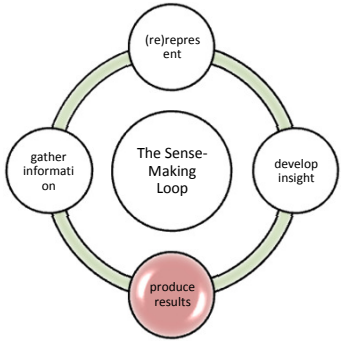
REPRODUCIBLE RESEARCH

PROVENANCE

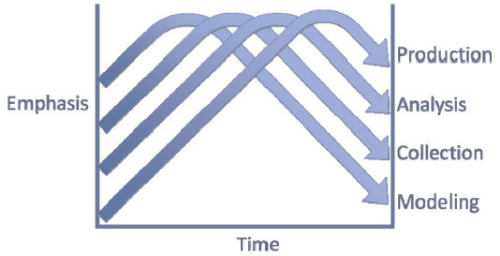
PETRA ISENBERG

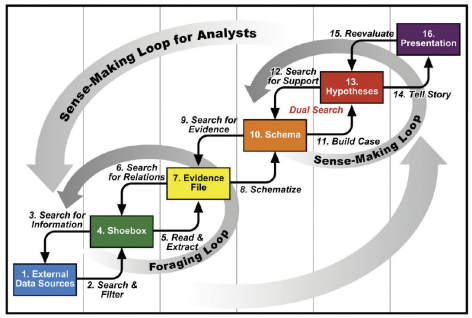
VISUAL ANALYTICS 07 Oct 2015

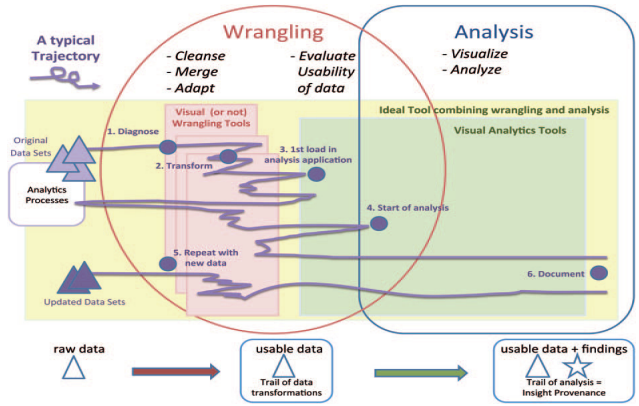
MANY MODELS FOR ANALYSIS



What do they have in common?







IN THIS LECTURE

you will learn that

- it is important to communicate what you've done
- in a way that someone else can reconstruct / understand what you've done
- details are important because there are so many steps involved in an analysis

HOW COULD YOU CONVEY THE PROCESS OF YOUR ANALYSIS?

- in words – tell it
- provide computer code, data, …
- write reports

WHY SHOULD YOU CONVEY THE PROCESS OF YOUR ANALYSIS?

- to show your findings are robust
- to highlight subjective decisions made
- to enable improvements on your methods
- to help someone learn how to do analysis
- ...

PROBLEMS

- even simple analyses not easy to describe
- often people don't have the right skills in computing, statistics, … to understand the analysis processes
- large datasets and complicated analyses create long analysis pipelines
- lots of trial and error in analysis

REPLICATION VS. REPRODUCIBILITY

REPLICATION

- ability of an entire experiment or study to be duplicated with independent / new
  - data
  - investigators
  - analysis methods
  - …
- ultimate standard for strengthening scientific evidence

Science 2 December 2011:
Vol. 334 no. 6060 pp. 1226-1227
DOI: 10.1126/science.1213847
+ Coursera MOOC – Reproducible Research

REPLICATION

- check if a finding is robust
  - is this claim true?
- especially important when studies have broad impact (e.g. on society)

REPLICATION

BUT sometimes you can't replicate because

- you don't have the time
- or the money
- or the resources
- or the situation is unique

e.g. how would you replicate the Sloan Digital Sky Survey?

IF YOU CAN'T REPLICATE?

- what else can you do?
- let a study/an analysis stand by itself?

IF YOU CAN'T REPLICATE?

- what else can you do?
- let a study/an analysis stand by itself?

REPRODUCIBILITY

REPRODUCIBILITY

- asks:
  - can we trust this analysis?
- should be minimum standard for any scientific study
- new investigators: same data, same methods

→ allow for validation of the data analysis

WHY?

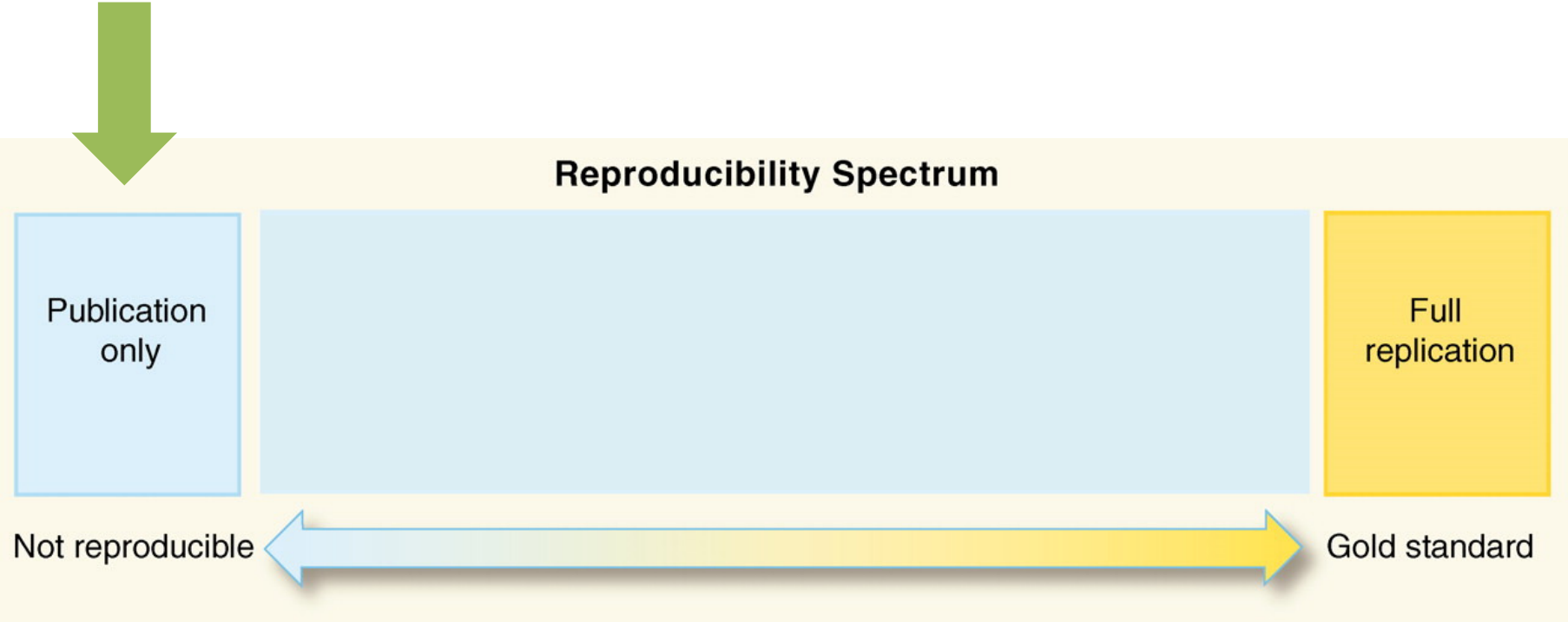
WHY?

Another video for you to look at at home

https://www.youtube.com/watch?v=eV9dcAGaVU8

("Deception at Duke")

Analysis (incl. data collection, cleaning, analytic methods, figure generation, …)



Roger D. Peng Science 2011;334:1226-1227

Science AAAS

Analysis (incl. data collection, cleaning, analytic methods, figure generation, …)

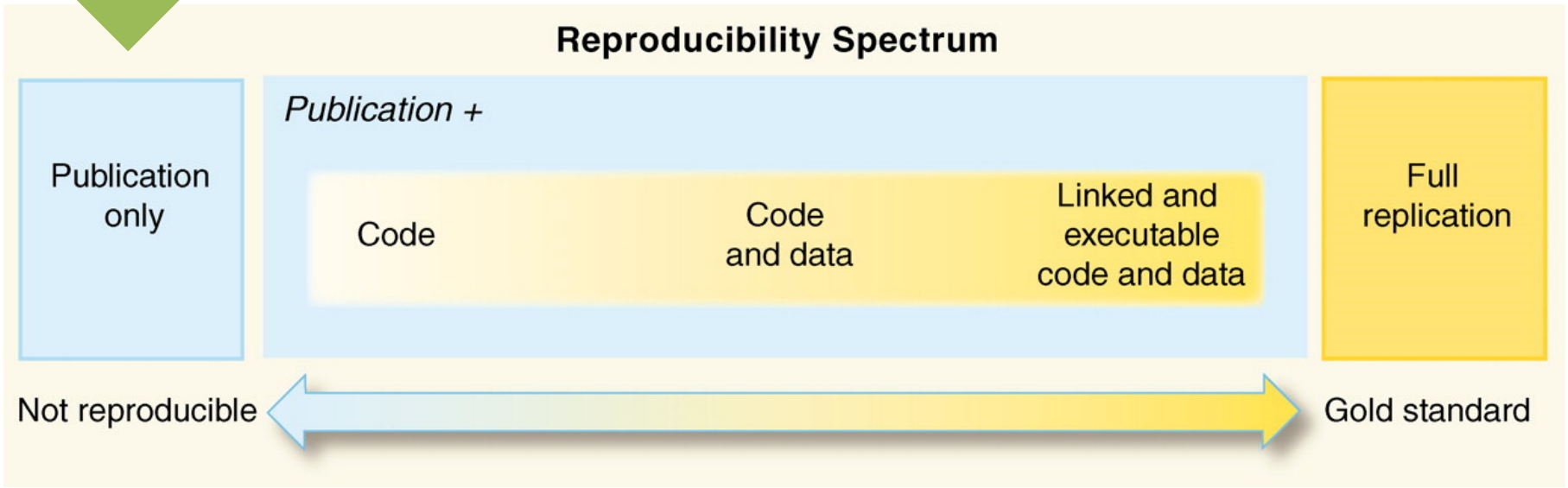
Roger D. Peng Science 2011;334:1226-1227

Science AAAS

WHAT TO DO?

- make your data available
  - analyze same data again (rather than analyzing independently collected data)
- make your analysis methods available
- document code and data
- use standard means of distribution

Analysis



Roger D. Peng Science 2011;334:1226-1227

Science AAAS

WHO IS INVOLVED?

- analysts
  - who want to make their work reproducible
- readers
  - who want to reproduce (or build on) the previous analysis

CHALLENGES

- what are good tools for analysts?
  - documentation is time-consuming
  - needs resources (web servers, etc.)
- what are good tools for reproduction?
  - how to piece together data & code
  - trying to understand what happened

REPRODUCIBILITY

- concept important to ANYONE conducting an analysis
- BUT: there is no agreed-upon notation for writing “instructions”

REPRODUCIBILITY

For coding environments – like R

BE ORGANIZED

BE ORGANIZED!

- you will deal with
  - data (raw + processed)
  - figures (exploratory + final)
  - code (raw, unused, final, bugged, debugged, …)
  - text (readme files, analysis report, documentation)

RAW DATA

- should be stored in your analysis folder
- should come with readme (for data provenance – see later slide)
  - if accessed from web, include url, description, and date accessed

PROCESSED DATA

sometimes you need to transform data (remember your data cleaning exercises)

- name processed data so you know which script generated it
- make a readme that says which script/procedure generated the file
- processed data should be ready for analysis

FIGURES

- you will generate many that you don't need
- make the final figured pretty and use proper labeling and color, possibly captions

SCRIPTS

- clearly comment your final scripts
  - what, when, why, how throughout
  - bigger comment blocks for whole sections
- include processing details
- clean the script to only include code needed to produce the final analysis

GENERAL RECOMMENDATIONS

- keep track of what you're doing
  - e.g. use version control systems
- save as much code as possible as little output as necessary
- save data in non-proprietary formats

PROBLEMS

- it takes a lot of effort to make data/results available
- readers must find your stuff and piece it together
- typically data, code, text are not linked

LITERATE PROGRAMMING

LITERATE PROGRAMMING

explanation of the program logic in a natural language, such as English, interspersed with snippets of macros and traditional source code (Wikipedia)

- You write code to do an analysis
  - compute results
  - generate data tables
  - …
- You also write a document – text chunks surrounding your analysis code
  - explain your analysis
  - format your results

LITERATE PROGRAMS

- use a documentation language (human readable)
- use a programming language (machine readable)
- have a pre-processor that:
  - weaves the doc to produce human-readable documents (pdf, html, …)
  - tangles the doc to produce machine-readable documents

EXAMPLES

- First:
  - WEB (by Donald Knuth, 1981): Pascal + TeX
- Sweave: R + Latex
- Knitr: R + Latex, Markdown, HTML

```
---
title: "Mayhem at DinoFunWorld"
author: "Petra Isenberg"
date: "October 5, 2015"
output: html_document
---

#Merging Data Files with R

##Loading Files

First we will load a file that contains attractions, their ids, and coordinates in the park
```{r}
coordinates <- read.csv("ParkCoordinates.csv")
head(coordinates)
```

Next we will load our data from the data cleaning exercise
```{r}
attractions <- read.csv("AttractionsOCR-txt.csv")
head(attractions)
```

```

# Mayhem at DinoFunWorld

*Petra Isenberg*

*October 5, 2015*

# Merging Data Files with R

## Loading Files

First we will load a file that contains attractions, their ids, and coordinates in the park

```
coordinates <- read.csv("ParkCoordinates.csv")
head(coordinates)
```

```
##               Attraction AttractionID  x  y
## 1 Wrightiraptor Mountain            1 47 11
## 2     Galactosaurus Rage            2 27 15
## 3     Auvilotops Express            3 38 90
## 4              TerrorSaur            4 78 48
## 5      Wendisaurus Chase            5 16 66
## 6   Keimosaurus Big Spin            6 86 44
```

Next we will load our data from the data cleaning exercise

```
attractions <- read.csv("AttractionsOCR-txt.csv")
head(attractions)
```

```
##   AttractionID      ParkArea             Attraction CategoryNames
## 1            1 Coaster Alley Wrightiraptor Mountain  Thrill Rides
## 2            2 Coaster Alley     Galactosaurus Rage  Thrill Rides
## 3            3   Tundra Land     Auvilotops Express  Thrill Rides
## 4            4      Wet Land              TerrorSaur  Thrill Rides
## 5            5   Tundra Land      Wendisaurus Chase  Thrill Rides
## 6            6 Coaster Alley   Keimosaurus Big Spin  Thrill Rides
```

PROS & CONS

- text and code all in one place
  - order is maintained
- results are automatically updated when data changes
- code needs to run to produce the document

PROS & CONS

- documents can become difficult to read, when there is a lot of code
- can be slow
  - but you can use things like caching

REPRODUCIBILITY

In Visual Analytics Tools

REPRODUCIBILITY FOR GUIS

- how do you make your analysis methods available in a GUI-tool?

FIRST IDEA...

Capture all interactions in a system and make them available

BUT

what interactions to capture and how?

HISTORY MODELS

- Maintain a graph of application states
- Nodes = states of the application (incl. settings & application content)
- Edges = action that transform states

Graphical Histories for Visualization: Supporting Analysis, Communication, and Evaluation, TVCG 2008, Heer et al.

HISTORY MODELS

- what do we store?
  - states?
  - actions?
  - both?

ACTION LOGGING

Also called: command object model

- command object holds interface action
  - typically provides undo and redo
- common in graphic design tools

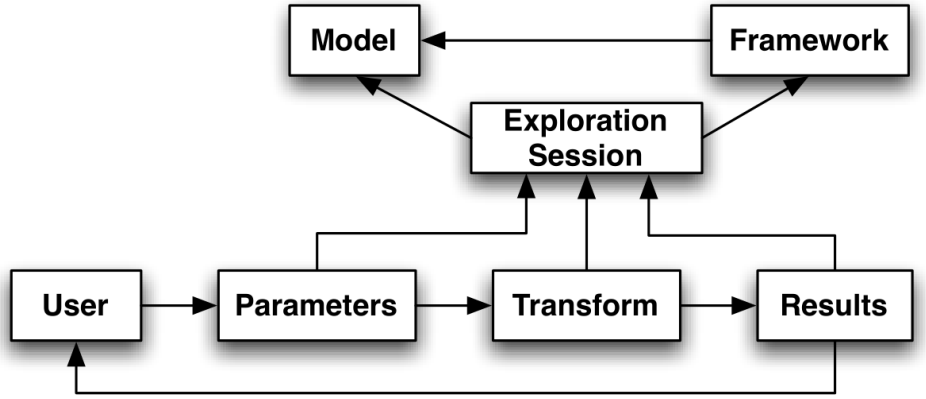
LOGGING STATES

- application can be restored to any stored configuration
- can be memory inefficient

- common in web browsing (states stored as URLs)

IN VISUALIZATION

- describe visualization as chain of visua encoding operat
- P-Set Model:
  - state = set of parameters & ac as transformations of these parameters



A Model and Framework
for Visualization Exploration
T.J. Jankun-Kellym TVCG 2007

PROBLEMS IN VISUALIZATION?

- dependence on underlying dataset
  - what if data is streamed or editable?

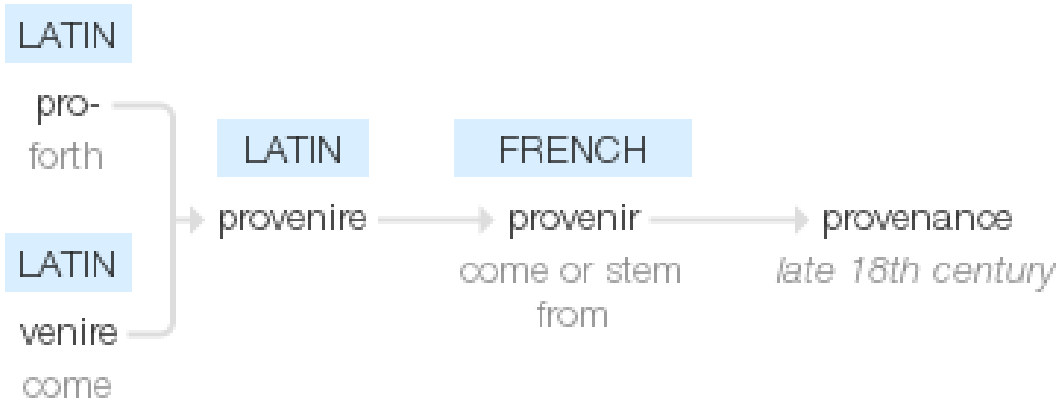
Is capturing interactions enough to allow for reproducibility?

PROVENANCE

The following slides are inspired by a lecture given by Remco Chang at Tufts

DEFINITION

- “origin, source”
- “the history of ownership of a valued object or work of art of literature”



source: Google

PROVENANCE

- Data provenance
- Information provenance
- Insight provenance
- Analytic provenance

DATA PROVENANCE

- description of the origins of a piece of data and the process by which it arrived in a database
- also called “lineage” or “pedigree”

Why and where: A characterization of data provenance,
ICDT 2001 Bunemann et al.

WHY?

- know about derivations of a data source
- experimental replay
- auditing
- fraud and malicious behavior detection
- quota and billing management

Towards a Secure and Efficient System for End-to-End Provenance, McDaniel et al. ; USENIX Workshop 2010

DATA PROVENANCE

- well researched topic in the database community

INFORMATION PROVENANCE

- know how a piece of information is modified as it propagates
- know how the owner of a piece of information is connected to its transmission

Provenance Data in Social Media
By Geoffrey Barbier, Zhuo Feng, Pritam Gundecha

INSIGHT PROVENANCE

a historical record of the process and rationale by which an insight is derived

"Characterizing users' visual analytic activity for insight provenance," in *VAST '08.* Gotz, D.; Zhou, M.X., doi: 10.1109/VAST.2008.4677365

ANALYTIC PROVENANCE

Goal:

- To understand a user's analytic reasoning process when using a (visual) analytical system for task-solving.

Benefits:

- Training
- Validation
- Verification
- Recall
- Repeated procedures
- Etc.

STAGES

- Recording what a user sees
- Capture interactions with the system
- Store the interactions
- Translate the interaction logs into something meaningful
- Reuse - reapply the interaction log to a different problem or dataset

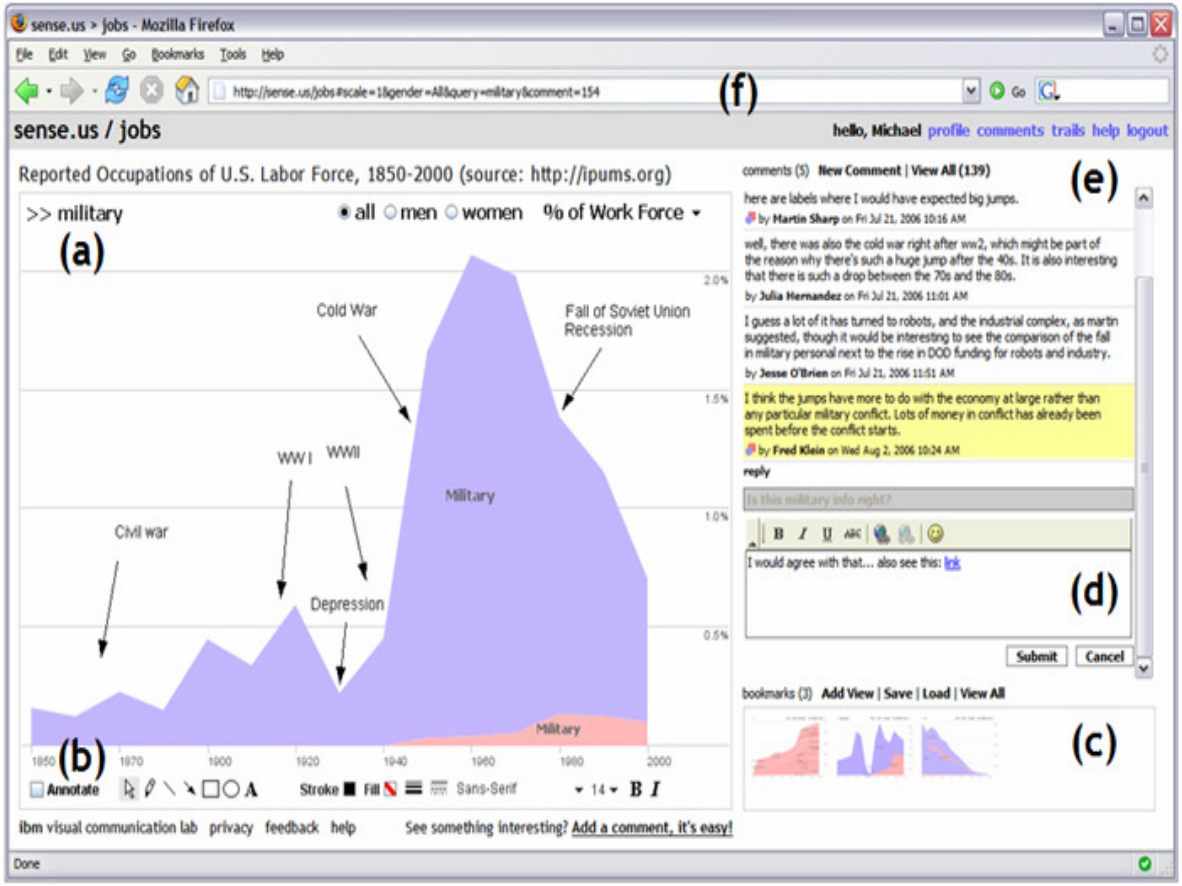
CAPTURE

- The “bread and butter” of analytic provenance
- Need to choose carefully about “what” to capture
  - Capturing at low level -> cannot decipher the intent
  - Capturing at high level -> not usable for other applications

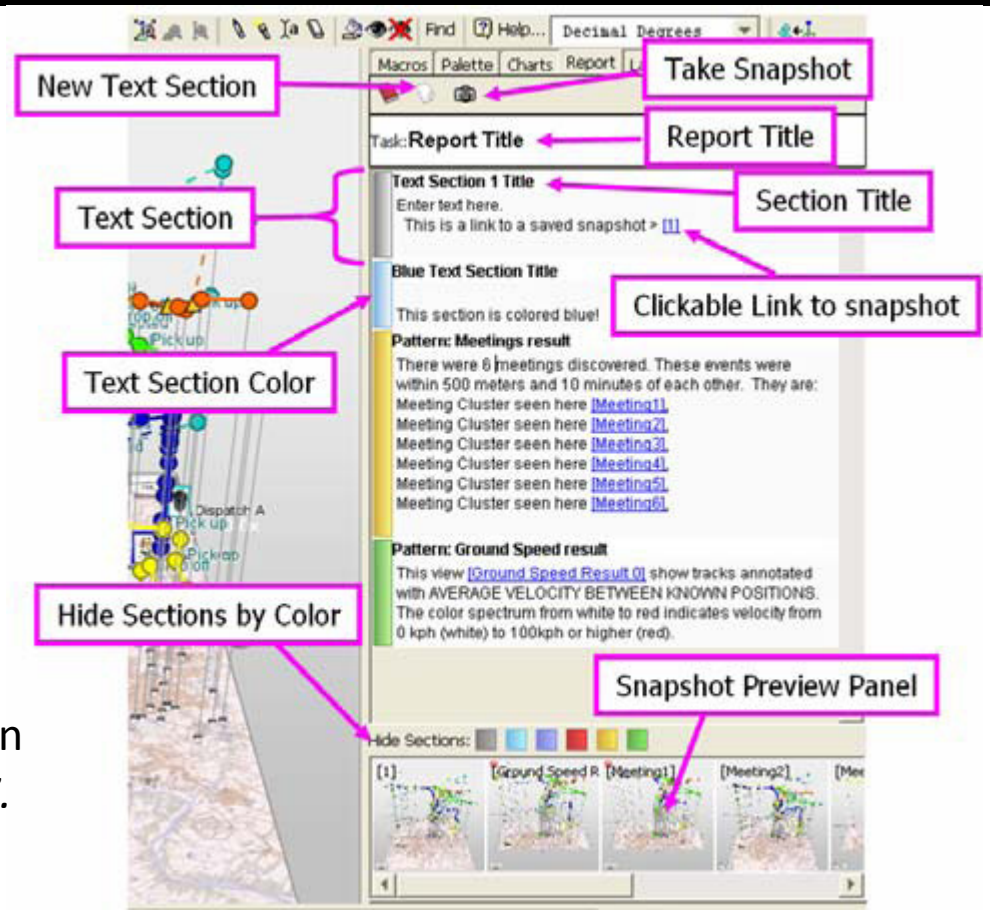
CAPTURING

- Manual Capturing – when in doubt, ask the user!
  - Annotations: directly edited text
  - Structured diagrams: illustrating analytical steps
  - Reasoning graph: reasoning artifacts and relationships

(MANUAL) ANNOTATIONS

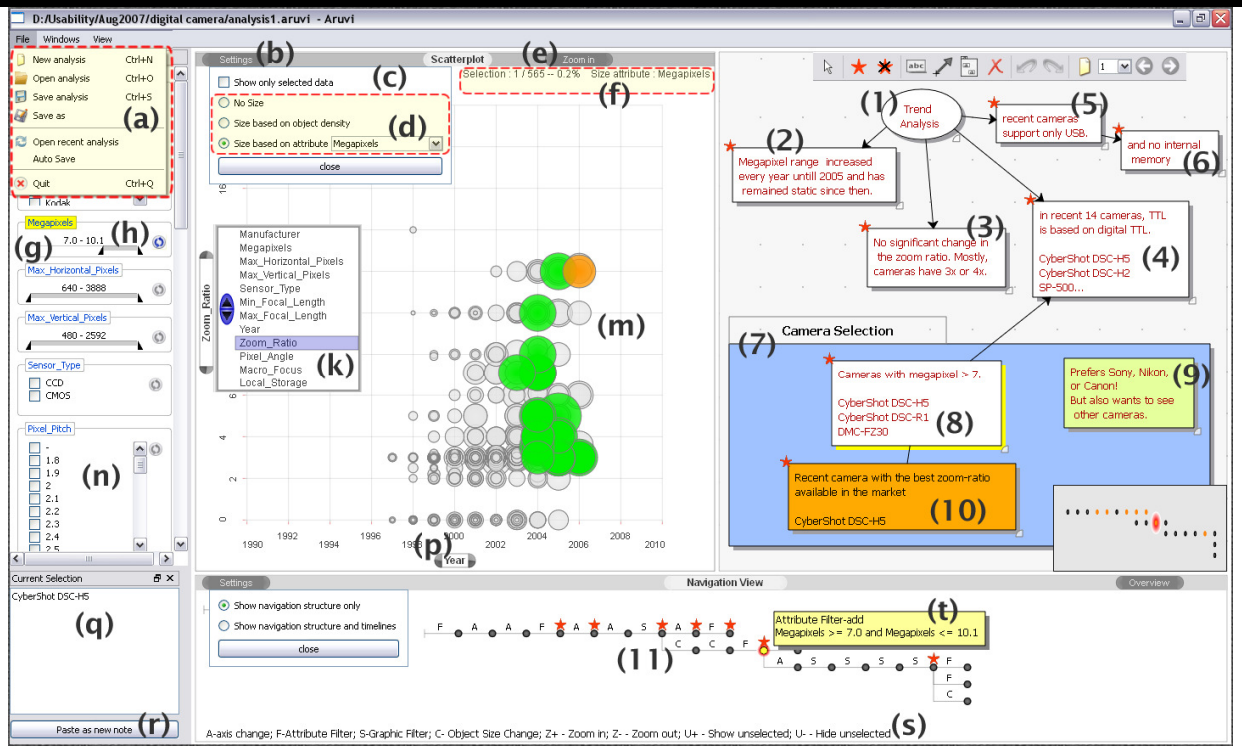


(MANUAL) ANNOTATIONS

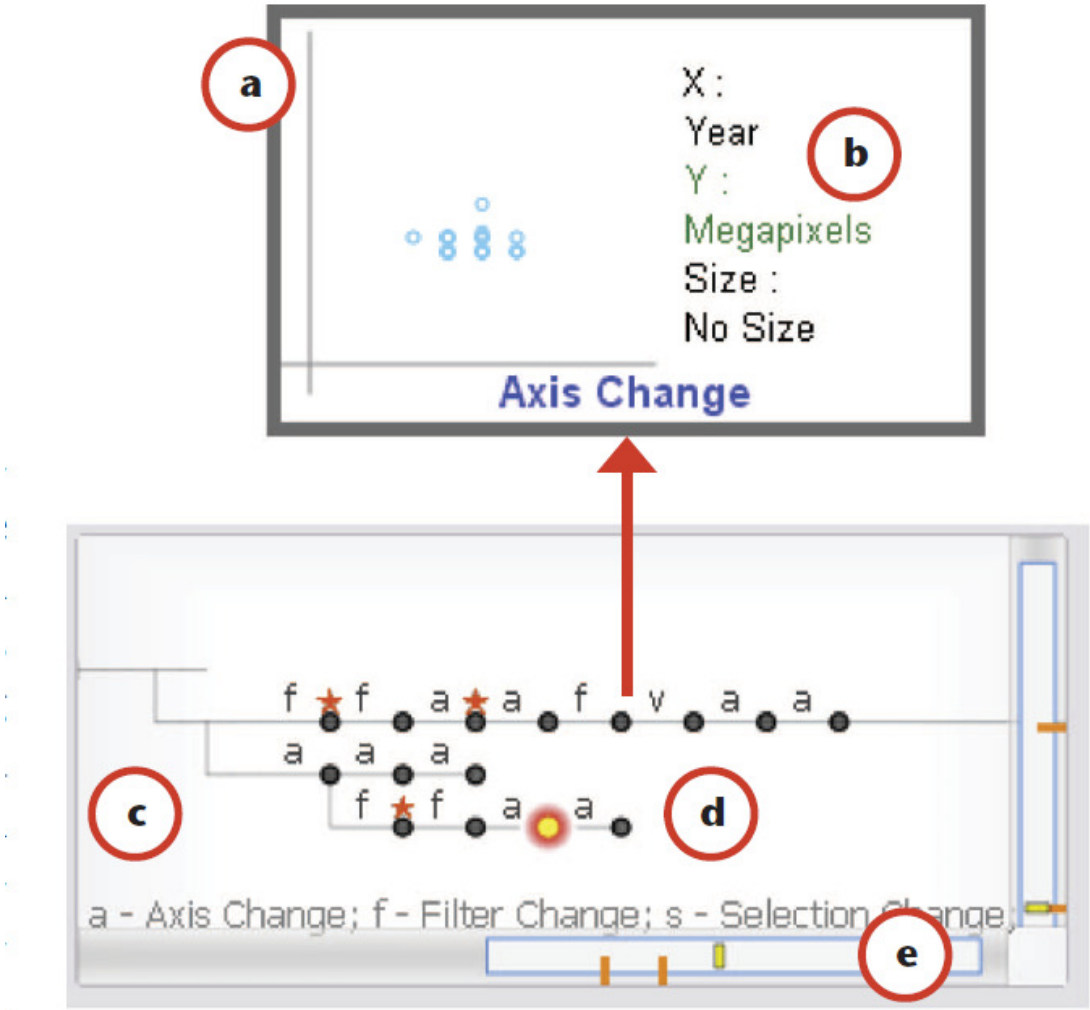


"Stories in GeoTime," in *VAST, 2007. VAST 2007. Eccles et al.*

(MANUAL) STRUCTURED DIAGRAMS



Shrinivasan and van Wijk. Supporting the Analytical Reasoning Process in Information Visualization. CHI 2008.

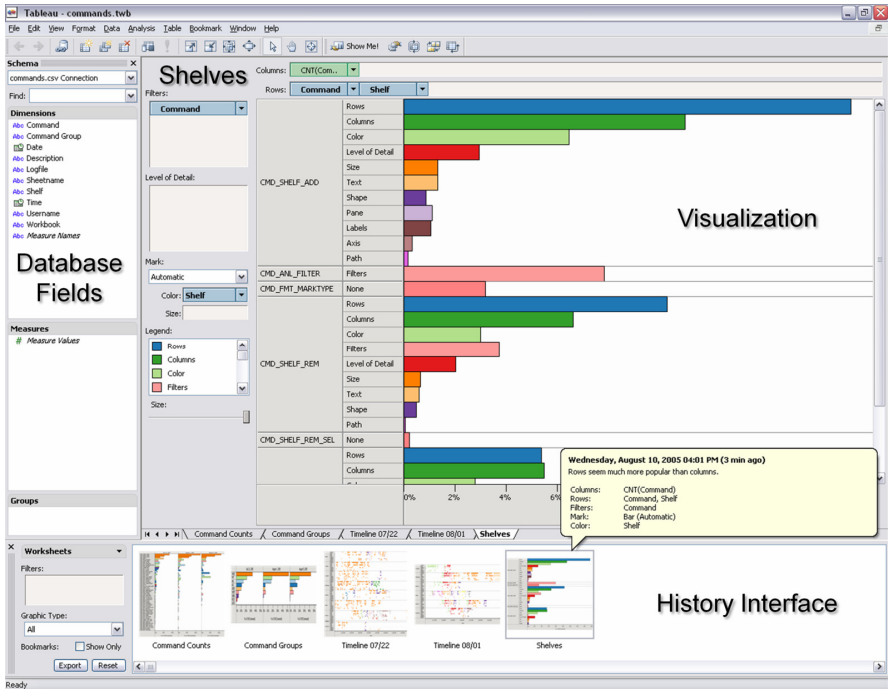

a
X :
Year
Y :
Megapixels
Size :
No Size
b
Axis Change
c
d
e
a - Axis Change; f - Filter Change; s - Selection Change

CAPTURING

Automatic Capturing

- Interactions: capture the mouse and key strokes
- Visualization States: capture the state of the visualization

VISUALIZATION STATE CAPTURING (TRANSITION)



Heer et al. Graphical Histories for Visualization: Supporting Analysis, Communication, and Evaluation. InfoVis 2008.

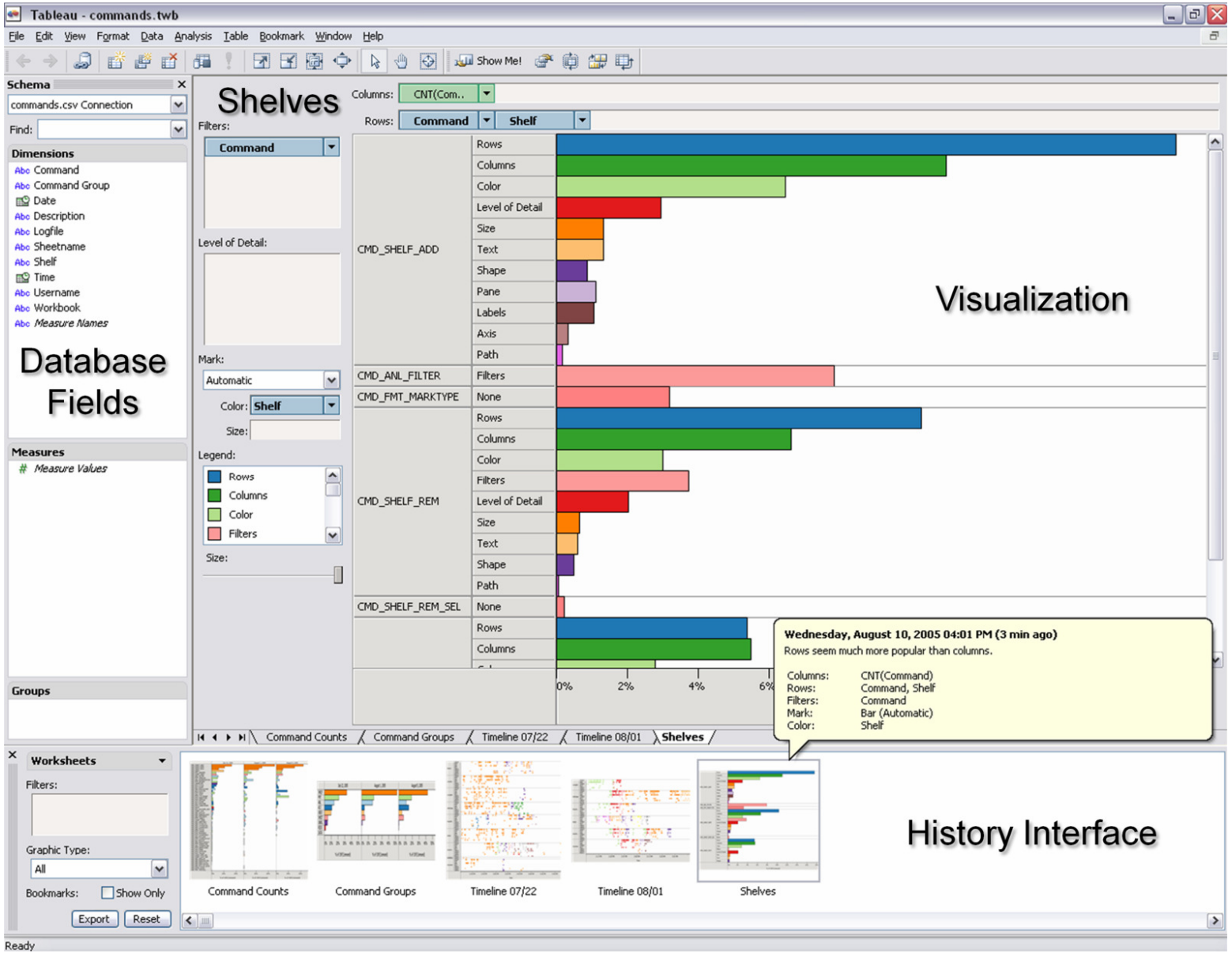
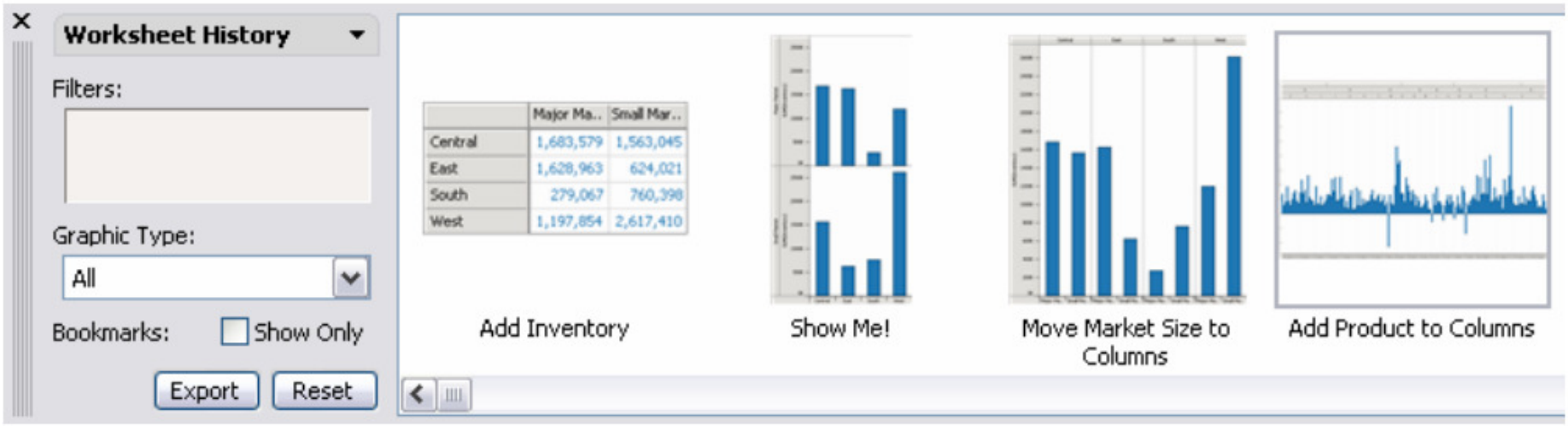


Tableau - commands.twb
File Edit View Format Data Analysis Table Bookmark Window Help
Show Me!
Schema
commands.csv Connection
Find:
Dimensions
Abc Command
Abc Command Group
Date
Abc Description
Abc Logfile
Abc Sheetname
Abc Shelf
Time
Abc Username
Abc Workbook
Abc Measure Names
Database Fields
Measures
Measure Values
Groups
Shelves
Columns: CNT(Com..
Rows: Command Shelf
Filters:
Command
Level of Detail:
Mark:
Automatic
Color: Shelf
Size:
Legend:
Rows
Columns
Color
Filters
Size:
CMD_SHELF_ADD
CMD_ANL_FILTER
CMD_FMT_MARKTYPE
CMD_SHELF_REM
CMD_SHELF_REM_SEL
Rows
Columns
Color
Level of Detail
Size
Text
Shape
Pane
Labels
Axis
Path
Filters
None
0%
2%
4%
6%
Visualization
Wednesday, August 10, 2005 04:01 PM (3 min ago)
Rows seem much more popular than columns.
Columns: CNT(Command)
Rows: Command, Shelf
Filters: Command
Mark: Bar (Automatic)
Color: Shelf
Command Counts
Command Groups
Timeline 07/22
Timeline 08/01
Shelves
Worksheets
Filters:
Graphic Type:
All
Bookmarks:
Show Only
Export
Reset
History Interface
Ready


Worksheet History
Filters:
Graphic Type:
All
Bookmarks:
Show Only
Export
Reset
Major Ma..
Small Mar..
Central 1,683,579 1,563,045
East 1,628,963 624,021
South 279,067 760,398
West 1,197,854 2,617,410
Add Inventory
Show Me!
Move Market Size to Columns
Add Product to Columns

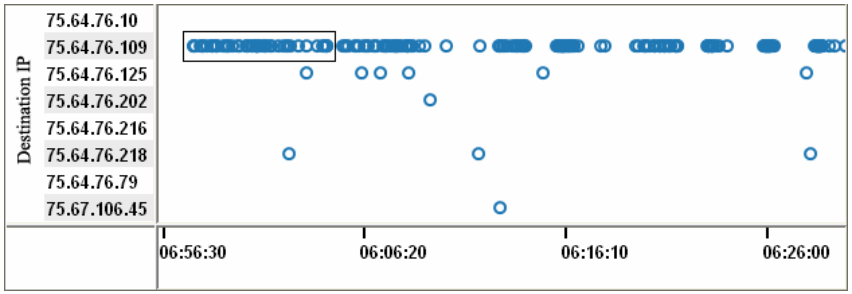
ENCODE

How do we store the captured interactions or visualization states?

- Encoding manually captured interactions: could be issues with different “languages”
- Encoding automatically captured interactions: more robust description of event sequences and patterns

ENCODING MANUAL CAPTURES (ONE EXAMPLE)

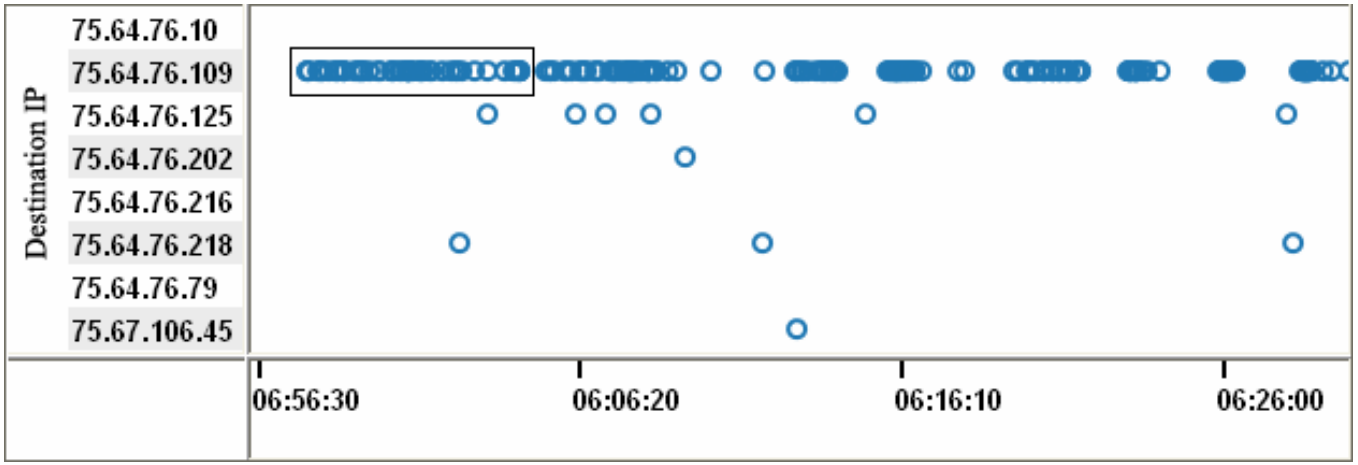
Network traffic visualization system
Analyst can create logical models of visual discoveries




```
WebCrawl(x1,x2,…) =
  time_sequence_30s(x1,x2,…) AND
  more_than_32_events(x1,x2,…) AND
  identical_source_AS_number(x1,x2,…) AND
  ( is_web_access_event(x1) AND
    is_web_access_event(x2) AND …)
```

Xiao et al. Enhancing Visual Analysis of Network Traffic Using a Knowledge Representation. VAST 2007.

ENCODING MANUAL CAPTURES



Here: HTTP requests from Google

1) select interesting pattern (burst)
2) system selects a set of predicates (from a list) that are true for these points

Xiao et al. Enhancing Visual Analysis of Network Traffic Using a Knowledge Representation. VAST 2007.

ENCODING MANUAL CAPTURES

destination_port_80, destination_Stanford, identical_source_asn, time_sequence_30s, time_sequence_60s, more_than_4_events, more_than_32_events

selected predicates

time_sequence_30s(x1,x2,…) AND
more_than_32_events(x1,x2,…) AND
identical_source_AS_number(x1,x2,…) AND
(is_web_access_event(x1) AND
is_web_access_event(x2) AND …)

analyst modifies list, adds conjunctions
and looks at visual feedback to see if pattern is correctly identified

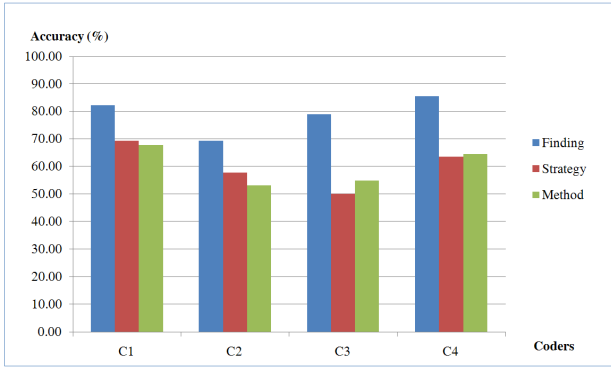
RECOVER

Given all the stored interactions, derive meaning, reasoning processes, and intent

- Manually: ask other humans to interpret a user's interactions
- Automatically: ask a computer to interpret a human's interactions

EXAMPLE: WIREVIS

MANUAL RECOVERY



From this experiment, we find that interactions contains at least:

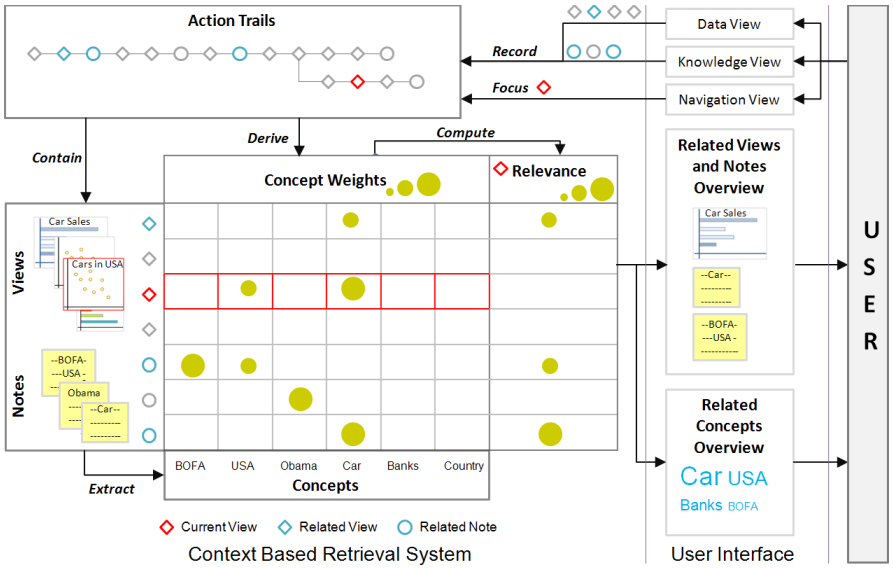
- 60% of the (high level) strategies
- 60% of the (mid level) methods
- 79% of the (low level) findings

R. Chang et al., Recovering Reasoning Process From User Interactions. IEEE Computer Graphics and Applications, 2009.
Jeong et al., Evaluating the Relationship Between User Interaction and Financial Visual Analysis. IEEE Symposium on VAST, 2008.

AUTOMATIC RECOVERY

Goal:

automatically identify notes, views, concepts from a user's past analyses that are most relevant to a view



Shrinivasan et al. Connecting the Dots in Visual Analysis. VAST 2009.

REUSE

Reapply the recovered user interactions, intent, reasoning process, etc. to a different dataset or problem

- Reuse user interactions: reapply the recorded interactions with some ability to recover from failures
- Reuse analysis patterns: reapply the “rules” learned from previous analysis

DISCUSSION

- Reuse is only applicable when some combinations of the previous stage(s) are successful
- More broadly speaking, does it make sense?
- (Familiar) example of reuse

PROVENANCE VS. REPRODUCIBILITY

PROVENANCE VS. REPRODUCIBILITY

- Goal of reproducibility: validate an analysis
  - by sharing data & code
- How can we validate a visual analysis?
  - by sharing interaction logs? by sharing manual analysis steps? …
  - how can this be done in a more general way across different GUI-based tools?

REPORTING

WHY?

- Sometimes you can't share code or even if you can:
  - make your analysis more understandable and reproducible by generating a good report
  - not everyone knows how to read code – so explain your analysis well

HOW TO MAKE A GOOD ANALYSIS REPORT

Adapt to your audience

- tl;dr – people are busy
- break it up into different levels of granularity

RESEARCH PAPER

- Title / Author List
- Abstract
- Body / Results
- Supplementary Materials
  (details details details!)
- Code / Data (even more details)

EMAIL

- Subject line / sender info
  - can you summarize findings in one sentence?
  - definitely add a subject line
- Email body
  - brief description of the problem / context
  - summarize findings / results (1-2 paragraphs)
  - if action necessary make concrete options
  - if needed try to make questions yes / no

EMAIL

- Attachment:
  - more detailed report
  - but stay concise
- Links to supplementary material
  - code / software / data
  - project website, repository (e.g. GitHub)

RESOURCES

- See scientific references on slides
- Reproducible Research MOOC Coursera.org (Roger Peng)

NEXT UP

AFTER THE BREAK

TUTORIAL 4 – REPRODUCIBLE RESEARCH IN R (+TABLEAU)

THIS AFTERNOON

YOUR PRESENTATIONS