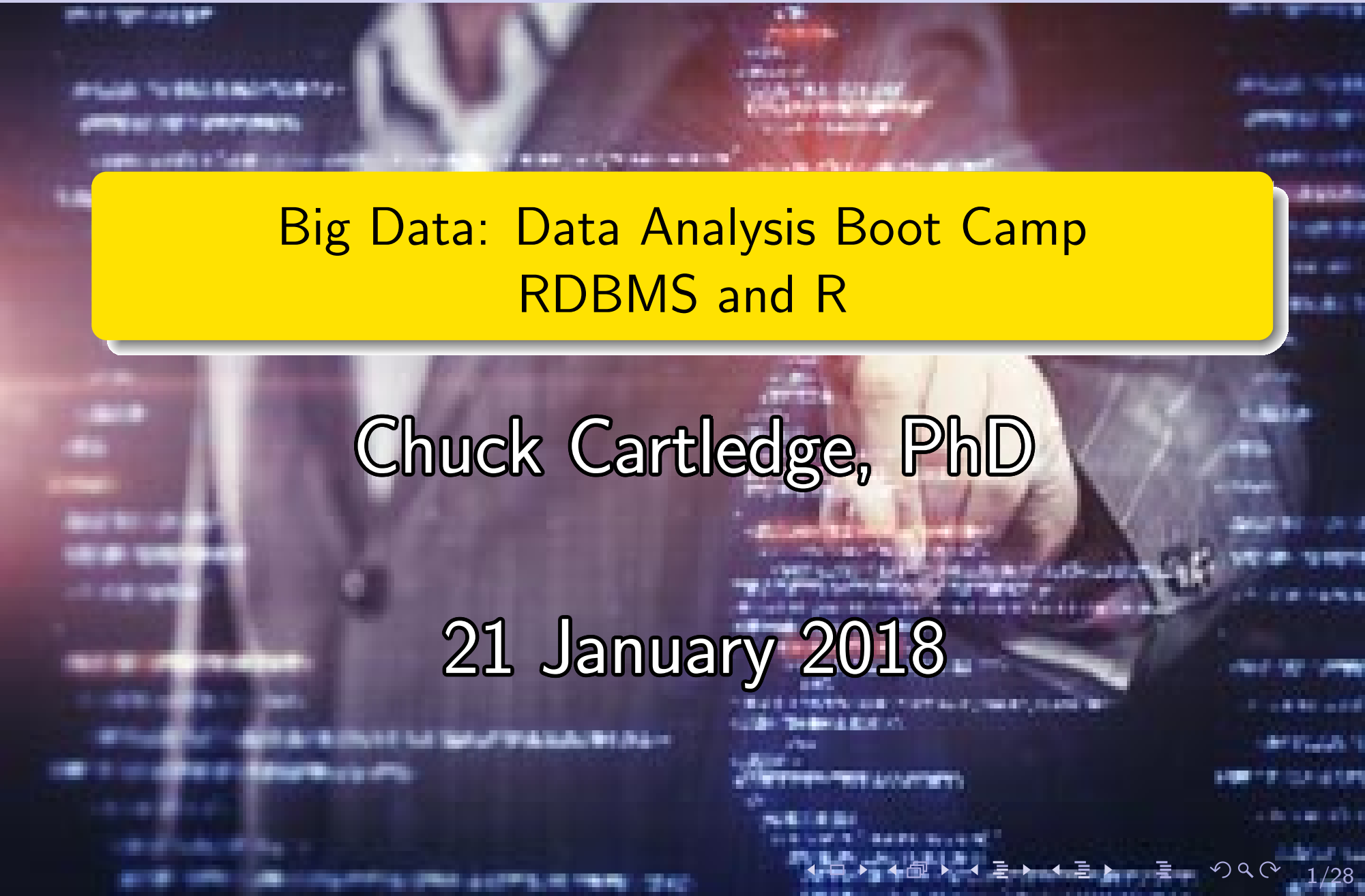


Big Data: Data Analysis Boot Camp
RDBMS and R

Chuck Cartledge, PhD

21 January 2018

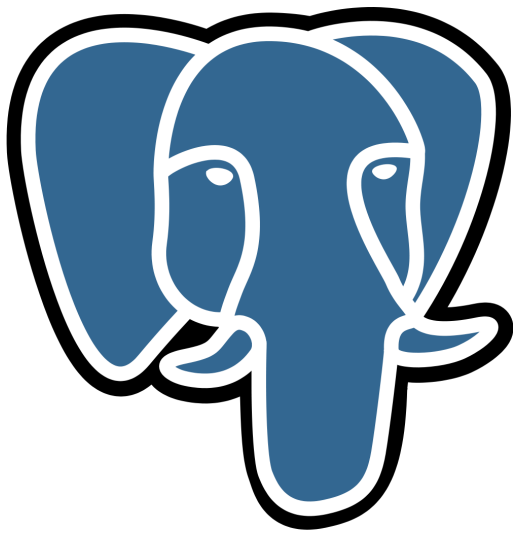
Table of contents (1 of 1)

What are we going to cover?

1. Talk about what a relational database management system (RDBMS) is and isn't.
2. Create a reasonably sized Postgres RDBMS table
3. Explore and modify an R program that accesses the Postgres database

PostgreSQL is old [3]

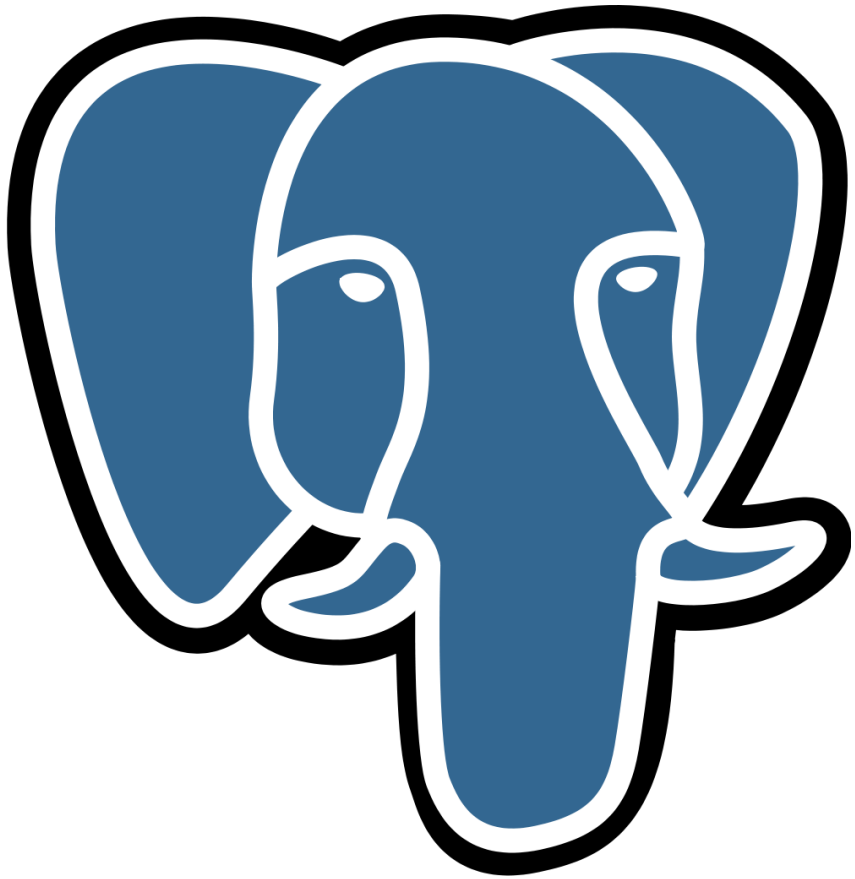
- Started at UC Berkeley as POSTGRES in 1986.
- DARPA and Army Research Office (ARO) project
- Demoed in 1987, presented in 1988, released in 1989
- Evolved into Postgres95
- Name changed in 1996 to PostgreSQL to reflect origin and new SQL capability
- Official name PostgreSQL, nicknamed Postgres



Postgres is Post(In)gres.

Bits and pieces.

- Open source software available 1 Aug. 1996 [5]
- Online presence at PostgreSQL.org since 22 Oct. 1996
- Under active development, major releases approx. yearly



An old and on going project. PostgreSQL is very widely available.

The origin.

"... The data model is a relational model that has been extended with abstract data types including user-defined operators and procedures, relation attributes of type procedure, and attribute and procedure inheritance. These mechanism can be used to simulate a wide variety of semantic and object-oriented data modeling constructs including aggregation and generalization, complex objects with shared subobjects, and attributes that reference tuples in other relations."

L. A. Rowe [2]

What does it mean?

"PostgreSQL is a relational database management system, which means its a set-theory-based system, implemented as two-dimensional tables with data rows and strictly enforced column types."

E. Redman [1]

Hypothetical Relational Database Model

PubID	Publisher	PubAddress
03-4472822	Random House	123 4th Street, New York
04-7733903	Wiley and Sons	45 Lincoln Blvd, Chicago
03-4859223	O'Reilly Press	77 Boston Ave, Cambridge
03-3920886	City Lights Books	99 Market, San Francisco

AuthorID	AuthorName	AuthorBDay
345-28-2938	Haile Selassie	14-Aug-92
392-48-9965	Joe Blow	14-Mar-15
454-22-4012	Sally Hemmings	12-Sept-70
663-59-1254	Hannah Arendt	12-Mar-06

ISBN	AuthorID	PubID	Date	Title
1-34532-482-1	345-28-2938	03-4472822	1990	Cold Fusion for Dummies
1-38482-995-1	392-48-9965	04-7733903	1985	Macrame and Straw Tying
2-35921-499-4	454-22-4012	03-4859223	1952	Fluid Dynamics of Aquaducts
1-38278-293-4	663-59-1254	03-3920886	1967	Beads, Baskets & Revolution

PostgreSQL is largely ANSI-SQL:2008 compliant

"The nice thing about standards is that you have so many to choose from."

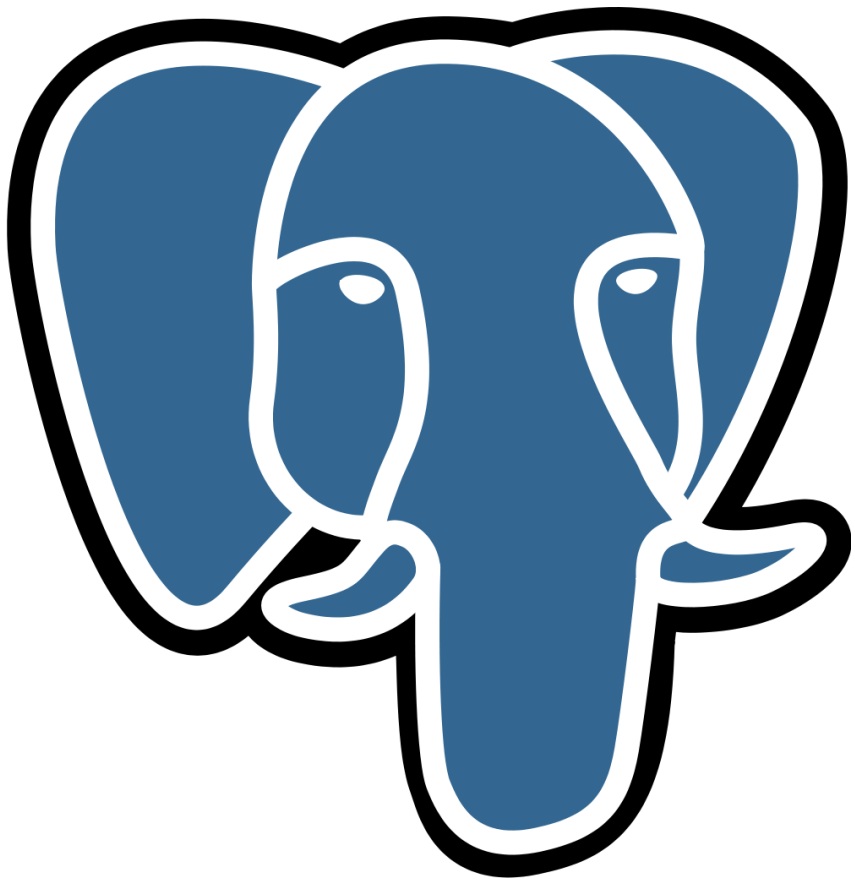
A. S. Tanenbaum [7]

ANSI-SQL:2011 adds many temporal related capabilities.

PostgreSQL is open source

"... is not only a powerful database system capable of running the enterprise, it is a development platform upon which to develop in-house, web, or commercial software products that require a capable DBMS."

PostgreSQL Staff [4]



Programmers are tool makers (among other things). When possible to extend something, they will.

What are extensions?

A way to define a collection of "loose" objects into a named entity.

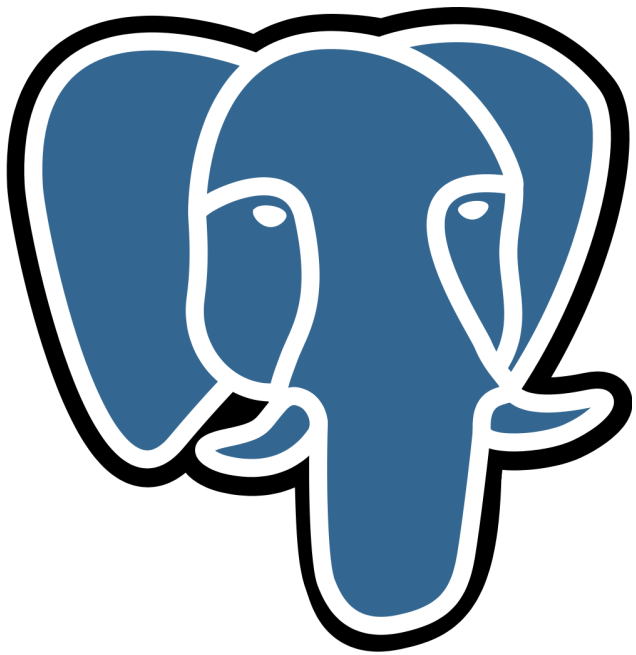
- A collection is called an "extension"
- An extension may have many internal objects
- An extension is loaded via the `CREATE EXTENSION` command
- An extension is dropped via the `DROP EXTENSION` command
- An extension object can be modified via the `CREATE FUNCTION` or `REPLACE FUNCTION` command
- `\dx` to list installed extensions
- `select * from pg_available_extensions() order by name;` is also available.

Where can I find information about an extension?

Like so many other things. It is in the documentation.[1]

Documentation is terse.

- A few sentences about the extension.
- A list of objects in the collection.
- Maybe an example.

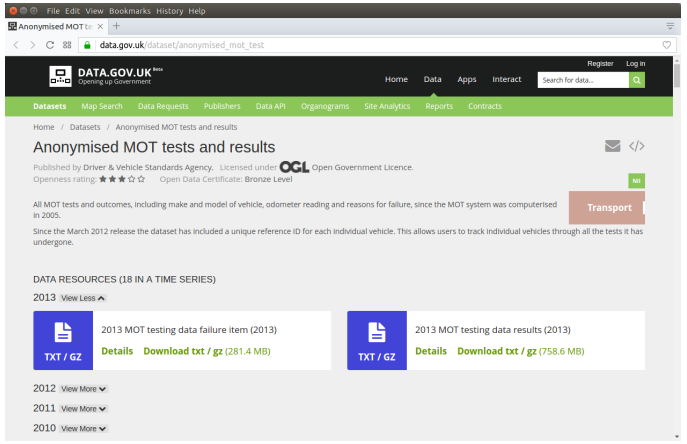


[1] http://www.postgresql.org/docs/9.3/static/contrib.html

Data source

"The MOT test (Ministry of Transport, or simply MOT) is an annual test of vehicle safety, road worthiness aspects and exhaust emissions required in Great Britain for most vehicles over three years old used on any way defined as a road in the Road Traffic Act 1988 . . ."

W. Staff [6]



`https://data.gov.uk/dataset/anonymised_mot_test`

Different classes of vehicles have different test periods.

Data we're interested in

- The text has a subset of the 2013 data. We will be using the complete 2013 data set.
- The compressed data set is 772,928,840 bytes long.
- The uncompressed data set is 3,410,125,747 bytes long.
- We'll download it once, and stick it into the database.
- Downloading the data and importing it will take a few minutes.
- Load the attached file.

Details about the data set:

The first few lines lines look like this:

```
17|28|2013-05-02|2|N|P|46414|BN|SUZUKI|UNCLASSIFIED|GREEN|P|398|1993-08-11
22|33|2013-06-07|2|N|P|15605|PE|UNCLASSIFIED|UNCLASSIFIED|BLACK|P|150|1962-01-01
44|49|2013-08-09|4|N|PRS|72694|SO|UNCLASSIFIED|UNCLASSIFIED|BLACK|P|998|2001-05-16
52|54|2013-04-19|4|PR|P|90255|PE|NISSAN|MICRA GX|GREEN|P|998|2000-03-31
53|54|2013-04-18|4|N|F|90255|PE|UNCLASSIFIED|UNCLASSIFIED|GREEN|P|998|2000-03-31
65|63|2013-02-04|4|PR|P|84821|CW|UNCLASSIFIED|UNCLASSIFIED|SILVER|P|1985|2000-05-12
66|63|2013-02-01|4|N|F|84821|CW|UNCLASSIFIED|UNCLASSIFIED|SILVER|P|1985|2000-05-12
110|93|2013-07-09|4|F|P|104188|S|BMW|318ti SE COMPACT|BLUE|P|1895|2000-09-20
111|93|2013-06-13|4|N|F|104188|S|UNCLASSIFIED|UNCLASSIFIED|BLUE|P|1895|2000-09-20
234|176|2013-01-17|4|PL|P|107447|SW|UNCLASSIFIED|UNCLASSIFIED|BLUE|P|1598|2000-01-01
```

Getting the data into Postgres

We'll create a table (with columns) and import the data like this:

```
1 statement <- sprintf("create table \"%s\" (testID
      varchar(12), vehicleID varchar(12),testDate varchar
      (12), testClassID varchar(12),testType varchar(12),
       testResult varchar(12),testMileage double
      precision , postcodeArea varchar(12),make varchar
      (120), model varchar(120),colour varchar(12),
      fuelType varchar(12),cylCapacity double precision ,
      firstUseDate varchar(12));", tableName)
2 res <- dbSendQuery(dbCon, statement)
3 statement <- sprintf("copy \"%s\" from '%s' delimiter
      '|' csv;",tableName, tempFile)
4 res <- dbSendQuery(dbCon, statement)
```

Queries in the attached file. (1 of 2)

- How many records are in the database?
  Answer is: "There are 37,390,457 records."
- What are the makes of vehicles and how did they do?
  Answer is:

	make	testresult	count
1	ABARTH	ABA	2
2	ABARTH	ABR	19
3	ABARTH	F	482
4	ABARTH	P	2413
5	ABARTH	PRS	134
6	AC	F	8

Queries in the attached file. (2 of 2)

- What was the average mileage of those vehicles that failed? Answer is:

	make	testresult	count	avg_miles
1	KASSBOHRER	F	1	847437.0
2	BOVA	F	1	746795.0
3	VAN HOOL	F	2	323021.0
4	METROCAB	F	512	283658.3
5	CARBODIES	F	686	282352.8
6	MAN / VW	F	9	277105.0

A strange way(?) to call functions.

Everything in R is an object. So once you know its name, you can "pass" it around and do different things with it.

```
1 funcs <- c(dbCountRows, dbMakeTestResults,
      dbAverageFailureMiles)
2 for (func in funcs)
3 {
4       print(system.time(func(con, tableName, verbose=TRUE
            )))
5 }
```

The loop calls three functions in succession and reports the execution time.

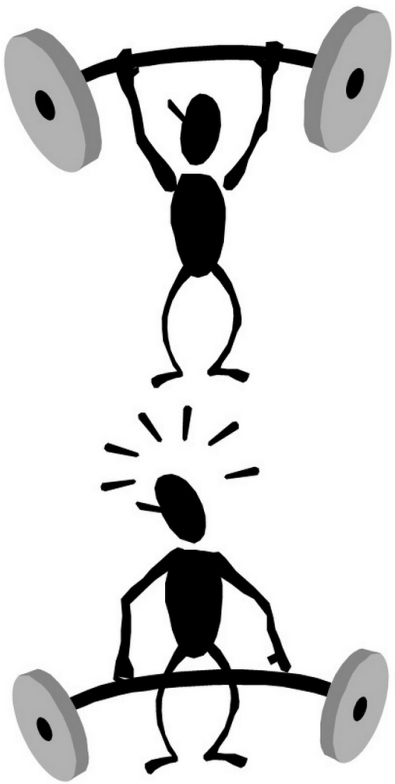
Ways to modify the MOT program.

A SELECT statement can be used to query the database, and R can be used for the numeric heavy lifting.

- What is the mileage distribution of vehicles that passed/failed?
- Is there a relationship between passing, test date, and the first use date?
- Has there been a change in the fuel type, or cylinder capacity as a function of time?
- Does the postcode affect the pass/fail rate?
- Does the color affect the pass/fail rate?
- What would a choroplethrmap of pass/fail results look like?

Good and not so good

- Strengths
  - Age, lots of years of active development
  - Lots of language specific drivers
  - Extensible
  - Open source
- Weaknesses
  - Partionability (re. CAP Theorem)
  - Data must be “neat and tidy”



Good for, and not so good for

- Good fit
  - Well structured data
  - Data known in advance
  - Data use not known in advance
- Not so good fit
  - Highly variable data
  - Hierarchical or “object oriented”
  - Extremely sparse data

Q & A time.

Q: How many hardware engineers does it take to change a light bulb?
A: None. We'll fix it in software.

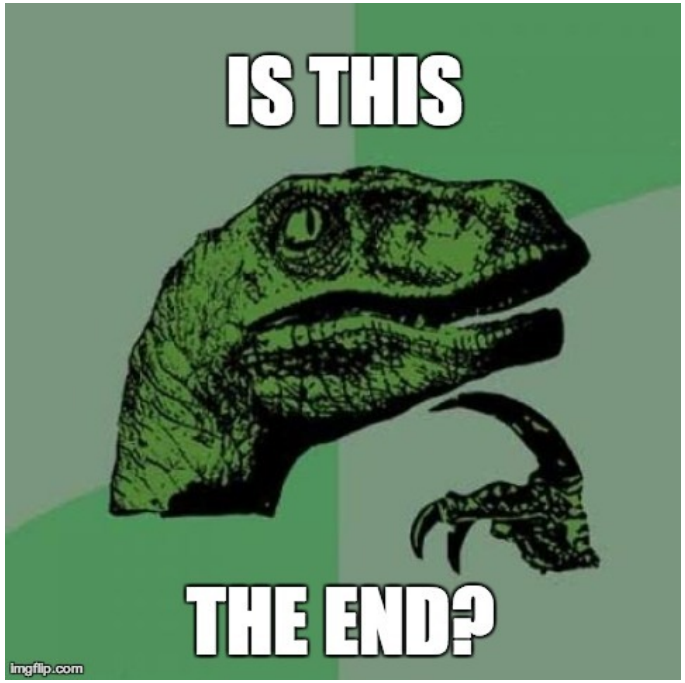
Q: How many system programmers does it take to change a light bulb?
A: None. The application can work around it.

Q: How many software engineers does it take to change a light bulb?
A: None. We'll document it in the manual.

Q: How many tech writers does it take to change a light bulb?
A: None. The user can figure it out.

What have we covered?

- Spent a little time talking about RDBMS and SQL
- Created and queried a reasonably sized Postgres table
- Modified an existing application to answer new questions



Next: BDAR Chapter 6, NoSQL and R

References (1 of 2)

[1] Eric Redmond and Jim R Wilson, Seven Databases in Seven Weeks, Pragmatic Bookshelf, 2012.

[2] Lawrence A. Rowe and Michael Stonebraker, The POSTGRES Data Model, Proceedings of the 13th International Conference on Very Large Data Bases (San Francisco, CA, USA), VLDB '87, Morgan Kaufmann Publishers Inc., 1987, pp. 83–96.

[3] PostgreSQL Staff, A Brief History of PostgreSQL, `http://www.postgresql.org/docs/9.0/static/history.html`, 2015.

[4] _____, About, `http://www.postgresql.org/about/`, 2015.

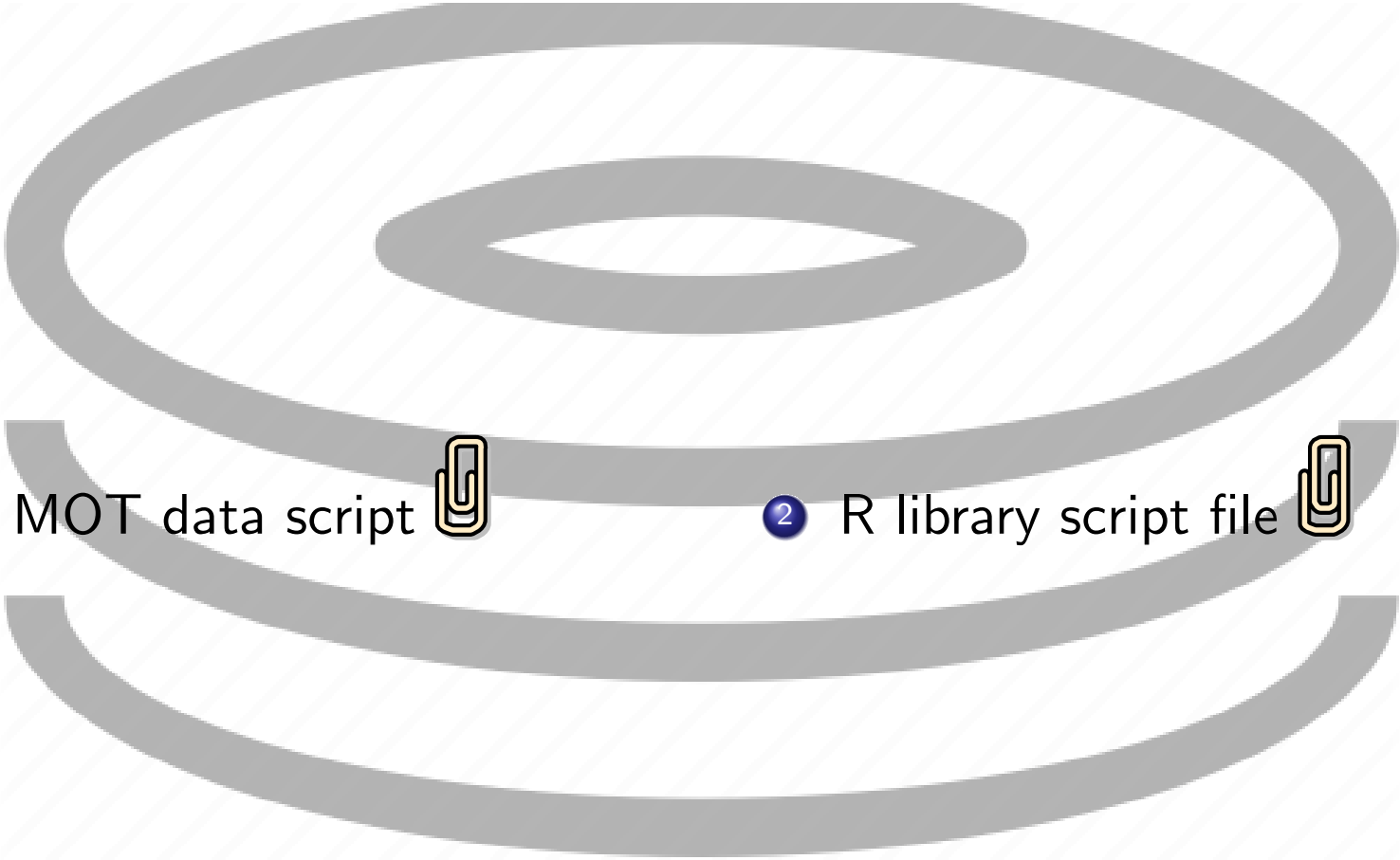
References (2 of 2)

[5] Wikipedia Staff, PosgreSQL, https://en.wikipedia.org/wiki/PostgreSQL, 2015.

[6] _____, MOT test, https://en.wikipedia.org/wiki/MOT_test, 2017.

[7] Andrew S Tanenbaum, Computer Networks, Prentice Hall, 2003.

Files of interest

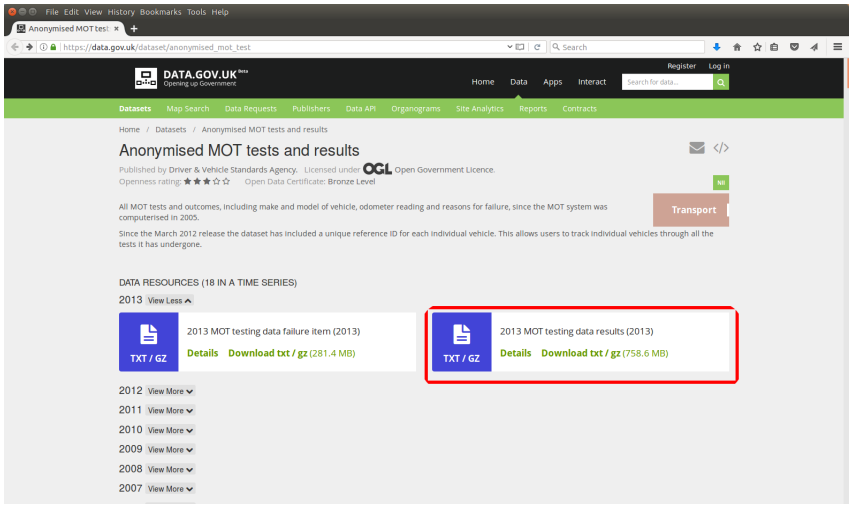


1. MOT data script
2. R library script file

UK Ministry of Transportation download page

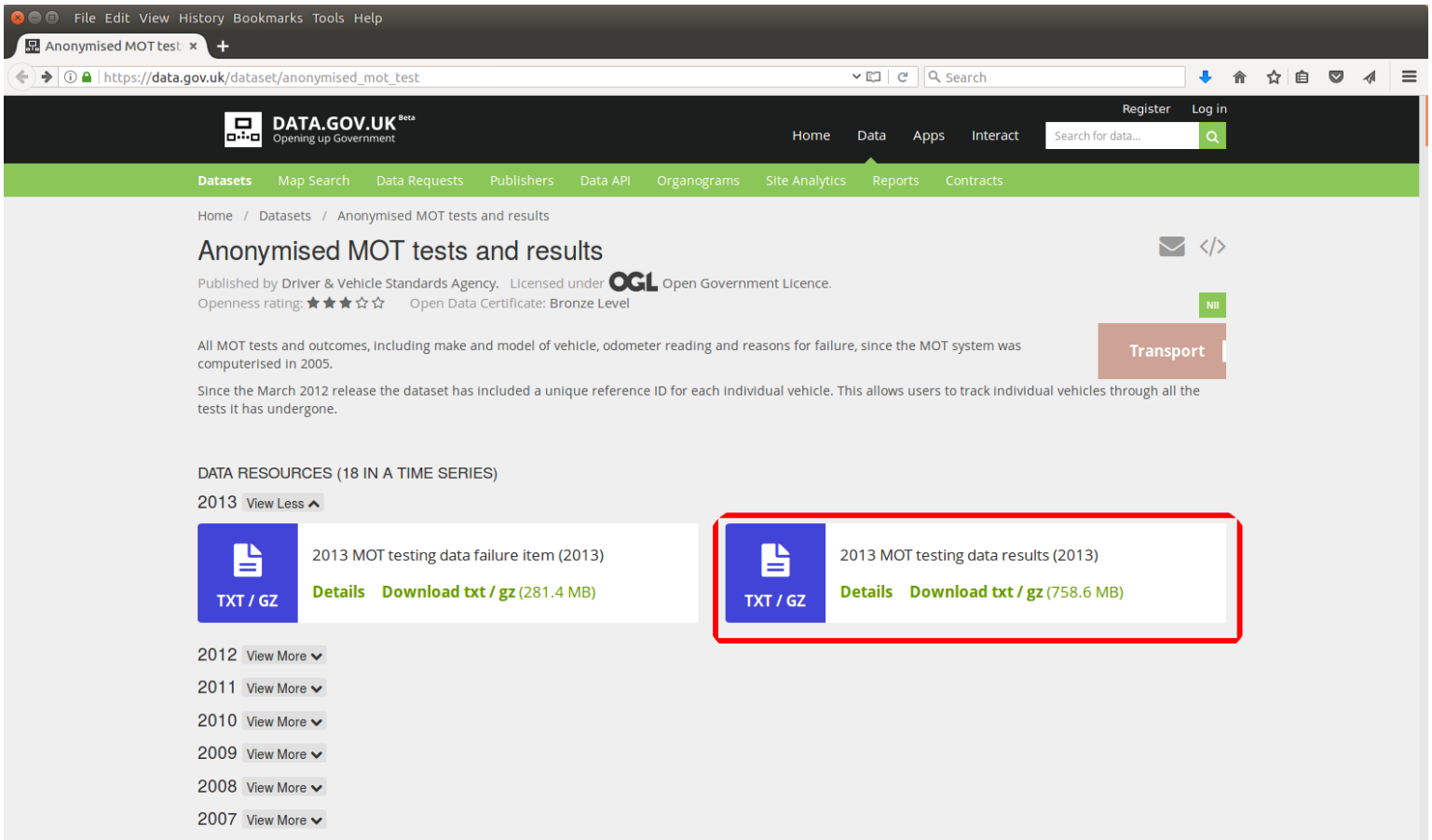
Anonymised data is downloaded from this page.
The download selection is outlined in red.
The download may take a few minutes.
File information (orig and uncompressed):

gzip	758,640,252
lines	37,390,457
words	110,992,552
bytes	3,410,125,747



`https://data.gov.uk/dataset/anonymised_mot_test`

Same image.



`https://data.gov.uk/dataset/anonymised_mot_test`