
Do Google, Bing, & Yahoo Differ? Contingency Table Analysis on Search Engines

June 30th, 2010 - SciPy Conference - Austin, TX

Anthony Scopatz

scopatz@gmail.com

Enthought, Inc.

Nuclear Engineering, UT Austin

Idea

- Search engines are large, complicated codes.

Idea

- Search engines are large, complicated codes.
- Most people understand what they do.

Idea

- Search engines are large, complicated codes.
- Most people understand what they do.
- Few people understand how they work.

Idea

- Search engines are large, complicated codes.
- Most people understand what they do.
- Few people understand how they work.
- Fewer still *actually* work on them.

Idea

- Search engines are large, complicated codes.
- Most people understand what they do.
- Few people understand how they work.
- Fewer still *actually* work on them.
- Moreover, developers each claim that their engine is the best for a variety of reasons.

Idea

- Search engines are large, complicated codes.
- Most people understand what they do.
- Few people understand how they work.
- Fewer still *actually* work on them.
- Moreover, developers each claim that their engine is the best for a variety of reasons.

“How do I, as an outside observer, determine if two search engines differ?”

Methodology Overview

- The answer is that we can gauge the strength of association between two or more engines by performing **Contingency Table** analysis on search engine results.

Methodology Overview

- The answer is that we can gauge the strength of association between two or more engines by performing **Contingency Table** analysis on search engine results.
- Setting up this analysis is tricky for a couple of reasons:

Methodology Overview

- The answer is that we can gauge the strength of association between two or more engines by performing **Contingency Table** analysis on search engine results.
- Setting up this analysis is tricky for a couple of reasons:
 - Contingency tables work best with at least partially stochastic systems.

Methodology Overview

- The answer is that we can gauge the strength of association between two or more engines by performing **Contingency Table** analysis on search engine results.
- Setting up this analysis is tricky for a couple of reasons:
 - Contingency tables work best with at least partially stochastic systems.
 - Search engines return sets of strings (URLs) for an input search term, not numerical values.

Methodology Overview

- The answer is that we can gauge the strength of association between two or more engines by performing **Contingency Table** analysis on search engine results.
- Setting up this analysis is tricky for a couple of reasons:
 - Contingency tables work best with at least partially stochastic systems.
 - Search engines return sets of strings (URLs) for an input search term, not numerical values.
- However, the benefit we get is that we no longer need to understand the internals of how any engine works! We only care about comparisons.

Analysis Methodology

- First, to simulate a ‘random’ variable, every hour we are going to grab the latest *Hot Trends* from Google.

Analysis Methodology

- First, to simulate a ‘random’ variable, every hour we are going to grab the latest *Hot Trends* from Google.
- For each new term, we’ll gather the search results from Google, Bing, & Yahoo.

Analysis Methodology

- First, to simulate a ‘random’ variable, every hour we are going to grab the latest *Hot Trends* from Google.
- For each new term, we’ll gather the search results from Google, Bing, & Yahoo.
- Results are defined as:
 - The top ten URLs returned for each term.

Analysis Methodology

- First, to simulate a ‘random’ variable, every hour we are going to grab the latest *Hot Trends* from Google.
- For each new term, we’ll gather the search results from Google, Bing, & Yahoo.
- Results are defined as:
 - The top ten URLs returned for each term.
 - The domain of each of these URLs

Analysis Methodology

- First, to simulate a ‘random’ variable, every hour we are going to grab the latest *Hot Trends* from Google.
- For each new term, we’ll gather the search results from Google, Bing, & Yahoo.
- Results are defined as:
 - The top ten URLs returned for each term.
 - The domain of each of these URLs
 - The total estimated number of results for each search term.

Analysis Methodology

- First, to simulate a ‘random’ variable, every hour we are going to grab the latest *Hot Trends* from Google.
- For each new term, we’ll gather the search results from Google, Bing, & Yahoo.
- Results are defined as:
 - The top ten URLs returned for each term.
 - The domain of each of these URLs
 - The total estimated number of results for each search term.
- To do this, we are going to borrow an analysis tool that is often used in Biology: Contingency Tables.

Contingency Tables

The 2×2 table is most common:

Table 2: Hair Color to Sex Contingency Table

	Blonde	Brunette	Totals
Female	18	17	35
Male	11	14	25
Totals	29	31	60

Contingency Tables

The 2×2 table is most common:

Table 2: Hair Color to Sex Contingency Table

	Blonde	Brunette	Totals
Female	18	17	35
Male	11	14	25
Totals	29	31	60

But doesn't this approach ignore the underlying biology?

Contingency Tables

The 2×2 table is most common:

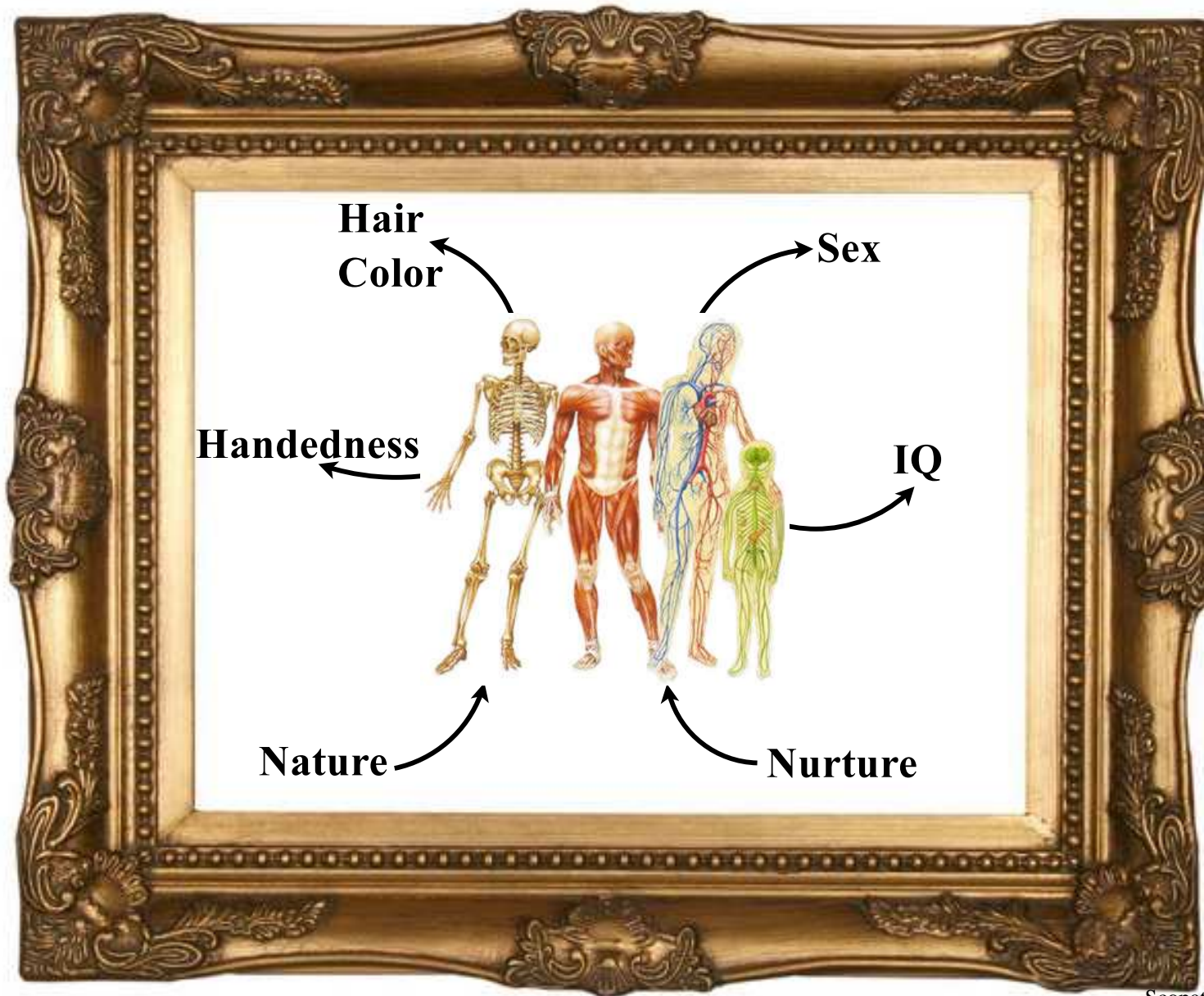
Table 2: Hair Color to Sex Contingency Table

	Blonde	Brunette	Totals
Female	18	17	35
Male	11	14	25
Totals	29	31	60

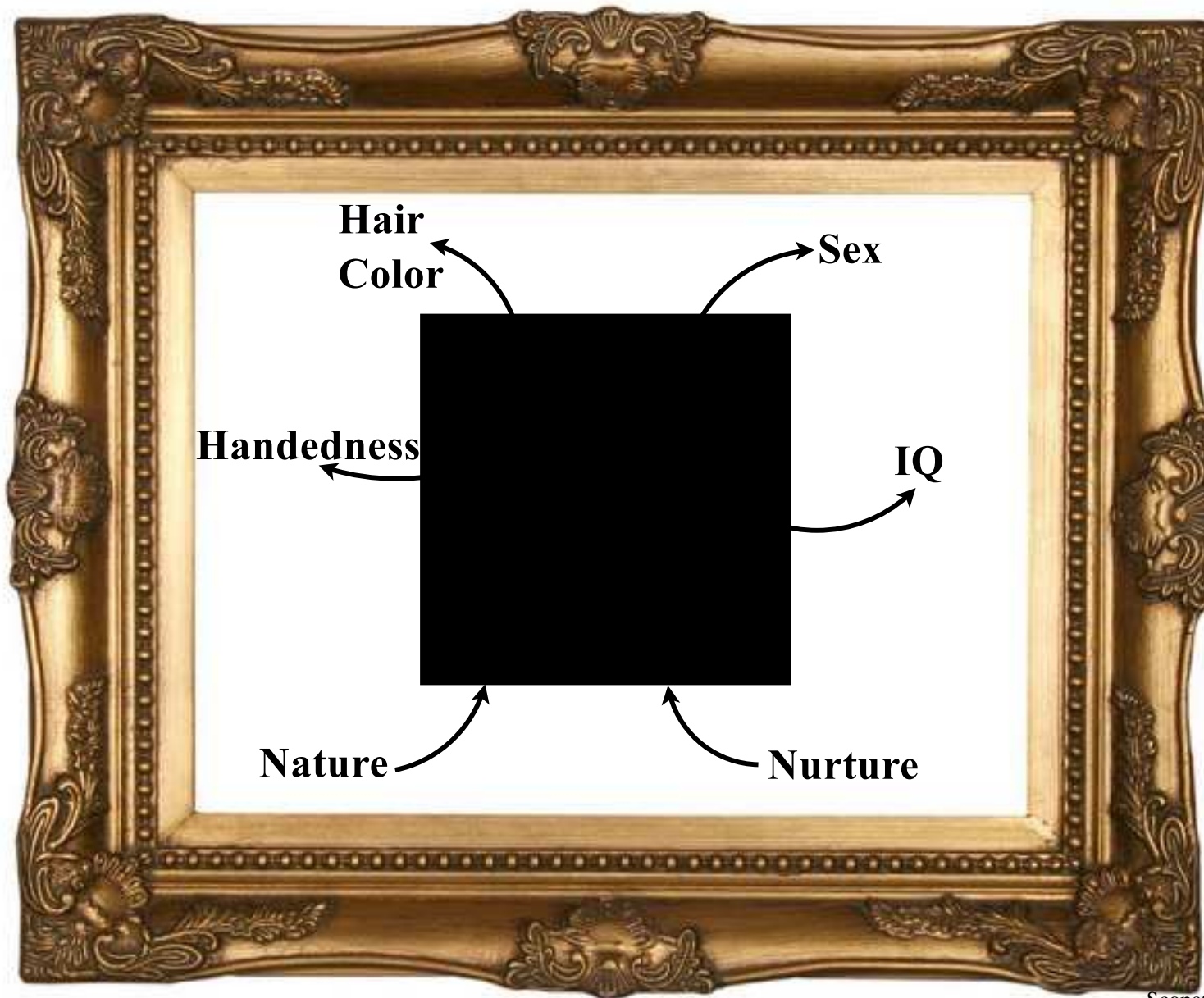
But doesn't this approach ignore the underlying biology?

Yes!

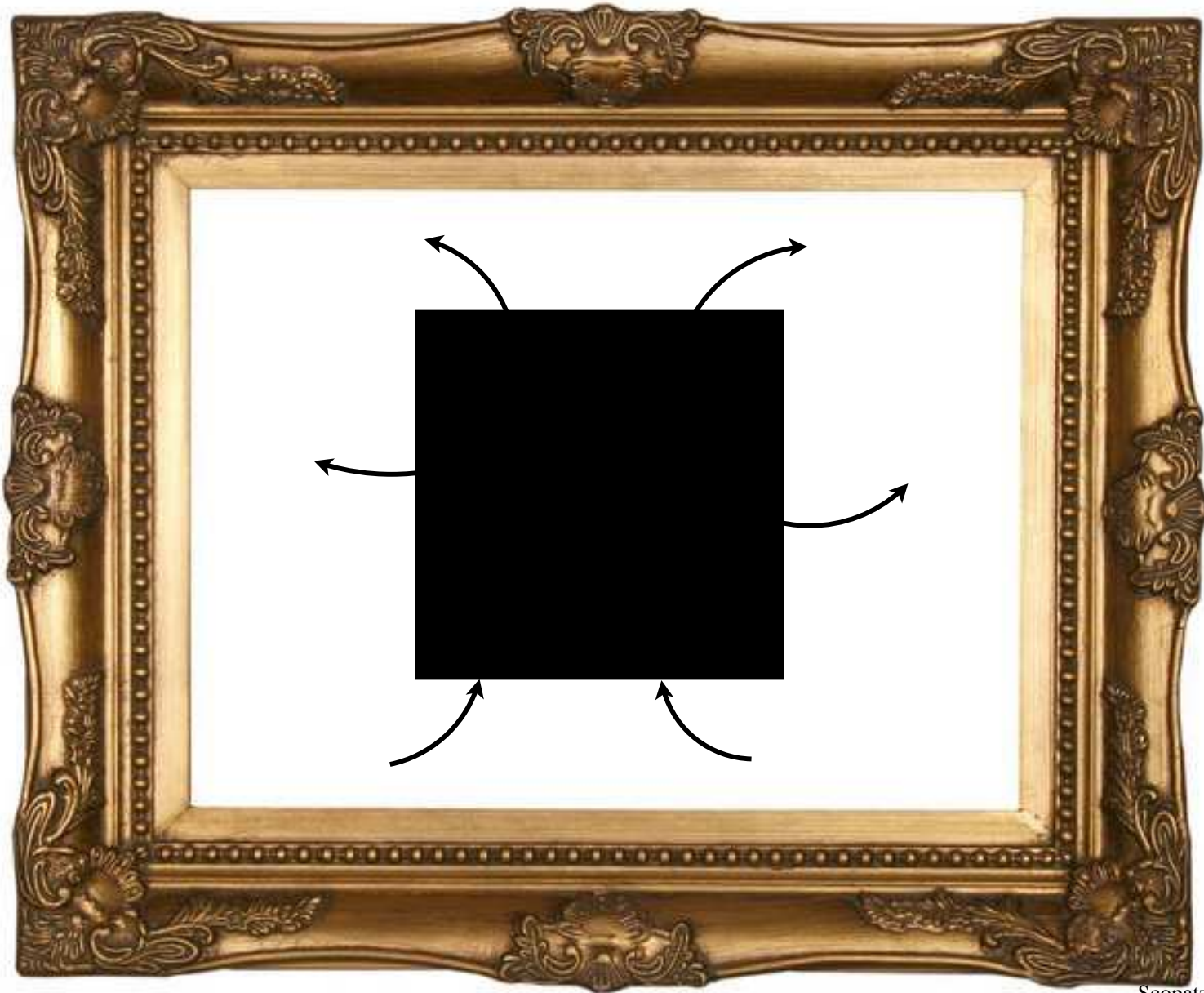
Contingency Tables



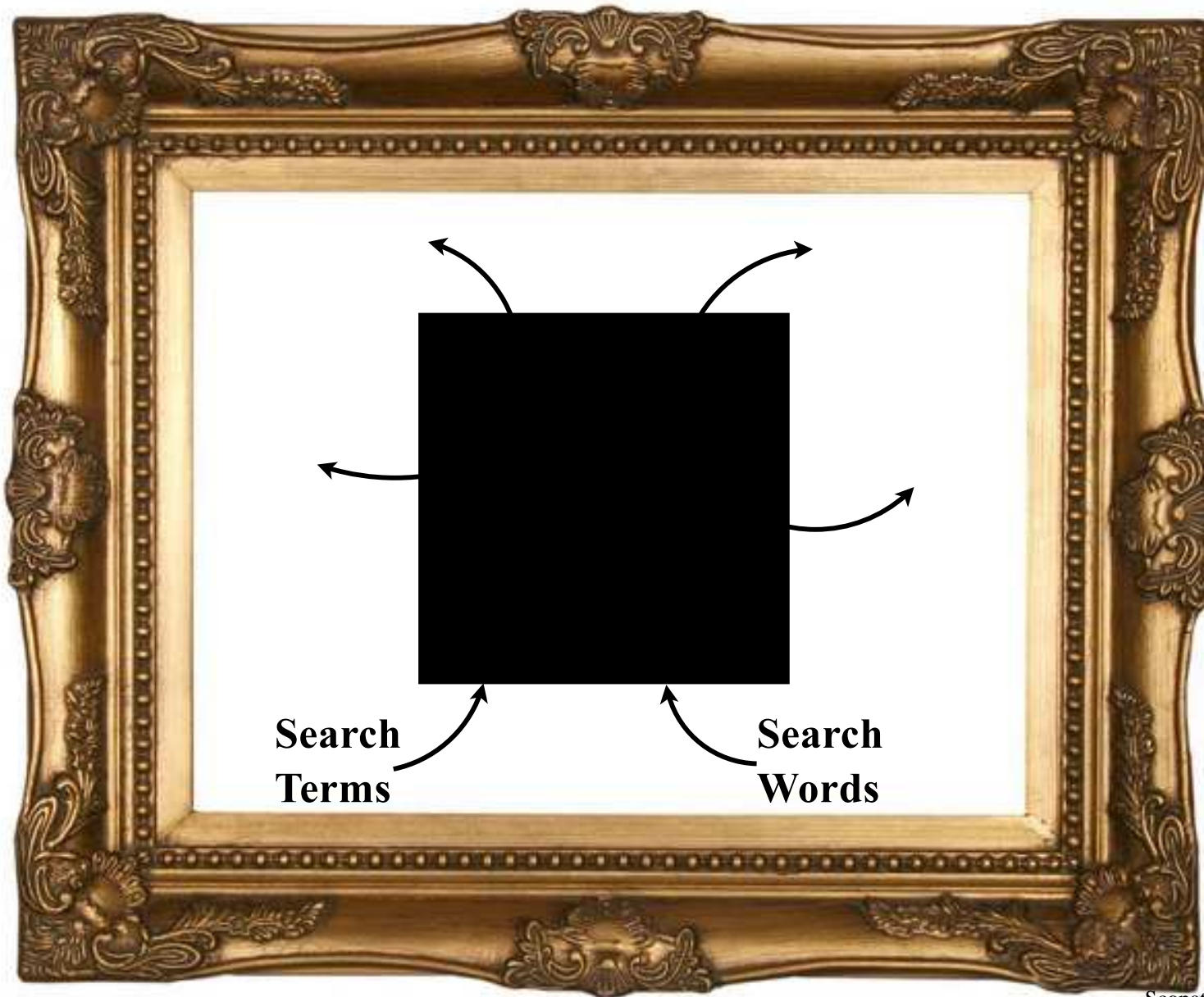
Contingency Tables



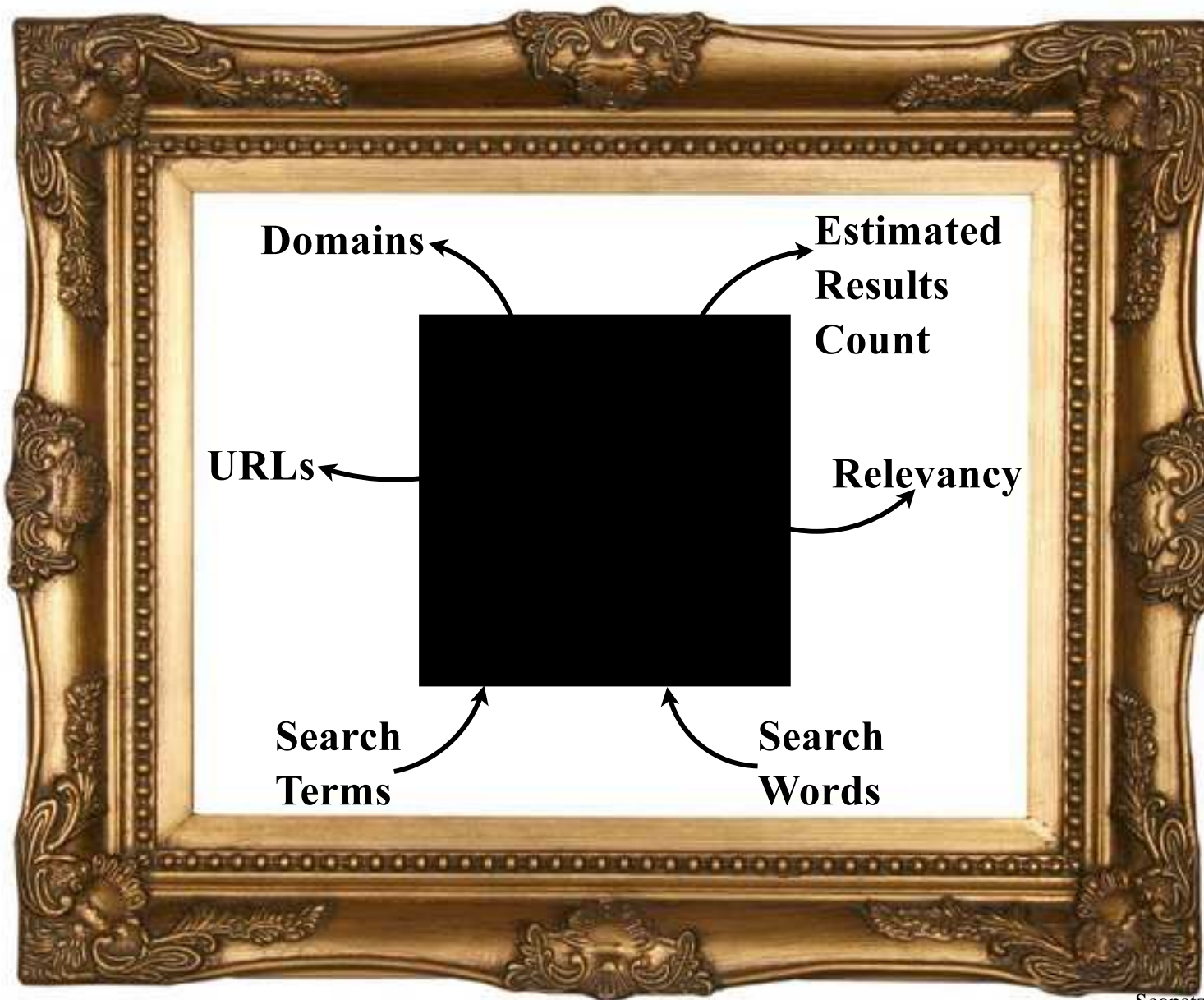
Contingency Tables



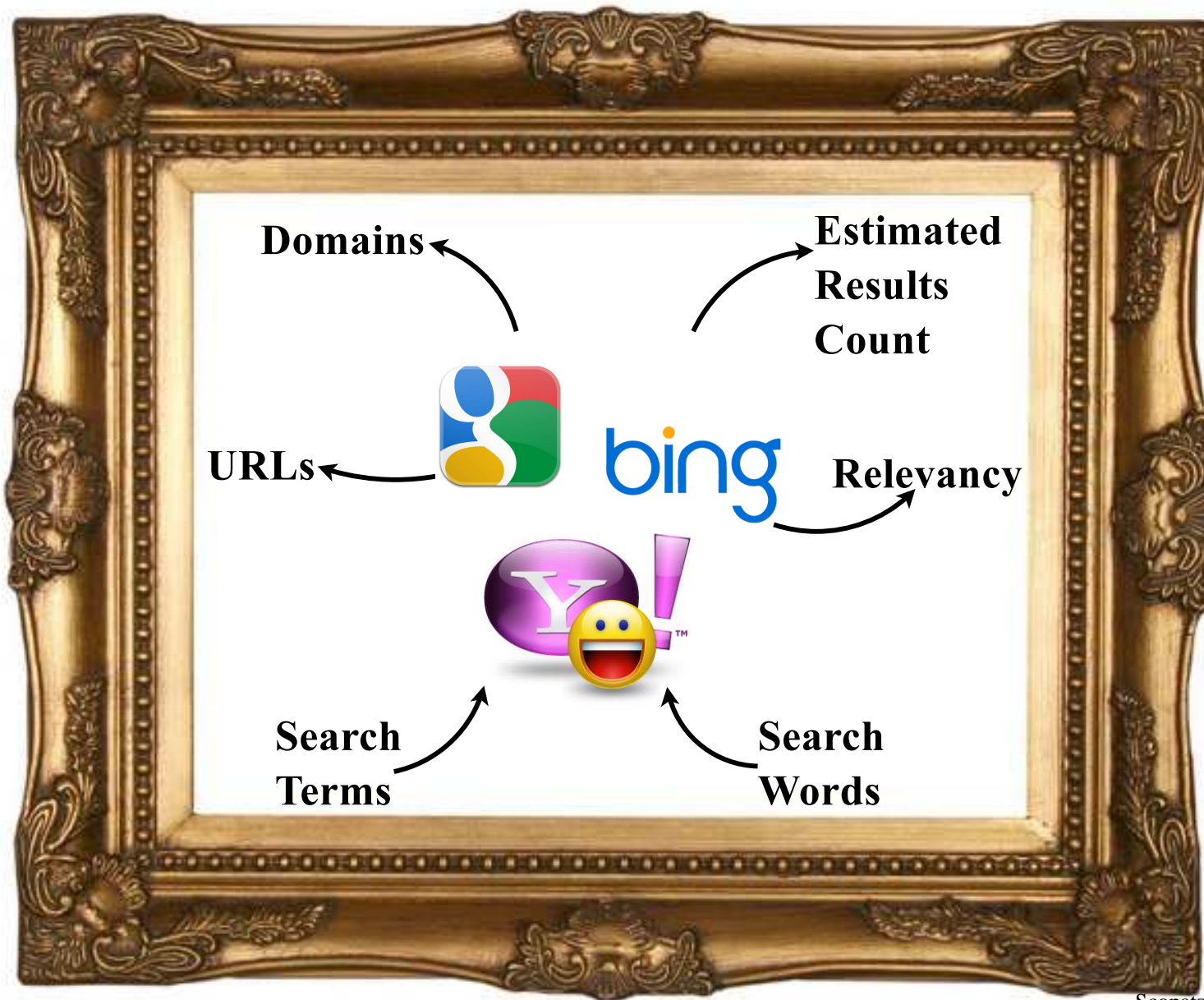
Contingency Tables



Contingency Tables



Contingency Tables



Fuel Cycle Contingency Table

For example, take the contingency table for the words in a term to the intersection of Google/Bing URLs:

Fuel Cycle Contingency Table

For example, take the contingency table for the words in a term to the intersection of Google/Bing URLs:

	words = 1	2	3	4	5	6	
G/B = 0	34	223	248	194	114	65	878
G/B = 1	81	482	360	291	125	97	1436
G/B = 2	137	730	393	257	123	79	1719
G/B = 3	206	899	371	204	89	59	1828
G/B = 4	195	731	283	171	62	29	1471
G/B = 5	155	549	213	93	27	11	1048
G/B = 6	99	233	90	49	15	7	493
G/B = 7	36	58	38	24	7	5	168
G/B = 8	4	16	10	2	1	0	33
G/B = 9	0	2	1	2	1	0	6
G/B = 10	0	1	0	0	0	0	1
	947	3924	2007	1287	564	352	9081

Contingency Table Statistics

- There are several metrics that have been developed to measure associations with contingency tables.

Contingency Table Statistics

- There are several metrics that have been developed to measure associations with contingency tables.
- The sample measure we will look at is the entropy.

Contingency Table Statistics

- There are several metrics that have been developed to measure associations with contingency tables.
- The sample measure we will look at is the entropy.
- The *Entropy* H is a measure of how evenly the data is spread out.

Contingency Table Statistics

- There are several metrics that have been developed to measure associations with contingency tables.
- The sample measure we will look at is the entropy.
- The *Entropy* H is a measure of how evenly the data is spread out.
- Moreover, we can calculate H for each search engine pair: G/B, G/Y, and B/Y.

Contingency Table Statistics

- There are several metrics that have been developed to measure associations with contingency tables.
- The sample measure we will look at is the entropy.
- The *Entropy* H is a measure of how evenly the data is spread out.
- Moreover, we can calculate H for each search engine pair: G/B, G/Y, and B/Y.
- Since we add terms to the database every hour, all of the measures can be represented by time series data.

Contingency Table Statistics

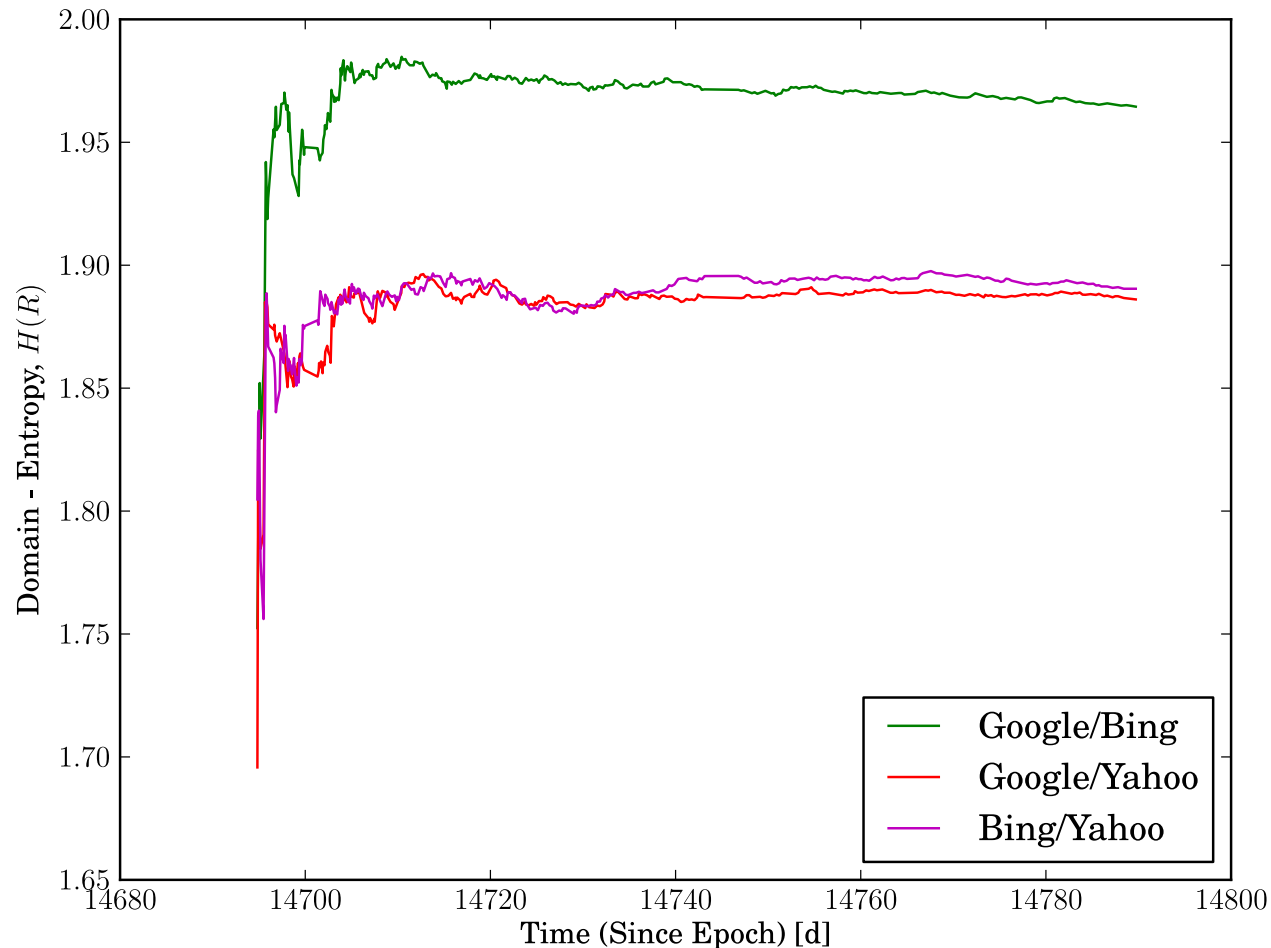
- There are several metrics that have been developed to measure associations with contingency tables.
- The sample measure we will look at is the entropy.
- The *Entropy* H is a measure of how evenly the data is spread out.
- Moreover, we can calculate H for each search engine pair: G/B, G/Y, and B/Y.
- Since we add terms to the database every hour, all of the measures can be represented by time series data.
- Presumably, after adding enough terms, the $H(t)$ for each pair should stabilize to some value...

Domain Entropy

Let's look at the entropy of the number of domain-based (not URL) intersections as a function of time.

Domain Entropy

Let's look at the entropy of the number of domain-based (not URL) intersections as a function of time.



Website

Since there is far too much data to show here, I encourage everyone to explore the data sets I have calculated.

Website

Since there is far too much data to show here, I encourage everyone to explore the data sets I have calculated.

You may find a presentation of this data at the SearchCompare website:

`http://bit.ly/ahpagi`

`http://www.scopatz.com/SearchCompare/`

Why Python Matters...

In case you were wondering, here is a list of packages that made this code possible:

NumPy	urllib
SciPy	urllib2
PyTables	json
MetaSci	feedparser
PyBing	HTMLParser
Matplotlib	Sphinx