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 Wisconsin DBMS Research Home Page

What's Here

- Information about database research and researchers at Wisconsin.
- Information and source code for UW database software, including CORAL, EXODUS, OO7, PARADISE, SHORE, ZetaSim and ZOO.
- Information about UW DBMS mailing lists.
- Pointers to other database information.

UW DBMS Research

THE WEB, 1.0

HYPER-DOCUMENTS

I.E. ... PROSE
THE WEB, 2.0alpha

COMMUNITIES

I.E. PEOPLE!
hmm...

people

WEB 2.0

HTTP://FLICKR.COM/PHOTOS/WORDFREAK/160963805/
HTTP://FLICKR.COM/PHOTOS/DAQUELLAMANERA/162104797/
HTTP://FLICKR.COM/PHOTOS/MYSTERYBEE/1659354500/

INFORMATION

COMPUTATION

oops!
WEB 1.0: INFORMATION? COMPUTATION?

☐ People compose web pages

☐ Computers extract structure and statistics

☐ Benefit: People get better access to web pages

THE NEXT INDUSTRIAL REVOLUTION: DATA

- UPC
- RFID
- GPS
- SensorNets
- Software Logs
-...

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/1108346563/
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) × Computation
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/8890561512/
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?

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/bankruptcystats/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/bankruptcystatistics/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 - 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/statistics.html - 19k - Cached - Similar pages
WHAT SAY THE ELEPHANTS?
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/400836453/
WHERE COULD THIS GO (PART 1)

"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)

Apathetic People

The World’s Data

Indifferent Leaders
WHERE THIS COULD GO (PART I)

Engaged People

The World’s Data

Accountable Leaders
WHERE COULD THIS GO (PART 2)

- Casual Data Users vs. the I.T. Fortress
- "Bottom-up" Business Intelligence
WHERE COULD THIS GO (PART 2)

☐ The Quantitative Internet

☐ People? Yes.

☐ But sub-communities, with opinions, agendas, and secrets.

☐ Information? Definitely.

☐ Computation? You bet.

☐ 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

Looks like Pac-man...

Does not look like Pac-man...

Presentation Zen blog
(http://www.presentationzen.com/presentationzen/2007/03/a_few_weeks_ago.html)
FUN?


- 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/234791161

http://www.flickr.com/photos/scmikeburton/517090571
A LITTLE HISTORY

- 1969: Edgar F. Codd publishes on the relational model
- Structured/Unstructured dichotomy established early
"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 ⇒ Precision
  - Data modeling
  - Integrity constraints
  - Normalization
  - Transactions
- Precision ⇒ Isolation
  - Warehousing & Federation
  - The challenges of data integration
Codd’s data independence was a revolution in software engineering:

Whenever: \( \frac{d\text{App}}{dt} \ll \frac{d\text{Env}}{dt} \)

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

<table>
<thead>
<tr>
<th>Documents: Relevance?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subjective Value</td>
</tr>
<tr>
<td>Search &gt;&gt; Query</td>
</tr>
<tr>
<td>Primacy of Ranking</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Internet: Search + Surf</th>
</tr>
</thead>
<tbody>
<tr>
<td>Autonomous Data Generation</td>
</tr>
<tr>
<td>Ease of Integration</td>
</tr>
<tr>
<td>Hyperlink: Content = Intent</td>
</tr>
</tbody>
</table>
A KEY METHODOLOGICAL DISTINCTION

- ENGINEERED STRUCTURE (DBS)
  vs.
- “FOUND” STRUCTURE (IR)
- WE WILL BE RETURNING TO THIS
AND YET...

THE DISTINCTIONS BECOME EVER BLURRIER

<table>
<thead>
<tr>
<th>I.R.</th>
<th>D.B.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tagged Fields</td>
<td>Full-Text Predicates</td>
</tr>
<tr>
<td>Information Extraction</td>
<td>Ranked SQL</td>
</tr>
<tr>
<td>Map-Reduce</td>
<td>Non-Transactional Update</td>
</tr>
</tbody>
</table>

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 Afield
  ☐ 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
  - STRUCTURISMS/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”

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

Really, engages in bricolage after all.
CONFESSION

☐ This talk is an exercise in Bricolage.

☐ Self-referentiality and recursion are part of the deconstructionist 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 Afield
  - 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**

**Target 3:** Ensure that, by 2015, children everywhere, boys and girls alike, will be able to complete a full course of primary schooling.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>East Asia</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>China, People's Rep. of</td>
<td>97</td>
<td>...</td>
<td>93 (2001)</td>
<td>95</td>
<td>...</td>
<td>95</td>
<td>95</td>
<td>...</td>
<td>95</td>
<td>96</td>
<td>...</td>
<td>95</td>
</tr>
<tr>
<td>Hong Kong, China</td>
<td>...</td>
<td>93</td>
<td>93 (2003)</td>
<td>...</td>
<td>91</td>
<td>90</td>
<td>95</td>
<td>...</td>
<td>95</td>
<td>100</td>
<td>...</td>
<td>100</td>
</tr>
<tr>
<td>Korea, Rep. of</td>
<td>100</td>
<td>97</td>
<td>99 (2004)</td>
<td>100</td>
<td>97</td>
<td>99</td>
<td>98</td>
<td>97</td>
<td>99</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Mongolia</td>
<td>90</td>
<td>91</td>
<td>94 (2004)</td>
<td>91</td>
<td>93</td>
<td>98</td>
<td>90</td>
<td>89</td>
<td>94</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Taiwan, China</td>
<td>98</td>
<td>99</td>
<td>93 (2004)</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td><strong>Southeast Asia</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brunei Darussalam</td>
<td>92</td>
<td>...</td>
<td>91</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>95</td>
<td>...</td>
<td>...</td>
<td>95</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Cambodia</td>
<td>69</td>
<td>91</td>
<td>92 (2004)</td>
<td>86</td>
<td>96</td>
<td>96</td>
<td>87</td>
<td>...</td>
<td>...</td>
<td>53</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Indonesia</td>
<td>97</td>
<td>91</td>
<td>94 (2004)</td>
<td>98</td>
<td>92</td>
<td>98</td>
<td>98</td>
<td>92</td>
<td>93</td>
<td>96</td>
<td>93</td>
<td>93</td>
</tr>
<tr>
<td>Lao PDR</td>
<td>63</td>
<td>91</td>
<td>94 (2004)</td>
<td>81</td>
<td>97</td>
<td>93</td>
<td>85</td>
<td>92</td>
<td>98</td>
<td>53</td>
<td>53</td>
<td>53</td>
</tr>
<tr>
<td>Malaysia</td>
<td>94</td>
<td>97</td>
<td>93 (2004)</td>
<td>94</td>
<td>97</td>
<td>94</td>
<td>94</td>
<td>97</td>
<td>93</td>
<td>96</td>
<td>93</td>
<td>96</td>
</tr>
<tr>
<td>Myanmar</td>
<td>96</td>
<td>92</td>
<td>93 (2004)</td>
<td>91</td>
<td>93</td>
<td>91</td>
<td>90</td>
<td>88</td>
<td>95</td>
<td>96</td>
<td>93</td>
<td>96</td>
</tr>
<tr>
<td>Philippines</td>
<td>97</td>
<td>93</td>
<td>94 (2004)</td>
<td>96</td>
<td>93</td>
<td>96</td>
<td>95</td>
<td>93</td>
<td>93</td>
<td>75</td>
<td>75</td>
<td>75</td>
</tr>
<tr>
<td>Singapore</td>
<td>96</td>
<td>...</td>
<td>96</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>97</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Thailand</td>
<td>76</td>
<td>80</td>
<td>86 (2004)</td>
<td>75</td>
<td>76</td>
<td>84</td>
<td>77</td>
<td>82</td>
<td>87</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Viet Nam</td>
<td>90</td>
<td>95</td>
<td>93 (2004)</td>
<td>86</td>
<td>92</td>
<td>94</td>
<td>94</td>
<td>97</td>
<td>...</td>
<td>...</td>
<td>86</td>
<td>87 (2002)</td>
</tr>
<tr>
<td><strong>South Asia</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bangladesh</td>
<td>71</td>
<td>89</td>
<td>94 (2004)</td>
<td>66</td>
<td>90</td>
<td>89</td>
<td>76</td>
<td>89</td>
<td>92</td>
<td>82</td>
<td>91</td>
<td>92</td>
</tr>
<tr>
<td>Bhutan</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>73</td>
<td>...</td>
<td>...</td>
<td>89</td>
<td>...</td>
<td>...</td>
<td>80</td>
<td>92</td>
</tr>
<tr>
<td>India</td>
<td>14</td>
<td>62</td>
<td>90 (2004)</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>92</td>
<td>92</td>
<td>92</td>
</tr>
</tbody>
</table>

Note: Data reflects enrollment ratios as of the specified years.
<table>
<thead>
<tr>
<th>Region</th>
<th>Country</th>
<th>Target</th>
<th>Metric Name</th>
<th>Subpopulation</th>
<th>Year</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>East Asia</td>
<td>China, People's Rep. of</td>
<td>Target 3</td>
<td>6. Net Enrollment Ratio in F Total</td>
<td></td>
<td>1990 b</td>
<td>97.4</td>
</tr>
<tr>
<td>East Asia</td>
<td>China, People's Rep. of</td>
<td>Target 3</td>
<td>6. Net Enrollment Ratio in F Total</td>
<td></td>
<td>2000 c</td>
<td>...</td>
</tr>
<tr>
<td>East Asia</td>
<td>China, People's Rep. of</td>
<td>Target 3</td>
<td>6. Net Enrollment Ratio in F Total</td>
<td></td>
<td>Latest Year</td>
<td>98.5</td>
</tr>
<tr>
<td>East Asia</td>
<td>China, People's Rep. of</td>
<td>Target 3</td>
<td>6. Net Enrollment Ratio in F Girls</td>
<td></td>
<td>-2003</td>
<td>...</td>
</tr>
<tr>
<td>East Asia</td>
<td>China, People's Rep. of</td>
<td>Target 3</td>
<td>6. Net Enrollment Ratio in F Girls</td>
<td></td>
<td>1990</td>
<td>96.3</td>
</tr>
<tr>
<td>East Asia</td>
<td>China, People's Rep. of</td>
<td>Target 3</td>
<td>6. Net Enrollment Ratio in F Boys</td>
<td></td>
<td>2000 c</td>
<td>...</td>
</tr>
<tr>
<td>East Asia</td>
<td>China, People's Rep. of</td>
<td>Target 3</td>
<td>6. Net Enrollment Ratio in F Boys</td>
<td></td>
<td>Latest Year</td>
<td>...</td>
</tr>
<tr>
<td>East Asia</td>
<td>China, People's Rep. of</td>
<td>Target 3</td>
<td>6. Net Enrollment Ratio in F Boys</td>
<td></td>
<td>1990</td>
<td>99.4</td>
</tr>
<tr>
<td>East Asia</td>
<td>China, People's Rep. of</td>
<td>Target 3</td>
<td>6. Net Enrollment Ratio in F Boys</td>
<td></td>
<td>Latest Year</td>
<td>...</td>
</tr>
<tr>
<td>East Asia</td>
<td>China, People's Rep. of</td>
<td>Target 3</td>
<td>6. Net Enrollment Ratio in F Boys</td>
<td></td>
<td>1990</td>
<td>86.0</td>
</tr>
<tr>
<td>East Asia</td>
<td>China, People's Rep. of</td>
<td>Target 3</td>
<td>6. Net Enrollment Ratio in F Boys</td>
<td></td>
<td>Latest Year</td>
<td>...</td>
</tr>
<tr>
<td>East Asia</td>
<td>China, People's Rep. of</td>
<td>Target 3</td>
<td>6. Net Enrollment Ratio in F Boys</td>
<td></td>
<td>-2001</td>
<td>...</td>
</tr>
<tr>
<td>East Asia</td>
<td>China, People's Rep. of</td>
<td>Target 3</td>
<td>6. Net Enrollment Ratio in F Boys</td>
<td></td>
<td>1990 f</td>
<td>78.3</td>
</tr>
<tr>
<td>East Asia</td>
<td>China, People's Rep. of</td>
<td>Target 3</td>
<td>6. Net Enrollment Ratio in F Boys</td>
<td></td>
<td>Latest Year</td>
<td>...</td>
</tr>
<tr>
<td>East Asia</td>
<td>China, People's Rep. of</td>
<td>Target 3</td>
<td>6. Net Enrollment Ratio in F Boys</td>
<td></td>
<td>1990 f</td>
<td>57.6</td>
</tr>
<tr>
<td>East Asia</td>
<td>China, People's Rep. of</td>
<td>Target 3</td>
<td>6. Net Enrollment Ratio in F Boys</td>
<td></td>
<td>Latest Year</td>
<td>...</td>
</tr>
<tr>
<td>East Asia</td>
<td>China, People's Rep. of</td>
<td>Target 3</td>
<td>6. Net Enrollment Ratio in F Boys</td>
<td></td>
<td>1990 f</td>
<td>95.3</td>
</tr>
<tr>
<td>East Asia</td>
<td>China, People's Rep. of</td>
<td>Target 3</td>
<td>6. Net Enrollment Ratio in F Boys</td>
<td></td>
<td>Latest Year</td>
<td>...</td>
</tr>
</tbody>
</table>
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
## Types: A Piece of the Puzzle

### Bird flu in humans and poultry by country

**Data Summary**

Showing last 6 rows

<table>
<thead>
<tr>
<th>Country</th>
<th>Poultry outbreaks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thailand</td>
<td>1,137</td>
</tr>
<tr>
<td>Togo</td>
<td>1</td>
</tr>
<tr>
<td>Turkey</td>
<td>212</td>
</tr>
<tr>
<td>Ukraine</td>
<td>1</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>0</td>
</tr>
<tr>
<td>Vietnam</td>
<td>2,406</td>
</tr>
</tbody>
</table>

**Source**

UNICEF

**Summary**

Rows: 43
Columns: 4

**Categories**

Health Science Technology

**Tags**

OIE afghanistan albania avian azerbaiania bird cambodia

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

Updated as of September 10, 2007

---

### French country names

**Data Summary**

Showing last 6 rows

<table>
<thead>
<tr>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>Espagne</td>
</tr>
<tr>
<td>Suède</td>
</tr>
<tr>
<td>Royaume-Uni</td>
</tr>
<tr>
<td>États-Unis</td>
</tr>
<tr>
<td>Total OCDE</td>
</tr>
<tr>
<td>Slovénie</td>
</tr>
</tbody>
</table>

**Source**

OECD

**Summary**

Rows: 27
Columns: 1

**Popular Graphs**

- Bird flu deaths vs. cases
- Outbreaks by Country

---

ISO Country Codes – Swivel

http://www.swivel.com/data_sets/show/1006933
MDL TYPE INDUCTION

- Best type = Best compression

\[ dl(col_{\text{type}}) = \text{matches} \times \log(|\text{type}|) + 8 \sum_i \text{mismatch}_i \cdot \text{len} \]

- 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

**Data Summary**

Showing last 6 rows and first 4 columns

<table>
<thead>
<tr>
<th>Month</th>
<th>Google</th>
<th>MSN/Microsoft</th>
<th>Time Warner</th>
<th>Yahoo!</th>
<th>Ask</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apr 2006</td>
<td>43.1%</td>
<td>12.9%</td>
<td>6.9%</td>
<td>28.0%</td>
<td>5.8%</td>
</tr>
<tr>
<td>May 2006</td>
<td>44.1%</td>
<td>12.9%</td>
<td>6.7%</td>
<td>27.9%</td>
<td>5.3%</td>
</tr>
<tr>
<td>Jun 2006</td>
<td>44.7%</td>
<td>12.8%</td>
<td>5.6%</td>
<td>28.5%</td>
<td>5.1%</td>
</tr>
<tr>
<td>Jul 2006</td>
<td>43.7%</td>
<td>12.8%</td>
<td>5.9%</td>
<td>28.8%</td>
<td>5.4%</td>
</tr>
<tr>
<td>Aug 2006</td>
<td>44.1%</td>
<td>12.5%</td>
<td>5.6%</td>
<td>28.7%</td>
<td>5.5%</td>
</tr>
<tr>
<td>Sep 2006</td>
<td>45.1%</td>
<td>11.9%</td>
<td>5.6%</td>
<td>28.1%</td>
<td>5.8%</td>
</tr>
</tbody>
</table>

Download to spreadsheet

See entire table >
GRAPHS ARE NOT CREATED, THEY EXIST

HAVE INTRINSIC IDENTITY

EASILY SHARED

DECLARATIVE: MALLEABLE/COMPOSABLE

NATURALLY KNITS GRAPHS INTO THE WEB

INDEPENDENT OF IMAGE FORMATS, ETC.

THIS WILL BE KEY

HIGHLIGHTS MINING OPPORTUNITIES

MACKINLAY’S PHD
A SIMPLE GRAPHSCAPE

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

HTTP://WWW.FLICKR.COM/PHOTOS/MYHOBOSOUL/419702164/
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
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/
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?

*Ceci n'est pas une pipe.*
AN ASIDE

☐ Semi-structured data?

This is not semi-structured.