



2015全局数据白皮书

2015年11月



OMNILab

上海交通大学开放创新实验室

Topdigital
China TopDigital World



01

我们面对的世界

A

万物互联：人类连接方式的改变促进了数据的爆炸性增长，导致了社会驱动力变革

B

信息时代的驱动力变革：技术驱动-数据驱动-场景驱动



OMNILab
上海交通大学开放创新实验室

万物互联

人类连接方式的改变导致了数据体量的爆炸性增长

在人类文明的伊始，日光之下，人与人的第一声交流即意味着“连接”的开始。语言使人与人连接，连接即产生“数据”。而文字使这种连接沉积为信息与知识。

工业文明时代，“连接”开始通过无线电台、电报、电视的形式存在，但这样的“连接”产生的信息往往是单向性而缺乏互动的。

伴随着移动互联网的发展、物联网的逐渐普及，人类的沟通和交互成本趋向为零成本、零时差，万物互联的时代来临，使随时随地收集数据成为可能。

在整个农耕文明时代，“连接”仅是以语言沟通和书面文字沟通的形式存在。

互联网时代，人和人开始通过网络进行复杂交错的互动连接。社交网站、电子邮件、搜索引擎、聊天工具.....人类建立连接的方式开始多样化、多维化。人类社会产生的数据开始爆炸性增长。

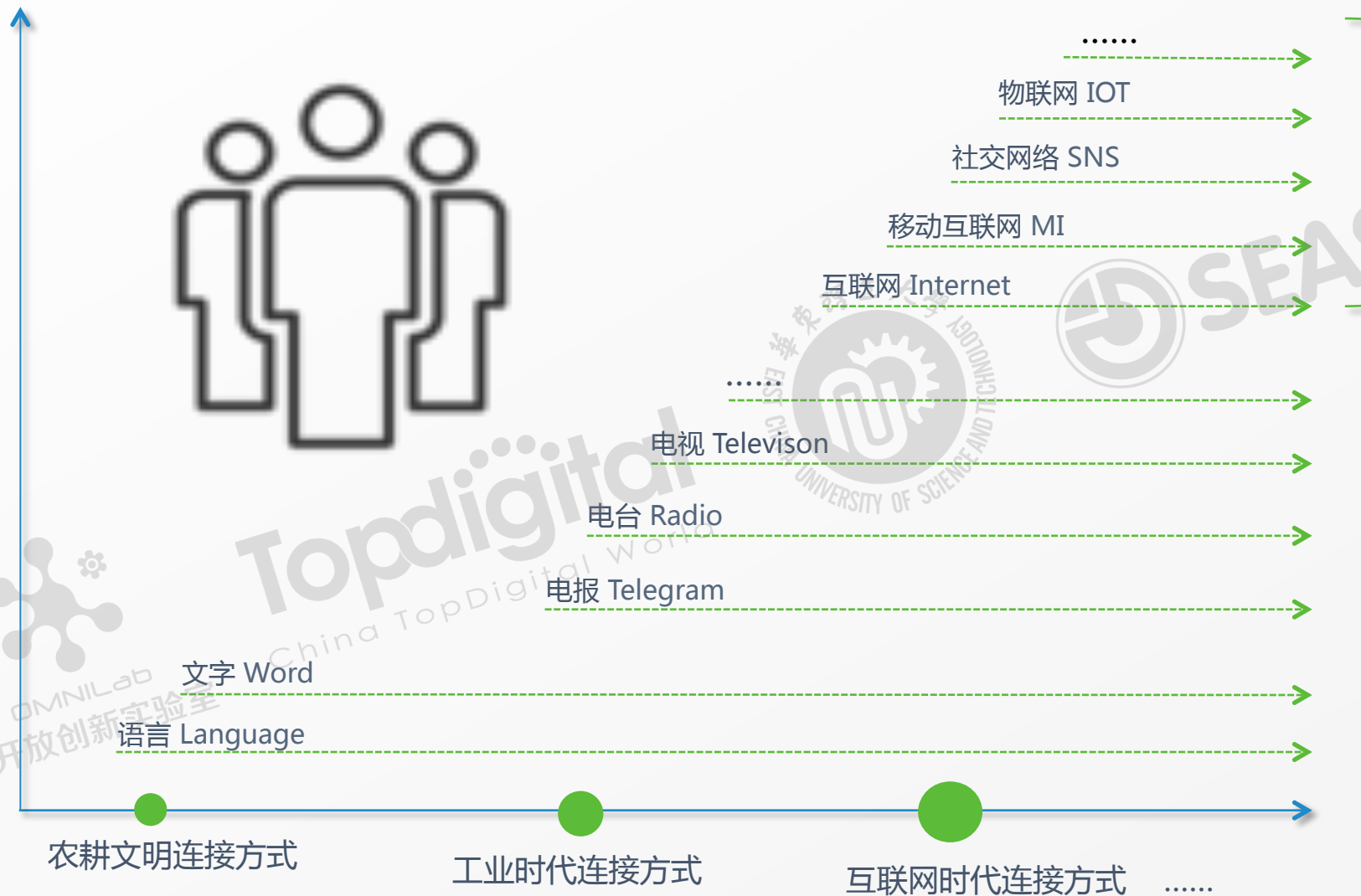


上海交通大学开放创新实验室

万物互联

人类连接方式的改变导致了数据体量的爆炸性增长

连接方式改变导致可收集、记录的数据体量增长



- 移动互联网、物联网的兴起，产业的数据化及技术的进步，让人类社会可被收集、记录的数据量产生了爆炸性增长。

万物互联

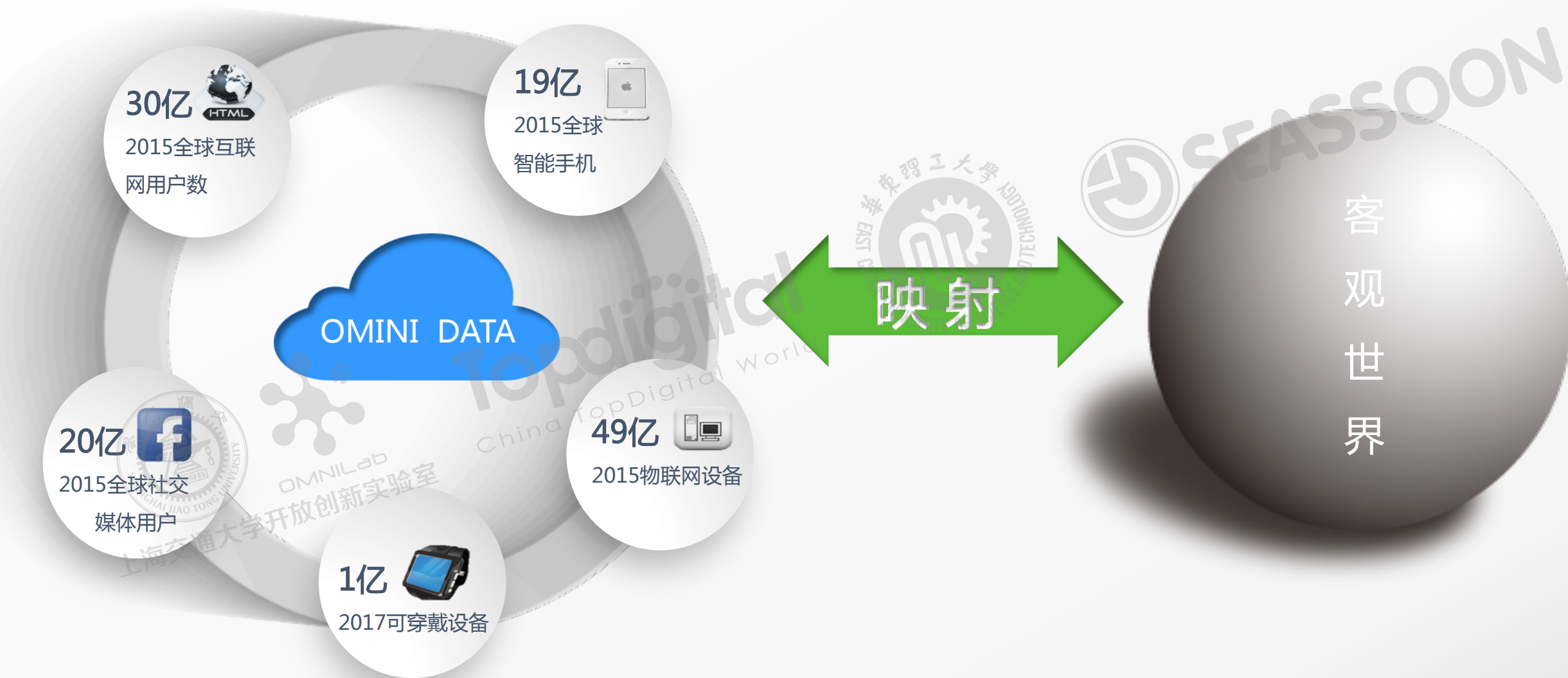
万物互联包括人与人、人与设备、设备与设备之间的连接



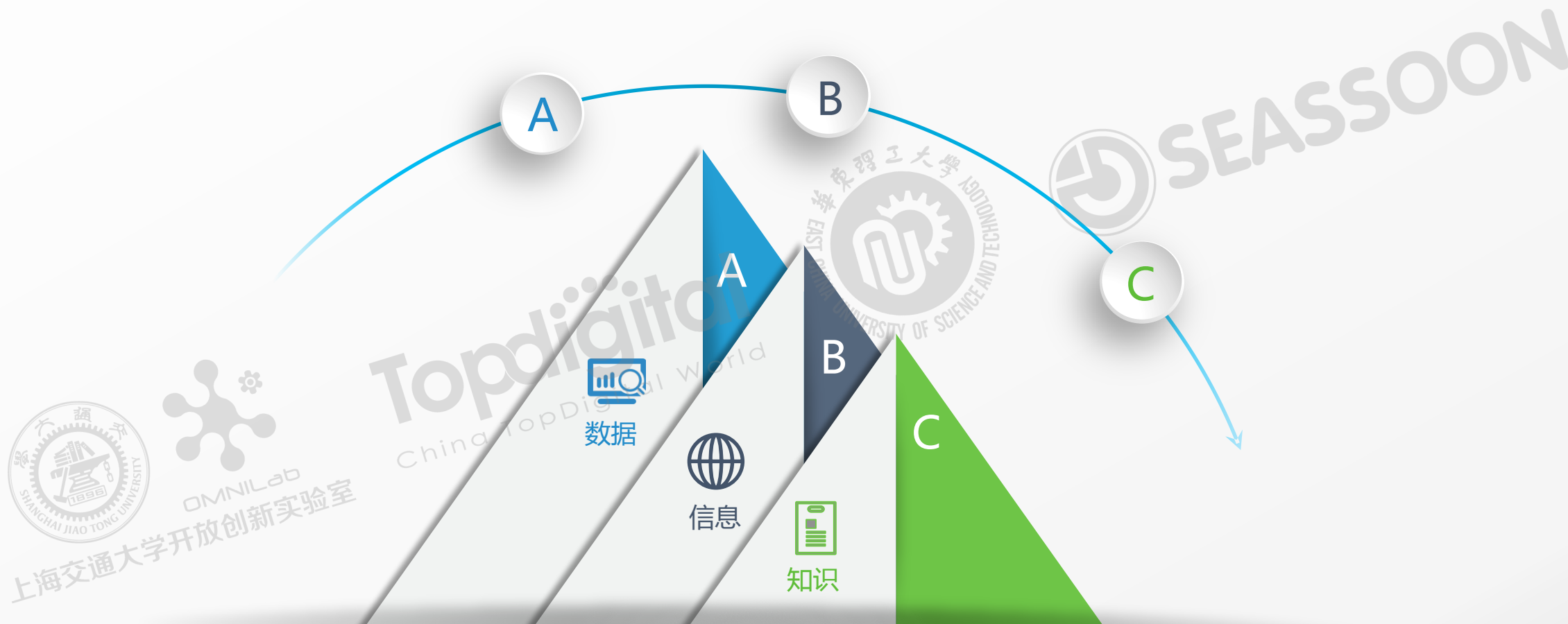
- 以移动设备为媒介，人类与人类之间、人类与设备之间、设备与设备之间建立了更密切、更即时的连接。

万物互联

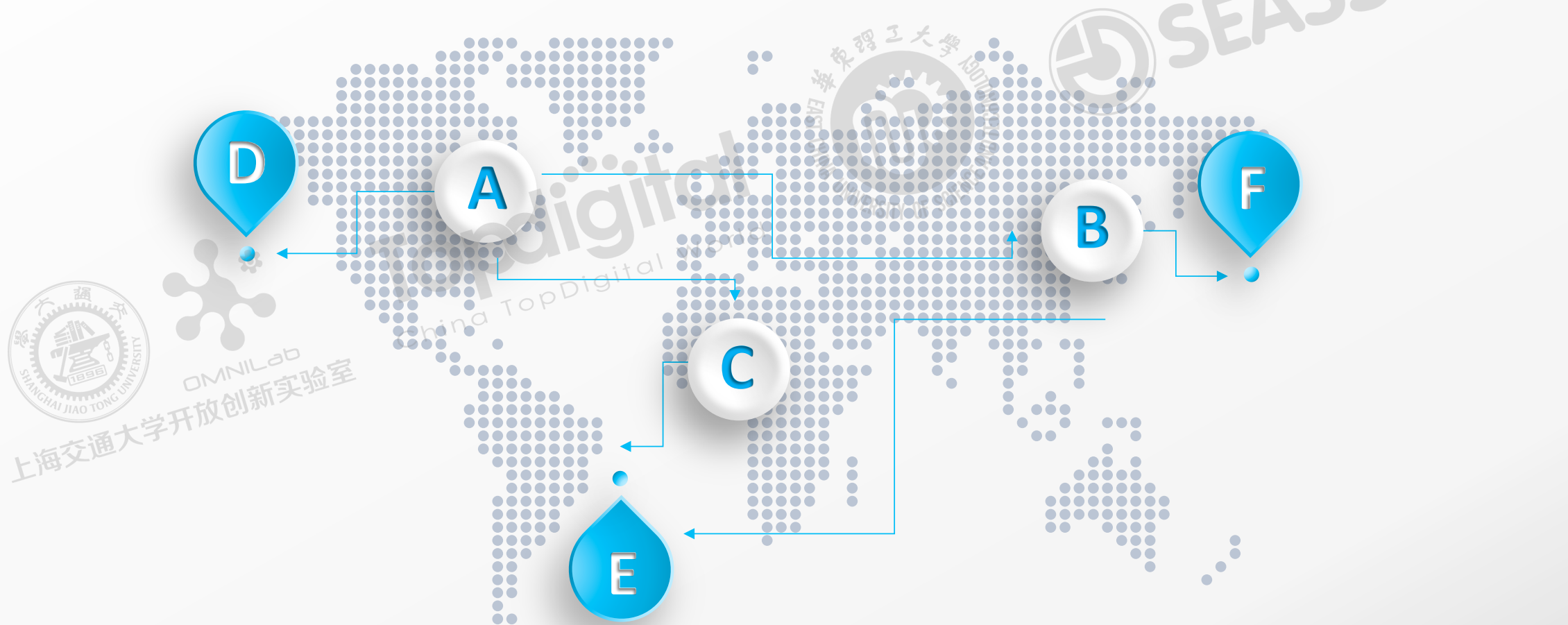
人类收集数据的渠道和手段极大丰富，数据开始真实还原客观世界的映射



- 数据是信息和知识的基础，数据的激增也导致了人类社会信息和知识的爆炸性扩大。



万物互联、数据爆炸的态势下， 什么在驱动着我们的未来？



从技术驱动到数据驱动，再到场景驱动

信息时代的驱动力变革



定位



测试优化

技术驱动

“用户参与”成为可能，收集用户初始化信息，根据信息执行策略并实施



精准画像



预测分析

数据驱动

收集的数据量大大增加，可直接通过数据进行趋势预测和决策分析，可绘制用户画像，进行针对性产品/服务推介



全渠道整合



场景化决策

场景驱动

可针对不同个体、群体所处情境，洞察情境中“人”的决策机制

- 技术发展驱动了数据收集量的暴增，而数据成为场景驱动的基础。
- 人类对数据质量的要求大大提升。人类需要经过治理的数据去做智慧的驱动和决策。

信息时代的驱动力变革

从技术驱动到数据驱动，再到场景驱动

- “场景驱动”的基础一定是经过治理的数据。

- “场景驱动”不仅会驱动人类的决策，亦会驱动机器的自我进化和升级*。

- “场景驱动”的“场景”对应的——一定是有应用价值的场景。

“场景驱动”三要素

- 人类掌控未来的核心破局点，是以什么样的数据、以何种方式驱动人和机器的进化。
- 我们需要经过治理的数据去做智慧的驱动，才不致于将人类社会引入歧途。

*未来，机器将从“模拟再现人类某些智能行为”进化到自主收集信息、自动优化和“深度学习”，甚至可以开始独立思考。喂食给他们什么质量的数据，关系到机器如何优化、学习和思考。

02

大数据需要进化

A

大数据应用的问题，关键是连接方式问题和数据质量问题。

B

万物互联、驱动力变革的时代，我们更需要数据治理导向下“开放、可度量、场景化”的全局数据



OMNILab
上海交通大学开放创新实验室

大数据的定义

01

“大数据指的是大小超出常规的数据库工具获取、存储、管理和分析能力的数据集.....但并不是说一定要超过特定TB值的数据集才能算是大数据。”

——麦肯锡《Big data: The next frontier for innovation, competition, and productivity》

02

“（大数据指）海量的数据规模（Volume）、快速的数据流转和动态的数据体系（Velocity）、多样的数据类型（Variety）、巨大的数据价值（Value）。”

——IDC，国际数据公司

03

“大数据是任何超过了一台计算机处理能力的数据量。”

——John Rauser，亚马逊大数据科学家

04

“巨量资料(big data)，或称大数据，指的是所涉及的资料量规模巨大到无法通过目前主流软件工具，在合理时间内达到撷取、管理、处理并整理成为帮助企业经营决策更积极目的的资讯。” ——维基百科

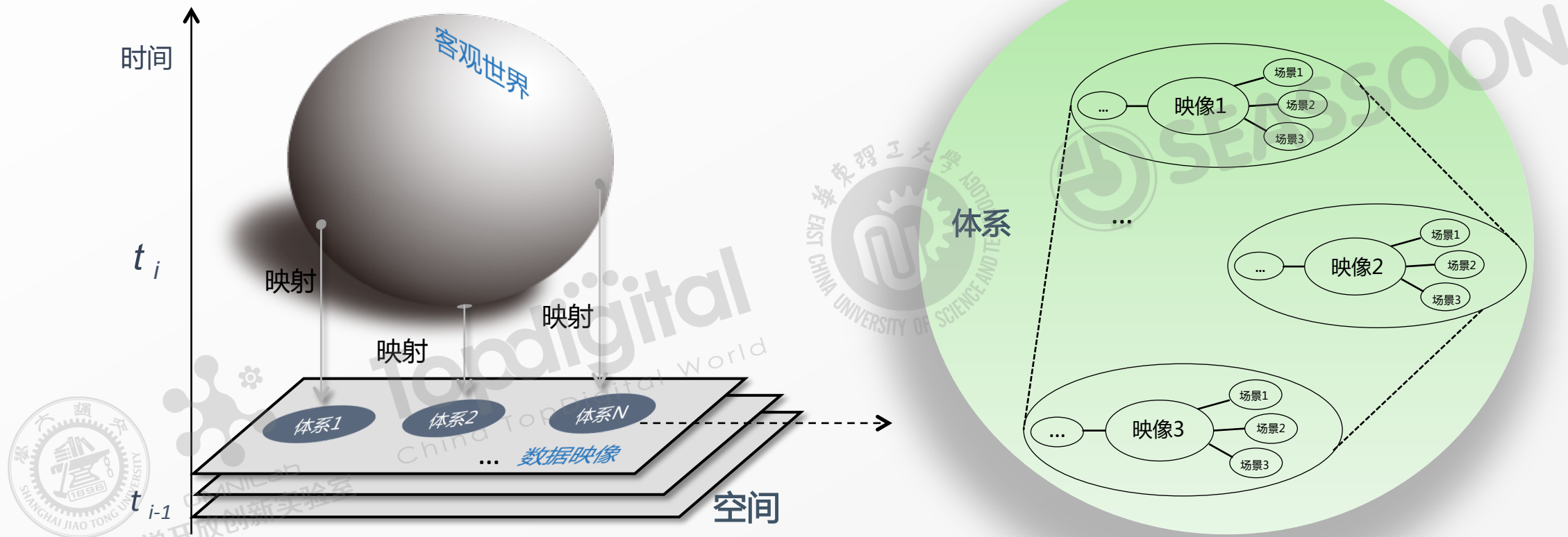


上海交通大学

BIG ?

仅仅大是不够的

数据映像与场景的逻辑关系



- 映像是由不同体系根据不同需求对客观世界的数字化抽象和沉积。在这种抽取和沉积中本身已包含了一种逻辑（映射本身即算法）。
- 不同的映像构成体系，而不同体系组合交叠成为客观世界。映像永远只是客观世界的一部分投影，而非全部。

BIG ?

仅仅是大，是不够的

数据映像与场景的逻辑关系



- 场景包括四部分：时间、空间、语义、语境（情绪、社群.....）。
- 这四部分在不同的场景中随机组合而存在。
- 场景是映像的子集，不同的场景组合而成映像。



BIG ?

仅仅大是不够的

什么是“场景四要素”

时间

表达事物的生灭排列。
其内涵是无尽永前，其
外延是一切事件过程长
短和发生顺序的度量。

空间

空间是与时间相对的一
种物质客观存在形式，
由长度、宽度、高度、
大小表现。通常指四方
上下。

语义

信息的含义。（信息是
被赋予意义的数
据）。本处我们定义为
特定场景中的内容表
达。

语境

语义产生的环境。不
同语境下，相同的语
义会产生不同的含义。
又可细分为不同要素，
如社群*、情绪*等。



OMNILab
上海交通大学开放创新实验室

TopDigital
China Digital World



SEASOON

仅仅大是不够的

问题一

分散的数据孤岛，无法建立连接，导致数据无法流动印证



典型的商业场景是，在BAT自有体系内，他们的数据量都是非常巨大的。但是，他们只能沉积和应用自己产品体系内的各种数据，不能相互关联与开放。

- 仅仅是体量大，而不能使数据之间打破体系，互相流动印证是不够的。
- 孤立的数据的价值远远小于全局的、广泛连接的数据价值。

- 数据的核心价值不来源“大”，而来源于其中蕴含的规律性与智慧性。而这需要打破体系的界限，让数据产生连接，从而导向更深度的洞察。

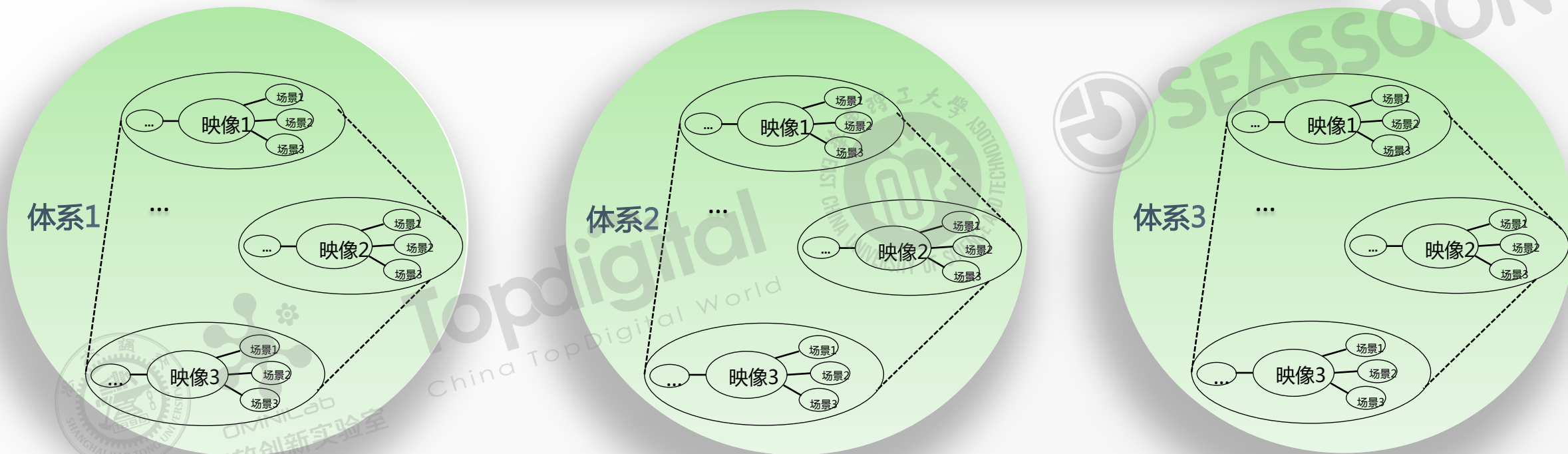
BIG ?

仅仅大是不够的

目前的大数据应用存在的问题，是连接方式和数据质量的问题

问题一

分散的数据孤岛，无法建立连接，导致数据无法流动印证

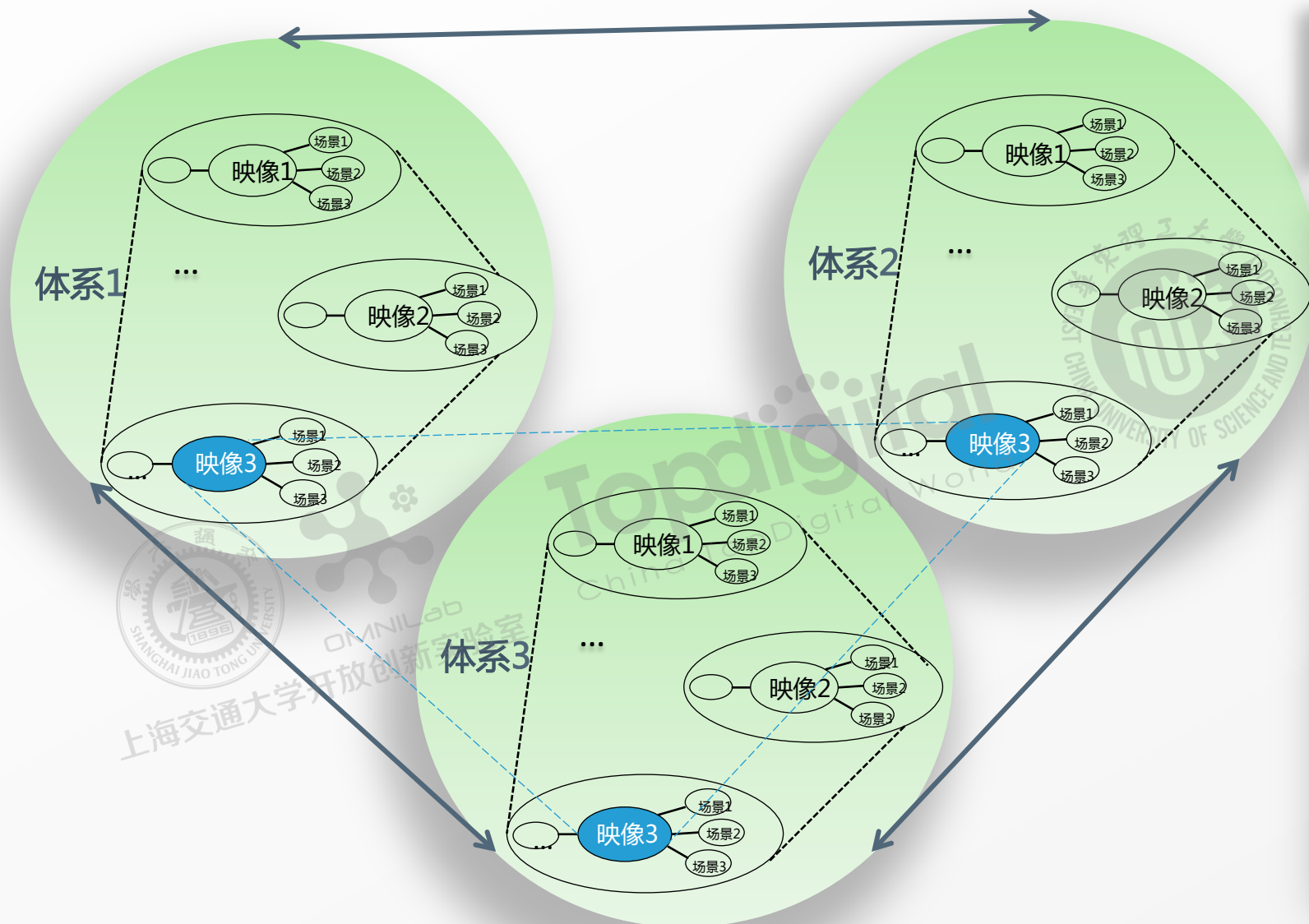


分散的数据孤岛，体系与体系间、映像与映像间、场景与场景间都没有连接

BIG ?

仅仅大是不够的

目前的大数据应用存在的问题，是连接方式和数据质量的问题



问题一

体系与体系间存在连接，
但连接过于粗放或过于细微

■ 过于粗放。

体系与体系之间的连接更多存在于体系间和映像间。但体系由多个映像构成，映像又由不同的场景组成，不同场景又由四要素随机组合，仅仅连接体系或映像仅会导致模糊结果，无法得出精准结论。

■ 过于细微。

另外一种可能的倾向是，体系和体系之间的连接仅是最基础的、分散的数据连接，这种连接只能导致管中窥豹的片面结论。

MASS ?

混杂带来的数据质量不高

新进阶的数据概念必须包含治理数据的能力，控制数据的量与质，避免错误连接

问题三

数据的质量不高，导致错误的场景重现、错误的映像，出现错误的连接和错误决策。



OMNILab

上海交通大学开放创新实验室

Topdigital
China TopDigital World



BETTER ?

想到，做不到

更“好”的数据必须具备以下特征：场景化、开放性、可度量

1 场景化



不同的场景连接形成对客观世界的不同映像。不同映像的组合形成不同的体系。如果连接只发生在体系与体系间、映像与映像间，这种数据之间的互相印证还是只能支持模糊结论而非精确结果的。解决精准还原，需要数据的“场景化”。

2 开放性



数据融合的价值远远大于数据割裂的价值，但是，大数据连接产生的三大连接问题使数据融合无法顺畅实现。造成了场景驱动的阻碍。打破连接问题，需要数据的“开放性”。

3 可度量



从技术驱动到数据驱动到场景驱动，我们一直面对的重大问题有两个：若开放数据，开放哪些数据？这些数据的维度和界限是否能被度量？以数据为基础进行的数据应用驱动效果是否可量化可衡量？解决效果评估，需要数据的“可度量”。

■ “开放”、“场景化”、“可度量”应成为数据“更好”的新标准。

03

全局数据新时代

A

从BIG DATA到OMINI DATA

B

OMNI DATA：特征 / 能力 / 界限



OMNILab
上海交通大学开放创新实验室

全新的概念：OMNI DATA

特征：场景化 开放性 可度量
及时性 价值化



OMNILab

上海交通大学开放创新实验室

全局数据的五大特征

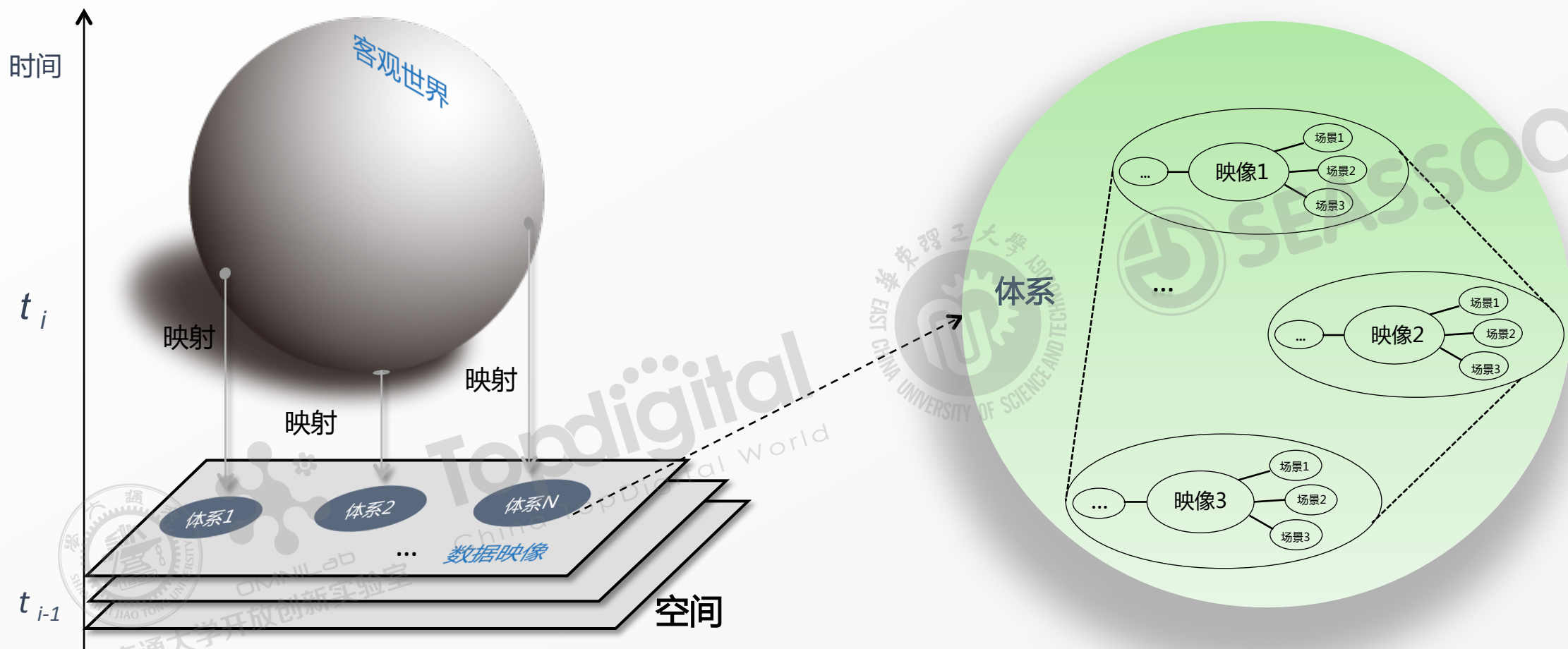


场景化 / 开放性 / 及时性 / 可度量 / 价值化



上海交通大学开放创新实验室

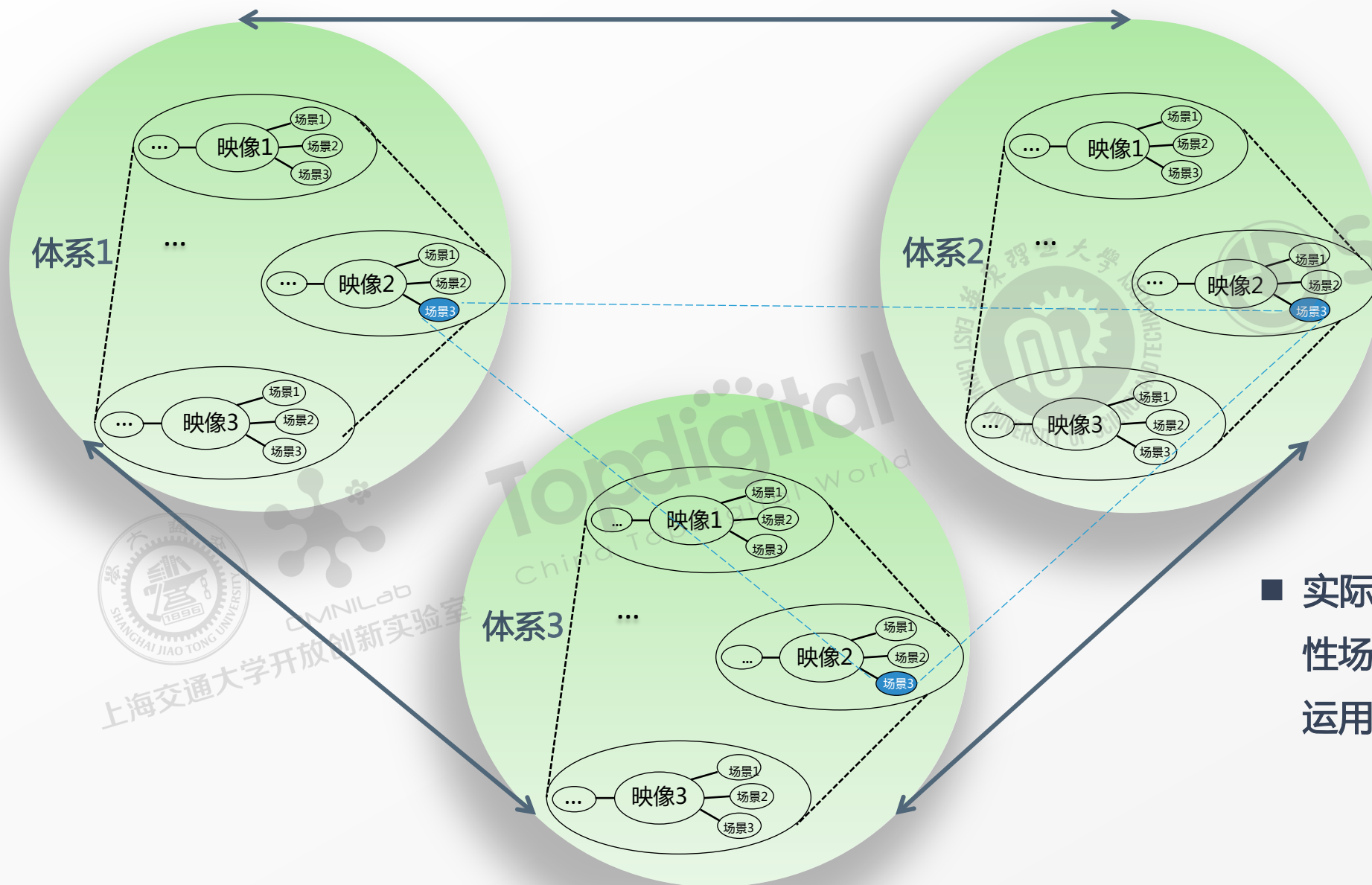
全局数据：是全局，而不是全部



- 如前所述，在任何体系内的数据抽取和沉积都只能是客观世界的映像，是客观世界的一部分，而不是全部。
- 全局数据并不意味着打破所有体系，拉平所有数据，再进行数据的挖掘和应用。

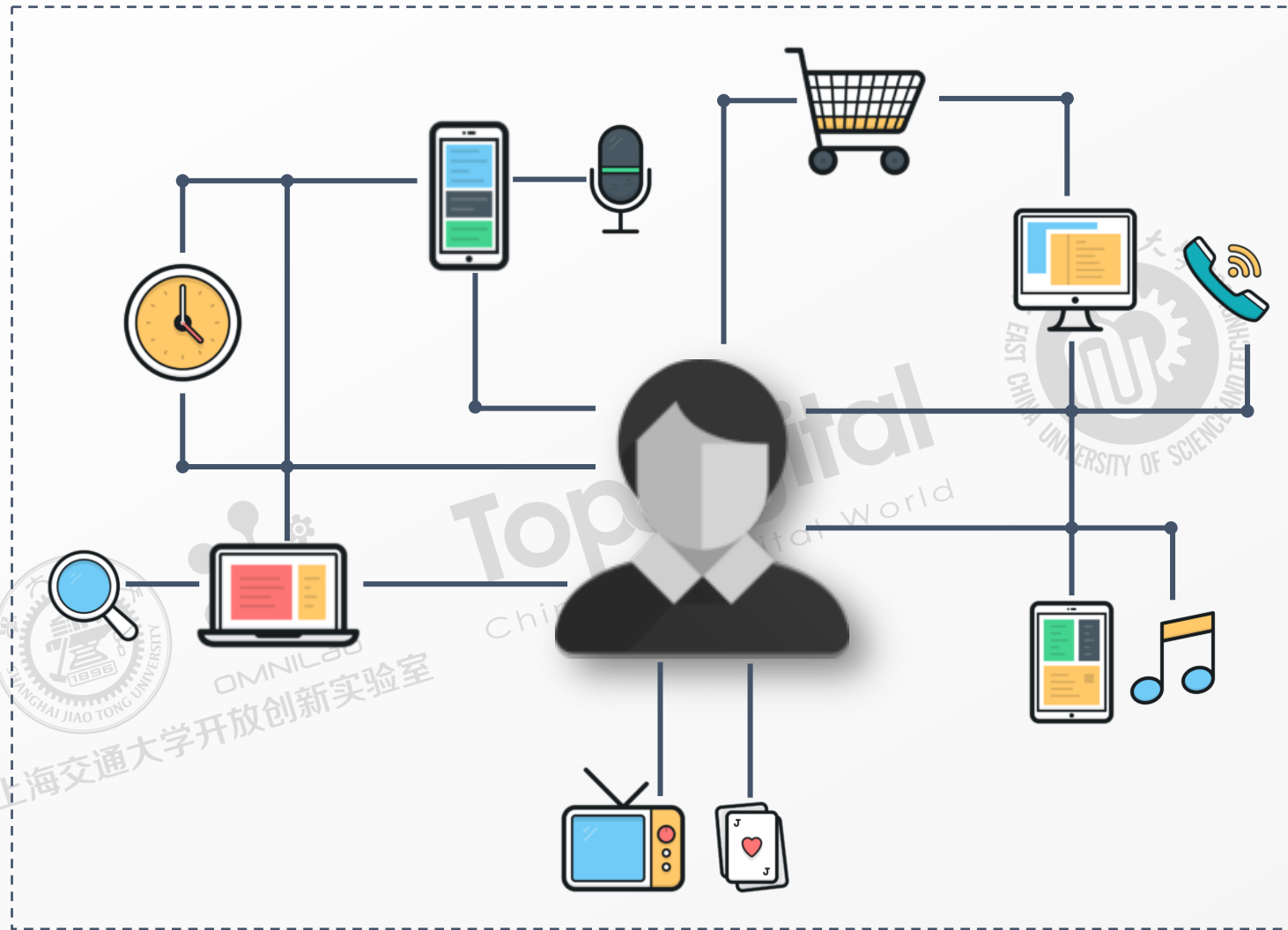
全局数据：全局化凸显场景化

场景化的数据连接，是全局数据的根本特征。



■ 实际上，全局仅需要将所有关联性场景进行匹配连接，进行挖掘运用。

全局数据：场景化



■ 全局数据时代，“人”成为场景的核心，“移动设备”成为沟通人与人、人与设备、设备与设备的介质，这是场景化连接的基础。

■ 场景化包括：场景构建、场景连接和场景洞察三部分。

全局数据：场景化

场景构建 / 场景连接 / 场景洞察

场景构建



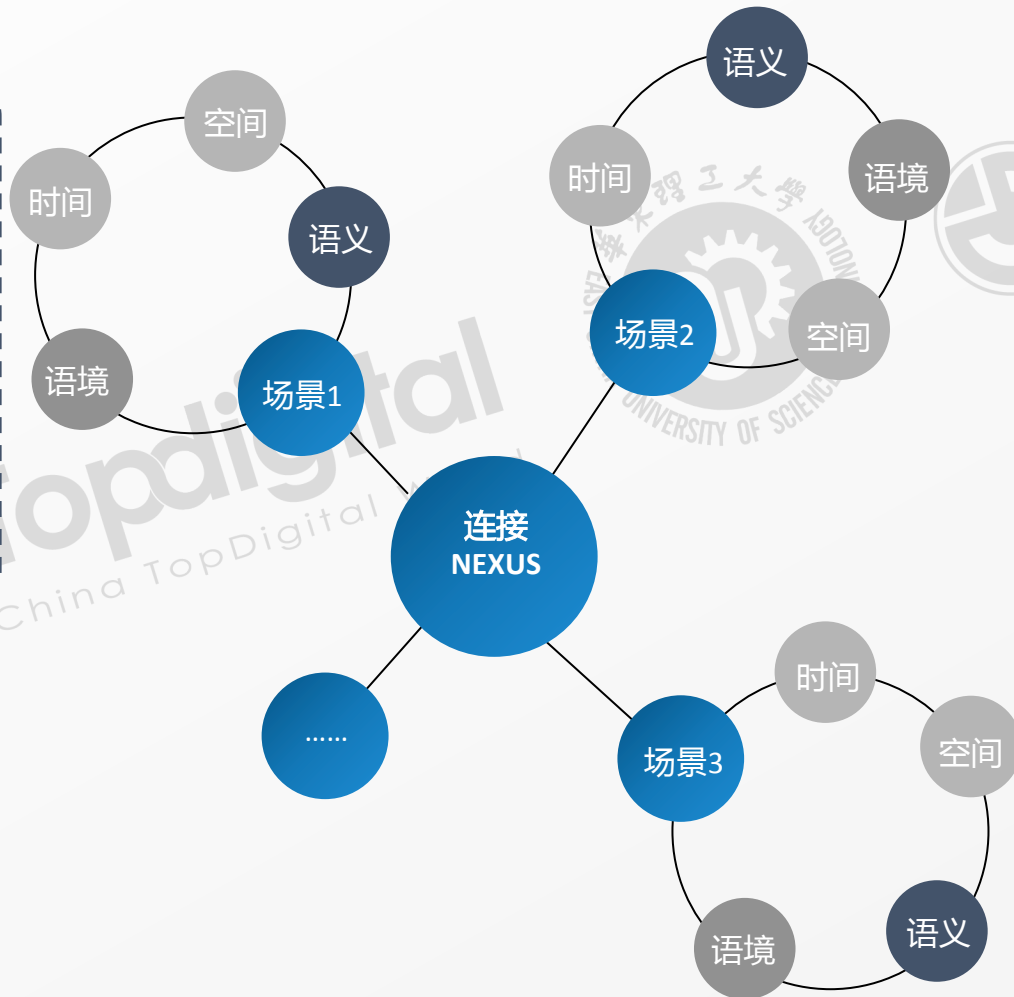
■ 一个完整的场景由时间、空间、语义、语境四要素组成。

全局数据：场景化

场景构建 / 场景连接 / 场景洞察

场景连接

- 网络与现实生活的不断融合，“人”在场景中所起的核心地位和移动设备的使用，使打通不同场景进行连接成为可能。这种连接创造的独特价值，能创造独特体验，促进消费，重构生活方式。



- 场景本身就是一种连接

- 场景连接可以是整个场景与场景的连接，亦可以是场景中的某个要素之间的连接。（如同一个人，在线上线下不同的场景中，藉由移动设备完成的线下扫码-线上支付的场景切换。）

- 场景连接与场景构建可以是同步进行的，即边进行不同情境下的场景构建，边使这些场景发生连接互动。

全局数据：场景化

场景构建 / 场景连接 / 场景洞察

场景洞察

- 场景洞察是建立在场景中产生的数据可被记录和分析的基础上。

- 而这些记录下来的数据，可以帮助我们形成最为清晰的用户画像；同时基于数据，我们可以进行场景的量化分析、综合洞察，更深入理解场景和场景中“人”的决策机制。

Situational
Awareness
态势感知

可量化

可分析

可复现

- 场景洞察使两个机制成为可能：

- 过去场景可追溯——完全了解过去场景人的决策机制
- 未来场景可感知——对未来进行动态、持续更新变化的“态势感知”，感知到未来各个场景的发展态势，从而根据需求，有针对性的构建和连接场景。

全局数据：开放性

01

- 数据的流动和交互是全局数据的基本特征

02

- 全局数据一定不是指同一体系内的数据，而是跨越、连接不同体系的数据。同时这种连接不是粗放的体系与体系之间、映射与映射之间的连接，而是场景与场景之间的连接

03

- 全局数据的开放性有两层含义：
1) 数据源的开放 2) 不同系统之间接口的开放



OMNILab

上海交通大学开放创新实验室

全局数据：开放性

数据源的开放 / 不同系统之间接口的开放

数据源的开放是全局数据“全局视野”的支柱之一,共包括四个层次的开放：

01

政府 Government

政府主导的数据库
建设和数据库开放

02

公共事业 Public Service

经个人同意，愿意公开分享
的非隐私性人口属性信息、
兴趣属性信息、行为信息等。

03

企业 Company

互联网、移动互联网公司、
物联网公司以及其他类型企
业共建的第三方开放数据库

04

个人 User

经个人同意，愿意公开分享的
非隐私性人口属性信息、兴趣
属性信息、行为信息等

“开放”不等同于“公开”：公开是一条条的、开放是一个系统一个系统的

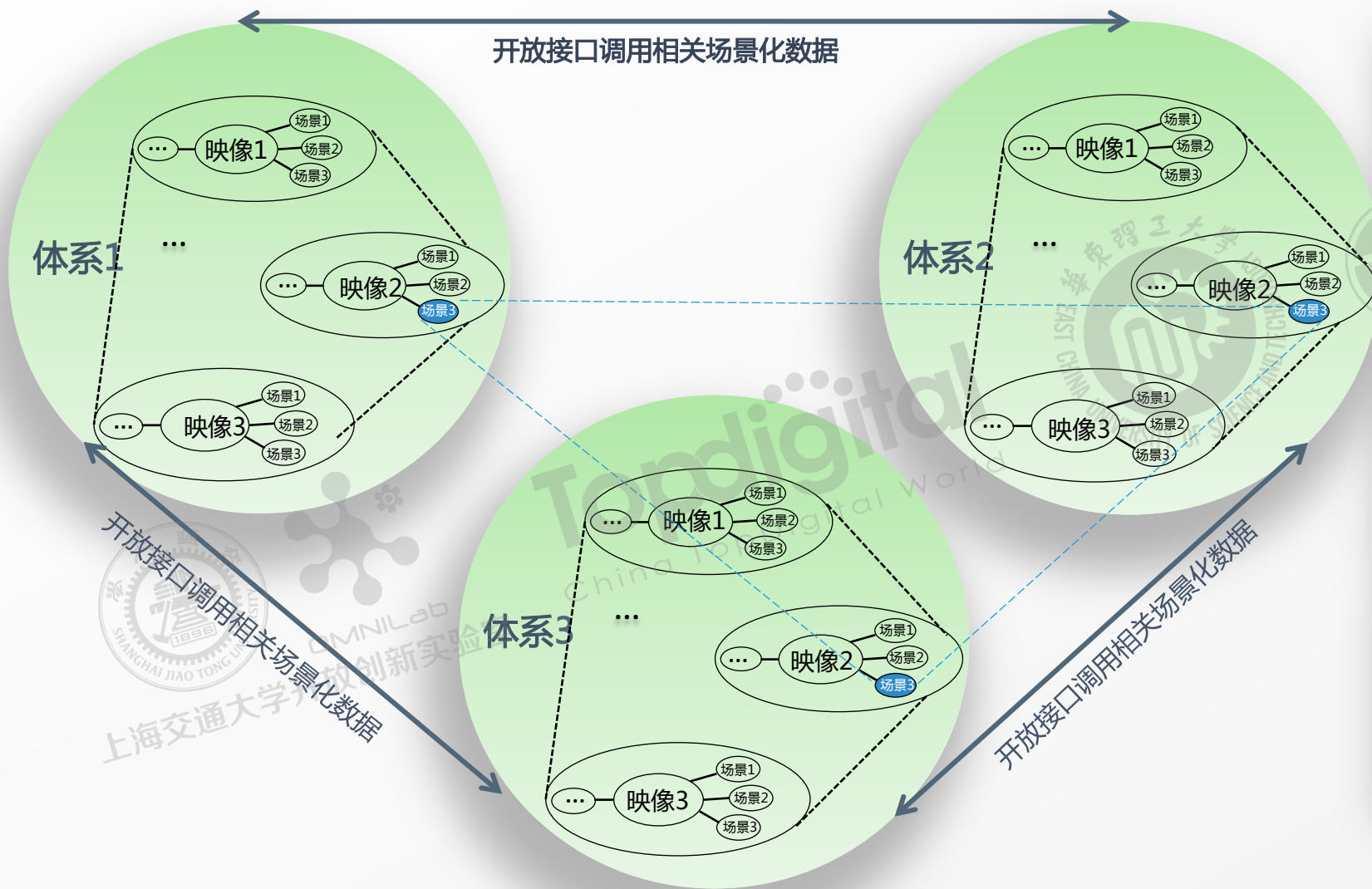
“公开”只是对人公开，开放的结果是让人通过机器更方便的理解数据

“开放”可以是有偿的：可以为数据开放支付费用

“开放”可以是部分的：可对部分人开放、可分层级开放

全局数据：开放性

数据源的开放 / 不同系统之间接口的开放



- 在不同体系之间通过开放接口调用场景化的数据进行匹配验证和应用。
- 这一方面保证了数据源方可度量数据（知道哪些数据被调用），保护了数据安全（数据的归属权仍在数据源方，应用方仅根据需求调用部分，调用的部分可以通过前期脱敏保证安全）。
- 一方面便利了数据应用方根据实际的场景化需求去匹配数据。同时，亦使即时处理数据成为可能。

全局数据：即时性



“即时性”的内涵

- 在线的即时智能数据收集，在入端即进行数据质量的管护。
- 在线的即时智能数据分析，根据结果反推分析机制。
- 在线的即时智能应用决策，数据价值即时发挥效用。



“即时性”先决条件

- “万物互联”，“时刻在行动中产生数据”将成为数据的新常态。
- 不同体系之间的接口开放
- 这两个先决条件使数据即时被收集、即时被分析、即时被应用决策成为可能。



“即时性”的意义

- 即时性使全局数据不再是固化的“资产”，而是活动性的“生产力”。
- 即时性使全局数据成为“活数据”
- 即时的数据 = 即时的认知 = 即时的洞察

全局数据：可度量、价值化

A



可度量

当体系与体系之间的开放是由标准化技术接口完成的时候，数据的收集就是完全可度量的。这包括：**数据收集标准的可度量、数据收集维度的可度量**。数据拥有方完全可以定义给出的数据标准、维度，最大限度保护隐私和安全。

B



价值化

数据的应用效果可度量则是指，根据全局数据得出的结论和决策是直接能够导致可量化的优化发生的。而这种可量化的优化就直接带来了一个影响：**全局数据是价值化的数据，能直接带来社会价值、商业价值、科学价值的提升。**

全新的概念：OMNI DATA

能力：收集数据的能力

治理数据的能力

应用数据的能力

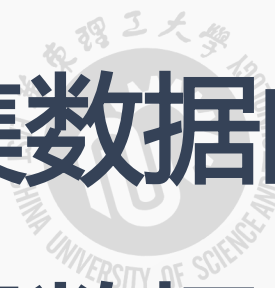
全局数据是被赋能的数据



OMNILab

上海交通大学开放创新实验室

Topdigital
China TopDigital



全局数据三大能力

收集数据的能力 / 治理数据的能力 / 应用数据的能力

- 收集数据就是连接决策相关的各种场景。
- 连接决策相关的场景化数据意味着连接用户与用户、产品与产品、行业、企业与政府。
连接与决策相关的每一个场景因素。
- 同时，全局数据的收集数据是通过各个体系间的开放接口进行的。



全局数据三大能力

收集数据的能力 / 治理数据的能力 / 应用数据的能力

理想中的全局数据来源：

各个源头的开放数据源（社交网、物联网、开放数据源）通过标准化接口联合汇入，构成基础数据层（Landing Area-Raw machine data collection, collect everything）



“The contents of the data lake stream in from a source to fill the lake.....”

全局数据三大能力

关于数据湖：

收集数据的能力 / 治理数据的能力 / 应用数据的能力

DATA

2010年，Pentaho 的 CTO James Dixon介绍了“数据湖Data Lake”的概念并且影响至今。Dixon认为数据湖是一种数据仓库的架构，他如此描述：

“如果你把数据市集看做一个贩卖瓶装水的商店——贩卖清洁的、被包装过的、标准化的可被轻易销售的产品——数据湖则是一个在更自然状态的大体量水体。数据的不同支流源源不断的丰富着湖泊，而各种各样的用户都可以到湖水中检验、潜水，或者抽取样品。”



“

If you think of a datamart as a store of bottled water – cleansed and packaged and structured for easy consumption – the data lake is a large body of water in a more natural state. The contents of the data lake stream in from a source to fill the lake, and various users of the lake can come to examine, dive in, or take samples.”

LAKE

全局数据三大能力

收集数据的能力 / 治理数据的能力 / 应用数据的能力

数据管理协会(DAMA)：数据治理是针对数据资产管理的具体实践，包括授权、控制和共享的决策（计划、监控和强制执行）等。数据治理包括了规则的制定、流程、工具以及实施审计数据资产管理相关的制度等。

数据是资产，通过服务产生价值，数据治理是在数据产生价值的过程中，治理团队对其的评价、指导和控制。（《数据治理白皮书》）

数据治理是确保数据质量最有效的方式。数据质量定义为数据的“适用性（fitness for use）”，即数据满足使用需要的合适程度。

我国数据治理分为三个层面：
体制层面：业务条线的纵向数据管控推动力、执行力
管理实践层面：数据标准建设、指标管理，一般归于主数据管理和业务元数据管理
技术层面：数据资源管理平台、定制化开发、一站式数据管理（《数据治理白皮书》）

全局数据三大能力

收集数据的能力 / 治理数据的能力 / 应用数据的能力

然而 并非所有的数据都具有价值，混杂的数据表现出“乱杂错丢骗”的特征，给数据治理带来难度。

乱 数据的无序性

- 采集标准乱
- 存储格式乱
- 应用机制乱
- 身份识别乱
- 时、空、身份、场景对应乱

杂 数据的非标准性

- 结构化数据和非结构化数据并存
- 非结构化数据的大小、内容、格式、用途可能完全不同，但非结构化数据的体量占据数据总量80%以上

错 数据的错误

- 数据本身的错误
- 采集手段的错误
- Outlier *

丢 数据的丢失遗漏

- 机械因素导致*
- 系统因素导致*
- 人为因素导致*

骗 伪造数据

- 基于商业或其他目的的制造的伪数据，故意导向错误结果

*Outlier：指在同样的采集方式和渠道下，数据中有一个或几个同类数值与其他数值相比差异较大，而差异究竟是由于随机张罗还是其他因素引发难以判断，如果处理不好将会造成系统误差。

*机械因素：由于机械原因导致的数据收集或保存的失败造成的数据缺失，比如数据存储的失败，存储器损坏，机械故障导致某段时间数据未能收集（对于定时数据采集而言）。

*系统因素：由于系统设置本身的缺失而带来的数据收集采集缺失。

*人为因素：由于人的主观失误、历史局限或有意隐瞒造成的数据缺失

全局数据三大能力

收集数据的能力 / 治理数据的能力 / 应用数据的能力

全局数据下的数据治理其实是一种体系，是一个关注于信息系统执行层面的体系，这一体系的目的是整合知识和意见，通过将流程、策略、标准和组织的有效组合，对信息化建设进行全方位的监管。

数据治理也是数据从基础层到可被智慧洞察运用的全局数据仓库的进化。在这个过程中，数据的入端通过治理保证了质量和管护，在出端保证了隐私和安全，保证了数据的时效性、可共享性、高质量、一致性，解决了数据的“乱杂错丢骗”。

解决数据的“乱杂错丢骗”才能保证真实的关联场景化连接，才能保证数据应用能是真正场景化、可度量、有价值的应用。

全局数据三大能力

收集数据的能力 / 治理数据的能力 / 应用数据的能力

全面的数据治理

4 所有需要使用的人都可实现自由查询与报告抽取

3 通过数据整理、机器学习、外部数据融合和拓展数据仓库来实现敏锐快捷的商业洞察

全局数据仓库

数据科学实验区

Data Lake —整合离散数据

Landing Area —全保真度的源数据

2 数据准备好可以变为信息：可被组织的、完善定义的、全面的

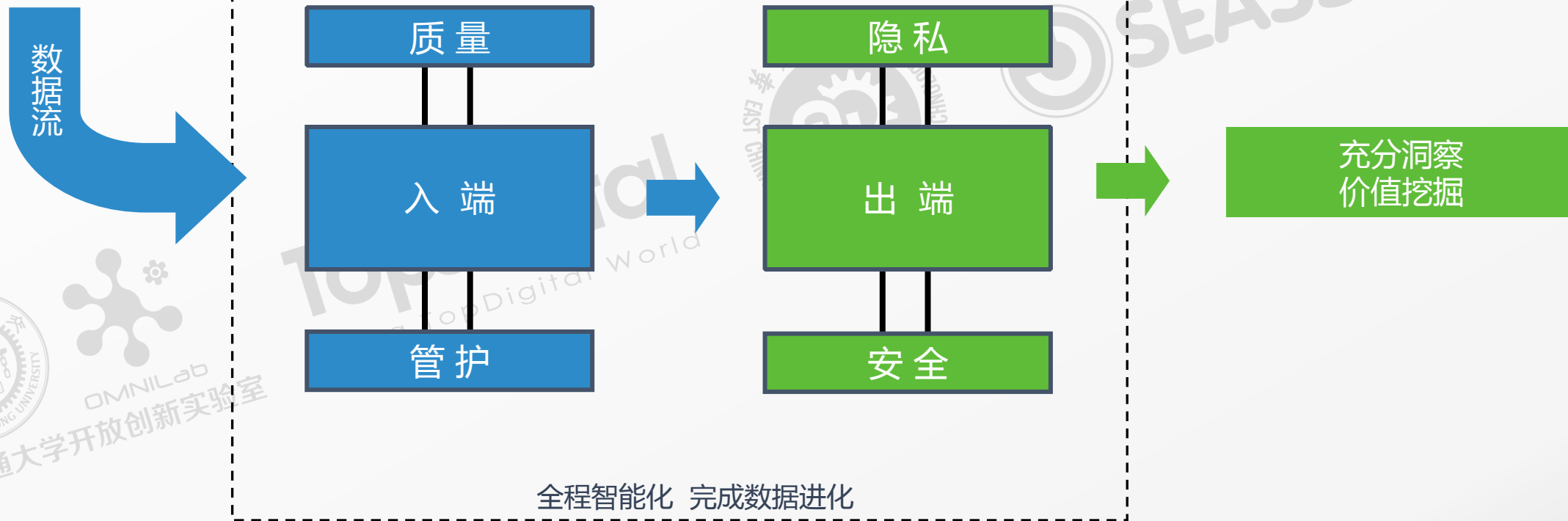
1 原始机器数据收集，收集一切可能获得的数据

全局数据理论体系下的“数据”，并不仅仅停留在收集状态，而包含了数据的“治理”这一重要概念。从全保真度的源数据、整合离散数据成为信息的数据湖、数据科学实验区到全局数据仓库之间，是有一个上升流动通路的。这一通路使处在自然状态的源数据**进化为**蕴含智慧规律、可被迅疾识别解读观点（无论是人还是机器）的全局数据仓库。而通路的动力，即为“治理”。

全局数据三大能力

收集数据的能力 / 治理数据的能力 / 应用数据的能力

- 入端保证数据质量和管护，出端保证数据隐私与安全
- 数据治理导向的是全局数据的充分洞察和价值挖掘



- 数据的充分治理意味着在入端就保证质量和管护，在出端保证隐私和安全，全程智能化完成数据进化，最后呈现出的是可被充分洞察和价值挖掘的数据。

全局数据三大能力

收集数据的能力 / 治理数据的能力 / 应用数据的能力

- 全局数据的数据应用是场景化的、可度量的、价值化的应用。

场景化的

正如前文所述，全局数据应用建立的是所有与决策相关的场景之间的连接，和之间的规律挖掘。

可度量的和价值化的

数据的获取通过API，入端可度量。应用按你的直接价值体现（收益）进行按比例收费，直接创造社会价值、商业价值、科学价值。

- 全局数据的应用场景涉及到生活的方方面面，小到消费决策，大到政府决策都可依赖与此。
- 全局数据概念下的应用需要能通过智慧化、可视化的技术揭露海量数据中隐藏的知识，让数据中的智慧能够以一种直观的形式流向决策者，无论决策者是人还是机器。
- 过去不再是不可追溯的，而是可以总结规律的。未来不再是混杂无序的，而是可被预测把控的。“日光之下，并无新事”。

全新的概念：OMNI DATA

界限：不危及国家安全

不侵犯公民隐私

不违背个体意愿



OMNILab

上海交通大学开放创新实验室

Topdigital
China TopDigital



SEASCOON

全局数据三大界限

不危及国家安全

不侵犯公民隐私

不违背个体*意愿



OMNILab
上海交通大学开放创新实验室

TopDigital
China's TopDigital World



SEASSOON

*个体包含了个体的“人”和作为“个体”的机构和企业。个人、机构和企业都有能力、渠道、权利去管理和自身相关的信息，可以决定什么时候公开，以及用什么方式公开信息。对自身的数据要有控制权。但我们并不提倡个人、机构和企业把数据视为自己的私有资产，我们倡导的是一种遵从个体意愿，同时有选择的公开部分数据，帮助整个体系进行全局数据价值挖掘和视野进化的理念。

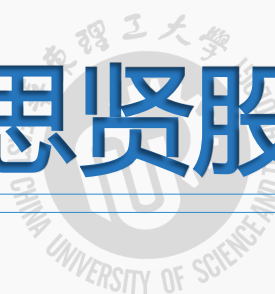
04

关于思贤股份



OMNILab
上海交通大学开放创新实验室

Digital
China Digital World



SEASSOON

关于思贤股份

上海思贤信息技术股份有限公司，成立于2010年。专注以全局数据为基础，为政府、行业客户提供智能洞察方案及服务。

思贤股份的数据来源于以下渠道：

- 与运营商合作收集的用户网络行为数据；互联网资讯数据；合作企业内部数据（仅针对合作企业本身进行挖掘运用，不做公开）；行业解决方案数据；政府开放数据。
- 最大限度的收集与企业决策相关的所有数据，并以强劲的技术实力保证，对数据进行充分的治理，解决数据的“乱杂错丢骗”，并致力对数据进行全局性洞察，导向智能化、价值化、可衡量的决策应用。

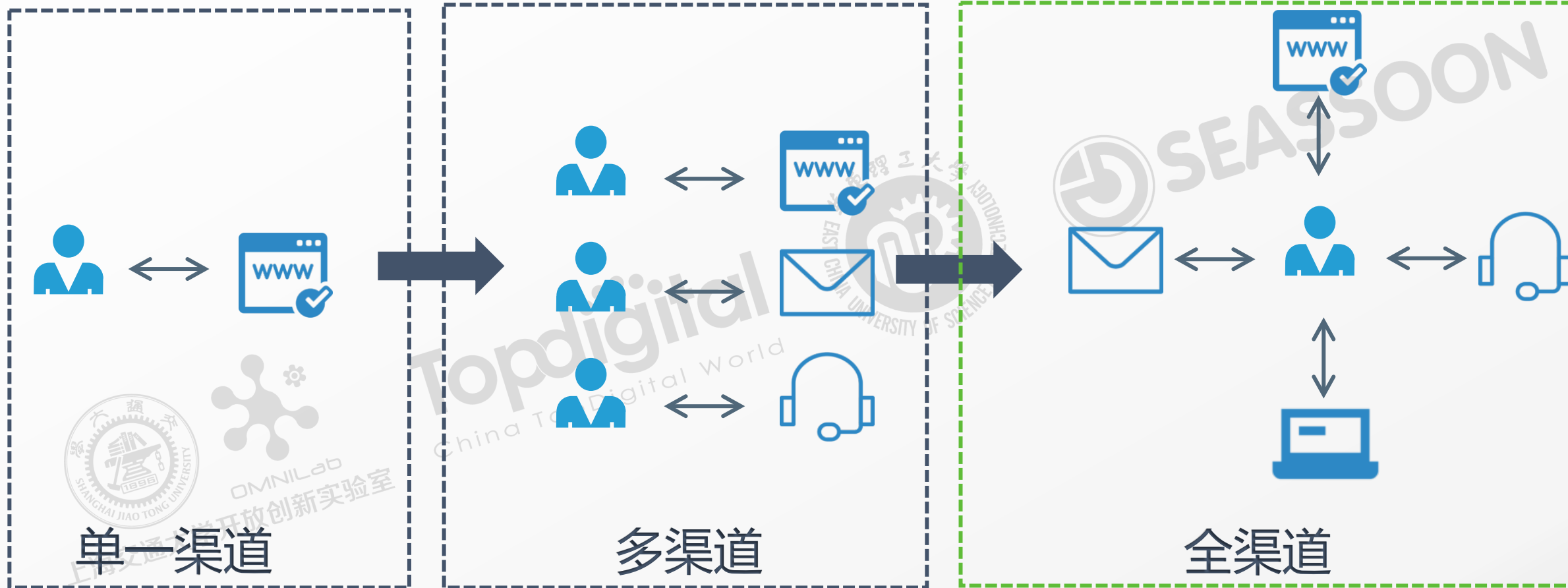
思贤股份目前有三大业务版块：

- 政府决策咨询智库：与政府部门合作，提供数据挖掘、洞察服务，是政府决策咨询智库。
- 多维全息用户画像：基于多维全息用户画像，为企业提供精细化营销、业务运营监控和用户体验优化服务。
- 企业级数据服务：为企业进行数据挖掘和智能洞见，提供旨在提升企业运营效率和营收的企业级数据产品，并进行行业咨询研究。

思贤全局数据收集

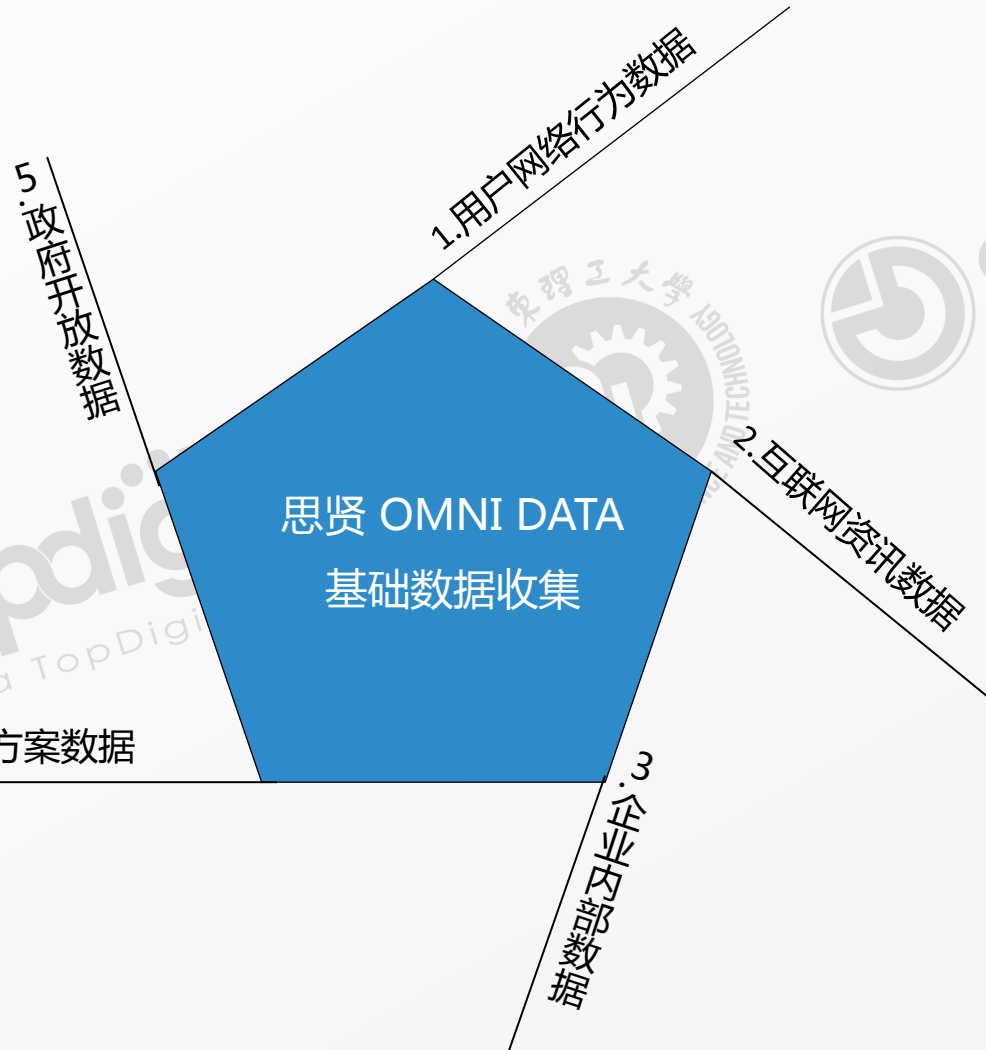
其他

思贤

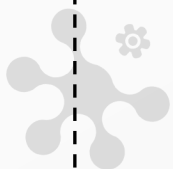


思贤全局数据收集

技术保证实时快速收集反应



上海交通大学开放创新实验室



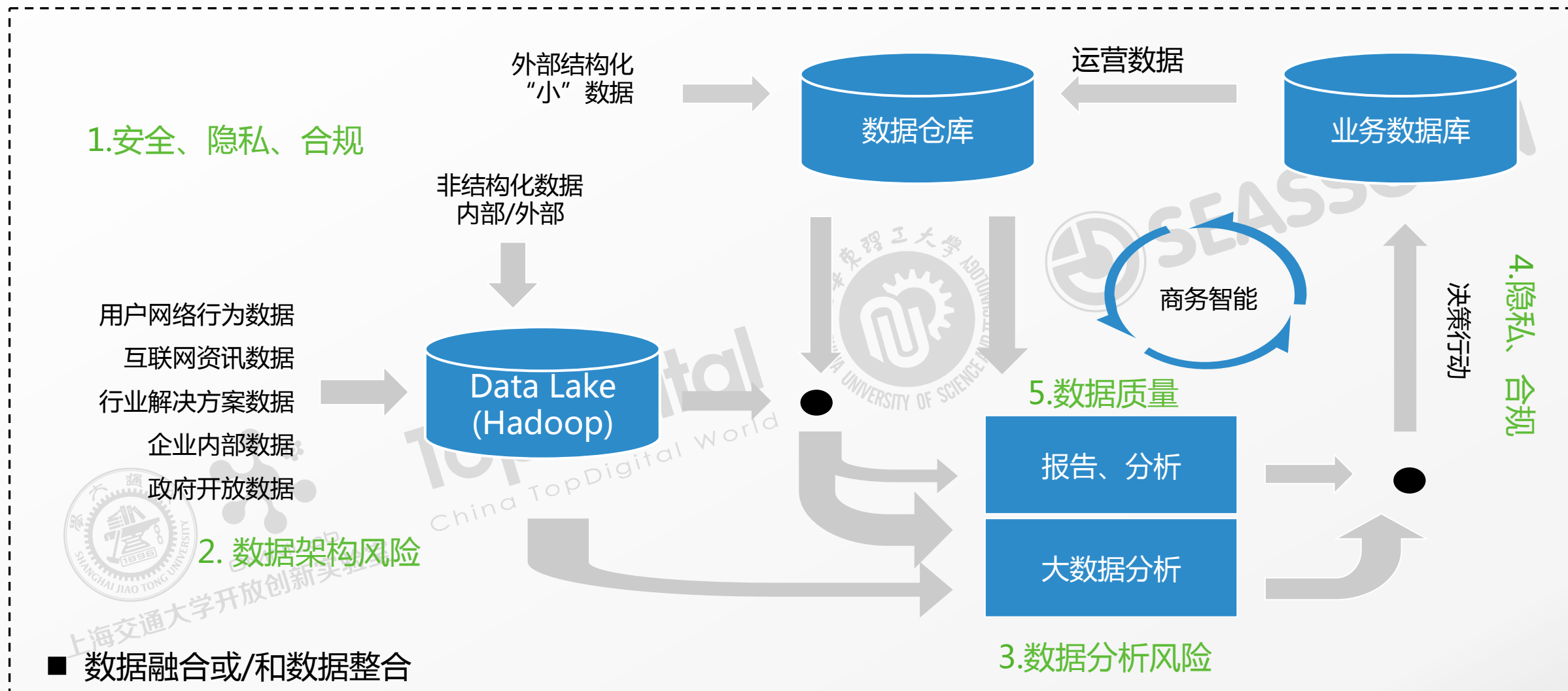
OMNILab

Topdig
China TopDigi



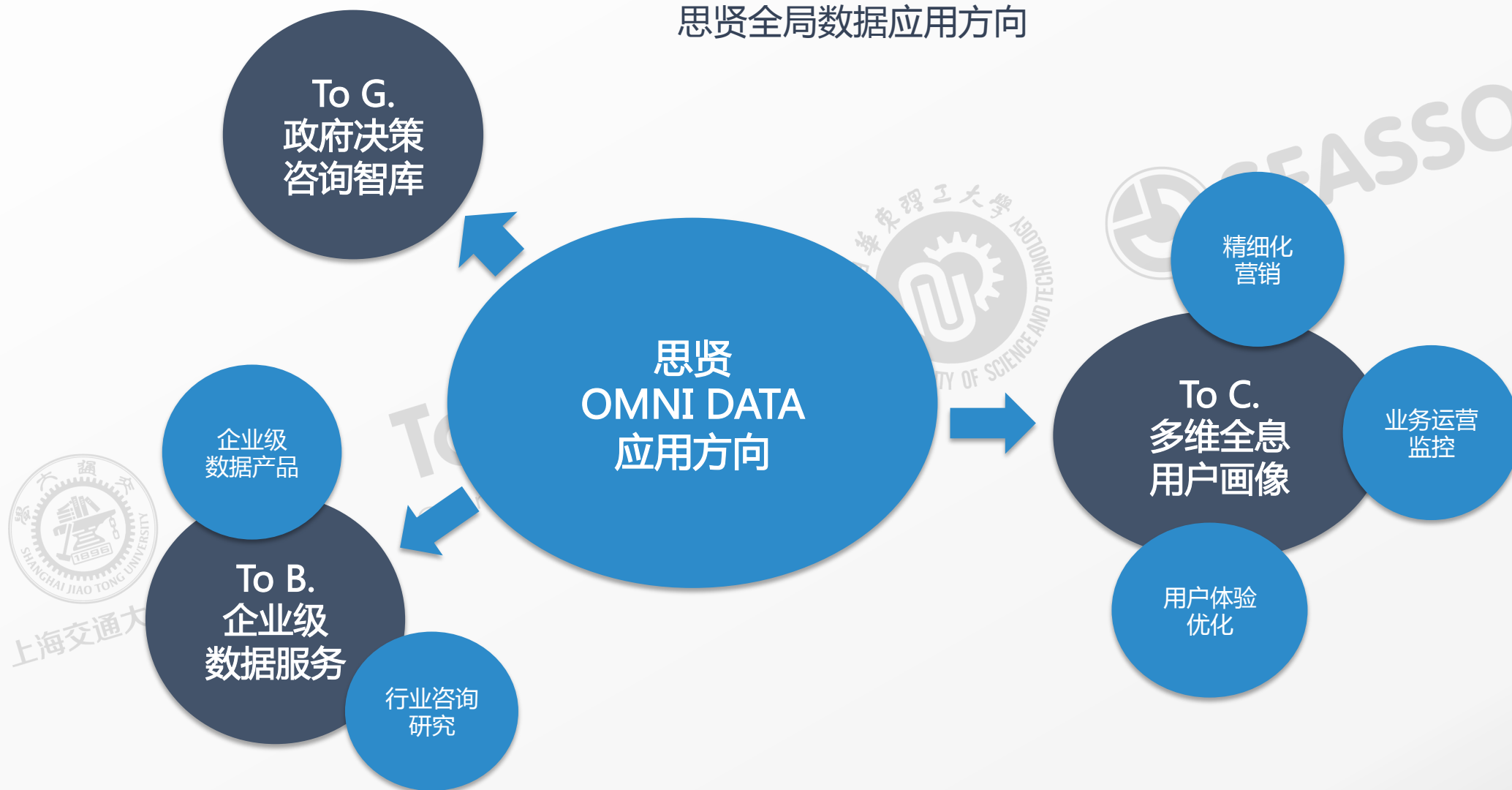
SEASSOON

思贤全局数据治理

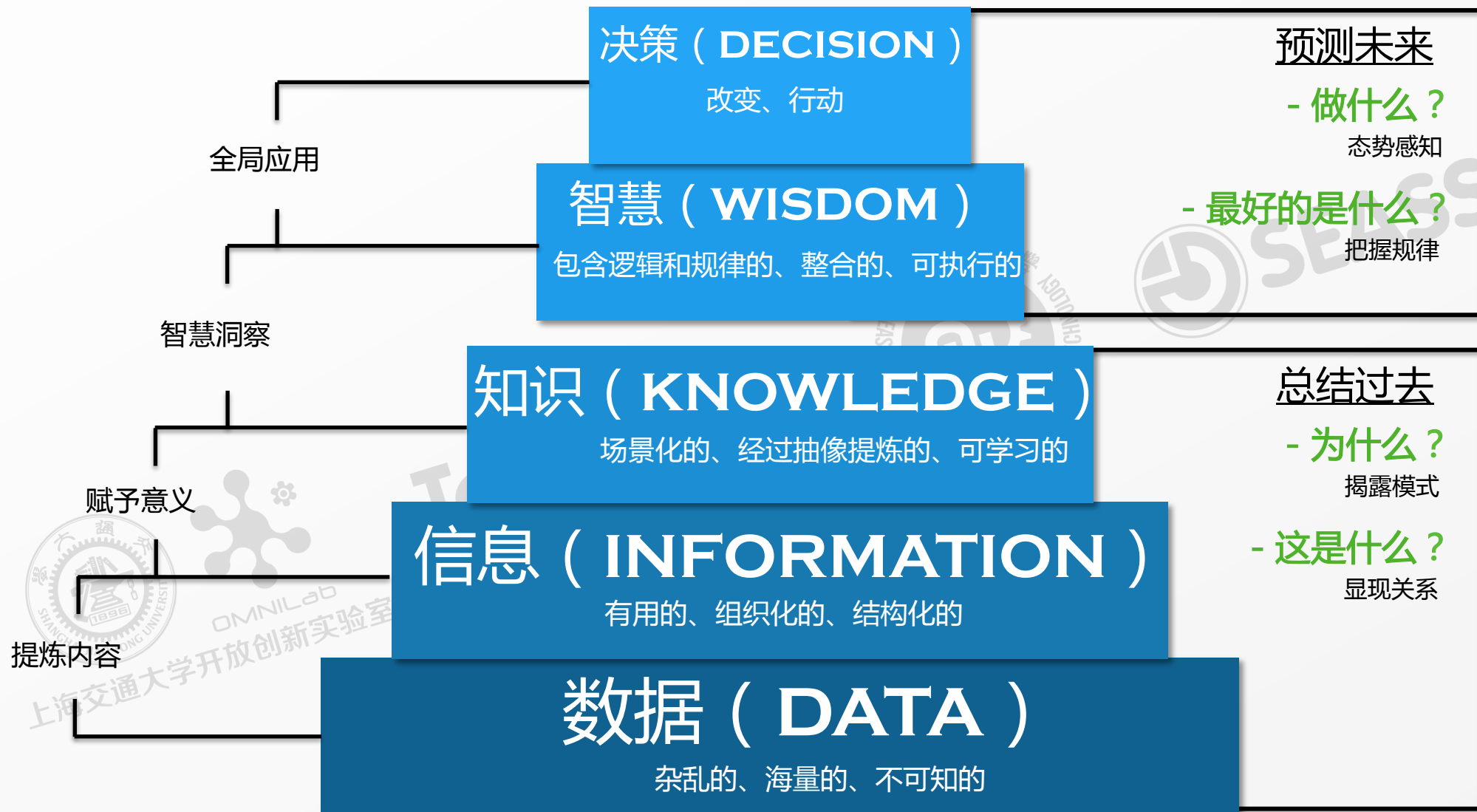


思贤全局数据应用

思贤全局数据应用方向



思贤全局数据应用



原创发布



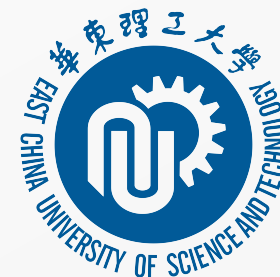
研究支持



学术支持



上海交通大学数联开放创新实验室



转载请注明出处，违者必究！最终法律解释权归思贤股份所有

OMINI DATA

全局数据



OMNILab

上海交通大学开放创新实验室

Topdigital
china digital world



SEASOON