REPRODUCIBLE RESEARCH

PROVENANCE

PETRA ISENBERG

VISUAL ANALYTICS

IN THIS LECTURE

YOU WILL LEARN ABOUT

- COMMUNICATING YOUR PROCESS
- DETAILS
- DIFFICULTIES

HOW TO CONVEY THE ANALYSIS PROCESS?

IN WORDS – TELL IT

PROVIDING DETAIL: CODE, DATA, …

GENERATE / WRITE REPORTS

WHY CONVEY THE ANALYSIS PROCESS?

SHOW YOUR FINDINGS ARE ROBUST

HIGHLIGHT SUBJECTIVITY

ENABLE IMPROVEMENTS

HELP SOMEONE LEARN ANALYZING

...

PROBLEMS

NOT EASY TO DESCRIBE

PEOPLE MAY NOT UNDERSTAND YOU

LONG ANALYSIS PIPELINES

LOTS OF TRIAL AND ERROR IN ANALYSIS

CONCEPTS

LETS FIRST DISCUSS TWO MAIN CONSIDERATIONS…

REPLICATION VS. REPRODUCIBILITY

REPLICATION

ABILITY OF AN ENTIRE EXPERIMENT / 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 WHY?

CHECK IF A FINDING IS ROBUST

IS THIS CLAIM TRUE?

ESPECIALLY IMPORTANT WHEN STUDIES HAVE BROAD IMPACT
(E.G. ON SOCIETY)

REPLICATION WHEN?

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?

Do Nothing $\longleftrightarrow$ Replication

IF YOU CAN'T REPLICATE?

WHAT ELSE CAN YOU DO?

LET A STUDY/AN ANALYSIS STAND BY ITSELF?

Do Nothing ⟵ Reproducibility ⟶ Replication

REPRODUCIBILITY

REPRODUCIBILITY

ASKS: CAN WE TRUST THIS ANALYSIS?

/SHOULD/ BE MIN STANDARD FOR ANY SCIENTIFIC STUDY

NEW INVESTIGATORS: SAME DATA, SAME METHODS

→ ALLOW FOR VALIDATION OF THE DATA ANALYSIS

WHY?

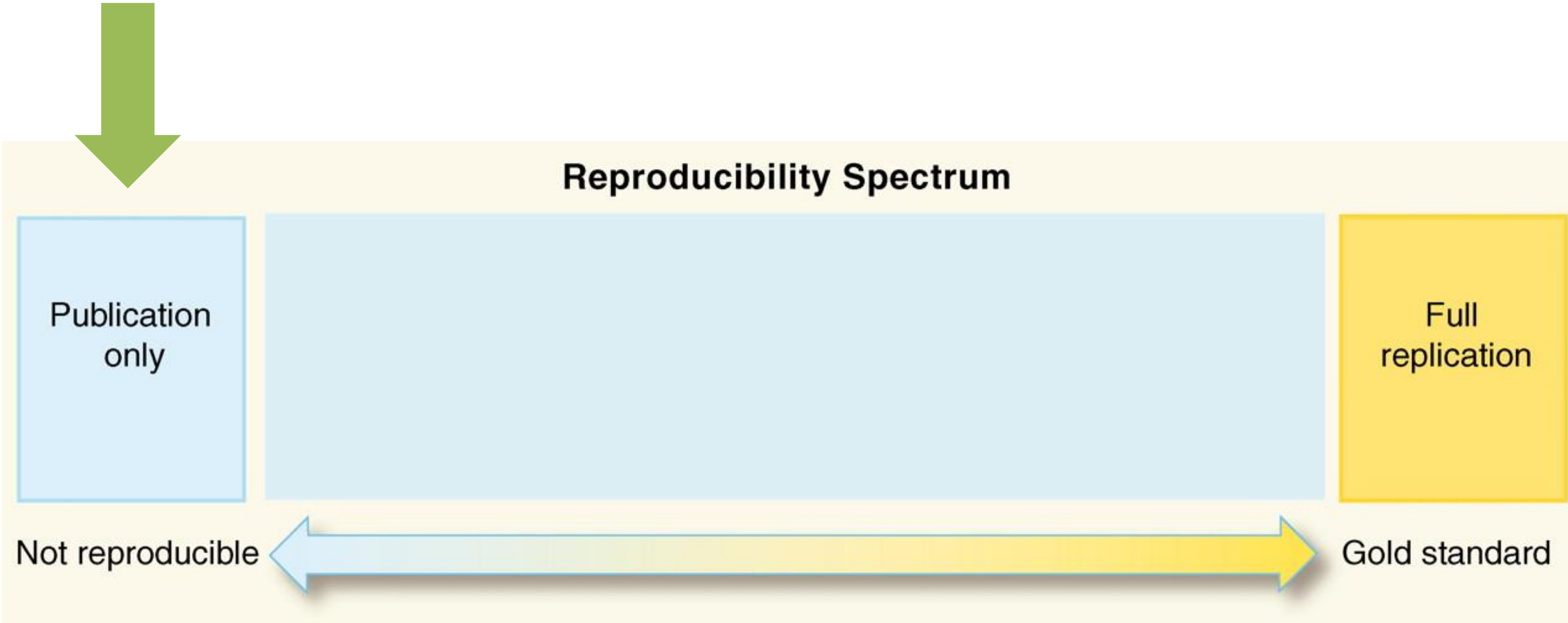
WHY?

ANOTHER VIDEO FOR YOU TO LOOK AT AT HOME

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

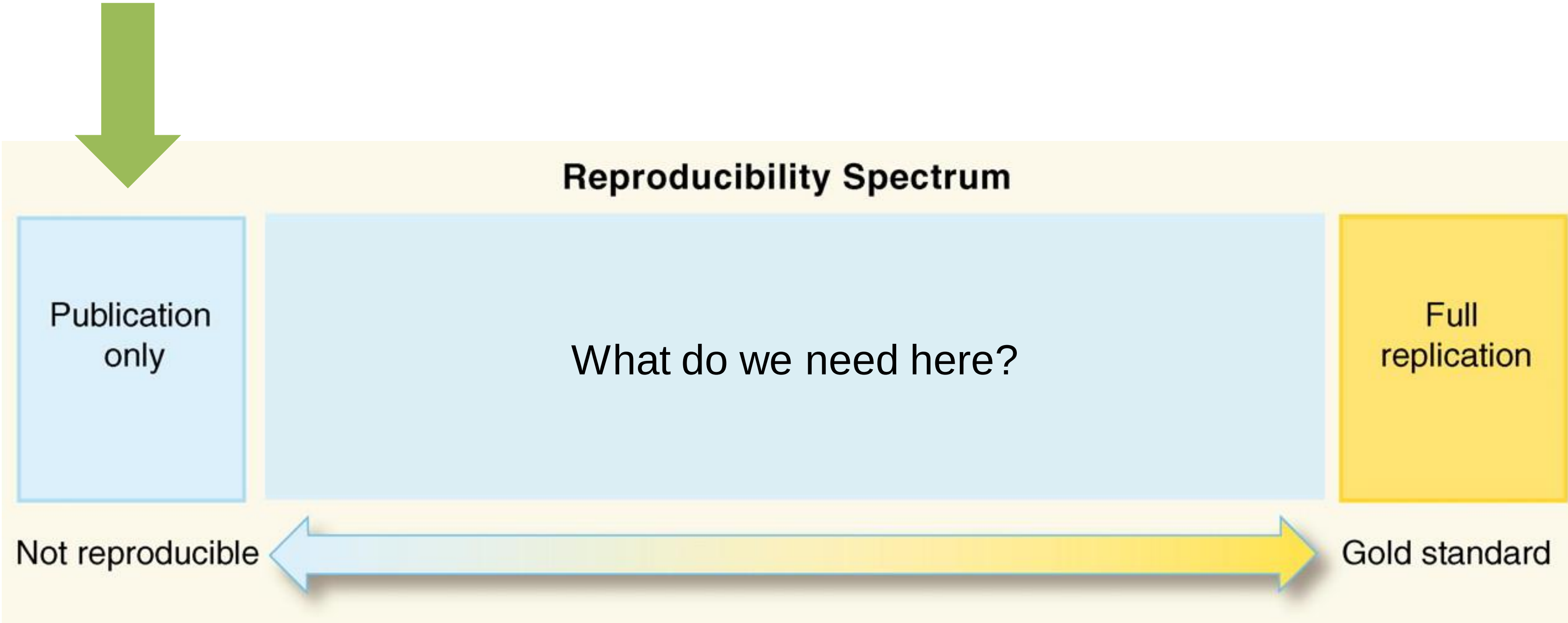
(“DECEPTION AT DUKE”)

ANALYSIS
(INCL. DATA COLLECTION, CLEANING, ANALYTIC METHODS, FIGURES, …)



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

ANALYSIS
(INCL. DATA COLLECTION, CLEANING, ANALYTIC METHODS, FIGURES, …)



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

WHAT TO DO?

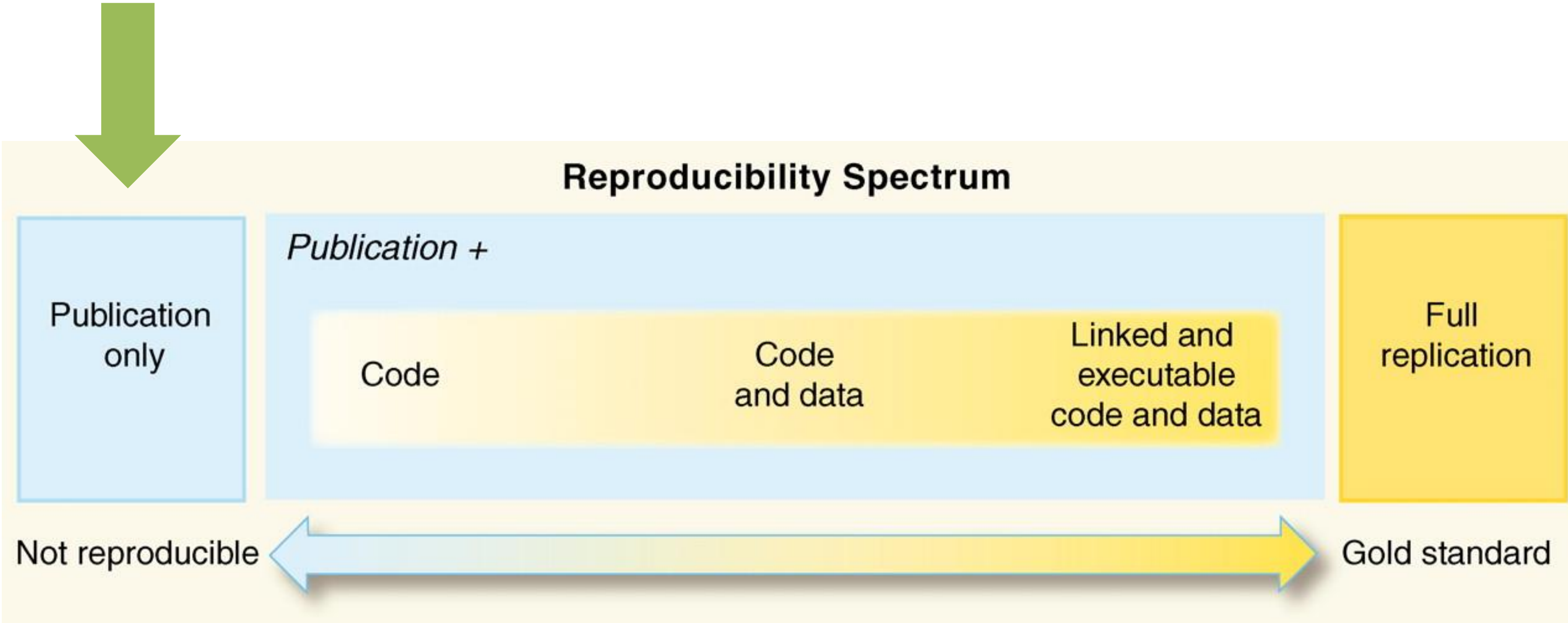
MAKE YOUR DATA AVAILABLE

MAKE YOUR ANALYSIS METHODS AVAILABLE

DOCUMENT CODE AND DATA

USE STANDARD MEANS OF DISTRIBUTION

ANALYSIS



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

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

- IF ACCESSED FROM WEB, INCLUDE URL, DESCRIPTION, AND DATE ACCESSED

PROCESSED DATA

REMEMBER YOUR DATA CLEANING EXERCISES?

SOMETIMES YOU NEED TO TRANSFORM DATA

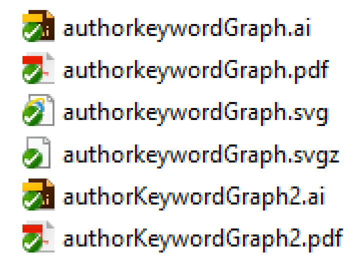
- NAME PROCESSED DATA TO KNOW WHICH SCRIPT GENERATED IT
- MAKE A README THAT SAYS WHICH SCRIPT/PROCEDURE GENERATED THE FILE
- PROCESSED DATA SHOULD BE READY FOR ANALYSIS

BAD EXAMPLE

coocurrence-author-level1.npy	3/20/2014 4:33 PM	NPY File	54,947 KB
coocurrence-author-level1-final clean.npy	3/20/2014 4:55 PM	NPY File	53,998 KB
coocurrence-author-level2.npy	5/7/2014 10:11 AM	NPY File	191 KB
coocurrence-PCS-all.npy	5/7/2014 10:11 AM	NPY File	127 KB
doc-term-level1.npy	3/20/2014 4:26 PM	NPY File	21,527 KB
doc-term-level1-final clean.npy	3/20/2014 4:48 PM	NPY File	21,341 KB
doc-term-level2.npy	5/7/2014 10:11 AM	NPY File	1,267 KB
equivalencematrix.npy	3/17/2014 1:06 PM	NPY File	54,039 KB
ieeecoocurrence.npy	2/11/2014 10:34 AM	NPY File	29,434 KB
inclusionmatrix.npy	3/17/2014 1:06 PM	NPY File	54,039 KB
inspec-controlled-coocurrence.npy	2/11/2014 1:31 PM	NPY File	10,369 KB
Matrix5.npy	3/10/2014 1:24 PM	NPY File	54,988 KB
Matrix5npy.npy	3/10/2014 12:46 PM	NPY File	54,988 KB
Matrix6.npy	3/10/2014 1:24 PM	NPY File	54,988 KB
Matrix6npy.npy	3/10/2014 11:52 AM	NPY File	54,988 KB
Matrix7.npy	3/10/2014 1:24 PM	NPY File	54,988 KB
Matrix7npy.npy	3/10/2014 11:52 AM	NPY File	54,988 KB
Matrix8.npy	3/10/2014 1:24 PM	NPY File	54,988 KB
Matrix8npy.npy	3/10/2014 11:52 AM	NPY File	54,988 KB
Matrix9.npy	3/10/2014 1:24 PM	NPY File	54,988 KB
Matrix9npy.npy	3/10/2014 11:52 AM	NPY File	54,988 KB

FIGURES

YOU WILL GENERATE MANY THAT YOU DON'T NEED



also name them properly

MAKE THE FINAL FIGURES PRETTY, 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
  - ONLY INCLUDE CODE FOR 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)

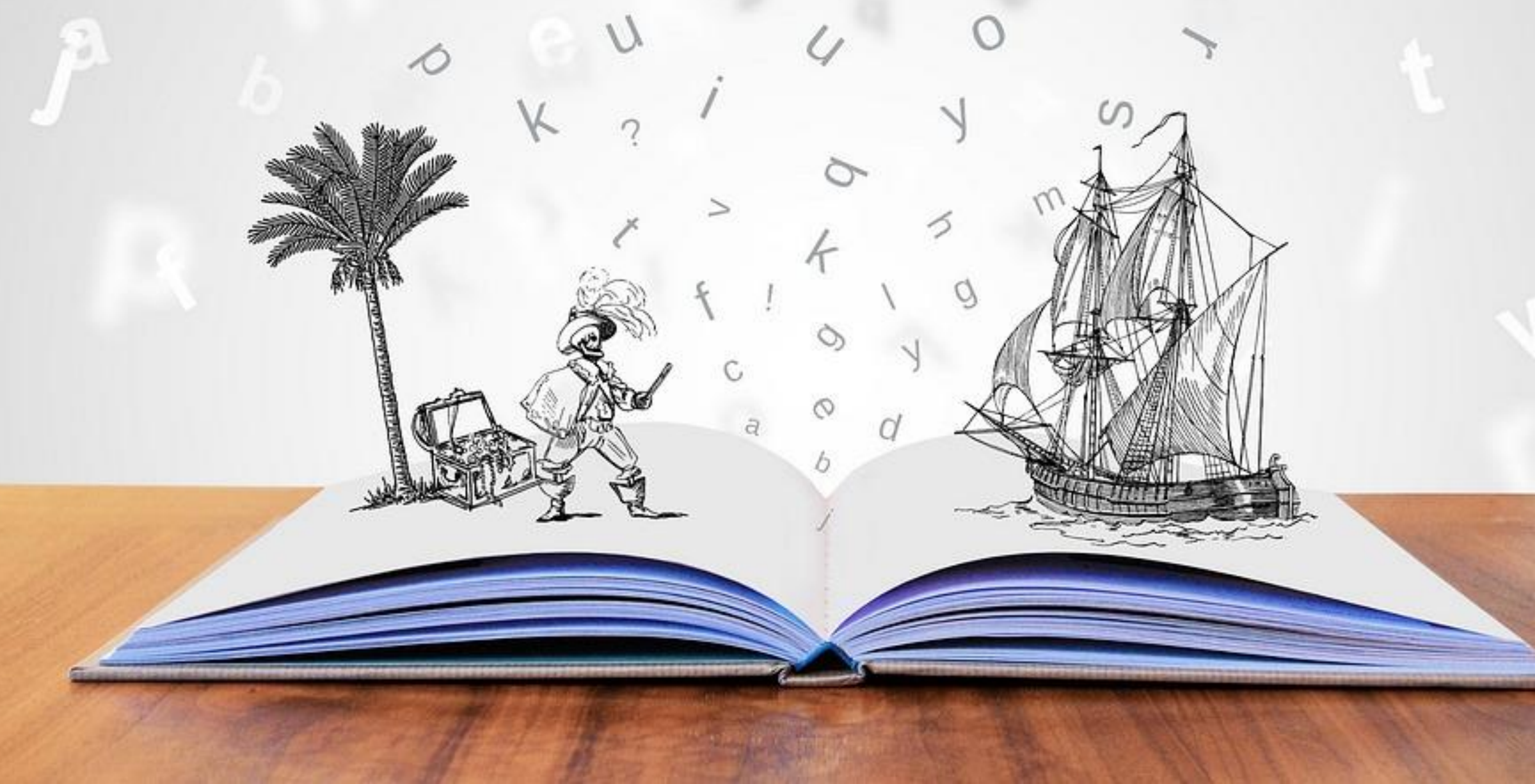
YOU WRITE CODE TO DO AN ANALYSIS

- COMPUTE RESULTS
- GENERATE DATA TABLES
- …

YOU ALSO WRITE A DOCUMENT – TEXT CHUNKS SURROUNDING YOUR ANALYSIS CODE

- EXPLAIN YOUR ANALYSIS
- FORMAT YOUR RESULTS

THINK OF IT AS STORYTELLING



LITERATE PROGRAMS

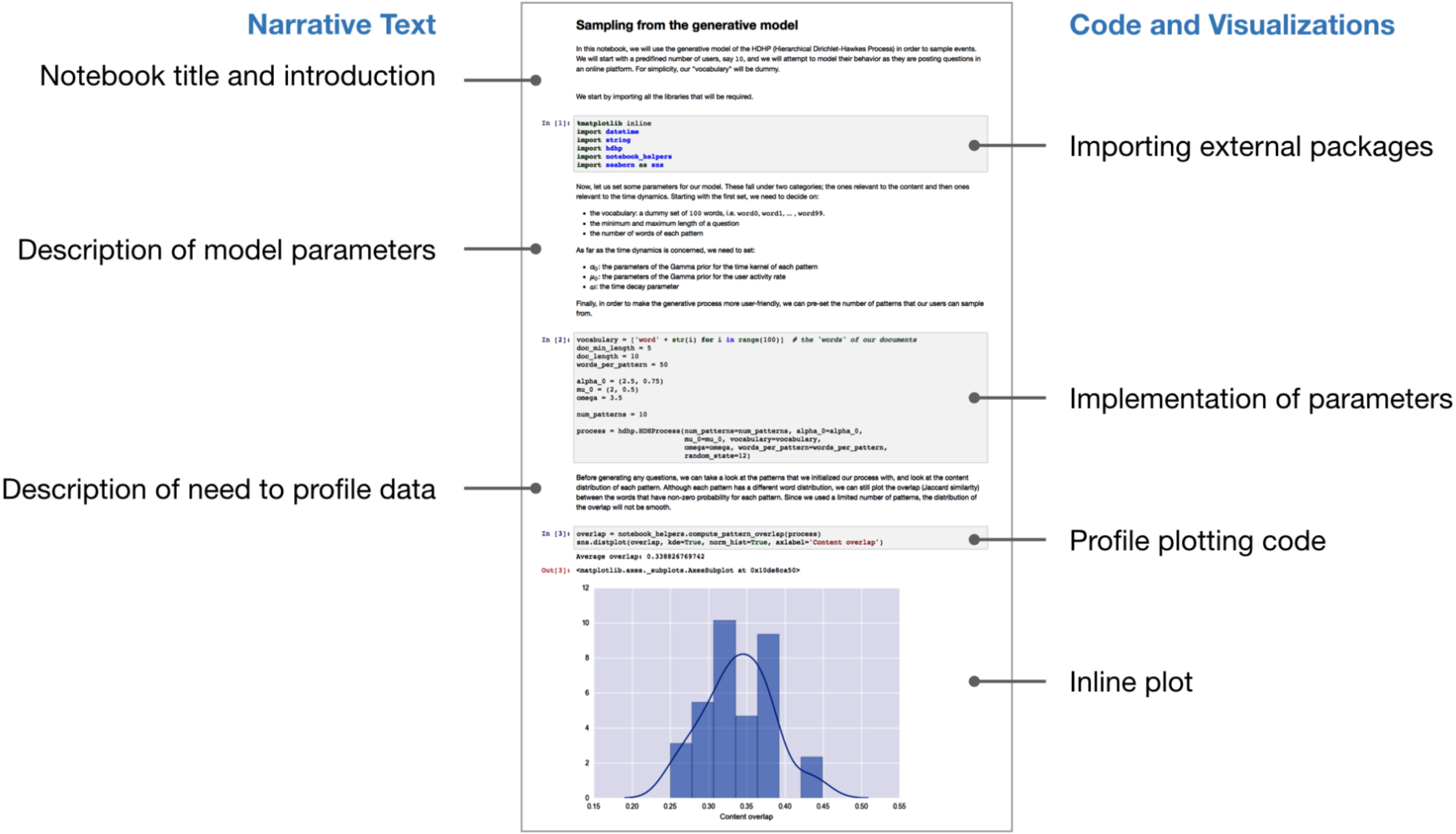
USE A DOCUMENTATION LANGUAGE
(HUMAN READABLE)

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

Adam Rule, Aurélien Tabard, James Hollan. Exploration and Explanation in Computational Note-books. ACM CHI Conference on Human Factors in Computing Systems, Apr 2018



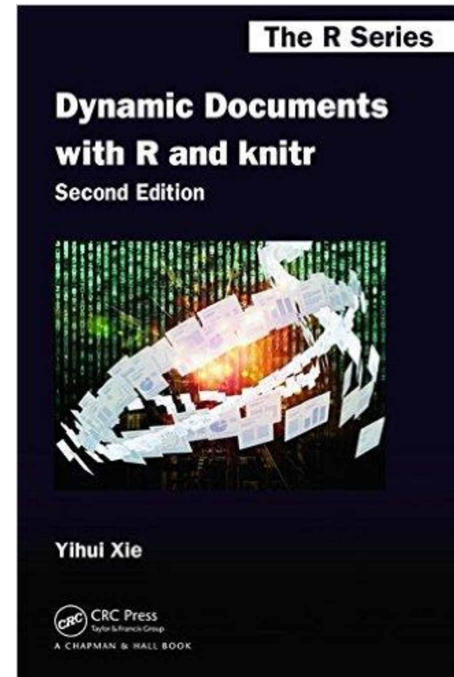
EXAMPLES

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
```

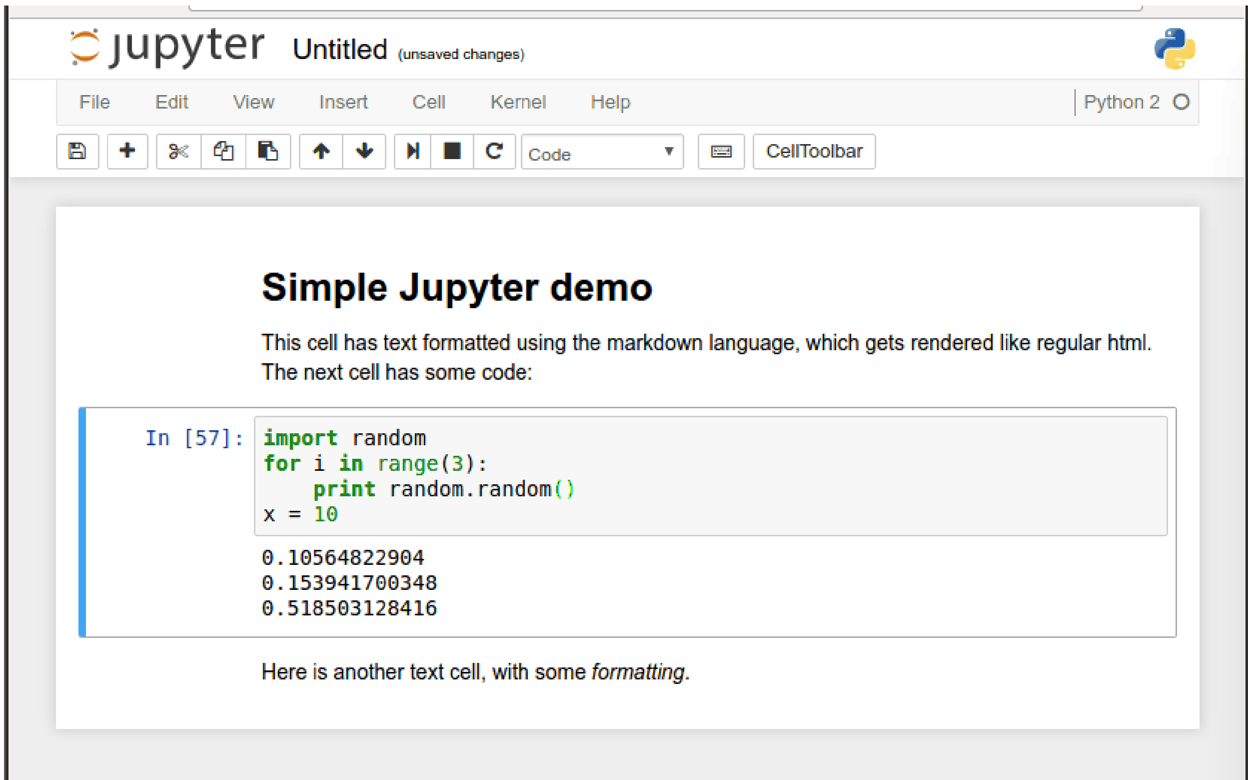
Next we will load our data from the data cleaning exercise

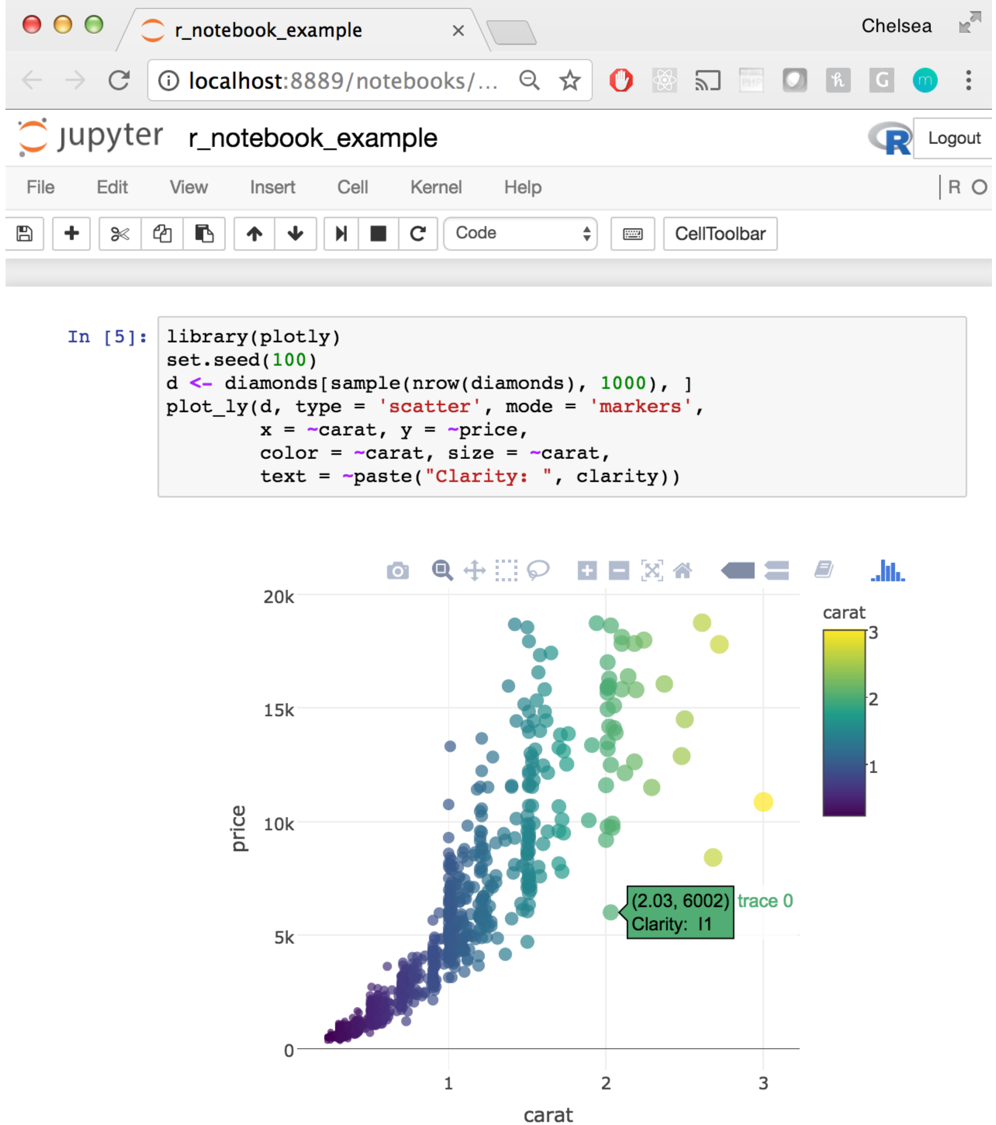
```
attractions <- read.csv("AttractionsOCR-txt.csv")
head(attractions)
```

```
##   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
```

EXAMPLES

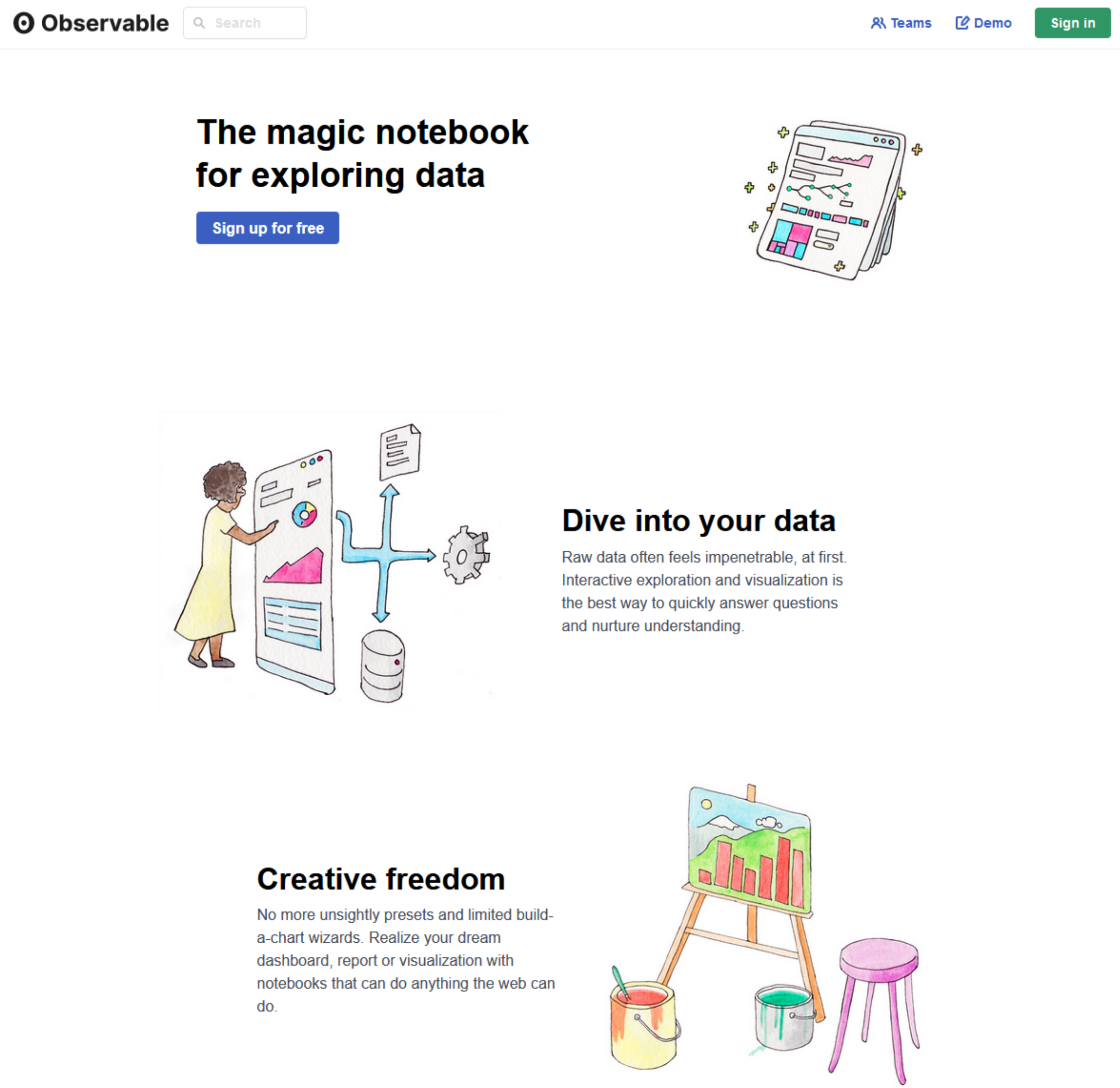
Jupyter Notebook

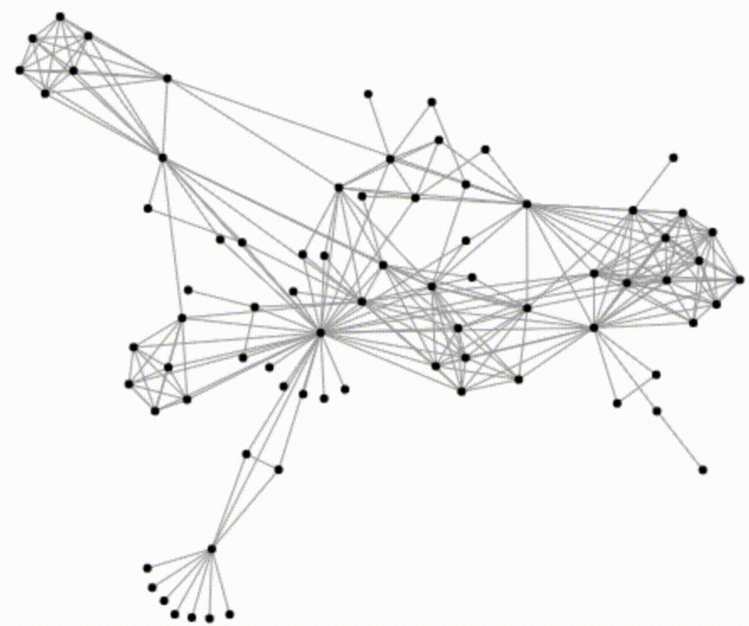



r_notebook_example
Chelsea
localhost:8889/notebooks/...
jupyter r_notebook_example
Logout
File
Edit
View
Insert
Cell
Kernel
Help
R
Code
CellToolbar
In [5]:
library(plotly)
set.seed(100)
d <- diamonds[sample(nrow(diamonds), 1000),]
plot_ly(d, type = 'scatter', mode = 'markers',
x = ~carat, y = ~price,
color = ~carat, size = ~carat,
text = ~paste("Clarity: ", clarity))
20k
15k
10k
5k
0
price
1
2
3
carat
carat
(2.03, 6002)
Clarity: I1
trace 0

EXAMPLES

https://observablehq.com/





```
simulation = d3.forceSimulation(data.nodes)
    .force("charge", d3.forceManyBody())
    .force("link", d3.forceLink(data.links).id(function(d) { return d.id; }))
    .force("center", d3.forceCenter(width / 2, height / 2))
```

https://medium.com/@mbostock/a-better-way-to-code-2b1d2876a3a0

EXAMPLES

https://www.gicentre.net/litvis

LITERATE VISUALIZATION

TELLING VISUALIZATION DESIGN STORIES

Openvis talk

Elm Europe talk

Paper *(IEEE VIS best paper honourable mention)*

Presentation slides from IEEE VIS 2018

litvis code, tutorials and examples

elm-vegalite

elm-vega

litvis.org

IEEE InfoVis 2018

Design Exposition with Literate Visualization

Jo Wood, *Member, IEEE*, Alexander Kachkaev and Jason Dykes

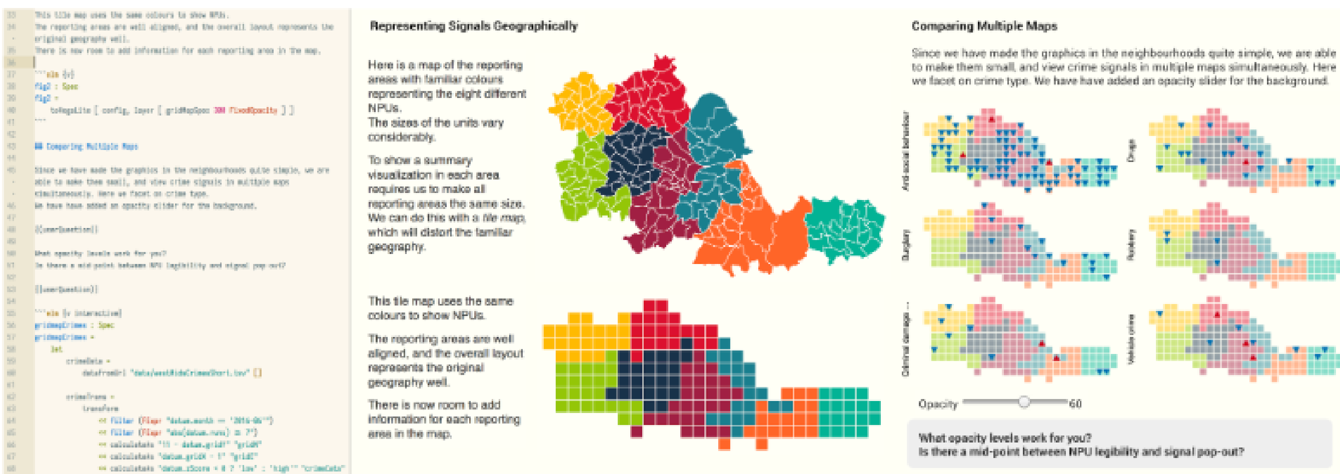


Fig. 1. Literate Visualization code (left) and output (centre and right) for a design exposition that elicits user feedback.

Abstract—We propose a new approach to the visualization design and communication process, *literate visualization*, based upon and extending, Donald Knuth's idea of literate programming. It integrates the process of writing data visualization code with description of the design choices that led to the implementation (design exposition). We develop a model of design exposition characterised by four visualization designer architypes: the evaluator, the autonomist, the didacticist and the rationalist. The model is used to justify the key characteristics of literate visualization: 'notebook' documents that integrate live coding input, rendered output and textual narrative;

DIFFERENCE

Focus on exposing design rationale (rather than analysis steps)

→ notebooks for designers

algebra.md ×

```
## Principle of Unambiguous Data Depiction

{( unambiguity |}

^^^elm {v=[(brexitMap Medium (LeaveBy 5) BySize Desc),(brexitMap Medium (LeaveBy 0) BySize Desc), (brexitMap Medium (LeaveBy -5) BySize Desc)]}^^^

{| unambiguity )}

{( unambiguityAssessment |}

Systematic shifts by 5% of votes cast towards leave or remain are easily detectable where they affect the majorty (shift between red and blue).
Systematic shifts that don't cross the 50% boundary are also detectable, although less obvious.
See for example size of blue circles in Scotland or red circles in Midlands/Northern England.
Therefore, no evidence for _confusers_ in design.

-   [x] passed?

{| unambiguityAssessment )}

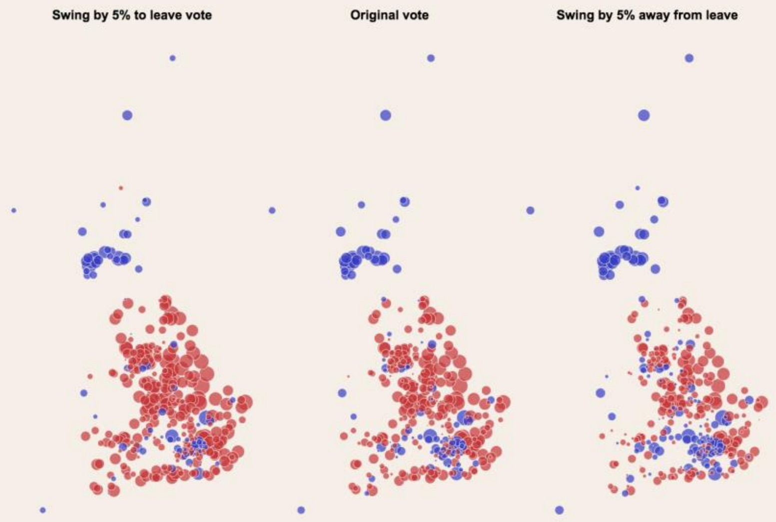
```elm {l=hidden}
brexitMap : MapSize -> DataChange -> OrderType -> SortOrder -> Spec
brexitMap mapSize dChange orderType oDirection =
    let
        orderParams =
            let
                sortOrder =
                    case oDirection of
                        Asc ->
                            oSort [ soAscending ]

                        Desc ->
                            oSort [ soDescending ]
                in
```

algebra.md preview ×

Principle of Unambiguous Data Depiction

$\omega = 1_V \Rightarrow \alpha = 1_D$: What are the smallest meaningful changes in data that should result in identifiable changes in the visualization?



Systematic shifts by 5% of votes cast towards leave or remain are easily detectable where they affect the majorty (shift between red and blue).
Systematic shifts that don't cross the 50% boundary are also detectable, although less obvious.
See for example size of blue circles in Scotland or red circles in Midlands/Northern England.
Therefore, no evidence for *confusers* in design.

- [x] passed?

Linter ×

Severity	Provider	Description	Line	File
Warning	Litvis	each description of visual-data correspondence should be followed by an assessment of it. (narrative-schema:rule-validation)	0:0	/Users/jwo/Dropbox/home_work_transfer/research/literateProgramming/litvisSandbox/algebra.md

~/Dropbox/home_work_transfer/research/literateProgramming/litvisSandbox/algebra.md ⓘ 0 ⚠ 1 ⓘ 0 64:1 • LF UTF-8 Markdown 0 files
```

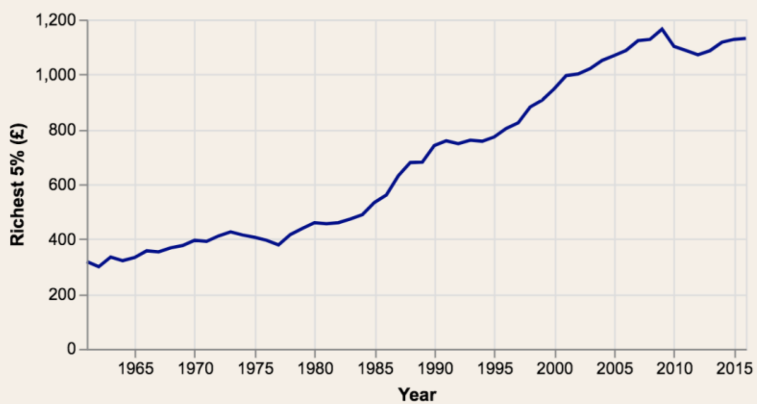
WHAT ARE DESIGN CHOICES?

An example…

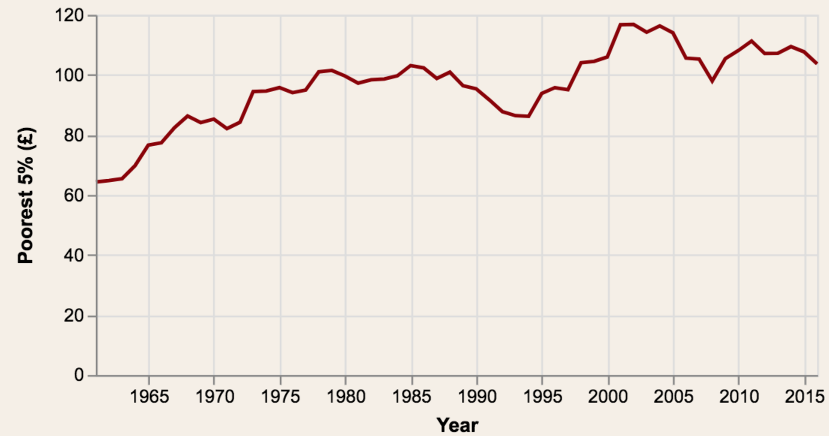
I want to plot income inequality and came up with this:

Income Inequality: Line Charts

Household income of the richest 5 percent after housing costs and adjusted for inflation:

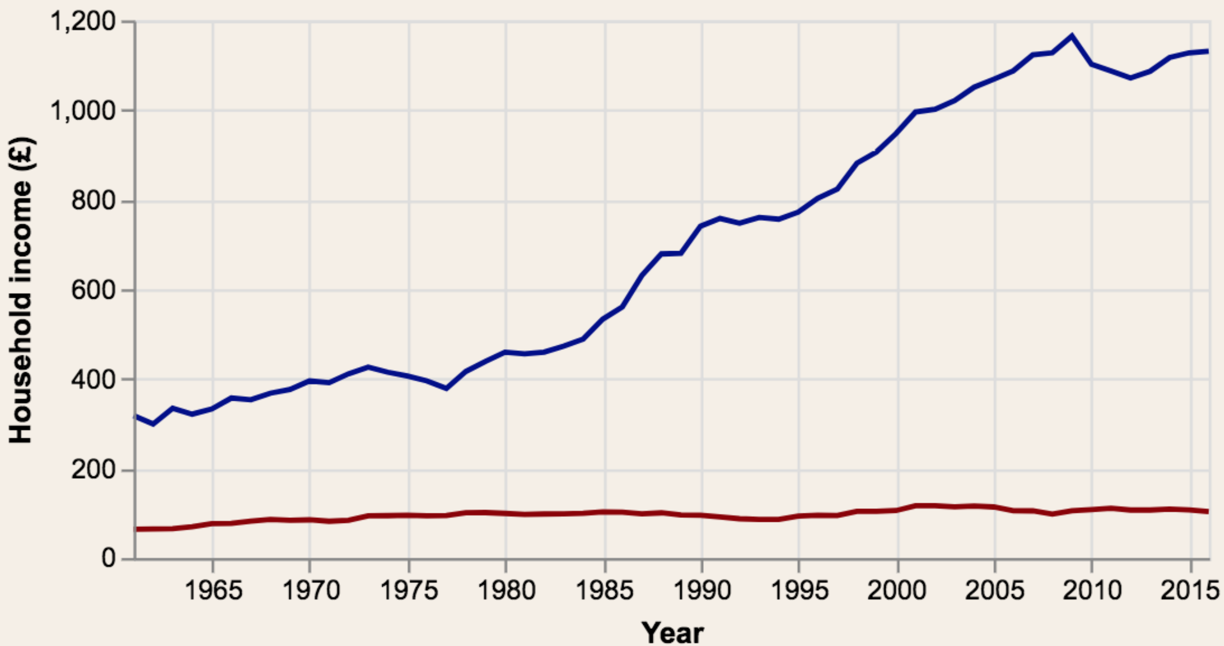


And the same for the poorest 5% (5th percentile):



Next try…

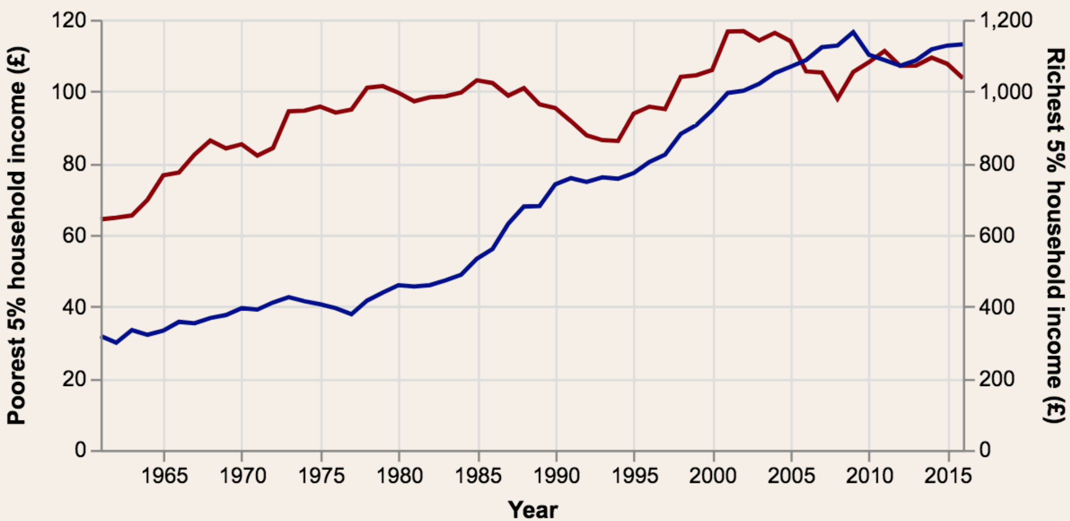
Comparison between the two is quite hard, so perhaps it would be easier on the same chart:



Next try…

Noting that the income of the richest 5% is an order of magnitude greater than the poorest 5%, while we can now compare both sets of figures, it is difficult to see any significant variation in the 5% line (in red).

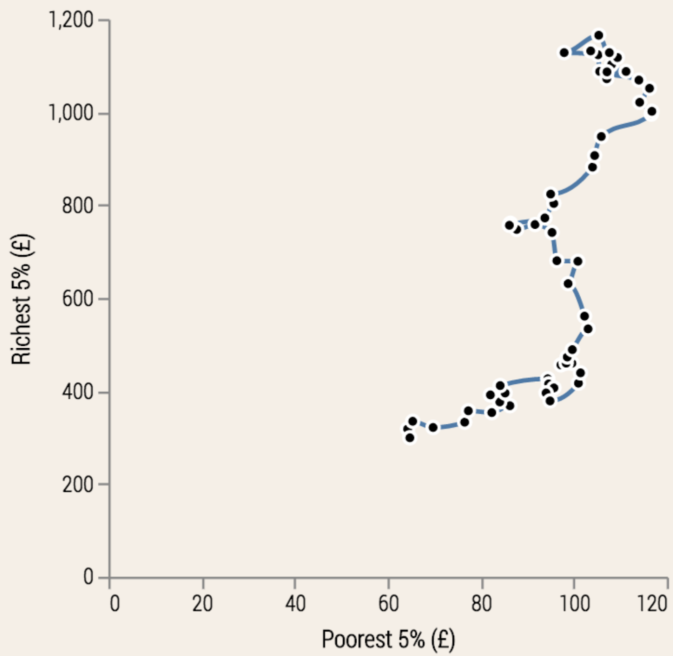
So perhaps it would be better to give each line its own scale on a dual-axis linechart:



Next try…

Income Inequality : Connected Scatterplots

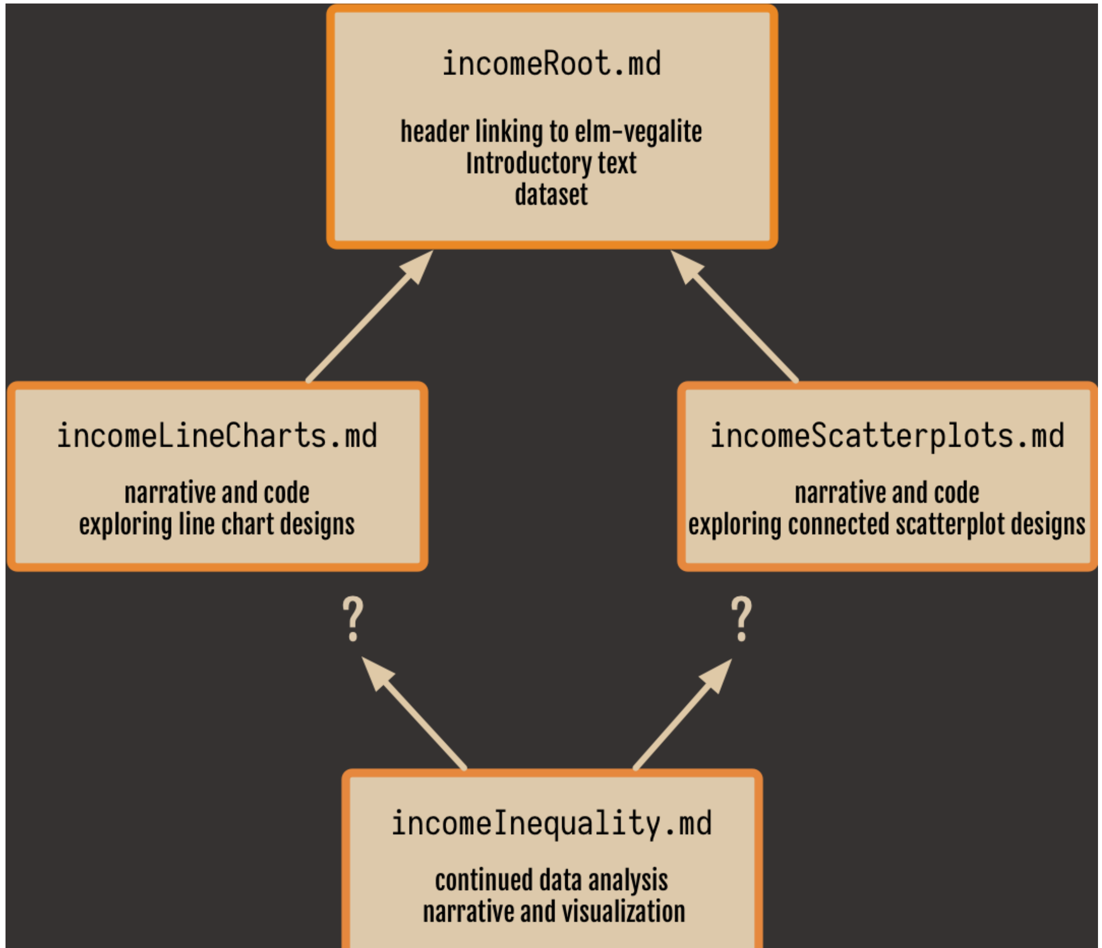
Rather than separate the 5% and 95% income quantiles, consider a connected scatterplot that joins the points in temporal order (1961 in bottom left, 2016 at top right):

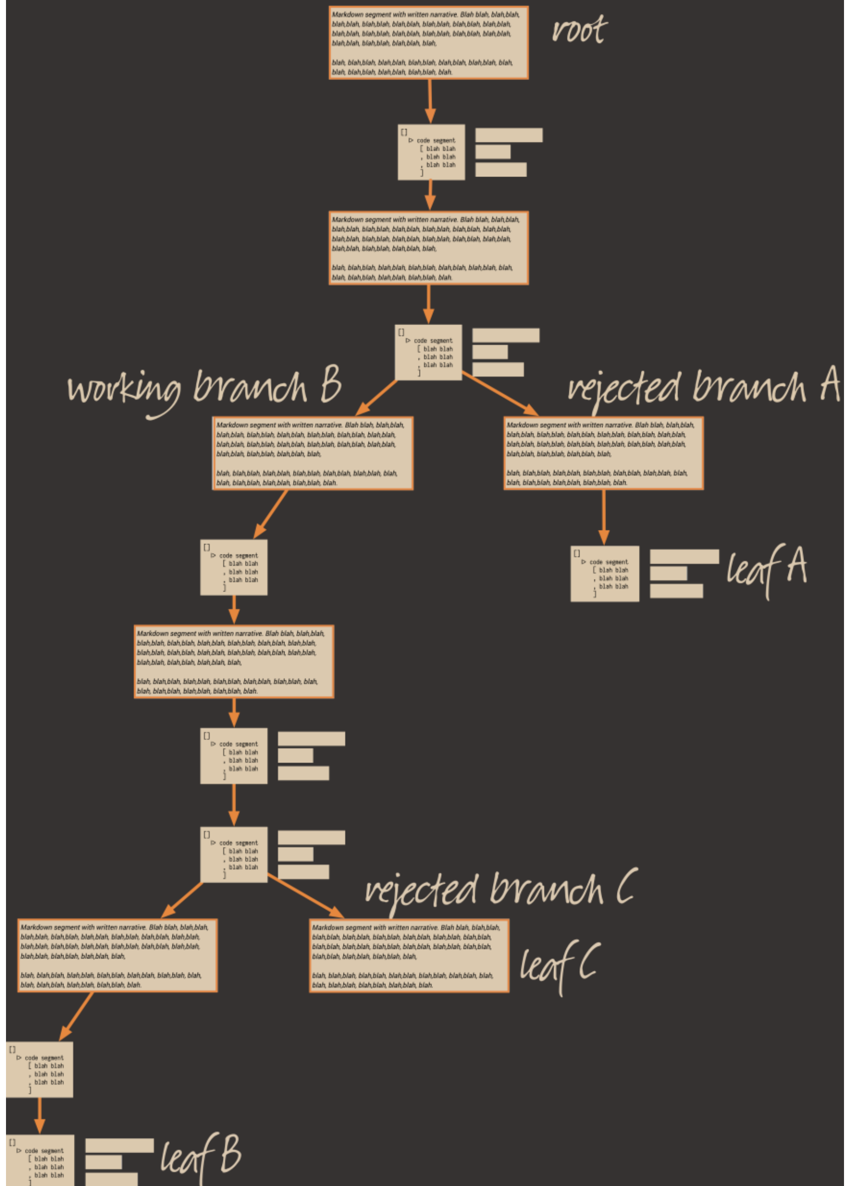


The plot still lacks important context (which dots refer to which years), so we can overlay some text labels indicating the year of every new Prime Minister:

Continue until you are satisfied with the chart

IDEA: CAPTURE ALL DESIGN ALTERNATIVES




root
working branch B
rejected branch A
leaf A
rejected branch C
leaf C
leaf B

codeReferencing.md — ~/litvisDemo

codeReferencing.md ×

````
---
elm:
  dependencies:
    gicentre/elm-vega: latest
---

```elm
import VegaLite exposing (..)
```

# Simple litvis chart

```elm
barChart : Spec
barChart =
    let
        data =
            dataFromUrl
            "https://vega.github.io/vega-lite/data/cars.json"
                []

        enc =
            encoding
                << position X [ PName "Horsepower", PmType
                Quantitative ]
                << position Y [ PmType Quantitative, PAggregate
                Count ]
    in
    toVegaLite [ data, enc [], mark Bar [] ]
```
````

codeReferencing.md pr...×

```
import VegaLite exposing (..)
```

Simple litvis chart

```
barChart : Spec
barChart =
    let
        data =
            dataFromUrl "https://vega.github.io/vega-lite/data/cars.json"
                []

        enc =
            encoding
                << position X [ PName "Horsepower", PmType Quantitative ]
                << position Y [ PmType Quantitative, PAggregate Count ]
    in
    toVegaLite [ data, enc [], mark Bar [] ]
```

+ × codeReferencing.md ⓘ 0 ⚠ 0 ⓘ 0 28:1 LF UTF-8 Markdown 0 files

“VALIDITY CRISIS” IN VISUALIZATION

[Wood et al., 2018]

How do we know the visual leads to the conclusions people draw?

How do design choices shape how we build our knowledge?

How do we learn from the visual design contributions of others?

+PROS

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

-CONS

DOCUMENTS CAN BECOME DIFFICULT TO READ

- WHEN THERE IS A LOT OF CODE

CAN BE SLOW

- BUT YOU CAN USE THINGS LIKE CACHING

IN PRACTICE...

Adam Rule, Aurélien Tabard, James Hollan. Exploration and Explanation in Computational Note-books. ACM CHI Conference on Human Factors in Computing Systems, Apr 2018

Exploration and Explanation in Computational Notebooks

Adam Rule
Design Lab, UC San Diego
La Jolla, CA
acrule@ucsd.edu

Aurélien Tabard
Univ Lyon, CNRS,
Université Lyon 1, LIRIS, UMR5205,
F-69622, Villeurbanne, France
aurelien.tabard@univ-lyon1.fr

James D. Hollan
Design Lab, UC San Diego
La Jolla, CA
hollan@ucsd.edu

ABSTRACT

Computational notebooks combine code, visualizations, and text in a single document. Researchers, data analysts, and even journalists are rapidly adopting this new medium. We present three studies of how they are using notebooks to document and share exploratory data analyses. In the first, we analyzed over 1 million computational notebooks on GitHub, finding that one in four had no explanatory text but consisted entirely of visualizations or code. In a second study, we examined over 200 academic computational notebooks, finding that although the vast majority described methods, only a minority discussed reasoning or results. In a third study, we interviewed 15 academic data analysts, finding that most considered computational notebooks personal, exploratory, and messy. Importantly, they typically used other media to share analyses. *These studies demon-*

tion. Analysts struggle to track which of the many versions of their code produced a particular result [11, 17]. Exploration often leads to dead-ends, prompting analysts to view code as being "throw-away" and see little point in annotating it [17]. Over time analysts produce dozens of similarly named scripts, figures, and files, which can be difficult to navigate [35]. Together, these factors complicate tracking and sharing of analyses, undermining replication and review.

Computational notebooks address these problems by combining code, visualizations, and text in a single document (Figure 1). While they have ties to Knuth's early work on literate programming [20], and have been available for decades in software such as Maple and Mathematica, the recent emergence of open-source computational notebooks has enabled rapid adoption by millions of researchers, data analysts, and journalists. Many users adopt computational we analyzed over **1 million computational notebooks** on GitHub, finding that one in four had **no explanatory text** but consisted entirely of visualizations or code

we examined over 200 academic computational notebooks, finding that although the vast majority described methods, only **a minority discussed reasoning or results**

REPRODUCIBILITY

What do we need to understand an analysis and its results?

WHAT ABOUT?

HUMAN PROCESSES SUCH AS

- INTERACTIONS WITH GUI SYSTEMS
- RESOURCE SHARING/COORDINATION
- INSIGHTS AND HYPOTHESES PRODUCED

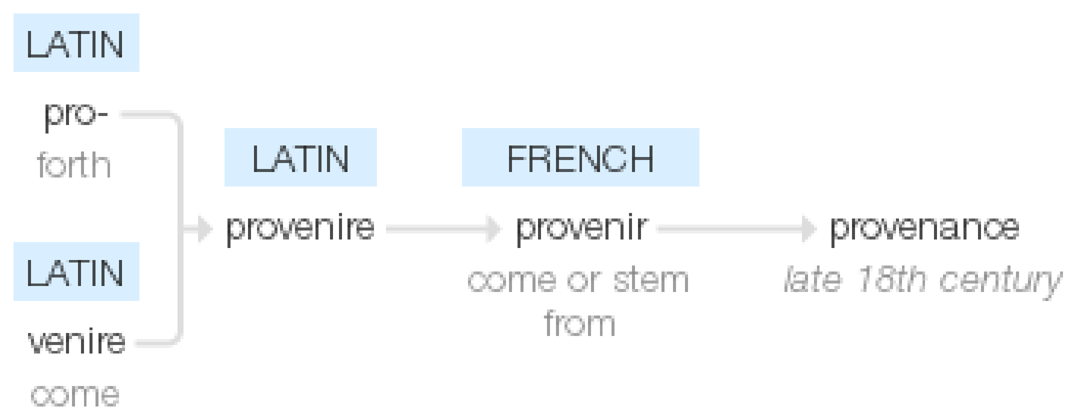
PROVENANCE

A broad concept of “history” in the analysis process

DEFINITION

"ORIGIN, SOURCE"

"THE HISTORY OF OWNERSHIP OF A VALUED OBJECT OR WORK OF ART OF LITERATURE"



source: Google

PROVENANCE IN VISUAL ANALYTICS

PROVENANCE OF:

- DATA
- VISUALIZATION
- INTERACTIONS
- INSIGHTS
- RATIONALE

Characterizing Provenance in Visualization and Data Analysis: An Organizational Framework of Provenance Types and Purposes

Eric D. Ragan, Alex Endert, Jibonananda Sanyal, and Jian Chen

Abstract— While the primary goal of visual analytics research is to improve the quality of insights and findings, a substantial amount of research in provenance has focused on the history of changes and advances throughout the analysis process. The term, *provenance*, has been used in a variety of ways to describe different types of records and histories related to visualization. The existing body of provenance research has grown to a point where the consolidation of design knowledge requires cross-referencing a variety of projects and studies spanning multiple domain areas. We present an organizational framework of the different *types* of provenance information and *purposes* for why they are desired in the field of visual analytics. Our organization is intended to serve as a framework to help researchers specify types of provenance and coordinate design knowledge across projects. We also discuss the relationships between these factors and the methods used to capture provenance information. In addition, our organization can be used to guide the selection of evaluation methodology and the comparison of study outcomes in provenance research.

Index Terms— Provenance, Analytic provenance, Visual analytics, Framework, Visualization, Conceptual model

1 Introduction

Data visualization and visual analytics combine the power of visualization with advanced data analytics to help people to better understand data and discover meaningful insights. While the goal of visualization research is ultimately to improve the quality of insights and findings, analytic processes are complicated activities involving technology, people, and real work environments. Practical applications encounter problems that extend beyond the integration of any system's analytic models, processing power, visualization designs, and interaction techniques. Visualization systems must also support human processes, which often involve non-standardized methodologies including extended or interrupted periods of analysis, resource sharing and coordination, collaborative work, presentation to different levels of management, and attempts at reproducible analyses [92, 52, 42].

For these reasons, a substantial amount of research in the areas of visualization, data science, and visual analytics has been dedicated to supporting *provenance*, which broadly includes consideration for the history of changes and advances throughout the analysis process (e.g., [34, 73, 37, 21]). It is clear that the research community agrees on the importance of supporting provenance, and many scholars have developed tools and systems that explicitly aim to help analysts record both computational workflows (e.g., [21, 5, 71]) and reasoning processes (e.g., [26, 37]). For example, *VisTrails* tracks steps of the computational workflow during scientific data analysis and visualization, and then provides graphical representations of the workflow through a combination of node diagrams and intermediary visual outputs [5, 14]. Groth and Streefkerk [39] presented another example with a system for recording and annotating stages of view manipulations during a 3D molecule-inspection task. As another example, Del Rio and da Silva [22] designed *Probe-It* to keep track of the data sets that contributed to the creation of map visualizations. Focusing on the provenance of insights, Gotz and Zhou described how the *HARVEST* system records the history of semantic actions during business and financial analysis activities [37]. These are just a few examples from a large number of visual analytics tools designed to support provenance across a wide range of domains and for different purposes.

As the body of research and existing tools has grown, the community's knowledge of the many factors and goals relevant for effective provenance support has also broadened. However, the variety of perspectives can make it challenging to assess the specific aspects and purposes of provenance that are targeted by any particular project. The term, *provenance*, has been used in a variety of ways to describe different types of origins and histories. For example, the scientific visualization community, especially the simulation and modeling communities, often interpret provenance as the history of computational workflow (e.g., [34]), while other interpretations focus on the history of insights and hypotheses (e.g., [70]). Although many researchers proactively provide clear definitions and explanations of their foci in the provenance research, this does not entirely resolve the challenge of consolidating the variety of interpretations and research outcomes across projects. Different perspectives and applications of concepts become problematic for interpreting and coordinating outcomes from different provenance projects, for communicating ideas within the visualization community, and for allowing new-comers to clearly understand the research space. In our work, we analyzed the different perspectives of provenance that are most relevant to areas of visualization and data analysis.

Our goal in this paper is to organize the different *types* of provenance information and *purposes* for why they are desired in information visualization, scientific visualization, and visual analytics. We present an *organizational framework* as a conceptual model that categorizes and describes the primary components of provenance types and purposes. Further, we discuss the relationships between these factors and considerations when capturing provenance information. Our organizational framework is intended to help researchers specify types of provenance and coordinate design knowledge across projects. In addition, our organization can be used to guide the selection of evaluation methodology and the comparison of study outcomes in provenance research.

- *Eric D. Ragan is with Texas A&M University. E-mail: ragane@acm.org.*
- *Alex Endert is with Georgia Tech. E-mail: endert@gatech.edu.*
- *Jibonananda Sanyal is with Oak Ridge National Laboratory. E-mail: sanyalj@ornl.gov.*
- *Jian Chen is with University of Maryland, Baltimore County. E-mail: jichen@umbc.edu.*

Manuscript received 31 Mar. 2015; accepted 1 Aug. 2015; date of publication xx Aug. 2015; date of current version 25 Oct. 2015. For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.

2 Existing Perspectives of Provenance

Analytic provenance is a broad and complex concept within the areas of information visualization, data analysis, and data science. In visual data analysis, the concept often includes aspects of the cognitive and interactive processes of discovery and exploration, and also the computational sequences and states traversed to arrive at findings or insights. Prior surveys have presented definitions, categorizations,

Ragan et al., VAST 2015

PROVENANCE OF DATA

HISTORY OF CHANGES AND MOVEMENT OF DATA

- SUBSETTING, MERGING, FORMATTING,…

COUPLED WITH WORKFLOWS

- CAPTURES ACTIONS ON DATA

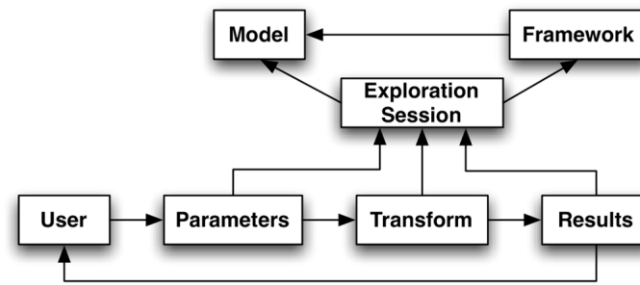
PROVENANCE OF VISUALIZATION

HISTORY OF GRAPHICAL VIEWS AND VISUALIZATION STATES

SAVE SCREENSHOTS OR PARAMETERS TO RECREATE VIEWS/STATES

VISUALIZATION STATES

DESCRIBE VISUALIZATION AS CHAIN OF VISUAL ENCODING OPERATORS

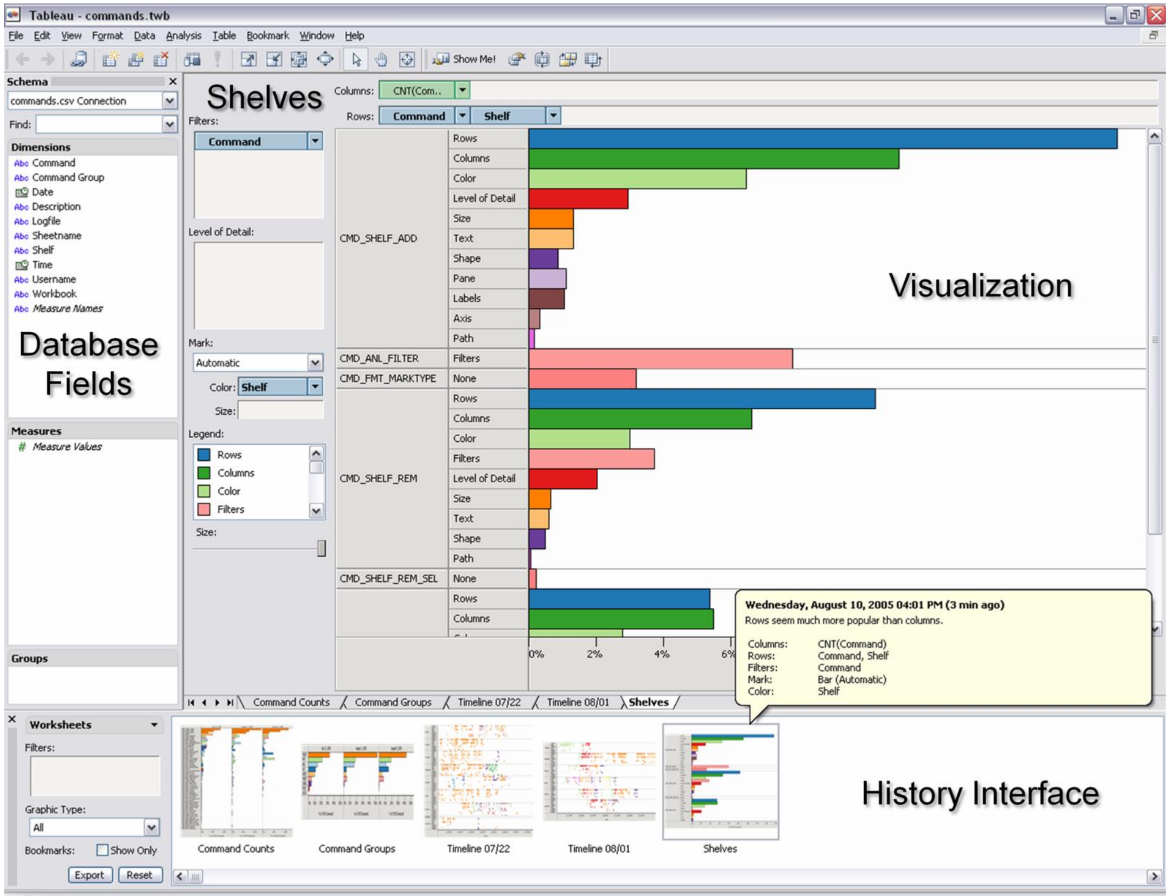


P-SET MODEL:

STATE = SET OF PARAMETERS & ACTIONS AS TRANSFORMATIONS OF THESE PARAMETERS

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

VISUALIZATION STATES



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

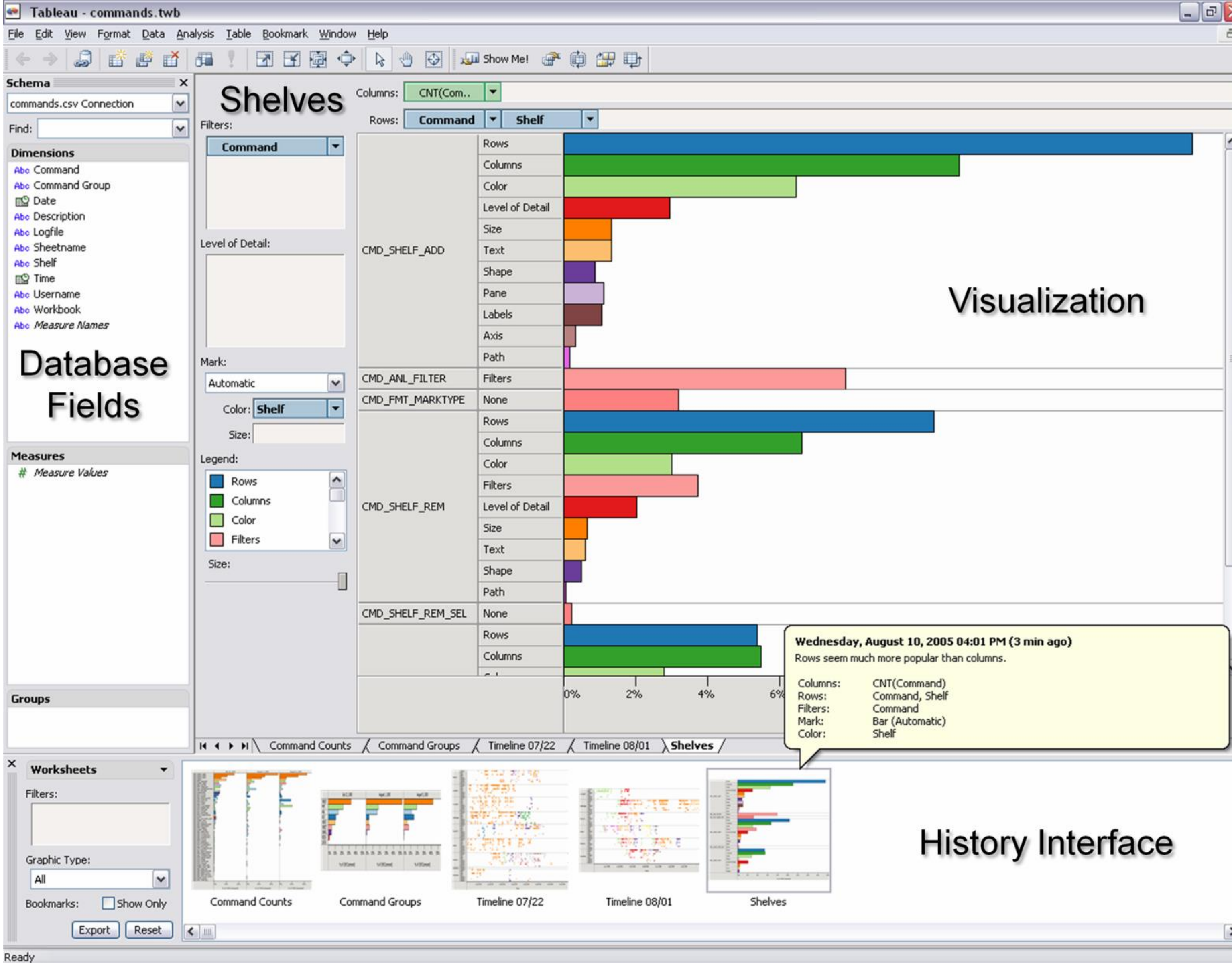
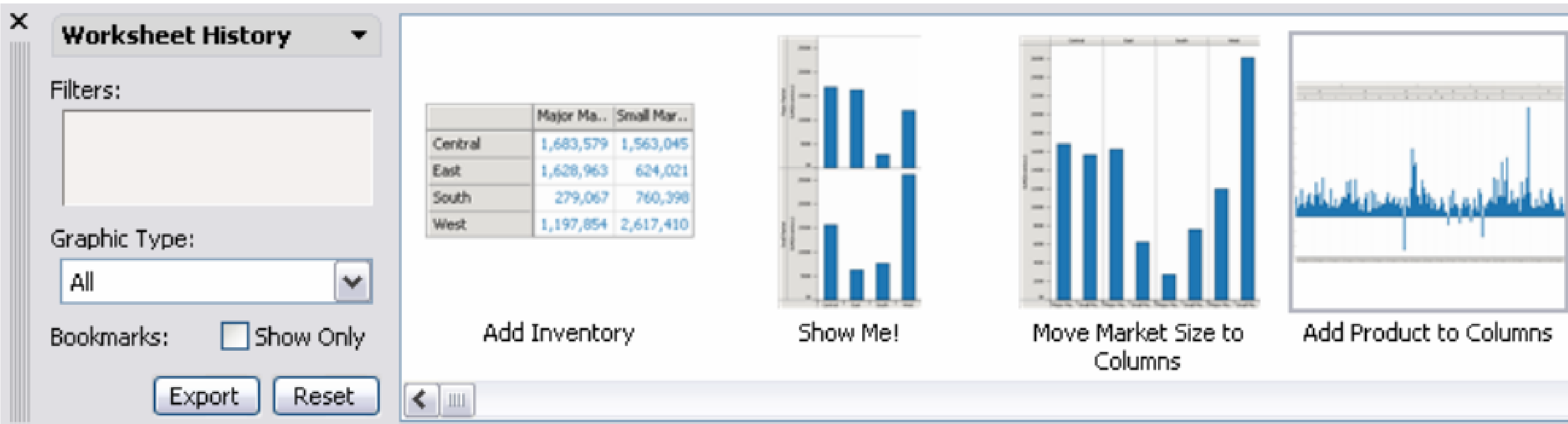
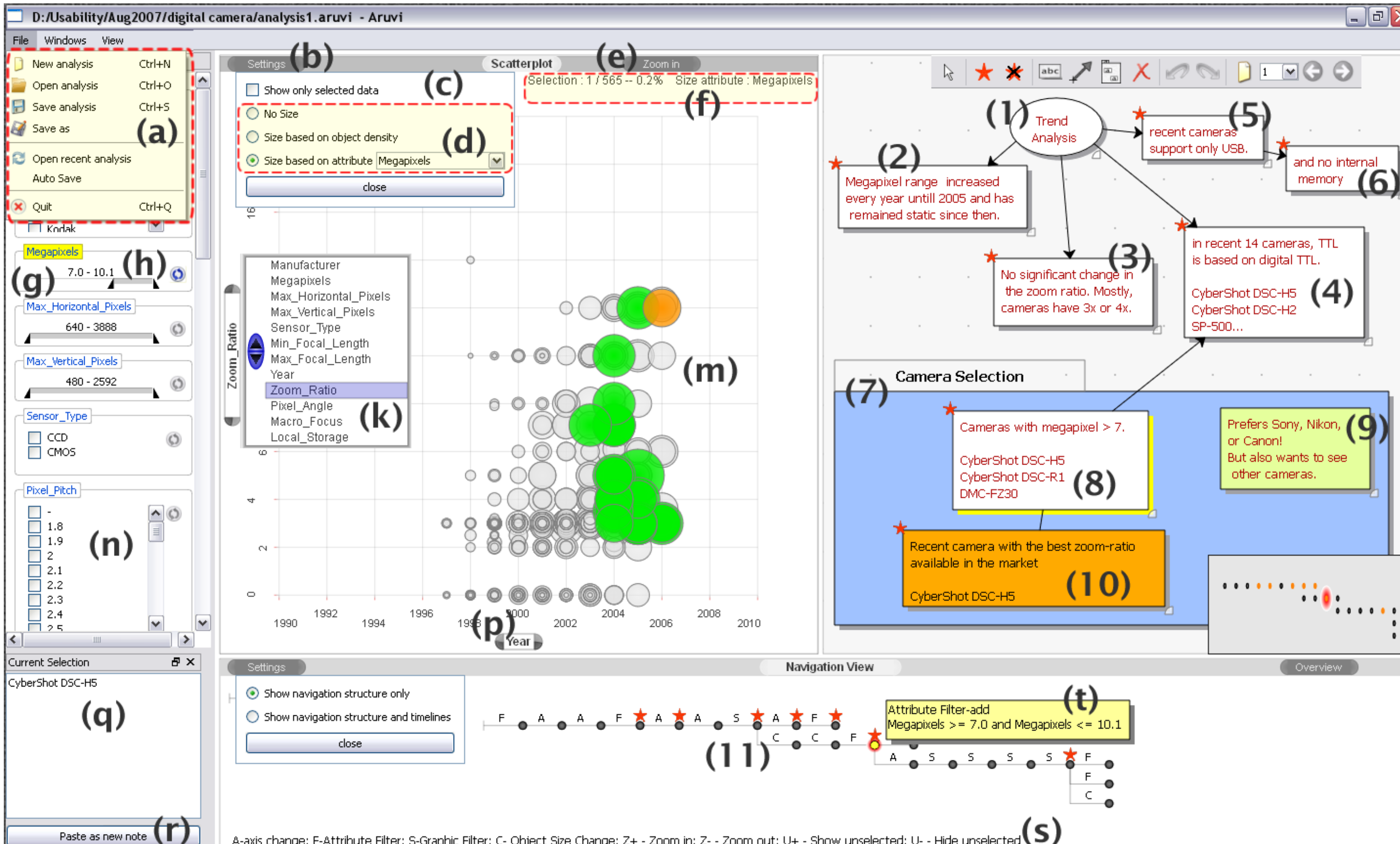


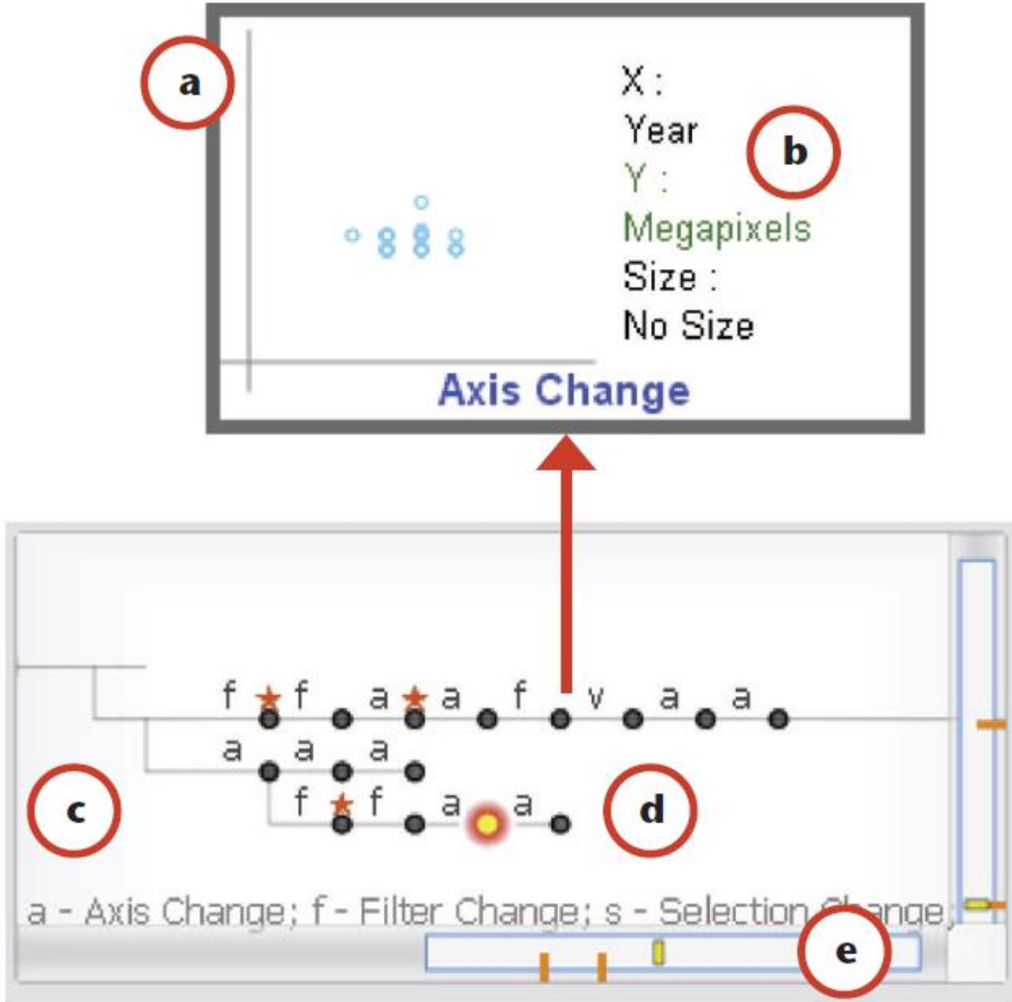
Tableau - commands.twb
File Edit View Format Data Analysis Table Bookmark Window Help
Show Me!
Schema
commands.csv Connection
Find:
Dimensions
Command
Command Group
Date
Description
Logfile
Sheetname
Shelf
Time
Username
Workbook
Measure Names
Database Fields
Measures
Measure Values
Groups
Shelves
Columns:
CNT(Com..
Rows:
Command
Shelf
Filters:
Command
Level of Detail:
Mark:
Automatic
Color:
Shelf
Size:
Legend:
Rows
Columns
Color
Filters
Size:
CMD_SHELF_ADD
Rows
Columns
Color
Level of Detail
Size
Text
Shape
Pane
Labels
Axis
Path
CMD_ANL_FILTER
Filters
CMD_FMT_MARKTYPE
None
CMD_SHELF_REM
Rows
Columns
Color
Filters
Level of Detail
Size
Text
Shape
Path
CMD_SHELF_REM_SEL
None
Rows
Columns
Visualization
0%
2%
4%
6%
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
Command Counts
Command Groups
Timeline 07/22
Timeline 08/01
Shelves
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

VISUALIZATION STATES



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


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

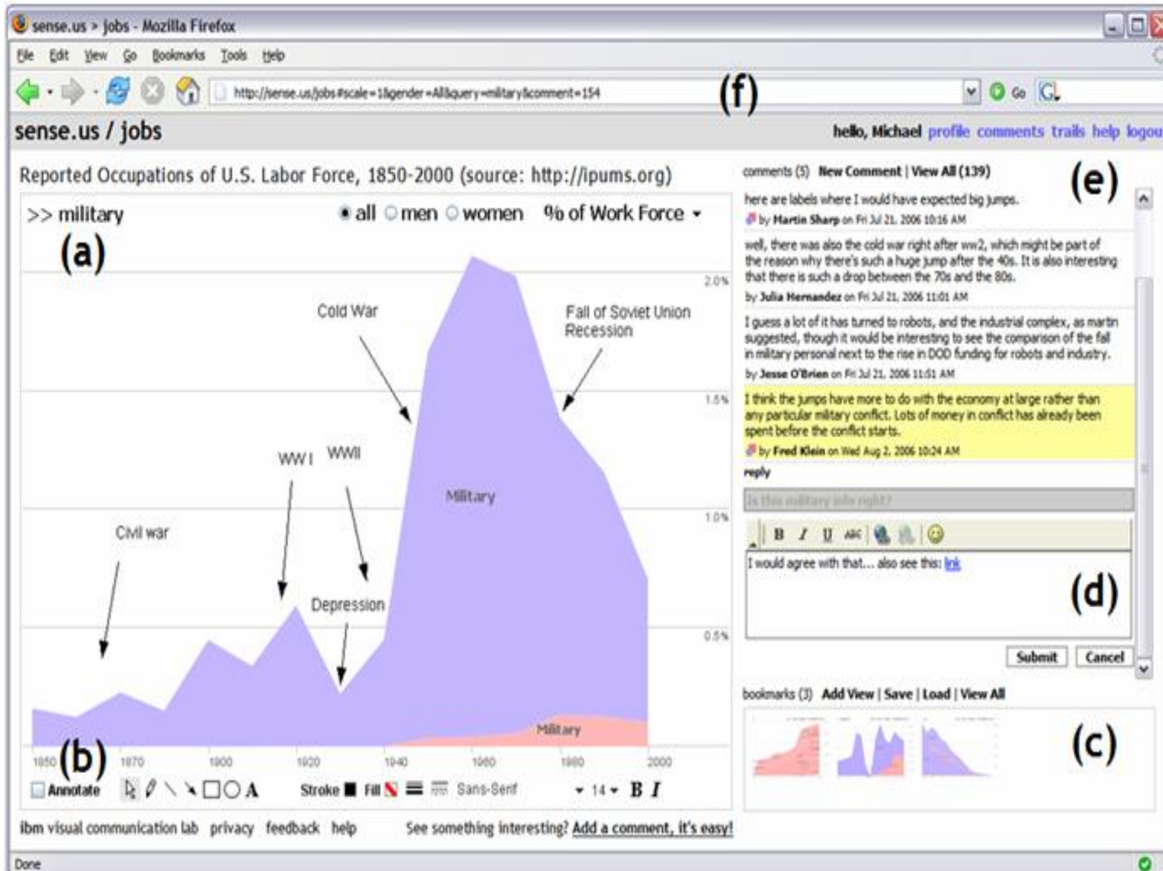
PROVENANCE OF INTERACTIONS WITH A GUI/VIS

HISTORY OF USER INTERACTIONS/COMMANDS

INCLUDES

- DATA EXPLORATION INTERACTION (E.G. QUERIES)
- ANNOTATION INTERACTIONS
- COMMAND HISTORY ACTION (E.G. UNDO/REDO)

(MANUAL) ANNOTATIONS



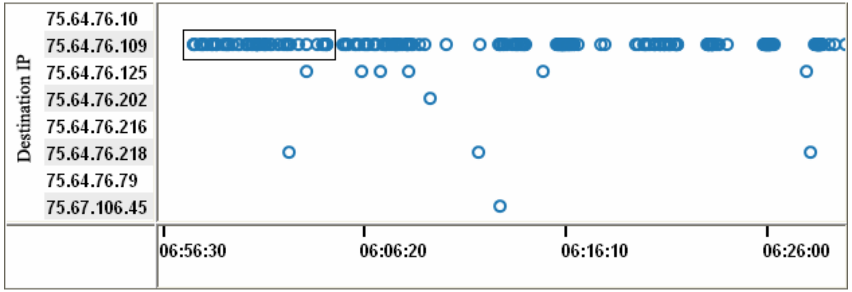
PROVENANCE OF INSIGHT

HISTORY OF COGNITIVE OUTCOMES FROM THE ANALYSIS

DIFFICULT TO CAPTURE, OFTEN MANUALLY ENTERED

CAPTURING INSIGHT

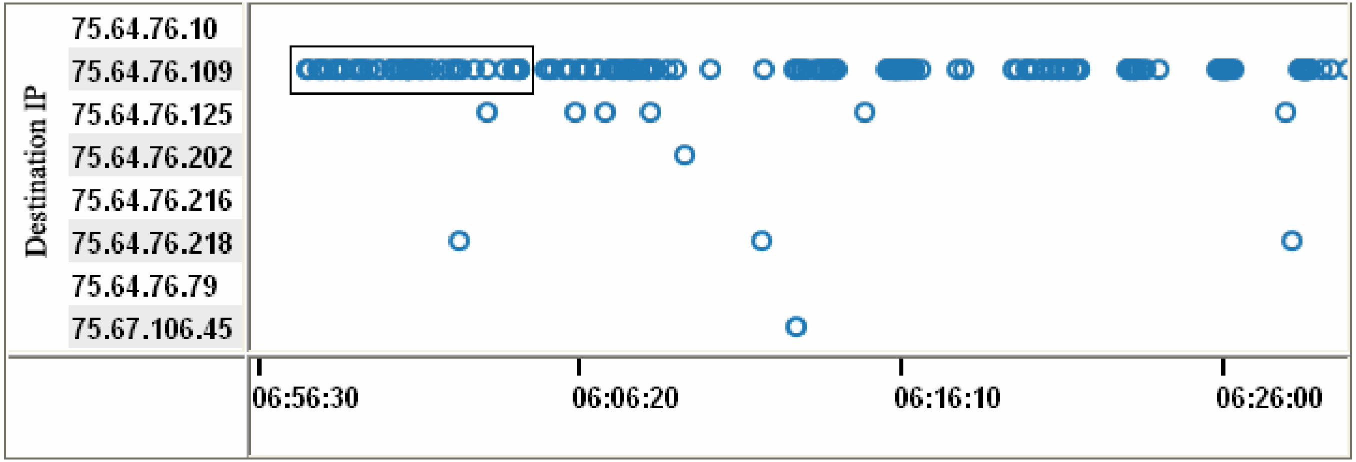
Network traffic visualization system
Analyst can create logical models of visual discoveries




```
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.

CAPTURING INSIGHT



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.

CAPTURING INSIGHT

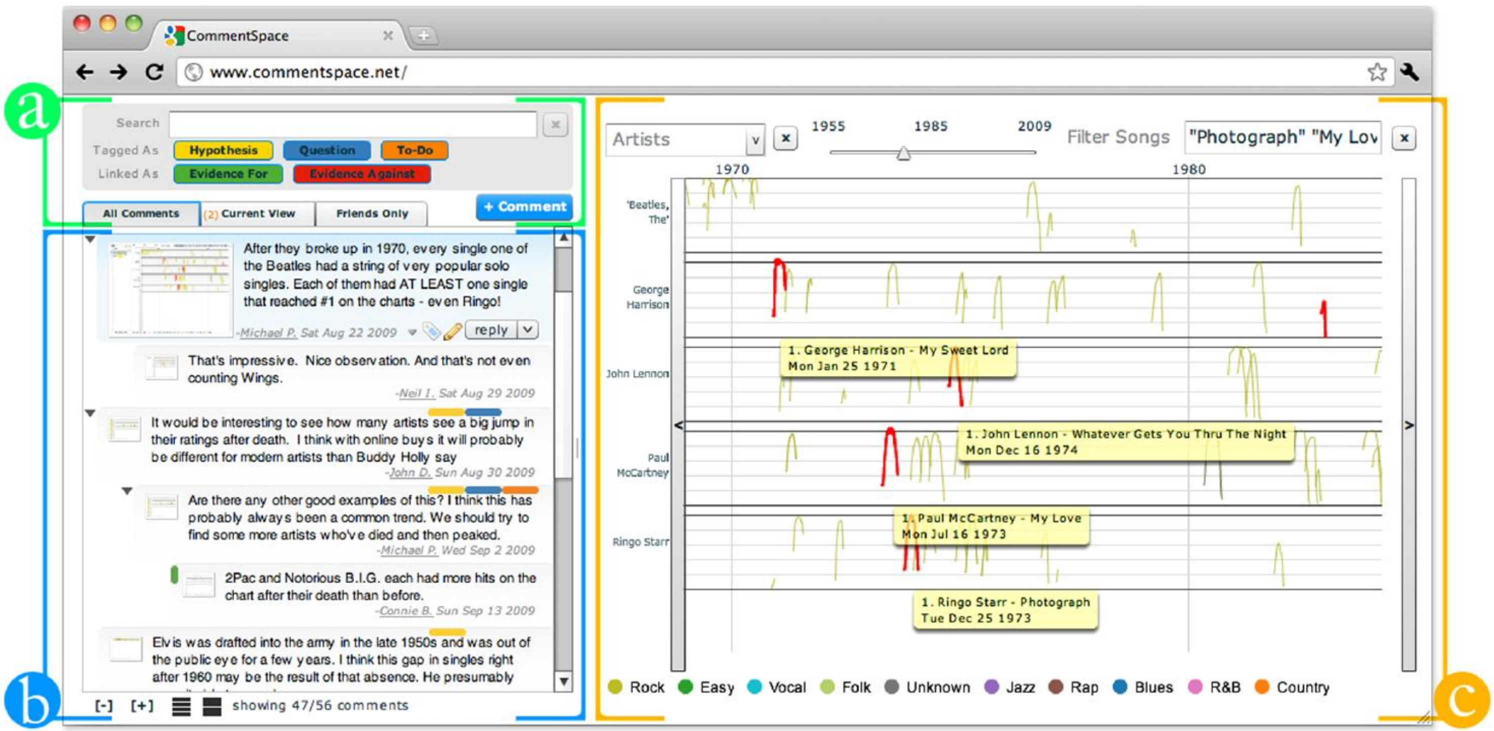
destination_port_80, destination_Stanford, identical_source_asn, time_sequence_30s, time_sequence_60s, more_than_4_events, more_than_32_events

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

CAPTURING INSIGHT



CommentSpace: Structured Support for Collaborative Visual Analysis
Wesley Willett, Jeffrey Heer, Joseph Hellerstein, Maneesh Agrawala
ACM Human Factors in Computing Systems (CHI), 2011

PROVENANCE OF RATIONALE

CAPTURE REASONING BEHIND DECISIONS, HYPOTHESES, INTERACTIONS

GOAL: IDEALLY FIGURE OUT SOMEONE'S ANALYTIC STRATEGY

PROVENANCE IN VISUAL ANALYTICS (RECAP)

PROVENANCE OF:

- DATA
- VISUALIZATION
- INTERACTIONS
- INSIGHTS
- RATIONALE

Characterizing Provenance in Visualization and Data Analysis: An Organizational Framework of Provenance Types and Purposes

Eric D. Ragan, Alex Endert, Jibonananda Sanyal, and Jian Chen

Abstract— While the primary goal of visual analytics research is to improve the quality of insights and findings, a substantial amount of research in provenance has focused on the history of changes and advances throughout the analysis process. The term, *provenance*, has been used in a variety of ways to describe different types of records and histories related to visualization. The existing body of provenance research has grown to a point where the consolidation of design knowledge requires cross-referencing a variety of projects and studies spanning multiple domain areas. We present an organizational framework of the different *types* of provenance information and *purposes* for why they are desired in the field of visual analytics. Our organization is intended to serve as a framework to help researchers specify types of provenance and coordinate design knowledge across projects. We also discuss the relationships between these factors and the methods used to capture provenance information. In addition, our organization can be used to guide the selection of evaluation methodology and the comparison of study outcomes in provenance research.

Index Terms— Provenance, Analytic provenance, Visual analytics, Framework, Visualization, Conceptual model

1 INTRODUCTION

Data visualization and visual analytics combine the power of visualization with advanced data analytics to help people to better understand data and discover meaningful insights. While the goal of visualization research is ultimately to improve the quality of insights and findings, analytic processes are complicated activities involving technology, people, and real work environments. Practical applications encounter problems that extend beyond the integration of any system's analytic models, processing power, visualization designs, and interaction techniques. Visualization systems must also support human processes, which often involve non-standardized methodologies including extended or interrupted periods of analysis, resource sharing and coordination, collaborative work, presentation to different levels of management, and attempts at reproducible analyses [92, 52, 42].

For these reasons, a substantial amount of research in the areas of visualization, data science, and visual analytics has been dedicated to supporting *provenance*, which broadly includes consideration for the history of changes and advances throughout the analysis process (e.g., [34, 73, 37, 21]). It is clear that the research community agrees on the importance of supporting provenance, and many scholars have developed tools and systems that explicitly aim to help analysts record both computational workflows (e.g., [21, 5, 71]) and reasoning processes (e.g., [26, 37]). For example, *VisTrails* tracks steps of the computational workflow during scientific data analysis and visualization, and then provides graphical representations of the workflow through a combination of node diagrams and intermediary visual outputs [5, 14]. Groth and Streefkerk [39] presented another example with a system for recording and annotating stages of view manipulations during a 3D molecule-inspection task. As another example, Del Rio and da Silva [22] designed *Probe-It* to keep track of the data sets that contributed to the creation of map visualizations. Focusing on the provenance of insights, Gotz and Zhou described how the *HARVEST* system records the history of semantic actions during business and financial analysis activities [37]. These are just a few examples from a large number of visual analytics tools designed to support provenance across a wide range of domains and for different purposes.

As the body of research and existing tools has grown, the community's knowledge of the many factors and goals relevant for effective provenance support has also broadened. However, the variety of perspectives can make it challenging to assess the specific aspects and purposes of provenance that are targeted by any particular project. The term, *provenance*, has been used in a variety of ways to describe different types of origins and histories. For example, the scientific visualization community, especially the simulation and modeling communities, often interpret provenance as the history of computational workflow (e.g., [34]), while other interpretations focus on the history of insights and hypotheses (e.g., [70]). Although many researchers proactively provide clear definitions and explanations of their foci in the provenance research, this does not entirely resolve the challenge of consolidating the variety of interpretations and research outcomes across projects. Different perspectives and applications of concepts become problematic for interpreting and coordinating outcomes from different provenance projects, for communicating ideas within the visualization community, and for allowing new-comers to clearly understand the research space. In our work, we analyzed the different perspectives of provenance that are most relevant to areas of visualization and data analysis.

Our goal in this paper is to organize the different *types* of provenance information and *purposes* for why they are desired in information visualization, scientific visualization, and visual analytics. We present an *organizational framework* as a conceptual model that categorizes and describes the primary components of provenance types and purposes. Further, we discuss the relationships between these factors and considerations when capturing provenance information. Our organizational framework is intended to help researchers specify types of provenance and coordinate design knowledge across projects. In addition, our organization can be used to guide the selection of evaluation methodology and the comparison of study outcomes in provenance research.

2 EXISTING PERSPECTIVES OF PROVENANCE

Analytic provenance is a broad and complex concept within the areas of information visualization, data analysis, and data science. In visual data analysis, the concept often includes aspects of the cognitive and interactive processes of discovery and exploration, and also the computational sequences and states traversed to arrive at findings or insights. Prior surveys have presented definitions, categorizations,

- *Eric D. Ragan is with Texas A&M University. E-mail: ragane@acm.org.*
- *Alex Endert is with Georgia Tech. E-mail: endert@gatech.edu.*
- *Jibonananda Sanyal is with Oak Ridge National Laboratory. E-mail: sanyalj@ornl.gov.*
- *Jian Chen is with University of Maryland, Baltimore County. E-mail: jichen@umbc.edu.*

Manuscript received 31 Mar. 2015; accepted 1 Aug. 2015; date of publication xx Aug. 2015; date of current version 25 Oct. 2015. For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.

Ragan et al., VAST 2015

WHAT TO DO WITH PROVENANCE INFORMATION?

PROVENANCE PURPOSES

RECALL

- MEMORY OF STATES OF ANALYSIS

REPRODUCIBILITY

- REPRODUCE STEPS/WORKFLOW

ACTION RECOVERY

- UNDO/REDO, BRANCHING

PROVENANCE PURPOSES

COLLABORATIVE COMMUNICATION

- SHARE INFO WITH OTHERS

PRESENTATION

- COMMUNICATE INSIGHT/PROGRESSION

META-ANALYSIS

- REVIEW THE ANALYTIC PROCESS

PROVENANCE VS. REPRODUCIBILITY

PROVENANCE VS. REPRODUCIBILITY

GOAL OF GENERAL 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?

RESOURCES

- SEE SCIENTIFIC REFERENCES ON SLIDES
- REPRODUCIBLE RESEARCH MOOC COURSERA.ORG (ROGER PENG)

NEXT UP

AFTER THE BREAK

TUTORIAL – REPRODUCIBLE RESEARCH IN R