# Technology and Application of Big Data

Qing LIAO(廖清)

School of Computer Science and Technology

HIT

# Course Details

- Instructor:
  - Qing LIAO, liaoqing@hit.edu.cn
  - Rm. 303B, Building C
  - Office hours: by appointment
- Course web site:
  - liaoqing.me
- Reference books/materials:
  - Big data courses from University of California
  - Book: BIG DATA: A Revolution That Will Transform How We Live, Work, and Think
  - Papers
- Grading Scheme:
  - Paper Report 30%
  - Final Exam 70%
- Exam:
  - 21st July(Friday), 14:00-16:00, A502

# What You Learnt: Overview

- Topics:
  1) Introduction of Big Data
  2) Characterizes of Big Data
  3) How to Get Value from Big Data
  4) Technologies of Big Data
  5) Applications of Big Data
- Prerequisites
  - Statistics and Probability would help
    - But not necessary
  - Machine Learning would help
    - But not necessary

# Prevision Section

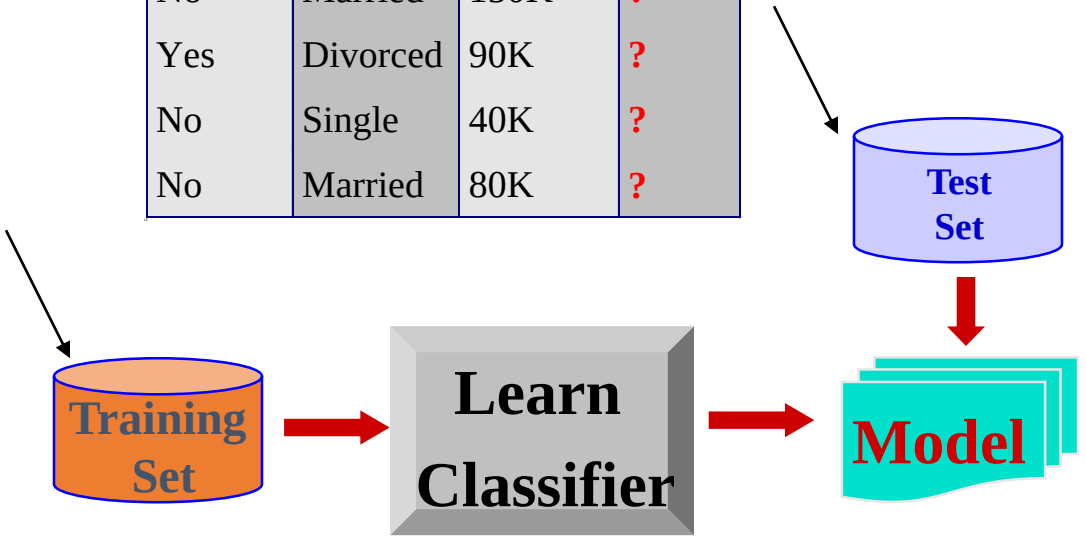
- Supervised learning (classification)
  - training data (observations, measurements, etc.) are accompanied by **labels** indicating the class of the observations
  - New data is classified based on the training set
- Unsupervised learning(clustering)
  - The class labels of training data is **unknown**
  - Given a set of measurements, observations, etc. with the aim of establishing the existence of classes or clusters in the data

# Classification: Tax Cheating

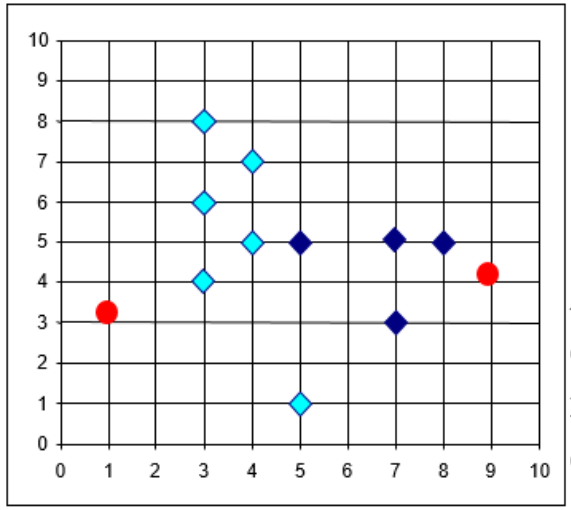
- Tax Income

| *Tid* | Refund (categorical) | Marital Status (categorical) | Taxable Income (continuous) | Cheat (class) |
|---|---|---|---|---|
| 1 | Yes | Single | 125K | No |
| 2 | No | Married | 100K | No |
| 3 | No | Single | 70K | No |
| 4 | Yes | Married | 120K | No |
| 5 | No | Divorced | 95K | Yes |
| 6 | No | Married | 60K | No |
| 7 | Yes | Divorced | 220K | No |
| 8 | No | Single | 85K | Yes |
| 9 | No | Married | 75K | No |
| 10 | No | Single | 90K | Yes |

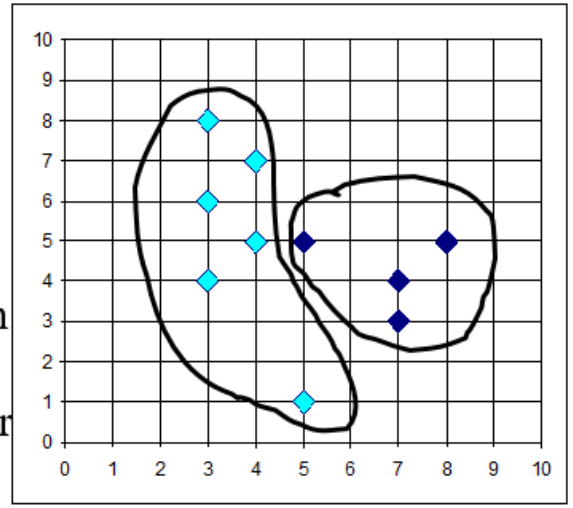
| Refund | Marital Status | Taxable Income | Cheat |
|---|---|---|---|
| No | Single | 75K | ? |
| Yes | Married | 50K | ? |
| No | Married | 150K | ? |
| Yes | Divorced | 90K | ? |
| No | Single | 40K | ? |
| No | Married | 80K | ? |



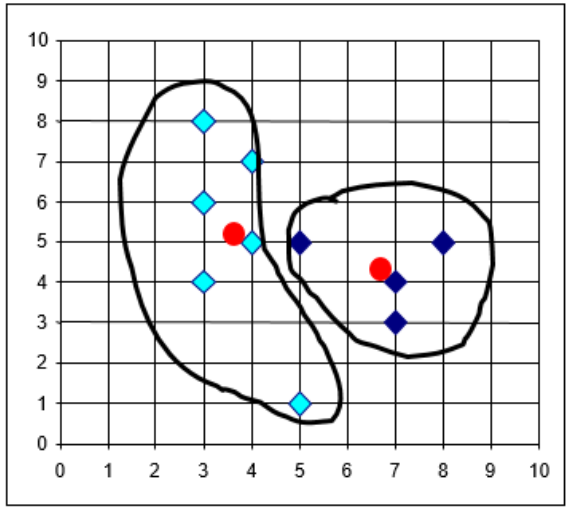
# Clustering: K-means Clustering



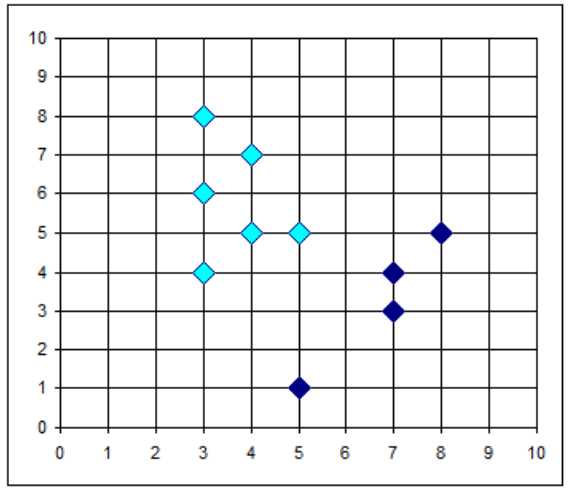
Assign each objects to most similar center



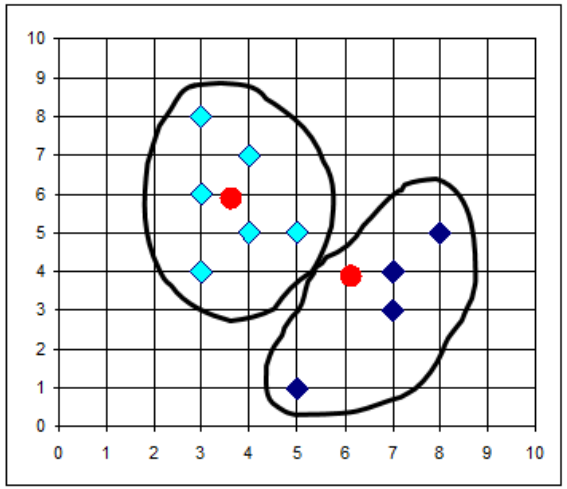
Update the cluster means



reassign



Update the cluster means



reassign

K=2

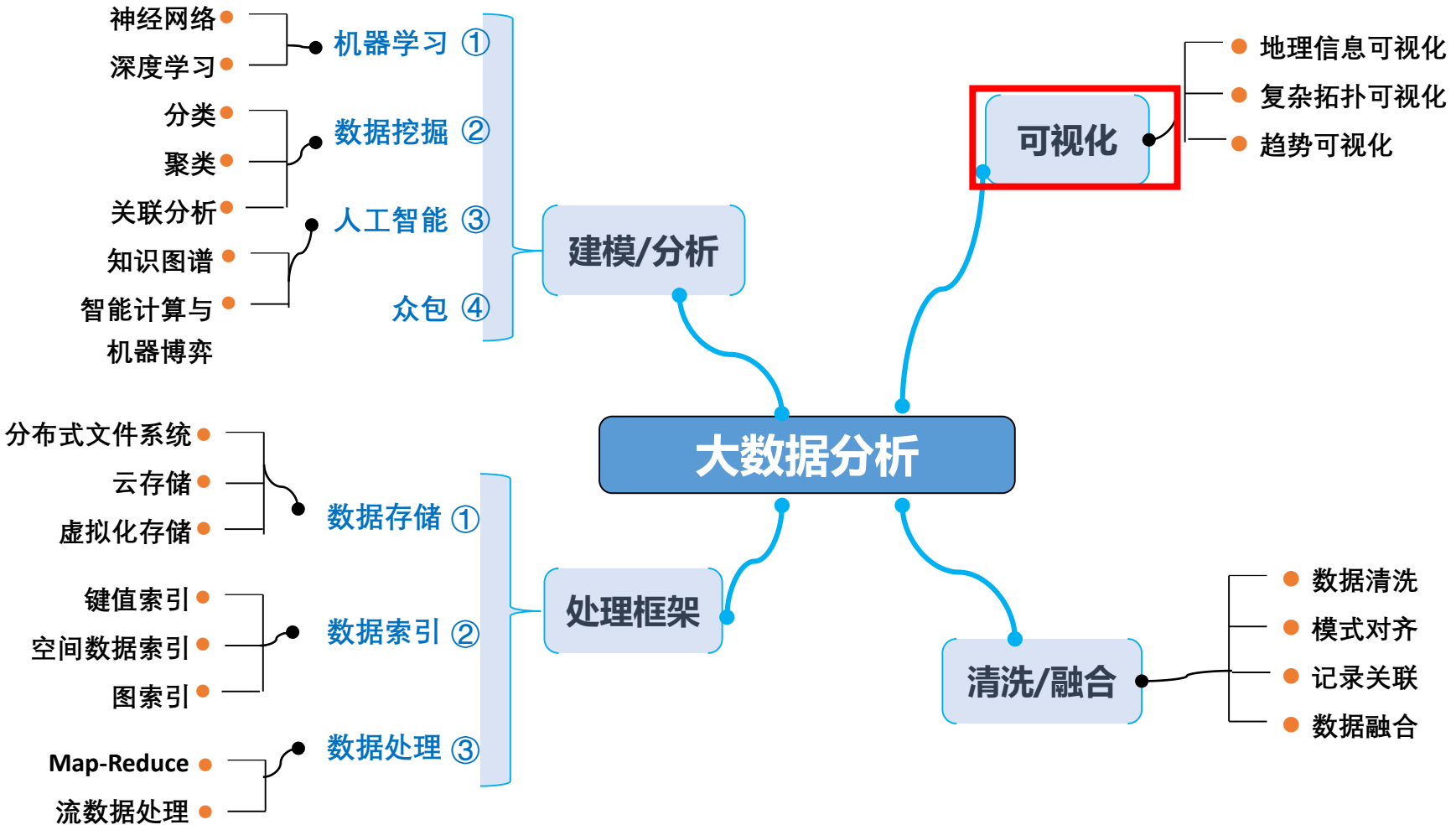
Arbitrarily choose K object as initial cluster center

The mean point can be a virtual point!

# Clustering Example: Market Segmentation

- Goal: Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs.
- Approach:
  - Collect different attributes of customers based on their geographical and lifestyle related information.
  - Find clusters of similar customers.
  - Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters.

# Technologies of Big Data




*Thanks NUDT Provides Slides*

# What is data visualization?

- “Data visualization is the creation and study of the visual representation of data” - wiki
- Input: data  Output: visual form  Goal: insight

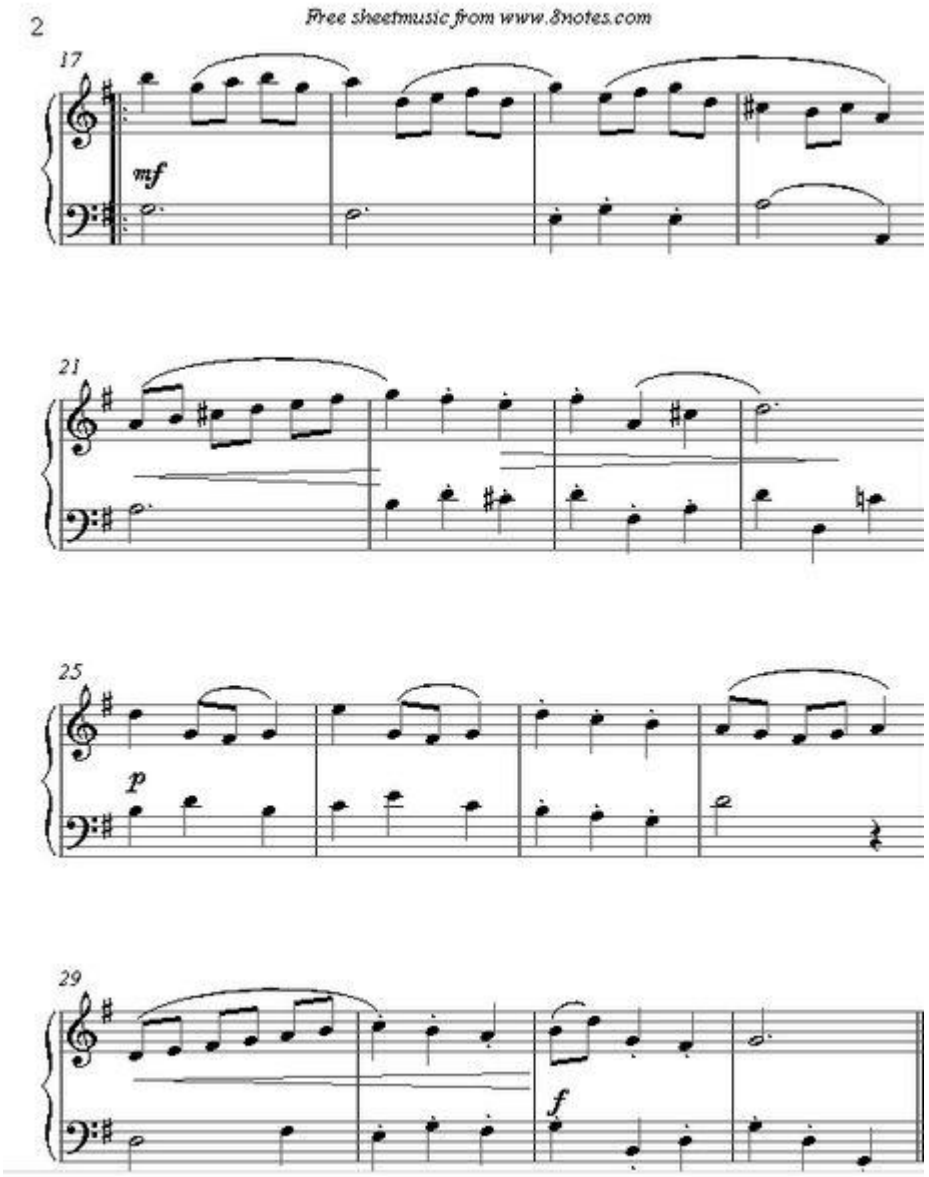
# Why visualization?

- Anscombe's Quartet: Four datasets

**Anscombes quarte**

| I | | II | | III | | IV | |
|---|---|---|---|---|---|---|---|
| *x* | *y* | *x* | *y* | *x* | *y* | *x* | *y* |
| 10.0 | 8.04 | 10.0 | 9.14 | 10.0 | 7.46 | 8.0 | 6.58 |
| 8.0 | 6.95 | 8.0 | 8.14 | 8.0 | 6.77 | 8.0 | 5.76 |
| 13.0 | 7.58 | 13.0 | 8.74 | 13.0 | 12.74 | 8.0 | 7.71 |
| 9.0 | 8.81 | 9.0 | 8.77 | 9.0 | 7.11 | 8.0 | 8.84 |
| 11.0 | 8.33 | 11.0 | 9.26 | 11.0 | 7.81 | 8.0 | 8.47 |
| 14.0 | 9.96 | 14.0 | 8.10 | 14.0 | 8.84 | 8.0 | 7.04 |
| 6.0 | 7.24 | 6.0 | 6.13 | 6.0 | 6.08 | 8.0 | 5.25 |
| 4.0 | 4.26 | 4.0 | 3.10 | 4.0 | 5.39 | 19.0 | 12.50 |
| 12.0 | 10.84 | 12.0 | 9.13 | 12.0 | 8.15 | 8.0 | 5.56 |
| 7.0 | 4.82 | 7.0 | 7.26 | 7.0 | 6.42 | 8.0 | 7.91 |
| 5.0 | 5.68 | 5.0 | 4.74 | 5.0 | 5.73 | 8.0 | 6.89 |

Table 1.1: Anscombe's quartet: four different datasets.



# Why visualization?

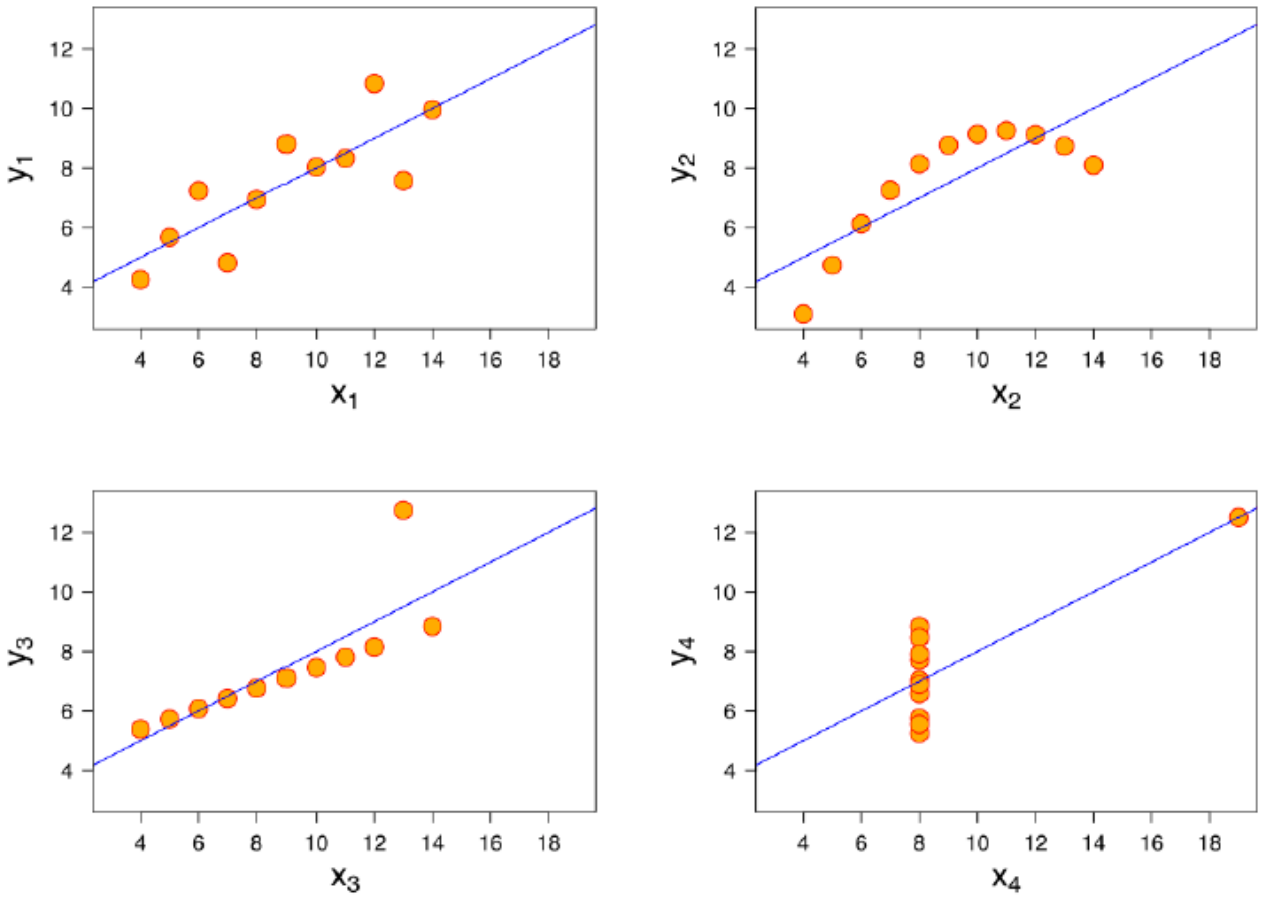
- Anscombe's Quartet: Statistics

| **Property (in each set)** | **Value** |
|---|---|
| Mean of x | 9.0 |
| Variance of x | 10.0 |
| Mean of y | 7.50 |
| Variance of y | 3.75 |
| Correlation between x and y | 0.898 |
| Linear regression line | $y = 0.5x + 3.0$ |

Table 1.2: Same statistics in Anscombe's quartet.

# Why visualization?

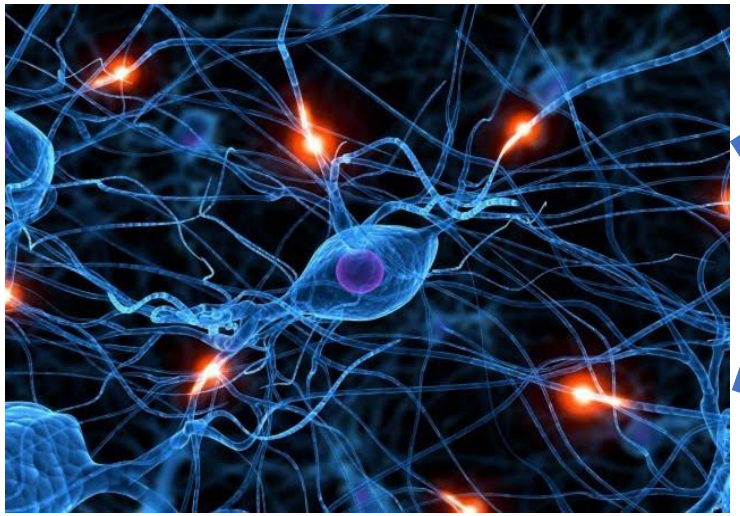
- Anscombe's Quartet: Statistics



# Visualization Advantages

- Data Analysis: Remarkable Progress

Deep Learning



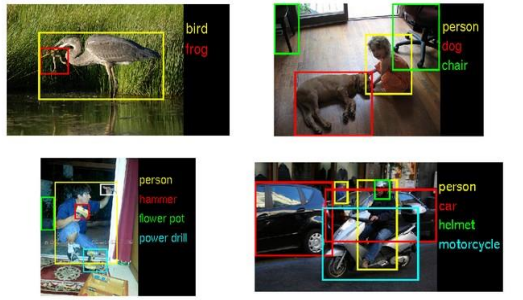
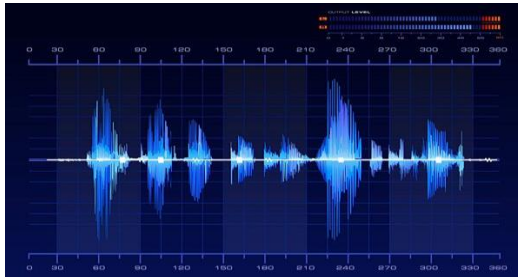
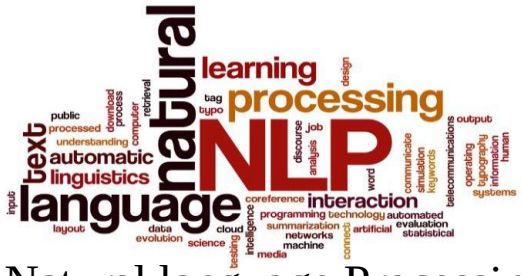



Image Recognition



Speech Recognition



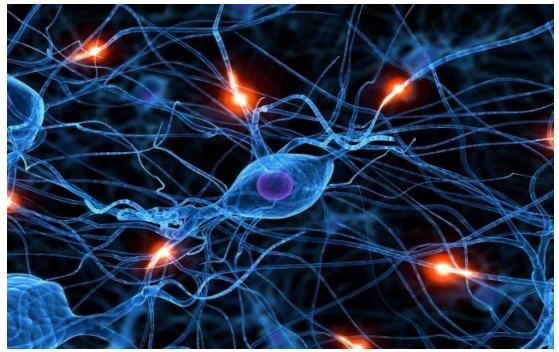

Natural language Processing

...

...

# Visualization Advantages

- Black Box
  - No clear understanding of the inner working mechanism
  - A substantial amount of trial-and-error procedures



Deep Learning


**Visualization**

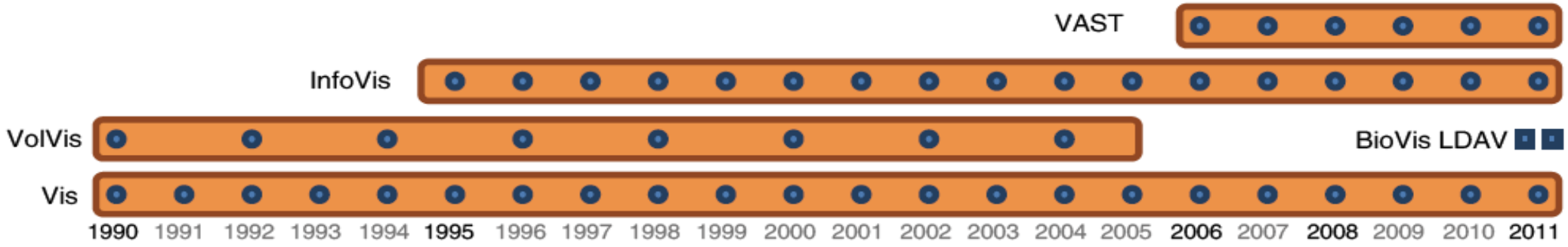
# Visualization Advantages

- Visualization on Big Data
  - Help understand deep learning models intuitively
  - Help train a better model efficiently
  - Make decisions more interpretable

**Visualization**  **Data Learning**

# Visualization is young

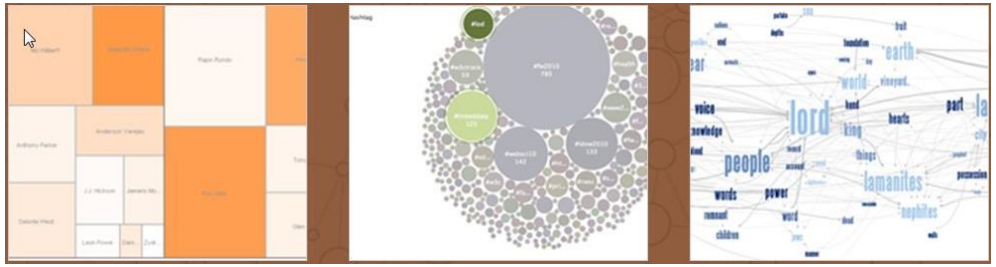
- VAST (Visual Analytics Science and Technology)
- InfoVis (Information Visualization)
- SciVis (Scientific Visualization)



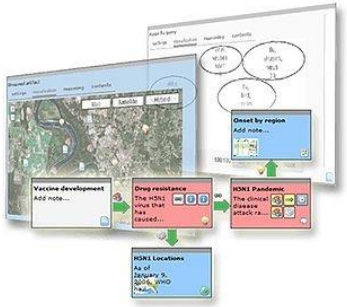
# Visualization Subfield

- Scientific Visualization (SciVis) – Spatial data

- Information Visualization (InfoVis)– Abstract data

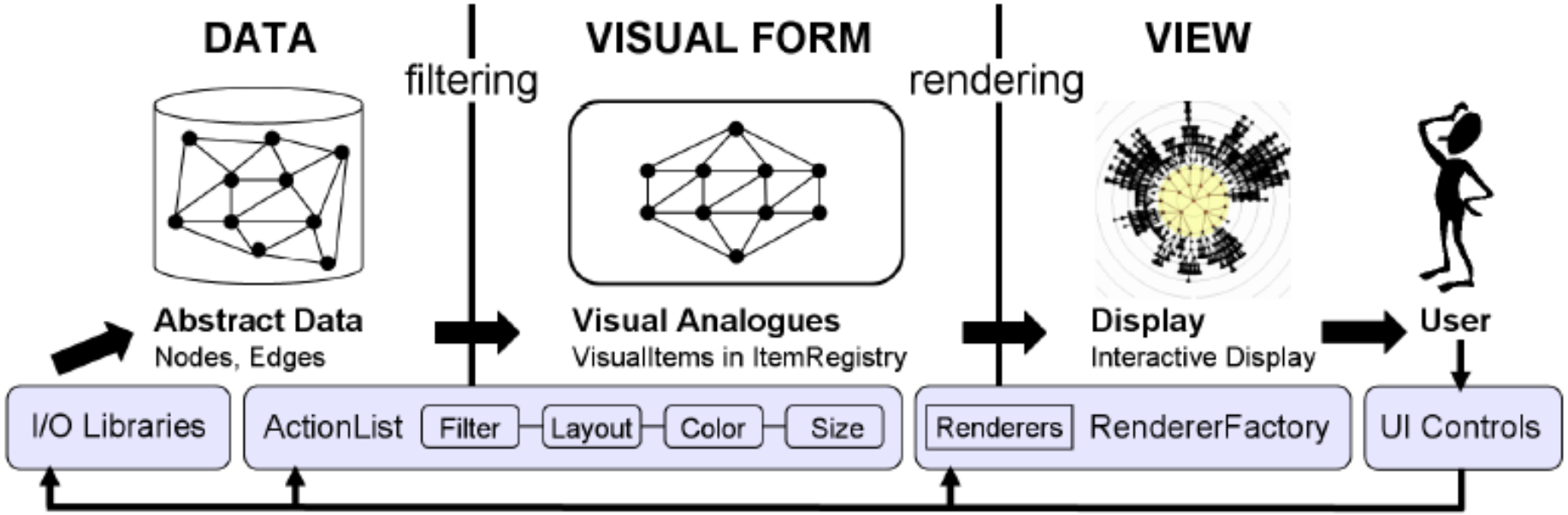



- Visual Analytics (VAST) – Analytical reasoning




# Visualization Process

- Visualization Pipeline




Engineering part  Visual design part

# What is Visualization research?

- Techniques/algorithms
- Applications
- Systems
- Evaluations
- Theory/models

# Visualization Taxonomy

- A taxonomy based on:
  - the challenges that learning methods faces
  - the purposes that visualization techniques serve

Challenges on Learning:

How a learning model works?

How to improve a learning model?

Purposes of Visualization:

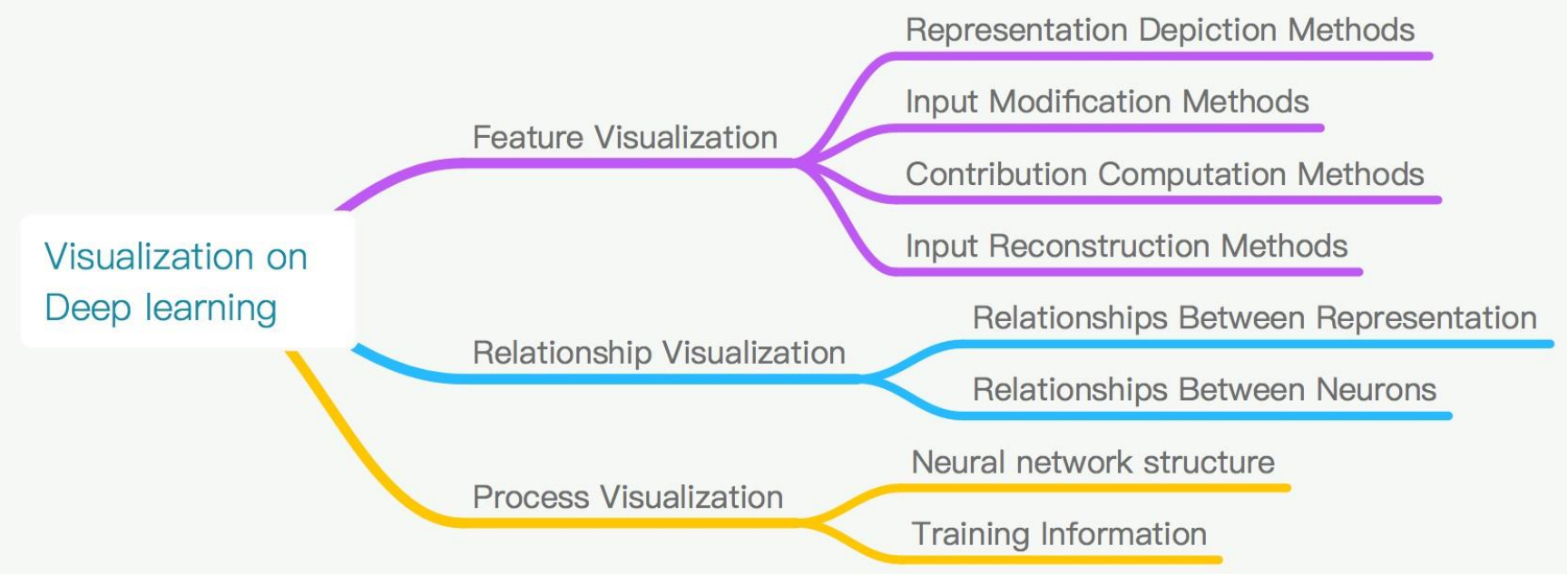
Visualize the features learned by a learning model

Visualize the relationships in a learning model

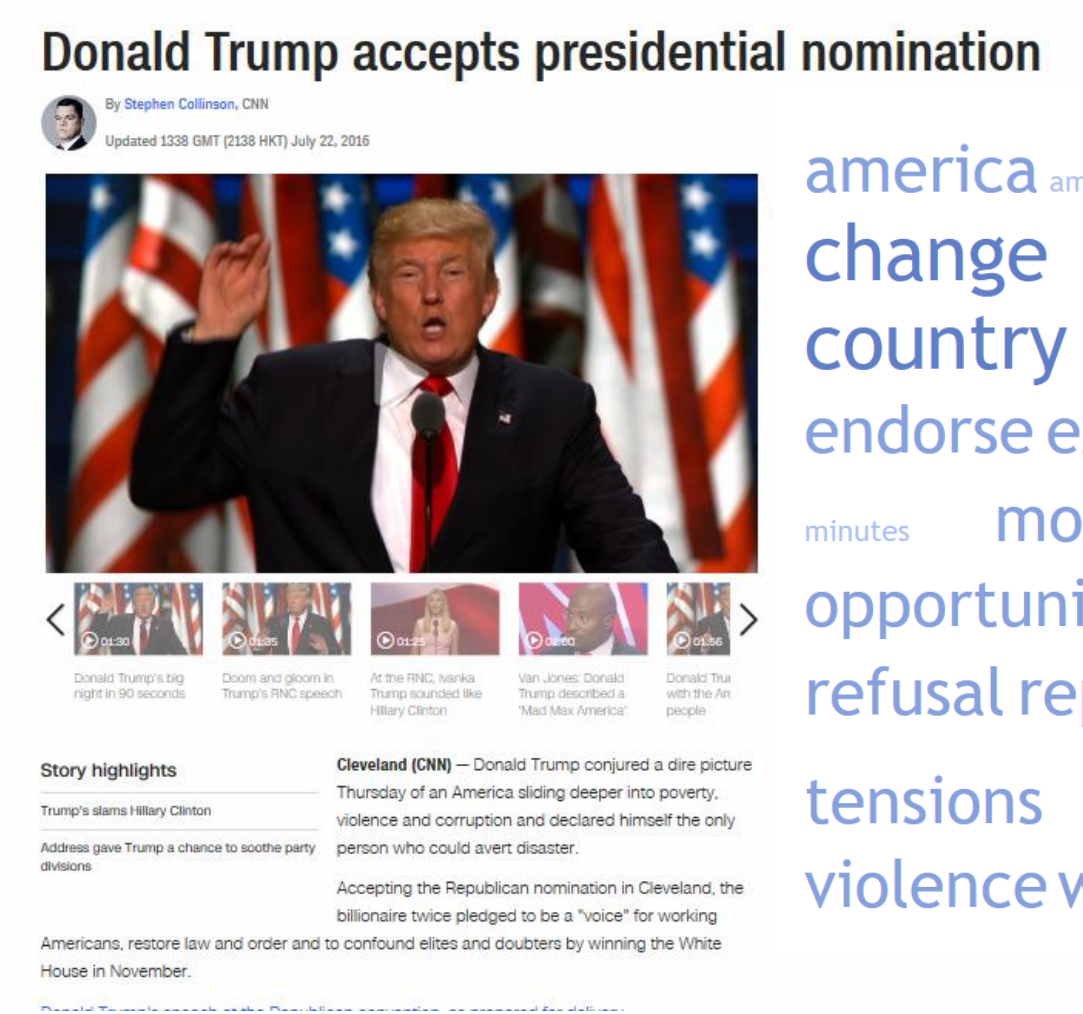
Visualize the whole process of a learning model

# Visualization Taxonomy

- A taxonomy based on:
  - the challenges that deep learning faces
  - the purposes that visualization techniques serve



# Feature Visualization Example



## Donald Trump accepts presidential nomination

By Stephen Collinson, CNN

Updated 1338 GMT (2138 HKT) July 22, 2016

Donald Trump's big night in 90 seconds

Doom and gloom in Trump's RNC speech

At the RNC, Ivanka Trump sounded like Hillary Clinton

Van Jones: Donald Trump described a 'Mad Max America'

Donald Tru with the An people

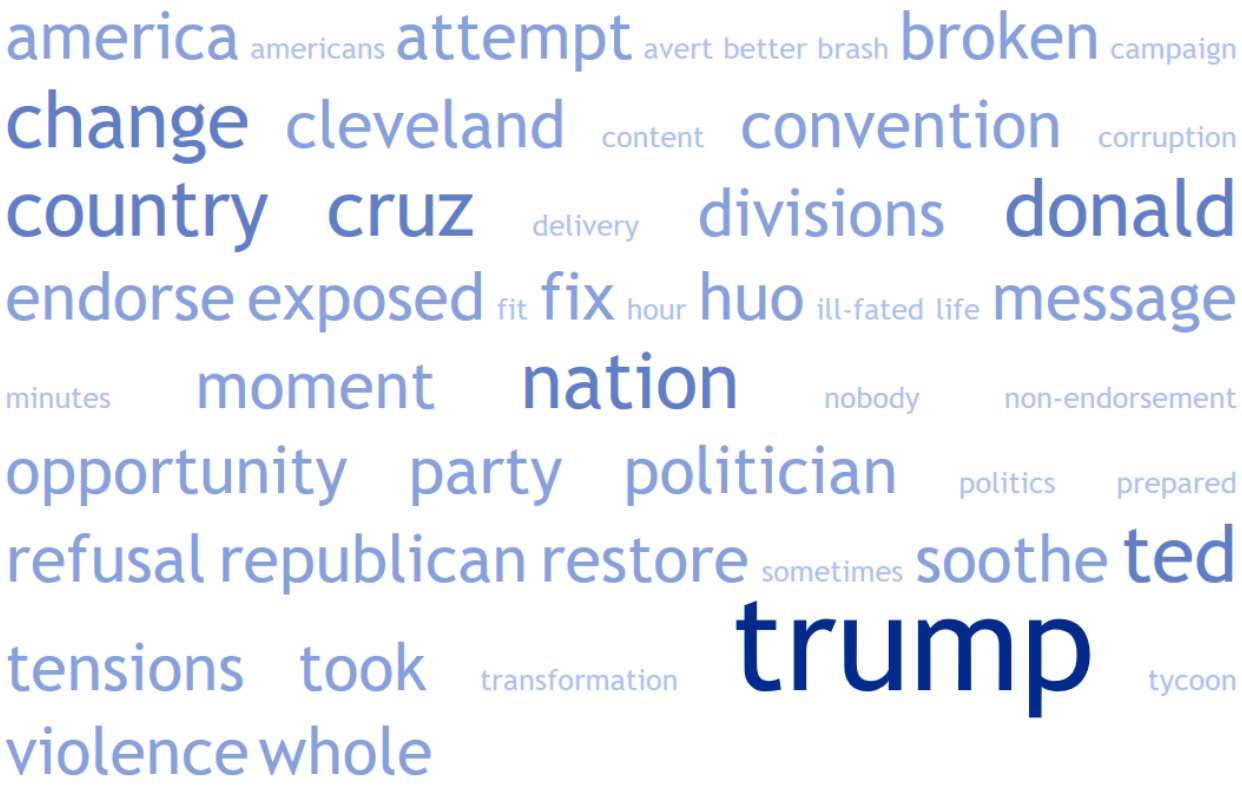
**Story highlights**

Trump's slams Hillary Clinton

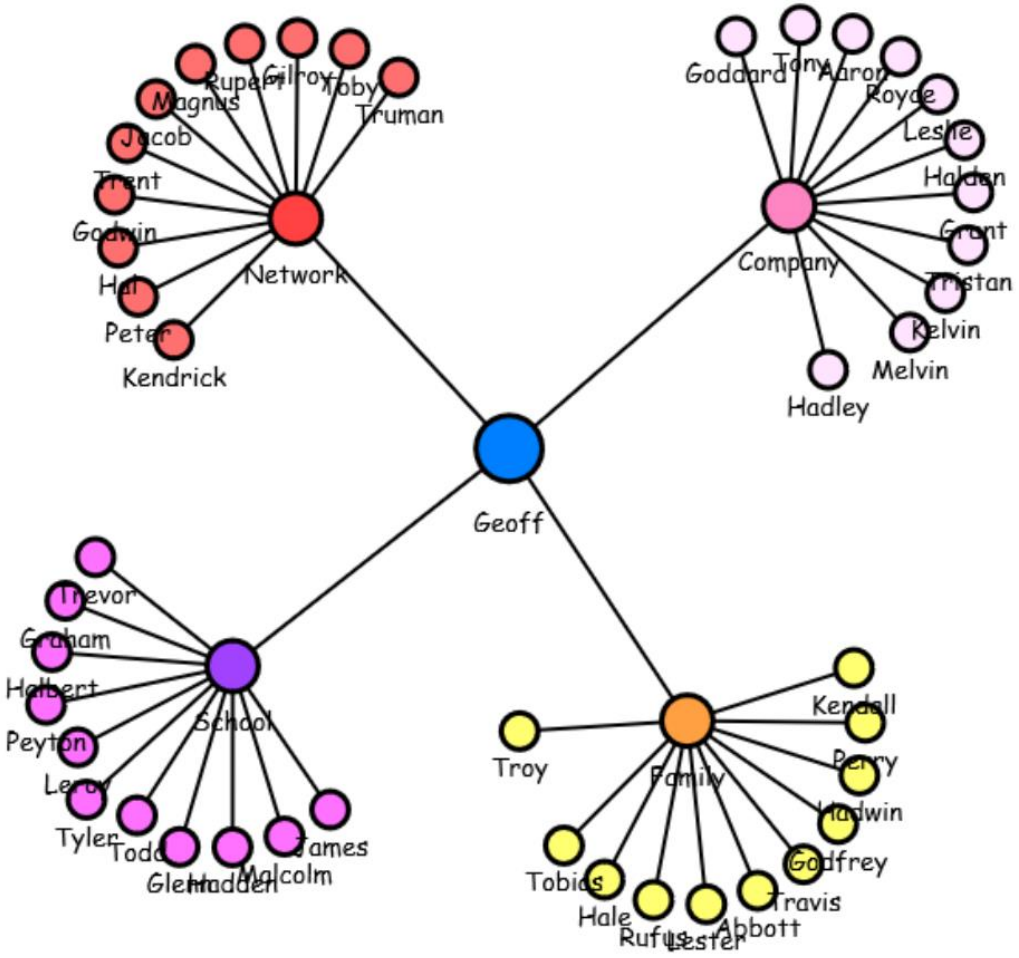
Address gave Trump a chance to soothe party divisions

**Cleveland (CNN)** — Donald Trump conjured a dire picture Thursday of an America sliding deeper into poverty, violence and corruption and declared himself the only person who could avert disaster.

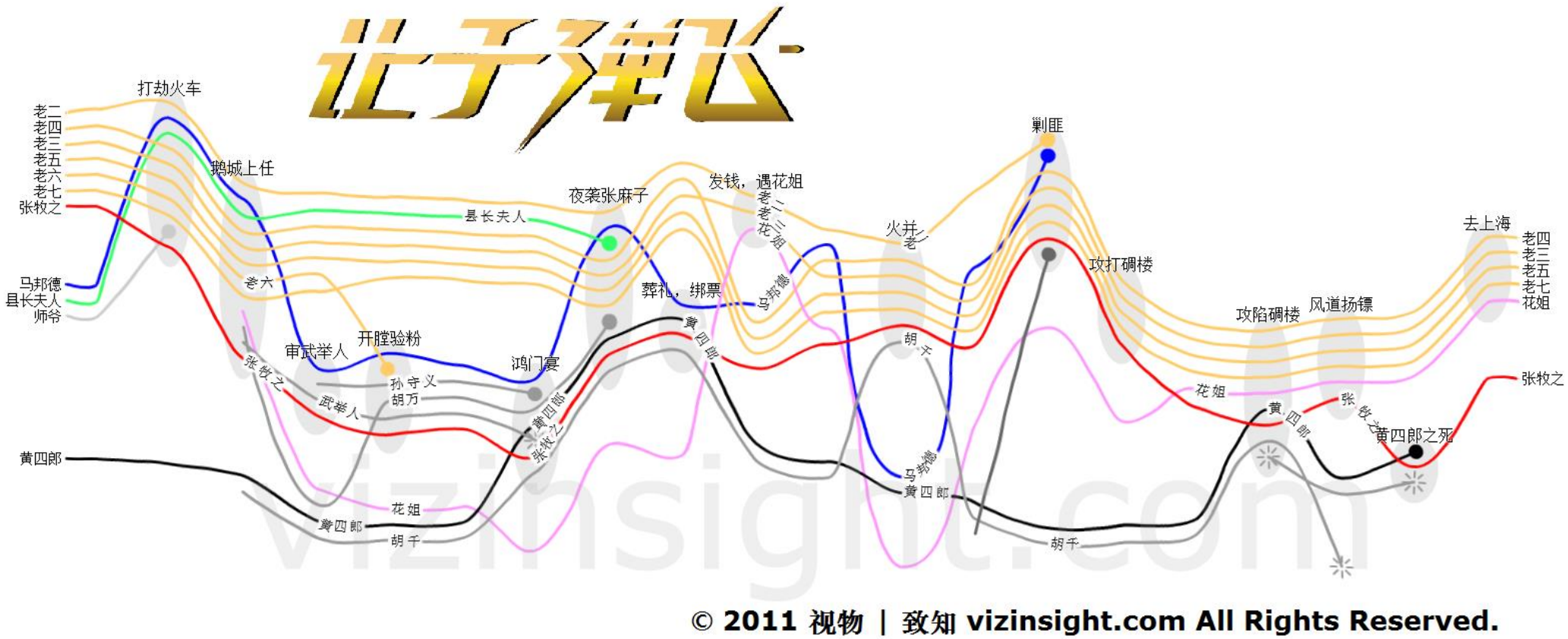
Accepting the Republican nomination in Cleveland, the billionaire twice pledged to be a "voice" for working Americans, restore law and order and to confound elites and doubters by winning the White House in November.



# Relationship Visualization Example



# Process Visualization Example



# Challenges for Visualization Research

- Better Scalability
  - ➢ Better Visualization
    - ○ Better overview & summarization
    - ○ Better data reduction
    - ○ Better visual encoding
    - ○ Better user interaction
    - ○ ….

# Visualization Tools

python

Google Developers CHARTS

Timeline JS

Beautifully crafted timelines that are easy and intuitive to use.