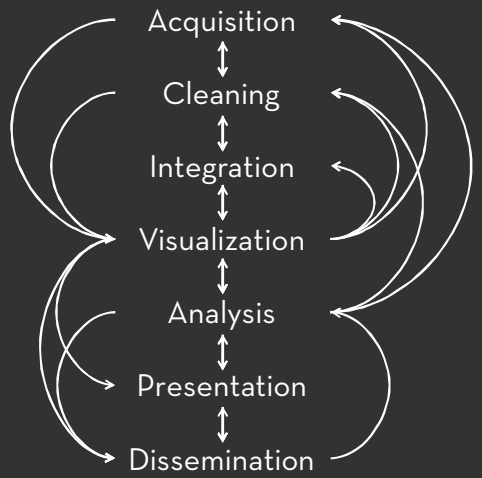
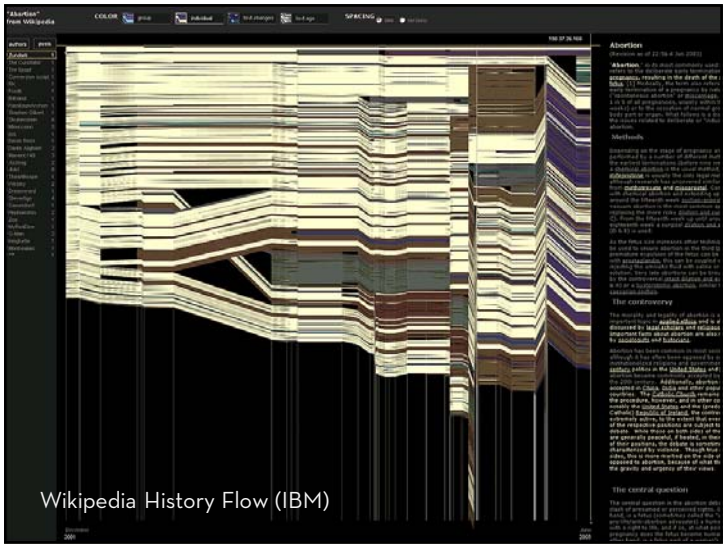
CS448G :: 4 Apr 2011

Data Collection

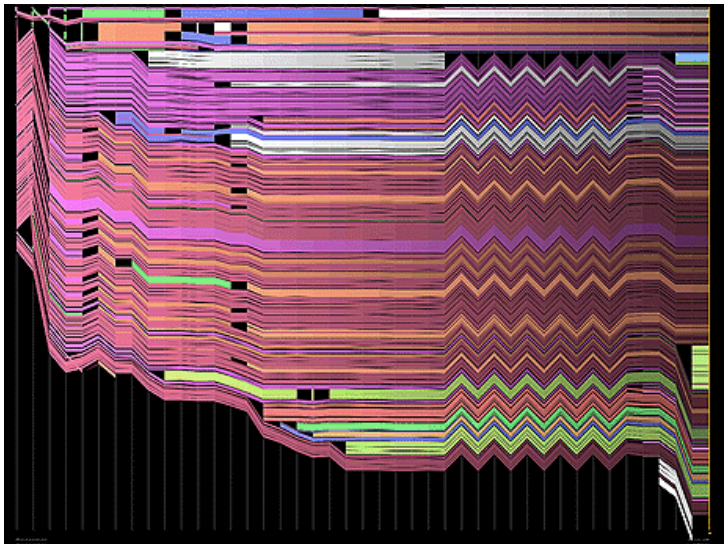
Jeffrey Heer Stanford University



cabspotting.org



Wikipedia History Flow (IBM)



Information Foraging

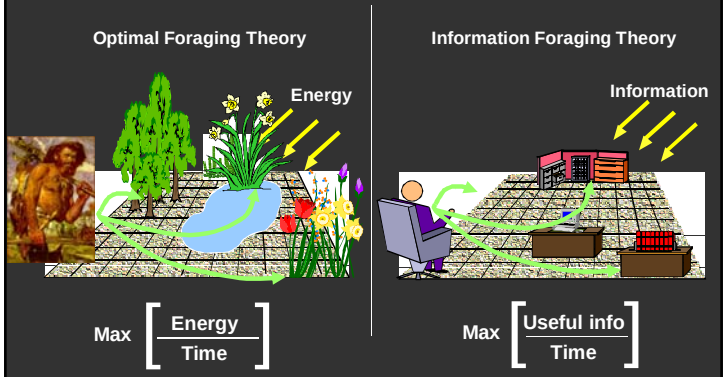
Exercise: Let's Find a Data Set

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)*



$$\text{Max}\left[\frac{\text{Energy}}{\text{Time}}\right] \qquad \text{Max}\left[\frac{\text{Useful info}}{\text{Time}}\right]$$

Information Foraging Theory

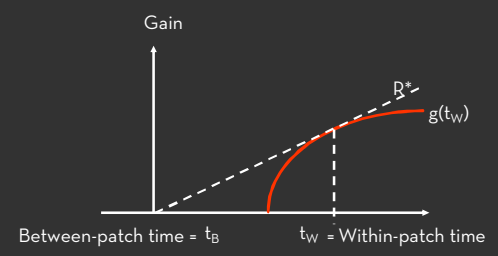
People are **information rate maximizers** of benefits/costs

Information has a **cost structure**

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

Charnov's Marginal Value Theorem

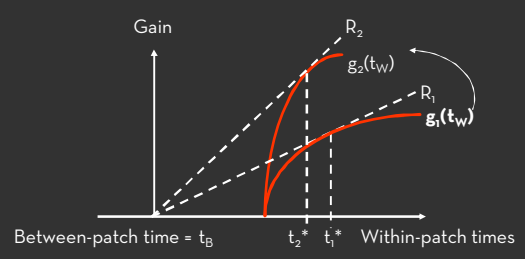
Maximum gain when slope of within-patch gain g = average gain R (tangent in diagram)



Within-Patch Enrichment

Behavior adapts to cost structure of environment.

Example: Better filtering of search hits

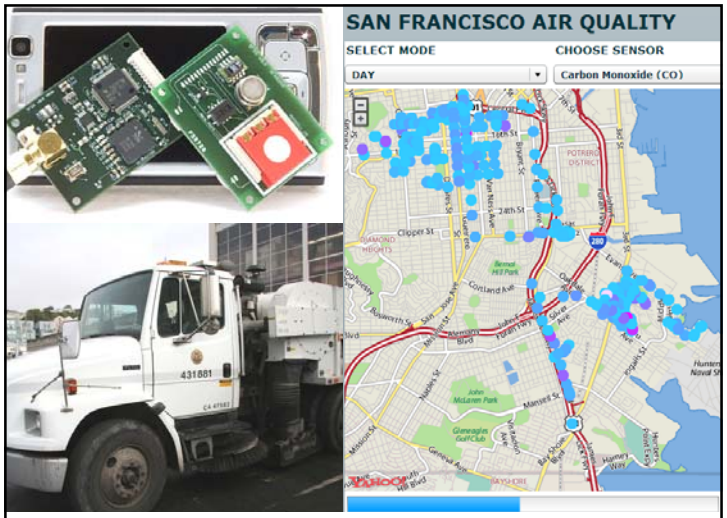


Enhancing Data Collection

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.



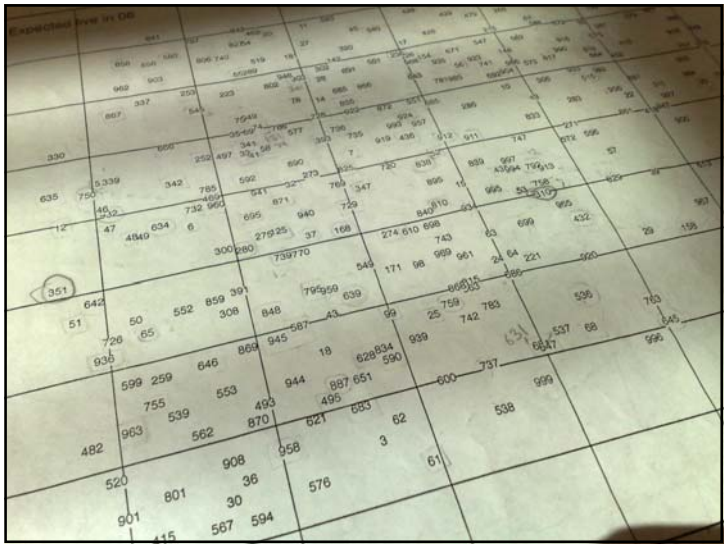
Aoki, Honicky, Mainwaring, Myers, Paulos, Subramanian & Woodruff, CHI 2009

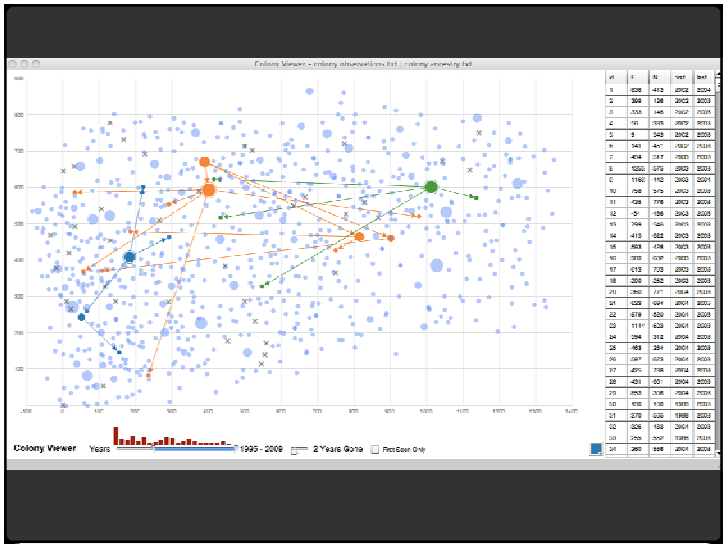
Insights for Data Collection

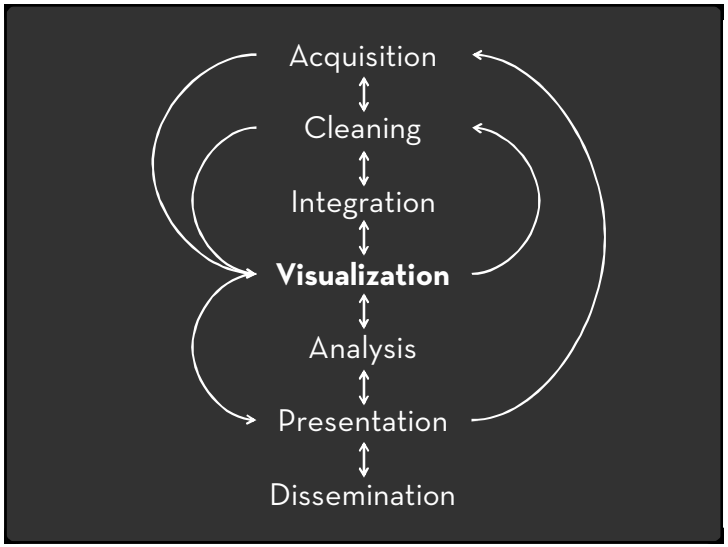
Foraging: decrease time between patches, increase information gain within patch.

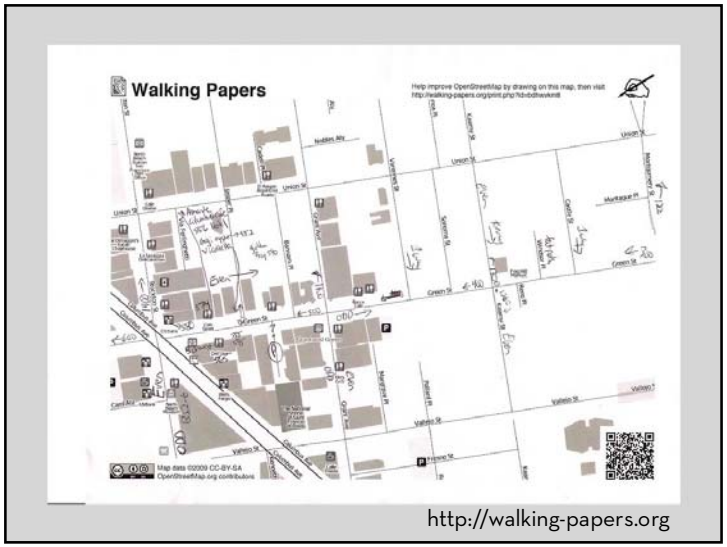
Ubiquitous networking and mobile devices open up new frontiers for data collection.

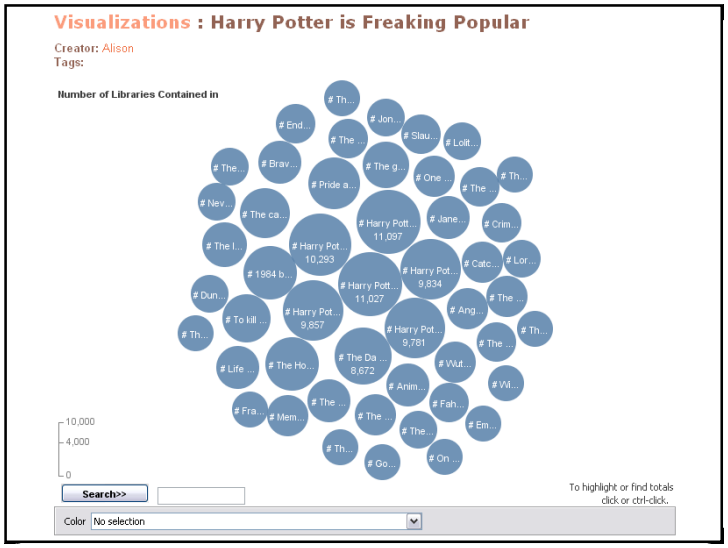
Visualizations are also input devices.

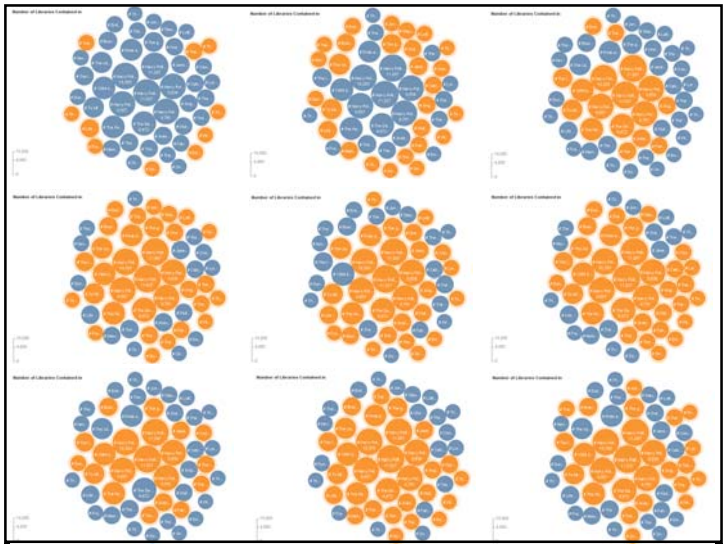



Colony Viewer
Years
1995 - 2009
2 Years Gone
Red Eaten Only


Acquisition
Cleaning
Integration
Visualization
Analysis
Presentation
Dissemination


Walking Papers
http://walking-papers.org





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, ...

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.

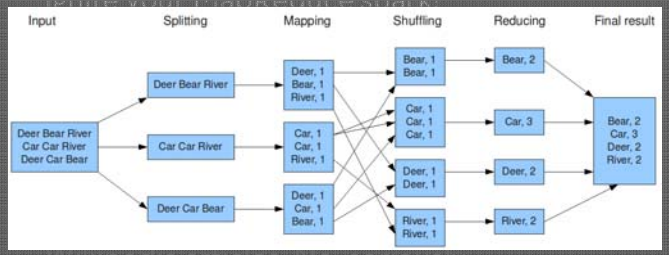
Assignment 2: Analyzing Big Data

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!



- 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 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