

The background features a large, faint logo of BK TP.HCM. It consists of a central white hexagon with the letters 'BK' in a large, bold, sans-serif font, and 'TP.HCM' in a smaller font below it. This central hexagon is surrounded by six other hexagons in shades of light blue and purple, arranged in a circular pattern.

# **SEMINAR**

# **BIG DATA VISUALIZATION**

Thái Hồ Phú Hào – 1670219

Trần Mạnh Kha – 13070237

Nguyễn Như Hải – 1670218

Nguyễn Thị Huyền – 1670225

# TÀI LIỆU THAM KHẢO

---

- [1] PROFESSOR CHING-YUNG LIN.: Slide Big Data Visualization, Columbia University**
- [2] Nguyen Thanh Tan and Insu Song.: Big Data Visualization. In ICISA 2016, 399-408**
- [3] Bauer, M.I., Johnson-Laird, P.N.: How diagrams can improve reasoning. Psychological Science 4(6), 372–378 (1993)**
- [4] Larkin, J.H., Simon, H.A.: Why a diagram is (sometimes) worth ten thousand words. Cognitive Science 11(1), 65–100 (1987)**
- [5] Mayer, R.E., Gallini, J.K.: When is an illustration worth ten thousand words? Journal of Educational Psychology 82(4), 715 (1990)**
- [6] Card, S.K., Mackinlay, J.D., Shneiderman, B.: Readings in information visualization: using vision to think. Morgan Kaufmann (1999)**
- [7] Colin, W.: Information visualization: perception for design. Morgan Kaufmann, San Francisco (2004)**
- [8] Ma, K.-L., Stompel, A., Bielak, J., Ghattas, O., Kim, E.J.: Visualizing very large-scale earthquake simulations. In: 2003 ACM/IEEE Conference on Supercomputing, pp. 48–48. IEEE (2003)**
- [9] Yi, J.S., ah Kang, Y., Stasko, J.T., Jacko, J.A.: Toward a deeper understanding of the role of interaction in information visualization. IEEE Transactions on Visualization and Computer Graphics 13(6), 1224–1231 (2007)**
- [10] Lamping, J., Rao, R., Pirolli, P.: A focus+ context technique based on hyperbolic geometry for visualizing large hierarchies. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 401–408. ACM Press/Addison-Wesley Publishing Co. (1995)**

# Nội dung

---

## 1. Giới thiệu

1.1 Hình dung là gì?

1.2 Tại sao chúng ta tạo sự hình dung ?

1.3 Các kỹ thuật hình dung hiện tại

## 2. Big Data Visualization

2.1 Thách thức

2.2 Kỹ thuật

## 3. Làm thế nào chúng ta có thể hình dung được dữ liệu lớn

3. 1 Kỹ thuật chính

3.2 Công cụ mã nguồn mở

3.3 Ví dụ

## 4. Phân tích trực quan dữ liệu lớn

---

# Giới thiệu



# 1. Giới thiệu

---

**Hình dung là gì?**



# Làm thế nào chúng ta có thể có được thông tin?

---

**Listen**



**Taste &  
Smell**



**Touch**



**Look**



# Tín hiệu quả ?

Tín hiệu hóa học

Tín hiệu vật lý

Taste &  
Smell



Touch



**Không thể ước lượng bằng lý thuyết thông tin**

# Tín hiệu quả ?

---

Tín hiệu âm thanh

Listen



Tín hiệu ánh sáng

Look



"Trực quan hoá thông tin, Nhận thức về Thiết kế"



# Tại sao hiệu quả?

## Tư duy sáng tạo

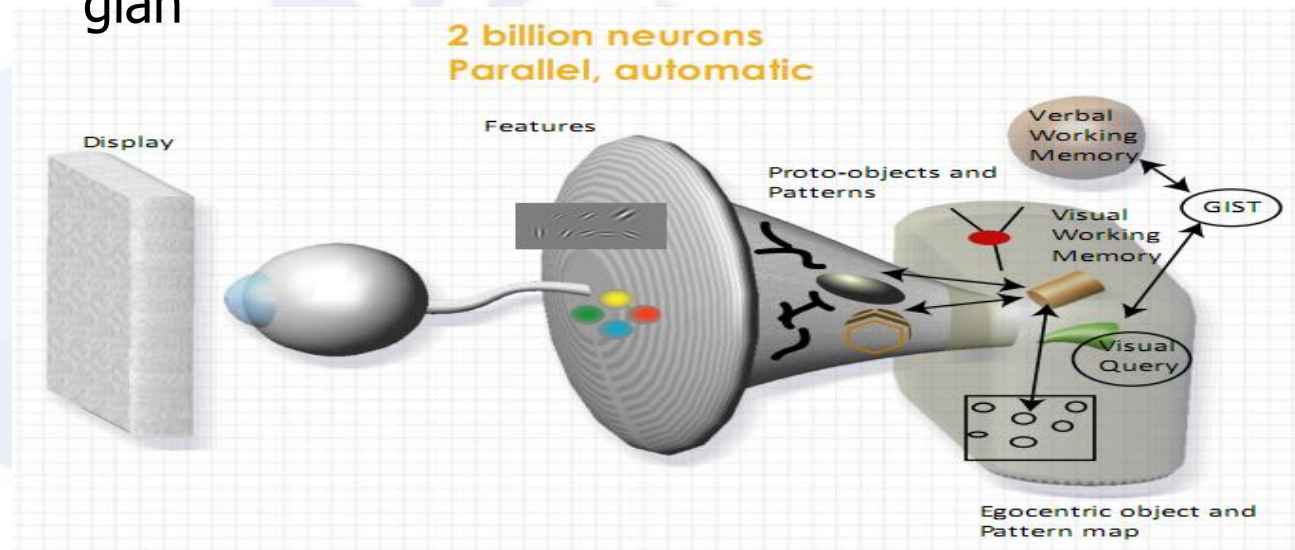
Xử lý song song để trích xuất các thuộc tính hình ảnh ở mức thấp như màu sắc, hình dạng, v.v ...

**Giai đoạn 1** → **Giai đoạn 2** → **Giai đoạn 3**

Sự phát hiện sớm, song song của màu sắc, kết cấu, hình dạng, thuộc tính không gian

Chia trường thị giác thành các vùng và các mẫu đơn giản

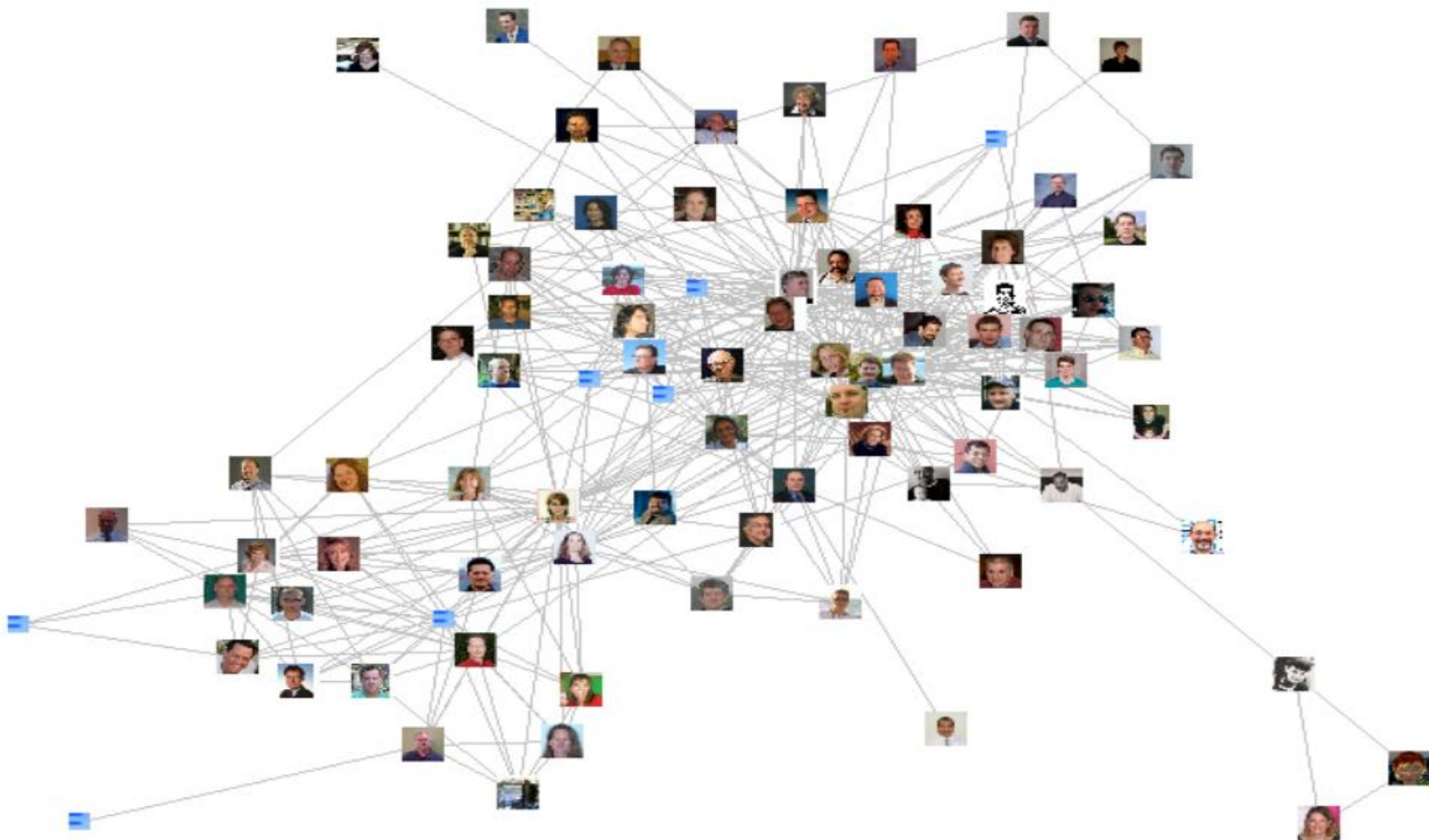
Giữ đối tượng trong bộ nhớ làm việc theo yêu cầu của sự chú ý tích cực



# Ví dụ



# Ví dụ





## Ví dụ



## Hình dung được sử dụng để giúp lý luận và ra quyết định

# Minh họa thông tin là gì?

---

"The action or fact of visualizing; the power or process of forming a mental picture or vision of something not actually present to the sight; a picture thus formed."

*-- Oxford English Dictionary*

"... finding the artificial memory that best supports our natural means of perception."

*-- Bertin, 1983*

**The use of computer-supported, interactive, visual representations of abstract data to amplify cognition**

*-- Cart, Mackinlay, Shneiderman, 1999*

---

**Tại sao chúng ta tạo sự hình dung ?**



## 1.2 Tại sao chúng ta tạo sự hình dung ?

---

**Đếm số 3 trong đoạn văn sau:**

1235693234870452973467  
0378937043679709102539

# Tìm mẫu

---

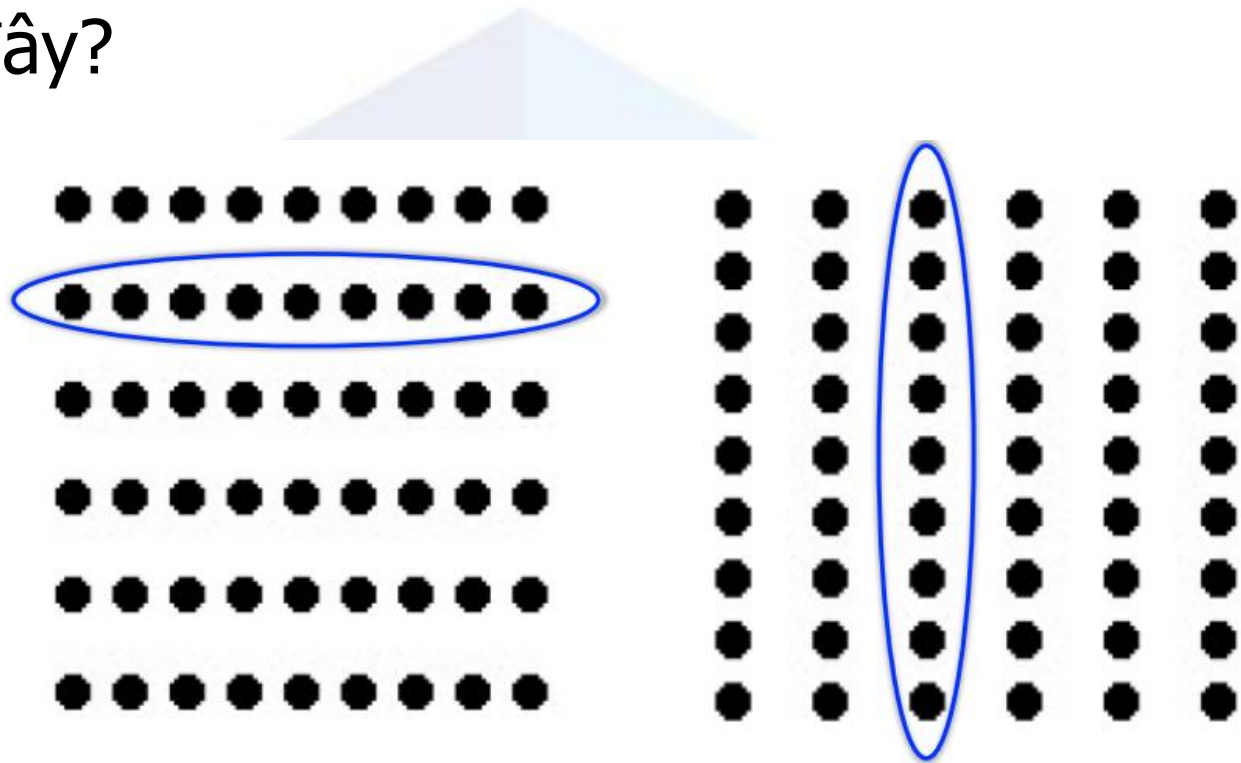
Đếm số 3 trong đoạn văn sau:

12**3**569**3**2**3**487045297**3**467  
0**3**789**3**704**3**6797091025**3**9



# Tìm mẫu

Bạn có thể xác định các nhóm các chấm trong các hình sau đây?



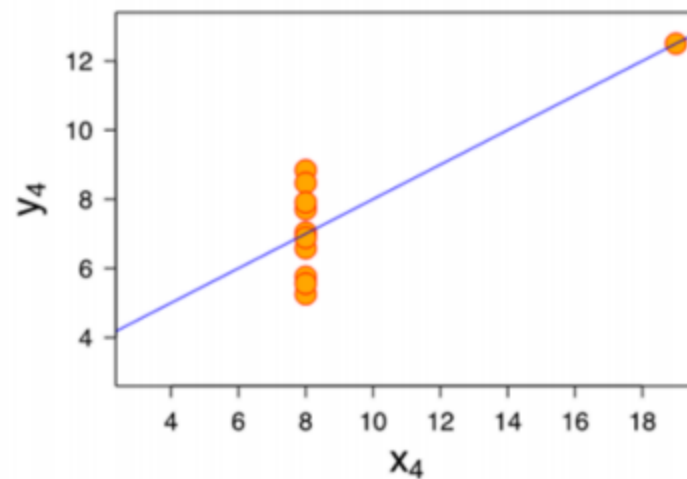
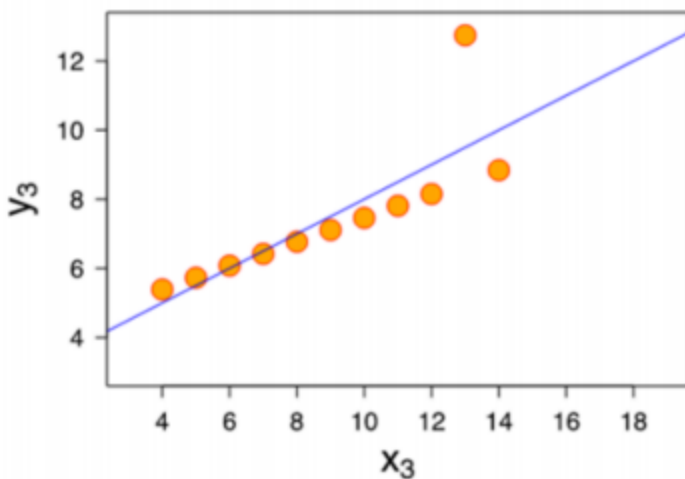
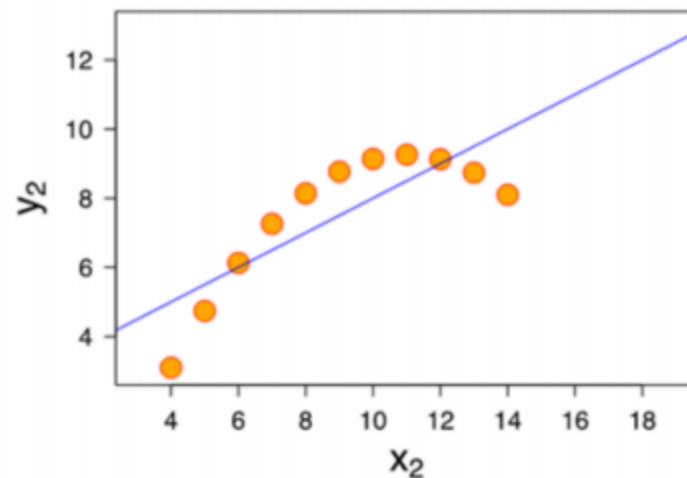
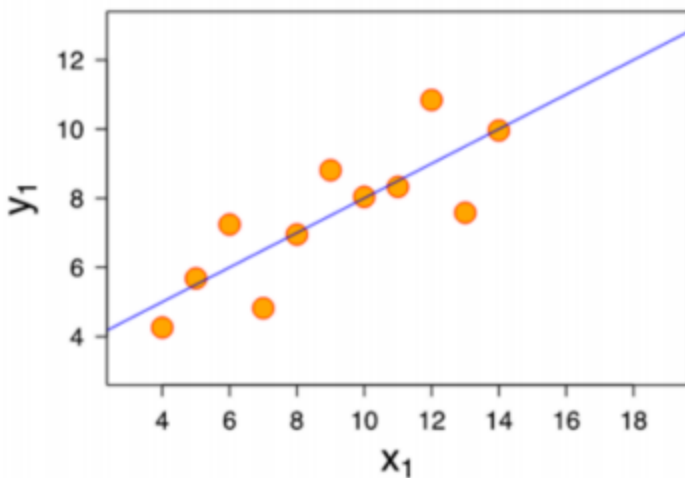
## Luật gần

Chúng ta có xu hướng nhóm các yếu tố gần nhau nhất

# 1.2 Tại sao chúng ta tạo sự hình dung ?

	Set A		Set B		Set C		Set D	
	X	Y	X	Y	X	Y	X	Y
0	10	8.04	10	9.14	10	7.46	8	6.58
1	8	6.95	8	8.14	8	6.77	8	5.76
2	13	7.58	13	8.74	13	12.74	8	7.71
3	9	8.81	9	8.77	9	7.11	8	8.84
4	11	8.33	11	9.26	11	7.81	8	8.47
5	14	9.96	14	8.10	14	8.84	8	7.04
6	6	7.24	6	6.13	6	6.08	8	5.25
7	4	4.26	4	3.10	4	5.39	19	12.50
8	12	10.84	12	9.13	12	8.15	8	5.56
9	7	4.82	7	7.26	7	6.42	8	7.91
10	5	5.68	5	4.74	5	5.73	8	6.89
mean	9.00	7.50	9.00	7.50	9.00	7.50	9.00	7.50
std	3.32	2.03	3.32	2.03	3.32	2.03	3.32	2.03
corr	0.82		0.82		0.82		0.82	
lin. reg.	$y = 3.00 + 0.500x$		$y = 3.00 + 0.500x$		$y = 3.00 + 0.500x$		$y = 3.00 + 0.500x$	

# Xem dữ liệu trong ngữ cảnh



# 1.2 Tại sao chúng ta tạo sự hình dung ?

## Một bức tranh đáng giá ngàn lời nói

News illustrated



### {GANGNAM STYLE!!! The 5 basic steps

The sudden explosion of a South Korean entertainer called Psy, has given the world Gangnam Style. It is setting the music and dance world on fire and has a set sequence. We simplify them for your perusal

#### ★ When to use the steps during the chorus ★

Step 1	Step 2
Oppa is Gangnam style, ahhhh... Gangnam style...	Oh, oh oh oh oh, Oppa is Gangnam style...
<b>Step 1</b> ahhhh... Gangnam style...	<b>Step 2</b> Oh, oh oh oh oh, Oppa is Gangnam style... Eeeehh- Sexy Lady...
<b>Step 1 or Step 2</b> (in the last chorus)	<b>Step 4</b> Oh, oh oh oh oh, Oppa is Gangnam style... Eeeehh- Sexy Lady oh oh oh oh. Oppa is Gangnam style.
	<b>Step 3 or Step 1</b> (in the last chorus)
	<b>Step 5</b> (only at the end)

#### Step 1

Riding the horse



Dress classy and dance cheesy

Cross your hands like taking the horse reins and pulse up and down

Do small jumps with your legs spread like you are riding a horse

Footsteps



#### Step 2

Lassoing the sexy lady



Lassoing motion with your right arm

Continue with the horse-riding movement

Footsteps



#### Step 3

Now everybody is looking at me



**A**  
Hands in pockets or waist and small hip side movements combined with the foot steps

Slight kick with the right leg. Alternate with small jumps with the left leg

**B**  
Finish this move dragging the right leg to the left leg.

Footsteps A

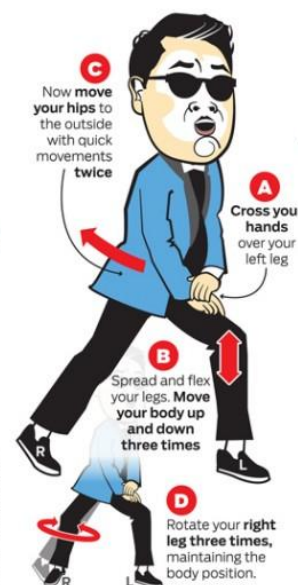


Footsteps B



#### Step 4

Combine a few 'sexy' moves



**C**  
Now move your hips to the outside with quick movements twice

**A**  
Cross your hands over your left leg

**B**  
Spread and flex your legs. Move your body up and down three times

**D**  
Rotate your right leg three times, maintaining the body position.

#### Step 5

Finish with a cool pose



Spread your arms and raise your right leg (position A). Now get down quickly on your right leg and flex the left one. Now rotate your right arm and with your hand touch your chin doing a "L" shape with your thumb and index fingers (position B)

Source: You Tube

HUGO A. SANCHEZ@Gulf News

# Một số lý do khác

---

- Xem dữ liệu trong ngữ cảnh
- Tìm mẫu
- Kể một câu chuyện
- Thu hút sự chú ý
- Giao tiếp với người khác
- Tóm tắt và giải thích
- Tính toán đồ họa
- Chi tiêu bộ nhớ
- Truyền cảm hứng cho mọi người

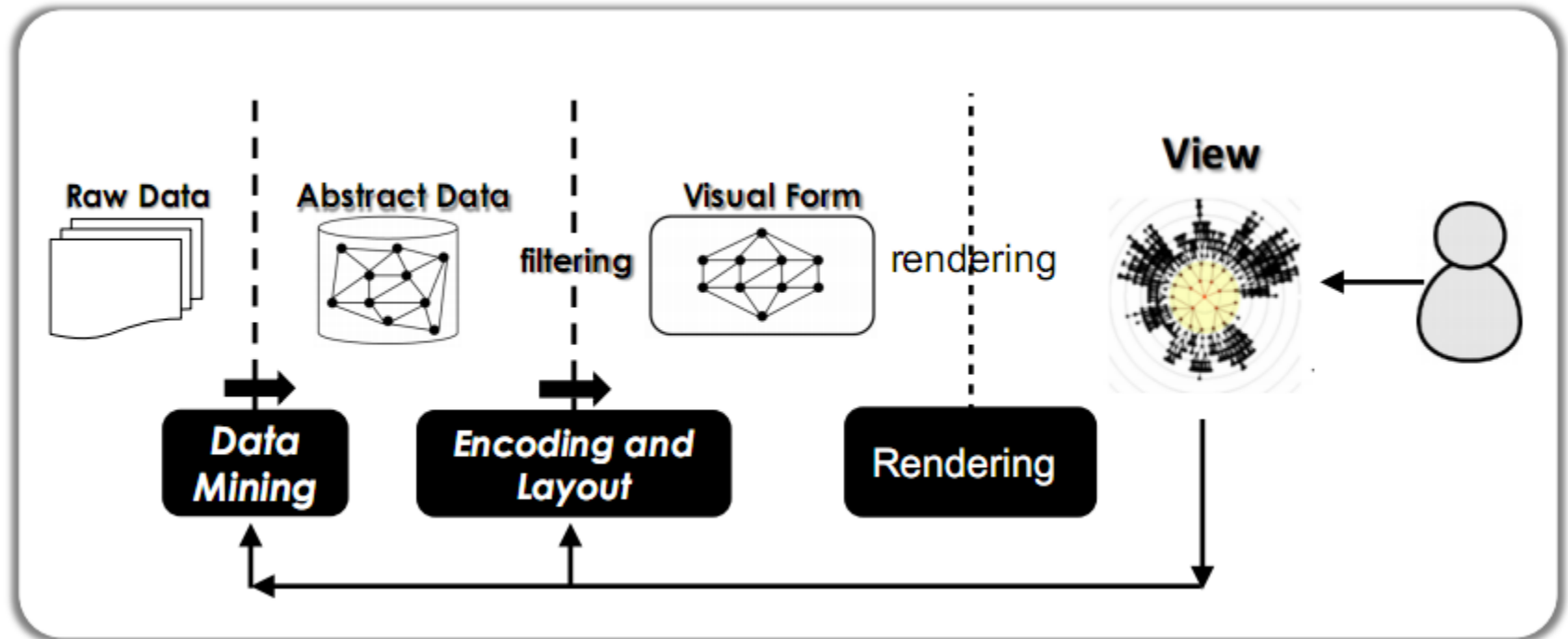
# 1. Giới thiệu

---

**Các kỹ thuật hiện hình hiện tại**

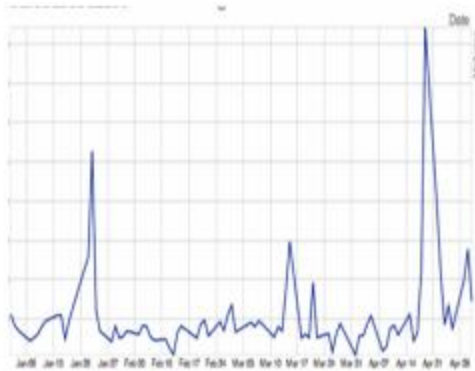


# Visualization & Visual Analysis Reference Model





# Phân loại theo loại dữ liệu



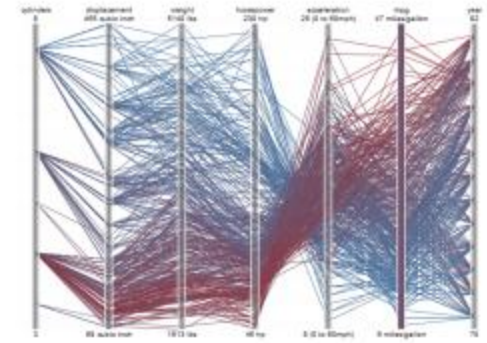
1D



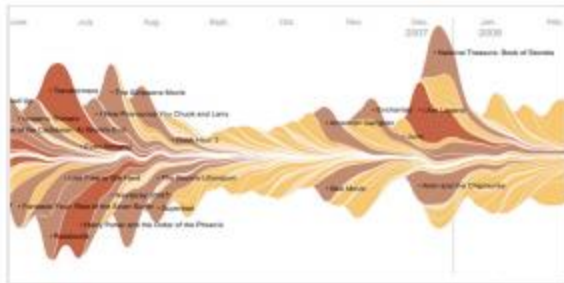
2D



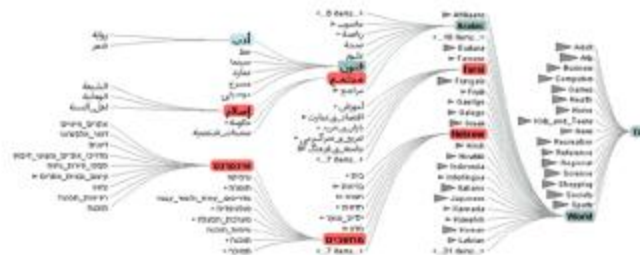
3D



Multi-D



Temporal



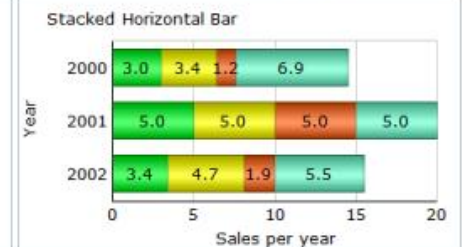
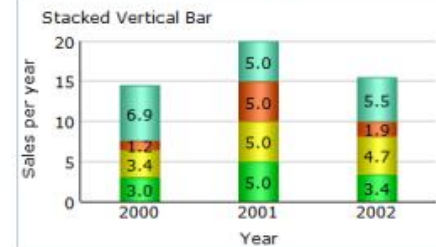
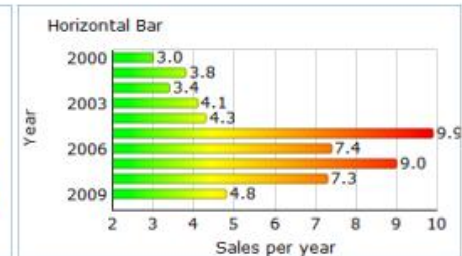
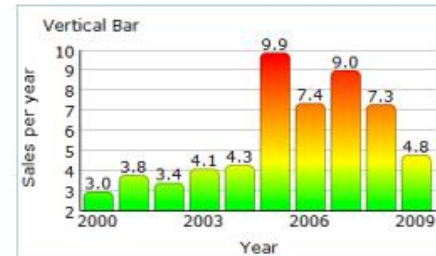
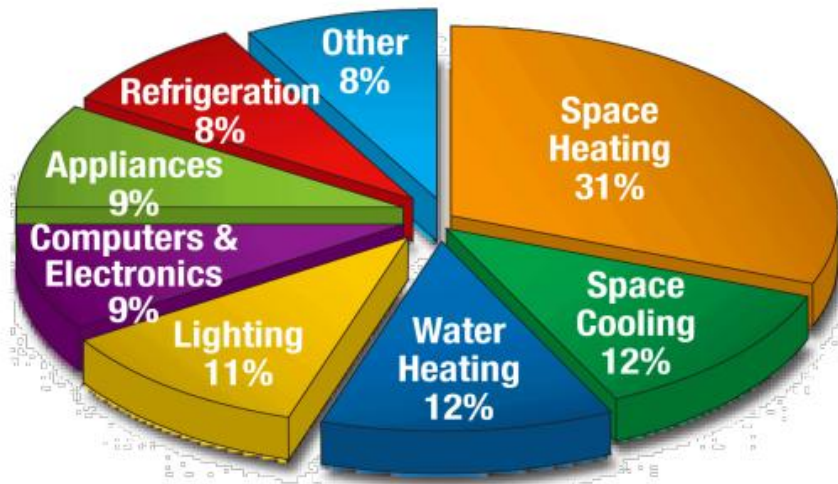
Tree



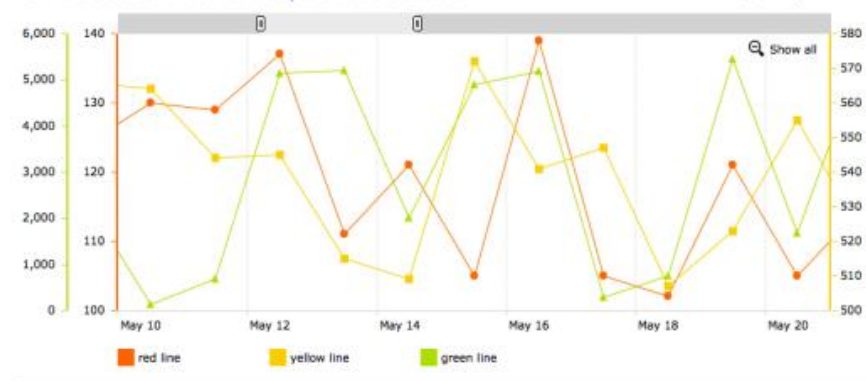
Graph



# Ví dụ: Hình dung dữ liệu số 1D



Line chart with multiple value axes



# Ví dụ: Dữ liệu 2D

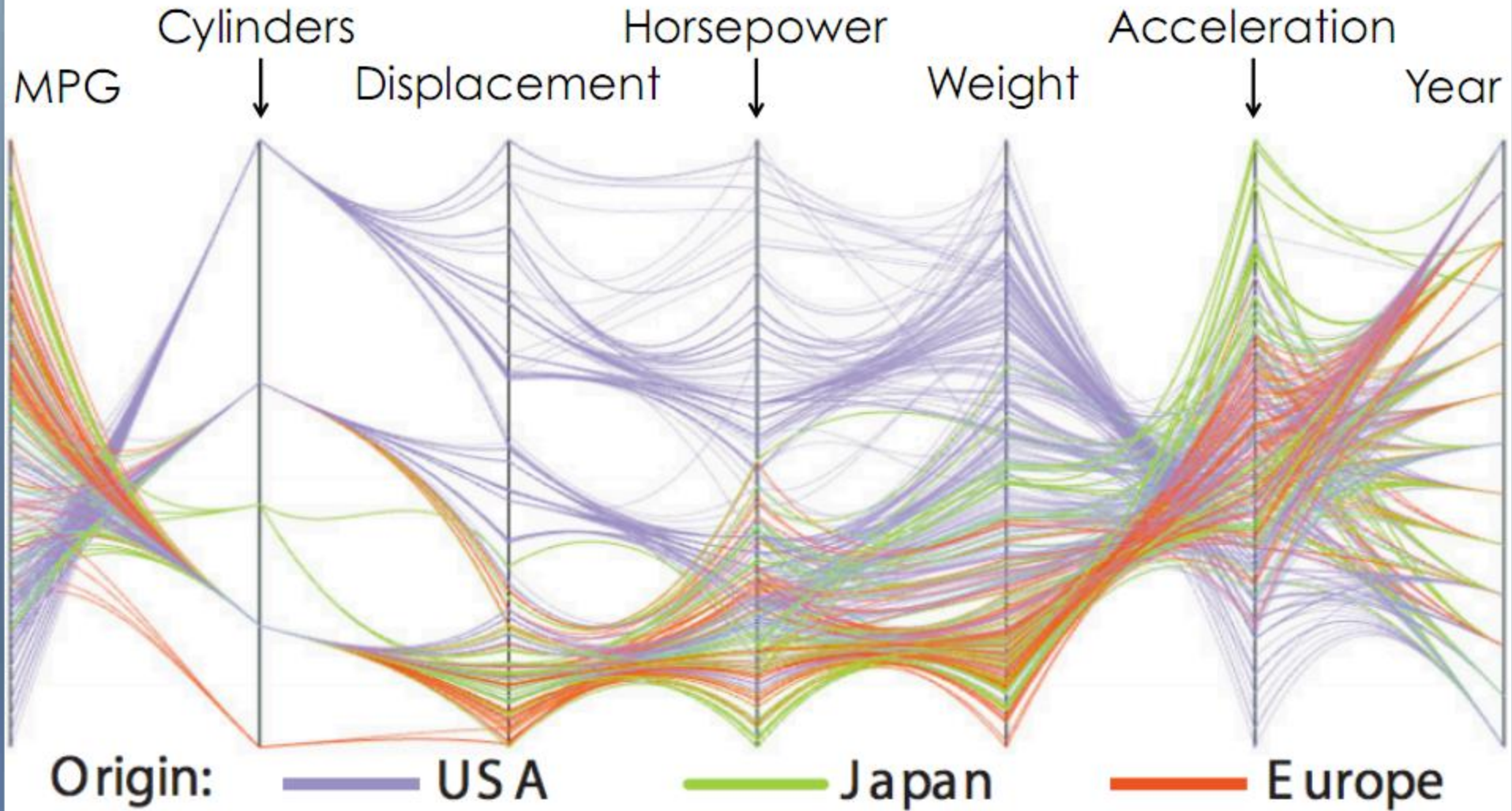


Kích thước của mỗi ô: Giá trị của thị trường chứng khoán

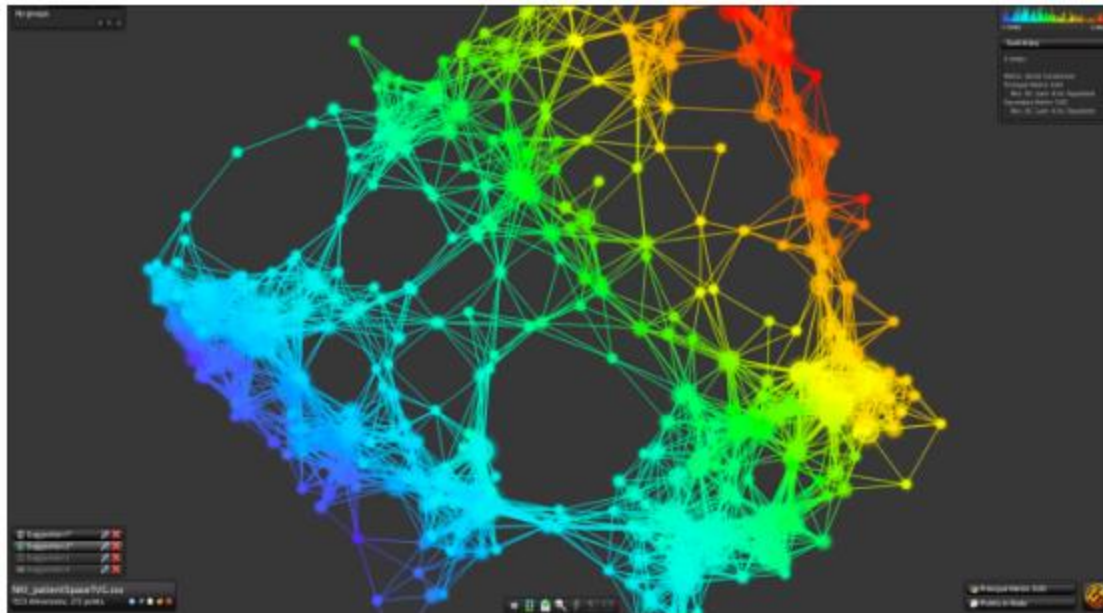
Màu sắc: Thay đổi Cổ phiếu

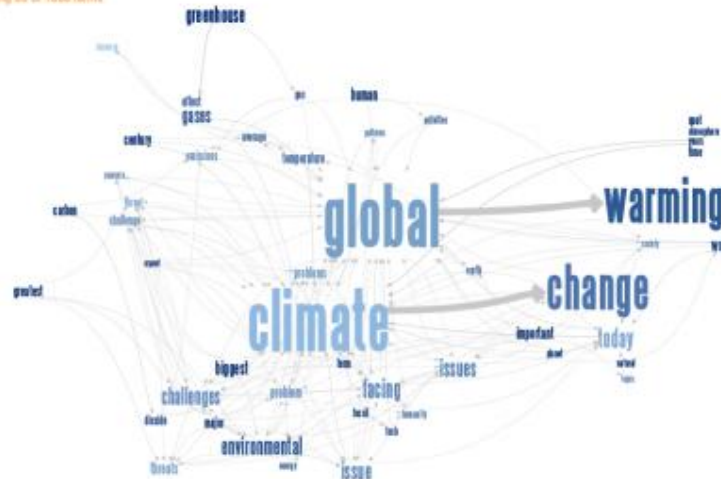
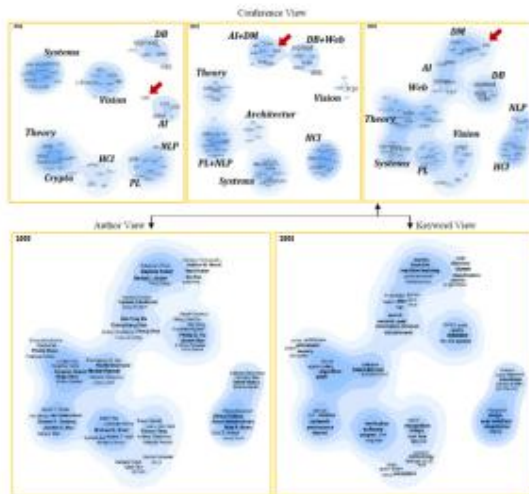
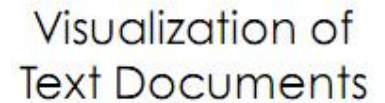


# Ví dụ: Dữ liệu Đa chiều



## Ví dụ: Hình dung dữ liệu cấu trúc



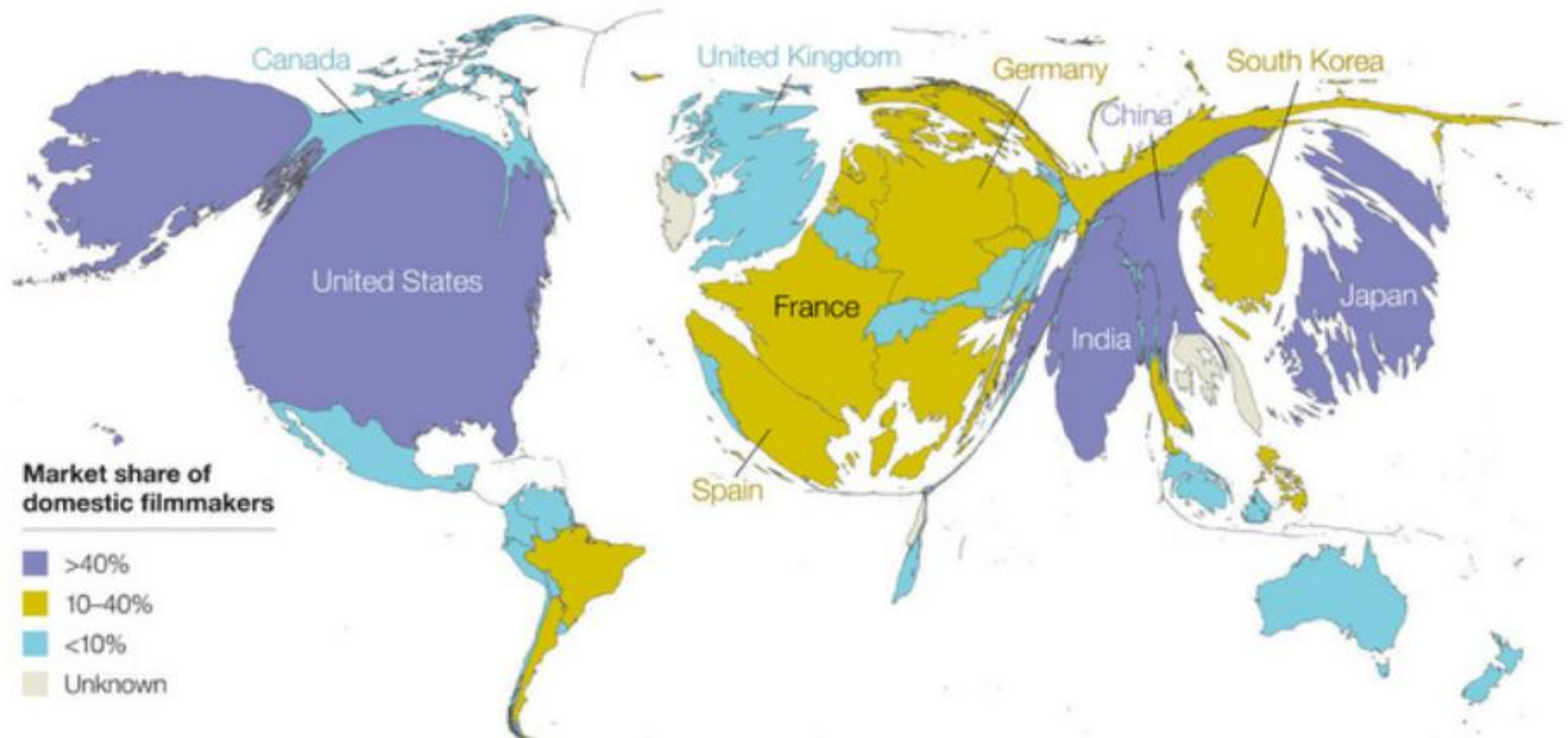




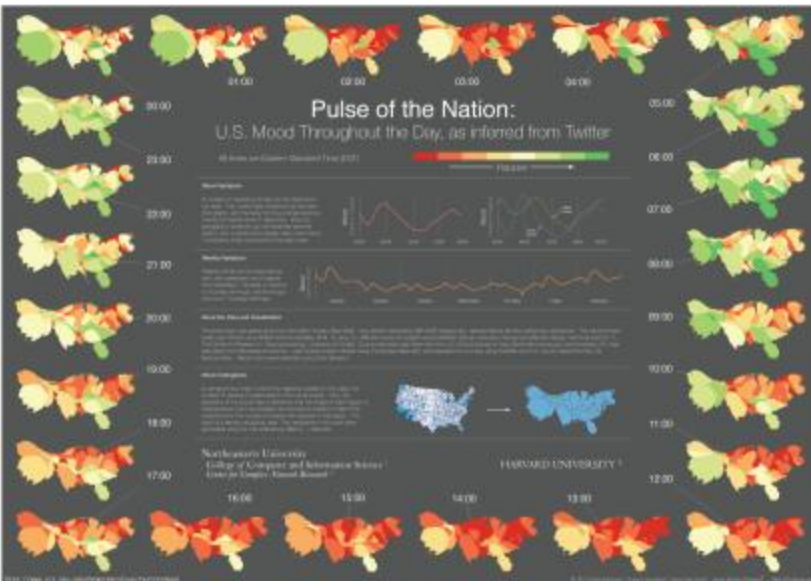
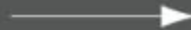
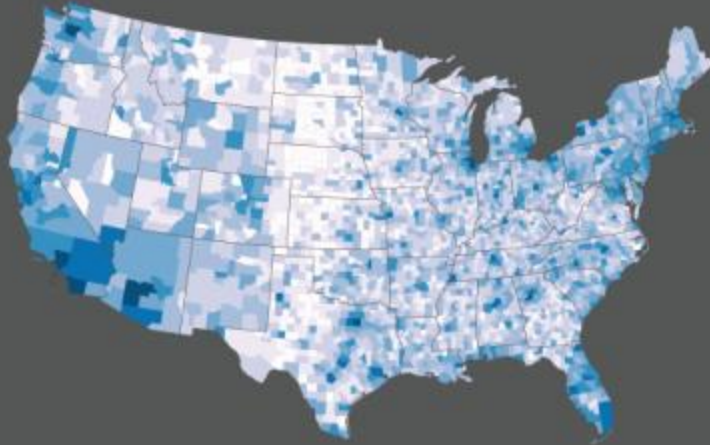
# Ví dụ: Không gian địa lý

**Larger cinema markets support stronger domestic film industries.**

Countries sized by relative share of worldwide box office revenue, 2009



# Ví dụ: Hình dung dữ liệu thời gian không gian



**Pulse of the Nation:**  
U.S. Mood Throughout the Day inferred from Twitter

Less Happy  More Happy

<http://www.ccs.neu.edu/home/amislove/twittermood>

# Ví dụ: Hình dung dữ liệu thời gian không gian

## wind map

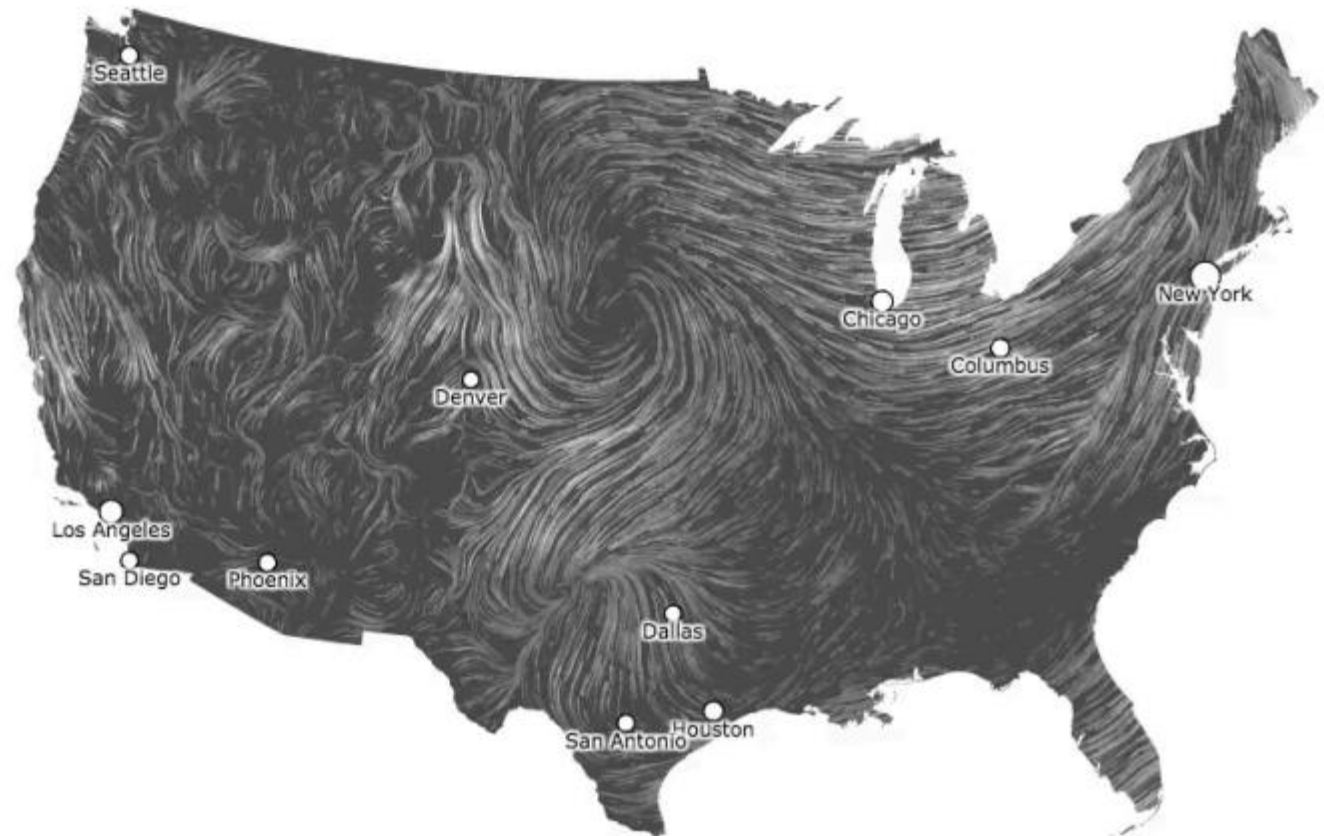
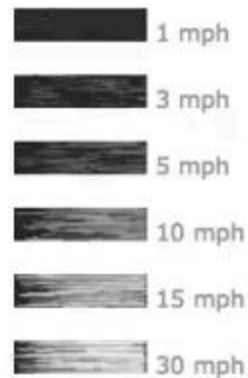
**Dec. 3, 2014**

11:35 am EST

(time of forecast download)

top speed: **31.5 mph**

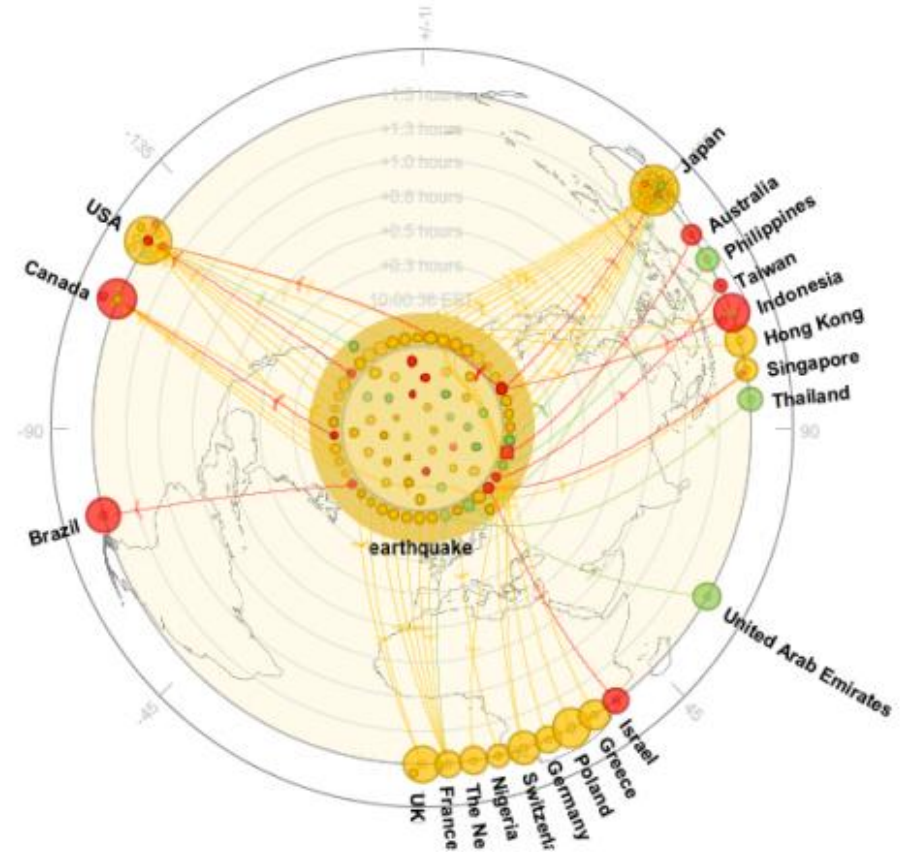
average: **8.2 mph**



<http://hint.fm/wind/>

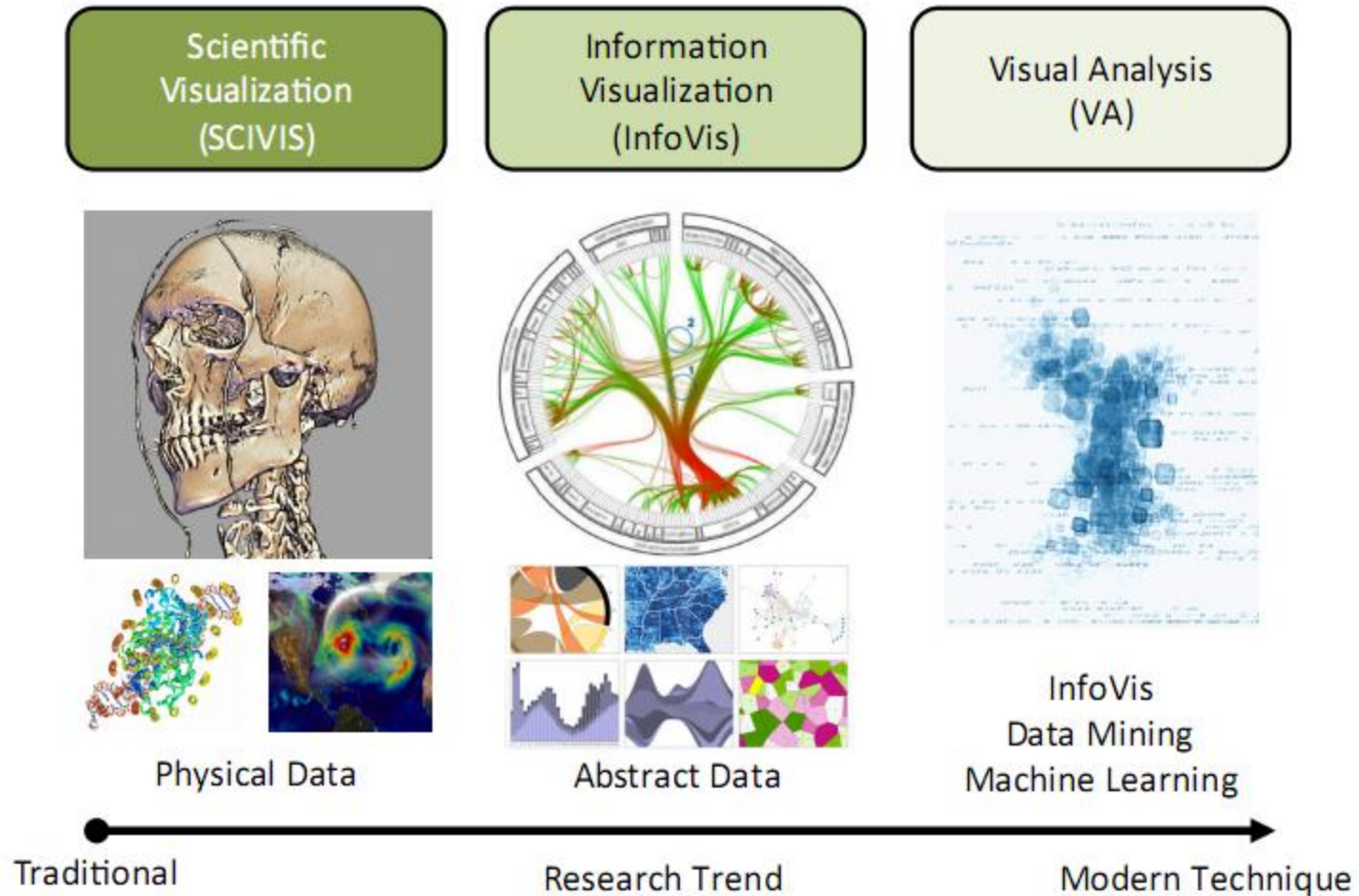


# Hình dung không chỉ là một bức tranh đẹp



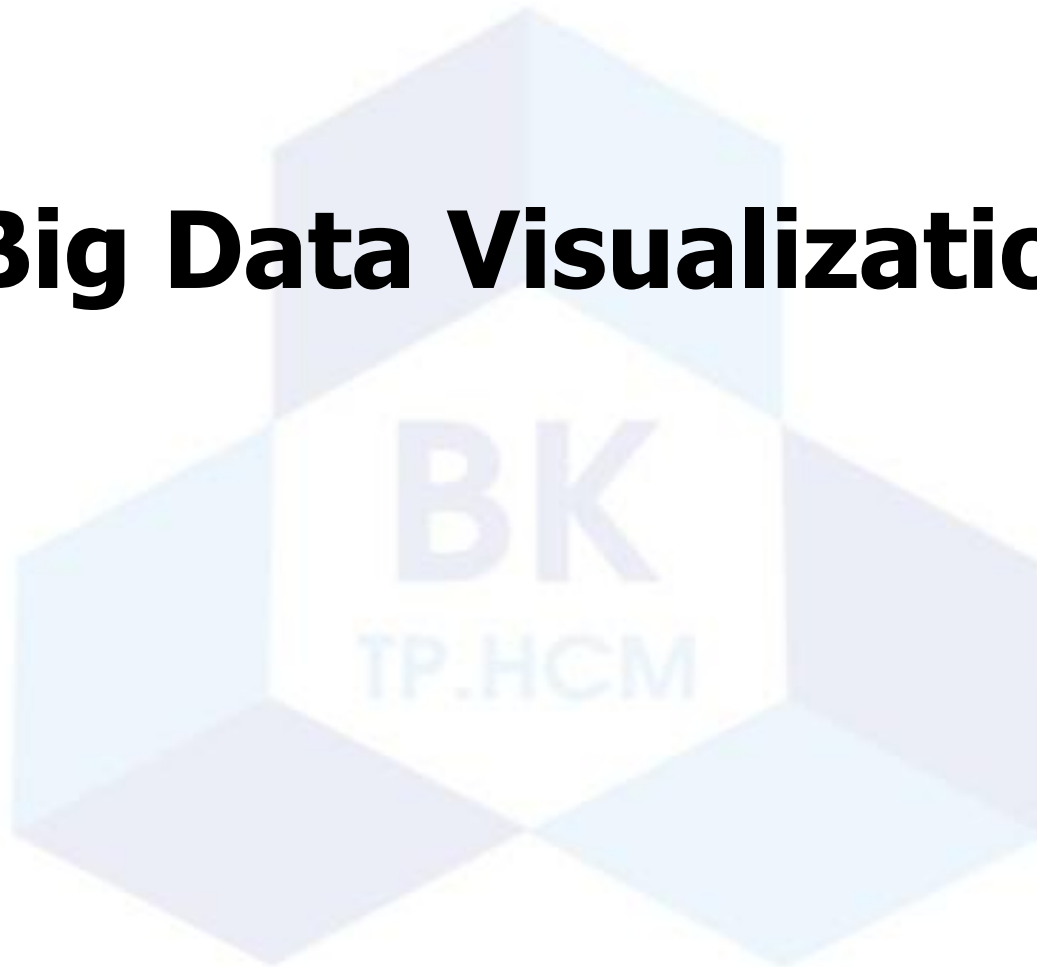
**Mục đích của hình dung là để lộ ra cái nhìn sâu sắc của dữ liệu**

# InfoVis v.s. Scientific Visualization



---

# Big Data Visualization



# Big Data Visualization

---



340 triệu lượt xem trong một ngày!



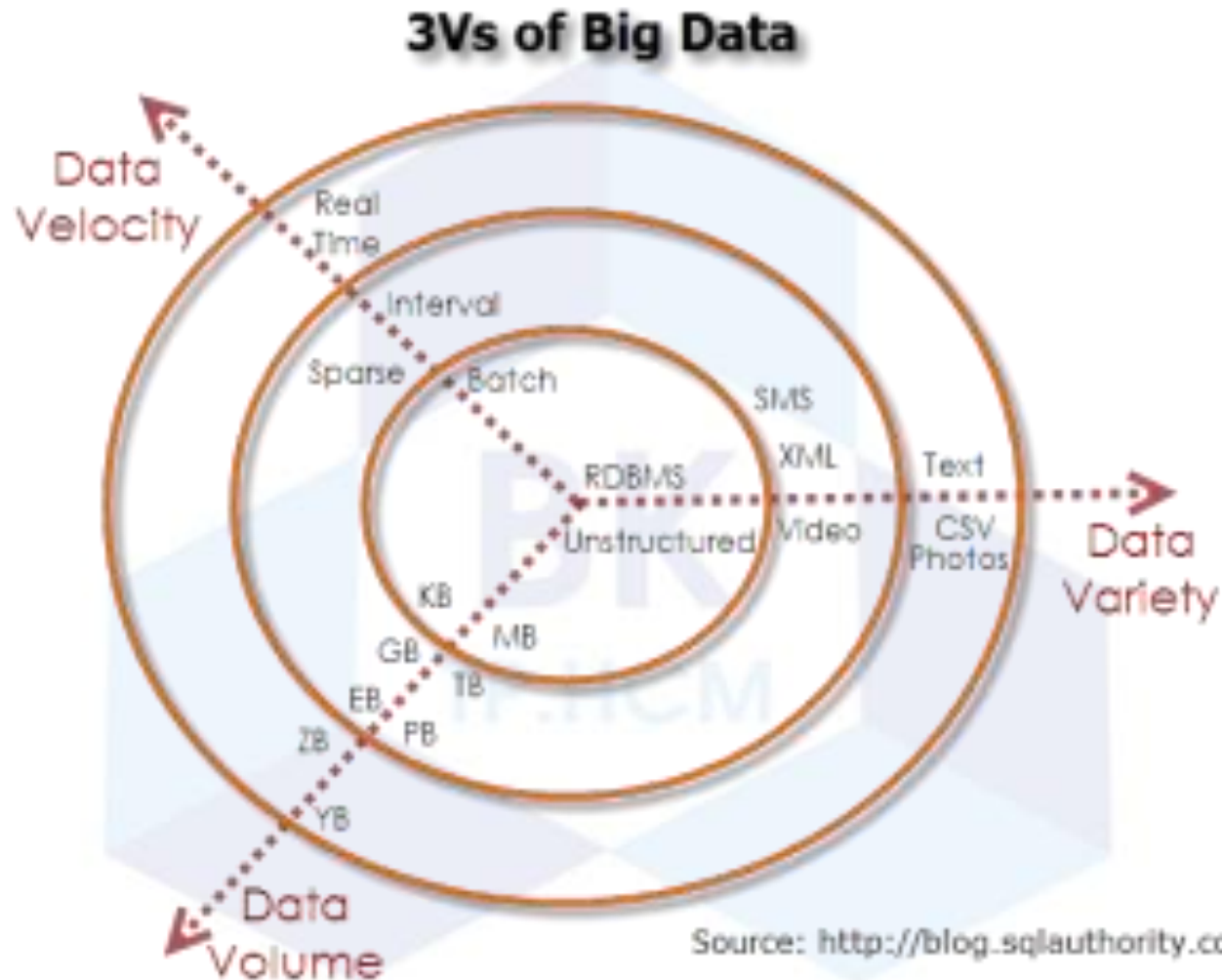
4 tỷ tin nhắn trong một ngày

# Thách thức

---

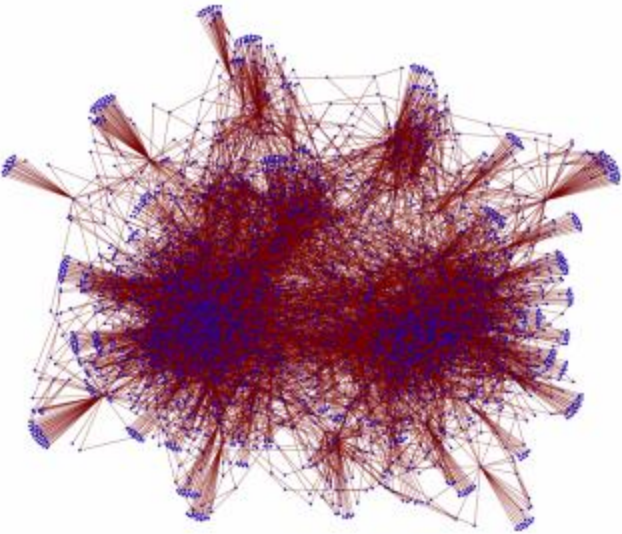
**Làm thế nào chúng ta có thể có được thông tin hữu ích từ rất nhiều các dữ liệu.**

# Đa dạng, tốc độ, khối lượng (3Vs)





# Thách thức



**Lộn xộn**



**Hiệu năng**



**Giới hạn nhận thức**

# Kỹ thuật (1): Pixel Oriented Visualization

---

## Mục dữ liệu

Thuộc tính 1	<input type="text"/>
Thuộc tính 2	<input type="text"/>
Thuộc tính 3	<input type="text"/>
Thuộc tính 4	<input type="text"/>
Thuộc tính 5	<input type="text"/>
Thuộc tính 6	<input type="text"/>

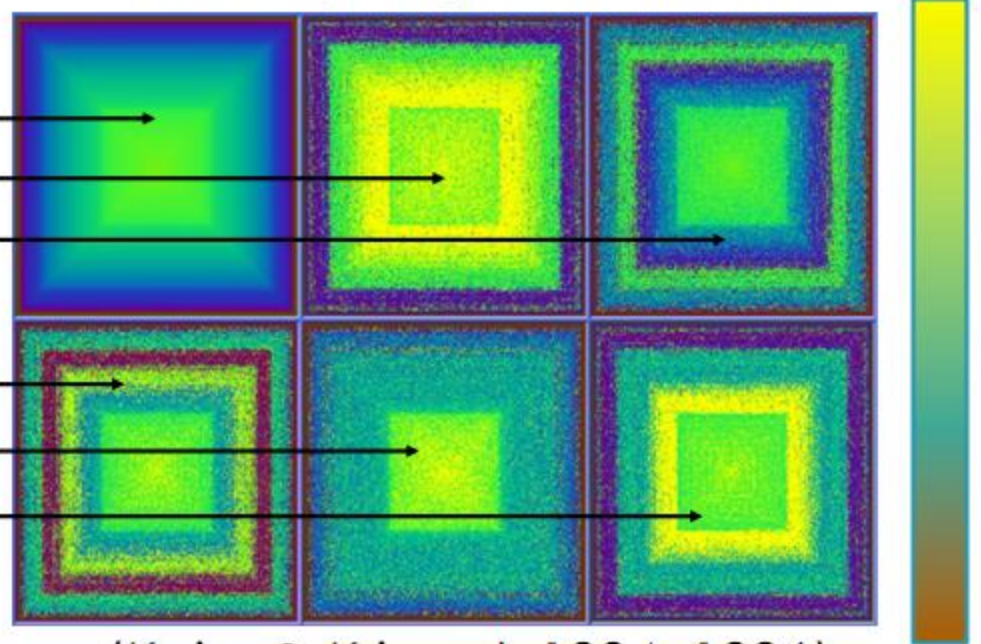
**Một mục dữ liệu đa chiều chứa 6 thuộc tính**



# Kỹ thuật (1): Pixel Oriented Visualization

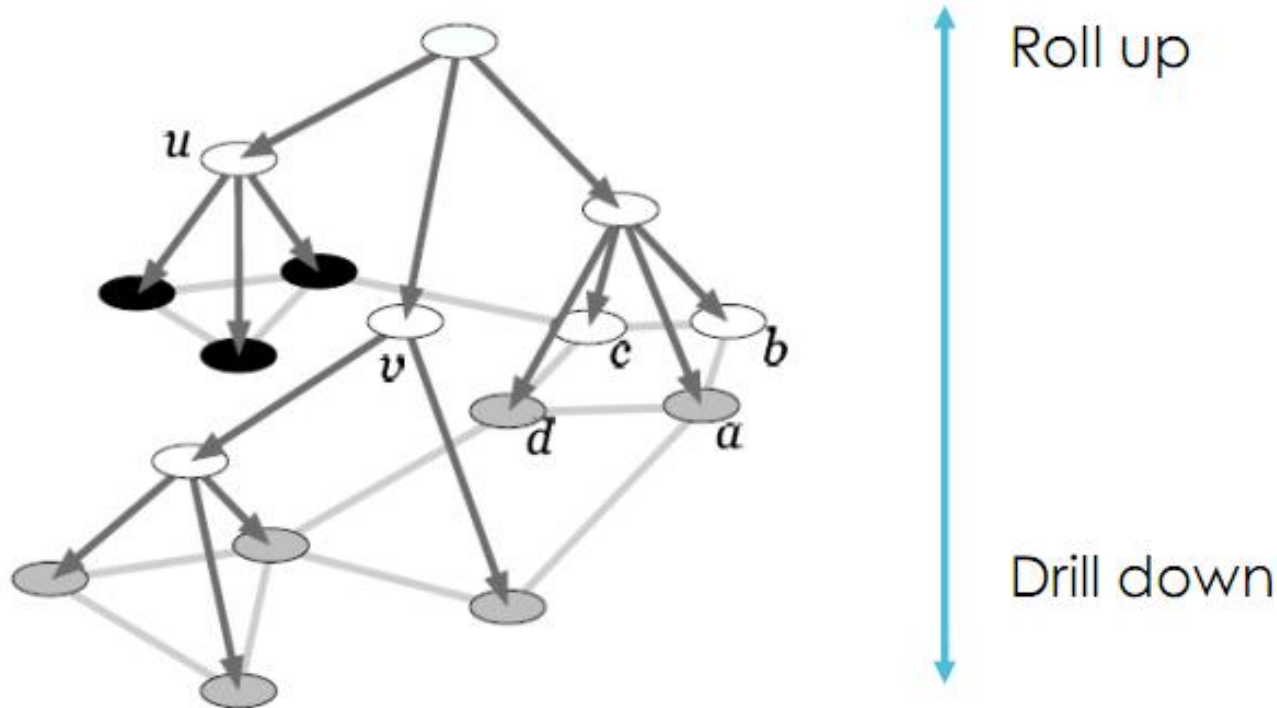
Jan	Feb	Mar	Apr	May	Jun
-99.99	-99.99	315.7	317.45	317.5	317.26
315.62	316.38	316.71	317.72	318.29	318.16
316.43	316.97	317.58	319.02	320.03	319.59
316.93	317.7	318.54	319.48	320.58	319.77
317.94	318.56	319.68	320.63	321.01	320.55
318.74	319.08	319.86	321.39	322.24	321.47
319.57	-99.99	-99.99	-99.99	322.24	321.89
319.44	320.44	320.89	322.13	322.16	321.87
320.62	321.59	322.39	323.87	324.01	323.75
322.06	322.5	323.04	324.42	325	324.09
322.57	323.15	323.89	325.02	325.57	325.36
324	324.42	325.64	326.66	327.34	326.76
325.03	325.99	326.87	328.14	328.07	327.66
326.17	326.68	327.18	327.78	328.92	328.57
326.77	327.63	327.75	329.72	330.07	329.09
328.55	329.56	330.3	331.5	332.48	332.07
329.35	330.71	331.48	332.65	333.09	332.25
330.4	331.41	332.04	333.31	333.96	333.6
331.75	332.56	333.5	334.58	334.87	334.34
332.93	333.42	334.7	336.07	336.74	336.27
334.97	335.39	336.64	337.76	338.01	337.89
336.23	336.76	337.96	338.89	339.47	339.29
338.01	338.36	340.08	340.77	341.46	341.17
339.23	340.47	341.38	342.51	342.91	342.25
340.75	341.61	342.7	343.57	344.13	343.35
341.37	342.52	343.1	344.94	345.75	345.32
343.7	344.5	345.28	347.08	347.43	346.79
344.97	346	347.43	348.35	348.93	348.25
346.3	346.96	347.86	349.55	350.21	349.54
348.02	348.47	349.42	350.99	351.84	351.25
350.43	351.73	352.22	353.59	354.22	353.79
352.76	353.97	354.68	355.42	355.67	355.13
353.66	354.7	355.39	356.2	357.16	356.23
354.72	355.75	357.16	358.6	359.33	358.24
356.08	356.72	357.81	359.15	359.66	359.25
356.7	357.16	358.38	359.46	360.28	359.6
358.37	358.91	359.97	361.26	361.68	360.95
359.97	361	361.64	363.45	363.79	363.26
362.05	363.25	364.02	364.72	365.41	364.97
363.18	364	364.56	366.35	366.79	365.62
365.33	366.15	367.31	368.61	369.3	368.87
368.15	368.87	369.59	371.14	371	370.35
369.14	369.46	370.52	371.66	371.82	371.7
370.28	371.5	372.12	372.87	374.02	373.3
372.43	373.09	373.52	374.86	375.55	375.41
374.68	375.63	376.11	377.65	378.35	378.13
376.79	377.37	378.41	380.52	380.63	379.57
378.37	379.69	380.41	382.1	382.28	382.13
381.38	382.03	382.64	384.62	384.95	384.06
382.45	383.68	384.23	386.26	386.39	385.87
385.07	385.72	385.85	386.71	388.45	387.64

Order by degree of interests max



(Keim & Kriegel, 1994; 1996) min

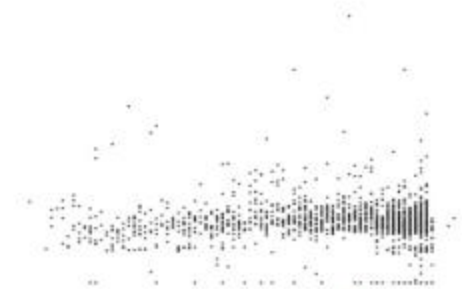
## Kỹ thuật (2): Tổng hợp và mức độ chi tiết



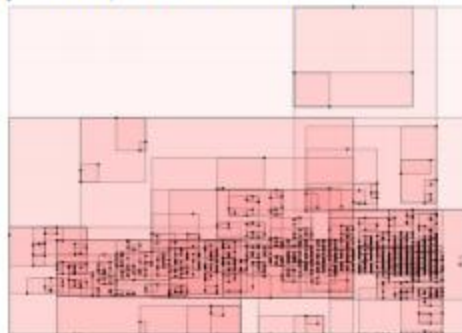
## **Xây dựng một cây để tổng hợp dữ liệu theo cách tiếp cận từ dưới lên hoặc từ trên xuống**

# Kỹ thuật (2): Tổng hợp và mức độ chi tiết

Scatter Plots

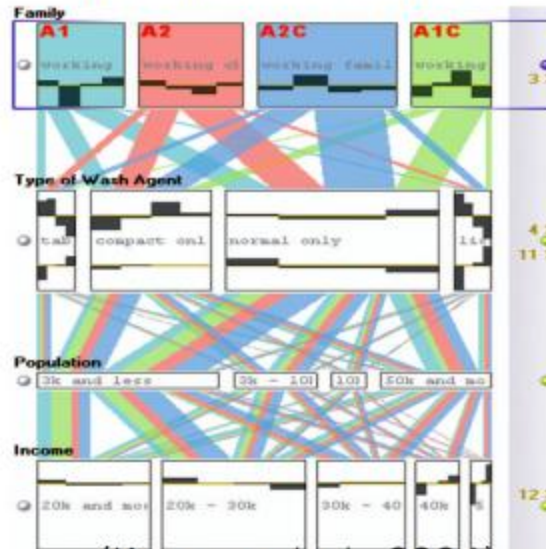


(Elmqvist & Fekete, 2010)

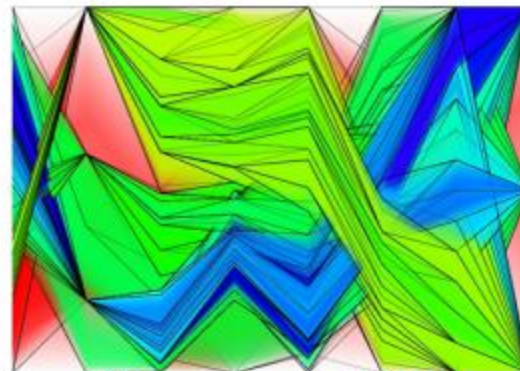


(Yang et al., 2003b)

Parallel Coordinates

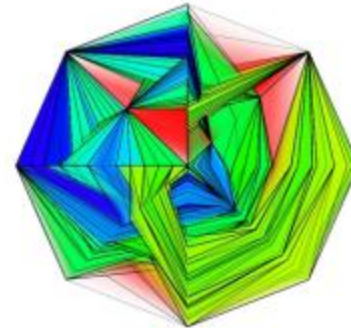
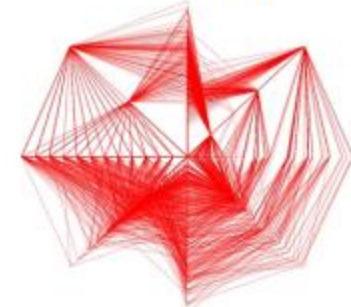


(Kosara et al., 2006)



(Fua et al. 1999)

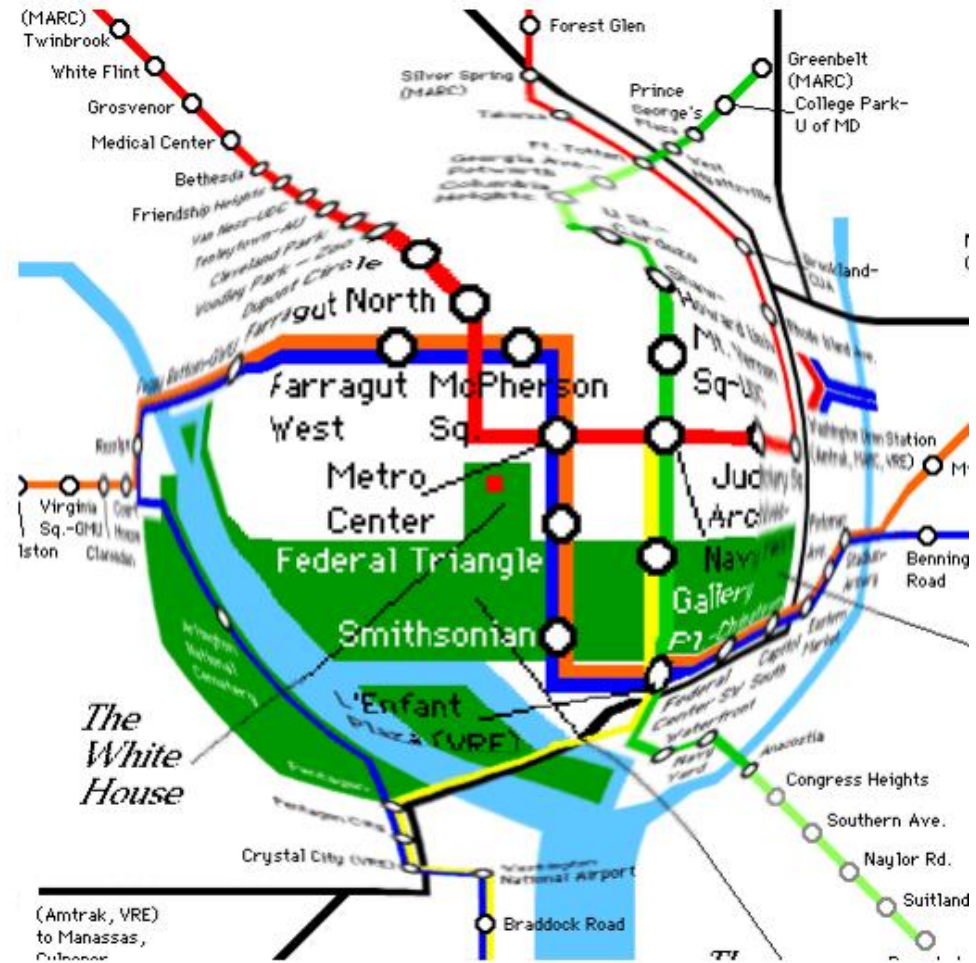
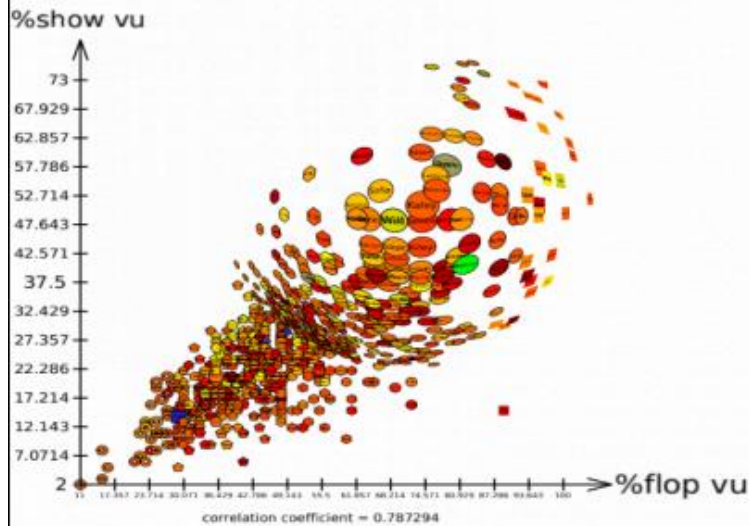
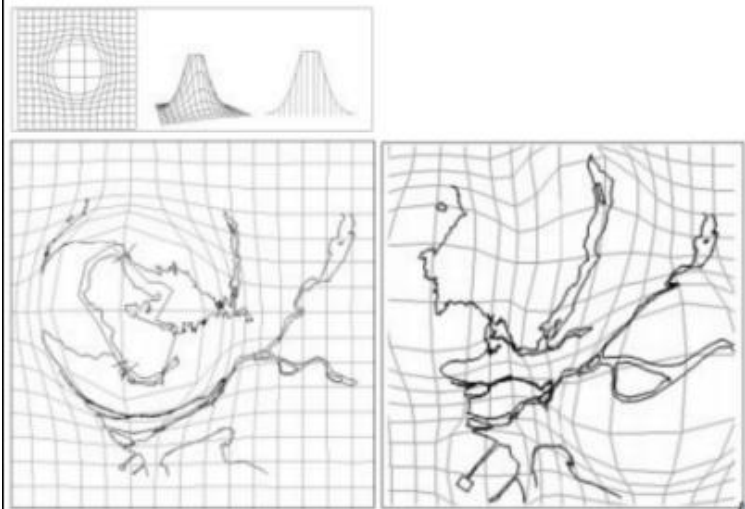
Star Plots



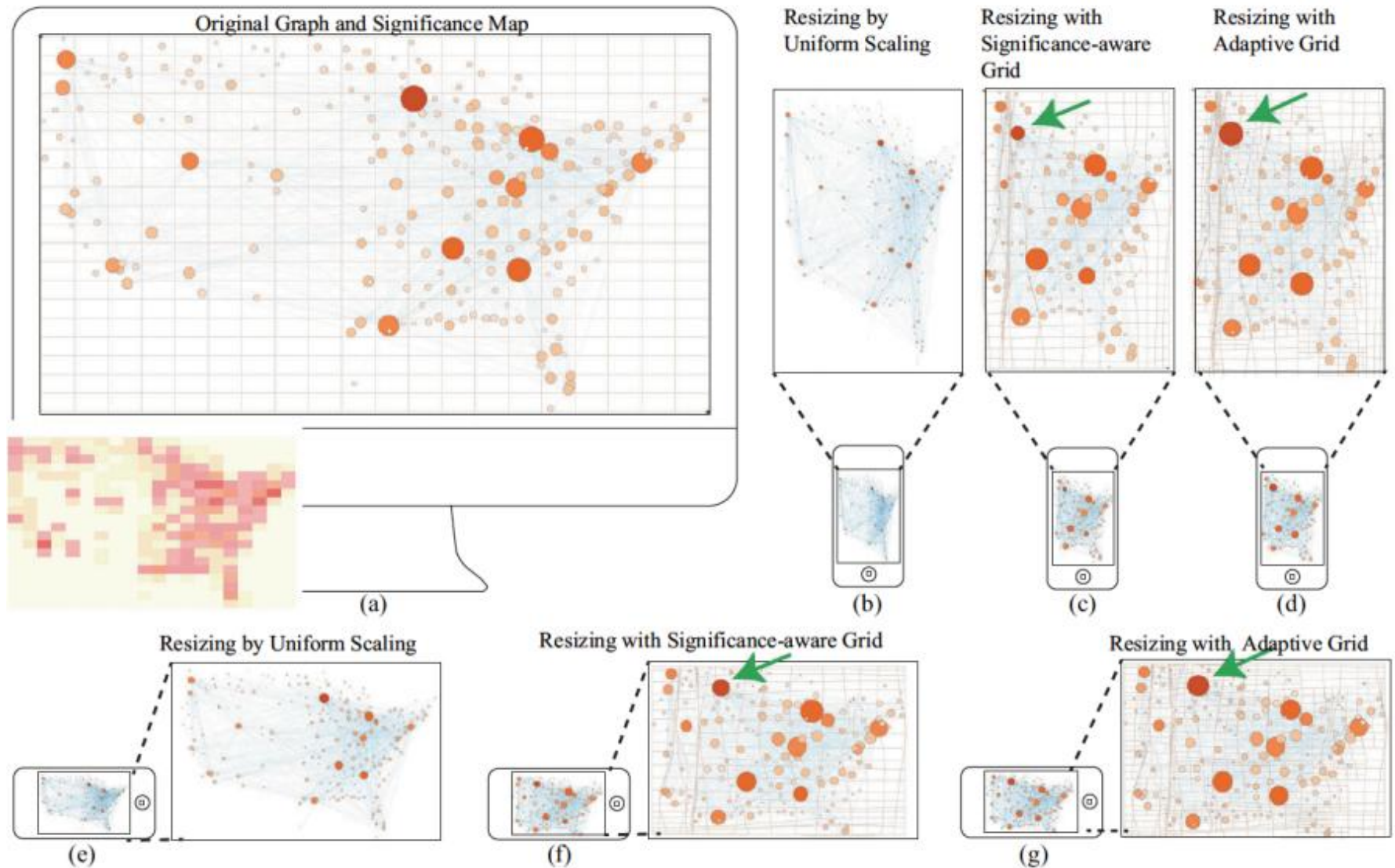
(Fua et al. 1999)



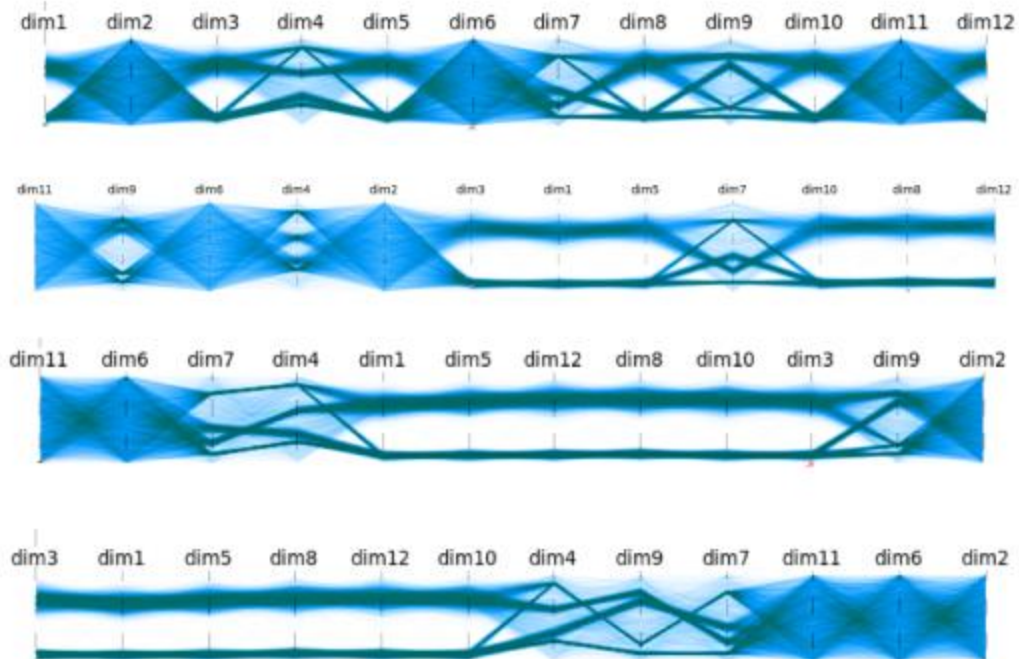
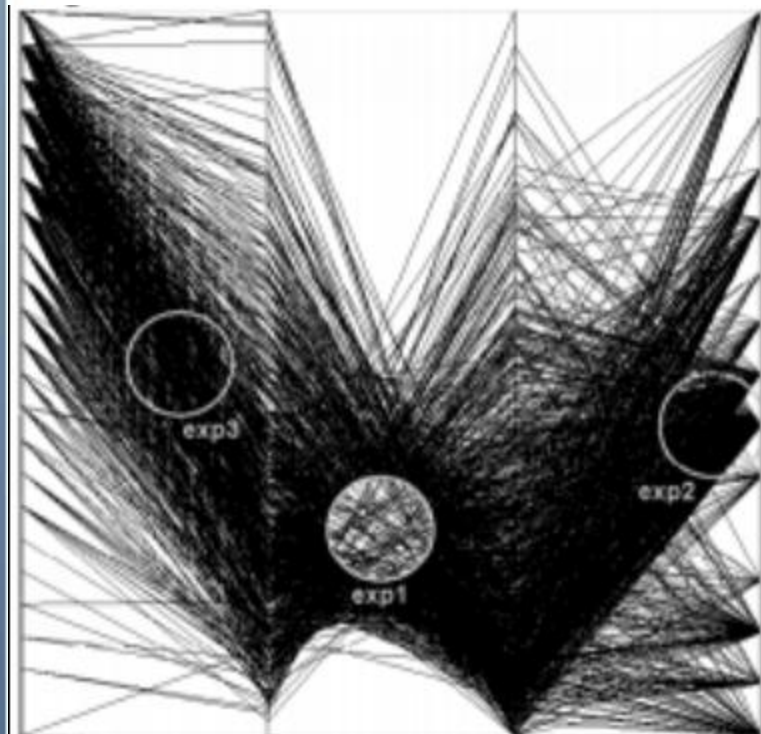
# Kỹ thuật (3): Distortion



# Kỹ thuật (3): Distortion



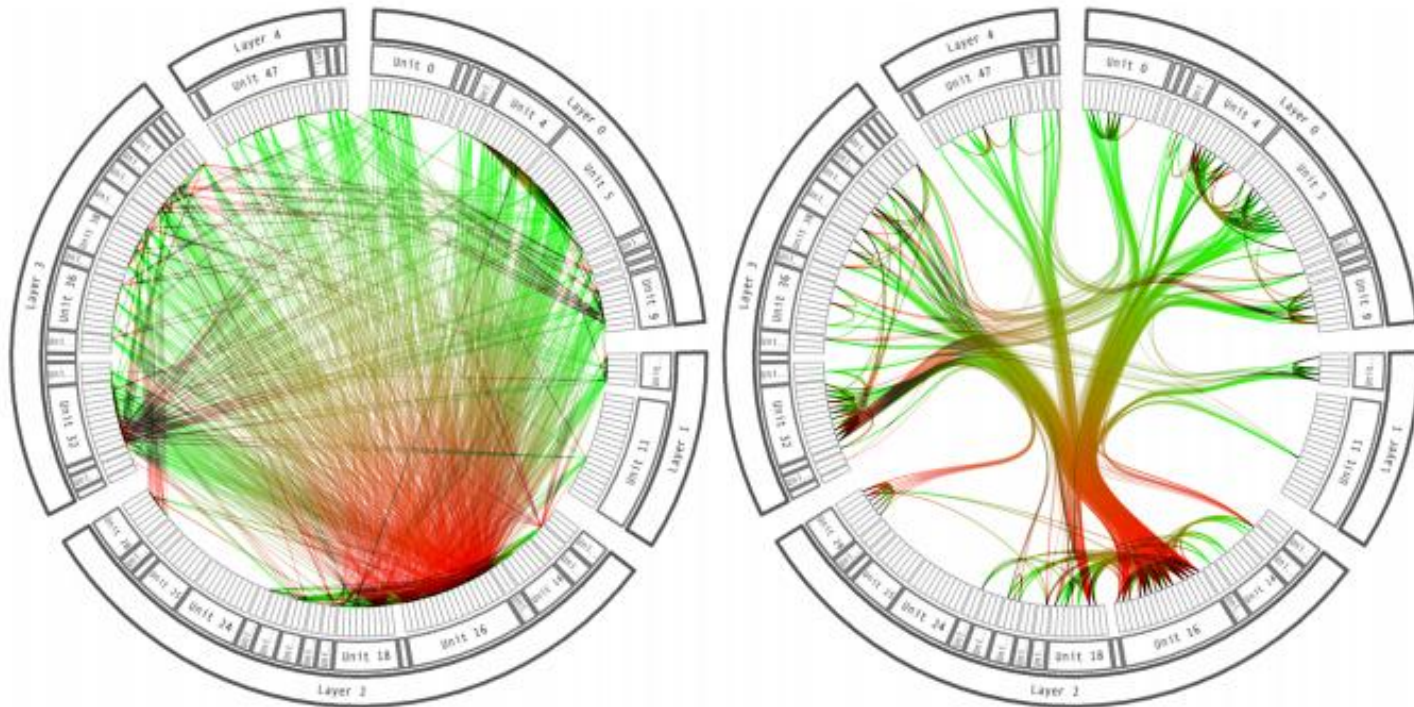
# Kỹ thuật (4): Giảm bớt lộn xộn





# Kỹ thuật (4): Giảm bớt lộn xộn

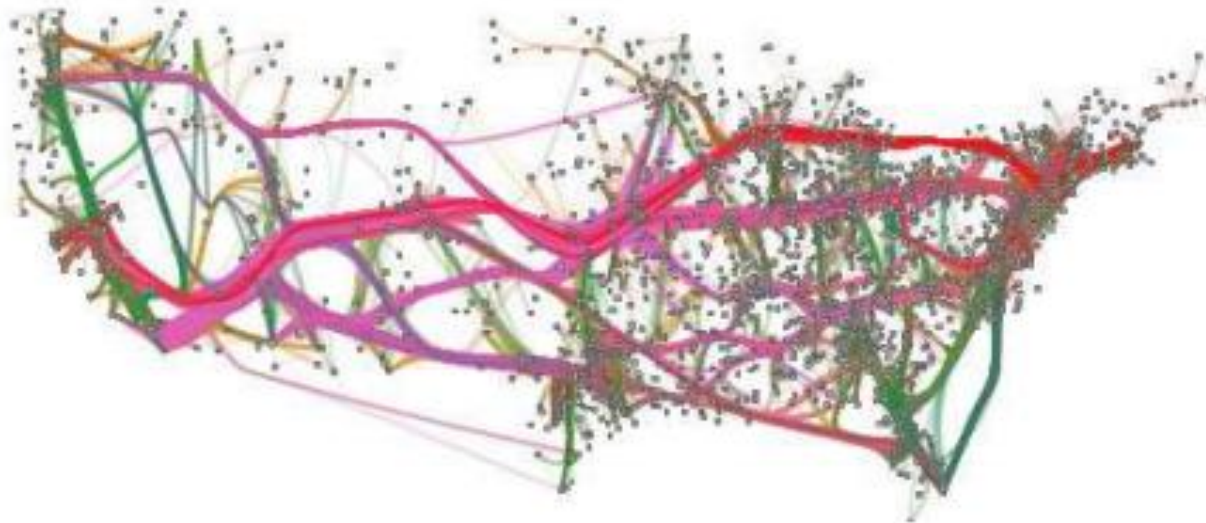
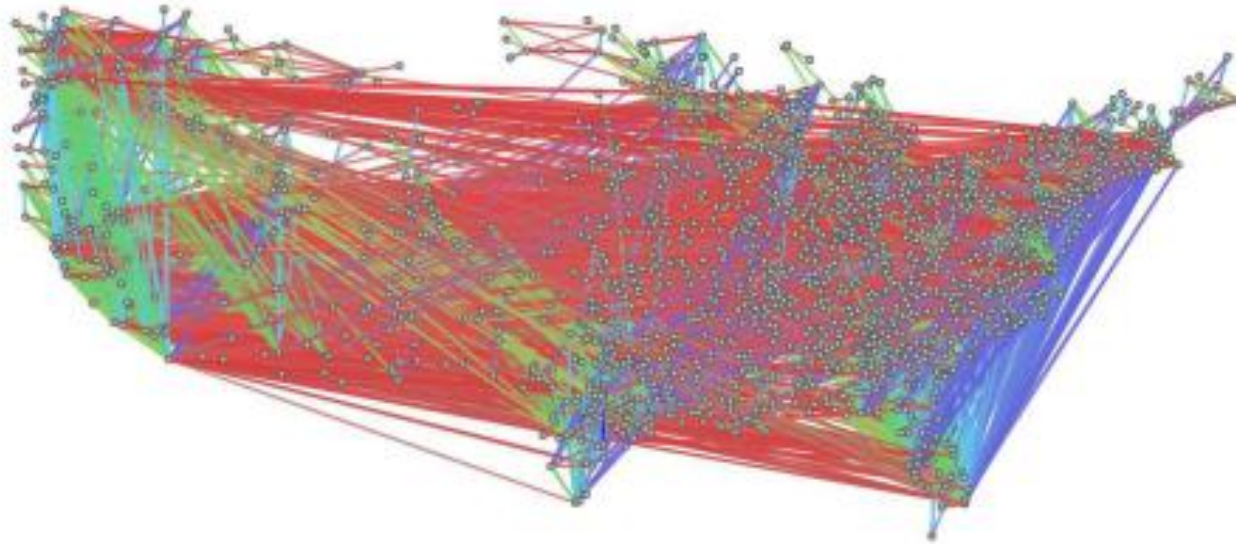
Edge Bundling



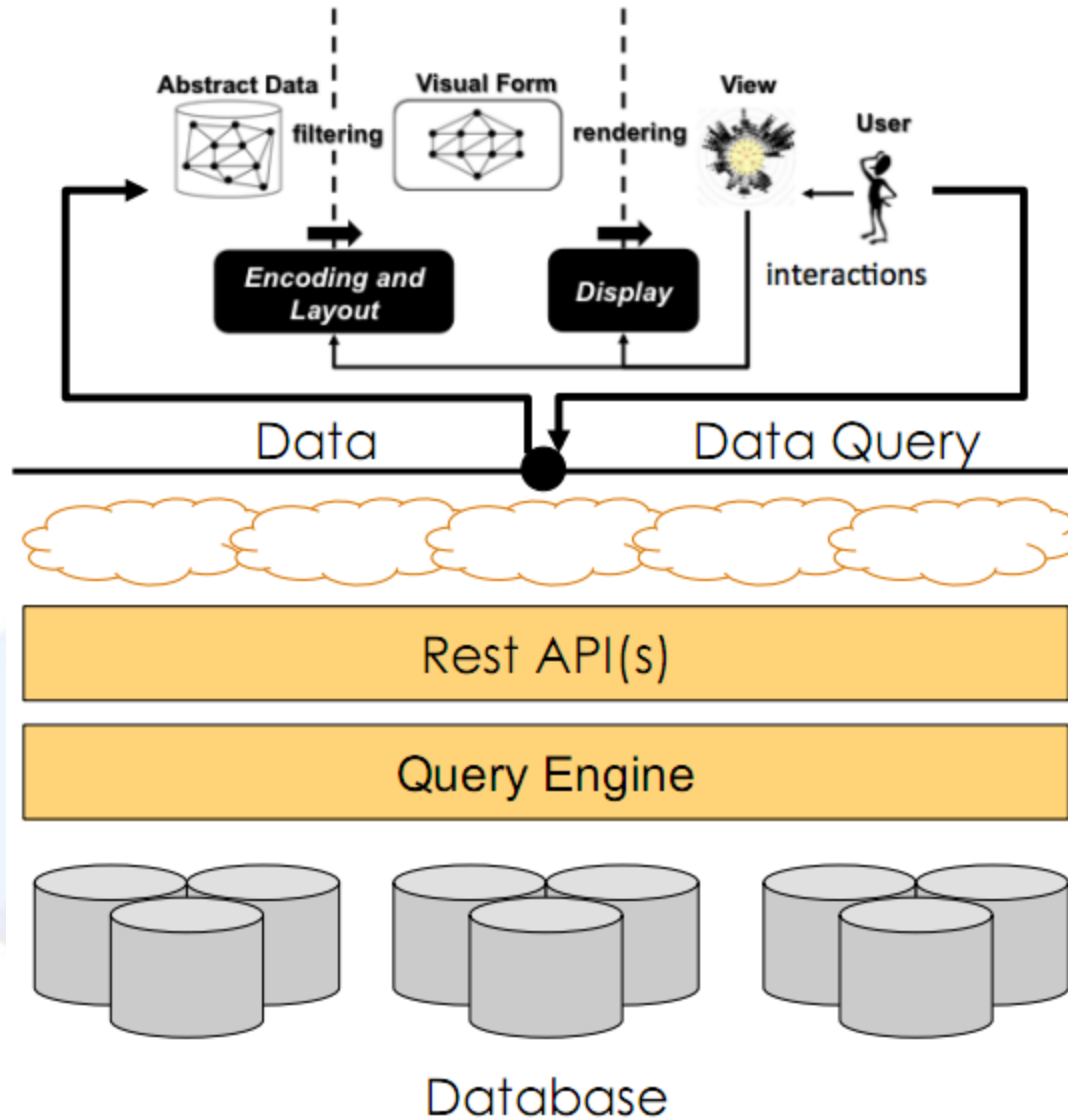


# Kỹ thuật (4): Giảm bớt lộn xộn

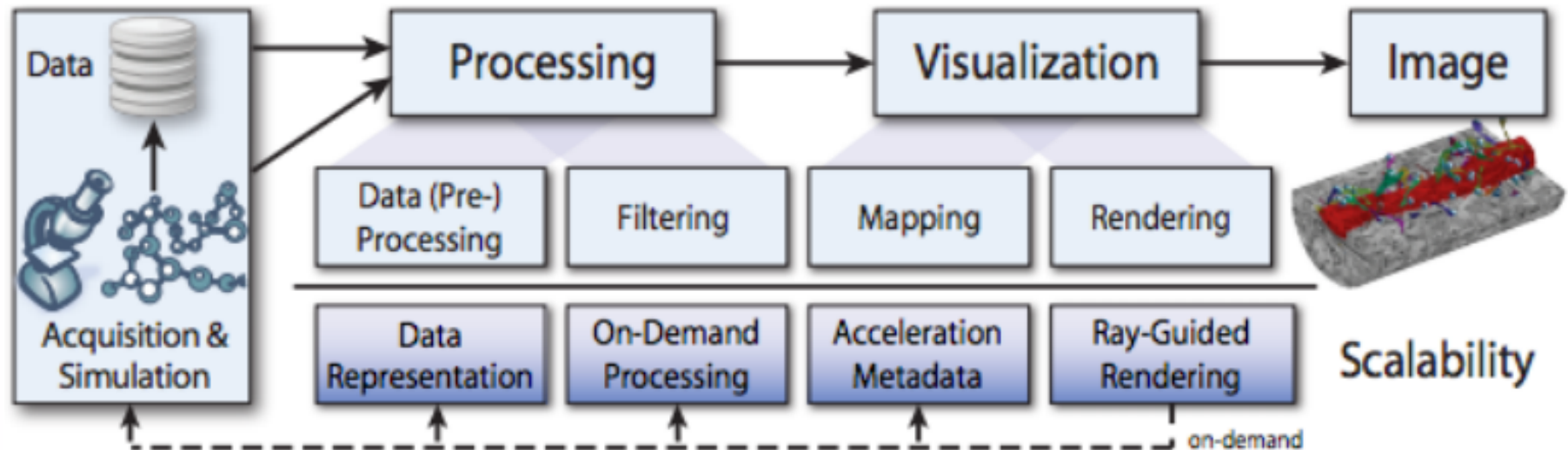
---



# Kỹ thuật (5): Truy vấn dựa trên trực quan



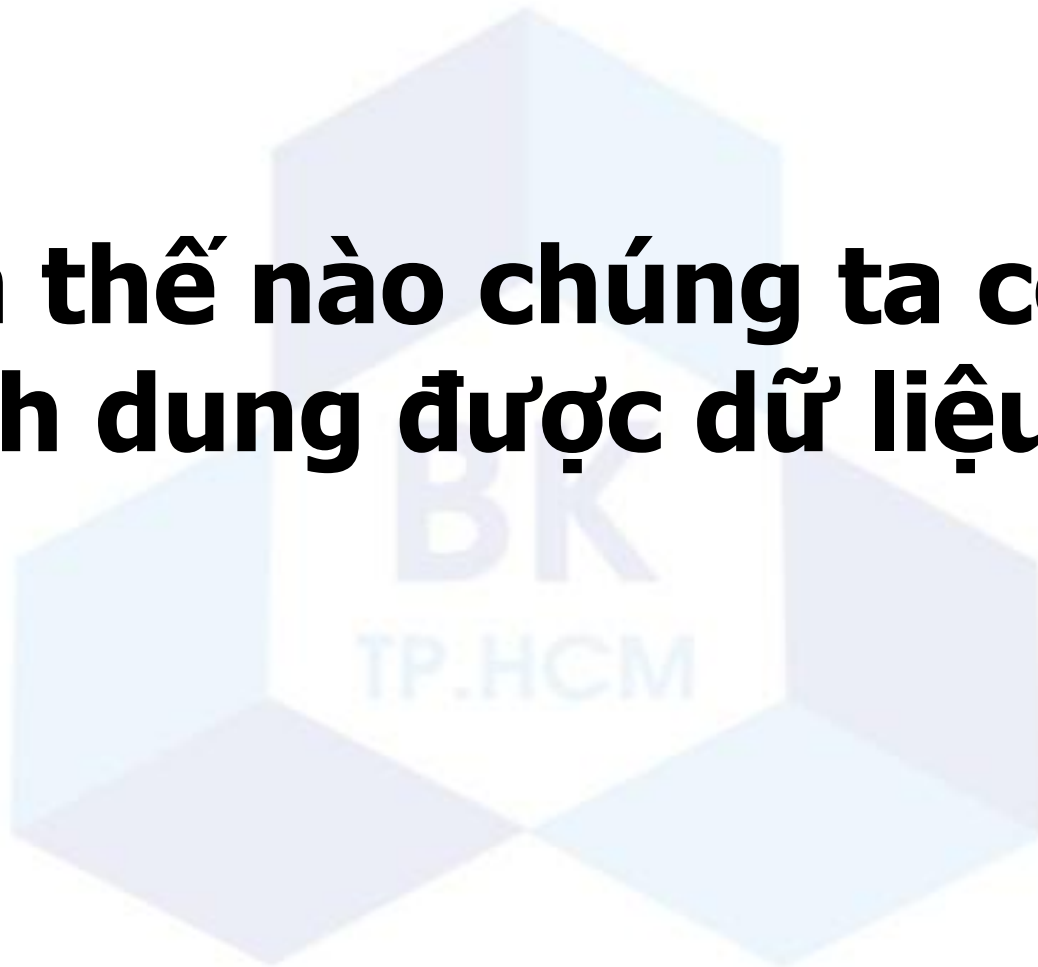
## Kỹ thuật (6): Tính toán song song qua GPU hoặc GUGPU



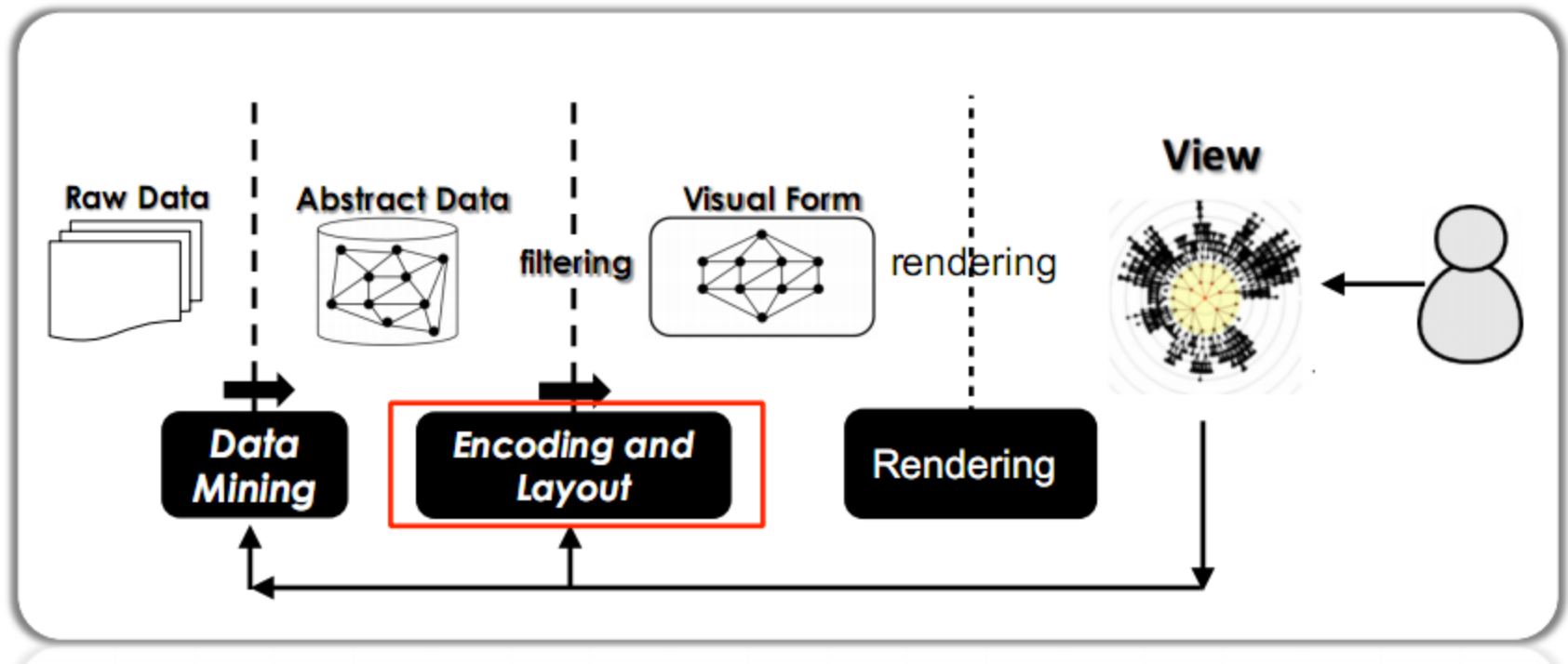
A Survey of GPU-Based Large-Scale Volume Visualization, EuroVis, 2014

---

**Làm thế nào chúng ta có thể  
hình dung được dữ liệu lớn**



# Hình ảnh và mô hình tham chiếu phân tích trực quan



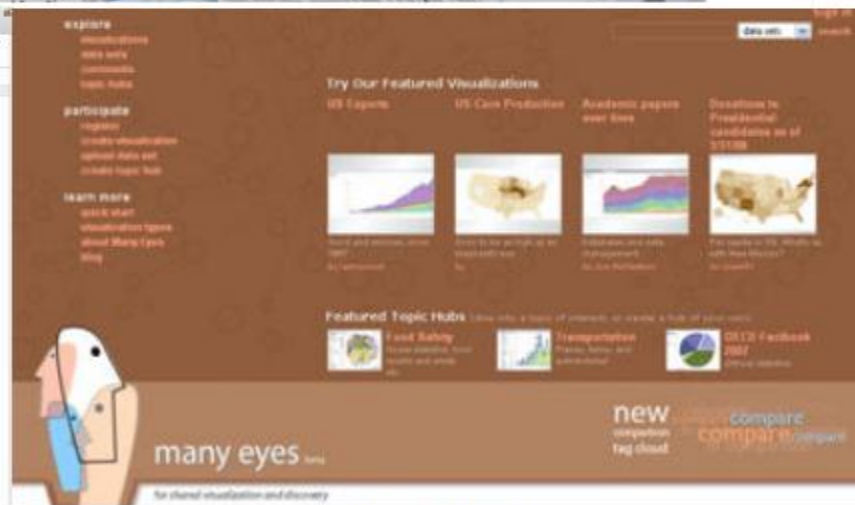
Mã hoá: Thiết kế Trực quan  
Kỹ thuật: Giao diện Thuật toán

# Sử dụng các công cụ hiện có

D3.js  
Data-Driven Documents



Tableau



ManyEyes



# Công cụ mã nguồn mở

---

## Python:

iGraph : <http://igraph.org/redirect.html>

Networkx : <https://networkx.github.io/>

## JavaScript:

D3.js (2D, SVG): <http://d3js.org/>

Tree.js (3D, WebGL): <http://threejs.org/>

## Java:

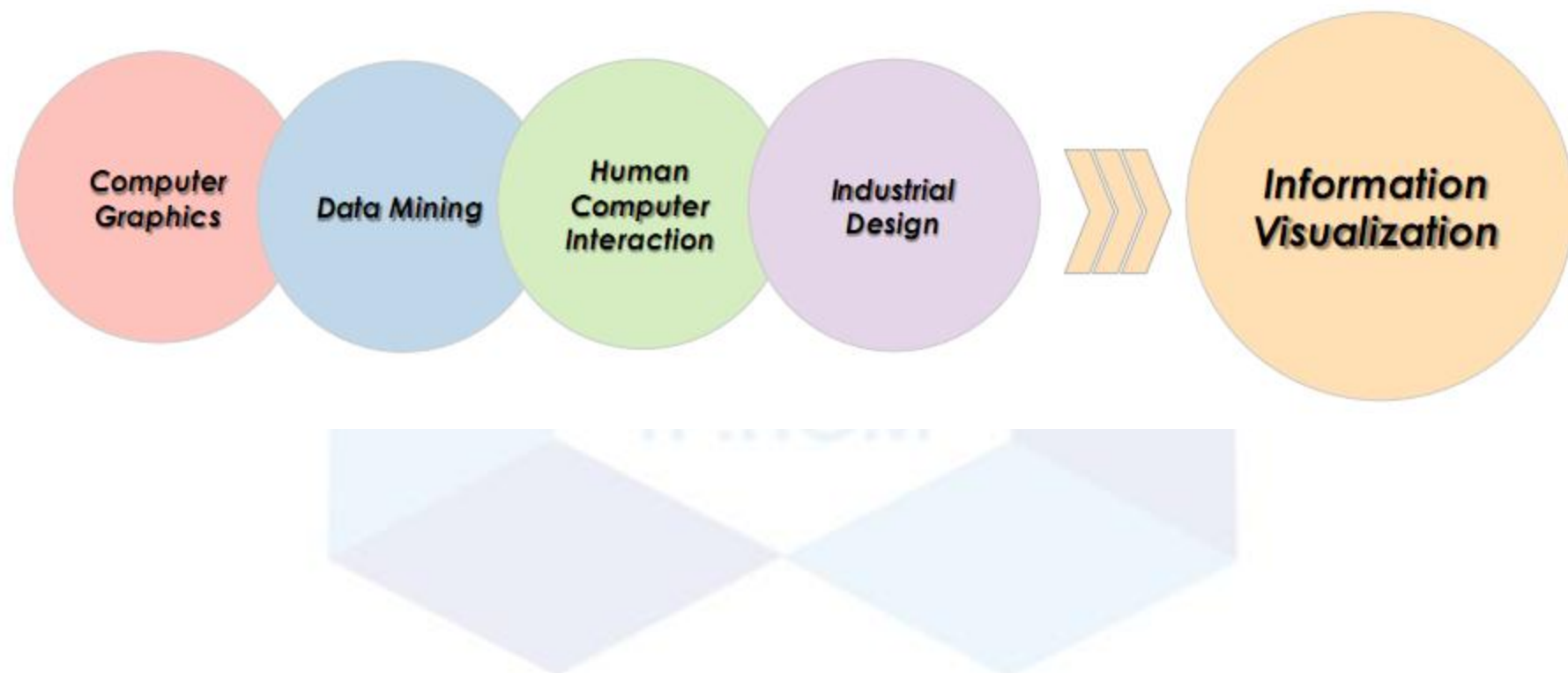
prefuse: <http://prefuse.org/>

InofVis Toolkit: <http://ivtk.sourceforge.net/>



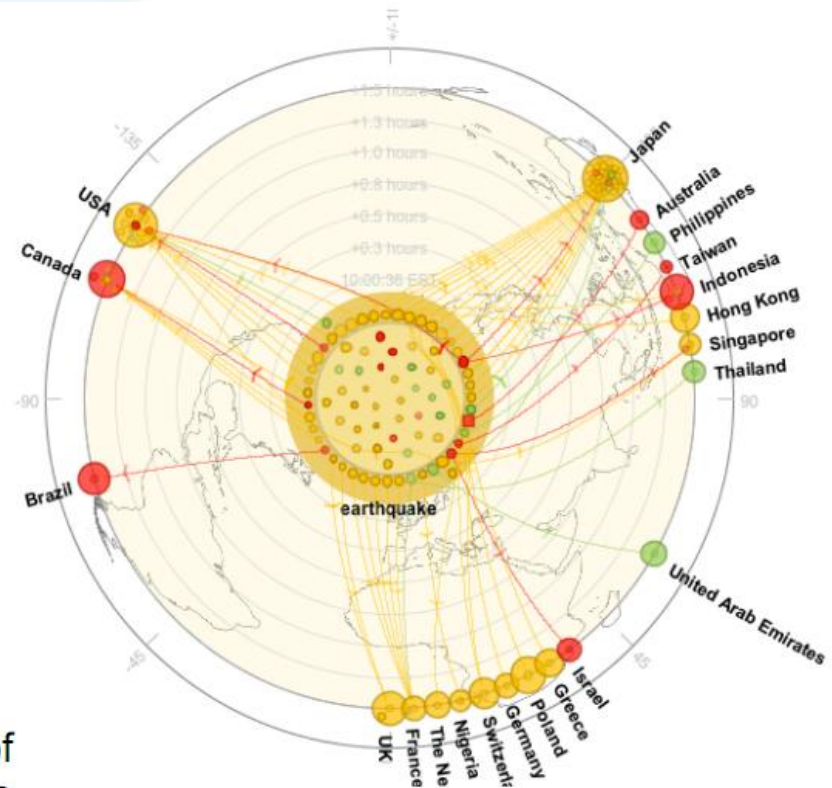
# Công cụ mã nguồn mở

**Việc phát triển những chương trình mới cần có kiến thức từ các lĩnh vực khác nhau**



# Làm thế nào chúng ta có thể hình dung được dữ liệu lớn

## Ví dụ 1: Visualising Streaming Data



Whisper: Tracing the Spatiotemporal Process of  
Information Diffusion in Real Time  
IEEE InfoVis 2012

## Ví dụ 2:

### Visualizing Large Text Corpus



FacetAtlas  
TVCG (InfoVis 2010)

The diagram is a circular visualization. The outer ring is divided into three colored segments: orange (top-left), yellow (top-right), and blue (bottom). Dashed arrows point from the labels 'symptom', 'treatment', and 'other facet' to these segments. Inside the ring, there are two main clusters of colored circles: a red cluster labeled 'Type-1-Diabetes' and a yellow cluster labeled 'Type-2-Diabetes'. Dashed arrows point from the label 'keyword clusters' to the outer ring and from 'topic clusters' to the inner clusters. Within the blue segment of the outer ring, there are two sub-clusters of circles: a blue one labeled 'Type-1-Diabetes' and a yellow one labeled 'Type-2-Diabetes'. Dashed arrows point from the label 'keywords' to these sub-clusters. The blue sub-cluster is further labeled with 'hunger', 'blurred vision', and 'increased thirst'.

SolarMap  
ICDM 2011

57

---

# Phân tích trực quan dữ liệu lớn



# Visual Analysis v.s. Data Mining

**Sức mạnh máy tính**



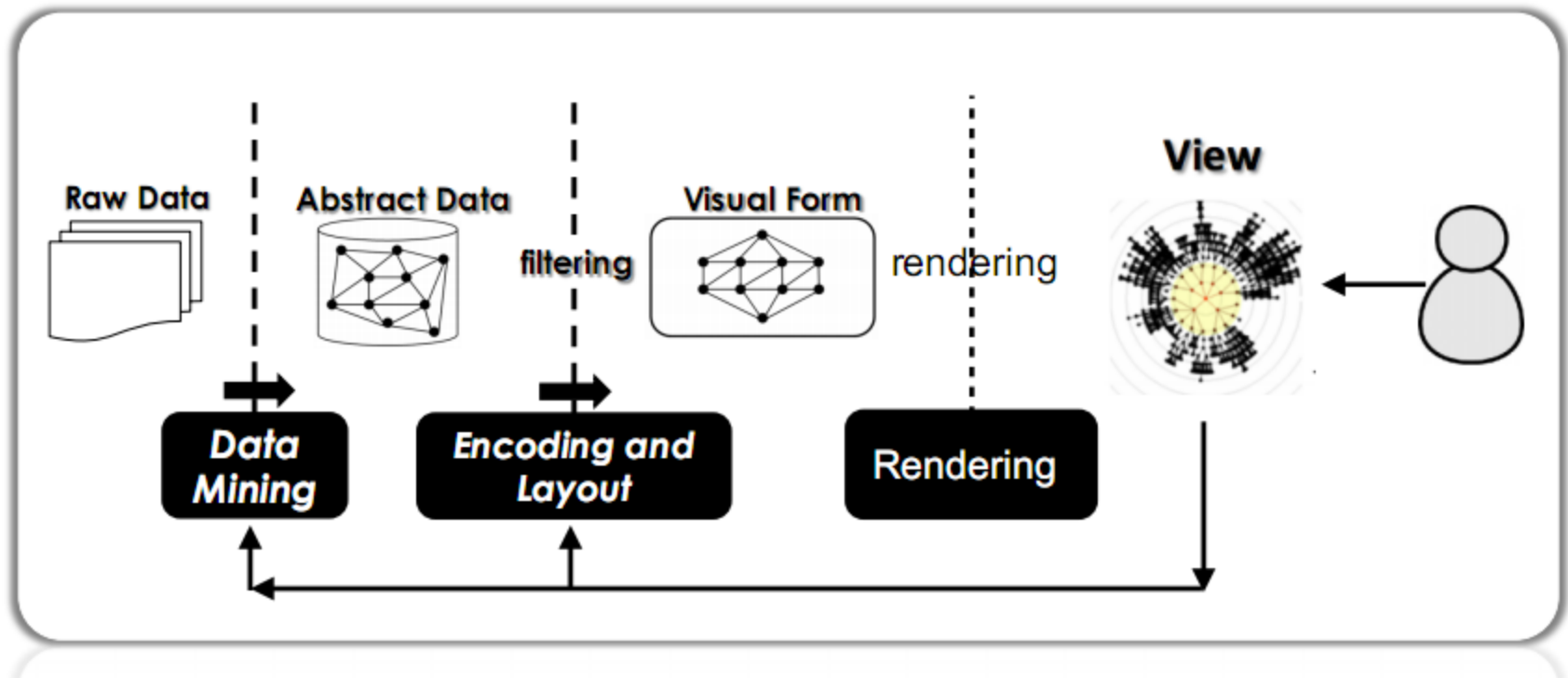
**Data mining**

**Trí tuệ con người**



**Phân tích trực quan**

# Visual Analysis v.s. Data Mining



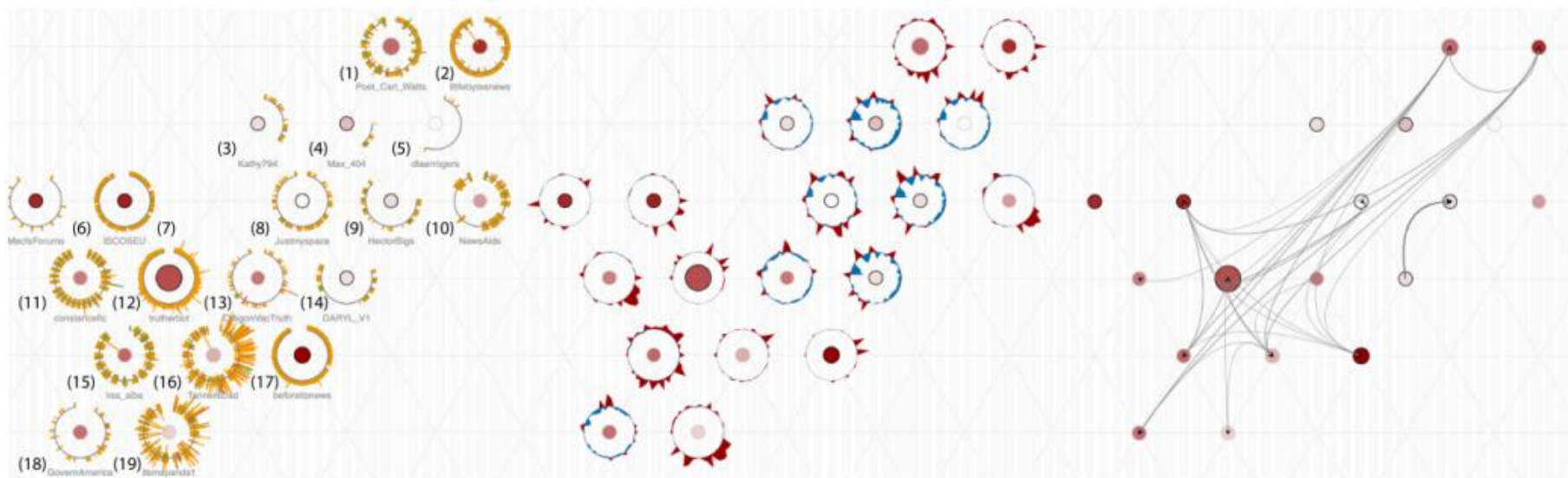
Phân tích + Hình dung + tương tác



# Phân tích trực quan dữ liệu lớn

Ví dụ 3:

Phát hiện người dùng bất thường trong Twitter



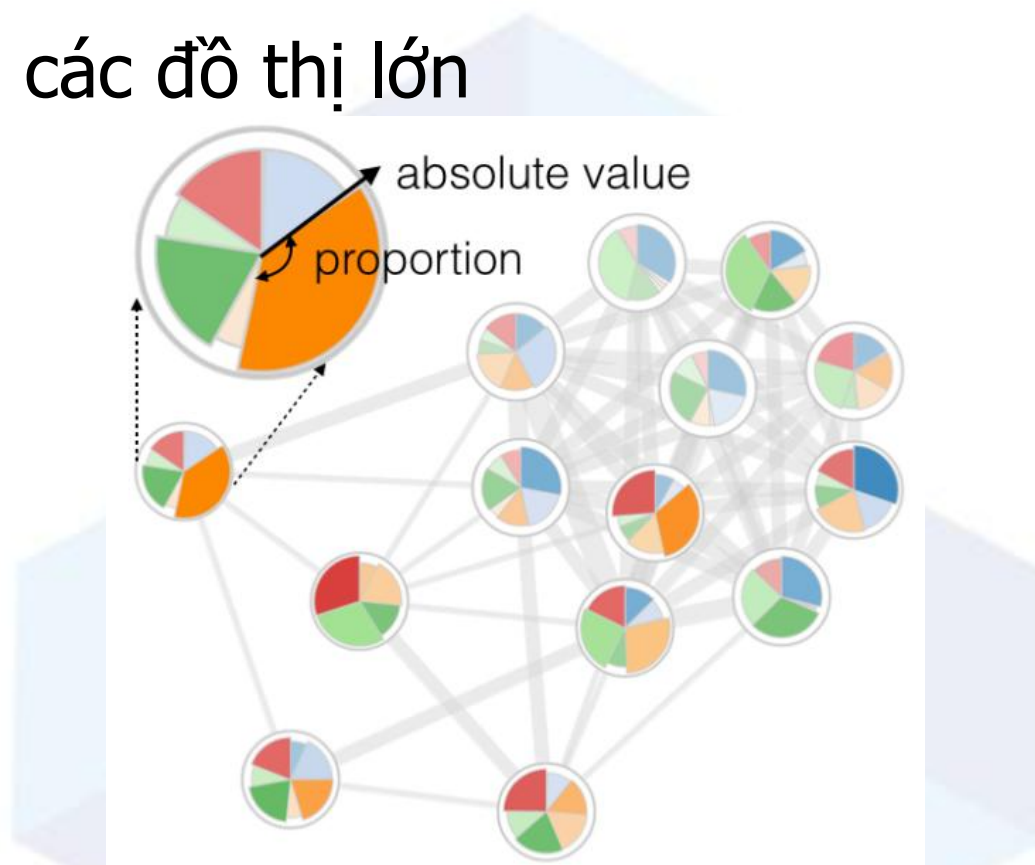
TargetVue: Visual Analysis of Anomalous User Behaviors in Online Communication Systems, IEEE Transactions on Visualisation and Computer Graphics (VAST'15)



# Phân tích trực quan dữ liệu lớn

Ví dụ 4:

Hình dung các đồ thị lớn



**g-Miner: Interactive Visual Group Mining on Multivariate Graphs, ACM CHI 2015**

# BIG DATA VISUALIZATION

---

## HỎI VÀ ĐÁP



# BIG DATA VISUALIZATION

---

**CÁM ƠN THẦY VÀ CÁC BẠN**

