

Introduction to Map Reduce

Continuum Analytics



MapReduce

- A story about **Big Data** ?

MapReduce

- A story about **Big Data** ?
- I Prefer Large Data
 - No Hype Machine
- *Data of An Unusual Size*
 - C. Titus Brown

Large Relative to What?

Large Relative to What?

- Can't fit into Excel

Large Relative to What?

- Can't fit into Excel
 - Increase Memory

Large Relative to What?

- Can't fit into Excel
 - Increase Memory
- Can't fit into R

Large Relative to What?

- Can't fit into Excel
 - Increase Memory
- Can't fit into R
 - Increase Memory

Large Relative to What?

- Can't fit into Excel
 - Increase Memory
- Can't fit into R
 - Increase Memory
- Can't fit into Memory
 - Increase Memory

Large Relative to What?

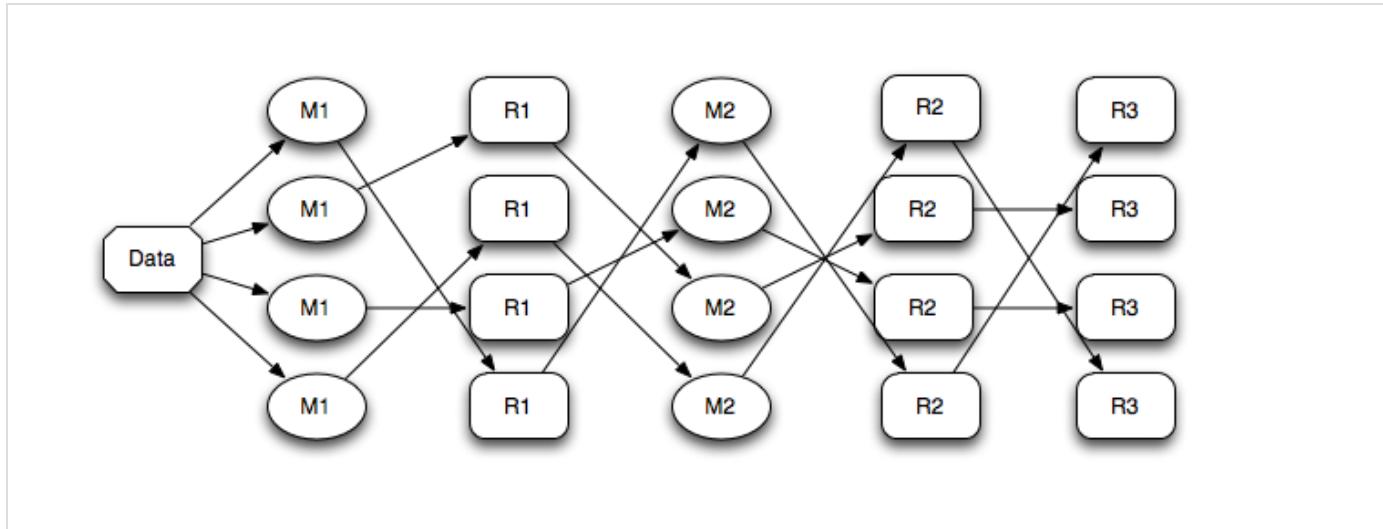
- Can't fit into Excel
 - Increase Memory
- Can't fit into R
 - Increase Memory
- Can't fit into Memory
 - Increase Memory
- Can't fit on a single disk
 - Distributed Filesystem: SAN, HDFS/DDFS, AWS: S3, Redshift, etc.

MapReduce

Framework to help solve the problem of distributed computation for distributed data

- A mass of data: records
- Split/**Map** records into key-values pairs
- Collect/Partition kv pairs (Optional Sort)
- Buckets are passed to **Reduce** function
- Result is returned

MapReduce Workflow



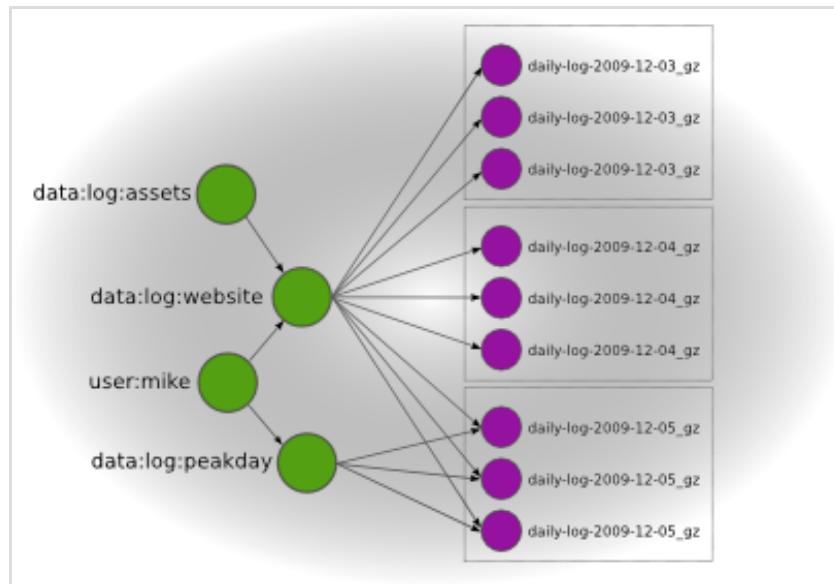
- Push Code to Data
- Lots of Network Traffic

MR Implementations

- Disco: Python + Erlang
 - Distributed FileSystem: DDFS
- Hadoop: Java
 - Streaming with Python
 - Dumbo
 - MRJob
 - Hadoopy

DDFS

- Tag based system



- ddfs push tag:mytag /the/path/data
- Replication spreads data across distributed filesystem
- Comes with chunker

MapReduce: It's a Party



Batteries Included

- NumPy
- SciPy
- pandas
- scikits-learn
- OpenCV
- ...

Canonical Example

```
1  from disco.job import Job
2  from disco.core import result_iterator
3
4  class WordCount(Job):
5
6      partitions = 3
7      input=["sherlock.txt","poirot.txt","clouseau.txt"]
8
9      @staticmethod
10     def map(line, params):
11         import string
12         for word in line.split():
13             yield word, 1
14
15     @staticmethod
16     def reduce(iter, params):
17         from disco.util import kvgroup
18         for word, counts in kvgroup(sorted(iter)):
19             yield word, sum(counts)
20
21 if __name__ == "__main__":
22     from disco_words import WordCount
23
24     wordcount = WordCount().run()
25
26     for (word, counts) in result_iterator(wordcount.wait(show=True)):
27         print word, counts
28
```

Bitly Data

- URL shortening service
 - URLs with .gov/.mil are made public
- Typical Clickstream Data
 - JSON encoded
 - url
 - GeoLoc
 - Time
 - OS/Browser
 - ...
- Thank you Wes McKinney

Demo 1

Questions

- What are the top 10 websites are for the day, month, quarter, etc.
- Most popular government institutions
 - NIH vs NASA vs CDC?
- Do most people look at links on mobile devices?
 - what is the break down of Windows vs OSX?
- Click through rate from Facebook vs twitter?

Regional Questions

- NIH only popular in DC, NY, LA ?
- what links are popular outside of the US

WikiLog Data

- Hourly summary statistics
- 4 fields: projectcode, pagename, pageviews, byte size

```
ca Casino_de_Manresa 1 8334
ca Caspar_David_Friedri 2 20242
ca Casquet_Glacial_Pata 1 8640
ca Castell_d%27Eramprun 6 11885
```

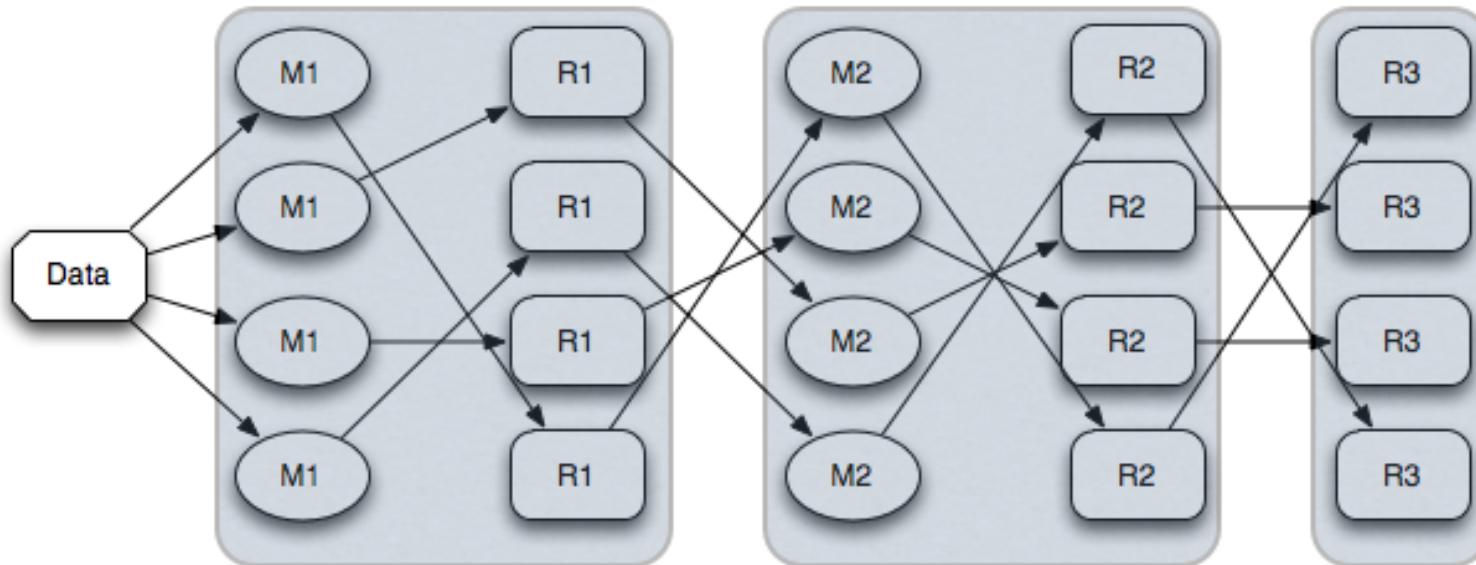
- Project codes can *almost* be tied to country/region

Demo 2

ETL Thoughts

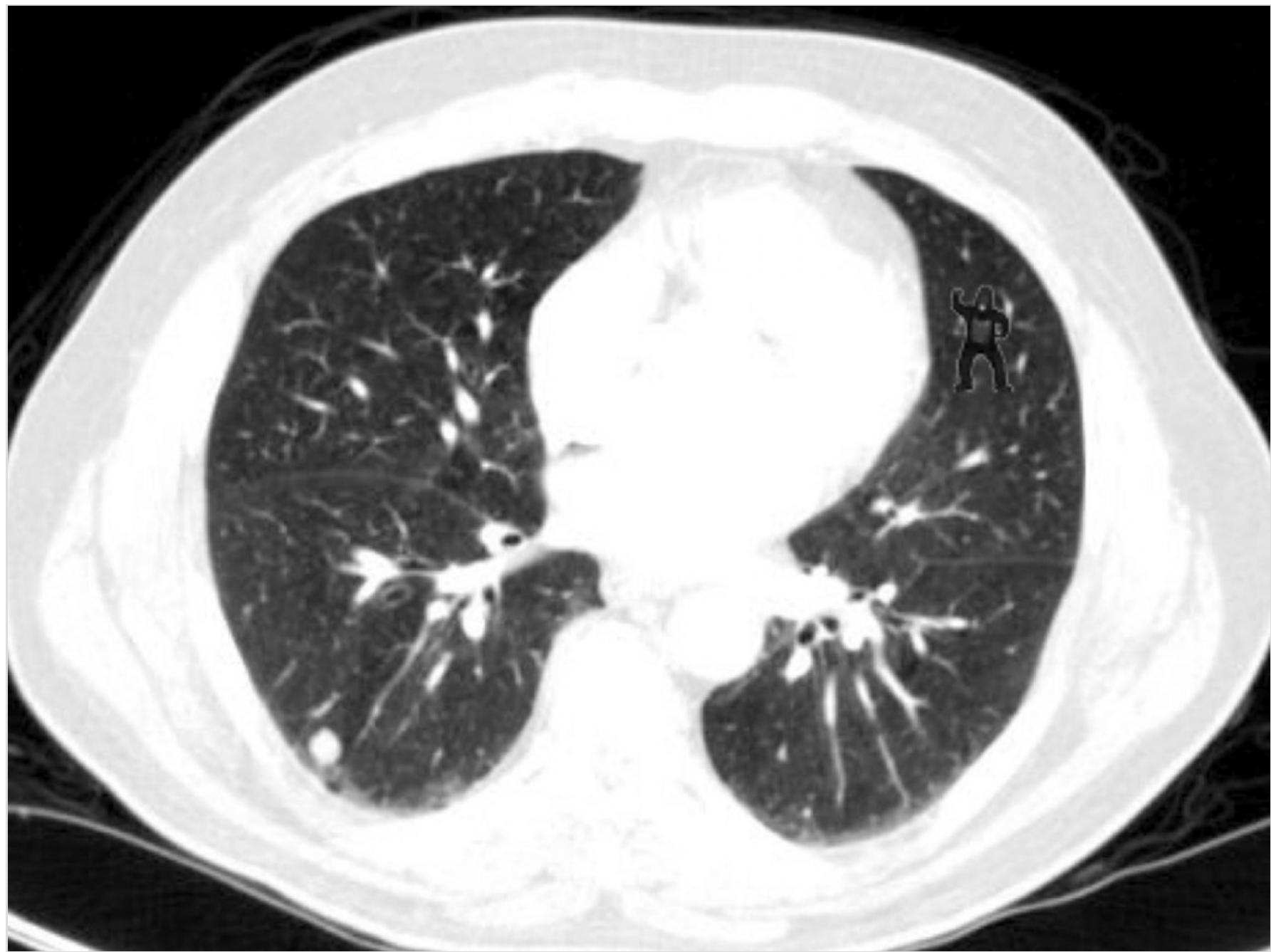
- Try not to use MapReduce for ETL for every job
- Often not tied to SQL
- Be conscience of file storage
 - JSON
 - CSV
 - HDF5
 - AVRO
 - NETCDF4
 - NPZ -Compression
 - IOPS are expensive, GBs/TBs/PBs are cheap

Chaining Jobs

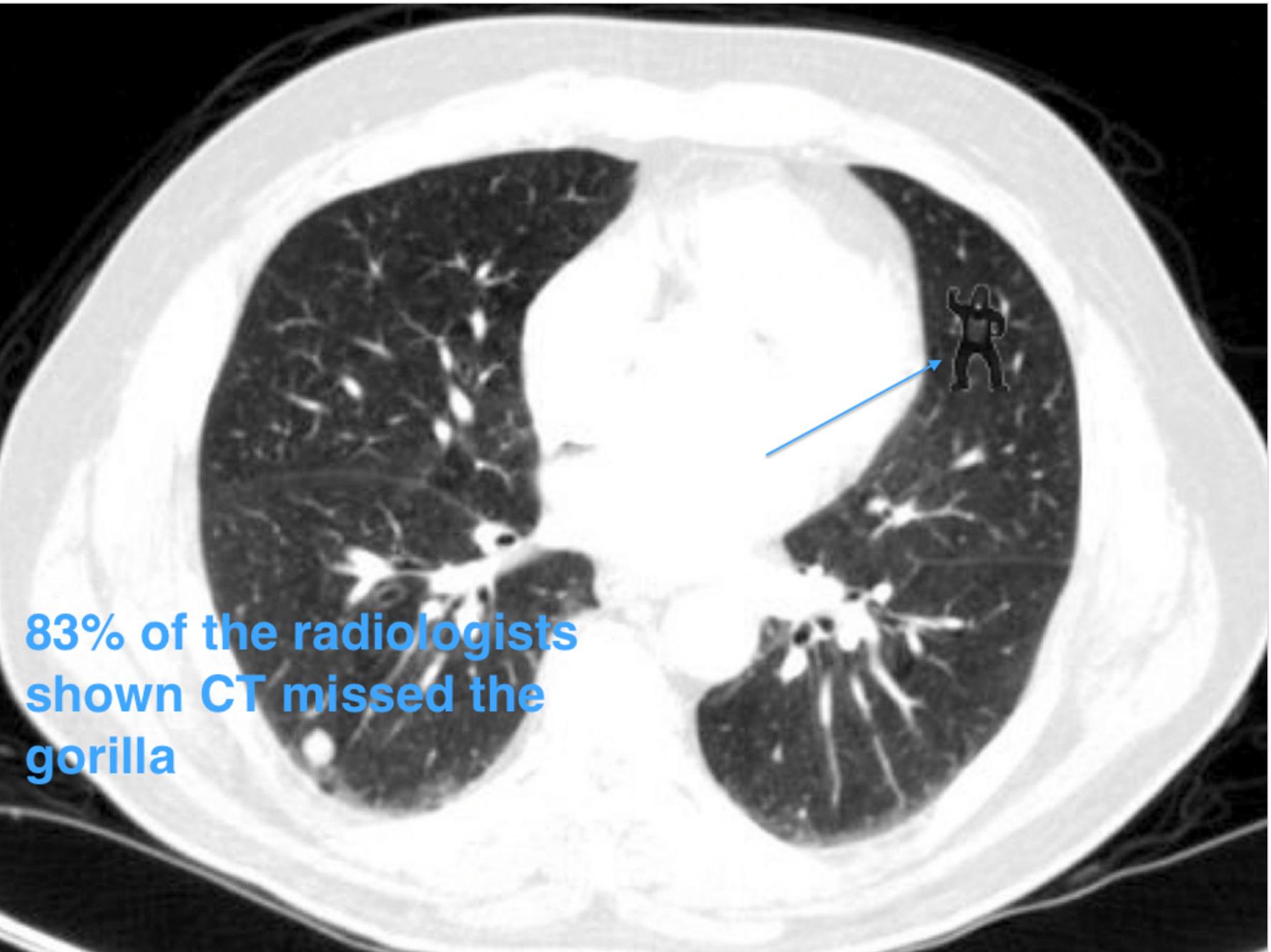


MapReduce Thoughts

- Data Cleansing
 - Everyone's pain point
- Task Deconstruction
 - Good for code management
 - Hides -- in a good way -- data management
- Can Be Inefficient
 - Network traffic
 - Job organization



M



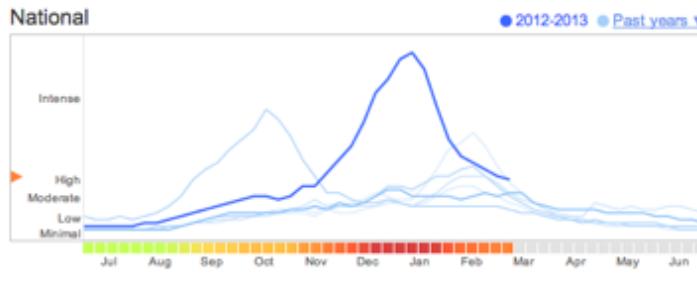
**83% of the radiologists
shown CT missed the
gorilla**

M

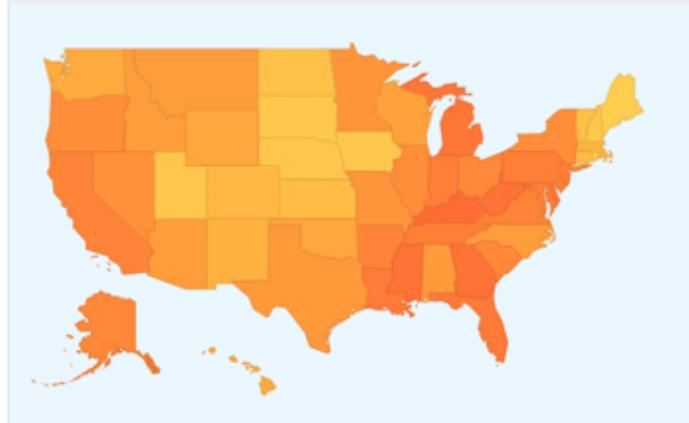
Google Flu

Explore flu trends - United States

We've found that certain search terms are good indicators of flu activity. Google Flu Trends uses aggregated Google search data to estimate flu activity. [Learn more ↗](#)



[States](#) | [Cities](#) (Experimental)

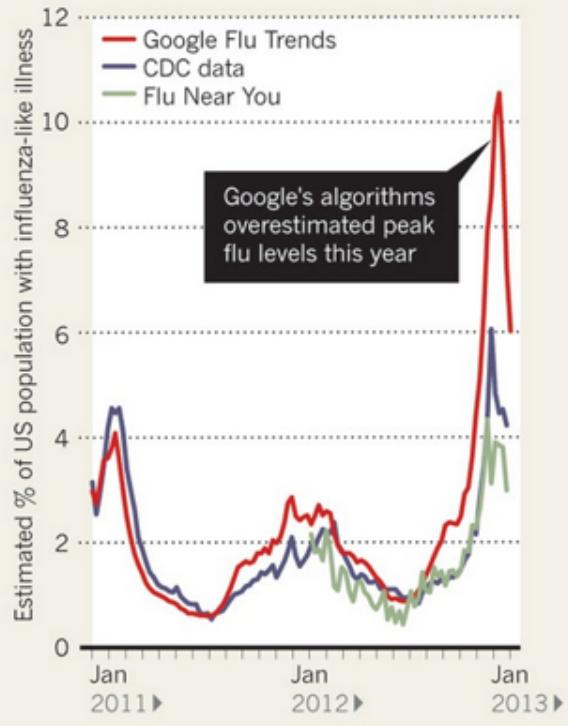


- Data Mining
- Faster than CDC

Google Gets It Wrong

FEVER PEAKS

A comparison of three different methods of measuring the proportion of the US population with an influenza-like illness.



- Typically, prediction is great!
- This year not so much
- Google: No comment!
- Feedback mechanism from hype-up media

Data Philosophy

- Invisible Gorillas will stay Invisible
 - Inattentional Blindness
- Machine Learning without Oversight
 - Turnkey analytics is dangerous
- Good Analysis
 - Requires iterative exploration
 - Peer review and collaboration