貴州師範大學
Gui zhou Normal University

Big data era and the application of big data in economic and social development in Guizhou

Prof. Xie Xiaoyao, tutor of doctorate student

August 3, 2015

Contents

The era of big data and cloud computing has come. They are two sides of one problem: the one is the question, and the other is the answer to the question. Big data analysis through cloud computing will make the decision more accurate and release more value behind the data.

Data science, the new frontier of human exploration in twenty-first Century, is being discovered and conquered.

Outline

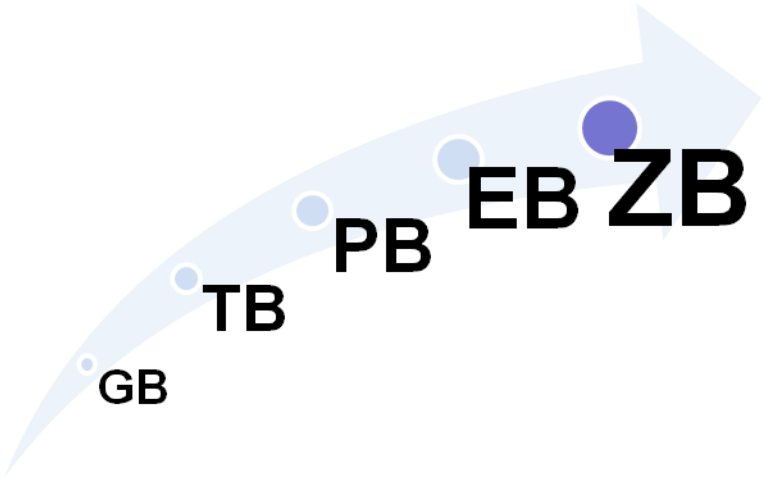
1. Data size in the Internet Era
2. The origin, definition, characteristics and analysis techniques of big data
3. Driving force for the era of big data
4. The practical application of big data in Guizhou
   - Big data monitoring of bridge safety
   - Data calculation of highway fee
   - In-Depth mining in large data of the middle school and college entrance examination
   - Cultural heritage protection in the era of big data
   - The analysis of biological and gene big data
5. Perspective of big data

1 Data size in the Internet Era

Storage Unit

1KB=1024 byte =10^3
1MB=1024KB=10^6
1GB=1024MB=10^9
1TB=1024GB=10^{12}
1PB=1024TB=1,048,576GB=10^{15}
1EB=1024PB=1,073,741,824GB=10^{18}
1ZB=1024EB=1,099,511,627,776GB=10^{21}

Explosive growth of big data



The total amount of data on the earth::

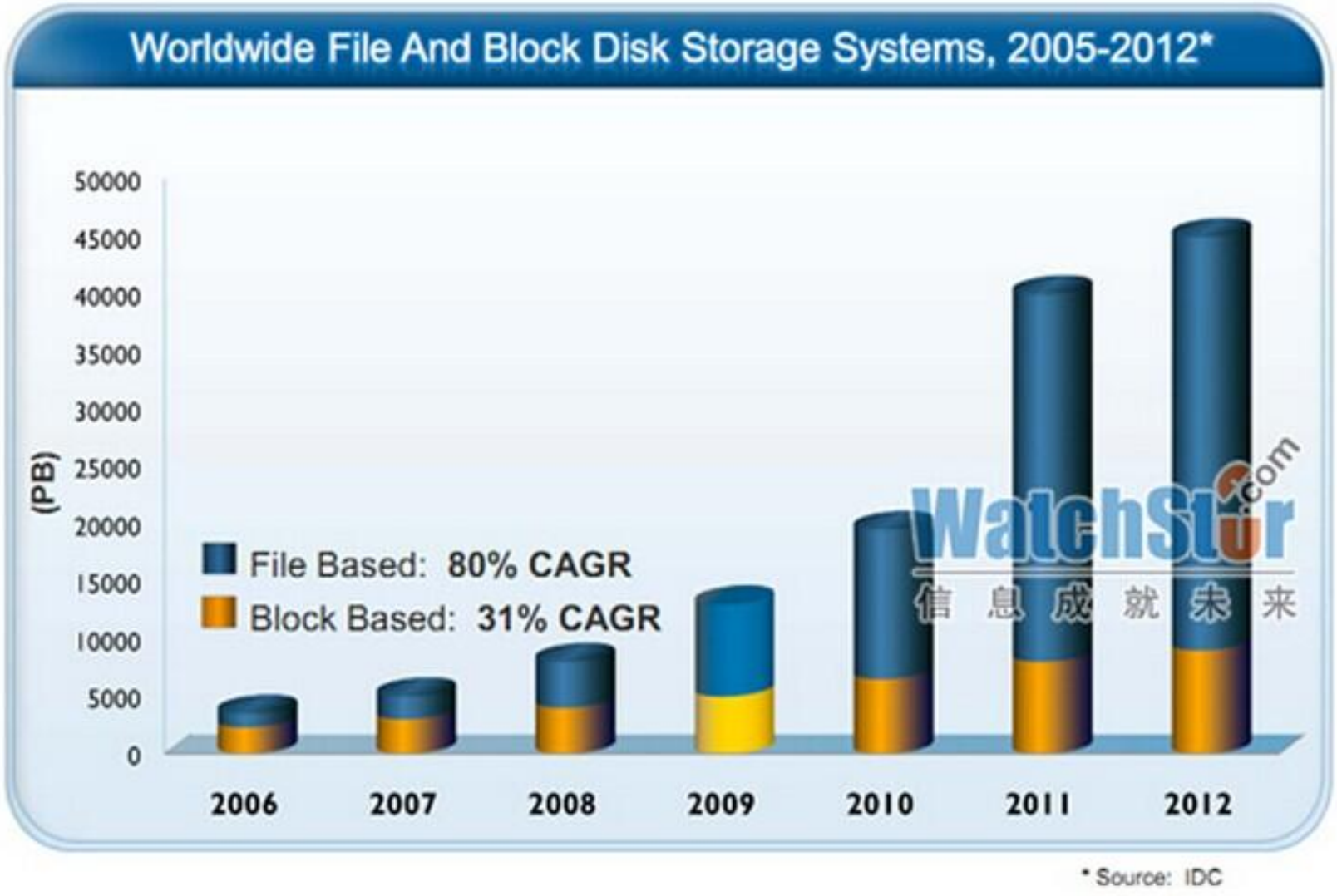
In 2006, the data has reached 180EB in total;

In 2011, this number reached 1.8ZB

By 2020, the **total amount of data** will grow 44 times, to 35.2ZB (1 ZB = 1 billion TB)!

1GB = 2^30byte
1TB = 2^40byte
1PB = 2^50byte
1EB = 2^60byte
1ZB = 2^70byte

High quantity of computation and storage

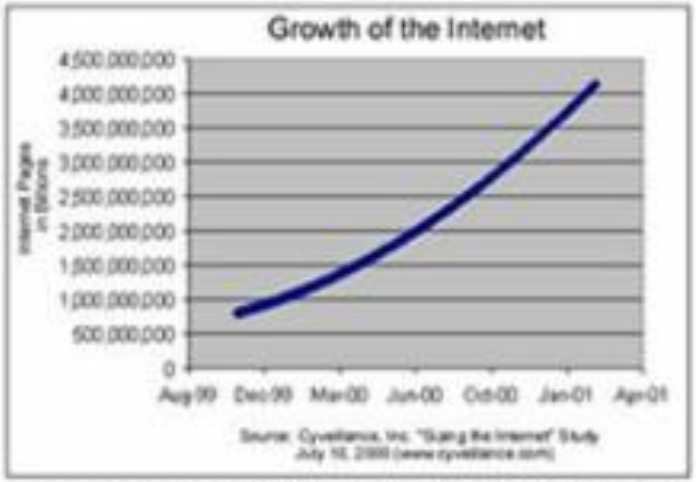


By 2020, global data volume will grow by 50 times and 80% of them are unstructured data

The origin, definition, characteristics and analysis techniques of big data

2

大数据的起源与断代



1999-2000网页数量从5亿增长到40亿，每天新增700万！PB级非结构化数据

互联网内容的暴增是第一推动力量

大数据起源

2000年前后，基于海量数据分析的搜索引擎逐步发展

大数据时代

2006后，电商、广告、SNS应用，2011大数据提出

突变

分布式文件系统
分布式并行计算
分布式数据库

数据分析

1990，提出数据挖掘和商业智能

数据库

1960，数据与应用分离，事务处理数据库技术发展

数据耦合

1950，计算机诞生，数据与应用捆绑，都存储在文件中

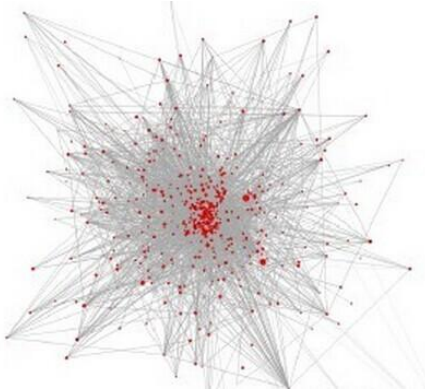
今天：随着应用领域扩展，技术也在不断演进，为更广泛应用提供条件

Definition of big data

Big data is a collection of data that can not be captured, managed and processed by traditional database software tools

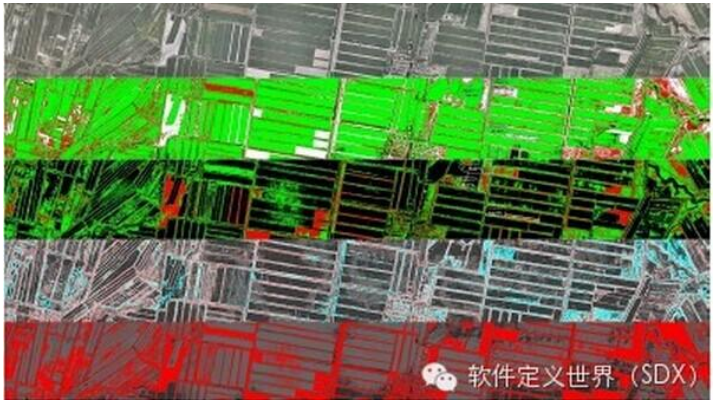
Typically, Big data have "4 features and 1 properties"

large volume



Complex logics

Quick response



New forms (such as GIS, satellite data or images of special device ）

1. Volume

Huge amount

The world officially entered the ZB era in 2010, IDC is expecting that the world will have a total of 35ZB data by 2020.

2. Variety

Structured data, semi-structured data and unstructured data

Today there are a variety of data types , such as text form, orders, logs and audio

3. Value

Low value density

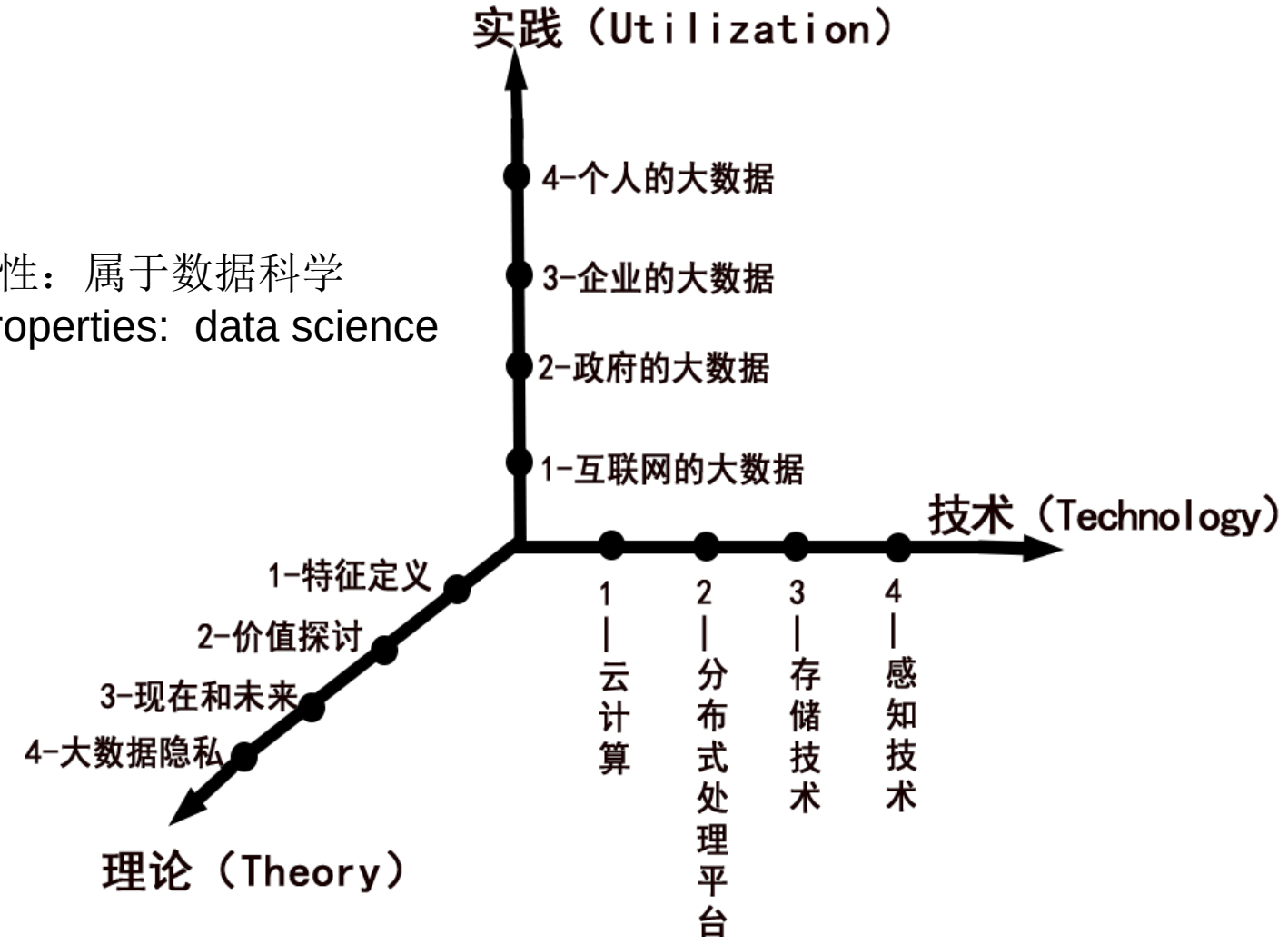
The data of usefulness may be a piece of one or two seconds in an video continuous for an hour. How to purify the value of the data with a powerful machine algorithm is a problem to be solved urgently.

4. Velocity

Real-time access

In the face of such a huge amount of data, the efficiency of processing is the life of the enterprise.
That's why Big data is distinguished from the traditional data

1属性：属于数据科学
1 properties: data science



Application of big data

Big data are those formed in the social production and lives as well as management and service process and whose acquisition, transmission and summarizing relies on modern information technology y instead of traditional data processing system . They have the characteristics of large amount , various types, fast processing speed, etc. through the integration of sharing, cross multiplexing, extraction and analysis to obtain new knowledge and create new value.

Data production	Data aggregation	Data analysis	Data utilization
1.Structured data in the internal business system database 2. Other unstructured data in the internal business system 3. External structured data, in the external service system4. other unstructured data from External websites, mobile applications, social networks, sensors and video surveillance equipment	1.Gather data as much as possible 2. Build master model 3. Realized data warehouse 4. Gather and deal with unstructured documents and knowledge 5. Gather and deal with the external unstructured data from social networks, streaming media, sensors and other sources	1.Public behavior pattern analysis 2. Public analysis 3. Legal analysis 4. Market analysis 5. Performance analysis 6. Risk analysis 7. Situation prediction 8. Cultural analysis	1.statements 2. Reports 3. Visual chart 4. Social network sharing

Big data analysis technology

- A/B testing(split testing、bucket testing) A/B测试
- Association rule learning关联式规则
- Classification分类法
- Cluster analysis群集分析
- Crowdsourcing众包
- Data fusion and data integration数据融合&数据集成
- Data mining数据挖掘
- Ensemble learning集成学习
- Genetic algorithm遗传算法
- Machine learning
- Natural language processing(NLP)自然语言处理
- Neural networks神经网络
- Network analysis网络分析
- Optimization最优化
- Pattern recognition模式识别
- Predictive modeling预测模型
- Regression回归分析
- Sentiment analysis
- Signal processing信号处理
- Spatial analysis空间分析
- Supervised learning监督学习
- Simulation仿真
- Time series analysis时序分析
- Unsupervised learning无监督学习
- Visualization可视化

Driving force for the era of big data

3

In the 19th century, the entrepreneurial opportunity centered on the European market and traditional industry; in the 20th century, entrepreneurial opportunities centered on the United States market and IT industry, business opportunities in the 21st century concentrated in the Chinese market and emerging Internet industry.

In the 20th century, the industries dominating global economic development are petrochemical, steel, machinery, automobiles, aircraft, trains, ships, aerospace and military industry.

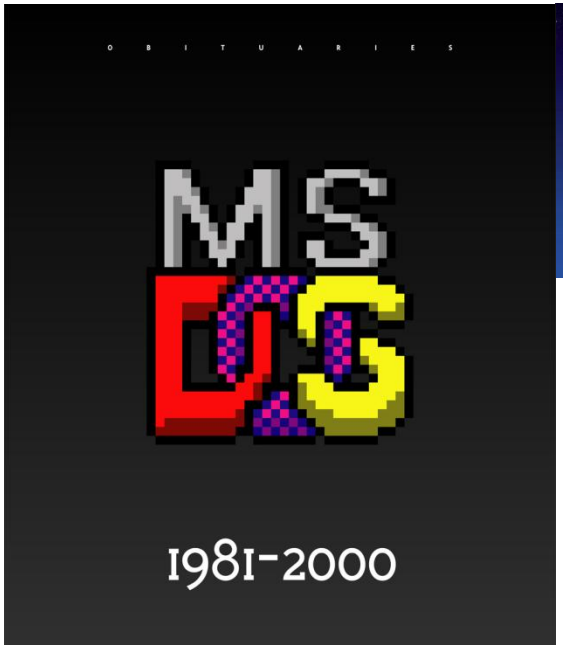
Twenty-first Century is a historic turning point in the history of global industry development. IT industry becomes the most dazzling star on the center of the stage.

（1）In 1911, IBM was established in New York. For a hundred years since it’s establishment, IBM‘s business focus is constantly changing. His success stem from adapting to the market, excellent management, keeping keen perception of the market changes by scientific and technological innovation and quickly adjustment of the direction into the newly emerging markets.

Blue Giant

（2）In 1975, Bill Gates started Microsoft. In 1980, IBM executed an extensive outsourcing of parts in order to launch a new personal computer to compete with apple. Consequently, Microsoft won its crucial operating system contract.

Microsoft's success tell us, standing on the shoulders of giants, is an important and convenient solution for small companies to host.



(3) In 1947, Shockley, an engineer who worked in Bell Laboratory, invented the transistor. In 1968, Noyce and Moore set up Intel.

In1971, Intel launched its and also the world's first micro processor, which become standard parts of IBM PC chip .

Afterwards, with the rise and thrive of the PC industry, it continues to grow and develop. Intel's success, is the same as standing on the shoulders of giants.

(4) In 1970, a researcher Edgar Caunt at IBM laid the theoretical foundation for relational database software. In 1976, Allison founded the Oracle Corp. Oracle Corp has been leading the development of the database software market and continue to grow with wide application of database software among enterprise.

Oracle's success tells us that once we identify opportunity in market, the implementation is more important than the R&D.

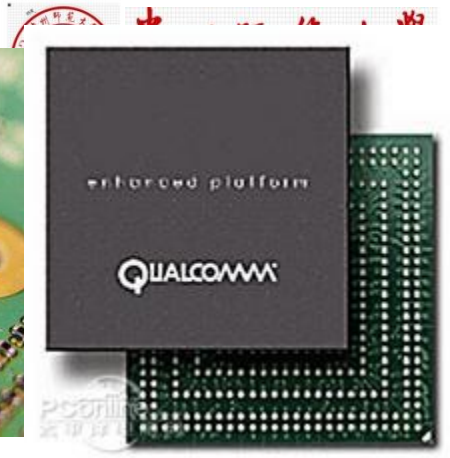
(5) In 1976, two DIY fancier Jobs and Wozniak founded Apple Corp in their garage. ，In 1997, the founder and the soul of apple Jobs returned to keep the spirit of innovation, and constantly refresh the world with the perfection of technology products.

The success of the Apple Corp brings us not only the inspiration of innovation, but also the difficulty of how to keep continuous innovation.

(6) In 1984, Cisco Systems, Inc. was founded by two Directors of Stanford University Computer Center. They unified the incompatible local area networks at campus into a network and began the age of the Internet.

CISCO's success is the success of the innovator as well as the success of the acquisition strategy.

QUALCOMM®



(7) In 1985, Qualcomm company was established. In its early age, it is mainly engaged in the service of mobile communication technologies in transportation industry. With the advent of smart phones, the growing of popularity of the Android system leading Qualcomm towards an overlord of mobile chip industry.

The Corp. who can adapt to the times and introduced the earliest IT technology into the traditional industries, will win the ability to subvert the giant.

(8) In 1963, Carrefour launched in Europe the business model of supermarket which was soon copied to the United States by WAL-MART. In 1980s, WAL-MART took the lead in introducing the new IT technology to perform a comprehensive process reform in the logistics and inventory. Soon ,it became the new generation of retail king in 1990 .

WAL-MART's success has its root in adapting to the market, introducing new business models and IT technology timely , and timely introduction of the latest IT technology, strengthening process management and reducing costs.

(9) Home depot was established in 1978. In 1983 started to introduce calibration technique and transform the traditional building materials distribution channel through mode of building materials chain stores and process reengineering of warehousing and logistics.

The home depot in just spent 20 years becoming the global retail giants only secondary to WAL-MART. The secret of its success is just like WAL-MART 's: the introduction of new chain store business model and IT technology to improve efficiency and reduce costs, and finally achieved continuous growth through scale advantages.

Obama at Home depot

(10) VISA's predecessor is the payment business sector of Bank of America Corp.

In 1958, they launched the Bank Americard card with new credit function.

In 1976, it was reorganized into VISA bank card company to maintain international business expansion.

VISA's success is representative of the introduction of IT technology into financial industry, constantly opening up new applications and moving towards Blue ocean market.

(11) Samsung Electronics Company was established in 1938 as subsidiary of Samsung Corp founded in 1969 and entered Mobile industry in 1997.

From 1997 to 2010, Samsung has been waiting for 13 years until the come of smart mobile phone era. Apple may subvert the traditional mobile phone industry, but stick in the mud with changeless product. Samsung seize the opportunity to launch a smart phone with large screen. This product meet the needs of the diversified market. From then on, they rise as the new generation of industry leaders.

(12) Established in 1995, Amazon Co is one of the world's first electronic Business Companies.

After 2000, Amazon grew into a retail giant competing with WAL-MART and the world's largest electronic Business Company through global expansion, construction of the logistics center and utilization of cloud computing technology.

Kindle

(13) As one of the first Internet search engines, Google was founded in 1998 and soon became a leader in the industry. They replaced the traditional media forms step by step and developed in to the earliest industry leader in the Internet era.

Since 2000, Google has maintained a high degree of attention to innovation and launched a series of great products, such as Google maps, Android operating system. Still, they continue to explore in advanced IT areas such as robotics, unmanned vehicles and wearable devices.

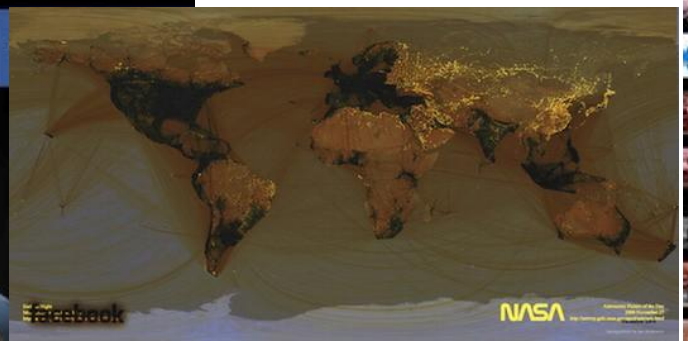
Google™

Android operating system

Google glasses

(14) The latest rise of IT Internet giant is Facebook who was founded in 2004. After 2006, it developed from the campus market to the whole social network. They went beyond MySpace in 2008 and has been leading the western social networking market since then.

Facebook's success is the success of continuous micro innovation and continuous improvement of customre experience so that Facebook defeat the industry leaders one after another and become a new generation of Internet giants.



（15） HUAWEI is China's main supplier for information and communications solutions with headquartered in Shenzhen City, Guangdong province. HUAWEI was registered and established in 1987 with its sphere of business spread in telecommunications networks, business networks, consumers and cloud computing.

- HUAWEI has engaged in 64 of 130 LTE commercial networks worldwide in 2012.
- In 2013, it reached a revenue of 240 billion RMB and a profit of 29.4 billion RMB.
- Currently, HUAWEI products have been applied to more than 170 countries and for serving the 1/3 global population.

（16） Tencent was founded in 1998 when it was a follower to the first instant messaging products ICQ. However, it has become the world's first social networking platform through innovation in product and business model. After 2003, Tencent became the world's largest network entertainment company.

In 2011, Tencent developed WeChat, a mobile platform for communication in a timely manner and considered to hold the world's largest customer base.

Tencent's success relies on ,firstly, innovation in business model and secondly, adapting to the market and transformation in a timely manner.

（17） In 1999, Alibaba was established. It was originally supplying information for China's foreign trade industry. In 2003,it launched Taobao ,an online trading market, and quickly gained the market popularity through the free C2C transaction business model.

Since 2010, Alibaba launched a series of new applications, such as Tmall, Ali cloud, Rookie network, YuEbao, online insurance Mass security, continuously expanded in internet business, logistics and finance, and became one of the most successful internet giants in the world.

Alibaba's success, relies on innovative business model and adaptation to constant seek for new blue ocean market.

The practical application of big data in Guizhou

4

Changes in scientific research

Empirical science （past）

Experiment → observation → Experiment Correction

Computation Science（present）

Model → Computation → Model amendment

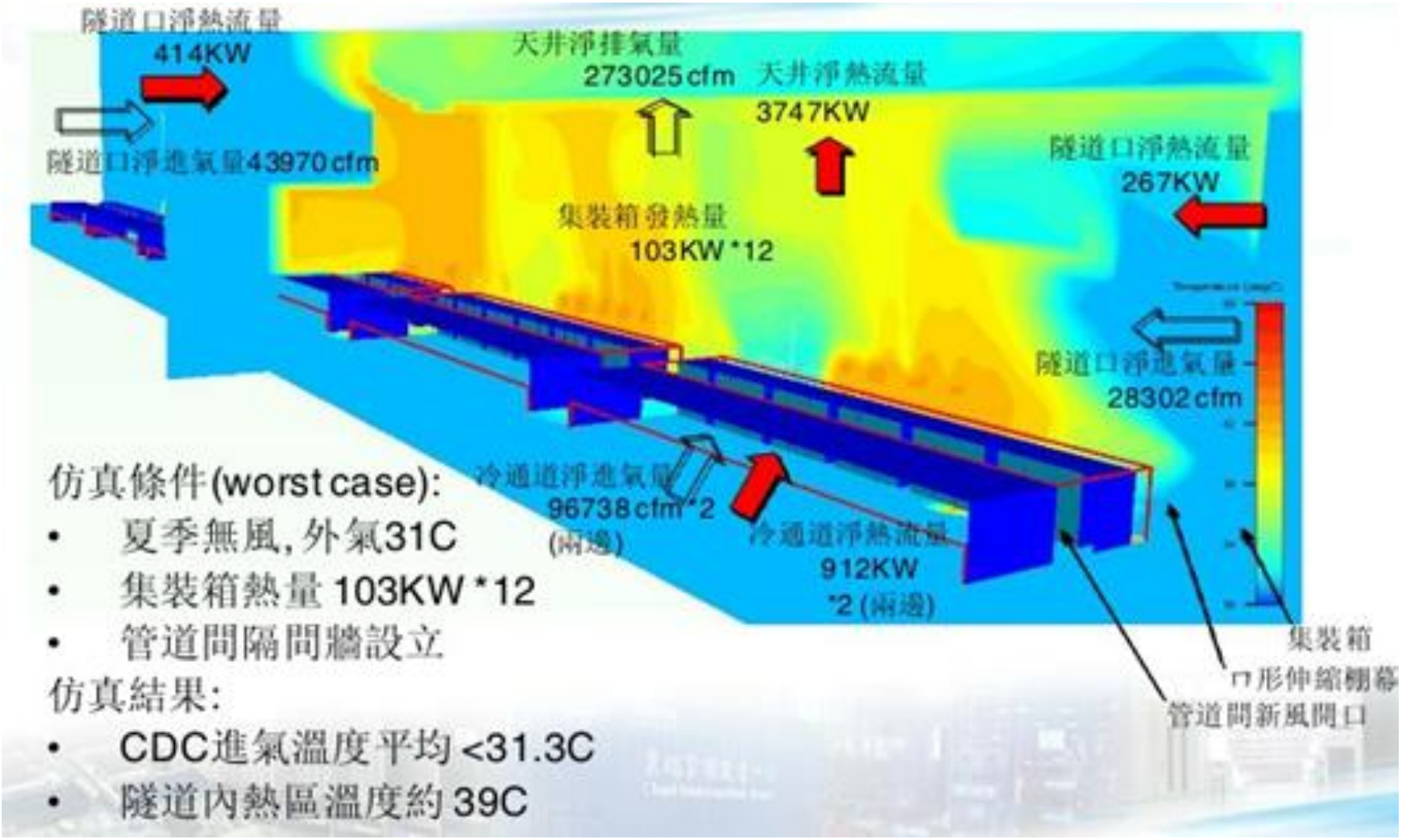
Data Science（future）

Data → Computation → Conclusion

(1) Green Data Center

Gui An trust investment-FOXCON: Green data center like tunnel

Relying on karst topography, we build the tunnel like data center in the pass between mountains.
Its energy consumption indicator PUE （Power Usage Effectiveness） is <1.1 which stands in the forefront of the world.



Heat flow simulation of tunnel like green data center, GuiAn trust investment-FOXCON

Guizhou Information Park of cloud computing of China Telecom

Data center utilized high efficient double cold source system cooled by water chilling unit of centrifugal type and natural cooling system . PUE is forced to a low value .

China Mobile data center (Guizhou)

Refrigeration method of China Mobile data center (Guizhou) combines the water chilling unit of centrifugal type and indirect cooling mode by cooling tower. The average PUE value is about 1.393.

Cloud computing base of China Unicom (Gui An)

Cloud computing base employed central air-conditioning system with efficient cold water unit and water pump. The systems will adjust the air blower according to load changes of engine room, to reduce the power of the fans. The ultimate PUE value was 1.6

Xinpu High-tech Industrial Park of Xiamen SunShine group

Projects is constructed at accelerated speed. The industry focuses on R & D and production of 3D TV, tablet PC, 3G/4G mobile phone, mobile phone touch screen and other supporting end products. The output value will reach 5 billion RMB.

(2) FAST project : Five hundred meters Aperture Spherical Telescope

FAST Into big data : Five-hundred-meter Aperture Spherical radio Telescope

satellite map of FAST station

Site：Tai WoTa Village at Pingtang county In Guizhou

Impression drawing

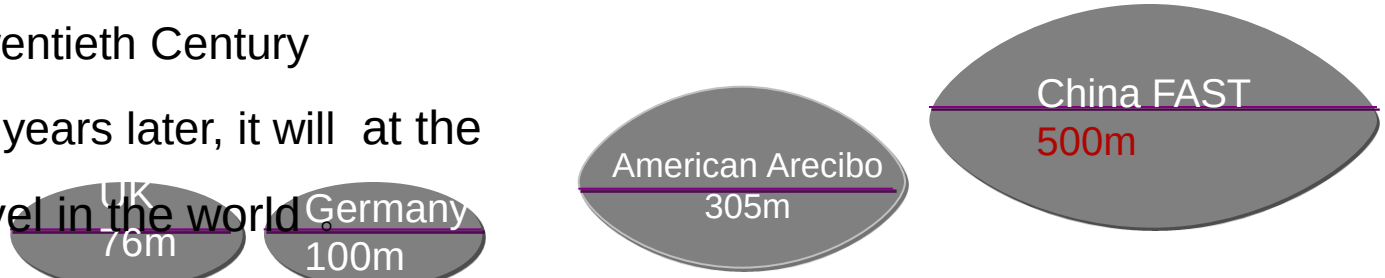
FAST Into big data : Five-hundred-meter Aperture Spherical radio Telescope

Contrast of foreign Aperture Spherical Telescope

- High sensitivity
- High cruising speed
- The comprehensive performance is 10 times than that of Arecibo the world's largest single aperture radio telescope in American, which is the first of the ten major works of mankind in twentieth Century
- Even 20-30 years later, it will at the advanced level in the world。
- Make an analogy, standing at Duyun FAST will see clearly all jewelry in jewelry's at wall street.

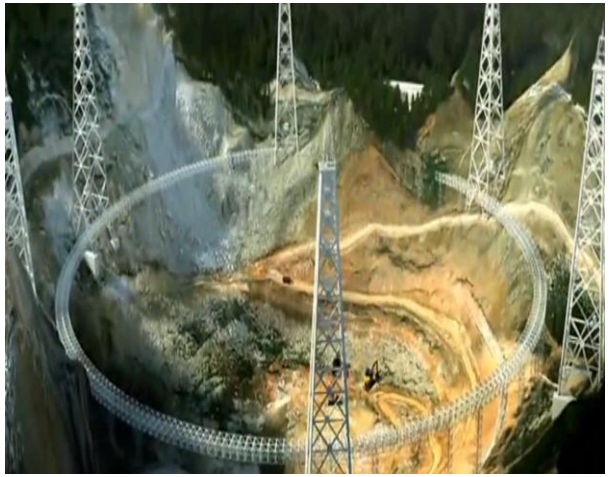
美国：Arecibo

中国：FAST



FAST Into big data : Five-hundred-meter Aperture Spherical radio Telescope

Working progress



FAST Into big data : Five-hundred-meter Aperture Spherical radio Telescope

FAST under **construction**

FAST project

FAST :working site

FAST: drainpipe of 1.2km

FAST Into big data : Five-hundred-meter Aperture Spherical radio Telescope

FAST under construction

FAST: Installation of f support tower for feedback source

FAST: working site

FAST Cable net installation

FAST into big data : Five-hundred-meter Aperture Spherical radio Telescope

FAST under construction

FAST: Installation of reflector panel mounting

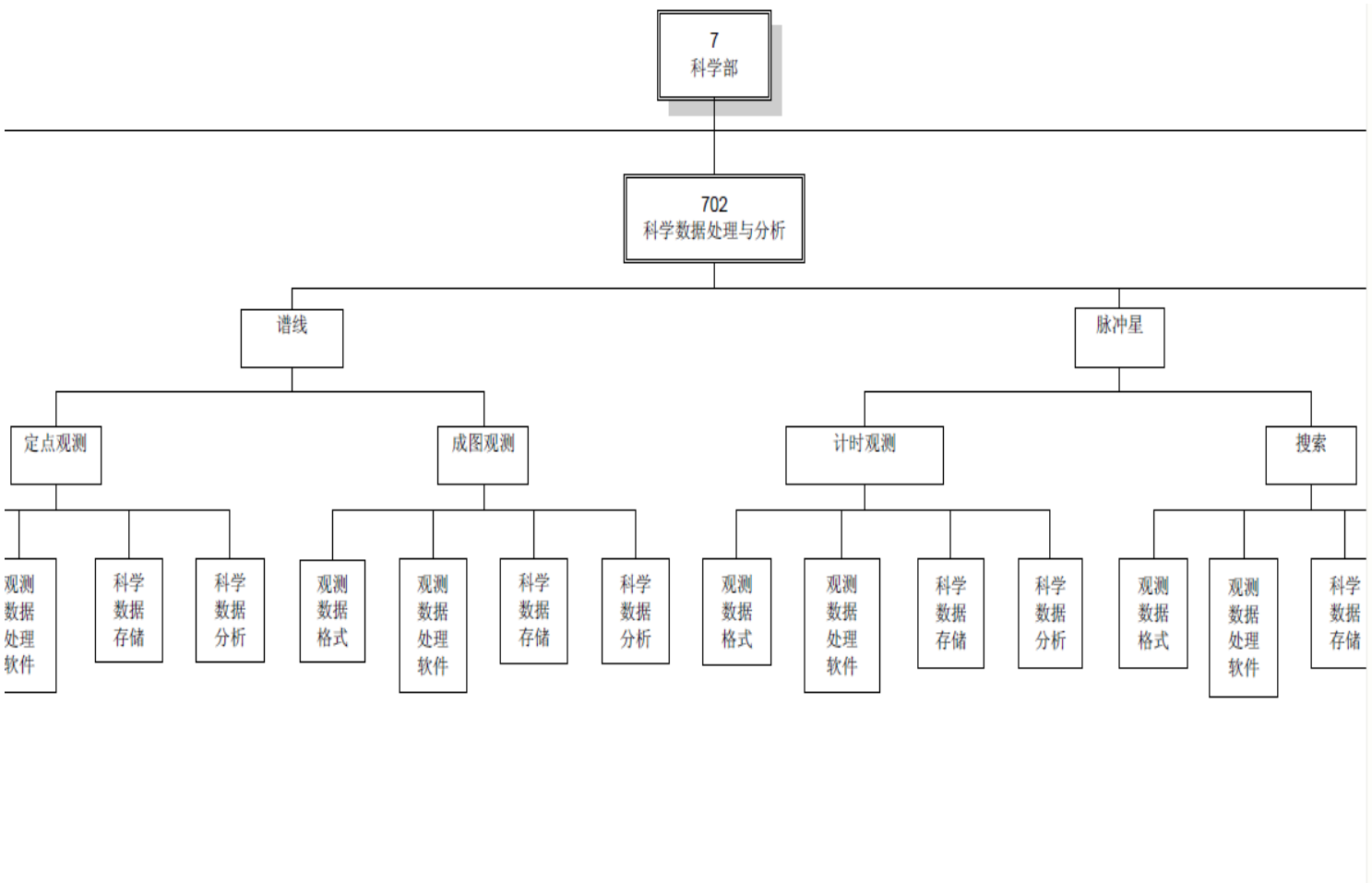
FAST into big data : Five-hundred-meter Aperture Spherical radio Telescope

FAST: Scientific significance and frontier issues

- Extend the neutral hydrogen observation to the edge of the universe; detect interstellar molecules; study star formation; reproduce the image of early universe; explore the origin of space life, look for extraterrestrial civilization.
- Establish the pulse star time array and participate in the autonomous navigation and gravitational wave detection in the future.
- Explore the physical structure and the physical law of the extreme state and acquire the astronomical ultra fine structure
- High resolution microwave inspection and weak spatial signal detection.。

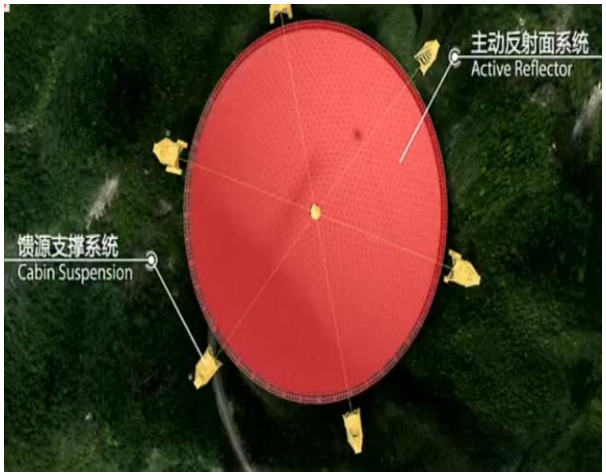
FAST Into big data : Five-hundred-meter Aperture Spherical radio Telescope

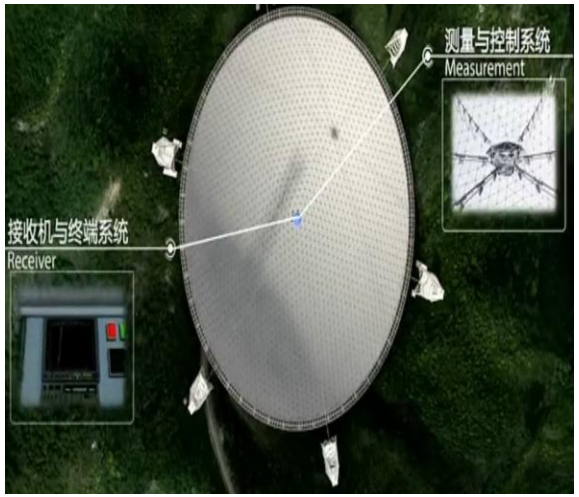
FAST: Subsystem of scientific data processing



FAST Into big data : Five-hundred-meter Aperture Spherical radio Telescope

FAST:Subsystem of scientific data processing receives the space signal,

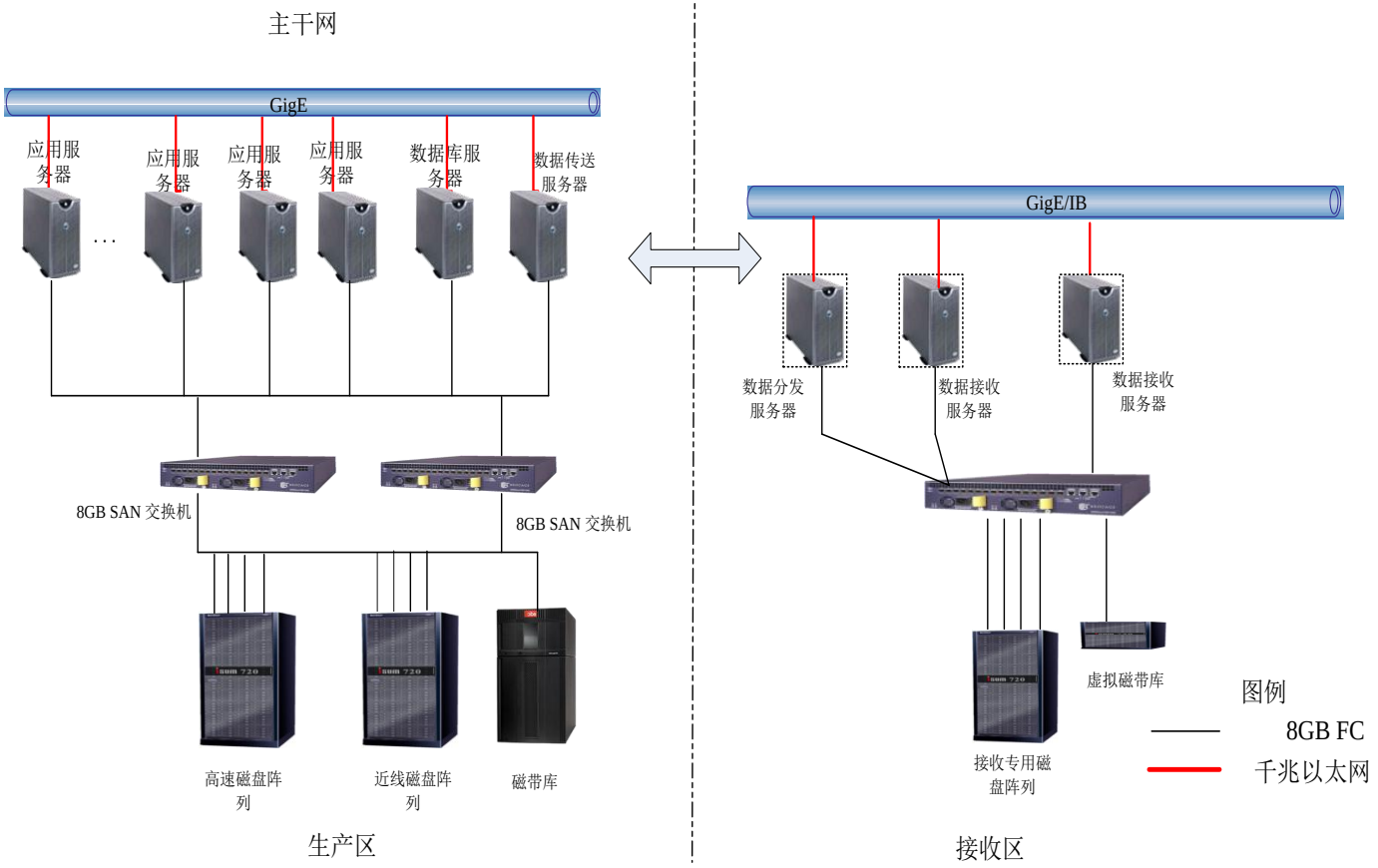




each subsystem coordinates to
receive signals from space and search the sky

FAST Into big data : Five-hundred-meter Aperture Spherical radio Telescope

FAST: Design of scientific data storage system



FAST Into big data : Five-hundred-meter Aperture Spherical radio Telescope

- 80MB/S*19~1.6GB/s~5TB/h
- Observation~8h/day, ~40TB/day
- Medium: disk, tape
- Scheme
  - For short term storage :disk (1PB) and for long-term storage magnetic tape (20PB)
  - Disk
- Tai Wo Taipa to Guiyang
- Raw data
  - Survey of 8 hours per day
  - 2GB/s*19=38GB/s ~ 1PB/8h
- Pulsar survey
  - ~40TB/8h（reduce ~25 times）
- Transmission
  40TB/24h ~ 0.5GB/s

FAST Into big data : Five-hundred-meter Aperture Spherical radio Telescope

FAST: Calculation requirement

- Observe singe beam data in 5 min
- 8k chan
- 0.1ms Time resolution
- 8k*10000*8bit=80MB/s(160MB/s if 2pol)
- 8k*3M*8bit=24GB
- The first round of the galactic plane survey 10deg*70deg
- Data volume: 17000*19*24B ~ 8PB
- Overall demand ~ 20PB

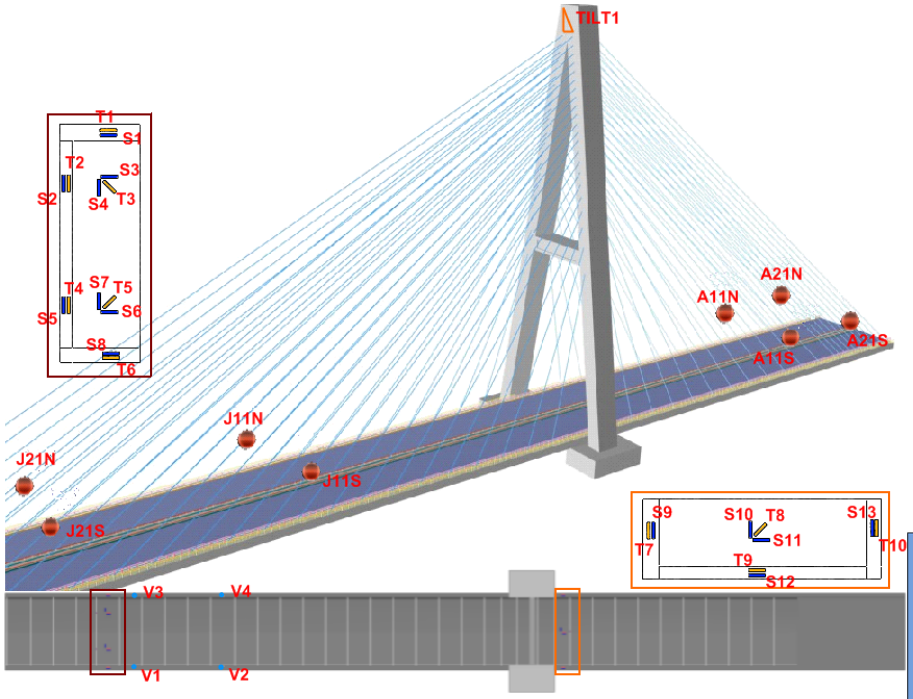
- Mainly from the pulsar search
- Still limited by hardware
  - Huge amount of FFT calculation
  - IO constraint
- Advantage： can be simply parallelized
- FAST basic requirement 200T FLOPS
- 1000CPU ~500 2-socket server
- Assume GPU~10CPU(?),then ~ 25 2-GPU server

(3) Big data monitoring of bridge safety

① Random sampling

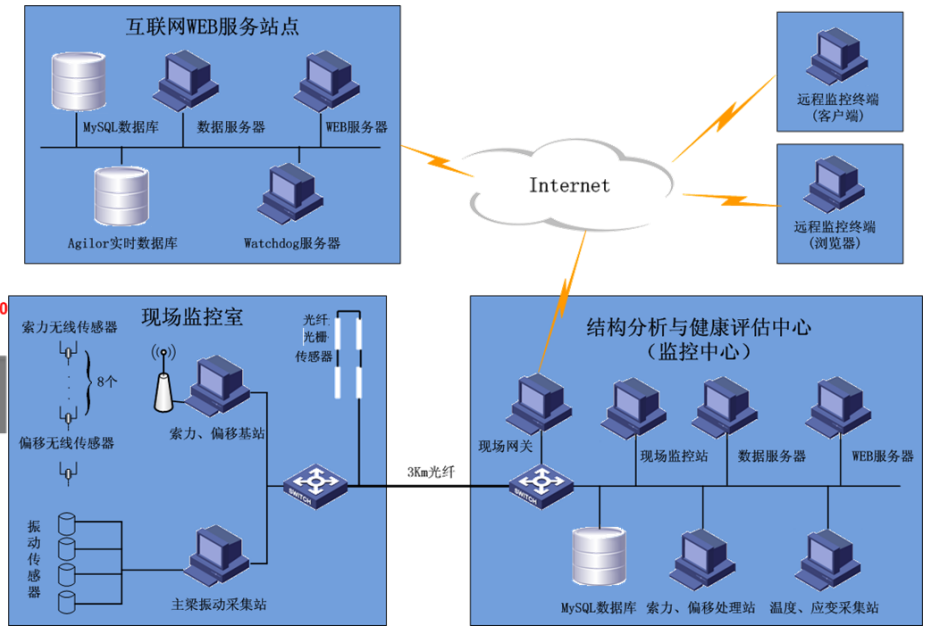
Sampling: Get more information with the least data

Bridge health monitoring system of Hongfeng Lake Bridge



Sensor distribution

System deployment



Health monitoring of bridge of Hongfeng Lake Bridge

监测项目	测点位置	传感器类型	传感器数量	监测方式
塔顶偏移	索塔塔顶	双轴倾斜仪	1	实时
斜拉索索力	斜拉索	无线加速度传感器	8	实时
动态特性	主梁	振动加速度传感器	4	实时
主梁应变	主梁控制截面	光纤光栅应变传感器	13	实时
温度监测	主梁控制截面	光纤光栅温度传感器	10	实时

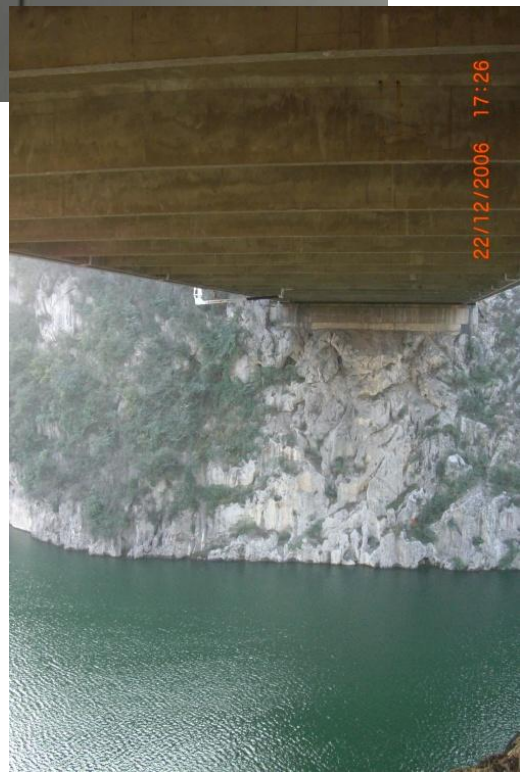
The design principle was extracted from health system of bridge safety maintenance in Guizhou Province. The design was performed according to function-first and cost control principle. Monitoring content were determined by analysis of the data of the bridge operation and typical Problem distribution, according to the characteristics of Hongfeng Lake bridge

Financial and hardware investment costs has been significantly reduced by utilizing a number of innovative technology(Cable force prediction, integration of multi sensor information, limited information evaluation, etc.).

Installation of vibration and optical fiber sensor


22/12/2006 16:55


禁止驶入
NO ENTRY
22/12/2006 9:59


22/12/2006 17:26


22/12/2006 11:18


22/12/2006 19:04

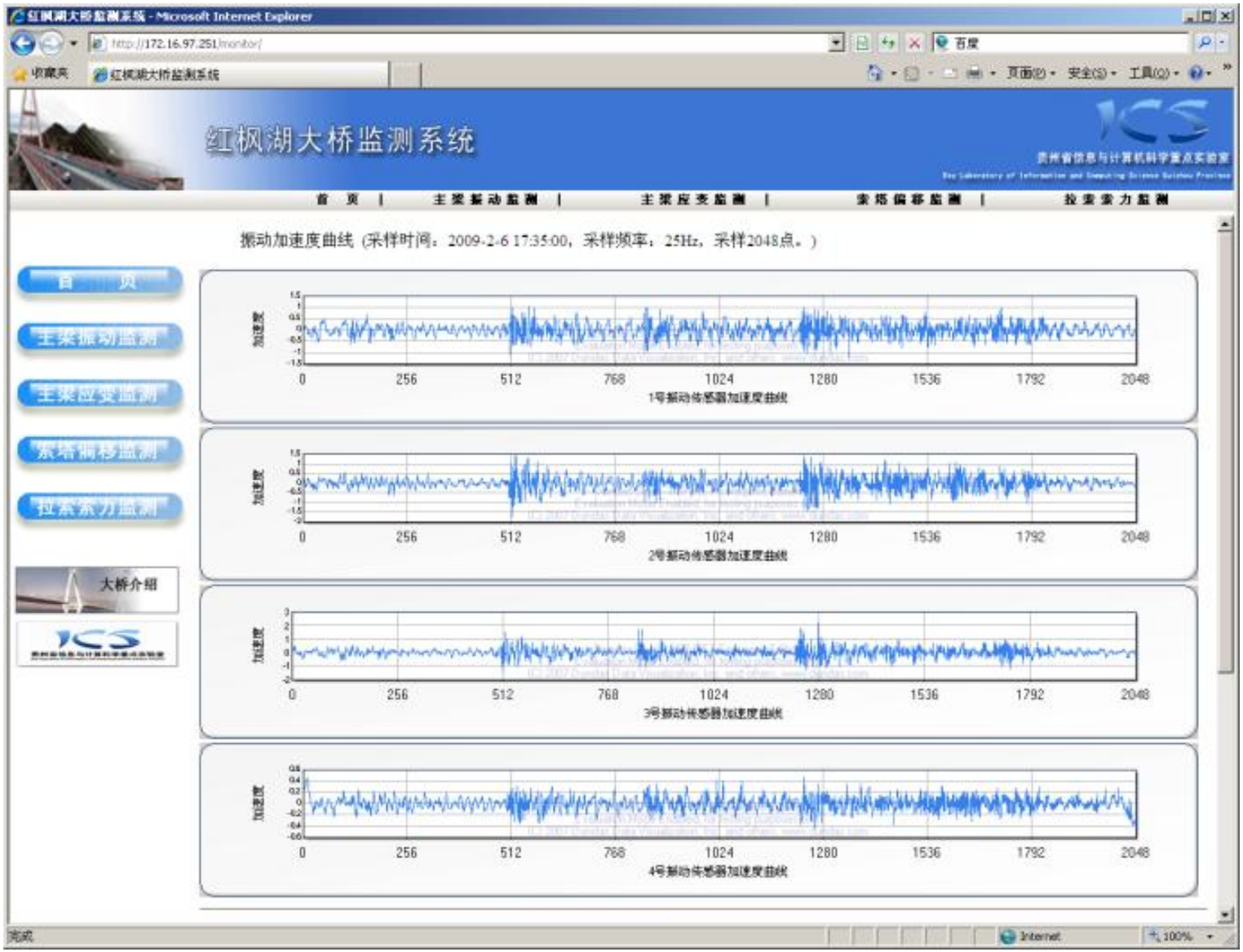

22/12/2006 18:49


22/12/2006 12:39

Installation of Cable force sensor

貴州師範大學
Guizhou Normal University

Monitored data



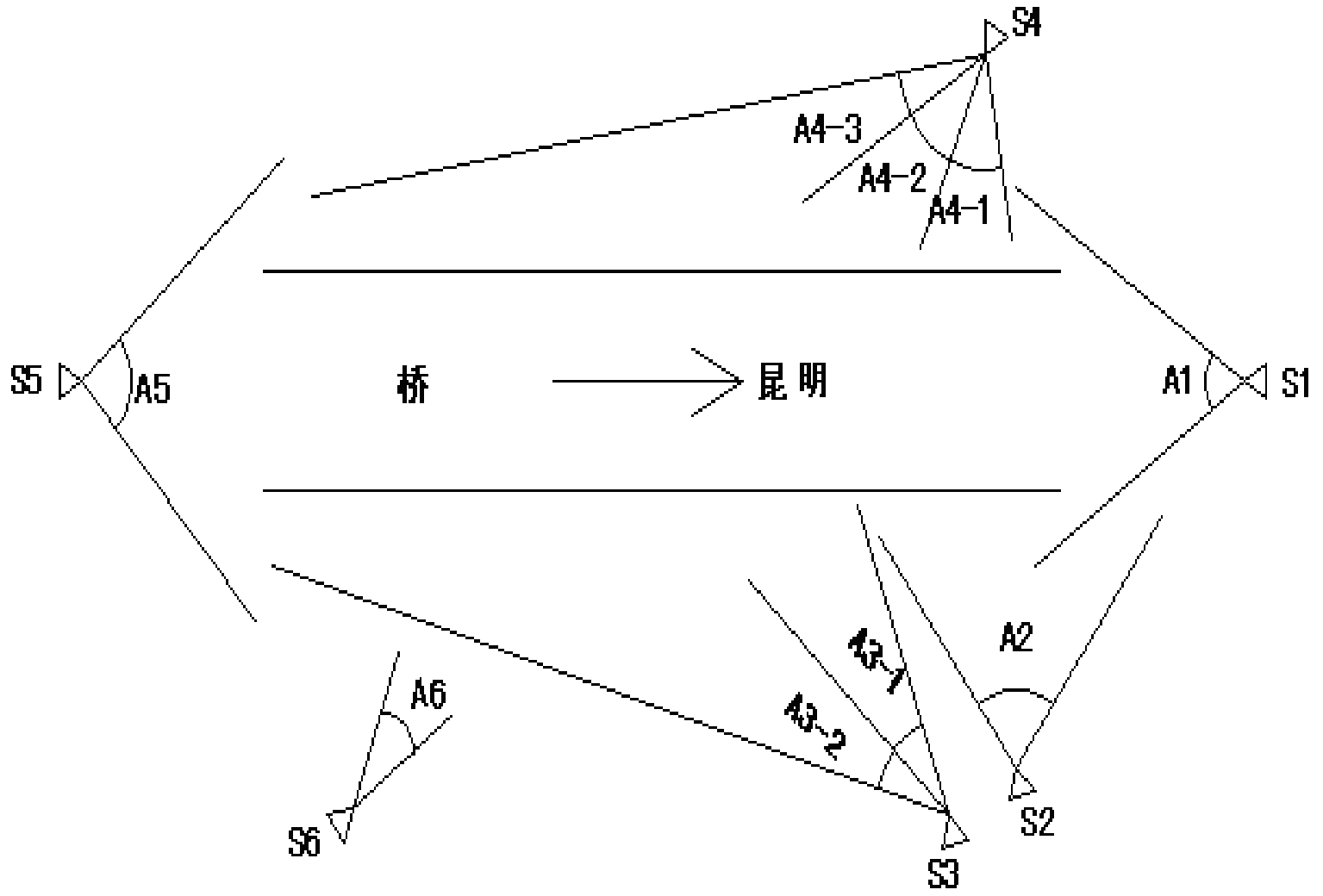
- Sampling: Get more information with the least data
- But when we can get massive amounts of data, this approach should not be so important 。
- Big data era: things changed：
  - Population instead of sampling。
    - We can analyze more data independent of random sampling. ；
  - Efficiency instead of precision
    - We are no longer interested in pursing accuracy in light of such huge amount of data.
  - Correlation instead of causality 。
    - We are seeking for the correlation with litter interest in causal relationship..

② Big data process

- 3D information acquisition method used in monitoring Beipanjiang bridge and Jiaxiulou
  - 3D laser scanner： ILRIS-3D

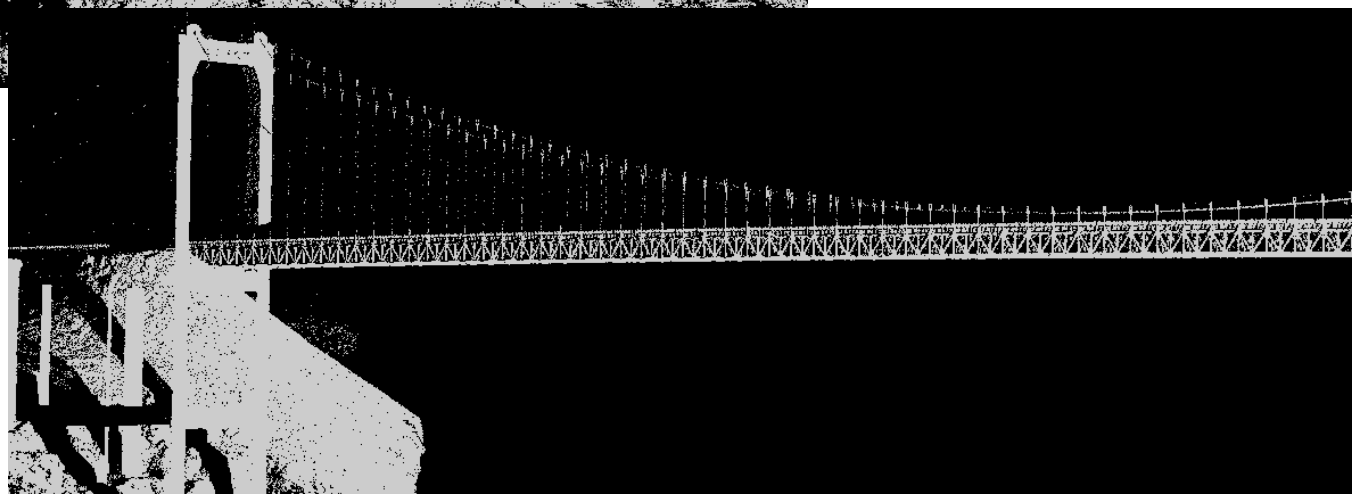
Fast speed
High precision
Extensive scope
Large amount

Layout of sampling point in Beipanjiang bridge：6 points with 9 scanning , acquiring a total of 10 million points with 5G size data

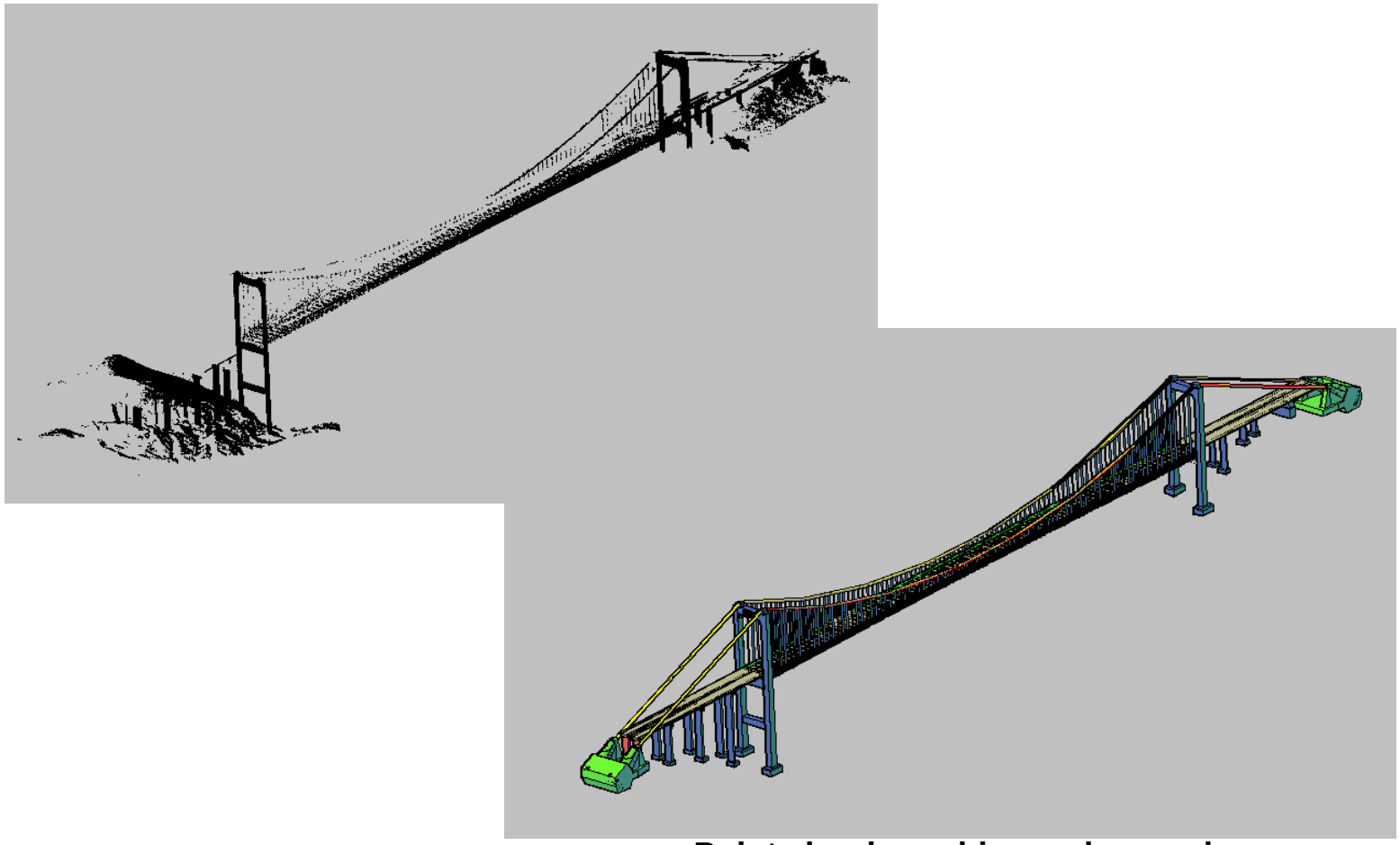


- Scanning scenario

- Scanning data

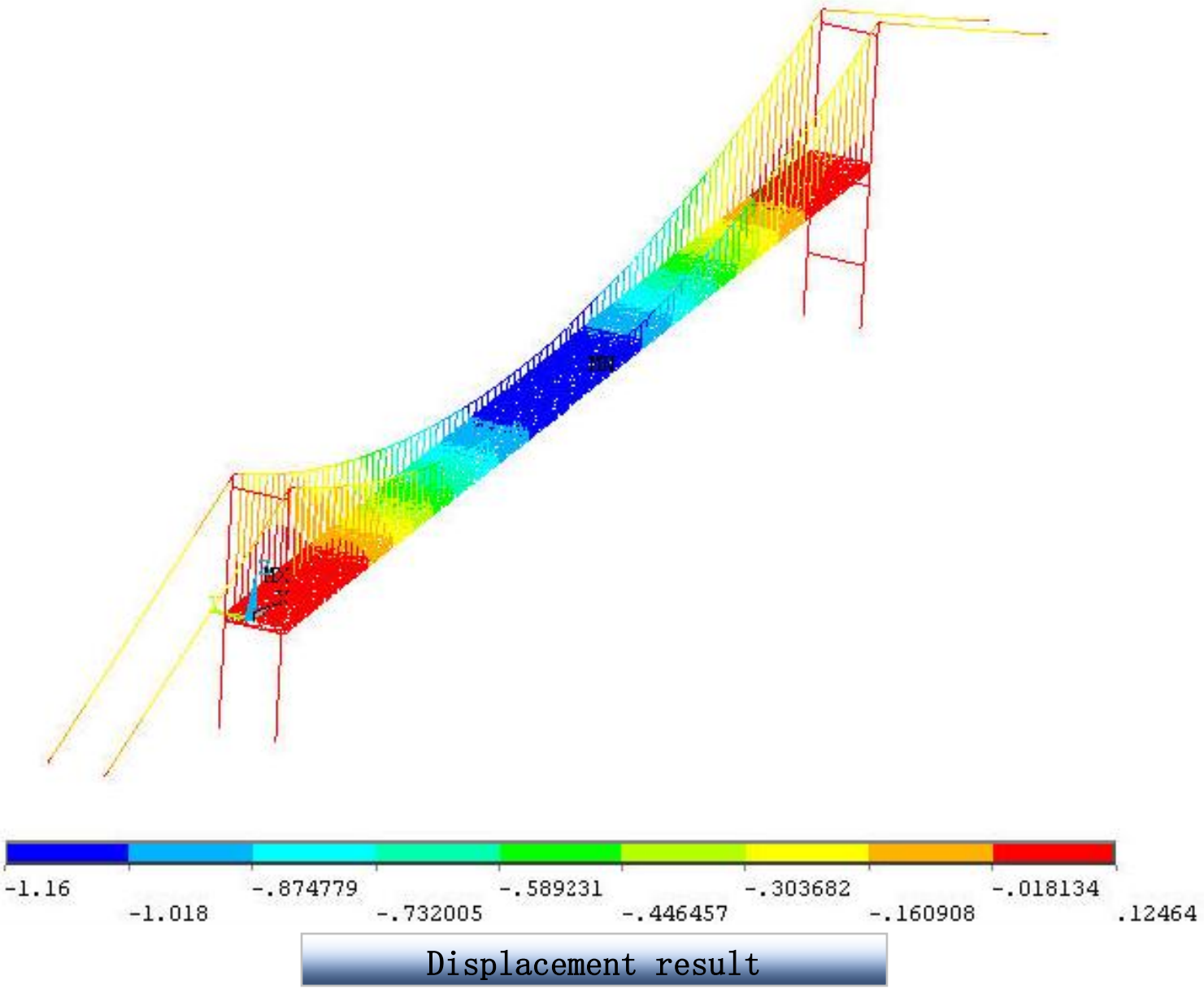


- **Traditional finite element calculation was carried out based upon Ideal model at design stage.**
- **Our computation utilized the real-time bridge model**
- **The results are more close to actual state。**



Point cloud graphics and reversing model

Calculation result



Displacement result

(4) Data calculation of highway fee

Highway fees Calculation for Guizhou province

- Highway track in use of guizhou is about 3483 km, of which containing 291 charged bridge (223,812 meters) and 222 charged tunnels (318,525 meters) .
- According to Twelfth Five-Year Plan of Guizhou:by the end of 2015, Highway track in use will reach 4,500 kilometers; by 2020 reach 6,000 km ,by2030 reach 7,000 kilometers.
- Since the topological structure of road network goes complex and the presence of discount problem, highway fees calculation require High performance calculation more and more.

Highway fees Calculation for Guizhou province

- the relationship between tolls station and path number

  – no flag station

$$R = S \bullet (S - 1)$$

  – flag station

$$R = S \bullet (S - 1) + S^2 \bullet \sum_{k=1}^{M} \frac{M!}{(M - k)!} \quad 。$$

：

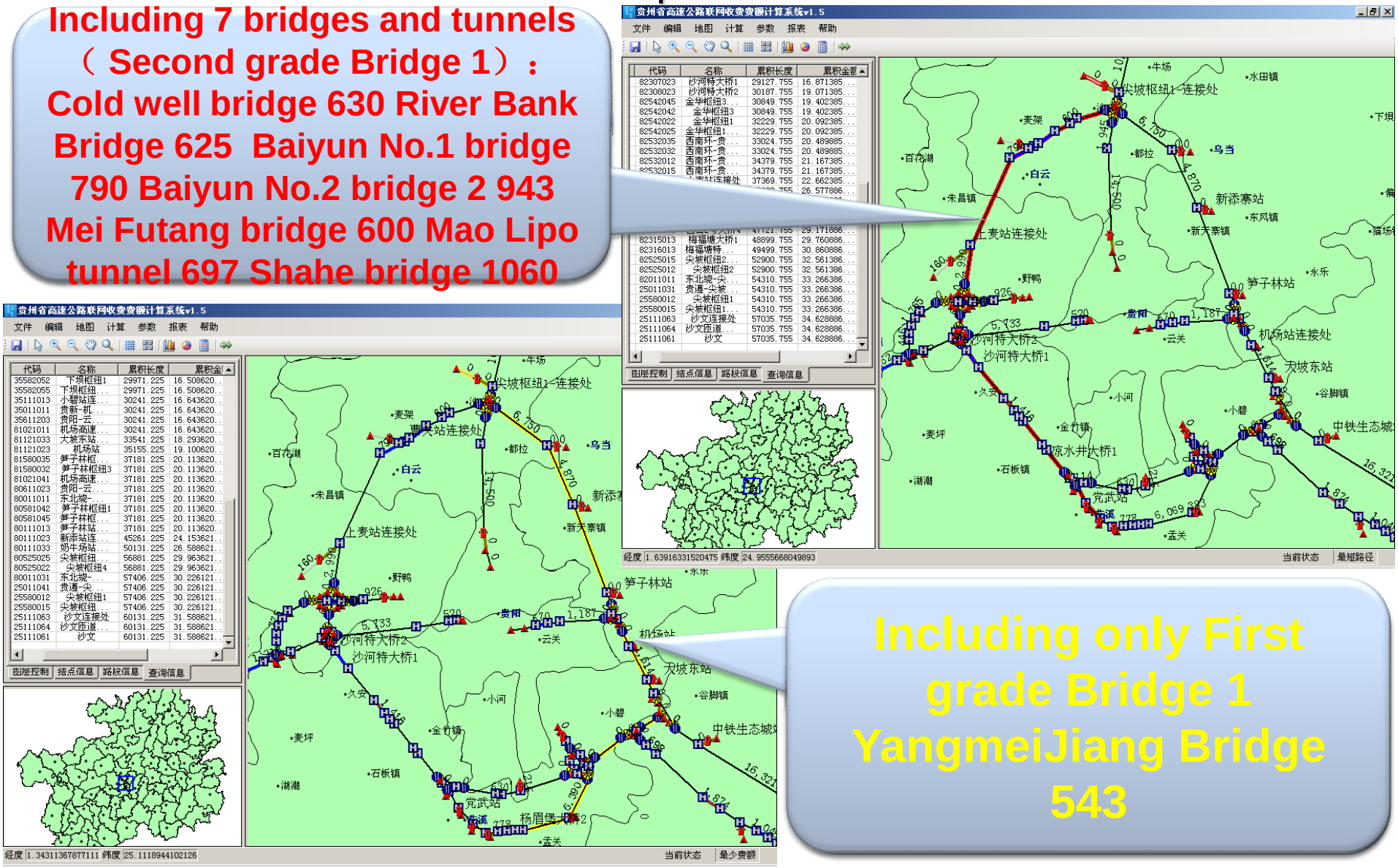
S	Number of tolls；
M	Number of marking stations；
R	Number of road paths。

Highway fees Calculation for Guizhou province

- Each calculation process generated more than 100 millon records, and exported nearly 30millions(V5.0 version) data,
  - The section table of the path with least fee has 6295642 records
  - The section table of the shortest path has 6265920 records
  - The section table of ambiguous path with the least fee has 3,901,012 records in the tables.
  - The section table the shortest ambiguous path routing have 3,876,564 records
- Each calculation generated data with more than 30,000 pages
- The entire computation on traditional platform spent more than 16 hours, but with the high-performance computing platform the entire process will end in 2 hours 。

Searching the path of minimum charge, splitting the shortest path

Including 7 bridges and tunnels（ Second grade Bridge 1）: Cold well bridge 630 River Bank Bridge 625 Baiyun No.1 bridge 790 Baiyun No.2 bridge 2 943 Mei Futang bridge 600 Mao Lipo tunnel 697 Shahe bridge 1060



Including only First grade Bridge 1 YangmeiJiang Bridge 543

Completed highway

1. Sinan to Jianhe highway
2. Qingshan to Mengzhiqioa highway
3. Zunyi to Suiyang highway
4. Shuicheng to Pan highway
5. Zunyi to Sinan highway
6. Bailakan Maitai highway
7. Yuping to Kaili highway
8. Congxihe to Zunyi highway
9. Guiyang to DuYun highway
10. Kaili to Majing highway
11. Shuikou to Duyun highway
12. Guiyang to Zunyi highway
13. Qingzhen to Huangguoshu highway
14. Guiyang to Xinzhai highway
15. Zhenning to Shengjingguan highway
16. Mawei to Jiaou highway
17. Guiyang to Huangguoshu highway
18. Guiyang Northeast highway
19. Guiyang Airport highway
20. Guiyang Southwest highway
21. Guiyang South highway
22. Guiyang to Qingzhen highway
23. Liping to Luoxiang highway
24. Banba to Jiangdi highway
25. Zunyi to Bijie highway
26. Tongren to Dalong highway
27. Huishui to Xingren highway
28. Anshun highway
29. Anshun ti Puan highway
30. Qinglong to Xingyi highway
31. Daxing to Sinan highway
32. Guiyang to Huishui highway
33. Renhuan to Chishui highway
34. Liuzhi to Zhenning highway
35. Qianxi to ZHijin highway
36. Bijie to Weining highway

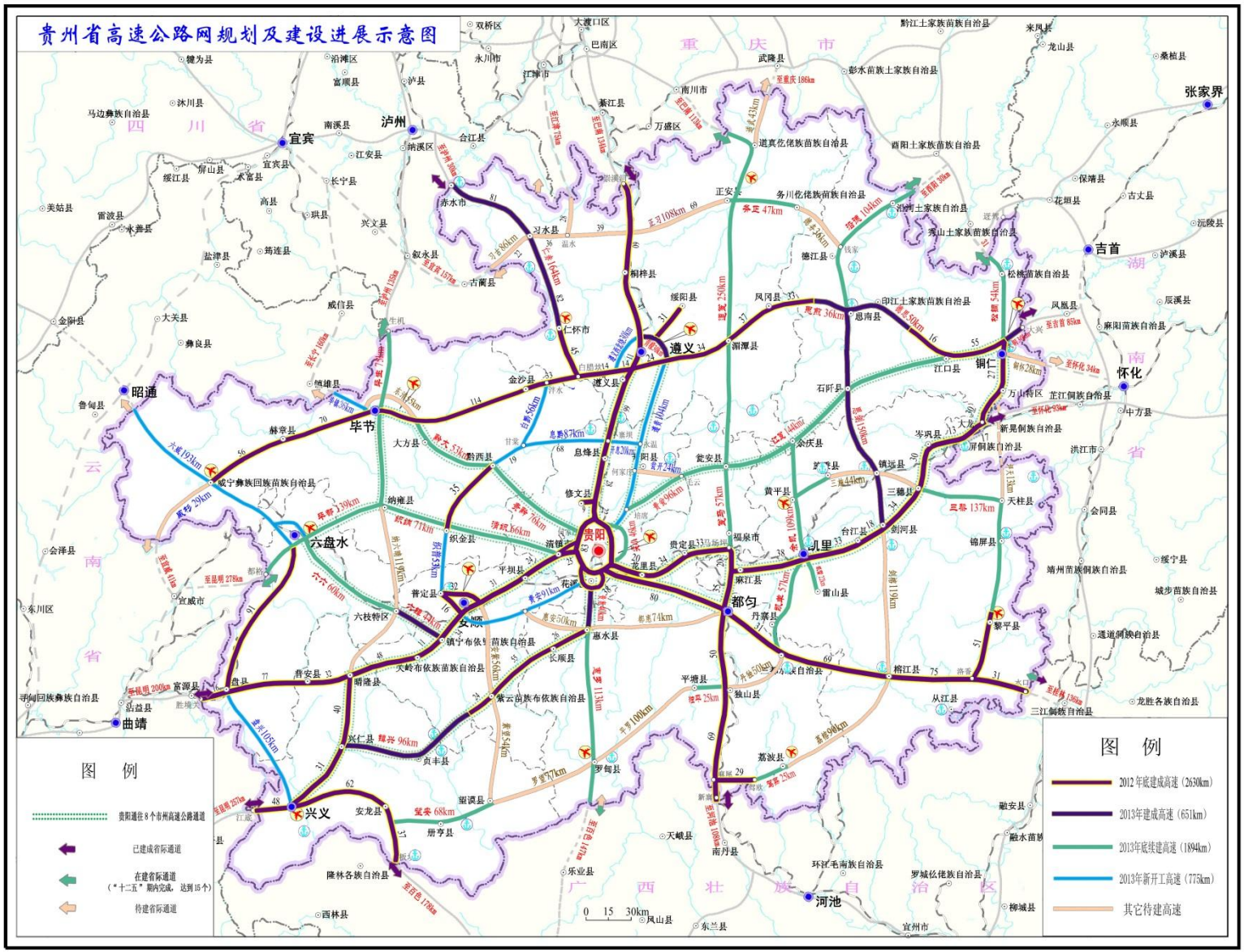
Three years of Guizhou construction
Schedule of highway constructed in 2014

No.	Project name	Passing County (city, district)	connected	Company in charge	Track in use (km)	Track completed in 2014 (km)	Invers ment (hundred million)	Land (Hectare)	The proportion of bridges and tunnels	Expected completion
1	Kaili to Yangjia	Kaili、danzhai		Guizhou Expressway Group Co., Ltd.	56	56	41.79	338	23%	December 20th
2	Yuqing to Kaili （Yuqing to Huaingping 44km+Shibing partly 23km+Huzhuang to Yatang 10km (Road toll mode online)	YuQing,shibing,Huangping，Kaili,Yuqing	Yuqing,Shibing,Huangping	Guizhou Expressway Group Co., Ltd.	110	77	84.66	584	26%	December 30th
3	LiuPanshui to Liuzhi	Shuichen.Liuzhi		Guizhou Expressway Group Co., Ltd.	60	60	62.58	382	37%	September 18th
4	Qingzhen to Zhijin	(Qingzhen,Pingba,Zhijin		Guizhou Expressway Group Co., Ltd.	66	66	50.95	420	32%	December 16th (Except Sanchahe Bridge)
5	Sansui to Liping	Sansui,Tianzhu,jinping,Liping	Tianzhu,Jinping	Guizhou Southeast Expressway Investment Ltd.	138	138	101.18	718	32%	December 10th
6	Zhijin to NaYong	Zhijin,Nayong	Nayong	Guizhou Expressway Group Co., Ltd.	71	71	66.49	420	33%	October 30th (Except Wuzuo he Bridge, Nayong Bridge, Longjing River Bridge, Nayong tunnel)
7	Bijie to Duge （Bijie to dongguan 6km）	Qixingguan,Nayong,，ShuiQiao		Guizhou Expressway Group Co., Ltd.	140	26	141.37	734	45%	Ddecember 20th

Three years of Guizhou highway construction
Schedule of highway constructed in 2014

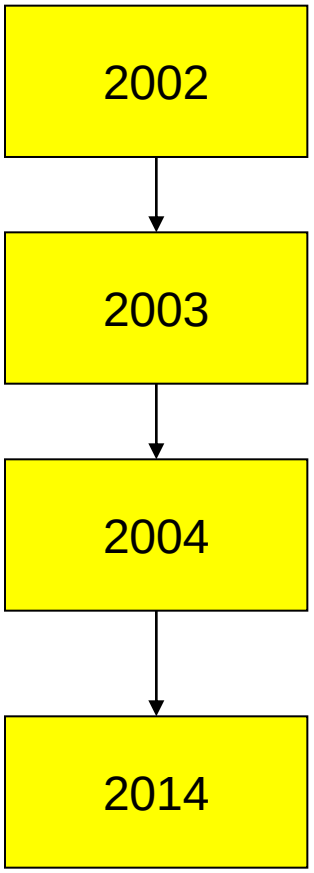
NO.	Project name	Passing County (city, district)	通县	Company in charge	Track in use (km)	Track completed in 2014 (km)	Invers ment (hundred million)	Land (Hectare)	The proportion of bridges and tunnels	Expec compl n
8	Jiaou to Libo	Libo	LiBo	Guizhou Expressway Group Co., Ltd.	26	26	24.96	112	50%	Decen 24th
9	Bijie to Shengji(Connection Hangrui, Bi Wei, Bidu)	QIxingguan,Shuangshan		Guizhou Expressway Group Co., Ltd.	74	7	67.45	507	32%	Decemb 22th
10	Wengan to Machangping	Wengan,Fuquan	Wengan	Guizhou bridge construction Refco Group Ltd	56	56	48.29	376	24%	Decemb 25th
11	Shiban to Dongguan（Shiban to Datu 30km+Liulong to Dongguan 11km）	Qianxi,Dafang		Guizhou bridge construction Refco Group Ltd	52	41	54.27	359	31%	Decemb 25th
12	Wuchuan to Zhengan（Wuchuan to Hujiaba 10km+Xiaoshuigou to Peiyang 12km）	Wuchuan,Zhengan	Wuchuan	Guizhou Provincial Highway Bureau	48	22	48.59	196	54%	Decemb 26th
13	Mengzhiqiao to Leli	Huichuan,Honghuagan,Zunyi		Zunyi Guizhou Highway Construction Investment Co., Ltd.	30	30	24	177	27%	Decemb 27th
14	Songtao to Tongren （Songtao North to Jiangjun Mount 37km）	Songtao	Songtao	Guizhou Highway Engineering Group Co., Ltd.	50	37	53	290	51%	Decemb 19h
15	Dushan to Pingtang （8km in Pingtang partial ）(Road toll mode online))	Dushan,Pingtang	Pingtan	Guizhou Luqiao Group Co., Ltd.	24	8	26.44	142	55%	Decemb 28th
	Total		11		1001	721				

Planning map of highway network of Guizhou province



(5) In-Depth mining in large data of High school and college entrance examination

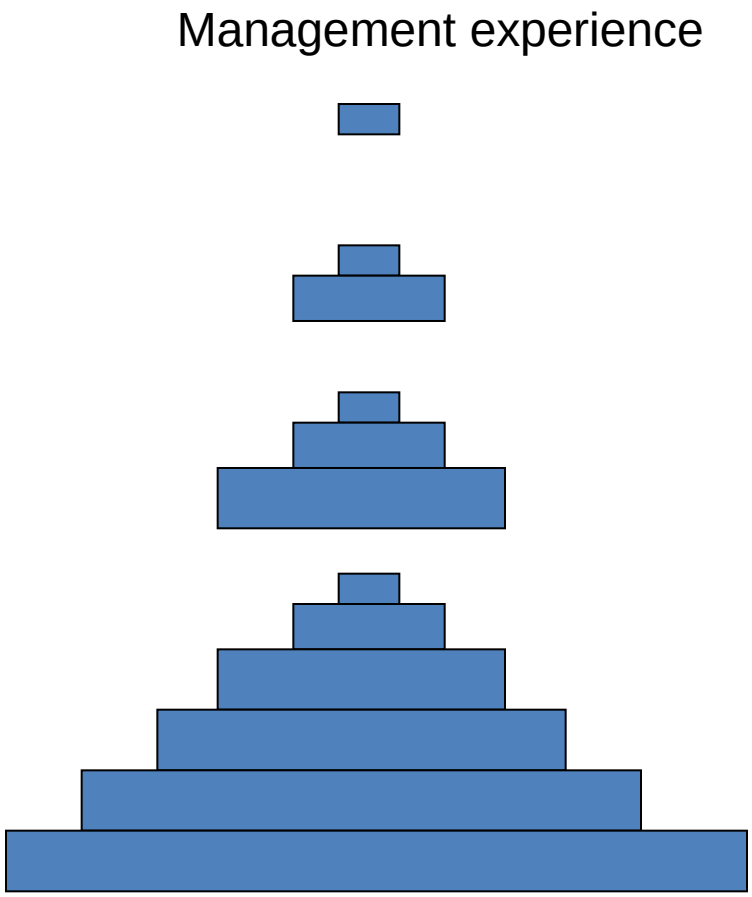
Process of paperless marking of Guizhou Normal University



Contact

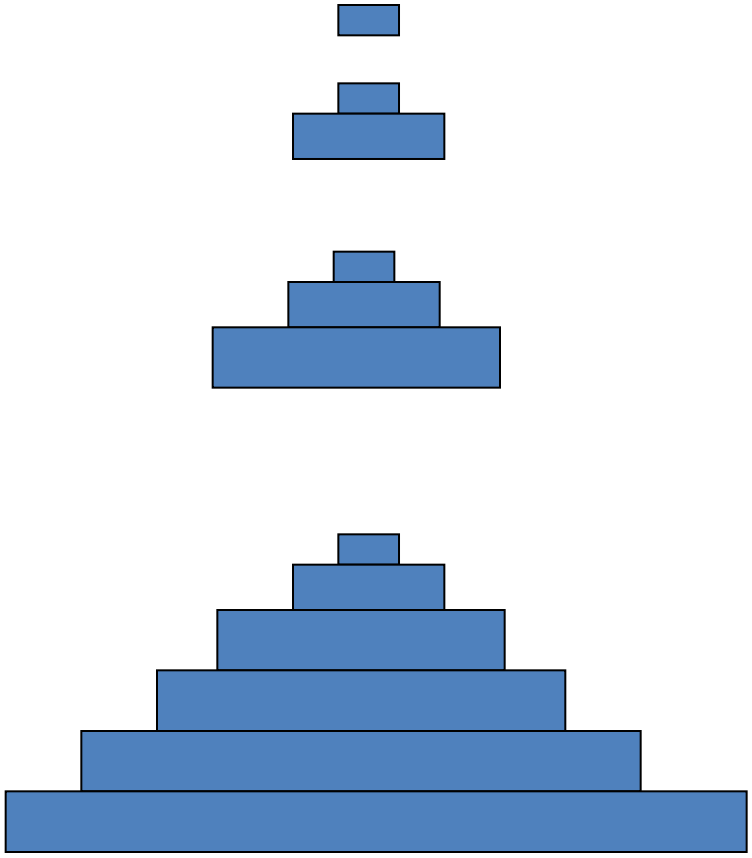
Initiative

Application



Online paperless marking system developed independently in Guizhou normal university

Schedule of development and application



2006 Survey

2007 Development and test

2008 Runing on high school entrance examination of Guiyang

2009 Simulation test for college entrance of Guiyang, extended to the Qianxi, Qiannan

2014

Analysis of data from online marking System of Guizhou Normal University

At the end of 2007, Guizhou Normal University completed the independent development of the online paperless marking system, and successfully applied it in service to all kinds of tests. But the most important and valuable core lies in analysis of test data to improve the quality of paper, to find the weakness of students and to guide teachers' instruction for many exams.

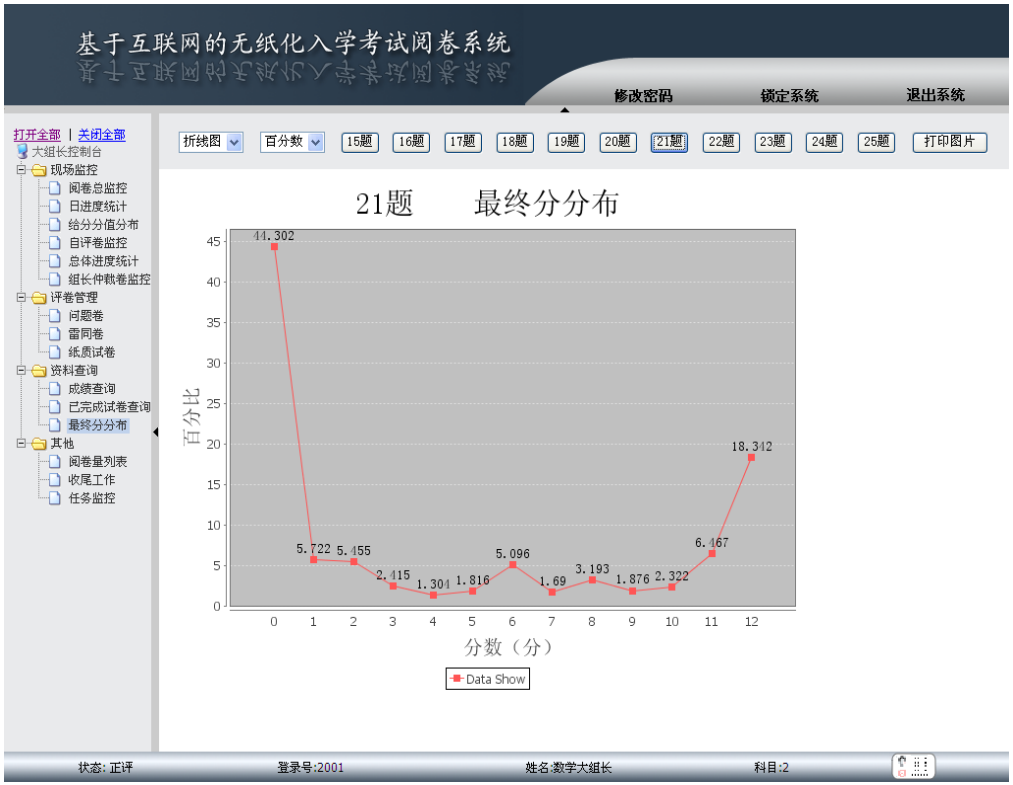
From 2008 on , RD team of online marking system of Guizhou Normal University has been constantly developing and improving the marking data analysis systems, with in-depth cooperation with Guiyang Education Institute

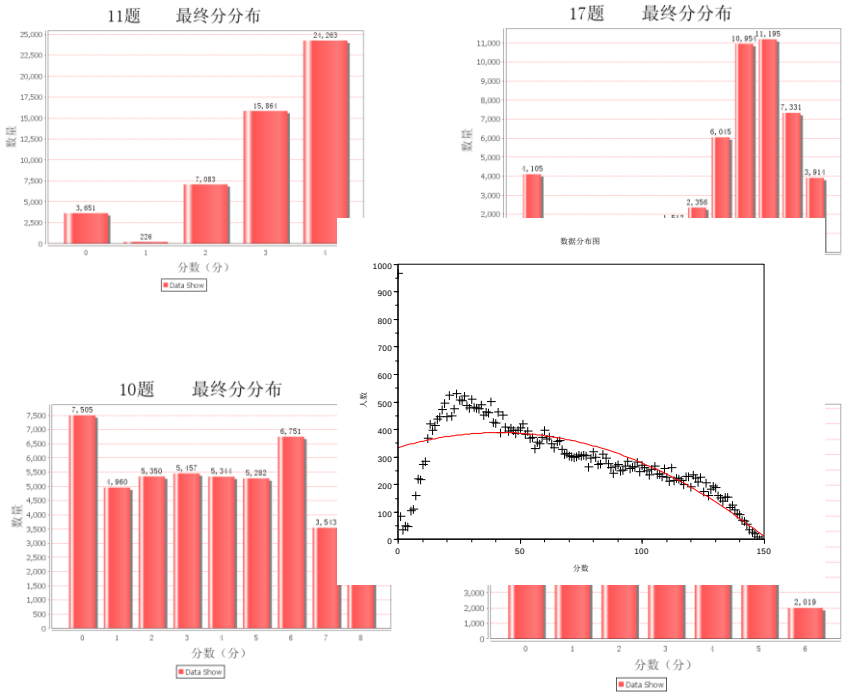
In-Depth mining in large data from online marking system of Guizhou Normal University

Since 2007, as the use of the online marking system with proprietary intellectual property rights ,Guizhou normal university has completed exam 37 times online marking including high school entrance exam, adaptation test for college entrance examination,and diagnostic test for college entrance examination; the total number of papers reaches 7.28 million. The total number of student's answers exceeds 200 million.

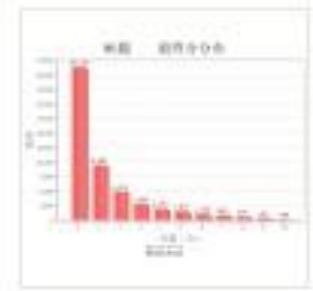
In order to improve the teaching quality of high school teachers, and to improve the students exam performance, RD team utilized the high performance computing platform at Guizhou Key Laboratory of information and Computing Science 's to mine and analyze the test data, extracting a large amount of valuable information for guiding teachers' proposition and teaching.

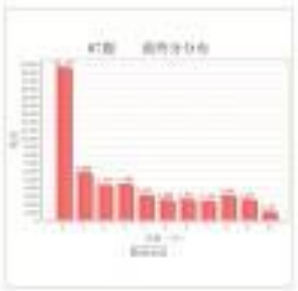
Data Analysis and Statistics

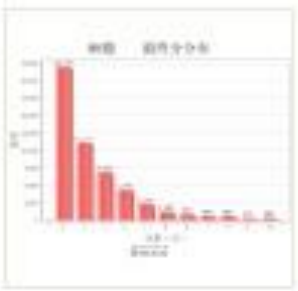


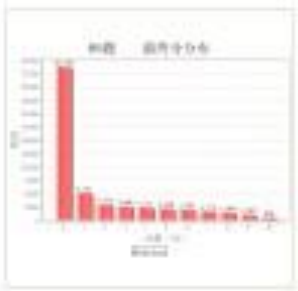


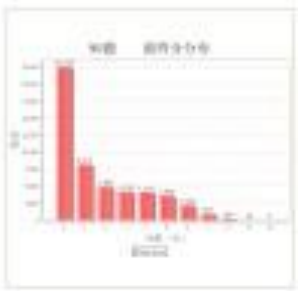
Analysis chart of exam quality


86.png

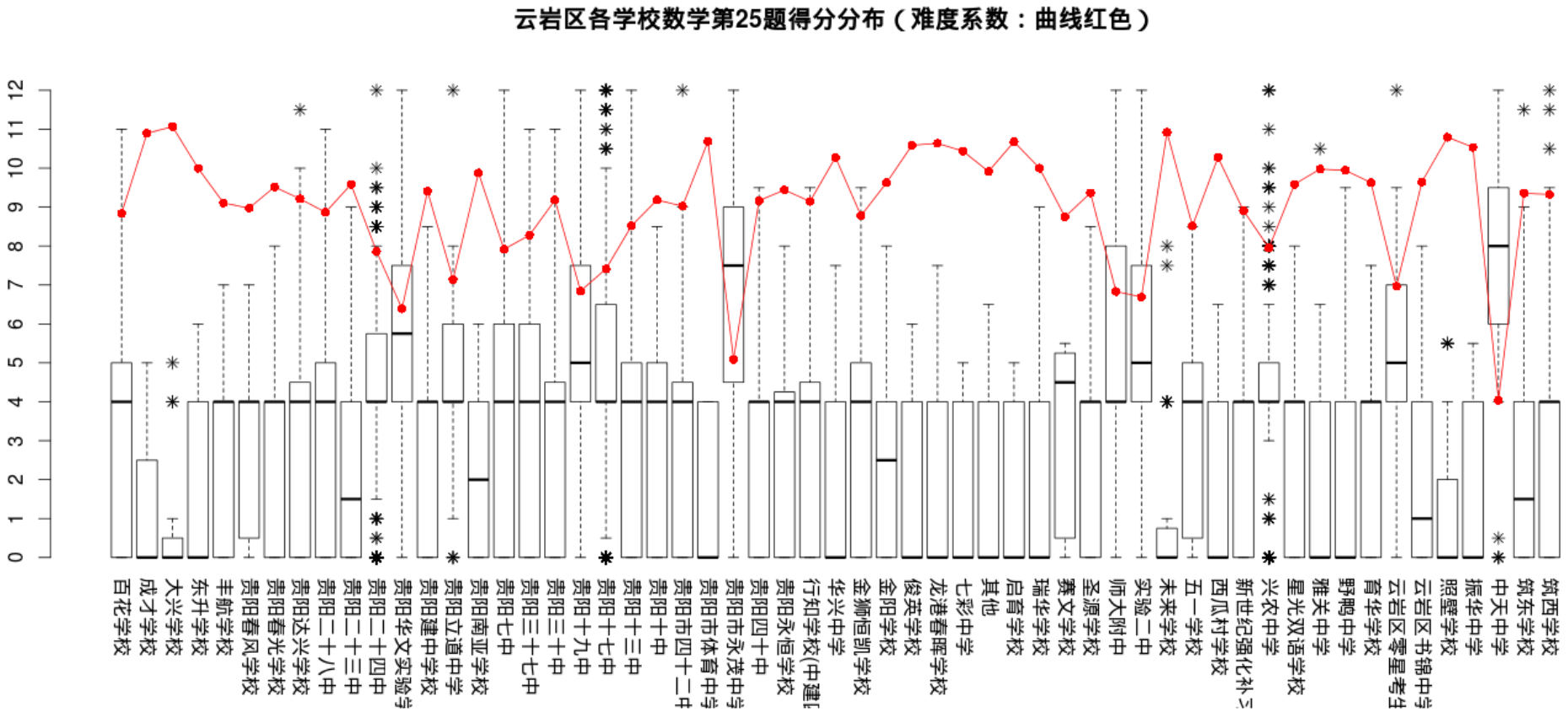

87.png


88.png


89.png


90.png

Data Analysis and Statistics

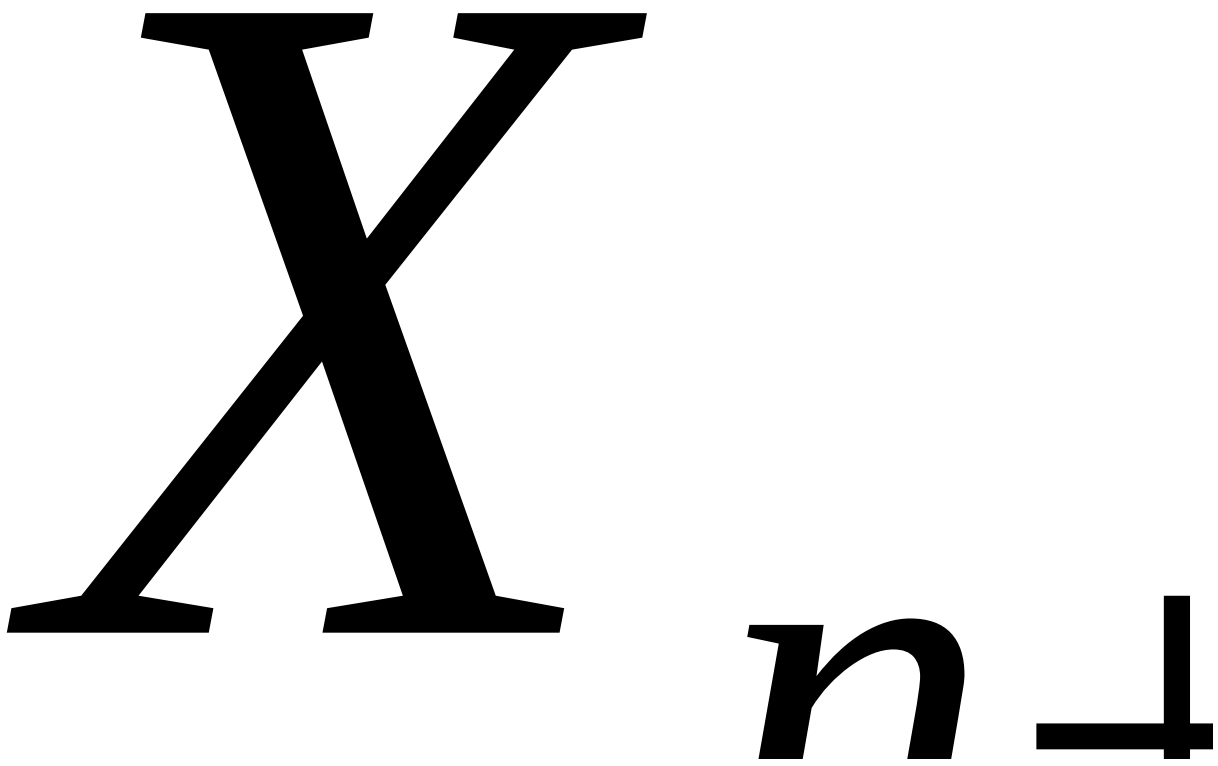


Standard deviation, difficulty degree, discrimination degree, variance and other statistics of the examnation

Guizhou Normal University

科目	题目	总分	标准差	难度	区分度	高分组难度	低分组难度	平均分	最高分	最低分	方差
10	16	20	5.6957812	0.2720281	0.2699290	0.6051979	0.0653398	5.4405620	20	0	32.4419230
10	17	12	3.1990475	0.1694244	0.2387773	0.4953447	0.0177901	2.0330930	12	0	10.2339052
10	18	12	3.5578494	0.2761100	0.2701687	0.6020317	0.0616942	3.3133204	12	0	12.6582926
10	19	12	4.9675469	0.5036552	0.3715895	0.8853589	0.1421799	6.0438621	12	0	24.6765219
10	20	12	2.3904237	0.1463572	0.1668765	0.3585214	0.0247684	1.7562858	12	0	5.7141257
10	21	12	2.2600342	0.0707116	0.1157353	0.2377605	0.0062899	0.8485395	12	0	5.1077546
10	22	10	1.4169133	0.0344613	0.1121376	0.2325980	0.0083228	0.3446131	10	0	2.0076433
10	23	10	3.1251560	0.3358543	0.2410037	0.5267210	0.0447137	3.3585434	10	0	9.7665998
10	24	10	2.1104624	0.1364903	0.1510924	0.3522651	0.0500803	1.3649025	10	0	4.4540516
11	16	6	1.1892672	0.3887867	0.1188875	0.5192586	0.2814836	2.3327201	6	0	1.4143565
11	17	4	0.9985386	0.3599373	0.1555290	0.5342046	0.2231467	1.4397491	4	0	0.9970793
11	18	6	1.2787544	0.3489728	0.1270855	0.4902445	0.2360734	2.0938365	6	0	1.6352129
11	19	5	0.7794616	0.3668524	0.0785113	0.4547801	0.2977575	1.8342619	5	0	0.6075604
11	20	6	1.4248491	0.1903711	0.1440665	0.3695022	0.0813693	1.1422264	6	0	2.0301949
11	21	5	1.5981400	0.6487585	0.1017736	0.8395408	0.6359937	3.2437923	5	0	2.5540515
11	22	6	0.9787025	0.2226725	0.1114003	0.3858950	0.1630943	1.3360353	6	0	0.9578586
11	23	6	1.3355681	0.3551942	0.1273534	0.5486288	0.2939220	2.1311649	6	0	1.7837420
11	24	8	1.5339148	0.3598186	0.0983043	0.5083307	0.3117221	2.8785485	8	0	2.3528946
11	25	5	1.4333771	0.6903448	0.1547367	0.8057086	0.4962352	3.4517241	5	0	2.0545699
11	26	6	1.0824457	0.8035920	0.1254697	0.8821496	0.6312102	4.8215517	6	0	1.1716887
11	27	6	1.2870146	0.4625659	0.1337873	0.5515267	0.2839522	2.7753951	6	0	1.6564066
11	28	8	1.3959151	0.5099093	0.1149946	0.5841665	0.3541774	4.0792744	8	0	1.9485790
11	29	6	1.1295348	0.3867982	0.0634164	0.4643798	0.3375470	2.3207891	6	0	1.2758489
11	30	5	0.8234568	0.1365220	0.0666488	0.2138004	0.0805028	0.6826098	5	0	0.6780811
11	31	60	10.3531265	0.6681045	0.0604731	0.7426872	0.6217409	40.0862679	59	0	107.1872279
12	16	20	5.6061822	0.3184358	0.2756584	0.6221264	0.0708096	6.3687151	20	0	31.4292785
12	17	12	3.6389160	0.4889844	0.2790051	0.7795072	0.2214970	5.8678122	12	0	13.2417099
12	18	12	4.2035480	0.3569259	0.3224143	0.7303285	0.0854999	4.2831104	12	0	17.6698157
12	19	12	3.2829392	0.5083540	0.2138236	0.7338761	0.3062290	6.1002483	12	0	10.7776896
12	20	12	2.9734960	0.2709262	0.2214671	0.5162303	0.0732960	3.2511145	12	0	8.8416784
12	21	12	2.5275382	0.1301116	0.1800427	0.3680201	0.0079348	1.5613397	12	0	6.3884494

Formula of fit index



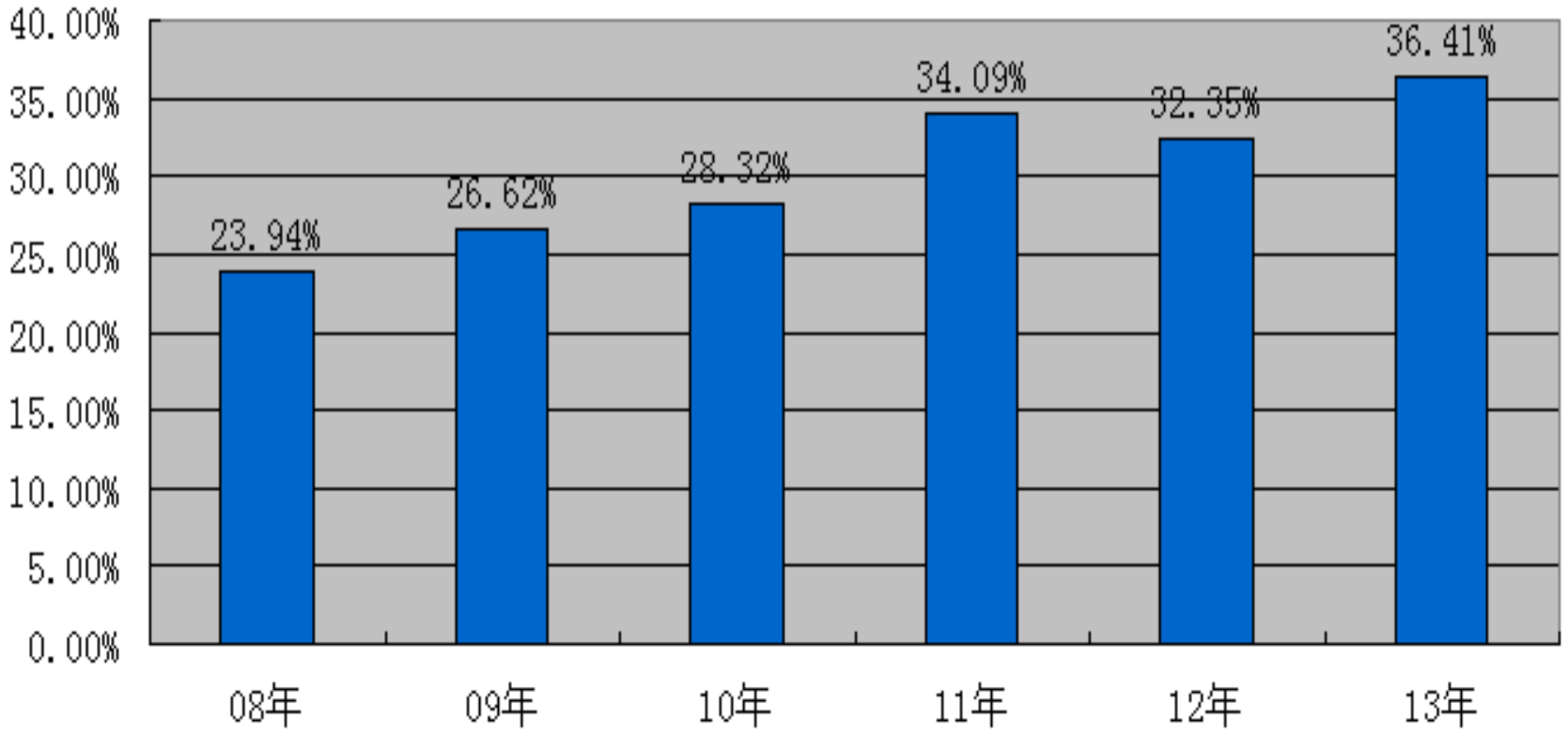
High-performance computing platform in data mining and statistical Analysis of the examnation data

貴州師範大學
Guizhou Normal University

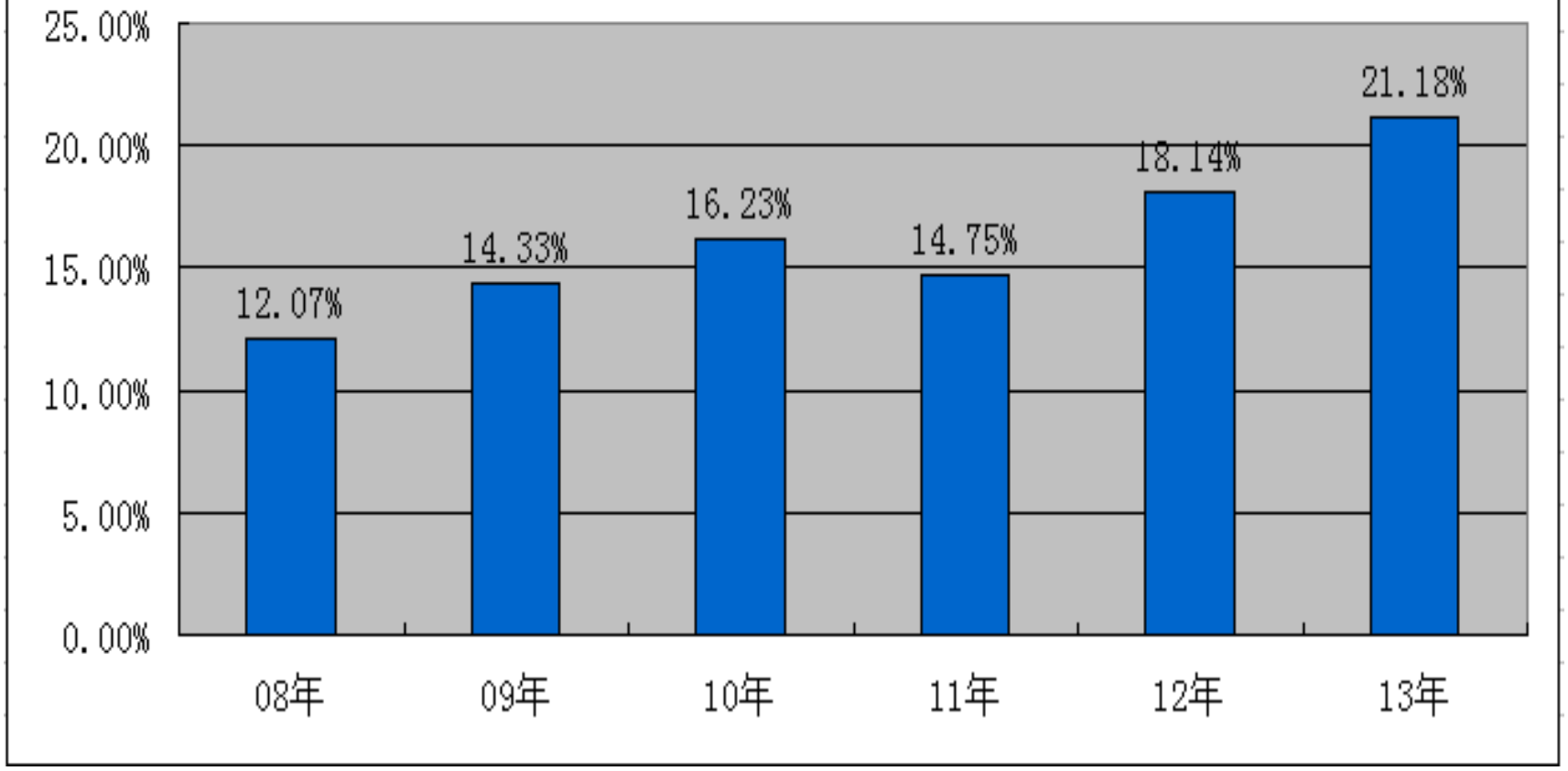
The making scene

Paper Scanner

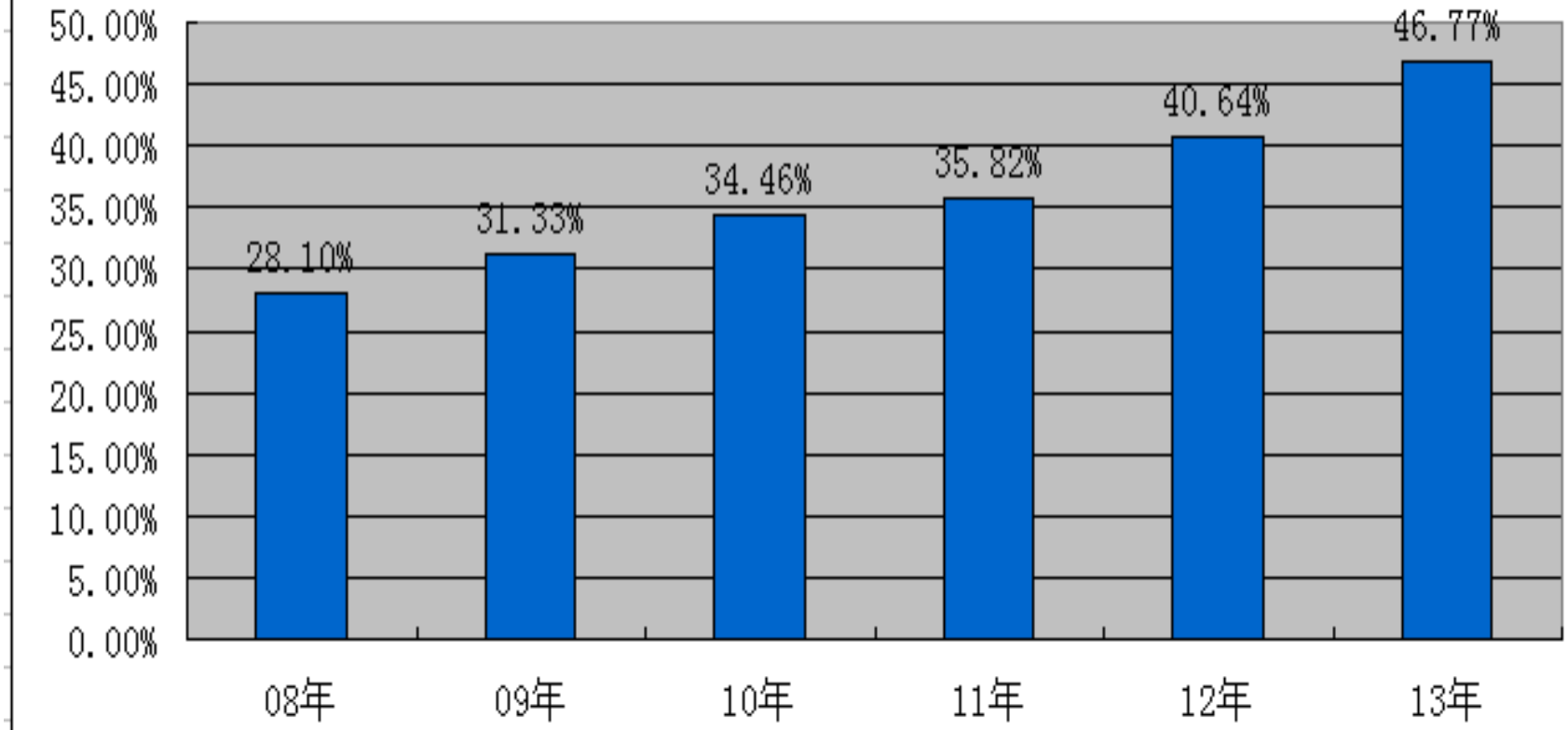
600分以上占贵州省的比例



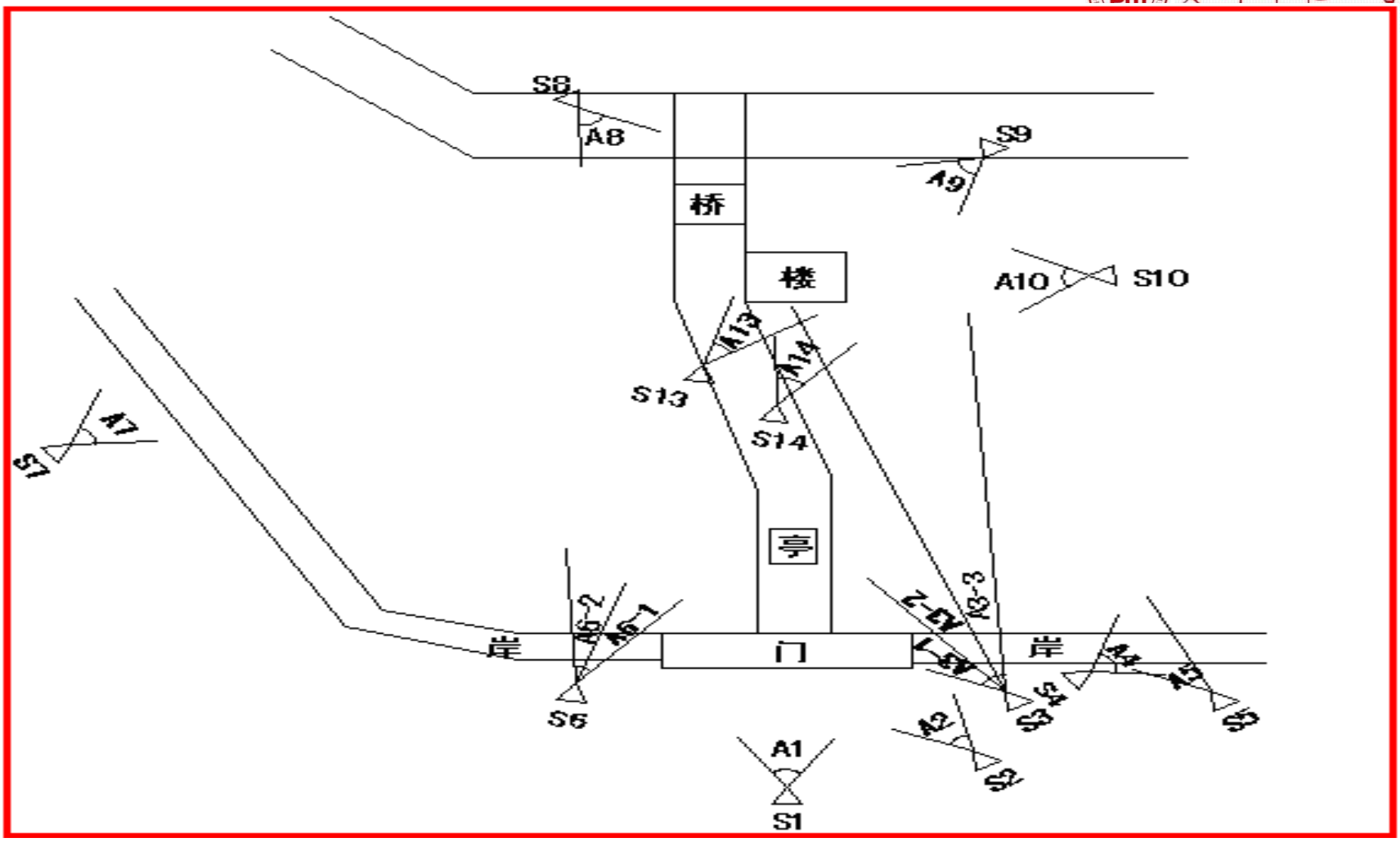
一本上线百分比


25.00%
20.00%
15.00%
10.00%
5.00%
0.00%
12.07%
14.33%
16.23%
14.75%
18.14%
21.18%
08年
09年
10年
11年
12年
13年

二本上线百分比

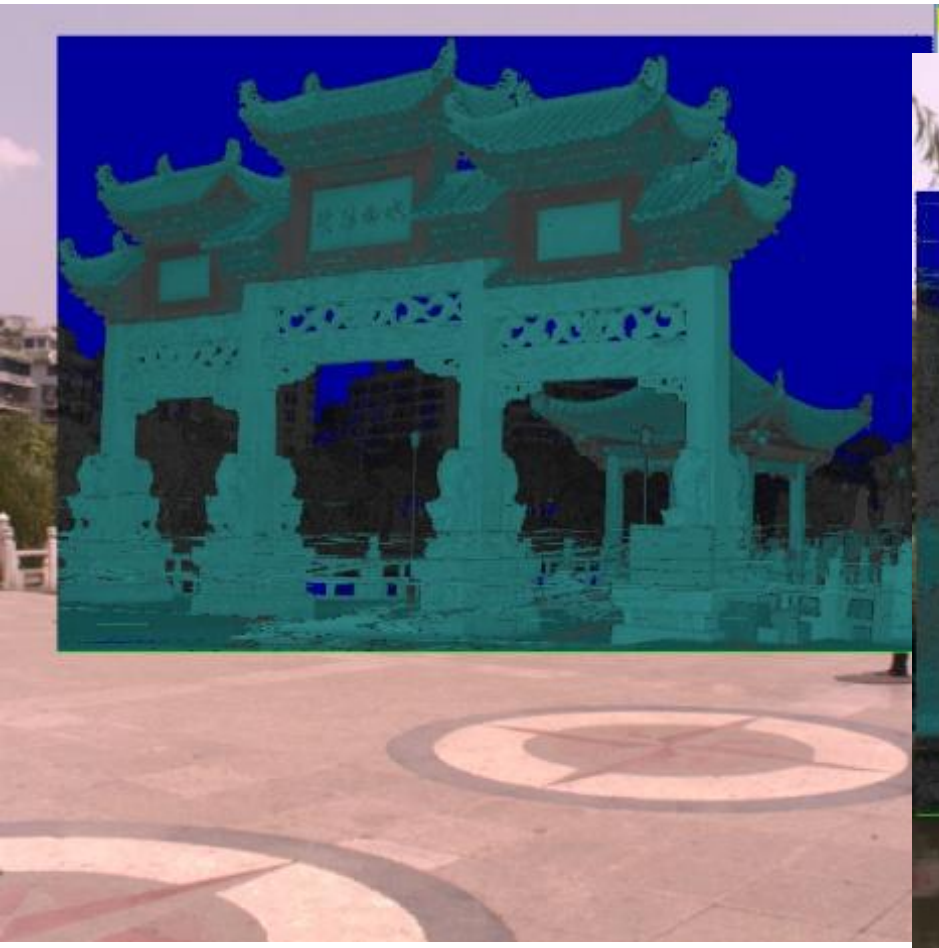

50.00%
45.00%
40.00%
35.00%
30.00%
25.00%
20.00%
15.00%
10.00%
5.00%
0.00%
28.10%
31.33%
34.46%
35.82%
40.64%
46.77%
08年
09年
10年
11年
12年
13年

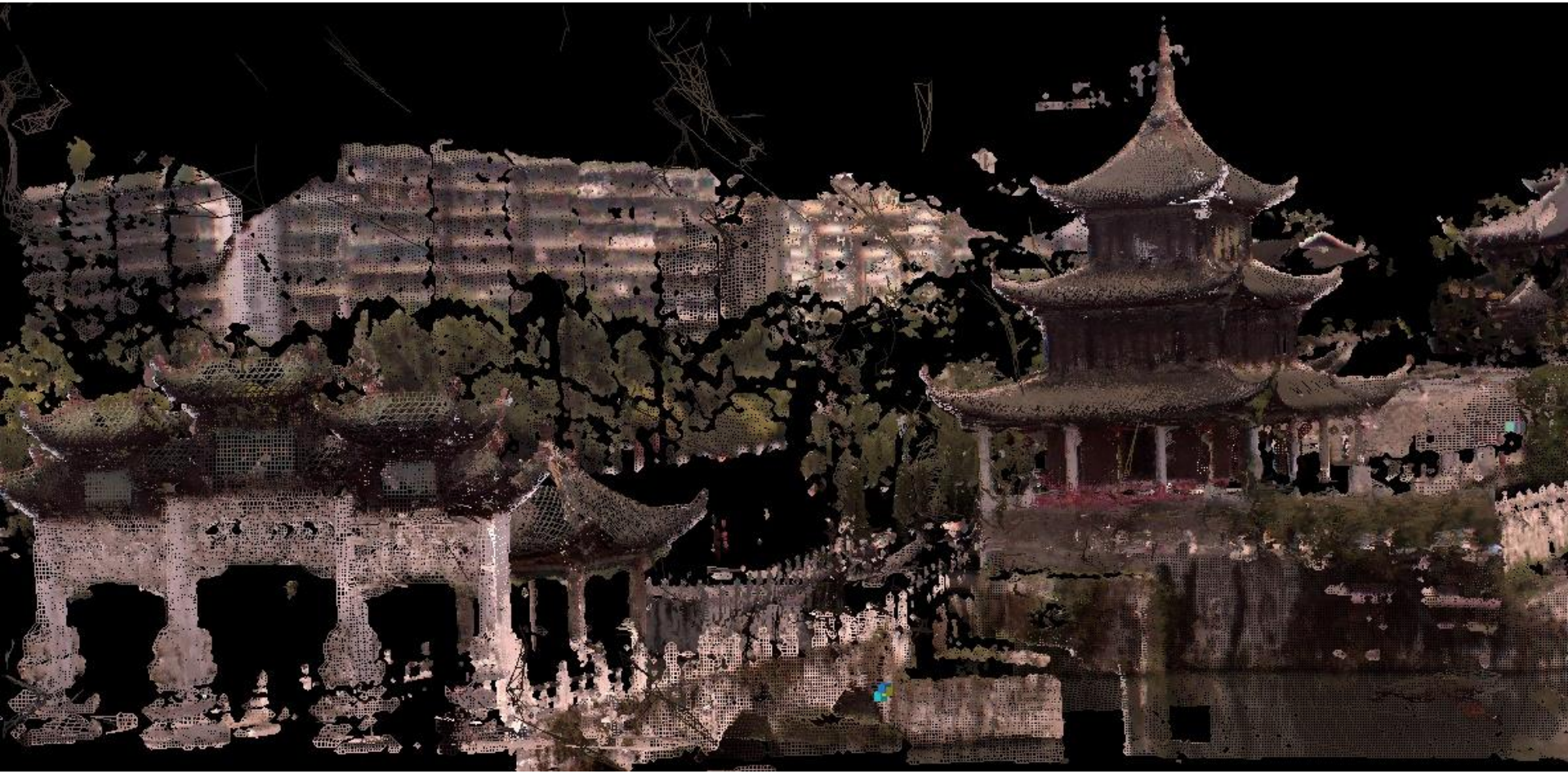
(4) Cultural heritage protection in the era of big data



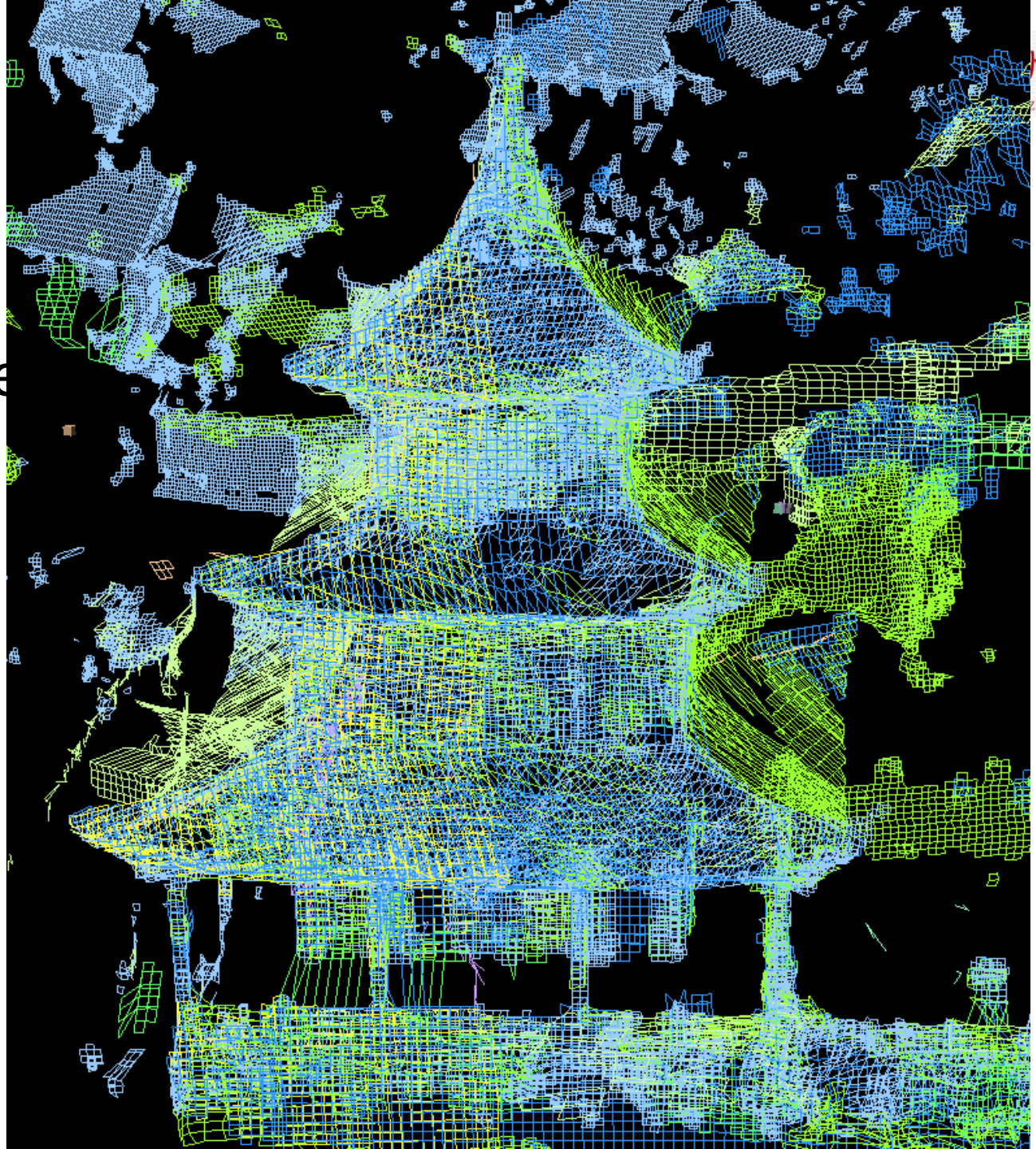
Scanner distribution of JiaXiu Lou： 17 points , 27 times of scan, 32 million resulting data point with size of 17G

- Scanning scene

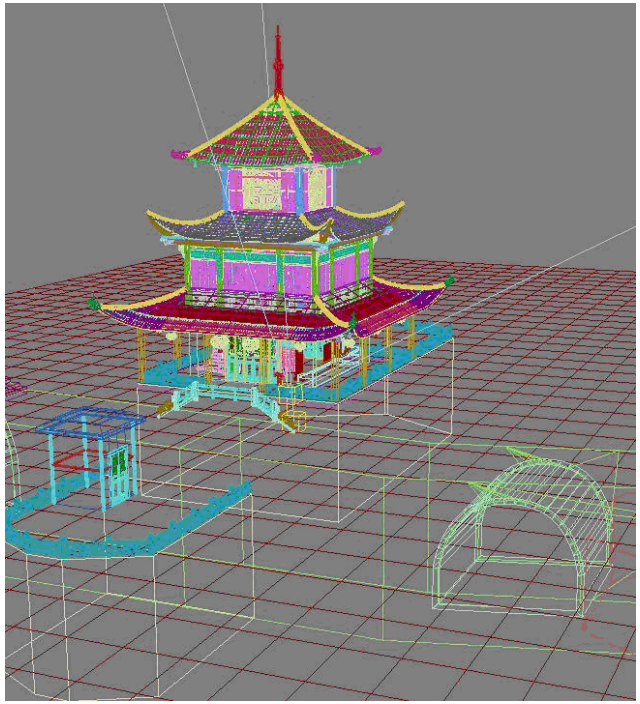




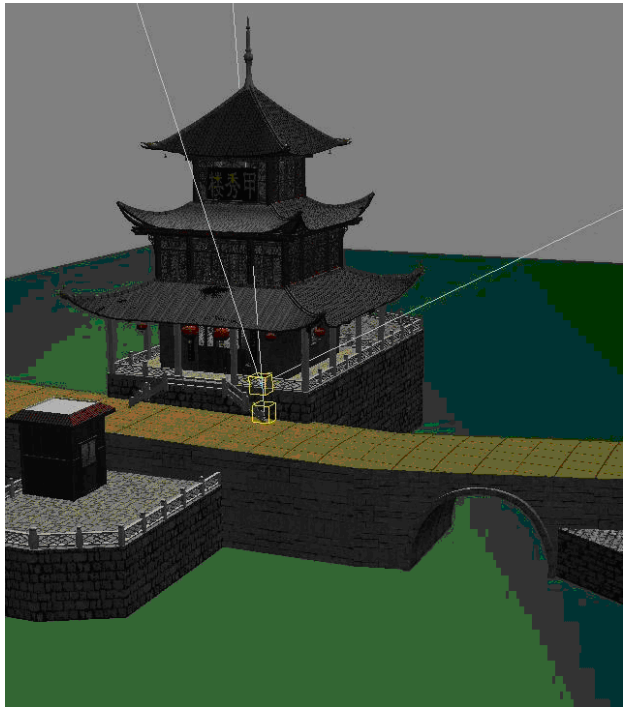
Plolygon model in the reverse modeling of JiaXiu Lou



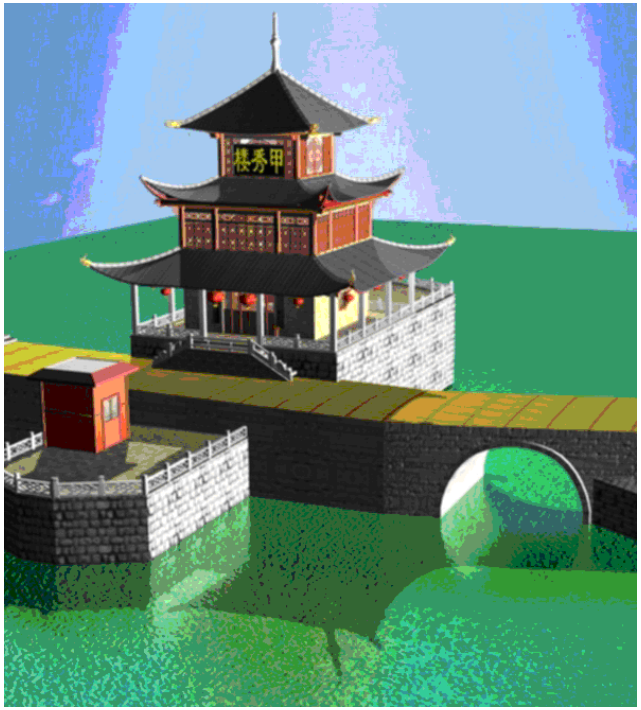
3D modeling of JiaXiu Lou



S3-3 Graphic of point, line and frame



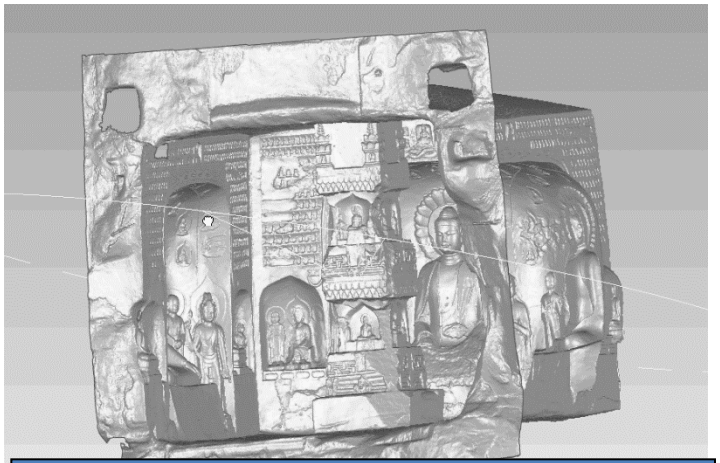
S3-3 Reconstructed graphic



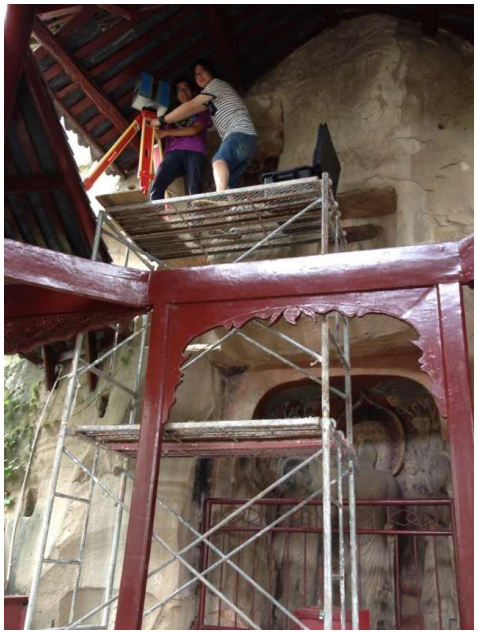
S3-3 3D Rendering graphic

National cultural relics protection units: Grotto image reconstruction at Sichuan Guangyuan Huang Zesi Temple

10,669,363 surfaces，3.1G



15,298,074 surfaces , 4G



1,178,129 surfaces 3.2G

(5) The analysis of biological and gene big data

Sequencing platform based on Miseq: 4T per year

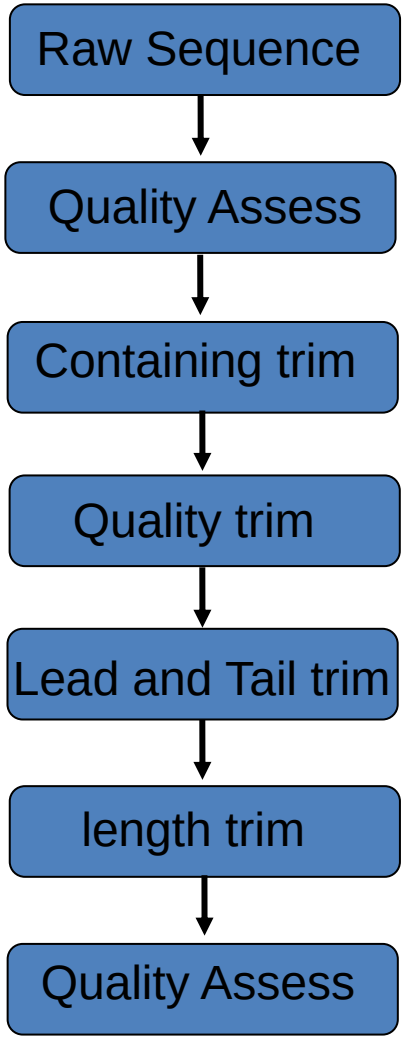
- 全基因组重测序
- De Novo测序
- RNA seq测序
- Chip seq
- 转录组测序
- 宏基因组测序

MiSeq squencing performence

Mode/lengt	Speed	Clusters	Bp (G*)	Data size(G)	Q30 %
1 × 36 bp	~4 hrs	12～15	0.54-0.65	1.4-1.6	90
2 × 25 bp	~5.5 hrs	24～30	0.75-0.85	1.9-2.1	90
2 × 75 bp	~24hrs	44～50	3.30-3.80	8.3-9.5	85
2 × 150 bp	~24 hrs	24～30	4.50-5.10	11.3-12.8	80
2 × 250 bp	~39 hrs	24～30	7.50-8.50	18.8-21.2	75
2 × 300 bp	~65 hrs	44～50	13.2-15.0	33.0-37.5	70

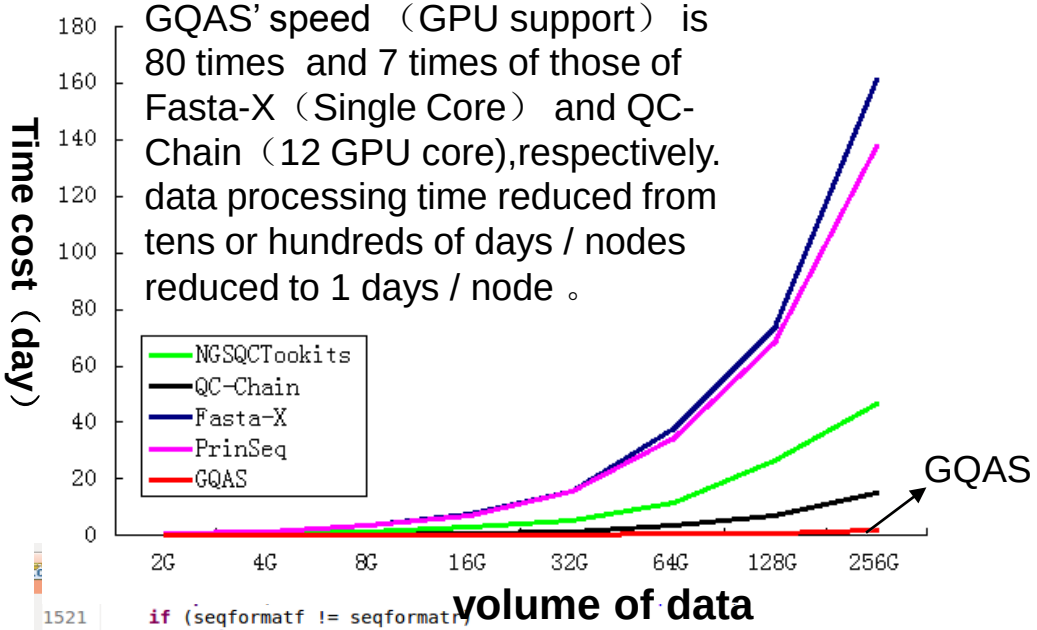
*：**Human genome size is 3G. 30x resequencing of human genome will last 15 day**

GQAS:the DNA sequence quality filter software based on the GPU



Quality control processes of DNA sequence

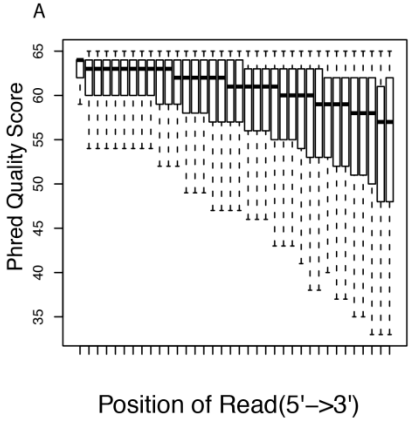
GQAS' speed （GPU support） is 80 times and 7 times of those of Fasta-X（Single Core） and QC-Chain（12 GPU core),respectively. data processing time reduced from tens or hundreds of days / nodes reduced to 1 days / node 。

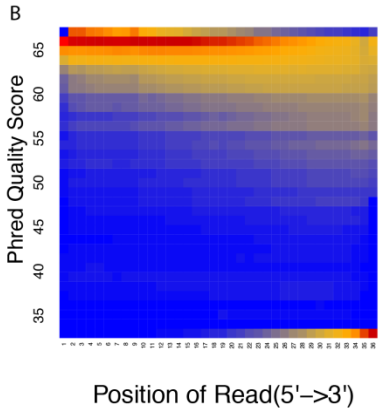


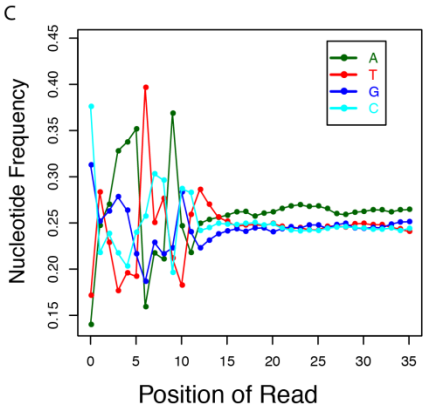

```
1521    if (seqformatf != seqformatr)
1522        printf("Format of the two inputfiles must be same!");
1523
1524    seqformat = seqformatf;
1525    int valmax = 0, valmin = 0;
1526
1527    printf("seqformatf: %d ,seqformatr: %d \n", seqformatf, seqformatr);
1528
1529    if (seqformat == 1) {
1530        subval = 33;
1531        valmin = 33, valmax = 73;
1532        printf("Input FASTQ file format: Sanger\n");
1533    }
1534    if (seqformat == 2) {
1535        subval = 64;
1536        valmin = 59, valmax = 105;
1537        printf("Input FASTQ file format: Solexa\n");
1538    }
1539    if (seqformat == 3) {
1540        subval = 64;
1541        valmin = 64, valmax = 105;
1542        printf("Input FASTQ file format: Illumina 1.3+\n");
1543    }
1544    if (seqformat == 4) {
1545        subval = 64;
1546        valmin = 66, valmax = 105;
1547        printf("Input FASTQ file format: Illumina 1.5+\n");
1548    }
```

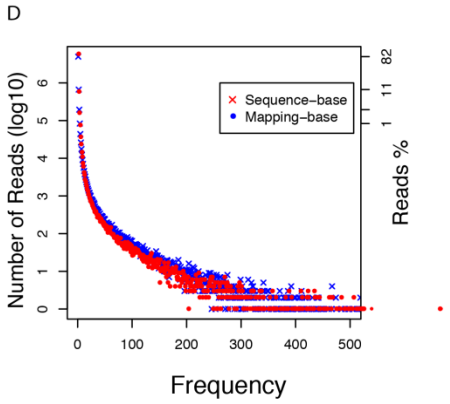
GQAS:the DNA sequence quality filter software based on the GPU

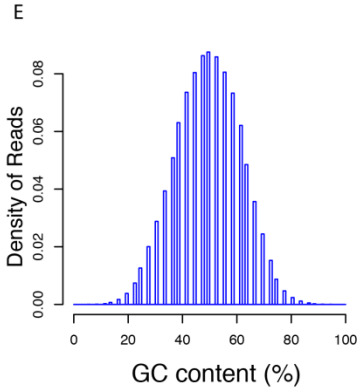
Guizhou Normal University

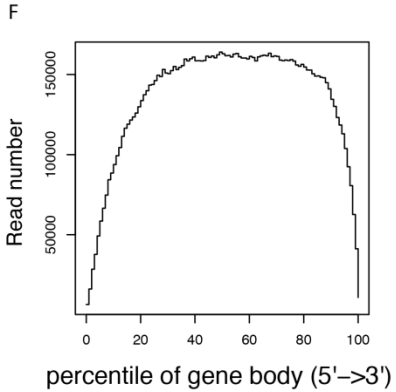


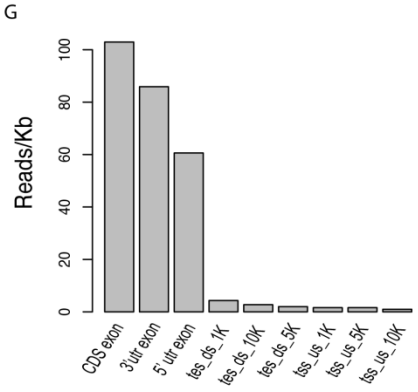


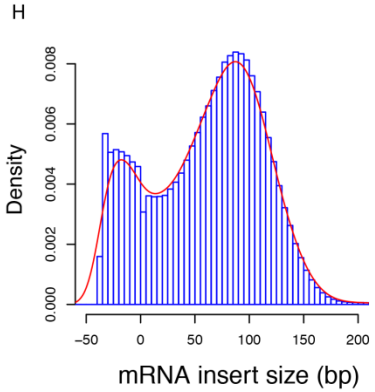


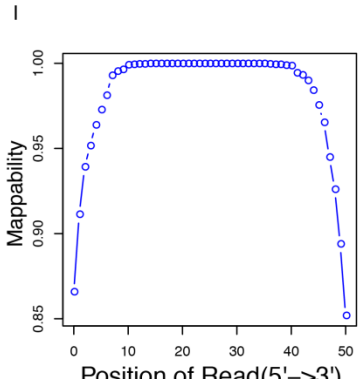


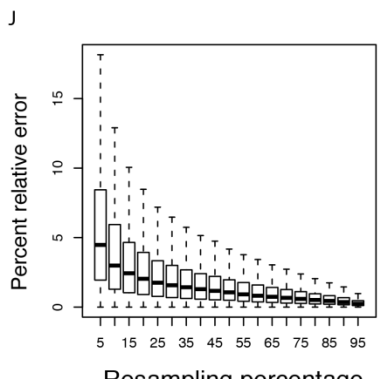


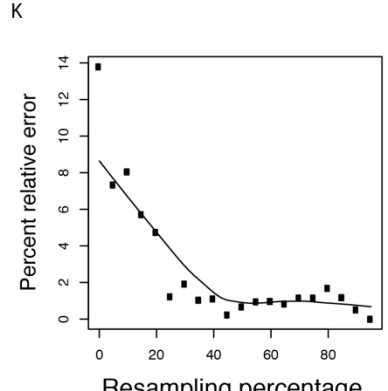








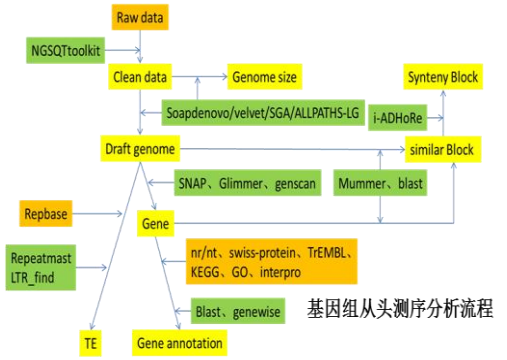




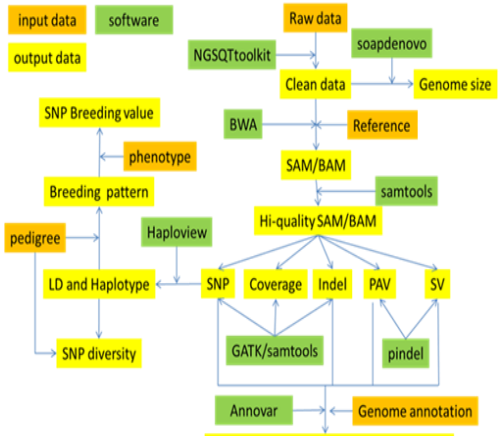
Genome De novo and resequencing system

貴州師範大學 Guizhou Normal University

Coordinated efforts from Guizhou Normal University, Huazhong Agricultural University, Australia NSW DPI, to provide sustainable development for drought-resistant and quality breeding of canola



基因组从头测序分析流程



基因组重测序分析流程和程序

Suggested project: The genome of *Brassica carinata* and its impact in modifying the genome of major Brassica oilseed crop, *B. napus*

De novo genome sequencing One line:

Yellow-BcDH64, a DH line which was used as one parent of YW DH mapping population with yellow flower, yellow seed coat and crossability to Brassica rapa. The original accession was named as CGN03964, collected from Ethiopia plateau. Transcriptome sequencing would be parallel carried out. The genome of B. carinata will be compared with the genome of *B. napus* (Darmor, cooperated with Boulos) and, perhaps, with the genome of *B. juncea* (cooperation with some Chinese scientists).

The Whole Genome Shotgun (WGS) strategy will employed to finished the genome sequence. Accord to the experiment from other complex genome project and the sequencer equip by HAU (Huazhong Agricultural University) and GZNU (Guizhou Normal University), seven type of library will construct, the sequence model and sequence data coverage was listed in the follow table, the total cost for de novo genome is estimated about $20,000, including all of the necessary reagent and consume.

The assembly strategy will combine the advantage of ALLPATH-LG (with many special design and optimization for complex genome assembly, http://www.broadinstitute.org/software/allpaths-lg/blog/?page_id=12) and SOAPDenovo (The best memory efficient genome assembly software for complex genome: http://soap.genomics.org.cn/soapdenovo.html) to archiev an optimal asseblies.

Library will be constructed for de novo assembly of Yellow-BcDH64:

Insert size	Library#	Bases	model (bp)	Platform*	Coverage	Cost ($)
500(400?)	4	30G	2 * 300	MiSeq	30X	2600 (2 run)
800	3	24G	2 * 300	MiSeq	24X	2000 (1.5 run)
2k	4	30G	2 * 90	HiSeq 2000	30X	2500 (1 lane)
5k	4	30G	2 * 90	HiSeq 2000	30X	2700 (1 lane)
10k	4	30G	2 * 90	HiSeq 2000	30X	2700 (1 lane)
20k	4	30G	2 * 90	HiSeq 2000	30X	2700 (1 lane)
40k	4+2	30G + 15G	2 * 90	HiSeq 2000	30X + 15X	2700 + 1350 (1.5 lane)
Total	27 + 2	219G	-	-	219X	20,000

* GZNU have equip one Miseq, and HAU have equip one Hiseq2000.

Pan-genome sequencing Four lines of B. carinata and four lines of new type *B. napus*:

1. White-BcDH76， the another parent of YW DH mapping population with white flower, brown seed coat and low crossability to Brassica rapa. The original accession was named as CGN03976, collected from an island in the Gulf of Guinea near the equator.
2. Two accessions of *B. carinata* which were used as parents of new type *B. napus*.
3. One accessions of B. carinata which were identified by Harsh with unique trait.
4. Four lines of new type *B. napus* with various introgression of *B. carinata* developed in HAU. Two lines would be those used for constructing segmental introgression populations in HAU, and another two would be those with very high combinability in hybrid breeding.

Library will be constructed for Pan genome assembly (cost for one line):

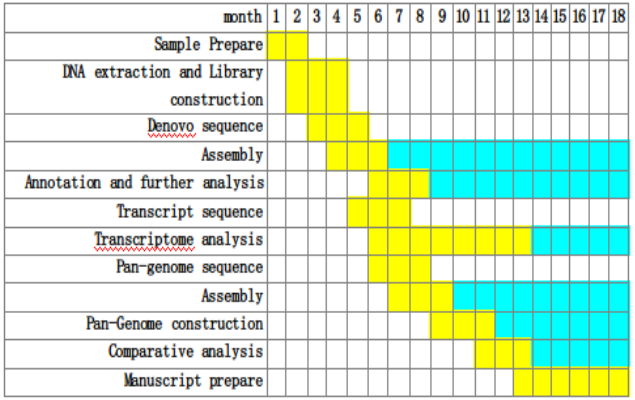
Insert size	Library#	Bases	model (bp)	platform	Coverage	cost
500	4	30G	2 * 300	MiSeq	30X	$2600 (2 run)
800	3	24G	2 * 300	MiSeq	24X	$2000 (1.5 run)
2k	4	30G	2 * 90	HiSeq2000	30X	$2500 (1 lane)
5k	4	30G	2 * 90	HiSeq2000	30X	$2700 (1 lane)
Total	15	114G	-	-	95~112X	$10,000

Total Pan genome sequence cost was estimated to about $80,000 (8 lines).

The strategy for Pan genome will adjust according to the assembly result of the Yellow-BcDH64.

Time line

Sample prepare: 2 month (HAU, NSW DPI); **DNA extraction and library construction**: 3 month? (HAU, GZNU); **Denovo Sequence**: 3 month (HAU, GZNU); **Assembly**: 3 month (GZNU, HAU, NSW DPI); **Annotation and further analysis**: 3 month (GZNU, HAU, NSW DPI); **Transcriptome sequence**: 3 month (HAU, GZNU); **Transcriptome analysis**: 8month (GZNU, HAU, NSW DPI); **Pan genome sequence**: 3 month (HAU,GZNU); **Assembly**: 3 month (GZNU, HAU, NSW DPI); **Pan genome construction**: 3month (GZNU, HAU, NSW DPI); **Comparative genome analysis and other**: 3 month (GZNU, HAU, NSW DPI); **Manuscript**: 6 month (HAU,GZNU, NSW DPI)



Foundation

Prof. Xiaoyao Xie, Prof. Zhijie Liu and Dr. Ruiyuan Li Guizhou Normal University will provide 100,000RMB, and try to gain additional 100,000RMB from the local government, for the project and as a major player for sequencing and analysis.

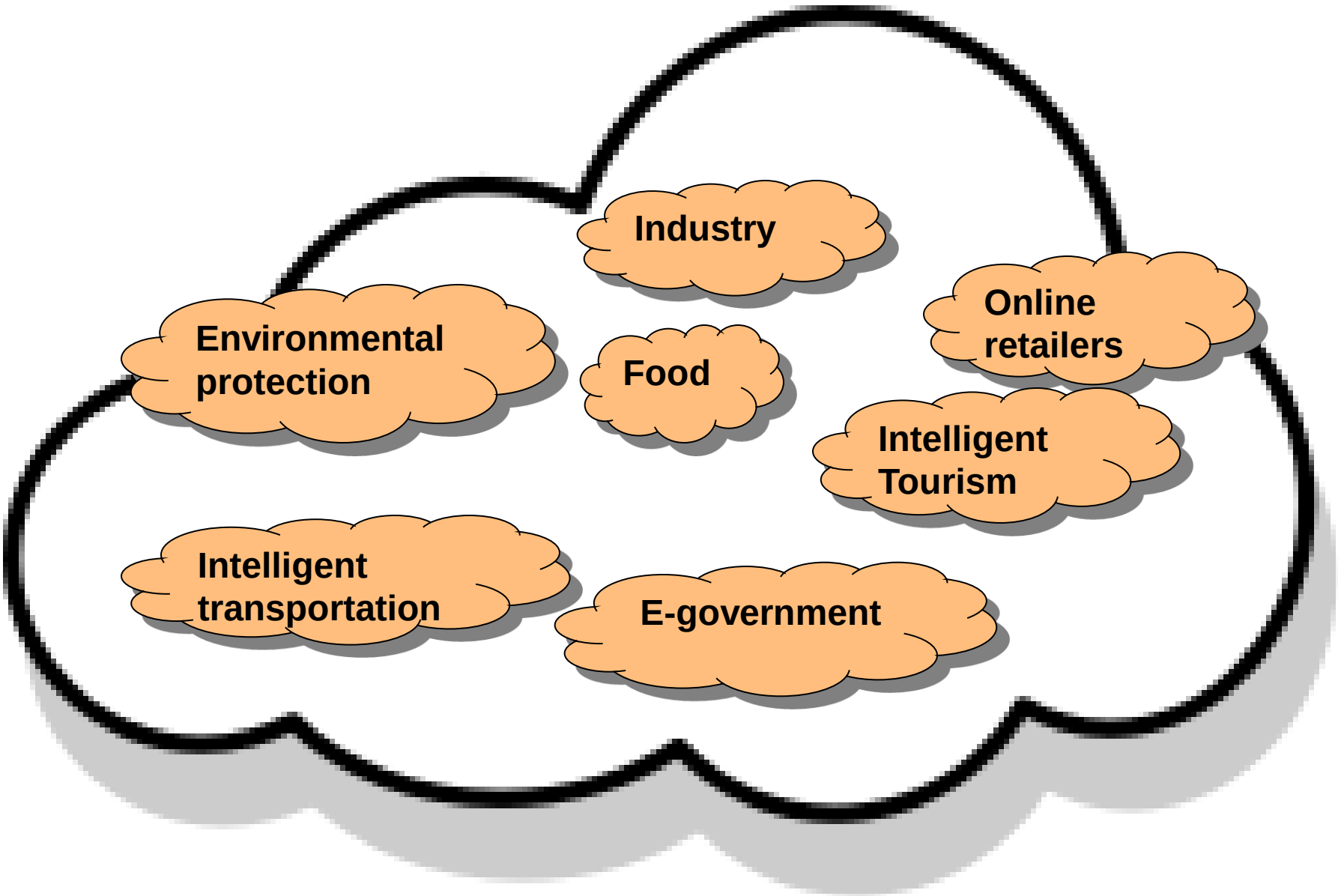
Dr. Harsh Raman from NSW DPI，Australia, would like to provide $50,000 to support the de novo sequencing/pan-genome sequencing of *B carinata*.

Dr. Jun Zou/Prof. Jinling Meng from HAU will provide about 60,000RMB, and seek money (about 150,000?) to support the pan-genome sequencing of lines of new type *B. napus*.

High-performance computing platform in GPU based gene analysis

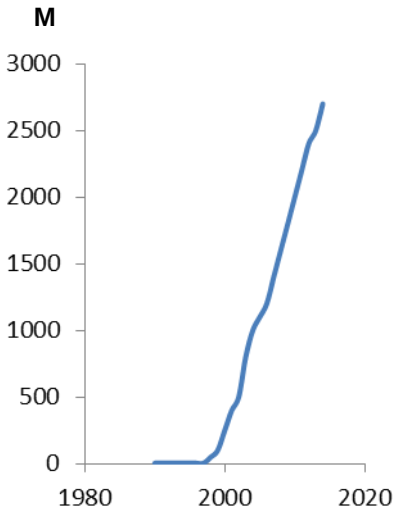
Guizhou Normal University

Clouds of Guizhou

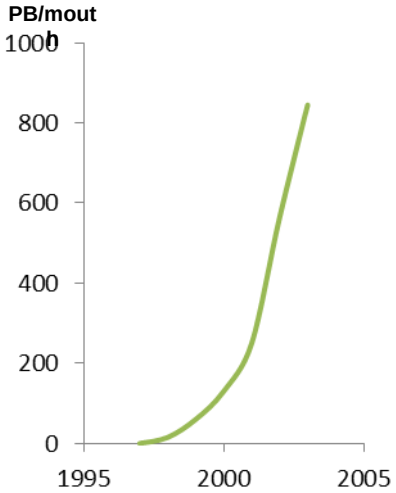


4 Perspective of big data

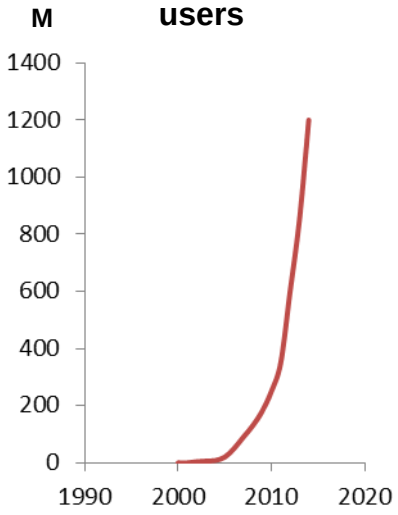
Development trend of Big Data



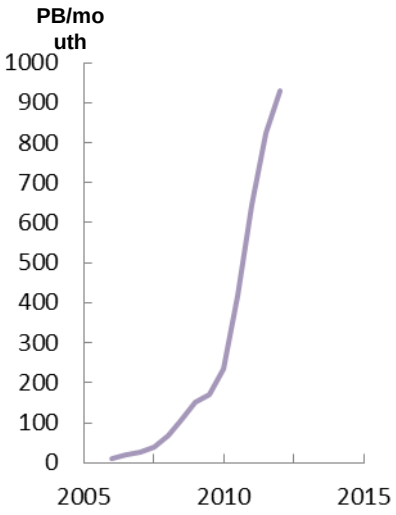
Number of Internet users



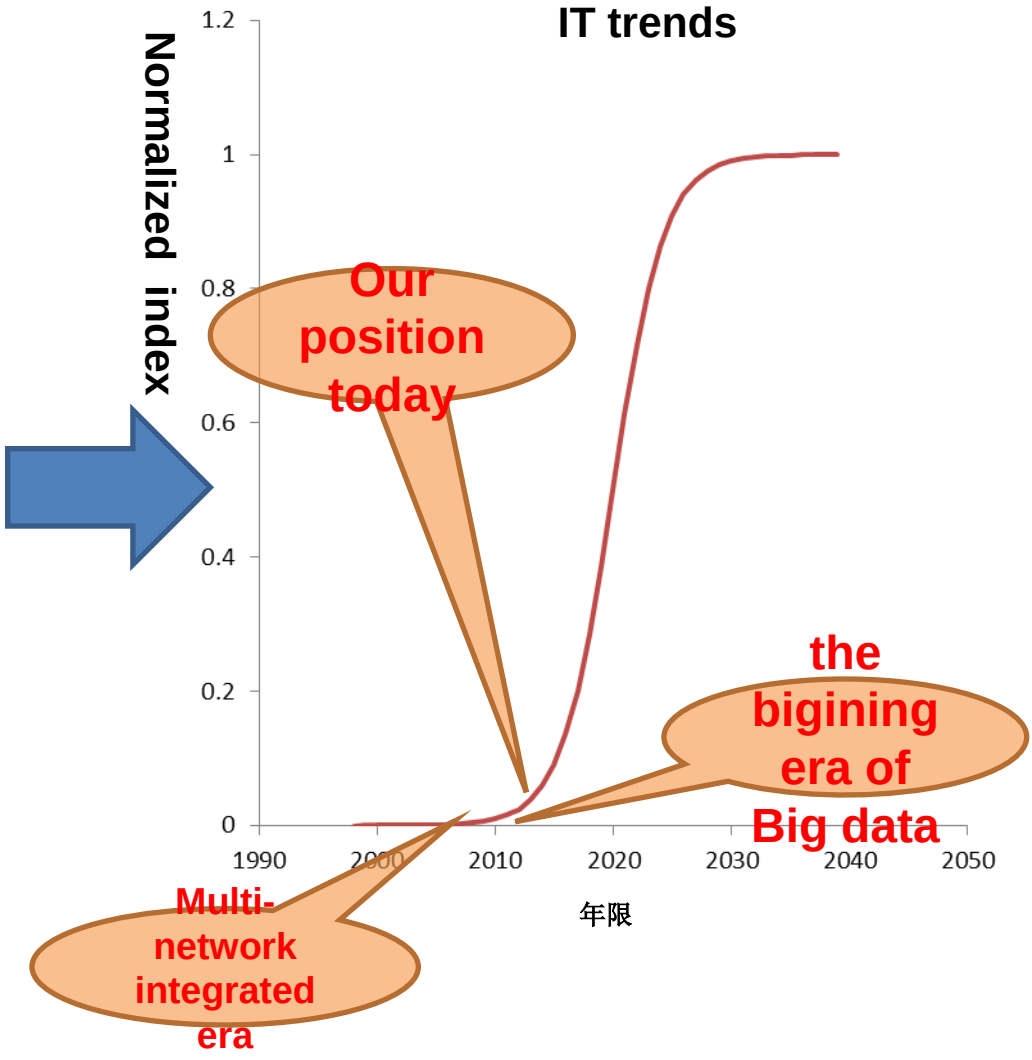
Fixed network traffic



Smart phone shipments



Mobile terminal traffic



Integration of production and marketing：

Prosumer

- Pioneer： Alvin Toffler 2006
- The integration of producer and consumer considers that the future of producers and consumers must be unified , or that the different identity labels of the producer or consumer becomes fuzzy.

- Active consumption patterns
  - Consumers design their own products
  - Family generated electricity / Mutual Production
  - Public comment
- Wiki model
  - Non professional expert /Volunteer Community
- Maker mode
  - Creative Workshop / Creative Workshop
- Freelancers Union model
  - Free Lancers Union

Flow： The integration of the 3 new networks

- The first One： physical network, which is passed through by traffic flow, parcels flow and cash flow
- The second One： Online electronic commerce
- The Third One： Social networks

- Flow is the superposition of the flows through the 3 networks. It is the streaming of data and the streaming of ideas.It is the data currency that integrate the three networks .

Digital industry shocks all business

- Alibaba's rise
  - 11.11 last year, the turnover of one day exceeded more than 30 billion and that of WAL-MART
  - Broke 100 million at the first minute, 300 million at the second minute and one billion at the sixth minute

- Mobile has a powerful communication resources, but ignored the power of fusion(data from Alipay)
  - Wechart rise as a rival, Does Mobile still have time?
- Alibaba's Alipay/Yu Ebao is coming
  - Internet banking has come, how long can closed bank can stick to?

The core driving force of big data is the human's desire to measure, record and analyze the world. Revolution of Information technology can be seen everywhere, but past IT revolution focuses on the T (Technology), rather than the I (information).Now, it is the time to place our eyes on the "I", to center on the information itself.

——《Times of Big Data》

Reference

1. Tu Zipei.The top data: big data revolution, history, reality and future

2. Wu Jun. Beauty of Mathematics

3. Li Hang. Statistical learning method

4. Wu Zhuhua. Analysis of core technology of cloud computing

5. Zhao Guodong. Historical opportunity in the era of big data: Industrial Revolution and data science

6 Kelly Kevin. Out of control: the ultimate fate and outcome of the whole mankind

7 Zhao Gang. Big data: technical and practical guide

8 T.M. Mitchell.computer science books: Machine Learning

9. lan H.Witten. Data mining: practical machine learning tools and techniques

10 Tu Zipei. Big data: the data revolution is coming

11 Hastie Trevor. Statistical learning foundation: data mining, inference and prediction

12 Duan Yongzhao. Big data flow

Thank you

xyx@gznu.edu.cn