

# What is data?

---

Jeff Leek

@jtleek

[www.jtleek.com](http://www.jtleek.com)

**What is data?**

“Data are values of qualitative or quantitative variables, belonging to a set of items.”

“Data are values of qualitative or quantitative variables, belonging to a set of items.”

**Set of items:** Sometimes called the population; the set of objects you are interested in

“Data are values of qualitative or quantitative **variables**, belonging to a set of items.”

**Variables:** A measurement or characteristic of an item

“Data are values of **qualitative** or **quantitative** variables, belonging to a set of items.”

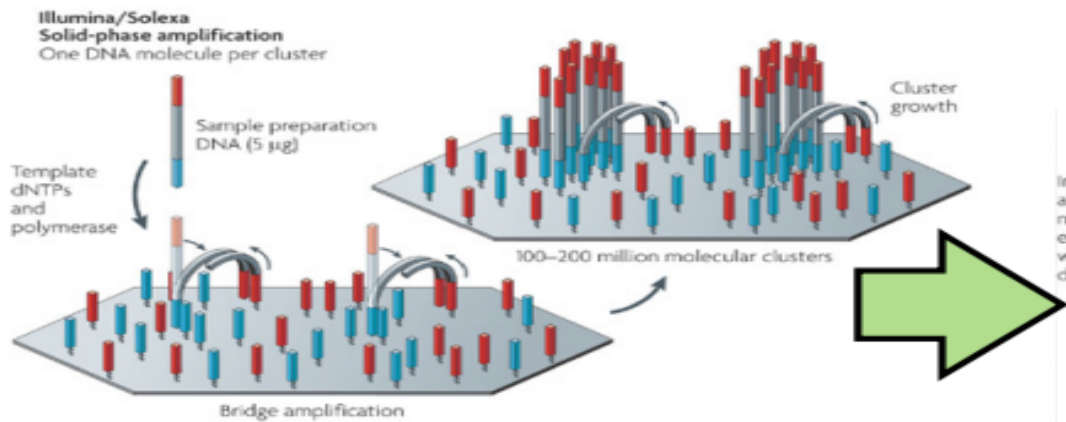
**Qualitative:** Country of origin, sex, treatment

**Quantitative:** Height, weight, blood pressure

**An example of a processing pipeline**







Source: Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet. 2010

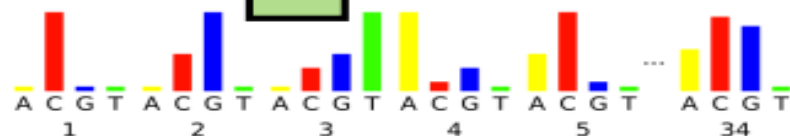
```

@HWI-EAS146:5:1:1:961#0/1
TCCGAGGCCAACCGAGGCTCCGCGCGCTGNNNNNNNNNNNNNN
+
BBBB>A7B@;@BBBBBAA-BA-A~~~~~
@HWI-EAS146:5:1:1:1595#0/1
TCAGGAAGCAGGAAGAGCTGGTCAGCAGNNNNNNNNNNNGNNNN
+
B9B@B<;BAA<@B@9=1>~~~~~
@HWI-EAS146:5:1:1:1048#0/1
CTGGACTGCATCCTACCACCAACTCGTCCAANNNNNNNNNNNN
+
A-B767:>B@>A>79<:>747~~~~~
@HWI-EAS146:5:1:1:1607#0/1
CTCCTCTCAAGGTCCTCCAGAAAGCACAGCCAANNNNNANTNCTN
+
BBCCCCCBBB<B7CB<7>+<>=B<BCB~~~~~
@HWI-EAS146:5:1:1:1719#0/1
CACGATCTGGGTTTATTGTAACTCCGCTC>NNNNNGNTAAAGNN
+
BCC7+<B=7BBS=ABA7BGBB@4BB7B~~~~~
@HWI-EAS146:5:1:2:947#0/1
CCCAGGAGAAAGCCATGTTCACTTCAGGCGCNAMANCCTGANNN
+
BB@@TTAT>A@>T@>T@>B7B~~~~~
@HWI-EAS146:5:1:2:563#0/1

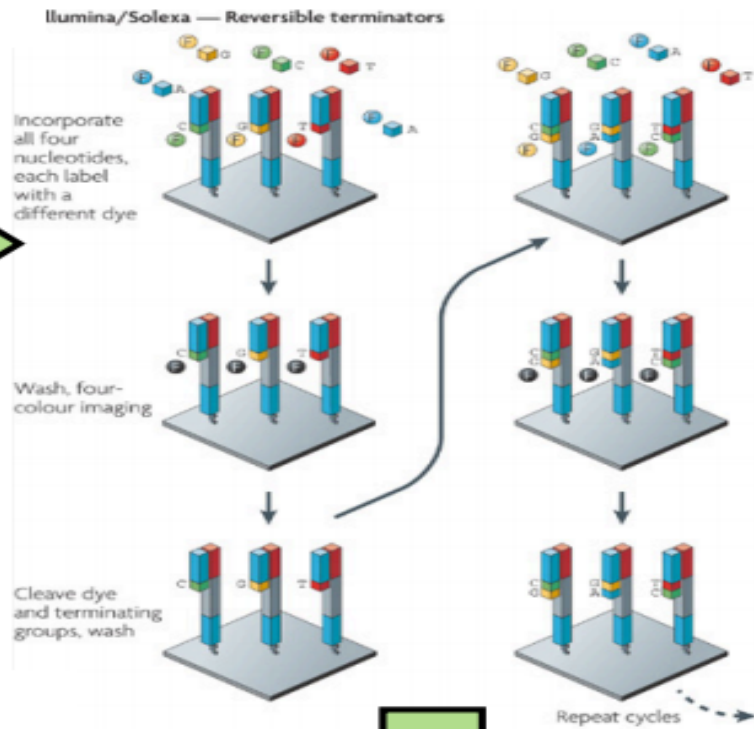
```

name  
sequence  
quality scores

x 100s of  
millions



Source: Whiteford et al. Swift: primary data analysis for the Illumina Solexa sequencing platform. Bioinformatics. 2009



Source: Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet. 2010

**Raw data**

**= no computations**

**Processed data**

**= final, tidy dataset**

**The goal is to be interactive**



**Elle McDonagh**

@ElleMcDonagh



Follow

Robert Gentleman, Genentech: "make big data as small as possible as quick as is possible" to enable sharing [#bigdatamed](#)

↩ Reply ↻ Retweet ★ Favorite ⋮ More

RETWEETS

4

FAVORITES

7



11:34 AM - 21 May 2014

# Don't use Hadoop - your data isn't that big

Mon 16 September 2013 [big data](#) / [buzzwords](#) / [hadoop](#)

 Follow @stucchio  Tweet 2,403

 Like  Share 1,365 people like this. [Sign Up](#) to see what your friends like.

 +581 Recommend this on Google

"So, how much experience do you have with Big Data and Hadoop?" they asked me. I told them that I use Hadoop all the time, but rarely for jobs larger than a few TB. I'm basically a big data neophyte - I know the concepts, I've written code, but never at scale.

The next question they asked me. "Could you use Hadoop to do a simple group by and sum?" Of course I could, and I just told them I needed to see an example of the file format.

They handed me a flash drive with all 600MB of their data on it (not a sample, everything). For reasons I can't understand, they were unhappy when my solution involved `pandas.read_csv` rather than Hadoop.

Hadoop is limiting. Hadoop allows you to run one general computation, which I'll illustrate in pseudocode:

Scala-ish pseudocode:

```
collection.flatMap( (k,v) => F(k,v) ).groupBy( _. _1 ).map( _.reduce( (k,v) => G(k,v) ) )
```

SQL-ish pseudocode: