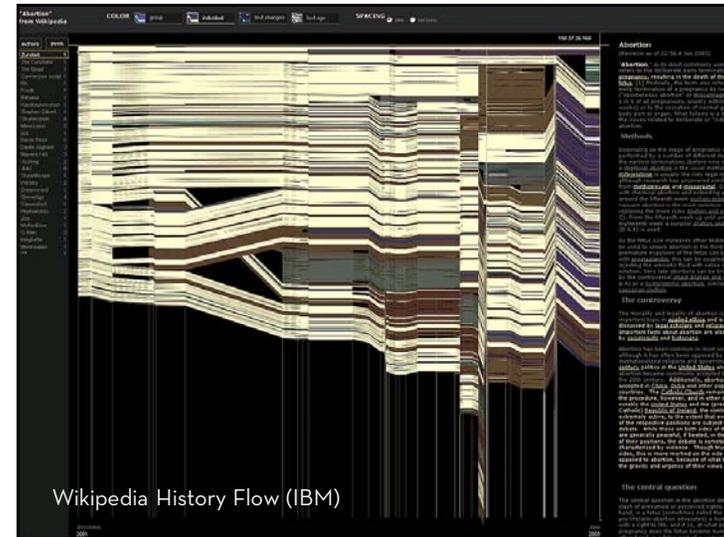
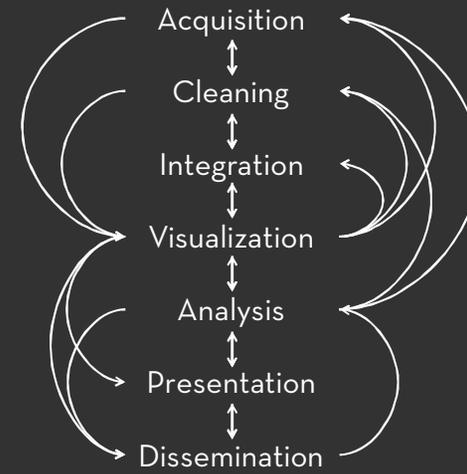


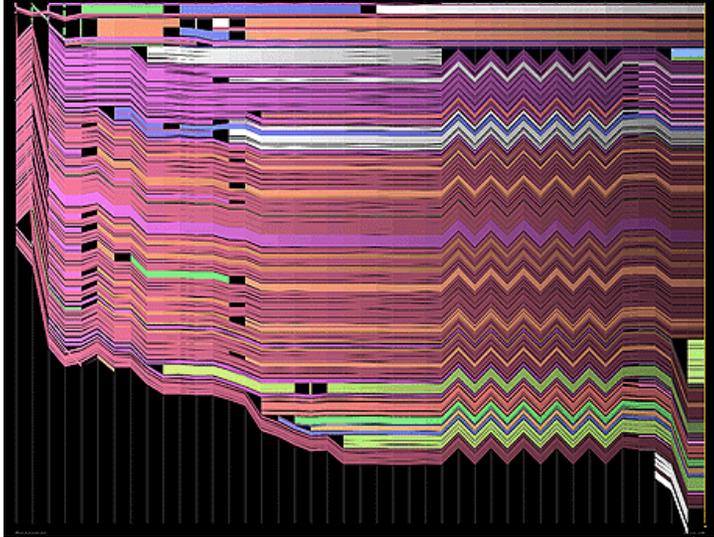
CS448G :: 4 Apr 2011

Data Collection



Jeffrey Heer Stanford University





Information Foraging

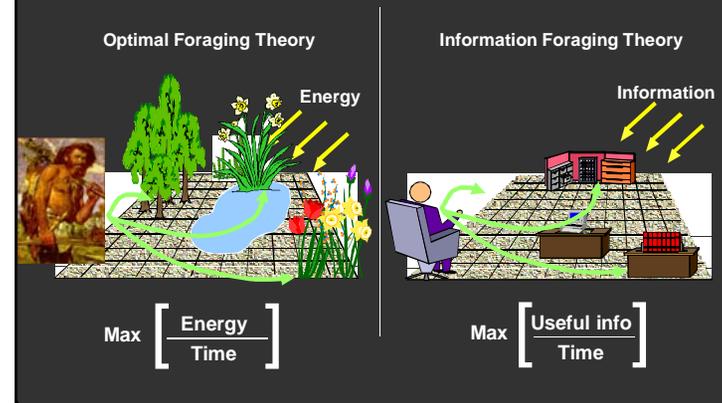
Exercise: Let's Find a Data Set

Think of a topic that you are *passionate* about and/or is of *great societal interest*.

What *pre-existing data* might provide insight?

Now let's try to find it on the web...

Foraging Theory (Pirolli & Card)



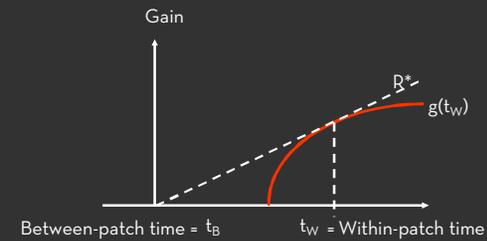
Information Foraging Theory

People are **information rate maximizers** of benefits/costs
 Information has a **cost structure**

$$R = \frac{G}{T_B + T_W} = \frac{\text{Gain}}{T_{\text{Between-patch}} + T_{\text{Within-patch}}}$$

Charnov's Marginal Value Theorem

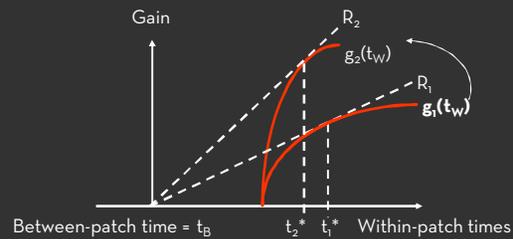
Maximum gain when slope of within-patch gain g = average gain R (tangent in diagram)



Within-Patch Enrichment

Behavior adapts to cost structure of environment.

Example: Better filtering of search hits



Enhancing Data Collection

Insights for Data Collection

Foraging: decrease time between patches, increase information gain within patch.

Ubiquitous networking and mobile devices open up new frontiers for data collection.



Aoki, Honicky, Mainwaring, Myers, Paulos, Subramanian & Woodruff, CHI 2009

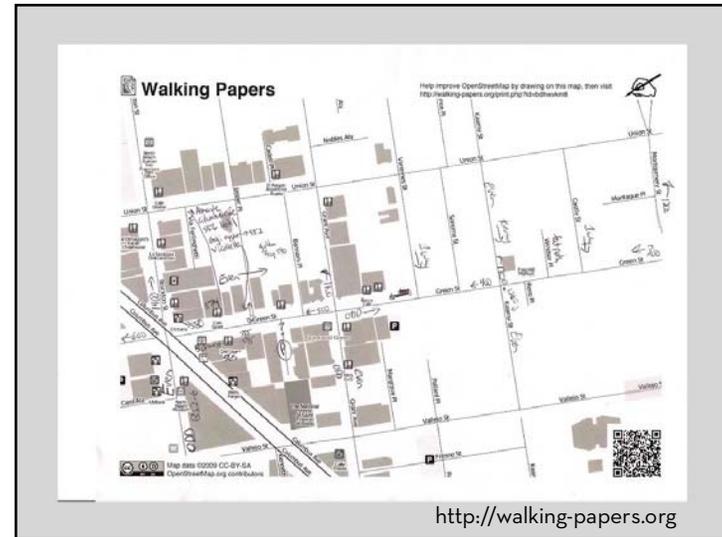
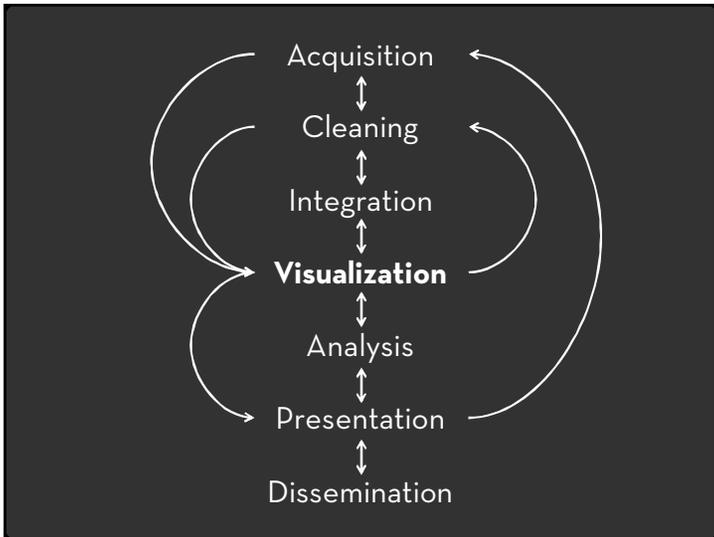
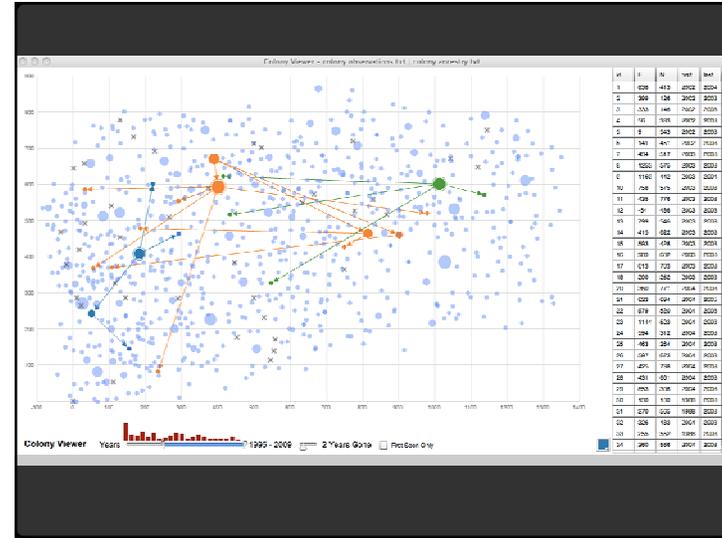
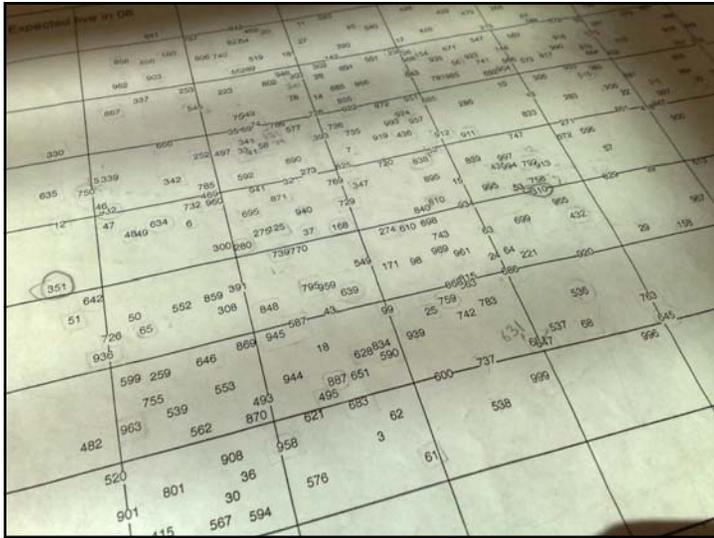
Insights for Data Collection

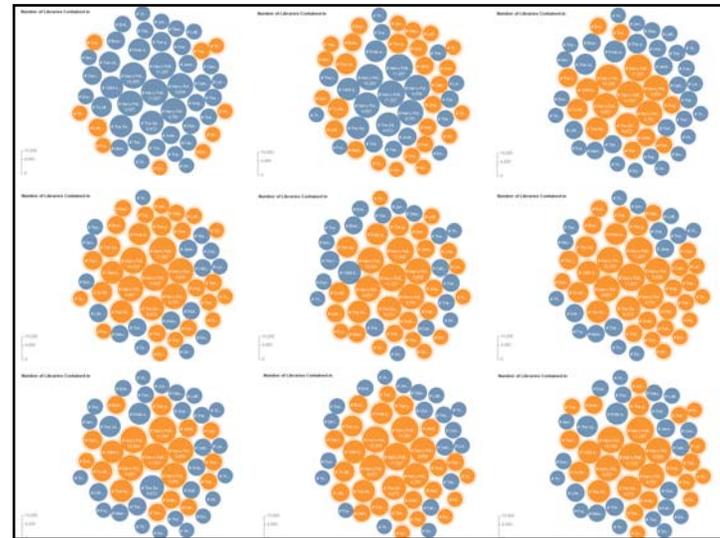
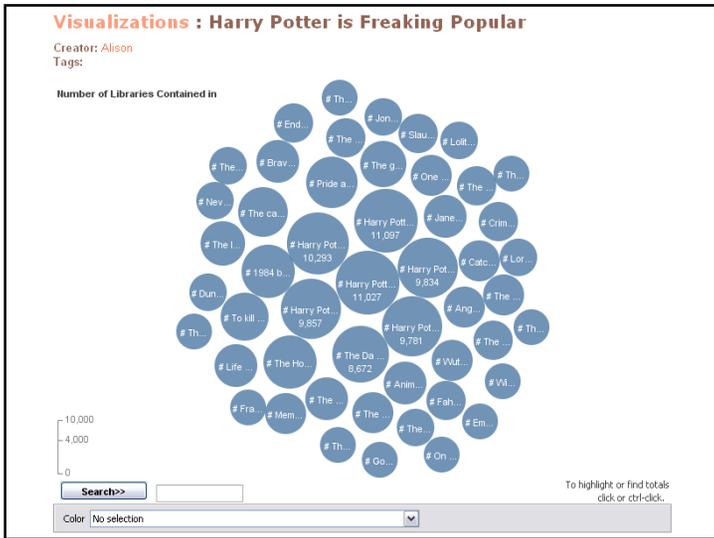
Foraging: decrease time between patches, increase information gain within patch.

Ubiquitous networking and mobile devices open up new frontiers for data collection.

Visualizations are also input devices.







Insights for Data Collection

Foraging: decrease time between patches, increase information gain within patch.

Ubiquitous networking and mobile devices open up new frontiers for data collection.

Visualizations are also input devices.

Think ahead: act at the time of capture to improve subsequent cleaning & analysis.

The challenges of **data documentation**:
 creating, accessing, browsing, ...

IPUMS USA

Home Select Data FAQ Contact Us Login

Data Cart
 Your data extract
 0 variables
 0 samples

Variables Select Options
 Personal A-Z Search Samples Help

Race, Ethnicity, and Ancestry variables - PERSON [202]

Variable	Variable Label	Type	Code	2009	2008	2007	2006	2005	2004	2003	2002	2001	2000	1990	1980	1970	1960	1950	1940	1930	1920	1910	1900	1880	1870	1850	
RACE	Race	P	codes	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
BPL	Birthplace	P	codes	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	
MBPL	Mother's birthplace	P	codes	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	
FBPL	Father's birthplace	P	codes	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	
PRCORPAR	Puerto Rican Parentage Indicator	P	codes																								
NATIVITY	Foreign birthplace or parentage	P	codes																								
NACDISE	Nativity and parentage, Puerto Rico	P	codes																								
ANCESTR1	Ancestry, first response	P	codes	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	
ANCESTR2	Ancestry, second response	P	codes	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	
CITIZEN	Citizenship status	P	codes	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	
YEMAZUR	Year naturalized	P	codes	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	
YEMANG	Year of immigration	P	codes	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	
YEMUS1	Years in the United States	P	codes	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	
YEMUS2	Years in the United States, intervalled	P	codes	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	
YEMARR	Year of immigration to Puerto Rico	P	codes																								
YEMPR	Years in Puerto Rico	P	codes																								
YEMDSE	Year of immigration to Puerto Rico (intervalled)	P	codes																								
MCDIENGL	Mother tongue	P	codes																								
MKDIENGL	Mother's mother tongue	P	codes																								
FKDIENGL	Father's mother tongue	P	codes																								

http://usa.ipums.org/usa-action/variables/group/race_eth

Potential Project Ideas

Novel data acquisition & annotation interfaces
Identify currently ill-supported tasks
Improve data quality, prioritize collection
Improve metadata at the point of capture

The design of a data search engine
... when searching for data on a new topic.
... for complementary data within an analysis.

Better data assessment tools: data quality,
coverage, integrated documentation, ...

Wednesday: Guest Panel

Selina Tobaccowala, Philip Garland (SurveyMonkey)
Administrators of surveys to millions on the web.

Kuang Chen (UC Berkeley)
Creator of Usher & Shreddr, new user interfaces for
enhancing and accelerating data entry.

Christine Robson (IBM & UC Berkeley)
Researcher exploring data tools for citizen science.

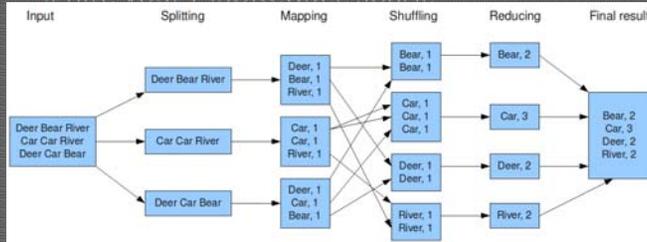
Assignment 2: Analyzing Big Data

A2 Goals

- Ignite your MapReduce spark!
- Learn about Amazon Web Services
 - EC2 = Elastic Cloud Computing
 - S3 = Simple Storage Service
 - EMR = Elastic MapReduce
- Count the word frequencies for all current Wikipedia articles. No, seriously.
- Design your own experiment to run using MapReduce on AWS

A2 Goals

- Ignite your MapReduce spark!



- Design your own experiment to run using MapReduce on AWS

A2 Goals

- Ignite your MapReduce spark!
- Learn about Amazon Web Services
 - EC2 = Elastic Cloud Computing
 - S3 = Simple Storage Service
 - EMR = Elastic MapReduce
- Count the word frequencies for all current Wikipedia articles. No, seriously.
- Design your own experiment to run using MapReduce on AWS

A2 Goals

- Ignite your MapReduce spark!
- Learn about Amazon Web Services
 - EC2 = Elastic Cloud Computing
 - S3 = Simple Storage Service
 - EMR = Elastic MapReduce
- Count the word frequencies for all current Wikipedia articles. No, seriously.
- Design your own experiment to run using MapReduce on AWS

A2 Goals

- Ignite your MapReduce spark!
- Learn about Amazon Web Services
 - EC2 = Elastic Cloud Computing
 - S3 = Simple Storage Service
 - EMR = Elastic MapReduce
- Count the word frequencies for all current Wikipedia articles. No, seriously.
- Design your own experiment to run using MapReduce on AWS

A2 Structure

Part 1

- Learn how to run an EMR workflow
- Write and edit MapReduce scripts that sort and filter output
- Count word frequencies in current Wikipedia

Part 2

- Design your own experiment to run on Amazon EMR. Go wild!
- May use own data.

A2 Evaluation

Part 1 (due 04/11)

- Correctness
- Deliverables are code + results

Part 2 (due 04/18)

- Correctness
- Experiment design and justification
- Deliverables are code + results + write up

Nitty Gritty Admin

- A2 is now live at
 - <http://hci.stanford.edu/courses/cs448g/a2>
- Links to MapReduce and AWS Setup tutorials
- Credit Coupon → check your inbox
- AWS Setup Bash
 - Tuesday 04/05
 - 4-6pm
 - Location TBA

Discussants

Philip Guo
Sean Holbert