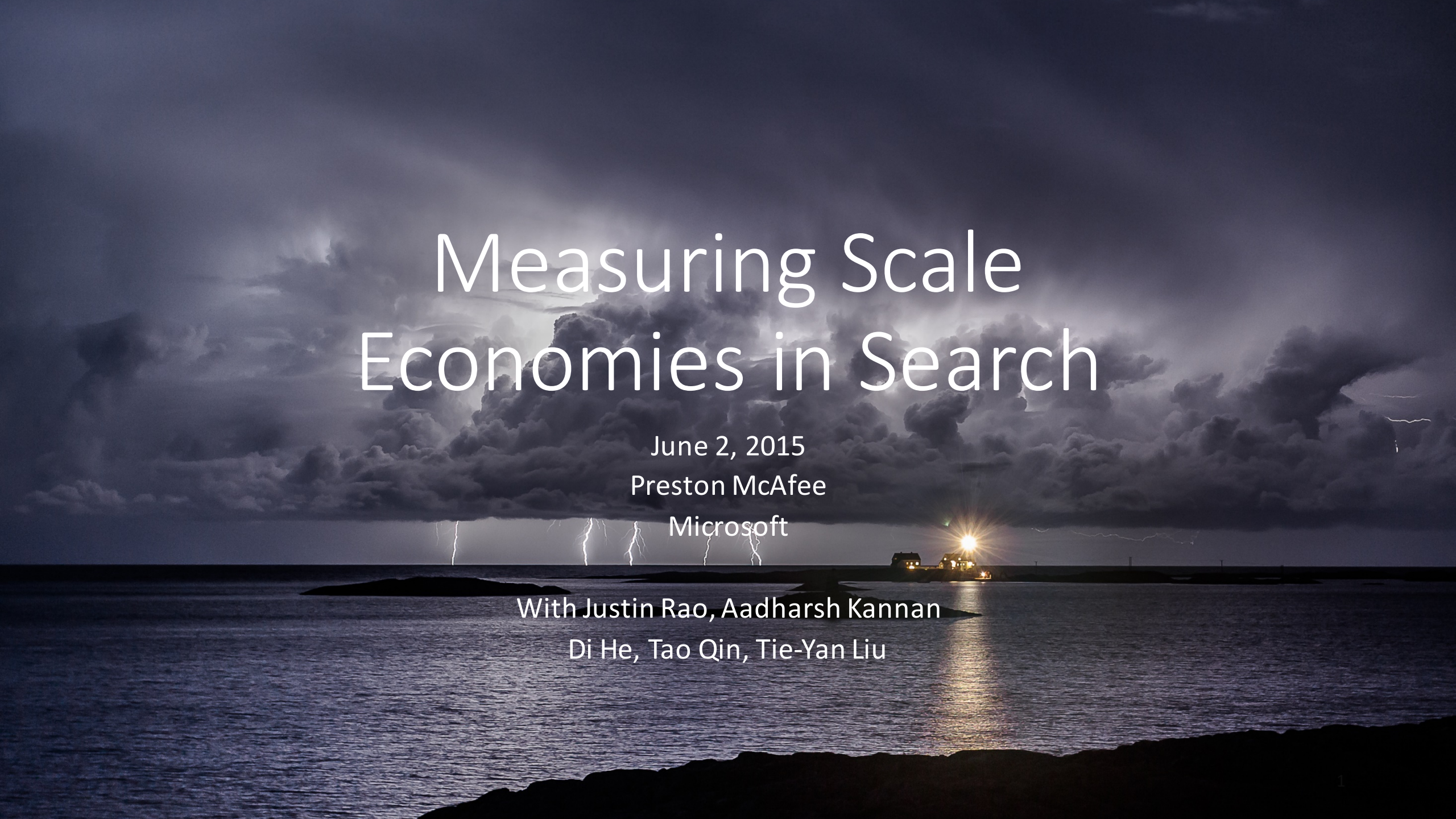


Measuring Scale Economies in Search

June 2, 2015

Preston McAfee
Microsoft

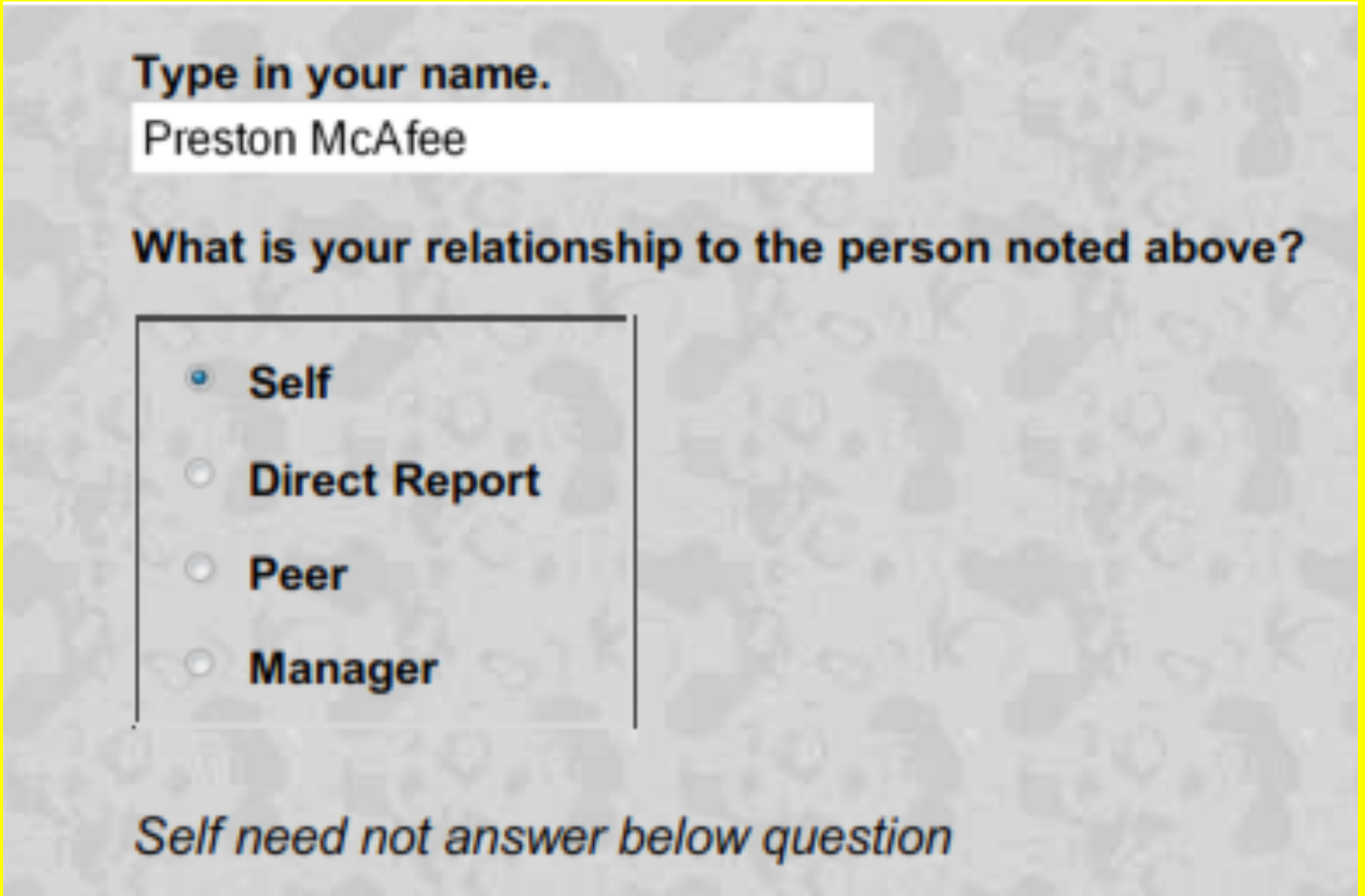
With Justin Rao, Aadharsh Kannan
Di He, Tao Qin, Tie-Yan Liu

Email From the Texas Higher Education Coordinating Board:

Your last name is your password.

If you have any questions or have forgotten your password, please contact the Coordinating Board.

Self Assessment



Value of Data in Algorithmic Search

- Search engines do not answer queries (mostly)
- Search engines are enormous matching programs, **matching** billions of URLs to hundreds of billions of queries
- Process uses sophisticated algorithms, probably the most complicated algorithms ever built
- Data is used to initialize and update algorithms
- How important is data at modern web scale?
- With hundreds of billions of observations, does a doubling or even a 20X increase actually matter?

How Large are the Scale Economies in Search?

- Statistically speaking, a trillion observations, a billion right hand side variables is still a trillion degrees of freedom
- Most queries are rare
  - 50% of Bing queries unique in 2014, 8% of searches
- But rare queries have related queries
  - Pasadena Ethiopian Restaurant related to Pasadena Restaurant, Ethiopian Restaurant
- $1/\sqrt{n}$ errors, where *n* is the amount of data – but what is *n*?
- Hasn't anyone measured this before?
  - Need billions of observations

Analytic Problems to Overcome

- Data used indirectly, makes whole search engine better
  - Not directly visible in results data
- Common queries are often easier
  - Common queries have less ambiguity
  - Common queries more likely to be navigational rather than informational

Overview and Summary

- Rare query trend analysis. As data on rare queries comes in, the quality of search page results improves, for Bing and Google.
  - Because of Google's greater scale, it improves faster.
- Direct and indirect view count analysis. Data on related (indirect) queries can help improve the quality of response to new queries. Indirect data has similar effect to direct; but many queries have little indirect data.
  - Google's greater scale means that it acquires data on new queries more quickly and that it has more data on indirect queries on which it can draw.
- Click position analysis. Quality of ranking improves with more data.
  - Not just improvements in website quality

Rare Query Trend analysis

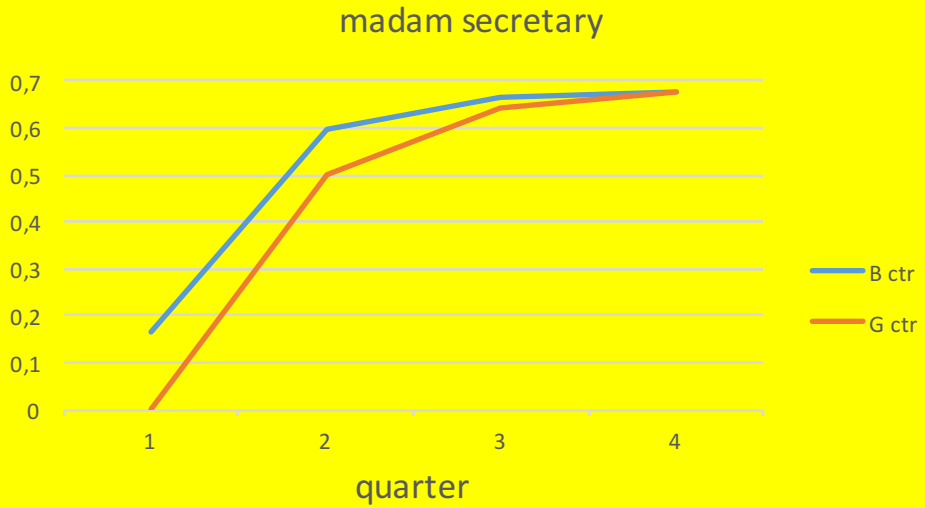
- Even if scale does not have a significant effect on the quality of responses to common queries, additional data on rare queries may improve the quality of algo search results.
- If so, then a search engine with greater scale will learn faster.
- Conservative: should understate data scale effects
  - Missing indirect data and external effects

Experimental Design #1: Rare Query Trend

- For a search engine, consider its IE logs in 2014
- Use 1/2014 to 3/2014 as benchmark data
- Use 4/2014 to 12/2014 as predicting data
  - #Clicks to define historical clicks
  - Use sliding windows in the next period to compute future CTR
- Definition of rare queries
  - #Clicks<200 in benchmark period
  - 1000<=#Clicks<2000 in prediction period
- With this setting, for the buckets [100, 900], we have exactly the same set of queries
  - Start getting different mixes as number goes up.

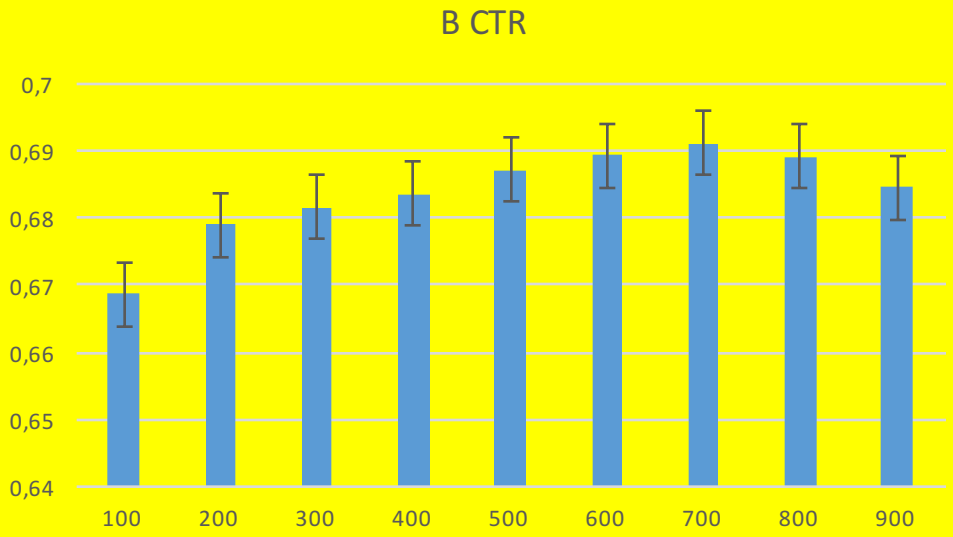
Examples of the Rare Query Trend

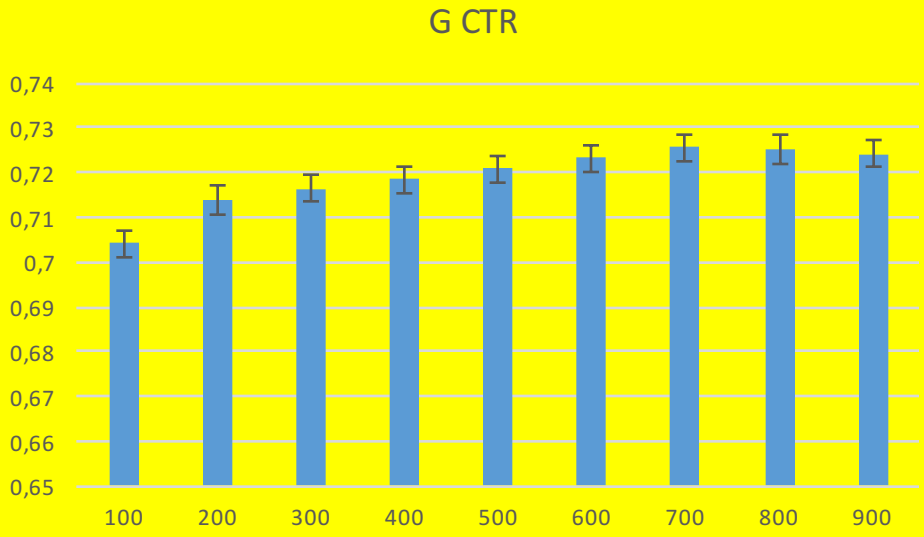
- Madam Secretary – a CBS show
- Letter Garden – an online game



CTR v.s Historical Data Volume: US

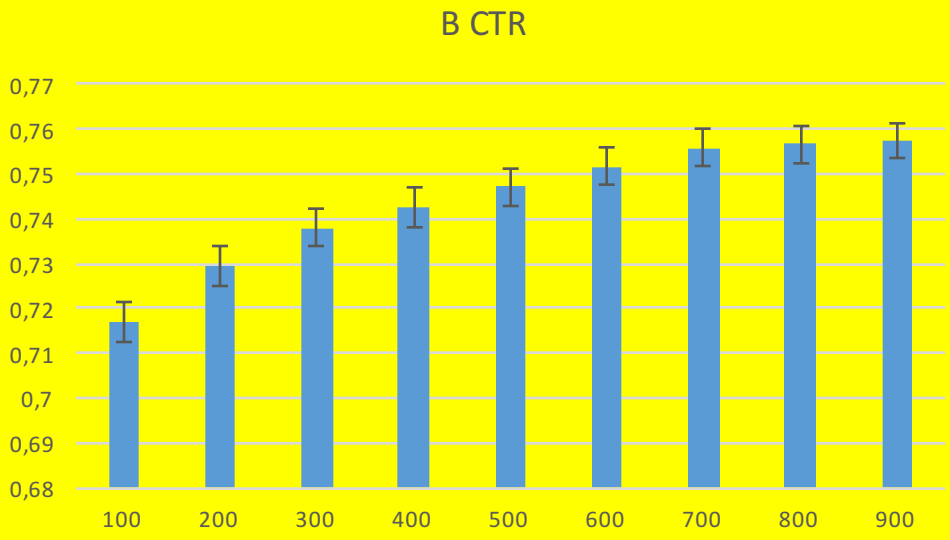
- Overall, we observe CTR growth for rare queries for both Bing and Google
- Non-monotonicities not statistically significant
- Holds query mix constant
- Levels not comparable between Google and Bing
  - Only observe portion of Google data but all of Bing
  - Different queries for Bing and Google (meetings the 200/1000 requirement)

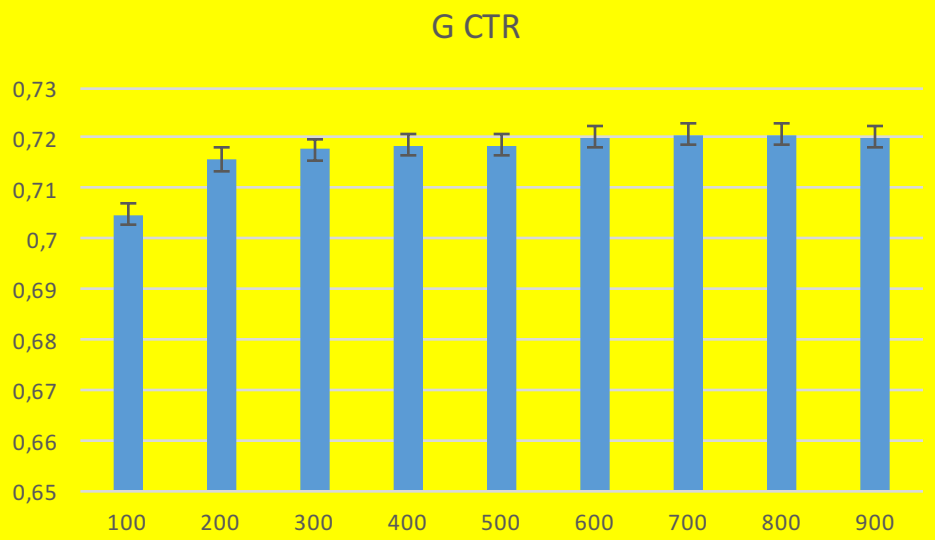




CTR v.s Historical Data Volume: EU

- Aggregated results
- Overall, we observe CTR growth for rare queries for both Bing and Google





How Much Data Do We Have?

- >50% of queries are unique in the year
  - That does not mean there is no data!
  - Historical data on “Pasadena restaurant” useful for “Pasadena Ethiopian restaurant”
  - Learn authoritative sites from past queries and then do text matching
- How many queries have “little” relevant data?
- This is a second, independent approach to assessing the importance of data

Direct and indirect view count analysis

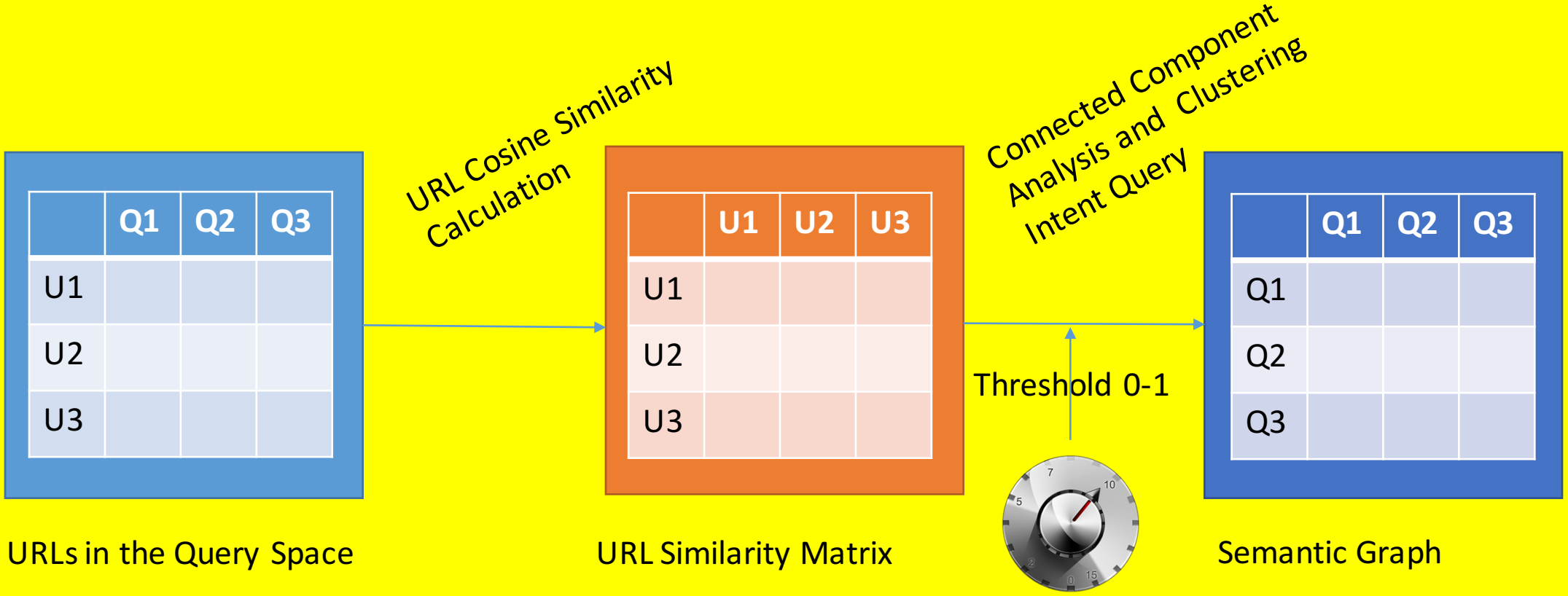
- If data on related (indirect) queries are useful in responding to new queries, then a search engine with greater scale will acquire data on new queries faster and it will also have more indirect data that it can use to improve the quality of responses to new queries.
- We can assess this empirically with Bing data
- We can estimate the effects of direct and indirect data on Bing quality

Approach

- Use data from Bing/Yahoo only
  - Query trend used IE logs; this is Bing logs
- Build semantic graph
  - Using click similarity
  - Exclude observations where graph not completed
  - Conservative because true singletons are excluded
- Assess how much indirect data is available

Semantic Graph Illustration

Algorithm to Build Semantic Graph



- Queries leading the same URL frequencies are similar
- URLs with the same queries are similar
- Recognized methodology: Baeza-Yates and Tiberi, 2007, KDD.

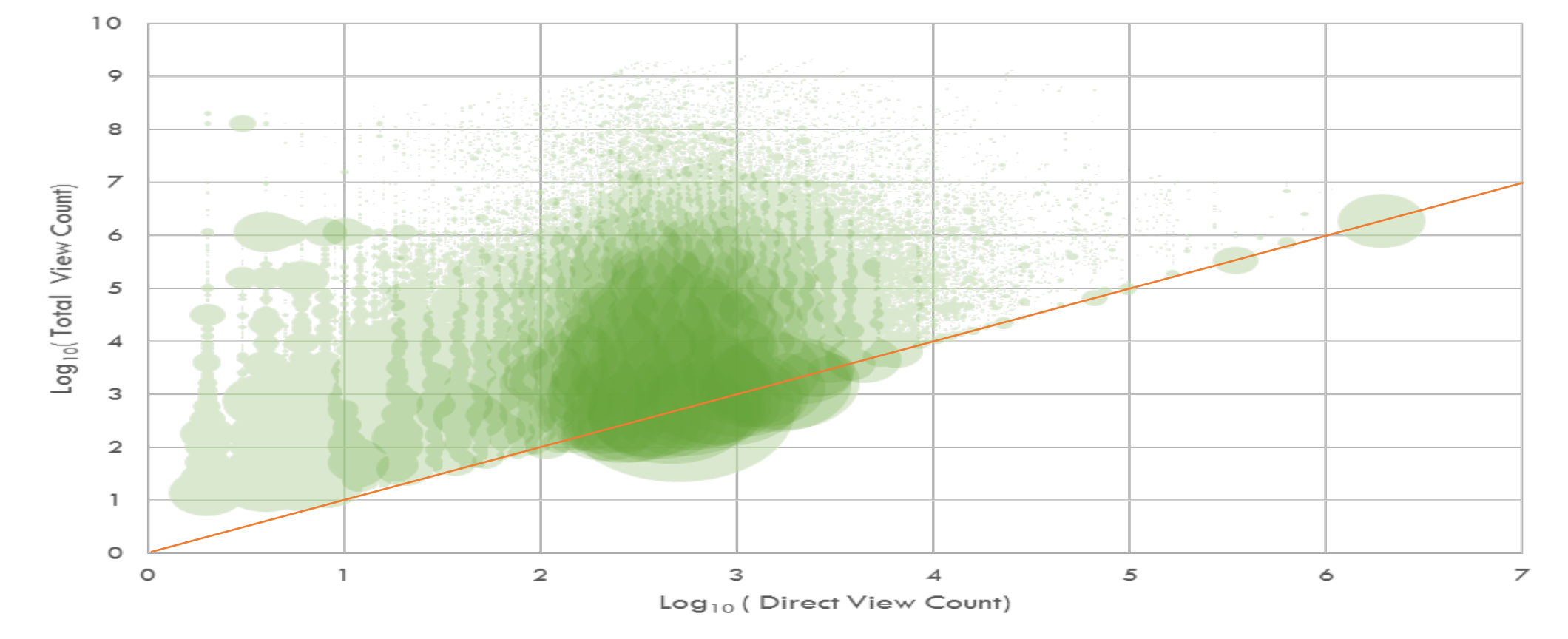
Stats

100B searches = 4.5B queries

Type of query	Number	% of queries	% of traffic
Queries that could be clustered	2.6B	53.3%	92.3%
Unclustered queries	1.9B	42.7%	7.7%
Total	4.5B	100.0%	100.0%

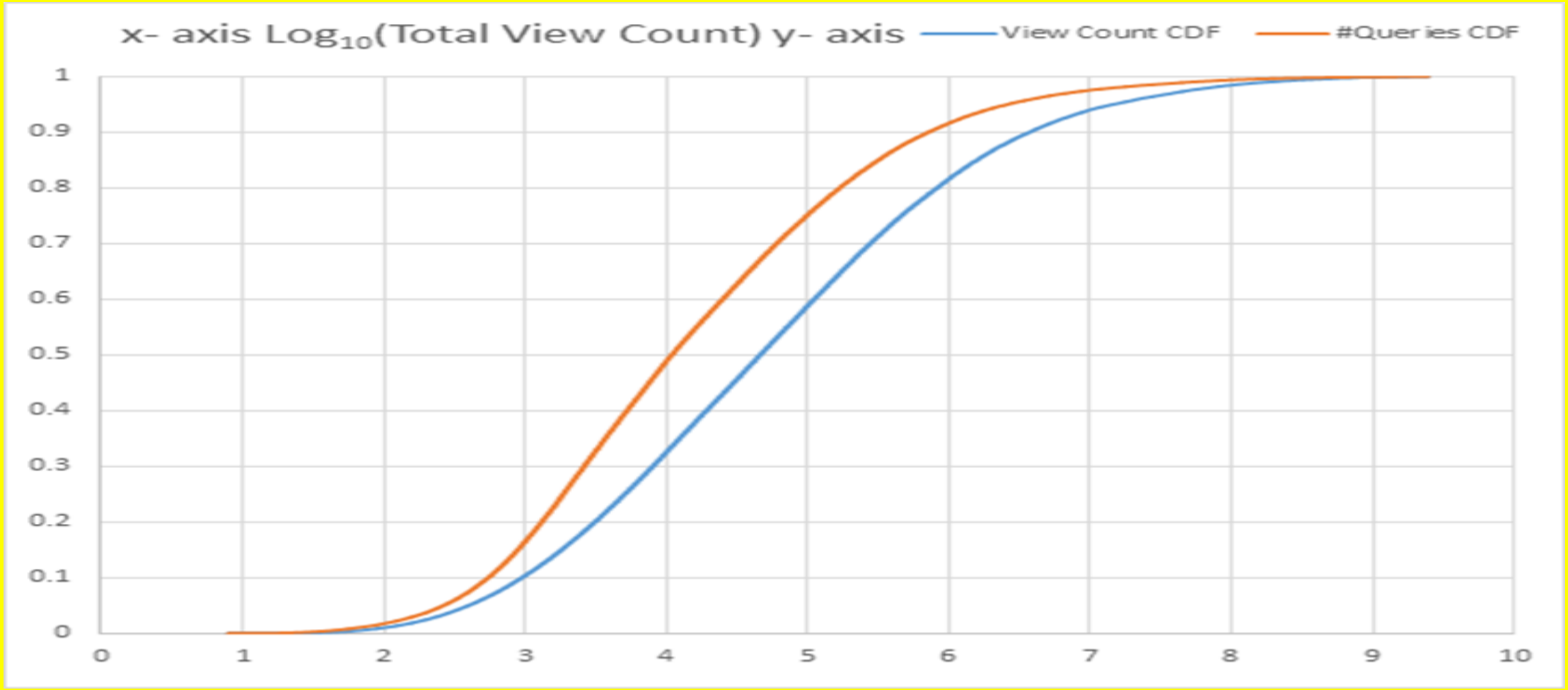
2.6B queries mapped to 128M clusters

Many Queries Have Limited Indirect Data: Indirect Data is Not a Panacea



Circle Radius ≡ #Query Instances

Total (Direct + Indirect) Data Frequency
17% of Queries, 10% of Searches Have < 1000



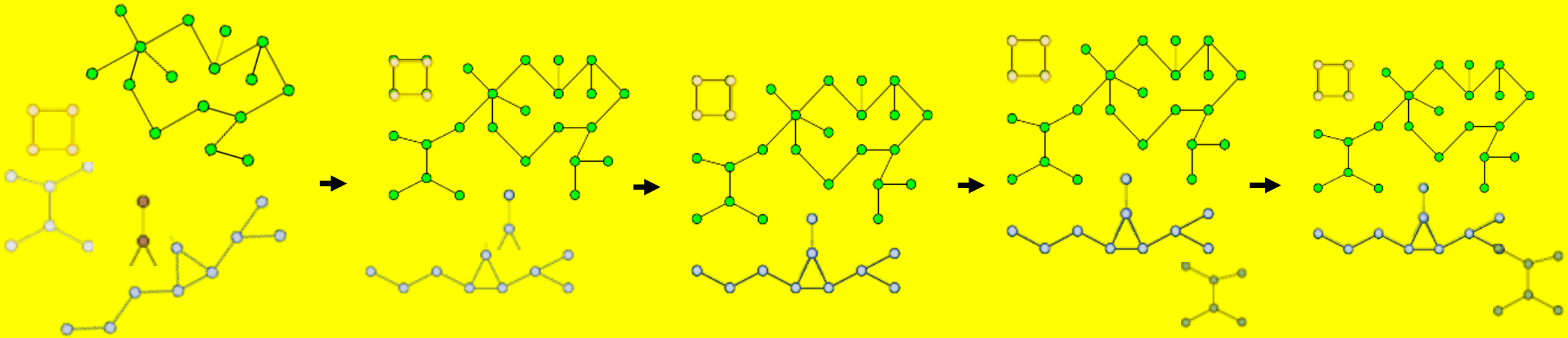
Data Accumulation

- Define new queries
  - Zero observed in 2012
  - Received 10+ per month in 2013
  - Results in 17K queries; 7800 after deduplication
- Build Graphs successively over time
  - New data allows cluster-joining

Examples

Query	Reason it appears (based on human judgment)
minecraft miniplex	The actual search is for Minecraft Mineplex, a Minecraft online server service launched in Oct 2013.
despicable me training wheels	This short movie was released in Oct 2013.
wwe 2k14 xbox 360 controls	The game's release date was 29th Oct 2013.
ipad air pictures	iPad air was launched in Nov 2013 while pictures were released/leaked around Oct 2013.
gta 5 online funny moments	GTA 5 game was launched on 17th Sep 2013 and it took some time for folks to get online and make funny moments. This query started appearing around Oct 2013 timeframe.

Illustrative graphics for construction of progressive clusters



We built the cluster progressively with data increasing each month for 12 months

Effect of direct & indirect view count on success (long) click through rate

We regressed the line given below for every query in the sample and averaged the coefficients and constant

$$y_{SuccessCTR} = \alpha + \beta_1 x_{IndirectViewCount} + \beta_2 x_{DirectViewCount}$$

β_1 = 2.251 E -05 [2.79 E-07 to 4.48 E-05] +

β_2 = 1.109 E -05 [5.28 E-06 to 1.69 E-05] +

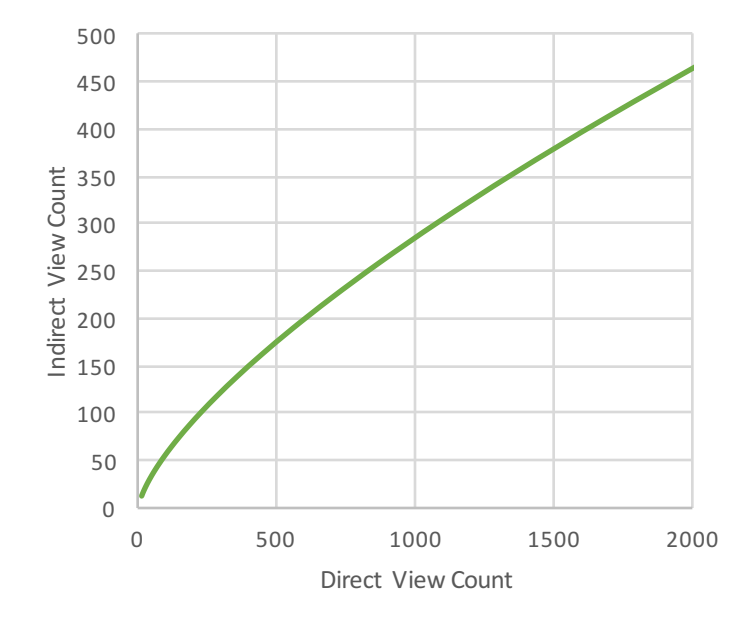
α = 0.742 [0.740 to 0.745] +

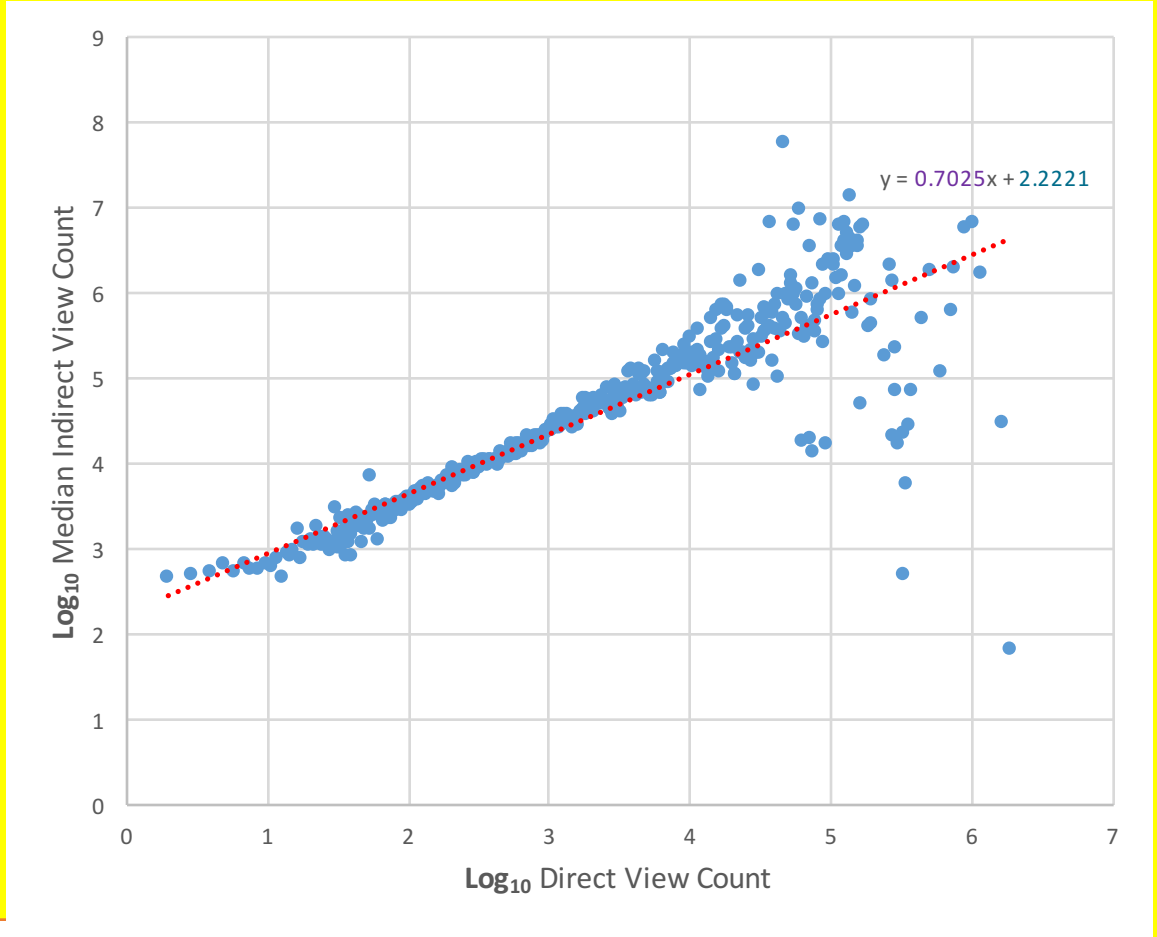
Inference 1:
Both the view counts contribute positively to the increase in Success CTR.

* We took queries with Pearson's coefficient of 0.7 or lesser deriving reliable coefficients for multiple regressions. +85% CI for all intervals

Effect of direct view count on Indirect view count (Collinearity)

$$x_{IndirectViewCount} = 2.221\, x^{0.7025}_{DirectViewCount}$$



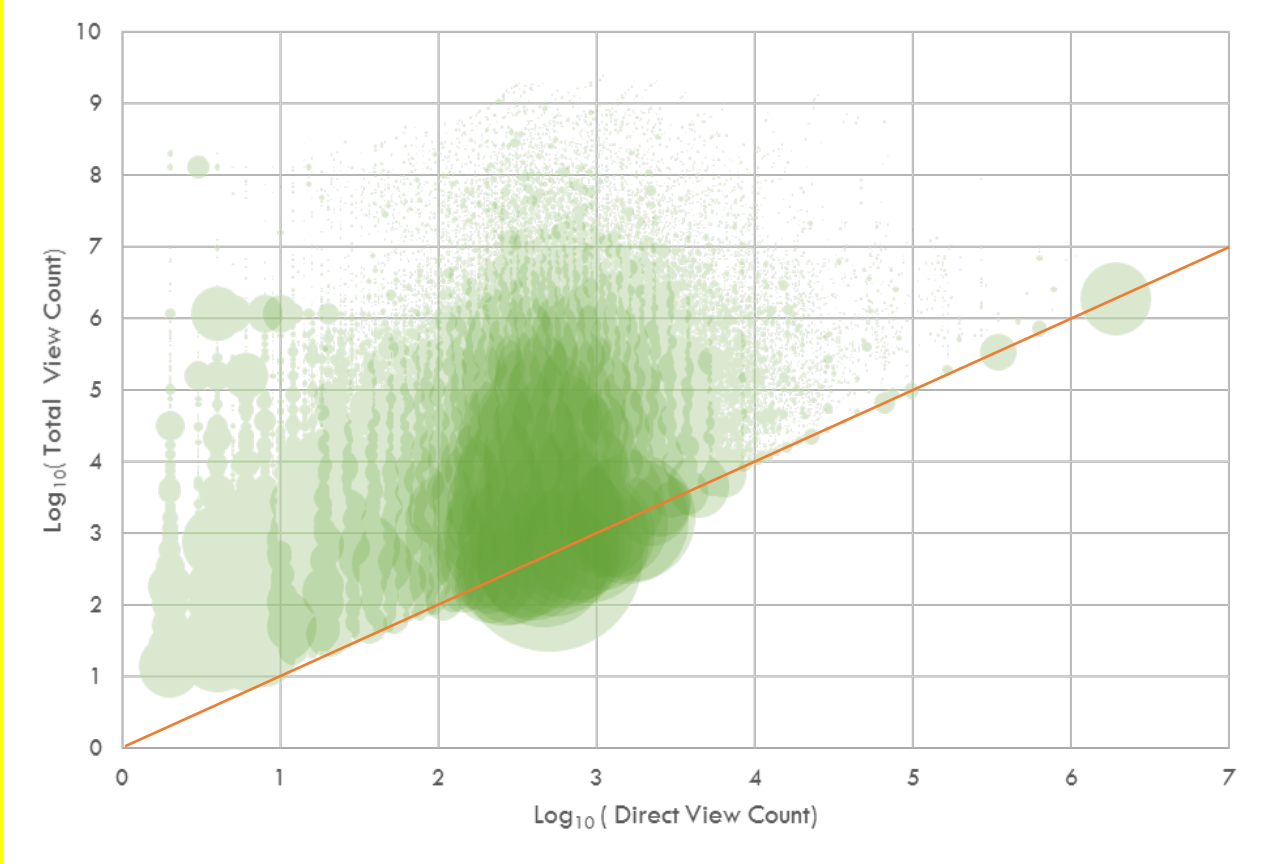


Inference 2:
For direct view count increase, a positive growth of the indirect view count occurs.
The growth in indirect view count is faster during the initial increase of the direct view count.

Effect of direct view count on Indirect view count

Circle Radius ≡ #Query Instances

Line of Direct View Count = Total View Count

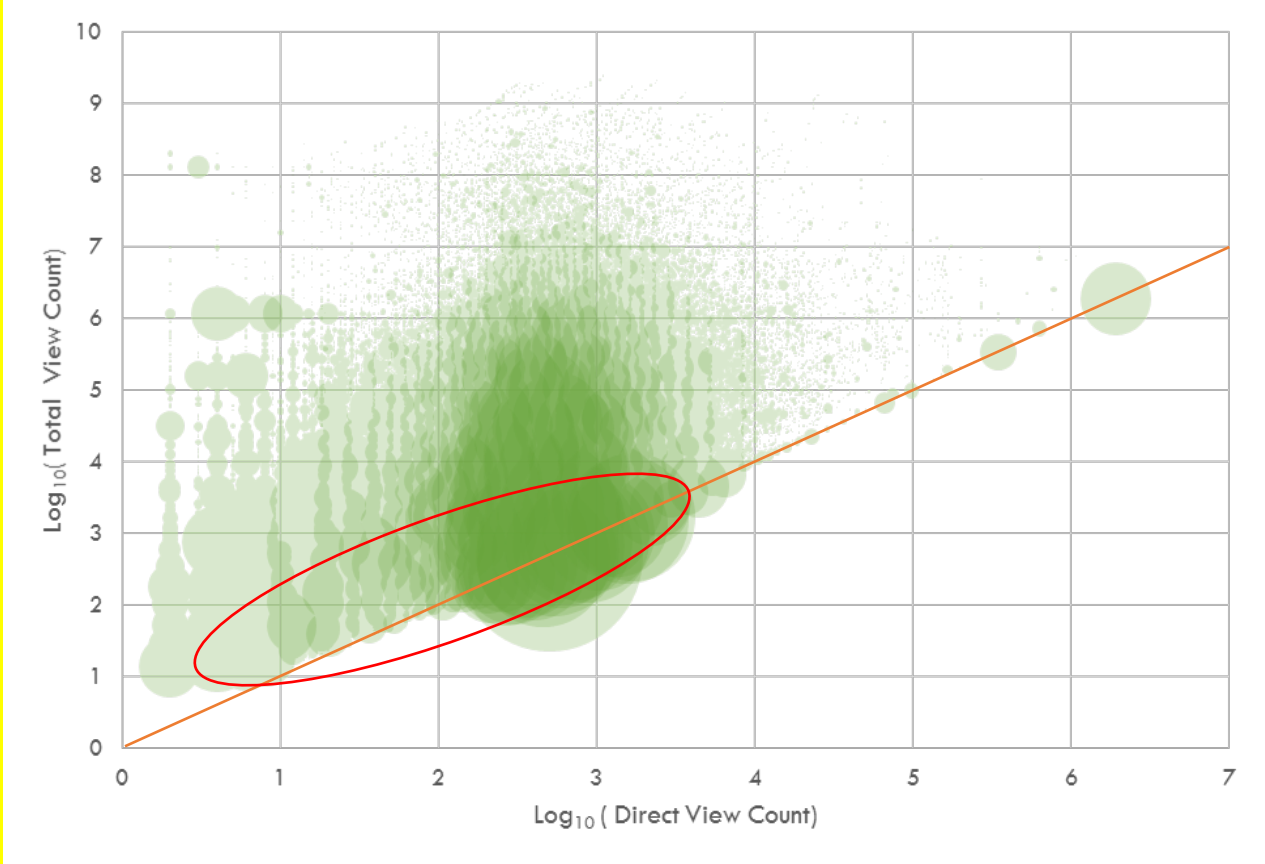


Inference 2:
For direct view count increase, a positive growth of the indirect view count occurs.
The growth in indirect view count is faster during the initial increase of the direct view count.

Effect of direct view count on Indirect view count

Circle Radius ≡ #Query Instances

Line of Direct View Count = Total View Count



Inference 3:
Even after adjusting for indirect data, many low data query clusters remain.

Bing Gets Better as More Data Becomes Available

We regressed the line given below for every query in the sample and averaged the coefficients and constant

Average click position rises as more data becomes available

$$y_{AverageClickPosition} = \alpha_{ClickPos} + \beta_{ClickPos}\, x_{DirectViewCount}$$

$\beta_{ClickPos}$ = (-)1.034131E -03 [(-)1.07 E-03 to (-)1.00 E-03] +

$\alpha_{ClickPos}$ = 2.597 [2.524 - 2.670] +

Inference 4:
Means Click Position decreases as View Count increases, a ranking improvement.

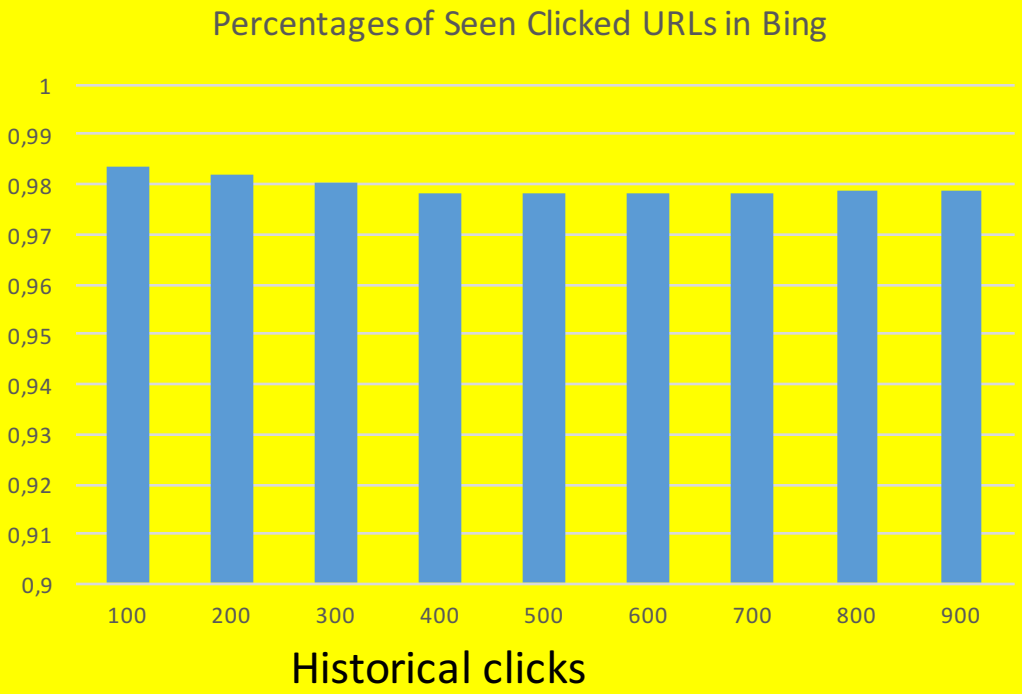
+85% CI for all intervals

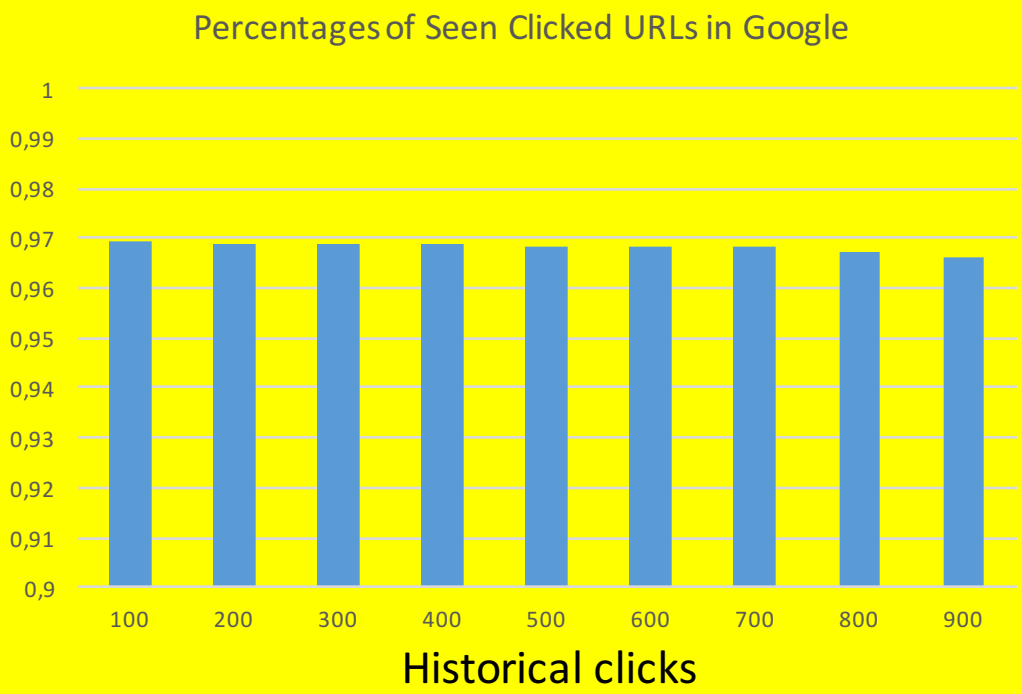
Does Search Improve Only Because Content Improved?

- Increased data is increasing clicks, suggesting better performance. But in principle, search engines could be static, and just the available results are improving.
- Mostly existing URLs, not newly created ones
- Clicks migrate upward, showing better search results, not just better content

IE Logs: Few New URLs in Rare Queries

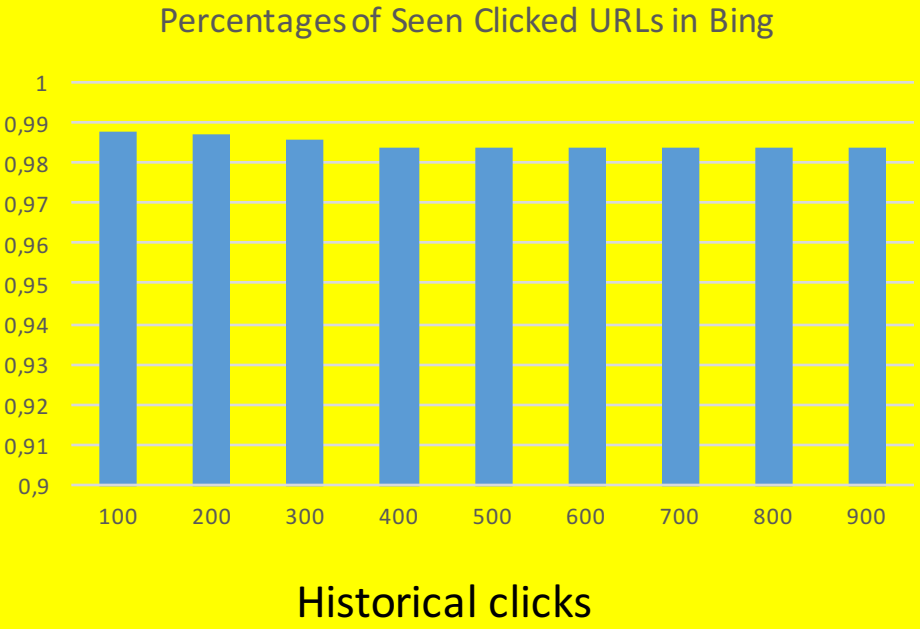
- Perhaps improvements due to discovery of new URLs?
- No, most URLs already existed (note Google has different query mix)

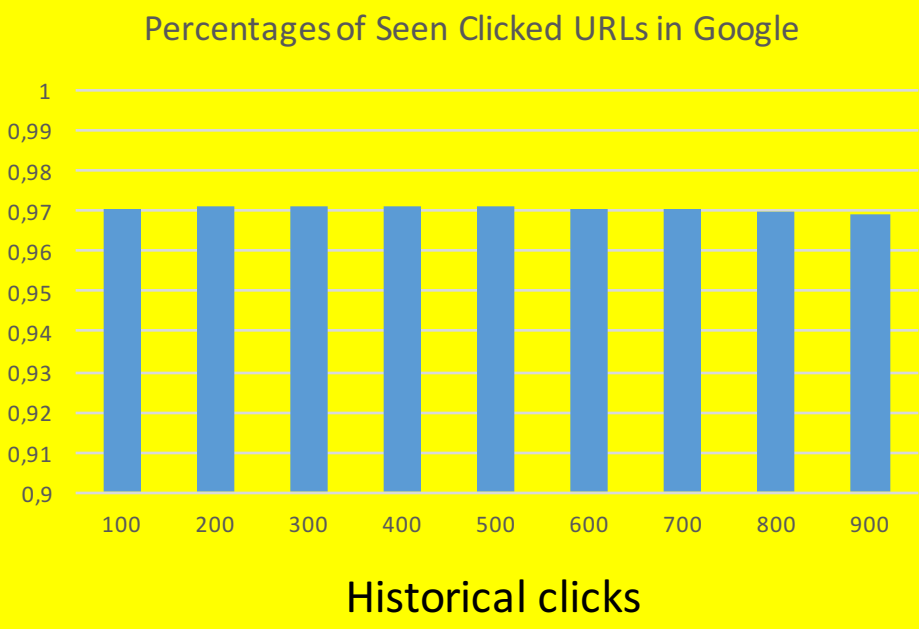




Same in the EU

- The same analysis for EU market, with similar observations
  - The percentages for EU market are around 97~98.5%;
  - The real percentages should be even higher due to low coverage of IE logs.





Bing: Effect of direct view count on URL Position

We regressed the lines given below for every query in the sample and averaged the coefficients and constant

$$y_{DistributionOfClickPosition\boldsymbol{k}} = \alpha_{\text{ClickPos}\boldsymbol{k}} + \beta_{\text{ClickPos}\boldsymbol{k}}\, x_{DirectViewCount}$$

$\beta_{\text{ClickPos1}}$ = 3.836 E -04 [5.28 E-04 to 1.69 E-04] +	$\beta_{\text{ClickPos2}}$ = (-)1.5911 E -05 [(-)2.04 E-05 to (-)1.15 E-05] +	$\beta_{\text{ClickPos3}}$ = (-)5.996 E -06 [(-)1.33 E-05 to 1.33 E-06] +	$\beta_{\text{ClickPos3}}$ = (-)4.321 E -05 [(-)4.76 E-05 to (-)3.89 E-05] +
$\alpha_{\text{ClickPos1}}$ = 0.5681 [0.541 – 0.596] +	$\alpha_{\text{ClickPos2}}$ = 0.2952 [0.290 – 0.300] +	$\alpha_{\text{ClickPos3}}$ = 0.2698 [0.260 – 0.279] +	$\alpha_{\text{ClickPos3}}$ = 0.1800 [0.173 – 0.187] +

Inference 5:
Means Click Position decreases as View Count increases.
This means better URLs are pushed to the top.

+85% CI for all intervals

Conclusion

- We measured effects of more data on new queries
  - Both Google and Bing
  - More data makes both search engines get better
- We examined related queries and websites
  - Proxied by similarity
  - Dropped singletons as a conservative measure
  - Many queries have little indirect data
  - More data makes the results better (regression)
- We provided evidence that URL position rises with more data
  - Search engine results improve with data

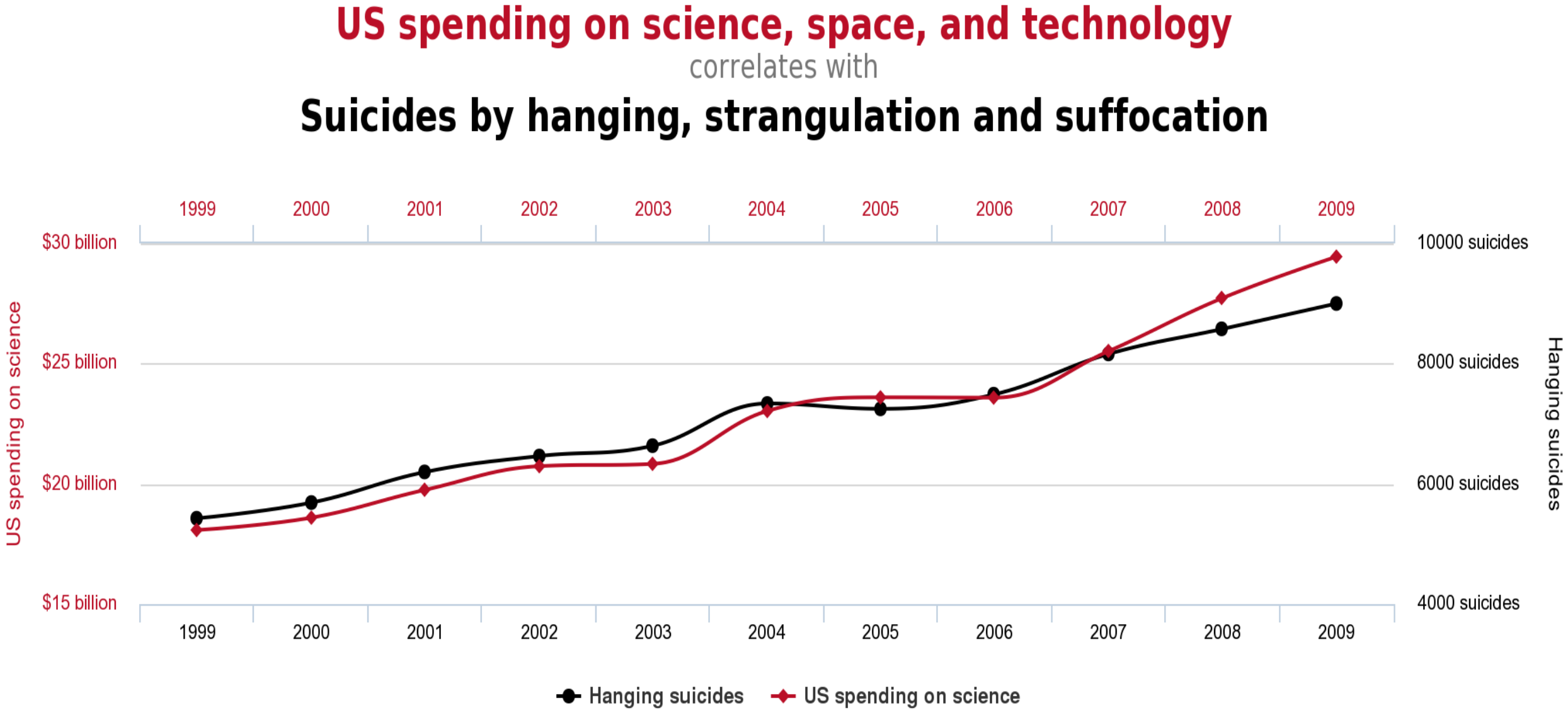
Conclusion, Continued

- Web scale involves billions of searches
  - Nearly unlimited degrees of freedom
- But, webscale involves matching billions of queries to billions of websites, searching a space with 10^{20} possibilities
- Even at web scale, more data makes search better

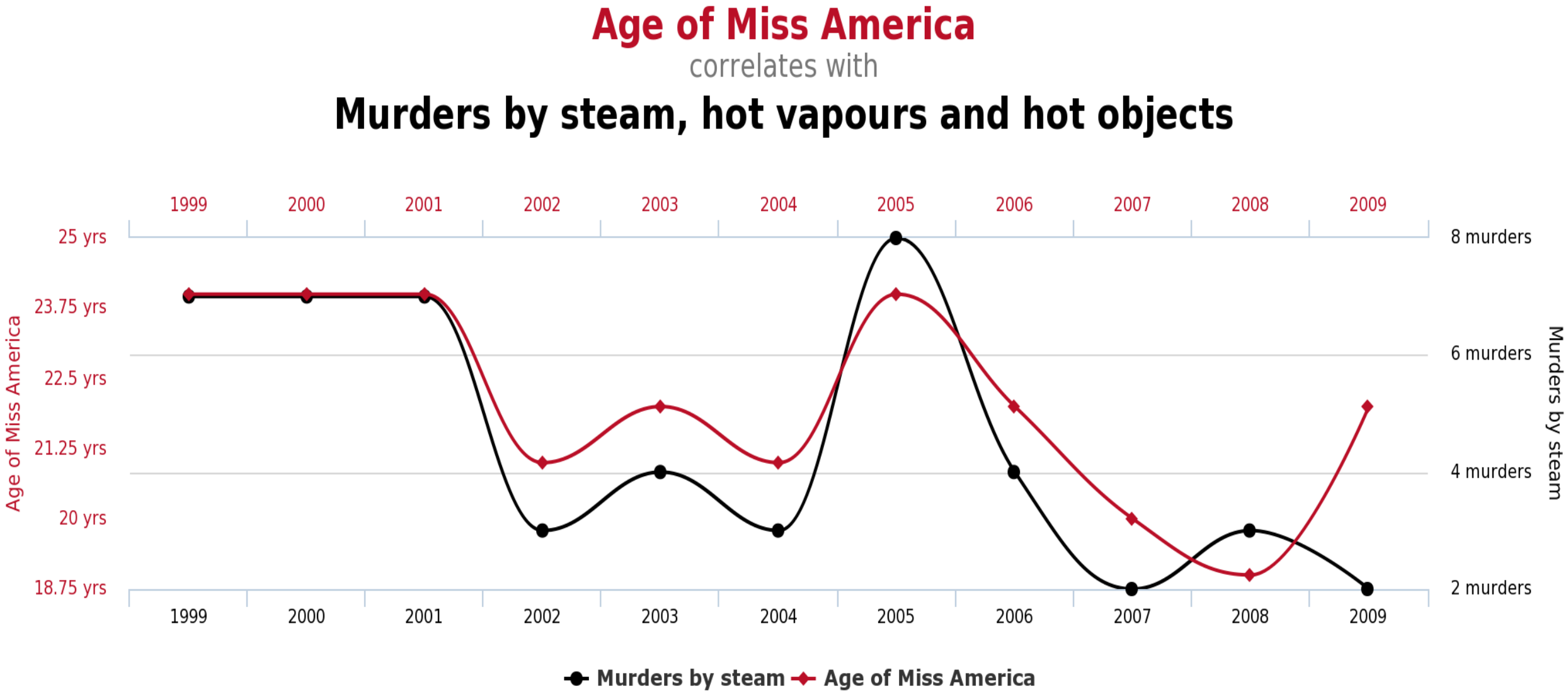
Thank You!


Trust
Partnership
Innovation
Performance
OUR VALUES
THOMSON REUTERS

Strongly Correlated (99.8%) From tylervigen.com



87% Correlated From tylervigen.com



Backup Slides

Rare Query Trend Examples: Data

		Bing			Google		
query	Q	search	#click	ctr	#search	#click	ctr
letter garden	1	228	167	0.732	222	189	0.851
	2	339	284	0.837	286	255	0.891
	3	508	441	0.868	751	675	0.898
	4	1305	1158	0.887	1075	1025	0.953
madam secretary	1	6	1	0.166	1	0	0
	2	35	21	0.6	16	8	0.5
	3	487	323	0.663	419	269	0.642
	4	1283	869	0.677	1151	781	0.678