Principles of Information and Database Management 198:336 Week 1 – Jan 24

Matthew Stone

Today

Preliminaries

Motivation

Overview

Themes

Preliminaries

Course information

- Web page
  http://www.cs.rutgers.edu/~mdstone/class/336
  linked off my web page
  lecture notes, regular announcements, etc.
  will appear here
- Email list
  336-stone@rams.rutgers.edu
  everyone registered should have received a
  "welcome" message with web page

Preliminaries

People:

Matthew Stone

Matthew.Stone@Rutgers.edu

Regular office hours: 4-6pm Tues, C328

Special tomorrow: 4-6pm, Psych A103.

Preliminaries

People:

Vladislav Shkapenyuk

vshkap@cs.rutgers.edu

Preliminaries

Class conduct

- Power point, for outline
- Ask questions at any time
- Please introduce yourself

Motivation

"Knowledge is power."

- Sir Francis Bacon

"Predictive technology"

You're WalMart. A hurricane is coming. What should you send your stores?

“Predictive technology”

You’re WalMart. A hurricane is coming. What should you send your stores?

Answer: pop-tarts and beer.

How do you know? Because that’s what people bought last time.

Data at WalMart

460 Terabytes (460K Gb)

- Larger than the WWW
- Every product sold, all of inventory.

“Retail link”

- Suppliers see how their own products sell.
- Cost $4B to develop.

Data at WalMart

Coming in the next couple years
- Every shipment tagged with RFID

Eventually
- Every item tagged.

A bit about DBMS

"database management system"
- a software package designed to store and manage databases

Why would you need such a thing?

Why DBMS

Data independence and efficient access.
- Tedious business of bigness

Data integrity and security
- Tedious business of enforcing policies

Concurrent access, recovery from crashes
- Tedious business of robustness

Why Data

"Value of information"
- Not just a metaphor. Actual $$$.

WalMart

WalMart is sending a truck to Florida.

- They can either
  - A: load it with the goods people buy on average
  - B: load it with pop-tarts and beer
- How exactly does data help WalMart decide?

Information

WalMart knows that when the hurricane comes, people will come into the store looking for one thing in particular.

If they send the right thing, they sell it all: get \$10K/truckload.

Otherwise they just sell half of usual stuff: get \$5K/truckload.

Value of Information

WalMart is uncertain about what we want.

P (want flashlights | hurricane) = .1

P (want poptarts | hurricane) = .1

P (want clothes | hurricane) = .1

P (want games | hurricane) = .1 …

Value of information

If they send a little of everything:

- They guess right about one item.
  Get 1/10 truckload times 10K/truckload
- They guess wrong about nine items.
  Get 9/10 truckload times 5K/truckolad
- Overall: $5500

Value of information

If they guess randomly – say “board games”

- 10% of the time they are right, and get $10K.
- 90% of the time they are wrong. They just sell the usual amount of games and get $500.
- Overall they can expect $1450.

Value of information

What if they know people want pop-tarts?

- They know they’re getting the whole $10K

That truckload now looks like it’s worth $8550 more than before you knew.

That truckload is now the best thing you can send – by $4500.

Data could be really expensive

And WalMart would still want it.

But the cheaper the data is, the more we want!

Man, being the servant and interpreter of Nature, can do and understand so much and so much only as he has observed in fact or in thought of the course of nature. Beyond this he neither knows anything nor can do anything.

- Sir Francis Bacon

The obsession of data

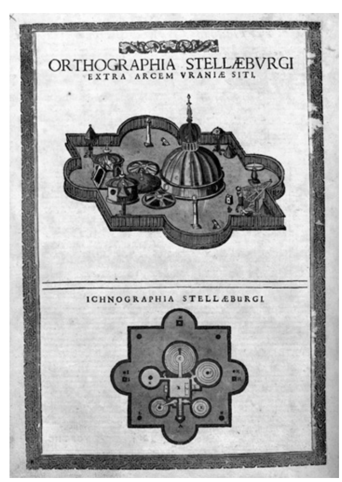
Here's the human genome project web site:
http://www.ncbi.nlm.nih.gov/genome/seq/

You can download it – all 3B base pairs.
By the end, each letter only cost 9¢

The obsession of data

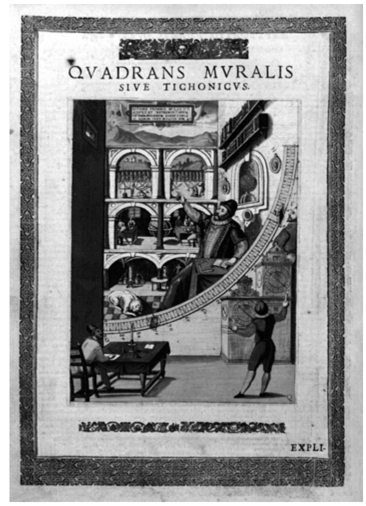
Spending for data is ancient

Tycho Brahe built two castle-observatories for himself in the 1580s.



The obsession of data

He decorated his living room with a 3 meter bronze quadrant for reading the angles of celestial bodies.



The obsession of data

His best instruments required two people to work simultaneously for each sighting.

The obsession of data

He sighted the planets every clear night for decades, with readings accurate to 1/1000 radian.

(A couple hundredths of an inch a yard away.)

The obsession of data

Kepler worked from Brahe's observation books to figure out the elliptical orbits of the planets.

Computers take it to a new level

Google indexes

- 8B web pages
- 800M images
- 6600 catalogs
- Etc.

Computers take it to a new level

This laser-scanned model of Michaelangelo’s David contains

8M polygons

(2mm resolution)

The raw data is

2B polygons

7K images

32G of disk

Computers take it to a new level

All the sightings of the Arecibo radio telescope over the next five years will be stored in a single Petabyte database

(=1M Gb)

Data in My Research

“Motion capture”

– Recording all aspects of how someone acts in conversation, to build an empirical model of face-to-face dialogue.


tricky
74 mph

Performance in our animation

Motion	Voice	Content
#091	#041	that was ugly
#122	#172	dude
#214	#174	you didn’t manage
#185	#155	to set up your landing

True in 1620 and still true today

No search has been made to collect a store of particular observations sufficient either in number, or in kind, or in certainty, to inform the understanding.

What in observation is loose and vague, is in information deceptive and treacherous.

Sir Francis Bacon

Information management

Computer systems now deal with wide variety of information

- Stored in heterogeneous formats
- Created by different organizations
- Requiring different kinds of access
- Plus an over-arching scheme for communicating and integrating information

Overview of the course

Kinds of information

Structured data

Relational model, schemas, queries, transactions
SQL, Oracle

Semi-structured data

XML, hypertext

Unstructured data

Text – vector space model and text retrieval

General knowledge

conceptual models, logic, data mining

Overview of the course

Methodology

- Understanding data, data-intensive problems
  - Building conceptual models of data
  - Designing data-intensive applications
- Using data management tools
  - Designing relational schemas
  - Integrating information sources

Timeline – Approximate

Week 2 – Logic, entities and relationships
Week 3 – Conceptual modeling
Week 4 – SQL, schemas and constraints
Week 5 – Description logic, datalog, views
Week 6 – Inference, triggers, integrity
Week 7 – Accessing DBMS from the web
Week 8 – Midterm

Timeline – Approximate

Week 9 – Sequence data: Text, etc.
Week 10 – Tree data: XML
Week 11 – Graph data: the web
Week 12 – Data mining
Week 13 – Transactions and concurrency

Requirements

Homeworks and exams

- Come to office hours twice.
- Short exercises
  (written, interactive, programming)
- Project
  (real DB application, in parts, teams of 2)
- Midterm
- Final

Themes

The excitement of real data

- Inherent coolness
- Value to individuals and organization
- Heterogeneity
- Effort to collect and creativity to use
- Social implications – privacy, security

Themes

Data management as a “growth field”

- New kinds of data
- New kinds of tools
- Fundamental to science, business, policy
- How computers affect society
- How computers become intelligent

The social science of computing

Themes

Database design as a real-world problem

- Causality: making sure data has real meaning
- Constraints: modeling, reasoning with data
- Collaboration: fitting data into organizations

Function of information manager

Start from a formal language

- Set of sentences
- Each is a symbolic structure with an intended interpretation – as information

Support operations

- TELL a sentence to the IM
  - Give the IM the information that S is true

Function of information manager

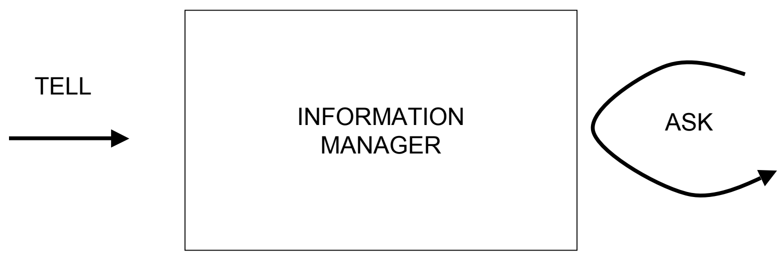
Start from a formal language

- Set of sentences
- Each is a symbolic structure with an intended interpretation – or as an information need

Support operations

- ASK a question to the IM
  - Express an information need to the IM
  - Get back a sentence representing the IM's answer

Pictorial view



Symbol table

Keep track of tokens encountered

- Tell the IM that some token has been seen
- Ask if some token has been encountered
- Get answer: yes or no.

Symbol table

Precise description:

TELL has tokens as its sentences.

ASK has tokens as its questions

ASK gets yes or no as its answers

Symbol table

Question answering

IM is a set of words

TELL(w, IM) : IM $\leftarrow$ {w} $\cup$ IM

ASK(w, IM) : if w $\in$ IM then “yes” else “no”

Symbol table

Abstracts from implemenation

- Can be hash table, linked list, binary tree
- Choice can be optimized based on use
- Choice can be changed

Design issues

What information can be told?

- What info does the teller actually have?
- How can we characterize it precisely?
- How do we formalize it to implement it?

But that’s not all…

Those who have handled sciences have been either men of experiment or men of dogmas. The men of experiment are like the ant, they only collect and use; the reasoners resemble spiders, who make cobwebs out of their own substance. But the bee takes a middle course: it gathers its material from the flowers of the garden and of the field, but transforms and digests it by a power of its own.

Sir Francis Bacon

Design issues

What information can be asked?

- What information do applications need?
- How do we characterize and specify this info?
- Can we “digest” what’s in the IM to give this?

Final design issue

How do we keep the IM on track?

- How do we catch bugs?
- How do we make sure it gets good info?
- How do we document the design?

DEFINE how the IM should behave

- Use sentences to make explicit constraints on what can be told and what can be asked.

Survey

1. Hometown
2. Used a DBMS?
3. Philosophy?
4. Economics/business?
5. Data from research lab experiment?
6. Dinner before or after?

FYI

Sir Francis Bacon

http://www.luminarium.org/sevenlit/bacon/

Novum Organum

1620

http://www.constitution.org/bacon/nov_org.txt