A Brain-Friendly Guide

# Head First
# Data Analysis

**A learner's guide to big numbers, statistics, and good decisions**

Predict your raise with linear regression

Experiment to discover who your customers *really* are

Load important statistical concepts directly into your brain

Sell more toys by optimizing your business model

Overcome your cognitive biases

Clean messy data for efficient analysis

O'REILLY®

Michael Milton

# Table of Contents (Summary)

| | | |
|---|---|---|
| | Intro | xxvii |
| 1 | Introduction to Data Analysis: *Break It Down* | 1 |
| 2 | Experiments: *Test Your Theories* | 37 |
| 3 | Optimization: *Take It to the Max* | 75 |
| 4 | Data Visualization: *Pictures Make You Smarter* | 111 |
| 5 | Hypothesis Testing: *Say It Ain't So* | 139 |
| 6 | Bayesian Statistics: *Get Past First Base* | 169 |
| 7 | Subjective Probabilities: *Numerical Belief* | 191 |
| 8 | Heuristics: *Analyze Like a Human* | 225 |
| 9 | Histograms: *The Shape of Numbers* | 251 |
| 10 | Regression: *Prediction* | 279 |
| 11 | Error: *Err Well* | 315 |
| 12 | Relational Databases: *Can You Relate?* | 359 |
| 13 | Cleaning Data: *Impose Order* | 385 |
| i | Leftovers: The *Top Ten Things (We Didn't Cover)* | 417 |
| ii | Install R: *Start R Up!* | 427 |
| iii | Install Excel Analysis Tools: *The ToolPak* | 431 |

# Table of Contents (the real thing)

## Intro

**Your brain on data analysis.** Here *you* are trying to *learn* something, while here your *brain* is doing you a favor by making sure the learning doesn't *stick*. Your brain's thinking, "Better leave room for more important things, like which wild animals to avoid and whether naked snowboarding is a bad idea." So how *do* you trick your brain into thinking that your life depends on knowing data analysis?

| | |
|---|---|
| Who is this book for? | xxviii |
| We know what you're thinking | xxix |
| Metacognition | xxxi |
| Bend your brain into submission | xxxiii |
| Read Me | xxxiv |
| The technical review team | xxxvi |
| Acknowledgments | xxxvii |

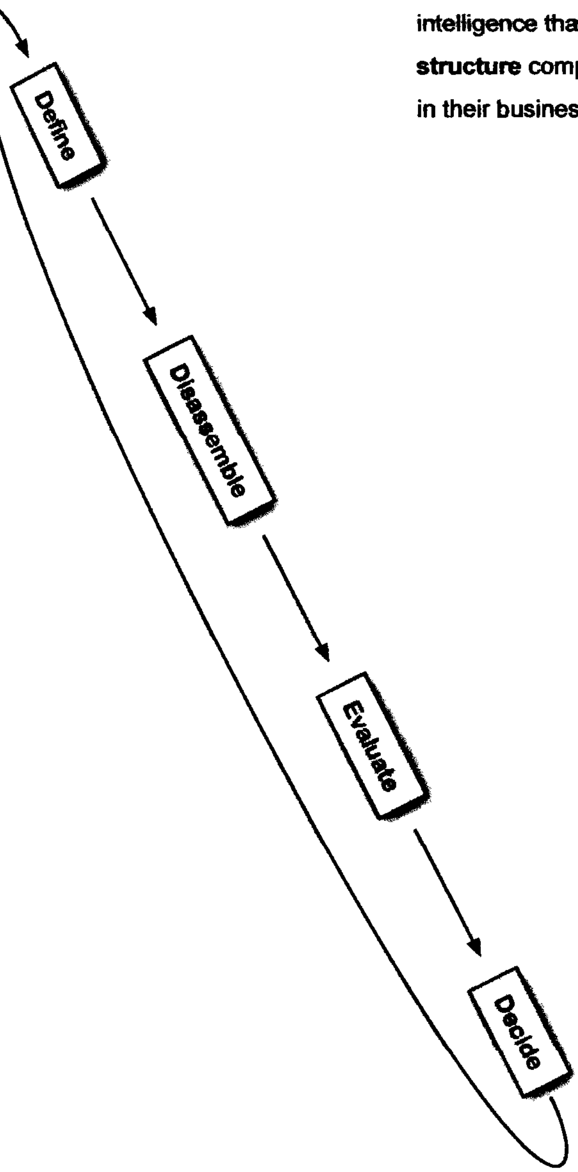
Download at Boykma.Com

# introduction to data analysis

## 1 Break it down

### Data is everywhere.

Nowadays, everyone has to deal with mounds of data, whether they call themselves "data analysts" or not. But people who possess a toolbox of data analysis skills have a **massive edge** on everyone else, because they understand what to *do* with all that stuff. They know how to translate raw numbers into intelligence that **drives real-world action**. They know how to **break down and structure** complex problems and data sets to get right to the heart of the problems in their business.



Acme Cosmetics needs your help 2
The CEO wants data analysis to help increase sales 3
Data analysis is careful thinking about evidence 4
Define the problem 5
Your client will help you define your problem 6
Acme's CEO has some feedback for you 8
Break the problem and data into smaller pieces 9
Now take another look at what you know 10
Evaluate the pieces 13
Analysis begins when you insert yourself 14
Make a recommendation 15
Your report is ready 16
The CEO likes your work 17
An article just came across the wire 18
You let the CEO's beliefs take you down the wrong path 20
Your assumptions and beliefs about the world are your mental model 21
Your statistical model depends on your mental model 22
Mental models should always include what you don't know 25
The CEO tells you what he doesn't know 26
Acme just sent you a huge list of raw data 28
Time to drill further into the data 31
General American Wholesalers confirms your impression 32
Here's what you did 35
Your analysis led your client to a brilliant decision 36

Download at Boykma.Com

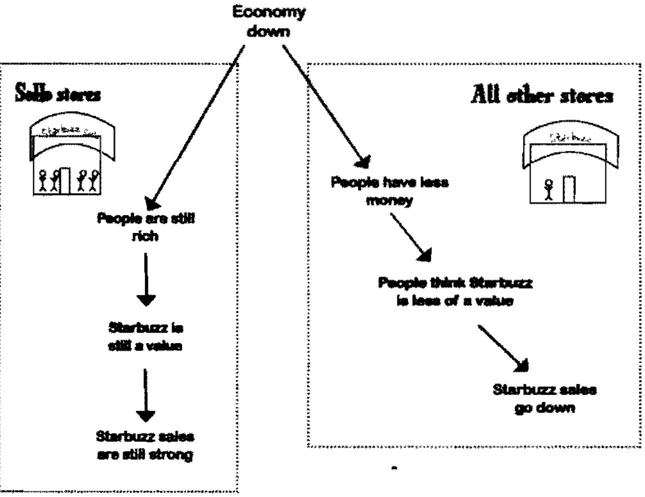
## 2 experiments

# Test your theories

### Can you show what you believe?

**In a real empirical test? There's nothing like a good experiment to solve your problems and show you the way the world really works. Instead of having to rely exclusively on your observational data, a well-executed experiment can often help you make causal connections. Strong empirical data will make your analytical judgments all the more powerful.**

It's a coffee recession! 38
The Starbuzz board meeting is in three months 39
The Starbuzz Survey 41
Always use the method of comparison 42
Comparisons are key for observational data 43
Could value perception be causing the revenue decline? 44
A typical customer's thinking 46
Observational studies are full of confounders 47
How location might be confounding your results 48
Manage confounders by breaking the data into chunks 50
It's worse than we thought! 53
You need an experiment to say which strategy will work best 54
The Starbuzz CEO is in a big hurry 55
Starbuzz drops its prices 56
One month later… 57
Control groups give you a baseline 58
Not getting fired 101 61
Let's experiment ~~again~~ for real! 62
One month later… 63
Confounders also plague experiments 64
Avoid confounders by selecting groups carefully 65
Randomization selects similar groups 67
Randomness Exposed 68
Your experiment is ready to go 71
The results are in 72
Starbuzz has an empirically tested sales strategy 73



Download at Boykma.Com

optimization

# 3 Take it to the max

## We all want more of something.

And we're always trying to figure out how to get it. *If* the things we want more of—profit, money, efficiency, speed—can be represented numerically, then chances are, there's an tool of data analysis to help us tweak our *decision variables,* which will help us find the **solution** or *optimal point* where we get the most of what we want. In this chapter, you'll be using one of those tools and the powerful spreadsheet **Solver** package that implements it.

You're now in the bath toy game 76

Constraints limit the variables you control 79

Decision variables are things you can control 79

You have an optimization problem 80

Find your objective with the objective function 81

Your objective function 82

Show product mixes with your other constraints 83

Plot multiple constraints on the same chart 84

Your good options are all in the feasible region 85

Your new constraint changed the feasible region 87

Your spreadsheet does optimization 90

Solver crunched your optimization problem in a snap 94

Profits fell through the floor 97

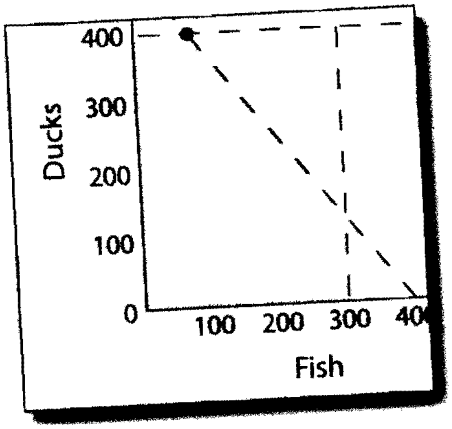
Your model only describes what you put into it 98

Calibrate your assumptions to your analytical objectives 99

Watch out for negatively linked variables 103

Your new plan is working like a charm 108

Your assumptions are based on an ever-changing reality 109



Download at Boykma.Com

## data visualization

# 4 Pictures make you smarter

### You need more than a table of numbers.

**Your data is brilliantly complex, with more variables than you can shake a stick at. Mulling over mounds and mounds of spreadsheets isn't just boring; it can actually be a waste of your time. A clear, highly multivariate visualization can, in a small space, show you the forest that you'd miss for the trees if you were just looking at spreadsheets all the time.**

New Army needs to optimize their website 112

The results are in, but the information designer is out 113

The last information designer submitted these three infographics 114

What data is behind the visualizations? 115

Show the data! 116

Here's some unsolicited advice from the last designer 117

Too much data is never your problem 118

Making the data pretty isn't your problem either 119

Data visualization is all about making the right comparisons 120

Your visualization is already more useful than the rejected ones 123

Use scatterplots to explore causes 124

The best visualizations are highly multivariate 125

Show more variables by looking at charts together 126

The visualization is great, but the web guru's not satisfied yet 130

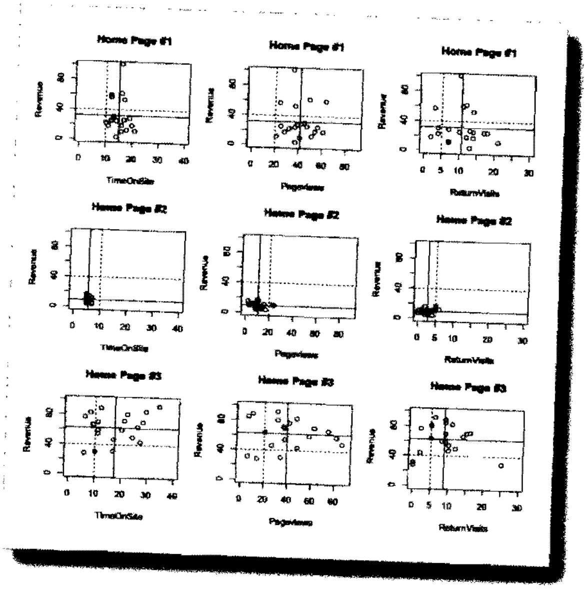
Good visual designs help you think about causes 131

The experiment designers weigh in 132

The experiment designers have some hypotheses of their own 135

The client is pleased with your work 136

Orders are coming in from everywhere! 137



Download at Boykma.Com

# 5 hypothesis testing

## Say it ain't so

**The world can be tricky to explain.**

**And it can be fiendishly difficult when you have to deal with complex, heterogeneous data to anticipate future events. This is why analysts don't just take the obvious explanations and assume them to be true: the careful reasoning of data analysis enables you to meticulously evaluate a bunch of options so that you can incorporate all the information you have into your models. You're about to learn about falsification, an unintuitive but powerful way to do just that.**

Gimme some skin... 140
When do we start making new phone skins? 141
PodPhone doesn't want you to predict their next move 142
Here's everything we know 143
ElectroSkinny's analysis does fit the data 144
ElectroSkinny obtained this confidential strategy memo 145
Variables can be negatively or positively linked 146
Causes in the real world are networked, not linear 149
Hypothesize PodPhone's options 150
You have what you need to run a hypothesis test 151
Falsification is the heart of hypothesis testing 152
Diagnosticity helps you find the hypothesis with the least disconfirmation 160
You can't rule out all the hypotheses, but you can say which is strongest 163
You just got a picture message... 164
It's a launch! 167

Download at Boykma.Com

## 6 bayesian statistics

## Get past first base

### You'll always be collecting new data.

**And you need to make sure that every analysis you do incorporates the data you have that's relevant to your problem. You've learned how *falsification* can be used to deal with heterogeneous data sources, but what about straight up probabilities? The answer involves an extremely handy analytic tool called Bayes' rule, which will help you incorporate your base rates to uncover not-so-obvious insights with ever-changing data.**

The doctor has disturbing news 170

Let's take the accuracy analysis one claim at a time 173

How common is lizard flu really? 174

You've been counting false positives 175

All these terms describe conditional probabilities 176

You need to count false positives, true positives, false negatives, and true negatives 177

1 percent of people have lizard flu 178

Your chances of having lizard flu are still pretty low 181

Do complex probabilistic thinking with simple whole numbers 182

Bayes' rule manages your base rates when you get new data 182

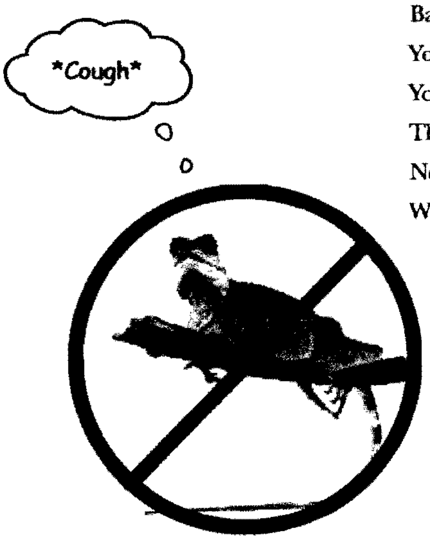
You can use Bayes' rule over and over 183

Your second test result is negative 184

The new test has different accuracy statistics 185

New information can change your base rate 186

What a relief! 189



Download at Boykma.Com