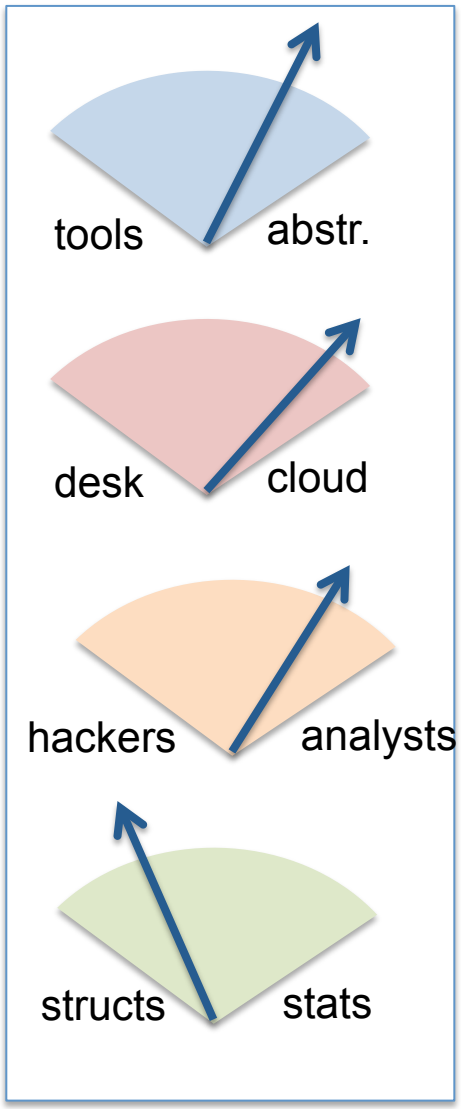
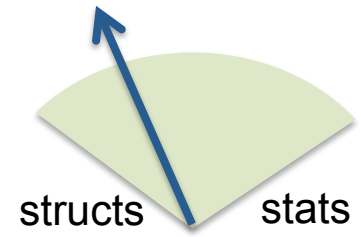
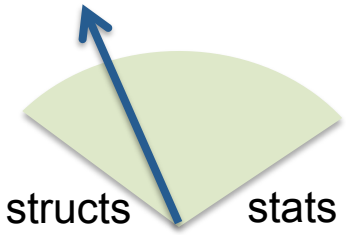
This Course





"80% of analytics is sums and averages"

-- Aaron Kimball, wibidata



Three types of tasks:

1) Preparing to run a model

"80% of the work"

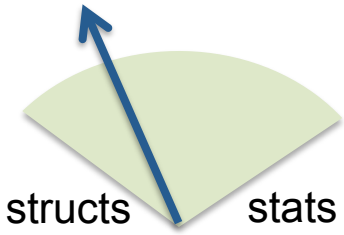
-- Aaron Kimball

Gathering, cleaning, integrating, restructuring, transforming, loading, filtering, deleting, combining, merging, verifying, extracting, shaping, massaging

2) Running the model

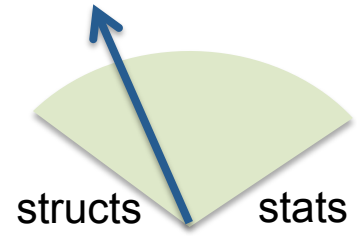
3) Interpreting the results

The other 80% of the work?



"…no greater barrier to effective data management will exist than the variety of incompatible data formats, non-aligned data structures, and inconsistent data semantics."

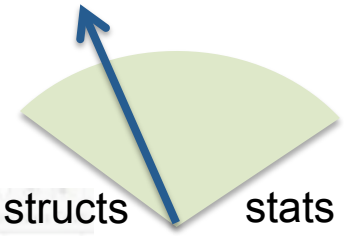
Doug Laney, "3-D Data Management: Controlling Data Volume, Velocity and Variety", Gartner, 2001



Problem

How much time do you spend "handling data" as opposed to "doing science"?

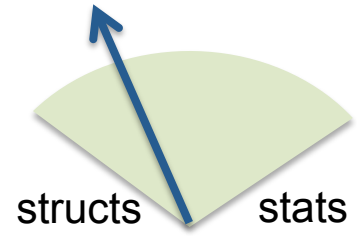
Mode answer: "90%"



- Databases and Statistical Packages
  - Many analysts download data to use in Excel/SAS/ Matlab/R or their favorite programming language?
    FORTRAN??
  - Use matrix/vector operations
  - Most of these stat packages require data to fit in RAM
    - Taking samples from the full data to fit into ram results in loss of precision
  - External toolkits may also lack parallelism

src: Christian Grant, MADSkills

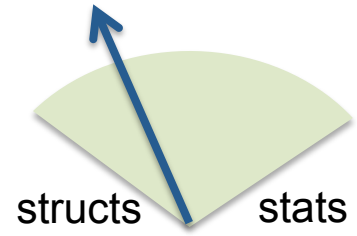
(Sparse) Matrix Multiply in SQL




```
SELECT A. row_number,  B.column_number, SUM(A.value * B.value)
FROM A, B
WHERE A.column_number = B.row_number
GROUP BY A.row_number,  B.column_number
```

src: Christian Grant, MADSkills

Aside: Schema-on-Write vs. Schema-on-Read



- A schema* is a shared consensus about some universe of discourse
- At the frontier of research, this shared consensus does not exist, *by definition*
- Any schema that does emerge will change frequently, *by definition*
- Data found "in the wild" will typically not conform to any schema, *by definition*
- But this doesn't mean we have to live with ad hoc scripts and files
- A good approach: Schema-later! Schemas are important, but not a prerequisite to processing.

* ontology/metadata standard/controlled vocabulary/etc.