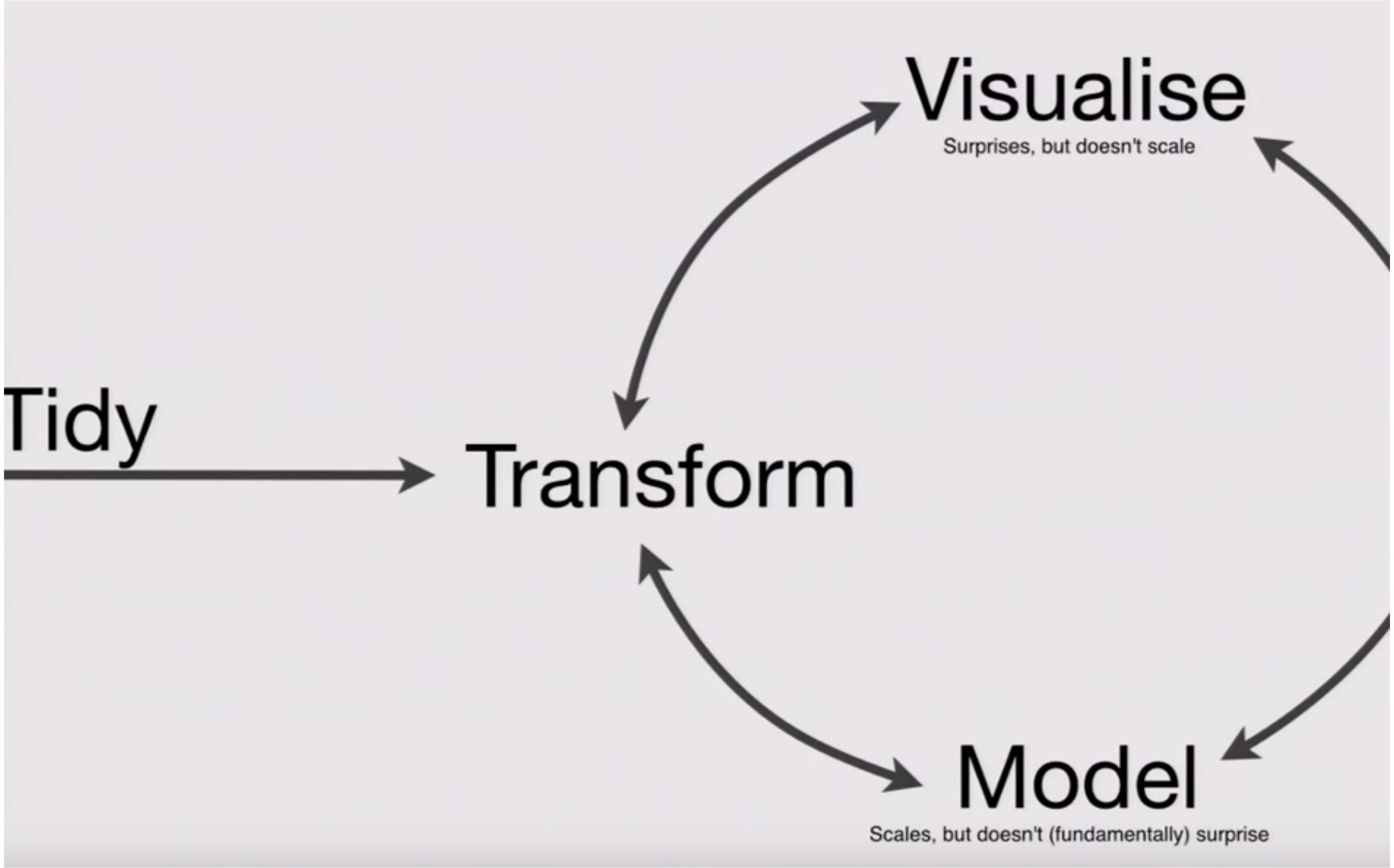
Systems for Interactive Data Analysis

CSC 630

What’s data analysis?
What systems are out there?
Why does it need to be interactive?
What can we do about it?

Data Analysis



Visualise
Surprises, but doesn't scale
Tidy
Transform
Model
Scales, but doesn't (fundamentally) surprise

https://www.youtube.com/watch?v=40tyOFMZUSM

Slide by Hadley Wickham

Data Analysis

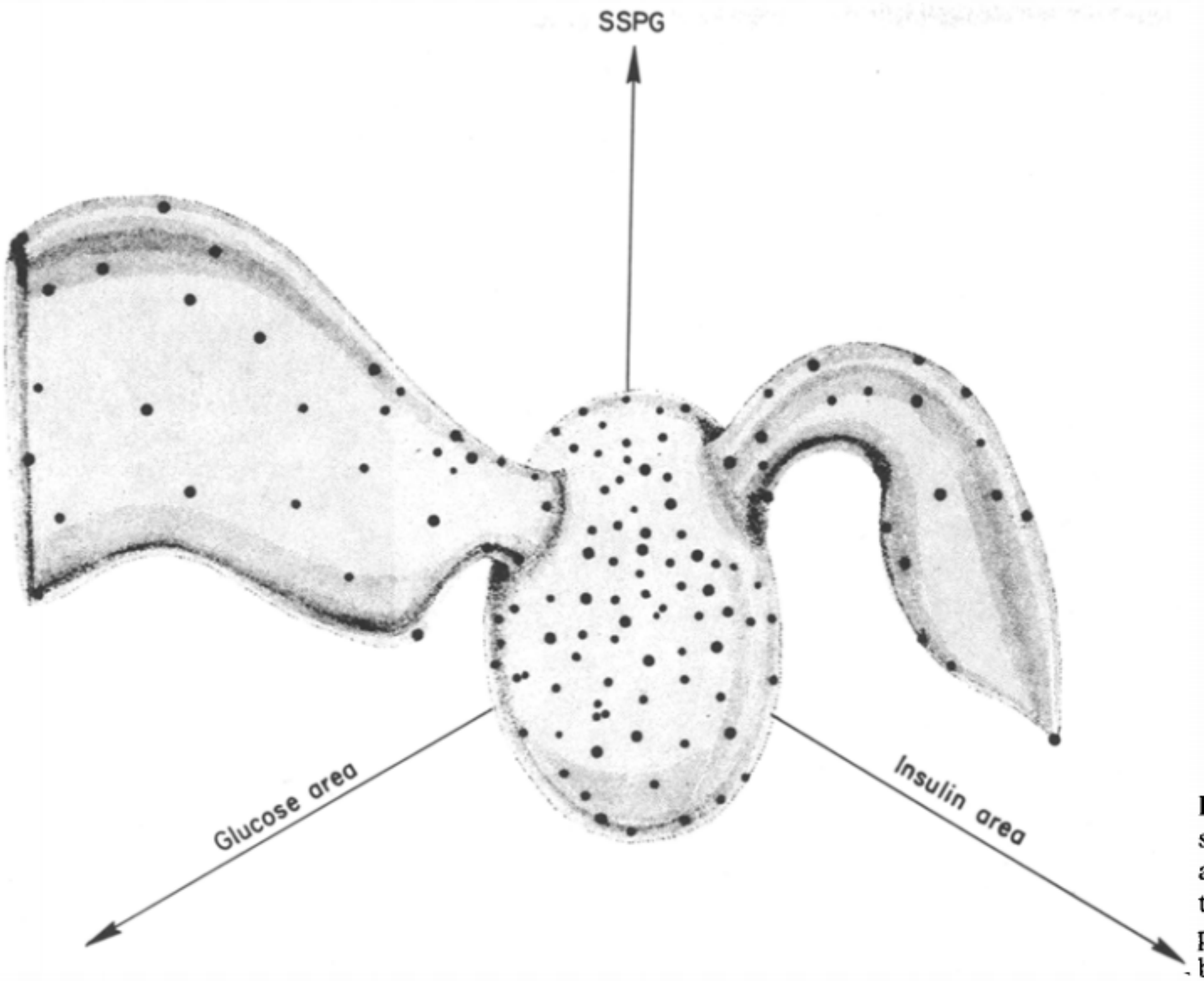


Fig. 1. Artist's rendition of data as seen in three dimensions. View is approximately along 45° line as seen through Prim 9 program on the computer; coordinate axes are in the background

http://www.math.utk.edu/~fernando/Students/GregClark/pdf/Miller-Reaven-Diabetes.pdf

(What’s Prim9?)

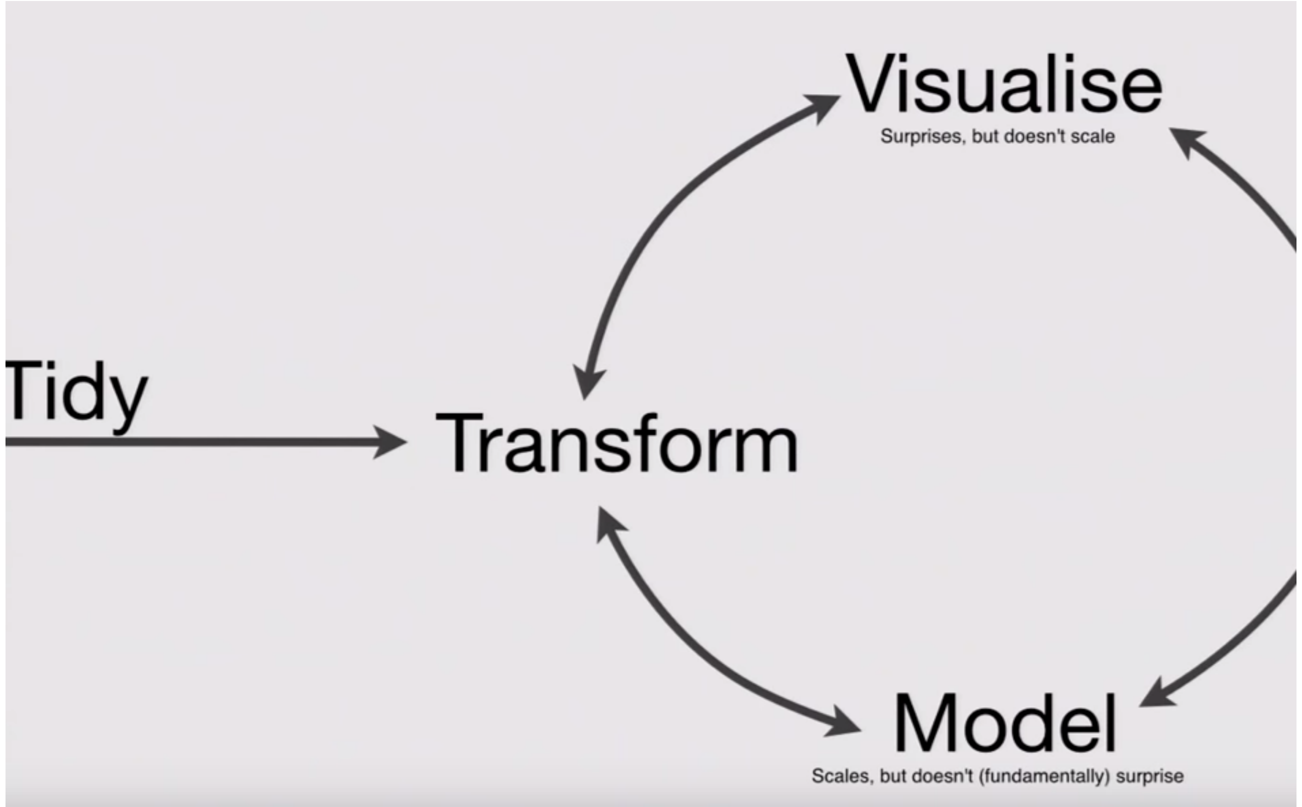
https://www.youtube.com/watch?v=B7XoW2qiFUA

Infrastructure for Interactive Data Analysis

- Software libraries so that people can think better

- Faster software so that people can keep thinking

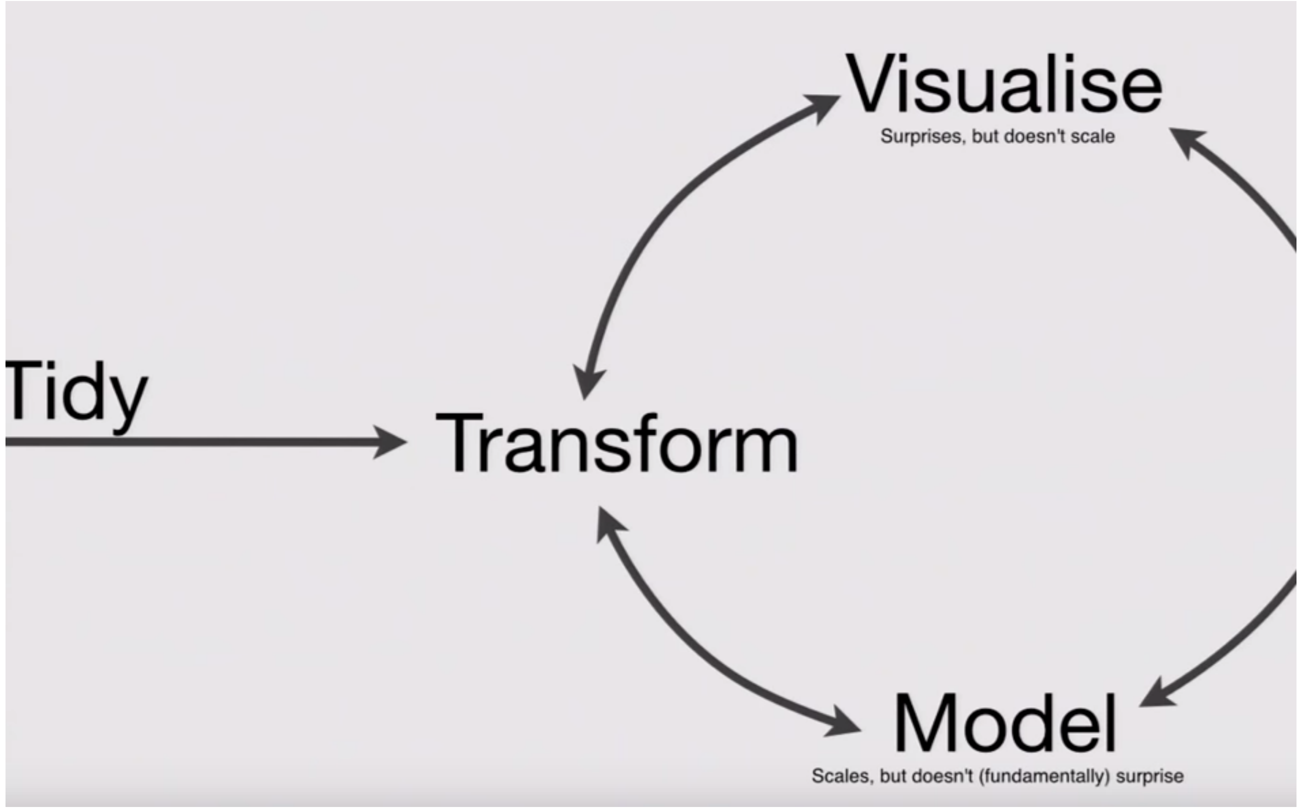
- Software so that people can think about their data at all


Visualise
Surprises, but doesn't scale
Tidy
Transform
Model
Scales, but doesn't (fundamentally) surprise

Why Interaction?

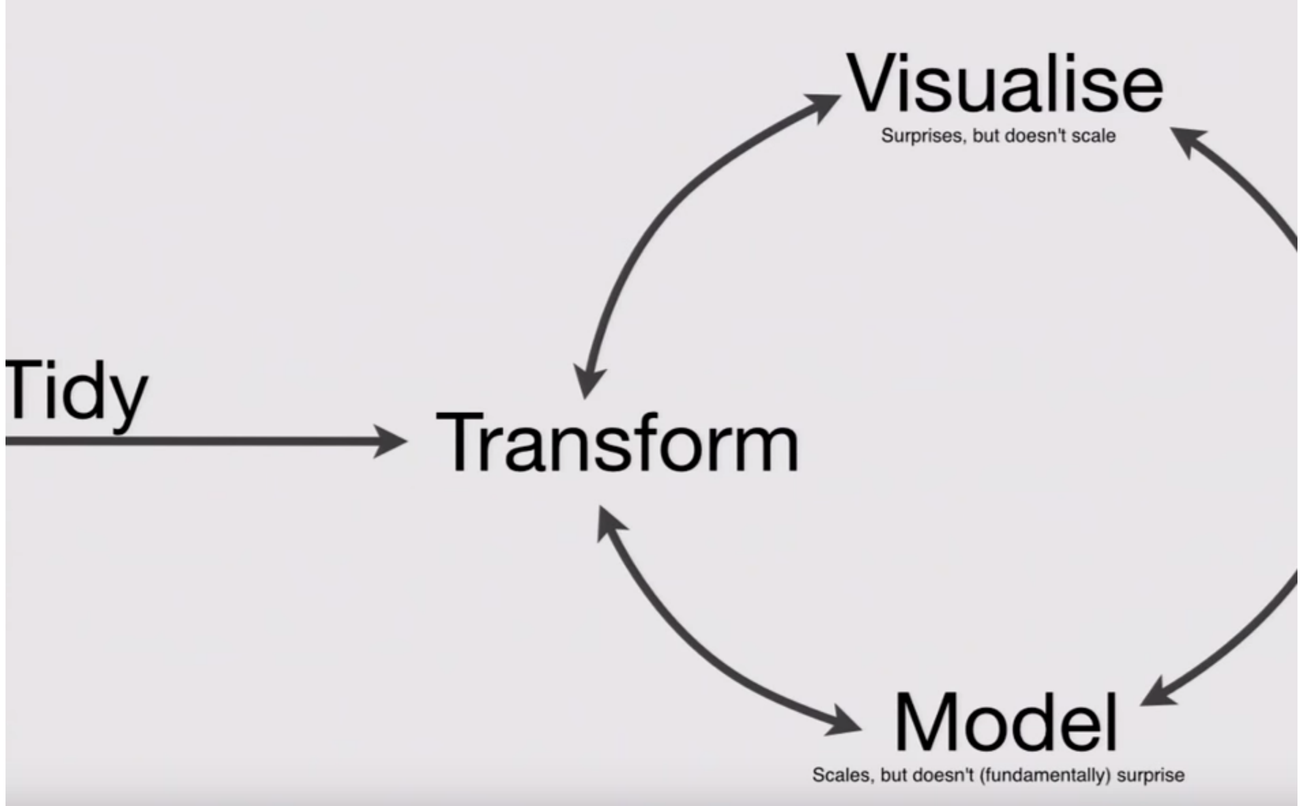
- What matters is people, and what they can do.

- Take your fanciest machine learning algorithm. It produced a result. **Now What?**


Visualise
Surprises, but doesn't scale
Tidy
Transform
Model
Scales, but doesn't (fundamentally) surprise

“Often, the hardest part is just getting the data”

- What are the best systems for cleaning data today?
  - Wrangler: now trifacta.


Visualise
Surprises, but doesn't scale
Tidy
Transform
Model
Scales, but doesn't (fundamentally) surprise

https://vimeo.com/19185801

http://vis.stanford.edu/files/2011-Wrangler-CHI.pdf

“Visualizations can surprise you”

- Concern: How do we help people build better visualizations?
    - **Make it easier to write code.**
- Progress over 5 years
    - http://prefuse.org/ Java
    - http://flare.prefuse.org/ (flash, web!)
    - http://mbostock.github.io/protovis/ (Javascript!)
    - http://d3js.org/ (More than just plots, **flexibility**)

What’s the problem with d3?

- still too much code
    - vega
    - vega-lite
- too slow!
    - **Use other technologies: WebGL**
    - http://threejs.org/ https://github.com/cscheid/lux

What's the problem with d3?

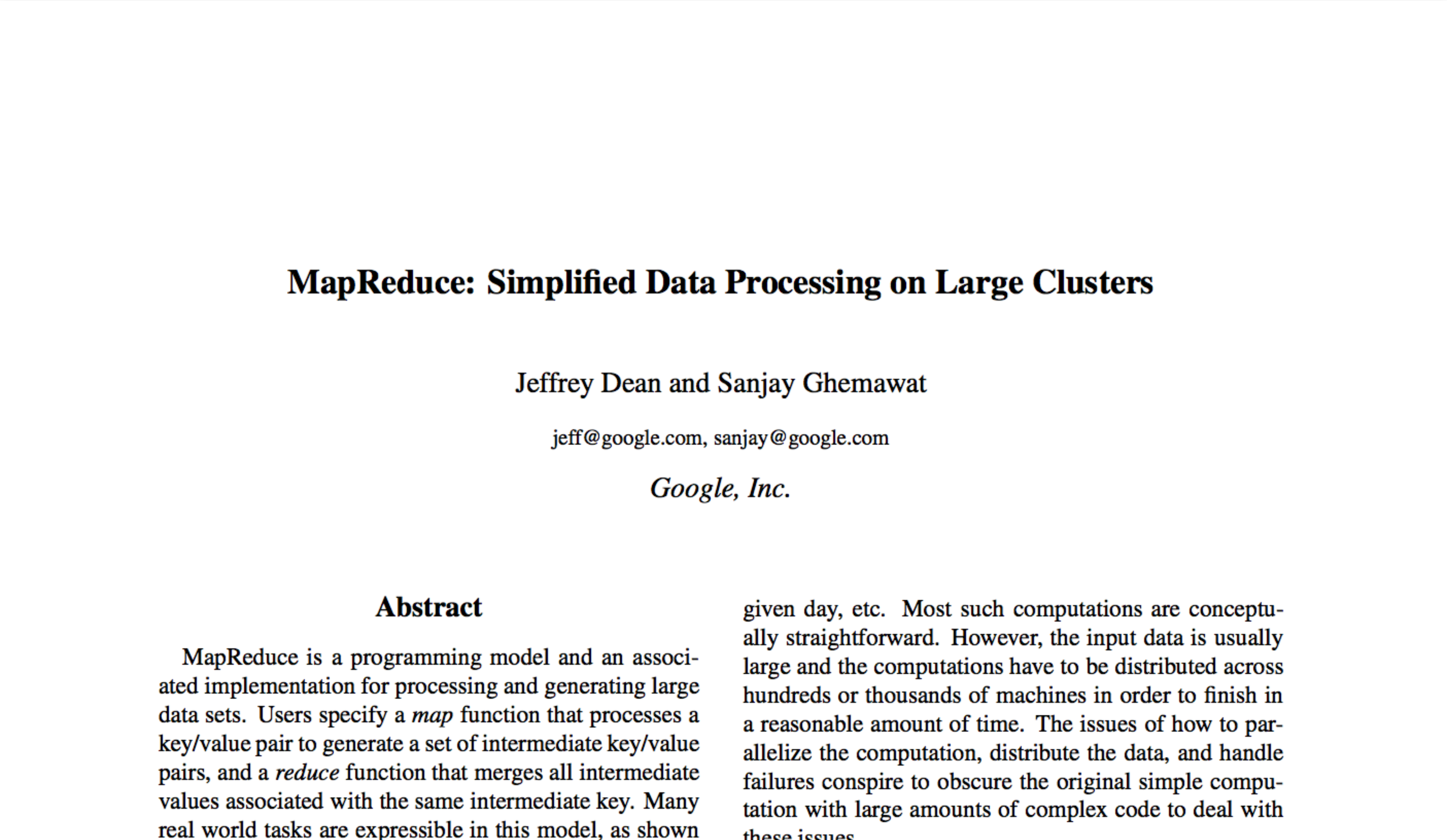
- Hard to write
    - **Visual debugger for d3**
    - **Visual debugger for data analysis**
- (example: http://benvanik.github.io/WebGL-Inspector/)

Modeling

- What kind of software can we write to help people with modeling?
- we can make it faster

MapReduce, Dremel, Hadoop, Hive, Pig, oh my

http://static.googleusercontent.com/media/research.google.com/en//archive/mapreduce-osdi04.pdf



MapReduce: Simplified Data Processing on Large Clusters

Jeffrey Dean and Sanjay Ghemawat

jeff@google.com, sanjay@google.com

Google, Inc.

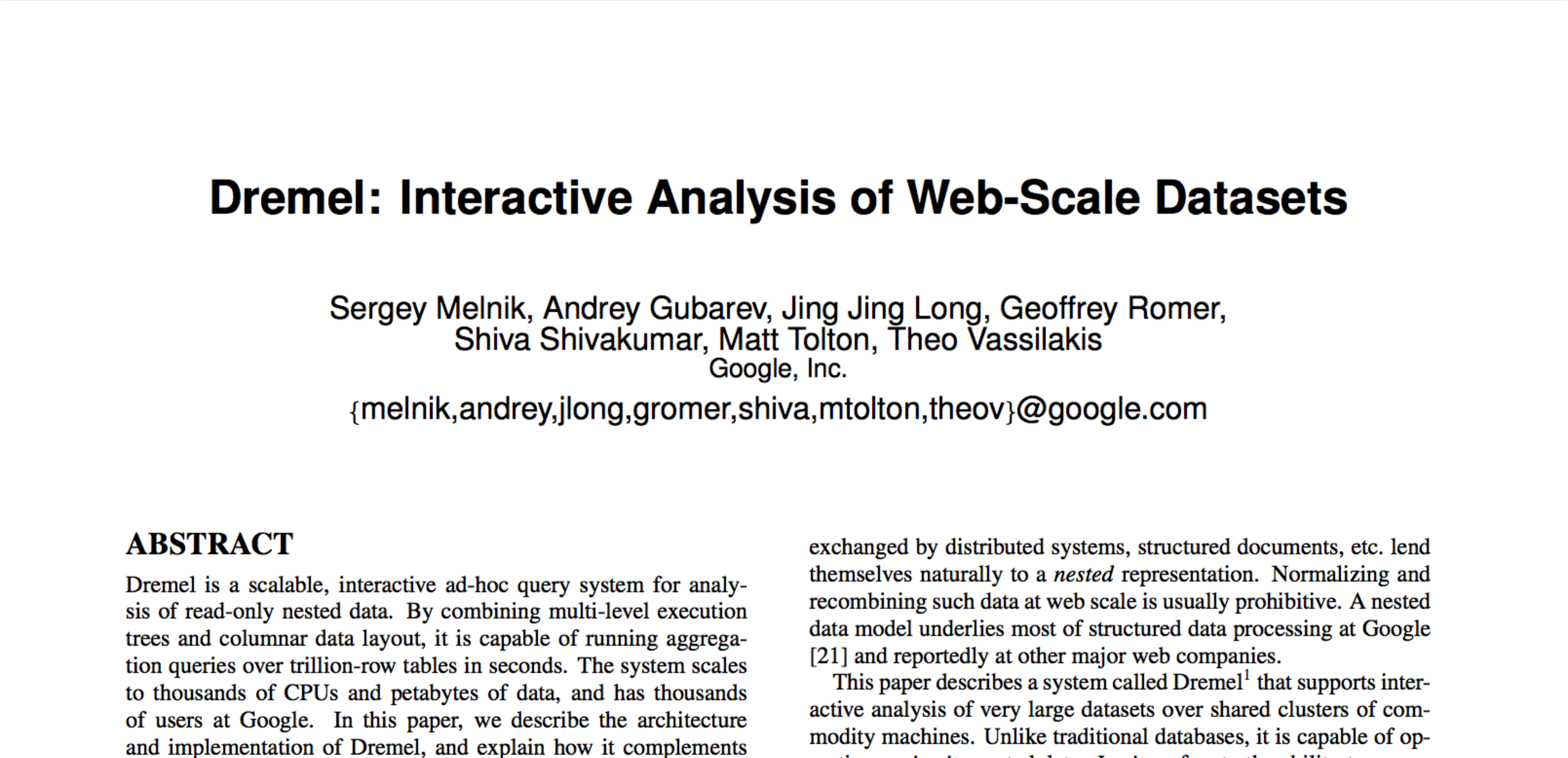
Abstract

MapReduce is a programming model and an associated implementation for processing and generating large data sets. Users specify a *map* function that processes a key/value pair to generate a set of intermediate key/value pairs, and a *reduce* function that merges all intermediate values associated with the same intermediate key. Many real world tasks are expressible in this model, as shown

given day, etc. Most such computations are conceptually straightforward. However, the input data is usually large and the computations have to be distributed across hundreds or thousands of machines in order to finish in a reasonable amount of time. The issues of how to parallelize the computation, distribute the data, and handle failures conspire to obscure the original simple computation with large amounts of complex code to deal with these issues.

MapReduce, Dremel, Hadoop, Hive, Pig, oh my

http://static.googleusercontent.com/media/research.google.com/en//pubs/archive/36632.pdf



Dremel: Interactive Analysis of Web-Scale Datasets

Sergey Melnik, Andrey Gubarev, Jing Jing Long, Geoffrey Romer,
Shiva Shivakumar, Matt Tolton, Theo Vassilakis
Google, Inc.
{melnik,andrey,jlong,gromer,shiva,mtolton,theov}@google.com

ABSTRACT

Dremel is a scalable, interactive ad-hoc query system for analysis of read-only nested data. By combining multi-level execution trees and columnar data layout, it is capable of running aggregation queries over trillion-row tables in seconds. The system scales to thousands of CPUs and petabytes of data, and has thousands of users at Google. In this paper, we describe the architecture and implementation of Dremel, and explain how it complements

exchanged by distributed systems, structured documents, etc. lend themselves naturally to a *nested* representation. Normalizing and recombining such data at web scale is usually prohibitive. A nested data model underlies most of structured data processing at Google [21] and reportedly at other major web companies.

This paper describes a system called Dremel[1] that supports interactive analysis of very large datasets over shared clusters of commodity machines. Unlike traditional databases, it is capable of op-

But we’re not going to out-google Google

So what can we do?

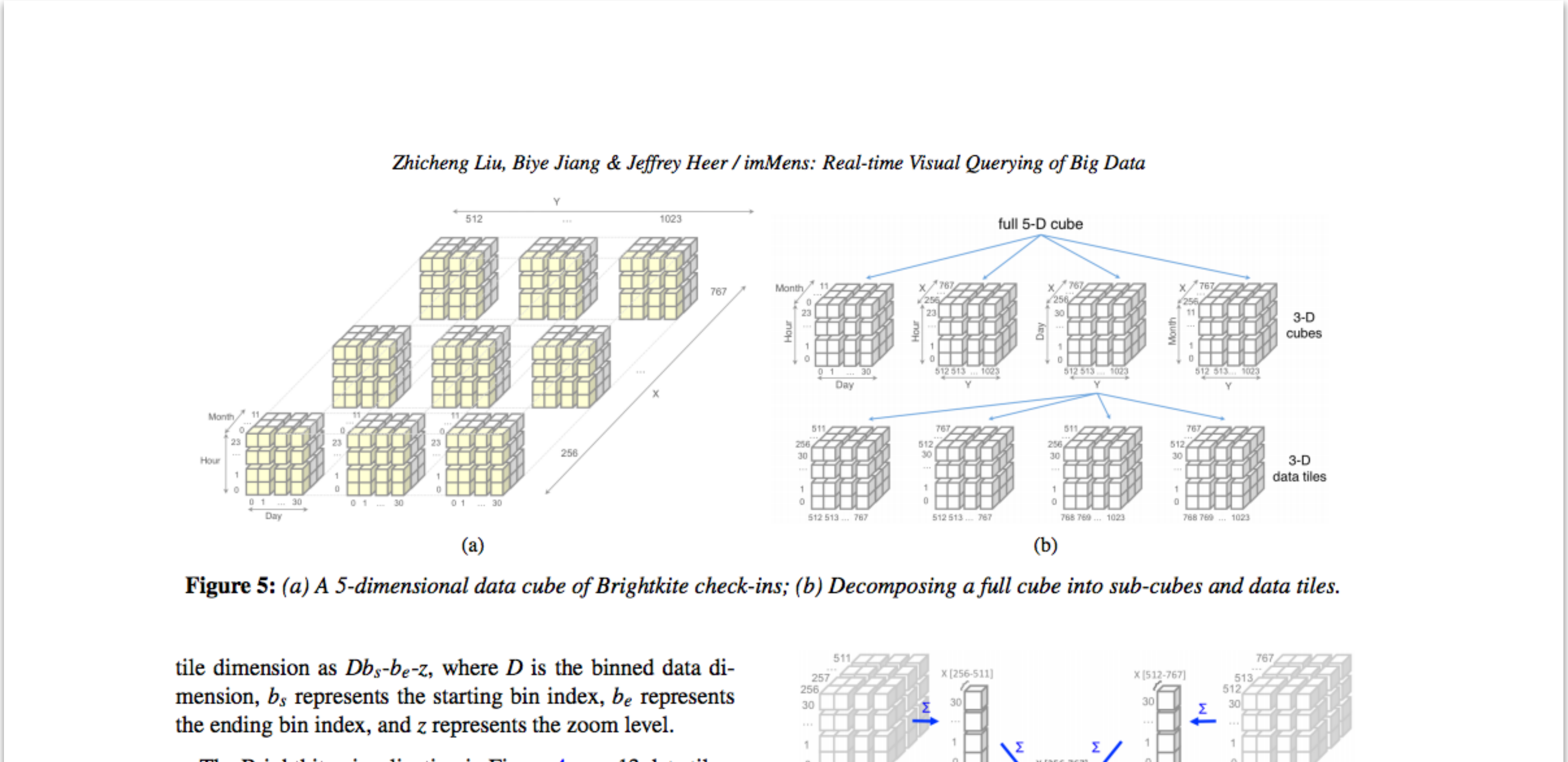
Find smaller chunks
of the same problem

Datavore

- “in-browser database engine” http://vis.stanford.edu/projects/datavore/
- “written in part to support Profiler, a system for integrated statistical and visual data analysis”
  - http://vis.stanford.edu/papers/profiler

Immens: Real-Time Visual Querying of Big Data

- http://vis.stanford.edu/papers/immens



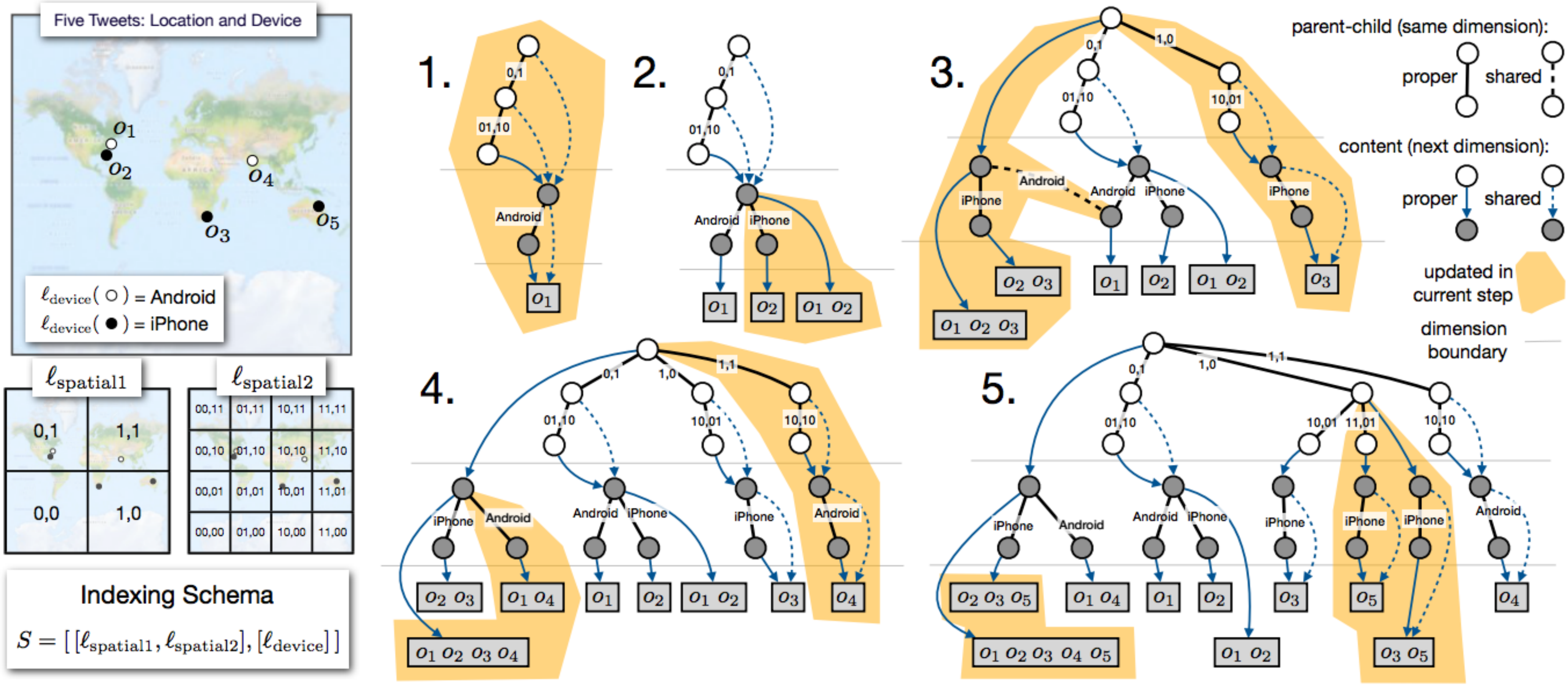
Zhicheng Liu, Biye Jiang & Jeffrey Heer / imMens: Real-time Visual Querying of Big Data

Figure 5: *(a) A 5-dimensional data cube of Brightkite check-ins; (b) Decomposing a full cube into sub-cubes and data tiles.*

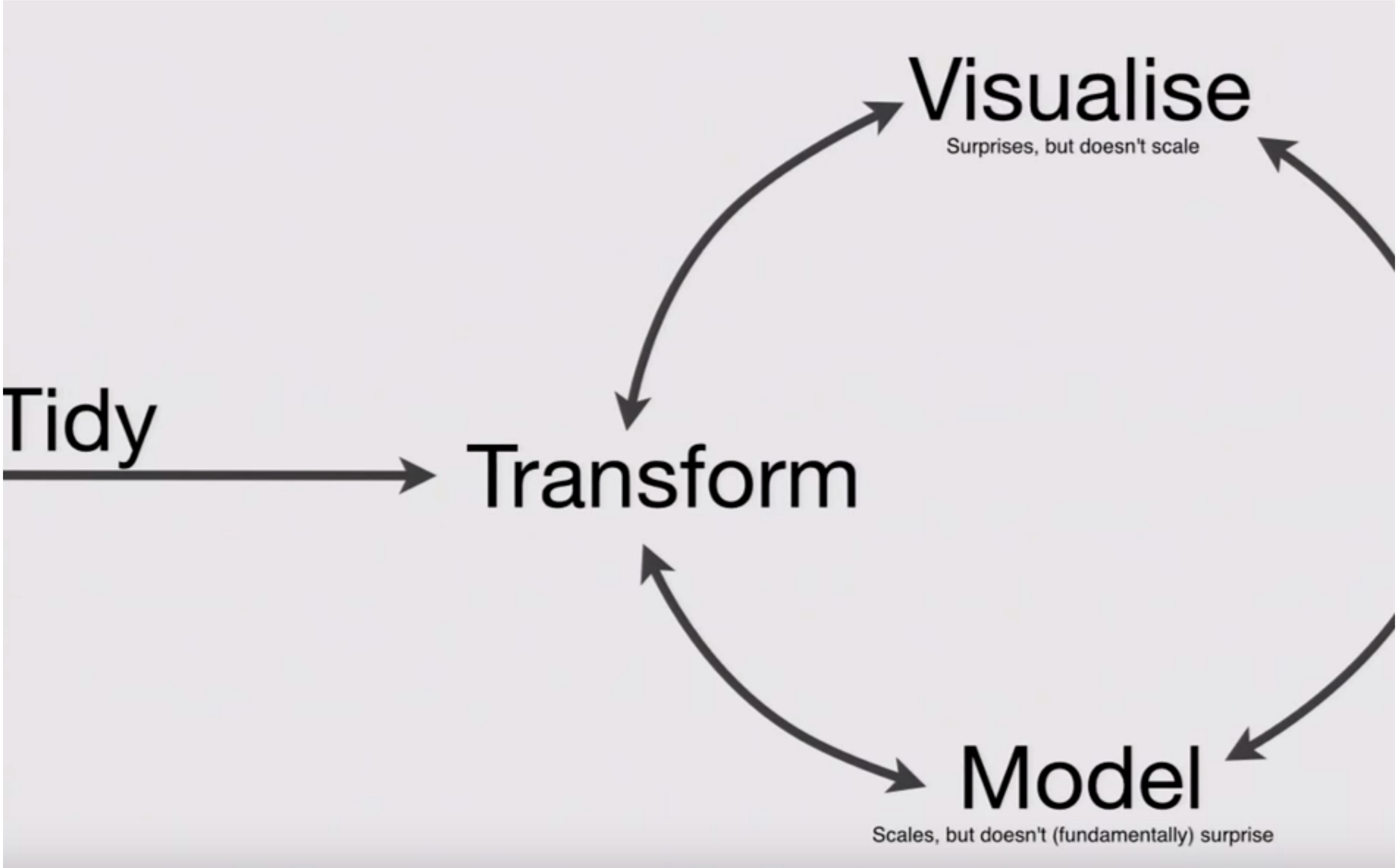
tile dimension as Db_s-b_e-z, where D is the binned data dimension, b_s represents the starting bin index, b_e represents the ending bin index, and z represents the zoom level.

Nanocubes

- http://nanocubes.net



Homework for next week



Visualise

Surprises, but doesn't scale

Tidy

Transform

Model

Scales, but doesn't (fundamentally) surprise