What is data?

Jeff Leek

@jtleek

www.jtleek.com

What is data?

http://en.wikipedia.org/wiki/Data

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

http://en.wikipedia.org/wiki/Data

"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

http://en.wikipedia.org/wiki/Data

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

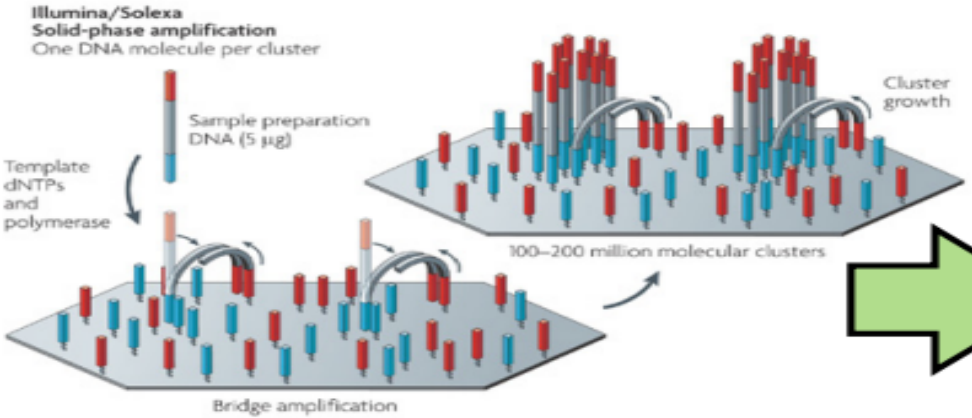
Variables: A measurement or characteristic of an item

http://en.wikipedia.org/wiki/Data

"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

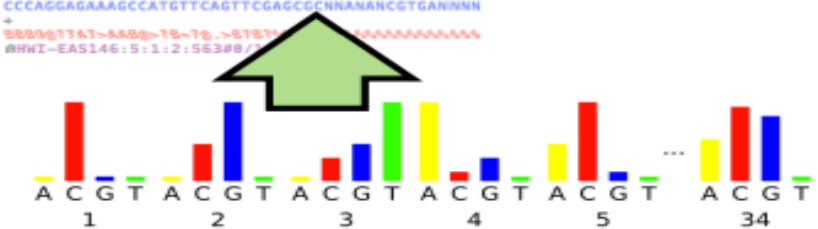
An example of a processing pipeline



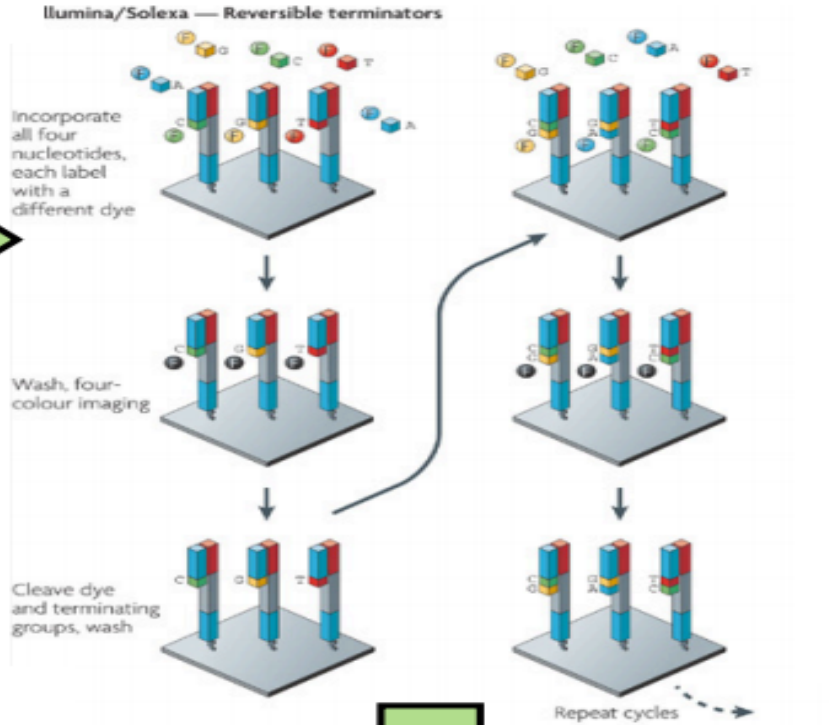
Source: Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet. 2010

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

https://twitter.com/EllieMcDonagh/status/469184554549248000

http://www.chrisstucchio.com/blog/2013/hadoop_hatred.html

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

Mon 16 September 2013 big data / buzzwords / hadoop

"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 neophite - 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: