Big Data

How to work with large datasets.

Jonathan Callahan
Mazama Science

MAZAMA SCIENCE

Data – Information – Knowledge

Data and metadata are different.

For this talk I will use **'data'** to refer to measurements with:

- numeric representation
- units

We also have **'metadata'** – information associated with measurements:

- numeric with units:
  - latitude, longitude, depth, time
- character strings:
  - instrument ID, city name, contaminent name

Be clear about what is your 'data'.

What is “Big Data”?

A dataset is “Big” when it is challenging to work with.

Different fields have different challenges

Challenges are determined by:

- data structure
- data format
- available computer hardware (memory)
- available software tools
- **employee skill set**

A lot of “big data” becomes small if you have the right skills and tools.

Computer memory is important.

Reading and writing from disk is slow.

Working with data in Random-Access Memory is fast.

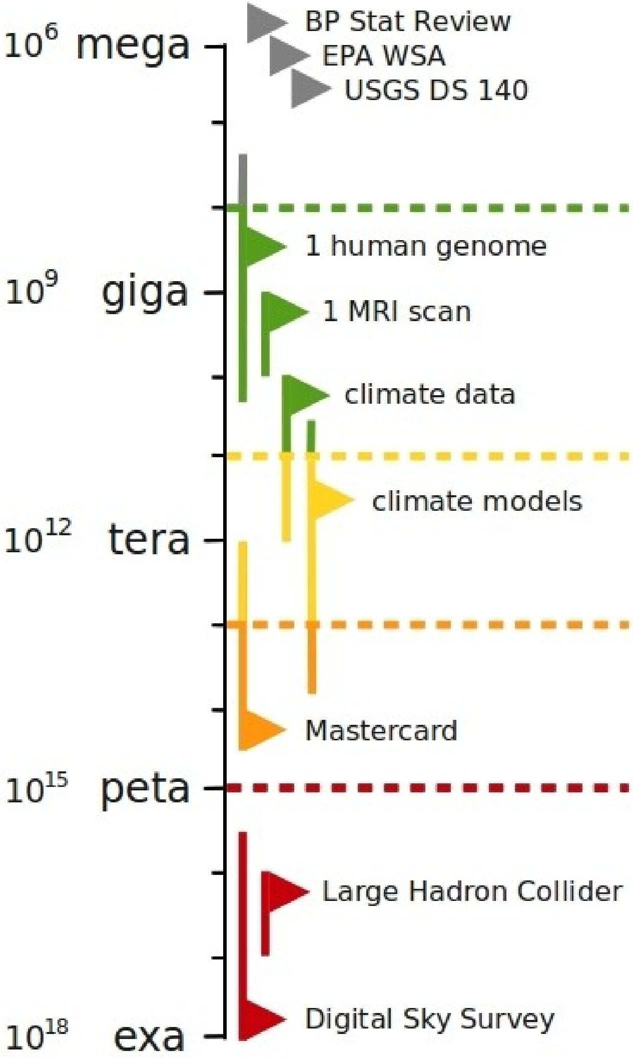
Modern laptops are pretty amazing machines:

– Apple MacBook Pro – **8 GB** of RAM

– Lenovo Thinkpad W540 – **8 GB** of RAM

If your dataset fits into 10% of RAM on your laptop, it's not big.

How big is BIG?



<1 Gbyte:	no special actions
>1 Gbyte:	special formats and software
>1 Tbyte:	special hardware
>1 Pbyte:	special teams

Some datasets may have special needs.

Environmental Sensor Data

Typical features:

- Numeric measurements
- Sampling at regular intervals
- Every sample has a timestamp
- Multiple sensors
- Metadata for each sensor
- More data every day

Sensor data has the potential to get big.

Do your *in-memory* math!

Data math for numbers

$$\textit{measurement} = 1000\,\textit{sta.} \times \frac{1\,\textit{float}}{\textit{sta.} \cdot \textit{hour}} \times \frac{4\,\textit{bytes}}{\textit{float}} \times \frac{24\,\textit{hours}}{\textit{day}} \times \frac{365\,\textit{days}}{\textit{year}} = 35\,\textit{Megabytes}$$

Metadata math for character strings

$$\text{"}\textit{Station Identifier}\text{"} = 1000\,\textit{sta.} \times \frac{18\,\textit{char}}{\textit{sta.}} \times \frac{1\,\textit{byte}}{\textit{char}} = 18\,\textit{Kilobytes}$$

$$\left(\times \frac{24\,\textit{hours}}{\textit{day}} \times \frac{365\,\textit{days}}{\textit{year}} = 157\,\textit{Megabytes}\,!!! \right)$$

Large datasets need to separate data and metadata.

Do your *on-disk* math!

Binary

- ~ kilobyte header
- 4 bytes per number
- □ □ □ □

CSV

- < kilobyte header
- ~ 8 bytes per number
- `“125.034,”`

XML

- kilobytes of structure
- ~ 14 byes per number
- `“<v>125.034</v>”`

Really bad XML

- many kilobytes of structure
- ~ 32 bytes per number
- `“<PM2.5Value>...</PM2.5Value>”`

Large datasets require compact formats.

EPA > AirData > Hourly PM2.5

Data Location:

- http://aqsdr1.epa.gov/aqsweb/aqstmp/airdata/download_files.html

Data File:

- daily_88101_2013.zip 3.6 Megabytes

Expanded:

- hourly_88101_2013.csv 665.3 Megabytes

Quick peek at header and first few lines:

```
"State Code","County Code","Site Num","Parameter Code","POC","Latitude","Longitude","Datum ...
"01","073","0023","88101",3,33.553056,-86.815,"WGS84","PM2.5 - Local Conditions","2013-02- ...
"01","073","0023","88101",3,33.553056,-86.815,"WGS84","PM2.5 - Local Conditions","2013-02- ...
"01","073","0023","88101",3,33.553056,-86.815,"WGS84","PM2.5 - Local Conditions","2013-02- ...
"01","073","0023","88101",3,33.553056,-86.815,"WGS84","PM2.5 - Local Conditions","2013-02- ...
```

Lots of repeated metadata.

Examining EPA AirData

Rearranging for readability:

header	first record
"State Code"	"01"
"County Code"	"073"
"Site Num"	"0023"
"Parameter Code"	"88101"
"POC"	3
"Latitude"	33.553056
"Longitude"	-86.815
"Datum"	"WGS84"
"Parameter Name"	"PM2.5 - Local Conditions"
"Date Local"	"2013-02-18"
"Time Local"	"13:00"
"Date GMT"	"2013-02-18"
"Time GMT"	"19:00"
"Sample Measurement"	**7.4**
"Units of Measure"	"Micrograms/cubic meter (LC)"
"MDL"	2
"Uncertainty"	""
"Qualifier"	""
"Method Type"	"FEM"
"Method Name"	"Thermo Scientific 5014i or FH62C14-DHS w/VSCC - Beta Attenuation"
"State Name"	"Alabama"
"County Name"	"Jefferson"
"Date of Last Change"	"2013-06-17"

This is the measurement!

We need to separate data from metadata.

Reshaping EPA AirData

What is the native structure of the data?

- hourly sampling X 335 instruments

$$335\ sta. \times \frac{4\ bytes}{sta. \cdot hour} \times \frac{24\ hours}{day} \times \frac{365\ days}{year} = 11.7\ Megabytes$$

What is the native structure of the metadata?

- 335 instruments X 22 pieces of information

$$335\ sta. \times 22\ parameters \times \sim 20\ bytes\ each = 147\ Kilobytes$$

We can make this data MUCH smaller.

Open Source R

http://www.r-project.org

R is:

- Free
- Open source
- Widely used
- Extremely powerful
- Hard to learn

Hard to learn … But worth it!

75 lines of R code

A script with 75 lines of code can convert the EPA data:

- 25 comment lines
- 25 blank lines for readability
- 25 actual lines of code

Input:

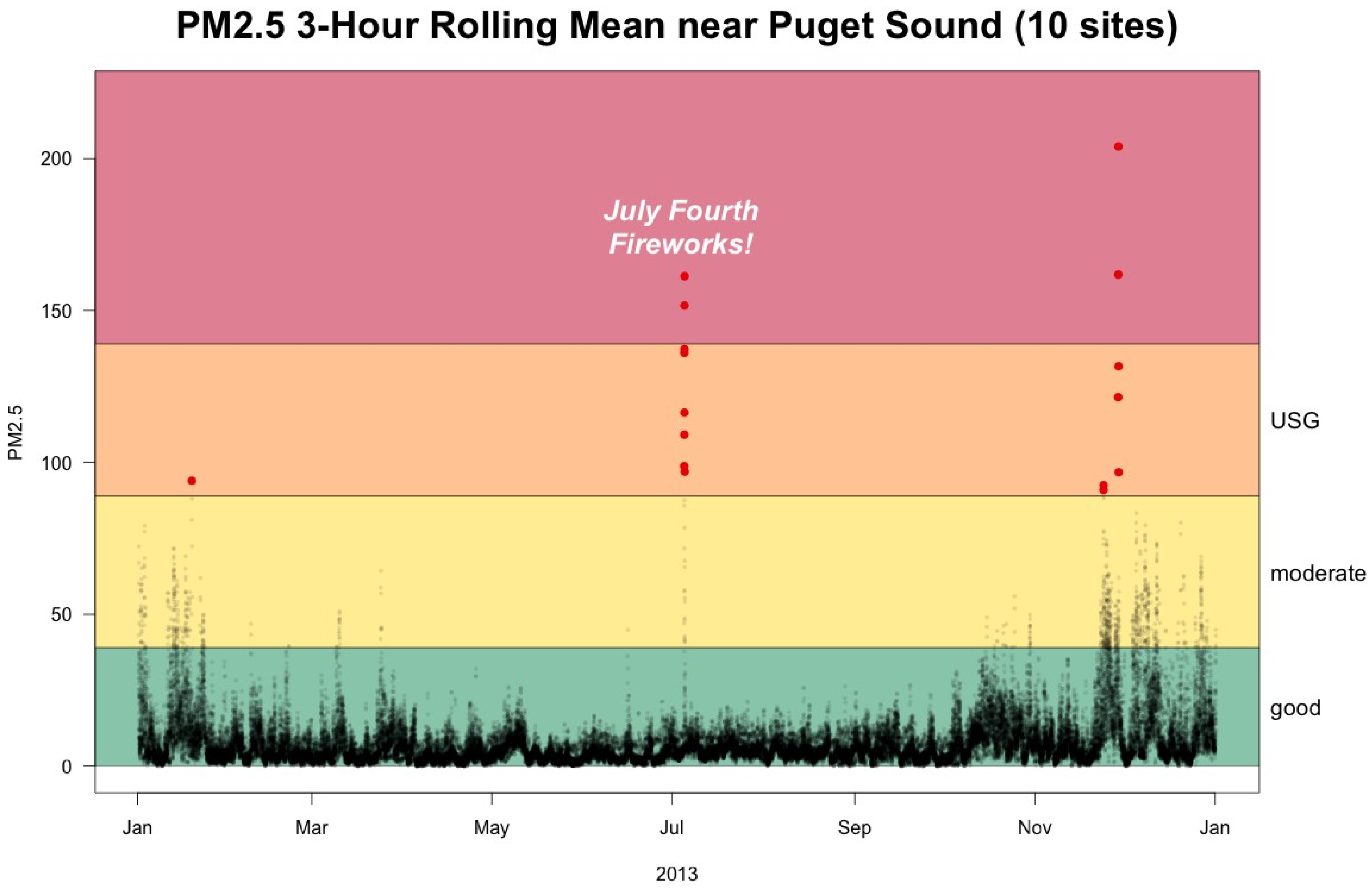
- ASCII CSV – 2,516,035 rows X 23 columns = **665 Megabytes**

Output (after 5 minutes):

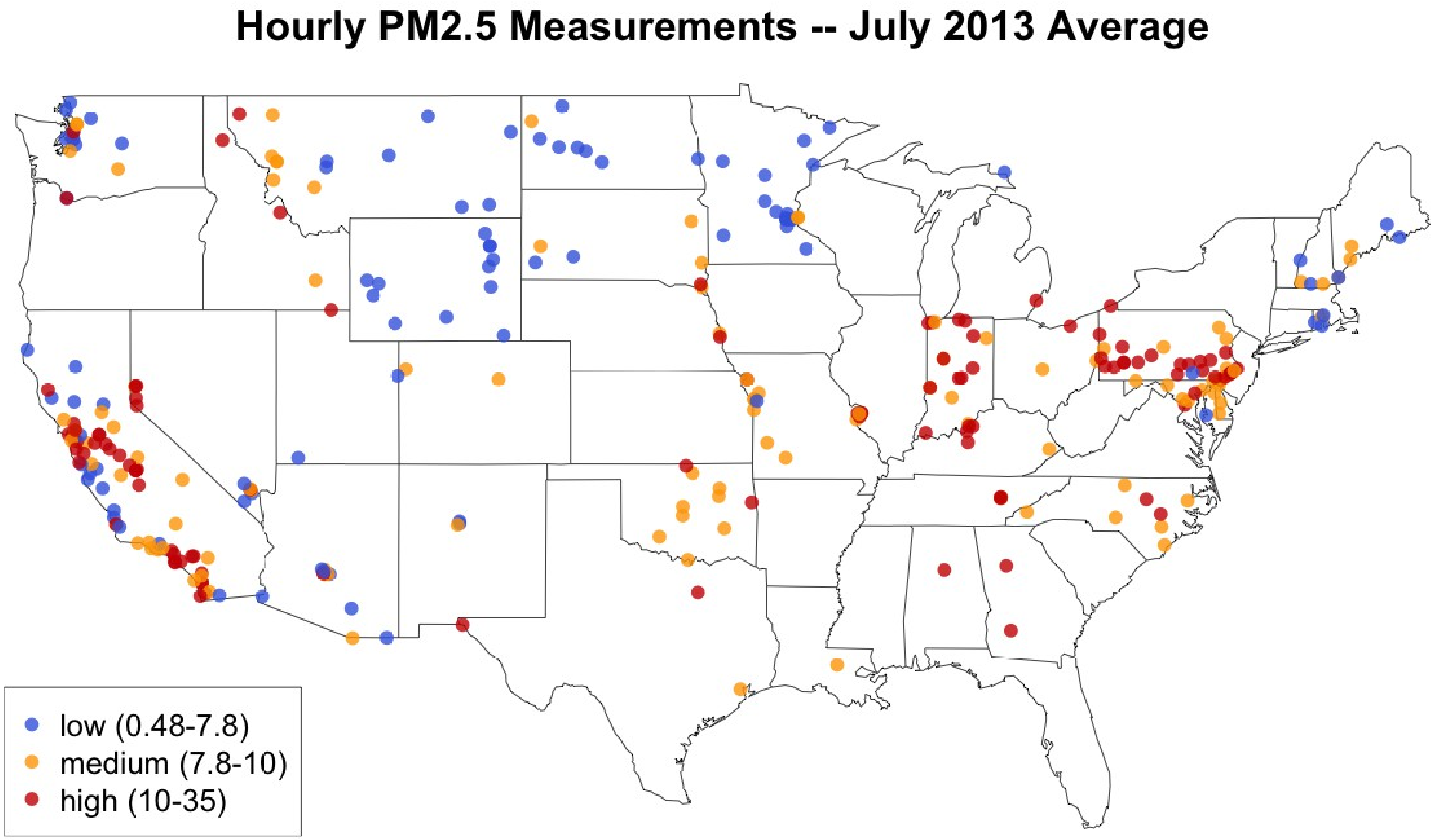
- Data .RData file – 8765 hours X 334 sites = **3.3 Megabytes**
- Metadata .RData file – 334 sites X 12 unique params = **12 Kilobytes**

With the right skills and tools, this becomes a very small dataset.

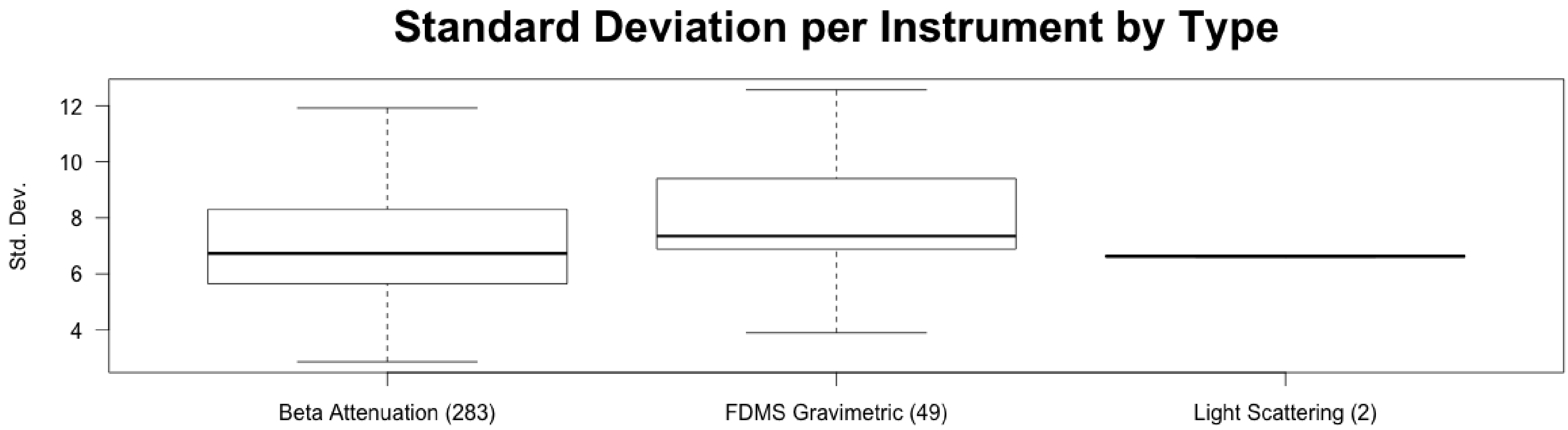
Now we can work with the entire dataset!



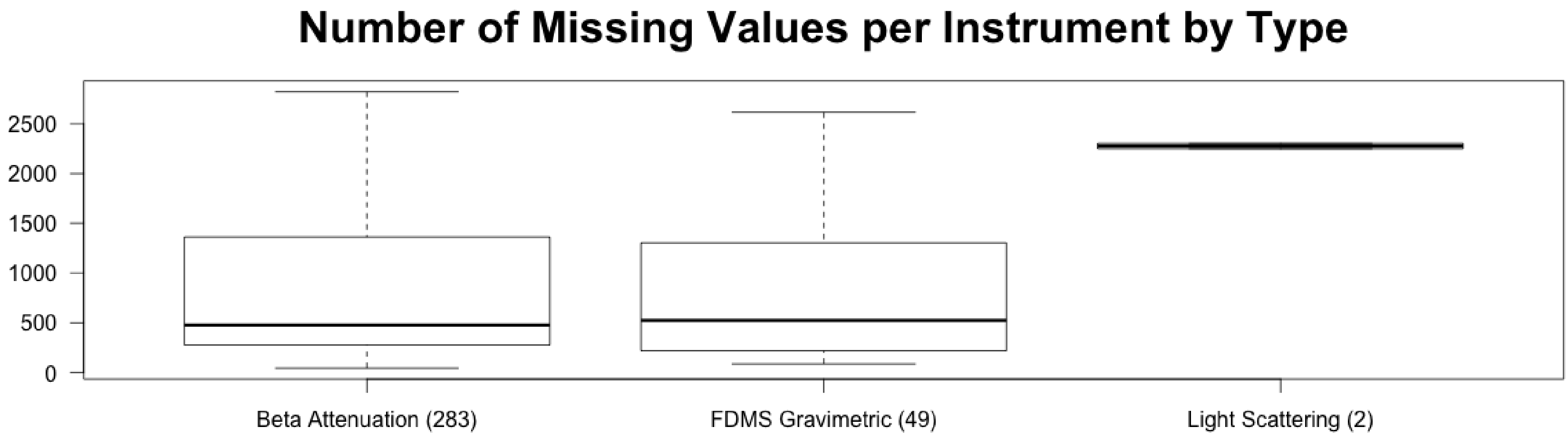
We can thoroughly interrogate the data.



We can search both data and metadata.



Standard Deviation per Instrument by Type



Number of Missing Values per Instrument by Type

Not so Big Data

How to ~~work with~~ make large datasets small

Jonathan Callahan
Mazama Science

MAZAMA SCIENCE

Data – Information – Knowledge

Take Home Message

To work productively with large datasets you should:

- Have a computer with 8+ GB of RAM
- Have good data analysis software (not a spreadsheet)
- Learn how to use your software
- Understand the structure of the *data* (not the format)
- Keep data and metadata in separate tables
- Store data in a compact format

Like most things, it's easy once you know how.

Don't be afraid of 'big' data!

MAZAMA SCIENCE

Data – Information – Knowledge