Institutes of Science and Development
Chinese Academy of Sciences

MAPPING RESEARCH FUNDING BY T-SNE EMBEDDING

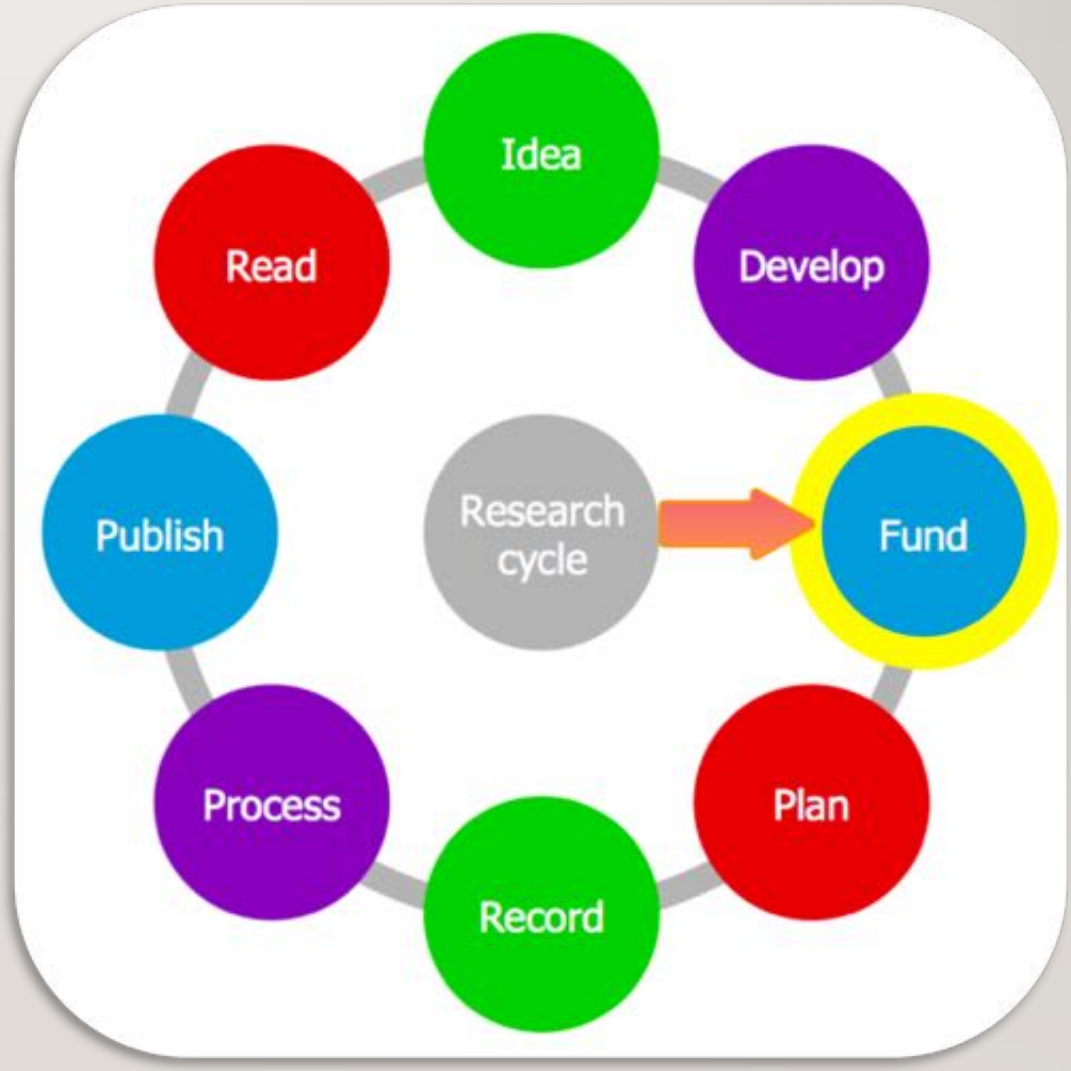
Ting Chen， Li Guopeng, Wang Xiaomei

Institutes Of Science And Development, CAS

GTM, Leiden, 2018

Why do we analysis funding data

- Research projects funded by grants represent the latest research ideas, maybe it could reveal research frontier quicker than published paper;
- Published paper focus on the research detail, funding application is more about ideas;
- Researches supported by national funding agencies may contain the future development direction of the country;



Objectives of this study

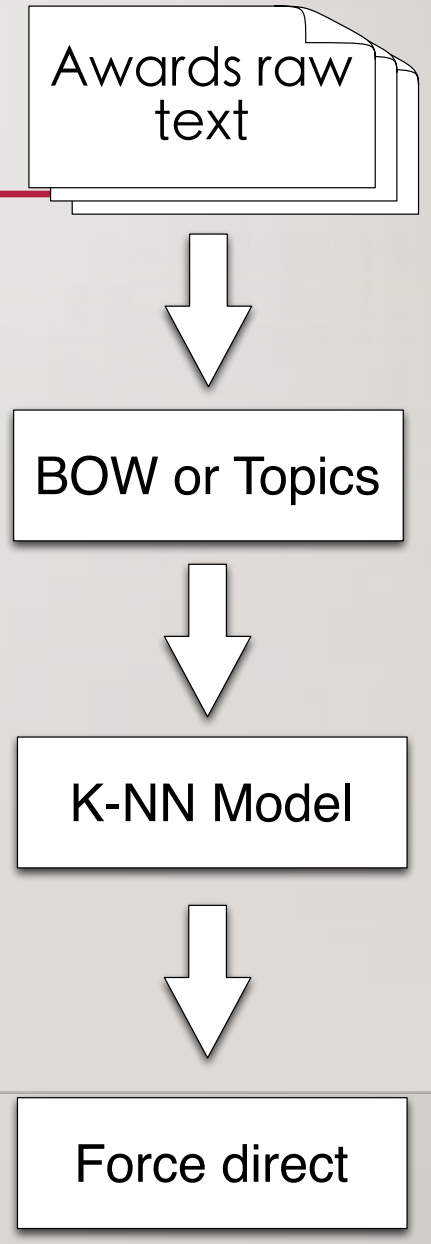
- Data visualizations are the best way to communicate the insights derived from data analysis;
- In this study we compared different mapping methods, discover a suitable one for visualizing unstructured funding aplication data;
- Create the landscape of research funding, and exploring research hotspots and gaps;

Test Data

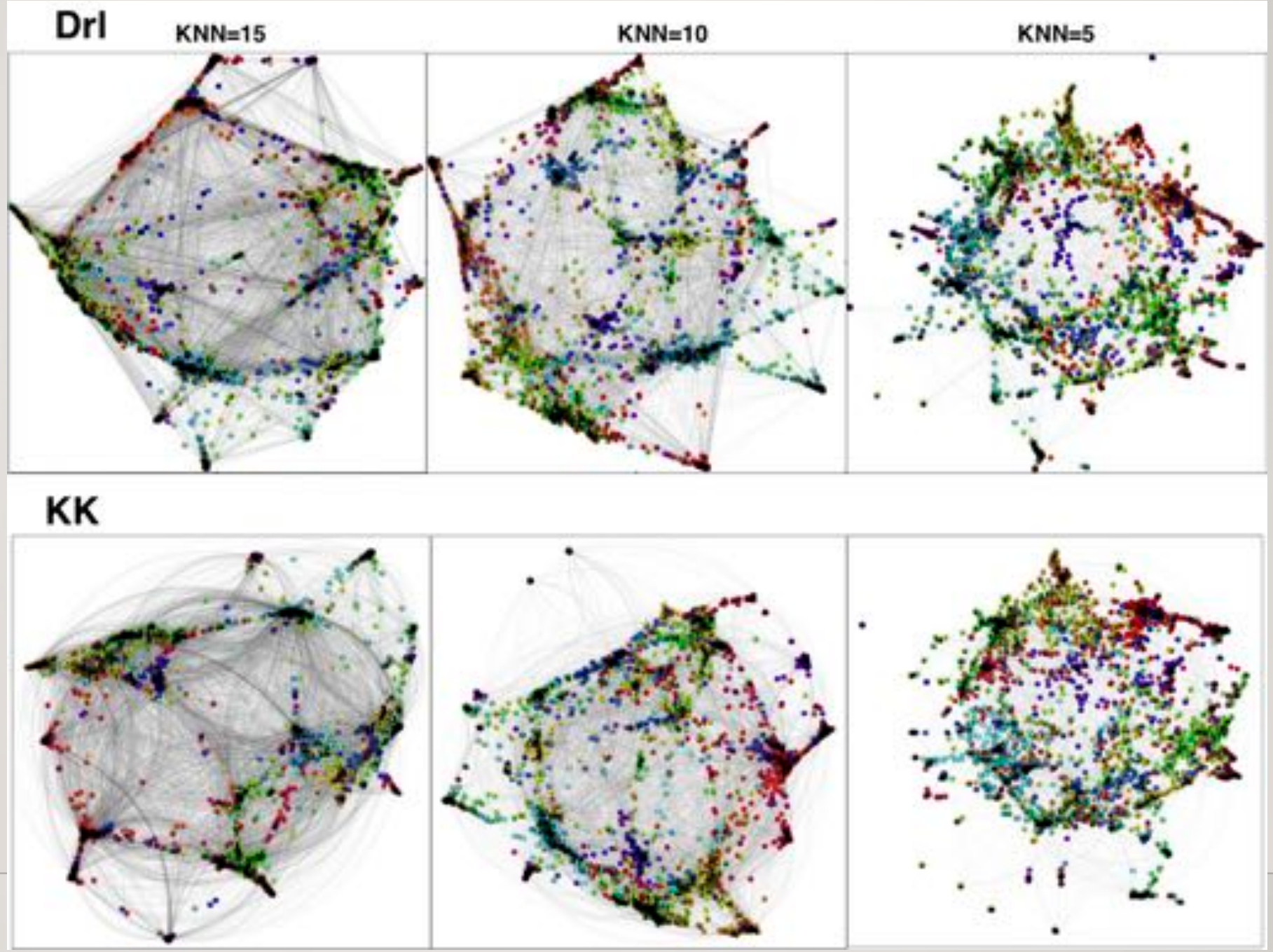
- 4669 NSF Information and Intelligent Systems department funded awards from 2008 to 2017
- **Create a labeled test dataset:**
  - Use k-means to divided awards into 70 small clusters, smaller clusters, better homogeneity
  - Human-read each cluster, combined some similar clusters into one, make sure the test set also has good completeness
  - Total 21 topics have been labeled, we will test our mapping methods by using 21 topic labels. Topics include NLP, data retrieval, database, image recognition, voice recognition, motion monitoring, robotics, brain-computer interface, etc.

First Try: Network Graph

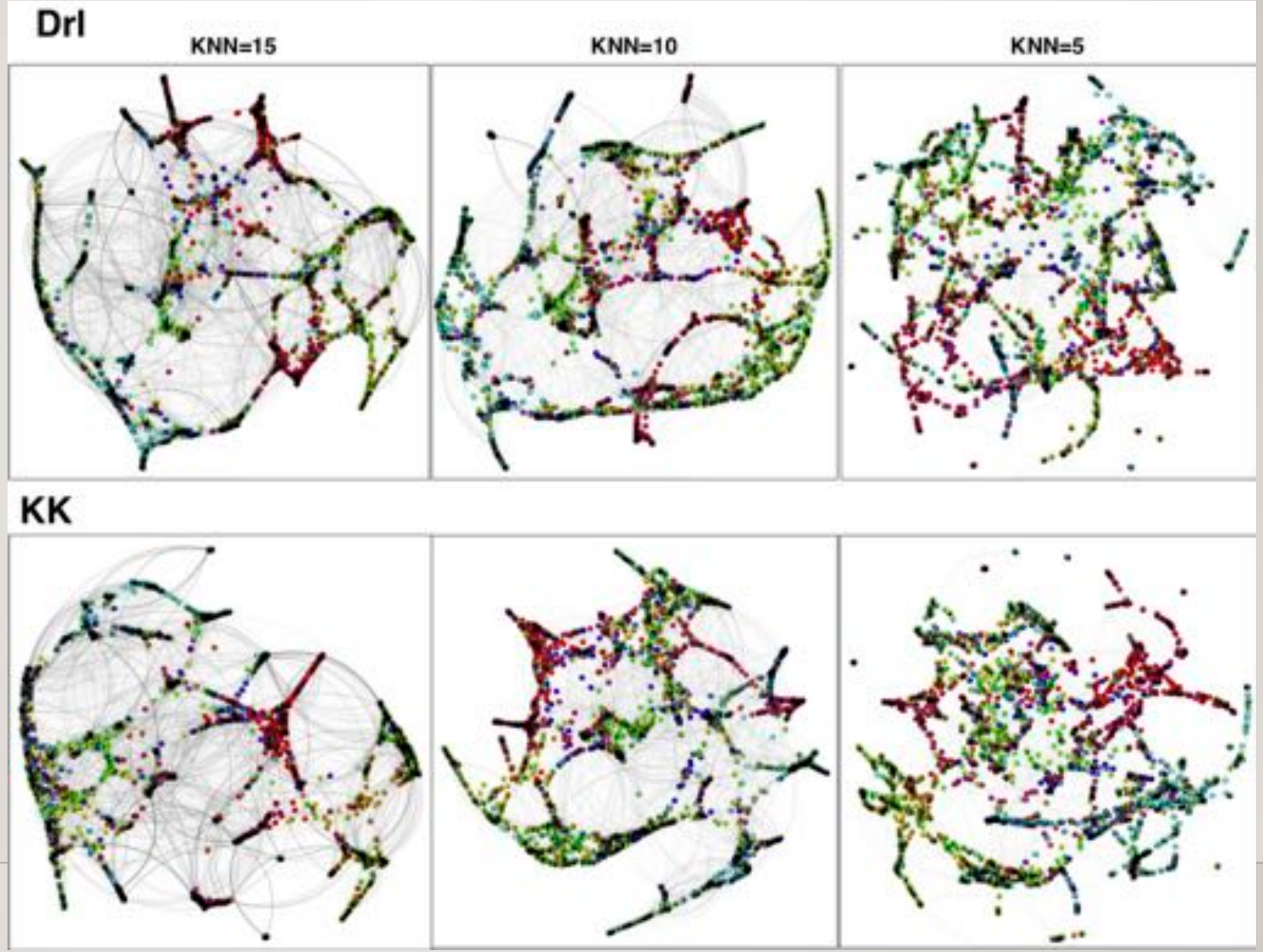
1. Standard NLP: Stop words, stemming and lemmatization...
2. Feature extraction: BOW and topic model LSA
3. Create a graph network: KNN model, K=5,10,15
4. Force direct layout: Two most common medium-sized networks layouts Drl (OpenOrd based on it) and Kamada-Kawai (KK) were applied



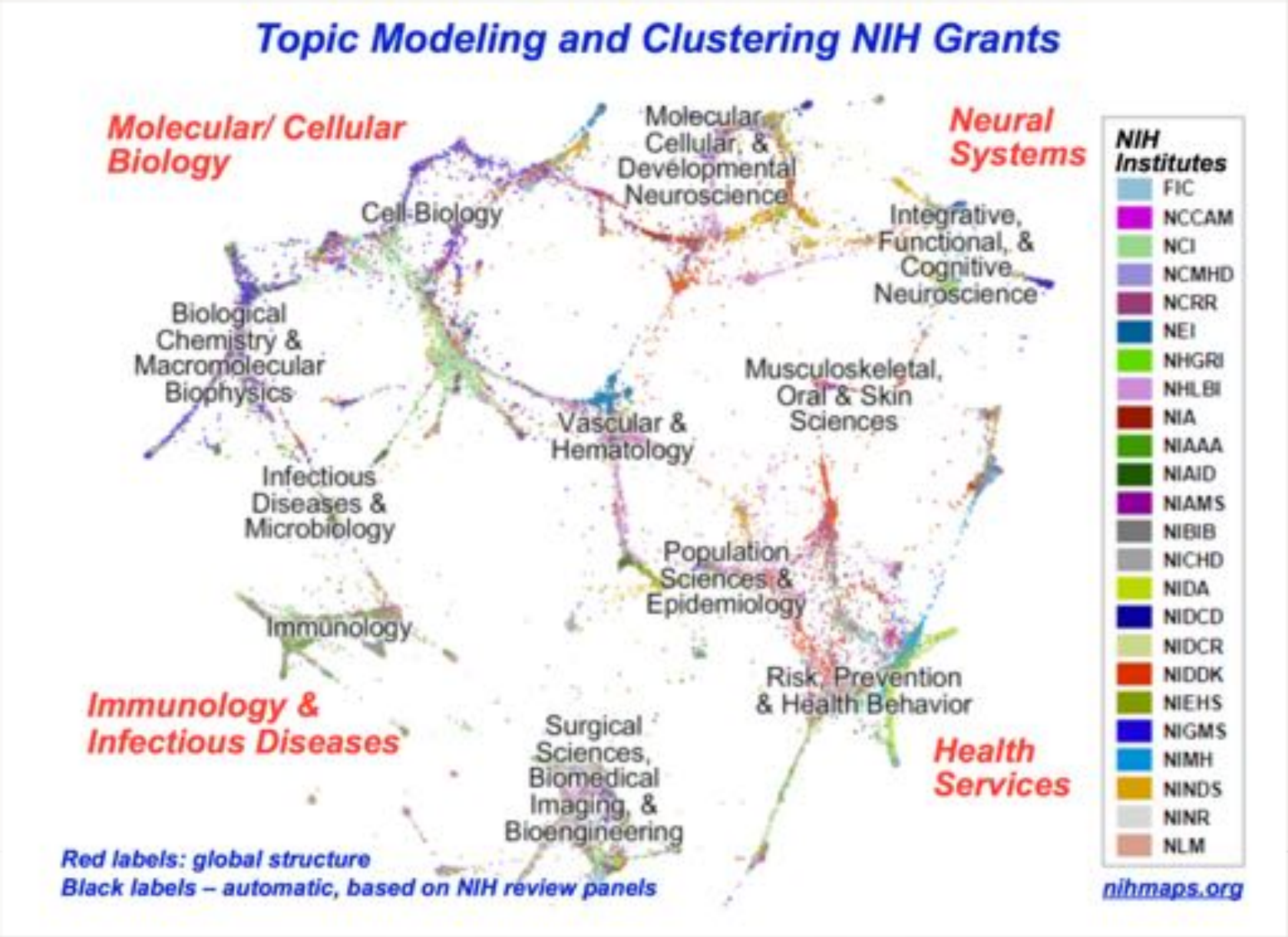
Graphs created by:
tf-idf similarity (7000 features)



Graphs create by:
Topic model
LSA similarity
(20 LSA features)



Ii B W H, Talley E M, Burns G A P C, et al. The NIH Visual Browser: An Interactive Visualization of Biomedical Research[C]// Information Visualisation, 2009, International Conference. IEEE, 2009:505-509.



Best graph:

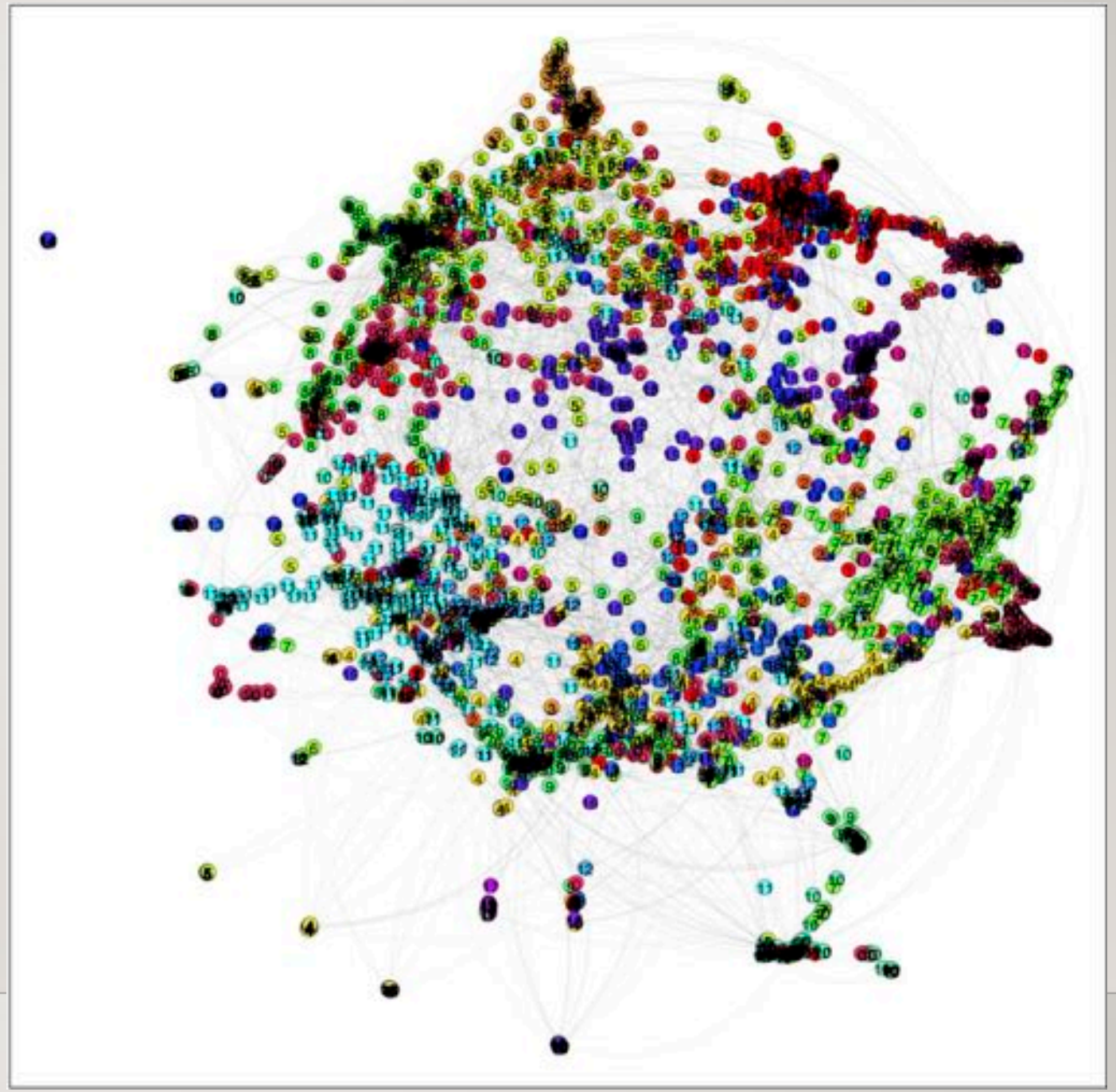
- BOW-Tf-idf features
- K-Nearest neighbor K = 5
- Kamada-Kawai force direct layout
- Use it as a base map for this research

Pros

- Good global structure, some natural clusters appeared
- Degree, betweenness, centrality. etc
- Very fast

Cons:

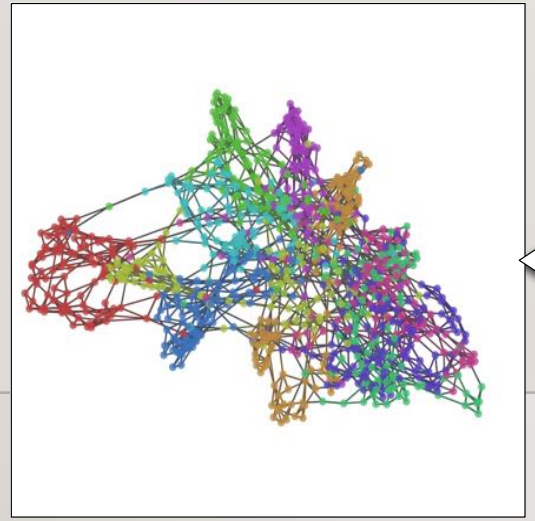
- Poor topic-separability (the local detail)
- No real networks for funding data, we have to convert vector features into similarity network (distance matrix)
- The choice of number of links is extremely critical



Second Try: Dimensionality Reduction (Embedding)

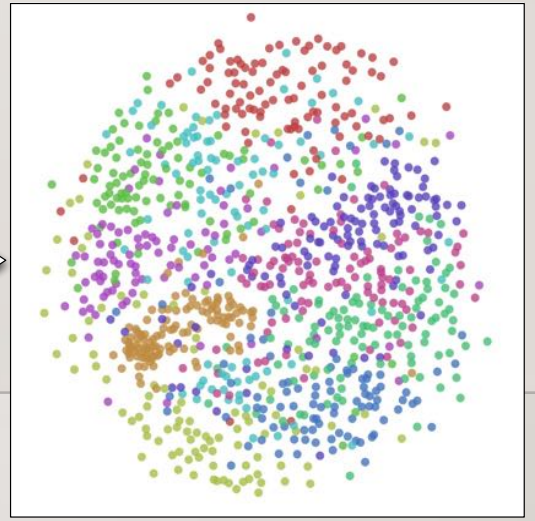
- High-dimensional datasets can be very difficult to visualize. To aid visualization of the structure of a dataset, the dimension must be reduced in some way.
- Dimensionality Reduction methods were used for translating high-dimensional funding textual data into lower dimensional data;

Graph



Hand writing digits

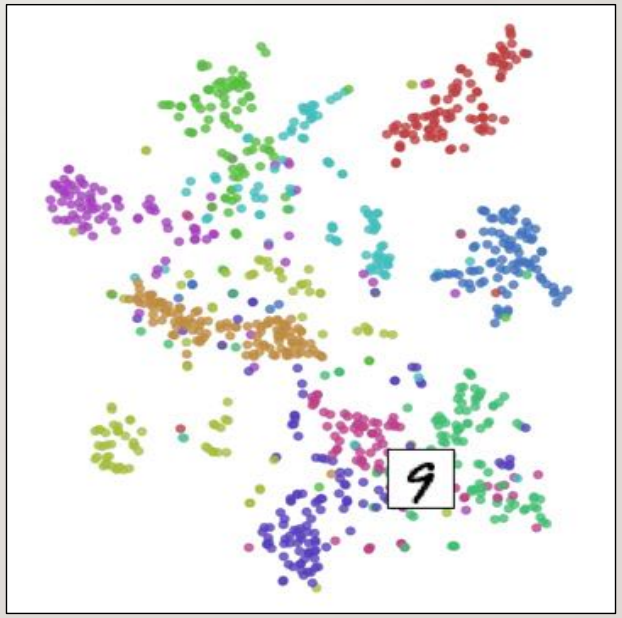
MDS



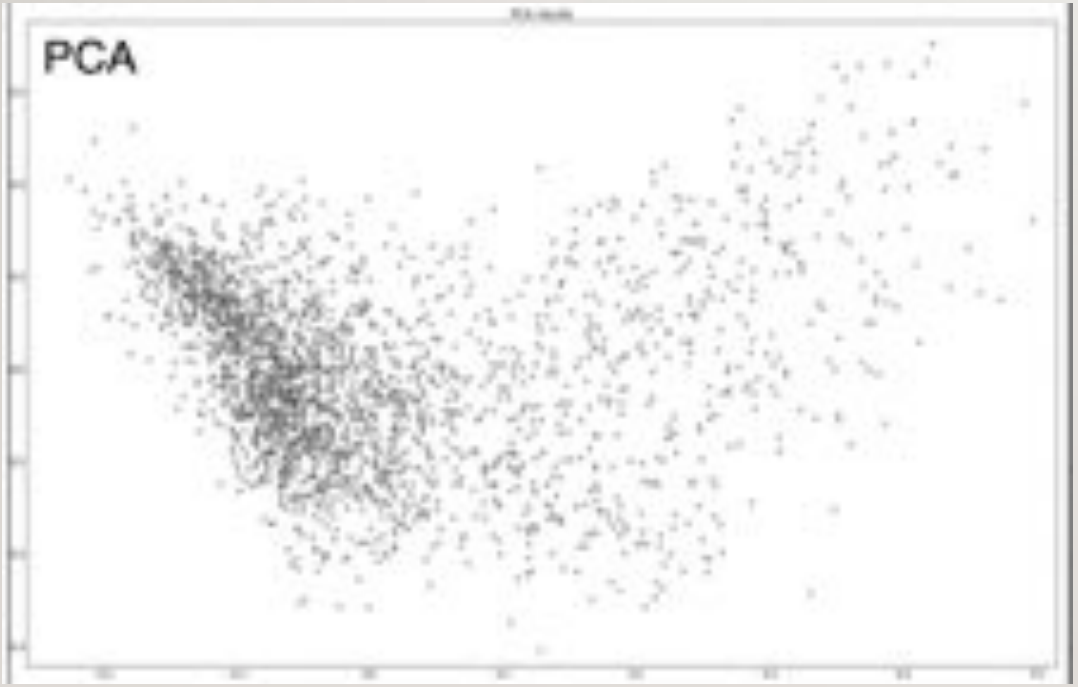
State of Art: t-distributed stochastic neighbor embedding (t-SNE)

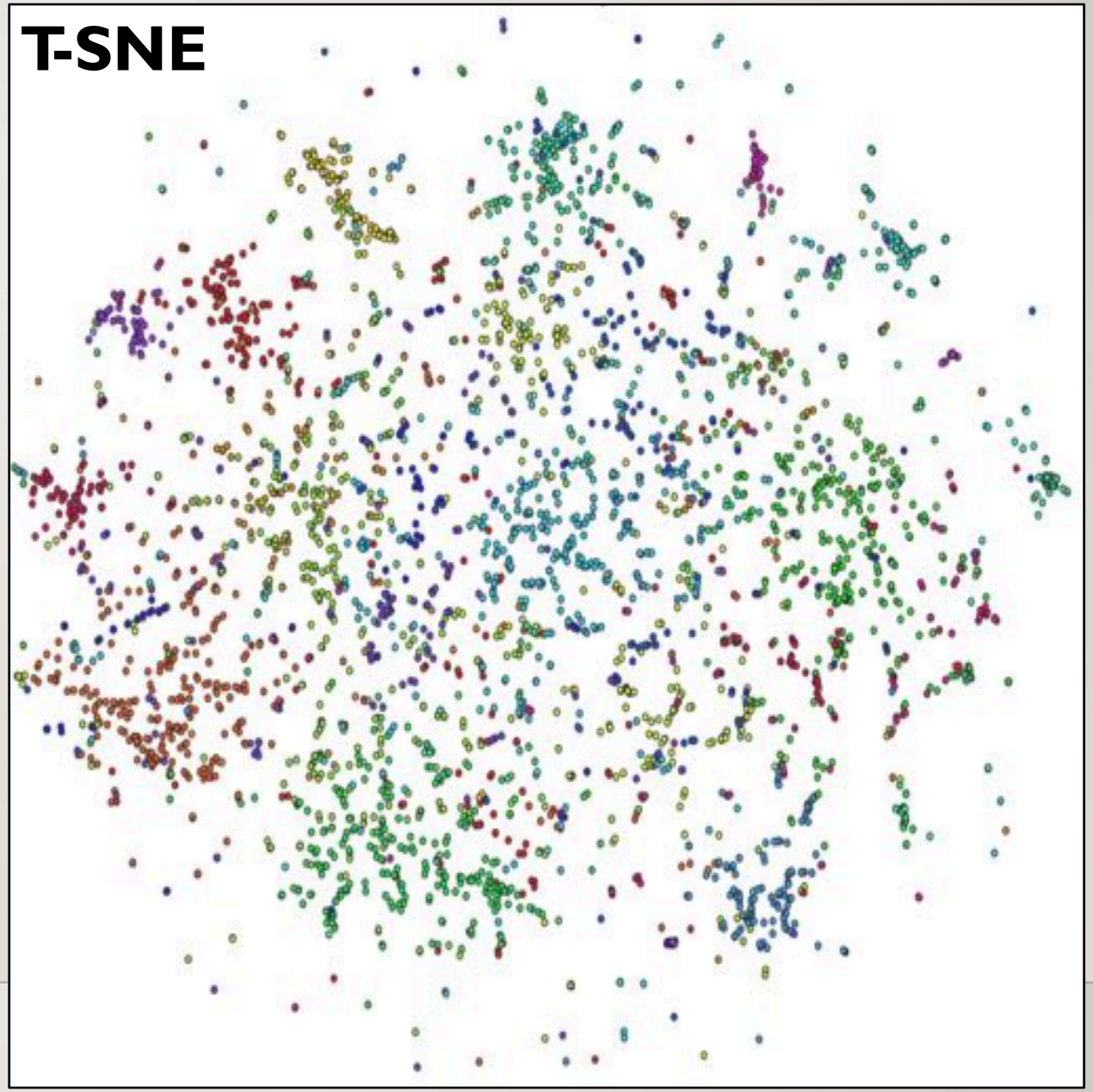
- Van der Maaten, L.J.P.; Hinton, G.E. (Nov 2008)
- t-SNE tends to preserve local structure and at the same time preserving the global structure as much as possible
- Others try to preserve the global structure but missed a lot of local details

T-sne



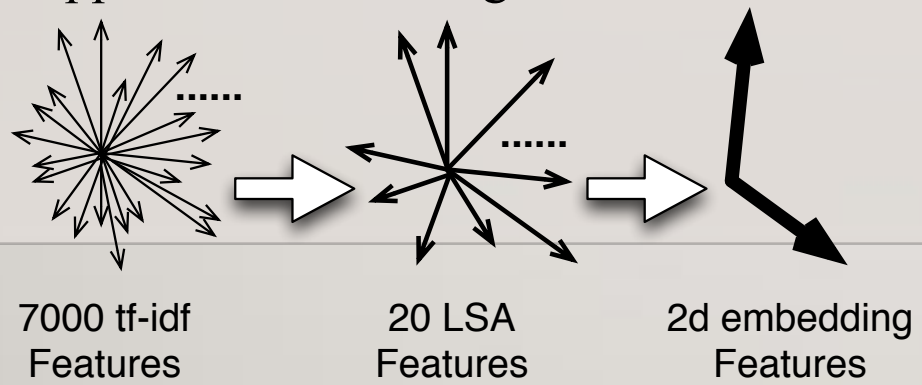
tf-idf 7000 features vector space

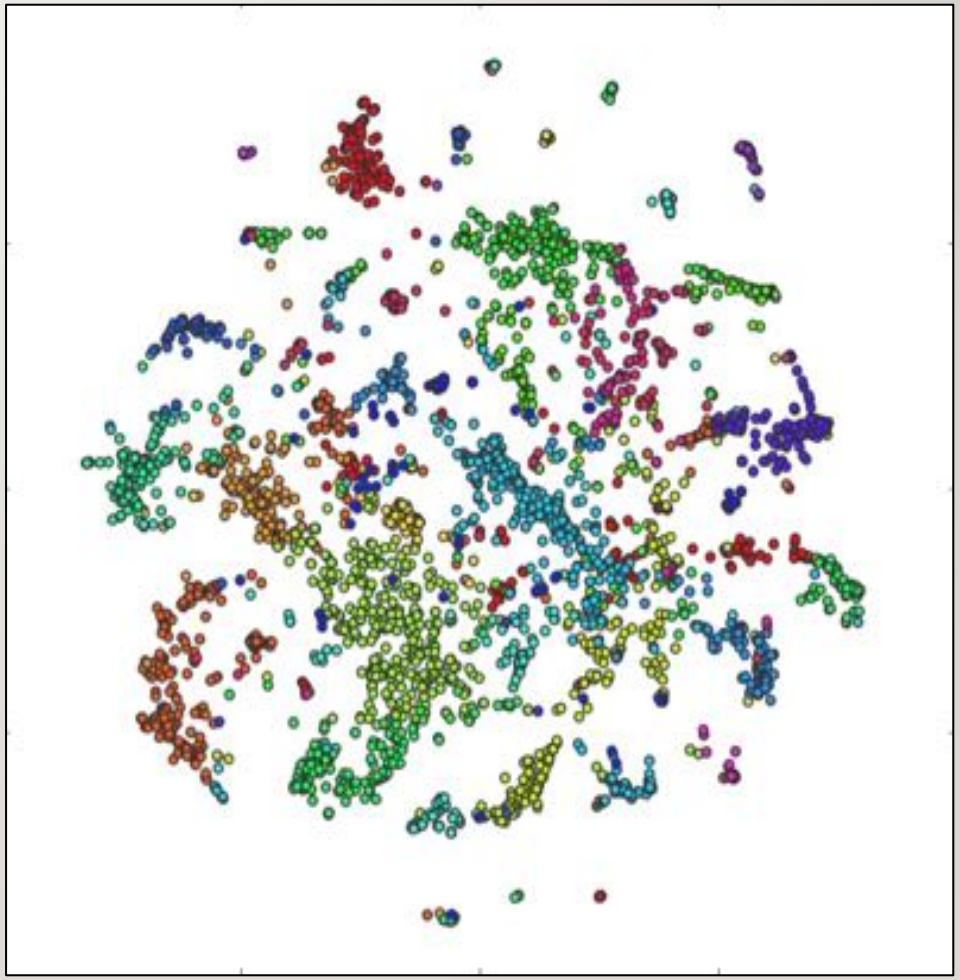




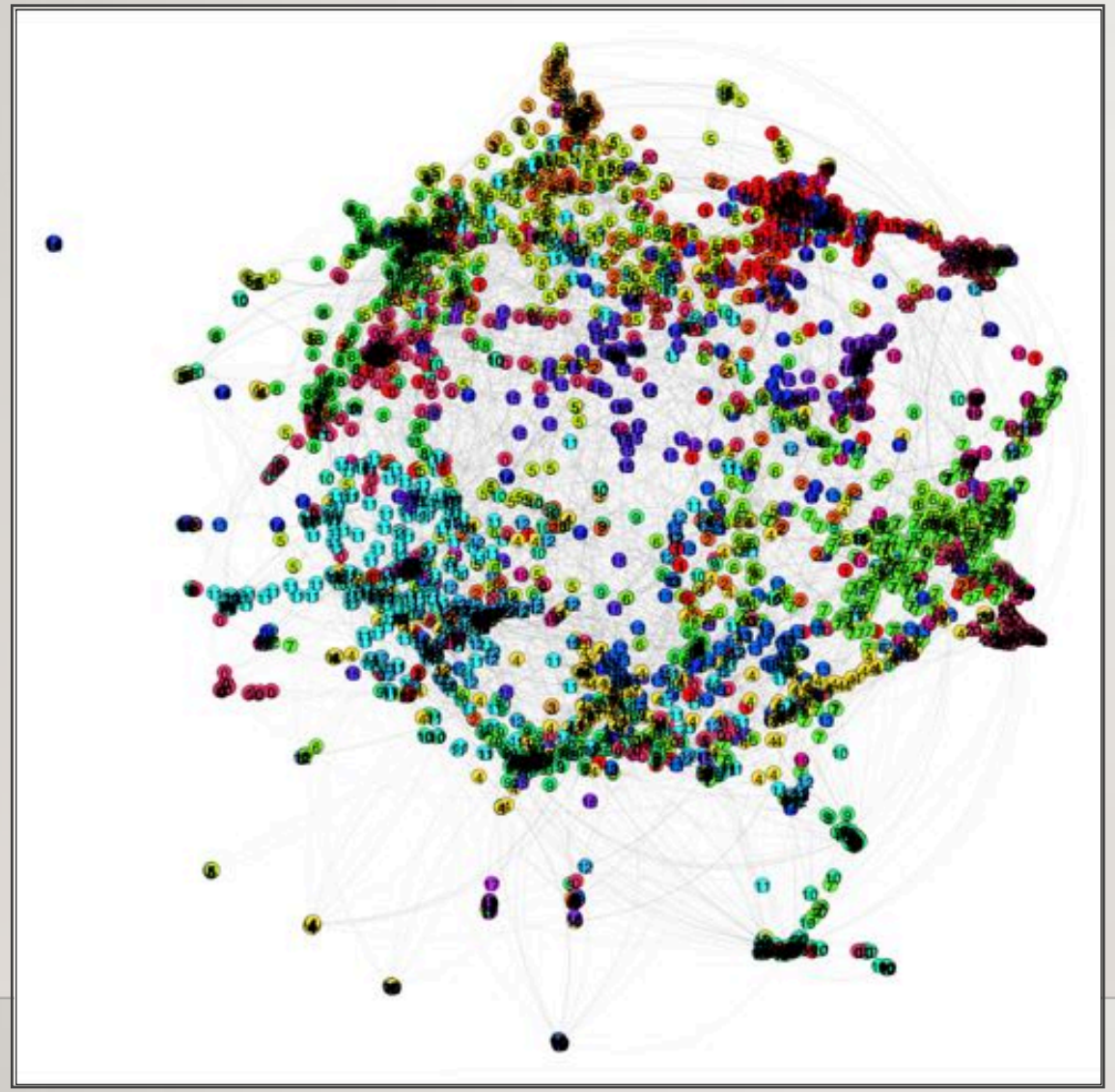
t-SNE funding map

- 7000 tf-idf features are too high (Dimensional disaster);
- Add topic features between tf-idf and 2d space;
- Now, 20 topics features embedding to 2d space;
- well-separated clusters even in non-clustered data appeared on the map, even some sub-topics appeared in some larger cluster

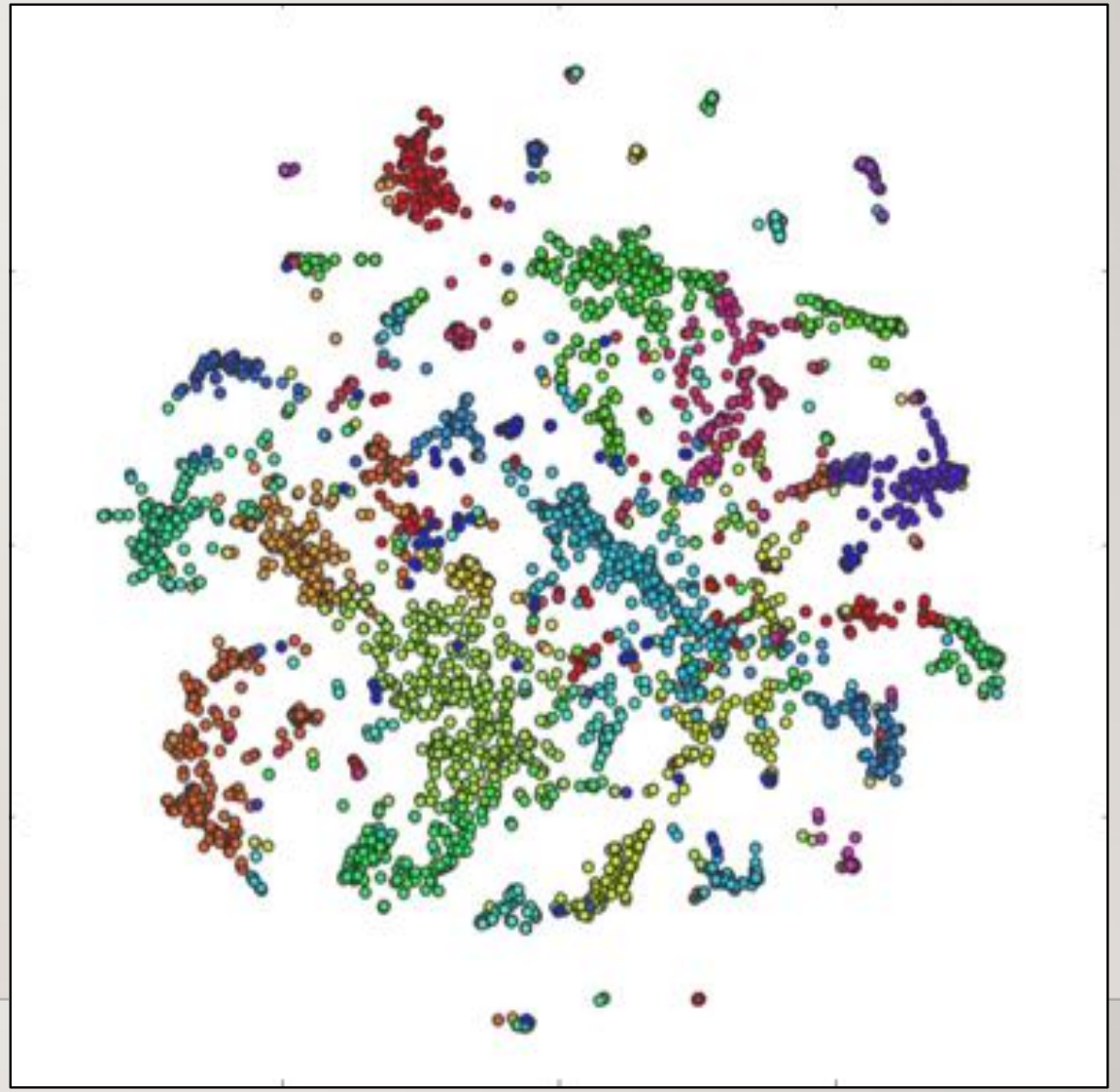




Graph funding map (Base map)



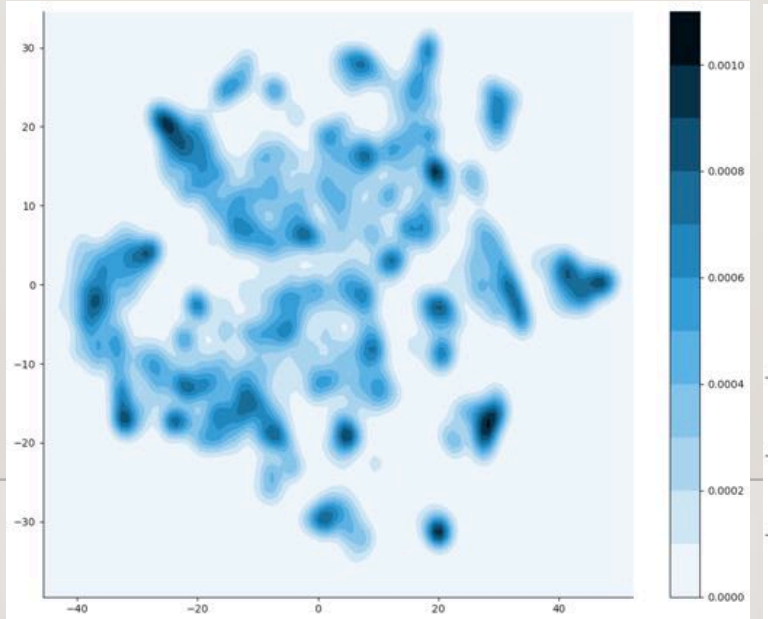
Embedding funding map



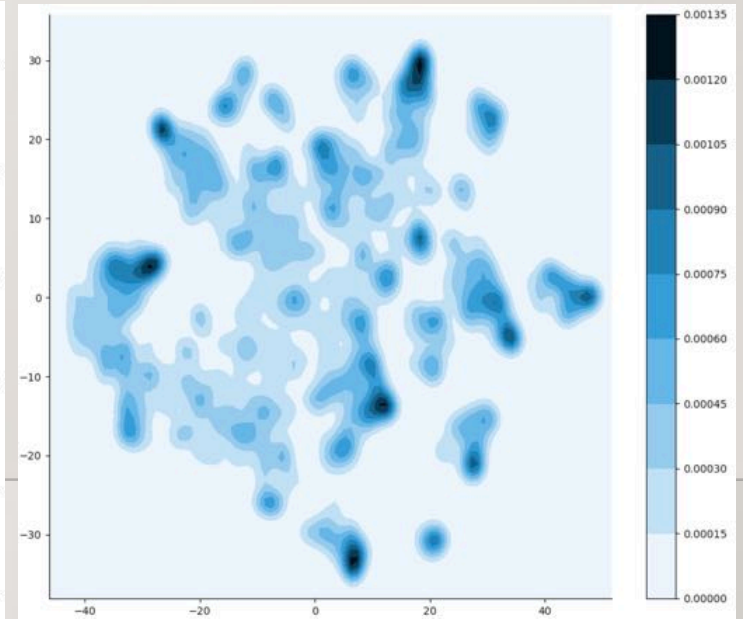
Applications

- **Hotspots detection:** The map provides good local structure, it is able to detect well-separated clusters even in non-clustered data. So one of the applications I suggested is funding hotspots in a global research landscape, the clustering is not necessary, we could directly apply the density function on the map. The high-density area (hot funding spots) would appear on the map.

NASA funding map 2000 ~ 2008

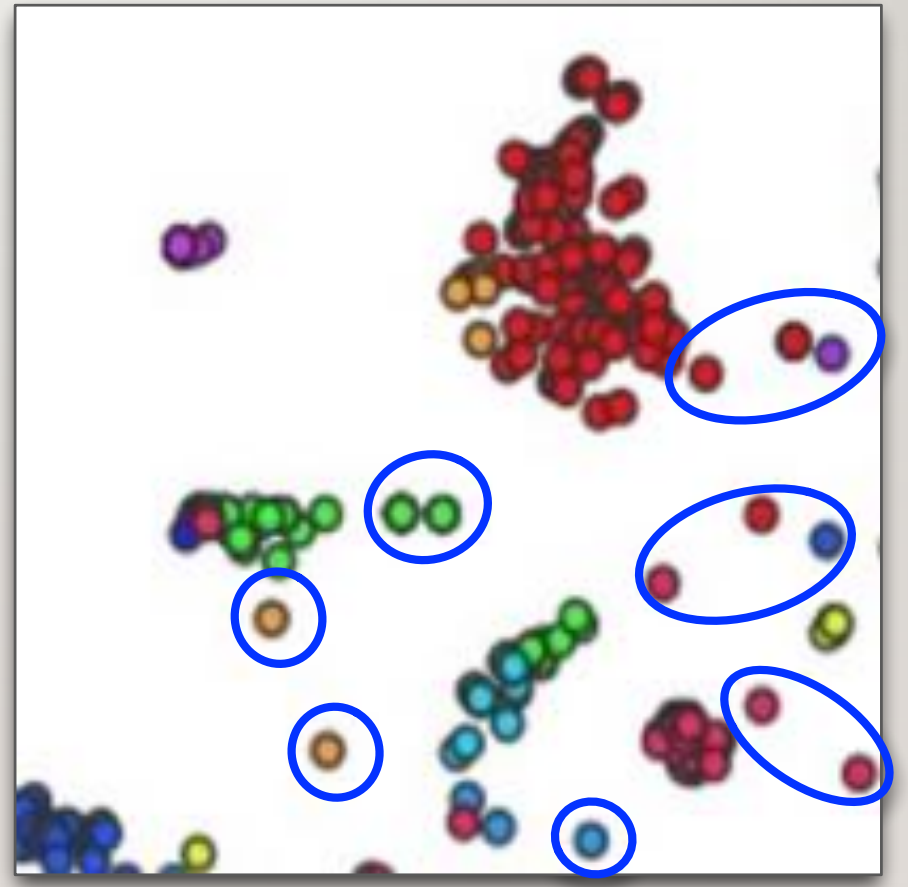


NASA funding map 2009 ~ 2017



Applications

- **Novelty/outlier detection**: Unlike the network graph, embedding won't lose the outliers (node without links), it will place the outlier at an appropriate position on the map even without any links.
- Funding agency more likely to support novel research.
- Some novelty/outlier detection method can be used based on the funding map, such as one-class SVM, IsolationForest



Discussion

- Both graph and embedding funding maps are good at revealing the global structure;
- The embedding map has the capability for retaining the local structure of the funding data, it seems to display natural clusters and sub-clusters very well;
- Text representing cannot be too high, better features will get a better map;
  - Tf-idf, BM25, LSA, LDA, NMF, Word2vec average/sum, doc2vec
- The cost of t-SNE algorithm is O(n 2), not very fast and scalable;

Next step

- Collect the larger amount of funding text data, train a doc2vec model for funding text
- Test different text representation models with t-SNE for creating the map, find the best combinations;
- Try to apply the funding map with multiple funding agencies' data, NSF/EURO Horizon 2020, maybe we will find some differences between counties and agencies;
- Test the method with other data sources, maybe patent or policy dataset;

Institutes of Science and Development
Chinese Academy of Sciences

Thanks very much!

Ting Chen
chenting@casisd.cn