G. A. REYNOLDS

THINK DATA

A CONCEPTUAL INTRODUCTION TO BIG DATA

FOR THE SKEPTICAL, THE PESSIMISTIC, AND THE MILDLY DISTURBED

Remark 1 Two tasks: conceptual and computational. Understand stats first, then learn how to *do* stats using software (statistics machines).

Contents

1
Introduction

What is data?

Wrong question. The right questions are: what is the role of data? What functions does it serve? How is it used? etc. Our practices involving data tell us what data "is"; there can be no antecedently defined notion of data. (I.e. in spite of the etymology, no data are "given".)

> Myth of the Given
>
> Sellars

See also:
Introduction to Data Technologies (Free online version available)
Clojure Data Analysis Cookbook

Part I

Modeling

Ed. note 1.0.1 *Modeling the world as data; it really should be called "world modeling". E.g. in industry, a "standard data model" is a standard way of representing some domain of interest, such as electrical connectors or transmission lines. TODO: clear, simple examples. Maybe Pipeline Open Data Standard (PODS)? "The PODS Pipeline Data Model provides the database architecture pipeline operators use to store critical information and analysis data about their pipeline systems, and manage this data geospatially in a linear-referenced database which can then be visualized in any GIS platform. The PODS Pipeline Data Model houses the asset information, inspection, integrity management, regulatory compliance, risk analysis, history, and operational data that pipeline companies have deemed mission-critical to the successful management of natural gas and hazardous liquids pipelines."*

TODO: how is this concept of data model related to statistical notions of "model" and modeling?

Part II

Description

Ed. note 1.0.2 *Description of data, not data as description of world - that's modeling*

In principle, every datum must be associated with a complete description that determines both its form and its meaning.

Syntax

Semantics

Type Same as syntax?

Traceability esp. given ability to rename, split, and generally slice and dice, traceability is critical.

2
Datum: description of types and individuals

3
Data: description of data aggregates

Descriptive statistics. The technical statistical side of such description is covered in detail in the companion volume. Here we cover ...?

4
Variable

- Statistical Variables and Characteristics (UNECE METIS wiki)
- "A variable is a characteristic of a statistical unit being observed that may assume more than one of a set of values to which a numerical measure or a category from a classification can be assigned." (Statistics Canada)
- "A variable defines the concept of an observation (or measurement) for a given statistical unit type. The variable describes the concept of the observation that the data item results from. Thus, the variable is always associated with a contextual variable that describes the concept of the variable in the context of a particular statistical activity." (Neuchâtel Terminology Model)
- "Definition: A characteristic of a unit being observed that may assume more than one of a set of values. Context: A variable in the mathematical sense, i.e. a quantity which may take any one of specified set of values. It is convenient to apply the same word to denote non-measurable characteristics, e.g., 'sex' is a variable in this sense since any human individual may take one of two 'values', male or female. It is useful, but far from being the general practice, to distinguish between a variable as so defined and a random variable (The International Statistical Institute, "The Oxford Dictionary of Statistical Terms", edited by Yadolah Dodge, Oxford University Press, 2003)." (SDMX Statistical Guidelines Part IV Metadata Common Vocabulary)

5
Metadata

6
Codebook

Variable name

Variable semantics e.g. “age in years”

Variable label or description (often an informal definition)

Variable “level” nominal, ordinal, interval, ratio

Variable types *numeric* *integer*

real

Unit of Measure e.g. years

enumerations - code, label, meaning

dates

durations

etc.

free text

missing data

Semantic constraints e.g. range constraints on age

7
Quality

How is data quality defined?

Part III

Management

Ed. note 7.0.3 *Here the OAIS model is relevant.*

ICPSR Guide to Social Science Data Preparation and Archiving

8
Acquisition

8.1 Conversion

I.e. input adapters, import, reformatting, etc.

Conversion occurs at both ends, not just dissemination.

9
Curation

AKA Data Management.

Databases, metadata, ingest/disseminate procs, etc.

9.1 Metadata

9.2 Provenance

I.e. traceability

9.3 Preservation

9.4 Discovery

9.5 Access

e.g. ACLs; e.g. NORC data enclave

9.6 Confidentiality

Applies to both respondents and users of data

10
Dissemination

10.1 Packaging

10.2 Conversion

I.e. export to formats

Conversion occurs at both ends, not just dissemination.

10.3 Presentation

Part IV

Transformation

Ed. note 10.3.1 *Why separate Data Transformation from Data Management? Because you might need ad-hoc transformations. You might need to clean data before archiving it, but such data munging is conceptually distinct from data management.*

UNECE Statistical Data Editing
UNECE GLOSSARY OF TERMS ON STATISTICAL DATA EDITING

11

Normalization: Cleaning, Imputation, etc.

12
Anonymization

A major issue for Social Science data.

Anonymisation: managing data protection risk code of practice (UK ICO)

UNECE Statistical confidentiality and disclosure protection

12.1 Disclosure Risk Analysis

13
Munging

ref: the PADS project

Data Management - a comparison of the commands used for common data "management" (actually, munging) tasks in R, SAS, SPSS and Stata.

See also OpenRefine "a powerful tool for working with messy data, cleaning it, transforming it from one format into another, extending it with web services, and linking it to databases"

13.1 Misc

Remark 2 examples taken from R, so far

plyr: "plyr is a set of tools for a common set of problems: you need to split up a big data structure into homogeneous pieces, apply a function to each piece and then combine all the results back together"

R Data Manipulation Manipulating Data

13.2 Subsetting Data

Subset data in R

13.3 Weights

13.4 Aggregation

Aggregate Data in R Using data.table - here "aggregate" seems to be synonymous with "compute statistic of". Misnomer; an aggregate is not a summary.

13.5 Recoding

Recode variables (examples)

New vars, recoding vars, renaming vars

Recoding vars: categorical, continuous, new

Recode data in R: replacement, recoding

Recode one column, output values into another column: transform(), replace()

The Recode Command From the Package Car

13.6 New vars

"copy of an existing field. Sometimes you don't want to recode data but instead just want another column containing the same data."

Recode into A New Field Using Data From An Existing field And Criteria from Another Field

13.7 Replacing data

"replace the data in an existing field when you want to replace the data for every row (no criteria)." http://rprogramming.net/recode-data-in-r/

13.8 Renaming vars

Part V

Visualization

Data Visualization in Sociology
Visualizing Categorical Data (Friendly, M.)

Part VI

Exploration

EDA as involving the combination of statistical description and visualization.

The Split-Apply-Combine Strategy for Data Analysis

Part VII

Tools & Techniques

14 Overview

14.1 Standards

OAIS

DDI

SMDX

IHSN Metadata Standards and Models

14.2 Editors

Emacs

Vi

Eclipse

etc

14.3 Languages

- Computing Trends Lead to New Programming Languages
- Twelve New Programming Languages: Is Cloud Responsible?
- The Popularity of Data Analysis Software (2014) A very detailed data-driven analysis of the relative popularity of various languages.

Remark 3 Organize by category?

15 Standards

Ed. note 15.0.1 *TODO: standards v. data models*

15.1 OAIS

Reference Model for an Open Archival Information System (OAIS)

15.2 Statistical Standards

- UN Global Inventory of Statistical Standards
- Statistical Standards Program - National Center for Education Statistics
- OECD Glossary of Statistical Terms
- Standards for Statistical Interpretation of Data (British Standards Institute)
- OMB Statistical Programs and Standards

15.2.1 DDI

15.2.2 SDMX

SDMX - Statistical Data and Metadata eXchange

16
R

R is a very popular hideous language.

Resources:

- Programming in R
- R Programming Wiki
- Why R is hard to learn

16.1 Overview

16.1.1 Lispiness

R is lispy, just like javascript is lispy.

"R presents a friendlier interface to programming than Lisp does, at least to someone used to mathematical formulas and C-like control structures, but the engine is really very Lisp-like. R allows direct access to parsed expressions and functions and allows you to alter and subsequently execute them, or create entirely new functions from scratch."http://cran.r-project.org/doc/manuals/R-lang.html#Computing-on-the-language

"There are three kinds of language objects that are available for modification, calls, expressions, and functions."

16.1.2 Meta-types

Ed. note 16.1.1 *Note the weird language. The "language objects" seem to meta-types used in the parse tree, or something like that. Syntactic, meta-objects, not objects "in" the language.*

From http://cran.r-project.org/doc/manuals/R-lang.html#Objects

"There are three types of objects that constitute the R language.[1] They are calls, expressions, and names...These objects have modes "call", "expression", and "name", respectively. They can be created directly from expressions using the quote mechanism and converted to and from lists by the as.list and as.call functions. Components of the parse tree can be extracted using the standard indexing operations."

That can't be quite right; at any rate, calls, expressions, and names are not defined in the section on basic types. Apparently these are "language objects", or meta-objects rather than objects in the language.

"Parsed expressions are stored in an R object containing the parse tree. A fuller description of such objects can be found in Language objects and Expression objects."http://cran.r-project.org/doc/manuals/R-lang.html#Parser

[1] Why call them objects? Why not "types of expression in the language"?

Language "objects":

Call " The most direct method of obtaining a call object is to use quote with an expression argument, e.g.,

```
> e1 <- quote(2 + 2)
> e2 <- quote(plot(x, y))
```

Then both e1 and e1 have "mode" of "call"; so apparently "call object" is r-speak for "function application expression".

Name "The components of a call object are accessed using a list-like syntax, and may in fact be converted to and from lists using as.list and as.call"

"All the components of the call object have mode "name" in the preceding examples. This is true for identifiers in calls, but the components of a call can also be constants—which can be of any type, although the first component had better be a function if the call is to be evaluated successfully—or other call objects, corresponding to subexpressions. Objects of mode name can be constructed from character strings using as.name, so one might modify the e2 object as follows"

Expression "An expression contains one or more statements. A statement is a syntactically correct collection of tokens. Expression objects are special language objects which contain parsed but unevaluated R statements. The main difference is that an expression object can contain several such expressions. Another more subtle difference is that objects of type "expression" are only evaluated when explicitly passed to eval, whereas other language objects may get evaluated in some unexpected cases."http://cran.r-project.org/doc/manuals/R-lang.html#Expression-objects

Expression objects "are very similar to lists of call objects."

16.1.3 Metaprogramming

"It is possible for a function to find out how it has been called by looking at the result of sys.call...However, this is not really useful except for debugging ...More often one requires the call with all actual arguments bound to the corresponding formals. To this end, the function match.call is used...The primary use of this technique is to call another function with the same arguments, possibly deleting some and adding others."

"The call can be treated as a list object where the first element is the name of the function and the remaining elements are the actual argument expressions, with the corresponding formal argument names as tags."

"Two further functions exist for the construction of function calls, namely call and do.call."

"It is often useful to be able to manipulate the components of a function or closure. R provides a set of interface functions for this purpose."

body Returns the expression that is the body of the function.

formals Returns a list of the formal arguments to the function. This is a pairlist.

environment Returns the environment associated with the function.

body<- This sets the body of the function to the supplied expression.

formals<- Sets the formal arguments of the function to the supplied list.

environment<- Sets the environment of the function to the specified environment.

"It is also possible to alter the bindings of different variables in the environment of the function, using code along the lines of evalq(x <- 5, environment(f)). It is also possible to convert a function to a list using as.list. The result is the concatenation of the list of formal arguments with the function body. Conversely such a list can be converted to a function using as.function. This functionality is mainly included for S compatibility. Notice that environment information is lost when as.list is used, whereas as.function has an argument that allows the environment to be set."

16.1.4 Evaluation

Lazy Eval

"Promise objects are part of R's lazy evaluation mechanism. They contain three slots: a value, an expression, and an environment. When a function is called the arguments are matched and then each of the formal arguments is bound to a promise. The expression that was given for that formal argument and a pointer to the environment the function was called from are stored in the promise. Until that argument is accessed there is no value associated with the promise. When the argument is accessed, the stored expression is evaluated in the stored environment, and the result is returned. The result is also saved by the promise. The substitute function will extract the content of the expression slot. This allows the programmer to access either the value or the expression associated with the promise. Within the R language, promise objects are almost only seen implicitly: actual function arguments are of this type. There is also a delayedAssign function that will make a promise out of an expression. There is generally no way in R code to check whether an object is a promise or not, nor is there a way to use R code to determine the environment of a promise."http://cran.r-project.org/doc/manuals/R-lang.html#Promise-objects

"A formal argument is really a promise, an object with three slots, one for the expression that defines it, one for the environment in which to evaluate that expression, and one for the value of that expression once evaluated."

16.1.5 Control

Mapping/iteration: the *apply family:

Do loops?

Recursion?

16.1.6 Types

Ed. note 16.1.2 *Note the difference between class() and mode(). E.g.:* > *class(baseball[,3:5]) [1] "data.frame"* > *mode(baseball[,3:5]) [1] "list"* >

"Symbols refer to R objects. The name of any R object is usually a symbol. Symbols can be created through the functions as.name and quote. Symbols have mode "name", storage mode "symbol", and type "symbol". They can be coerced to and from character strings using as.character and as.name. They naturally appear as atoms of parsed expressions, try e.g. as.list(quote(x + y))."

"In R one can have objects of type "expression". An expression contains one or more statements. A statement is a syntactically correct collection of tokens. Expression objects are special language objects which contain parsed but unevaluated R statements. The main difference is that an expression object can contain several such expressions. Another more subtle difference is that objects of type "expression" are only evaluated when explicitly passed to eval, whereas other language objects may get evaluated in some unexpected cases. An expression object behaves much like a list and its components should be accessed in the same way as the components of a list."

"In R functions are objects and can be manipulated in much the same way as any other object. Functions (or more precisely, function closures) have three basic components: a formal argument list, a body and an environment...A function's environment is the environment that was active at the time that the function was created. Any symbols bound in that environment are captured and available to the function. This combination of the code of the function and the bindings in its environment is called a 'function closure', a term from functional programming theory. In this document we generally use the term 'function', but use 'closure' to emphasize the importance of the attached environment. When a function is called, a new environment (called the evaluation environment) is created, whose enclosure (see Environment objects) is the environment from the function closure. This new environment is initially populated with the unevaluated arguments to the function; as evaluation proceeds, local variables are created within it."

16.2 Data Acquisition

a/k/a inputting and importing data. See R Data Import/Export (CRAN)

The "read" family (package:utils):

read.table

read.csv

read.csv2

read.delim

read.delim2

Troubleshooting:

- Error: more colums than column names. Try using read.csv instead of read.table.

16.2.1 Table data

16.2.2 CSV data

16.2.3 Spreadsheet data

16.2.4 SQL data

16.2.5 Imports

SPSS files

SAS files

Stata files ?

Minitab files

Other

16.3 Data Exploration

Use View(mytbl) to get a spreadsheet view.

print

head/tail

ncol, nrow

colnames (= names)

row.names (for dataframes) rownames (for arrays)

dim

summary() - for numerics:

```
       Q3
Min.   :18.00
1st Qu.:32.75
Median :40.50
Mean   :40.83
3rd Qu.:52.25
Max.   :60.00
```

for factors:

```
                                Q15
Barber shop                   : 2
Beauty salon                  : 2
Retail store                  : 2
Barber                        : 1
car body shop x painting cars: 1
Clothes factory               : 1
(Other)                       :31
```

str(): brief description of data, e.g. str(b652)

```
'data.frame': 40 obs. of  41 variables:
 $ DataEntry: Factor w/ 1 level "GAR": 1 1 1 1 1 1 1 1 1 1 ...
 $ Serial   : Factor w/ 40 levels "0301A","0325A",..: 1 33 40 39 37 35 38 34 26 36 ...
 $ Rev      : Factor w/ 1 level "6.5.2": 1 1 1 1 1 1 1 1 1 1 ...
 $ Q0       : int  1 1 2 2 2 1 2 2 2 2 ...
 $ Q1       : int  1 1 1 1 1 1 1 1 1 1 ...
 ... etc. ...
```

16.4 Data Munging

16.4.1 Recoding

Cookbook for R: Recoding data

lots of examples at Recode Data in R

Package: Car recode

Package: plyr

rename revalue

mapvalues

16.5 Categorical Data

"factor" - I don't know the origin of this terminology, but it looks like a portmanteau word combining f-something with "vector". In any case, "free vector" is a pretty good description of what a factor is: a vector whose values are free to vary (unlike, say, an integer vector).

To get a col vector from dataframe foo we have several options:
By colname:

- foo[,"bar"]

By numeric index:

- foo[[n]]
- foo[,n]

Examples:
factor(foo[x]) => error factor(foo[,x]) => ok
levels(foo[x]) => error levels(foo[,x]) => ok

17
Python

Resources:

- A Roadmap for Rich Scientific Data Structures in Python
- On the growth of R and Python for data science
- Getting Started With Python For Data Science
- Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython (O'Reilly book) "Python for Data Analysis is concerned with the nuts and bolts of manipulating, processing, cleaning, and crunching data in Python. It is also a practical, modern introduction to scientific computing in Python, tailored for data-intensive applications. This is a book about the parts of the Python language and libraries you'll need to effectively solve a broad set of data analysis problems. This book is not an exposition on analytical methods using Python as the implementation language."

Numpy

Scipy

IPython

Pandas high-performance, easy-to-use data structures and data analysis tools for the Python programming language.

18
Ruby

- SciRuby
- StatSample

19 Java

Resources:

JavaNumerics

Parallel Colt

Colt a set of Open Source Libraries for High Performance Scientific and Technical Computing in Java.

Processing

JFreeChart

Weka a collection of machine learning algorithms for data mining tasks

20
Clojure

Incanter

21
Scala

Scala is a java-based language.

Resources:

Scala as a platform for statistical computing and data science good blog post discussing a feature list for statistical computing languages

Brief introduction to Scala and Breeze for statistical computing

22
Julia

Types and "multiple dispatch" are "the core unifying features of Julia: functions are defined on different combinations of argument types, and applied by dispatching to the most specific matching definition." http://docs.julialang.org/en/latest/manual/introduction/

"Although it seems a simple concept, multiple dispatch on the types of values is perhaps the single most powerful and central feature of the Julia language."http://docs.julialang.org/en/release-0.2/manual/methods/#id2

23
Other

23.1 OCaml

Objective Caml for Scientists

23.2 F#

Using F# for Data Science

23.2.1 Experimental/Research

- Chapel
- X10 IBM Research; “Both its modern, type-safe sequential core and simple programming model for concurrency and distribution contribute to making X10 a high-productivity language in the HPC and Big Data spaces.”

24

J

J is a hideous language. But "J is particularly strong in the mathematical, statistical, and logical analysis of data."(http://www.jsoftware.com/) If you can stomach the syntax, not to mention the metalanguage (an "adverb" is unary op? a "monad" is a unary function? really?), not to mention the mental model, you may find it useful.

In J, all data is an array. etc.

Here's an example. The following produces all the factorizations of a number:

```
ext=: [: ~. ,&.> , ;@:(tu&.>)
tu =: ] <@:(/:~)@:*"1 [ ^ </\"1@=@]
af =: ext/ @ q:
```

Holy Brainfuck!!! For comparison, here is Hello World written in that infamous language:

```
++++++++[>++++[>++>+++>+++>+<<<<-]>+>+>->>+[<]<-]>>.>---.+++++++..+++.>>.<-.<.+++.----
```

The obvious question: is the benefit of becoming fluent in such a syntax worth the cost? More to the point: does it really provide anything that you cannot find in some other less alienating language?

25
Commercial

Part VIII

Appendices

Appendices

A
Bibliography

B
Bibliography