



大数据智能概述：大数据、人工智能、自然语言处理

Big Data Intelligence: Big Data, Artificial Intelligence and Natural Language Processing

张华平 副教授 博士

Email: kevinzhang@bit.edu.cn



<http://www.nlpir.org/>

@ICTCLAS张华平博士

大数据搜索与挖掘实验室 (BDSM@BIT)

2018-9

大数据分析与应用/张华平



北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

从Apple密码门开始...



大数据

江大学
OF TECHNOLOGY



从棱镜手机监控看大数据...



大数

工大學

BEIJING INSTITUTE OF TECHNOLOGY

大数据：一切信息都将成为呈堂证供

三峡大坝固若金汤，可以抵挡万年一遇洪水

<http://www.sina.com.cn> 2003年06月01日08:54 金羊网-新快报

三峡大坝可抵御百年一遇特大洪水

2006年05月22日09:02 新华网 我要评论 (35)

字号： T | T

新华社记者王璐、万后德、余国庆、姚力刚三峡报道

凤凰网资讯 > 大陆 > 南方多省遭暴雨袭击 > 正文

三峡集团董事长：三峡大坝防汛标准为千年一遇

进行了最后的蓄水将是安全

2010年07月19日 21:02 新华网 【大 中 小】 【打印】 共有评论4条

遇万年一遇洪峰 三峡大坝仍安全

2011/06/20 00:43 来源：YNET.com 北青网 北京青年报

工程院院士：三峡大坝抵御百年一遇洪水是肯定的

2011-05-26 13:10:40 来源：中国网 有0人参与 手机看新闻 转发到微博 (0)

这里止住奔腾的脚步

浇筑而成的大坝坝顶海拔高程185米，大坝实际浇筑最大高度为100层楼房那么高。正

搜索 浏览器医



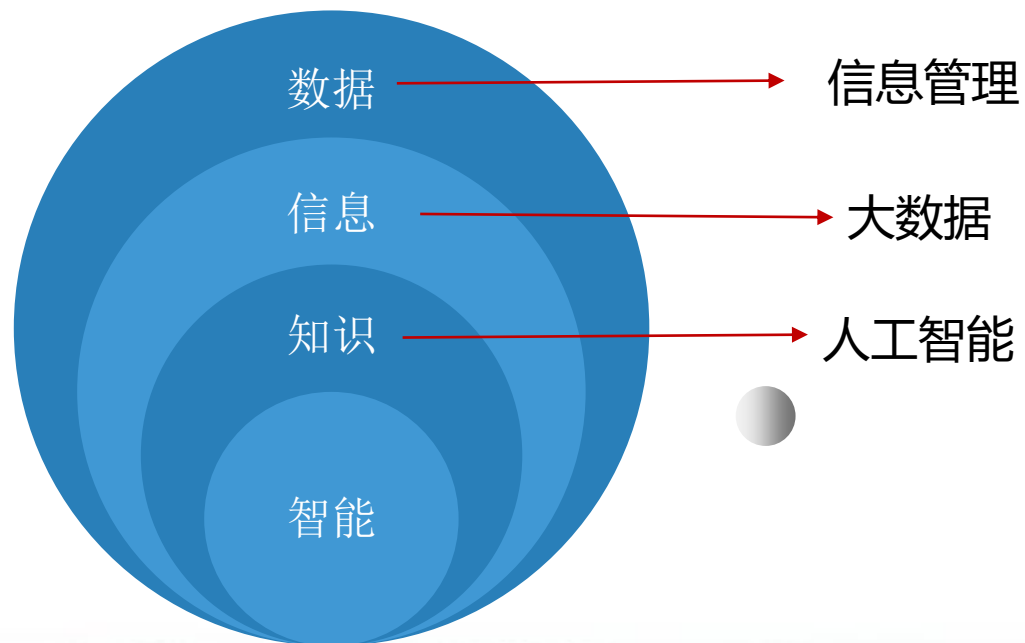
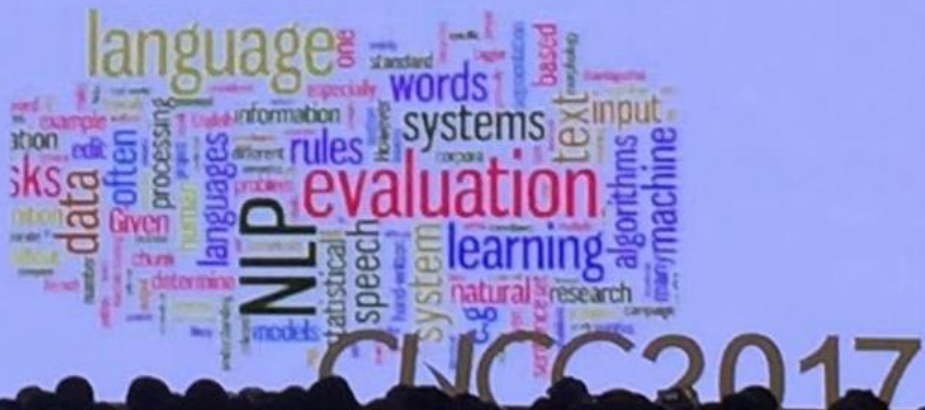
人工智能，现代科学皇冠上的明珠；
自然语言处理则是人工智能皇冠上的明珠。
-周明，ACL主席

人工智能的突破在自然语言理解，懂语言者得天下

-沈向洋，微软全球执行副总裁

在下一个十年...

④ 人工智能的突破在自然语言理解——“懂语言者得天下”



北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY



认识大数据





什么是大数据

- **Wiki:** **Big data** is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications.
- 维克托 《大数据时代》：大数据指不用随机分析法（抽样调查）这样的捷径，而采用所有数据的方法。



➤ 我们的见解:

- 大数据智能是指从客观变化的微观数据中，等技术抽取知识，转
- 是一场新的科技革命数据说话；承认并客代因果推断)



源异构、实时
索、机器学习

全量分析，让
目关性挖掘替



这是显微镜下的世界



这是望远镜中的宇宙



杨达才启示：1+1>>2才是大数据



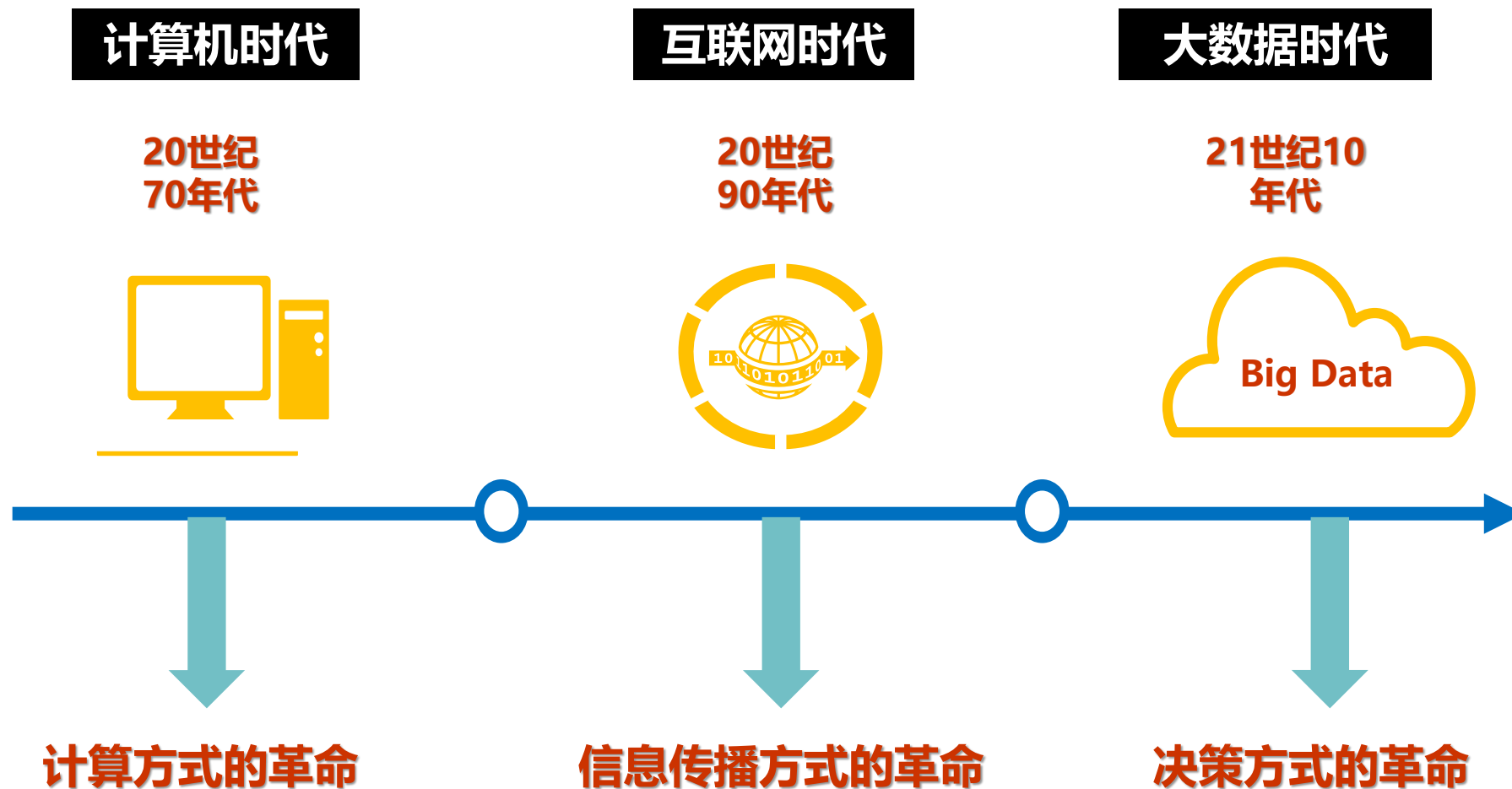
大数据分析与应用/张华平



北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY



近半世纪来的三次革命



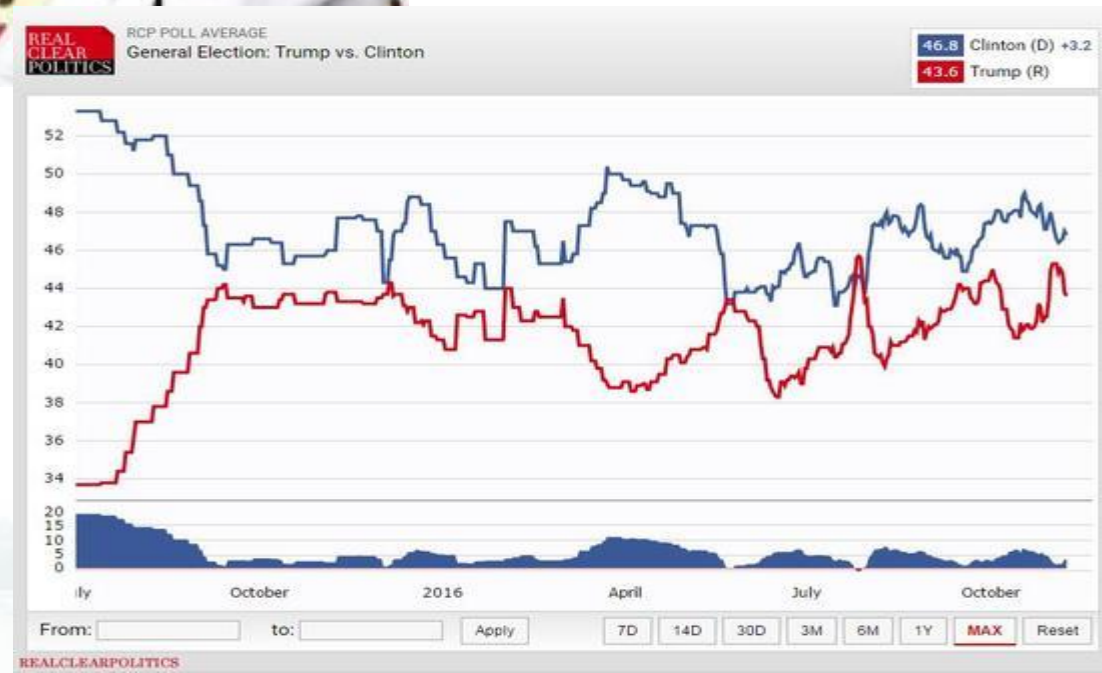
大数据颠覆决策模式



大数据分析与应用/张

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

小数据精英 VS 大数据庶民



大数据分析与应用/张华平



小数据精英 VS 大数据庶民



华平



大學
NOLOGY



大数据时代的特征

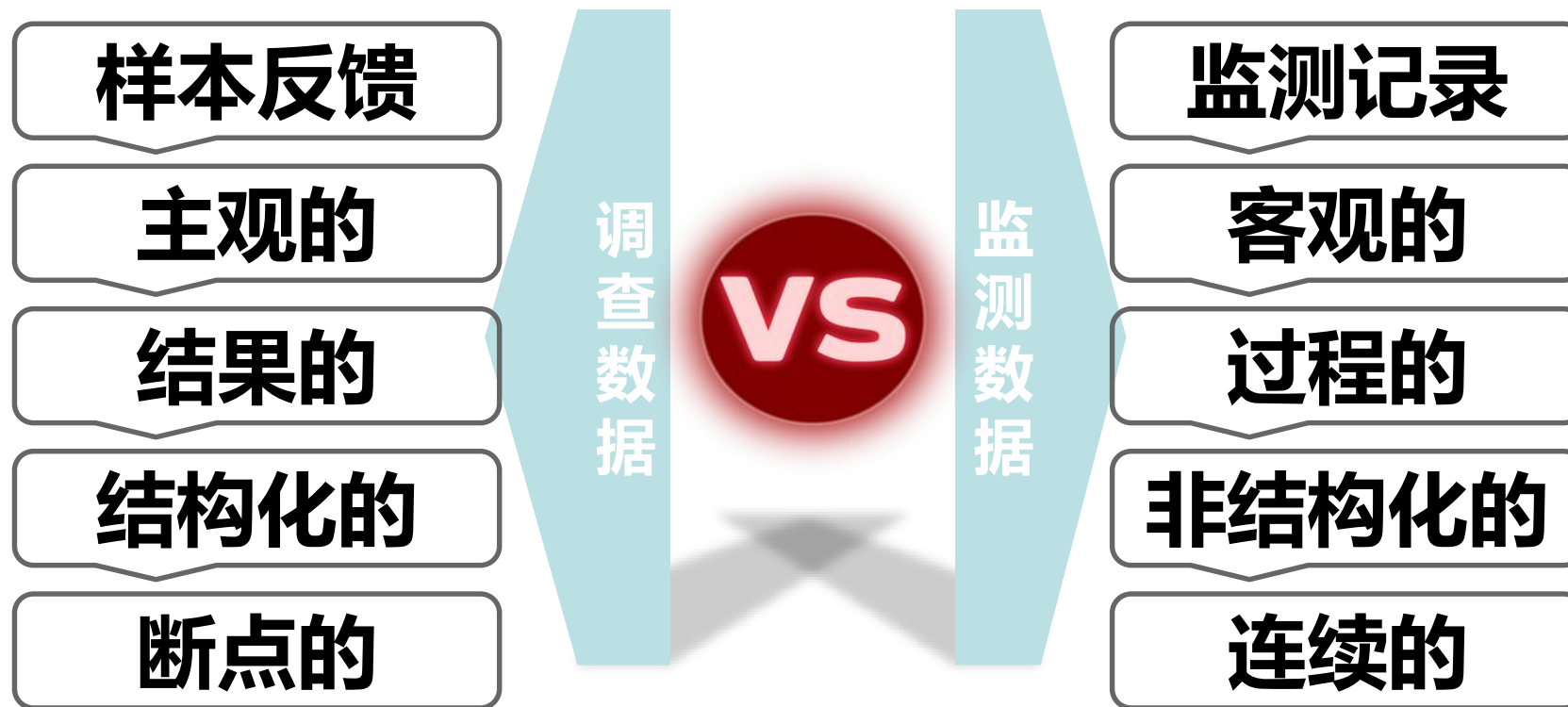


大数据分析与应用/张华平



北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

小数据和大数据的区别





WHAT IS A.I.?

认识人工智能

大数据分析与应用/张华平

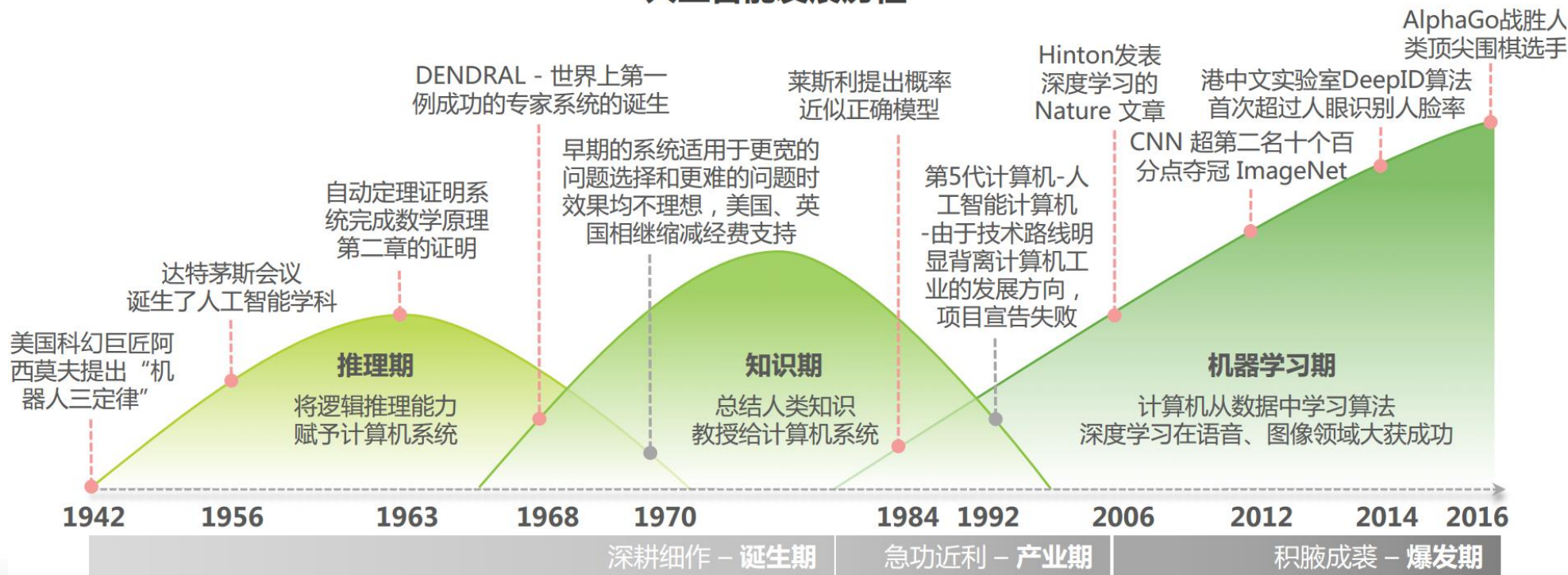


北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

人工智能定义与发展历程

➤ 1956年，约翰麦卡锡(John McCarthy)在达特茅斯会议上首次提出人工智能（Artificial Intelligence: AI）的定义：使一部机器的反应方式像一个人在行动时所依据的智能。

人工智能发展历程



注：图来自于艾瑞收集整理

大数据分析与应用/张华平



北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

人工智能的三个境界

超人工智能

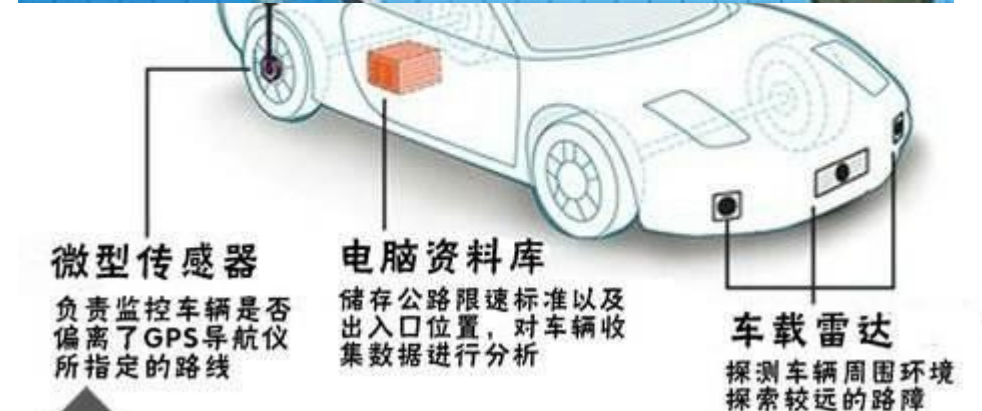
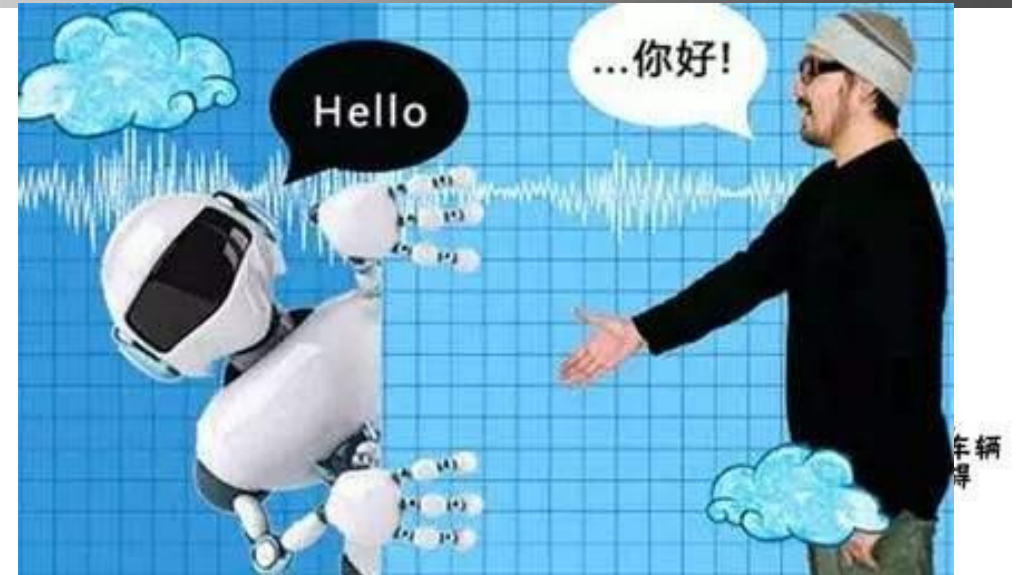
- 认知能力：与人的交流、交互与交融
- 自然语言理解、知识推理

强人工智能

- 感知能力：受限的环境
- 自动驾驶、人脸识别、传感器等

弱人工智能

- 计算能力：人工定义的严格规则
- AI剪枝优化决策，大数据存储与计算



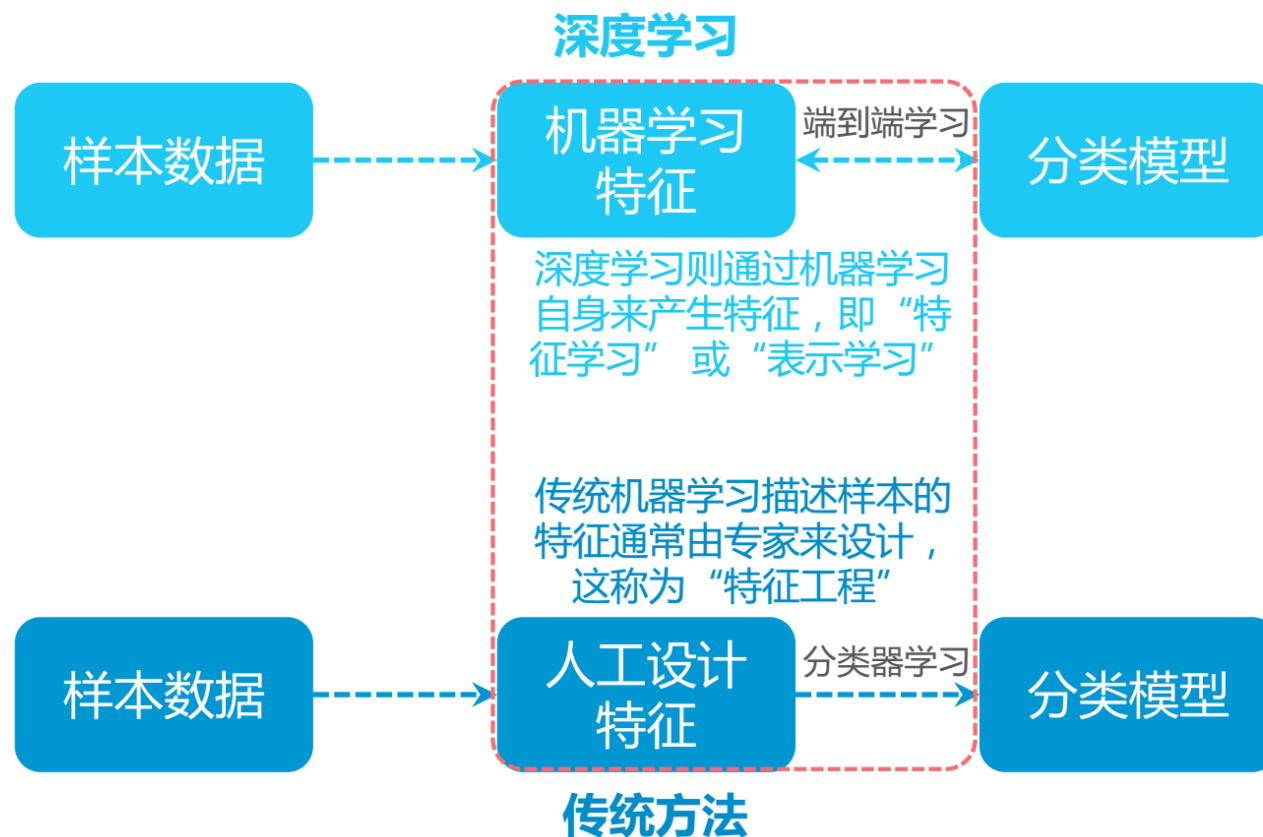


深度学习vs机器学习vs人工智能

深度学习 < 机器学习 < 人工智能



深度学习与传统方法的区别



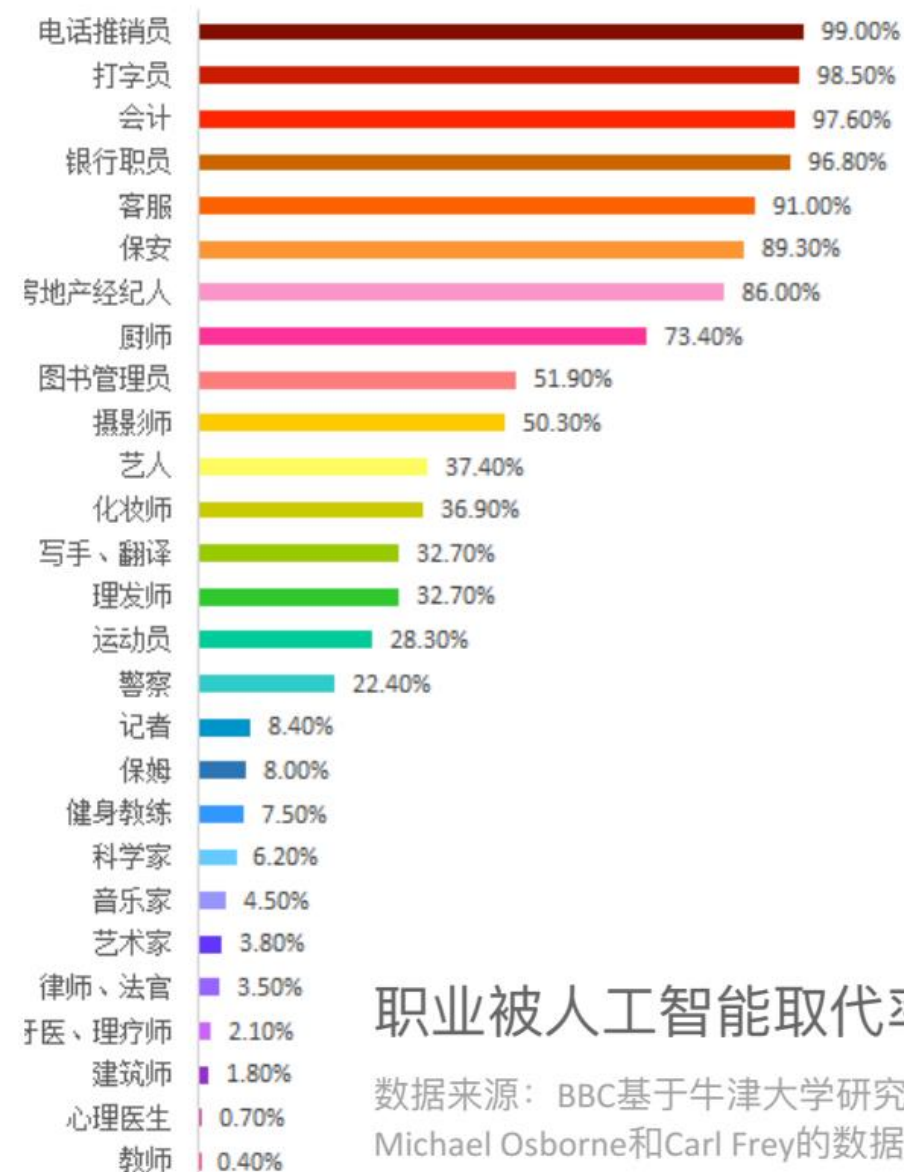
注：图来自于艾瑞收集整理

大数据分析与应用/张华平



北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

AI将取代大量人力工作



职业被人工智能取代率

数据来源：BBC基于牛津大学研究者
Michael Osborne和Carl Frey的数据体系分析



翻译



记者



司机



客服



保姆



助理



保安



交易员

甚至还有...



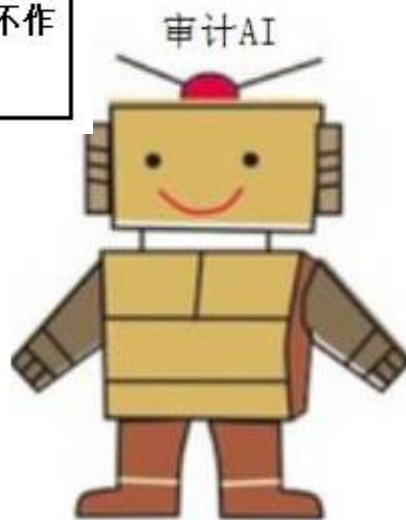
法律咨询师



放射科医生



不行！根据国际会计准则IFRS
XXX, &*%\$#@#\$%&*%\$\$\$~&&*%~
%()(*&&&&&~%\$\$\$%~&&&~*(*(*
((()(*&#@#@!\$\$\$\$@~#@\$\$\$%\$~%~
&~*%&&(*&*%&*%&(&*%())+__+&
~&~%~%\$\$\$#@#@, 所以你要是不作
调整我就会去公安局举报你！



审计猿

➤ 问题1： 审计猿是否将大部分被AI替代？

➤ 问题2：审计工作如何适应人工智能新时代？

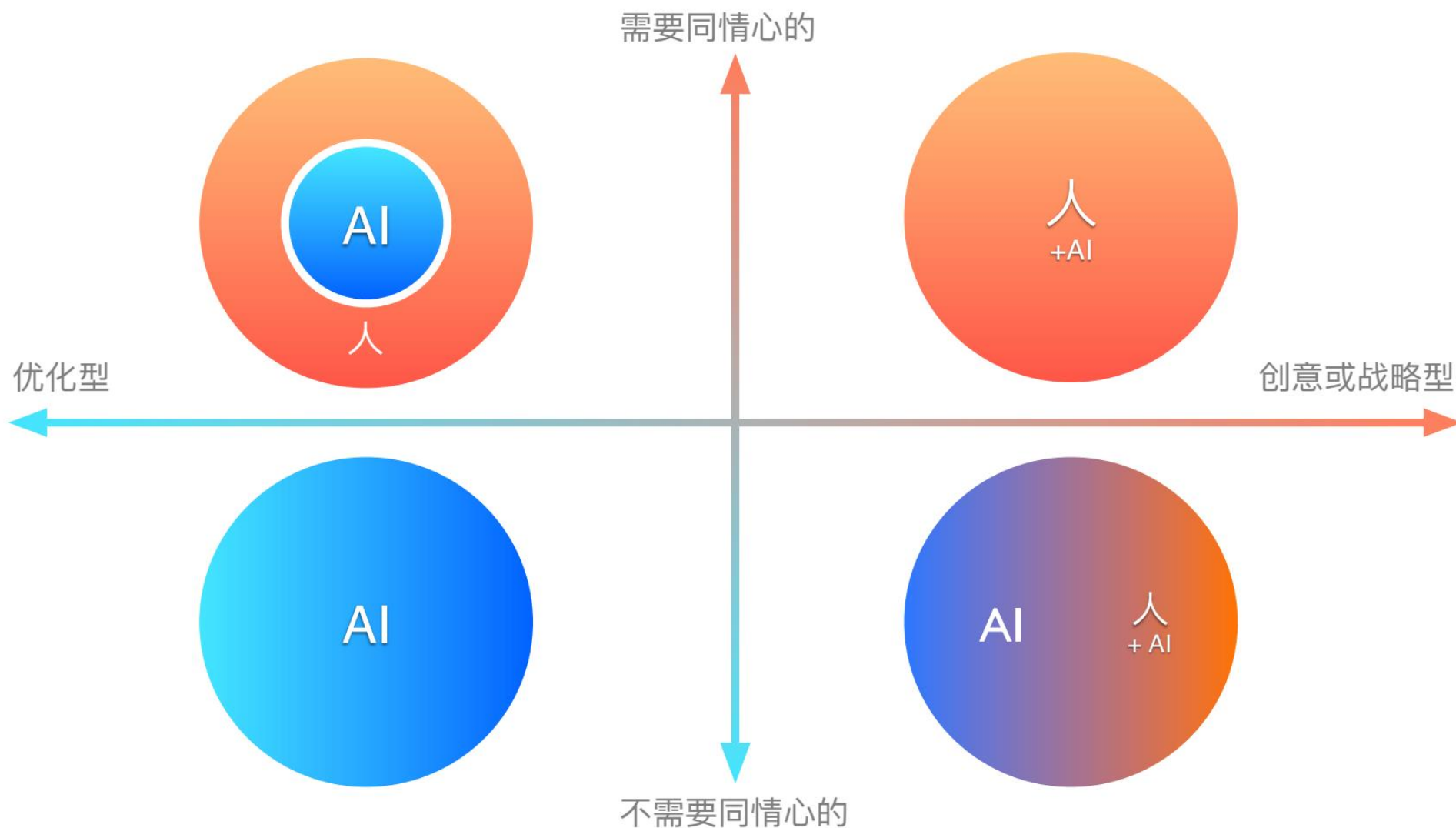
大数据分析与应用/张华平



北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY



人工智能替代性分析

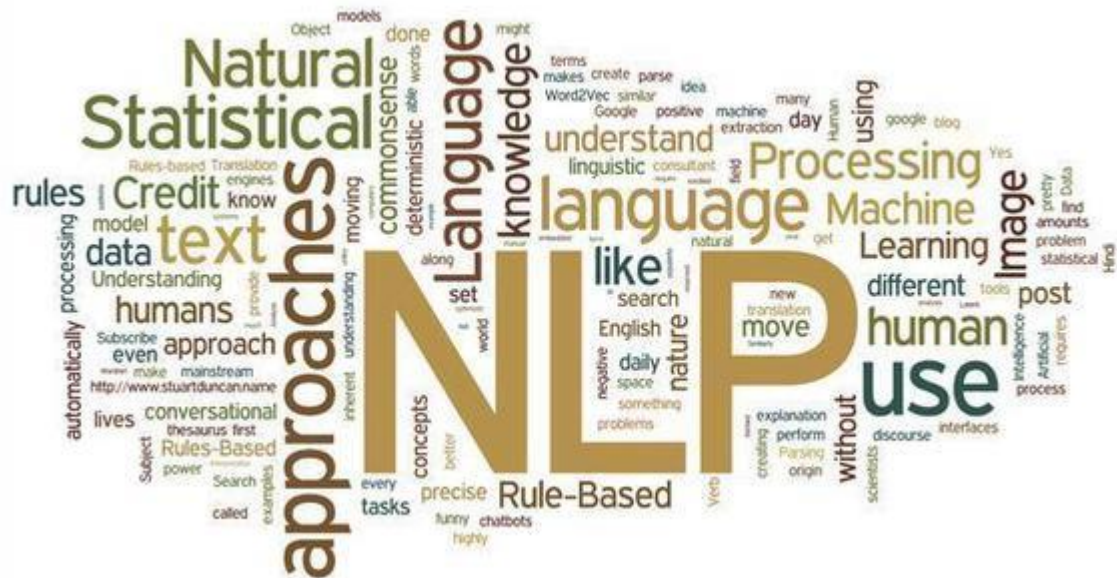




人工智能不能干什么？

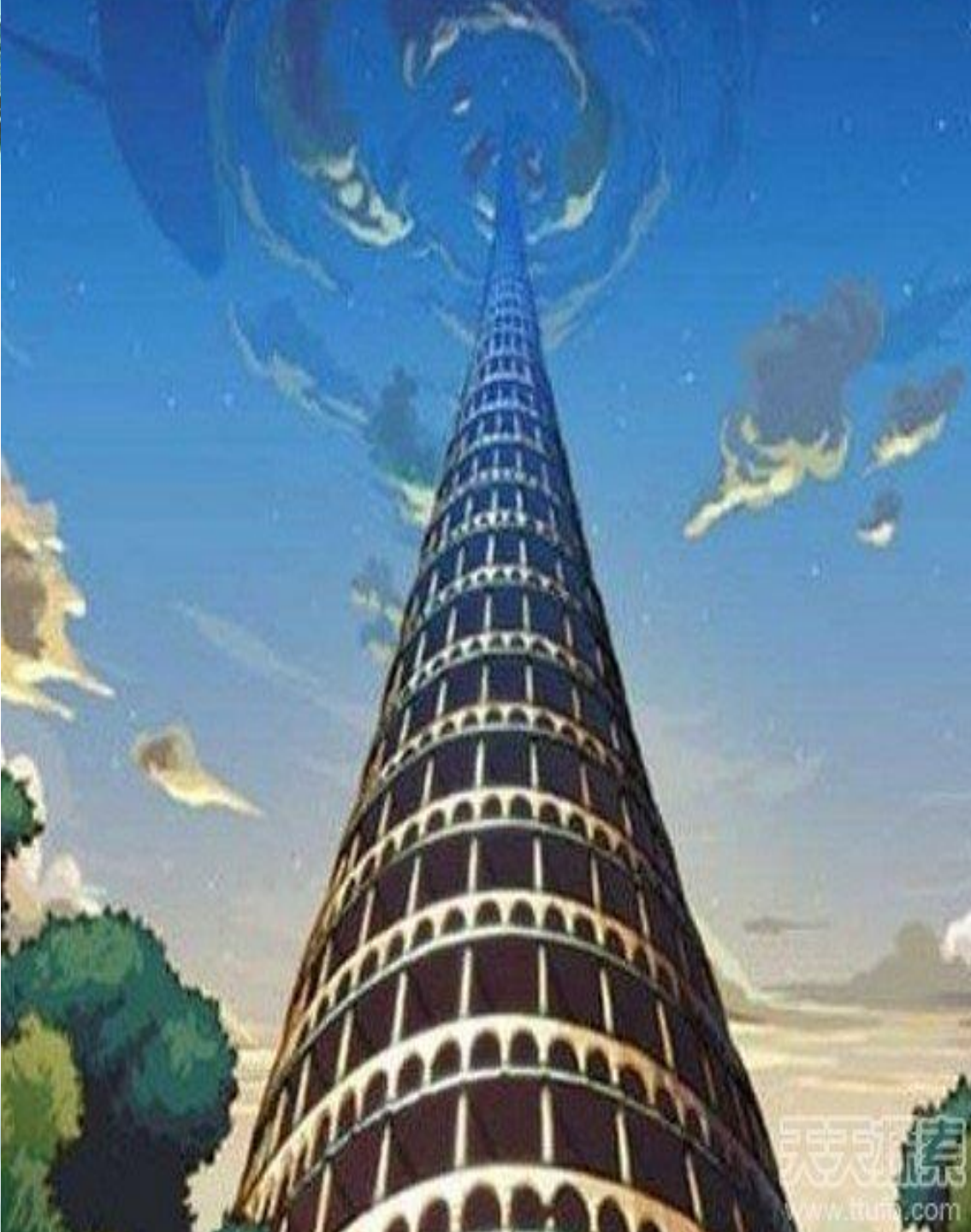
- 跨领域推理
- 抽象能力
- 审美
- 知其然知其所以然
- 常识推断
- 自我意识
- 情感





认识自然语言处理





巴别通天塔之惑

NLP两大挑战:

歧义消解
人类知识

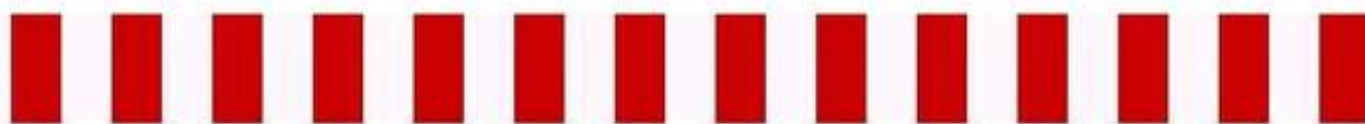


北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

人工智能？

➤ 足球队和乒乓球队：一个谁也打不过，一个谁也打不过。





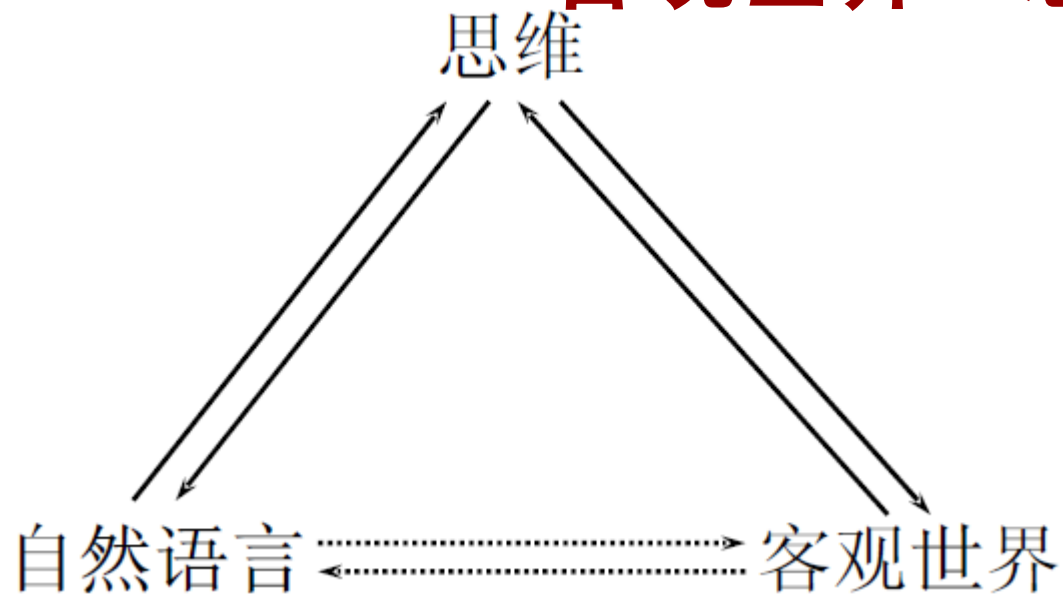
右面的鸡才是最好的



我们行我们当我们要把鸡侧打成右派!!!



客观世界->思维->自然语言



➤ 衰减效应：

- 思维最多只能反映80%的客观世界；
- 自然语言只能反映80%的思维：词不达意，答非所问；
- 听众最多只能听懂80%；
- 听懂的部分只有80%能反映到思维中；
- 分析客观世界的最多只能利用80%。





相关概念

➤ 自然语言处理：技术概念

- is a field of computer science, artificial intelligence and computational linguistics concerned with the interactions between computers and human (natural) languages, and, in particular, concerned with programming computers to fruitfully process large natural language corpora.

➤ 计算语言学：学科概念

- Computational linguistics is an interdisciplinary field concerned with the statistical or rule-based modeling of natural language from a computational perspective.

➤ 文本挖掘：应用概念

- is the process of deriving high-quality information from text.





计算语言学定义

计算语言学是一门以计算为手段对自然语言进行研究和处理的科学。

Computational Linguistics

Natural language processing

"Wherever there is Artificial Intelligence,
there is Artificial Stupidity."

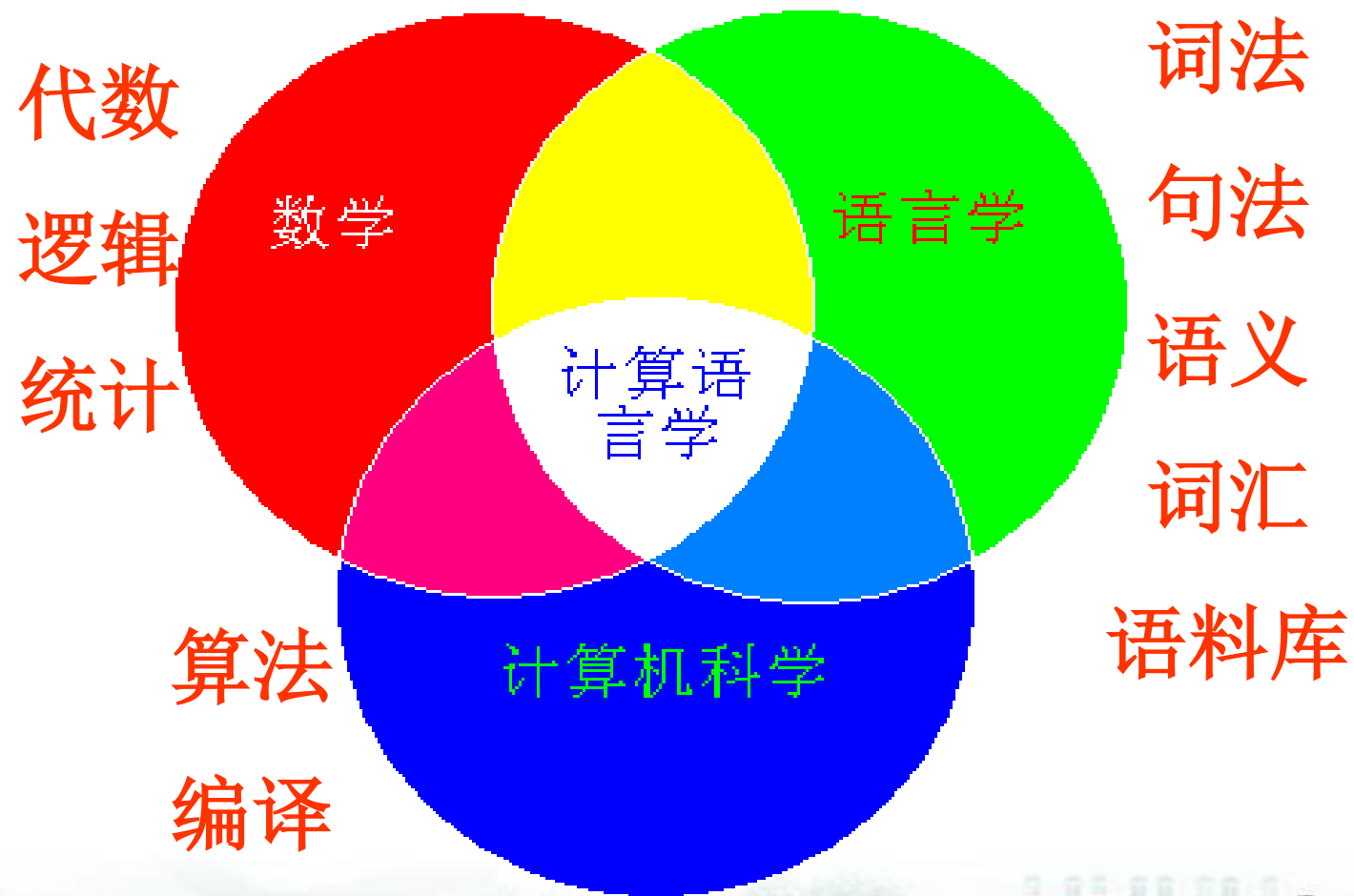
“哪里有人工智能，哪里就有人工愚蠢”。

(从一到万, seed→see+ed)



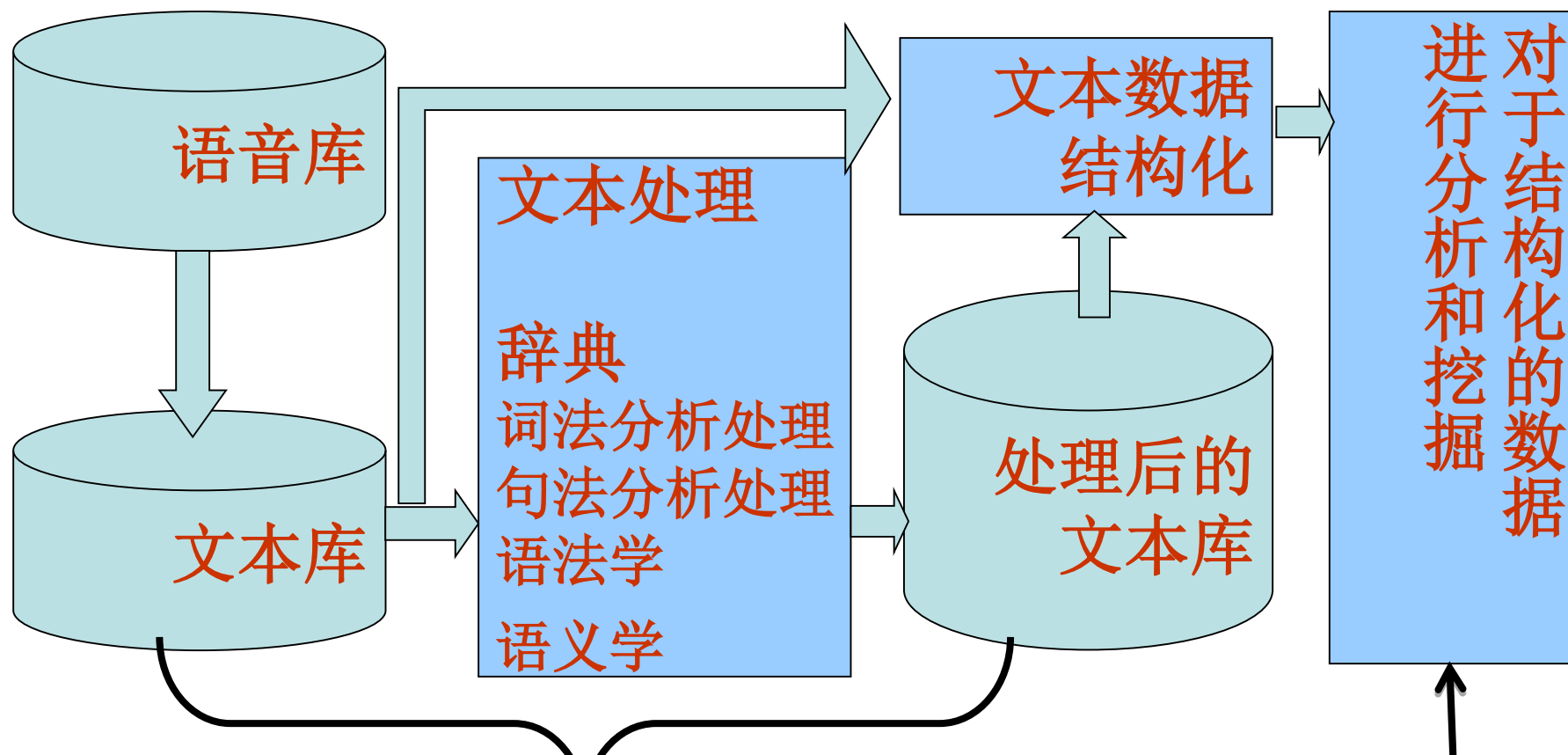


计算语言学与其他学科的关系





文本挖掘的框架



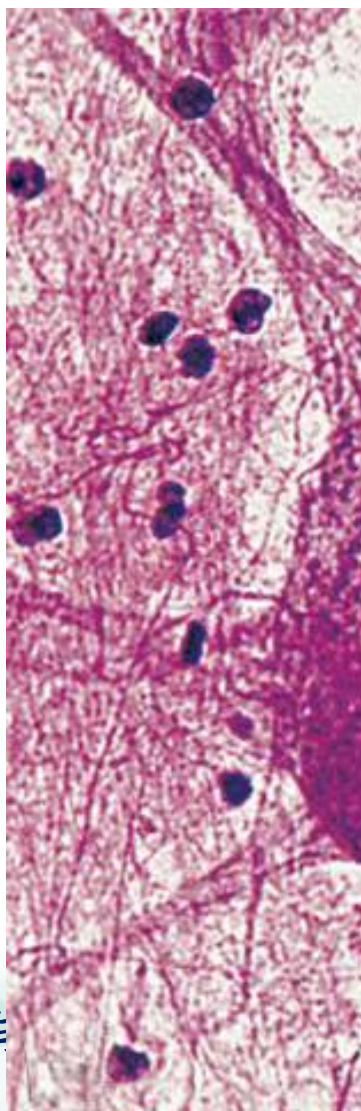
自然语言处理

数据挖掘





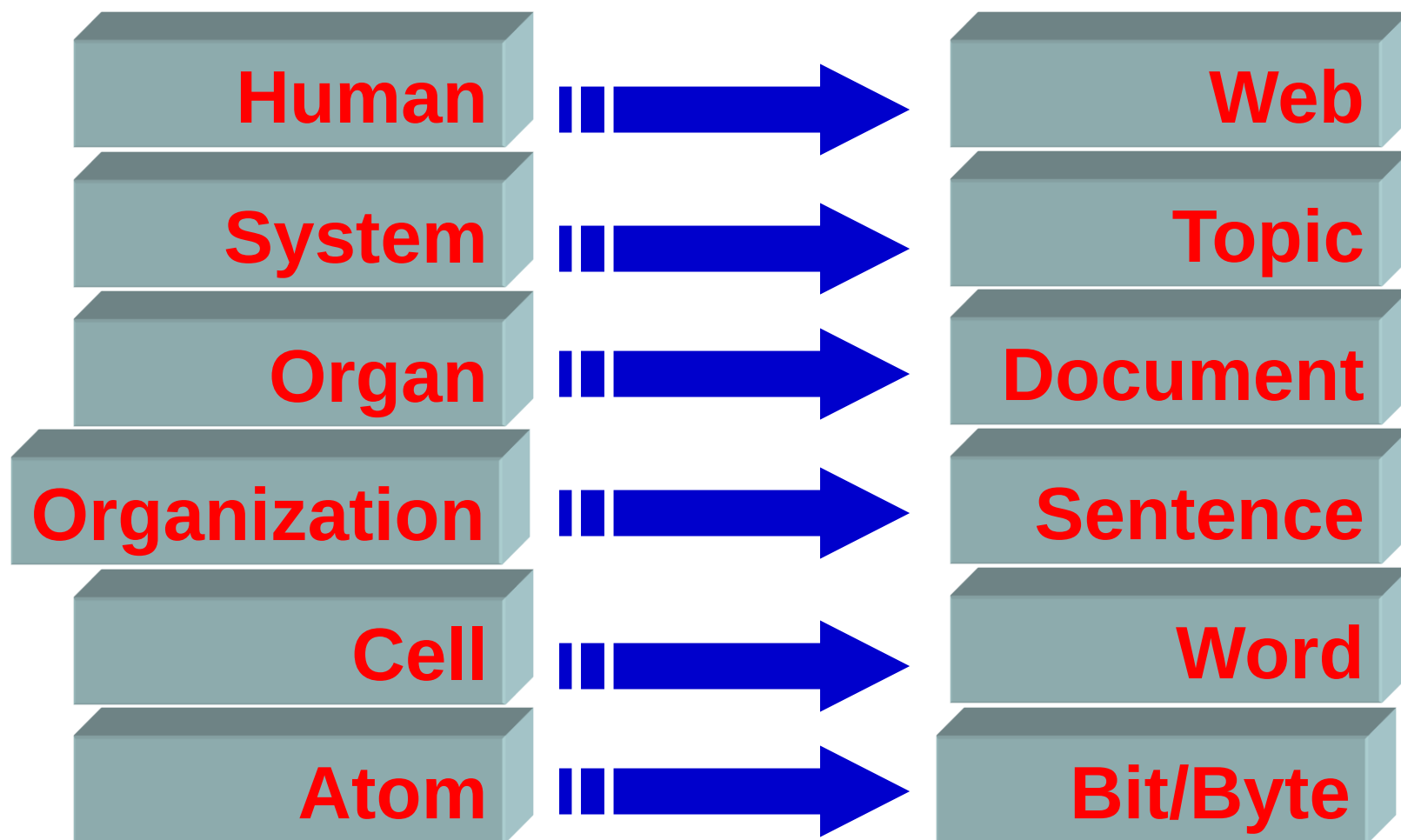
自然语言处理框架



大数据分析



自然语言处理框架



➤ 基础理论

- 自动机 形式逻辑 统计机器学习, 汉语语言学
形式语法理论

➤ 语言资源

- 语料库 词典

➤ 关键技术

- 汉字编码 词法分析 句法分析 语义分析 文本生成 语音识别

➤ 应用系统

- 人机对话, 机器翻译, 社交网络分析, 分类, 聚类,
检索, 过滤, 信息抽取, 音字转换

大数据分析与应用/张华平



北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY



Byte层：字符编码系统

D4 F0 C8 CE BF C9 D2 D4 C8 C3 CE D2 C3 C7 BD AB
CA C2 D7 F6 CD EA D5 FB A3 AC E0 AE BF C9 D2 D4
C8 C3 CE D2 C3 C7 BD AB CA C2 C7 E9 D7 F6 BA C3
A1 A3 5F 5F 5F 5F D4 F0 C8 CE EA CD BO AE 20

粽 禡眈 店蠅蔚岬酖俚涿匚乾禡眈 店蠅蔚
岬 酖疑〔____粽 睿乾

责任可以让我们将事做?瞰 梢匀梦

颐墙 虑樽龔谩?____责任和爱

责任可以让我们将事做完整，爱可以

让我们将事情做好。____责任和爱





字符层面：简繁转换

八皇后问题后来开发成游戏软件了

八皇后問題後來開發成遊戲軟體了

启发：

- 不能简单的按照简体字和繁体字进行简单的映射
- 解决思路：分词，随后按照词为基本单位，进行映射替换





NLP主要算法体系-编码识别为例

➤ 理性主义（规则方法）：Rule-Based

- 自动机 形式逻辑 统计机器学习，汉语语言学
形式语法理论

➤ 经验主义（统计方法）：Statistics-Based Language Model

- 线性：N-Gram; Bayes; Maximum Entropy; Hidden Markov Model; Conditional Radom Fields; Support Vector Machine;
- 网络：神经网络，深度神经网络; CNN; RNN; LSTM;

➤ 两者结合：Hybrid

大数据分析与应用/张华平



北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY



结语：三个论断

➤ **人工智能**：现代科学皇冠上的明珠；

➤ **自然语言处理**：**人工智能**皇冠上的明珠

➤ 机器一思考，人类就发笑；人类一思考，上帝就发笑。



- Level Easy: 数据大与大数据的区别?
- Level Normal: 大数据思维的本质?
- Level Hard: 人工智能的三个层次, 自然语言处理的难点?



感谢关注聆听！



张华平

Email: kevinzhang@bit.edu.cn

微博: @ICTCLAS张华平博士

实验室官网:

<http://www.nlpir.org>



大数据千人会

