Big Data: Data Wrangling Boot Camp
Publicly Available Sources of BD

Chuck Cartledge, PhD

16 September 2016

Table of contents I

What are we going to cover?

The world is awash in Big Data, and a lot of it is freely available.
We're going to talk about:

- Different ways to get Big Data,
- Different formats that Big Data can come in,
- Costs associated with "free" Big Data, and
- Sources of Big Data.

You can create your own.

Nicholas Felton has been collecting and publishing personal data since 2005.
(You don't have to publish the data to make use of it.)

Image from [1].

You can collect from your sensors.

Aaron Parecki is the co-founder of IndieWebCamp, and maintains oauth.net. He is known for having tracked his location at 5 second intervals since 2008.

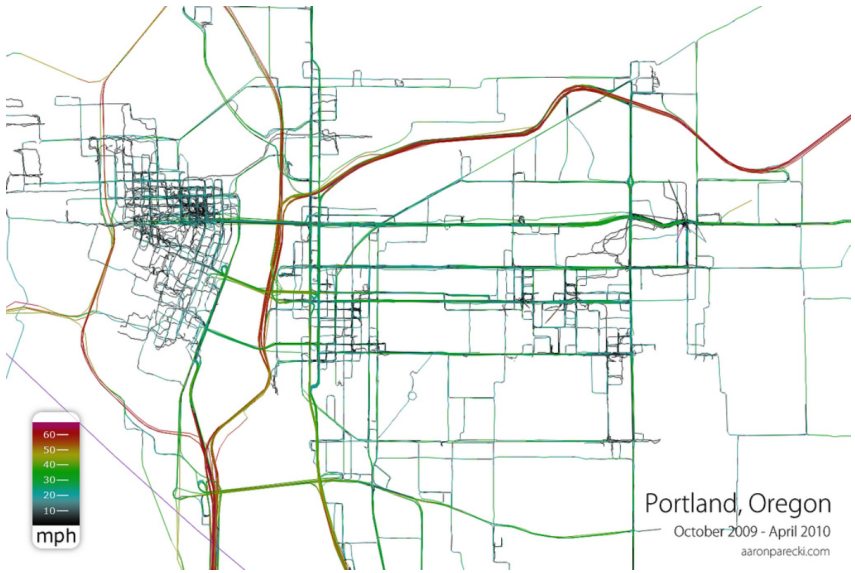


Image from [2].

You can collect it from your wearables.

If you are wearing a Fitbit (or other wearable sensor), you are creating data all the time.

Requires fitbit developer's access (`https://dev.fitbit.com/`).

You can collect from your phone

1. Motion sensors that can tell the difference between walking and driving,
2. A barometer for measuring atmospheric pressure,
3. A gesture sensor that detects hand movements through infrared rays,
4. Gyroscope to measure acceleration,
5. Magnetometer to measure magnetic lines of flux,
6. GPS to tell where you are around the world,
7. WIFI to connect to the world, and also to tell how close you are to a broadcast station or router,
8. Camera(s) to see,
9. Microphone(s) to listen,
10. Speaker(s) to speak,
11. Temperature and pressure (on the screen).

Lots of sensors.

You can download a file from somewhere.

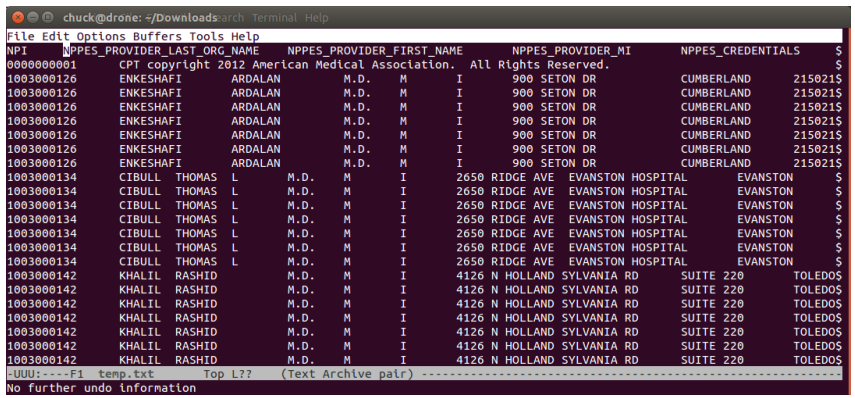
Centers for Medicare and Medicaid Service, part of the Department of Health and Human Services (HHS).

Makes available a vast array of data relating to all their programs.

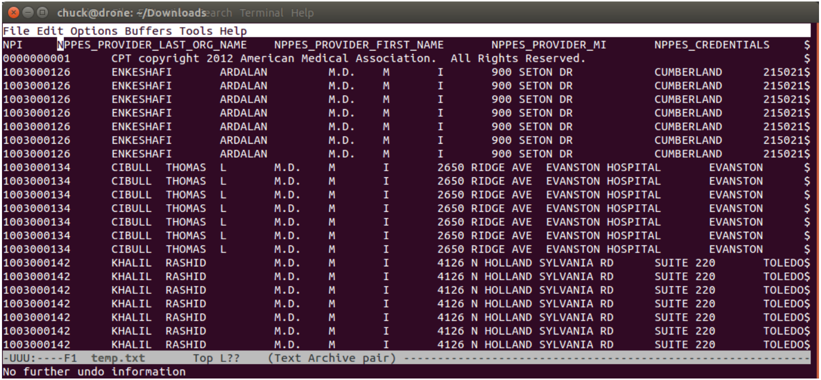
Including Medicare payments per calendar year.

CMS Medicare Physician and Other Supplier Public Use File (PUF), CY2013

1. The ZIP single file contains three other files
2. It is 497,014,400 bytes of compressed data.
3. The data file has a header record.
4. The data fields are delimited by tab characters that are normally invisible
5. The a single tab is between each data field. The editor can make the tabs visible

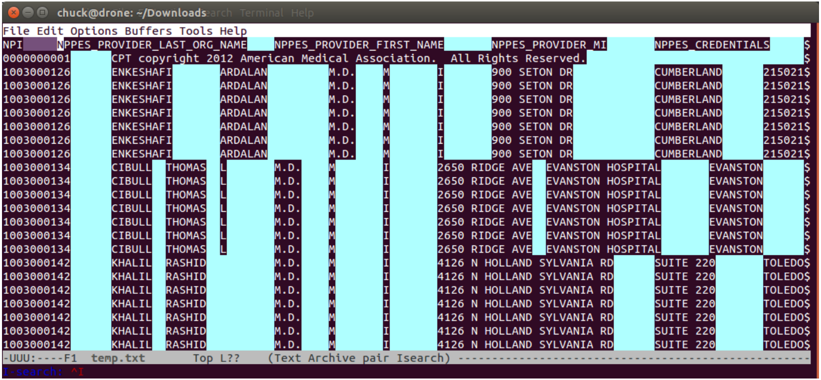


Same image.



```
chuck@drone: ~/Downloads   Search  Terminal  Help
File Edit Options Buffers Tools Help
NPI       NPPES_PROVIDER_LAST_ORG_NAME     NPPES_PROVIDER_FIRST_NAME       NPPES_PROVIDER_MI       NPPES_CREDENTIALS      $
0000000001      CPT copyright 2012 American Medical Association.  All Rights Reserved.                                 $
1003000126      ENKESHAFI       ARDALAN         M.D.    M       I       900 SETON DR            CUMBERLAND      215021$
1003000126      ENKESHAFI       ARDALAN         M.D.    M       I       900 SETON DR            CUMBERLAND      215021$
1003000126      ENKESHAFI       ARDALAN         M.D.    M       I       900 SETON DR            CUMBERLAND      215021$
1003000126      ENKESHAFI       ARDALAN         M.D.    M       I       900 SETON DR            CUMBERLAND      215021$
1003000126      ENKESHAFI       ARDALAN         M.D.    M       I       900 SETON DR            CUMBERLAND      215021$
1003000126      ENKESHAFI       ARDALAN         M.D.    M       I       900 SETON DR            CUMBERLAND      215021$
1003000126      ENKESHAFI       ARDALAN         M.D.    M       I       900 SETON DR            CUMBERLAND      215021$
1003000134      CIBULL  THOMAS  L       M.D.    M       I       2650 RIDGE AVE  EVANSTON HOSPITAL       EVANSTON      $
1003000134      CIBULL  THOMAS  L       M.D.    M       I       2650 RIDGE AVE  EVANSTON HOSPITAL       EVANSTON      $
1003000134      CIBULL  THOMAS  L       M.D.    M       I       2650 RIDGE AVE  EVANSTON HOSPITAL       EVANSTON      $
1003000134      CIBULL  THOMAS  L       M.D.    M       I       2650 RIDGE AVE  EVANSTON HOSPITAL       EVANSTON      $
1003000134      CIBULL  THOMAS  L       M.D.    M       I       2650 RIDGE AVE  EVANSTON HOSPITAL       EVANSTON      $
1003000134      CIBULL  THOMAS  L       M.D.    M       I       2650 RIDGE AVE  EVANSTON HOSPITAL       EVANSTON      $
1003000134      CIBULL  THOMAS  L       M.D.    M       I       2650 RIDGE AVE  EVANSTON HOSPITAL       EVANSTON      $
1003000142      KHALIL  RASHID          M.D.    M       I       4126 N HOLLAND SYLVANIA RD      SUITE 220       TOLEDO$
1003000142      KHALIL  RASHID          M.D.    M       I       4126 N HOLLAND SYLVANIA RD      SUITE 220       TOLEDO$
1003000142      KHALIL  RASHID          M.D.    M       I       4126 N HOLLAND SYLVANIA RD      SUITE 220       TOLEDO$
1003000142      KHALIL  RASHID          M.D.    M       I       4126 N HOLLAND SYLVANIA RD      SUITE 220       TOLEDO$
1003000142      KHALIL  RASHID          M.D.    M       I       4126 N HOLLAND SYLVANIA RD      SUITE 220       TOLEDO$
1003000142      KHALIL  RASHID          M.D.    M       I       4126 N HOLLAND SYLVANIA RD      SUITE 220       TOLEDO$
1003000142      KHALIL  RASHID          M.D.    M       I       4126 N HOLLAND SYLVANIA RD      SUITE 220       TOLEDO$
-UUU:----F1  temp.txt        Top L??    (Text Archive pair) -----------------------------------------------------------
No further undo information
```

Visible tabs



Contents of the 2013 PUF.

Length (bytes)	File name
24,011	CMS_AMA_CPT_license_agreement.pdf
3,650	Medicare-Physician-and-Other-Supplier-PUF-SAS-Infile.sas
2,209,344,403	Medicare_Provider_Util_Payment_PUF_CY2013.txt

The payment file is over 2.2Gigabytes in size and has 9,287,878 lines of data.

Project Gutenberg has downloadable books

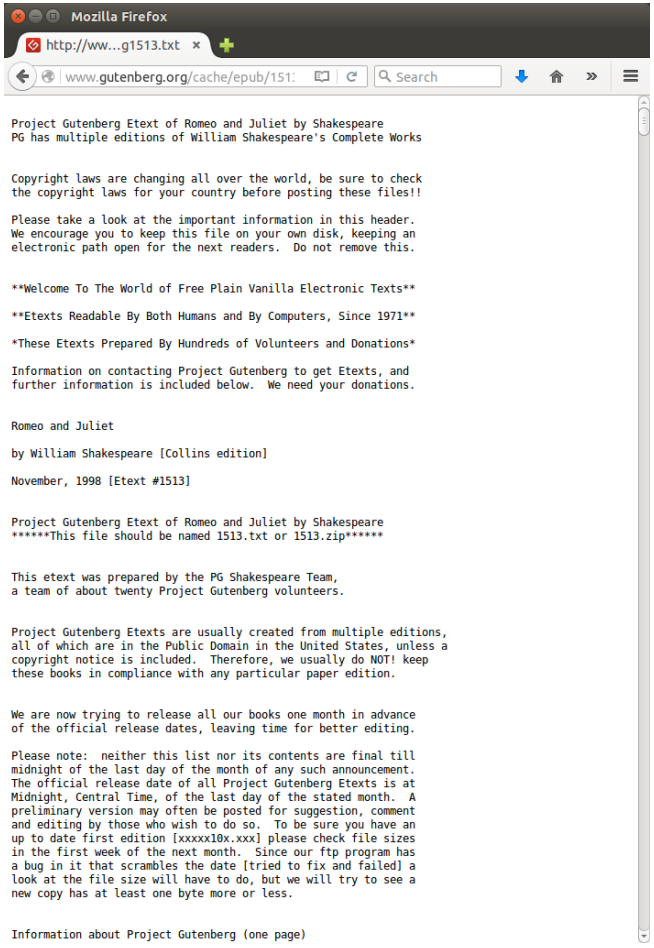
Most of the items in its collection are the full texts of public domain books. The project tries to make these as free as possible, in long-lasting, open formats that can be used on almost any computer.

Image from [3].

Project Gutenburg's version of Romeo and Juliet

Some particulars about the PG version of Romeo and Juliet:

1. It has 5,557 lines.
2. It has 27,424 words.
3. It has 153,666 characters.
4. It has a PG specific header that is 289 lines long.



You can use an Application Program Interface (API)

- The "thing" that wants something is called the client.
- The "thing" that does the work is called the server.
- The client has to talk to the server in the right way.
- The server (usually) will return something to the client.
- A browser is a client, a web site is a server.
- A person is a client, an ATM is a server.

We will use an API to get tweets from Twitter.

Data is available everywhere.

1. Aggregator
2. Aviation
3. Developers
4. Education
5. General
6. Geographic information
7. Government
8. Social
9. Weather
10. Zip

Looking for Big Data (BD) in the "Wild"

Tidewater Big Data Enthusiasts
Chuck Cartledge
Developer

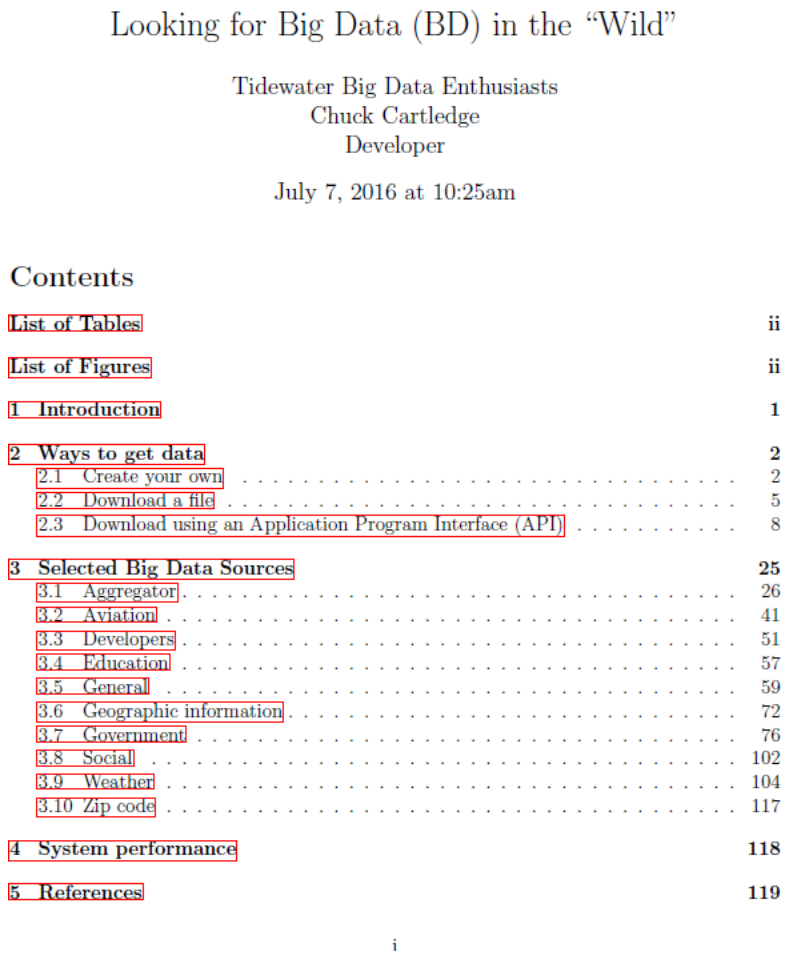
July 7, 2016 at 10:25am

Contents

i

The report is attached.

Same image.



Looking for Big Data (BD) in the "Wild"

Tidewater Big Data Enthusiasts
Chuck Cartledge
Developer

July 7, 2016 at 10:25am

Contents

i

The report is attached.

Even if it doesn't cost money, you still pay.

Things to think about when looking at data:

1. Not all data is created equally (source of data)
2. Fact checking (reliability)
3. Readability, cleanliness, and longevity (maintainability)
4. Where and how to store your data (local, cloud, SQL, NoSQL)

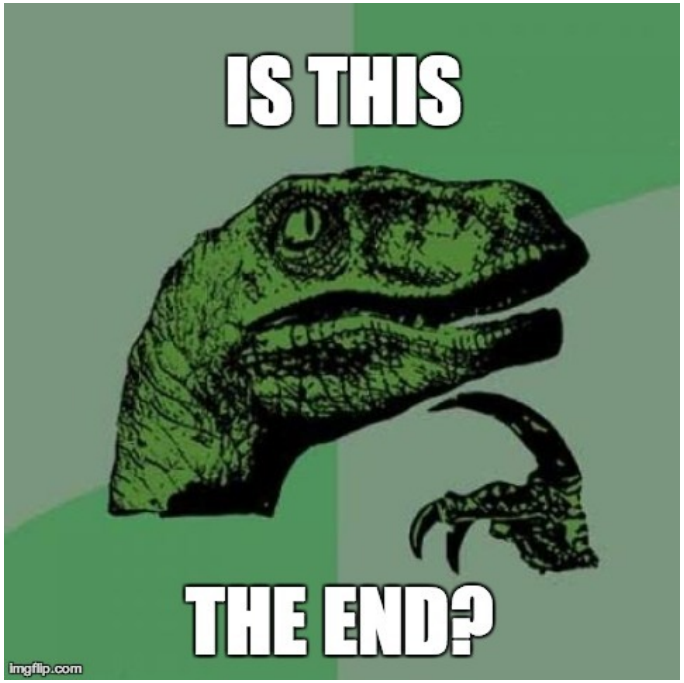
Each of these items costs time, and time is money.
All the BD Vs come into play.

Q & A time.

"'The Answer to the Great Question . . . Of Life, the Universe and Everything . . . is . . . forty-two,' said Deep Thought, with infinite majesty and calm."
Douglas Adams, The Hitchhiker's Guide to the Galaxy

What have we covered?

- Some of the many ways we can get Big Data
- Some of the many places we can get Big Data
- Some of the hidden costs associated with free Big Data



Next: Overview of Big Data tools and techniques.

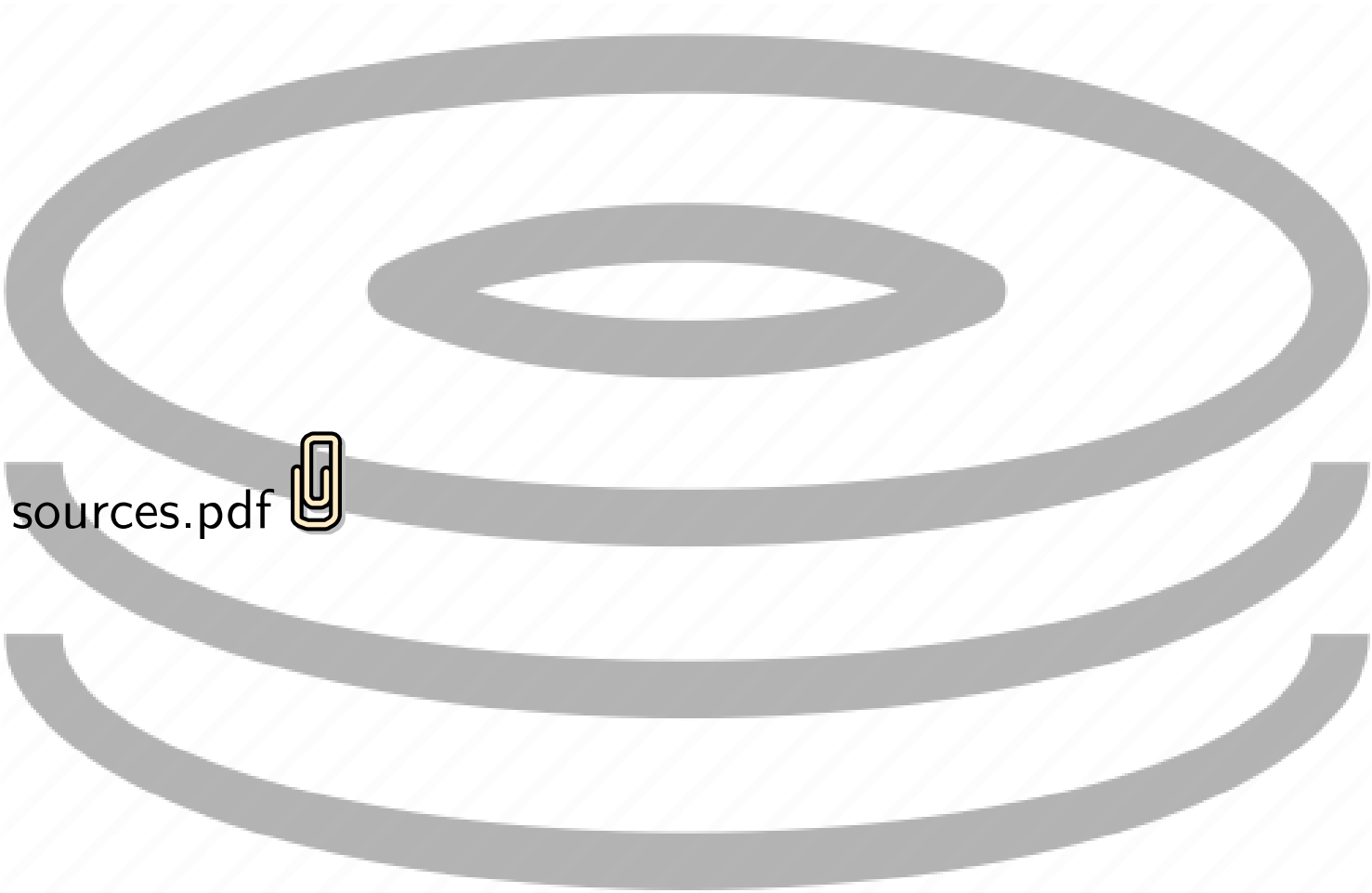
References I

[1] Nicholas Felton, Nicholas feltron personal site, http://feltron.com/, 2014.

[2] Aaron Parecki, Aaron parecki personal site, http://aaronparecki.com/, 2015.

[3] Gutenberg Staff, Free ebooks by project gutenberg, https://www.gutenberg.org/, 2016.

Files of interest



1. sources.pdf