Graphical Inference for Infovis

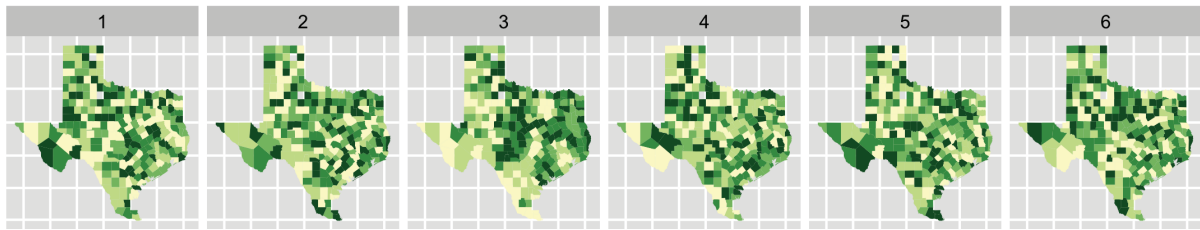## Hadley Wickham, Dianne Cook, Heike Hofmann, and Andreas Buja

Fig. 1. One of these plots doesn't belong. These six plots show choropleth maps of cancer deaths in Texas, where darker colors = more deaths. Can you spot which of the six plots is made from a real dataset and not simulated under the null hypothesis of spatial independence? If so, you've provided formal statistical evidence that deaths from cancer have spatial dependence. See Section 8 for the answer.

Abstract— How do we know if what we see is really there? When visualizing data, how do we avoid falling into the trap of apophenia where we see patterns in random noise? Traditionally, infovis has been concerned with discovering new relationships, and statistics with preventing spurious relationships from being reported. We pull these opposing poles closer with two new techniques for rigorous statistical inference of visual discoveries. The "Rorschach" helps the analyst calibrate their understanding of uncertainty and the "line-up" provides a protocol for assessing the significance of visual discoveries, protecting against the discovery of spurious structure.

Index Terms—Statistics, visual testing, permutation tests, null hypotheses, data plots.

✦

1 INTRODUCTION

What is the role of statistics in infovis? In this paper we try and answer that question by framing the answer as a compromise between curiosity and skepticism. Infovis provides tools to uncover new relationships, tools of curiosity, and much research in infovis focuses on making the chance of finding relationships as high as possible. On the other hand, most statistical methods provide tools to check whether a relationship really exists: they are tools of skepticism. Most statistics research focuses on making sure to minimize the chance of finding a relationship that does not exist. Neither extreme is good: unfettered curiosity results in findings that disappear when others attempt to verify them, while rampant skepticism prevents anything new from being discovered.

Graphical inference bridges these two conflicting drives to provide a tool for skepticism that can be applied in a curiosity-driven context. It allows us to uncover new findings, while controlling for apophenia, the innate human ability to see pattern in noise. Graphical inference helps us answer the question "Is what we see really there?"

The supporting statistical concepts of graphical inference are developed in [1]. This paper motivates the use of these methods for infovis and shows how they can be used with common graphics to provide users with a toolkit to avoid false positives. Heuristic formulations of these methods have been in use for some time. An early precursor is [2], who evaluated new models for galaxy distribution by generating samples from those models and comparing them to the photo-

graphic plates of actual galaxies. This was a particularly impressive achievement for its time: models had to be simulated based on tables of random values and plots drawn by hand. As personal computers became available, such examples became more common.[3] compared computer generated Mondrian paintings with paintings by the true artist, [4] provides 40 pages of null plots, [5] cautions against over-interpreting random visual stimuli, and [6] recommends overlaying normal probability plots with lines generated from random samples of the data. The early visualization system Dataviewer [7] implemented some of these ideas.

The structure of our paper is as follows. Section 2 revises the basics of statistical inference and shows how they can be adapted to work visually. Section 3 describes the two protocols of graphical inference, the Rorschach and the line-up, that we have developed so far. Section 4 discusses selected visualizations in terms of their purpose and associated null distributions. The selection includes some traditional statistical graphics and popular information visualization methods. Section 5 briefly discusses the power of these graphical tests. Section 8 tells you which panel is the real one for all the graphics, and gives you some hints to help you see why. Section 7 summarizes the paper, suggests directions for further research, and briefly discusses some of the ethical implications.

2 WHAT IS INFERENCE AND WHY DO WE NEED IT?

The goal of many statistical methods is to perform inference, to draw conclusions about the population that the data sample came from. This is why statistics is useful: we don't want our conclusions to apply only to a convenient sample of undergraduates, but to a large fraction of humanity. There are two components to statistical inference: testing (is there a difference?) and estimation (how big is the difference?). In this paper we focus on testing. For graphics, we want to address the question "Is what we see really there?" More precisely, is what we see in a plot of the sample an accurate reflection of the entire population? The rest of this section shows how to answer this question by providing a short refresher of statistical hypothesis testing, and describes how testing can be adapted to work visually instead of numerically.

Hypothesis testing is perhaps best understood with an analogy to

- *Hadley Wickham is an Assistant Professor of Statistics at Rice University, Email: hadley@rice.edu.*
- *Dianne Cook is a Full Professor of Statistics at Iowa State University.*
- *Heike Hofmann is an Associate Professor of Statistics at Iowa State University.*
- *Andreas Buja is the Liem Sioe Liong/First Pacific Company Professor of Statistics in The Wharton School at the University of Pennsylvania.*

the criminal justice system. The accused (data set) will be judged guilty or innocent based on the results of a trial (statistical test). Each trial has a defense (advocating for the null hypothesis) and a prosecution (advocating for the alternative hypothesis). On the basis of how evidence (the test statistic) compares to a standard (the *p*-value), the judge makes a decision to convict (reject the null) or acquit (fail to reject the null hypothesis).

Unlike the criminal justice system, in the statistical justice system (SJS) evidence is based on the similarity between the accused and known innocents, using a specific metric defined by the test statistic. The population of innocents, called the null distribution, is generated by the combination of null hypothesis and test statistic. To determine the guilt of the accused we compute the proportion of innocents who look more guilty than the accused. This is the *p*-value, the probability that the accused would look this guilty if they actually were innocent.

There are two types of mistakes we can make in our decision: we can acquit a guilty dataset (a type II error, or false negative), or falsely convict an innocent dataset (a type I error, or false positive). Just as in the criminal justice system, the costs of these two mistakes are not equal and vary based on the severity of the consequences (the risk of letting a guilty shoplifter go free is not equal to the risk of letting a guilty axe-murderer go free). Typically, as the consequences of our decisions become bigger, we want to become more cautious, and require more evidence to convict: an early-stage exploratory analysis is free to make a few wrong decisions, but it is very important not to approve a possibly dangerous drug after a late-stage clinical trial. It is up to the analyst to calculate and calibrate these costs.

To demonstrate these principles we use a small simulated example, based on an experiment designed to compare the accuracy of condition one vs. condition two in a usability study. Here, the defense argues that there is no difference between the two groups, and the prosecution argues that they are different. Statistical theory tells us to use the difference of the group means divided by the pooled standard deviation as the measure of guilt (the test statistic), and that under this measure the population of innocents will have (approximately) a *t*-distribution. Figure 2 shows this distribution for a sample of 10,000 innocents, a one-side two-sample t-test. The value of the observed test statistic is represented as a vertical line on the histogram. Since we have no a-priori notion of whether the difference between groups will be positive or negative, it is better to compare the accused to the absolute value of the innocents, as shown in the bottom plot, a one-sided two-sample t-test. As you can see, there are few innocents (about 3%) who appear as guilty as (or more guilty than) the accused and so the decision would be to convict.

These principles remain the same with visual testing, except for two aspects: the test statistic, and the mechanism of computing similarity. The test statistic is now a plot of the data, and instead of a mathematical measurement of difference, we use a human judge, or even jury.

Figure 3 illustrates a graphical alternative to the traditional *t*-test. The accused, a plot of the real data, is hidden among eight innocents, plots of data generated from the null distribution. We need some new terminology to make this description more concise: A **null dataset** is a sample from the null distribution, i.e. an example of an innocent dataset, and a **null plot** is a plot of a null dataset, showing what an innocent might look like. So Figure 3 hides the real plot amongst eight null plots. Can you spot the suspect? If so, then there is some evidence that the accused is different from the innocents, and we might move to convict. (See Section 8 for the solution.)

This example shows the analogy between a traditional numerical test and a new visual test, but the purpose of this work is *not* to supplant traditional tests. Traditional statistical tests are well studied, well-formulated and work best when data is well-behaved, following a known distribution in relatively simple scenarios. But as researchers in infovis have no doubt experienced, traditional statistical tests do not cover all of the complexities that arise when exploring data. The benefit of visual inference is that it can be used in complex data analysis settings that do not have corresponding numerical tests.
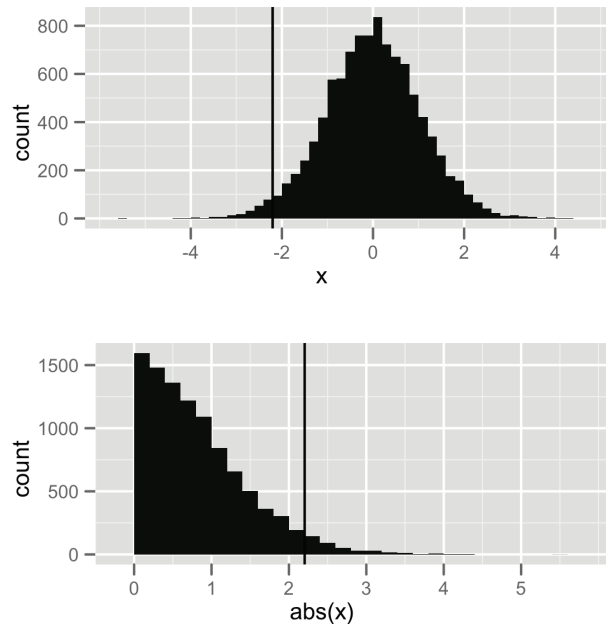

Fig. 2. (Top) Distribution of group means under null-hypothesis (distribution of innocents). Vertical line indicates value of observed test statistic. This is a one-sided two-sample t-test. (Bottom) Distribution of absolute value of group means. This is more appropriate comparison for the two-sided test, where magnitude of difference is more important than direction of difference. This is a two-sided two-sample t-test.
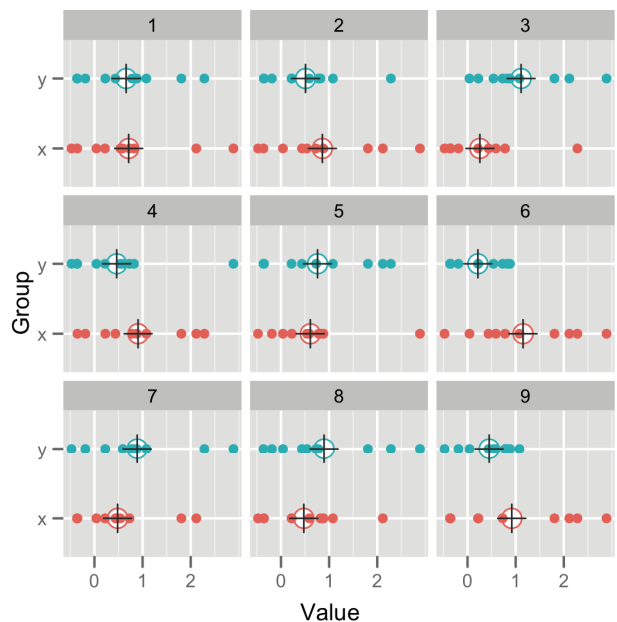

Fig. 3. A visual *t*-test. For each data set, the observations are shown as points and the group means as crosses. The accused is hidden amongst eight innocents. Can you spot him?

3 PROTOCOLS OF GRAPHICAL INFERENCE

This section introduces two new rigorous protocols for graphical inference: the "Rorschach" and the "line-up". The Rorschach is a calibrator, helping the analyst become accustomed to the vagaries of random data, while the line-up provides a simple inferential process to produce a valid *p*-value for a data plot. We describe the protocols and show examples of how they can be used, and refer the reader to [1] for more detail.

3.1 Rorschach

The Rorschach protocol is named after the Rorschach test, in which subjects interpret abstract ink blots. The purpose is similar: readers are asked to report what they see in null plots. We use this protocol to calibrate our vision to the natural variability in plots in which the data is generated from scenarios consistent with the null hypothesis. Our intuition about variability is often bad, and this protocol allows us to reduce our sensitivity to structure due purely to random variability.

Figure 4 illustrates the Rorschach protocol. These nine histograms summarize the accuracy at which 500 participants perform nine tasks. What do you see? Does it look like the distribution of accuracies is the same for all of the tasks? How many of the histograms show an interesting pattern? Take a moment to study these plots before you continue reading.
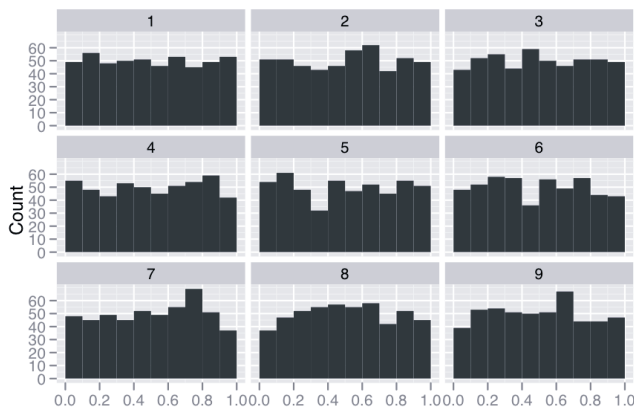

Fig. 4. Nine histograms summarizing the accuracy at which 500 participants perform nine tasks. What do you see?

It is easy to tell stories about this data: in task 7 accuracy peaks around 70% and drops off; in task 5, few people are 20-30% accurate; in task 9, many people are 60-70% accurate. But these stories are all misleading. It may come as a surprise, but these results are all simulations from a uniform distribution, that is, the distribution of accuracy for all tasks is uniform between 0 and 1. When we display a histogram of uniform noise, our expectation is that it should be flat. We do not expect it to be perfectly flat (because we know it should be a little different every time), but our intuition substantially underestimates the true variability in heights from one bar to the next. It is fairly simple to work out the expected variability algebraically (using a normal approximation): with 10 observations per bin, the bins will have a standard error of 30%, with 100 observations 19% and 1000, observations 6%. However, working through the math does not give the visceral effect of seeing plots of null data.

To perform the Rorschach protocol an administrator produces null plots, shows them to the analyst, and asks them what they see. To keep the analyst on their toes and avoid the complacency that may arise if they know all plots are null plots [8] the administrator might slip in a plot of the real data. For similar reasons, airport x-ray scanners randomly insert pictures of bags containing guns, knives or bombs. Typically, the administrator and participant will be different people, and neither should know what the real data looks like (a double-blinded scenario). However, with careful handling, it is possible to

self-administer such a test, particularly with appropriate software support, as described in Section 6.

Even when not administrated in a rigorous manner, this protocol is still useful as a self-teaching tool to help learn which random features we might spuriously identify. It is particularly useful when teaching data analysis, as an important characteristic of a good analyst is their ability to discriminate signal from noise.

3.2 Line-up

The SJS convicts based on difference between the accused and a set of known innocents. Traditionally the similarity is measured numerically, and the set of known innocents are described by a probability distribution. The line-up protocol adapts this to work visually: an impartial observer is used to measure similarity with a small set of innocents.

The line-up protocol works like a police line-up: the suspect (test statistic plot) is hidden in a set of decoys. If the observer, who has not seen the suspect, can pick it out as being noticeably different, there is evidence that it is not innocent. Note that the converse does not apply in the SJS: failing to pick the suspect out does not provide evidence they are innocent. This is related to the convoluted phraseology of statistics: we "fail to reject the null" rather than "accepting the alternative".

The basic protocol of the line up is simple:

- Generate $n-1$ decoys (null data sets).
- Make plots of the decoys, and randomly position a plot of the true data.
- Show to an impartial observer. Can they spot the real data?

In practice, we would typically set $n = 19$, so that if the accused is innocent, the probability of picking the accused by chance is $1/20 = 0.05$, the traditional boundary for statistical significance. Comparing 20 plots is also reasonably feasible for a human observer. (The use of smaller numbers of n in this paper is purely for brevity.) More plots would yield a smaller *p*-value, but this needs to be weighed against increased viewer fatigue. Another way of generating more precise *p*-values is to use a jury instead of a judge. If we recruit K jurors and k of them spot the real data, then the combined *p*-value is $P(X \le k)$, where X has a binomial distribution $B(K, p = 1/20)$. It can be as small as 0.05^K if all jurors spot the real data ($k = K$).

Like the Rorschach, we want the experiment to be double-blind - neither the person showing the plots or the person seeing them should know which is the true plot. The protocol can be self-administered, provided that it is the first time you've seen the data. After a first viewing of the data, a test might still be useful, but it will not be inferentially valid because you are likely to have learned some of the features of the data set and are more likely to recognize it. To maintain inferential validity once you have seen the data, you need to recruit an independent observer.

The following section shows some examples of the line-up in use, with some discussion of how to identify the appropriate null hypothesis for a specific type of plot and figure out a method of generating samples from the appropriate null distribution.

4 EXAMPLES

To use the line-up protocol, we need to:

- Identify the question the plot is trying to answer.
- Characterize the null-hypothesis (the position of the defense).
- Figure out how to generate null datasets.

This section shows how to approach each of these tasks, and then demonstrates the process in more detail for two examples. Section 4.1 shows a line-up of a tag cloud used to explore the frequency distribution of words in Darwin's "Origin of Species" and Section 4.2 shows a line-up of a scatterplot used to explore the spatial distribution of three point throws in basketball.

Table 1. A selection of common statistical and infovis plots and the question that they might have been created to answer. Each plot maybe be used to answer multiple questions and each question has a matching null hypothesis.

Plot	Question
Chloropleth maps	Is there a spatial trend?
Tag cloud	Is this document the same as that document?
Tag cloud	Is this document unusual?
Treemap	Does the distribution of sizes follow a power law?
Treemap	Is the distribution within higher-level categories the same?
Histogram	Is the underlying distribution smooth?
Histogram	Is the underlying distribution uniform (or normal or ...)?
QQPlot	Do the points lie along the line? (i.e. does the data match the distribution?)
Residual plot	Are residuals normally distributed?
Scatterplot	Are the two variables associated?
Scatterplot with points colored	Are points clustered by colour?
Facetted plots	Is the distribution of the data the same in all facets?
Time series	Does the mean change over time?
Time series	Does the variability change over time?

It's usually easy to identify the question a plot is trying to answer, because certain types of plots are used for specific tasks and for a particular data set the analyst typically chooses to make a plot with a question in mind. Table 1 lists the questions associated with some common plots. Note that some plots can be used to answer multiple questions, and so there may be different null hypotheses depending on the circumstances.

The null hypothesis is the least interesting answer to a question, that is, an answer which is "what we see is really nothing". In the language of the criminal justice system, the null hypothesis is the argument of the defense, that the suspect is innocent and nothing untoward occurred. If we are using a scatterplot to answer the question "is there a relationship between x and y?", then the null hypothesis is that there is no relationship, or that the two variables are independent. We only change our minds if we have evidence to the contrary.

The next step of the process is to determine how to generate null datasets from the null hypothesis. There are two techniques that apply in many circumstances:

- Resampling. This is how null samples are generated for permutation tests [9; 10] and similarly for confidence intervals using bootstrap samples [6]. In designed experiments it entails recreating the same randomization performed in the experiment. This technique is used in Section 4.1.

- Simulation. We might be interested in a more specific set of hypothesis: does time increase linearly with distance from target? Does accuracy decrease exponentially as number of distractors increases? In those cases we have a probabilistic model and we can generate null data sets by sampling from the distribution implied by the model. This approach is used in Section 4.2.

4.1   Tag clouds

A tag cloud (or text cloud) can be used to visualize frequency of words in a document (typically with common "stop" words removed). Words are arranged in various ways, often alphabetically, with size proportional to their frequency. Tag clouds are primarily descriptive, but if we look closely we can discover some inferential uses.

Firstly, there is an implicit question when we look at a tag cloud - are there any unusually frequent (or infrequent) words? This question carries with it some notion of "usual" frequency distribution. Secondly, tag clouds are also used for comparison. For example, `http://chir.ag/projects/preztags/` has a tag cloud for each major presidential speech, with the implication that we can use the tag clouds to spot the differences between speeches. Manyeyes [11] has an explicit comparison tool.

When comparing two texts with a tag cloud, we can imagine the data as two columns (word and source) with a row for every word. In the rawest form words are repeated reflecting their frequency. The null hypothesis for a comparison tag cloud is that the two documents are equivalent, the frequency of words is the same in each document. In other words, the word and source column are independent. It is simple to generate new datasets under this null: we just randomly permute one of the columns. Figure 5 uses this technique to compare the 1st and 6th editions of Darwin's "Origin of Species" (selected words displayed due to space constraints). Word frequencies from the two editions are displayed side-by-side, distinguished by colour (red = 1st ed, blue = 6th ed). Can you spot the accused? (See Section 8 for the solution.)

Permutation of a single column can be used to address questions about independence in other types of plots. The elegant feature of this approach to generating null data is that the marginal distribution of each variable is preserved while breaking any dependence. Here are some common examples:

- If we are interested in the spatial trend in a data map, then the null hypothesis might be that location and value are independent. To generate null datasets we permute the value column. This is how Figure 1 was generated.

- In a scatterplot, an initial hypothesis might be that there is no relationship between x and y. We can generate null hypotheses by permuting either the x or y variables.

- If we have clustered the data and are displaying the results with a coloured scatterplot, we might be interested to know if the clusters are well separated. Hence the null hypothesis is that cluster membership and position are independent, and we can generate null datasets by permuting the cluster id column.

4.2   Scatterplot

A scatterplot displays the relationship between two continuous variables, and answers the question: are x and y related in some way? The scatterplot can reveal many different types of relationships, e.g., linear trends, non-linear relationships and clustering. A strong null hypothesis is that there is no relationship x and y variables, that is, they are independent.

In many cases the assumption of independence is too strong: it is obvious that the two variables are related, and we want to investigate a specific functional form. For example, we might believe that three point attempts in basketball follow a quadratic distribution in space: as the angle between the player and basket increases the player moves closer to ensure success.

Figure 6 tests this hypothesis using data on all three pointers attempted by the Los Angeles Lakers in the 2008/09 season (data from `http://www.basketballgeek.com/`). Can you spot the real data? (See Section 8 for the solution.) It is embedded among null plots generated under the hypothesis that the relationship really is quadratic. These datasets are constructed by fitting the model, producing predictions and residuals, then adding rotated residuals [12] back on to the predictions.

For more statistically-minded audiences, we could instead display the residuals from the model. From the model definition, we expect the residuals to be distributed normally with mean 0 and standard deviation 1, which makes null datasets easy to generate: sample from the standard normal distribution. Figure 7 shows the results of this procedure. Is it easier or harder to spot the suspect compared to the previous figure?
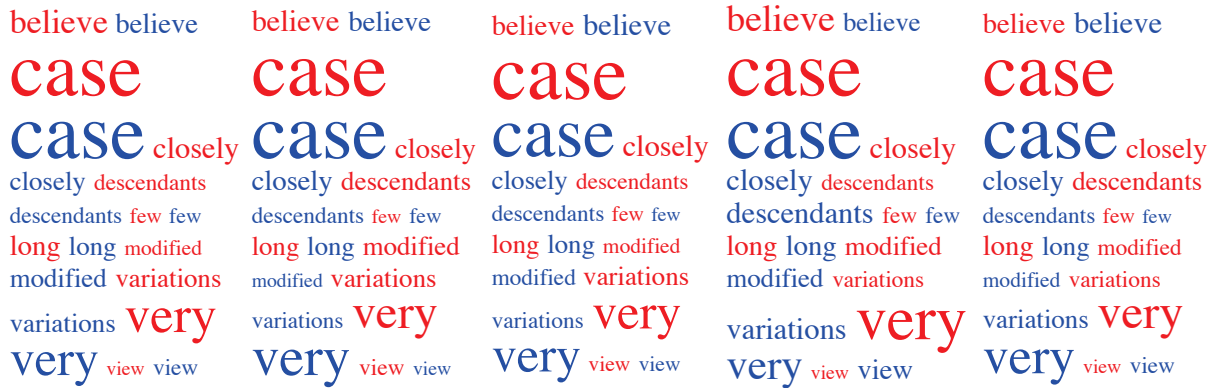
Fig. 5. Five tag clouds of selected words from the 1st (red) and 6th (blue) editions of Darwin's "Origin of Species". Four of the tag clouds were generated under the null hypothesis of no difference between editions, and one is the true data. Can you spot it?
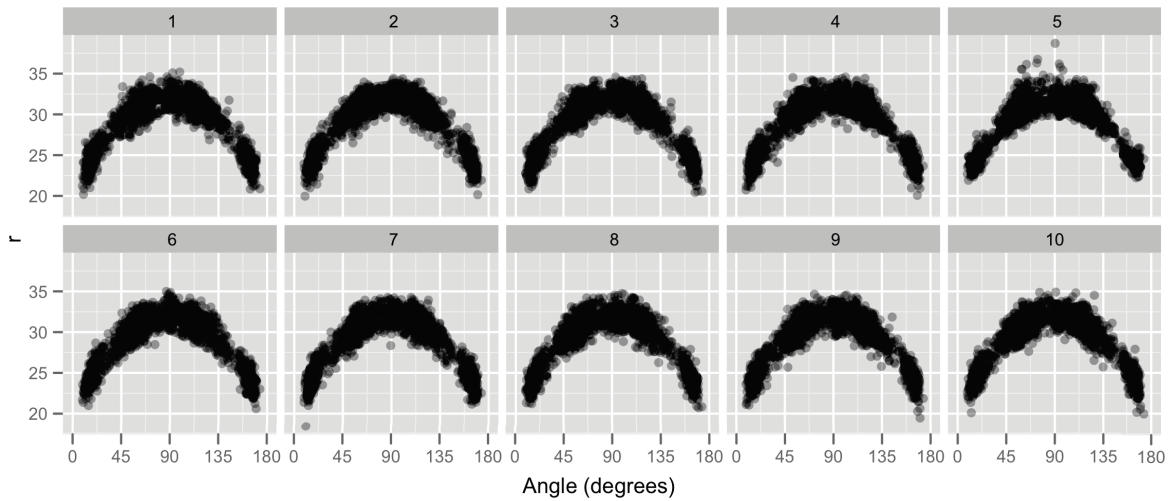

Fig. 6. Scatterplot of distance vs. angle for three pointers by the LA Lakers. True data is concealed in line-up of nine plots generated under the null hypothesis that there is a quadratic relationship between angle and distance.
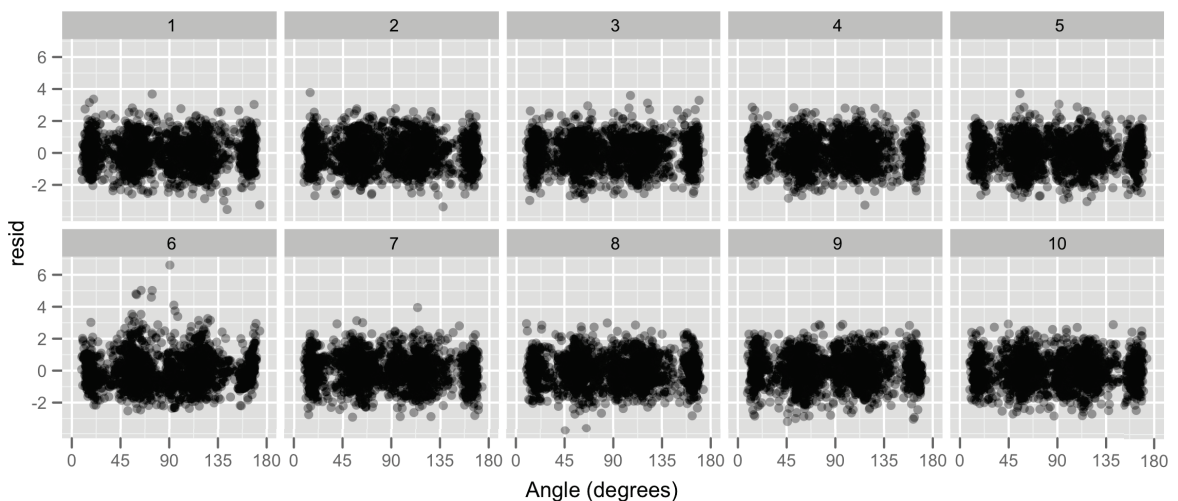

Fig. 7. Scatterplot of model residuals vs. angle for three point attempts by the LA Lakers. True data is concealed in line-up of nine plots generated under the null hypothesis of standard normally distributed residuals.

Figures 6 and 7 both test the same null hypothesis. Which should we use? The next section discusses the issue of power (the probability of spotting a real difference) in more detail.

5 POWER

The power of a statistical test is the probability of correctly convicting a guilty data set. The capacity to detect specific structure in plots can depend on many things, including an appropriate choice of plot, which is where the psychology of perception is very important. Existing research [13; 14] provides good suggestions on how to map variables to perceptual properties to maximize the reader's chances of accurately interpreting structure. For example, we should map our most important continuous variables to position along a common scale, and to use pre-attentive attributes, such as color, to represent categorical information like groups.

For large datasets, aggregation can make a big difference. Figure 8 displays a line-up for examining the relationship between arrival delays and wind direction at Phoenix airport (data from `http://stat-computing.org/dataexpo/2009`). The top plot is a "natural" and elegant plot of the raw data, mapping wind direction to angle and arrival delay to radius. The bottom plot displays the same data, but in the much aggregated form of a boxplot. It is much easier to spot the plot that is different from the others (panel 4), which, indeed, corresponds to the real data: unusually high delays at Phoenix airport are significantly associated with SW winds. One of the reasons making the real data difficult to detect in the first plot is that focus is on the outliers rather than the "hole" in the SW direction.

6 USE

To semi-automate the protocols we have created a new R [15] package called `nullabor`, available from `http://github.com/hadley/nullabor`. It operates simply. The user specifies how many decoy plots to create, and a mechanism to generate null datasets. For the line-up, `nullabor` generates the decoys, labelled with a new `.sample` variable, appends them to the real data set, and randomly chooses a position for the accused. The Rorschach protocol is similar, but the true data is only included with small probability. The package is bundled with methods to generate null datasets from common null hypotheses (independence, specified model, and specified distribution), while also allowing the user to add their own.

The following code shows two examples of `nullabor` in use. The first line specifies the type of plot, in this case a scatterplot for both. The second line specifies the protocol, the line-up, and the mechanism for generating null datasets: permutation in the first example, and simulation from a model in the second. By default, the line-up will generate a plot with 20 panels, but this can be specified by the n argument to the `lineup` function. The third line specifies the grid layout of the decoys. The position of the true data is encrypted and output to the screen so that the user can later decrypt the message and learn which panel shows the true data.

```
qplot(radius, angle, data = threept) %+%
  lineup(permute("response"), threept, n = 9) +
  facet_wrap(~ .sample)

qplot(radius, angle, data = threept) %+%
  lineup(model_null(radius ~ angle), threept) +
  facet_wrap(~ .sample)
```

This package makes it convenient for R users to administer the protocols as a normal step in their data analyses. We started with a command-line user interface because it is what many statisticians are most comfortable with, but it is not suitable for most analysts. We hope that others will integrate graphical inference in to their tools, using the open-source `nullabor` code to aid implementation.

This package enables the analyst to be the uninvolved observer, making inferentially valid judgements on the structure that is present in the real data plot by automatically generating a line-up before the analyst has seen the real data plot. At times the analyst will need to engage the services of an uninvolved observer. Services like Amazon Mechanical Turk [16] might be useful here. This problem requires a keen pair of human eyes to evaluate the line-ups, and the Turk offers a supply of workers who might even enjoy this type of task.

7 CONCLUSION

This paper has described two protocols to bring rigorous statistical inference to freeform data exploration. Both techniques center around identifying a null hypothesis, which then generates null datasets and null plots. The Rorschach provides a tool for calibrating our expectations of null data, while the line-up brings the techniques of formal statistical hypothesis testing to visualization.

Graphical inference is important because it helps us to avoid (or at least calibrate the rate of) false convictions, when we decide a relationship is significant, when it is actually an artifact of our sampling or experimental process. These tools seem particularly important for visualizations used in the VAST community, because the consequences of false conviction of data can be so severe for the people involved.

We have provided a reference implementation of these ideas in the R package `nullabor`. We hope others can build upon this work to make tools that can be used in a wide variety of analytic settings.

8 SOLUTIONS

These are the solutions to each of the line-ups shown in the paper.

- Figure 1: the real data is in panel 3. The features that might clue the reader in to this plot being different include spatial clustering, clumps of dark and light, and the prevalence of light polygons in the south-west edge of the state. In the other plots the dark and light coloring is scattered throughout the state.

- Figure 3: the real data is in panel 3. Features that the reader might pick up include bigger differences between the means, and fairly consistent but shifted spread from one group to another.

- Figure 5: the real data is second from the right. Hint: look at believe, variations, view and very.

- Figure 6: the real data is panel 5. More outliers from center court give the data away.

- Figure 7: the real data is panel 6. A few large outliers make this plot different from the others.

REFERENCES

[1] A. Buja, D. Cook, H. Hofmann, M. Lawrence, E.-K. Lee, D. F. Swayne, and H. Wickham, "Statistical inference for exploratory data analysis and model diagnostics," *Royal Society Philosophical Transactions A*, vol. 367, no. 1906, pp. 4361–4383, 2009.

[2] E. L. Scott, C. D. Shane, and M. D. Swanson, "Comparison of the synthetic and actual distribution of galaxies on a photographic plate," *Astrophysical Journal*, vol. 119, pp. 91–112, Jan. 1954.

[3] A. M. Noll, "Human or machine: A subjective comparison of piet mondrian's "composition with lines" (1917) and a computer-generated picture," *The Psychological Record*, vol. 16, pp. 1–10, 1966.

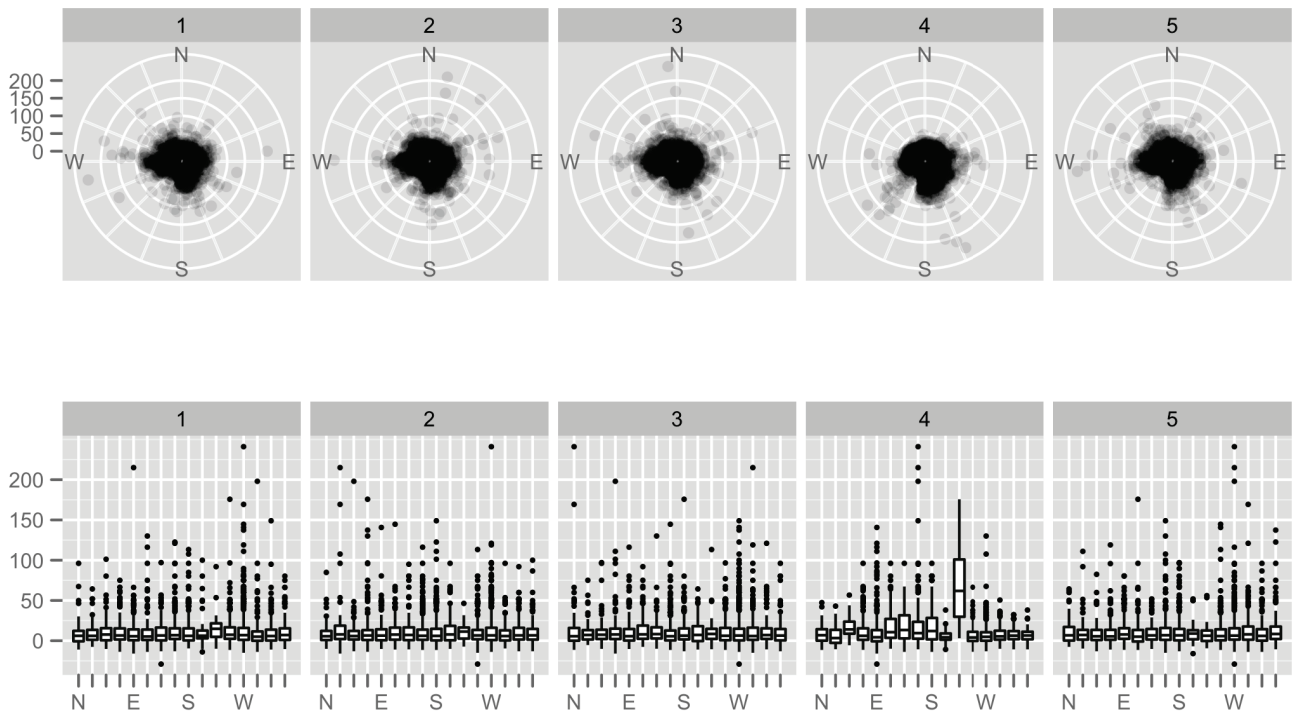[4] C. Daniel, *Applications of Statistics to Industrial Experimentation*. Hoboken, NJ: Wiley-Interscience, 1976.

Fig. 8. Two graphics showing the relationship between wind direction and arrival delays for incoming flights to Phoenix airport. (Top) A scatterplot with wind direction mapped to angle and delay to radius. (Bottom) Compass directions summarized with boxplots. Which of the two plots makes it easier to spot the real data?

[5] P. Diaconis, "Theories of data analysis: From magical thinking through classical statistics," in *Exploring Data Tables, Trends and Shapes* (D. Hoaglin, F. Mosteller, and J. Tukey, eds.), pp. 1–36, New York: Wiley, 1983.

[6] A. C. Davison and D. V. Hinkley, *Bootstrap Methods and their Applications*. Cambridge, UK: Cambridge University Press, 1997.

[7] A. Buja, D. Asimov, C. Hurley, and J. A. McDonald, "Elements of a viewing pipeline for data analysis," in *Dynamic Graphics for Statistics*, Wadsworth, Inc., 1988.

[8] J. M. Wolfe and M. J. V. Wert, "Varying target prevalence reveals two dissociable decision criteria in visual search," *Current Biology*, vol. 20, no. 2, pp. 121–124, 2010.

[9] E. J. G. Pitman, "Significance tests which may be applied to samples from any populations," *The Journal of the Royal Statistical Society*, vol. 4, pp. 119–130, 1937.

[10] P. Good, *Permutation, Parametric, and Bootstrap Tests of Hypotheses*. New York: Springer, 2005.

[11] F. Viègas, M. Wattenberg, F. van Ham, J. Kriss, and M. McKeon, "Manyeyes: A site for visualization at internet scale," *Transactions on Visualization and Computer Graphics*, vol. 13, pp. 1121–1128, 2007.

[12] Ø. Langsrud, "Rotation tests," *Statistics and Computing*, vol. 15, no. 1, pp. 53–60, 2005.

[13] W. S. Cleveland and R. McGill, "Graphical perception: Theory, experimentation and application to the development of graphical methods.," *Journal of the American Statistical Association*, vol. 79, no. 387, pp. 531–554, 1984.

[14] C. Healey, "Perception in visualisation," 2009.

[15] R Development Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2009. ISBN 3-900051-07-0.

[16] Amazon, "Mechanical Turk," 2008.

[17] H. Wickham, *ggplot2: Elegant graphics for data analysis*. useR, Springer, July 2009.