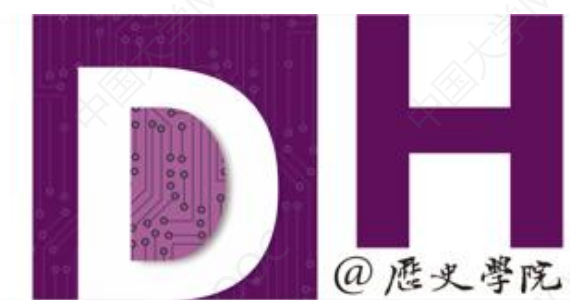




南京大學
人文社會科學高級研究院
Institute of Advanced Studies in Humanities and Social Sciences



Digital Humanities Initiative
IAS, Nanjing University
南京大学高研院 数字人文创研中心



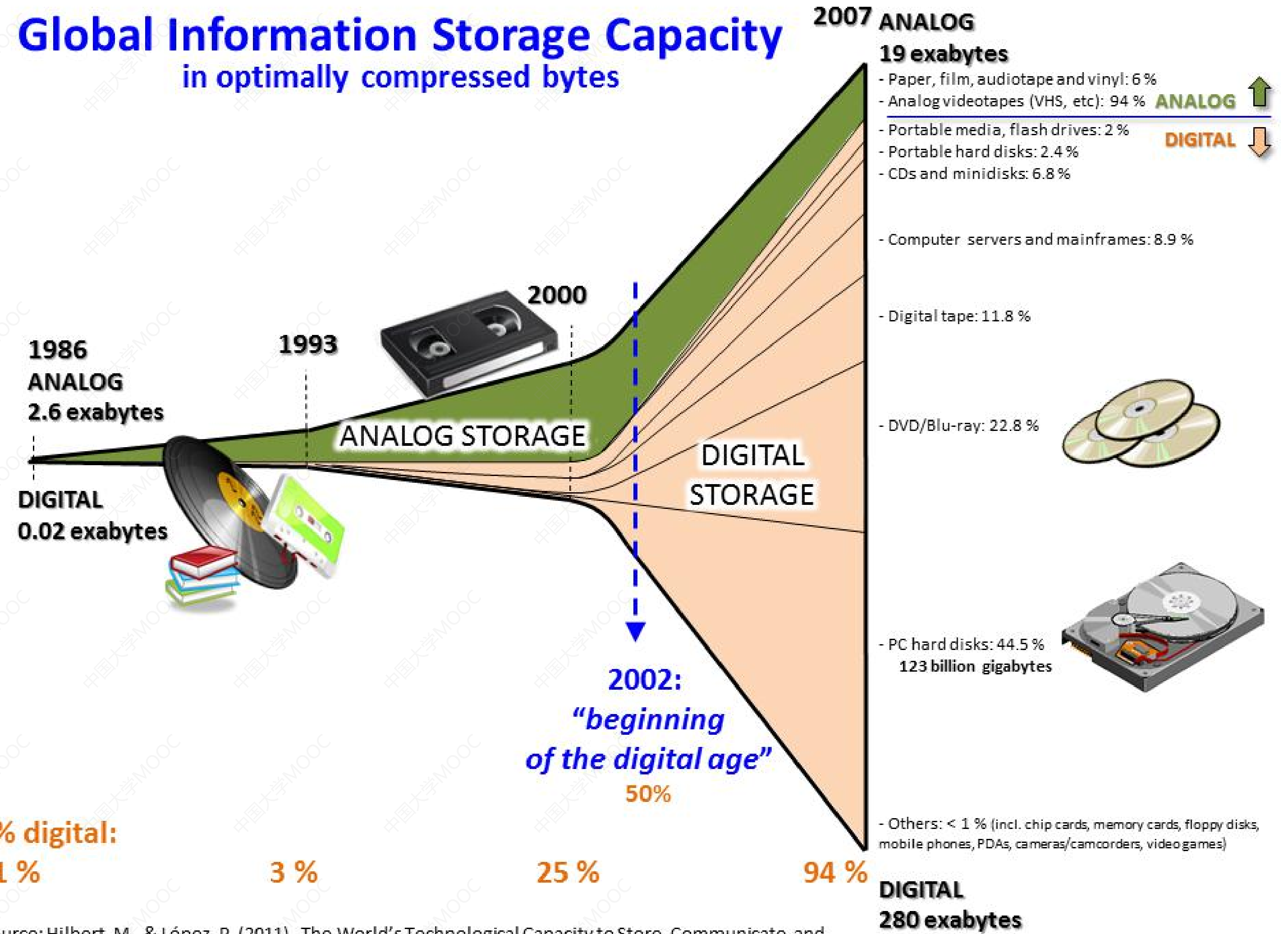
数字人文创新思维与方法

陈静、王涛、陈刚
2020年春

- 大数据与人文研究的关系；
- 什么样的数据是人文数据？
- 以数据驱动的人文研究有什么创新之处？

- 数据与数字技术
- information and communication technologies, ICTs
- 模拟数据与数字数据

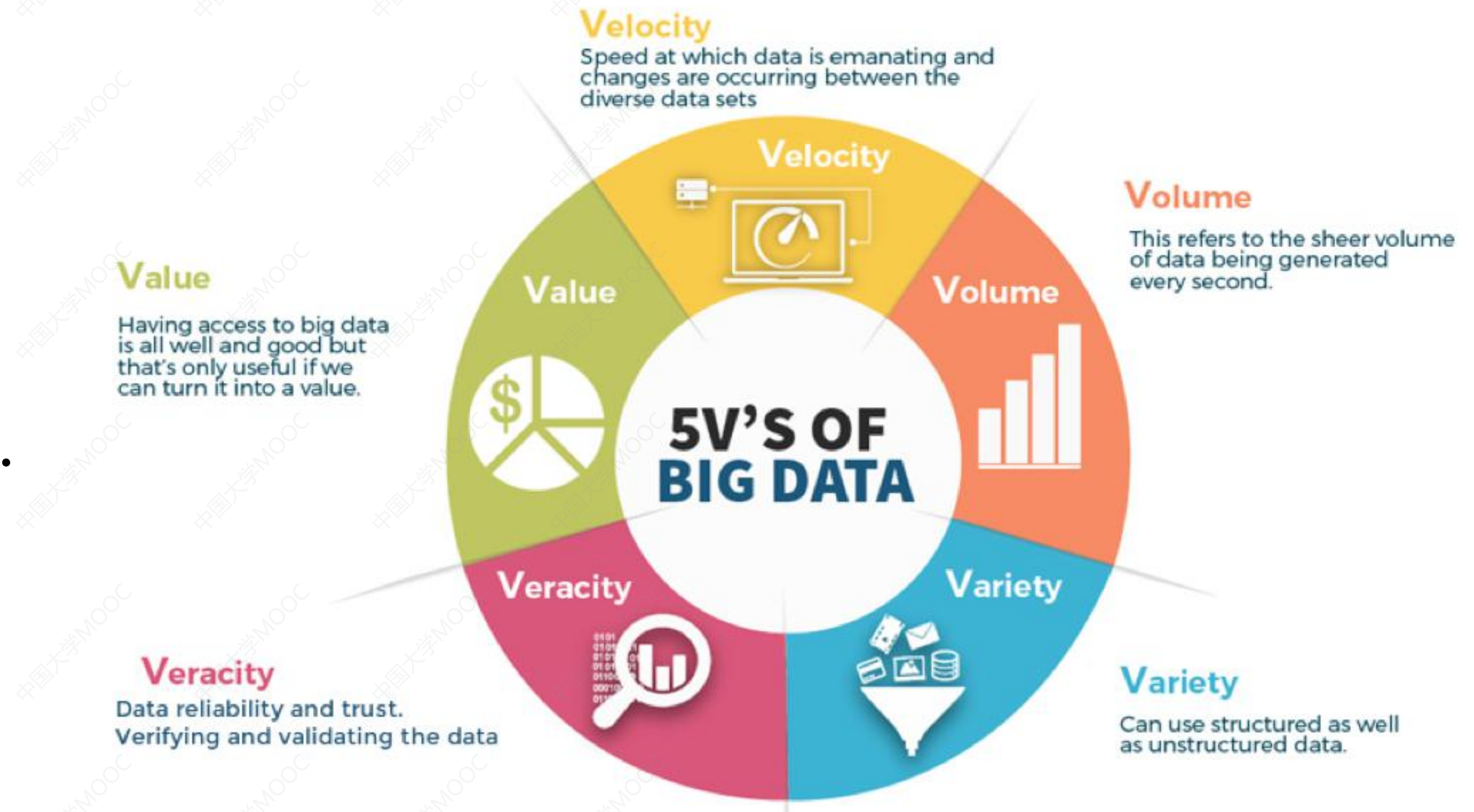
Global Information Storage Capacity in optimally compressed bytes



Source: Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025), 60–65. <http://www.martinhilbert.net/WorldInfoCapacity.html>

- Digital technology is transforming the assemblage and dissemination of historical information. Museums, libraries, archives, and universities increasingly modify their digital research infrastructures in order to make data open and available.
- Digital technology has been shaped the cyberinfrastructure, environment and condition of humanities studies.
 - digitized/born-digital text, image, video and data
 - computational analysis, multimedia, VR/AR, H(GIS) and 3D modeling tools and softwares
 - Popularization and personalization of digital technology

- “Big Data is a term applied to data sets whose size is beyond the ability of commonly used software tools to capture, manage, and process the data within a tolerable elapsed time. Big data sizes are a constantly moving target currently ranging from a few dozen terabytes to many petabytes of data in a single data set”



- Aside from how we define big data as a technological phenomenon, the wide variety of potential uses for big data analytics raises crucial questions about whether our legal, ethical, and social norms are sufficient to protect privacy and other values in a big data world. Unprecedented computational power and sophistication make possible unexpected discoveries, innovations, and advancements in our quality of life. But these capabilities, most of which are not visible or available to the average consumer, also create an asymmetry of power between those who hold the data and those who intentionally or inadvertently supply it.

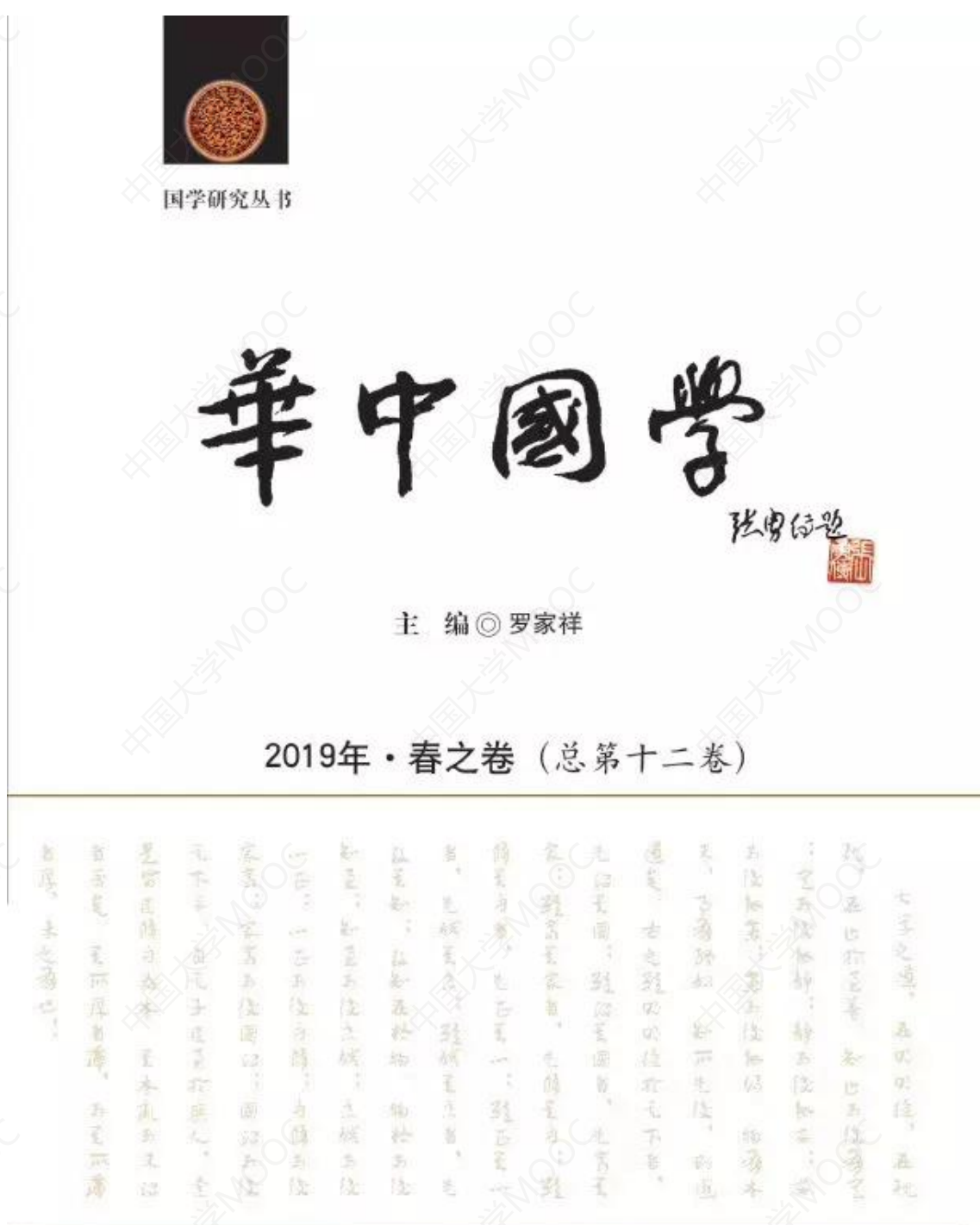
Executive Office of the President, *Big Data: Seizing Opportunities, Preserving Values*, 2014–5

- “The idea behind the Digging into Data Challenge is to address how ‘big data’ changes the research landscape for the humanities and social sciences. Now that we have massive databases of materials used by scholars in the humanities and social sciences—ranging from digitized books, newspapers, and music to transactional data like web searches, sensor data or cell phone records—what new, computationally-based research methods might we apply? As the world becomes increasingly digital, new techniques will be needed to search, analyze, and understand these everyday materials” .

“Digging into Data Challenge” , by the NEH together with a number of other research agencies in the United States, Canada, the UK, and the Netherlands

- In the twentieth century, the study of the social and the cultural relied on two types of data: “surface data” about lots of people and “deep data” about the few individuals or small groups.
- The first approach was used in all disciplines that adapted quantitative methods (i.e., statistical, mathematical, or computational techniques for analyzing data). The relevant fields include quantitative schools of sociology, economics, political science, communication studies, and marketing research.
- The second approach was used in humanities fields such as literary studies, art history, film studies, and history. It was also used in qualitative schools in psychology (for instance, psychoanalysis and Gestalt psychology), sociology (Wilhelm Dilthey, Max Weber, Georg Simmel), anthropology, and ethnography. The examples of relevant methods are hermeneutics, participant observation, thick description, semiotics, and close reading.

Lev Manovich, “Trending: The Promises and the Challenges of Big Social Data”, *Debates in the Digital Humanities*, <https://dhdebates.gc.cuny.edu/read/untitled-88c11800-9446-469b-a3be-3fdb36bfbd1e/section/65b49184-0ab4-4f29-859c-f7f02d875d8d>

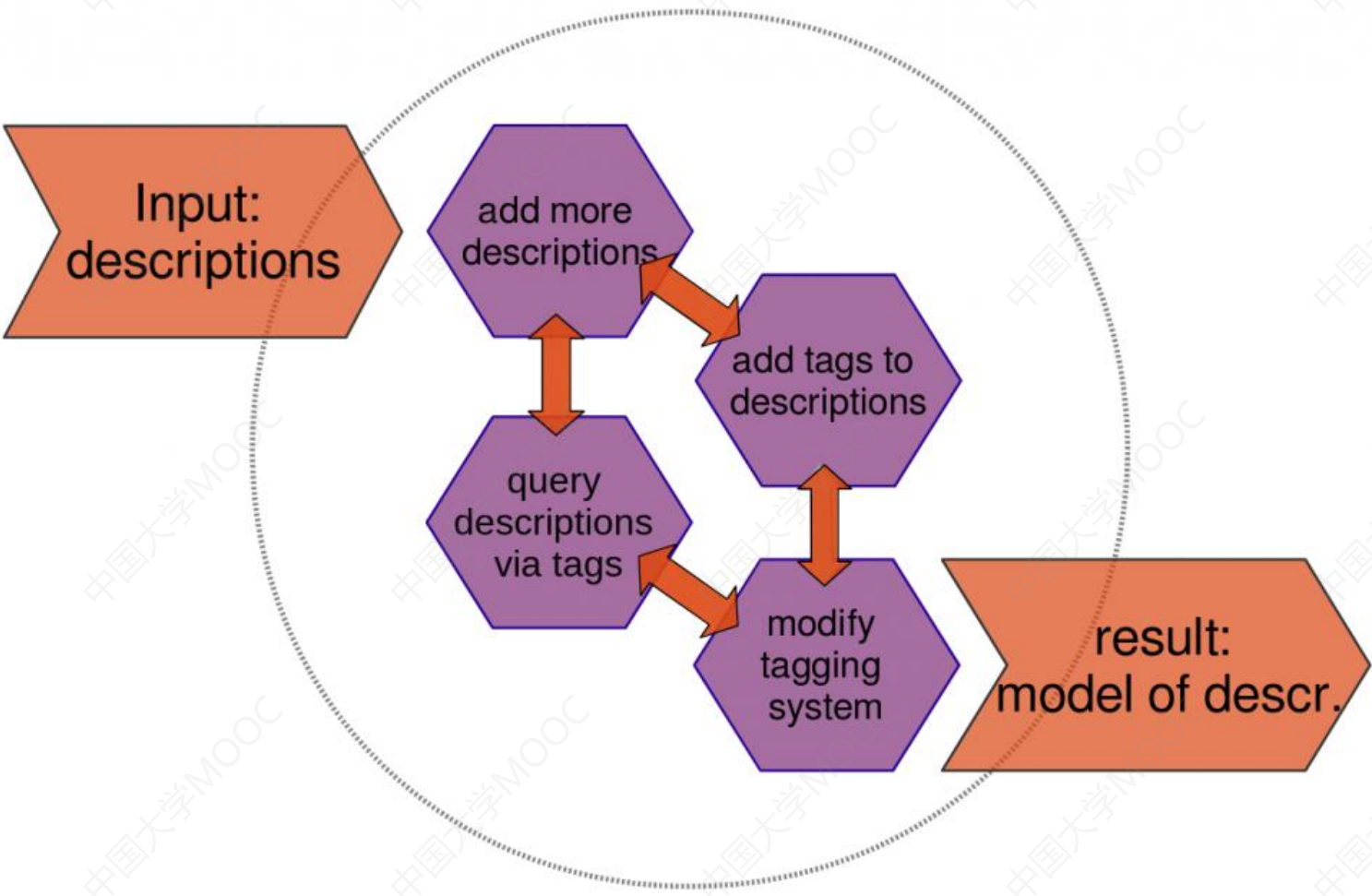


- 运用现代科学技术，我有两个衡量标准，那就是，首先它最后的精确度有没有其他方法加以验证；其次，它的结果有没有意义，能不能改变一个重大的学术论断。我发现大数据在历史研究中还是没有太大必要，因为我们掌握的数据不够，而且很多都是二手甚至三手数据，盲目运用的结果就是可信度越来越低，误差也会越来越大，到最后还是需要我们来做出判断和取舍，这是没有必要的。还比如天气预报，它是大数据的典型，现在相当一部分大部头的电脑都用在这上面，复杂的卫星云图用高速计算机几秒钟就可以被计算出来，但是它算出来的只是一种可能，最后的取舍也还是需要人根据历史经验去做出判断，所以为什么我们历史气候资料重要呢，就是这个原因。

——葛剑雄，《新科技与历史研究》

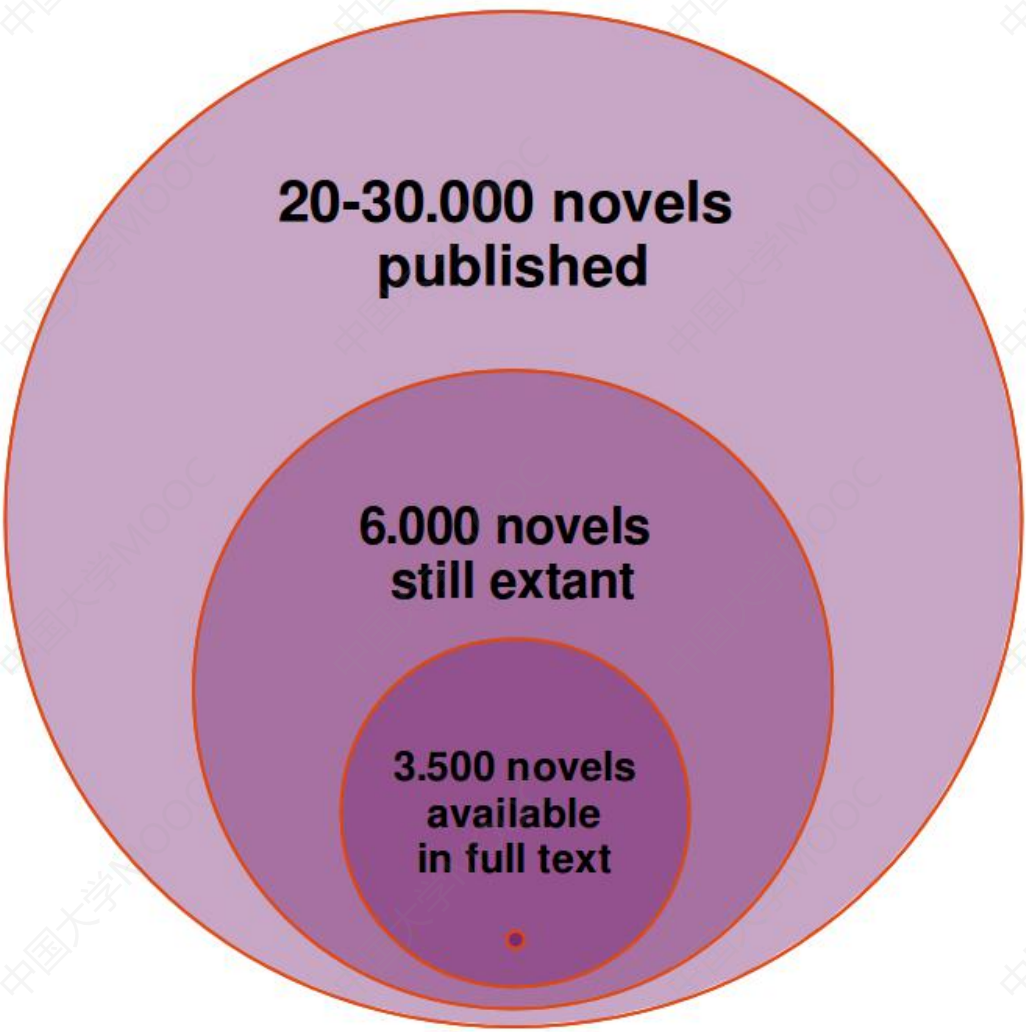
CHRISTOF SCHÖCH, “Big? Smart? Clean? Messy? Data in the Humanities”, *Journal of Digital Humanities* 2(3): 2-13.

- Data in the humanities could be considered a digital, selectively constructed, machine-actionable abstraction representing some aspects of a given object of humanistic inquiry.



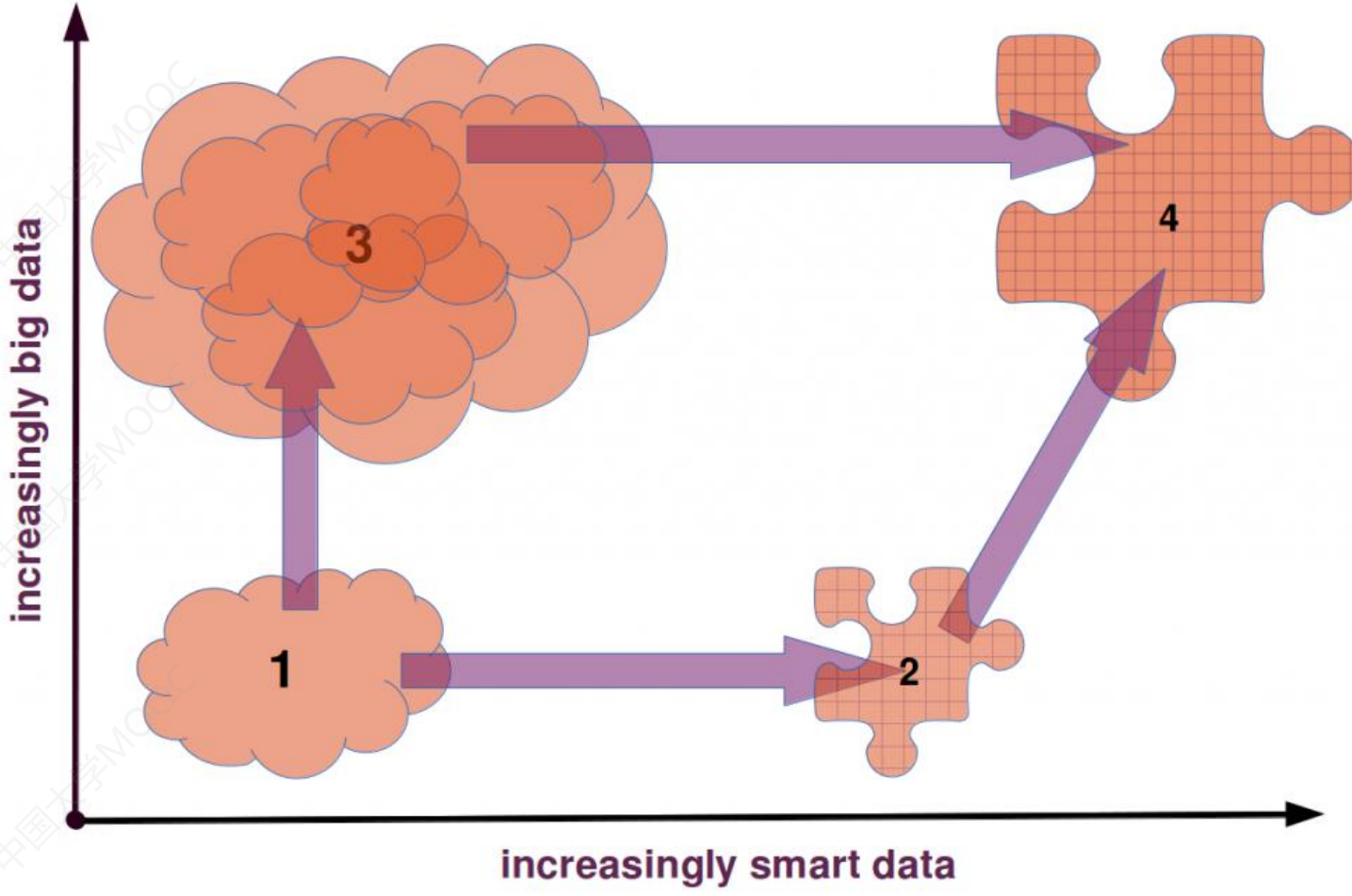
The iterative process of modeling descriptive techniques with keywords. Image published with a CC-BY license.

- Smart data as we find it in scholarly digital editions, in annotated linguistic corpora and in carefully curated image collections is immensely useful...it does not scale well...ultimately smart data depends on manual work by real people. This means that it is very time-consuming to create large volumes of smart data.



Digitization and availability of British Nineteenth-Century Novels. Image published under a CC-BY license.

- Big data in the humanities is not the same as big data in the natural sciences or in economics.
- Velocity does not play a key role in big humanities data right now. Also, the large “volume” is less usefully defined in the humanities by a shift from databases to distributed computing. Variety of formats, complexity or lack of structure does come into play.



The story of smart and big data

messy

big

data

small

smart

- 海量数据：相对非结构化的、混乱的和模糊的，数量相对巨大，且形式相对多样；
- 智慧数据：半结构或结构化的、整洁和明晰的，数量相对较小且异质性有限。

- 文本与图像都已经是数据
- 无论是文字文本还是图像，都是一种模拟的、非连续性的数据，无法以计算的方式来分析或转化；
- 语言、文本、绘画和音乐都是符号系统，它们具有某些超出物理度量的维度，某些关乎于“语境中的意义”的语义学或语用学的维度。
- 批评：人文文本如何量化？
 - 基于经验的解释
 - 研究者及其研究目标的情境性

- 数字化:
digitalization/digitization
- The Oxford English Dictionary (OED) traces the first uses of the terms ‘digitization’ and ‘digitalization’ in conjunction with computers to the mid-1950s.
- In the OED, digitization refers to “the action or process of digitizing; the conversion of analogue data (esp. in later use images, video, and text) into digital form.” Digitalization, by contrast, refers to “the adoption or increase in use of digital or computer technology by an organization, industry, country, etc.”

- digitization: the material process of converting individual analogue streams of information into digital bits.
- digitalization: the way in which many domains of social life are restructured around digital communication and media infrastructures.

Daniel Kreiss and Scott Brennen, International Encyclopedia of Communication Theory and Philosophy <http://culturedigitally.org/2014/09/digitalization-and-digitization/>

- The range of digitized material captured as digital images is as broad as the range of material held in libraries, archives, museums and private collections, including (but not limited to) printed books, printed journals, manuscripts, maps, photographs, photographic transparencies, music manuscripts, woodcuts, line drawings, paintings, archaeological site plans, archaeological finds, blueprints and architectural illustrations or plans, medical illustrations, documents, correspondence, newspapers and papyri. If processable, searchable text is required, further work is needed to generate this from an image of a text. Text can be ‘rekeyed’, where images of individual pages are used as the basis for an operator manually typing in content.
- Melissa Terras, “Digitization and digital resources in the humanities”, *Digital Humanities in Practice*, Facet Publishing, 2012

- Anything visual that is accessible to photography can be digitized, and there are many digitization processes, including: image scanning, microfilming and then scanning the microfilm; photography, followed by scanning of the photographic surrogates; rekeying (typing in) of textual content; OCR (Optical Character Recognition) of scanned textual content; encoding textual content to create a markedup digital resource; and advanced imaging techniques for large format or specialist items (Deegan and Tanner, 2002, 34).
- Sound and moving images can also be digitized, by rerecording video and audio onto digital media. The digitization process creates the core content for a digital resource, but additional infrastructure (such as a database, a website front end and an explanatory apparatus or additional teaching material) is required, in order to deliver the content successfully to users. Digitization is never a substitute for proper, structured collections management; although the development of detailed, up-to-date catalogues and metadata of institutional holdings may be a byproduct of digitization, this should not be the primary motive for carrying out the digitization in the first place (Hughes, 2004, 51 – 2).

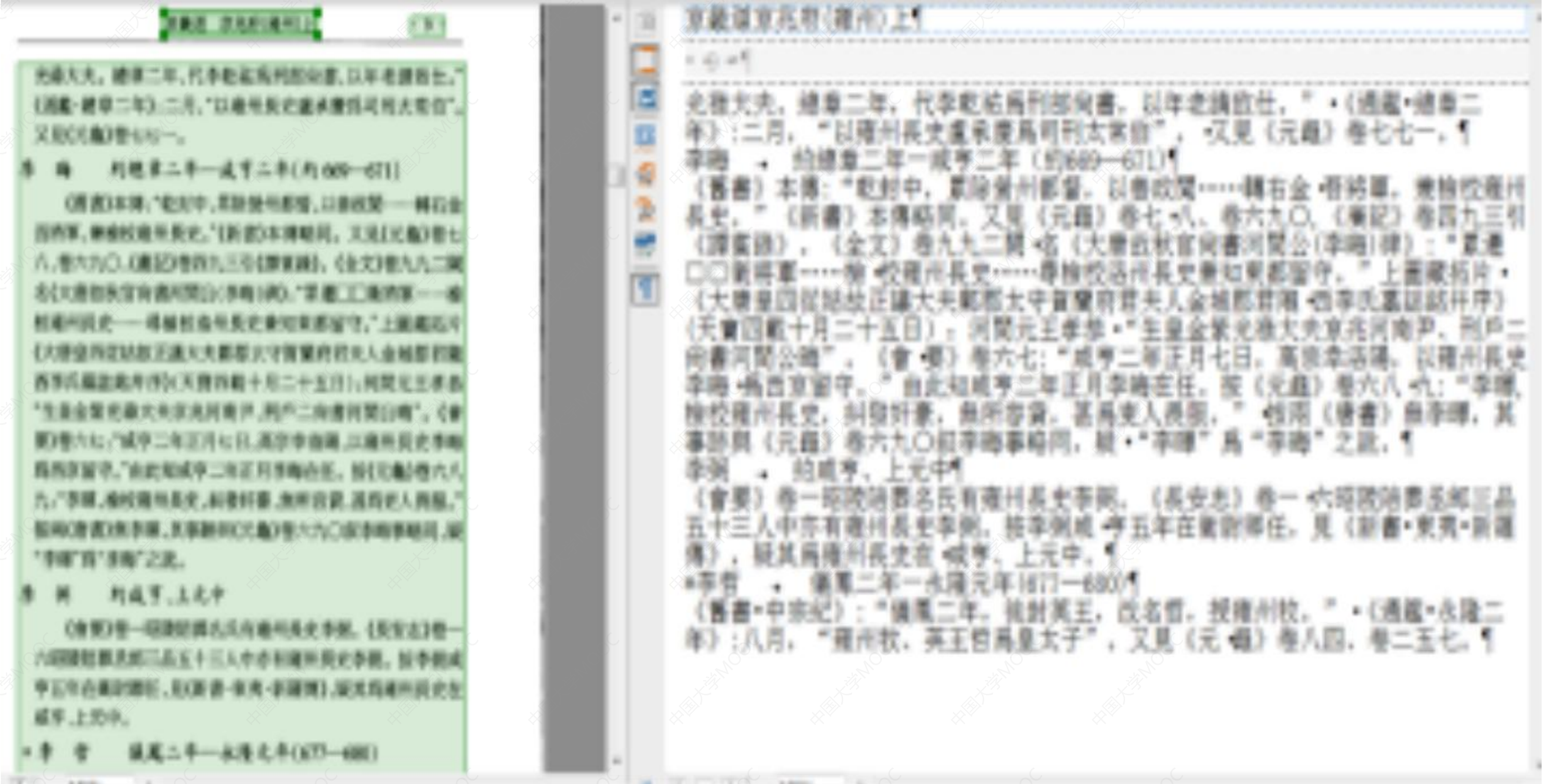
ImageTrac 6300 Class Scanners



- Harvard Law School
- 286 pages per minute
- 3 cents per page

• 以上内部资料引用，获得CBDB项目组授权使用

OCR Software: Abbyy



Proofreading the formatted data

河南道 陳州(淮陽郡) · 863 ·	
徐 向 大曆初	徐向:大曆初
李 芄 大曆十一年—十四年(776—779)	李芄:大曆十一年—十四年(776—779)
薛 珏 約大曆十四年—建中四年(約 779—783)	薛5玉:約大曆十四年—建中四年(約779—783)
薛 寶 建中四年(783)	薛寶:建中四年(783)

Collecting data points (with the help of regular expressions)

name	altname	office	page	time_proofreaded	time span1_Nianhao	time span2—split
郑元璜		太常卿	55	武德元年-四年（618—621）	武德元年	四年
窦诞		太常卿	56	武德四年-贞观元年（621—627）	武德四年	贞观元年
萧瑒		太常卿	57	约贞观元年-四年（约627—630）	约贞观元年	四年
萧瑒		太常卿	57	贞观六年-八年（632—634）	贞观六年	八年
杨师道		太常卿	57	约贞观八年-十年（约634—636）	约贞观八年	十年
韦挺		太常卿	58	贞观十年-约十八年（636—约644）	贞观十年	约十八年
尉迟敬德		太常卿	59	贞观十八年-十九年（644—645）	贞观十八年	十九年
杨师道		太常卿	59	贞观十九年-二十一年（645—647）	贞观十九年	二十一年
李勣		太常卿	60	贞观二十二年（648）	贞观二十二年	
李道宗		太常卿	60	贞观二十二年-永徽四年（648—653）	贞观二十二年	永徽四年
辛彦武		太常卿	61	贞观中？	贞观中？	
张文收		太常卿	61	永徽四年-五年（653—654）	永徽四年	五年
柳亨		太常卿	61	永徽五年-六年（654—656）	永徽五年	六年
高履行		太常卿	61	永徽末-显庆元年（？—656）	永徽末	显庆元年
薛景山		太常卿	62	显庆间？	显庆间？	
源直心		太常卿	62	龙朔二年（662）	龙朔二年	
李宽	丙宽	太常卿	62	高宗时	高宗时	
唐松龄		太常卿	62	高宗时？	高宗时？	
田仁会		太常卿	63	总章二年-咸亨元年（669—670）	总章二年	咸亨元年
裴明礼		太常卿	63	仪凤二年（677）	仪凤二年	
王德真		太常卿	63	约高宗末-嗣圣元年（？—684）	约高宗末	嗣圣元年
武承嗣		太常卿	63	嗣圣元年（684）	嗣圣元年	
窦孝谲		太常卿	64	约垂拱中	约垂拱中	
武重规		太常卿	64	约垂拱中-天授元年（？—690）	约垂拱中	天授元年

Disambiguate the data
Personal names
Alternative names
Dates of posting
Kinship relations ... etc.

name	inner disambiguate	fy	ly	office	full raw
李就		898		太常卿	○○《新书•宗室世系表上》大郑王房：“太常卿就。”乃太子太○傅同平章事李福子。按李福同平章事在咸通中，计其子仕历当○在昭宗时。○卷二太常寺○○○○

c_personid	c_name	c_name_chn	c_birthe ar	c_deathy ear	c_deathy ear_rang e	c_index_ year	唐代墓誌 pages	offices	Kins
187579	Li Jiu	李就		902	2	902		太常卿	李福(父);李彦孫(子)

Coding the Data (with Excel formulas, Macros, machine learning)

name	csk_id	duplicates	csk_by	csk_fulltext	cbdb_id	cbdb_by	office	kinNoRel	kin	officeMapping	officeMappingHlt	kinMapp	kinMag	yearDi	mark	disambiguation	.f
葛非元禮	csk_2000		716	2028-葛非元禮	376199	716	臺南府			1	臺南府	0		0	★★★★		
鄭濟	csk_3522		709	8310-鄭濟【break	191580	709	州刺史	鄭氏元真鄭濟	鄭氏(女兒)元真	1	州刺史	1	鄭氏	0	★★★★★		
李懷宣	csk_586		689	5488-李懷宣開元	189344	689	州刺史	李璿李璿李璿	李璿(父)李璿子	1	州刺史	1	李璿	0	★★★★★		
崔弘禮	csk_330		784	2125-崔弘禮【be	159925	784		崔弘禮	崔弘禮(子)	0		1	崔弘禮	0	★★★★		
郭孝恪	csk_1042		596	1782-郭孝恪【be	376534	596	太府少卿			0		0		0	★		
支訥	csk_12983		818	13119-支訥【be	142101	818		支訥支訥王氏	支訥(父)支訥(弟)	0		0		0	★		
楊國忠	csk_71		709	11172-楊國忠【	172579	709	司空	李璿楊國忠楊國	李璿(女婿)楊國	0		0		0	★		
李義	csk_5788		596	11047-李義約貞	32026	596		李義李義仁	李義(父)李義仁	0		1	李義仁	0	★★★★		
宋璟	csk_6009		707	12894-宋璟？【	31298	707		宋璟	宋璟(父)	0		1	宋璟	0	★★★★		
王思禮	csk_1025		714	1667-王思禮【be	184935	715	司空	王處城	王處城(父)	1	司空	0		1	★★★★		
郭鴻	csk_11994		812	12085-郭鴻【be	92701	811	州刺史			1	州刺史	0		1	★★★★		
高隱行	csk_13		606	2979-高隱行(真	376056	606	太常卿臺南府			1	太常卿	0		1	★★★★		
蔣璠	csk_1382		608	4319-蔣璠總章	94698	609	都散大夫殿中少監太常卿太僕卿太子右衛率			4	殿中少監太僕卿太	0		1	★★★★		
賈思直	csk_153		774	629-賈思直【be	172260	775	節度使太常卿	賈思直賈思直賈思直子賈思		2	節度使太常卿	0		1	★★★★		
李適之	csk_1589		686	2232-李適之【be	32128	685	兵部尚書大都督李適	李適(父)	6	大都督府長史太守	0		1	★★★★			
于邵	csk_1706		716	9884-于邵【brea	33083	717	比部郎中兵部郎中于邵誠于尹紀于邵誠父于尹			2	交廣副使州刺史	0		1	★★★★		
許世緒	csk_1882		581	3306-許世緒武周	376472	580	司農卿太府少卿			2	司農卿太府少卿	0		1	★★★★		
姚明敬	csk_2417		739	11085-姚明敬【	94162	738	觀察使司農卿州長史州刺史			2	州長史州刺史	0		1	★★★★		
源南	csk_2625		707	10295-源南【be	184051	706	給事中	源光伯	源光伯(父)	1	給事中	0		1	★★★★		
封昌道	csk_2691		596	2780-封昌道貞觀	172368	595	州刺史	封倫封思敬	封倫(父)封思敬	1	州刺史	0		1	★★★★		
薛訥	csk_275		685	4351-薛訥【brea	180663	684	州刺史	薛實亂薛用薛用薛實亂父薛用		1	州刺史	0		1	★★★★		
鄭休遠	csk_2991		668	4726-鄭休遠開元	376100	667	太常少卿			1	太常少卿	0		1	★★★★		
樊貞	csk_3430		645	5796-樊貞【brea	171231	644	州刺史	樊治	樊治(父)	1	州刺史	0		1	★★★★		
元祐壽	csk_3503		653	8279-元祐壽【be	184381	654	州刺史	元義端元愷	元義端(父)元愷	1	州刺史	0		1	★★★★		
崔玄同	csk_3513		681	5485-崔玄同約貞	94512	680	司封郎中刑部侍郎州刺史			1	州刺史	0		1	★★★★		
尚衡	csk_3524		714	3525-尚衡【brea	93736	715	兵部侍郎節度使右散騎常侍御史大夫州刺史			3	節度使御史大夫州	0		1	★★★★		
姚廷	csk_407		641	668-姚廷(姚班)	30974	640	戶部郎中戶部郎中姚處年姚廷	姚處年(父)姚廷	4	郡公太子詹事州刺	0		1	★★★★			
許文寶	csk_4440		575	5567-許文寶(許	376324	576	太僕少卿			1	太僕少卿	0		1	★★★★		
崔瑒	csk_444		707	5376-崔瑒約開元	94581	706	採訪使太守	崔玄暉	崔玄暉從父伯叔	2	採訪使太守	0		1	★★★★		
李茂貞	csk_657		854	1112-李茂貞【be	145345	855	節度使	郭氏宋瑛李繼	郭氏(妻子)宋瑛	1	節度使	0		1	★★★★		

Search or Browse Resources

Search

View Only Datasets

View Only Recipes

View Both Resource Types

Title	Description	Resource Type	Full Record
"Genre, Computation, and the Varieties of Twentieth-Century U.S. Fiction" derived data	Code and data to support analysis reported in "Genre, ...	dataset	Full Record
"Linguistic Markers of Status in Food Culture: Bourdieu's Distinction in a Menu Corpus" data and scripts Version 1.0	Supporting data and python and R scripts for the ...	dataset	Full Record
"The Migrant Letter Digitised: Visualising Metadata" datasets	Datasets used to support analysis reported in "The Migrant ...	dataset	Full Record
"Topic Modeling, Epistemology, and the English and German Novel" data	Novel text files, spreadsheets, stopword lists, and Python notebooks ...	dataset	Full Record
"Towards a Poetics of Strangeness: Experiments in Classifying Language of Technological Novelty" data	Code and data for "Towards a Poetics of Strangeness: ...	dataset	Full Record
"Understanding Gender and Character Agency in the 19th Century Novel" derived data	Code and data to support analysis reported in "Understanding ...	dataset	Full Record
2015 New Year's Resolutions	A Twitter sentiment analysis of users' 2015 New Year's ...	dataset	Full Record
20th Century Poetry	A table of derived word counts from a collection ...	dataset	Full Record
72 Hours of #Gamergate	Unfortunately, as it turns out, distributing the contents and ...	recipe	Full Record
A TEI XML edition of James Joyce's A Portrait of the Artist as a Young Man, corrected and revised	This is the open collaborative scholarly edition of James ...	dataset	Full Record

Showing 1 to 10 of 125 rows

10 rows per page

<

1

2

3

4

5

...

13

>

https://humanitiesdata.com/resources

20th Century Poetry

Description

A table of derived word counts from a collection of 75,297 English-language poems. A table with the top 20K words is located here and three tables of POS, Hypernyms, and word counts.

Resource Fields

Resource Type:
dataset
Submitted By:
Matthew Lavin
Date Submitted:
2018-09-07 17:00:23
Access URL:
https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/YVN6IW

Tags

- poetry
- word-counts

Project Open Data Required Fields (version 1.1)

Modified
[No data]
Publisher
[No data]
Contact Name
Andrew Piper
Unique Identifier
[No data]
Public Access Level
[No data]

Project Open Data Additional Fields (version 1.0)

Contact email
[No Data]
Endpoint
[No Data]
Format
csv, tab

Project Open Data Required-if-Applicable Fields (version 1.1)

Access Level Comment
[No Data]
Bureau Code
[No Data]
Program Code
[No Data]
License
[No Data]
Rights
[No Data]
Spatial
[No Data]
Temporal
[No Data]

.txtLAB

PROJECTS ▾

PUBLICATIONS

NEWS

txtLAB (McGill University)

A laboratory for cultural analytics

Harvard Dataverse > txtLAB > 20C Poetry

 Contact  Share



20C Poetry

Version 1.0

Andrew Piper, 2018, "20C Poetry", <https://doi.org/10.7910/DVN/YVN6IW>, Harvard Dataverse, V1, UNF:6:qxIF/P1xhAixReh9a8ACXg== [fileUNF]



Cite Dataset ▾

Learn about [Data Citation Standards](#).

Dataset Metrics

37 Downloads 

Description

This is a table of word counts for a collection of 75,297 English-language poems. (2018-08-07)

Subject

Arts and Humanities

Keyword

Poetry, Document term matrix

Related Publication

This data was used in Andrew Piper, Enumerations: Data and Literary Study (Chicago 2018). [isbn: 9780226568751](#)

Files

Metadata

Terms

Versions

Search this dataset...



Find

Filter by

File Type: All ▾

Access: All ▾

Sort ▾



1 to 2 of 2 Files



Download ▾



20CPoetryAll_Authors.tab

Tabular Data - 81.3 KB - Aug 7, 2018 - 25 Downloads

27 Variables, 455 Observations - UNF:6:qxIF/P1xhAixReh9a8ACXg==
Metadata of authors.



Explore ▾



Download ▾



20CPoetryAll_DTM.csv.zip

ZIP Archive - 28.6 MB - Aug 7, 2018 - 12 Downloads

MD5: f11328d80de0fc22c85f35e9d9ef2bf7

Word counts of twentieth century poems.



Download

Location	Title	Author	Edition	Features
ctext				🔍 ⬇
ctext				🔍 ⬇
ctext				🔍 ⬇
ctext			古今逸史	🔍 ⬇ 🖼
ctext			民國四明叢書	🔍 ⬇ 🖼
ctext			知不足齋叢書	🔍 ⬇ 🖼
ctext			欽定四庫全書	🔍 ⬇ 🖼
ctext				🔍 ⬇
ctext				🔍 ⬇
ctext				🔍 ⬇
ctext				🔍 ⬇
ctext			正統道藏	🔍 ⬇ 🖼
ctext			正統道藏	🔍 ⬇ 🖼
ctext			正統道藏	🔍 ⬇ 🖼
ctext			正統道藏	🔍 ⬇ 🖼
ctext			正統道藏	🔍 ⬇ 🖼
ctext			正統道藏	🔍 ⬇ 🖼
ctext			正統道藏	🔍 ⬇ 🖼
ctext			正統道藏	🔍 ⬇ 🖼
ctext			正統道藏	🔍 ⬇ 🖼
ctext			十萬卷樓叢書	🔍 ⬇ 🖼
ctext			正統道藏	🔍 ⬇ 🖼
ctext			正統道藏	🔍 ⬇ 🖼
ctext			正統道藏	🔍 ⬇ 🖼

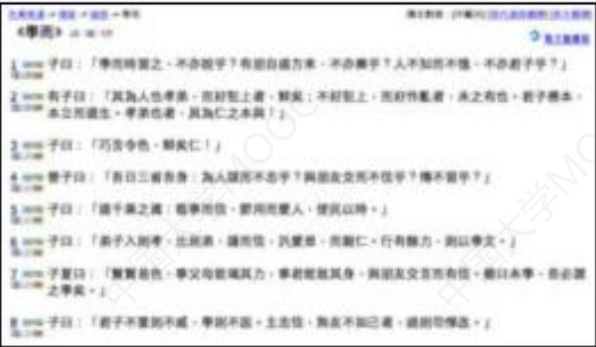
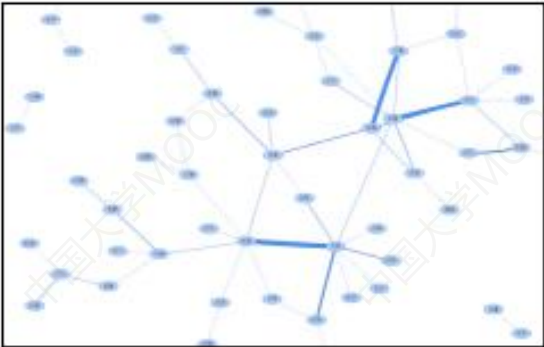
諸子 Chinese Text Project

Welcome

Welcome to the Chinese Text Project homepage. The Chinese Text Project is an online open-access digital library that makes pre-modern Chinese texts available to readers and researchers all around the world. The site attempts to make use of the digital medium to explore new ways of interacting with these texts that are not possible in print. With over thirty thousand titles and more than five billion characters, the Chinese Text Project is also the largest database of pre-modern Chinese texts in existence.

You may wish to [read more about the project](#), view the [pre-Qin and Han](#), [post-Han](#) or [Wiki](#) tables of contents, or [consult the instructions](#), [FAQ](#), or list of [tools](#). If you're looking for a particular Chinese text, you can [search for texts by title](#) across the main textual sections of the site.

網站有中、英文版本，也有繁、簡體版，可通過每頁左上角的連結隨時調整。



This site is provided as an open-access resource. **If you use this site in your teaching or research, please ask your university library to [subscribe](#).** [Donations](#) are also greatly appreciated.

Latest additions

Date	Content
2016-10-10	Harvard Yenching Library Chinese materials added Thanks to the support of Harvard Yenching Library, over 5 million pages of scanned materials from the Yenching Library collection have been added to the Library section of the site, including high quality images from the Chinese Rare Books Collection. Approximate transcriptions created using the ctext.org OCR procedure have also been added to the Wiki, making these materials full-text searchable. In future we hope to collaborate with other libraries to include materials from their Chinese language collections.
2015-07-02	Support for Unicode 8.0 adds new characters A new version of the Unicode standard has been released, defining thousands of additional rarely used and variant Chinese characters. Support for these has been added to the dictionary section of the site; to view these characters, please install the latest version of the Hanazono font. Many new characters belong to "CJK Extension E" - you can confirm system support for these from the Font Test Page.

Previous technical updates are listed in the ["Latest updates"](#) forum. Please note that only major technical updates are listed here; for content updates and recently added texts (which occur daily and hourly), please [log in](#) and refer to the [Wiki Recently Updated](#) sections.

Recent discussion topics

Date	Content
2020-03-01 by ethanyet	戰國策的版本存在非常多的錯誤 貴站的戰國策文本存在很多注音打字的同音錯誤以及許多錯誤。建議換一個版本。 [More...]
2020-02-24 by haenim	There are so many translation errors on Mengzi (Mencious) For example; https://ctext.org/mengzi/liang-hui-wang-i [More...] [2 replies]
2020-02-23 by haenim	重複頁 "古本大學" 4~17 頁, 18~33 頁, 相同 [More...]

Fonts and Compatibility

In order to view the Chinese text on this site, you will need to have a Chinese font and compatible web browser installed. In addition, because the ancient texts on this site incorporate many now obsolete and rarely-used characters, you will need to install a Unicode font which provides support for such characters (often referred to as CJK Extensions A through D). A suitable font is the freely downloadable "[Hanazono](#)" font; "[Han Nom](#)" also includes many (though not all) of the required characters. To confirm whether your system properly supports these characters, please see the [Font test page](#). Proper support for Unicode (including so-called 'surrogates') is also needed, which may cause additional problems for users with older operating systems.

The site is designed to comply to relevant W3C standards, and so should be usable in any standards-compliant web browser. However, as the emphasis on this site is on usability and not compatibility with legacy browsers, it is recommended that you upgrade to a browser with good CSS support such as Mozilla Firefox to get the most out of the site.



Visual Geometry Group



People



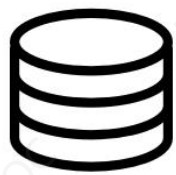
Research



Publications



Demos



Data



Software



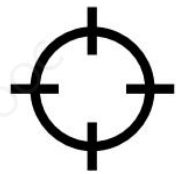
Practicals



Projects



Internal



Jobs

Visual Geometry Group

Department of Engineering Science, University of
Oxford.

About

- Location
- Contact

Socials

- Twitter
- Blog

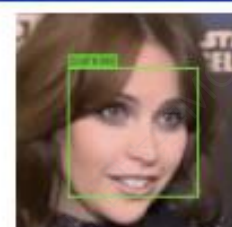
HOME PEOPLE RESEARCH PUBLICATIONS DEMOS DATA SOFTWARE PRACTICALS PROJECTS JOBS INTERNAL

Data



Click on a dataset category to expand/collapse it or [click here to expand/collapse ALL categories](#)

Audio-visual Face



VoxCeleb Speaker Identification Dataset



VoxCeleb2 Dataset



BBC-Oxford Lip Reading Sentences Dataset



BBC-Oxford Lip Reading Dataset

Face Recognition



VGG Face Dataset



VGG Face 2 Dataset



Celebrity Together Dataset



Celebrity in Places Dataset



Labeled Ancestral Origin Faces in the Wild

Video-based Recognition



Sherlock Character Identification Dataset



Looking at Each Other Annotations



Arrow of Time



Automatic Naming of Characters in TV Video



TV Human Interactions Dataset

The Paintings Dataset



Elliot J. Crowley, Ernesto Coto and Andrew Zisserman

Overview

The dataset covers ten of the categories present in [PASCAL VOC](#), and is split into training, validation, and test sets. Objects have a variety of sizes, poses and depictive styles, and can be partially occluded or truncated.

The paintings have been obtained from [Your Paintings](#). The annotations have been provided by the public as part of the [Tagger Project](#) as well as having been extracted from the painting titles.

Downloads

The URLs for the paintings and their labels are provided in the following spreadsheets:

- [Spreadsheet of Data \(March 2018\)](#): "Your Paintings" has been completely moved to [ArtUK](#). This new spreadsheet contains the updated URLs to the images and the web pages with information about them. Unfortunately, please be aware that **56 images from the original dataset are now inaccessible**. They are marked as "N/A" in the spreadsheet. In addition, all [ArtUK](#) URLs are pointing to images **with a fixed width of 550px**.
- [Spreadsheet of Data \(old version\)](#)

Class Examples

Aeroplane

Bird

Boat

Chair

Cow

Dining-table

Dog

Horse

Sheep

Train

Painting Statistics

The table below shows the number of images in the dataset containing an instance of a particular class, as well as the total number of images for each subset.

	Aero	Bird	Boat	Chair	Cow	Dtable	Dog	Horse	Sheep	Train	Total
Train	74	319	862	493	255	485	483	656	270	130	3463
Val	13	72	222	140	52	130	113	127	76	35	865
Test	113	414	1059	569	318	586	549	710	405	164	4301
Total	200	805	2143	1202	625	1201	1145	1493	751	329	8629

The State of the Art: Object Retrieval in Paintings using Discriminative Regions

Elliot J. Crowley
elliott@robots.ox.ac.uk

Andrew Zisserman
az@robots.ox.ac.uk

Visual Geometry Group
Department of Engineering Science
University of Oxford

Abstract

The objective of this work is to recognize object categories (such as animals and vehicles) in paintings, whilst learning these categories from natural images. This is a challenging problem given the substantial differences between paintings and natural images, and variations in depiction of objects in paintings.

We first demonstrate that classifiers trained on natural images of an object category have quite some success in retrieving paintings containing that category. We then draw upon recent work in mid-level discriminative patches to develop a novel method for re-ranking paintings based on their spatial consistency with natural images of an object category. This method combines both class based and instance based retrieval in a single framework.

We quantitatively evaluate the method over a number of classes from the PASCAL VOC dataset, and demonstrate significant improvements in rankings of the retrieved paintings over a variety of object categories.

1 Introduction

The question we investigate in this paper is: can paintings containing an object category (e.g. a train or a bird) be retrieved starting from a model learnt from natural images? At first sight, we might not be optimistic since natural images (i.e. everyday photos taken with a camera) and paintings can have very different low level statistics, and paintings vary considerably in depiction style from photo-realistic renderings through particular movements (e.g. impressionism, pointillism – which are almost designed to disrupt the fine scale measurement of local gradients in a HOG or SIFT feature) to more abstract depictions (Fauvism, Cubism).

Apart from the challenge in its own right, this goal of automatically obtaining paintings with a particular object is of much interest to Art Historians who currently find paintings manually or from memory [8, 22, 34]. They can then study the change in the depiction style over time [25], or determine when an object first appeared in paintings.

The problem is essentially one of domain adaptation (also referred to as domain transfer) [12, 24, 31] from natural images to paintings. The problem of generalizing across depiction styles has been studied recently by Wu and Hall [35]. They took the interesting approach of building a multi-layer depiction invariant graph model that was shown to be capable of generalizing to drawings and cartoons in particular. However, a limitation of the method was

that it was restricted in both training and testing to uncluttered Caltech101 [15] style images – where the object of interest fills the image against a uniform background. In our work, both training and testing images are PASCAL VOC style [13] where the object may only occupy a small part of the image, and can be partially occluded, see figure 1. Others [6, 32] have recently considered the problem of using a (single) natural image to retrieve paintings, in particular for the case of a specific building, rather than a class of objects: Shrivastava *et al.* [32] use an Exemplar SVM [26] for retrieval, and Aubry *et al.* [6] improve on this method by employing mid-level discriminative patches (MLDPs) [7, 23, 27, 33] to allow for more variation. We build on and extend the method of [6] from *specific* buildings to *classes* of objects, and overcome its two principal limitations: that the training images must have a very similar pose to the target object, and that the training images have the object (a building in their case rendered from a 3D model) segmented. Others have investigated classification and retrieval in paintings, e.g. the interesting analysis of [9], but have not considered this domain adaptation aspect.

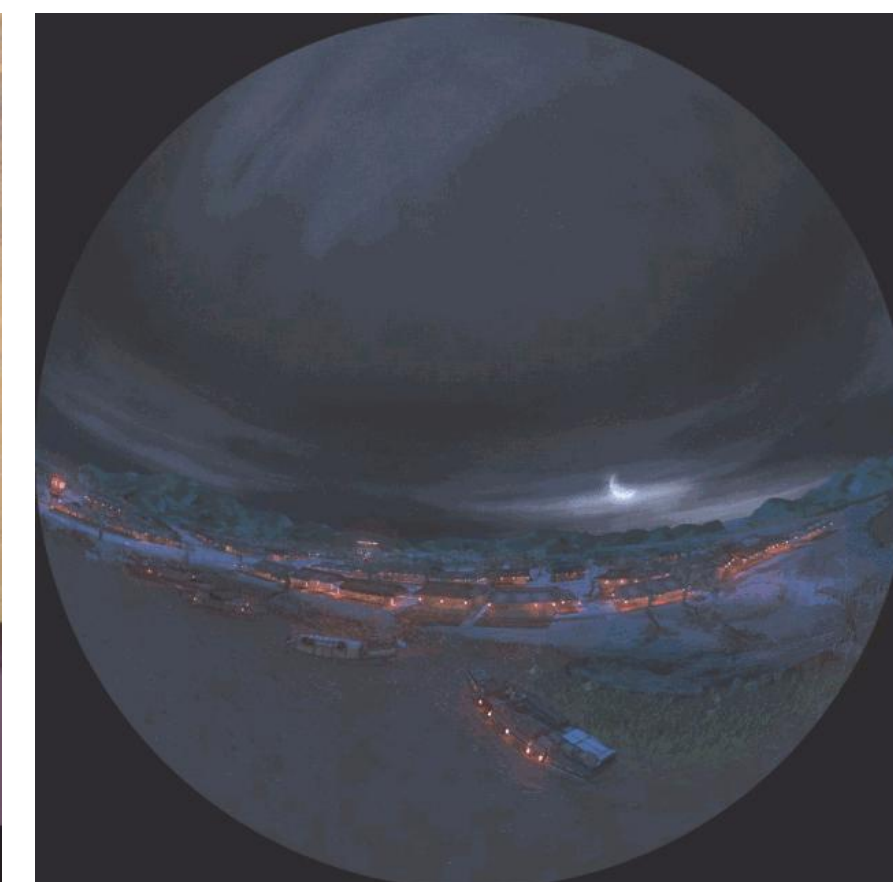
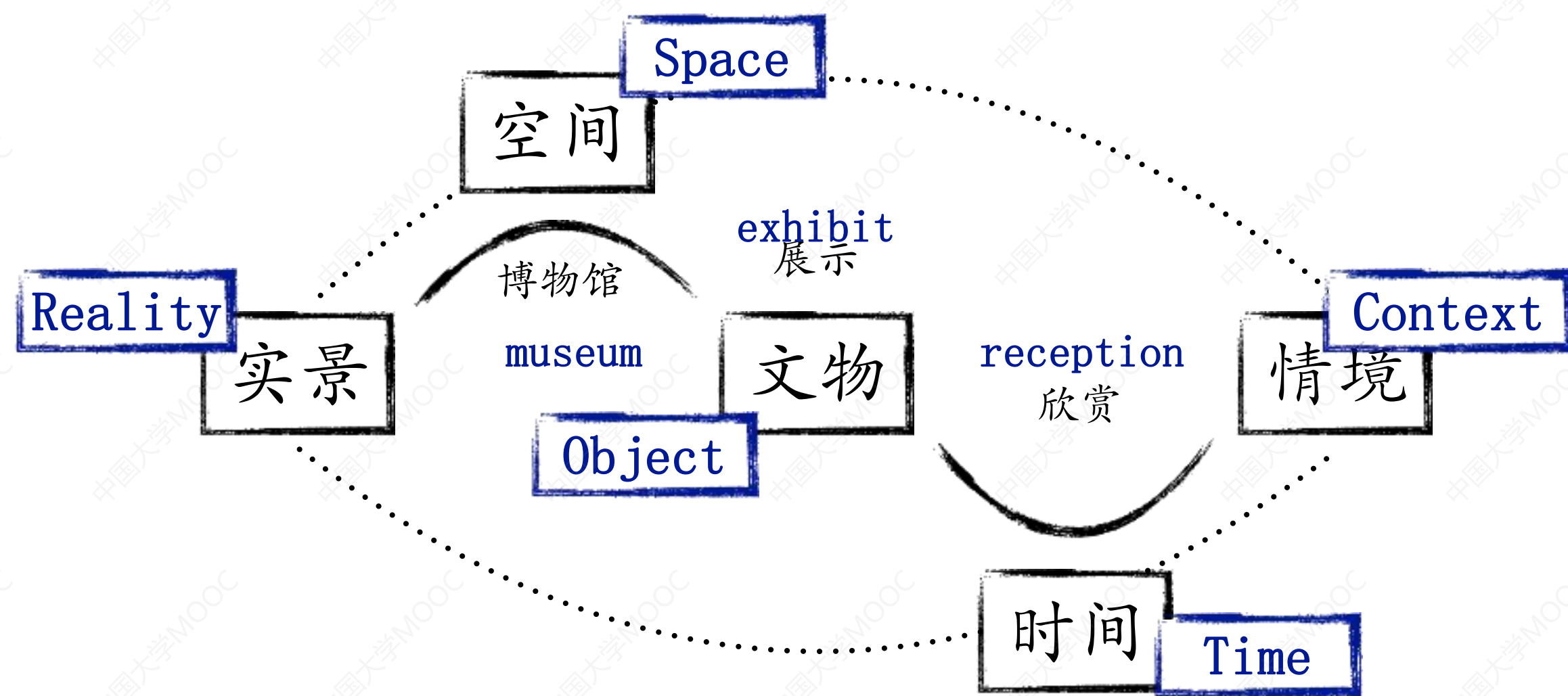
We make the following contributions: (i) we show, somewhat surprisingly, that image classifiers and object detectors learnt from PASCAL VOC images can retrieve paintings containing an object class (section 2); (ii) we introduce a method of re-ranking that is based on spatial consistency of MLDP correspondences (section 3), and show that the precision of low ranked paintings (i.e. the ones that would appear on the first webpage in an image search) can be significantly improved based on how spatially consistent the paintings are with the natural images used to train the classifiers (section 4); and (iii) we compare other methods of training and re-ranking including training from Google images (where the images are more Caltech101 style) and using a DPM [16] detector, and also investigate hybrid re-ranking strategies.

Note, although using spatial consistency to re-rank is standard practice in large scale object instance retrieval [21, 30], using it in this manner for object categories is novel, and contrasts to the spatial consistency implicit in Spatial Pyramids and DPMs.



Figure 1: Example class images from the Paintings Dataset. >From top to bottom row: dog, horse, train. Notice that the dataset is challenging: objects have a variety of sizes, poses and depictive styles, and can be partially occluded or truncated.

- Environmental representation: 360°、180° ；
- Physical records: video and image record, 3D modeling, AR and textual description/影像记录、3D建模、AR和文字描述；
- Restoration of context: VR、AR、audio and 3D Space modeling/VR、AR、声音和3D空间建模



- Semantic Web（语义网）
- Ontology Model（艺术品的本体模型）
 - Conceptual Reference Model（CIDOC-CRM）
 - Europeana Data Model（EDM）
- Visual Resources Association（VRA Core）



故宫博物院
THE PALACE MUSEUM

首页

导览

展览

教育

探索

学术

文创

关于

人

中

Q

三

首页 » 探索 » 藏品 » 陶瓷



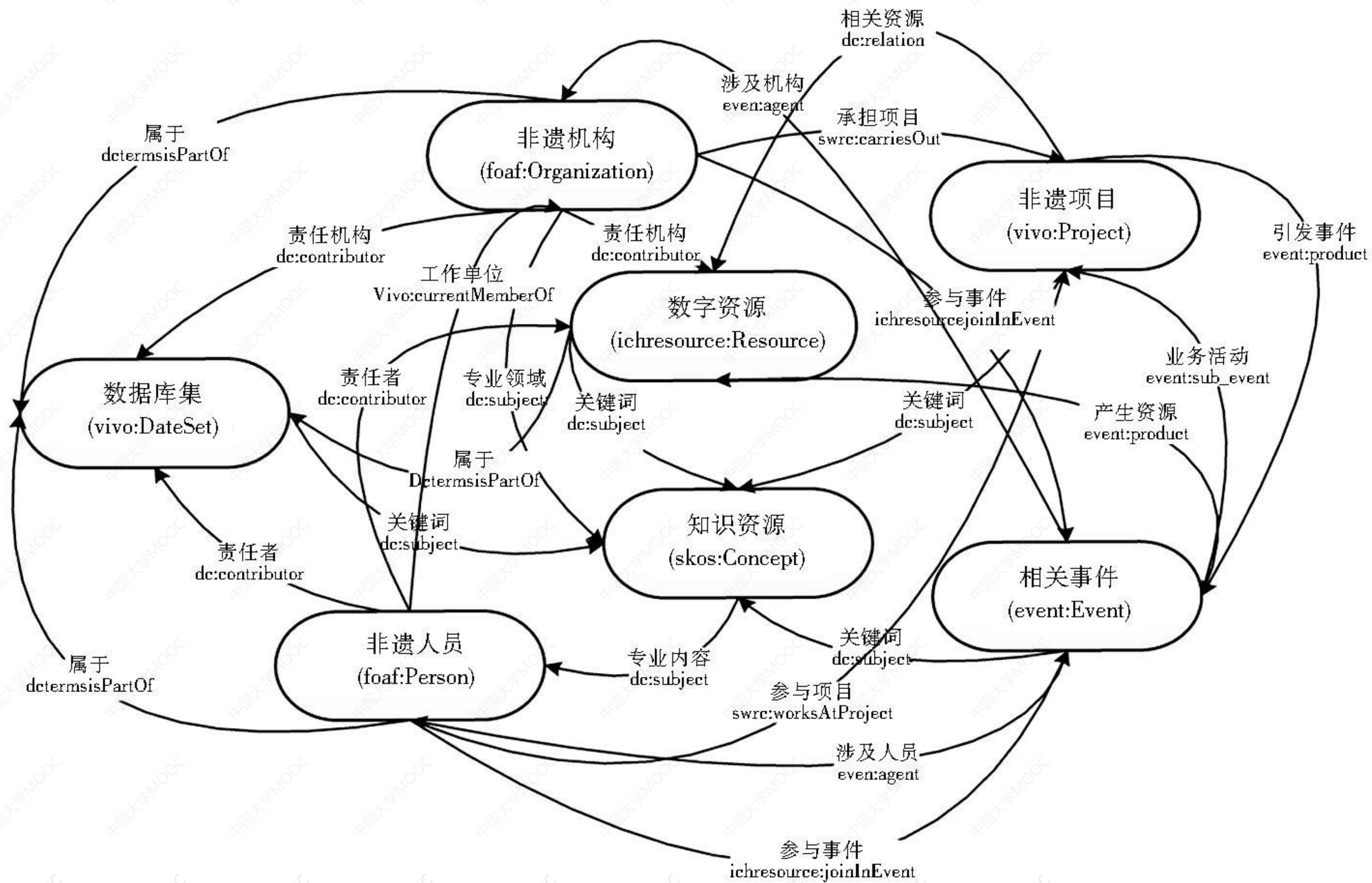
【白陶刻几何纹甗】

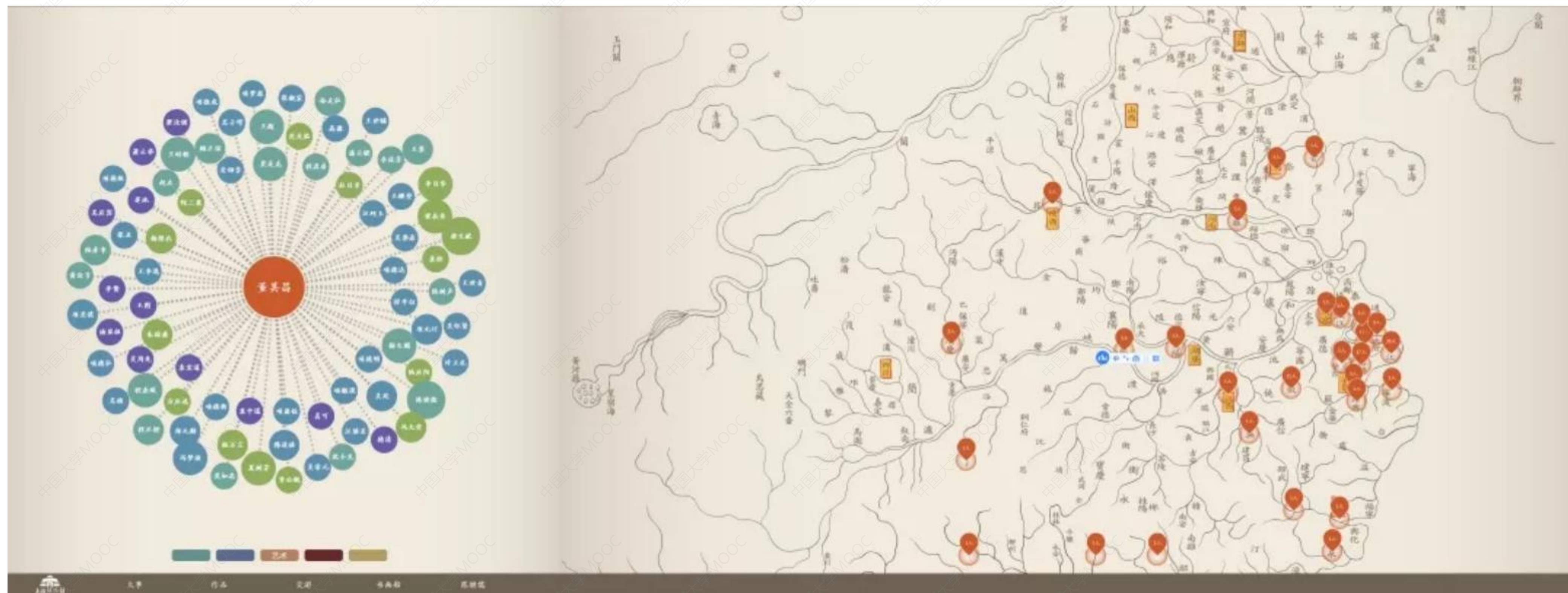
白陶刻几何纹甗，商，高20cm，口径18.6cm，足径15cm。

甗（bù 音不）唇口外卷，溜肩，鼓腹，腹下渐收，近足处外撇，圈足。通体雕刻纹饰，以精细的回纹作地衬托几何纹，主次分明，错落有致，显得格外庄重精美。

白陶早在新石器时代晚期就已出现。至商代，由于烧成温度提高，原料的淘洗亦较精细，致使白陶质地更加洁白细腻。商代早期白陶器形以鬲（guī 音归）、盂（hé 音禾）、爵为主，纹饰有人字形纹、拍印的绳纹和附加堆纹等。商代中期，器物增加了豆、罐、钵等，其装饰除少数绳纹外，以素面磨光为多。商代后期是白陶烧制的鼎盛期，在黄河流域的商代晚期遗址与墓葬中均发现不少白陶，其中以河南安阳殷墟出土的白陶最具特点，器物有觶（zhì 音制）、壶、尊、卣（yǒu 音有）等酒器和鼎、豆、盘、簋（guǐ 音轨）等食器。纹饰常见有云雷纹、漩涡纹、饕餮（tāo tiè 音涛帖）纹、蝉纹、曲折纹、

白陶刻几何纹甗，
<http://www.dpm.org.cn/explore/collections.html>





陈晴. 大数据分析在博物馆场景中的应用——以上海博物馆数据中心为例[J]. 科学教育与博物馆, 2018, 4(3): 188-199.

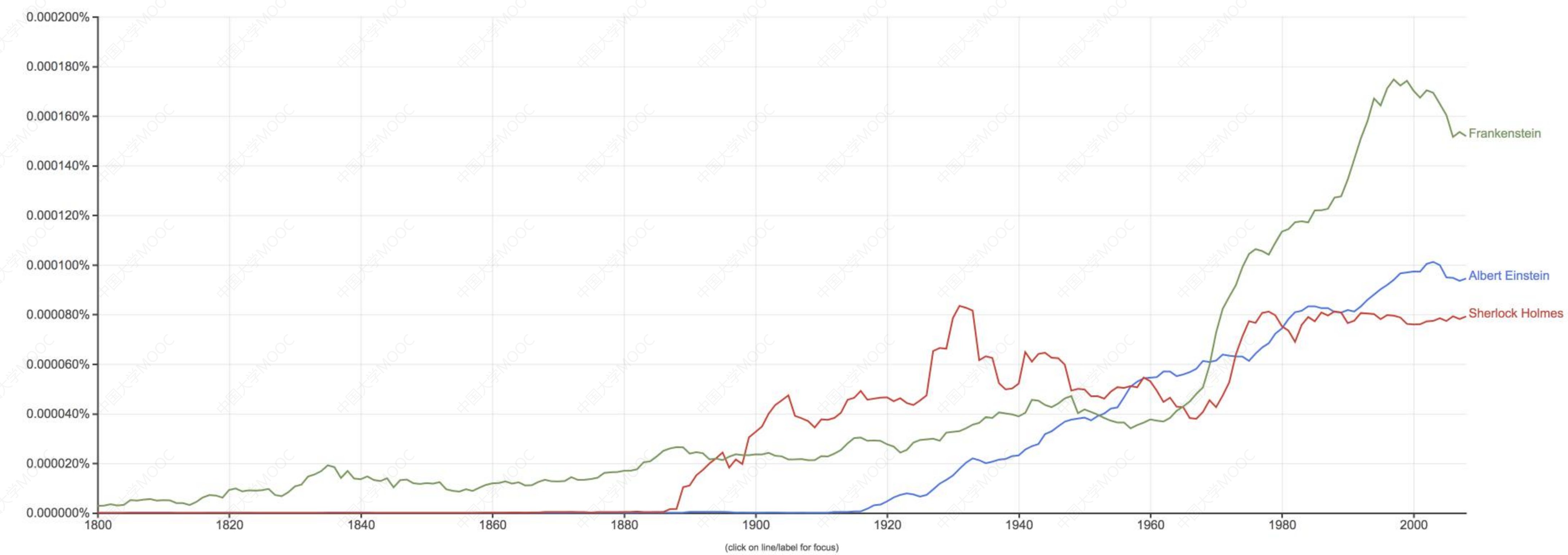
Graph these comma-separated phrases: ☐ case-insensitive

between and from the corpus with smoothing of .

[Search lots of books](#)

[Tweet](#)

[Embed Chart](#)



Run your own experiment! Raw data is available for download [here](#).

- 让-巴蒂斯特·米歇尔为首的研究团队与谷歌图书合作开展的“基于百万数字图书的文化量化分析”就是基于5,195,769本数字化书籍的语料库的计算分析（computational analysis），以量化的方式呈现文化趋势。研究团队以自然语言处理中较常用的n-gram模型方法，以单个词或多个词为单位，来自全世界的大学图书馆的1500万本数字化图书中选择了从1800年到2000年间的500万本，共计7种语言500亿字的文本量进行了统计分析，对包括英语词汇量变化、英语语法的变迁、集体记忆与健忘、大众声望、审查检测等等文化议题进行解读。因为是基于两百年间的词频波动观察，所以得出的一些结果还是非常具有启发性的。这种通过对大数据集进行定量分析从而学习人类文化的方式被命名为“文化测量”模式，并于2011年在《科学》杂志发表。此后，不少学者也将此模式也施用于了不同的文化数据集。比如卡莱弗·李塔鲁（Kalev H. Leetaru）对三十年间全球的本地新闻进行了调性和地理分析，并成功预测了2011年在阿拉伯半岛发生的重大政治事件以及该事件发生的地点。

- 2011年之后，多位学者用谷歌数据集和Ngram Viewer进行了基于英语及其它语言的语汇分析。参见Alexander M. Petersen, Joel Tenenbaum, Shlomo Havlin & H. Eugene Stanley, “Statistical Laws Governing Fluctuations in Word Use from Word Birth to Word Death”, Scientific Reports, Vol. 2, 2012; Roth, S., Fashionable Functions. “A Google Ngram View of Trends in Functional Differentiation (1800–2000)”, International Journal of Technology and Human Interaction, Vol. 10 No. 2, 2014

近三百年中国城市的国际知名度 基于大数据的描述与回归

社会
2015 · 5
CJS
第 35 卷

陈云松 吴青熹 张 翼

摘 要：本文利用谷歌图书的百万书籍大数据，以中国地级以上（含）城市近300年来英语书籍中出现的词频来展示和分析城市国际知名度的变迁及其特征。研究发现，北京、香港、上海、广州、南京、澳门、天津、台北、重庆和拉萨依次为近300年国际知名度的前十强。在此基础上，本文进一步对基于书籍大数据的国际知名度和媒体提及率进行基于时间序列回归的格兰杰因果检验。研究表明，近代中国大陆城市在国际媒体中的提及率显著影响其国际知名度，而港澳台城市的国际知名度和媒体提及率之间则不存在这种关联。这意味着近代以来大陆城市的国际传播主要通过媒体报道进入西方社会。本文最后总结了中国城市国际知名度获得过程的模式和特征。

关键词：大数据 国际知名度 城市 文化传播

- 另一种从宏观视角观察文化历史的模式是由美国德克萨斯州大学达拉斯分校的马克西米利安·席希（Maximilian Schich）所带领开展的两百年间世界（主要是欧洲和北美大陆）的文化历史“元叙事”（meta-narrative）问题。
- We collected data from Freebase.com (FB) (3), the General Artist Lexicon (AKL) (4–6), and the Getty Union List of Artist Names (ULAN) (7), representing spatiotemporal birth and death information of notable individuals, spanning a time period of more than twomillennia. The data sets are included in the supplementary materials (SM), accompanied by an explanation of their nature and data preparation (8) (tables S1 and S2).

AAAS

Become a Member

Log In

ScienceMag.org

Search

Science

Contents

News

Careers

Journals

Advertisement

SEQ SUCCESS WITH CRISPR SCREENING

Minimize QC with the robust KAPA HyperPrep DNA library prep workflow

Roche

Download here

Supplementary Materials

A network framework of cultural history

Maximilian Schich, Chaoming Song, Yong-Yeol Ahn, Alexander Mirsky, Mauro Martino, Albert-László Barabási, Dirk Helbing

Materials/Methods, Supplementary Text, Tables, Figures, and/or References

Download Supplement

Materials and Methods

Supplementary Text

Figs. S1 to S17

Tables S1 and S2

References

Movie legends S1 and S2

Database captions S1 to S4

Data S1

SchichDataS1_FB.xlsx (Freebase.com)

Data S2

SchichDataS2_AKL.xlsx (Allgemeines Künstlerlexikon)

Data S3

SchichDataS3_ULAN.xlsx (Getty Union List of Artist Names)

Data S4

SchichDataS4_WCEN.xlsx (Winckelmann corpus)

Images, Video, and Other Other Media

Movie S1

European birth-death network dynamics, offering a meta-narrative of cultural history based on the FB dataset, 0 to 2012 CE (67 seconds). The dynamically applied color scheme indicates birth sources (blue) and death attractors (red) in correspondence to Fig. 1C. Individuals in the videos appear as particles gravitating towards their death locations, indicating collective directions of flow. The video is rendered with one frame per year at 30 frames per second. Further characterization of the movie content is given in the text and Fig. 3A.

Movie S2

North American birth-death network dynamics, offering a metanarrative of cultural history based on the FB dataset, 1620 to 2012 CE (13 seconds). As in Movie S1, the dynamically applied color scheme indicates birth sources (blue) and death attractors (red) in correspondence to Fig. 1C. Individuals in the videos appear as particles gravitating towards

MORE FROM SCIENCE

Current Table of Contents

First Release Science Papers

Archive

Collections

Book and Media Reviews

About Science

Mission and Scope

Editors and Advisory Boards

Editorial Policies

Information for Authors

Information for Reviewers

Staff

Contact Us

Sciencehound

Custom Publishing

Awards

Order a Single Issue

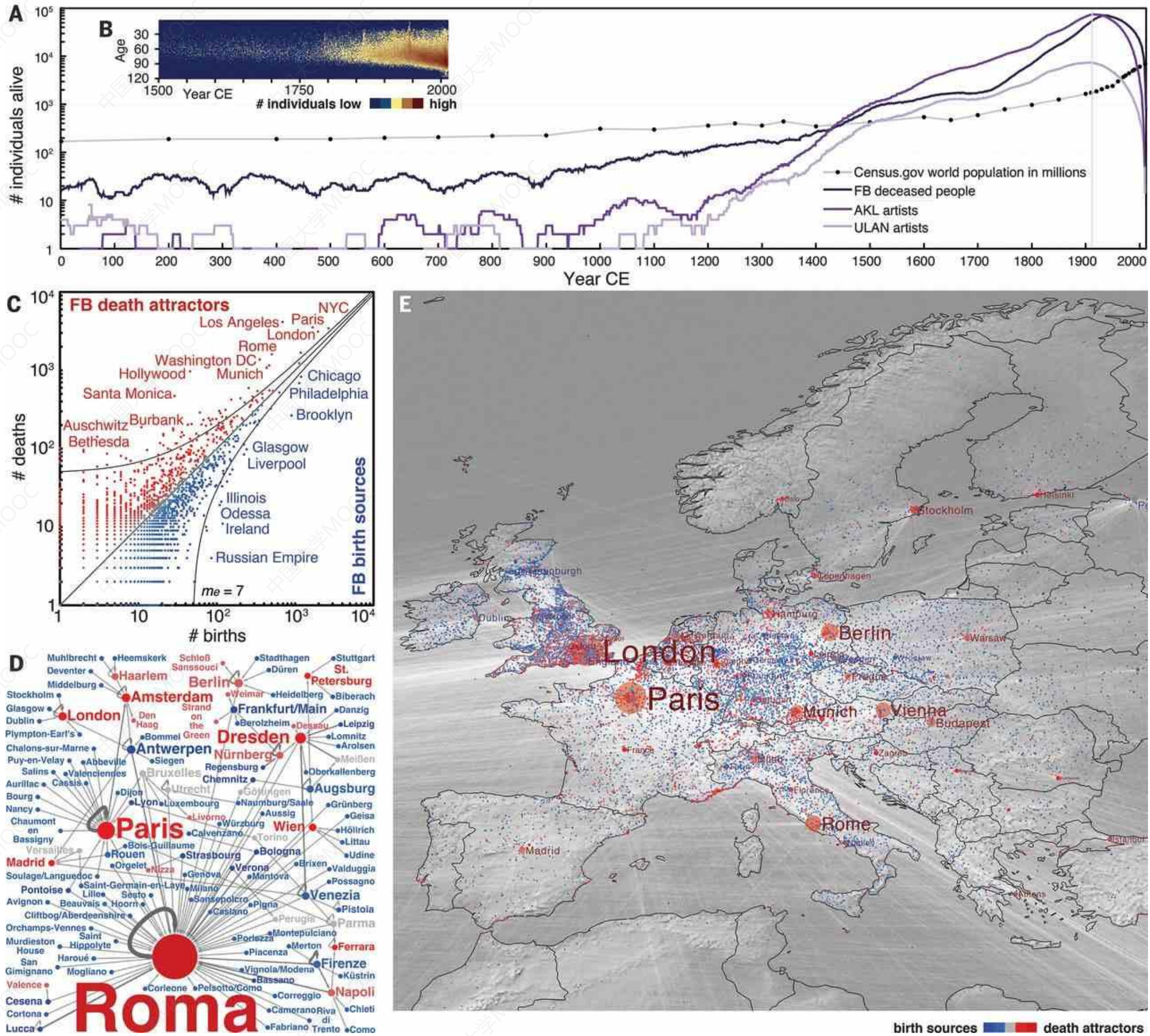
Get the Science eTOC Alert

Submit

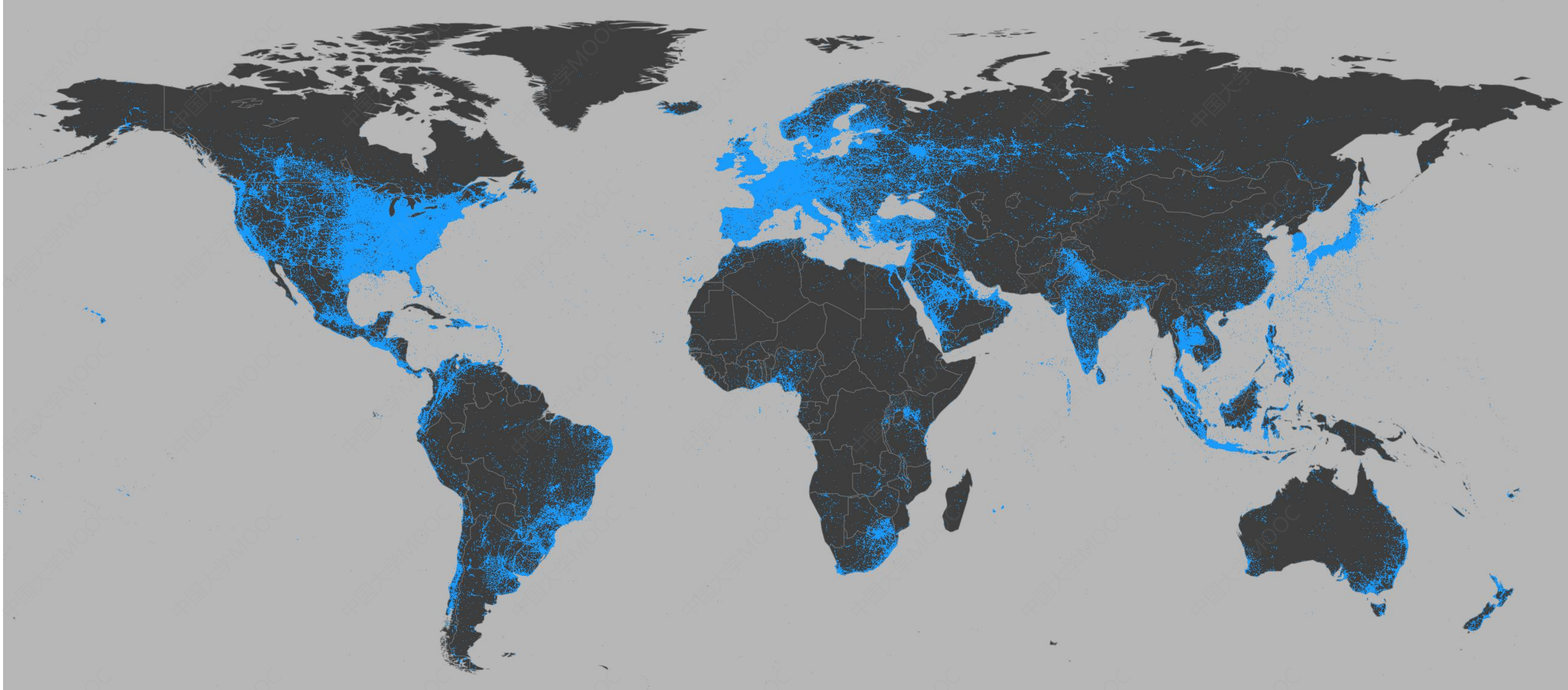
Advertisement

The most important ROOM in the office.

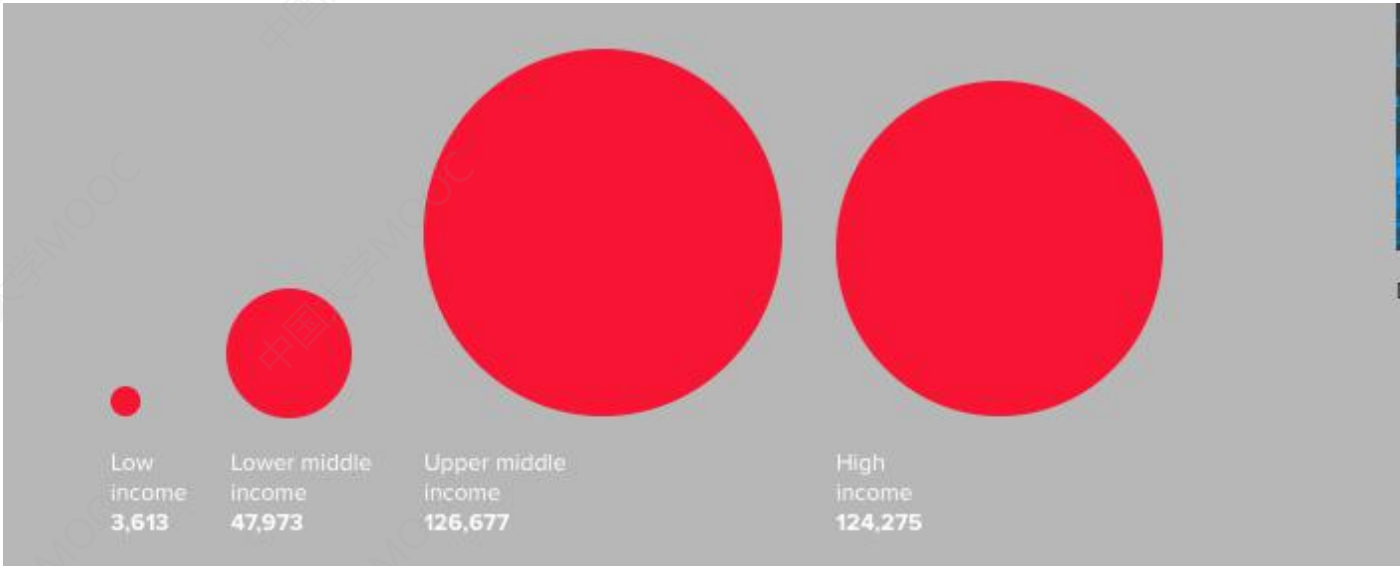
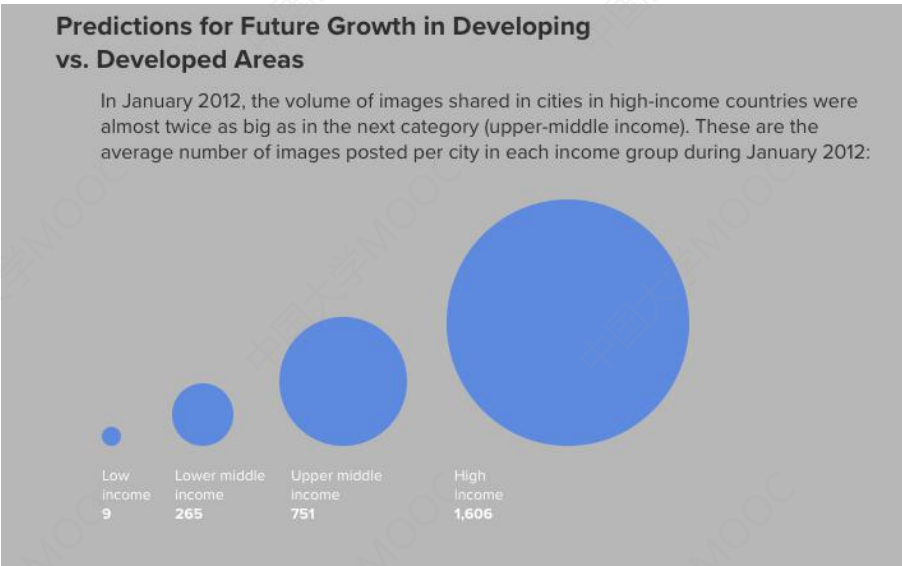
GET A ROOM



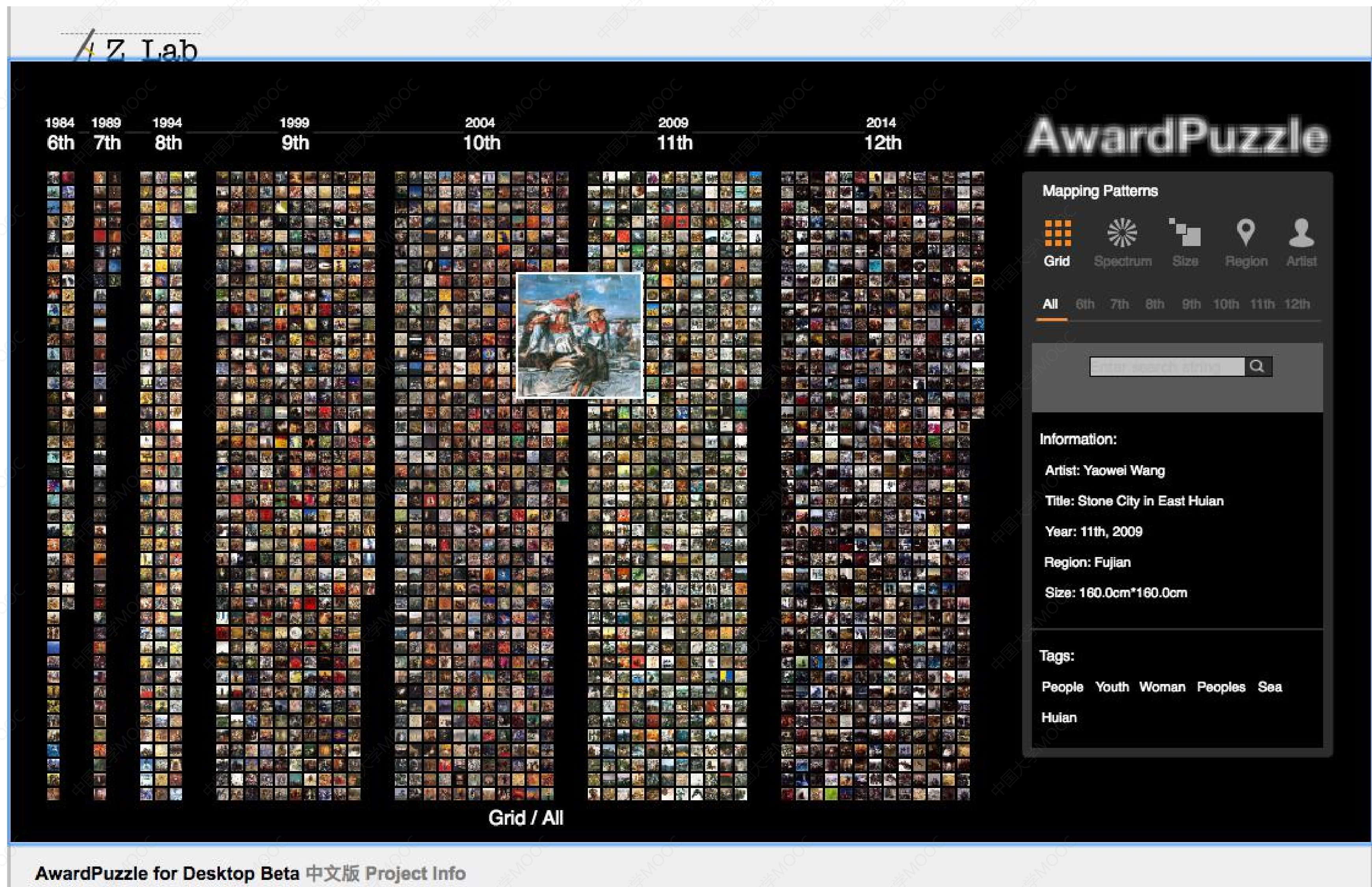
<https://science.sciencemag.org/content/suppl/2014/07/30/345.6196.558.DC1>



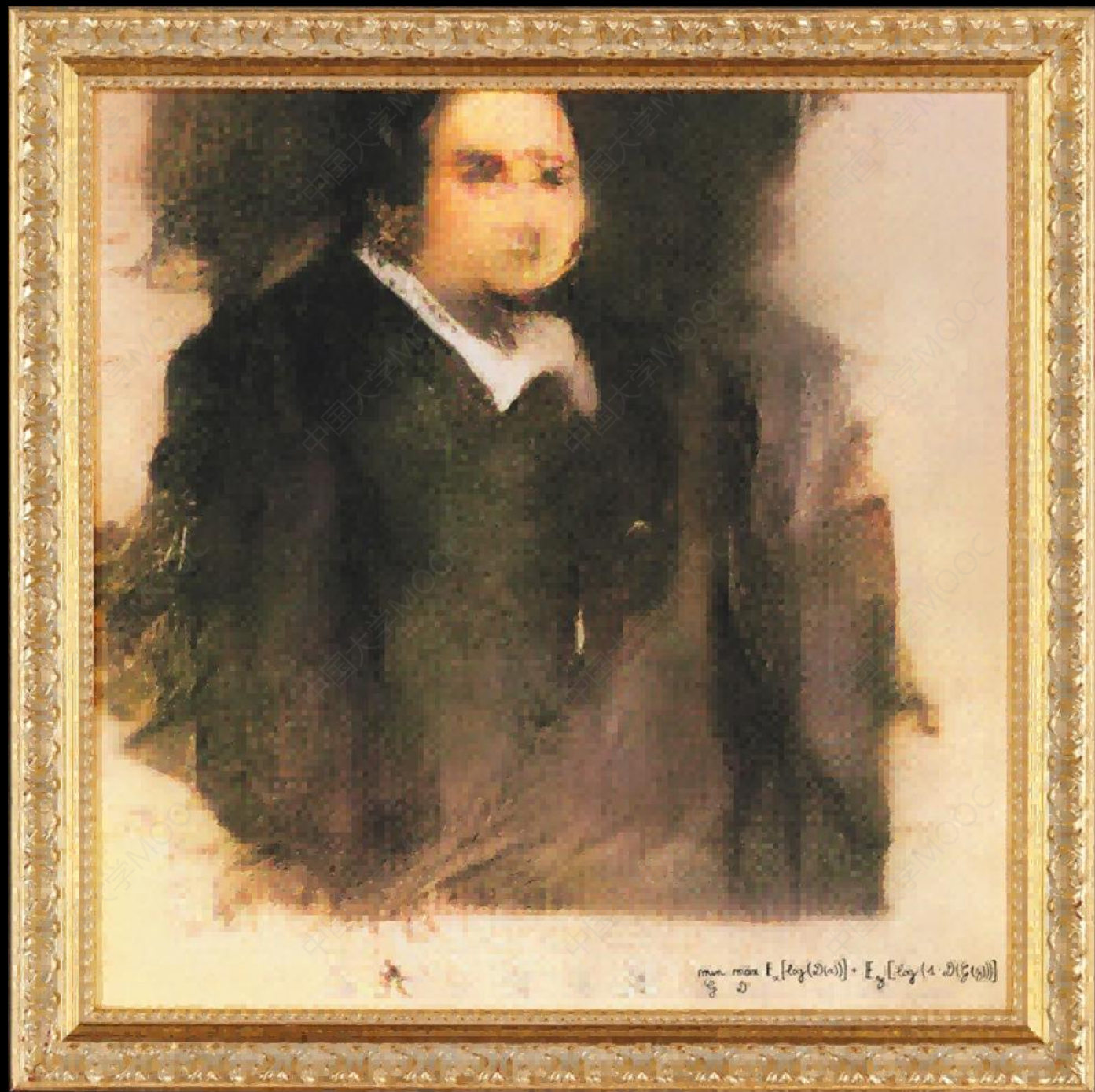
City	Country	Economy	Images shared
London	United Kingdom	High	1,200,958
Tokyo	Japan	High	776,317
Jakarta	Indonesia	Lower middle	722,753
Istanbul	Turkey	Upper middle	690,564
Paris	France	High	609,826
Mexico City	Mexico	Upper middle	553,389
Manila	Philippines	Lower middle	521,027
Bangkok	Thailand	Upper middle	436,941
Barcelona	Spain	High	411,035
Bogota	Colombia	Upper middle	406,964



Visual Earth: <http://manovich.net/index.php/exhibitions/visual-earth>



向帆、朱舜山，“AwardPuzzle”，<http://zeelab.cn/AP/desktopEnglishNew.html>



唐宋文学作品按地区统计				
城市	诗	词	赋	文
北京	150	9	0	8
天津	19	0	0	0
石家庄	6	0	0	6
秦皇岛	0	0	0	0
邯郸	487	3	0	59
邢台	12	1	0	0
保定	12	0	0	2
承德	8	0	0	0
沧州	18	0	0	6
衡水	4	0	0	0
定州	37	3	1	95
太原	22	0	0	105
大同	6	0	0	0
阳泉	1	0	0	0
长治	33	0	0	5
晋城	2	0	0	1
朔州	1	0	0	0
晋中	4	0	0	4
运城	194	0	3	43
忻州	1	0	0	0
临汾	2	0	0	0
吕梁	1	0	0	1
呼和浩特	1	0	0	0
赤峰	7	0	0	3
鄂尔多斯	1	0	0	0
巴彦淖尔	0	0	0	0
朝阳	6	0	0	0
烟台	0	0	0	0
威海	0	0	0	0
青岛	1	0	0	0
济南	1	0	0	3
淄博	1	0	0	0
潍坊	1	0	0	1
济宁	3	0	0	0
德州	1	0	0	0
滨州	104	0	0	43
东营	1	0	0	1
菏泽	1	0	0	0



量化分析唐诗及诗人

2017年3月上线5个月，《唐宋文学编年地图》网站日流量已经突破了700万；

中南民族大学王兆鹏教授带领100多人的团队，用了五年时间打造出来的检索系统，目前已将公元603年-1315年之间的156位诗人的诗文、年谱、生平考订论文等数据录入其中。

数据来源：《全唐诗》《全唐文》的作家小传、《中国文学家大辞典·唐代卷》、年谱及生平考证类论著

采集作家数据：时间层面的生卒年、享年、进士及第年、进士年岁、创作起始年、创作终止年等；空间层面的出生地（籍贯）、定居地、去世地、活动地、贬谪地等；以及身份层面的帝王、宰相、将军、帅臣、幕僚、官员、隐士、布衣、僧道、歌妓等等

类似的研究：台湾元智大学罗凤珠教授的“唐代诗人行吟地图”、台湾中山大学简锦松教授的“在地研究”。

“唐宋文学编年地图”，王兆鹏，来源：搜韵网，<https://souyun.com/poetlifemap.html>，访问时间：2018-01-14

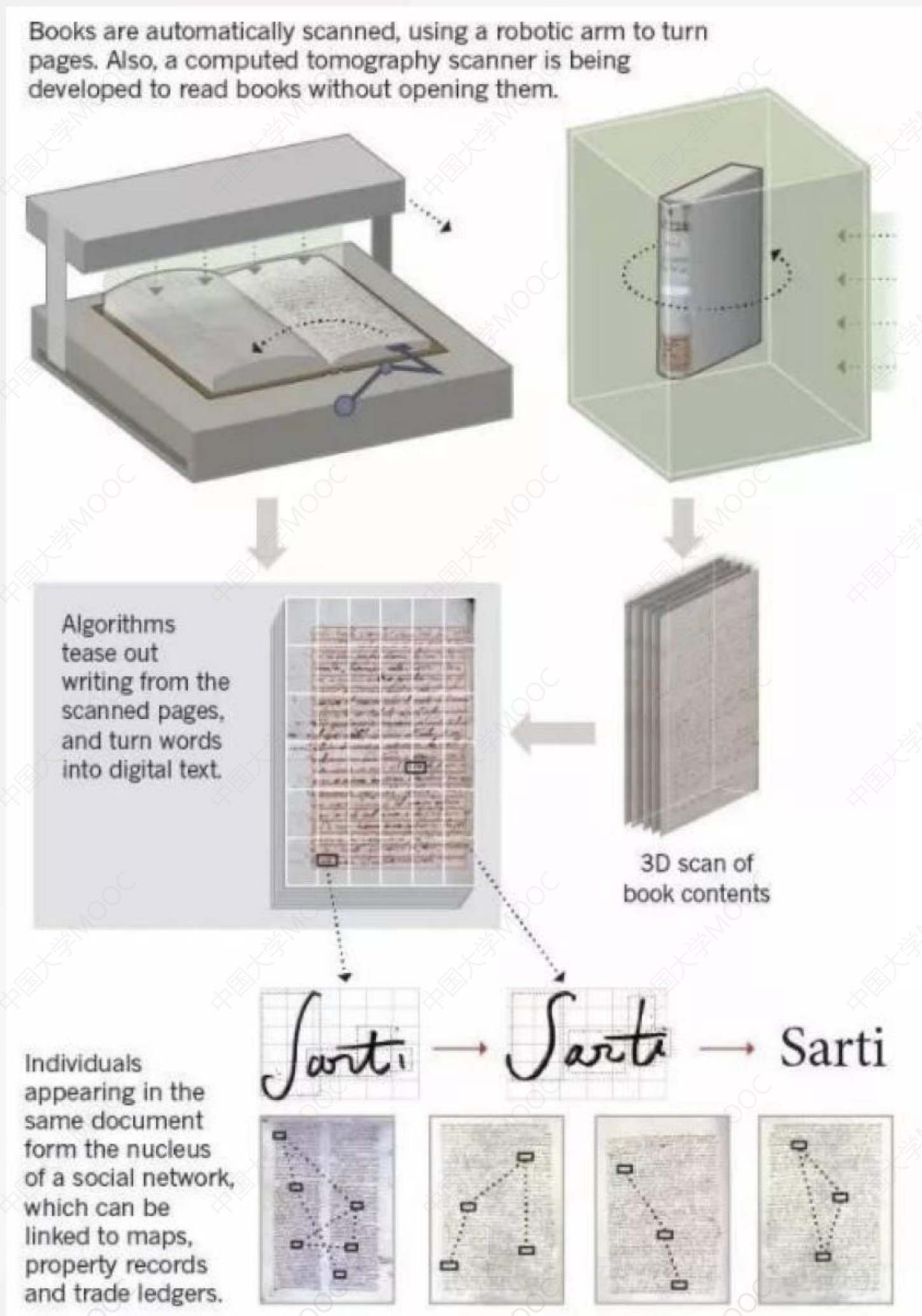
- 今天的未来遗产——Venice/Europe Time Machine



- AI technologies
- vast amounts of information from complex historical data sets.
- the transformation of fragmented data - with content ranging from medieval manuscripts and historical objects to smartphone and satellite images - into useable knowledge for industry
- a large-scale computing and digitisation infrastructure
- cultural heritage investment/an important economic driver across industries
- a constant source of new knowledge/an economic motor

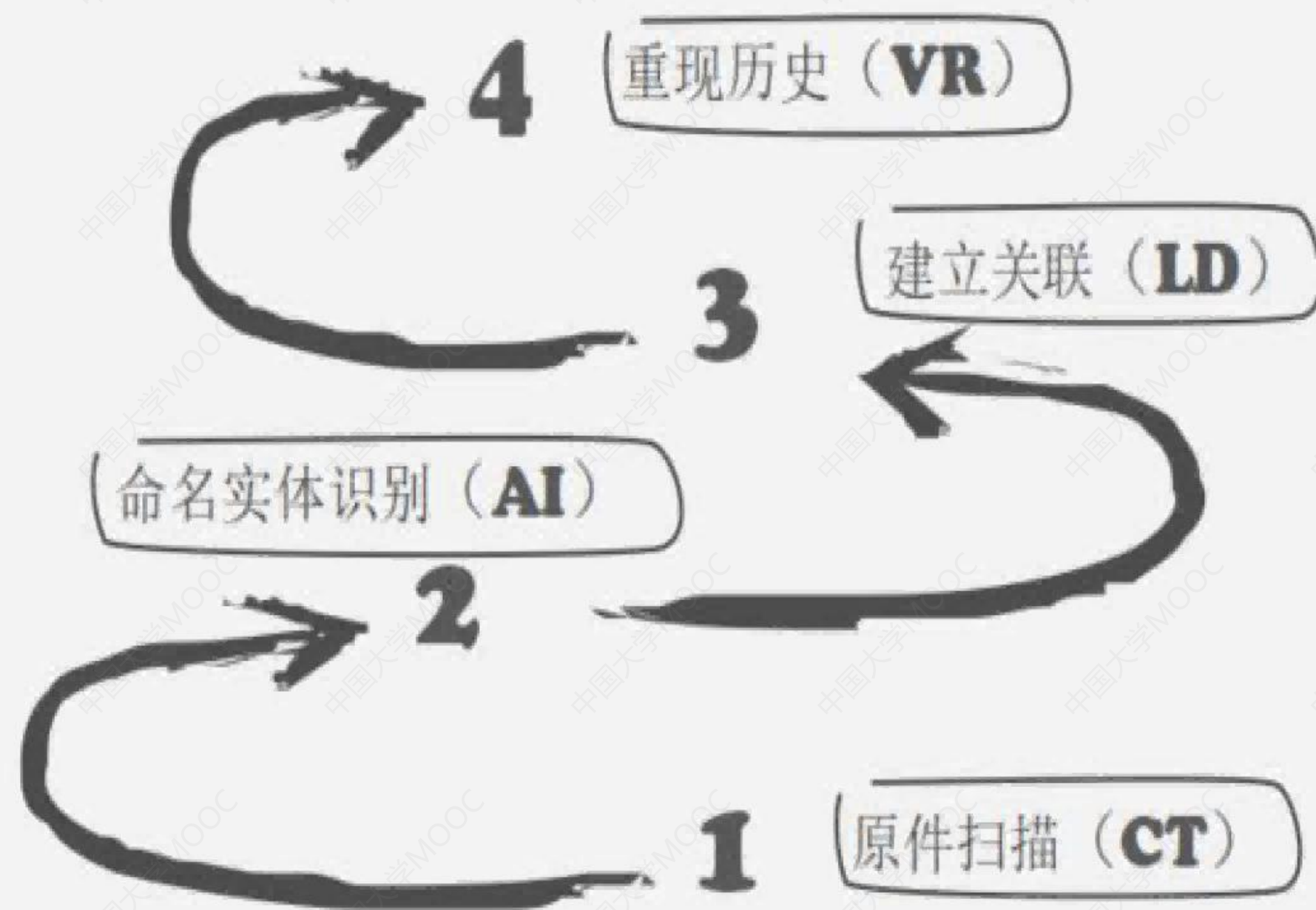
威尼斯时间机器项目

VENICE TIME MACHINE PROJECT



威尼斯时间机器项目

VENICE TIME MACHINE PROJECT



文献：档案、地图、照片、音视频

实体：人、道路、商铺、建筑、空间

关系：实体关系、社会网络关系、时空关系

语义：形式化编码(RDF)机器可理解

模型：领域模型、3D模型、时空模型

数字人文的“数据基础设施”建设



图档博在“数字人文”大潮中的可为和能为



1. 背景

研究范式、知识生产方式、知识交流模式



3. 需求

共建共享、开放互联、知识融通、版本迭代



2. 定义

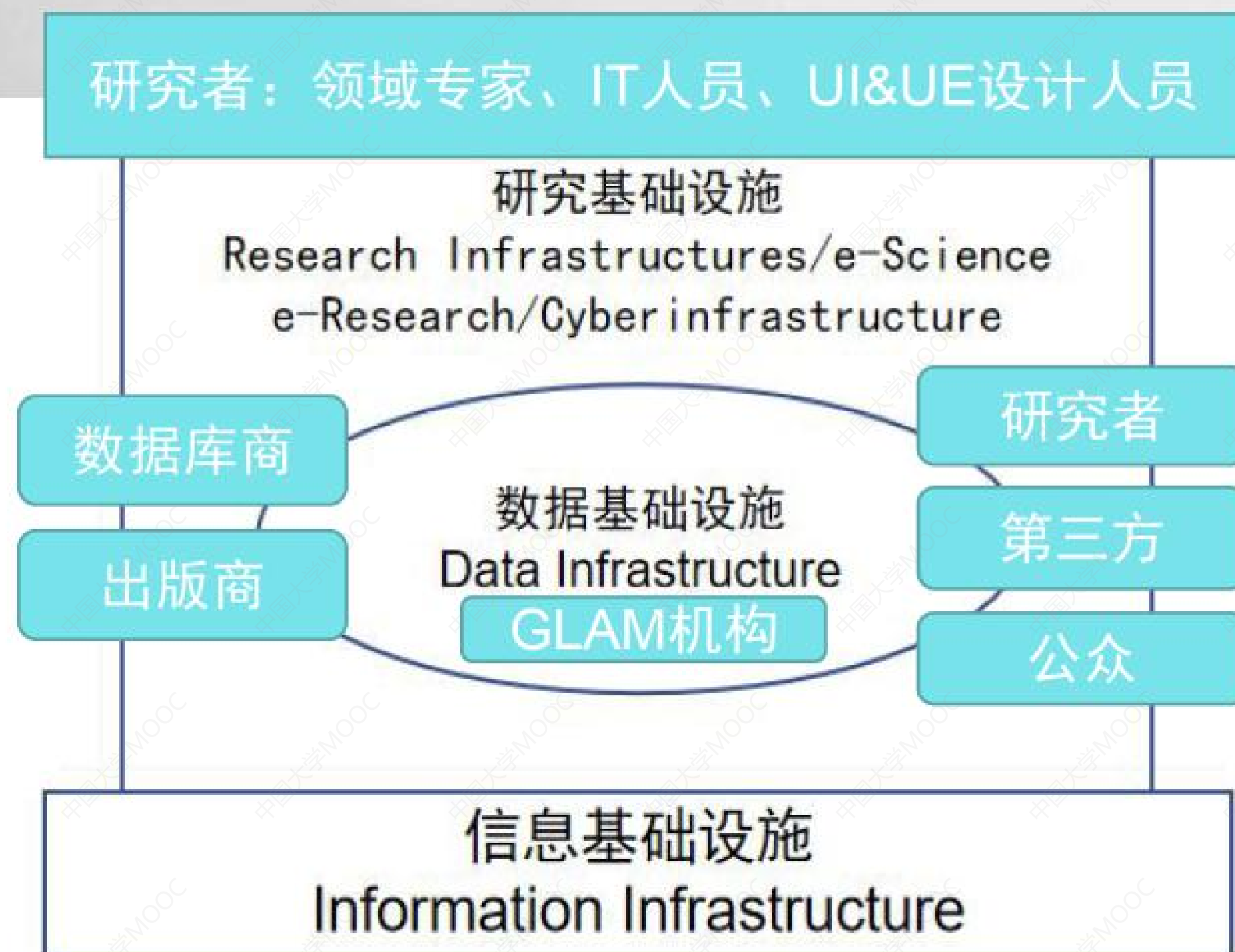
研究基础设施 vs. 数据基础设施



4. 定位

GLAM的公共性、开放性、可持续性
数据基础、方法论基础

数据基础设施



- 以“数据”为基本的组成单位，对于人文研究来说，“数据”可理解为可被机器处理的信息单元，如文献或实物资源对象、概念、术语、实体（人物、机构、团体）或其结构化的描述信息（包括变量、数值、文字符号或事实等）。
- 充分体现数据规模大、覆盖时间长、地域范围广、描述粒度小、维度多等特点，遵循通用的知识组织规范和开放共享规范，具备开放性、公共性和持续性的，独立于具体应用开发和特定领域研究之外、介于信息基础设施“后台” 特定领域研究“前台”之间的“数据中台”。

夏翠娟. 面向历史人文的“数据基础设施”建设
——试论图书馆学的方法论贡献. 中国图书馆学报

上海图书馆支撑“上海年华”的“数据基础设施”

service for people

Rest API for machine

从武康路出发

红色旅游

老电影之旅

上海年华 (Shanghai Memory)

古籍联合目录与循证平台

红色文献库

老电影知识库

家谱知识服务平台

盛宣怀档案库

名人手稿档案库

文献知识库

Open API

上海物质文化遗产名录

上海历史建筑库

影戏院

红色旅游景点

上海历史地理名词表 (地名志)

藏印知识库

避讳字知识库

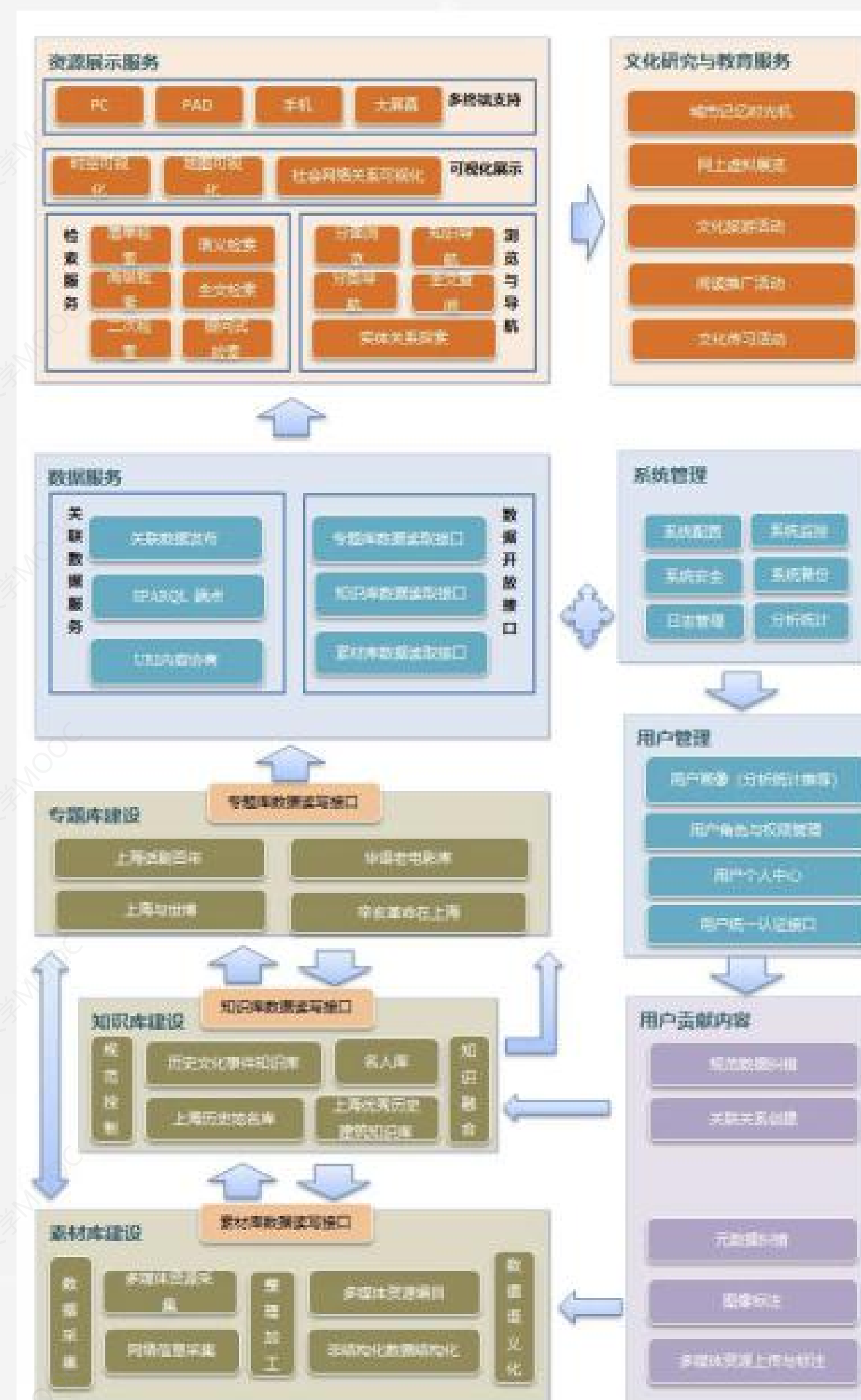
人名规范库

地理名词表

历史纪年表

历史文化事件知识库

基础知识库



- 作业：
- 选择一个你感兴趣的题目，收集/数字化2-3个代表性的文本/图像，对其中的具有的数据进行描述；
- 选择一个课上介绍的数据集/库案例，自行进行学习，完成一个500字的笔记。