About me

学习科研经历 :

-2004　　系统工程，大连理工大学
2004-2007 知识科学，北陆先端科学技术大学院大学
2007-　　IME， ASA　国际应用系统分析研究所

项目经验 :

- ✓Energy efficiency and Risk Management in Public Buildings (EnRiMa)
- ✓New Energy Externalities Developments for Sustainability (NEEDS)
- ✓Web-based Emission Trading System
- ✓MCA-based Global Energy Assessment
- ✓Intuitive Decision Analysis support System (Prototype)
- ✓COE: Knowledge Management system, E-science Environments, Domain Knowledge Ontology Construction

大数据分析及其在 Web3.0 中的应用

Hongtao Ren

renh@iiasa.ac.at
International Institute for Applied Systems Analysis
A-2361 Laxenburg, Austria

ECUST, 8-12 Oct, 2012

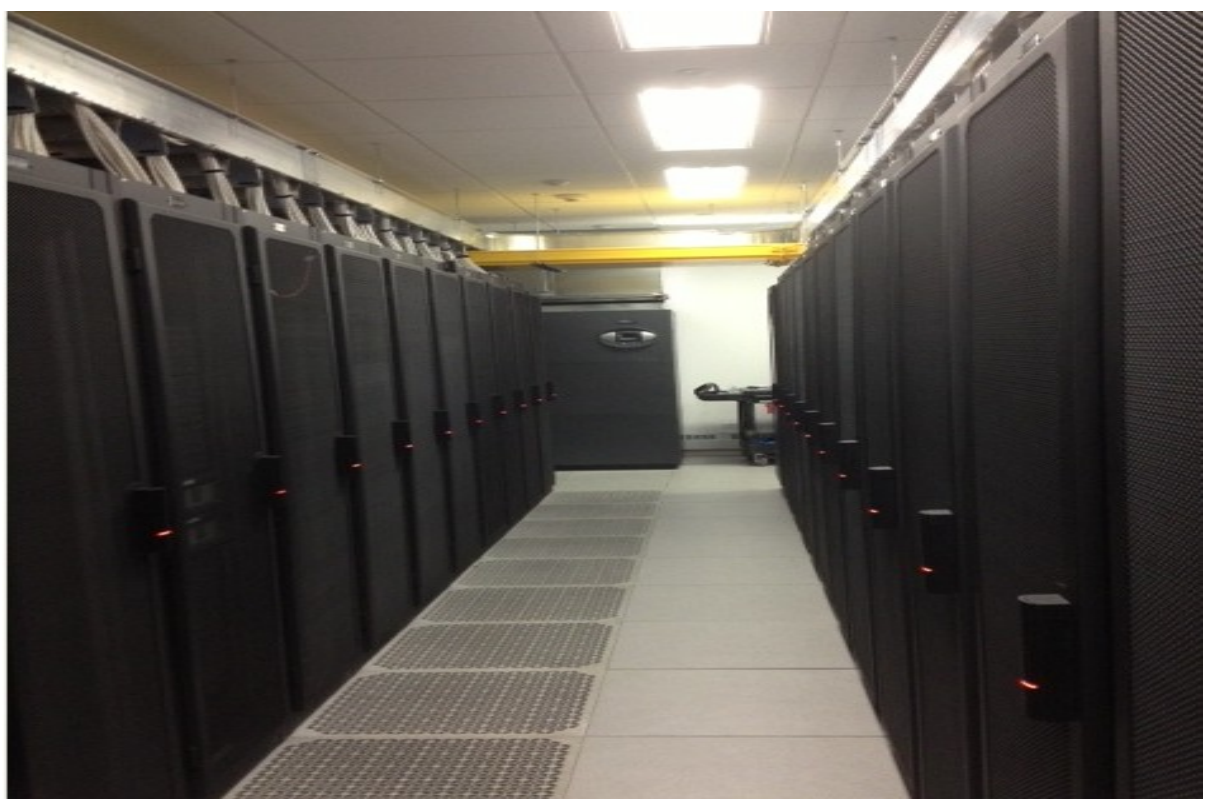
大数据定义

大数据定义：

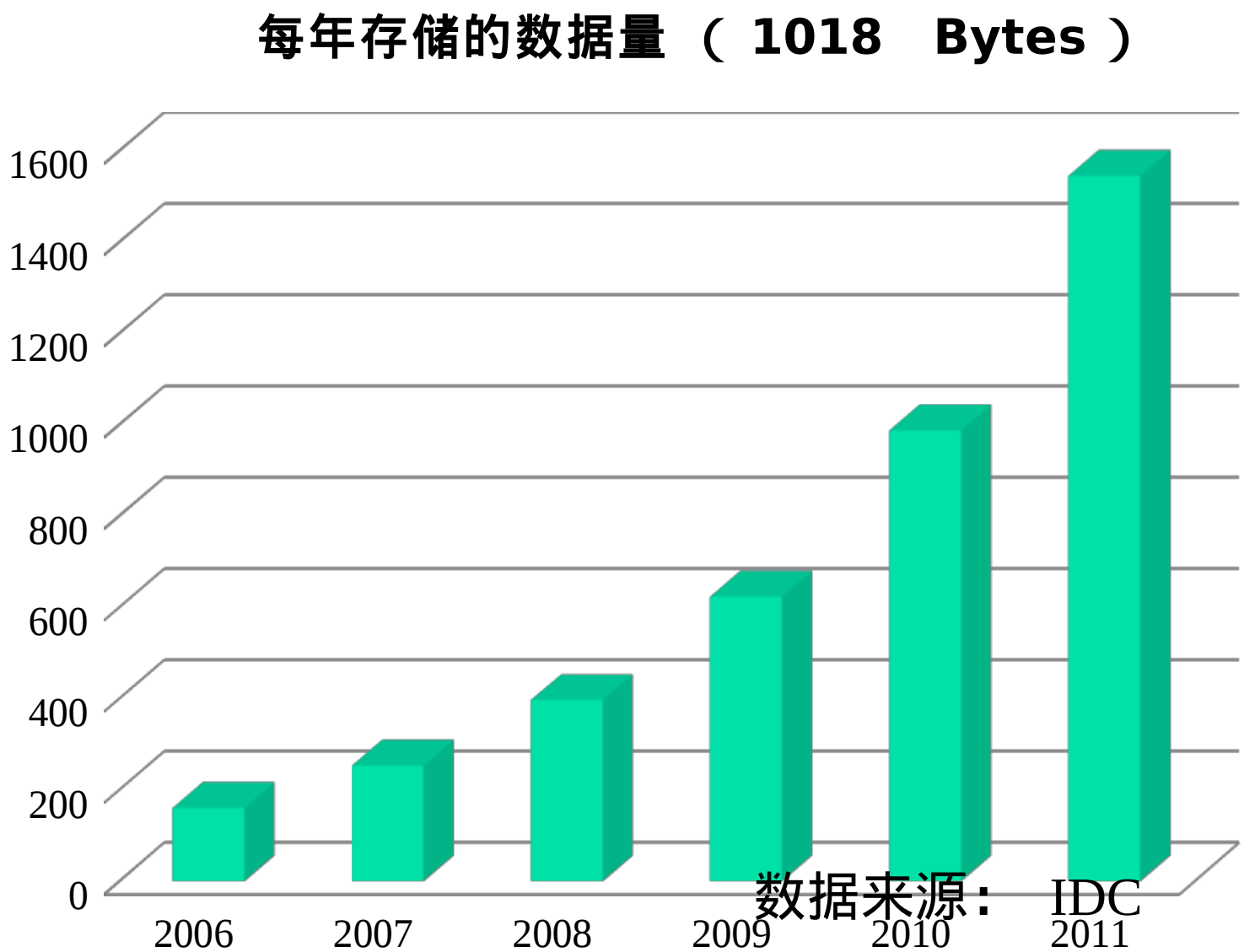
大数据是指需要**即时处理的、数据集容量非常庞大的、结构非常复杂的数据。**
(volume, velocity, and variability)

数据处理包括：

- 数据捕捉
- 数据存储
- 数据搜索
- 数据分享
- 数据分析
- 数据可视化



数据量


每年存储的数据量（ 1018 Bytes ）
1600
1400
1200
1000
800
600
400
200
0
2006
2007
2008
2009
2010
2011
数据来源： IDC

大数据应用领域

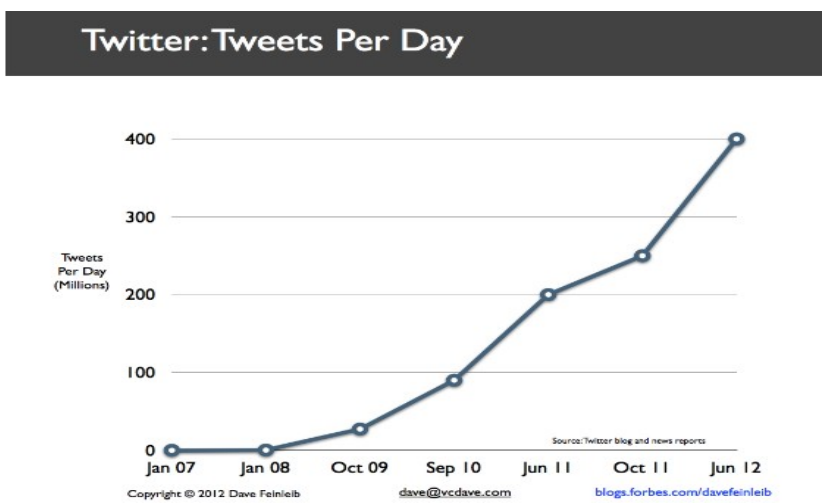
典型的大数据应用领域：

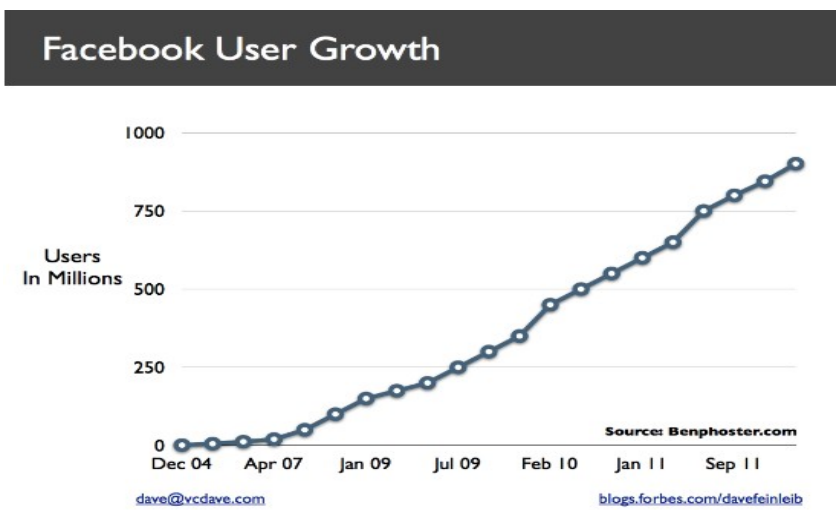
✓ 网络日志：

沃尔玛 2012 年平均每小时要处理大于 100 万客户的交易数据，数据容量约 2.5×10^6 G 字节的数据；

亿赞普每天要对 40×10^3 G 字节的日志进行数据挖掘、精确广告投送

✓社会网络：



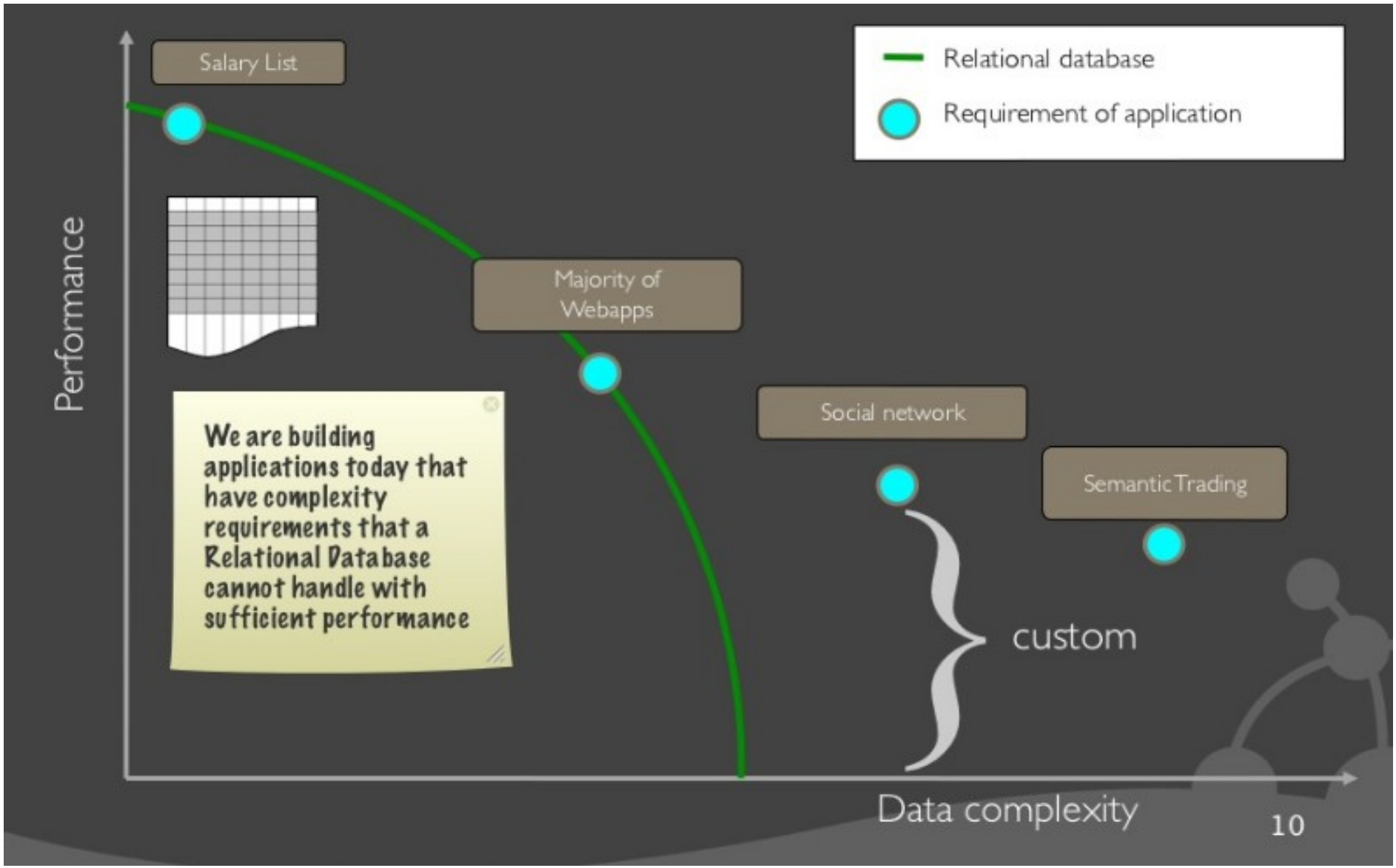


✓基因学，天文，地理信息

大数据结构复杂性

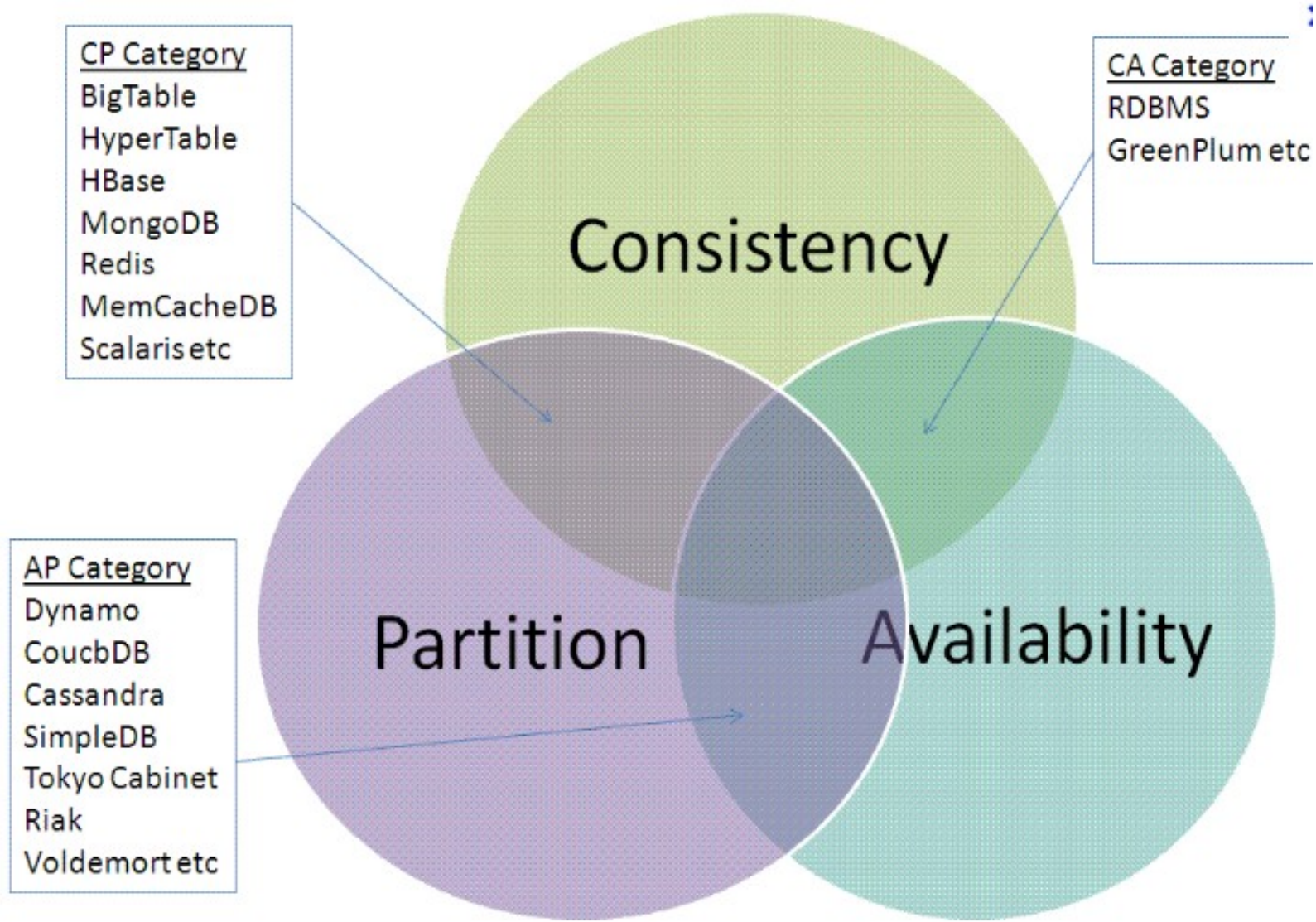
- ✓**结构化数据：**
  - 1）传统关系型数据库 Oralce， MySQL, postgresSQL,MSSQL
  - 2）二维表存储实体、关系模型
  - 3） 1NF， 2NF， 3NF 来确保数据无冗余以及数据一致性
- ✓**半结构化数据：**
  - 1）面向对象数据库（ Versant 、 UNISQL ）
  - 2） XML， RSS， JSON
- ✓**非结构化数据：**
  - 1）网页
  - 2）语义网， RDF
  - 3） Ontology
  - 4）标签、注释

结构复杂性与查询效率



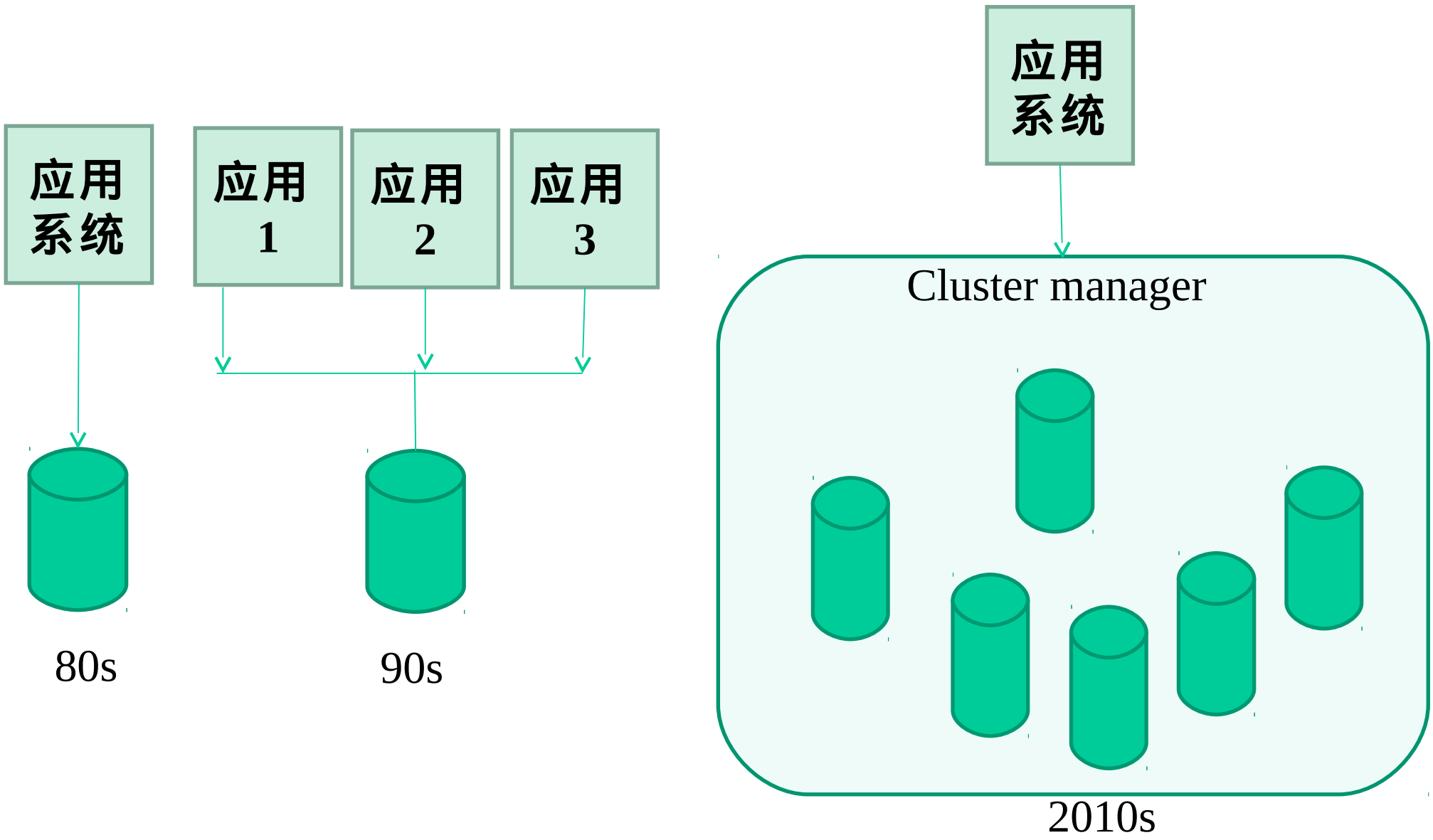
来源： Noe4j

数据库系统比较

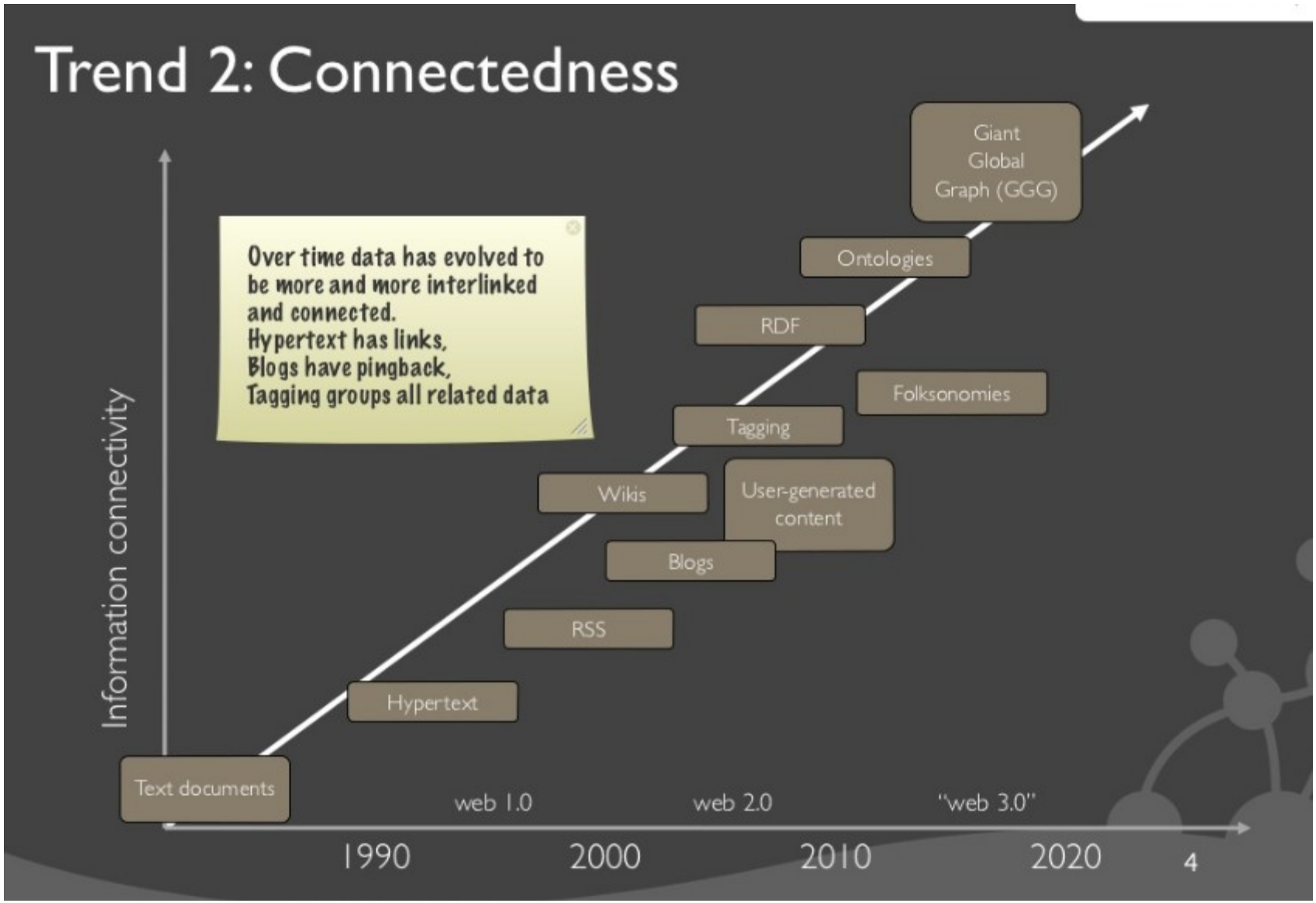


CAP theorem

数据库系统框架



大数据定义


Trend 2: Connectedness
Over time data has evolved to be more and more interlinked and connected.
Hypertext has links,
Blogs have pingback,
Tagging groups all related data
Information connectivity
Giant Global Graph (GGG)
Ontologies
RDF
Folksonomies
Tagging
Wikis
User-generated content
Blogs
RSS
Hypertext
Text documents
web 1.0
web 2.0
"web 3.0"
1990
2000
2010
2020
4

大数据定义

多准则问题：

大数据定义

大数据定义

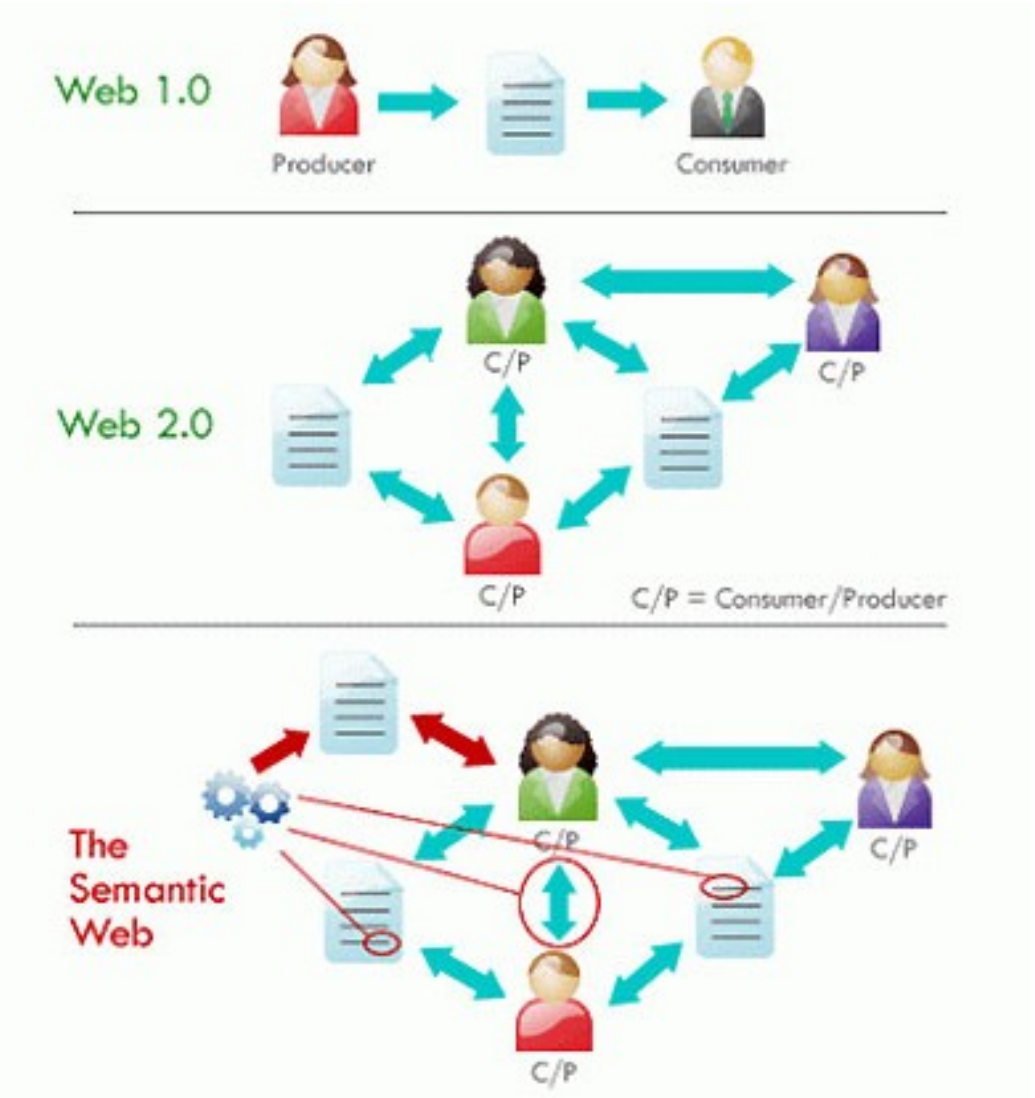
大数据定义

大数据定义

大数据定义

Web 1.0	Web 2.0	Web 3.0
"the mostly read only web"	"the wildly read-write web"	"the portable personal web"
45 million global users (1996)	1 billion+ global users (2006)	focused on the individual
focused on companies	focused on communities	lifestream
home pages	blogs	consolidating dynamic content
owning content	sharing content	the semantic web
Britannica Online	Wikipedia	widgets, drag & drop mashups
HTML, portals	XML, RSS	user behavior ("me-onomy")
web forms	web applications	iGoogle, NetVibes
directories (taxonomy)	tagging ("folksonomy")	user engagement
Netscape	Google	advertainment
pages views	cost per click	
advertising	word of mouth	

大数据定义


Web 1.0
Producer
Consumer
Web 2.0
C/P
C/P
C/P
C/P = Consumer/Producer
The
Semantic
Web
C/P
C/P
C/P

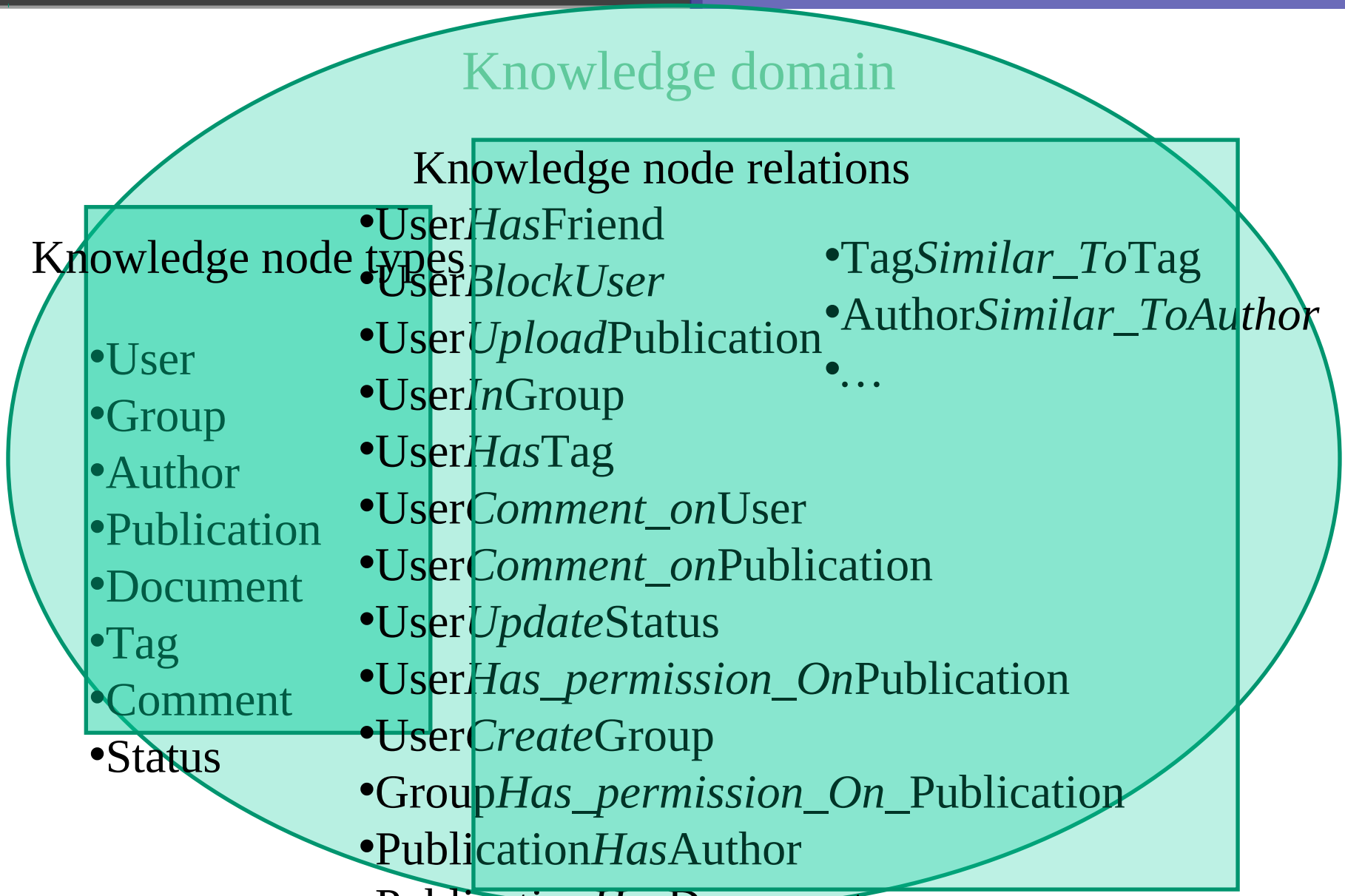
大数据定义

- A web were the context of the content defined as data
- A web capable of reading and understanding content and context
- A web can better satisfy the requests of the people and machine
- A web capable of filtering the content that of interest to the user

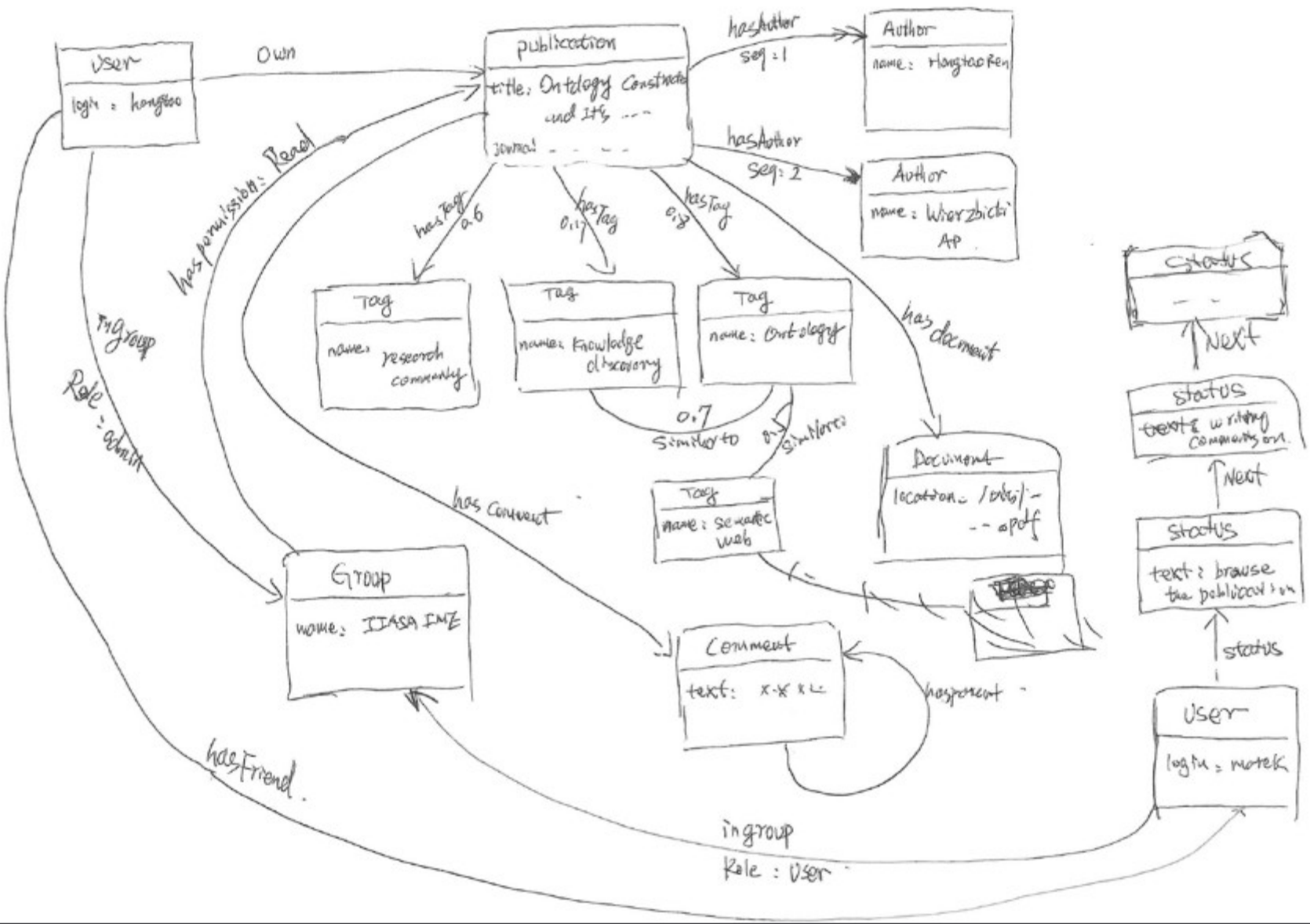
Challenges

- Connectedness: All different type of knowledge node are interlinked and connected
- Data size: For example, if we want to store 10 thousand publications, we may need store over 1 million related entities
- Semi-structure: Individualization of the knowledge node, the property of the knowledge node should be user definable
- Extracts implicit, potentially useful information from the data
- Knowledge visualization

大数据定义


Knowledge domain
Knowledge node types
•User
•Group
•Author
•Publication
•Document
•Tag
•Comment
•Status
Knowledge node relations
•User*Has*Friend
•User*BlockUser*
•User*Upload*Publication
•User*In*Group
•User*Has*Tag
•User*Comment_on*User
•User*Comment_on*Publication
•User*Update*Status
•User*Has_permission_On*Publication
•User*Create*Group
•Group*Has_permission_On*_Publication
•Publication*Has*Author
•Publication*Has*Document
•Publication*Has*Tag
•Tag*Similar_To*Tag
•Author*Similar_To*Author
•…

大数据定义


User
login : hongtao
Own
publication
title: Ontology Construct and its ...
journal ...
hasAuthor
seq : 1
Author
name : Hongtao Ren
hasAuthor
seq : 2
Author
name : Wierzbicki Adam
hasPermission: Read
hasTag
0.6
hasTag
0.1
hasTag
0.3
Tag
name: research community
Tag
name: Knowledge discovery
Tag
name: Ontology
0.7
similar to
0.5
similar
has document
Document
location: /obj/...
...pdf
Tag
name: semantic web
has Comment
Group
name: IIASA IMZ
Comment
text: x x x ...
hasComment
TriGroup
Role : admin
hasFriend
ingroup
Role : User
User
login : matek
status
Status
text : browse the publication
Next
status
text : writing comments on ...
Next
Status
Next

大数据定义

- ✓ JDK7
- ✓C++, PL solver
- ✓Tomcat7
- ✓Sun Grid Engine
- ✓Springframework
- ✓Axis2 Web services
- ✓Jfreechart
- ✓Hibernate
- ✓Oracle10g
- ✓Ajax

Layers:
- Java persistence layer
- Object-relational mapping abstraction layer
- Data access layer
- Services layer
- Process and integration layer
- Interfaces layer

Modules
- Problem
- Instance
- Analysis

大数据定义

大数据定义

Lessons from science-policy interactions:

- Much more demanding than pure science
- There is no golden key: modeling requires a combination of science, craft, art, experience
- Modeling for (interdisciplinary) knowledge integration and creation

A. Einstein:

Everything should be made as simple as possible, but not one bit simpler