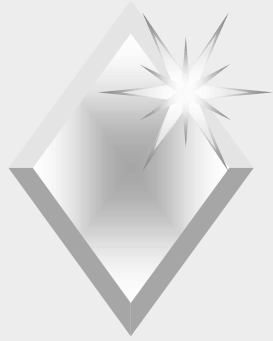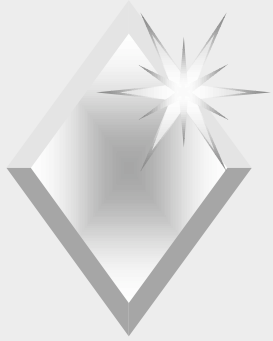# Controlling Information Technologies In Biological Sciences

Presented by

Antoine Danchin

HKU-Pasteur Research Centre.

December 15, 2000
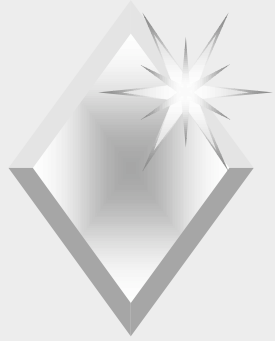
Genome initiatives

◆ **Techno-political**: the Human Genome Project, the Yeast European Initiative

◆ **Scientific**: bacterial genomes, and more recently, the Drosophila genome
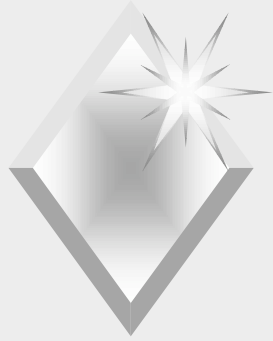
The Human Genome Program

◆ The Atomic Bomb Casualty Commission

◆ The Alta Summit (dec 1984)

◆ Sinsheimer / Dulbecco / De Lisi

http://www.hkupasteur.hku.hk/hkuip/Rattling.html

Data Banks (late seventies)

◆ European Molecular Biology Laboratory (Heidelberg)

◆ GenBank at the Department of Energy (Los Alamos)

◆ SwissProt (Geneva)

Genome information: the present state of affairs

◆ **One** DNA data library with **three** entry points

- ◆ GenBank NCBI (Bethesda, USA)
- ◆ EMBL/EBI EBI (Hinxton, Europe)
- ◆ DDBJ NIG (Mishima, Japan)

◆ Several protein data libraries but **only one** reference (with mirrors in Australia, Beijing and Taiwan)

- ◆ SwissProt (Geneva, Swizerland)
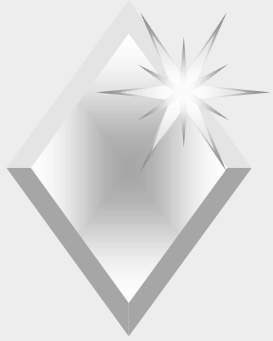
PubMed Central

The NLM initiative, PubMed Central, aims at placing on-line all the literature pertaining to Medical Sciences. While this is very convenient, a single entry point is prone to network failures, and, above all, to political or commercial uncertainty
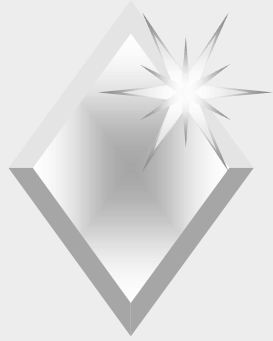
The Blair-Clinton Declaration
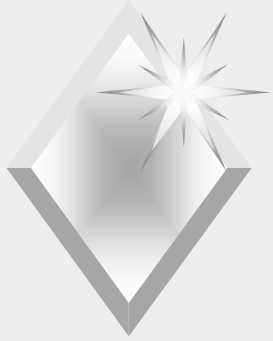
- 1945-1985 The Cold War (alliance with Japan)
- 1985-present The USA against Japan (and later, other Asian countries)

=> Federal incentive to create companies (in particular in Biotechnology: 85% public support)
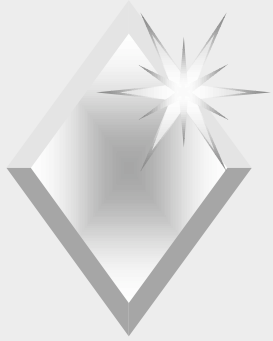
- Volte-Face against Celera
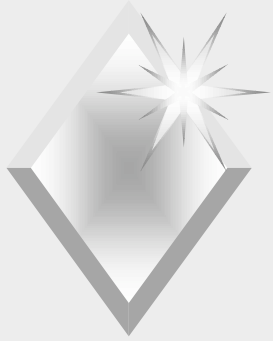
Conclusion

◆ A large body of knowledge is confidential and in the hands of private companies, mainly in the USA

◆ Urgent need for other complementary initiatives

A world-wide initiative

◆ The best solution is the world-wide sharing of information

◆ This makes accessibility to knowledge robust to accidents (voluntary and involutary)

◆ America, Europe and Asia should cooperate

◆ China is an obvious entry point: need for explicit political support (government funds)

HKU-Pasteur Research Centre

Why China? Background

- ◆ **Population: 1.30 billion**
- ◆ **Urban: 29% (= 370 millions)**

- ◆ **Education: Compulsory, 7 - 17**
  - ♥ **Literacy (1996): 82%**

Age Distribution (%)

- 60 + 10%
- < 19 32%
- 40-59 22%
- 20-29 18%
- 30-39 18%
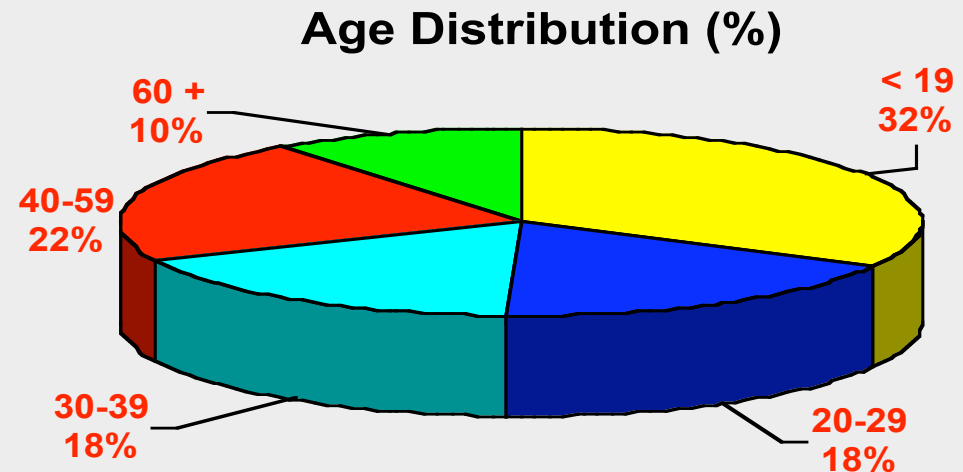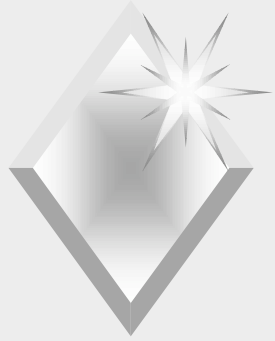
Why China? Cultural Background

The unified written Chinese language has been the major driving force in shaping the Chinese national identity and cultural characters. There exists a large corpus of scientific literature in Chinese (usually not exploited with modern means)
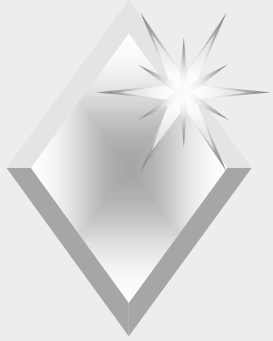
Why China? Cultural Background

◆ **In China, the ability to read and write is highly respected and confers prestige to people**

◆ **Teaching children to recite and learn to appreciate ancient poems and articles constitutes the most important part of pre-school education**

◆ **Ancient Chinese texts written 2000 years ago are far more closer to modern Chinese people than the ancient Greek texts to modern Western people**

=> Support to activities pertaining to literacy should be well perceived
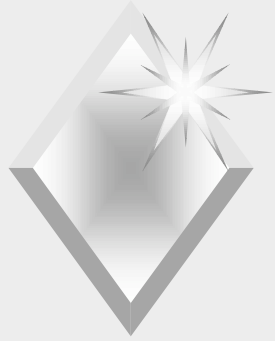
Why China? Economic Background
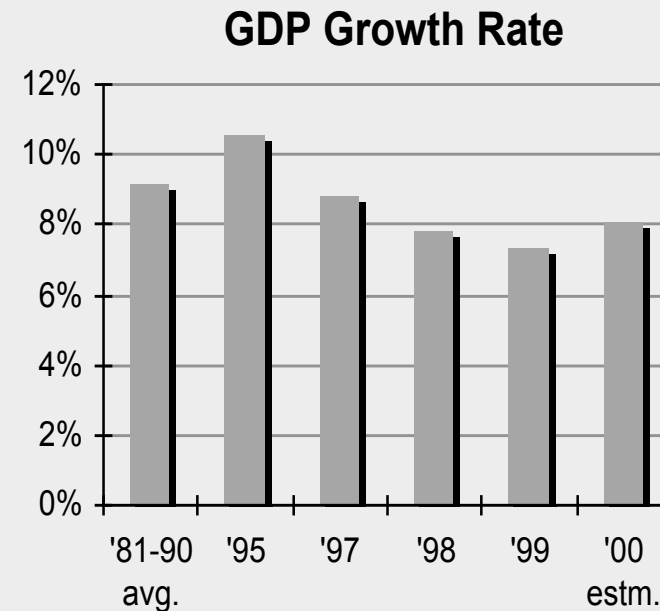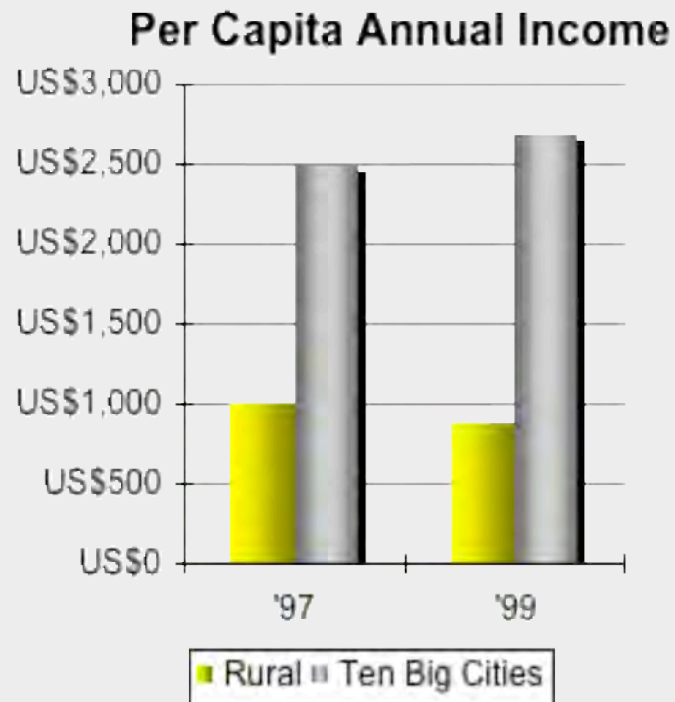
Based on its unified cultural heritage, today's China could constitute the largest single media market in the world. This makes access to education and knowledge through the Internet a plausible opportunity
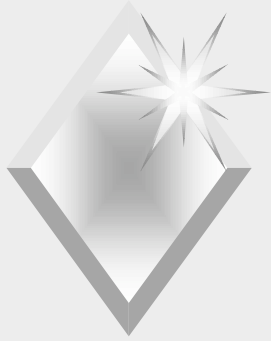
Why China? Economic Indicators

◆ **Even amid the financial crisis sweeping the whole Asia, Chinese economy growth kept at very high rates.**

Per Capita Annual Income

| | US$3,000 |
| US$2,500 |
| US$2,000 |
| US$1,500 |
| US$1,000 |
| US$500 |
| US$0 |

'97   '99

■ Rural ■ Ten Big Cities

GDP Growth Rate

12%
10%
8%
6%
4%
2%
0%

'81-90 avg.   '95   '97   '98   '99   '00 estm.

◆ Source: IMF

HKU-Pasteur Research Centre

Why China? Economic Indicators

Household Ownerships

Legend: Rural (yellow), Ten Big Cities (grey)

Source: Gallup 1999
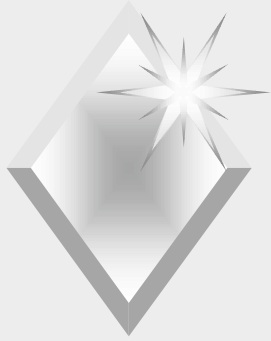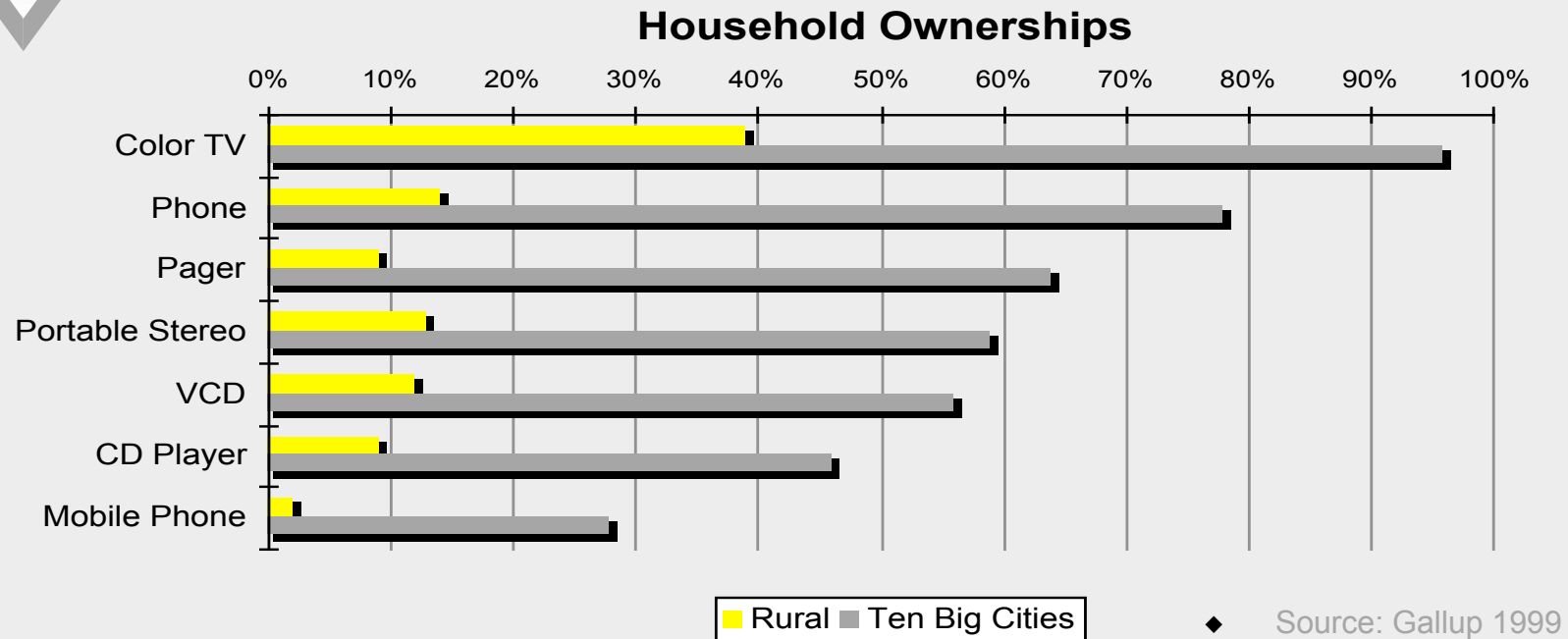
- ◆ **So, what is next for people to own?**
- ◆ **Computer and Internet access!**

HKU-Pasteur Research Centre
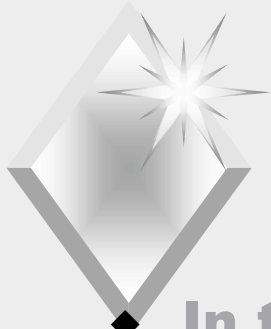
Why Internet in China?

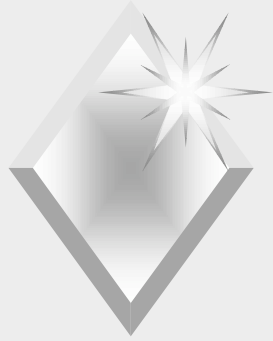◆ **In the past years, number of Internet users has been doubled every half year. Internet is forming the fastest growing business in China.**

◆ **China will constitute the largest Internet market of the whole Asia-Pacific region in a few years**

Number of Internet users in China

(Millions)

Dec '96	Jun '97	Dec '97	Jun '98	Dec '98	Jun '99	Dec '99	Jun '00	Dec '00	Dec '02
0.1	0.2	0.5	0.9	2.0	4.6	8.9	16.9	28.0	60.0

Forecast (Dec '00, Dec '02)
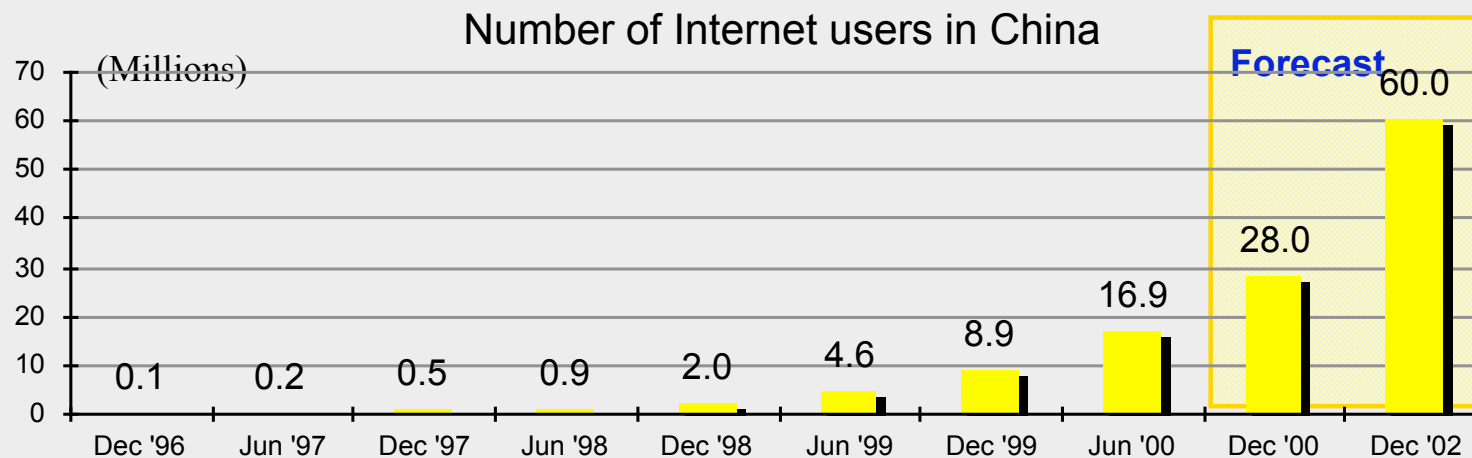
◆ Source: CNNIC, August 2000

HKU-Pasteur Research Centre

The virtuous cycle of basic research and information

….. to create enough knowledge- and education-related information resources ……

…… to create very large communities for extracting information from research laboratories and (on-line) teaching and learning ……

**….. to provide permanent education and offer various training programs, as well as provide resources for research and help discovery (knowledge bases) ….**

HKU-Pasteur Research Centre

Data Banks and Data Bases

- A **Data Bank** is similar to a library; it is a collection of records, usually with appropriate index files

- A **Data Base** allows one to make multicriteria queries on the data. In a Data Base, records are linked to each other following a **Data Schema,** which represents the way investigators see Data in the domain of interest
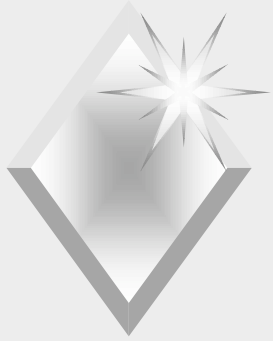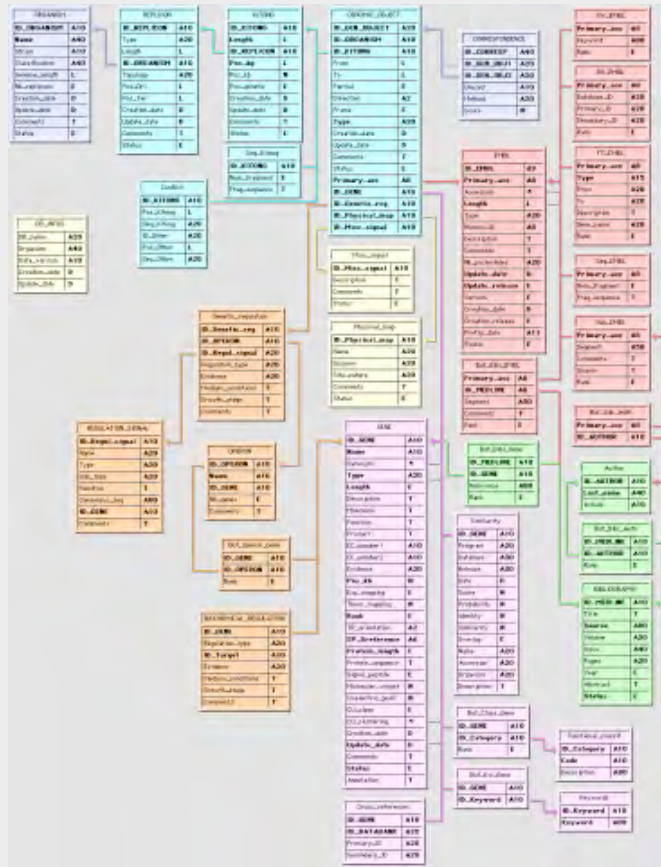
What is a Data base?

- ◆ Data (sequences, annotations, literature…)
- ◆ Data Schema (Data structure)
- ◆ Data Base Management System (commercial or public) and Query Language
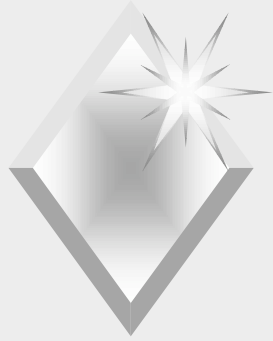- ◆ User Interface (stand alone or WWW)
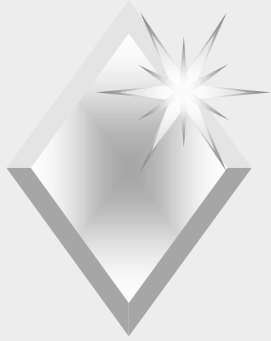
A Data Base Schema
(invisible to the user!)

◆ The **SubtiList** DataBase collects the reference information on the Gram-positive model *Bacillus subtilis*

**http://genolist.pasteur.fr/
SubtiList/**

Why knowledge control is important?

Compared with other products, knowledge might be the most sustainable and competitive product in the e-Business era
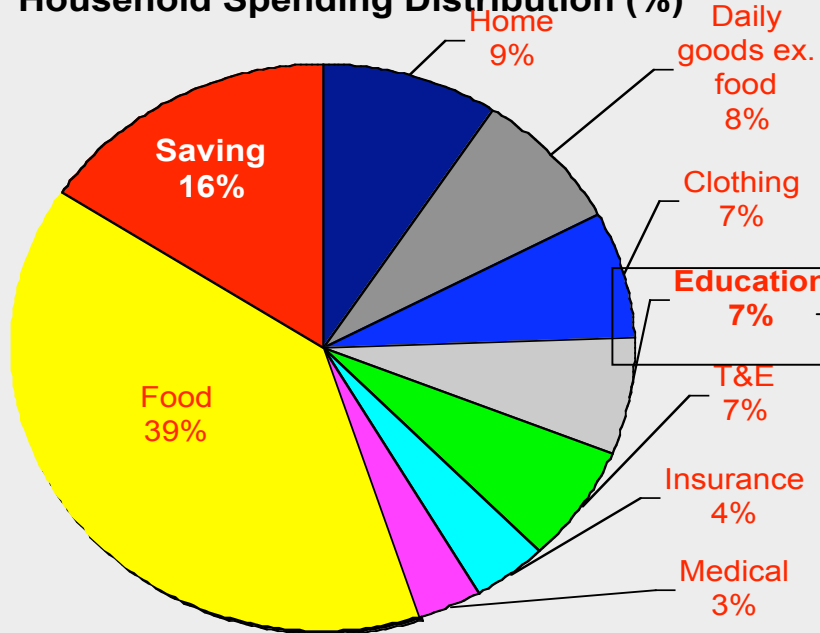
- ◆ Knowledge products are more relevant for people to improve their life quality
- ◆ Knowledge products do not require a complicated system of distribution
- ◆ An on-line / off-line mixed approach could be easily applied to knowledge products. Coupling with off-line materials is easy: basic and clinical research
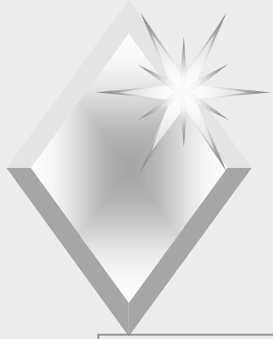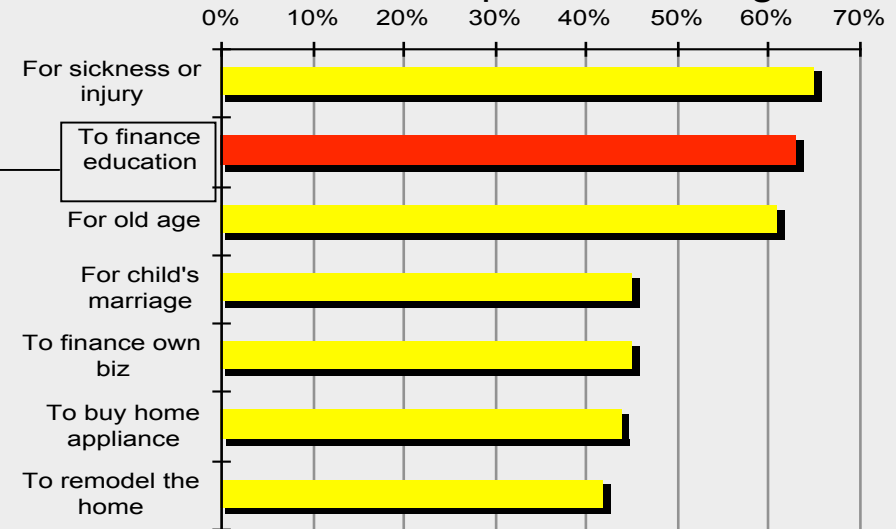
Why expect support?

◆ **To finance education is among the top priorities for Chinese people**

Household Spending Distribution (%)

- Home 9%
- Daily goods ex. food 8%
- Clothing 7%
- **Education 7%**
- T&E 7%
- Insurance 4%
- Medical 3%
- Food 39%
- **Saving 16%**

The Purpose for Saving

0% 10% 20% 30% 40% 50% 60% 70%

- For sickness or injury
- To finance education
- For old age
- For child's marriage
- To finance own biz
- To buy home appliance
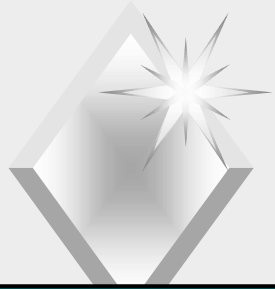- To remodel the home

◆ Source: Gallup 1999
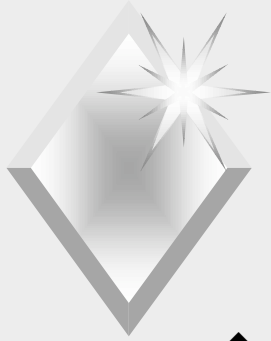
First steps

- Create a very large data library of various academic resources based on multiple partners (more than 1,000 Universities in China) and mirror daily major Data Banks
- Find a topic for intelligent information extraction from literature
- Daily campus news of both domestic and international education institutions
- On-line editions of both Chinese and European academic journals and other academic resources
- Who-is-Who directories of outstanding Chinese and European professors and researchers
- On-line directories of library resources
- E-zines of various disciplines and interests
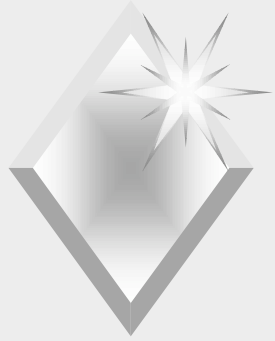
On-line education Offerings

◆ Selected programs in emerging topics in biology

◆ On-line seminars in bio-sciences, hosted by the world leading scientists

◆ Selected programs on ethical and general issues pertaining to biology

On-line knowledge access Summary

Literature E-publications

Education programmes

WWW interfaces

Data Base Management Systems

Specialized Analysis Platforms

Data Mining

Data Libraries

Knowledge Bases