Chapter 11. Visualization

11.1. An Exploratory Data Analysis of the US Housing Crisis[*]

Introduction

The US housing crisis has undermined the world economy in wide reaching and poorly understood ways. Although there is a lot of speculation over the causes and the effects of the housing crisis, most of these ideas come from opinionated blogs or news articles that do not list their sources. This lack of data becomes perilous as the US government invests trillions of dollars based on untested hypotheses concerning the crisis. Our PFUG's focus is to compile, clean, and analyze data pertaining to the housing crisis to get a clearer picture of what is actually going on.

Overview and Motivation

Real Estate Bubble: Around 2006, house prices rose much higher than their true value. Eventually, housing prices became so high, it was difficult for current owners to afford their house. As foreclosure rates increased, house prices began to plummet. This has largely affected the global economy.

Little Public Organized Data: There is a lot of speculation over the causes and the effects of the housing crisis. Unfortunately, most of these ideas come from opinionated blogs or news articles that don't list their sources. Therefore, it is difficult to collect reliable information.

Government Expenditures: The government has already exhausted millions of dollars in order to aid those affected by housing crisis. With such little public data about the crisis, we are left wondering what data the government is using.

Still Unfolding: It is important to realize that the housing crisis in ongoing. This allows us to track its progression and hopefully make predictions for the upcoming years.

Large Data Sets: The housing crisis serves as a perfect model for visualizing large data sets. Most data sets we collect usually cover multiple years, counties and variables.

Problems with Large Data

Hard To Find: All of the data we have collected come from multiple sources. Currently, there is no central repository where data can be found.

Licenses and Fees: Some of the data sets have licenses that do not allow us to reproduce or publish any of our findings. Also many of the data sets cost large amounts of money to purchase.

Size: Some data sets were as large as 10 GB. In order to work around this problem, we were able to extract certain parts of the data sets without having to completely download them.

Dirty: Most of the data sets we find are what we call “dirty.” They are usually unorganized and practically unreadable.

Data Sets

To view our most current data sets and work, please visit our PFUG's website: **http://github.com/hadley/data- housing-crisis**. Some of our major data sets include...

- American Community Survey
- Case-Shiller House Price Index (HPI)
- Census 2007
- Construction of Housing Units
- Market Value of 1 month rent in a Room
- Vacancies
- Mortgage Rates
- Federal Housing Finance Agency HPI

Cleaning and Analysis

To facilitate sharing data, we have conducted both data cleaning and analysis with the open source statistical software R, which is available free of charge at **http://www.r-project.org**. We use the program R to clean our data sets. R is considered a statistical standard among statisticians. There are several advantages to using R. We are able to manipulate extremely large data sets (>2GB) on a normal desktop. It also allows us to produce impressive graphics with minimal coding.

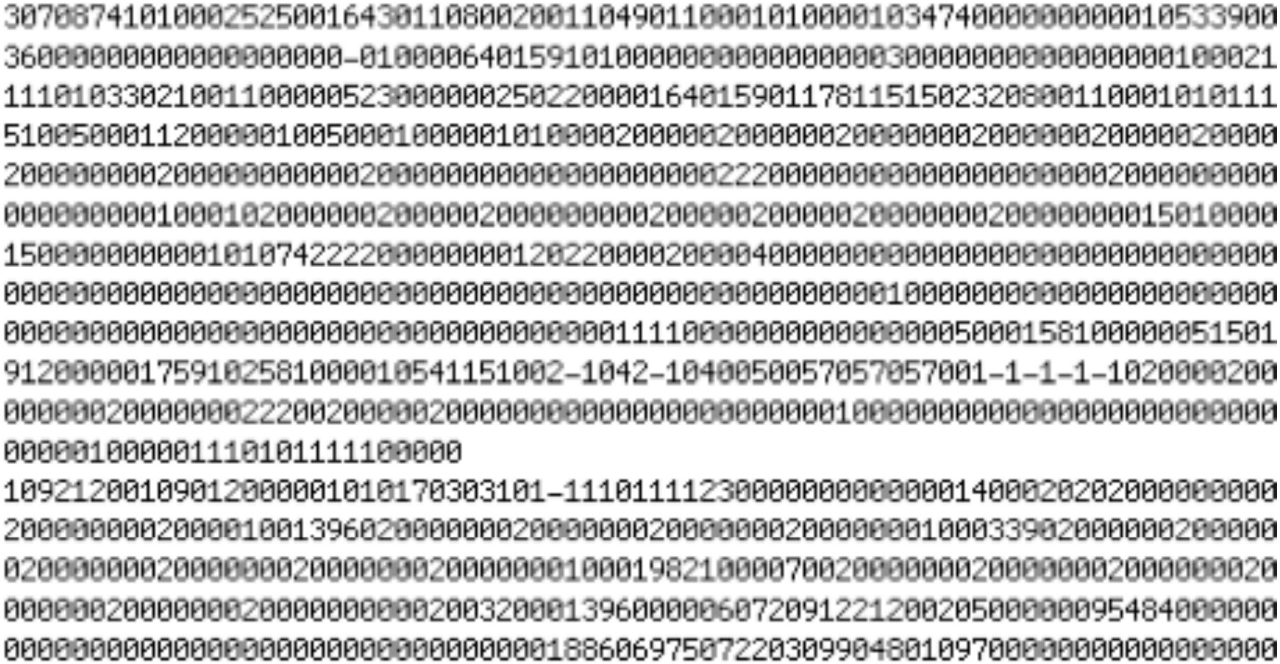
Clean Data is...

- **Consistent**: In a few data sets county names change over the course of a few years. This affects how we compare yearly data.
- **Concise**: Some data sets contained only parts of information we needed. For example, the American Community Survey contains over 200 questions. We were only interested in the answer to one of those questions.
- **Complete**: One of the data sets that was collected was missing around 80\% of the data.
- **Correct**: We must assume that the data we collect is not corrupt and was recorded properly. Some smaller data sets contained unusual observations. We used our own discretion when deciding what data sets were correct.

Cleaning Process

1. First we start with ``dirty'' data. (Fig.1)

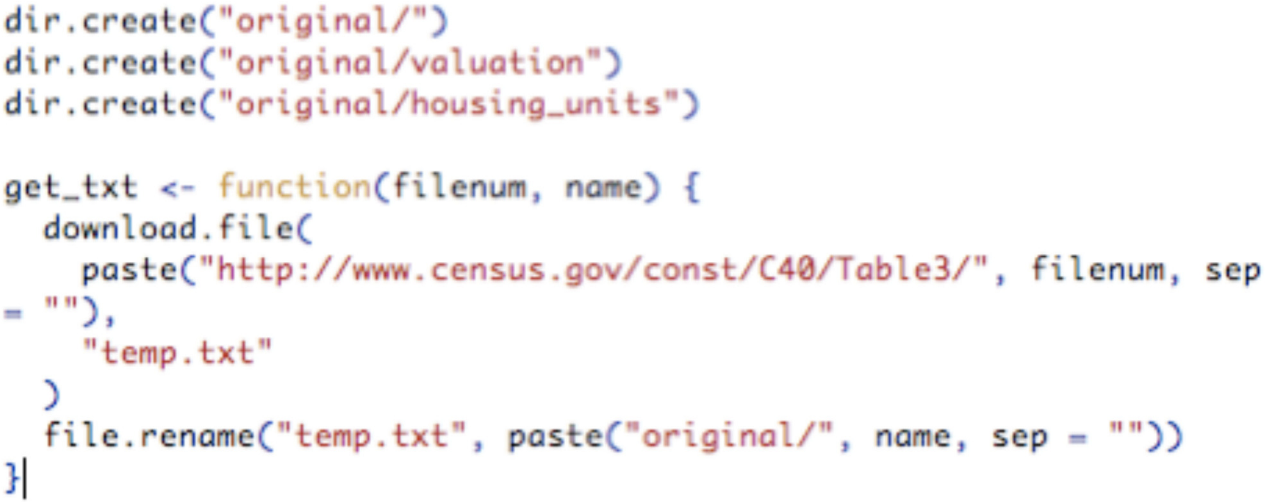
Figure 11.1.



```
3070874101000252500164301108002001104901100010100001034740000000000010533900
36000000000000000000-0100006401591010000000000000000300000000000000000100021
1110103302100110000052300000025022000016401590117811515023208001100010101111
5100500011200000100500010000010100002000000200000020000000200000002000000020000
2000000002000000000000200000000000000000000222000000000000000000000200000000
0000000001000102000000200000200000000002000002000000200000000200000000015010000
1500000000001010742222000000001202200002000040000000000000000000000000000000
0000000000000000000000000000000000000000000000000000000100000000000000000000000
00000000000000000000000000000000000001111000000000000000005000158100000051501
912000001759102581000010541151002-1042-1040050057057057001-1-1-1020000200
000000200000000222002000002000000000000000000000000001000000000000000000000000
00000100000011101011111000000
10921200109012000001010170303101-1110111123000000000000140002020200000000
20000000002000010013960200000000200000002000000002000000010003390200000020000
0200000000200000000200000000200000000100019821000070020000000020000000020000000020
000000200000000200000000000200320001396000006072091221200205000000954840000000
00000000000000000000000000000000001886069750722030990480109700000000000000000
```

2. Next we must download the data. A section of download code is shown below. (Fig. 2)

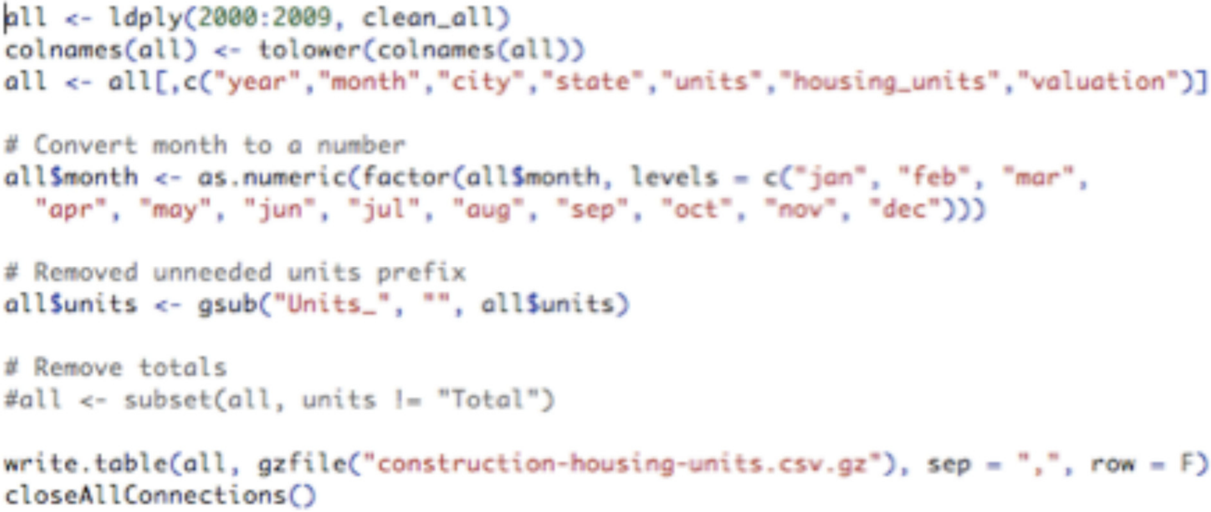
Figure 11.2.



```
dir.create("original/")
dir.create("original/valuation")
dir.create("original/housing_units")

get_txt <- function(filenum, name) {
  download.file(
    paste("http://www.census.gov/const/C40/Table3/", filenum, sep
= ""),
    "temp.txt"
  )
  file.rename("temp.txt", paste("original/", name, sep = ""))
}
```

3. Once we have the data, we clean the data as best we can according to the rules describing clean data above. A section of cleaning code is shown below. (Fig. 3)

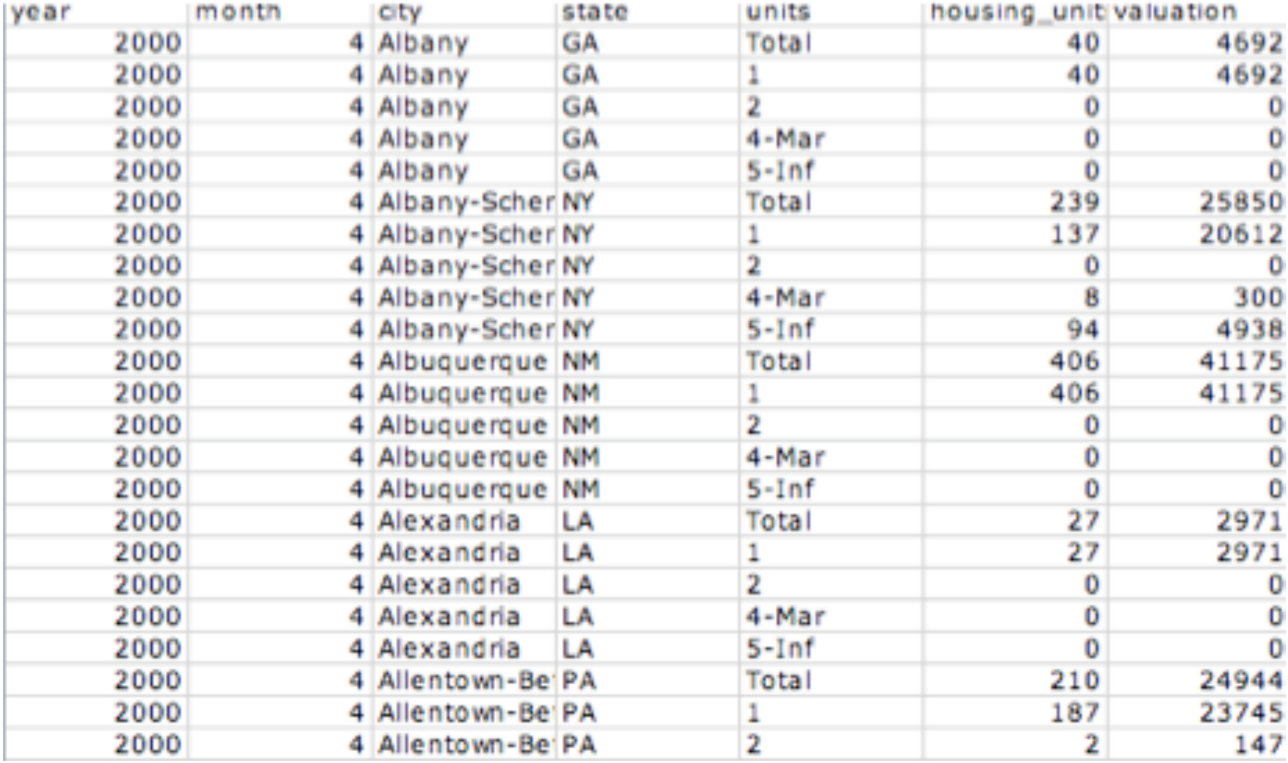
Figure 11.3.



```
all <- ldply(2000:2009, clean_all)
colnames(all) <- tolower(colnames(all))
all <- all[,c("year","month","city","state","units","housing_units","valuation")]

# Convert month to a number
all$month <- as.numeric(factor(all$month, levels = c("jan", "feb", "mar",
  "apr", "may", "jun", "jul", "aug", "sep", "oct", "nov", "dec")))

# Removed unneeded units prefix
all$units <- gsub("Units_", "", all$units)

# Remove totals
#all <- subset(all, units != "Total")

write.table(all, gzfile("construction-housing-units.csv.gz"), sep = ",", row = F)
closeAllConnections()
```

4. Now that the data has been cleaned, it may look like the top part of the data below. (Fig. 4)

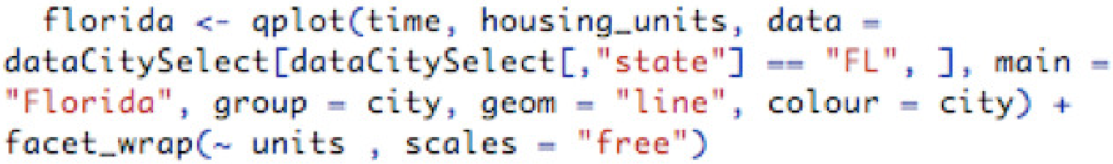
Figure 11.4.



year	month	city	state	units	housing_unit	valuation
2000	4	Albany	GA	Total	40	4692
2000	4	Albany	GA	1	40	4692
2000	4	Albany	GA	2	0	0
2000	4	Albany	GA	4-Mar	0	0
2000	4	Albany	GA	5-Inf	0	0
2000	4	Albany-Scher	NY	Total	239	25850
2000	4	Albany-Scher	NY	1	137	20612
2000	4	Albany-Scher	NY	2	0	0
2000	4	Albany-Scher	NY	4-Mar	8	300
2000	4	Albany-Scher	NY	5-Inf	94	4938
2000	4	Albuquerque	NM	Total	406	41175
2000	4	Albuquerque	NM	1	406	41175
2000	4	Albuquerque	NM	2	0	0
2000	4	Albuquerque	NM	4-Mar	0	0
2000	4	Albuquerque	NM	5-Inf	0	0
2000	4	Alexandria	LA	Total	27	2971
2000	4	Alexandria	LA	1	27	2971
2000	4	Alexandria	LA	2	0	0
2000	4	Alexandria	LA	4-Mar	0	0
2000	4	Alexandria	LA	5-Inf	0	0
2000	4	Allentown-Be	PA	Total	210	24944
2000	4	Allentown-Be	PA	1	187	23745
2000	4	Allentown-Be	PA	2	2	147

5. With clean data, we are able to explore it. The code below (Fig. 5) is the command used to produce the plot in figure Fig. 6.

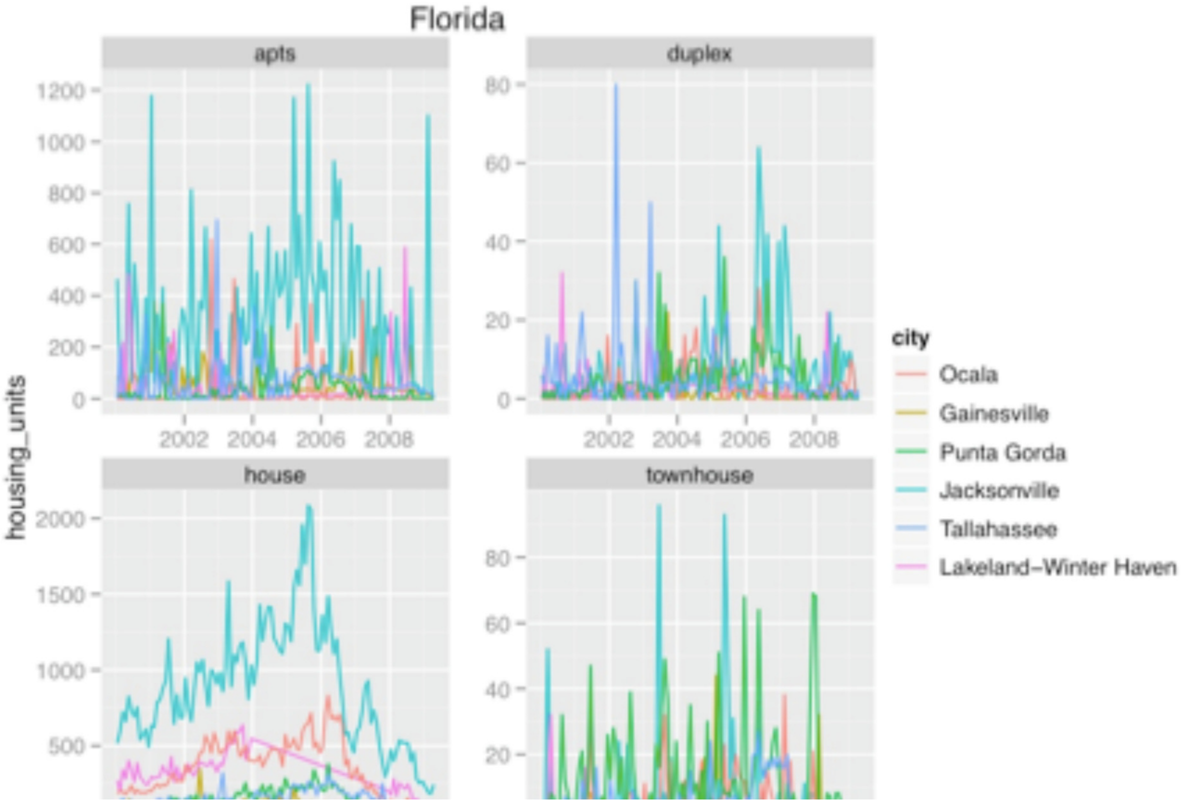
Figure 11.5.

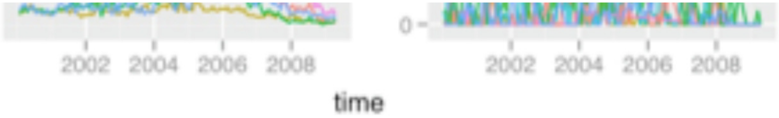


```
  florida <- qplot(time, housing_units, data =
dataCitySelect[dataCitySelect[,"state"] == "FL", ], main =
"Florida", group = city, geom = "line", colour = city) +
facet_wrap(~ units , scales = "free")
```

6. With R code we are able to produce complex plots with minimal amount of code. (Fig. 6)

Figure 11.6.


Florida
apts
duplex
house
townhouse
housing_units
1200
1000
800
600
400
200
0
80
60
40
20
0
2000
1500
1000
500
80
60
40
20
2002
2004
2006
2008
2002
2004
2006
2008
city
Ocala
Gainesville
Punta Gorda
Jacksonville
Tallahassee
Lakeland–Winter Haven



Interesting Findings

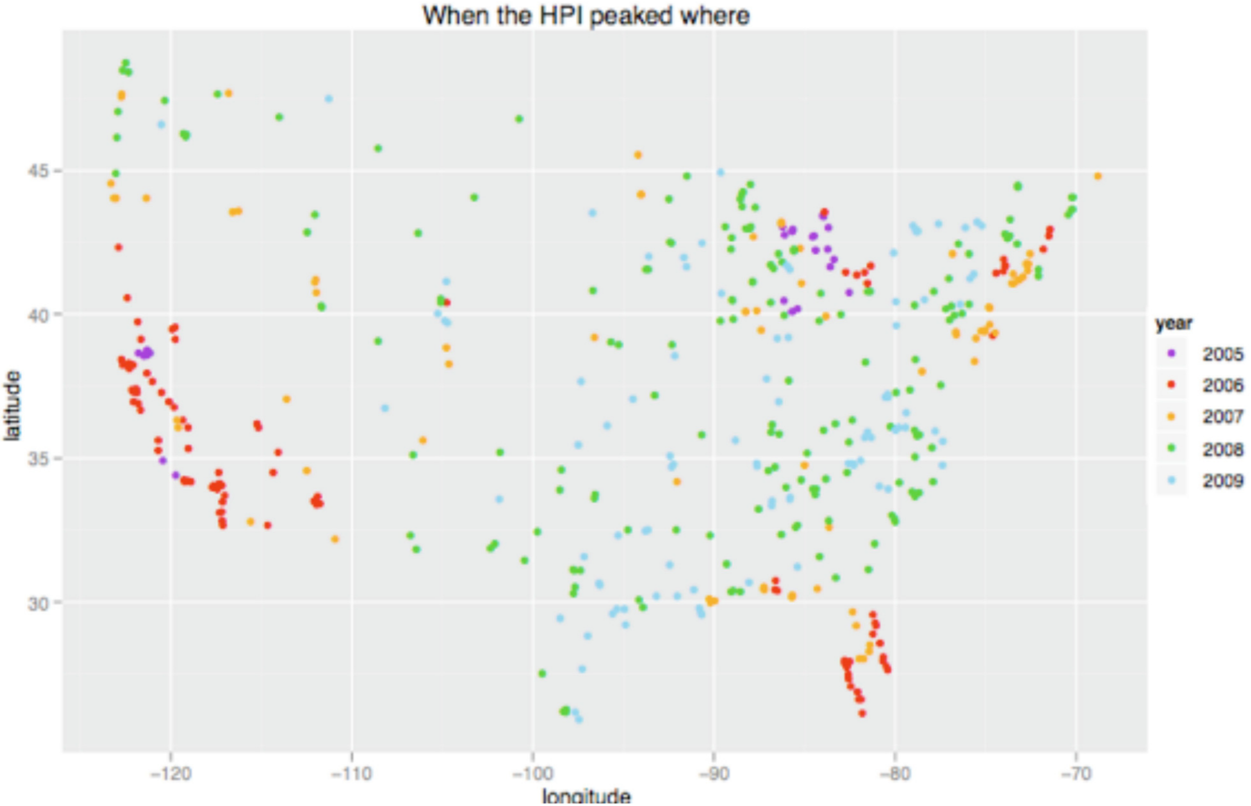
Location, Location, Location...

The data graphed (Fig. 7 & Fig. 8) is from the Federal Housing Finance Agency (FHFA) house price index (HPI). Both of these graphs analyze what time the HPI peaked for each metropolitan statistical area (MSA).

Looking at both graphs we believe that timing seems to be very significant. If a state peaked earlier than 2006 or later than 2007, their HPI was not as greatly affected. This also supports the claim that California and Florida were impacted the greatest.

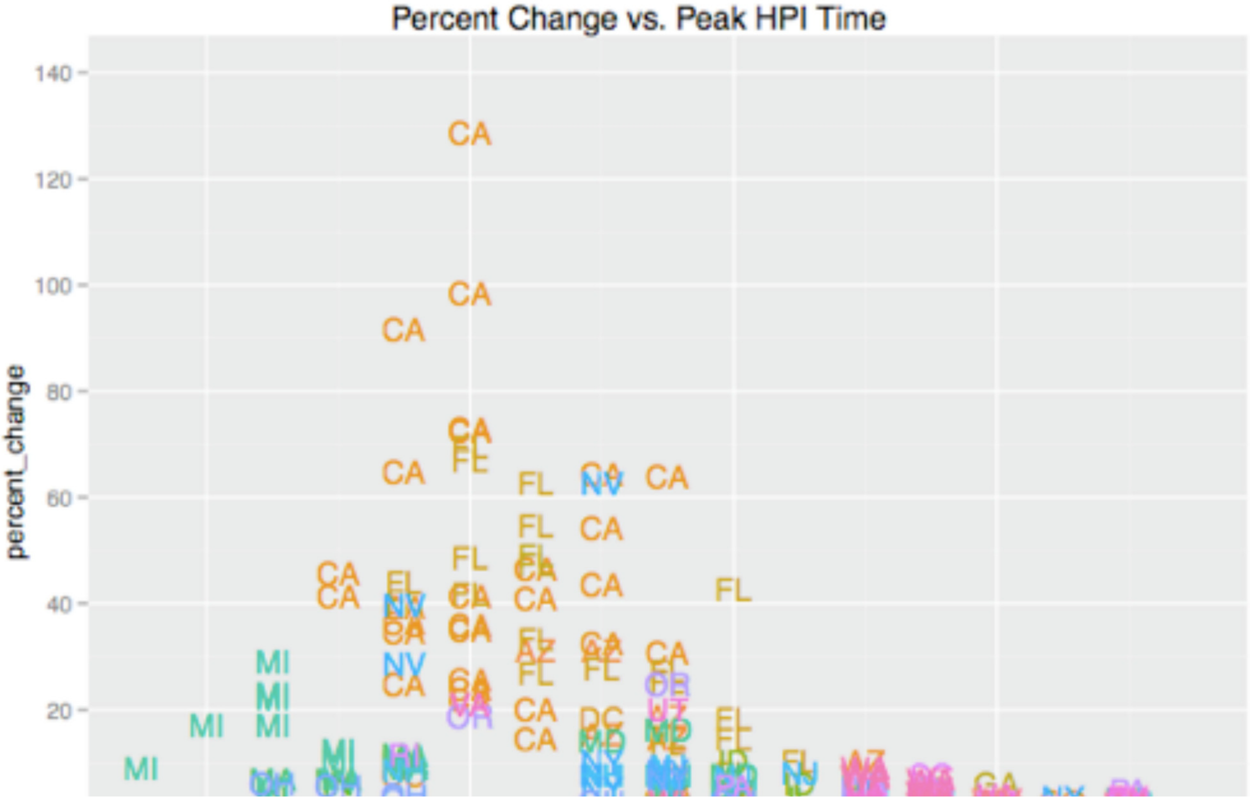
In Figure 7, you can see that both California and Florida peaked around the same time. The graph shows in what year each MSA area reached its maximum housing price.

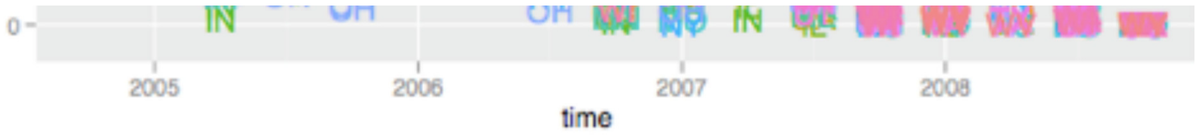
Figure 11.7.


When the HPI peaked where
latitude
45
40
35
30
longitude
-120
-110
-100
-90
-80
-70
year
● 2005
● 2006
● 2007
● 2008
● 2009

In Figure 8, every point is a MSA and labeled by state. It graphs the peak HPI time versus the percent change in HPI between then maximum HPI to 2009, quarter 1 HPI. This graph shows that if HPI peaked between 2006 and 2007, then that state typically experienced a much larger percent change in HPI.

Figure 11.8.


Percent Change vs. Peak HPI Time
percent_change
140
120
100
80
60
40
20
CA
FL
MI
NV
AZ
OR
DC
MD
UT
ID
PA
NJ
GA
WA
NY



Merced, CA

The city with the greatest percent change in the FHFA HPI was Merced, CA. This observation is very unusual of small cities. Further research into Merced showed that University California of Merced has finished construction in late 2005. Using both Figures 9 and 10, we hypothesize that the construction increased due to the necessity of housing for UC Merced students and employees.

Figure 11.9.

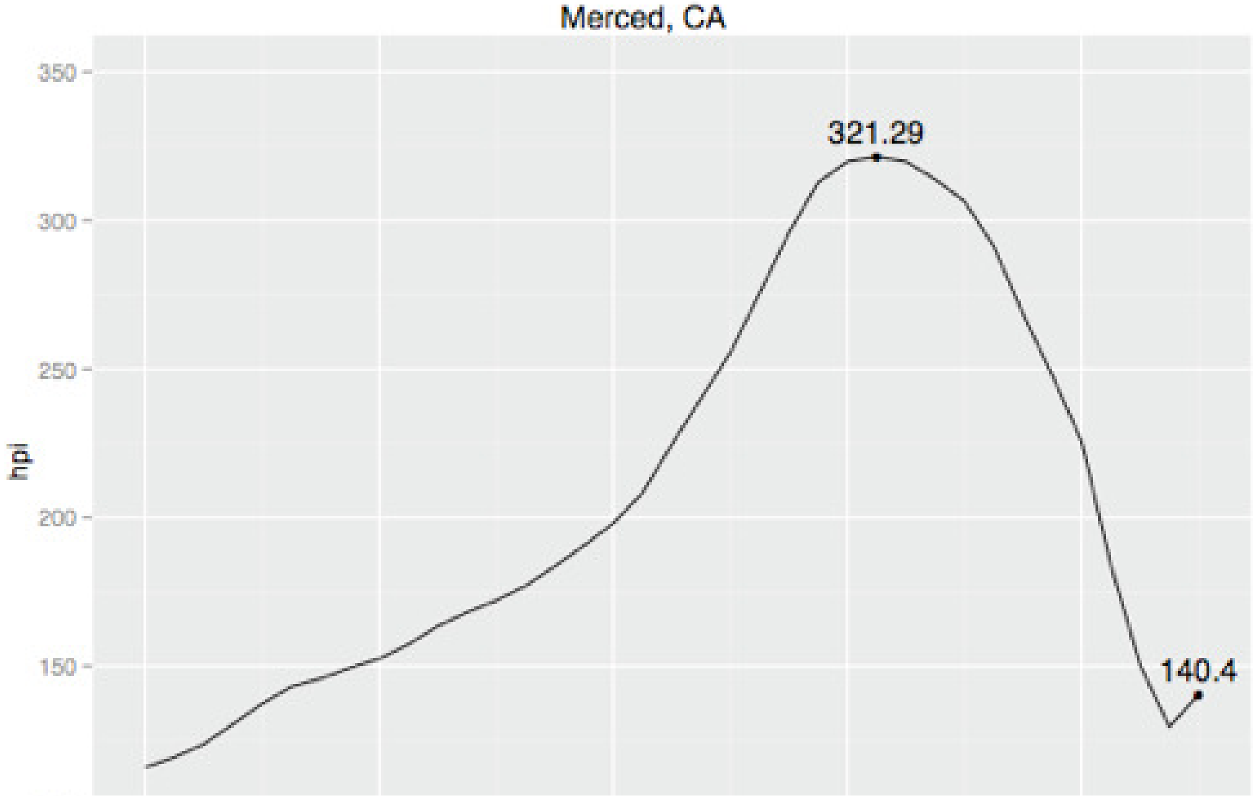
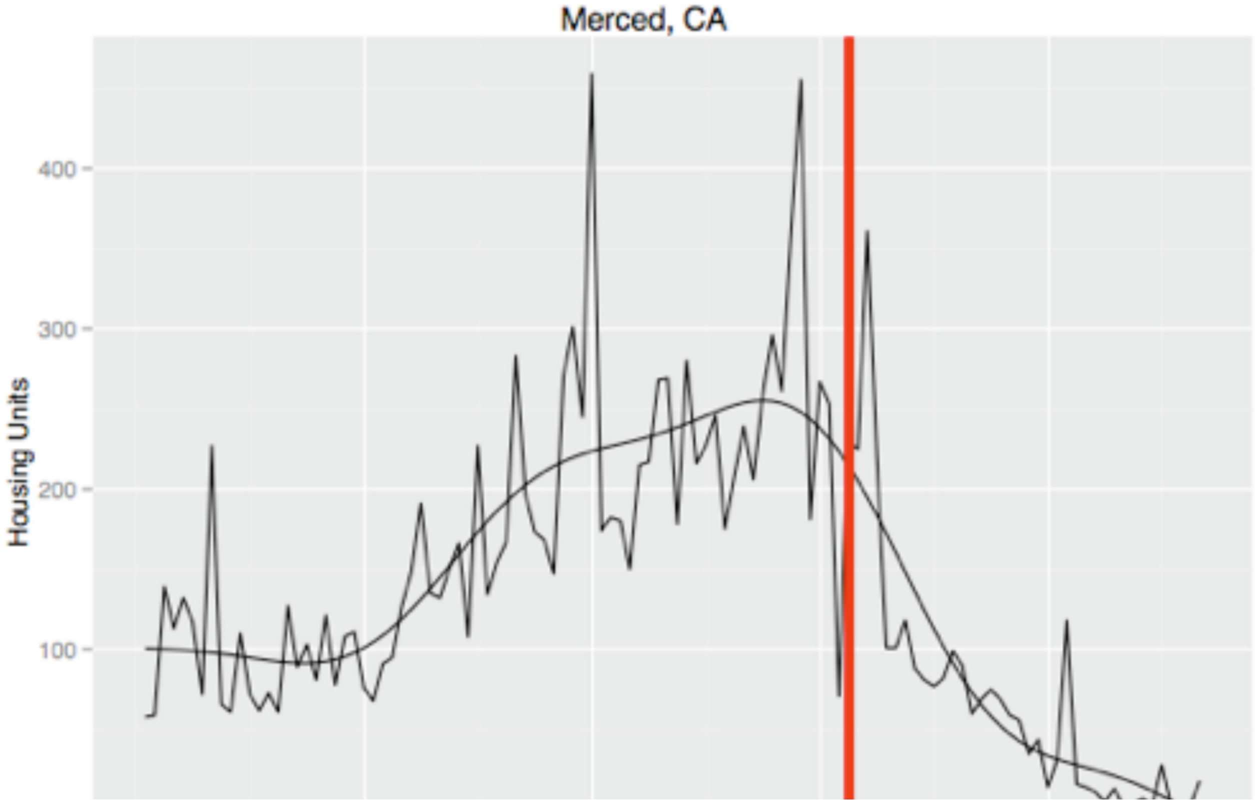
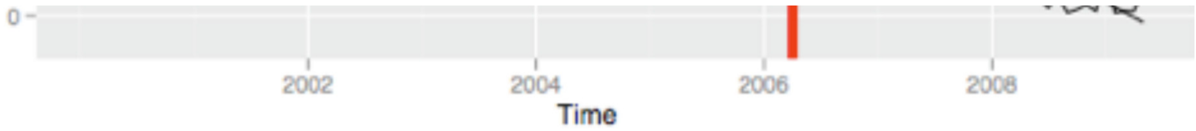

Merced, CA
350
300
250
200
150
hpi
321.29
140.4

Figure 11.10.

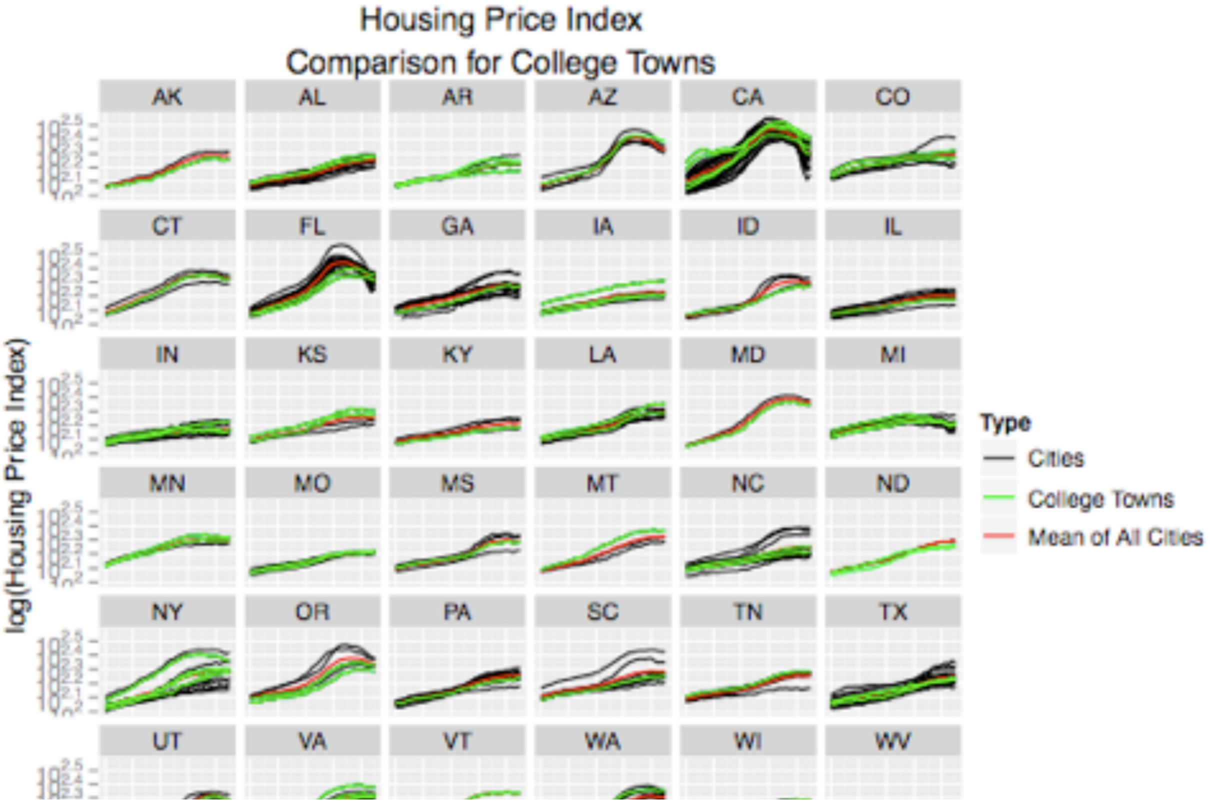

Merced, CA
Housing Units
400
300
200
100

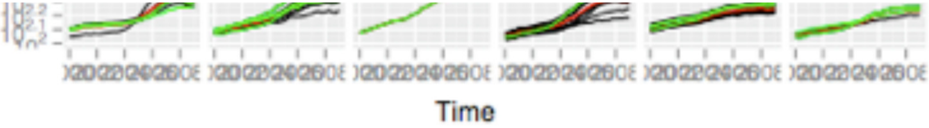


Myth Busters

After discovering Merced, CA we decided to look more closely at college towns. Contrary to belief, college towns were not greatly impacted by the housing crisis. They were affected more by the location that they were in, rather than being a ``college town''. (Fig. 11)

Figure 11.11.


Housing Price Index
Comparison for College Towns
AK
AL
AR
AZ
CA
CO
CT
FL
GA
IA
ID
IL
IN
KS
KY
LA
MD
MI
MN
MO
MS
MT
NC
ND
NY
OR
PA
SC
TN
TX
UT
VA
VT
WA
WI
WV
10^2.5
10^2.4
10^2.3
10^2.2
10^2.1
10^2
log(Housing Price Index)
Type
Cities
College Towns
Mean of All Cities



Other Explorations

- **Vacation Spots**: Are areas where people own a second home more affected?
- **Renting vs. Owning**: Is is better to rent or own a house?
- **Migration**:Are cities that experienced massive population change affected?
- **Gross Domestic Product**: Can we categorize a certain city by industry? Is there a relationship between cities that were hit by the housing crisis?

Communication and Future Work

It is extremely important that all of our data cleaning and findings are reproducible. We've made both the data and programming code available to the public through our PFUG's website on http://github.com/hadley/data-housing-crisis . Github is a very advance website that is able to track changes made to data and code from multiple individuals.

Github is advantageous to both our research group and to the general public. Firstly, we are able to freely store large amounts of data. Also it allows us to work on the same data without having to e-mail changes back and forth. In addition, others can view and download our data for free. We hope that by keeping the code transparent and self-replicating, others are able to easily build off our work.

We would like to develop a website that will allow users to easily access the data they are interested in, which would otherwise be a daunting task for those who wish to use a data set of this size. Because our analysis and findings also involve large amounts of information, (such as construction price time series for each US metropolitan area) we are exploring interactive graphical methods for displaying this information. Our future research will involve using the internet application Many Eyes, **http://manyeyes.alphaworks.ibm.com**, and then eventually the program Protovis,**http://vis.stanford.edu/protovis**, to create this website.

Acknowledgements

This Connexions module describes work conducted as part of Rice University's VIGRE program, supported by National Science Foundation grant DMS--0739420.

11.2. A theoretical model for the data analysis process based on cognitive science*

MOTIVATION

Data analysis is the process by which we glean understanding from data. While the origins of data analysis extend at least as far back as Francis Bacon and certainly further, the term “Data Analysis” was first introduced as a field of academic study in 1962 by John Tukey.

Improvements in technology have increased both the amount of data that we can store and the speed with which we can analyze it (Friedman 1997). With each improvement, data analysis becomes more relevant. Modern commentators now claim we live in the midst of a "data deluge," where we no longer have the cognitive power to understand all of the data available (Hey 2003). Further advances in data collection technology will require further advances in data analysis methods.

The fields of Machine Learning, Data Mining, InfoVis, and Visual Analytics are all attempts to improve upon Data Analysis to better meet our analytical needs. But even with the research already done in these areas, scientists claim that there is very little Data Analysis theory to build upon, and that the theory that is available is hard to access (Unwin 2001, Mallows 2006, Cox 2007). This lack of theoretical understanding stymies improvement in the field. Many academic disciplines create innovations by extending existing theory in new ways. Data analysis appears to proceed through a trial and error process.

Researchers have offered multiple suggestions to remedy this. Cox and Mallows propose reviewing data analysis case studies to induce a general pattern of analysis. Unwin suggests creating a pattern language of Data Analysis similar to the pattern language first proposed by architects Alexander, Ishikawa, and Silverstein (1977), and used successfully in the field of software engineering (Coplien 1996). While we are intrigued by Unwin's proposition, we do not presently have the resources to define a complete pattern language. However, we begin our examination of data analysis by reviewing the data analysis case studies that exist in the literature of statistical consulting, as suggested by Cox and Mallows.

RESEARCH QUESTION

Can the sensemaking model of cognitive science provide a theoretical model for data analysis?

PREVIOUS MODELS OF DATA ANALYSIS

Past efforts to describe data analysis reveal a lack of consensus about the process. Below are three illustrations of the process provided by Box (1976), Box, Hunter, and Hunter (1978), and Wild and Pfannkuch (1999).

Figure 11.12.

B. Data Analysis and Data Getting in the Process of Scientific Investigation[a]

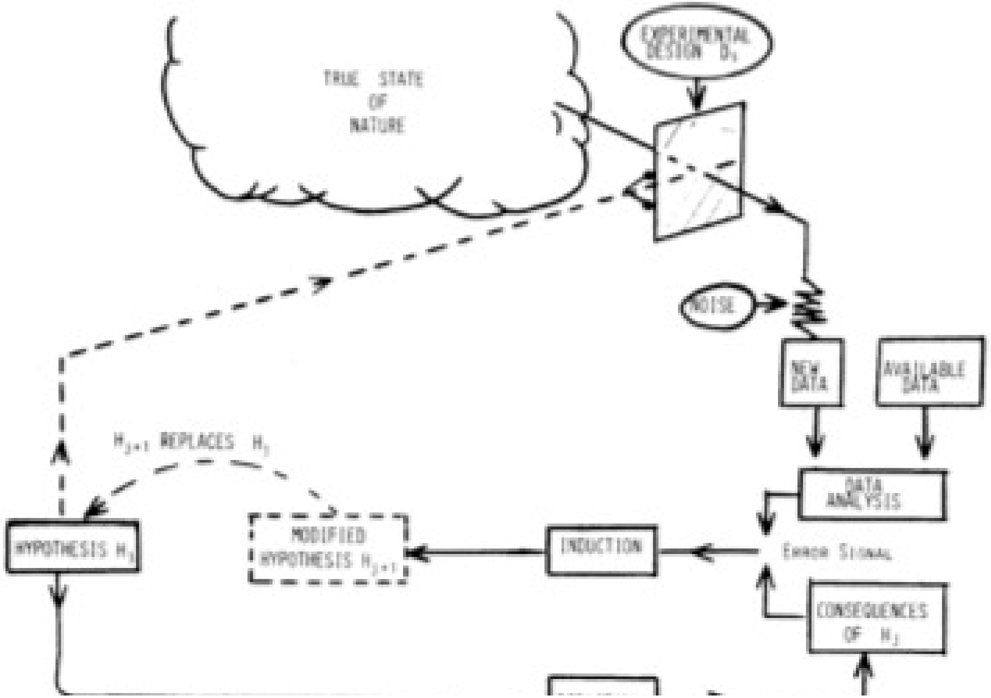


Figure 11.13.

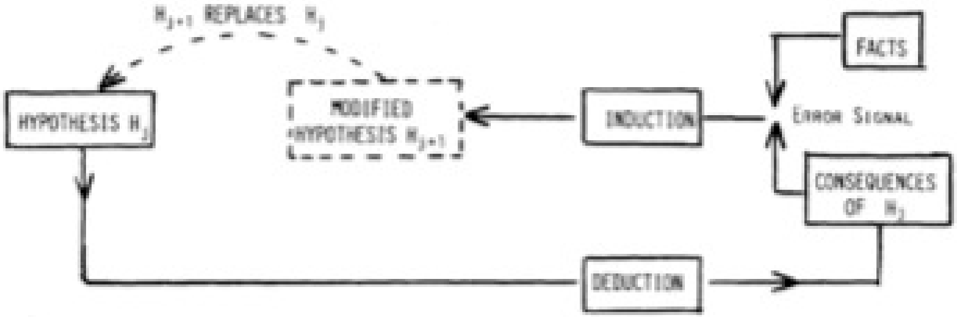
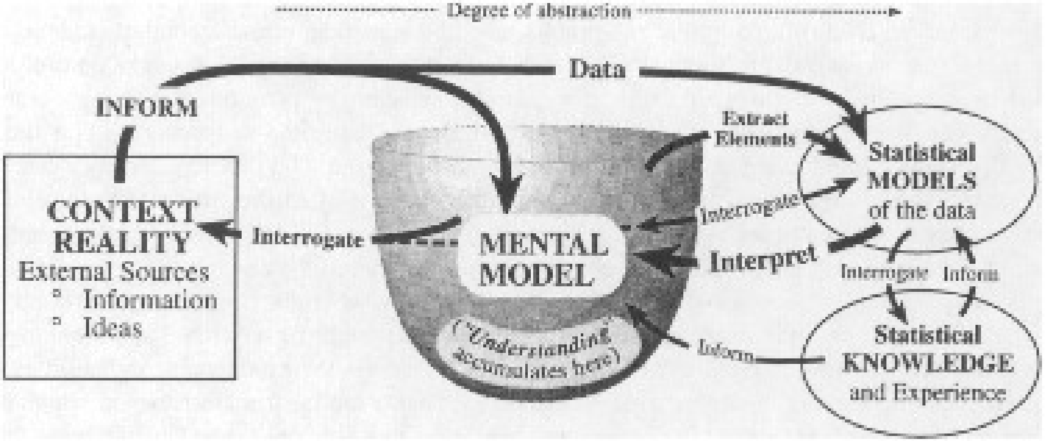


Figure 11.14.



While different, the three diagrams suggest some salient aspects of the data analysis process:

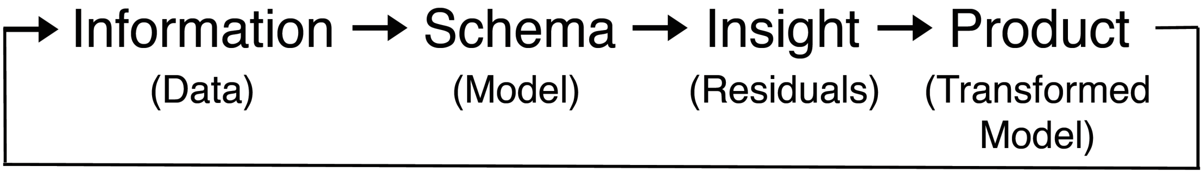
- It is an iterative process
- It uses observable data to adjust a mental model
- It alternates between an inductive phase and a deductive phase
- The aim of data analysis is to create understanding

Data analysis shares these features with a process that has been well studied by cognitive scientists: sense making.

SENSE-MAKING

Sensemaking is an area of cognitive science that examines how the human brain creates understanding from its surroundings. It began in the 1970's as an extension of communication theory, but was then adopted by experimental and theoretical psychologists. According to sensemaking research, the human brain continuously scans its environment for data and builds this data into a mental model that explains its surroundings. A couple of sensemaking models exist to explain how this occurs (e.g, the cost structure model, the data-frame model), but each has the same basic components.

Figure 11.15.



The sensemaking process (Pirolli and Card 2005) with matching data analysis stages provided in parenthesis.

The brain begins with a tentative theory, which is also called a model, a schema, or a frame. This theory suggests to the brain what is and what is not relevant data. The brain then constructs this data from the external stimuli it receives through the sense organs. An important facet of sensemaking is that the mind does not automatically accept all present stimuli as data. It instead decides which stimuli would be relevant, searches for them, and then synthesizes them into a piece of data.

The brain compares its currently held theory to the data it has collected. It confirms the theory if the theory accurately fits the data. Otherwise, it will modify the theory to better fit the data or completely reject the theory in favor of a new one. The process occurs continuously; the brain constantly refines existing theories against new data.

A theory provides understanding by describing the relationships between data. These relationships assign meaning to the data points and also allow predictions of unobserved data from observed data. A theory also allows the mind to encode data more efficiently than just storing the raw bits. In this way, sensemaking resembles parametric modeling. The brain retains the theory instead of the raw data, but retains the information contained in the data in the parameters of the theory. Different types of theories can describe different types of relationships among data. Mental maps describe spatial relationships, stories describe temporal and causal relationships, scripts describe roles, plans describe an intended sequence of events, etc. (Klein et al. 2003)

WHY SENSE-MAKING?

Sensemaking shares all of the salient features of data analysis noted above, but there are other reasons to suspect that cognitive science may offer a theoretical foundation for data analysis.

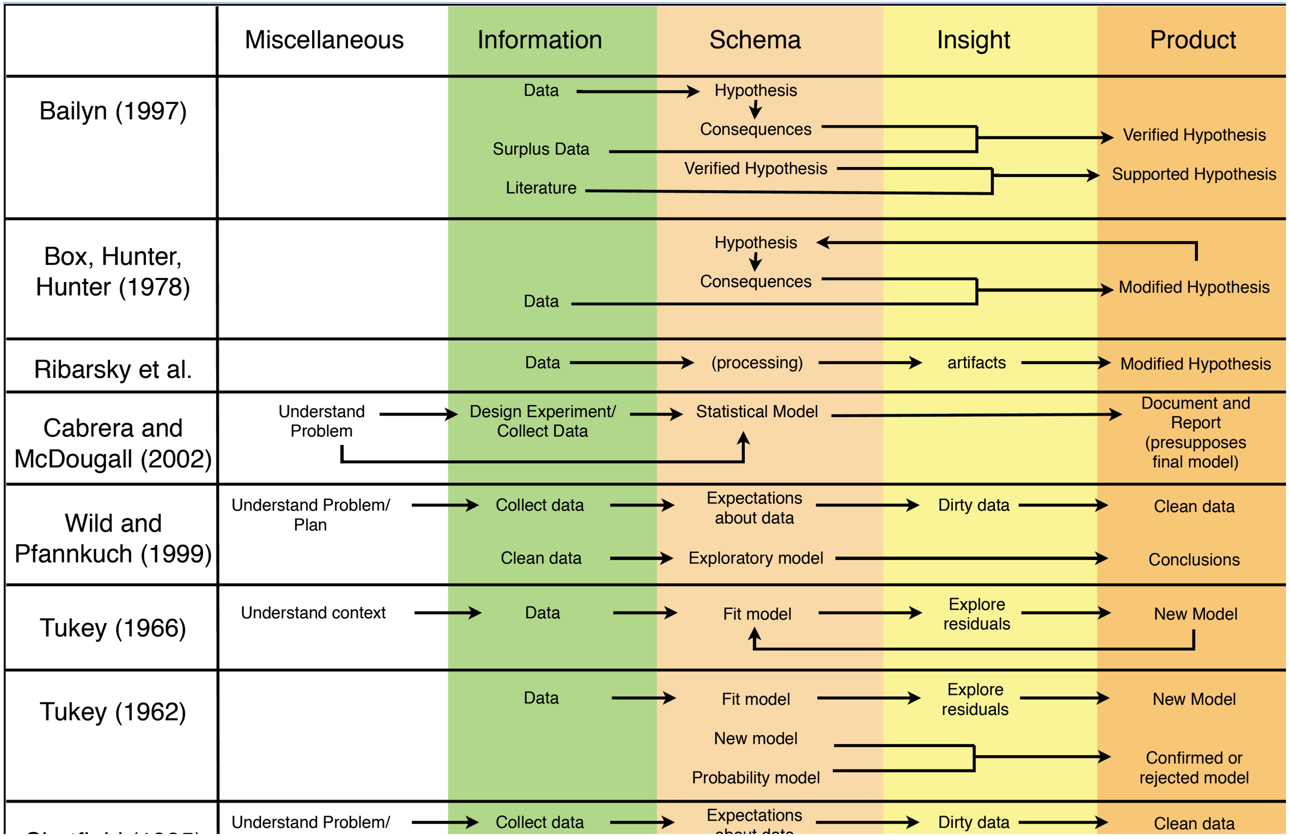
Almost all data analysis is conducted by humans in order to improve their understanding of the world. Hence, data analysis extends the sensemaking process. Moreover, data analysts may use their internal reasoning processes as a model for their data analysis.

As Velleman (1997) points out, data analysis is a revival of Francis Bacon's scientific method and could be considered the modern incarnation of that method. The history of this method resembles a movement from an internal sensemaking process, which can often be subjective, to an external sensemaking process that tries to be objective. If so, we should expect data analysis to display a foundation based on sensemaking with added safeguards against the biases that sensemaking is vulnerable to.

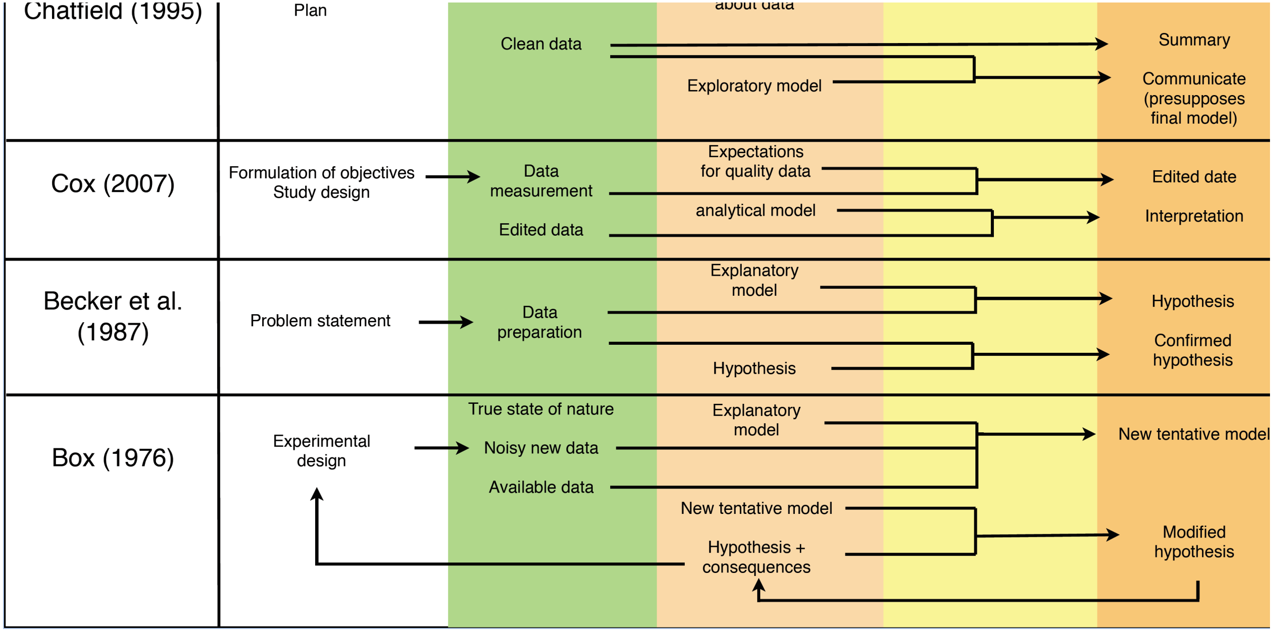
PRELIMINARY RESULTS

I followed Cox and Mallows suggestions and compared data analysis case studies and suggestions available in the statistical literature to the sensemaking model. In all cases most of the data analysis prescriptions fell into one of the four sensemaking steps. The remaining prescriptions were all "meta-steps" which dealt with the data analysis process itself (e.g, plan, understand the problem). These meta-techniques may be evidence that data analysis has incorporated safeguards against the vulnerabilities of the internal sensemaking process. A visual description of the compliance of 11 papers:

Figure 11.16.



	Miscellaneous	Information	Schema	Insight	Product
Bailyn (1997)		Data Surplus Data Literature	Hypothesis Consequences Verified Hypothesis		Verified Hypothesis Supported Hypothesis
Box, Hunter, Hunter (1978)		Data	Hypothesis Consequences		Modified Hypothesis
Ribarsky et al.		Data	(processing)	artifacts	Modified Hypothesis
Cabrera and McDougall (2002)	Understand Problem	Design Experiment/ Collect Data	Statistical Model		Document and Report (presupposes final model)
Wild and Pfannkuch (1999)	Understand Problem/ Plan	Collect data Clean data	Expectations about data Exploratory model	Dirty data	Clean data Conclusions
Tukey (1966)	Understand context	Data	Fit model	Explore residuals	New Model
Tukey (1962)		Data	Fit model New model Probability model	Explore residuals	New Model Confirmed or rejected model
Chatfield (1995)	Understand Problem/	Collect data	Expectations about data	Dirty data	Clean data



LOOKING FORWARD

This preliminary analysis supports the hypothesis that sensemaking may provide a theoretical model for data analysis. Further study must address the question, "How can we provide a rigorous demonstration that data analysis follows a sensemaking model?” As Cox points out, only a small number of data analysis case studies are available in the statistical literature. Future research may employ more direct methods such as observing actual data analyses or scouring computer code used to perform data analyses.

if a cognitive basis is demonstrated, cognitive science may provide opportunities to improve the activity of data analysis. Do current data analysis methods provide adequate safeguards to the well documented list of sensemaking biases?

Finally, a firmly established model for data analysis can be used to expand the academic understanding of the sub-field. The author originally embarked on this study to address the lack of well defined objectives for data visualization techniques. A better definition of the purpose of data analysis methods may provide new opportunities to optimize data analysis techniques.

ACKNOWLEDGEMENTS

- The National Science Foundation
- Dr. Hadley Wickham

REFERENCES

Alexander, et al. (1977). A pattern language: towns, buildings, construction. Oxford University Press, USA.

Bailyn (1977). ‘Research as a cognitive process: Implications for data analysis’. Quality and Quantity **11**(2):97–117.

Becker, et al. (1987). ‘Dynamic Graphics for Data Analysis’. *Statistical Science* **2**(4):355–383.

Box (1976). ‘Science and Statistics’. Journal of the American Statistical Association **71** (356):791–799.

Box, et al. (1978). Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building. John Wiley & Sons.

Cabrera & McDougall (2002). *Statistical consulting*. Springer Verlag.

Chatfield (1995). Problem solving: a statistician’s guide. Chapman & Hall/CRC.

Coplien (1996). *Software patterns*. Citeseer.

Cox (2007). ‘Applied statistics: A review’. *Annals of Applied Statistics* **1**(1):1–16.

Friedman (1997). ‘Data mining and statistics: what’s the connection? ’Computing Science and Statistics: Proceedings of the 29th Symposium on the interface.

Hey & Trefethen (2003). ‘The Data Deluge: An e-Science Perspective’ pp. 809–824.

Klein, et al. (2003). ‘A Data/Frame Theory of Sense Making"’. In Expertise out of context: proceedings of the sixth International Conference on Naturalistic Decision Making, pp. 113–155.

Mallows (2006). ‘Tukey’s Paper after 40 years (with discussion)’. *Technometrics* **48**(3):319–325.

Pirolli & Card (2005). ‘The Sensemaking Process and Leverage Points for Analyst Technology as Identified Through Cognitive Task Analysis’. Proceedings of International Conference on Intelligence.

Ribarsky, et al. (2009). ‘Science of analytical reasoning’. *Information Visualization* **8**(4):254–262.

Tukey & Wilk (1966). ‘Data analysis and statistics: an expository overview’. In *Proceedings of the November 7-10, 1966, fall joint computer conference*, pp. 695– 709. ACM.

Tukey (1962). ‘The Future of Data Analysis’. *The Annals of Mathematical Statistics* **33**(1):1–67.

Wild & Pfannkuch (1999). ‘Statistical thinking in empirical enquiry’. *International Statistical Review/Revue Internationale de Statistique* **67**(3):223–248.

Velleman (1997). The Philosophical Past and the Digital Future of Data Analysis. Princeton University Press.

The Art of the PFUG

Table of Contents

How to Compose a PFUG

1. Algebraic Geometry

2. Knot Theory

3. Calculus of Variations

4. Combinatorial Optimization

5. Computational Biology

6. Computational Finance

7. Computational Neuroscience

8. Imaging

9. Physics of Strings

10. Probability

11. Visualization

12. Outreach

13. Appendix

A. Attributions