Big Data Visual Analytics

Jean-Daniel Fekete
AVIZ/INRIA
Jean-Daniel.Fekete@inria.fr
http://www.aviz.fr/~fekete

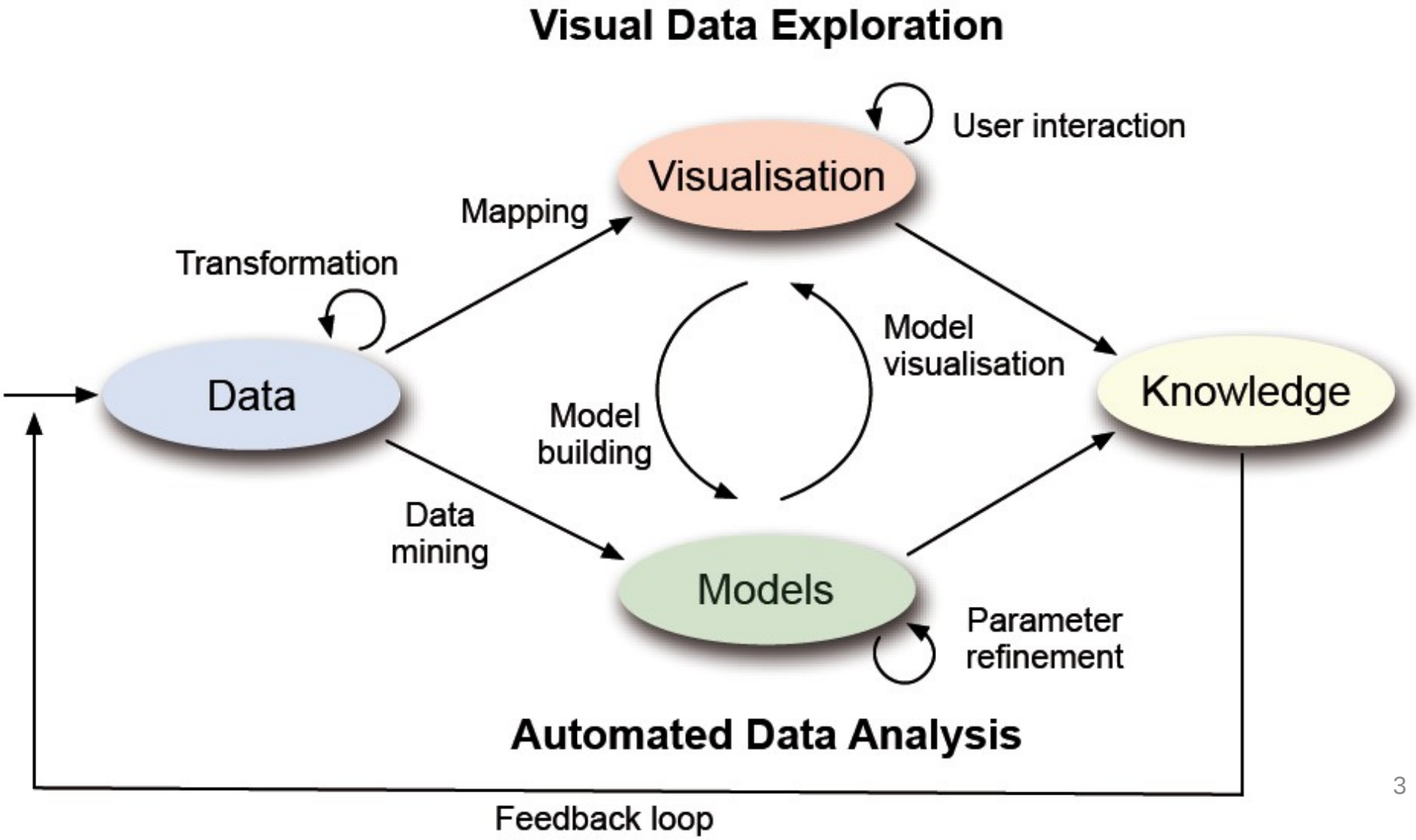
Big Data

- Volume
  - Like “really big”, has evolved with time from Tb to Pb
- Variety
  - Many types, e.g. text, image, tables
- Velocity
  - Acquisition/input speed, output speed
- Variability, Veracity...Vatever

- Traditionally used with *predictive analytics*

The Visual Analytics Process

- D. A. Keim, J. Kohlhammer, G. Ellis and F. Mansmann. Mastering The Information Age - Solving Problems with Visual Analytics. Eurographics, 2010.



Exploration and Latency

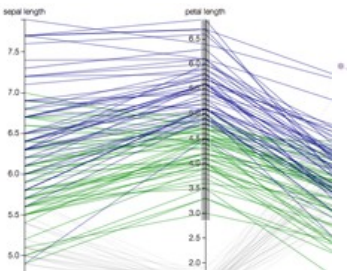
3 types of latency to consider for HCI:

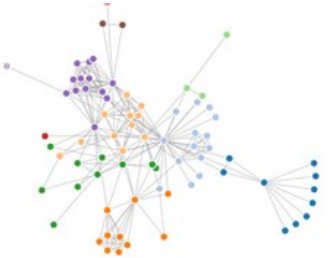
1. *Continuity Preserving Latency*: ~0.1s user feel that the system is reacting instantaneously
2. *Flow Preserving Latency*: ~1s user's flow of thought to stay uninterrupted
3. *Attention Preserving Latency*: ~10s keeping the user's attention focused on the dialogue

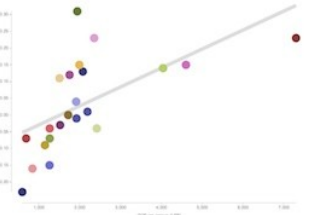
- R. B. Miller. Response time in man-computer conversational transactions. In Proceedings of the December 9-11, 1968, Fall Joint Computer Conference, Part I, AFIPS '68 (Fall, part I), pages 267–277, New York, NY, USA, 1968. ACM.
- J. Nielsen. Response times: The 3 important limits, https://www.nngroup.com/articles/response-times-3-important-limits/
- B. Shneiderman. Response time and display rate in human performance with computers. ACM Comput. Surv., 16(3):265–285, Sept. 1984.

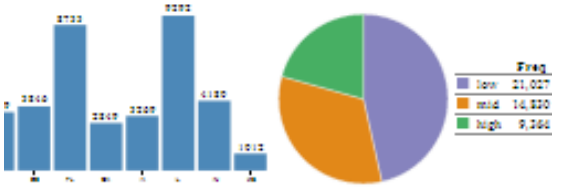
Scaling Visualization

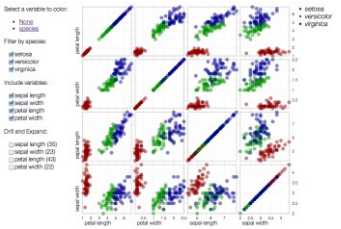
- Vis. does not scale well
  - Not in number of items
  - Not in number of dimensions
- It needs additional methods such as:
  - Sampling (of items/dim.)
  - Aggregation
  - Dimensionality Reduction
- These methods introduce artifacts
  - Their results should be explored too, to be validated!

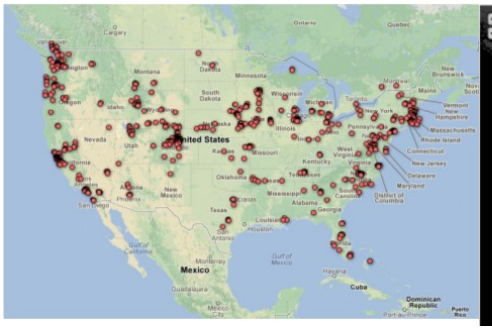


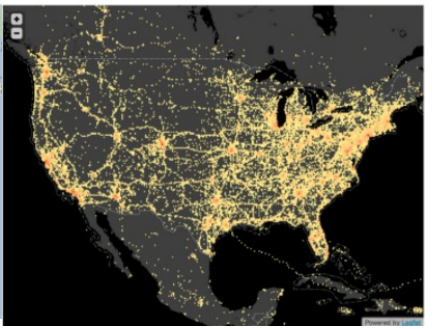








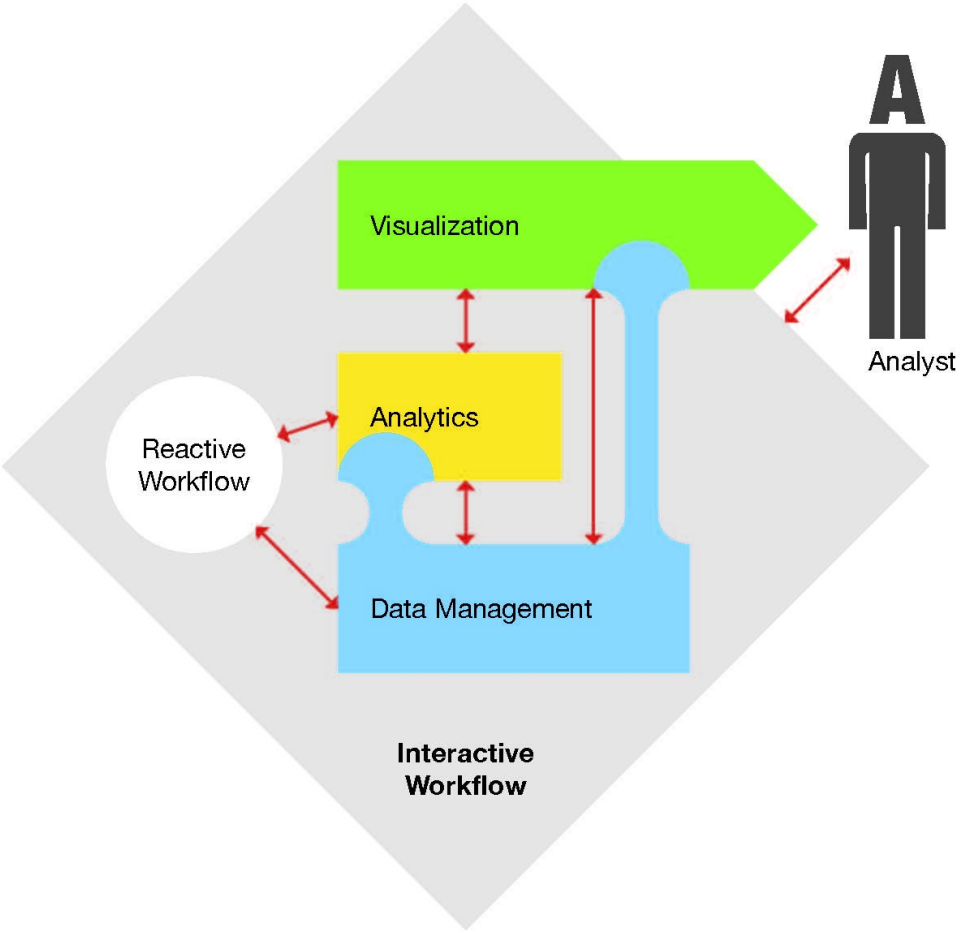




Layers of Visual Analytics

Three Layers:

- Data Management
- Analytics
- Visualization+ Interaction



Examples

- Hierarchical Clustering Explorer
- WikiReactive
- HAL Deduplication Framework
- Real-time sentiment analysis
- Nanocubes
- Progressive tSNE

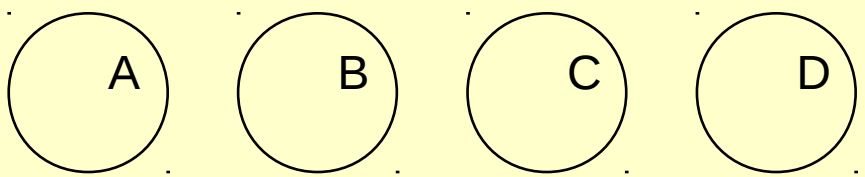
Hierarchical Clustering Explorer (Seoh & Shneiderman 2002)

http://www.cs.umd.edu/hcil/hce/

- Data
  - Multidimensional (n numerical dimensions)
- Task
  - Find clusters that clearly reflect properties in the data
- Volume: In memory
- Variety: none
- Velocity: none

Hierarchical Clustering

Initial Data Items

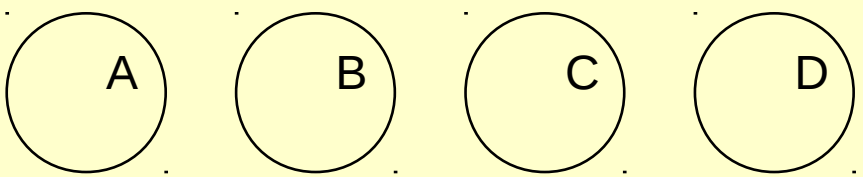


Distance Matrix

Dist	A	B	C	D
A		20	7	2
B			10	25
C				3
D				

Hierarchical Clustering

Initial Data Items



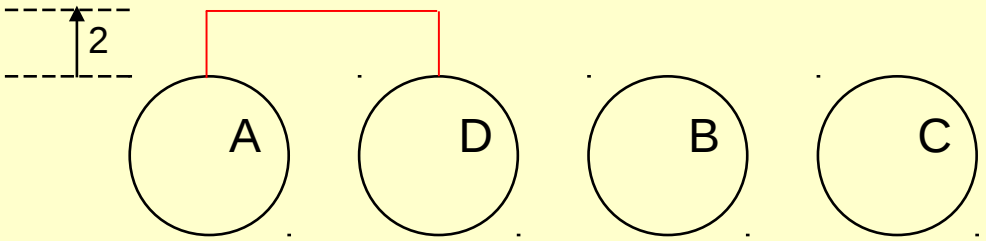
Distance Matrix

Dist	A	B	C	D
A		20	7	2
B			10	25
C				3
D				

Hierarchical Clustering

Single Linkage

Current Clusters



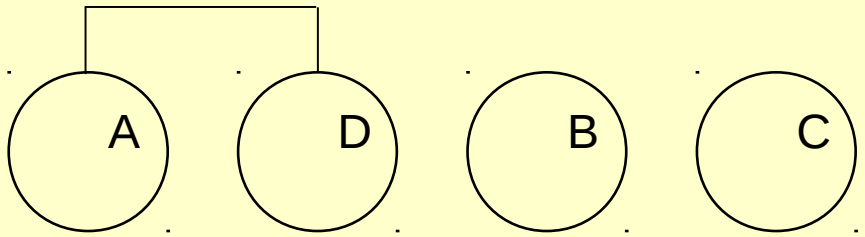
Distance Matrix

Dist	A	B	C	D
A		20	7	2
B			10	25
C				3
D				

Hierarchical Clustering

Single Linkage

Current Clusters



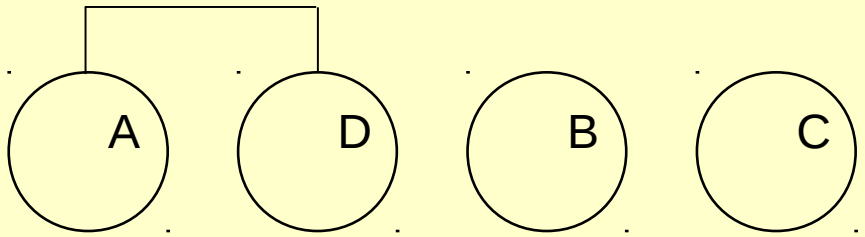
Distance Matrix

Dist	AD	B	C	
AD		20	3	
B			10	
C				

Hierarchical Clustering

Single Linkage

Current Clusters



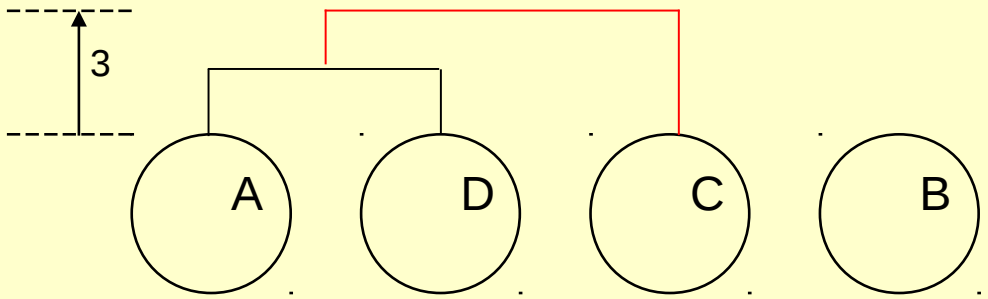
Distance Matrix

Dist	AD	B	C	
AD		20	3	
B			10	
C				

Hierarchical Clustering

Single Linkage

Current Clusters



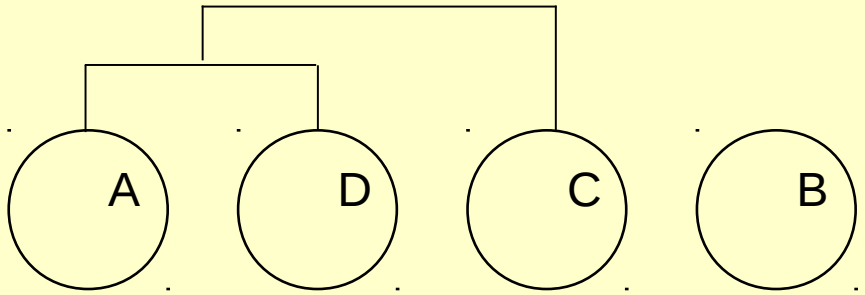
Distance Matrix

Dist	AD	B	C	
AD		20	3	
B			10	
C				

Hierarchical Clustering

Single Linkage

Current Clusters



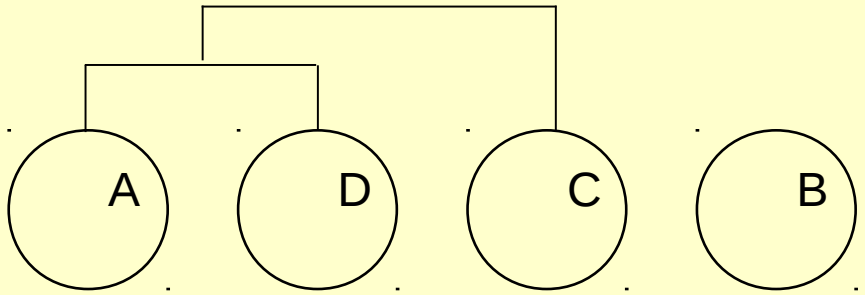
Distance Matrix

Dist	ADC	B		
ADC		10		
B				

Hierarchical Clustering

Single Linkage

Current Clusters



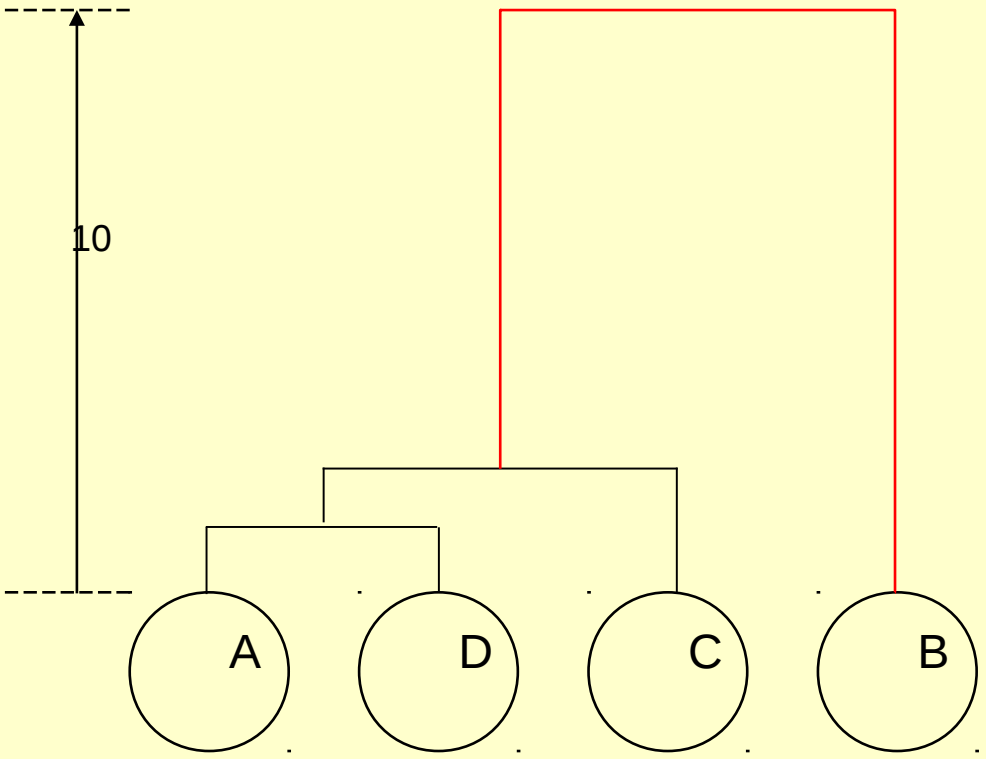
Distance Matrix

Dist	ADC	B		
ADC		10		
B				

Hierarchical Clustering

Single Linkage

Current Clusters



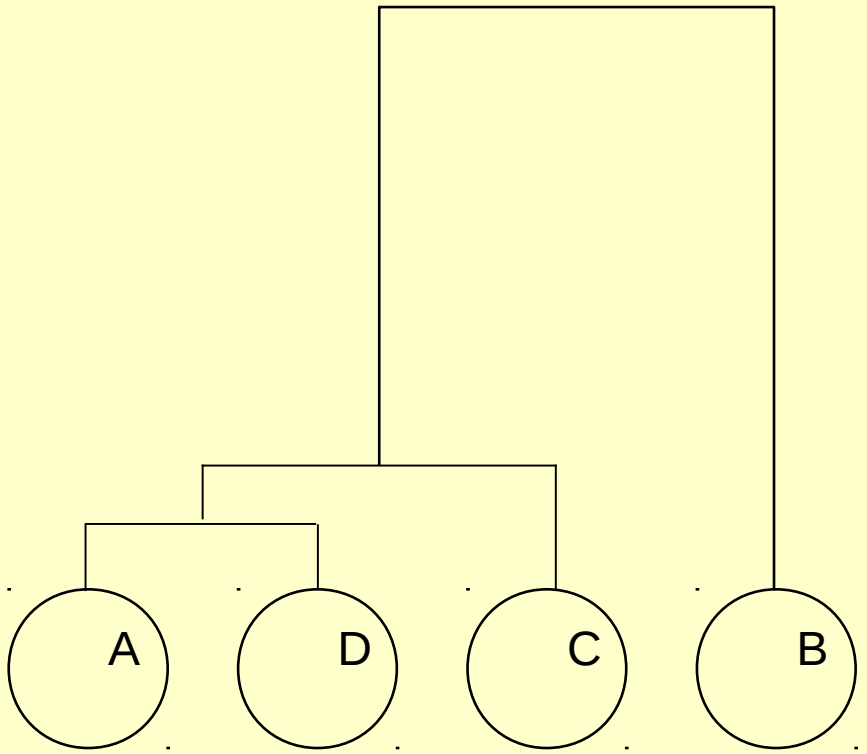
Distance Matrix

Dist	ADC	B		
ADC		10		
B				

Hierarchical Clustering

Single Linkage

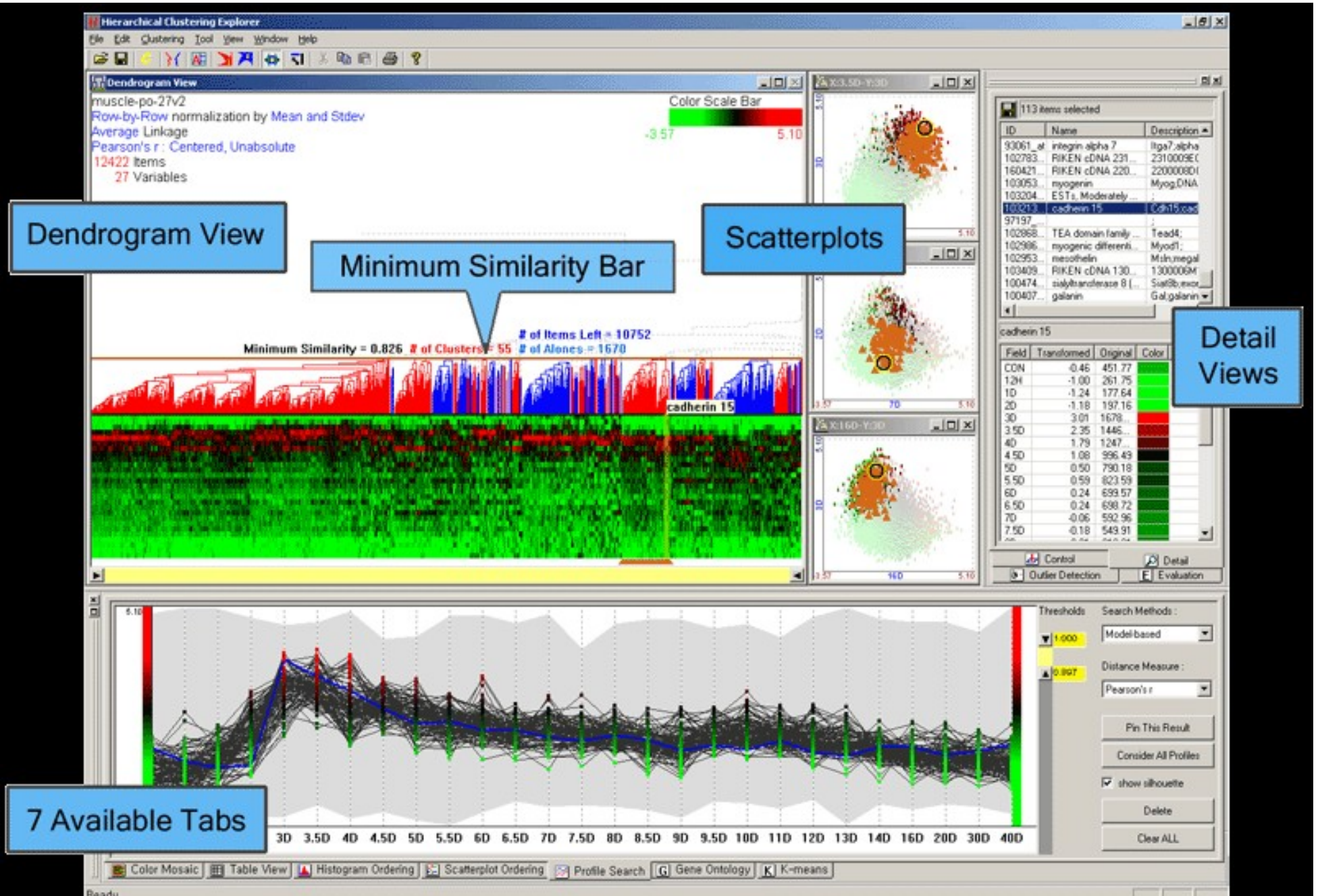
Final Result



Distance Matrix

Dist	ADC B			
ADC B				

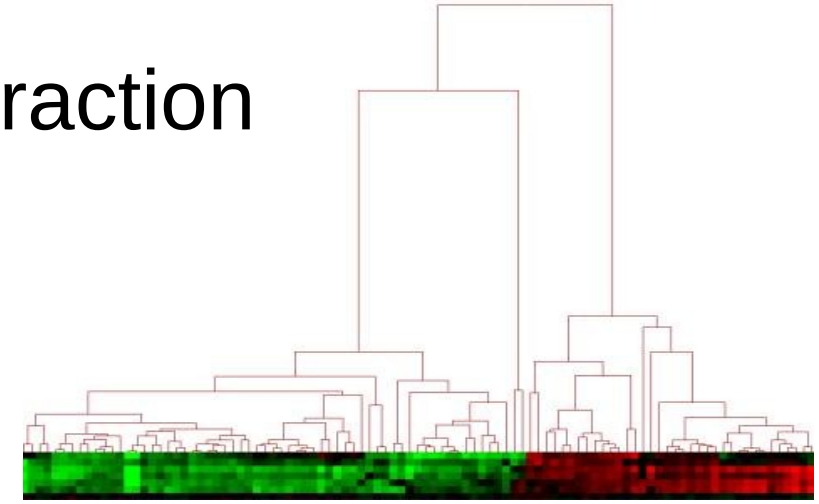
Hierarchical Clustering Explorer

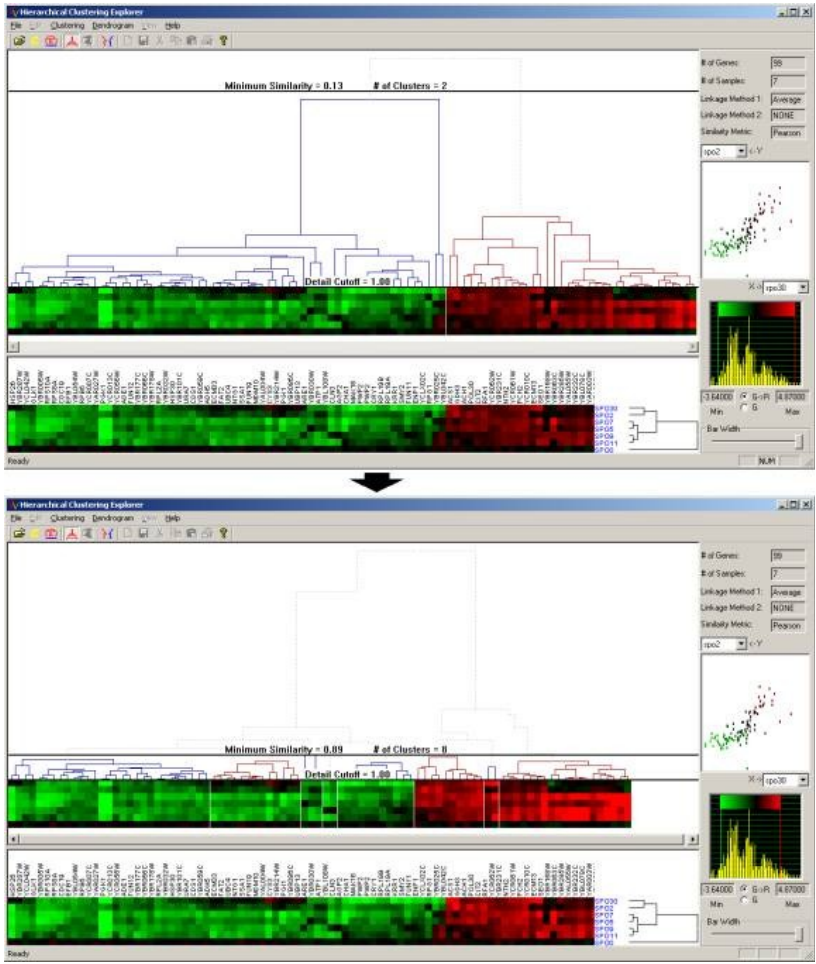


http://www.cs.umd.edu/hcil/hce/

Hierarchical Clustering and Interaction

- How many cluster?
  - Many criteria
- Explore interactively
  - Vary height (distance)
  - Vary number of clusters
  - Vary distance function
- Are they good in the end
  - Many way to assess but linear!





What if the number of vectors increases?

- From 1,600 to 10,000?
  - ~100,000,000 entries for the distance matrix
  - Memory and Computation still OK
- From 10,000 to 100,000,000?
  - Memory and Computation not OK

What if the number of vectors increases?

- 100,000,000 vectors could fit in memory?
- The distance matrix cannot fit in memory
  - It will take hours to compute
- Interaction is not possible any more

- What can we do about it?

Strategies to cope with Big Data and Visual Analytics

1) Increase the memory?
2) Use a distributed systems?
3) Use a parallel system (HPC)?
4) Use tricks?

Increase the memory for Big Data Visual Analytics

- How much?
  - Say for n = 1,000,000 (10^6)
- Dataset + distance matrix + hierarchy
- Memory = ???
- Time to compute the distance matrix?
  - Assume 10^6 operations per second

- Conclusion?

Use a distributed system

- How many machines to perform the computation quickly?
  - Say 10s
- Distributed system have a high latency
  - Usually > 10s, around 30s to minutes
- Not good for interaction
- But can compute results ahead of time

Use a Parallel System (HPC)

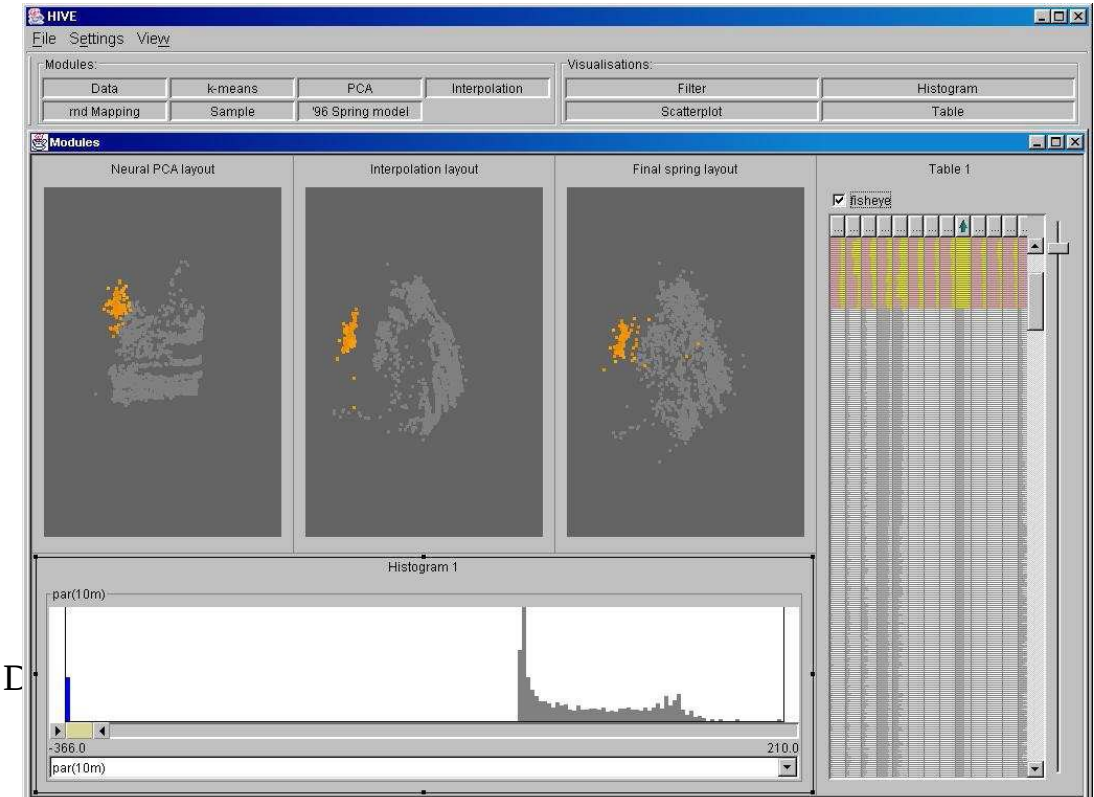
- Much more expensive than a distributed system
  - But faster
- Do you really need a special architecture?

Use Tricks: Hybrid Algo.

- Clustering a huge dataset?
- HC is quadratic: not possible
- K-Means is linear but requires a good K
- Sample -> HC -> Estimate good K -> k-Means
- Need a good sampling

Ross, G. and Chalmers, M. (2003) A visual workspace for constructing hybrid MDS algorithms and coordinating multiple views. Information Visualization, 2 (4). pp. 247-257.

Does not work well for Text mining



Big Data Visual Analytics

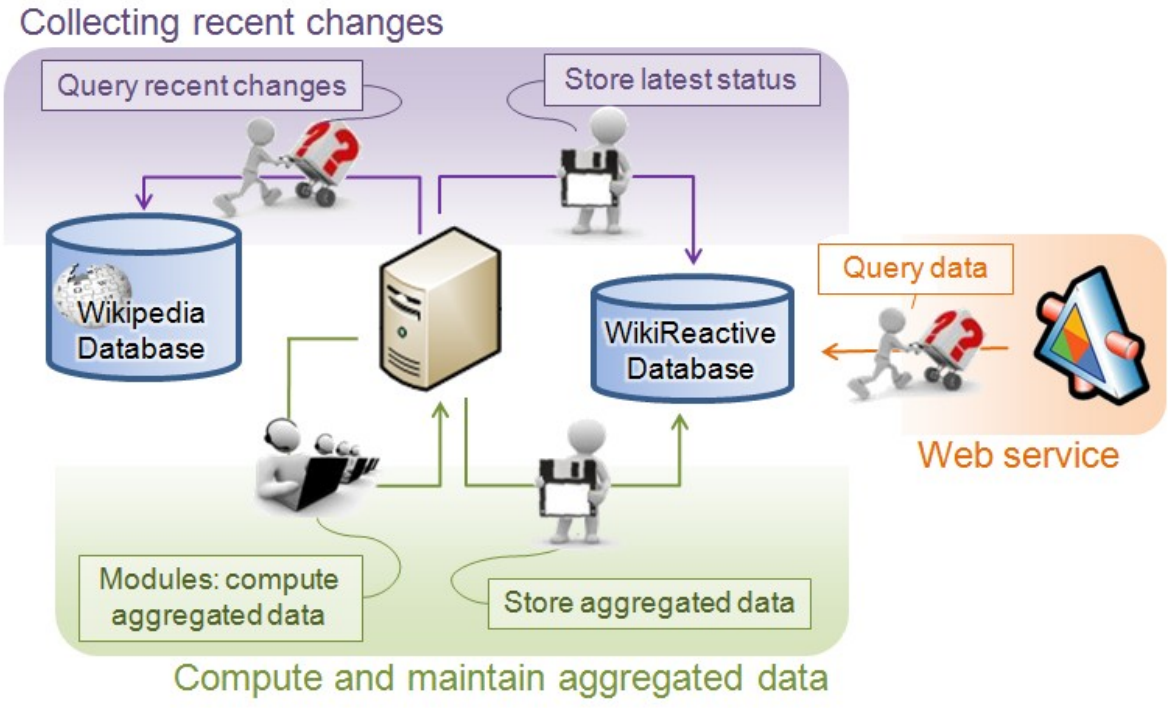
- 3 situations according to Hardley Wickham https://peadarcoyle.wordpress.com/2015/08/02/interview-with-a-data-scientist-hadley-wickham/
- When data does not fit in memory (1TB):

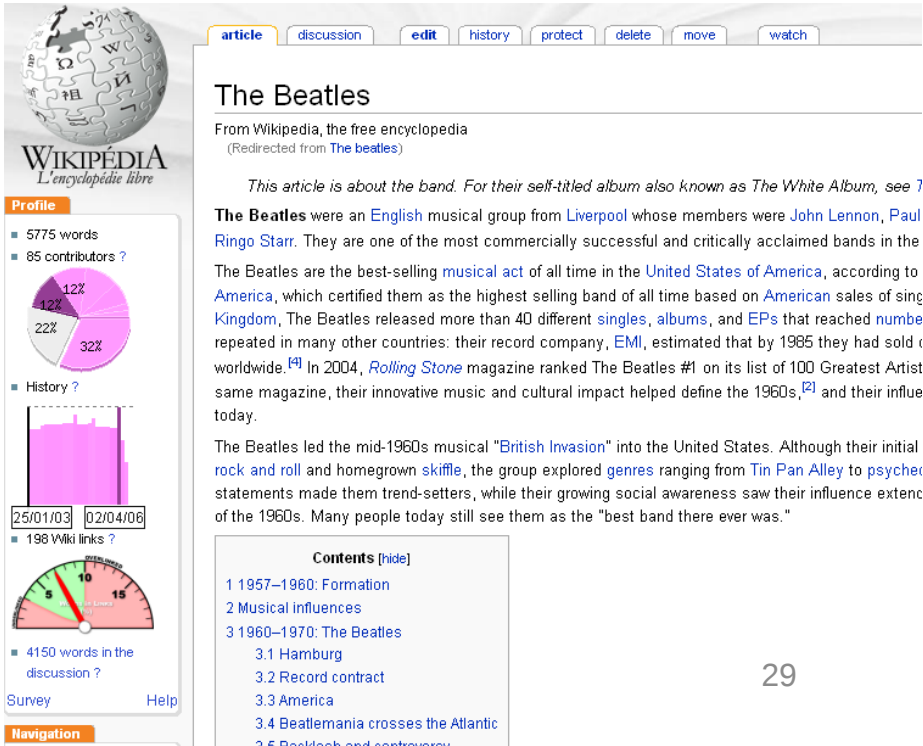
1) Data can be filtered/selected to become small
   - actually small data problems, once you have the right subset/sample/summary
2) Analysis can be split into independent chunks
   - actually lots and lots of small data problems
3) Don't know how to filter/split, hard case!
   - irretrievably big
   - Research is working on it

WikiReactive

N. Boukhelifa, F. Chevalier and J.D. Fekete Real-time Aggregation of Wikipedia Data for Visual Analytics. In Proceedings of Visual Analytics Science and Technology. VAST '10. 147-154. 2010

- Collect wikipedia changes and computes derived information
  - Diffs, user contributions, user per character



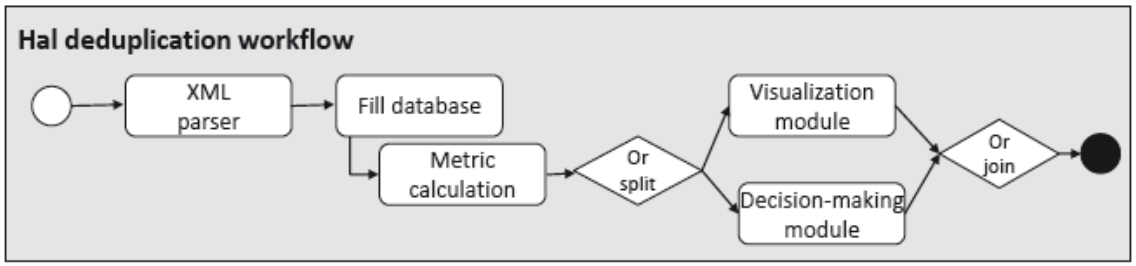


WikiReactive

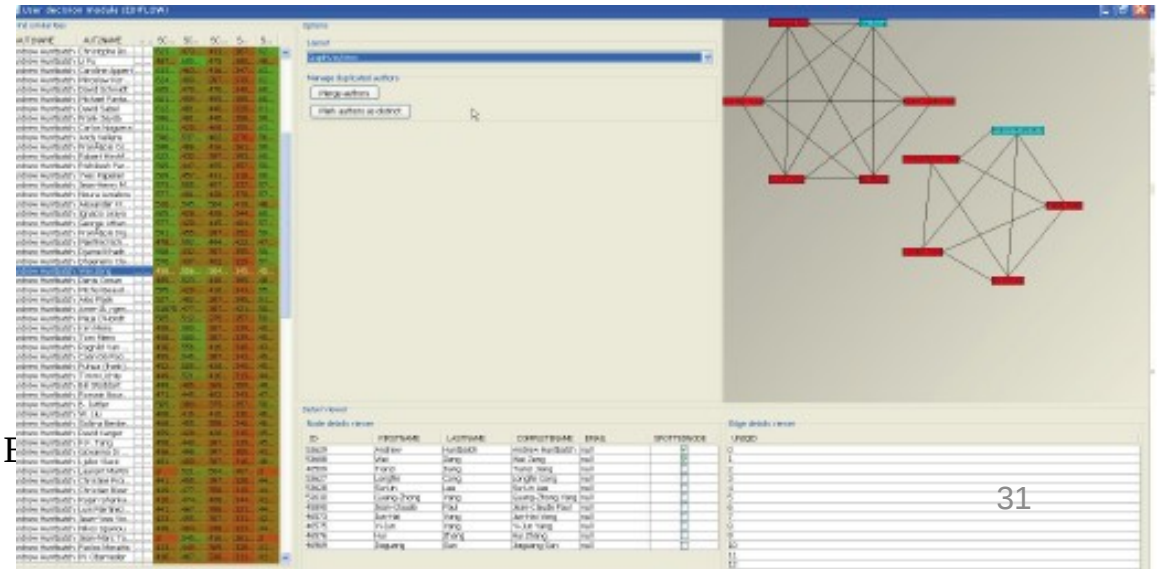
- Volume
  - 5 million articles in English, many TB of text
- Variety
  - Text + previous versions, structure
  - Users (id), Talks, categories, stats
- Velocity
  - About 100 changes per second
  - But each article does not change every second
- HW Category (1, 2, or 3)?

HAL Deduplication framework

- For each article author added to the HAL database



- Computes similarity with all other authors
- Resolve simple case (< or > threshold)
- Show an interface for the other cases



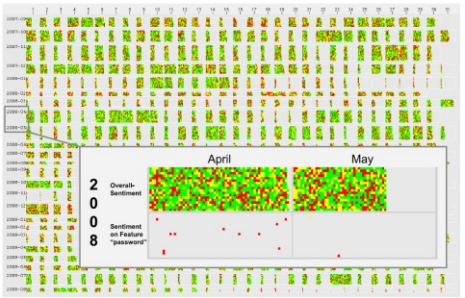
HAL Deduplication framework

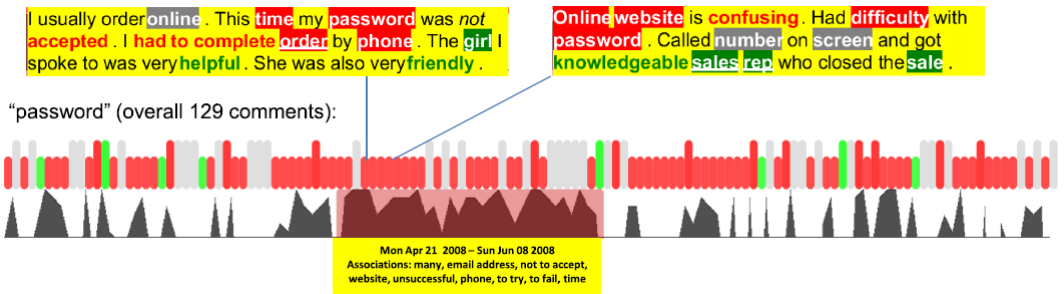
- Volume
  - 3 million articles, many TB of text, 3 million authors
- Variety
  - Users (id, email, institution, lab, date)
- Velocity
  - About 1 change per second
- HW Category (1, 2, or 3)?

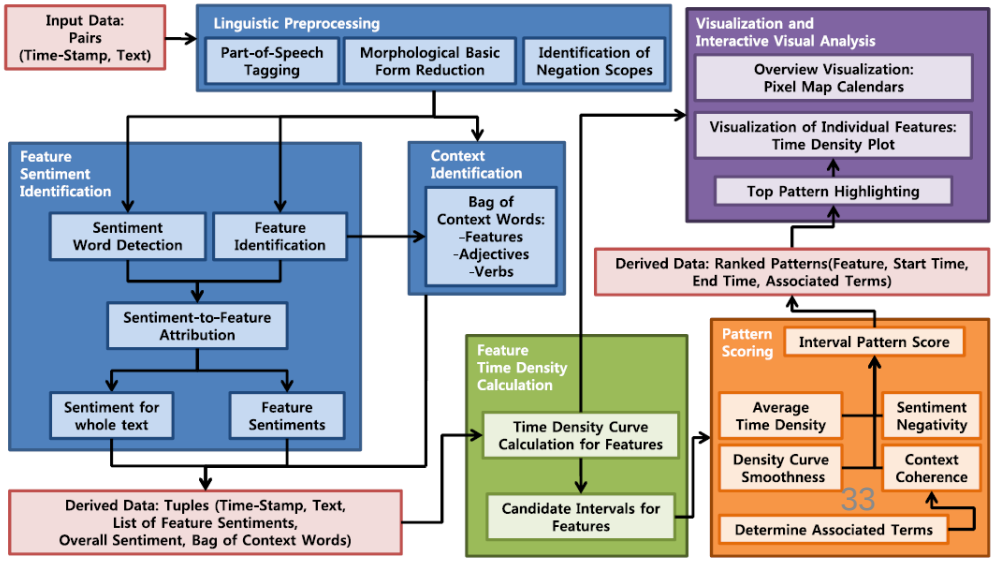
Real-Time Sentiment Analysis

- Christian Rohrdantz, Ming C. Hao, Umeshwar Dayal, Lars-Erik Haug, and Daniel A. Keim. 2012. Feature-Based Visual Sentiment Analysis of Text Document Streams. *ACM Trans. Intell. Syst. Technol.* 3, 2, Article 26 (February 2012), 25 pages.

- For each new document scrapped
- Compute part-of-speech tagging, lemmatization, negation detection, feature extraction, sentiment detection, sentiment-to-feature mapping





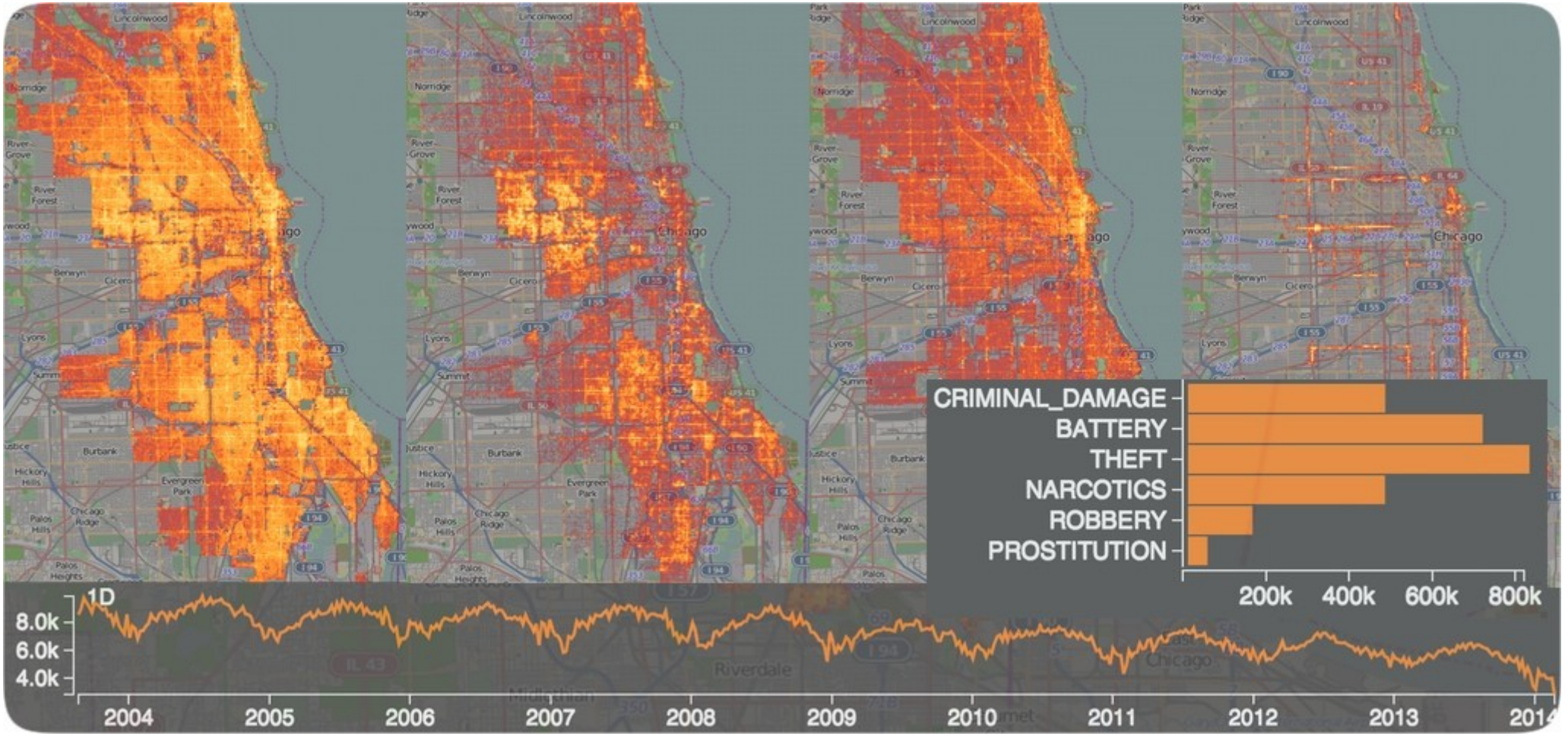


Real-Time Sentiment Analysis

- Volume
  - many million articles read continuously
- Variety
  - Time-stamp, text
- Velocity
  - As the crawler can work
- HW Category (1, 2, or 3)?

Nanocubes (Lins et al. 2013)

http://nanocubes.net/



Lauro Lins, James T. Klosowski, and Carlos Scheidegger. Nanocubes for Real-Time Exploration of Spatiotemporal Datasets. Visualization and Computer Graphics, IEEE Transactions on 19, no. 12 (2013): 2456-2465.

Nanocubes

- Create a spatio-temporal index
- Quickly retrieve distributions from range-queries
  - Over time
  - Over space
  - Over values
- Index creation can take hours

Nanocubes

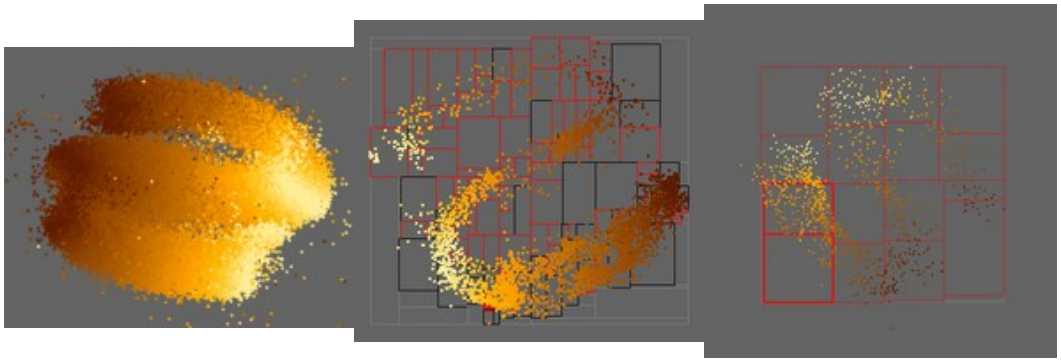
- Volume
  - Many (200) million points
- Variety
  - Spatio-temporal data
- Velocity
  - Static
- HW Category (1, 2, or 3)?

Beyond Pre-Computation: Bounding Time and Quality

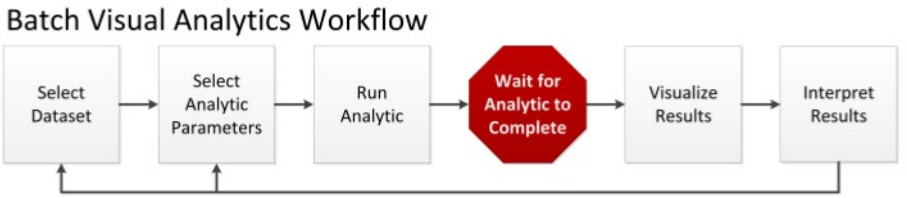
- Visualization is User Centric
  - Visualization will only show a small amount of data
  - Visualization need interactive time
  - How can we address the scale in interactive time?
- Analysis is Program Centric
  - Analysis will read data, process it and store its results in the end
  - Analysis will produce unbounded amounts of data in unbounded time
  - How can we get something in a bounded time?
- Databases is Data Centric
  - Databases will store and retrieve unbounded amounts of data in unbounded (but fast) time
  - How can we bound time with a specified level of quality?

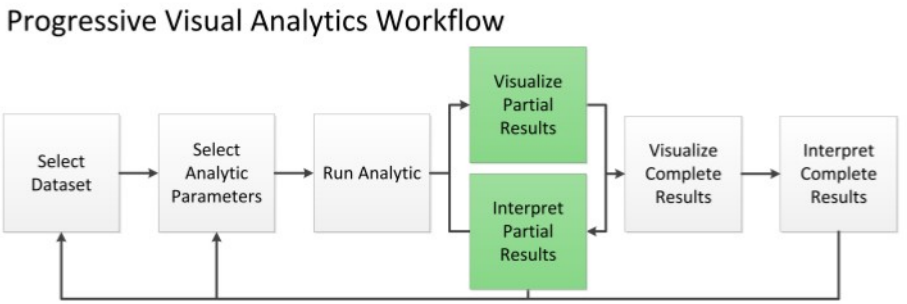
Progressive VA

- Allow Exploratory tools to work while the computation is being done



Williams, M.; Munzner, T., "Steerable, Progressive Multidimensional Scaling," in *INFOVIS 2004.*





Charles D. Stolper, Adam Perer, and David Gotz. **Progressive Visual Analytics**. *IEEE TVCG* (Volume 20, Issue 12, 2014).

Progressive MDS

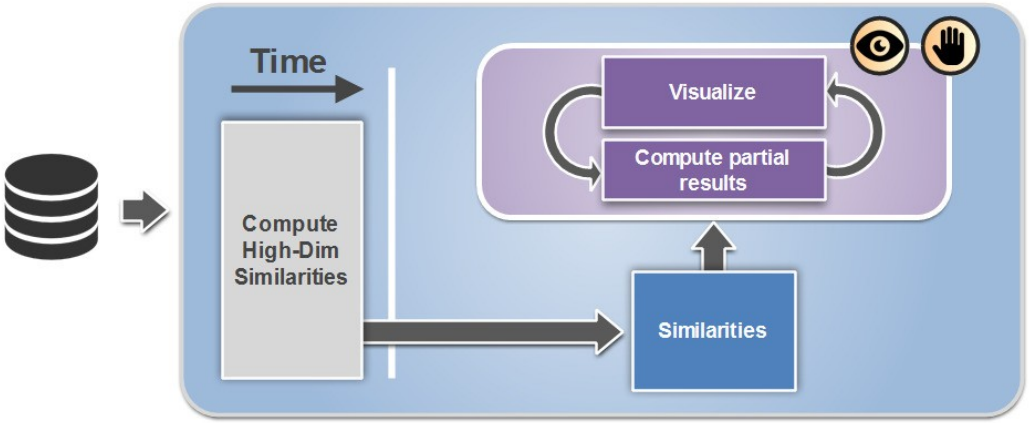
Progressive tSNE (Pezzoti et al. 2016)

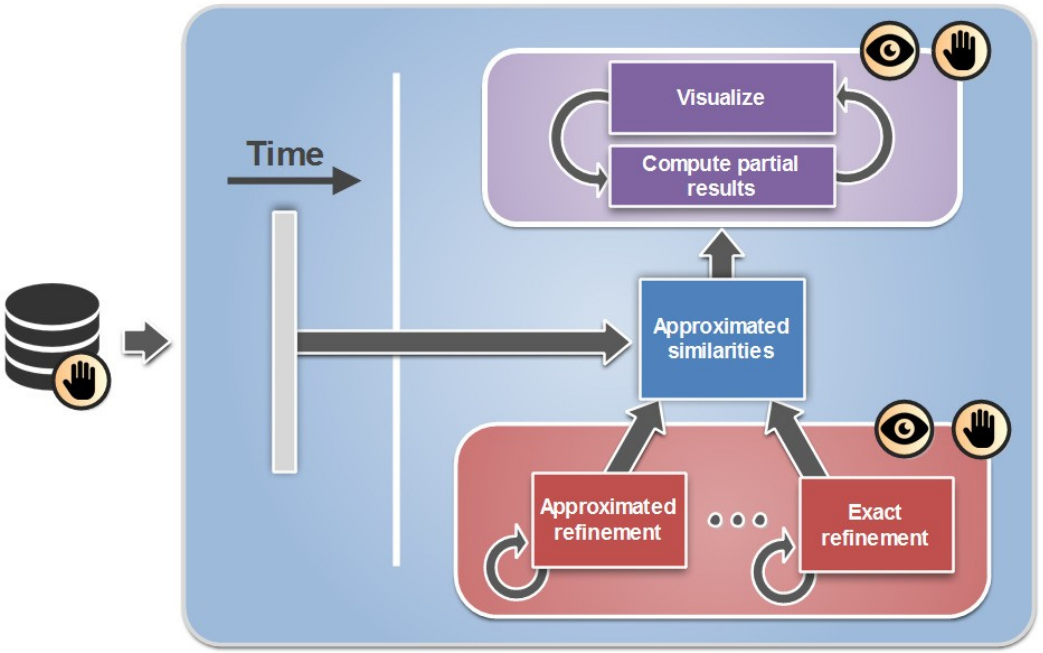
- Multidimensional projection method
- Input: points in nD
- Output: points in 2D
- Similar points nearby

https://lvdmaaten.github.io/tsne/

Progressive tSNE

- Compute distances
- Iterate to converge





Progressive tSNE

- Volume
  - many million points
- Variety
  - N-D points
- Velocity
  - Static or dynamic
- HW Category (1, 2, or 3)?

Bibliography

- R. B. Miller. Response time in man-computer conversational transactions. In Proceedings of the December 9-11, 1968, Fall Joint Computer Conference, Part I, AFIPS ’68 (Fall, part I), pages 267–277, New York, NY, USA, 1968. ACM.
- J. Nielsen. Response times: The 3 important limits, https://www.nngroup.com/articles/response-times-3-important-limits/
- B. Shneiderman. Response time and display rate in human performance with computers. ACM Comput. Surv., 16(3):265–285, Sept. 1984.
- Jinwook Seo, Ben Shneiderman, "Interactively Exploring Hierarchical Clustering Results," IEEE Computer, Volume 35, Number 7, pp. 80-86, July 2002. http://www.cs.umd.edu/hcil/hce/
- Ross, G. and Chalmers, M. (2003) A visual workspace for constructing hybrid MDS algorithms and coordinating multiple views. Information Visualization, 2 (4). pp. 247-257.
- N. Boukhelifa, F. Chevalier and J.D. Fekete Real-time Aggregation of Wikipedia Data for Visual Analytics. In Proceedings of Visual Analytics Science and Technology. VAST '10. 147-154. 2010 http://www.aviz.fr/Research/Wikireactive
- Christian Rohrdantz, Ming C. Hao, Umeshwar Dayal, Lars-Erik Haug, and Daniel A. Keim. 2012. Feature-Based Visual Sentiment Analysis of Text Document Streams. ACM Trans. Intell. Syst. Technol. 3, 2, Article 26 (February 2012), 25 pages.
- Lauro Lins, James T. Klosowski, and Carlos Scheidegger. Nanocubes for Real-Time Exploration of Spatiotemporal Datasets. Visualization and Computer Graphics, IEEE Transactions on 19, no. 12 (2013): 2456-2465. http://nanocubes.net/
- Matt Williams and Tamara Munzner. 2004. Steerable, Progressive Multidimensional Scaling. In Proceedings of the IEEE Symposium on Information Visualization (INFOVIS '04). IEEE Computer Society, Washington, DC, USA, 57-64. DOI=http://dx.doi.org/10.1109/INFOVIS.2004.60 http://www.cs.ubc.ca/~tmm/papers/mdsteer/
- N. Pezzotti, B.P.F. Lelieveldt, L. van der Maaten, T. Höllt, E. Eisemann, and A. Vilanova. Approximated and User Steerable tSNE for Progressive Visual Analytics. Transaction on Visualization and Computer Graphics. http://nicola17.github.io/ https://lvdmaaten.github.io/tsne/