VISUAL ANALYTICS INTRODUCTION TO R

TUTORIAL 1

Petra Isenberg

informatics mathematics
Inria

DATA ANALYSIS

Challenge

BIBLIOMETRICS

Study of measuring and analysing science, technology and innovation

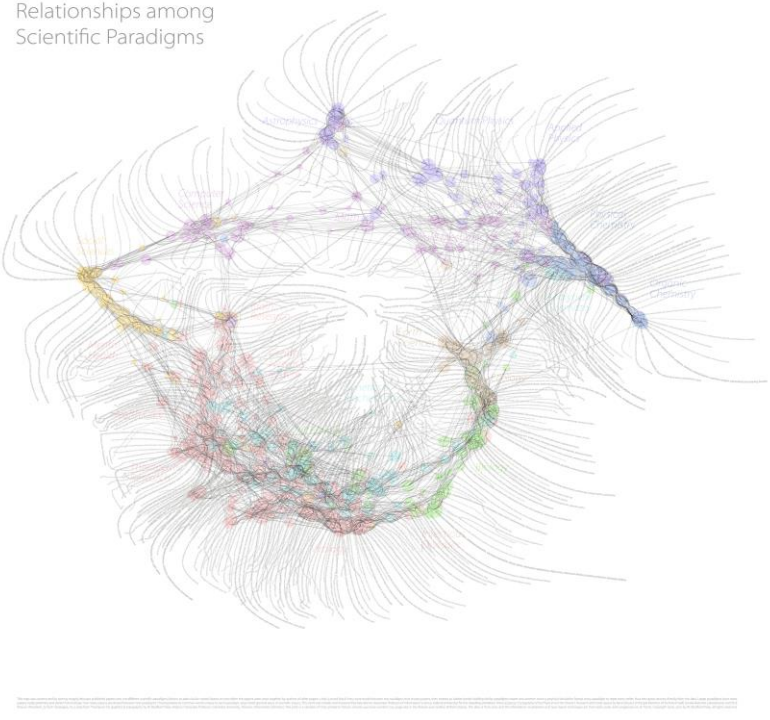
BIBLIOMETICS

the application of mathematical and statistical methods to books and other media of communication (Pritchard, 1969)

Scientometrics: the science of measuring and analyzing science

WHY?

- to understand science



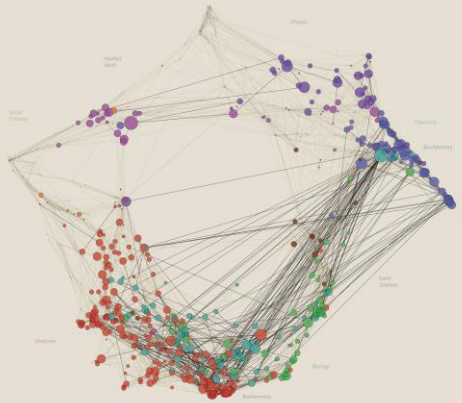
http://wbpaley.com/brad/mapOfScience/

The Scientific Paradigms that Support Patent Generation

This network diagram directs attention to the areas of science that support patents. Each circular node represents scientific publications on a topic (more accurately called a paradigm). There are 775 nodes in all, distributed in a ring-like structure; the largest colored ones are the ones that represent scientific paradigms that generate many patents. General groupings of paradigms include those that concern physics (at approximately 1:00), Chemistry (2:00), nodes that form the jutting peninsula of Organic Chemistry (at 3:00), clusters of nodes abutting with Earth Sciences and Biology (4:00), an interesting span of teal Biochemistry (from 4:30 to 6:00), linking Chemistry to Medicine (from 6:00 to 8:00), the Social Sciences (9:00) and Applied Math (including Computer Science, at 10:00).

The size of each colored node represents the number of patents that build from that area of science. Patents emerge from these general areas of science. There is a concentration of network of research in the computer science (around 12:00), a larger band of science in physics in chemistry (between 12:30 and 3:00), and a very large band of science in the medical sciences that yield patents. There are relatively fewer patents emerging from the social sciences (9:00), astrophysics (12:00), and the earth sciences (5:30).

The edges (lines between nodes) represent papers that explore multiple scientific paradigms; the black edges are the ones where patents are created. Chemistry and medicine seem to have many edges between them. This represents the large number of pharmaceutical patents that must do interdisciplinary work to bridge these disparate scientific paradigms. There are a particularly large number of links between areas of chemistry dealing with analytical methods (at 2:30) and the treatment of infectious diseases (9:00). There are also many patents that build from disparate paradigms within medicine (between 6:00 and 8:00).



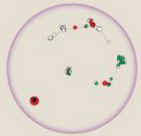
Drilling Down for Additional Insights

Paradigm 365

A paradigm in computer science has been pulled out of the diagram. Additional edges, showing relationships between this paradigm and other areas of science, illustrate that this paradigm links to the social sciences, chemistry, biology and medical science.

Research Communities within Paradigm 365

There are 85 unique research communities within this paradigm that are further grouped into four research areas. The size of each circle represents the amount of research in the area. The color represents performance: white for average, red for high-vitality science, green for patent-producing communities, black for communities that both get patents and practice high-vitality science.



Author Communities within Paradigm 365

The size of each circle represents the amount of research from each the 152 author clusters in this paradigm. The color represents performance, as mentioned at left. Author communities are much more dispersed (only 29% of the authors could be clustered). Authors that are members of clusters are much more likely to be associated with high-performing science.

Themes within Paradigm 365

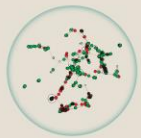
Another way to gain insights into this paradigm is to cluster the keywords in the 3150 scientific publications. The 127 keyword clusters shown below are more interconnected than the other two diagrams. The largest and most interconnected theme (the large black circle near the middle) deals with the use of network diagrams to understand the development of the World Wide Web.

Paradigm 735

One of the largest and most important paradigms in Chemistry has been pulled out of the diagram in order to show its impact in medicine. There are 4982 publications in this paradigm. The strongest edges connecting it and other paradigms are shown to illustrate the wide variety of scientific disciplines with which it shows publications.

Research Communities within Paradigm 735

There are 410 unique research communities within paradigm 735. The relationships between these clusters suggest a relatively complex but highly structured set of research topics. The dispersion of patent-supporting science (green and black) suggest areas to focus on and areas where the science is not as likely to generate patents.



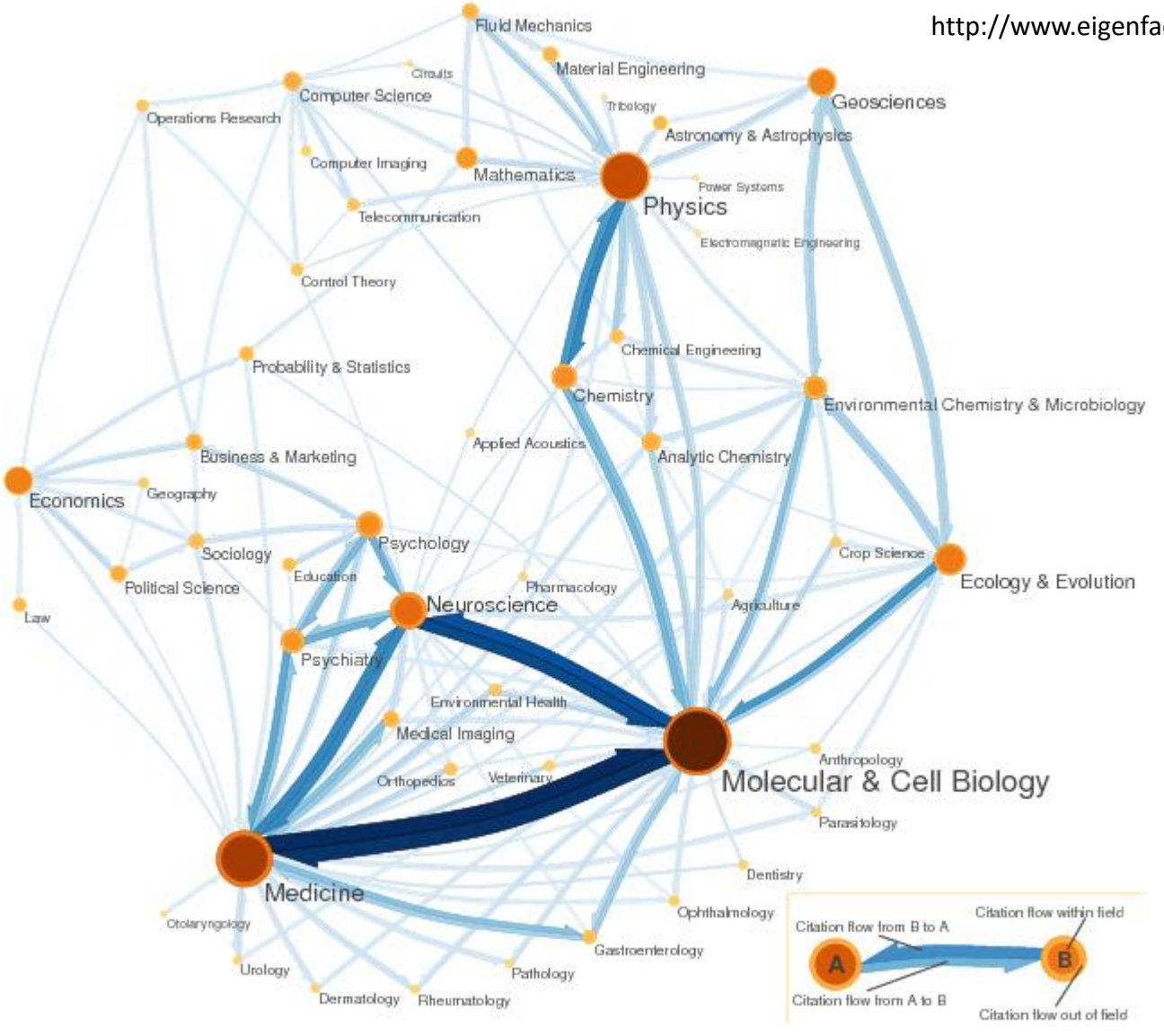
Author Communities within Paradigm 735

The size of each circle represents the amount of research from the 762 author clusters in this paradigm. Author communities are more dispersed that the paradigm shown above (only 28% of the authors are members of an author community). However, authors that are members of these clusters are much more likely to be associated with high-performing science.

Themes within Paradigm 735

The 890 clusters of keywords shown below form a very tight star pattern. This suggests a central theme to this research area. The largest and most interconnected theme in the paradigm deals with pancreatitis. Overall, the paradigm deals with many of the new work in chemistry that is being applied to drug development.

http://www.eigenfactor.org/map/maps.php



WHY?

- to understand science
- to manage science / research
  - ranking of scholarly output of researchers / institutions
  - identifying the centers of excellence

WHY IMPORTANT?

- Globalization of research
- Availability of large databases
- Increased research output → need for awareness
- Quickly evolving research fields

HOW WILL WE ANALYZE SCIENCE?

- through the study of scientific publications
- in the domains of Visual Analytics and Visualization
- by using exploratory analysis using visual analytics techniques (& some statistics)

SCIENTIFIC PUBLICATIONS

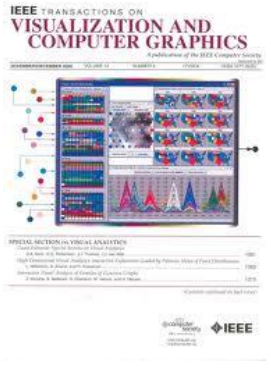
Why are they there?

1. Sharing scientific results/methods/processes
2. To show research performance
3. To allow validation of findings
4. To gain prestige and recognition

PUBLICATION VENUES

Conferences vs. Journals

- journals typical publication venues in most sciences
- in computer science (some) conference publications are highly regarded (with acceptance rates <25%)



RESEARCH QUESTIONS

- Simple & boring
  - Numbers of papers at IEEE VIS 2015
- Boring
  - Numbers of papers by P. Isenberg in 2015
- Interesting (unfortunately not simple)
  - In the domain of visual analytics growing or shrinking?
  - Are visual analytics and visualization the same community?
  - Are research interests of specific researchers changing?
  - What are new research trends in visual analytics?
  - To which university should I go to do a PhD in visual analytics?
  - Who are good reviewers for a certain topic?
  - Who should be in the program committee of VAST / VIS 2017?
  - How does a change in affiliation impact a researcher's interests?
  - I there a relation between affiliation and citations?

Exploring the Placement and Design of Word-Scale Visualizations

Pascal Goffin, Wesley Willett, Jean-Daniel Fekete *Senior Member, IEEE* and Petra Isenberg

Abstract—We present an exploration and a design space that characterize the usage and placement of word-scale visualizations within text documents. Word-scale visualizations are a more general version of sparklines—small, word-sized data graphics that allow meta-information to be visually presented in-line with document text. In accordance with Edward Tufte's definition, sparklines are traditionally placed directly before or after words in the text. We describe alternative placements that permit a wider range of word-scale graphics and more flexible integration with text layouts. These alternative placements include positioning visualizations between lines, within additional vertical and horizontal space in the document, and as interactive overlays on top of the text. Each strategy changes the dimensions of the space available to display the visualizations, as well as the degree to which the text must be adjusted or reflowed to accommodate them. We provide an illustrated design space of placement options for word-scale visualizations and identify six important variables that control the placement of the graphics and the level of disruption of the source text. We also contribute a quantitative analysis that highlights the effect of different placements on readability and text disruption. Finally, we use this analysis to propose guidelines to support the design and placement of word-scale visualizations.

Index Terms—Information visualization, text visualization, sparklines, glyphs, design space, word-scale visualizations

1 Introduction

Small high-resolution data graphics, included alongside words or word sequences in text documents, can often communicate information that could not be succinctly conveyed by the text itself. Examples include small stock charts embedded next to the name of a company, game statistics next to the name of a soccer team, or weather trends next to the name of a city. Traditionally, most of these *"word-scale visualizations"* have consisted of small line charts and bar charts and been placed in-line with text. Edward Tufte terms these word-scale visualizations "sparklines" [30], and provides some guidelines for their visual design. However, Tufte provides little guidance for *placing* word-scale visualizations with respect to text, suggesting only that they be placed in a "relevant context"—usually just after the word that they complement. However, the space of design and placement options for word-scale visualizations is actually quite large, and the consequences of placement decisions, in particular, are not well-understood.

In this paper, we provide design considerations for placing word-scale visualizations associated with words or word sequences (what we refer to as "entities") in a document. Our work is motivated by a close collaboration on digital note-taking with historians in the digital humanities. When visiting an archive, the historians we work with regularly take detailed notes on their findings. In these notes, they specifically tag entities such as the people, locations, or dates that occur in their document sources. The goal of tagging these entities is to help historians build an understanding of how entities relate to one another, where else the same entities appear in their notes, and what kinds of metadata are associated with them. Embedding this information using word-scale visualizations is a promising approach, because these small visualizations can add additional information in-context without distracting attention from the primary reading task.

In prior work, sparklines have typically been placed before or after the word they are related to. However, this is often not possible for the kinds of notes taken by our historians—e. g. when adding information to scanned documents and other immutable texts. Placing word-scale visualizations in-line with text may also be undesirable in other situations, as it requires reflowing the text and restricts the visualization's maximum height to that of the font—making visualizations hard to read when small font sizes were chosen. In-line visualizations can also disrupt sentences, making the text more difficult to read.

To better understand the options available for integrating word-scale visualizations in text documents, we outline a design space of possible placements relative to the text. In doing so, we relax some aspects of Tufte's original sparkline definition, imposing less restrictive size requirements and allowing the small visualizations to extend beyond strictly "word sized." Also, while Tufte did not restrict sparklines to specific visual encodings, the term "sparkline" does inherently suggest a "line-based" data encoding such as a line chart. In contrast, we specifically allow a variety of encodings, including geographical maps, heat maps, pie charts, and more complex visualizations and, thus, chose the term *word-scale visualizations*. We also formalize the notion of an *entity*—a concrete piece of text with associated metadata that can be encoded in a word-scale visualization. This explicit connection between an entity and a word-scale visualization directly affects the options for placing the visualization, and allows us to formally characterize the spatial relationship between text and graphic.

We begin our discussion by reviewing related work on small-scale and text visualizations. Then, in Section 3 we introduce the design space, its focus, and dimensions. Section 4 details several placement options and discusses trade-offs between word-scale visualization placement options. In Section 5 we discuss three examples that demonstrate the importance of the association between word-scale visualization and entity for the purpose of layout and interaction. Finally, in Section 6 we provide an in-depth analysis that examines how various placement options affect word-scale visualization placement in real documents. Based on this analysis, we provide recommendations that can help designers choose the right word-scale visualization given their own constraints.

2 Related Work

Our work relates closely to four research areas: (a) the use of sparklines and the design of word-scale visualizations (b) the integration of meta-data within text documents, (c) research on labeling in visualization, and (d) the readability of texts and visualizations.

2.1 Sparklines and Small-Scale Visualizations

According to Tufte [30] sparklines are "small, intense, simple, word-sized graphics with typographic resolution" that can be included anywhere a word or number can be—e. g. in a sentence, table, headline, map, spreadsheet or graphic. Tufte presents several examples of these embeddings. One example shows sparklines embedded in-line with text in order to provide metadata for a single word, for example glucose measurements next to the word glucose. In another, sparklines

- *Pascal Goffin is with Inria. E-mail: pascal.goffin@inria.fr.*
- *Wesley Willett is with Inria. E-mail: wesley.willett@inria.fr.*
- *Jean-Daniel Fekete is with Inria. E-mail: jean-daniel.fekete@inria.fr.*
- *Petra Isenberg is with Inria. E-mail: petra.isenberg@inria.fr.*

Manuscript received 31 Mar. 2014; accepted 1 Aug. 2014; date of publication xx xxx 2014; date of current version xx xxx 2014.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.

References

[1] A. Abdul-Rahman, J. Lein, K. Coles, E. Maguire, M. Meyer, M. Wynne, C. R. Johnson, A. Trefethen, and M. Chen. Rule-based visual mappings–with a case study on poetry visualization. *Computer Graphics Forum*, 32(3):381–390, 2013.

[2] E. Bertini, M. Rigamonti, and D. Lalanne. Extended excentric labeling. *Computer Graphics Forum*, 28(3):927–934, 2009.

[3] S. Bird, E. Klein, and E. Loper. *Natural Language Processing with Python*. O'Reilly Media Inc., 2009.

[4] R. Borgo, J. Kehrer, D. H. Chung, E. Maguire, R. S. Laramee, H. Hauser, M. Ward, and M. Chen. Glyph-based visualization: Foundations, design guidelines, techniques and applications. In *Eurographics 2013–State of the Art Reports*, pages 39–63. The Eurographics Association, 2012.

[5] M. Bostock, V. Ogievetsky, and J. Heer. D^3 data-driven documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2301–2309, 2011.

[6] U. Brandes, B. Nick, B. Rockstroh, and A. Steffen. Gestaltlines. *Computer Graphics Forum*, 32(3):171–180, 2013.

[7] B.-W. Chang, J. D. Mackinlay, P. T. Zellweger, and T. Igarashi. A negotiation architecture for fluid documents. In *Proceedings of the Conference on User Interface Software and Technology (UIST)*, pages 123–132. ACM, 1998.

[8] W. S. Cleveland, M. E. McGill, and R. McGill. The shape parameter of a two-variable graph. *Journal of the American Statistical Association*, 83(402):289–300, 1988.

[9] M. C. Dyson. How physical text layout affects reading from screen. *Behaviour & Information Technology*, 23(6):377–393, 2004.

[10] J.-D. Fekete and C. Plaisant. Excentric labeling: Dynamic neighborhood labeling for data visualization. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, pages 512–519. ACM, 1999.

[11] S. Few. Time on the horizon, Last read: March 2014. `http://www.perceptualedge.com/articles/visual_business_intelligence/time_on_the_horizon.pdf`.

[12] J. Fuchs, F. Fischer, F. Mansmann, E. Bertini, and P. Isenberg. Evaluation of alternative glyph designs for time series data in a small multiple setting. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, pages 3237–3246. ACM, 2013.

[13] P. Goffin, W. Willett, J.-D. Fekete, and P. Isenberg. Sparklificator, Last read: June 2014. `http://inria.github.io/sparklificator/`.

[14] B. Greenhill, M. Ward, and A. Sacks. The separation plot: A new visual method for evaluating the fit of binary models. *American Journal of Political Science*, 55(4):991–1002, 2011.

[15] J. Heer and M. Agrawala. Multi-scale banking to 45 degrees. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):701–708, 2006.

[16] J. Heer, N. Kong, and M. Agrawala. Sizing the horizon: The effects of chart size and layering on the graphical perception of time series visualizations. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, pages 1303–1312. ACM, 2009.

[17] M. R. Jakobsen and K. Hornbæk. Transient visualizations. In *Proceedings of the Conference on Computer-Human Interaction (OzCHI)*, pages 69–76. ACM, 2007.

[18] B. Lee, N. H. Riche, A. K. Karlson, and S. Carpendale. SparkClouds: Visualizing trends in tag clouds. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1182–1189, 2010.

[19] J. Pearson, G. Buchanan, and H. Thimbleby. Improving annotations in digital documents. In *Research and Advanced Technology for Digital Libraries*, pages 429–432. Springer, 2009.

[20] C. Perin, R. Vuillemot, and J.-D. Fekete. SoccerStories: A kick-off for visual soccer analysis. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2506–2515, 2013.

[21] P. Pirolli and S. Card. Information foraging. *Psychological Review*, 106(4):643–675, 1999.

[22] H. Reijner. The development of the horizon graph. In *Proceeding of Workshop From Theory to Practice: Design, Vision and Visualization Extended Abstracts of IEEE VisWeek*. Citeseer, 2008.

[23] T. Ropinski, S. Oeltze, and B. Preim. Survey of glyph-based visualization techniques for spatial multivariate medical data. *Computers & Graphics*, 35(2):392–401, 2011.

[24] T. Saito, H. N. Miyamura, M. Yamamoto, H. Saito, Y. Hoshiya, and T. Kaseda. Two-tone pseudo coloring: Compact visualization for one-dimensional data. In *Proceedings of the Conference on Information Visualization (InfoVis)*, pages 173–180. IEEE, 2005.

[25] H.-J. Schulz, T. Nocke, M. Heitzler, and H. Schumann. A design space of visualization tasks. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2366–2375, 2013.

[26] M. Steinberger, M. Waldner, M. Streit, A. Lex, and D. Schmalstieg. Context-preserving visual links. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2249–2258, 2011.

[27] M. Stone. In color perception, size matters. *IEEE Computer Graphics and Applications*, 32(2):8–13, 2012.

[28] J. Talbot, J. Gerth, and P. Hanrahan. An empirical model of slope ratio comparisons. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2613–2620, 2012.

[29] E. R. Tufte. *Envisioning Information*. Graphics Press, Cheshire, CT, 1990.

[30] E. R. Tufte. *Beautiful Evidence*. Graphics Press, Cheshire, CT, 2006.

[31] M. O. Ward. A taxonomy of glyph placement strategies for multidimensional data visualization. *Information Visualization*, 1(3-4):194–210, 2002.

[32] W. Willett, J. Heer, and M. Agrawala. Scented widgets: Improving navigation cues with embedded visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1129–1136, 2007.

[33] D. Yoon, N. Chen, and F. Guimbretière. TextTearing: Opening white space for digital ink annotation. In *Proceedings of the Conference on User Interface Software and Technology (UIST)*, pages 107–112. ACM, 2013.

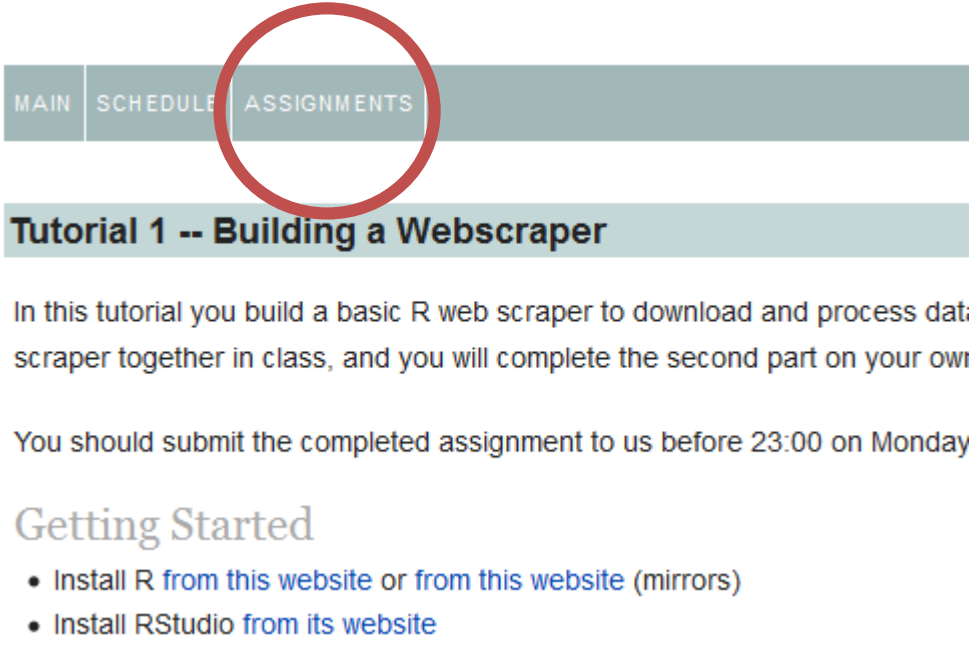
[34] P. T. Zellweger, S. H. Regli, J. D. Mackinlay, and B.-W. Chang. The impact of fluid documents on reading and browsing: An observational study. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, pages 249–256. ACM, 2000.

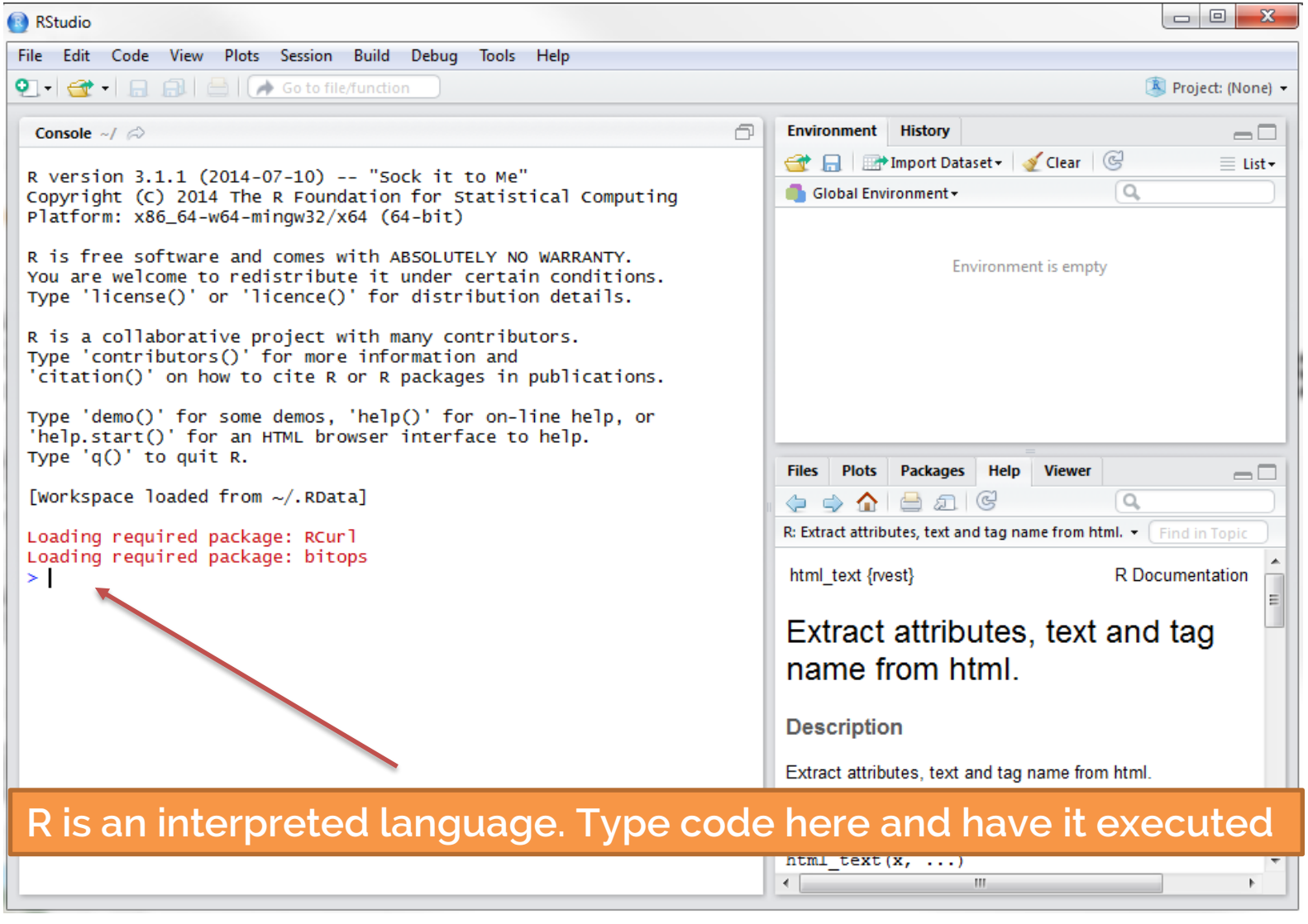
BUT FIRST WE NEED TO LEARN THE TOOLS

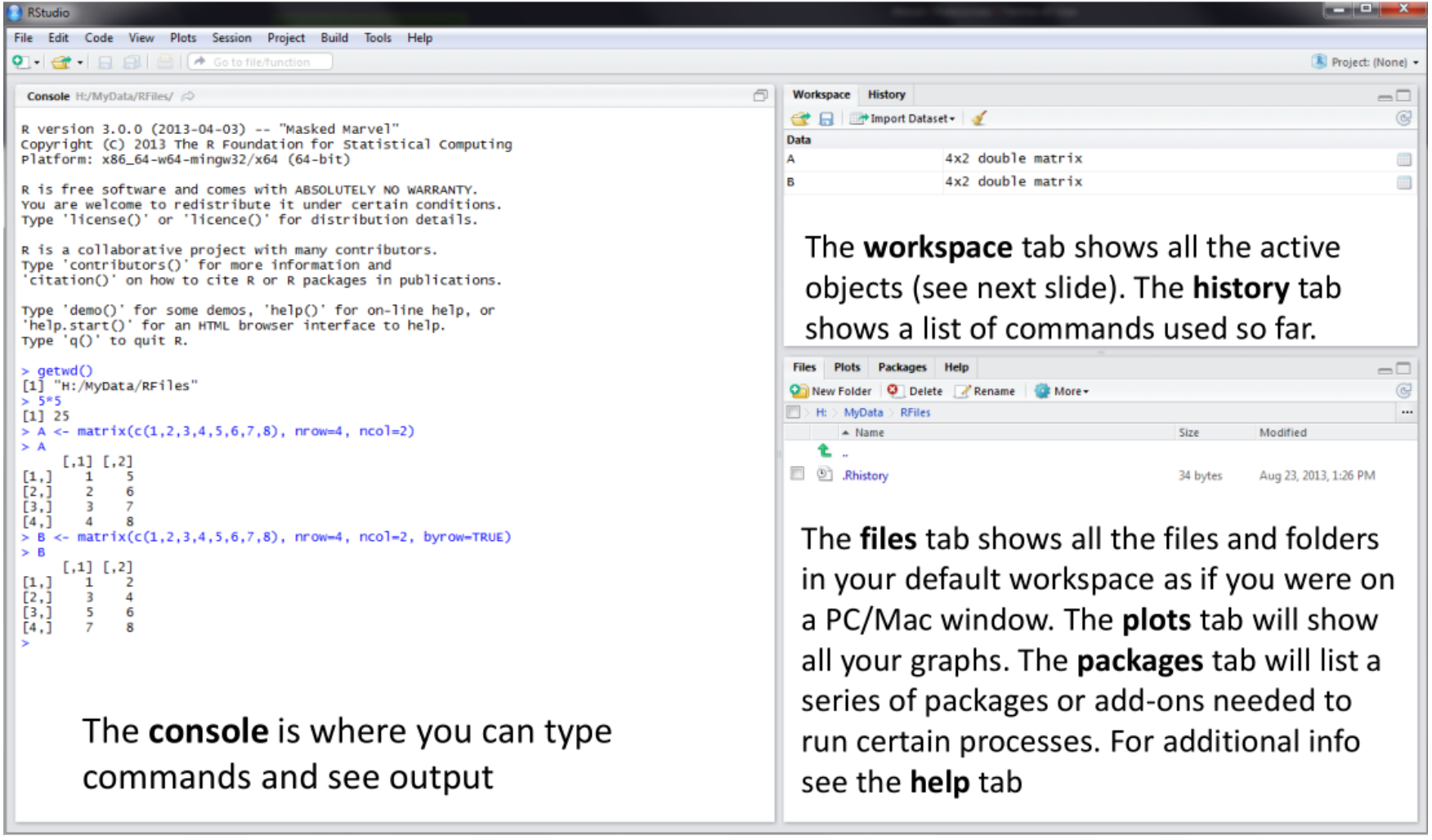
An introduction to R

INSTALLATION

http://tinyurl.com/VisualAnalytics2016




RStudio
File Edit Code View Plots Session Build Debug Tools Help
Go to file/function
Project: (None)
Console ~/
R version 3.1.1 (2014-07-10) -- "Sock it to Me"
Copyright (C) 2014 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
[Workspace loaded from ~/.RData]
Loading required package: RCurl
Loading required package: bitops
>
Environment History
Import Dataset
Clear
List
Global Environment
Environment is empty
Files Plots Packages Help Viewer
R: Extract attributes, text and tag name from html.
Find in Topic
html_text {rvest}
R Documentation
Extract attributes, text and tag name from html.
Description
Extract attributes, text and tag name from html.
html_text(x, ...)
R is an interpreted language. Type code here and have it executed



Picture from: http://dss.princeton.edu/training/

HELLO WORLD

- Type into your console

```
> print("Hello world!")

output:
  [1] "Hello world!"
```

QUICK R TUTORIALS

Let's get you to work:

```
> install.packages("swirl")

> library(swirl)
> install_from_swirl("R Programming")
> swirl()
```

Choose "R Programming"

If you are new to R complete the following lessons:
1, 2, 4, 7

If you are already a proficient R user pick a lesson that interests you

```
| When you are at the R prompt (>): |
•  Typing skip() allows you to skip the current question.
   |
•  Typing play() lets you experiment with R on your own;
   swirl | will ignore what you do... |
•  UNTIL you type nxt() which will regain swirl's
   attention. |
•  Typing bye() causes swirl to exit. Your progress will
   be | saved. |
•  Typing main() returns you to swirl's main menu. |
•  Typing info() displays these options again.
```