



[HTTP://FLICKR.COM/PHOTOS/BINARYAPE/671776081/](http://flickr.com/photos/binaryape/671776081/)



[HTTP://FLICKR.COM/PHOTOS/MCCOCK/441965856/](http://flickr.com/photos/mccock/441965856/)

# BRICOLAGE: DATA AT PLAY

JOE HELLERSTEIN, UC BERKELEY



# OUTLINE

---

- ☐ SIMULTANEOUS REVOLUTIONS

- ☐ WEB 2.0

- ☐ INDUSTRIAL REVOLUTION  
OF DATA

- ☐ TAPPING THE CONFLUENCE

- ☐ OPPORTUNITY

- ☐ CHALLENGE

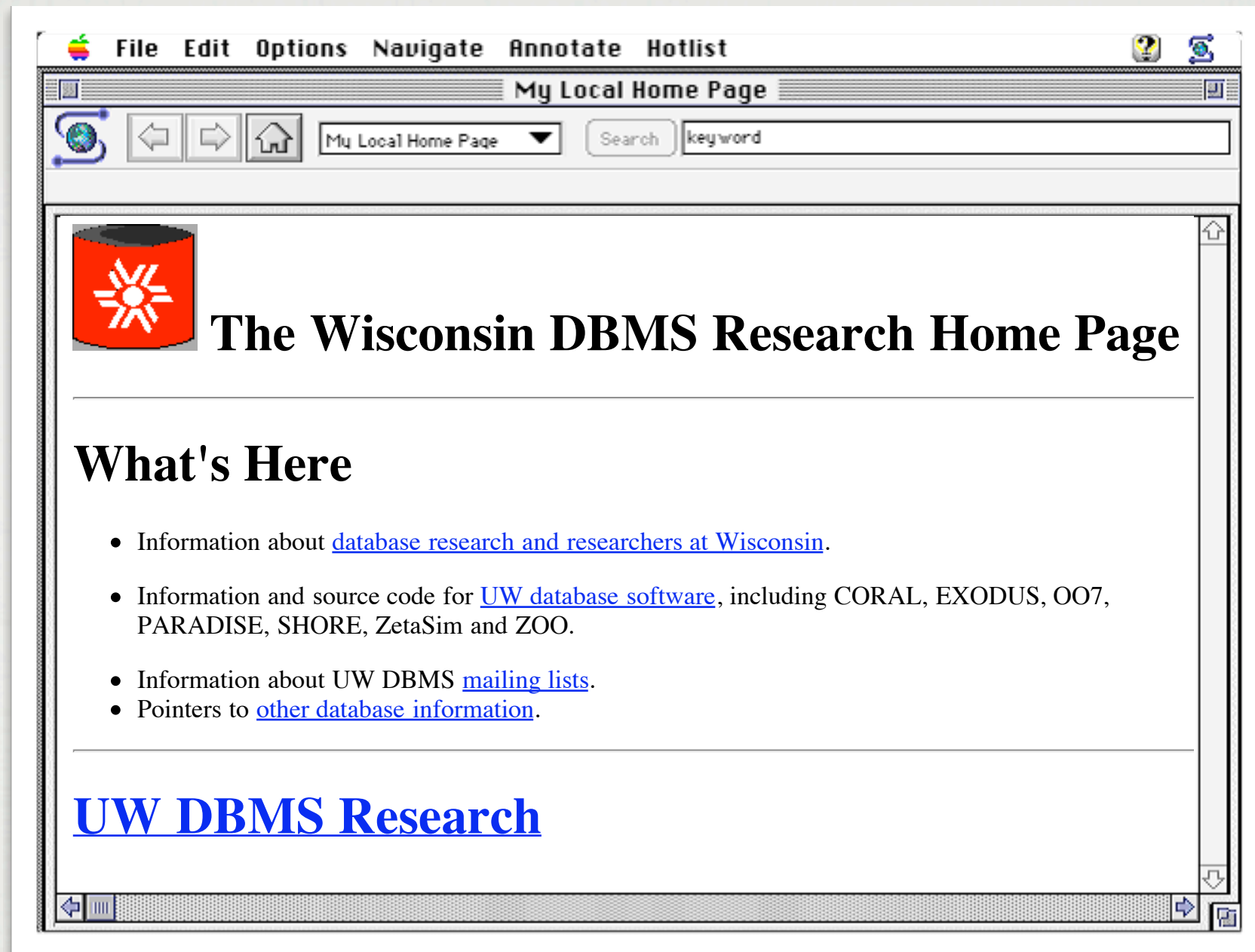
- ☐ INSPIRATION FROM A FIELD

- ☐ BRICOLAGE & PLAY

- ☐ EARLY DAYS OF DATA 2.0

- ☐ LIFECYCLE, CHALLENGES

- ☐ WHAT IS TO BE DONE?



# THE WEB, 1.0

HYPER-DOCUMENTS

I.E. ...  
PROSE



AND  
ADVERTISING

or is that a  
web 1.0  
leftover?

The screenshot shows a Facebook profile page with the following elements:

- Header:** Facebook logo, navigation tabs (Profile, Friends, Networks, Inbox), and links (home, account, privacy, logout).
- Left Sidebar:** Search bar, Applications (Photos, Groups, Events, Marketplace), and a Citi advertisement for myHomeEquity.com.
- Friend List:** A list of 26 friends, with the first three visible: Ron Avnur, Seth Bain, and Huned Botee. Each friend entry includes a profile picture, name, network, details, status, and interaction links (Send Message, Poke, View Friends, Remove Friend).

**Friend Details:**

Name	Network	Details	Status
Ron Avnur	Silicon Valley, CA	How do you know Ron Avnur?	Ron is In need of drink.
Seth Bain	East Bay, CA Stanford Alum '96	How do you know Seth Bain?	
Huned Botee	San Francisco, CA Duke Alum '01	You worked together. [ edit details ]	Huned is strangely craving fried chicken.

THE WEB, 2.0<sub>alpha</sub>

COMMUNITIES

I.E.  
PEOPLE!





[HTTP://FLICKR.COM/PHOTOS/WORDFREAK/1609963805/](http://flickr.com/photos/wordfreak/1609963805/)

PEOPLE

[HTTP://FLICKR.COM/PHOTOS/DAQUELLAMAMERA/162104797/](http://flickr.com/photos/daquellamamera/162104797/)



hmm...

INFORMATION



[HTTP://FLICKR.COM/PHOTOS/MYSTERYBEE/1659354500/](http://flickr.com/photos/mysterybee/1659354500/)

COMPUTATION

oops!

WEB 2.0



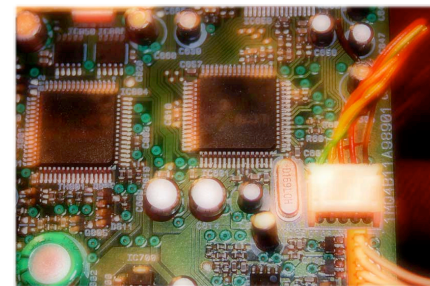
# WEB 1.0: INFORMATION? COMPUTATION?

☐ PEOPLE COMPOSE WEB PAGES



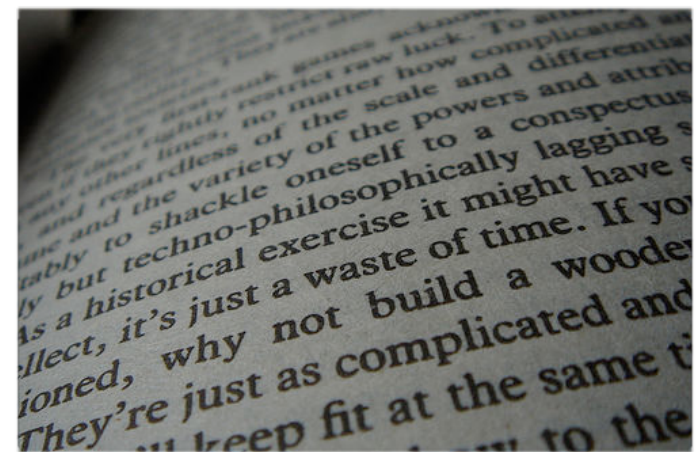
[TP://FLICKR.COM/PHOTOS/MRICON/1836673/](http://flickr.com/photos/mricon/1836673/)

☐ COMPUTERS EXTRACT  
STRUCTURE AND STATISTICS

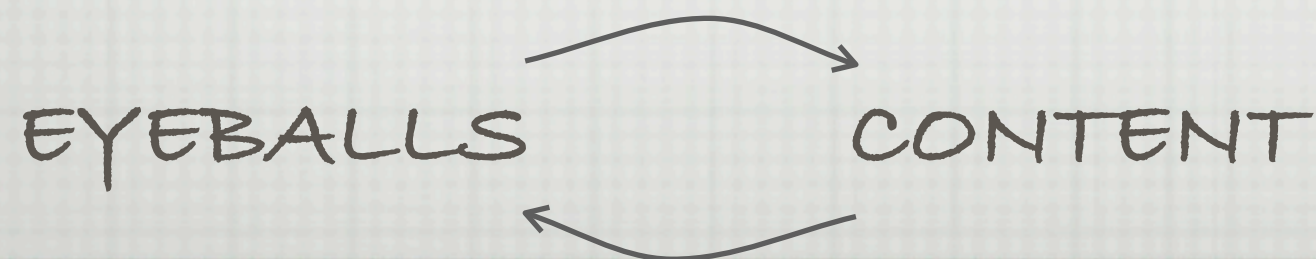


[HTTP://FLICKR.COM/PHOTOS/TIMCUMMINS/51065450/](http://flickr.com/photos/timcummings/51065450/)

☐ BENEFIT: PEOPLE GET BETTER  
ACCESS TO WEB PAGES



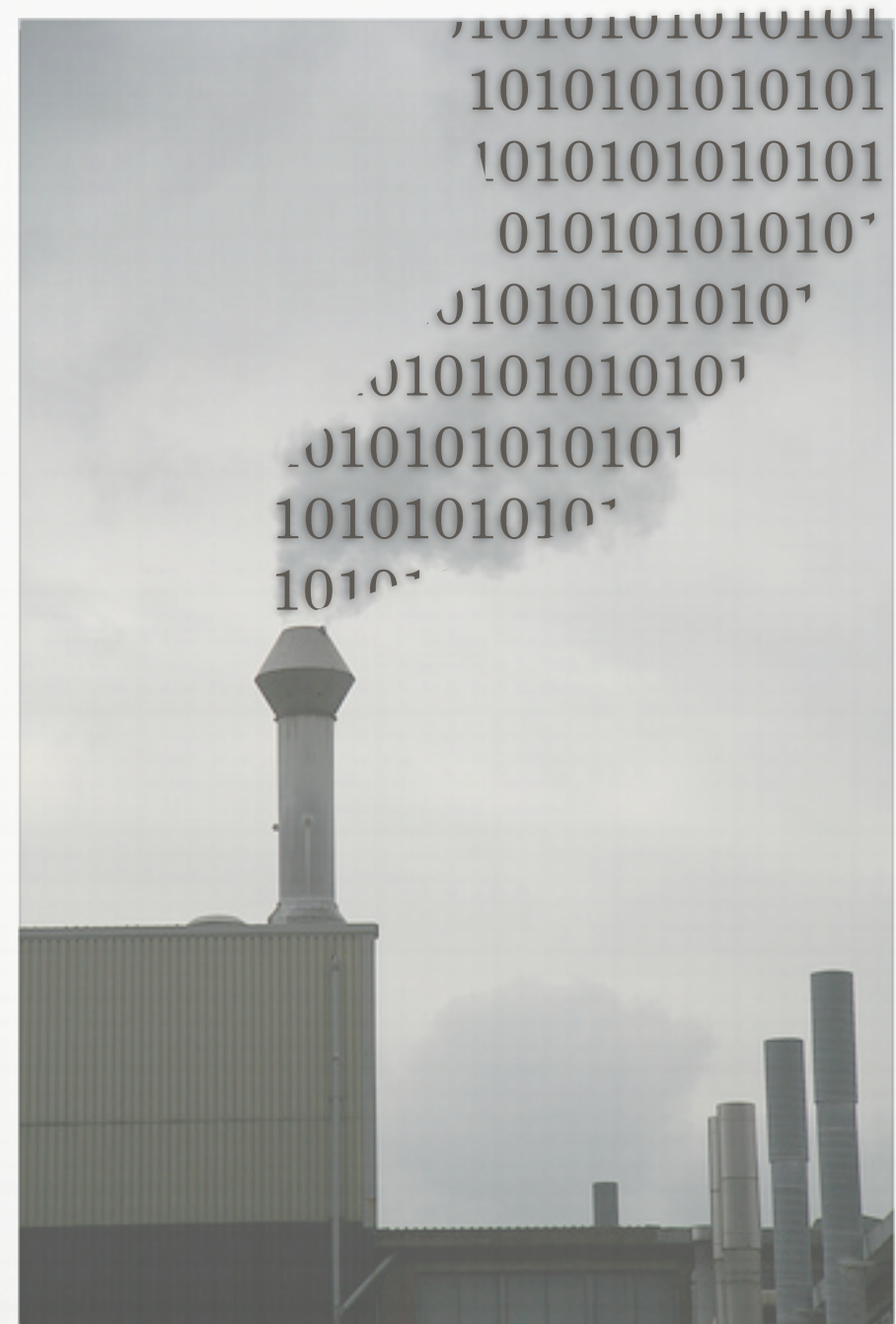
[HTTP://FLICKR.COM/PHOTOS/TIMO/20745748/](http://flickr.com/photos/timo/20745748/)





# THE NEXT INDUSTRIAL REVOLUTION: DATA

- ☐ UPC
- ☐ RFID
- ☐ GPS
- ☐ SENSORNETS
- ☐ SOFTWARE LOGS
- ☐ ...

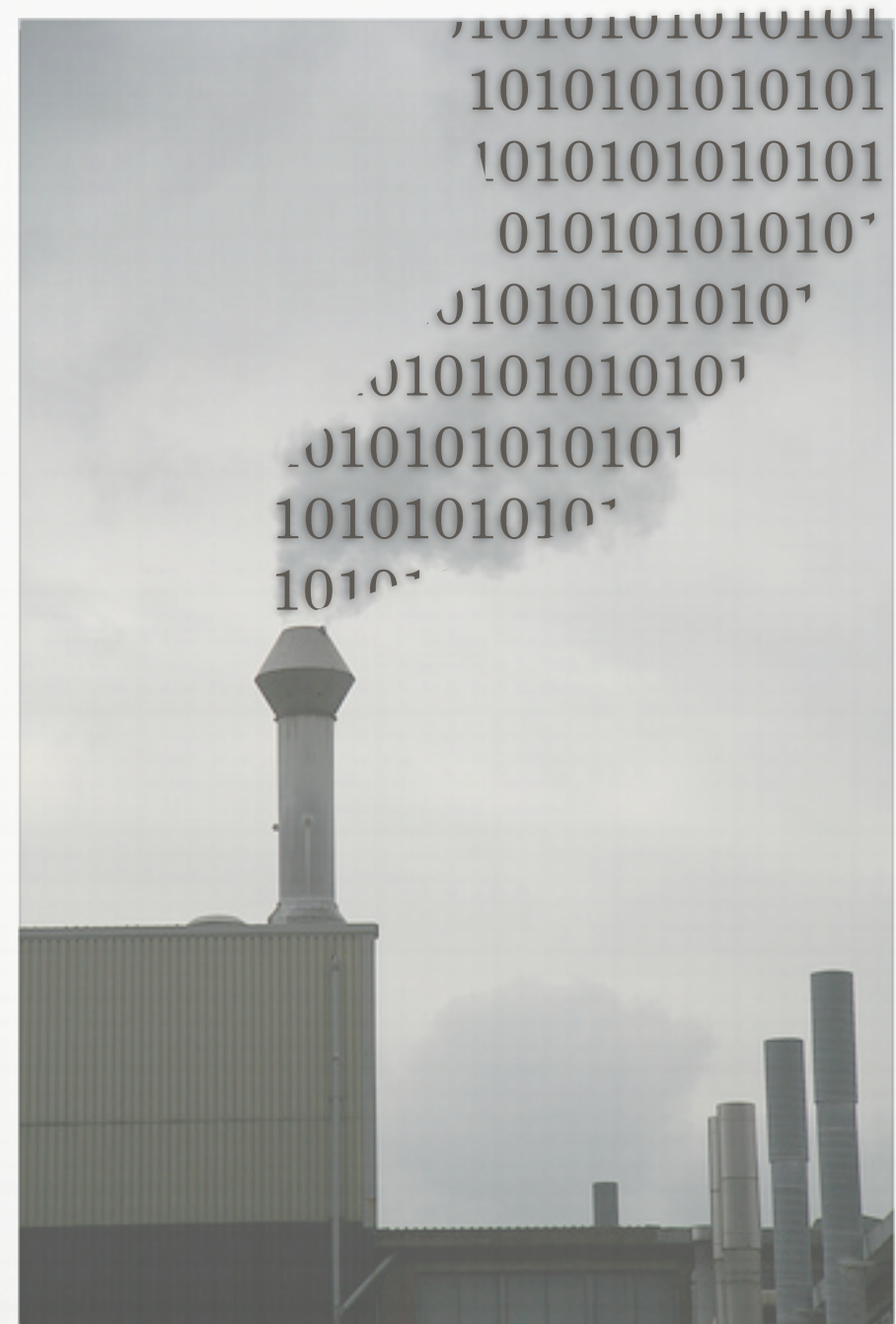


[HTTP://WWW.FLICKR.COM/PHOTOS/REVSORG/1168346563/](http://www.flickr.com/photos/revsorg/1168346563/)



# POST-INDUSTRIAL DATA

- ☐ STRUCTURED, STANDARDIZED,  
SIMPLE
- ☐ OR .... NOT?
- ☐ DATA INTEGRATION, MEET  
DATA FUSION
- ☐ NOISE, WASTE
- ☐ EVIDENCE, NOT DATA



[HTTP://WWW.FLICKR.COM/PHOTOS/REVSORG/1168346563/](http://www.flickr.com/photos/revsorg/1168346563/)



# OPPORTUNITY KNOCKS

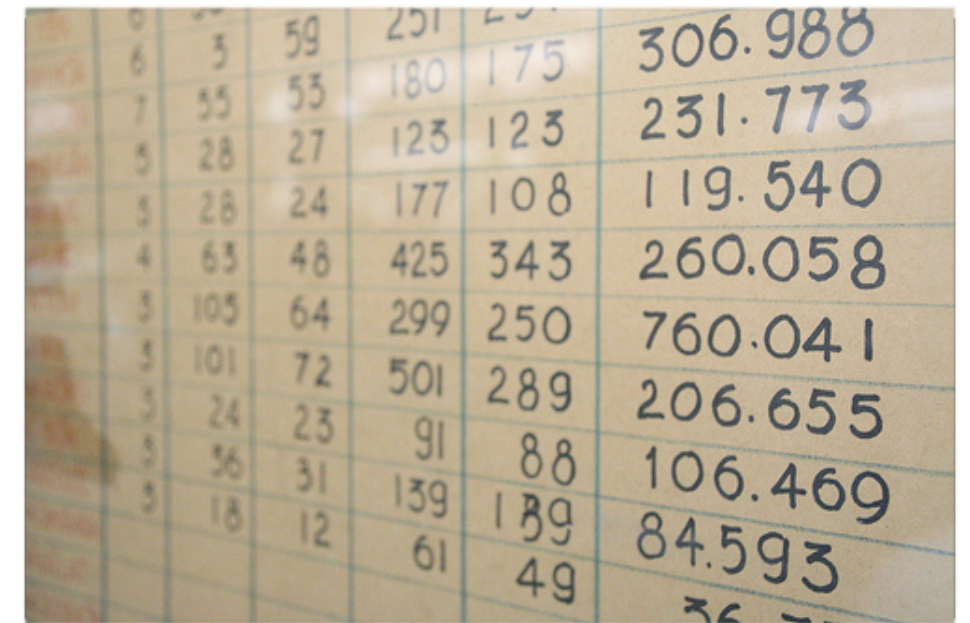
---

- ☐ CLEAR OPPORTUNITIES ON THE "PRODUCTION" SIDE
  - ☐ HW (SENSORS)
  - ☐ NETWORKING
  - ☐ INTELLIGENT DATA ACQUISITION...
  
- ☐ WHAT ABOUT "CONSUMER" SIDE?



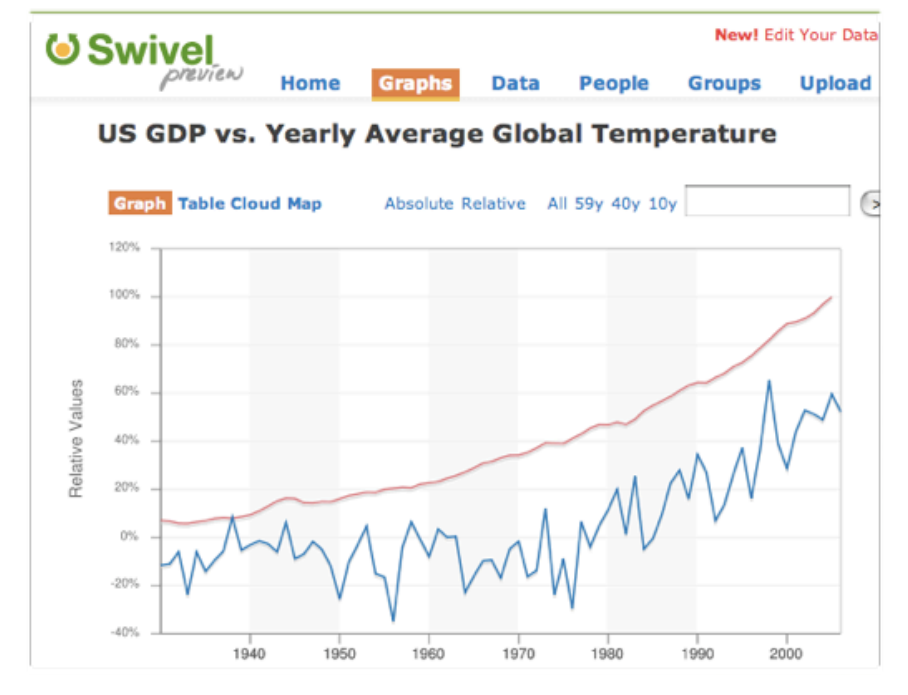
# ENRICHING THE SYMBIOSIS

- ☐ PEOPLE AND PROGRAMS BRING STRUCTURE AND STATISTICS
- ☐ WORKING TOGETHER, COMPUTERS & PEOPLE GENERATE WEB PAGES
- ☐ BENEFIT: PEOPLE GET BETTER INSIGHT & CONTROL OF THEIR STRUCTURE & STATISTICS
- ☐ (COLLECTIVE WISDOM) X COMPUTATION



6	3	59	251	175	306.988
7	55	53	180	123	231.773
8	28	27	123	108	119.540
9	28	24	177	108	260.058
4	63	48	425	343	760.041
5	105	64	299	250	206.655
5	101	72	501	289	106.469
5	24	23	91	88	84.593
5	56	31	139	139	
5	18	12	61	49	

[HTTP://FLICKR.COM/PHOTOS/JOFFLEY/118481375/](http://flickr.com/photos/joffley/118481375/)





# THE BIG QUESTION

---

- ☐ WHO CARES?
- ☐ WHO'S GOT DATA, WANTS TO ANALYZE WITH THEIR PALS?
- ☐ THIS DOESN'T SOUND LIKE AN ADVERTISING OPPORTUNITY...
- ☐ STEP BACK A SECOND:
  - ☐ IN 1993, WHO HAD A WEB PAGE?
  - ☐ COME TO THINK OF IT, WHAT DID WE USED TO THINK COMPUTERS WERE FOR?



[HTTP://WWW.FLICKR.COM/PHOTOS/STINKYPETER/889056151/](http://www.flickr.com/photos/stinkypeter/889056151/)



# OUTLINE

---

- ☐ SIMULTANEOUS REVOLUTIONS
  - ☐ WEB 2.0
  - ☐ INDUSTRIAL REVOLUTION OF DATA
- ☐ TAPPING THE CONFLUENCE
  - ☐ OPPORTUNITY
  - ☐ CHALLENGE
- ☐ INSPIRATION FROM A FIELD
  - ☐ BRICOLAGE & PLAY
- ☐ EARLY DAYS OF DATA 2.0
  - ☐ LIFECYCLE, CHALLENGES
- ☐ TOWARD A RESEARCH AGENDA



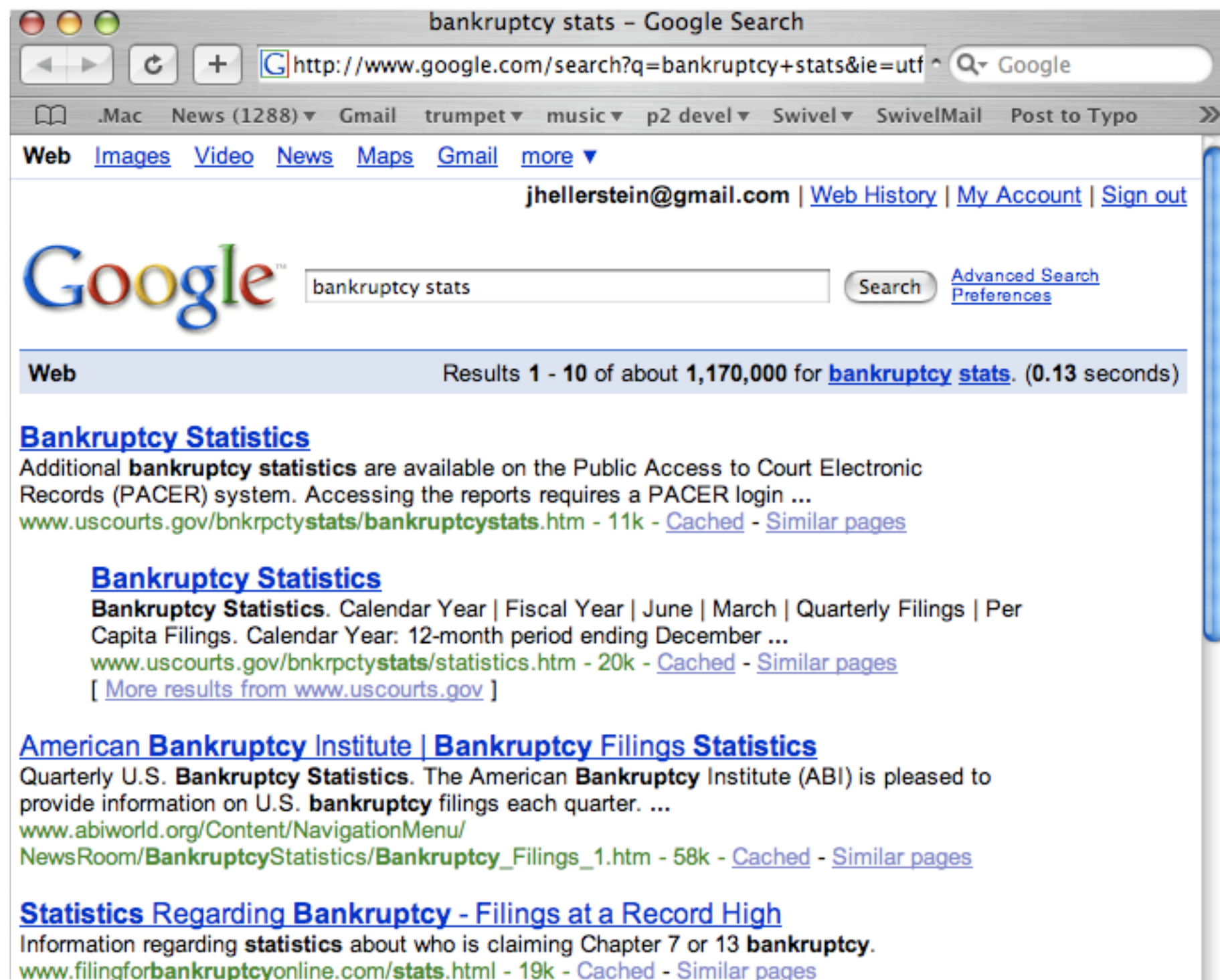
# THE DATA IS COMING

---

- ☐ POISED TO SWAMP THE HANDICRAFT TEXT
- ☐ WE HAVE EXAMPLES TODAY
  - ☐ HOW ARE WE DOING?



# WHAT SAY THE ELEPHANTS?



A screenshot of a Google search results page for the query "bankruptcy stats". The browser window title is "bankruptcy stats - Google Search". The address bar shows the URL "http://www.google.com/search?q=bankruptcy+stats&ie=utf8". The search bar contains the text "bankruptcy stats" and a "Search" button. The page shows the first 10 results out of approximately 1,170,000. The results are as follows:

**Web** Results 1 - 10 of about 1,170,000 for [bankruptcy stats](#). (0.13 seconds)

**[Bankruptcy Statistics](#)**  
Additional **bankruptcy statistics** are available on the Public Access to Court Electronic Records (PACER) system. Accessing the reports requires a PACER login ...  
[www.uscourts.gov/bnkrpctystats/bankruptcystats.htm](http://www.uscourts.gov/bnkrpctystats/bankruptcystats.htm) - 11k - [Cached](#) - [Similar pages](#)

**[Bankruptcy Statistics](#)**  
**Bankruptcy Statistics.** Calendar Year | Fiscal Year | June | March | Quarterly Filings | Per Capita Filings. Calendar Year: 12-month period ending December ...  
[www.uscourts.gov/bnkrpctystats/statistics.htm](http://www.uscourts.gov/bnkrpctystats/statistics.htm) - 20k - [Cached](#) - [Similar pages](#)  
[ [More results from www.uscourts.gov](#) ]

**[American Bankruptcy Institute | Bankruptcy Filings Statistics](#)**  
Quarterly U.S. **Bankruptcy Statistics.** The American **Bankruptcy** Institute (ABI) is pleased to provide information on U.S. **bankruptcy** filings each quarter. ...  
[www.abiworld.org/Content/NavigationMenu/NewsRoom/BankruptcyStatistics/Bankruptcy\\_Filings\\_1.htm](http://www.abiworld.org/Content/NavigationMenu/NewsRoom/BankruptcyStatistics/Bankruptcy_Filings_1.htm) - 58k - [Cached](#) - [Similar pages](#)

**[Statistics Regarding Bankruptcy - Filings at a Record High](#)**  
Information regarding **statistics** about who is claiming Chapter 7 or 13 **bankruptcy**.  
[www.filingforbankruptcyonline.com/stats.html](http://www.filingforbankruptcyonline.com/stats.html) - 19k - [Cached](#) - [Similar pages](#)



# WHAT SAY THE ELEPHANTS?

Bk2002\_1990Calendar.pdf (42 pages)

Drawer Previous Next Page 2 Back/Forward Zoom In Zoom Out Tool Mode

MD.....	35,388	35,573	0.5	31,240	31,670	1.4	35,288	39,191
NC,E....	13,917	15,072	8.3	12,034	12,969	7.8	17,591	19,694
NC,M....	10,908	11,822	8.4	10,281	10,890	5.9	19,154	20,086
NC,W....	8,887	9,488	6.8	7,708	9,621	24.8	15,350	15,217
SC.....	14,149	15,753	11.3	11,576	13,331	15.2	20,321	22,743
VA,E....	29,271	30,092	2.8	27,466	29,485	7.4	28,786	29,393
VA,W....	12,492	12,738	2	11,879	13,333	12.2	9,157	8,562
WV,N....	4,101	4,446	8.4	3,760	4,308	14.6	2,163	2,301
WV,S....	6,122	6,020	-1.7	6,449	5,788	-10.3	4,063	4,295
5TH...	126,156	129,580	2.7	112,886	121,190	7.4	175,424	183,814
LA,E....	10,236	9,750	-4.7	8,066	9,671	19.9	11,084	11,163
LA,M....	3,343	3,692	10.4	2,805	3,097	10.4	3,913	4,508
LA,W....	13,440	13,691	1.9	13,012	13,590	4.4	23,339	23,440
MS,N....	7,841	8,169	4.2	6,859	7,224	5.3	10,007	10,952
MS,S....	14,275	14,228	-0.3	12,697	14,021	10.4	16,902	17,109
TX,N....	27,146	28,084	3.5	24,186	26,679	10.3	37,440	38,845
TX,E....	11,504	12,175	5.8	10,823	8,718	-19.5	16,934	20,391
TX,S....	20,243	21,269	5.1	17,042	19,464	14.2	29,208	31,013
TX,W....	18,128	18,522	2.2	17,396	18,726	7.6	26,597	26,393
6TH...	204,435	224,908	10	175,474	209,052	19.1	234,665	250,521
KY,E....	11,550	12,208	5.7	11,159	9,503	-14.9	8,125	10,830
KY,W....	14,633	15,060	2.9	13,848	14,248	2.9	10,296	11,106
MI,E....	32,785	39,968	21.9	28,674	35,076	22.3	32,986	37,878
MI,W....	14,041	15,639	11.4	13,396	13,987	4.4	14,292	15,944
OH,N....	37,012	41,983	13.4	24,150	40,675	68.4	43,511	44,819
OH,S....	34,074	36,842	8.1	30,182	33,147	9.8	33,546	37,241
TN,E....	19,272	19,524	1.3	16,107	20,747	28.8	29,501	28,278
TN,M....	14,599	15,477	6	13,551	14,645	8.1	21,146	21,978
TN,W....	26,469	28,207	6.6	24,407	27,024	10.7	41,262	42,445
7TH...	144,730	161,146	11.3	132,346	149,635	13.1	116,140	127,651
IL,N....	51,348	57,598	12.2	49,389	53,947	9.2	43,177	46,828
IL,C....	14,465	15,649	8.2	13,346	14,837	11.2	9,134	9,946
IL,S....	8,770	9,084	3.6	8,043	8,593	6.8	7,730	8,221



# POWER TO THE PEOPLE?

## DATA 2.0

---

- ☐ SWIVEL.COM
- ☐ MANY-EYES.COM (IBM)
- ☐ DATA360.COM
- ☐ INSIGHT.BUSINESSOBJECTS.COM
- ☐ FREEBASE.COM

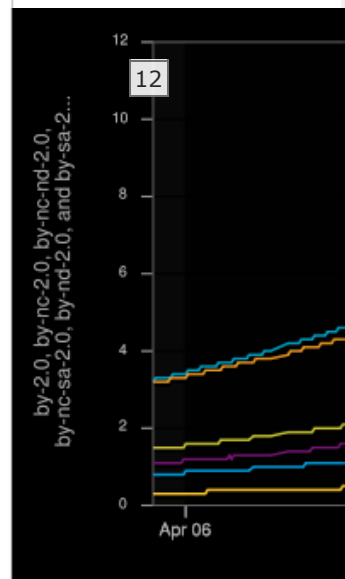


[HTTP://WWW.FLICKR.COM/PHOTOS/WADEY/400836753/](http://www.flickr.com/photos/wadey/400836753/)



## Growth of Creat photos)

Graph Table Cloud Map



Sources: Collected by Jared B (http://rediar.org/jared/blo...)

Swivel uses Creative Commons the purple line (by-2.0) for data set and column image to go for the more restrictive license. Attribution, Non-Commercial, No Derivatives (by-nc-nd-2.0), the blue line

### Comments (1 - 20 of 37)

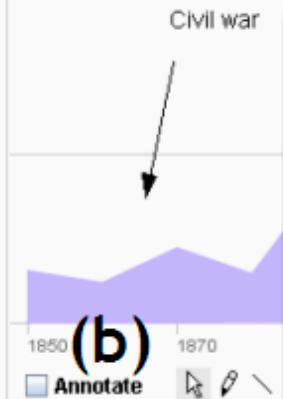
- tim** says  
what's this little cut between july and september by-nc-sa and by-nc-nd curves? ;)  
posted 6 months ago
- brian** says  
i was wondering that too. i assume the data is with a little spec of dirty data.  
posted 6 months ago
- rejon** says  
This is quite cool Brian. Yes, there is info on the which could be used to plot graphs:

Reported Occupations of U.S. Labor Force, 1850-2000 (source: http://ipums.org)

>> military

(a)

all men women % of Work Force



freebase alpha

## Barack Obama

Discuss "Barack Obama"

Hide Empty Fields



image 1 of 1

**Types:** [Person \(People\)](#), [US Senator \(Government\)](#), [US Politician \(Government\)](#), [Author \(Publishing\)](#), [Award Winner \(Award\)](#), [Book Subject \(Publishing\)](#)

**Also known as:** Barack Hussein Obama, Jr.

**Gender:** [Male](#)

**Date of Birth:** Aug 4, 1961

**Place of Birth:** [Honolulu, Hawaii](#)

**Country Of Nationality:** [United States](#)

**Profession:** [Politician](#), [Lawyer](#)

**Religion:** [United Church of Christ](#)

**Parents:** [Ann Dunham](#), [Barack Obama, Sr.](#)

**Children:** [Natasha Obama](#), [Malia Ann Obama](#)

**Siblings:** [double-click to add "Siblings"](#)

**Spouse (or domestic partner):** [Michelle Obama](#) • Oct 18, 1992

**Height:** 1.87 m

**Weight:** [double-click to add "Weight"](#)

comments (5) [New Comment](#) | [View All \(139\)](#)

here are labels where I would have expected big jumps.

by [Martin Sharp](#) on Fri Jul 21, 2006 10:16 AM

well, there was also the cold war right after ww2, which might be part of the reason why there's such a huge jump after the 40s. It is also interesting that there is such a drop between the 70s and the 80s.

by [Julia Hernandez](#) on Fri Jul 21, 2006 11:01 AM

I guess a lot of it has turned to robots, and the industrial complex, as mentioned, though it would be interesting to see the comparison of the fall in military personnel next to the rise in DOD funding for robots and industry

by [Jesse O'Brien](#) on Fri Jul 21, 2006 11:51 AM

I think the jumps have more to do with the economy at large rather than any particular military conflict. Lots of money in conflict has already been spent before the conflict starts.

by [Fred Klein](#) on Wed Aug 2, 2006 10:24 AM

[reply](#)

[Is this military info right?](#)

Keyword search Freebase

[Search](#)

[Home](#)

[Data](#)

[Apps](#)

[Discuss](#)

[Help](#)

Please sign in or register to contribute.

### Page History

**Created by** [Metaweb](#) Oct 22, 2006 10:15am

**Last edited by** [mw\\_prop\\_bot](#) Oct 5, 2007 11:49pm

### Web Link(s)

<http://www.barackobama.com/>

### Employment history

[University of Chicago](#) • [Lecturer](#) • 1993 • 2004

[Miner, Barnhill & Galland](#) • [Associate Attorney](#) • 1993 • 1996

[Sidley Austin](#) • [Associate Attorney](#) • 1988

[Business International Corporation](#) • 1983 • 1984

### Education

[Harvard Law School](#) • 1988 • 1991 • [Juris Doctor](#)



# WHERE COULD THIS GO (PART I)

---

"WITH A COLLABORATIVE SPIRIT, WITH A COLLABORATIVE PLATFORM WHERE PEOPLE CAN UPLOAD DATA, EXPLORE DATA, COMPARE SOLUTIONS, DISCUSS THE RESULTS, BUILD CONSENSUS, WE CAN ... ENGAGE PASSIONATE PEOPLE, LOCAL COMMUNITIES, MEDIA AND THIS WILL RAISE – INCREDIBLY – THE AMOUNT OF PEOPLE WHO CAN UNDERSTAND WHAT IS GOING ON.

AND THIS WOULD HAVE FANTASTIC OUTCOMES: THE ENGAGEMENT OF PEOPLE, ESPECIALLY NEW GENERATIONS; IT WOULD INCREASE KNOWLEDGE, UNLOCK STATISTICS, IMPROVE TRANSPARENCY AND ACCOUNTABILITY OF PUBLIC POLICIES, CHANGE CULTURE, INCREASE NUMERACY, AND IN THE END, IMPROVE DEMOCRACY AND WELFARE."

ENRICO GIOVANNINI, CHIEF STATISTICIAN, OECD. JUNE, 2007



# WHERE THIS COULD GO (PART I)

---





# WHERE THIS COULD GO (PART I)

---





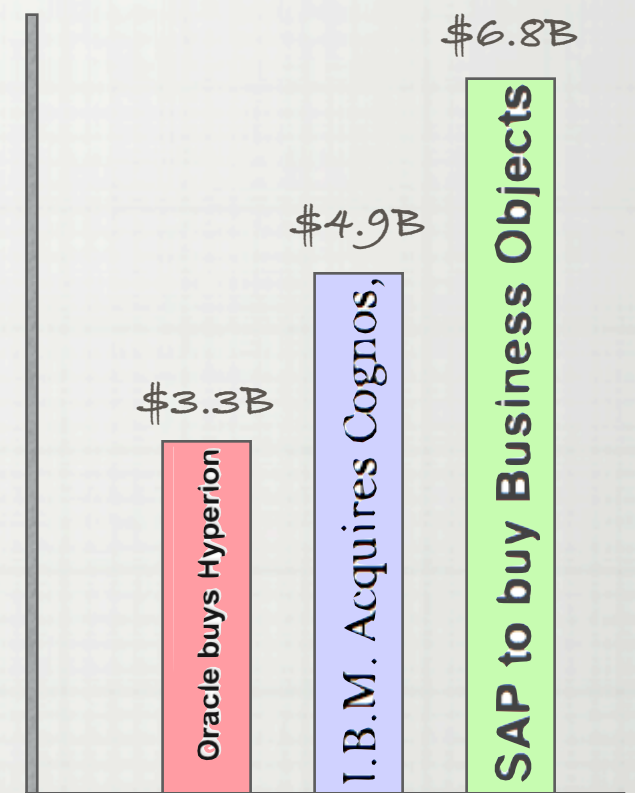
# WHERE COULD THIS GO (PART 2)



[HTTP://WWW.FLICKR.COM/PHOTOS/MATTPICIO/482766923/](http://www.flickr.com/photos/mattpicio/482766923/)



[HTTP://WWW.FLICKR.COM/PHOTOS/TANCREAD/35683040/](http://www.flickr.com/photos/tancread/35683040/)



- ☐ CASUAL DATA USERS VS. THE I.T. FORTRESS
- ☐ "BOTTOM-UP" BUSINESS INTELLIGENCE



# WHERE COULD THIS GO (PART 2)

---

☐ THE QUANTITATIVE INTERNET

☐ INFORMATION? DEFINITELY.

☐ PEOPLE? YES.

☐ COMPUTATION? YOU BET.

☐ BUT SUB-COMMUNITIES,  
WITH OPINIONS, AGENDAS,  
AND SECRETS.

☐ IN A CLOSED LOOP WITH  
PEOPLE.

VALUE:

LIMITED PUBLICATION, SHARING,  
COLLABORATIVE SENSEMAKING.



# CAN THIS WORK?

---

## ☐ EVIDENCE FOR:

☐ WIKIPEDIA

☐ YOUTUBE

☐ FLICKR

☐ FACEBOOK

## ☐ EVIDENCE AGAINST:

☐ CYC

☐ THE SEMANTIC WEB

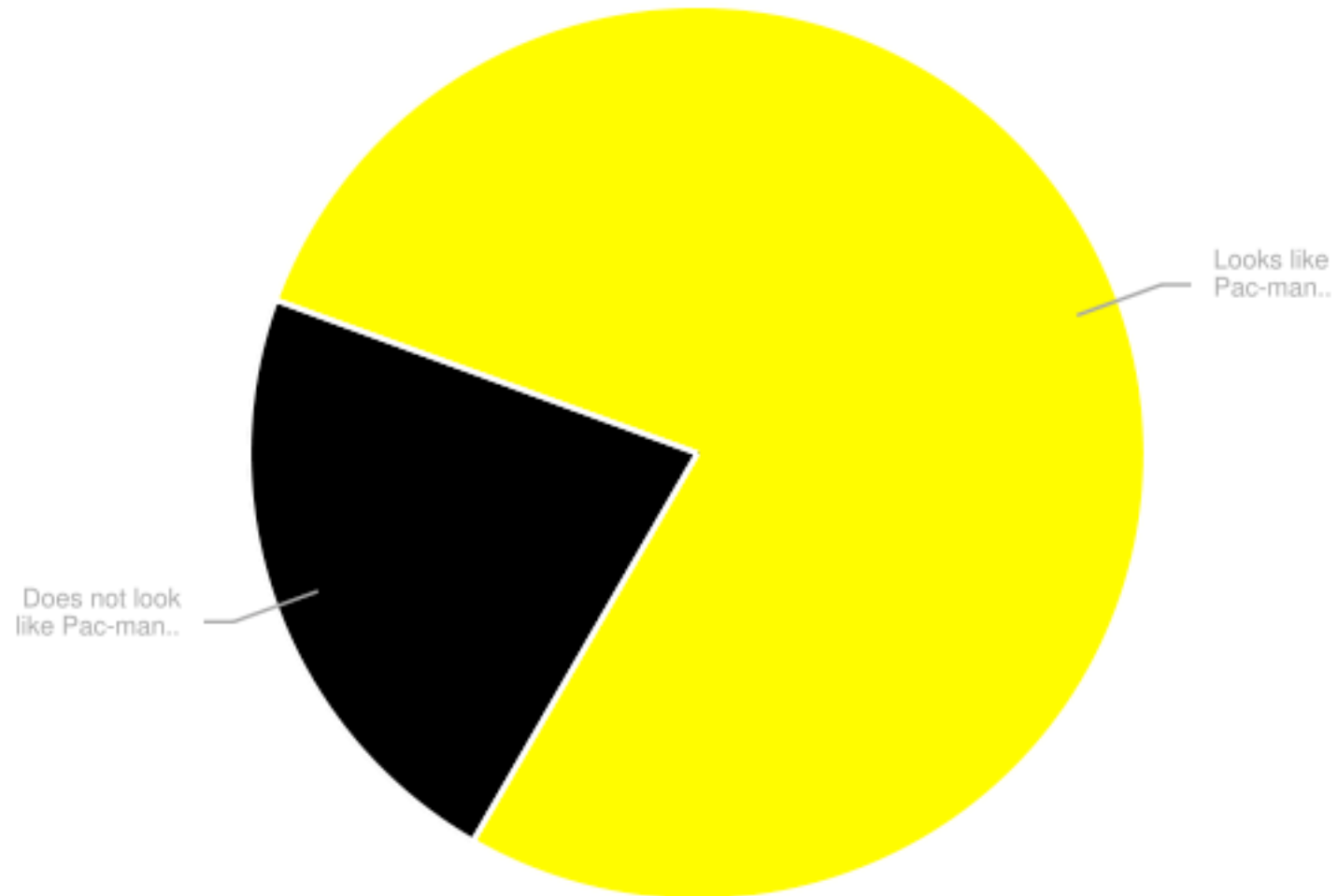
☐ EVERY DATA WAREHOUSE

☐ THE FUN FACTOR



# FUN?

Percentage of chart which looks like Pac-man

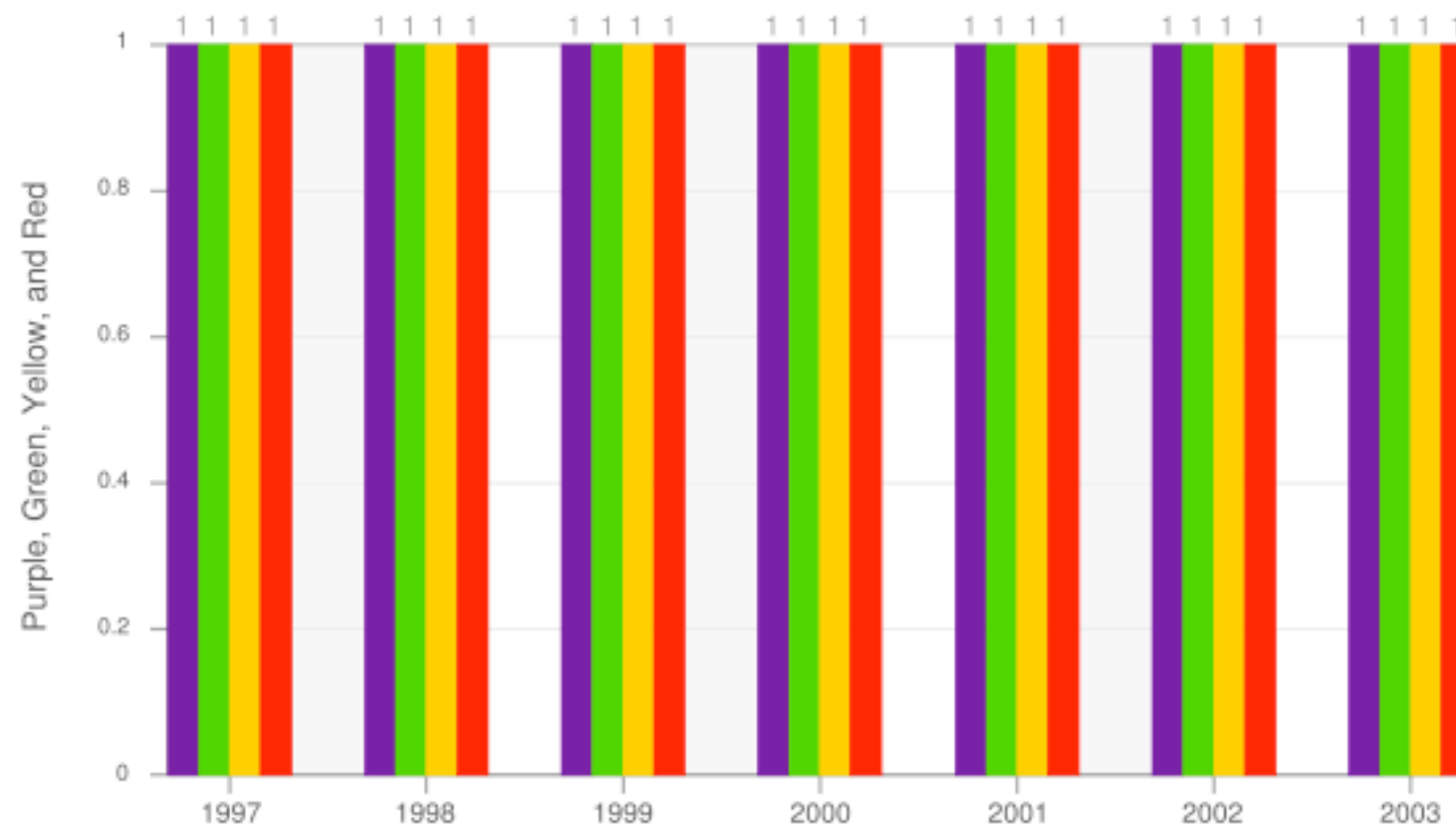




# FUN?

## Population (Teletubbyland - 1997/2003)

- Population by color (Teletubbyland) Purple
- Population by color (Teletubbyland) Green
- Population by color (Teletubbyland) Yellow
- Population by color (Teletubbyland) Red



Teletubbyland (<http://pbskids.org/teletubbies/teletubbyland.html>)



# THE REAL EVIDENCE AGAINST: DATA WAREHOUSING

---

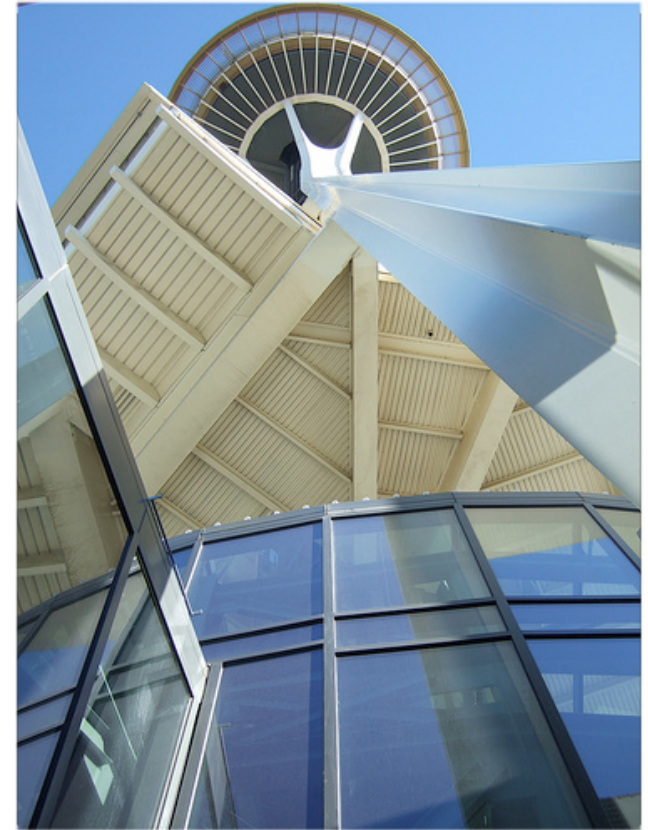
- ☐ DATA INTEGRATION AT CORPORATE SCALES IS A DISASTER
- ☐ MANY OPEN RESEARCH CHALLENGES IN DATA INTEGRATION.



# STRUCTURE & FREEDOM

---

- ☐ WHY HASN'T THIS BEEN A PROBLEM FOR THE WEB?
- ☐ STEPPING BACK FURTHER:
  - ☐ WHAT IS STRUCTURE?
  - ☐ WHAT IS FREEDOM?
  - ☐ WHAT DOES EACH PROVIDE?



[HTTP://FLICKR.COM/PHOTOS/LIFEASART/23479116/](http://flickr.com/photos/lifeasart/23479116/)



[HTTP://WWW.FLICKR.COM/PHOTOS/SCMIKEBURTON/517090571/](http://www.flickr.com/photos/scmikeburton/517090571/)



# A LITTLE HISTORY

---

- ☐ 1959: HANS P. LUHN DESCRIBES KEYWORD IN CONTEXT (KWIC).
- ☐ 1969: EDGAR F. CODD PUBLISHES ON THE RELATIONAL MODEL
- ☐ STRUCTURED/UNSTRUCTURED DICHOTOMY ESTABLISHED EARLY



# THE PILLARS OF MODERN INFOSYSTEMS

---

☐ "UNSTRUCTURED" DOCUMENT  
RETRIEVAL

☐ "STRUCTURED" DATABASES

☐ ASSERTION (FOLLOWING J.  
DERRIDA)

☐ THIS DICHOTOMY IS  
SIMULTANEOUSLY  
MEANINGLESS AND  
USEFUL

☐ LET US REVISIT EACH...



# STRUCTURED DATA: THE PRIMACY OF ACCURACY

---

☐ HIGH VALUE  $\Rightarrow$  PRECISION

☐ DATA MODELING

☐ INTEGRITY CONSTRAINTS

☐ NORMALIZATION

☐ TRANSACTIONS

☐ PRECISION  $\Rightarrow$  ISOLATION

☐ WAREHOUSING &  
FEDERATION

☐ THE CHALLENGES OF DATA  
INTEGRATION



# WE KNOW ABOUT STRUCTURED DATA

---

- CODD'S DATA INDEPENDENCE WAS A REVOLUTION IN SOFTWARE ENGINEERING:
- WHENEVER:  $d\underline{App}/dt \ll d\underline{Env}/dt$
- REQUIRES ENGINEERED STRUCTURE



# UNSTRUCTURED DATA

---

- IN MANY CASES, DATA WASN'T INTENDED FOR AN APP!
  - THEN FOR WHAT?
  - (SOYLENT GREEN IS ...)
    - PEOPLE!
- YET BEHIND ALL HUMAN DISCOURSE IS "DEEP STRUCTURE" (F. DE SAUSSURE)



# UNSTRUCTURED DATA: RELEVANCE & RELATIONSHIPS

---

## ☐ DOCUMENTS: RELEVANCE?

☐ SUBJECTIVE VALUE

☐ SEARCH >> QUERY

☐ PRIMACY OF RANKING

## ☐ INTERNET: SEARCH + SURF

☐ AUTONOMOUS DATA  
GENERATION

☐ EASE OF INTEGRATION

☐ HYPERLINK:  
CONTENT = INTENT



# A KEY METHODOLOGICAL DISTINCTION

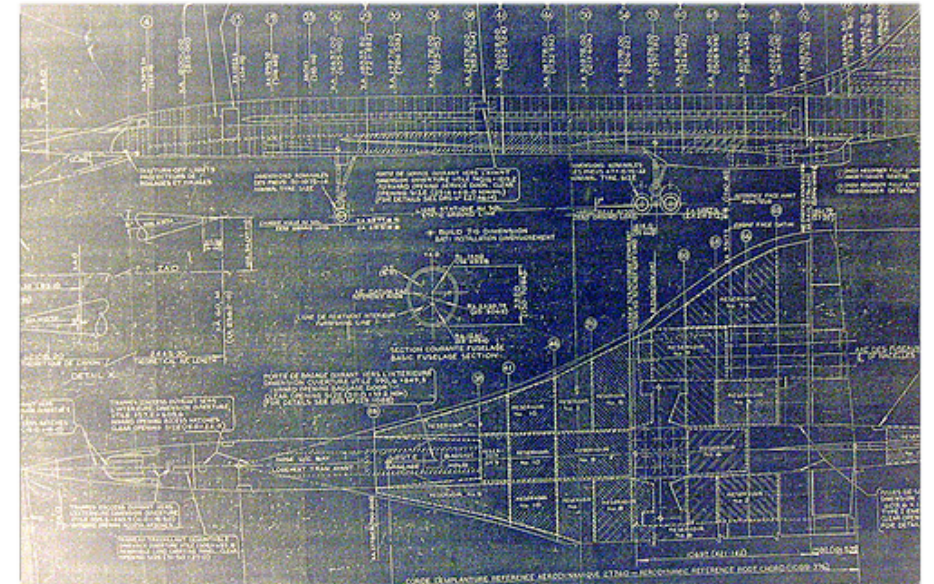
---

☐ ENGINEERED STRUCTURE (DBS)

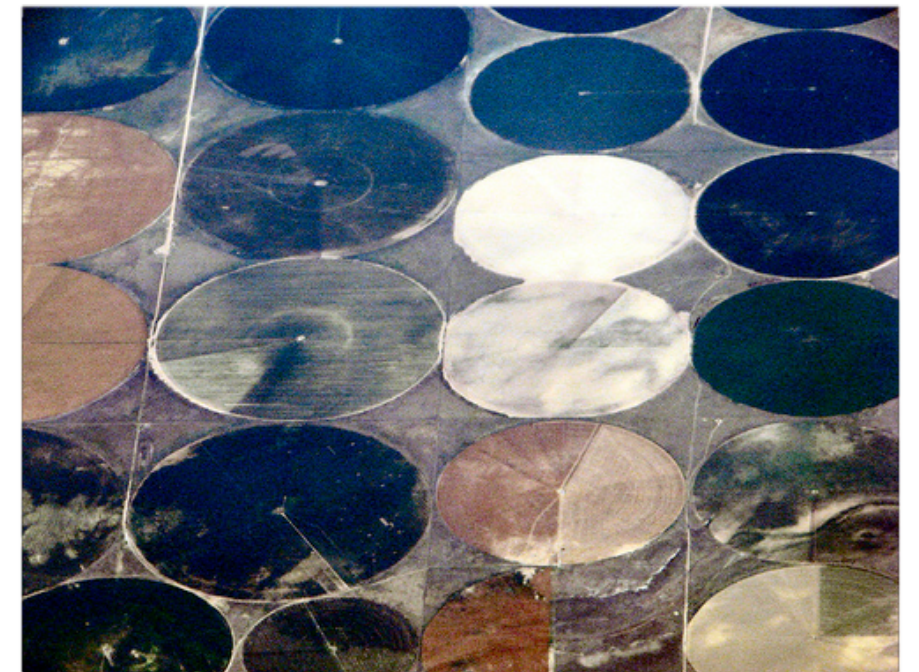
VS.

☐ "FOUND" STRUCTURE (IR)

☐ WE WILL BE RETURNING TO THIS



[HTTP://FLICKR.COM/PHOTOS/SANTINOBCAST/54285870/](http://flickr.com/photos/santinobroadcast/54285870/)



[HTTP://FLICKR.COM/PHOTOS/GOOSAMERPROMISE/636196238/](http://flickr.com/photos/goosamerpromise/636196238/)



# AND YET...

---

- THE DISTINCTIONS BECOME EVER BLURRIER



- ETC.



# WHERE DO WE GO FROM HERE?

---

- ☐ SUBVERT THE STRUCTURED/UNSTRUCTURED DICHOTOMY!?
- ☐ WITHOUT OPPOSITION, TERMS LOSE ALL MEANING!?
- ☐ AND YET, THE METHODOLOGIES MAY STILL BE USEFUL (DERRIDA, AGAIN)
- ☐ WHAT ARE THE METHODOLOGICAL LESSONS?



# A (?) BRIEF (?) DETOUR (?)

---

- A PEEK AT SOME 20TH CENTURY PHILOSOPHY/CRITICISM
- AND 21ST C. POP CULTURE!



# OUTLINE

---

- ☐ SIMULTANEOUS REVOLUTIONS

- ☐ WEB 2.0

- ☐ INDUSTRIAL REVOLUTION  
OF DATA

- ☐ TAPPING THE CONFLUENCE

- ☐ OPPORTUNITY

- ☐ CHALLENGE

- ☐ INSPIRATION FROM A FIELD

- ☐ BRICOLAGE & PLAY

- ☐ EARLY DAYS OF DATA 2.0

- ☐ LIFECYCLE, CHALLENGES

- ☐ TOWARD A RESEARCH AGENDA



# MANY HAVE WORRIED ABOUT STRUCTURE IN THE 20TH C

---

- ☐ DATABASES
  - ☐ STRUCTURED/UNSTRUCTURED
- ☐ PHILOSOPHY, LINGUISTICS, SOCIOLOGY, CRITICISM
  - ☐ STRUCTURALISM/DECONSTRUCTION
- ☐ ART
  - ☐ STRUCTURISM/BRICOLAGE
- ☐ MUSIC
  - ☐ COMPOSITION/IMPROVISATION



# DERRIDA ADDRESSED OUR DICHOTOMY

---

- ☐ (FOLLOWING CLAUDE LÉVI-STRAUSS)
- ☐ CONTRAST THE BRICOLEUR WITH THE ENGINEER
- ☐ THE BRICOLEUR POTTERS ABOUT WITH ODDS-AND-ENDS, PUTS THINGS TOGETHER OUT OF BITS AND PIECES. "TINKERER".
- ☐ THE ENGINEER FORMS STABLE STRUCTURES OUT OF "WHOLE CLOTH"

J. DERRIDA, "STRUCTURE, SIGN AND PLAY IN THE DISCOURSE OF THE HUMAN SCIENCES", 1966



# BRICOLEUR/ENGINEER

---

## ☐ BRICOLAGE

- ☐ JUXTAPOSITION WITHOUT  
REQUIRING RATIONALITY
- ☐ ENABLES WHAT DERRIDA  
CALLS "PLAY"
- ☐ ADDRESSING & AFFIRMING  
PROVISIONAL TRUTHS

## ☐ ENGINEERING

- ☐ STABLE STRUCTURES  
WITH LITTLE OR NO "PLAY"
- ☐ ENGINEER MUST BE AT  
CENTER OF HIS DISCOURSE
- ☐ A GOD-LIKE FIGURE.  
A MYTH.
- ☐ REALLY, ENGAGES IN  
BRICOLAGE AFTER ALL.



# CONFESSION

---

- ☐ THIS TALK IS AN EXERCISE IN BRICOLAGE.
- ☐ SELF-REFERENTIALITY AND RECURSION ARE PART OF THE DECONSTRUCTIONST MINDGAME...



BRICOLAGE: DATA AT PLAY

JOE HELLERSTEIN, UC BERKELEY



# IF THE ENGINEER IS REALLY A BRICOLEUR...

---

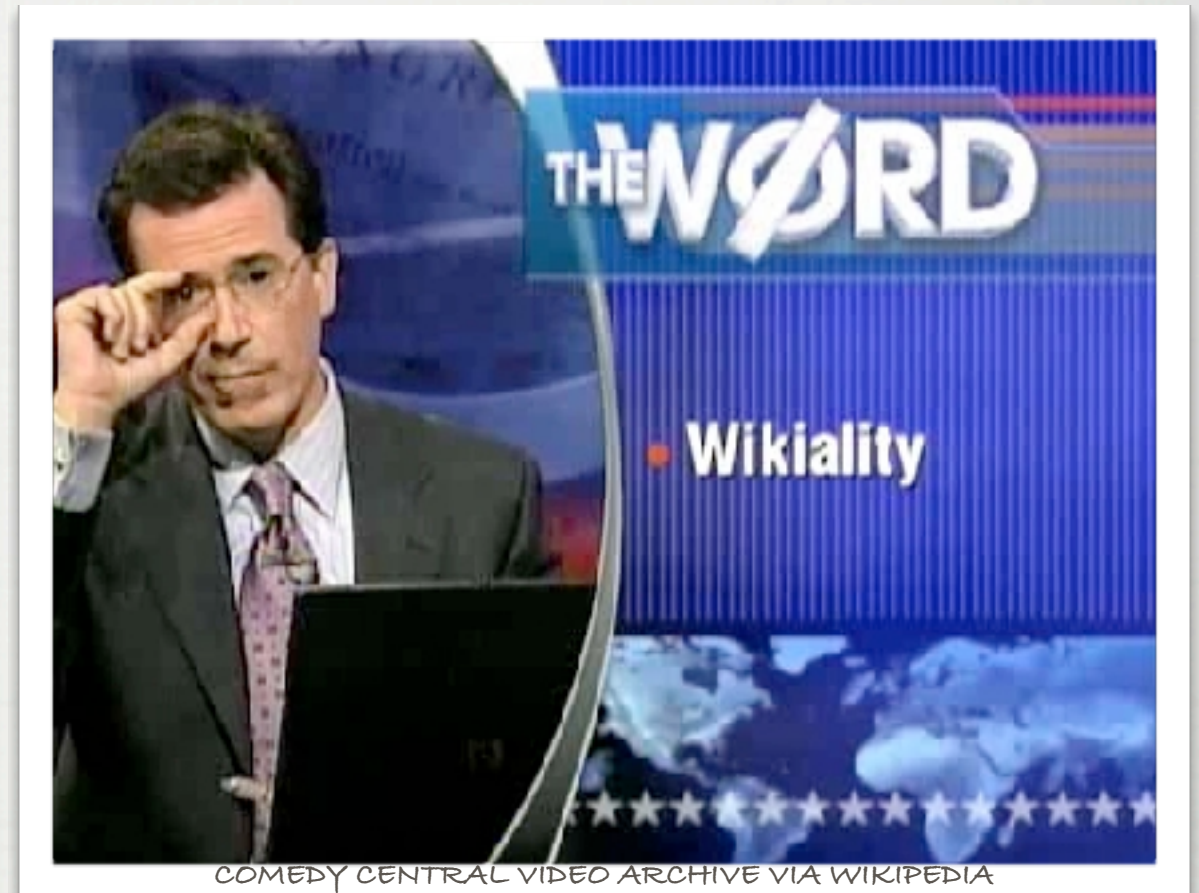
- ☐ THIS SUBVERTS THE DICHOTOMY BETWEEN ENGINEERING/BRICOLAGE
- ☐ JUST AS WE SAW WITH STRUCTURED/UNSTRUCTURED
- ☐ BUT THE DERRIDA RESPONSE IS TO AFFIRM THE PLAY IN THIS FALSE DICHOTOMY
- ☐ RATHER THAN MOURN THE LOSS OF SIMPLICITY



# 21ST C. POPULAR CULTURE

---

- ☐ STEVEN COLBERT'S WIKIALITY
- ☐ TOGETHER "WE CAN ALL CREATE A REALITY THAT WE ALL CAN AGREE ON; THE REALITY THAT WE JUST AGREED ON."
- ☐ "DEFINITIONS WILL WELCOME US AS LIBERATORS"
- ☐ DERRIDA'S "PROVISIONAL TRUTHS"!



... WITH THANKS TO  
PEDRO DEROSE,  
ANHAI DOAN, PHIL  
BOHANNON



# THAT'S ALL VERY NICE...

---

- ☐ ... AND IT MAKES SENSE FOR WIKIPEDIA
- ☐ BUT HOW DOES ONE PLAY WITH DATA?
- ☐ AND HOW DOES COMMUNITY FIT IT?
  - ☐ (SEE CLAUDE LÉVI-STRAUSS FOR REAL ANSWERS!)
- ☐ SOME EXAMPLES FROM THE FIELD
  - ☐ AND ATTENDING FOLLOW-ON QUESTIONS



# OUTLINE

---

- ☐ SIMULTANEOUS REVOLUTIONS
  - ☐ WEB 2.0
  - ☐ INDUSTRIAL REVOLUTION OF DATA
- ☐ TAPPING THE CONFLUENCE
  - ☐ OPPORTUNITY
  - ☐ CHALLENGE
- ☐ INSPIRATION FROM A FIELD
  - ☐ BRICOLAGE & PLAY
  - ☐ EARLY DAYS OF DATA 2.0
  - ☐ LIFECYCLE, CHALLENGES
- ☐ TOWARD A RESEARCH AGENDA



# 3 STAGES

---

LIBERATING DATA (UPLOAD/IMPORT)

EXPLOITING AGGREGATION

LEVERAGING COMMUNITY

WITH A BORROW FROM "POTTER'S WHEEL"  
(RAMAN/HELLERSTEIN VLDB 2001)



# LIBERATING USER DATA: STRUCTURE

Goal 2. Achieve Universal Primary Education															
Goal 2. Achieve Universal Primary Education															
Asian Development Bank (ADB) - Key Indicators 2006 (www.adb.org/statistics)															
Target 3															
Ensure that, by 2015, children everywhere, boys and girls alike, will be able to complete a full course of primary schooling															
6. Net Enrollment Ratio in Primary Education (%)															
DMC	Total <sup>a</sup>			Girls			Boys			Total <sup>c</sup>			Girls		
	1990 <sup>b</sup>	2000 <sup>c</sup>	Latest Year	1990	2000 <sup>c</sup>	Latest Year <sup>d</sup>	1990	2000 <sup>c</sup>	Latest Year <sup>d</sup>	1990 <sup>f</sup>	2000 <sup>g</sup>	Latest Year	1990 <sup>f</sup>	2000 <sup>g</sup>	
East Asia															
China, People's Rep. of	97	...	99 #####	95	...	...	99	...	...	86	...	98 (2001)	78	...	...
Hong Kong, China	...	93	93 #####	...	91	90	...	95	96	100	...	100 (2003)	...	...	...
Korea, Rep. of	100	97	99 #####	100	97	99	99	96	100	100	100	98 (2004)	100	100	...
Mongolia	90	91	84 #####	91	93	85	89	89	84	...	...	...	...	...	...
Taipei, China	98	99	98 #####	...	...	...	...	...	...	...	...	...	...	...	...
Southeast Asia															
Brunei Darussalam	92	...	...	91	...	...	93	...	...	95	93	...	...	...	99
Cambodia	69	91	98 #####	63	87	96	75	95	100	49	63	60 (2003)	42	63	...
Indonesia	97	94	94 #####	95	92	93	99	92	93	84	95	92 (2003)	...	100	...
Lao PDR	63	81	84 #####	58	78	82	67	85	87	53	53	63 (2003)	50	54	...
Malaysia	94	97	93 #####	94	97	93	94	97	93	97	...	98 (2002)	98	...	...
Myanmar	98	82	90 #####	96	82	91	100	82	89	...	55	70 (2004)	...	53	...
Philippines	97	93	94 #####	96	93	95	97	92	93	75	79	76 (2003)	...	83	...
Singapore	96	...	...	96	...	...	97	...	...	100	...	...	100	...	...
Thailand	76	80	85 #####	75	78	84	77	82	87	...	...	...	...	...	...
Viet Nam	90	95	93 #####	86	92	...	94	97	...	...	86	87 (2002)	...	83	...
South Asia															
Bangladesh	71	89	94 #####	66	90	95	76	89	92	...	66	65 (2003)	...	68	...
Bhutan	14	...	...	...	...	...	...	...	...	82	91	...	84	93	...
India	...	82	90 #####	...	73	87	...	89	92	59	59	79 (2003)	55	59	...



# LIBERATING USER DATA: STRUCTURE

Microsoft Excel - Swiveled.xls

File Edit View Insert Format Tools Data Window Help

100% Arial 10 B I U

Upload to swivel.com Replace...

K29

	A	B	C	D	E	F	G	H
1	Region	Country	Target	Metric Name	Subpopulation	Year	Value	
2	East Asia	China, People's Rep. of	Target 3	6. Net Enrollment Ratio in F	Total a	1990 b	97.4	
3	East Asia	China, People's Rep. of	Target 3	6. Net Enrollment Ratio in F	Total a	2000 c	...	
4	East Asia	China, People's Rep. of	Target 3	6. Net Enrollment Ratio in F	Total a	Latest Year	98.5	
5	East Asia	China, People's Rep. of	Target 3	6. Net Enrollment Ratio in F	Total a		-2003	
6	East Asia	China, People's Rep. of	Target 3	6. Net Enrollment Ratio in F	Girls	1990	95.3	
7	East Asia	China, People's Rep. of	Target 3	6. Net Enrollment Ratio in F	Girls	2000 c	...	
8	East Asia	China, People's Rep. of	Target 3	6. Net Enrollment Ratio in F	Girls	Latest Year d	...	
9	East Asia	China, People's Rep. of	Target 3	6. Net Enrollment Ratio in F	Boys	1990	99.4	
10	East Asia	China, People's Rep. of	Target 3	6. Net Enrollment Ratio in F	Boys	2000 c	...	
11	East Asia	China, People's Rep. of	Target 3	6. Net Enrollment Ratio in F	Boys	Latest Year d	...	
12	East Asia	China, People's Rep. of	Target 3 (Cont.)	7. Proportion of Pupils Star	Total e	1990 f	86	
13	East Asia	China, People's Rep. of	Target 3 (Cont.)	7. Proportion of Pupils Star	Total e	2000 g	...	
14	East Asia	China, People's Rep. of	Target 3 (Cont.)	7. Proportion of Pupils Star	Total e	Latest Year	98	
15	East Asia	China, People's Rep. of	Target 3 (Cont.)	7. Proportion of Pupils Star	Total e		-2001	
16	East Asia	China, People's Rep. of	Target 3 (Cont.)	7. Proportion of Pupils Star	Girls	1990 f	78.3	
17	East Asia	China, People's Rep. of	Target 3 (Cont.)	7. Proportion of Pupils Star	Girls	2000 g	...	
18	East Asia	China, People's Rep. of	Target 3 (Cont.)	7. Proportion of Pupils Star	Girls	Latest Year d	...	
19	East Asia	China, People's Rep. of	Target 3 (Cont.)	7. Proportion of Pupils Star	Boys	1990 f	57.6	
20	East Asia	China, People's Rep. of	Target 3 (Cont.)	7. Proportion of Pupils Star	Boys	2000 g	...	
21	East Asia	China, People's Rep. of	Target 3 (Cont.)	7. Proportion of Pupils Star	Boys	Latest Year d	...	
22	East Asia	China, People's Rep. of	Target 3 (Cont.)	7. Proportion of Pupils Star	Boys			
23	East Asia	China, People's Rep. of	Target 3 (Cont.)	8. Literacy Rate of 15-24 Y	Total	1990	95.3	
24	East Asia	China, People's Rep. of	Target 3 (Cont.)	8. Literacy Rate of 15-24 Y	Total	2000-2004 i	98.9	
25	East Asia	China, People's Rep. of	Target 3 (Cont.)	8. Literacy Rate of 15-24 Y	Total	2000-2004	98.9	



# LIBERATING USER DATA: STRUCTURE CHALLENGES

---

- ☐ A SIMPLE STRUCTURAL ALGEBRA
  - ☐ ACCOMODATES EXTRA-RELATIONAL OPERATIONS
- ☐ VISUALLY INTUITIVE
  - ☐ AFFORDANCES ENCOURAGING (RECOGNIZING) "GOOD" FORMATS
  - ☐ TRANSPARENCY OF CAUSE AND EFFECT
- ☐ ROLE OF AUTOMATION?





# CONTENT CHALLENGES

☐ DATA FORMATTING

☐ STRUCTURE AT THE CELL LEVEL

☐ DATA CLEANING

☐ ENTITY RESOLUTION

☐ OUTLIER DETECTION

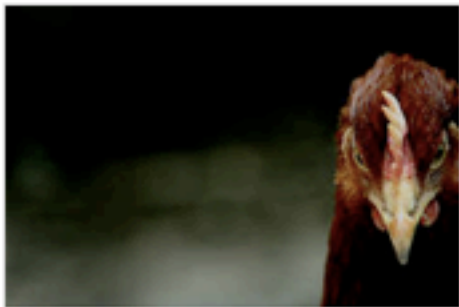
F	
Year	V
1990 b	
2000 c	...
Latest Year	
	1990
2000 c	...
Latest Year d	...
	1990
2000 c	...
Latest Year d	...
1990 f	
2000 g	...
Latest Year	
1990 f	
2000 g	...
Latest Year d	...
1990 f	
2000 g	...
Latest Year d	...
	1990
2000-2004 i	
2000-2004	



# TYPES: A PIECE OF THE PUZZLE

## Bird flu in humans and poultry by country

**Overview** Table Columns Comments



For more info, please visit:  
<http://un-influenza.org/>

Updated as of September 10, 2007

### Source

[UNSCIC](#)

### Summary

Rows **43**

Columns **4**

### Categories

[Health](#) [Science](#) [Technology](#)

### Tags

[OIE](#) [afghanistan](#) [albania](#) [avian](#)  
[azerbaijan](#) [bird](#) [cambodia](#)

### Data Summary

Showing last 6 rows

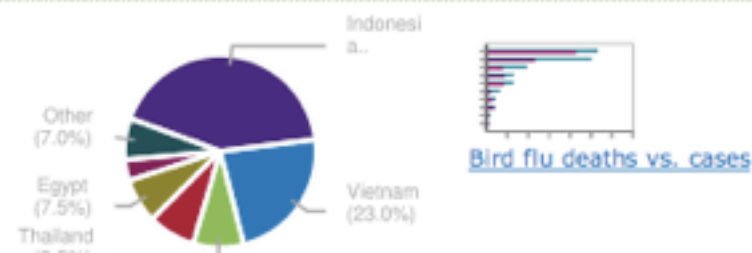
Country	Poultry outbreaks
<a href="#">Thailand</a>	1,137
<a href="#">Togo</a>	1
<a href="#">Turkey</a>	212
<a href="#">Ukraine</a>	40
<a href="#">United Kingdom</a>	1
<a href="#">Vietnam</a>	2,406
<a href="#">more...</a>	<a href="#">more...</a>

[Download to spreadsheet](#)

### Recent Comments

No one's commented yet. Have something to say?

### Popular Graphs



## french country names

**Overview** Table Columns Comments



No description provided

### Source

OECD

### Summary

Rows **27**

Columns **1**

### Data Summary

Showing last 6 rows

[Espagne](#)  
[Suède](#)  
[Royaume-Uni](#)  
[États-Unis](#)  
[Total OCDE](#)  
[Slovénie](#)  
[more...](#)

[Download to spreadsheet](#)

### Recent Comments

No one's commented yet. Have something to say?

### Popular Graphs

ISO Country Codes - Swivel

[http://www.swivel.com/data\\_sets/show/1006933](http://www.swivel.com/data_sets/show/1006933)

[Google](#)



# MDL TYPE INDUCTION

---

- ☐ BEST TYPE = BEST COMPRESSION

$$dl(col_{type}) = matches \log(|type|) + 8 \sum_i mismatch_i.len$$

- ☐ BALANCES AGAINST OVERFITTING
- ☐ WORKS FOR OPAQUE TYPES
- ☐ CHALLENGES
  - ☐ NON-CATEGORICAL TYPES
  - ☐ COMPOSITE TYPES
  - ☐ LOTS AND LOTS OF TYPES



# 3 STAGES

---

LIBERATING DATA (UPLOAD/IMPORT)

EXPLOITING AGGREGATION

LEVERAGING COMMUNITY



# EXPLOITING AGGREGATION: GRAPHSCAPE

Search Engines – Swivel


◀ ▶ ↺ +

http://www.swivel.com/data\_sets/show/1000489

🔍

News (701) ▾ Swivel HQ ▾ RoR ▾ Excel ▾ AC Transit: H P2 ▾ Ruby on AI

Search Engines – Swivel Swivel Mail Flickr Photo Download: ...

 **Swivel**  
*preview*


New! Swivel Toolbar for Excel Confectionary Blog Help Feedback Sign Out (


Home Graphs **Data** People Groups My Stuff Upload


Search


## Search Engines


Overview **Table** Columns Comments


 Google


 Yahoo


 Amazon.com

 Creative Commons

 Dictionary.com

 eBay

 Firefox Extensions

 Google Desktop

### Data Summary

Showing last 6 rows and first 4 columns

Month	Google	MSN/Microsoft	Time Warner	Yahoo!	Ask
Apr 2006	43.1%	12.9%	6.9%	28.0%	5.8%
May 2006	44.1%	12.9%	6.7%	27.9%	5.3%
Jun 2006	44.7%	12.8%	5.6%	28.5%	5.1%
Jul 2006	43.7%	12.8%	5.9%	28.8%	5.4%
Aug 2006	44.1%	12.5%	5.6%	28.7%	5.5%
Sep 2006	45.1%	11.9%	5.6%	28.1%	5.8%
<a href="#">more...</a>	<a href="#">more...</a>	<a href="#">more...</a>	<a href="#">more...</a>	<a href="#">more...</a>	<a href="#">more...</a>

[Download to spreadsheet](#)[See entire table >](#)



# SWIVEL PREVIEW

---

- ☐ GRAPHS ARE NOT CREATED, THEY EXIST

- ☐ HAVE INTRINSIC IDENTITY

- ☐ EASILY SHARED

MACKINLAY'S PHD

- ☐ DECLARATIVE: MALLEABLE/COMPOSABLE

- ☐ NATURALLY KNITS GRAPHS INTO THE WEB

- ☐ INDEPENDENT OF IMAGE FORMATS, ETC.

- ☐ THIS WILL BE KEY

- ☐ HIGHLIGHTS MINING OPPORTUNITIES



# A SIMPLE GRAPHSCAPE

- ☐ FEATURES OF AN EXCEL GRAPH?
  - ☐ DATA (POINTS AND LABELS)
  - ☐ VISUAL SEMANTICS
    - ☐ COORDINATE SPACE
    - ☐ MARKS
    - ☐ CONNECTIVITY OF MARKS
    - ☐ RELATIONSHIPS BETWEEN MULTIPLE SERIES



# GRAPHSCAPE & TRANSFORMATION

---

- ☐ GIVEN A TRANSFORMATION ALGEBRA
  - ☐ STRUCTURAL TRANSFORMS
  - ☐ RELATIONAL OPERATORS
- ☐ INHERENTLY SPANS MULTIPLE "DATA SETS"
  - ☐ THIS IS GOOD, WE NEED TO GO THERE
- ☐ NEIGHBORHOOD FUNCTION?



# GRAPHSCAPE: WHAT FOR?

---

- ☐ NAVIGATION (INCLUDING CREATION)
- ☐ SEARCH
- ☐ MASHUP
- ☐ DATA CLEANING
- ☐ SCHEMA MINING
- ☐ TREND ANALYSIS, PREDICTION
- ☐ ETC.



# 3 STAGES

---

LIBERATING DATA (UPLOAD/IMPORT)

EXPLOITING AGGREGATION

LEVERAGING COMMUNITY



# GRAPHSCAPE: NOW ADD COMMUNITY

---

- ☐ TAGS
- ☐ COMMENTS & SHOUT-OUTS
- ☐ ANCHOR TEXT (BLOG ENTRIES)
- ☐ SOCIAL NETWORK
- ☐ SEARCHES (DATA & BLING)
- ☐ MASHUPS



# OUTLINE

---

- ☐ SIMULTANEOUS REVOLUTIONS

- ☐ WEB 2.0

- ☐ INDUSTRIAL REVOLUTION  
OF DATA

- ☐ TAPPING THE CONFLUENCE

- ☐ OPPORTUNITY

- ☐ CHALLENGE

- ☐ INSPIRATION FROM A FIELD

- ☐ BRICOLAGE & PLAY

- ☐ EARLY DAYS OF DATA 2.0

- ☐ LIFECYCLE, CHALLENGES

- ☐ TOWARD A RESEARCH AGENDA



# BUILDING BLOCKS

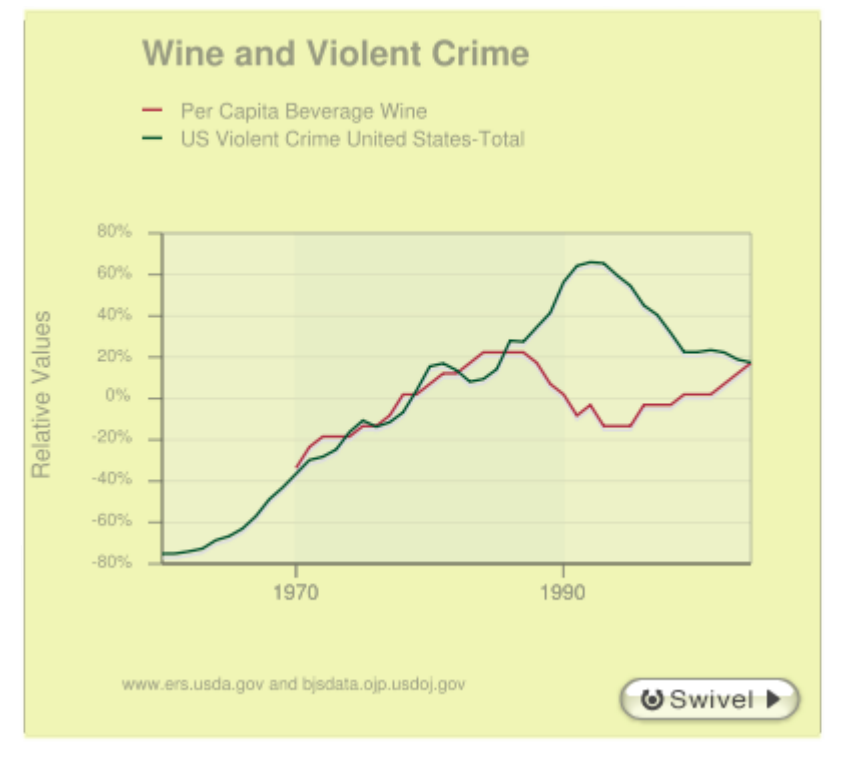
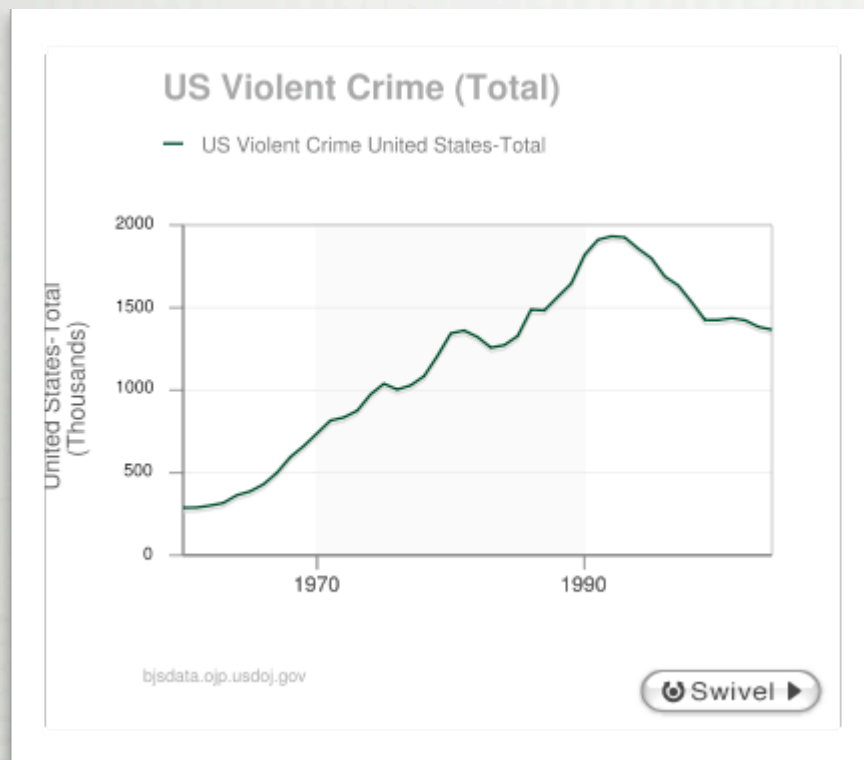
---

- ☐ GRAPHSCAPES
- ☐ COMMUNITY CODEBOOKS & TYPE INDUCTION
- ☐ MINING COLLABORATIVE BEHAVIOR ON VISUALIZATIONS
- ☐ PSEUDO-ENGINEERED WAREHOUSES
- ☐ SUPPORTING MULTIPLE WIKIALITIES
- ☐ .. SEE HEER/AGRAWALA VAST '2006 FOR VIZ DIRECTIONS



# ONE DIRECTION: NUMERACY

☐ KIDS + GOOD TOPICS + WEB 2.0



☐ THE CEO SCENARIO...



# SO MUCH TO DO HERE!

---

- ☐ EMERGING PHENOMENON
- ☐ BUILD IT, STUDY IT, USE IT
- ☐ SOCIAL/TECHNICAL,  
QUANTITATIVE/CREATIVE,  
STRUCTURED/UNSTRUCTURED
- ☐ COME PLAY.....



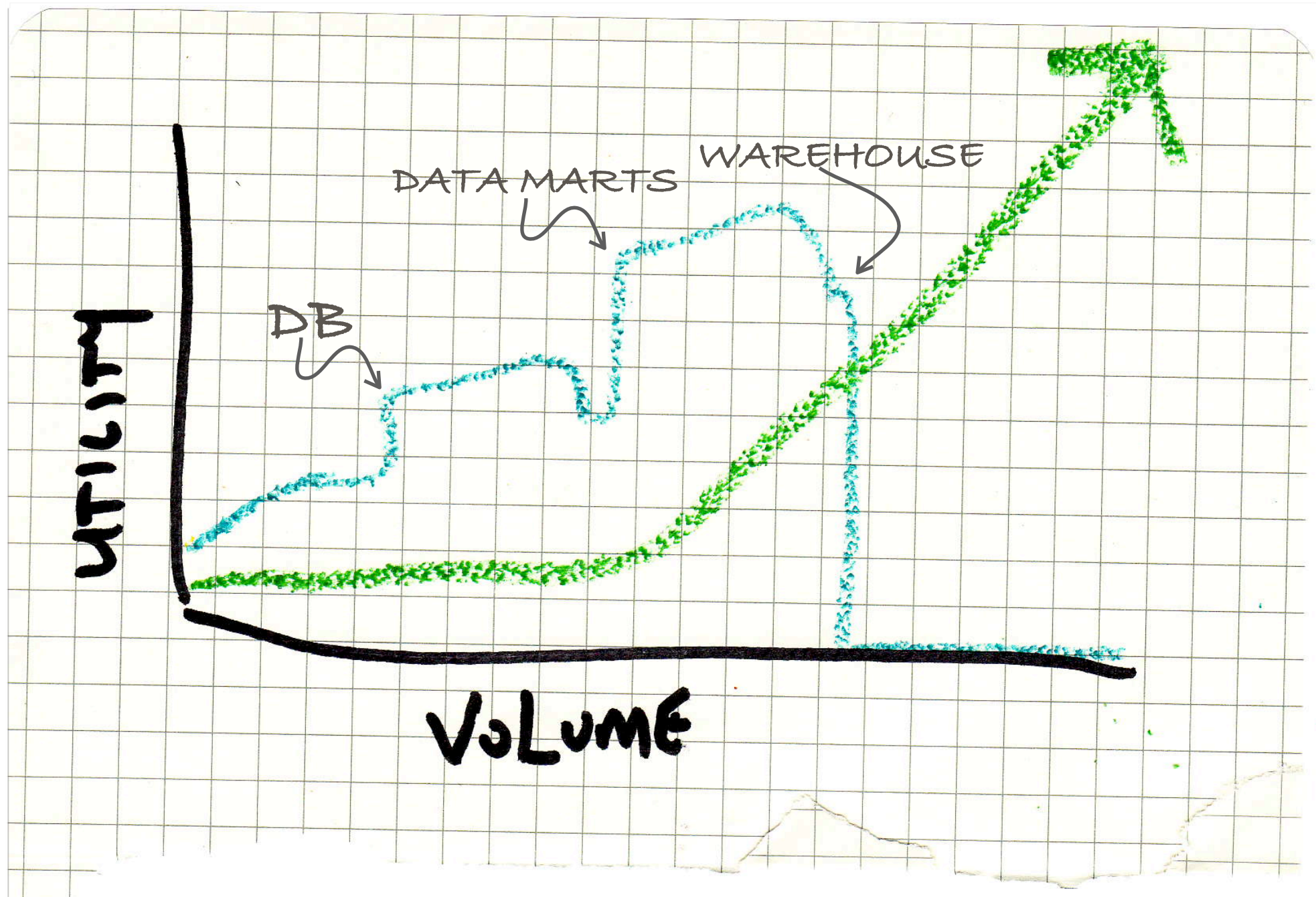
[HTTP://FLICKR.COM/PHOTOS/TIGGYWINKLE/166703632/](http://flickr.com/photos/tiggywinkle/166703632/)



ADDITIONAL SLIDES

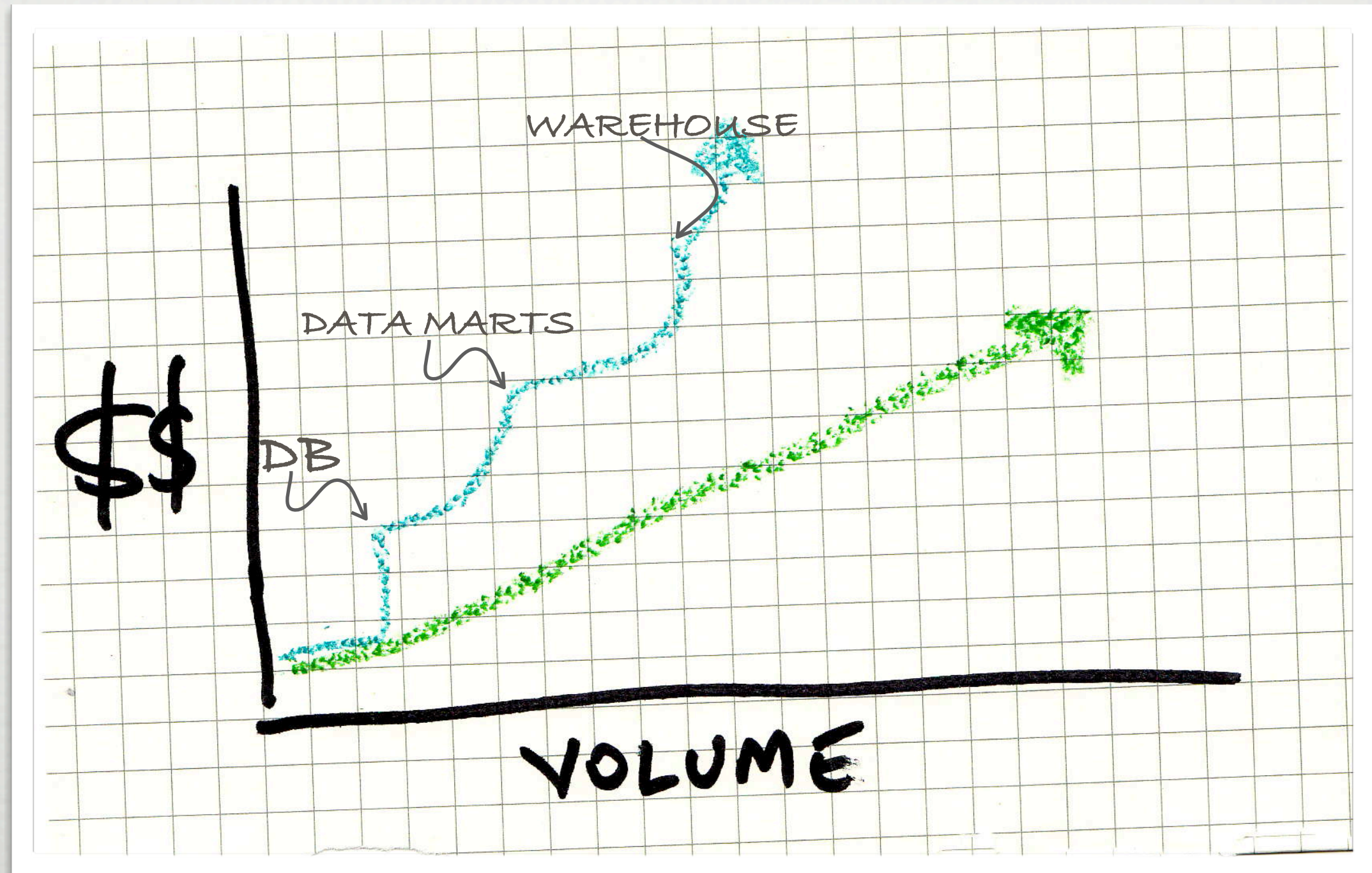


# VISUALIZING THIS SPACE





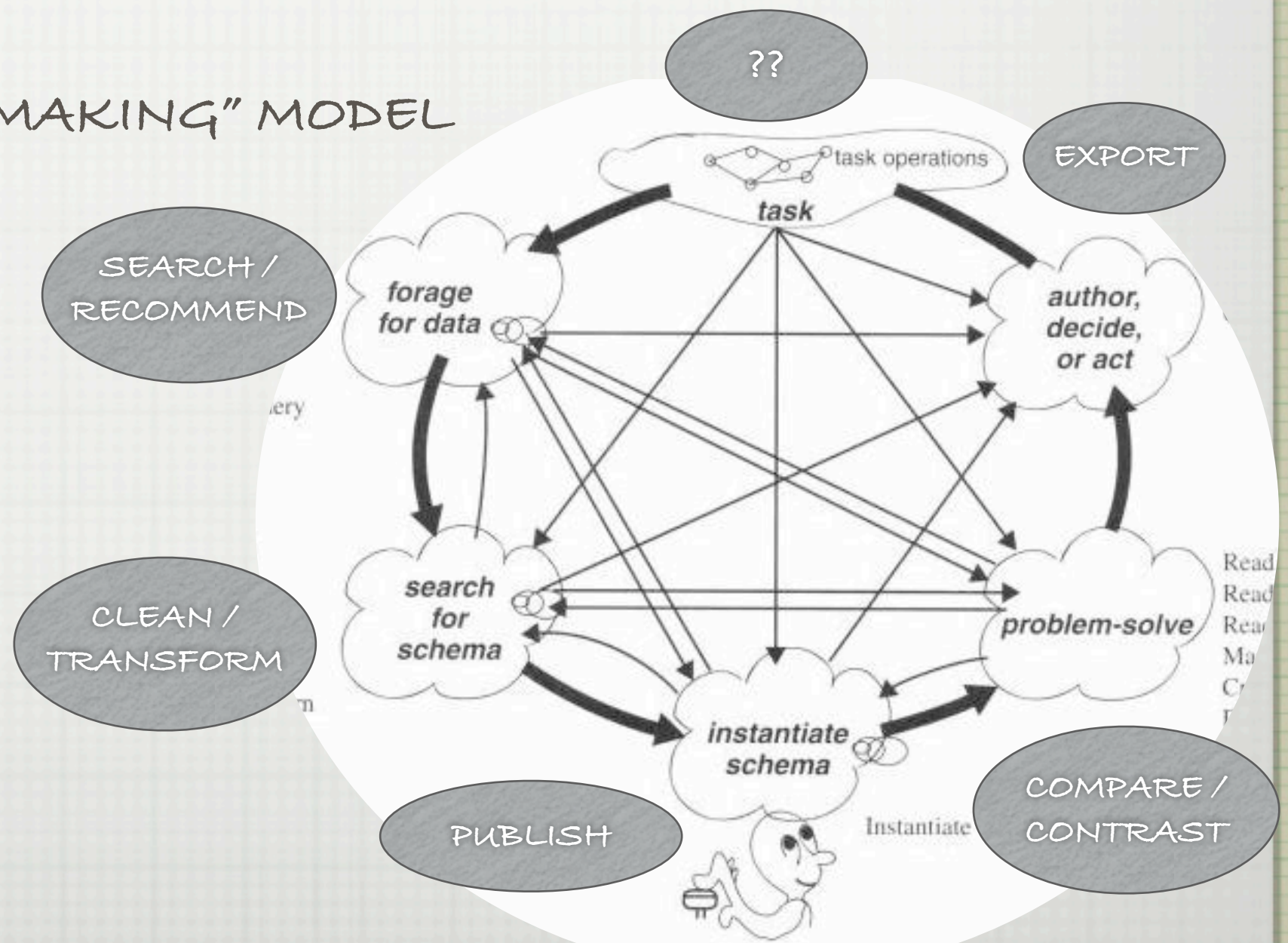
# VISUALIZING THIS SPACE





# PLAYING WITH STRUCTURE

- ☐ LIFECYCLE OF A COLLABORATIVE VISUALIZATION
- ☐ CARD'S "SENSEMAKING" MODEL
- ☐ COMMUNITY?
- ☐ MINING?





# COMMUNITY OPPORTUNITY & CHALLENGE

---

- ☐ COULD CRACK SOME BIG OPEN PROBLEMS
  - ☐ OPTIMISM IN THE WAREHOUSING SPACE
  
- ☐ BUT MANY CHALLENGES ARISE AT SCALE
  - ☐ NOISY USER INPUT (ERRORS, SPAM)
  - ☐ REDUNDANCY AND INCONSISTENCY IN DATA



# ENGAGING TECHNOLOGISTS

---

- ☐ NEW KIND OF CORPUS
  - ☐ BUT NOT JUST SWIVEL: SPREADSHEET SILOS IN LOTS OF ORGANIZATIONS
  - ☐ CHALLENGE PROBLEMS (KDD CUP?)
- ☐ SWIVEL AS A PLATFORM FOR DATA MINING FOLK
  - ☐ HOW DO TECHNOLOGISTS LEVERAGE CORPUS, USERBASE, ?
  - ☐ FUNCTIONALITY OF INTEREST



# AN ASIDE

---

☐ SEMI-STRUCTURED DATA?





# AN ASIDE

---

☐ SEMI-STRUCTURED DATA?

```
<key:text sfa:ID="SFWPFrame-53"
sf:layoutstyle="SFWPLayoutStyle-287">
  <sf:text-storage
sfa:ID="SFWPStorage-1284"
sf:kind="textbox" sf:excl="mstca">
  <sf:stylesheet-ref
sfa:IDREF="SFSStylesheet-32"/>
  <sf:text-body>
    <sf:p
sf:style="SFWPParagraphStyle-418"
sf:list-level="1">This is not semi-
structured.</sf:p>
  </sf:text-body>
</sf:text-storage>
</key:text>
```