse.cfm/2%

3
www.powderinc.com
https://www.powderinc.com/cgi-bin/bp
ACTION=enter&thispage=index...

4
Black Powder Hunting, Shooting, and
Count on Able Ammo to provide you
shooting supplies, gun accessories, a
www.ableammo.com/catalog/default.p

5
Black Dawge Cartridge - Black powd
Black Dawge Cartridges and black po
June 14, 2011. Changes in the .45 ca
.45 cal, 450 ...
www.blackdawgecartridge.com

6
Black Powder Ammunition - Compar
Black Powder Ammunition - 154 resu
Expolosion, 20 Sticks, Avery Outdoo
Series Black ...
www.nextag.com/black-powder-amm

7
Buffalo Arms - Ammunition - Black P
WARNING: Discharging firearms in po
handling ammunition may result in ex
www.buffaloarms.com/browse.cfm/2,3

8
Black powder ammunition | Shop bla
Black powder ammunition - Find the
ammunition on sale. Shop by price, c
coupons, and deals at ...
www.thefind.com/sports/info-black-po

9
Black Powder Ammunition - Sportsm
Black powder ammunition, bullets, b
outdoor superstore your one stop hun
www.sportsmansoutdoorsuperstore.co
ammunition

10
Murphy Powder Ammunition - W.A. M
black, powder, loaded, ammunition .
per box: 32-20 wcf 115 mfp: 26.50: 3
145 mfp
www.murphypowder.com/ammunition

...

$$\mathrm{RBP} = \sum_{i=1}^{n} rel_i (1-\vartheta)^{i-1}\vartheta$$

Relevance discounted by geometric distribution

Choosing Parameter Values

- Different approaches:
  - Predefine parameters
  - Use click log; fit a model to gaps between clicks (Zhang et al., IRJ, 2010)
  - Minimize variance in evaluation (Kanoulas & Aslam, CIKM ’09)

- All user models have parameters
  - Metrics evaluated at fixed parameter values
  - Evaluation w.r.t. an average user

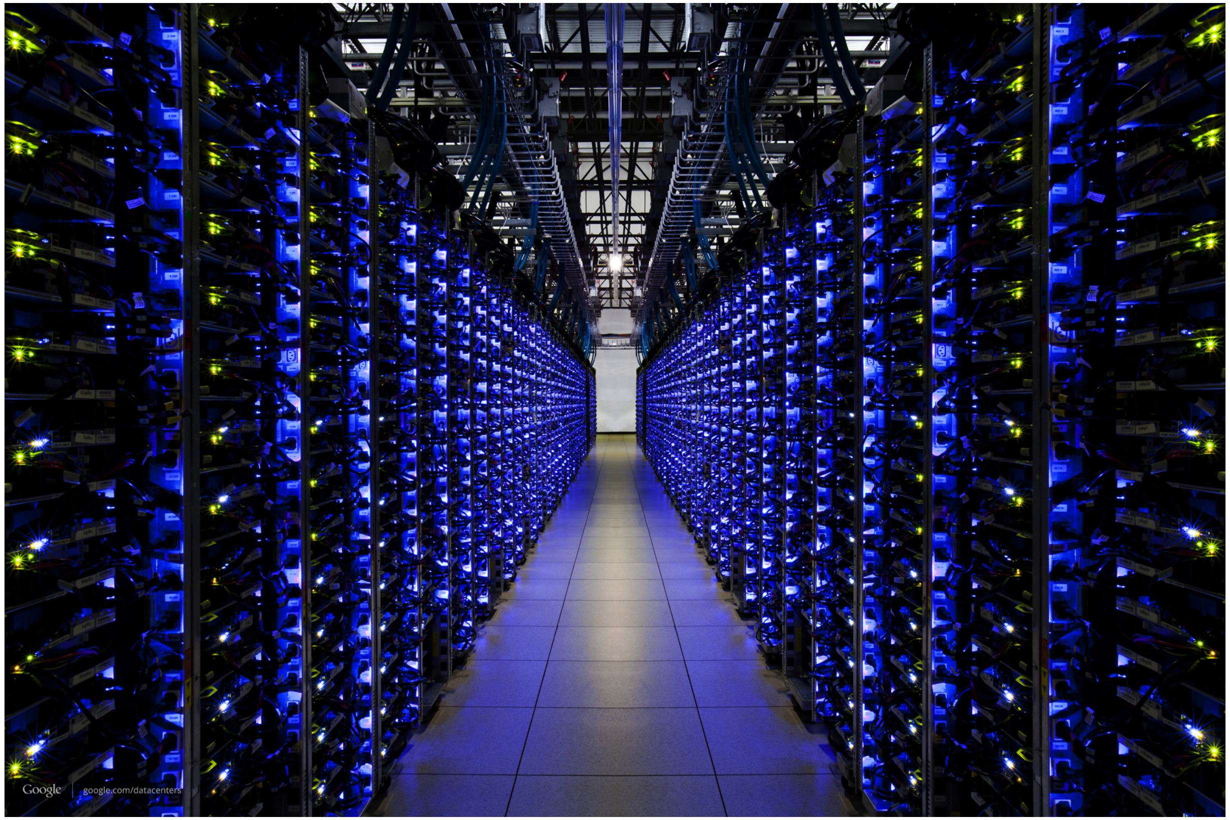
Choosing Parameter Values

- Users behave very differently when they search
  - Distribution of parameters (users) need to be considered

- A different approach
  - Mine Web Query logs
  - Learn a distribution of the parameters
  - Use this distribution to evaluate the quality of systems

Patience Distribution for RBP

- Goal: produce a posterior distribution for θ
- Start with a uniform distribution for θ
- Update it based on logged data



Posterior Distribution of Patience θ for RBP

$$P(\theta | E) = P(\theta | c) = \sum_{r=0}^{\infty} P(\theta | r, c) P(r | c)$$

The probability that user skips r document

$$P(\theta | r, c) \propto P(c | \theta, r) P(\theta | r$$

$$P(c | \theta, r) = NB(r, \theta)$$

$$P(\theta | r) = Beta(\alpha, \beta)$$

Probability distribution of the number of successes in a sequence of Bernoulli trials before **r** failures occurs.

- Start with uniform prior (α=β=1)

Posterior Distribution of Patience θ for RBP

$$P(\theta \mid r, E) \propto \boxed{P(E \mid \theta, r)}\boxed{P(\theta \mid r)}$$

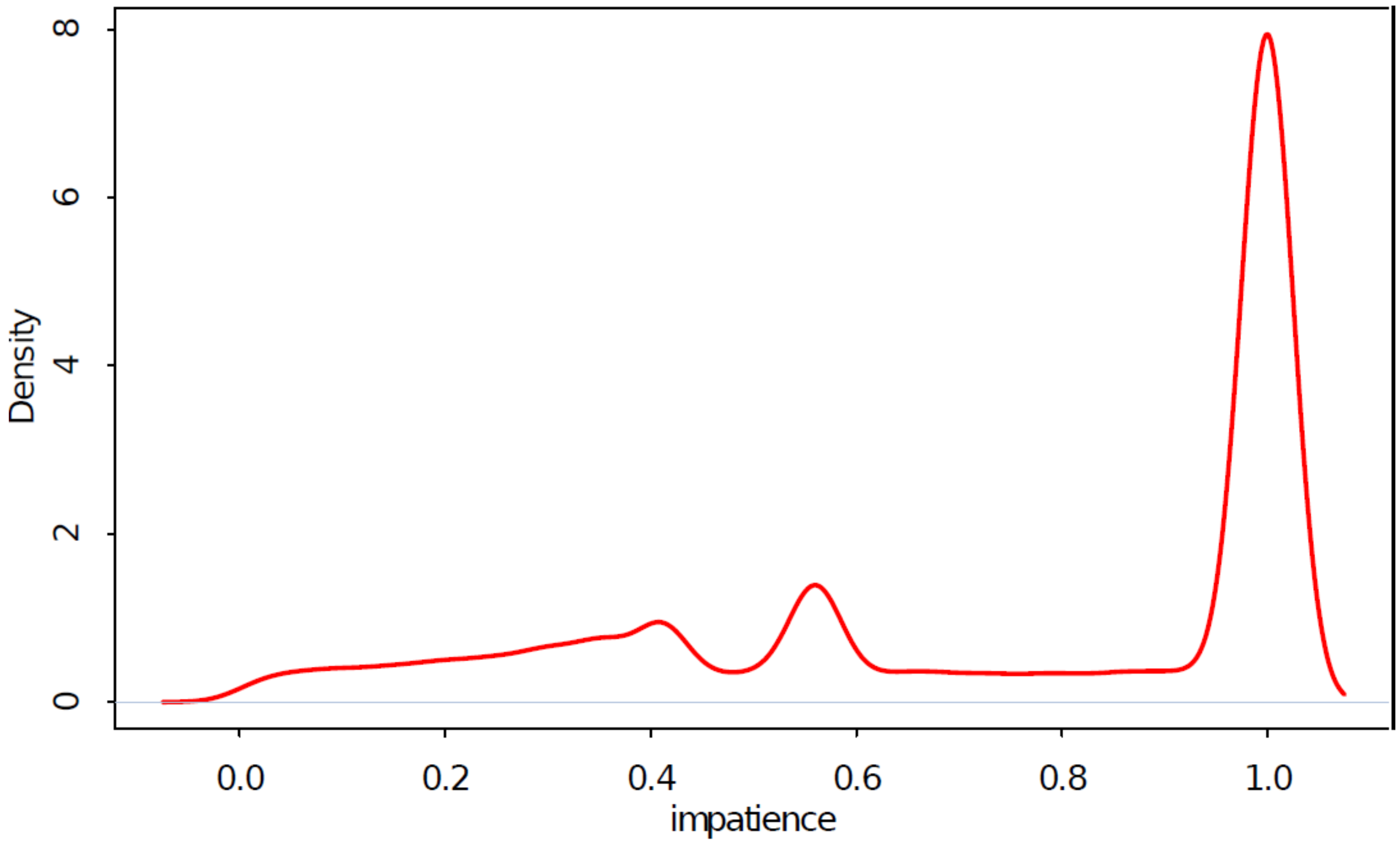
$$P(E \mid \theta, r) = NB(r, \theta) \qquad P(\theta \mid r) = Beta(\alpha, \beta)$$

- If there are m queries, with r number of failures, and c_i number of successes, i=1..m

$$P(\theta \mid r, E) = Beta(\alpha + \sum_{i=1}^{m} c_i, \beta + mr)$$

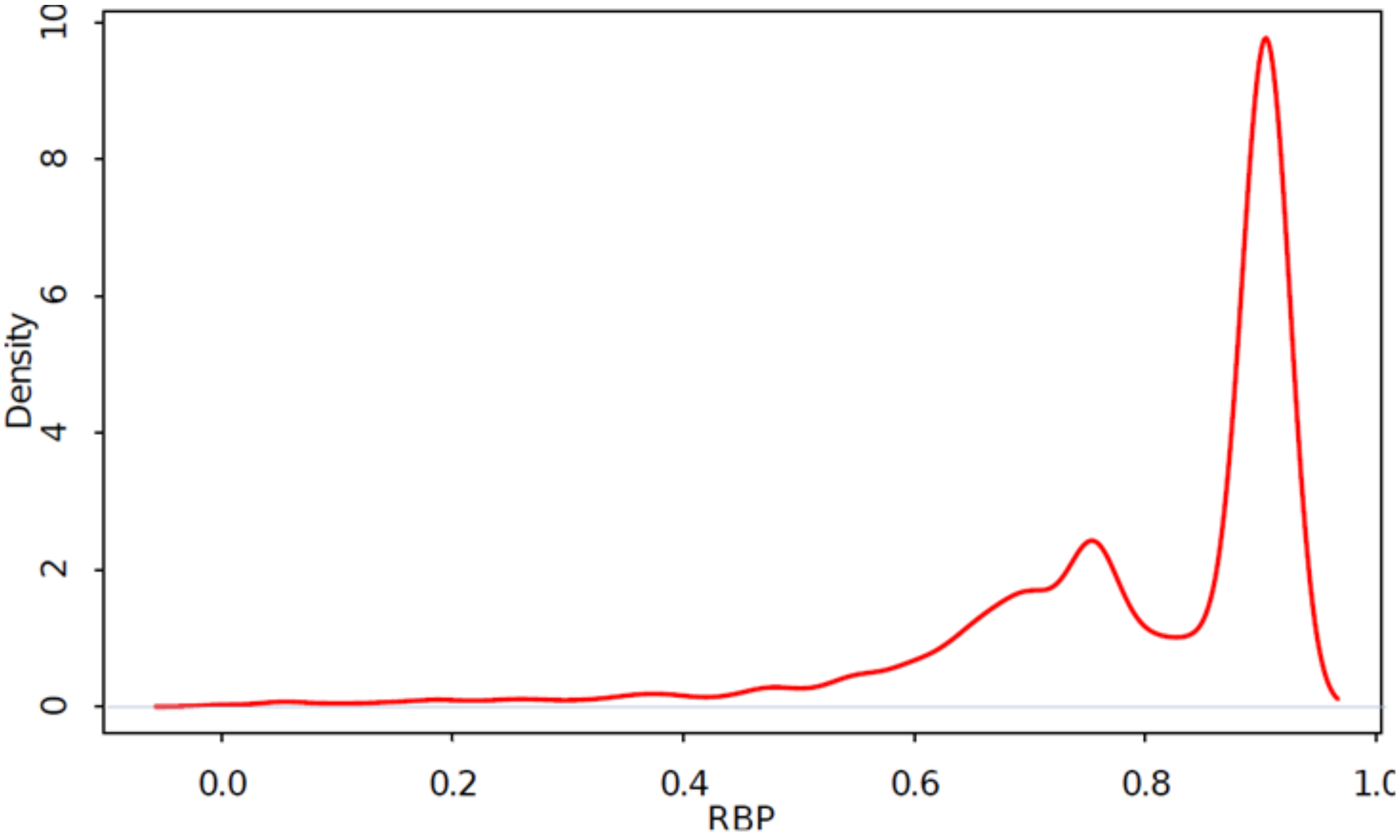
Posterior Distribution for Impatience: RBP

- Distribution of users using the AOL log



Distribution of RBP

- RBP values for different users for a single system and a single query



Mixed Effects Models

$$y_{ijk} = \alpha_i + (\beta_j + \phi_j p_k) + (\kappa_{ij} + \gamma_{ij} p_k) + \epsilon_{ijk}$$

y_{ijk}: value of a metric on topic j for system i with parameter p_k

α_i : effect of system i

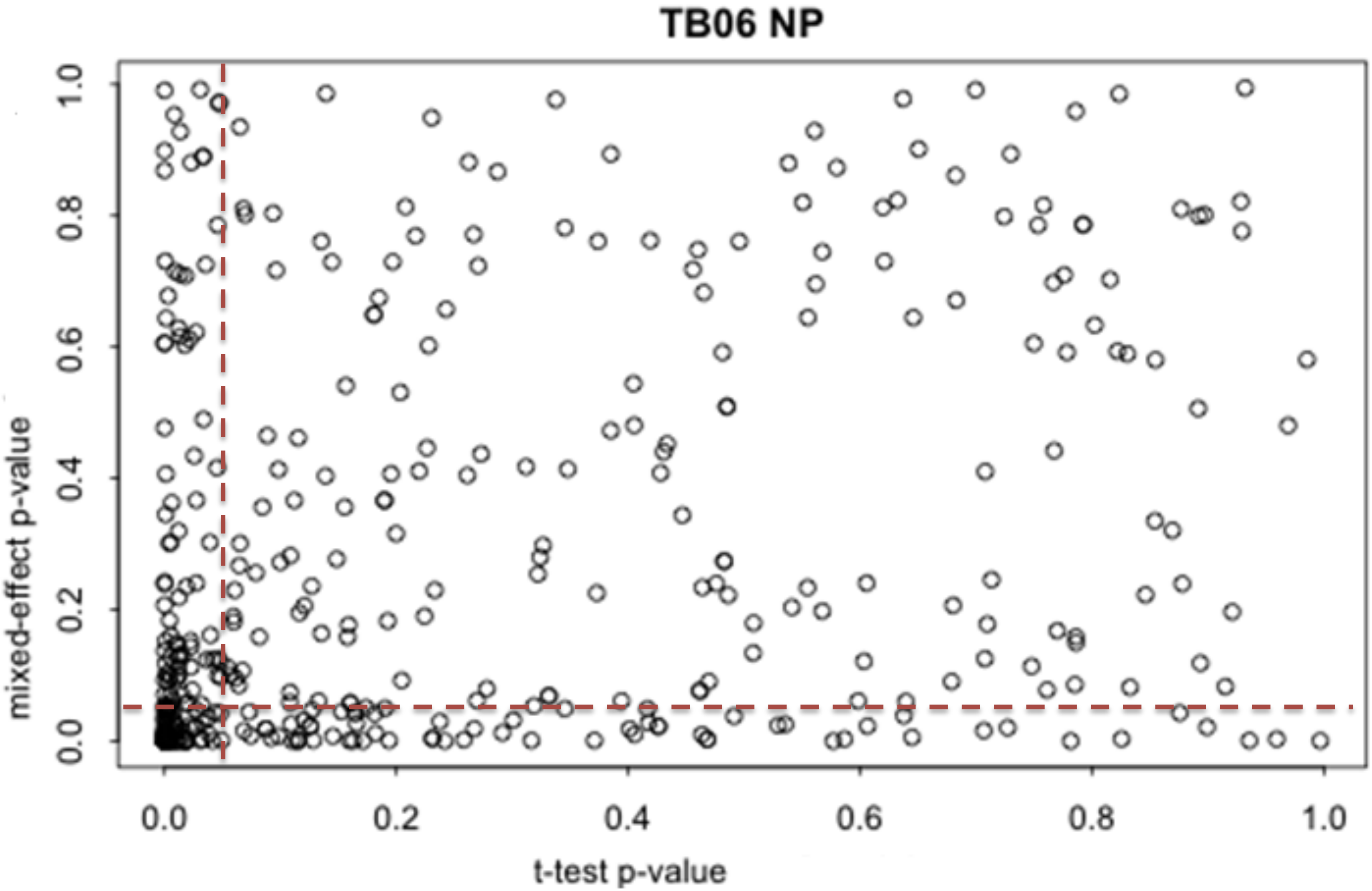
β_j : effect of topic j

$\varphi_j p_k$: interaction of topic with RBP parameter

κ_{ij} : system/topic interaction effect

$\gamma_\mathrm{ij} p_k$: interaction of system/topic with RBP parameter

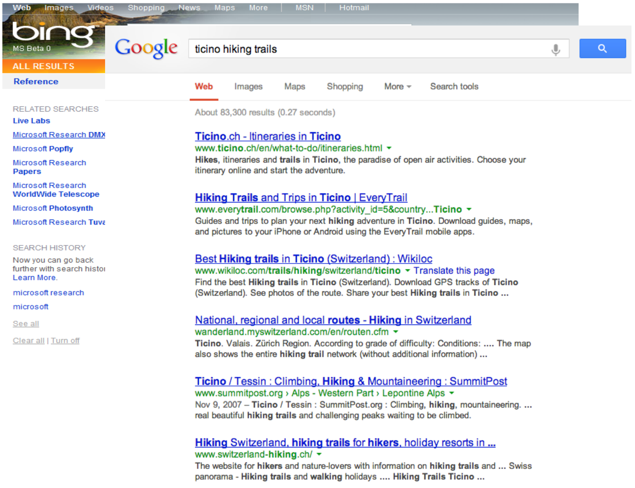
ε_ijk : system/topic/parameter interaction effect


TB06 NP
mixed-effect p-value
0.0
0.2
0.4
0.6
0.8
1.0
t-test p-value
0.0
0.2
0.4
0.6
0.8
1.0

Variance due to Query Intents

- uses of alternative dispute resolution
- job search vancouver washington
- poem of arrival of columbus

Results



Web Images Videos Shopping News Maps More MSN Hotmail
bing
MS Beta 0
ALL RESULTS
Reference
Google
ticino hiking trails
Web Images Maps Shopping More Search tools

RELATED SEARCHES
Live Labs
Microsoft Research DMX
Microsoft Popfly
Microsoft Research Papers
Microsoft Research WorldWide Telescope
Microsoft Photosynth
Microsoft Research Tuva

SEARCH HISTORY
Now you can go back further with search histo Learn More.
microsoft research
microsoft
See all
Clear all | Turn off

About 83,300 results (0.27 seconds)

Ticino.ch - Itineraries in Ticino
www.ticino.ch/en/what-to-do/itineraries.html
Hikes, itineraries and **trails** in **Ticino**, the paradise of open air activities. Choose your itinerary online and start the adventure.

Hiking Trails and Trips in Ticino | EveryTrail
www.everytrail.com/browse.php?activity_id=5&country...Ticino
Guides and trips to plan your next **hiking** adventure in **Ticino**. Download guides, maps, and pictures to your iPhone or Android using the EveryTrail mobile apps.

Best Hiking trails in Ticino (Switzerland) : Wikiloc
www.wikiloc.com/trails/hiking/switzerland/ticino Translate this page
Find the best **Hiking trails** in **Ticino** (Switzerland). Download GPS tracks of **Ticino** (Switzerland). See photos of the route. Share your best **Hiking trails** in **Ticino** ...

National, regional and local routes - Hiking in Switzerland
wanderland.myswitzerland.com/en/routen.cfm
Ticino. Valais. Zürich Region. According to grade of difficulty: Conditions: .... The map also shows the entire **hiking trail** network (without additional information) ...

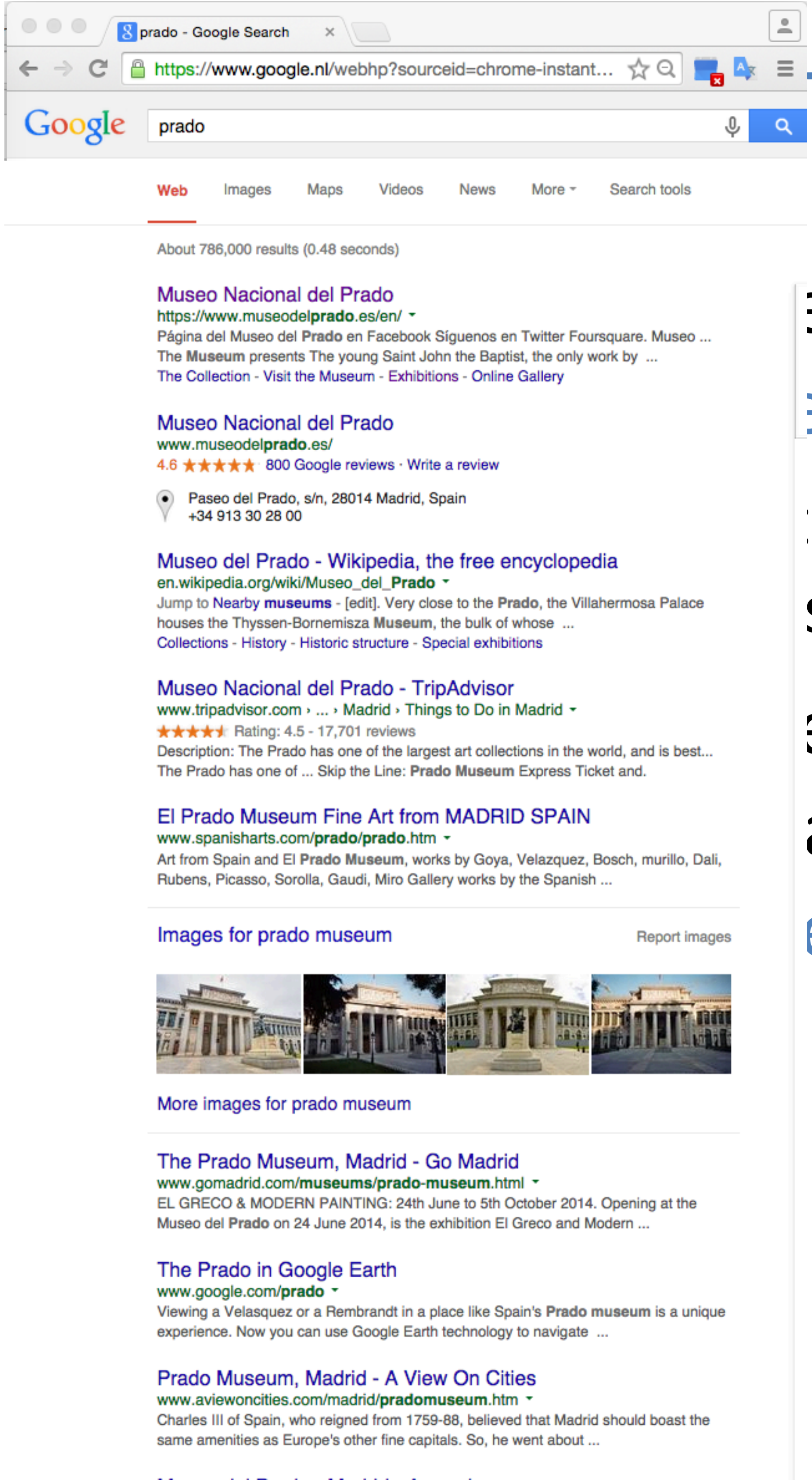
Ticino / Tessin : Climbing, Hiking & Mountaineering : SummitPost
www.summitpost.org › Alps - Western Part › Lepontine Alps
Nov 9, 2007 – **Ticino** / Tessin : SummitPost.org : Climbing, **hiking**, mountaineering. ... real beautiful **hiking trails** and challenging peaks waiting to be climbed.

Hiking Switzerland, hiking trails for hikers, holiday resorts in ...
www.switzerland-hiking.ch/
The website for **hikers** and nature-lovers with information on **hiking trails** and ... Swiss panorama - **Hiking trails** and **walking** holidays .... **Hiking Trails Ticino** ...

```
<topic number="19" type="ambiguous">
    <query>the current</query>
    <description>
        I'm looking for the homepage of The Current, a program
        on Minnesota Public Radio.
    </description>
    <subtopic number="1" type="nav">
        Take me to the homepage of The Current, a program on Minnesota
        Public Radio.
    </subtopic>
    <subtopic number="2" type="nav">
        I'm looking for the homepage of The Current newspaper in New Jersey.
    </subtopic>
    <subtopic number="3" type="nav">
        I want to find the homepage of The Current newspaper in Hartford.
    </subtopic>
    <subtopic number="4" type="nav">
        I want to find the homepage of The Current magazine in San Antonio.
    </subtopic>
</topic>
```

QUERY: job search vancouver washington

DESCRIPTION: I would like to find web page that aggregate job opportunities in the IT industry in Vancouver, Washington.



About 786,000 results (0.48 seconds)

Museo Nacional del Prado
https://www.museodelprado.es/en/
Página del Museo del **Prado** en Facebook Síguenos en Twitter Foursquare. Museo ... The **Museum** presents The young Saint John the Baptist, the only work by ...
The Collection - Visit the Museum - Exhibitions - Online Gallery

Museo Nacional del Prado
www.museodelprado.es/
4.6 ★★★★★ 800 Google reviews · Write a review
Paseo del Prado, s/n, 28014 Madrid, Spain
+34 913 30 28 00

Museo del Prado - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Museo_del_**Prado**
Jump to Nearby **museums** - [edit]. Very close to the **Prado**, the Villahermosa Palace houses the Thyssen-Bornemisza **Museum**, the bulk of whose ...
Collections - History - Historic structure - Special exhibitions

Museo Nacional del Prado - TripAdvisor
www.tripadvisor.com › ... › Madrid › Things to Do in Madrid
★★★★★ Rating: 4.5 - 17,701 reviews
Description: The Prado has one of the largest art collections in the world, and is best... The Prado has one of ... Skip the Line: **Prado Museum** Express Ticket and.

El Prado Museum Fine Art from MADRID SPAIN
www.spanisharts.com/**prado**/**prado**.htm
Art from Spain and El **Prado Museum**, works by Goya, Velazquez, Bosch, murillo, Dali, Rubens, Picasso, Sorolla, Gaudi, Miro Gallery works by the Spanish ...

Images for prado museum
Report images
More images for prado museum

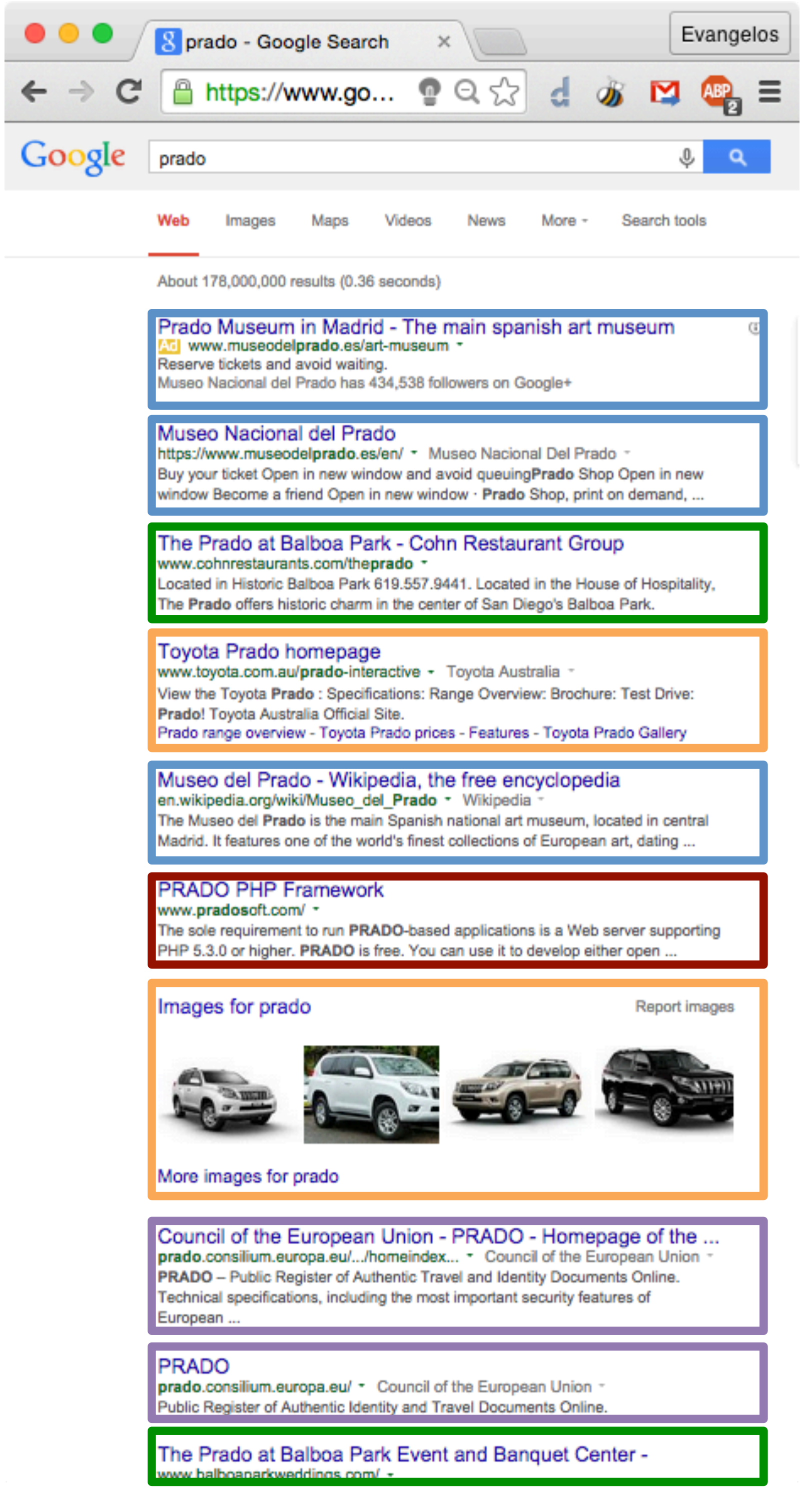
The Prado Museum, Madrid - Go Madrid
www.gomadrid.com/**museums**/**prado**-**museum**.html
EL GRECO & MODERN PAINTING: 24th June to 5th October 2014. Opening at the Museo del **Prado** on 24 June 2014, is the exhibition El Greco and Modern ...

The Prado in Google Earth
www.google.com/**prado**
Viewing a Velasquez or a Rembrandt in a place like Spain's **Prado museum** is a unique experience. Now you can use Google Earth technology to navigate ...

Prado Museum, Madrid - A View On Cities
www.aviewoncities.com/madrid/**pradomuseum**.htm
Charles III of Spain, who reigned from 1759-88, believed that Madrid should boast the same amenities as Europe's other fine capitals. So, he went about ...

ty?

ently
eted
em can
ser's intent
empts to
at may be
e of



About 178,000,000 results (0.36 seconds)

Prado Museum in Madrid - The main spanish art museum
Ad www.museodelprado.es/art-museum
Reserve tickets and avoid waiting.
Museo Nacional del Prado has 434,538 followers on Google+

Museo Nacional del Prado
https://www.museodel**prado**.es/en/ · Museo Nacional Del Prado
Buy your ticket Open in new window and avoid queuing**Prado** Shop Open in new window Become a friend Open in new window · **Prado** Shop, print on demand, ...

The Prado at Balboa Park - Cohn Restaurant Group
www.cohnrestaurants.com/the**prado**
Located in Historic Balboa Park 619.557.9441. Located in the House of Hospitality, The **Prado** offers historic charm in the center of San Diego's Balboa Park.

Toyota Prado homepage
www.toyota.com.au/**prado**-interactive · Toyota Australia
View the Toyota **Prado** : Specifications: Range Overview: Brochure: Test Drive: **Prado**! Toyota Australia Official Site.
Prado range overview - Toyota Prado prices - Features - Toyota Prado Gallery

Museo del Prado - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Museo_del_**Prado** · Wikipedia
The Museo del **Prado** is the main Spanish national art museum, located in central Madrid. It features one of the world's finest collections of European art, dating ...

PRADO PHP Framework
www.**prado**soft.com/
The sole requirement to run **PRADO**-based applications is a Web server supporting PHP 5.3.0 or higher. **PRADO** is free. You can use it to develop either open ...

Images for prado
Report images
More images for prado

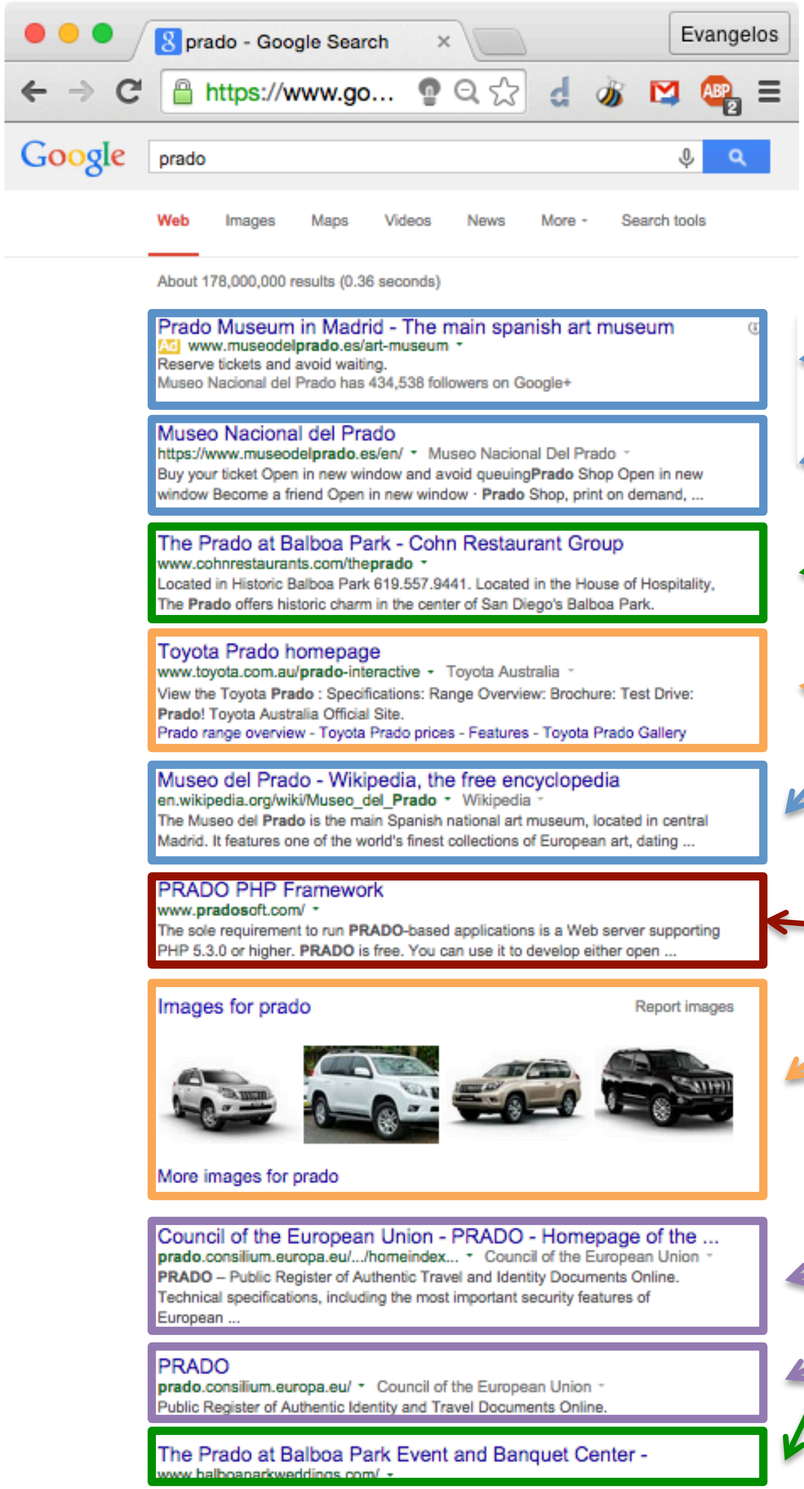
Council of the European Union - PRADO - Homepage of the ...
prado.consilium.europa.eu/.../homeindex... · Council of the European Union
PRADO – Public Register of Authentic Travel and Identity Documents Online. Technical specifications, including the most important security features of European ...

PRADO
prado.consilium.europa.eu/ · Council of the European Union
Public Register of Authentic Identity and Travel Documents Online.

The Prado at Balboa Park Event and Banquet Center -
www.balboaparkweddings.com/

Intent-Aware Measures

- Assume there is a probability distribution *P(i|Q)* over intents for a query Q
  - Probability that a randomly-sampled user means intent i when submitting query Q
- The intent-aware version of a measure is its weighted average over this distribution



prado

About 178,000,000 results (0.36 seconds)

Prado Museum in Madrid - The main spanish art museum
Ad www.museodelprado.es/art-museum
Reserve tickets and avoid waiting.
Museo Nacional del Prado has 434,538 followers on Google+

Museo Nacional del Prado
https://www.museodelprado.es/en/ - Museo Nacional Del Prado
Buy your ticket Open in new window and avoid queuingPrado Shop Open in new window Become a friend Open in new window · Prado Shop, print on demand, ...

The Prado at Balboa Park - Cohn Restaurant Group
www.cohnrestaurants.com/theprado
Located in Historic Balboa Park 619.557.9441. Located in the House of Hospitality, The Prado offers historic charm in the center of San Diego's Balboa Park.

Toyota Prado homepage
www.toyota.com.au/prado-interactive - Toyota Australia
View the Toyota Prado : Specifications: Range Overview: Brochure: Test Drive: Prado! Toyota Australia Official Site.
Prado range overview - Toyota Prado prices - Features - Toyota Prado Gallery

Museo del Prado - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Museo_del_Prado - Wikipedia
The Museo del Prado is the main Spanish national art museum, located in central Madrid. It features one of the world's finest collections of European art, dating ...

PRADO PHP Framework
www.pradosoft.com/
The sole requirement to run PRADO-based applications is a Web server supporting PHP 5.3.0 or higher. PRADO is free. You can use it to develop either open ...

Images for prado
Report images
More images for prado

Council of the European Union - PRADO - Homepage of the ...
prado.consilium.europa.eu/.../homeindex... - Council of the European Union
PRADO – Public Register of Authentic Travel and Identity Documents Online. Technical specifications, including the most important security features of European ...

PRADO
prado.consilium.europa.eu/ - Council of the European Union
Public Register of Authentic Identity and Travel Documents Online.

The Prado at Balboa Park Event and Banquet Center -
www.balboaparkweddings.com/

P(Prado Museum | Q) = 0.35

P(Prado Balboa | Q) = 0.10

P(Toyota Prado | Q) = 0.45

P(Prado PHP | Q) = 0.08

P(Prado EU | Q) = 0.02

Precision@10-IA = 0.35*0.3 + 0.10*0.2
+ 0.45*0.2 + 0.08*0.1 + 0.02*0.2 = 0.227

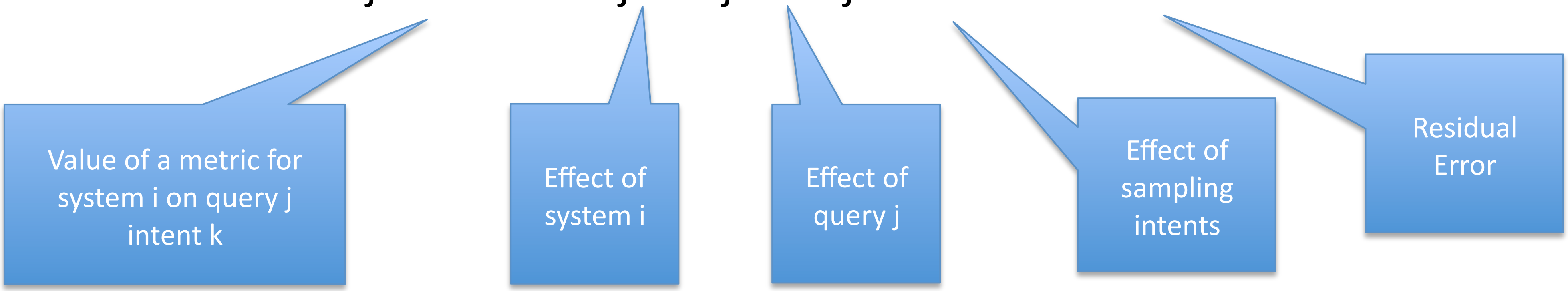
Variance due to Query Intents

- The intents are not fixed
  - may think of them as a sample
    - from some large/infinite universe of possible intents
  - now we have two simultaneous sampling processes
    - what does this (sample x sample) tell us about the (population x population)?

Mixed Effects Models

- Two sources of variance
  - Queries
  - Intents (within queries)

$$y_{ijk} = \beta_i + b_j + c_{ij} + \varepsilon_{ijk}$$



Value of a metric for system i on query j intent k

Effect of system i

Effect of query j

Effect of sampling intents

Residual Error

$b_j \sim N(0,\sigma_1^2)$, $c_{ij} \sim N(0,\sigma_2^2)$, $\varepsilon_{ijk} \sim N(0,\sigma^2)$

Experimental Design

- Experimental setup
  - How many queries vs. how many intents per query?

TREC 2010	Query Effect	Intent Effect
IA-MAP	0.0478	0.0312
IA-ERR	0.1429	0.0650

TREC 2011	Query Effect	Intent Effect
IA-MAP	0.0707	0.0607
IA-ERR	0.2058	0.0973

Conclusions

- Choose your measure carefully
- Choose your experimental setup carefully
  - Put your money where most of the variance comes from
  - ⇒ Increase statistical power
- Always do significance tests
  - Model all the effects
  - Check your assumptions
- Always take results of tests with a grain of salt
  - Especially when the effect size is low