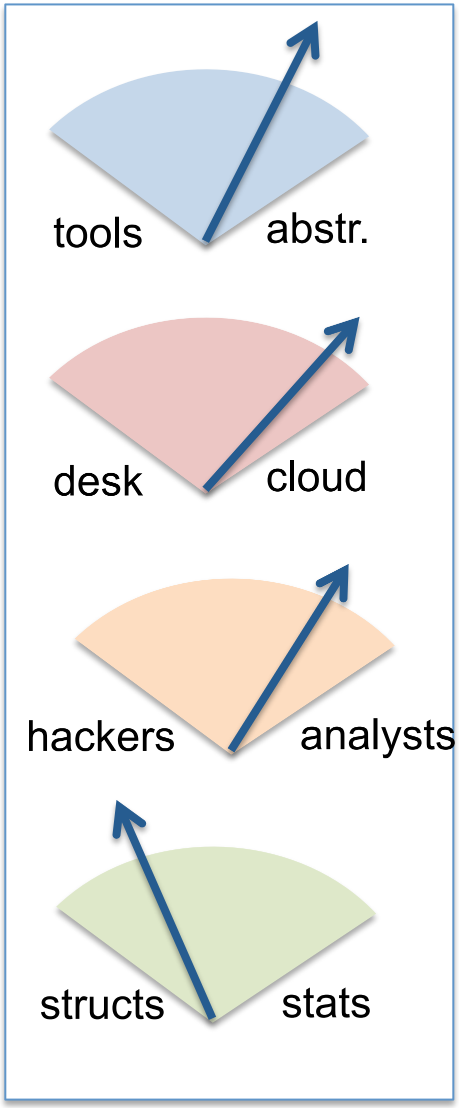
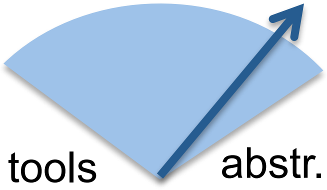
This Course

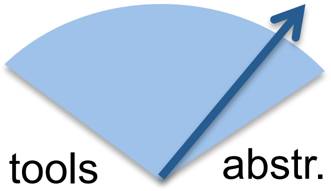


What goes around comes around



- Pre-2004: commercial RDBMS, some open source
- 2004 Dean et al. MapReduce
- 2008 Hadoop 0.17 release
- 2008 Olston et al. Pig: Relational Algebra on Hadoop
- 2008 DryadLINQ: Relational Algebra in a Hadoop-like system
- 2009 Thusoo et al. HIVE: SQL on Hadoop
- 2009 Hbase: Indexing for Hadoop
- 2010 Dietrich et al. Schemas and Indexing for Hadoop
- 2012 Transactions in HBase (plus VoltDB, other NewSQL systems)

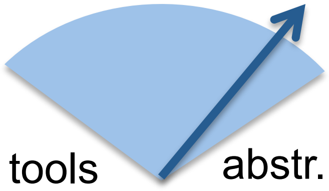
- But also some permanent contributions:
  - Fault tolerance
  - Schema-on-Read
  - User-defined functions that don’t suck



What are the *abstractions* of data science?

"Data Jujitsu"
"Data Wrangling"
"Data Munging"

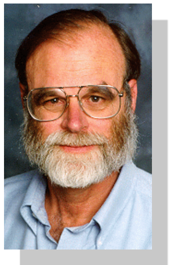
Translation: "We have no idea what this is all about"

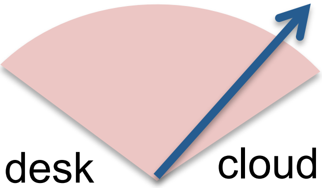


What are the *abstractions* of data science?

matrices and linear algebra?
relations and relational algebra?
objects and methods?
files and scripts?
data frames and functions?

Data Access Hitting a Wall





Current practice based on data download (FTP/GREP)
Will not scale to the datasets of tomorrow

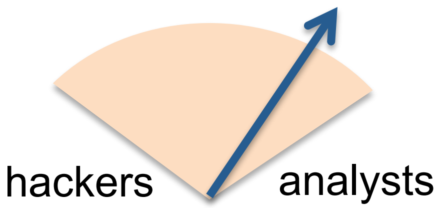
- You can GREP 1 MB in a second
- You can GREP 1 GB in a minute
- You can GREP 1 TB in 2 days
- You can GREP 1 PB in 3 years.

- Oh!, and 1PB ~5,000 disks

- At some point you need
  indices to limit search
  parallel data search and analysis
- This is where databases can help

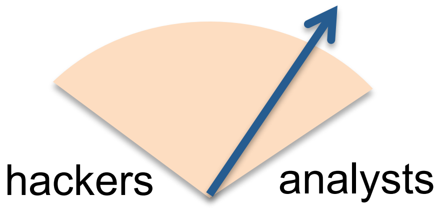
- You can FTP 1 MB in 1 sec
- You can FTP 1 GB / min (~1$)
- … 2 days and 1K$
- … 3 years and 1M$

[slide src: Jim Gray]



US faces shortage of 140,000 to 190,000 people “with deep analytical skills, as well as 1.5 million managers and analysts with the know-how to use the analysis of big data to make effective decisions.”

--Mckinsey Global Institute



Biologists are beginning to write very complex queries (rather than relying on staff programmers)

Example: Computing the overlaps of two sets of blast results

```
   SELECT x.strain, x.chr, x.region as snp_region, x.start_bp as snp_start_bp
      , x.end_bp as snp_end_bp, w.start_bp as nc_start_bp, w.end_bp as nc_end_bp
      , w.category as nc_category
      , CASE WHEN (x.start_bp >= w.start_bp AND x.end_bp <= w.end_bp)
     THEN x.end_bp - x.start_bp + 1
     WHEN (x.start_bp <= w.start_bp AND w.start_bp <= x.end_bp)
     THEN x.end_bp - w.start_bp + 1
     WHEN (x.start_bp <= w.end_bp AND w.end_bp <= x.end_bp)
     THEN w.end_bp - x.start_bp + 1
    END  AS len_overlap

    FROM [koesterj@washington.edu].[hotspots_deserts.tab] x
    INNER JOIN [koesterj@washington.edu].[table_noncoding_positions.tab] w
    ON x.chr = w.chr
    WHERE (x.start_bp >= w.start_bp AND x.end_bp <= w.end_bp)
    OR (x.start_bp <= w.start_bp AND w.start_bp <= x.end_bp)
    OR (x.start_bp <= w.end_bp AND w.end_bp <= x.end_bp)
    ORDER BY x.strain, x.chr ASC, x.start_bp ASC
```

We see thousands of queries written by non-programmers