Low-Level Components of Analytic Activity in Information Visualization

Robert Amar, James Eagan,
& John Stasko

Information Interfaces Research Group
College of Computing & GVU Center
Georgia Institute of Technology

Georgia Tech College of Computing

GVU center georgia tech

Motivating Question

- Are there (low-level) categories of analytic activities people perform when using information visualization systems?

  If so, what are they?

Existing Work

Wehrend & Lewis (Vis '90)

User tasks

identify
locate
distinguish
categorize
cluster
distribution
rank
compare within relations
compare between relations
associate
correlate

Roth & Mattis (CHI '90)

User information-seeking goals

value lookup
compare within relations
compare between relations
distribution
functional correlation
indexing needs

Shneiderman (VL '96)

Visualization task

overview
zoom
filter
details on demand
relate
history
extract

Existing Work

Zhou & Feiner (CHI '98)

User goals in creating multimedia presentations

Inform

Elaborate

Emphasize
Reveal

Summarize

Associate
Background
Categorize
Cluster
Compare
Correlate
Distinguish
Generalize
Identify
Locate
Rank

Enable

Explore

Search

Categorize
Cluster
Compare
Correlate
Distinguish
Emphasize
Identify
Locate
Rank
Reveal

Verify

Categorize
Compare
Correlate
Distinguish
Identify
Locate
Rank
Reveal

Compute

Sum

Correlate
Locate
Rank

Differentiate

Correlate
Locate
Rank

Shortcomings

- Focus on a generated presentation or infovis system as end-result
  - User tasks a subcomponent
- Issues with task sets (important ones left out)

Background

- Use "commercial tools" class assignment (early in class)
- Students generate questions to be answered using commercial infovis systems
- Data sets:

Domain	Data cases	Attributes	Questions Generated
Cereals	78	15	43
Mutual funds	987	14	14
Cars	407	10	53
Films	1742	10	47
Grocery surveys	5164	8	39

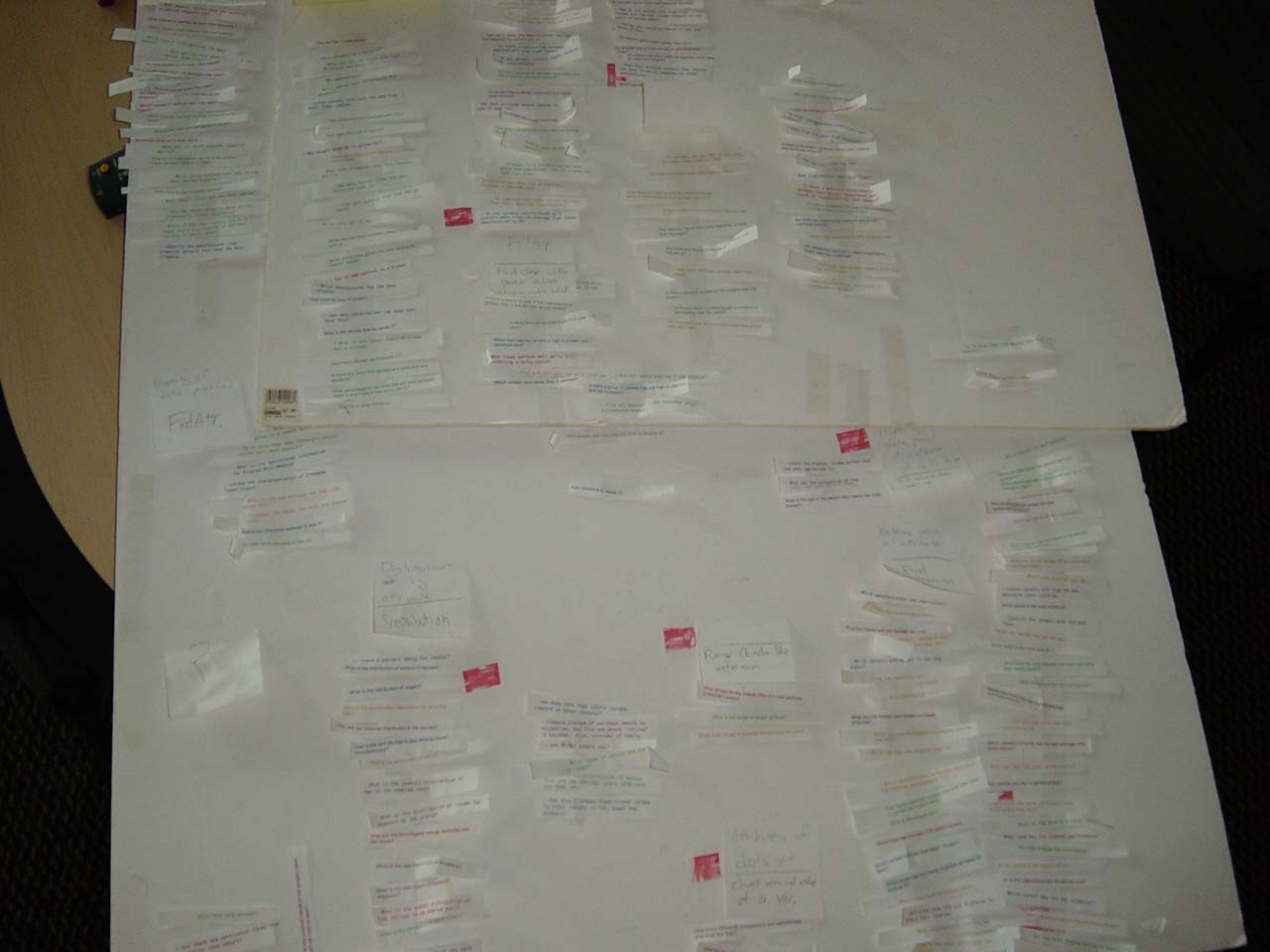
- Generated 196 total analysis tasks

Background

- Use "commercial tools" class assignment (early in class)
- Students generate questions to be answered using commercial infovis systems
- Data sets:

Domain	Data cases	Attributes	Questions Generated
Cereals	78	15	107
Mutual funds	987	14	41
Cars	407	10	153
Films	1742	10	169
Grocery surveys	5164	8	126

- Generated 596 total analysis tasks



extreme value of attribute

Find extremum

- Which manufacturers are healthiest?

Find the heaviest cars have the worst MPG?

Find the Fidelity with the highest net asset

Japanese, European and have the best MPG?

- Which cereals west in fat and sugar?

- Find the heaviest car.

5 accelerating cars.

Range (kinda like extremum)

What ranges do the middle 75% of funds perform in the first 3 years?

What is the range of length of films?

What is the range of possible horsepower for cars?

What are the highest and lowest purchase amounts?

Which cars have the highest horsepower and the best MPG?

- Which car has the biggest engine?

- Which manufacturer has the cars with the highest horsepower?

Find the shortest and longest films made after year X that are not music videos.

What is the longest film?

- Locate cereals determine their

Which cereal is the

- Identify the fiber.

Which car has the best

Which actor is the most pop

What were the most p they mostly recent?

What category

Identify the chain with the highest purchase amount.

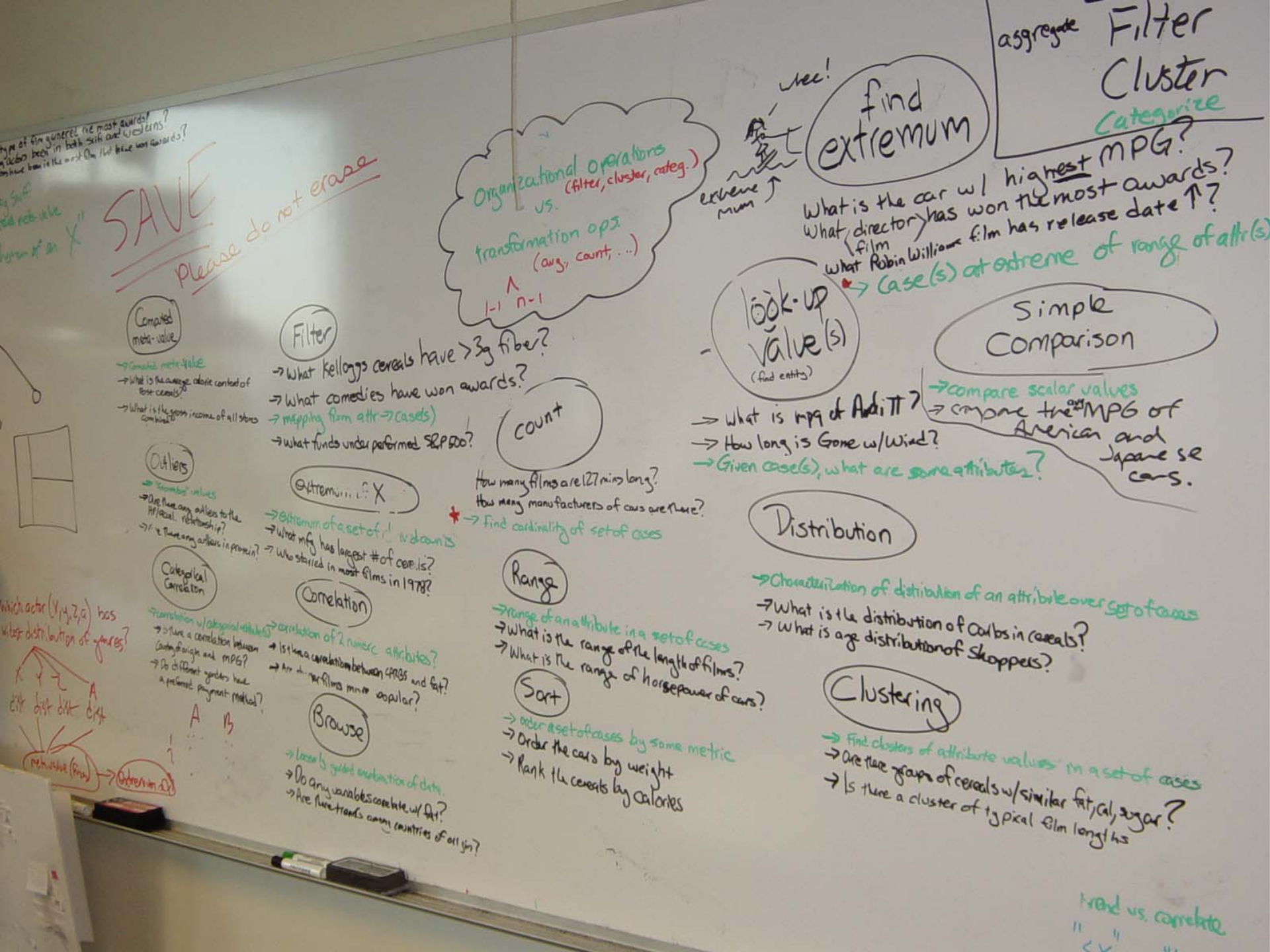
- Find the car with acceleration.

Which category of funds has the be performance?

- What car has the best accele

Which cereals are low in carbohydrates

Which of the more efficient accelerate the best?


SAVE
Please do not erase
Organizational operations
vs.
transformation ops
(aggs, count, ...)
1-1
n-1
find extremum
yee!
extreme mum
What is the car w/ highest MPG?
What director has won the most awards?
What film has release date ↑?
What Robin Williams film has release date ↑?
→ case(s) at extreme of range of attr(s)
look-up value(s)
(find entity)
What is mpg of Audi TT?
How long is Gone w/ Wind?
→ Given case(s), what are some attributes?
Simple Comparison
→ compare scalar values
→ compare the avg MPG of American and Japanese cars.
aggregate
Filter
Cluster
Categorize
Computed meta-value
Filter
→ What kelloggs cereals have > 3g fiber?
→ What comedies have won awards?
→ mapping from attr → case(s)
→ What funds underperformed S&P 500?
Outliers
Extremum
→ Extremum of a set of n values
→ What mfg has largest # of cereals?
→ Who starred in most films in 1978?
Count
How many films are 127 mins long?
How many manufacturers of cars are there?
→ Find cardinality of set of cases
Categorical Correlation
Correlation
→ correlation of 2 numeric attributes?
→ Is there a correlation between CARBS and fat?
Are ... films more popular?
Range
→ range of an attribute in a set of cases
→ What is the range of the length of films?
→ What is the range of horsepower of cars?
Distribution
→ Characterization of distribution of an attribute over set of cases
→ What is the distribution of carbs in cereals?
→ What is age distribution of shoppers?
Browse
→ Loosely guided exploration of data.
→ Do any variables correlate w/ fat?
→ Are there trends across countries or all yrs?
Sort
→ order a set of cases by some metric
→ Order the cars by weight
→ Rank the cereals by calories
Clustering
→ Find clusters of attribute values in a set of cases
→ Are there groups of cereals w/ similar fat/cal/sugar?
→ Is there a cluster of typical film lengths
trend vs. correlate

Terminology

- *Data case* – An entity in the data set
- *Attribute* – A value measured for all data cases
- *Aggregation function* – A function that creates a numeric representation for a set of data cases (eg, average, count, sum)

1. Retrieve Value

General Description:

Given a set of specific cases, find attributes of those cases.

Examples:

- What is the mileage per gallon of the Audi TT?
- How long is the movie Gone with the Wind?

2. Filter

General Description:

Given some concrete conditions on attribute values, find data cases satisfying those conditions.

Examples:

- What Kellogg's cereals have high fiber?
- What comedies have won awards?
- Which funds underperformed the SP-500?

3. Compute Derived Value

General Description:

Given a set of data cases, compute an aggregate numeric representation of those data cases.

Examples:

- What is the gross income of all stores combined?
- How many manufacturers of cars are there?
- What is the average calorie content of Post cereals?

4. Find Extremum

General Description:

Find data cases possessing an extreme value of an attribute over its range within the data set.

Examples:

- What is the car with the highest MPG?
- What director/film has won the most awards?
- What Robin Williams film has the most recent release date?

5. Sort

General Description:

Given a set of data cases, rank them according to some ordinal metric.

Examples:

- Order the cars by weight.
- Rank the cereals by calories.

6. Determine Range

General Description:

Given a set of data cases and an attribute of interest, find the span of values within the set.

Examples:

- What is the range of film lengths?
- What is the range of car horsepowers?
- What actresses are in the data set?

7. Characterize Distribution

General Description:

Given a set of data cases and a quantitative attribute of interest, characterize the distribution of that attribute's values over the set.

Examples:

- What is the distribution of carbohydrates in cereals?
- What is the age distribution of shoppers?

8. Find Anomalies

General Description:

Identify any anomalies within a given set of data cases with respect to a given relationship or expectation, e.g. statistical outliers.

Examples:

- Are there any outliers in protein?
- Are there exceptions to the relationship between horsepower and acceleration?

9. Cluster

General Description:

Given a set of data cases, find clusters of similar attribute values.

Examples:

- Are there groups of cereals w/ similar fat/calories/sugar?
- Is there a cluster of typical film lengths?

10. Correlate

General Description:

Given a set of data cases and two attributes, determine useful relationships between the values of those attributes.

Examples:

- Is there a correlation between carbohydrates and fat?
- Is there a correlation between country of origin and MPG?
- Do different genders have a preferred payment method?
- Is there a trend of increasing film length over the years?

Discussion/Reflection

- Compound tasks
  - "Sort the cereal manufacturers by average fat content"
    Compute derived value; Sort
  - "Which actors have co-starred with Julia Roberts?"
    Filter; Retrieve value

Discussion/Reflection

- What questions were left out?
  - Basic math
    "Which cereal has more sugar, Cheerios or Special K?"
    "Compare the average MPG of American and Japanese cars."
  - Uncertain criteria
    "Does cereal (X, Y, Z…) sound tasty?"
    "What are the characteristics of the most valued customers?"
  - Higher-level tasks
    "How do mutual funds get rated?"
    "Are there car aspects that Toyota has concentrated on?"
  - More qualitative comparison
    "How does the Toyota RAV4 compare to the Honda CRV?"
    "What other cereals are most similar to Trix?"

Discussion/Reflection

- Shares overlap with taxonomies discussed earlier
- Shares operations with spreadsheets or DB languages such as SQL

Concerns

- InfoVis tools may have influenced students' questions
- Graduate students as group being studied
  - How about professional analysts?
- Subjective – Not an exact science

Contributions

- Set of grounded low-level analysis tasks
- Potential use of tasks as a language/vocabulary for comparing and evaluating infovis systems
- Continue emphasis and focus on end-user goals and tasks

Thanks for your attention!

Acknowledgments:

- Research supported in part by NSF IIS-0414667
- Thanks to all the students from CS 7450

www.cc.gatech.edu/gvu/ii/vistasks