BIG DATA VISUALIZATION

Team Impossible

Peter Vilim, Sruthi Mayuram Krithivasan,

Matt Burrough, and Ismini Lourentzou

Let's begin with a story…

Dora the Data Explorer has a new job!

Dora's new job

Explore Yahoo's data

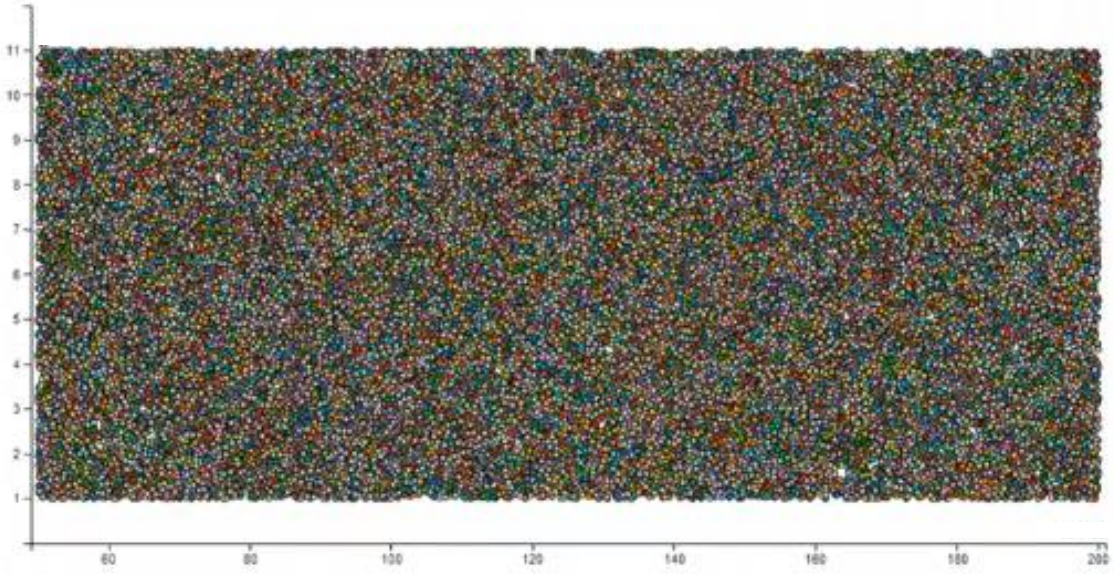
Yahoo has a ton of data though and they keep getting more

Hard to make sense of all the data

Dora runs into some challenges when working with this data

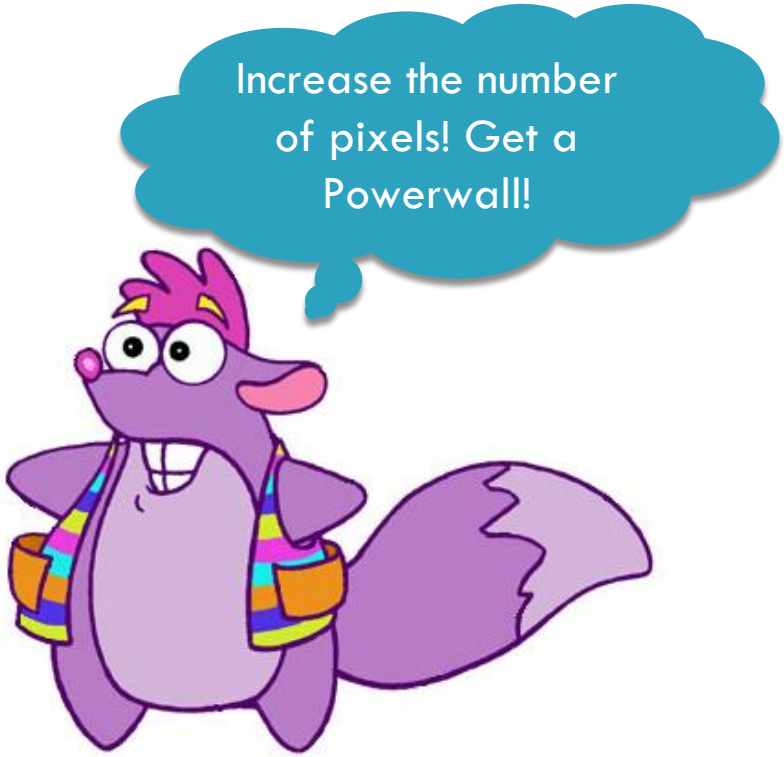
Screen resolution has its limitations!

40,000 data points from Yahoo data



1	Many data points

Tico the Squirrel has an idea!



53.7 million pixel Powerwall at the University of Leeds

Boots is laughing at how ineffective this is

- From the beginning of recorded time until 2003, humans had created 5 exabytes (5 billion gigabytes) of data.
- In 2011, the same amount was created every two days

YOU WILL SOON REACH A *NEW* LIMIT!

Besides….

Dora is too adventurous to stay in the office

But look at these tiny screens!

How do you gain insight from your data?

Even if you manage to fit all your data pixels on a screen

Humans don't think in terms of hundreds of numbers let alone tens of thousands

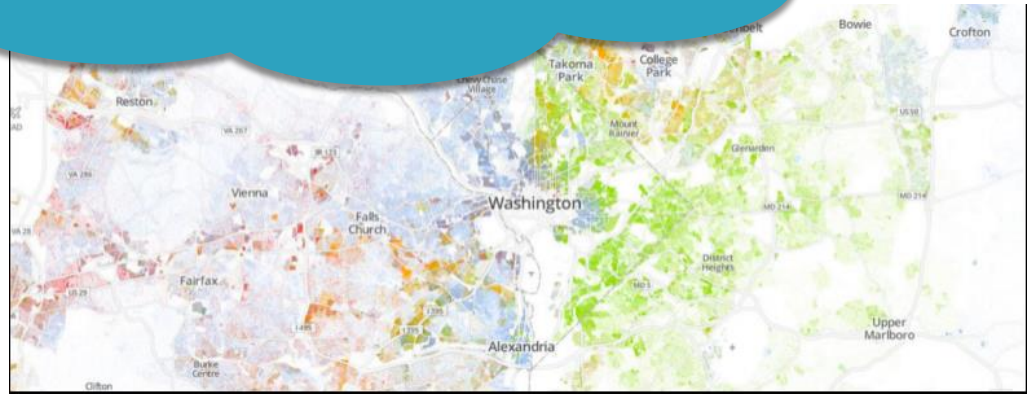
Can't draw insight from a large collection of points

We need to find ways to fully utilize/squeeze data points into the resolution we have…

And more importantly make sense out of our data as well!

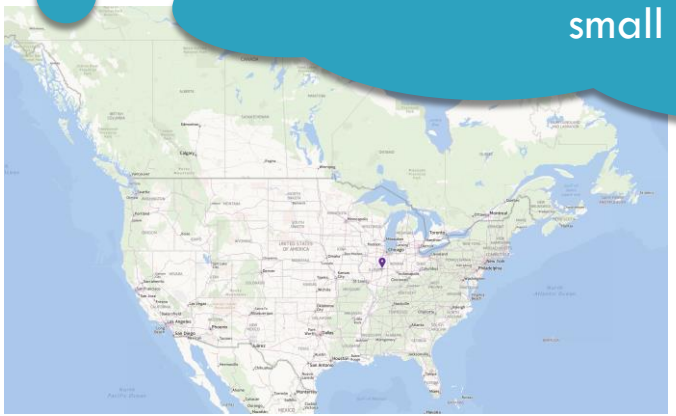
Boots is thinking…

How about zoom in to the points we are interested in? This is similar to how you navigate online maps.



Boots has an idea

How can Yahoo Maps show continents and also small islands? We need to be able to see the big picture and small outliers



2	Showing outliers

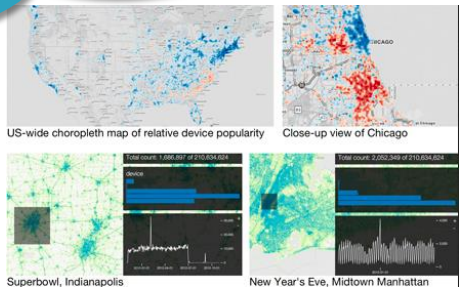
Boots is thinking…

But we have so many dimensions! We can't color everything differently! That won't work for all the dimensions we have.

3	Many dimensions

Hierarchies are a powerful way to explore

Yahoo Maps is so powerful! It can zoom in and out MULTIPLE times and does it FAST!
Can we do the same?



4	Interactive visualization

Can hierarchies help to better visualize big data?

Hierarchical Visualization Challenges

1	Many data points
Technique	
2	Showing outliers
3	Many dimensions
4	Interactive visualization
5	Integrating into working software

First Challenge

1	Many data points
Technique	Tree Maps
2	Showing outliers
3	Many dimensions
4	Interactive visualization
5	Integrating into working software

How Tree Maps Work

TreeMaps visualize **hierarchical structures** onto a rectangular region in a space-filling manner.

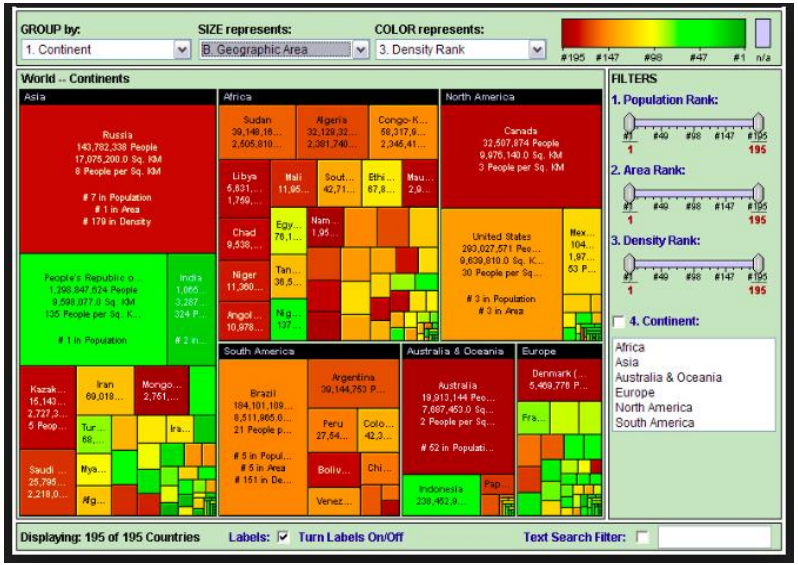
A node's weight (bounding box) determines its display size

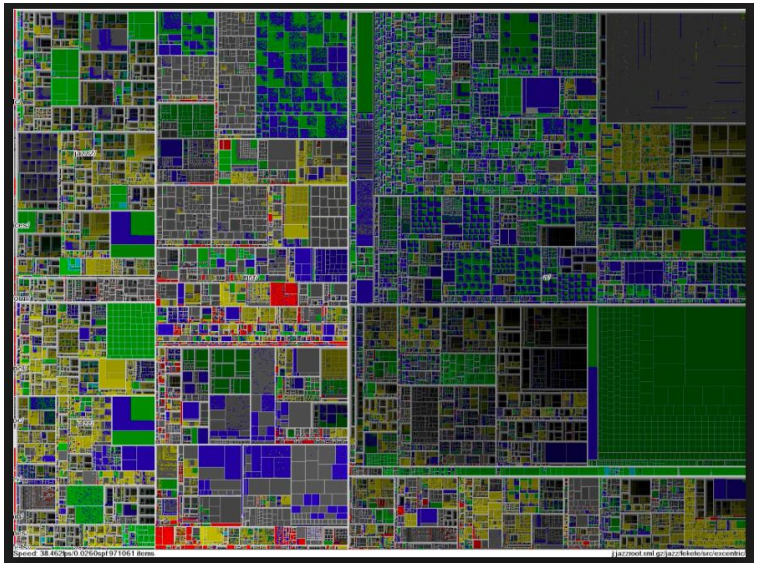
- Measure of importance or degree of interest

[1] Brian Johnson and Ben Shneiderman, Treemaps: a space-filling approach to the visualization of hierarchical information structures, 1991

Advantages and Disadvantages of Tree Maps

- ✓ 100% use of the available display space
- ✓ Allows users to set display properties (colors, borders, etc.)
- X Number and variety of domain properties visualized is limited
- X Cluttered





[1] Brian Johnson and Ben Shneiderman, Treemaps: a space-filling approach to the visualization of hierarchical information structures, 1991

And Dora is crying

I can't focus on one set of points only, now I am stuck with the whole dataset! And everything looks the same! Now I have to do this again for the a smaller set!

[1] Brian Johnson and Ben Shneiderman, Treemaps: a space-filling approach to the visualization of hierarchical information structures, 1991

Moving beyond just displaying data

Data exploration should maximize insight into a data set

Interactively

- ✓ Retrieve meaningful relations
- ✓ Extract inferences from the data
- ✓ Uncover underlying structure
- ✓ Detect patterns or anomalies
- ✓ Test underlying assumptions

Active process of discovery

NOT passive display

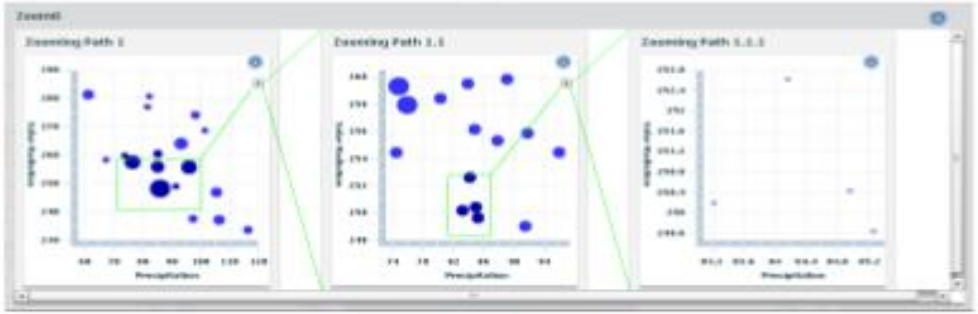
Second Challenge

1	Many data points
Technique	Tree Maps
2	Showing outliers
Technique	Zoom clustering
3	Many dimensions
4	Interactive visualization
5	Integrating into working software

How Zoom Clustering Works

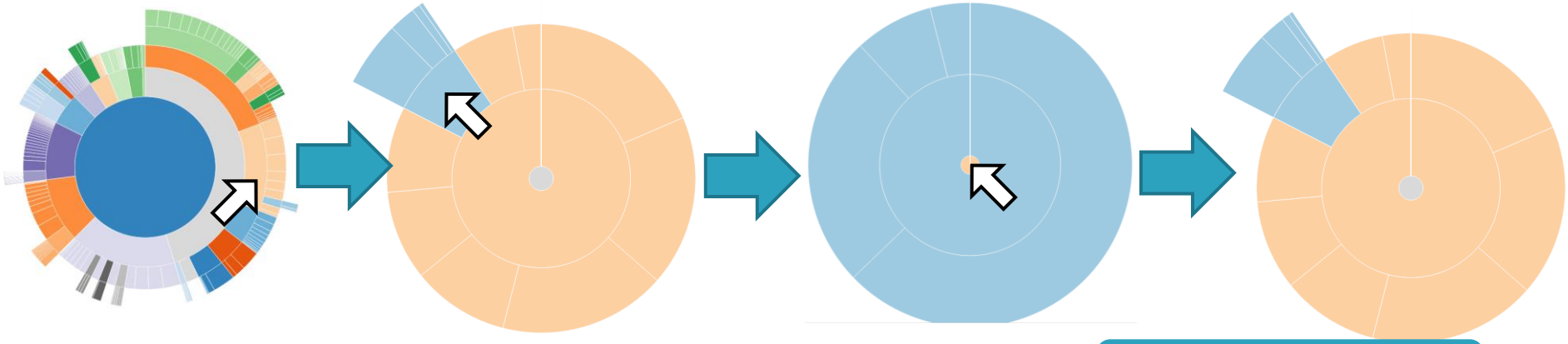
- Cluster data points
- Store clusters in tree structure
- Allow branching and zooming into different areas of tree
- Support back tracking in zoom tree

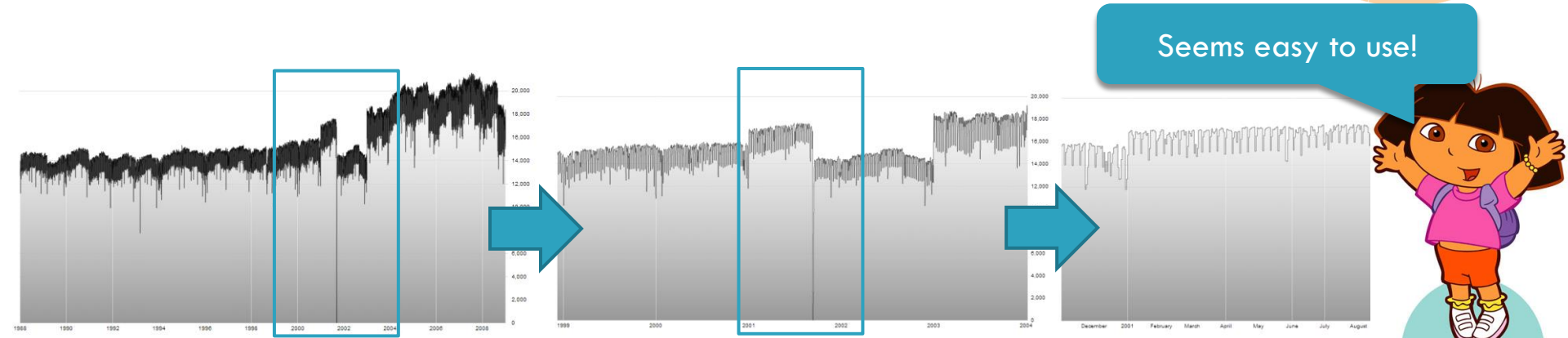
User zoom tree operation



[7] Baoyuan Wang, Gang Chen, Jiajun Bu & Yizhou Yu, Multiscale Visualization of Relational Databases Using Layered Zoom Trees and Partial Data Cubes, 2010

Zoom Tree Examples

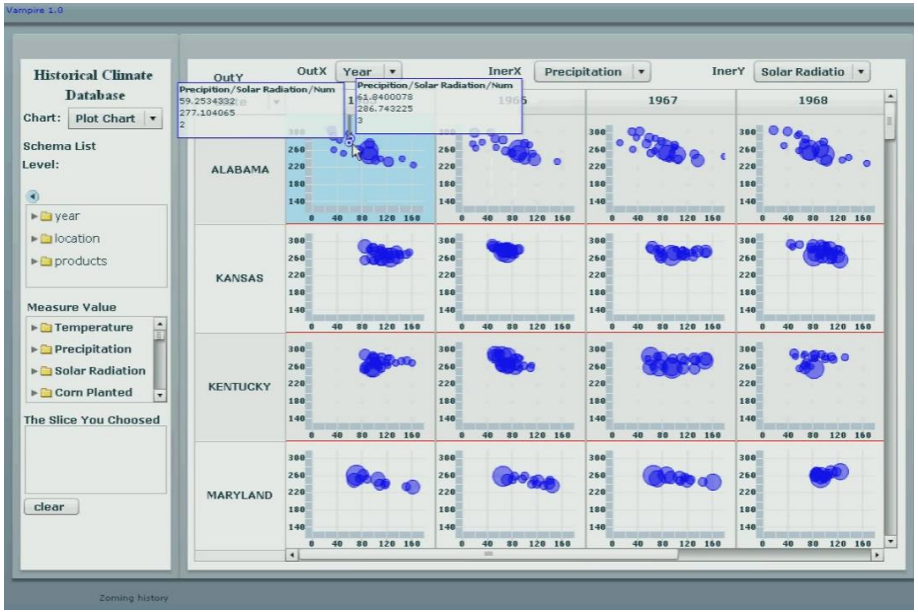




[12] Mike Bostock, Zoomable Sunburst, http://bl.ocks.org/mbostock/4348373, 2012; Mike Bostock, Zoomable Area, http://bl.ocks.org/mbostock/4015254, 2012.

Demo video: Zooming Plot Charts

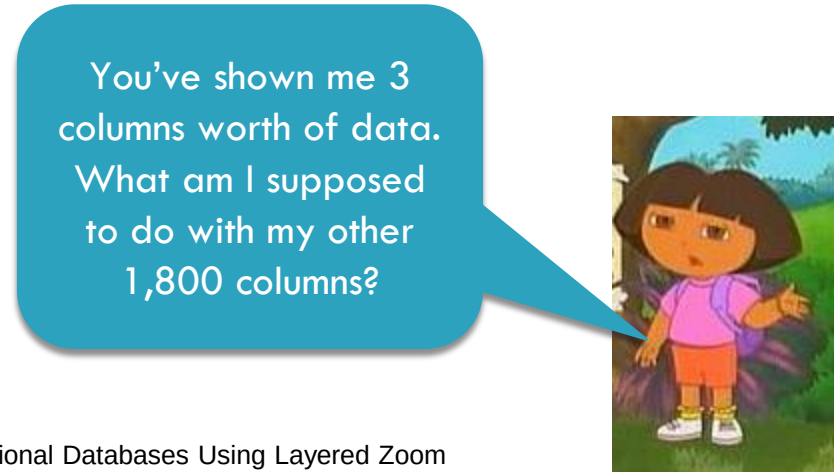
http://youtu.be/8dflke95xCM?t=3m52s



[7] Baoyuan Wang, Gang Chen, Jiajun Bu & Yizhou Yu, Multiscale Visualization of Relational Databases Using Layered Zoom Trees and Partial Data Cubes, 2010

Advantages and Disadvantages of Zoom Clustering

- Advantages
  - Not Overwhelmed with Data Points
  - Can Process Data in Parallel
  - Ability to Focus on Interesting Areas
- Disadvantages
  - Many-Dimensional Data

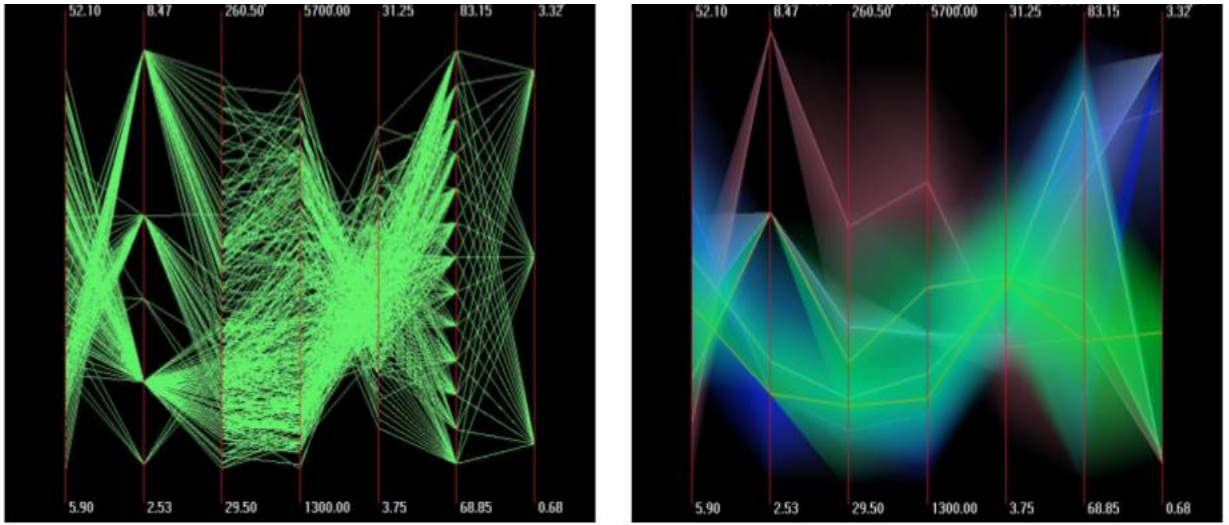


[7] Baoyuan Wang, Gang Chen, Jiajun Bu & Yizhou Yu, Multiscale Visualization of Relational Databases Using Layered Zoom Trees and Partial Data Cubes, 2010

Third Challenge

1	Many data points
Technique	Tree Maps
2	Showing outliers
Technique	Zoom clustering
3	Many dimensions
Technique	Parallel Coordinates
4	Interactive visualization
5	Integrating into working software

How Parallel Coordinates Work



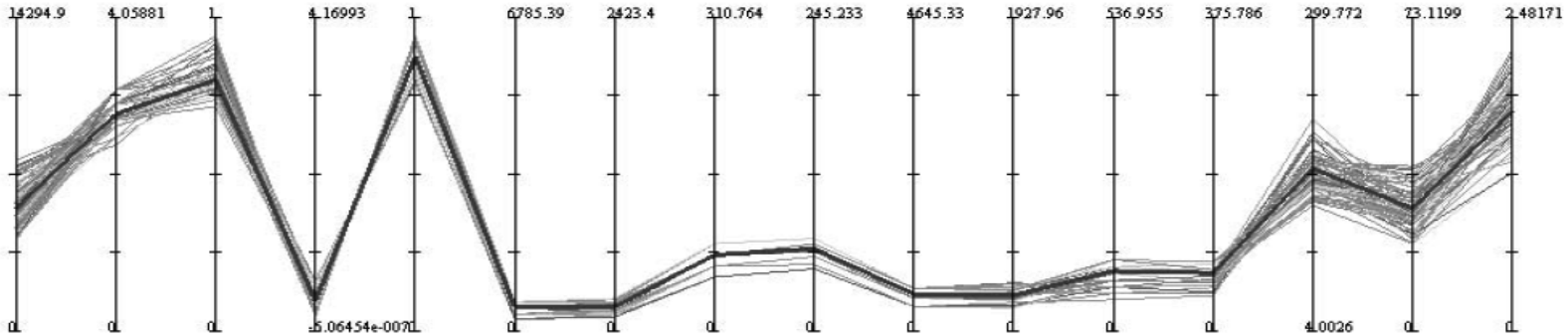
- ✓ Large number of dimensions
- ✓ Highlights relation between dimensions

Difficult to distinguish the overall structure when the number of tuples becomes very large

[9] Jimmy Johansson, Robert Treloar, and Mikael Jern, Integration of Unsupervised Clustering, Interaction and Parallel Coordinates for the Exploration of Large Multivariate Data, 2004

Self Organizing Map & Parallel Coordinates

SOM algorithm + Parallel Coordinates



Drill-down on the clusters reveals the original data elements and the weight vector.

Self Organizing Map:

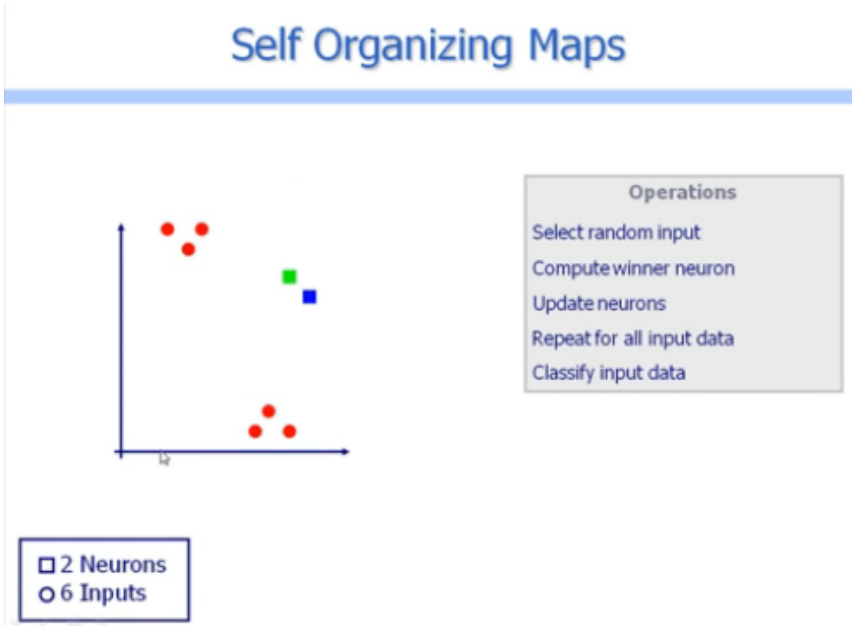
Nonlinear projection from m-dimensional space onto the two-dimensional display space.

Relies on distance, similarity and average.

[9] Jimmy Johansson, Robert Treloar, and Mikael Jern, Integration of Unsupervised Clustering, Interaction and Parallel Coordinates for the Exploration of Large Multivariate Data, 2004

Self Organizing Maps Explained

http://www.tubechop.com/watch/2698630

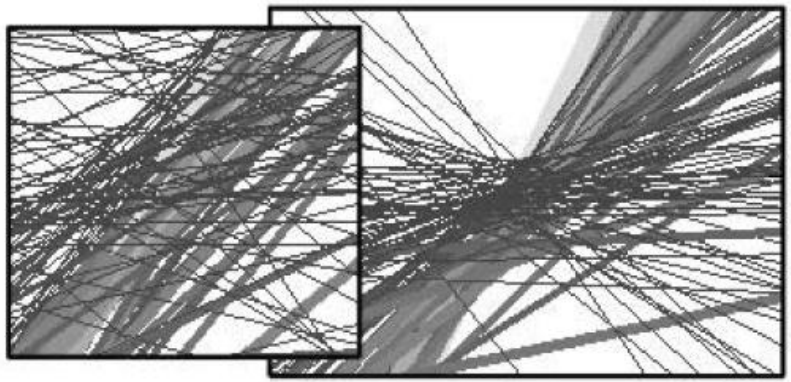


[9] Jimmy Johansson, Robert Treloar, and Mikael Jern, Integration of Unsupervised Clustering, Interaction and Parallel Coordinates for the Exploration of Large Multivariate Data, 2004\

Zooming in Parallel Coordinates

Interactive data analysis:

- Visual User Interface with drill down, filtering and zooming
- Zoom in by simply drawing a rectangle across the selected cluster bands.



Using the rectangle zoom for a more detailed view.

[9] Jimmy Johansson, Robert Treloar, and Mikael Jern, Integration of Unsupervised Clustering, Interaction and Parallel Coordinates for the Exploration of Large Multivariate Data, 2004

Advantages and Disadvantages of Parallel Coordinates

Cool machine learning method!

BUT…..

- X Parallel Coordinates does not provide a good overview as it becomes hard to see the structure in the data when the dataset gets large
- X Runs out of encoding possibilities as the number of dimensions increases.
- X Preprocessing or filtering the data is required
- X Not efficient for visualizing datasets with non-numerical data
- X Number of clusters/neurons predefined

[9] Jimmy Johansson, Robert Treloar, and Mikael Jern, Integration of Unsupervised Clustering, Interaction and Parallel Coordinates for the Exploration of Large Multivariate Data, 2004\

Fourth Challenge

1	Many data points
Technique	Tree Maps
2	Showing outliers
Technique	Zoom clustering
3	Many dimensions
Technique	Parallel Coordinates
4	Interactive visualization
Technique	Parallel sampling
5	Integrating into working software

Dora is frustrated!

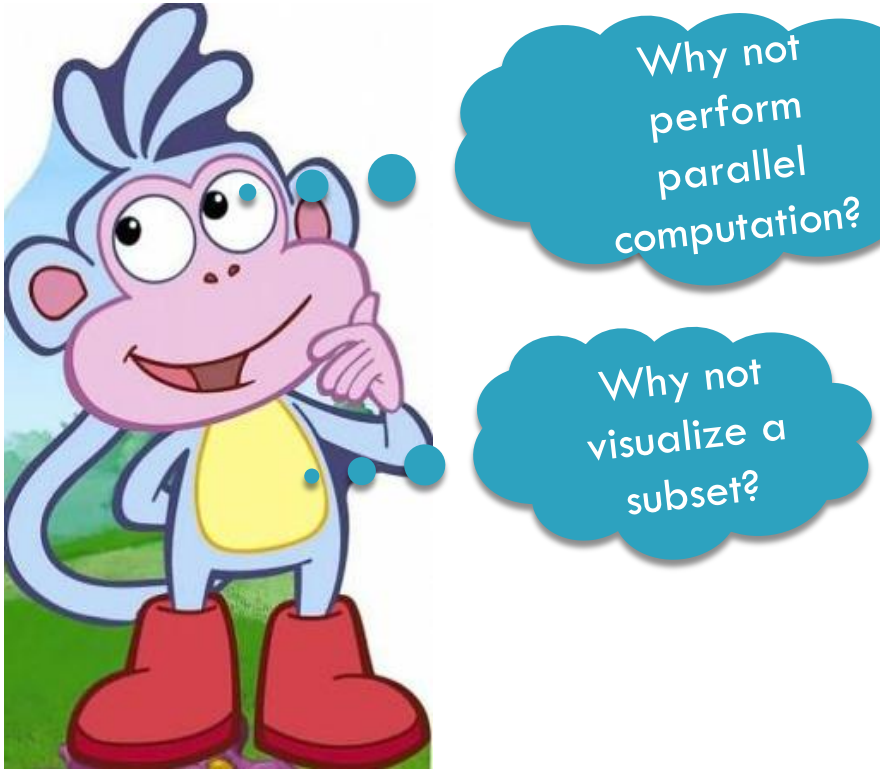
I can't wait this long!

I want to access the remote server faster.

How do I quickly visualize hierarchical data?

Which visualization approach is cost-effective?

Boots has some ideas about speed….



1. Subsampling and clustering

a) Can be very fast

b) Need to account for errors

c) Grid and multiresolution

2. Parallel Servers

a) Distributed computation (e.g. MapReduce, Spark, etc.)

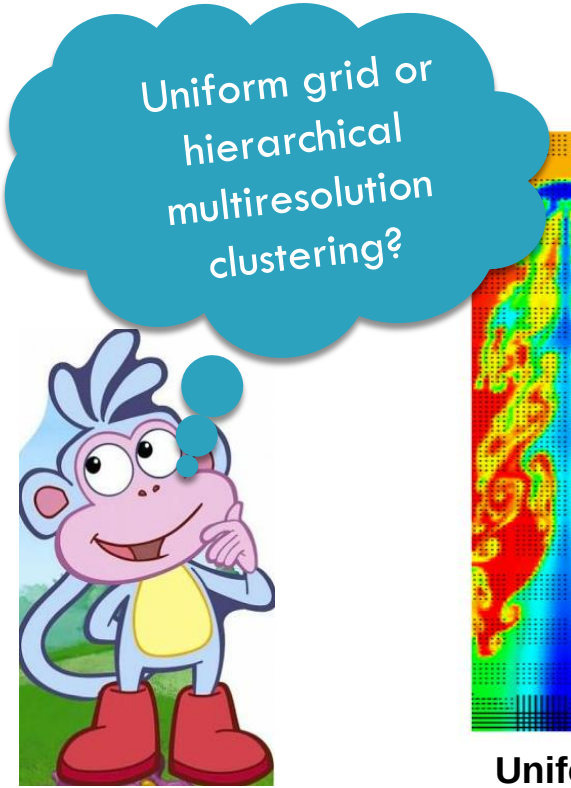
b) Don't have to account for errors

c) Aren't restricted to certain data types

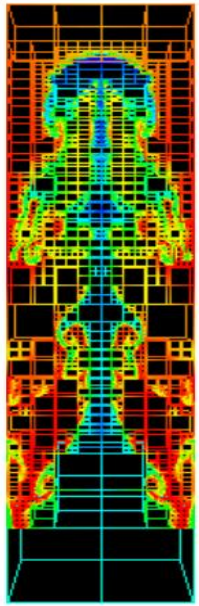
d) Technically challenging to implement

[4] Lori A. Freitag and Raymond M. Loy, Comparison of Remote Visualization Strategies for Interactive Exploration of Large Data Sets, 2001

Boots thinks sampling is a good idea…



Uniform



Hierarchical

1. **Uniform Grid**
   - Distributes points from original dataset into equal sized grids
   - Single level representation
   - Can be constructed quickly

2. **Hierarchical Multiresolution**
   - Multi level representation
   - Fewer approximation errors by showing more points where the data is changing rapidly
   - Quadratic complexity makes large problems difficult

[4] Lori A. Freitag and Raymond M. Loy, Comparison of Remote Visualization Strategies for Interactive Exploration of Large Data Sets, 2001

Isa comes to rescue Dora and Boots with another idea

Why don't we combine hierarchical clustering with parallel servers?

[4] Lori A. Freitag and Raymond M. Loy, Comparison of Remote Visualization Strategies for Interactive Exploration of Large Data Sets, 2001

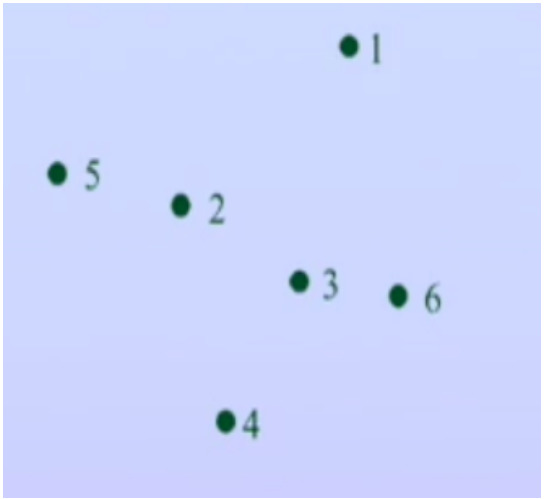
Data Explorers are excited about PINK

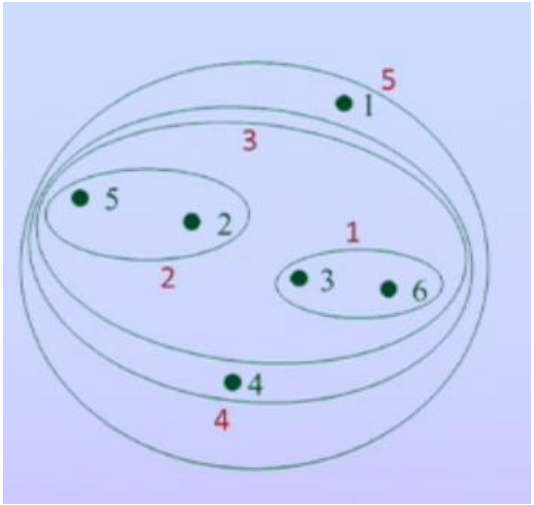
Why PINK (Parallel Single Linkage)?

- Scalable parallel algorithm for single-linkage hierarchical clustering
- Structure of single linkage problem can be exploited for parallelism
- Single linkage hierarchical clustering dendrogram for a dataset and the MST of the corresponding complete graph produce identical clusters

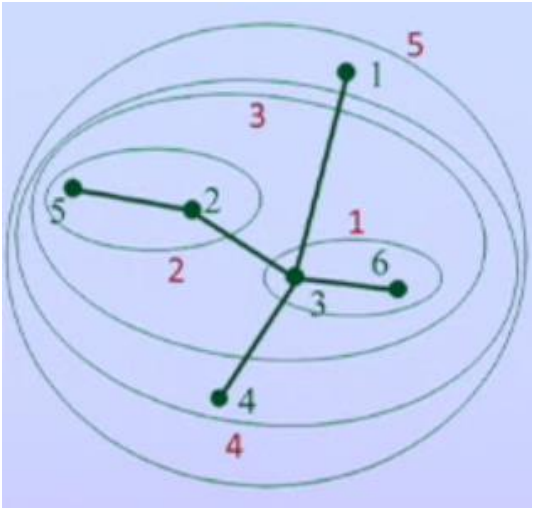
[4] Lori A. Freitag and Raymond M. Loy, Comparison of Remote Visualization Strategies for Interactive Exploration of Large Data Sets, 2001

Connection to Minimum Spanning Tree





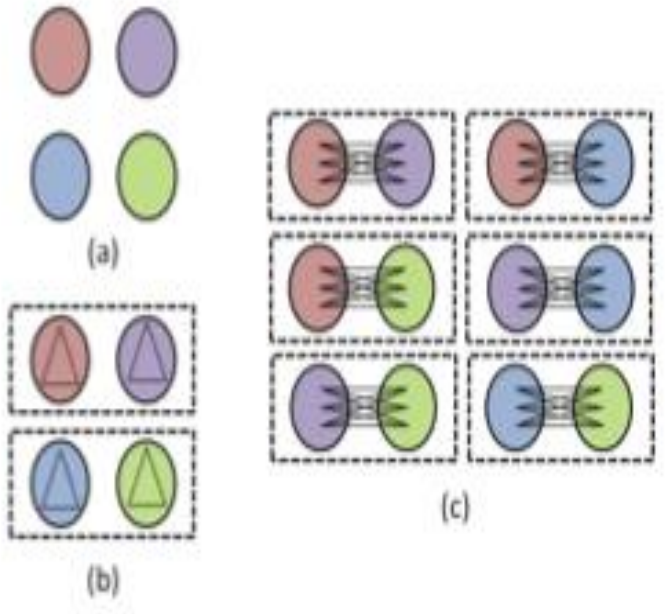
Single Linkage
Hierarchical Clustering



Minimum Spanning Tree

[4] Lori A. Freitag and Raymond M. Loy, Comparison of Remote Visualization Strategies for Interactive Exploration of Large Data Sets, 2001

How does PINK work?



Split Data Evenly

(a) Problem domain decomposition with *k* partitions

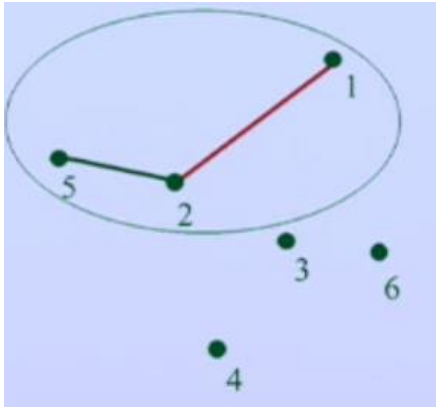
(b) Two processes are each assigned two complete subgraphs

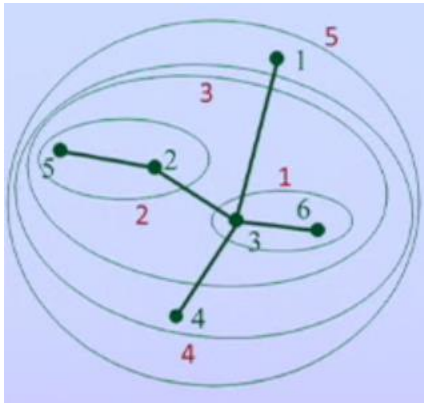
(c) Six processes are assigned one bipartite subgraph for the six pairs of partitions

(d) $K^2/2$ processors

[12] William Hendrix, Diana Palsetia, Mostofa Patwary, and Ankit Agrawal, A Scalable Algorithm for Single-Linkage Hierarchical Clustering on Distributed-Memory Architectures, 2013

Generating the Minimum Spanning Tree

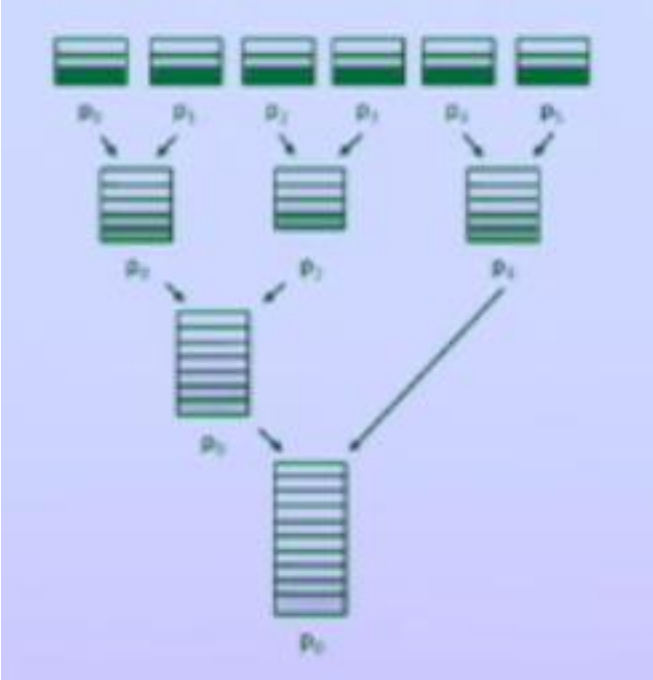




- Solve subproblems using prim's algorithm
- Combine partial solutions
- Subproblems may have edges not in MST
- Treat partial solutions as candidate edges
- Apply Kruskal's algorithm to candidate edges

[12] William Hendrix, Diana Palsetia, Mostofa Patwary, and Ankit Agrawal, A Scalable Algorithm for Single-Linkage Hierarchical Clustering on Distributed-Memory Architectures, 2013

Binary Merging of Partial Solutions

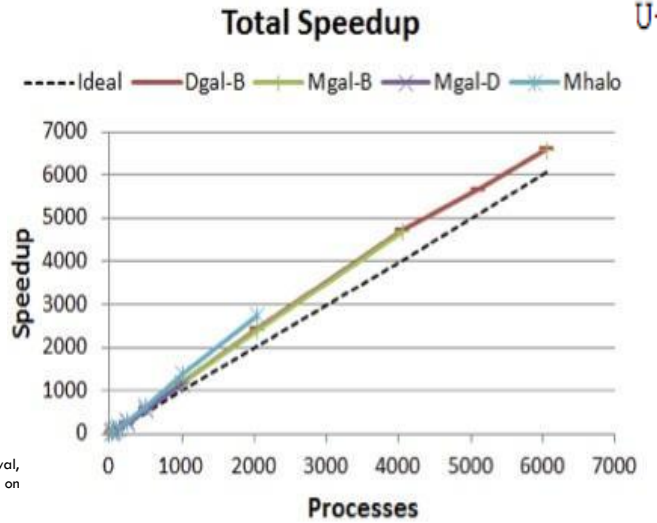


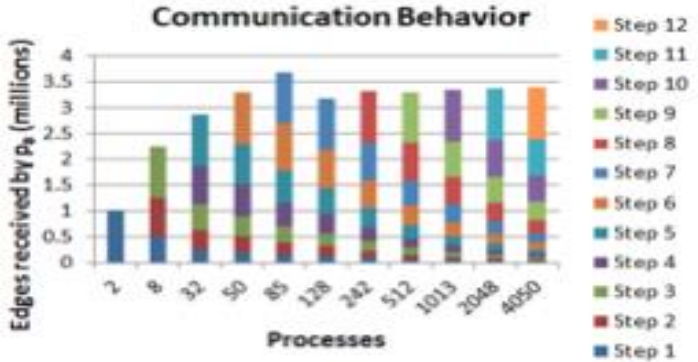
- Combine two MSTs at a time from consecutive processors
- Add an edge that does not join vertices that are already in the same component

Explorers are happy with PINK's performance

Memory Usage

Data	1 proc	50	512	8192	DM
U-100k-10	12.6	6.18	5.08	4.70	37.3 GB
U-500k-10	62.9	30.9	25.4	23.5	931 GB
U-1M-10	126	61.8	50.8	47.0	3.64 TB
U-1M-20	202	77.1	55.6	48.2	3.64 TB





[12] William Hendrix, Diana Palsetia, Mostofa Patwary, and Ankit Agrawal, A Scalable Algorithm for Single-Linkage Hierarchical Clustering on Distributed-Memory Architectures, 2013

ABC

- Why does PINK combine the dendrograms from consecutive processes?
  - Overlapping data partitions
  - Detect and eliminate edges sooner
  - Cuts down memory and communication cost
- What are the limitations of PINK algorithm?
  - Minimum processor requirement ($K^2/2$)
  - Binary Merge – the entire dendrogram must fit in one processor

Fifth Challenge

1	Many data points
Technique	Tree Maps
2	Showing outliers
	Zoom clustering
3	Many dimensions
	Parallel Coordinates
4	Interactive visualization
	Parallel sampling
5	Integrating into working software
Technique	????

Open Challenges

1. We have the back end of big data but no front end

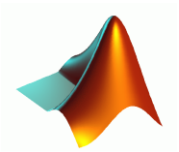

Google Dremel


2. Most big data tools are focused on batch processing which isn't good for visualization


This is changing

Open Challenges

3. Most front end tools don't integrate with the back end tools



This is improving

4. Most front end tools don't handle this type of data well (high dimensionality and many data points)

Comparison of Hierarchical Techniques

	Many data points 1	Showing outliers 2	Many dimensions 3	Interactive visualization 4	Working software 5
Tree Maps	●	○	◒	◒	◒
Parallel Coordinates	◒	◒	●	○	○
Zoom Clustering	◒	●	●	◒	◒
Parallel Sampling	N/A	N/A	N/A	●	N/A

How can we combine these?

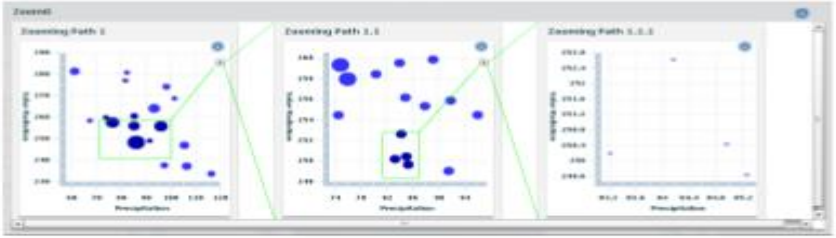
Can Dora use hierarchies to visualize big data?

Yes! With the right combination of techniques

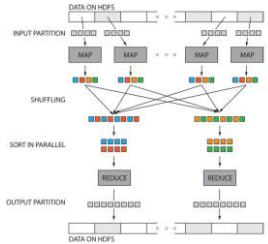
Our Hybrid Opinion

Reviewed several techniques with different advantages and disadvantages

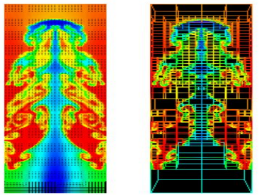
Use hierarchical clustering to tackle many data points



Use parallelization for performance



Use sampling because simple parallelization isn't enough



This also happens to be approach we picked for our project!