中国大数据发展面临的挑战

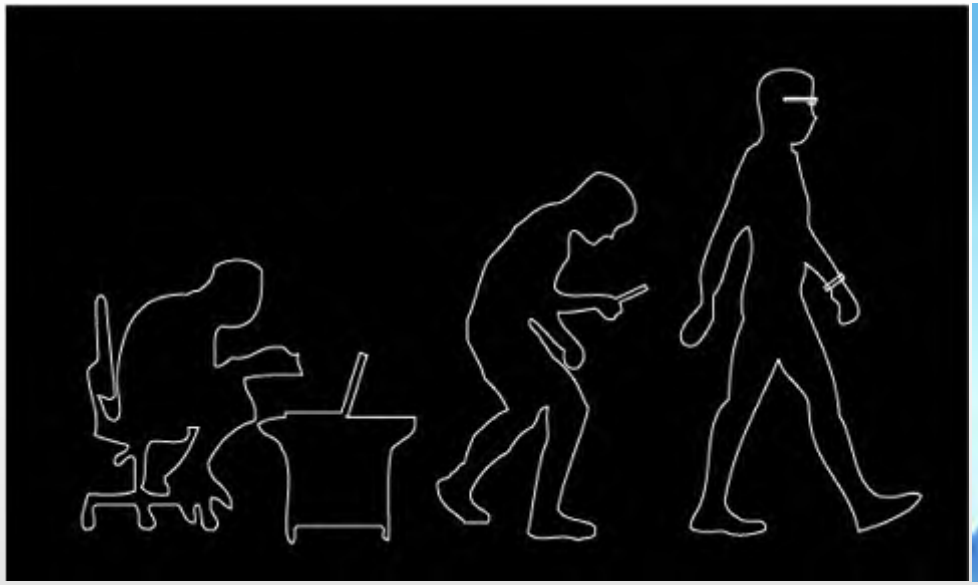
主讲人：沈浩

中国传媒大学新闻学院　　教授 博导
中国传媒大学调查统计研究所　　所长
大数据挖掘与社会计算实验室　　主任
中国市场研究行业协会　　会长

2016

这是一个令人兴奋的时代，也是一个大数据的时代，数据科学让我们越来越多地从数据中观察到人类社会的复杂行为模式。以数据为基础的技术决定着人类的未来，但并非是数据本身改变了我们的世界，起决定作用的是我们对可用知识的增加。

人类行为的93%是可预知的！



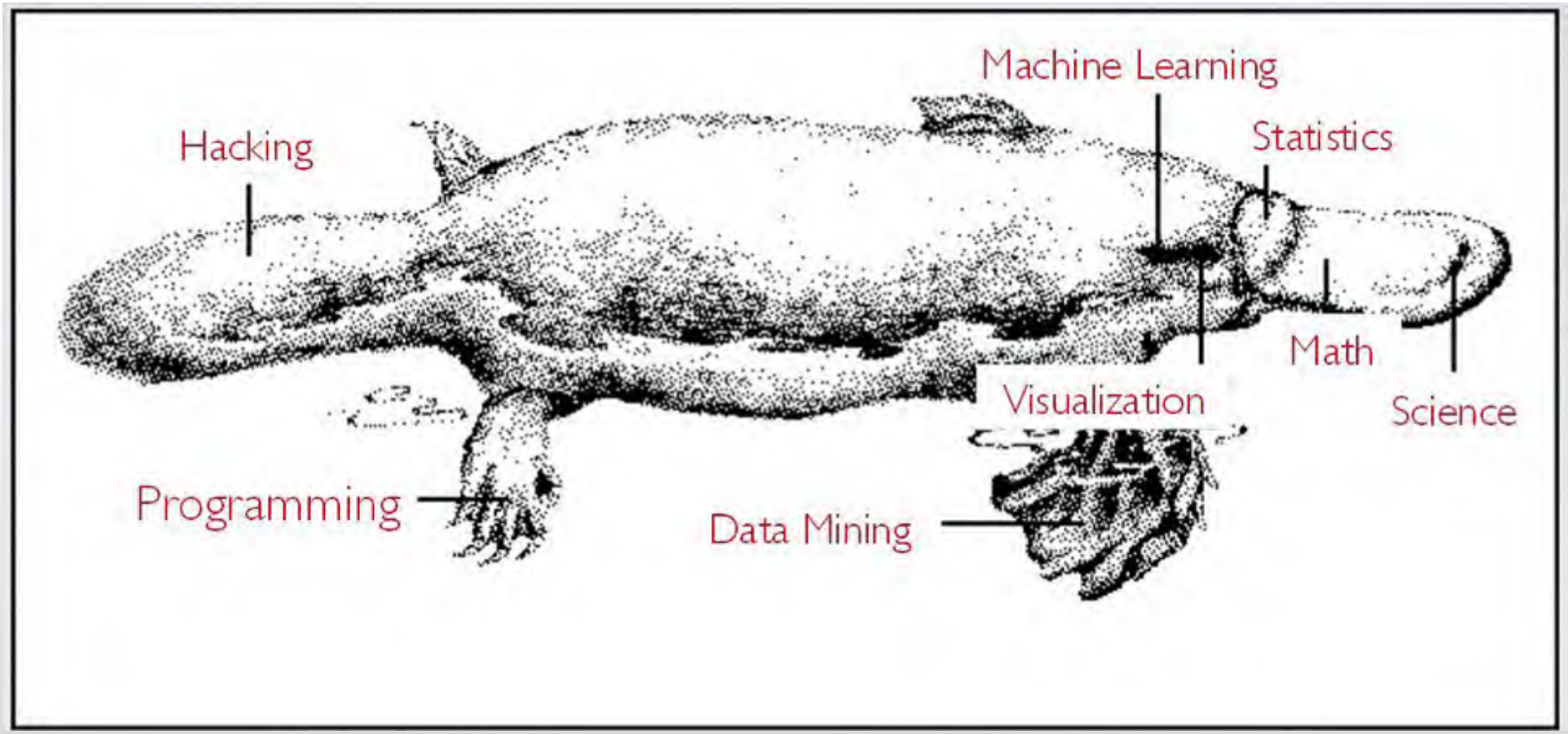
BIG DATA

什么是大数据（Big Data）？它将如何改变我们的社会生活？对政府管理、商业活动、媒介生态、个人生活将产生怎样深刻的影响？我们该如何拥抱大数据？

其实，没有多少人在真正的接触大数据，我们更愿意说这是一个大数据时代，或许是大数据太热了，不同学科背景的人都在谈论，不同行业的都在谈论，说明大数据时代到了，全球已经点燃了大数据时代。

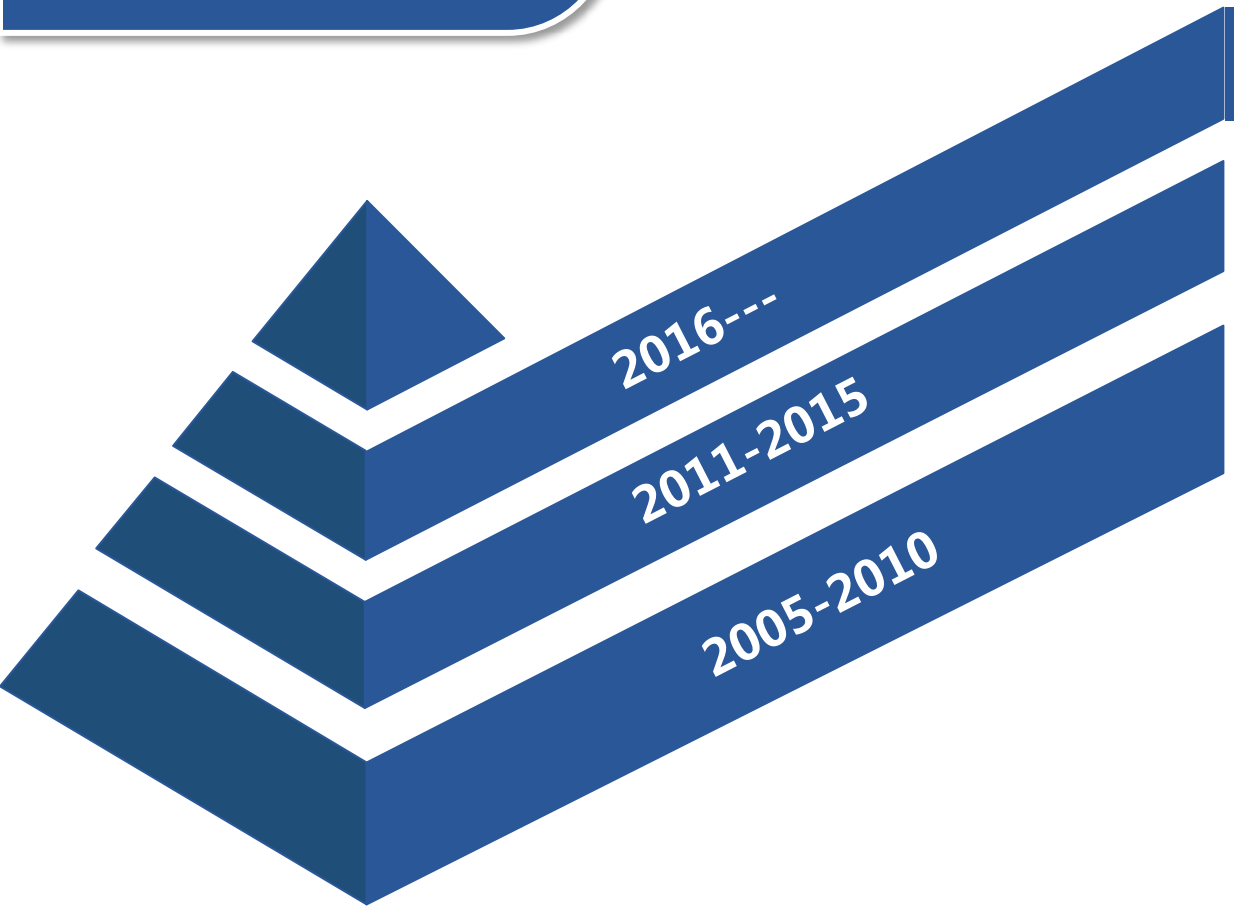
- Annual data creation in zettabytes (10007 bytes)
- 90% of the world's data created in the last 2 years

大数据冰山一角


Machine Learning
Hacking
Statistics
Math
Visualization
Science
Programming
Data Mining


BIG
DATA

大数据时代



大数据的下一个发展阶段？

- 2014，美国白宫：《大数据：抓住机遇、守护价值》
- 2012，Splunk成为第一家上市的大数据处理公司
- 2012，瑞士达沃斯：《大数据，大影响》
- 2011，麦肯锡《大数据：创新、竞争和生产力的下一个新领域》
- 2010，《经济学人》：“数据，无所不在的数据”
- 2009，美国政府：Data.gov（开放数据）
- 2008，计算社区联盟 ：《大数据计算：在商务、科学和社会领域创建革命性突破》
- 2005，Hadoop项目诞生

大数据时代

2009年开始,“大数据”作为中国互联网信息技术行业的流行词汇

2011年12 月，工信部发布物联网十二五规划，提出的海量数据存储、数据挖掘、图像视频智能分析都是大数据的重要组成部分

2012年7月，阿里巴巴集团在管理层设立“首席数据官”一职，负责全面推进“数据分享平台”战略

2013年，媒体称之为“大数据元年”

2016年3月5日，国务院总理李克强政府工作报告，专门提出“促进大数据、云计算、物联网广泛应用。” 这也是自2014年3月5日首次进入政府工作报告以来，大数据连续三年成为我国政府的聚焦点，甚至已被看做“新经济”的高效率引擎。

CONTENT

BIG DATA

更新技术架构

大数据4V

01 Volume(amount)

数据量大：数据量级已从TB（1012字节）发展至PB乃至ZB，可称海量乃至超量。

02 Variety(type&sources)

类型繁多：愈来愈多为网页、图片、视频、图像等半结构化和非结构化数据信息。

03 Velocity(speed)

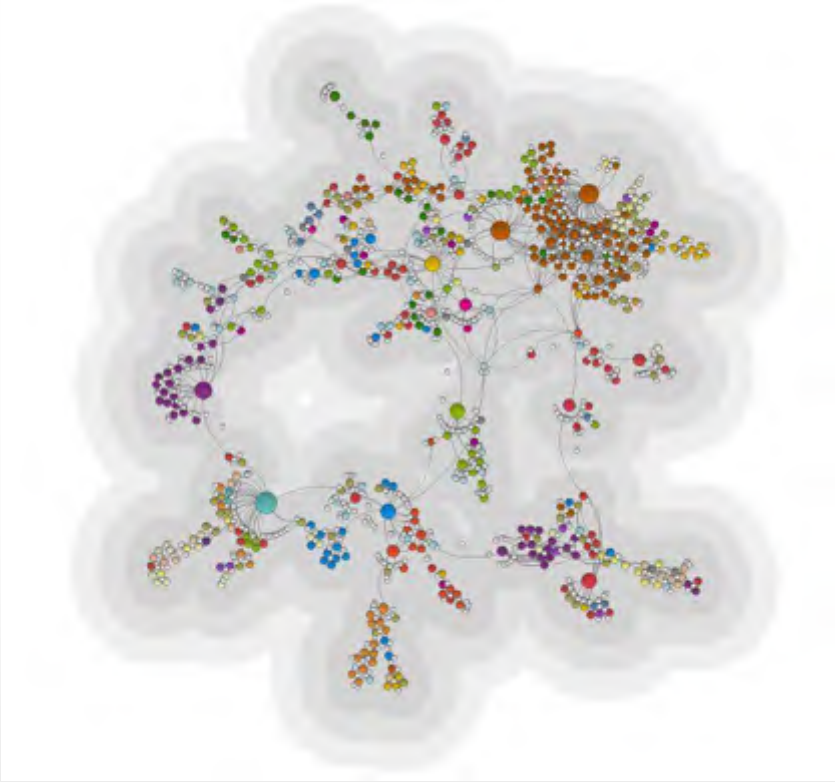
速度快时效高：大数据时代对其时效性要求很高。“1秒定律”：即要在秒级时间范围内给出分析结果。

04 Veracity(quality&trust)

价值密度的高低与数据总量的大小成反比。如何通过强大的机器算法更迅速地完成数据的价值“提纯”成为目前大数据背景下亟待解决的难题。

Visualization

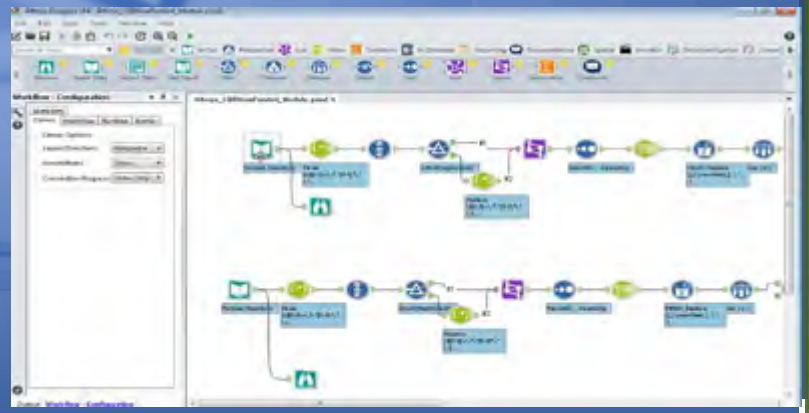
Visualization



让数据模式改变你的心理模式

数据科学 Data Science

- 数据挖掘
- 文本挖掘
- 意见挖掘
- 情感挖掘
- 语义挖掘



空间地理科学 Spatial Science

- GIS分析
- 空间地理信息匹配算法
- 城市规划
- 交通与智慧城市

网络科学 Network Science

- 复杂网络
- 社会网络分析
- 传播网络

可视化技术 Data Visualization

- 数据可视化
- 信息可视化
- 交互可视化
- 大屏实时展现

大数据处理中的特点

1）数据来自商业自动化

2）海量记录g、t、p级

3）数据维度和字典复杂

4）数据呈现时间维度和序列增长

5）商业属性特征向关系数据转换

6）数据更新实时，要求运算快速

7）支持实时和在线分析

8）数据非结构化、半结构化特征明显

9）公开易获得

因此，挖掘大数据价值需要技术更新

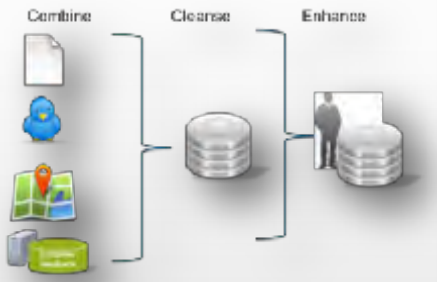
- 分布式文件系统和并行计算框架
- 复杂算法的高效迭代运算能力
- 数据结构从关系型到非关系型发生改变
- 存储、计算、应用一切尽在云端
- 非结构化和关系数据的分析方法发展
- 机器学习和人工智能大行其道
- 空间地理分析成为大数据分析重点
- 数据可视化重要性上升到极大高度

大数据技术更新

All Relevant Data

Enrich

Integrate



Analyze

Integrate any data source

Spark Streaming

Packaged Market & Customer Data & API & Census

App & Data

Rapid design of predictive analytics with unique spatial understanding

大数据技术更新

Visualization

Data, business intelligence, cloud

Big Data

Machine Learning, blending, extract-transform-load

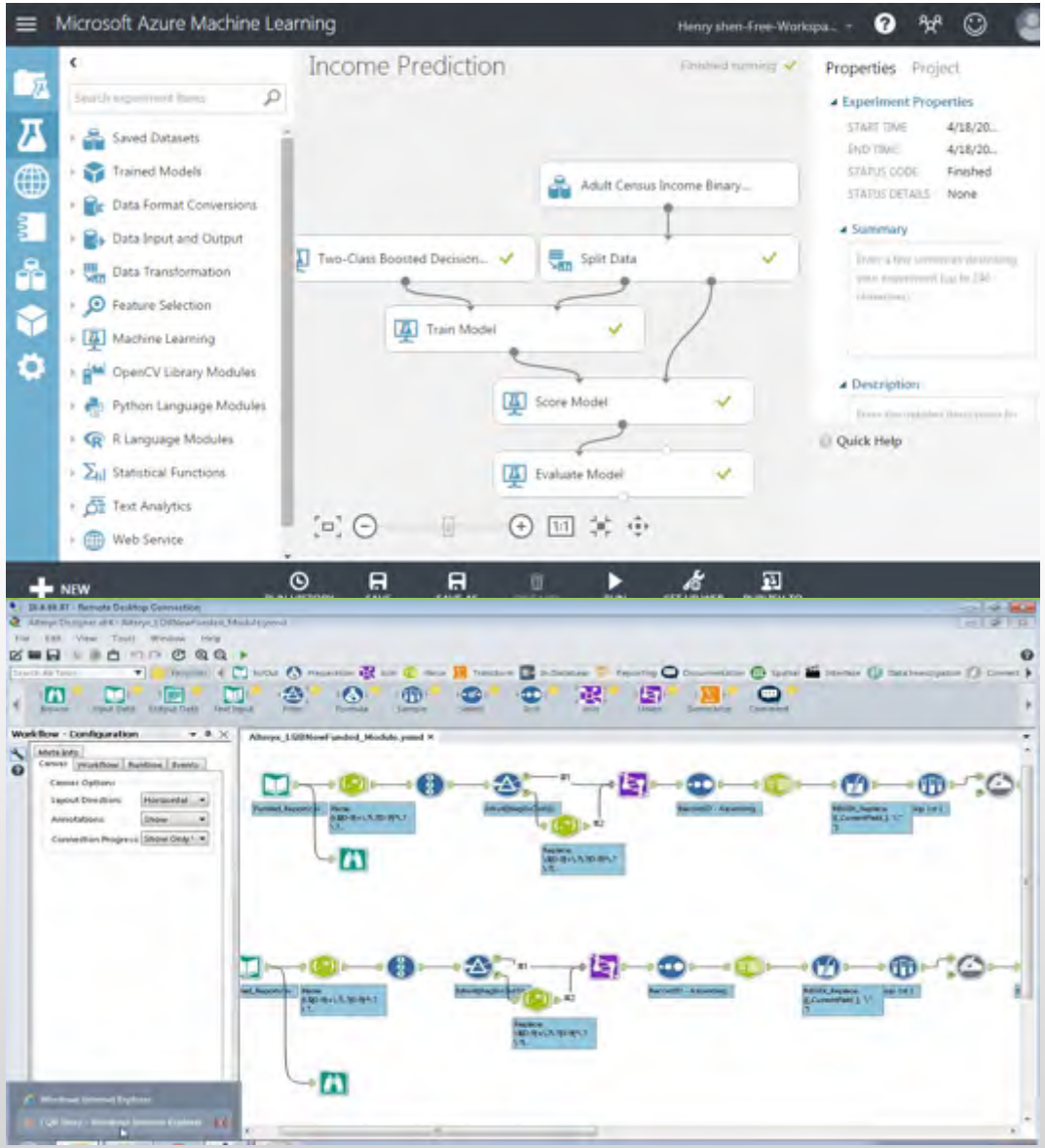
Predictive Analytics

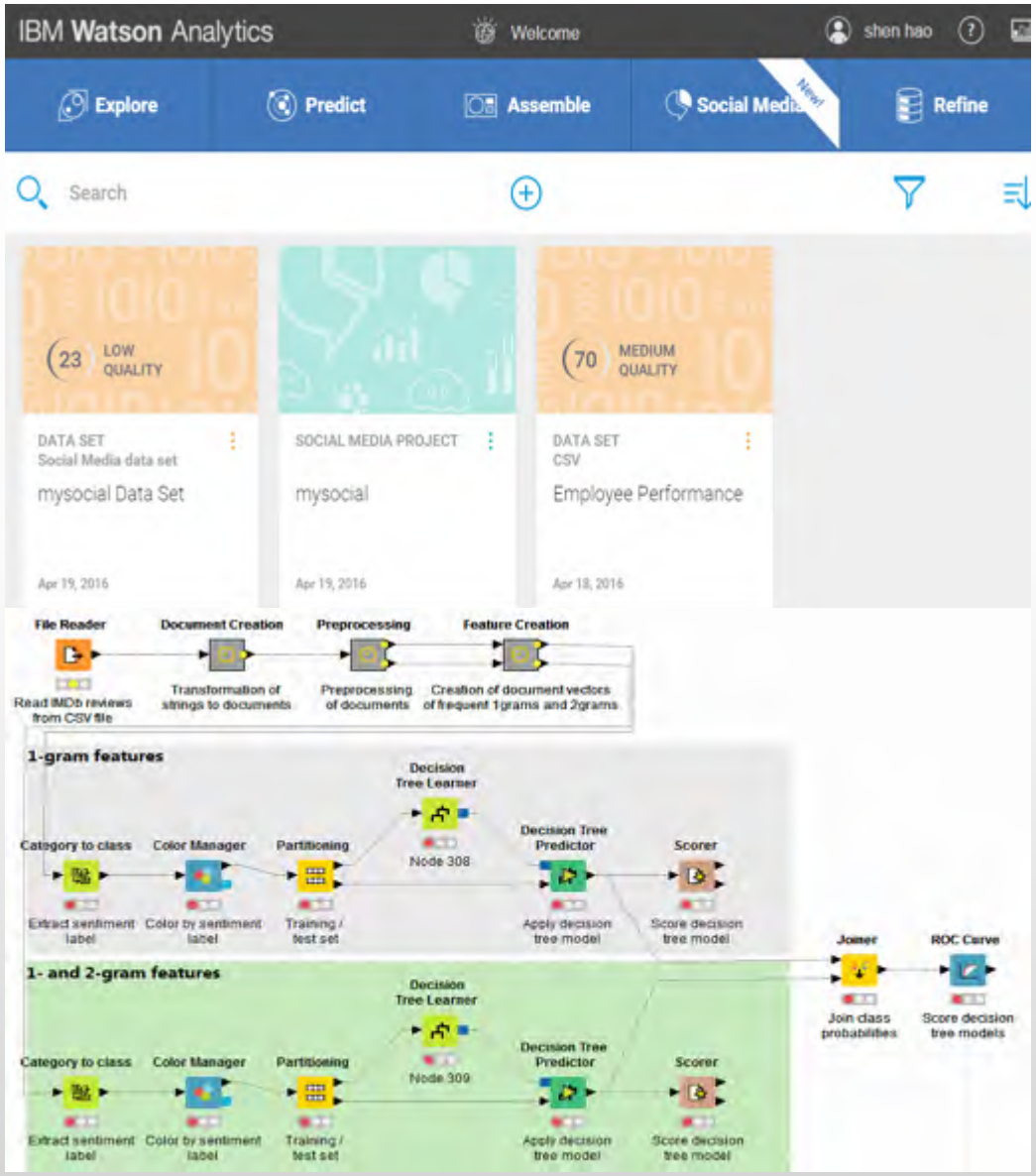
Models, one-click data mining, decision management

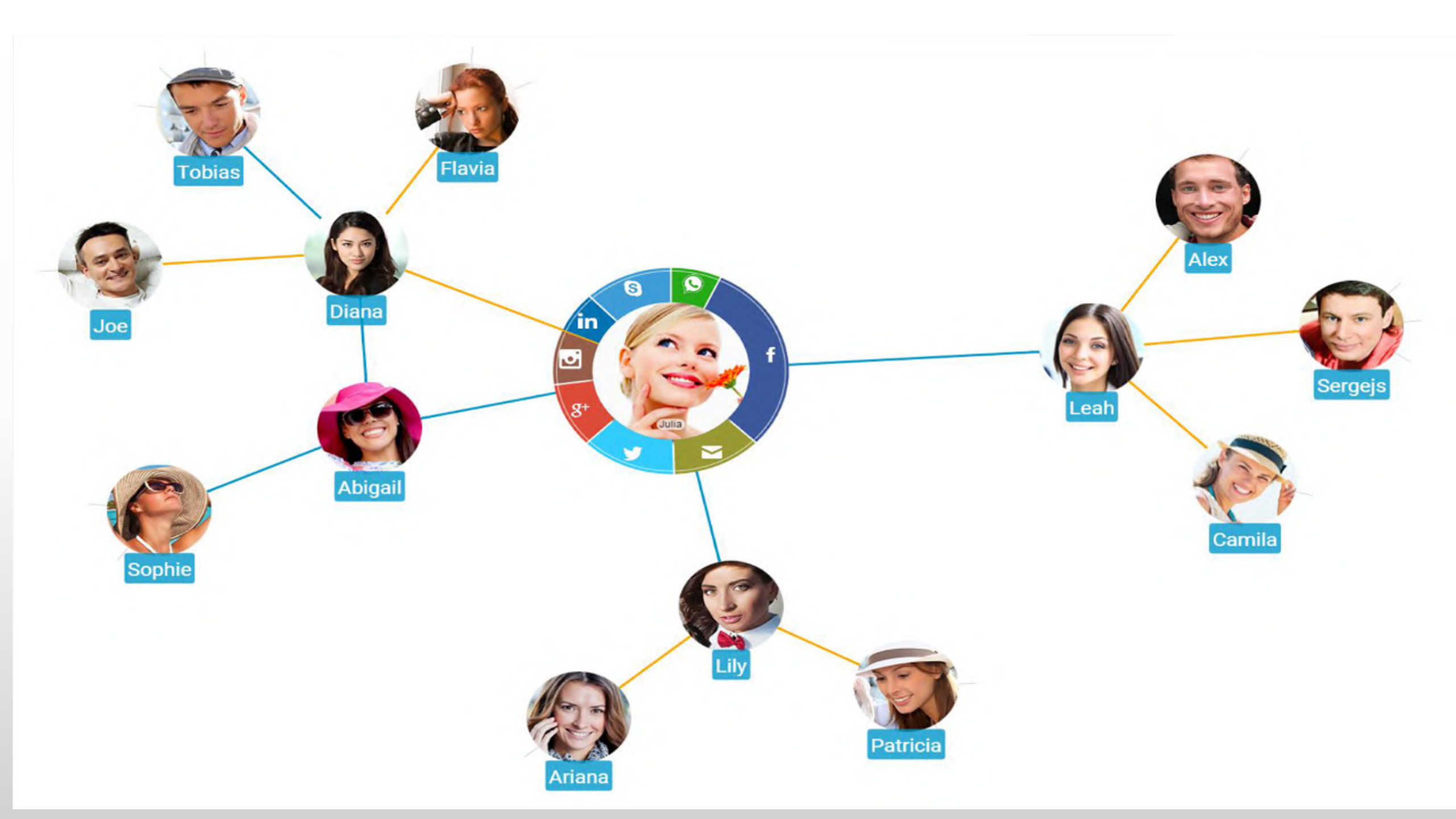
Data Warehousing

Data storage, data modeling

大数据技术更新


Microsoft Azure Machine Learning
Income Prediction
Properties
Project
Saved Datasets
Trained Models
Data Format Conversions
Data Input and Output
Data Transformation
Feature Selection
Machine Learning
OpenCV Library Modules
Python Language Modules
R Language Modules
Statistical Functions
Text Analytics
Web Service
Adult Census Income Binary...
Two-Class Boosted Decision...
Split Data
Train Model
Score Model
Evaluate Model
Quick Help
NEW

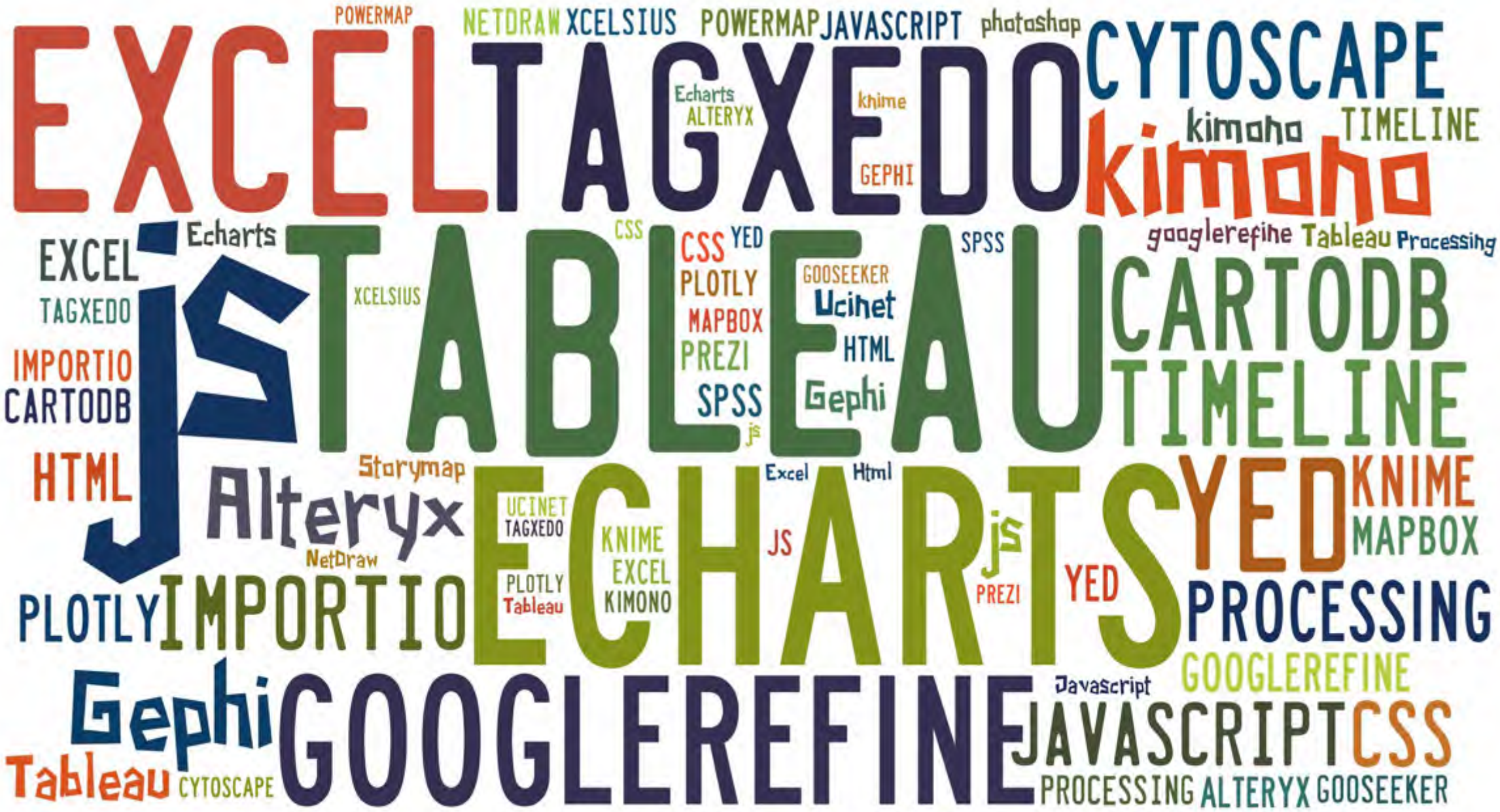

IBM Watson Analytics
Welcome
shen hao
Explore
Predict
Assemble
Social Media
Refine
Search
23 LOW QUALITY
70 MEDIUM QUALITY
DATA SET
Social Media data set
mysocial Data Set
Apr 19, 2016
SOCIAL MEDIA PROJECT
mysocial
Apr 19, 2016
DATA SET
CSV
Employee Performance
Apr 18, 2016
File Reader
Read IMDb reviews from CSV file
Document Creation
Transformation of strings to documents
Preprocessing
Preprocessing of documents
Feature Creation
Creation of document vectors of frequent 1grams and 2grams
1-gram features
Category to class
Extract sentiment label
Color Manager
Color by sentiment label
Partitioning
Training / test set
Decision Tree Learner
Node 308
Decision Tree Predictor
Apply decision tree model
Scorer
Score decision tree model
1- and 2-gram features
Node 309
Joiner
Join class probabilities
ROC Curve
Score decision tree models


Tobias
Flavia
Joe
Diana
Alex
Julia
Leah
Sergejs
Abigail
Sophie
Camila
Lily
Ariana
Patricia

掌握的主要技术和方法


文本聚类分类
KOL意见领袖
PMML模型与云端部署
API接口
情感词典建设
归并文本挖掘与网络挖掘
网络抓数据
规则建模推荐算法
RoambiAPP移动应用
MYSQL和HADOOP存储
用户词典构建
cytoscape可视化分析
Xcelsius仪表盘
GEPHI动态可视化分析
Khime和R语言挖掘
API插件和接口程序
R语言与分词
TABLEAU可视化分析
情感分析
网络分析

掌握的主要软件工具


EXCEL
TAGXEDO
CYTOSCAPE
kimoho
TIMELINE
JS
TABLEAU
CARTODB
TIMELINE
ECHARTS
YED
KNIME
MAPBOX
PROCESSING
IMPORTIO
PLOTLY
Gephi
GOOGLEREFINE
JAVASCRIPT
CSS
Tableau
POWERMAP
NETDRAW
XCELSIUS
POWERMAP
JAVASCRIPT
photoshop
Echarts
ALTERYX
khime
GEPHI
kimoho
TIMELINE
EXCEL
Echarts
CSS
CSS
YED
SPSS
googlerefine
Tableau
Processing
TAGXEDO
XCELSIUS
PLOTLY
GOOSEEKER
Ucinet
MAPBOX
HTML
PREZI
SPSS
Gephi
IMPORTIO
CARTODB
HTML
Storymap
Alteryx
Excel
Html
UCINET
TAGXEDO
KNIME
EXCEL
KIMONO
JS
js
PREZI
YED
NetDraw
PLOTLY
Tableau
PROCESSING
GOOGLEREFINE
Javascript
Gephi
GOOGLEREFINE
JAVASCRIPT
CSS
Tableau
CYTOSCAPE
PROCESSING
ALTERYX
GOOSEEKER

掌握的业务流程


1
Retail planner
需求预测与市场
2
Marketing
市场研究与
消费者洞察
3
Go-To-Market
市场定位
geospatial / upsell /
cross-sell prospects
4
Supply planner
销售监测与供应链
5
Commodities
成本控制与物流管理
6
Store stocks
产品与货架购物篮
7
Customer
消费者满意度与忠诚度
8
IT
个性化推荐
与AB测试
Connected device
Data sciences

BIG DATA

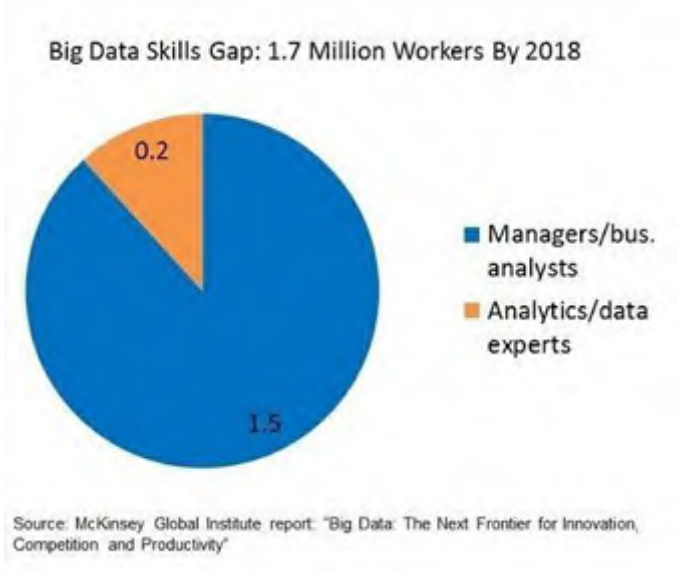
培养专业人才

大数据人才需求

Gartner

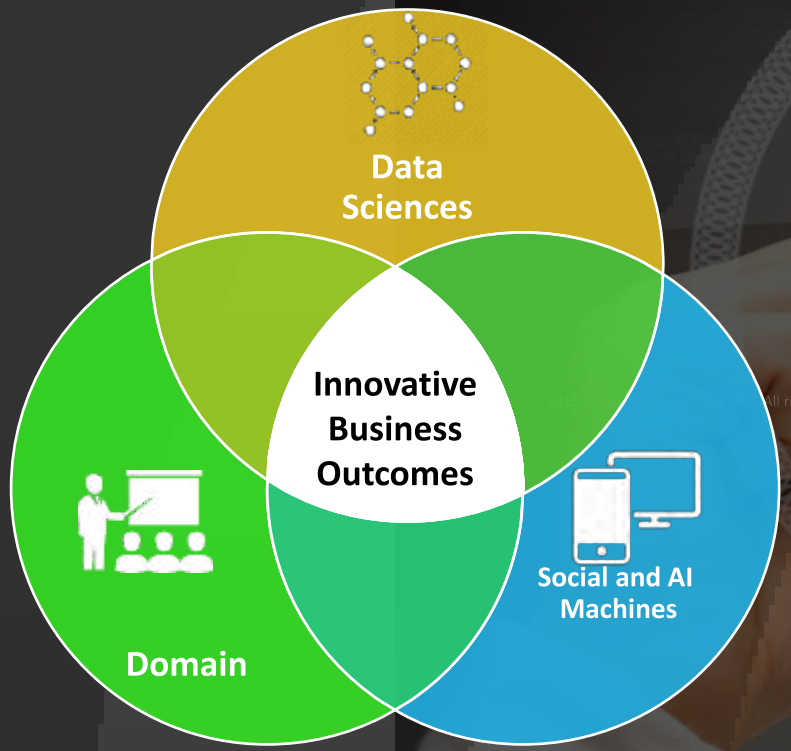
2015年，全球将新增440万个与大数据相关的工作岗位，且会有25%的组织设立首席数据官职位。未来，大数据将会出现约100万的人才缺口。

McKinsey&Company



麦肯锡全球研究院的研究预测在未来6年，仅在美国本土就可能面临缺乏14万至19万具备深入分析数据能力人才的情况，同时具备通过分析大数据并为企业做出有效决策的数据的管理人员和分析师也有150万人的缺口。

Data Artisans 数据工匠



培养专业人才

专业领域知识
与技能

知识管理
绩效管理
竞争分析
决策支持

数据库
数据挖掘
统计分析
预测分析

市场营销
市场研究
媒介研究
广告研究

传播与沟通
演讲与写作

可视化技能
网页设计

大数据人才培养

组建师资团队
开设专业课程
寻找实训基地

高 校

知识学习

&

企 业

技术实践

建立合作关系
开发合作项目
实现人才交流

保持紧密高效联系，实现联手双赢局面


BIG DATA
03
打破数据孤岛

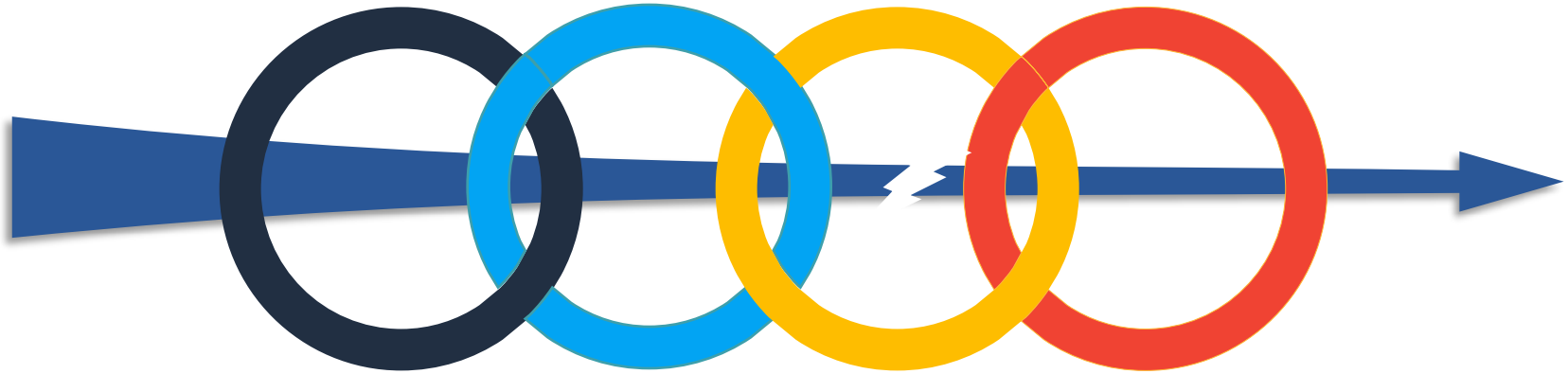

风景虽然美丽
现实问题严重

企业数据孤岛

数据孤岛：指的是一个个相对独立的不同类型的数据资源系统。

1.企业知道他们能够从信息和数据中获得更有价值的洞察，但不知道怎么做。

3. 数据散落在不同部门，存在不同的数据仓库中，不同部门的数据技术也有可能不一样，这导致企业内部自己的数据都没法打通。



2.虽然没有明确的大数据业务需求，但希望可以整合企业数据，保护数据资产。

4.大数据需要不同数据的关联和整合才能更好的发挥理解客户和理解业务的优势。将不同部门的数据打通，并且实现技术和工具共享，才能更好的发挥企业大数据的价值。

企业数据湖

数据湖：表面上看，数据都是承载在基于可向外扩展的HDFS廉价存储硬件之上的。但数据量越大，越需要各种不同种类的存储，并且，不是所有的企业数据都是适合存放在廉价的HDFS集群之上的。

数据湖近在眼前，有人持怀疑的态度，有人热情拥抱。怀疑者认为，数据湖无非另一种将全部数据整合至单一位置的存储形式，支持者认为，数据湖不仅预示着前所未见的存储效率，还让分析成为可能，让每个组织都可用。

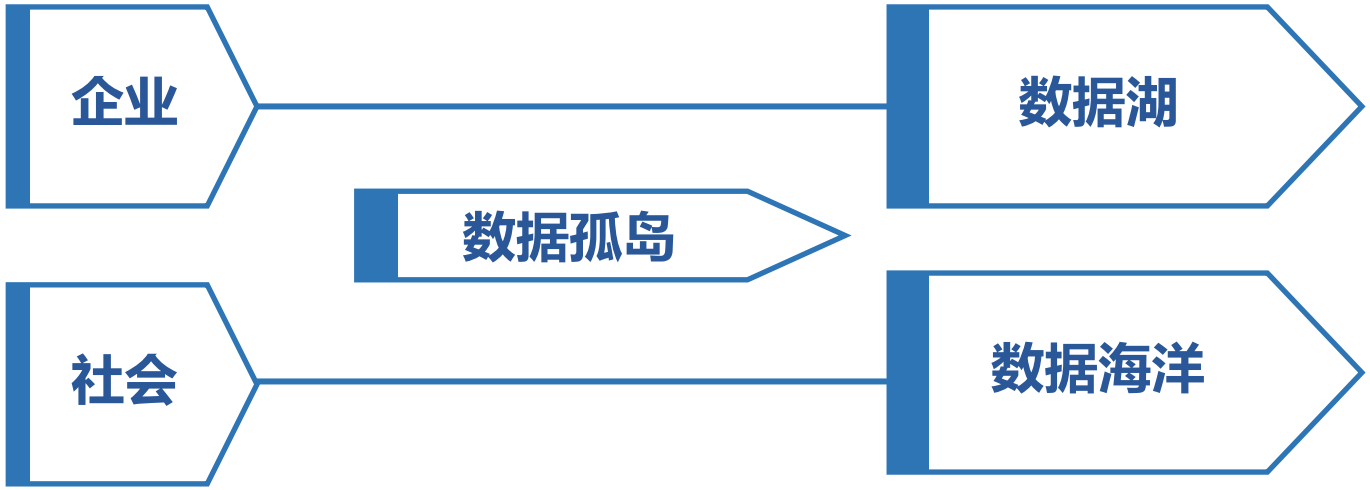
数据湖概念的提出，让我看到打破数据孤岛的重要性。

社会数据孤岛

从宏观层面上看，数据孤岛：**开放数据**

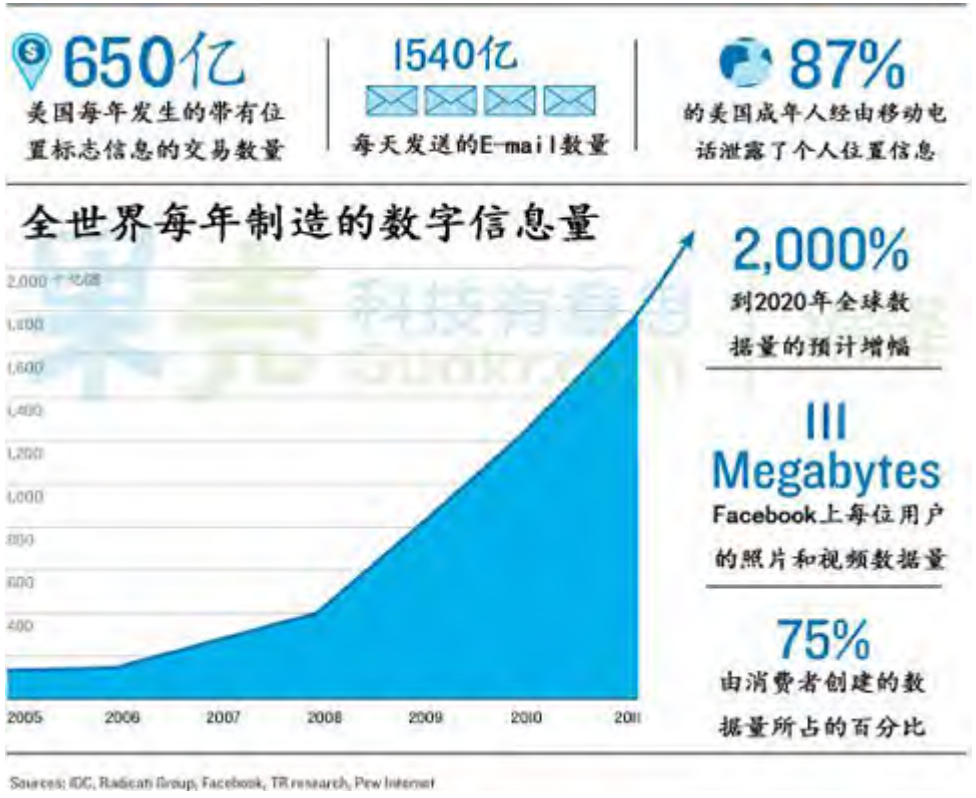
指的是政府、企业和行业信息化系统建设往往缺少统一规划和科学论证，系统之间缺乏统一的标准，形成了众多“数据孤岛”。

因此，它的意义不仅仅是数据开放，更重要的是能够制定一个可以遵循的**数据存储、读取的管理规则，为所有人使用。**




BIG DATA
04
权衡开放&隐私

数据开放与隐私



图片：MIT TechnologyReview（2013）

尽管这其中的大部分数据都是不可见的，似乎也并不携带任何个人信息，但事实并非如此。现代数据科学已经发现几乎任何类型的数据都可能用来识别创造它的人，实际上，数据越多，其中可以称得上隐私的就越少。

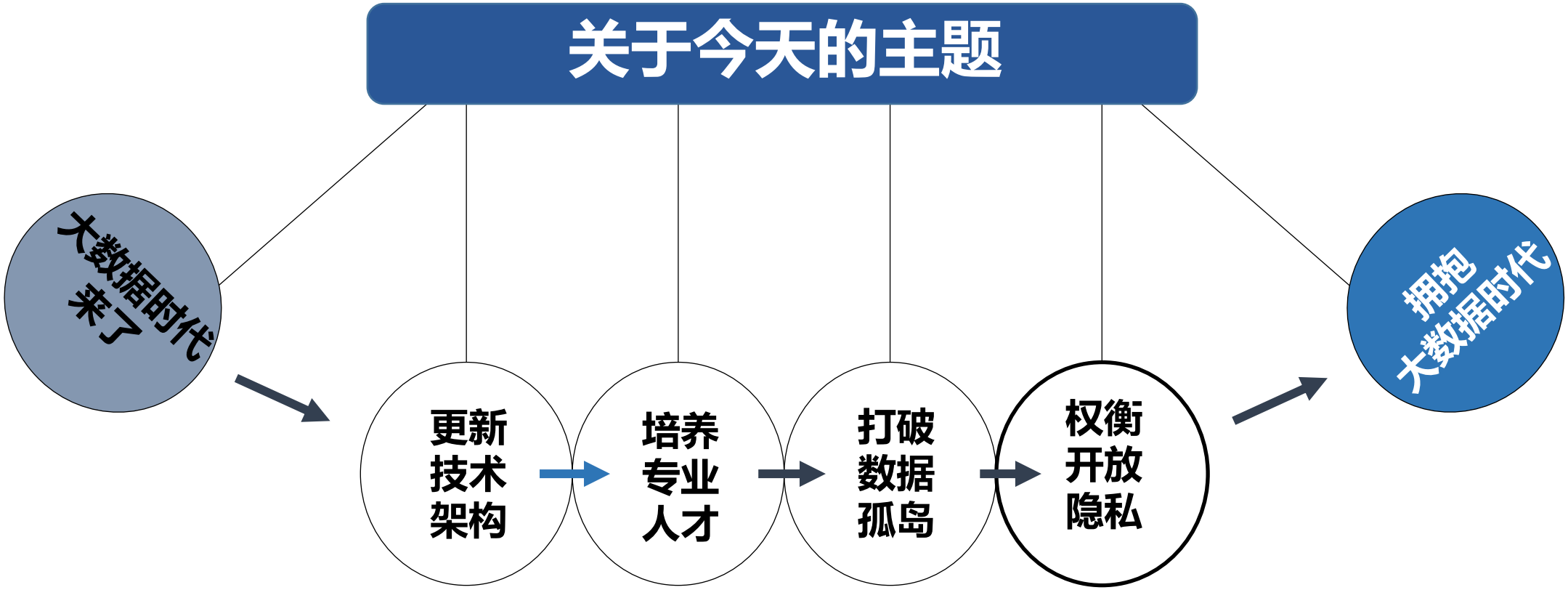
数据开放与隐私

“我点击了自己的名字，页面上出现了一张熟悉的照片——是我穿着一件蓝色衬衫的照片，旁边配有我的基本履历资料……我点开了一个最近更新的链接，地点是波士顿的马萨诸塞大街……两秒钟后，我在视频中看到了自己推开了地铁站那厚重的大门……每次看到自己出现在视频中，我都会浑身不自在。但现在可好，我的一举一动已经被LifeLinear网的系统给记录了下来……”

开放是基石，隐私是关键

加大立法制度

加强行业管控



大数据时代的到来，让我们认识到现有技术的缺陷，而新技术、新产品、新公司如雨后春笋般涌现；从不缺少技术的更新，当然也需要人才的跟进；当然，对数据资源的需求也在逐步增加，需要数据的开放与相互之间的连接；此时，我们应该回归本体，注重对数据隐私的保护。

我们生活在社会中，就不得不同数据打交道。我们也是数据的一部分，不论我们想不想与大数据牵扯到一起，数据都会找到我们，覆盖我们。

大数据时代已经来临，如何从海量数据中发现知识，寻找隐藏在大数据中的模式、趋势和相关性，揭示社会现象与社会发展规律，以及可能的商业应用前景，都需要我们拥有更好的数据洞察力。

中国大数据发展面临的挑战

THANKS.

2016