2022 MathWorks
中国汽车年会

大数据分析中工程专家和数据科学家的合作模式

李勇, 康明斯（中国）投资有限公司

MathWorks

内容

- 康明斯在中国
- 数据科学的能力要求
- 技术专家和数据科学家的“冲突”
- 技术专家和数据科学家的合作模式
- 大数据分析文化的建设

康明斯在中国

1975

与中国建立商业联系

12,000+

分布全国的员工

640K

2021中国区发动机产销量

(含发电机组及出口市场)

2,000+

授权经销商&代理商

3+2

3家技术研发中心（武汉、无锡和重庆）；
2个工程中心

∞

康明斯全球规模
长最快的海外市

此列表为不完全统计的市场，更多产品信息请访问www.cummins.com

DFM · FOTON 福田汽车 · CQME · LiuGong

JAC 江淮汽车 · TIANYUAN · Valvoline

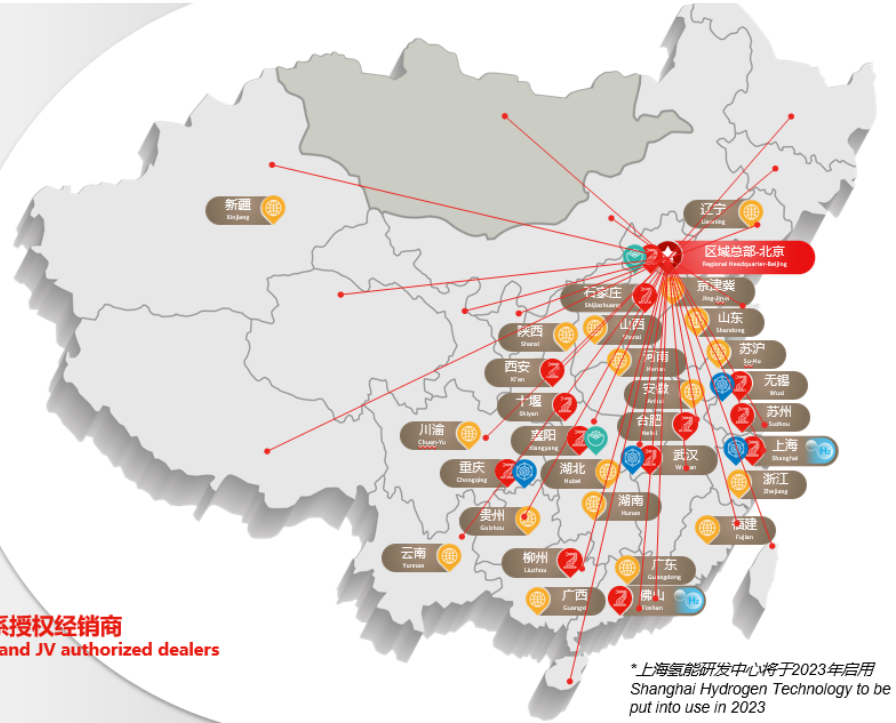
兴源控股 XINGYUAN HOLDINGS · XCMG · YUTONG · 中国重汽 SINOTRUK

SANY · ZOOMLION · Sinopec

** 本页列出的仅为部分本地合作伙伴，排名不分先后。*

康明斯中国全国机构分布
Cummins Entities in China

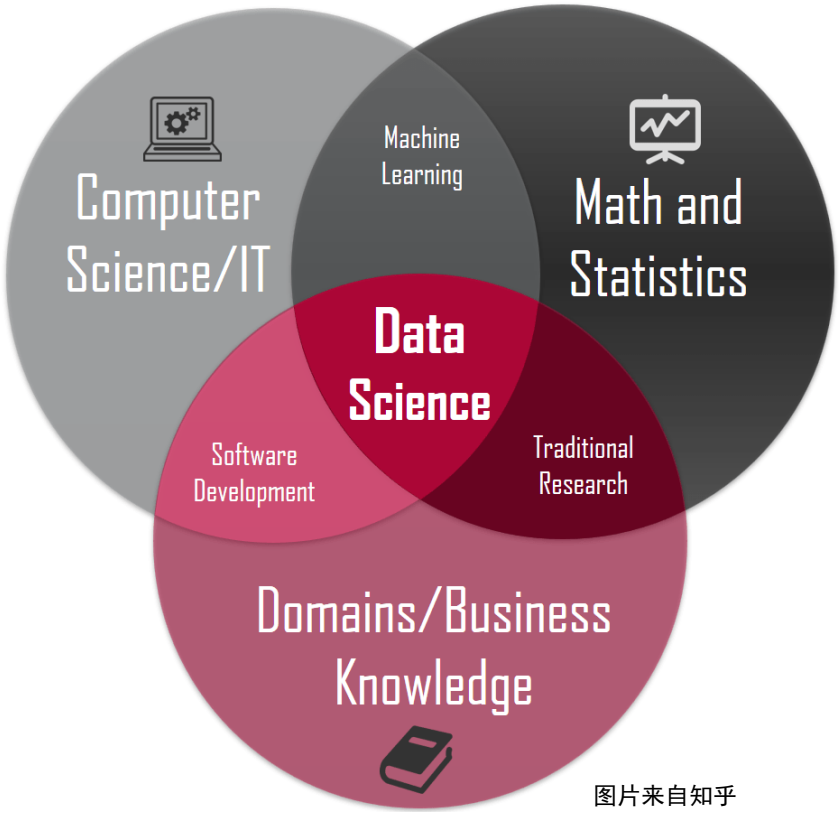
- 23家在华生产基地 / 23 Manufacturing & Operating Plants
- 4家技术研发中心 / 4 Tech Centers
- 2家工程中心 / 2 Engineering Centers
- 18家康明斯省级客户支持中心 / 18 Cummins Customer Support Hubs



3000多家康明斯及合资体系授权经销商
3000+ Cummins wholly-owned and JV authorized dealers

** 上海氢能研发中心将于2023年启用*
Shanghai Hydrogen Technology to be put into use in 2023

数据科学（大数据分析或建模）的能力要求



图片来自知乎

- 机器学习 = 数学&统计 + 计算机科学（编程）
- 传统的技术研究 = 数学&统计 + 领域/业务知识
- 数据科学 = 数学&统计 + 领域/业务知识 +计算机科学（编程）

技术专家和数据科学家的“冲突”

基于理论/原理的建模
First Principles Modeling

数据驱动的建模
Data-Driven Modeling

- 业务知识
- 领域工程经验 Know-How
- 编程能力 – MATLAB, VBA, C,…

- 机器学习
- 大数据技术 – Cloud, 数据湖, 并行计算,…
- 编程能力 – MATLAB, Python, R, …

真实环境促进了合作

基于理论/原理的建模
First Principles Modeling

数据驱动的建模
Data-Driven Modeling

- ◆ 物联网的数据频率低
- ◆ 物联网的数据质量差
- ◆ 云计算的优势

- ◆ 数据维度、样本数据的少
- ◆ 领域知识缺乏
- ◆ 边缘计算的优势

技术专家和数据科学家的合作模式

- 工作坊式的

图片来自图行天下

- 流水线式的（物联网场景）

图片来自51yuansu.com

工作坊式合作模式

参数选择/特征工程

常见挑战：

1. 基于算法的特征工程不能发挥作用；
2. 原始数据需要进一步计算处理；
3. 物联网特有的数据质量问题。

合作中各自的贡献：

- 工程专家：依据系统原理指导参数选择或数据再加工；
- 数据科学家：从数字化产品角度设计数据的预处理（流式计算和批处理）。

模型开发和训练

常见挑战：

1. 理解业务需求的过程；
2. ROC曲线的落点；
3. 模型性能和成本之间的平衡。

合作中各自的贡献：

- 工程专家：数据科学家和业务代表之间的桥梁；
- 数据科学家：高性价比算法的选择。

道路测试和验收

常见挑战：

1. 道路试验方案的设计；
2. 验收标准的确定；
3. 加速验证方案（如果必要）。

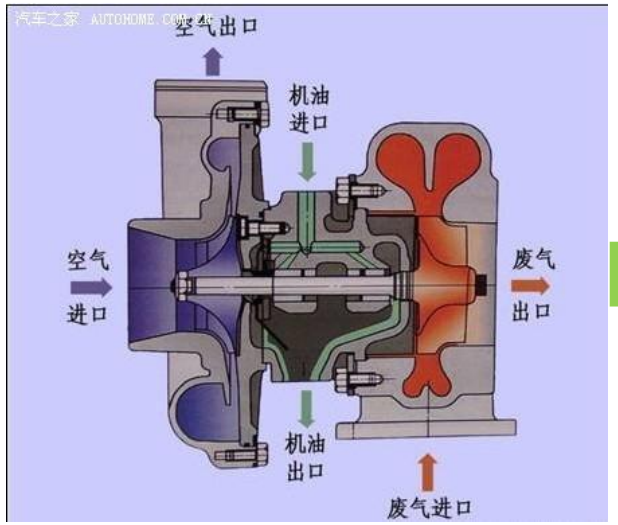
合作中各自的贡献：

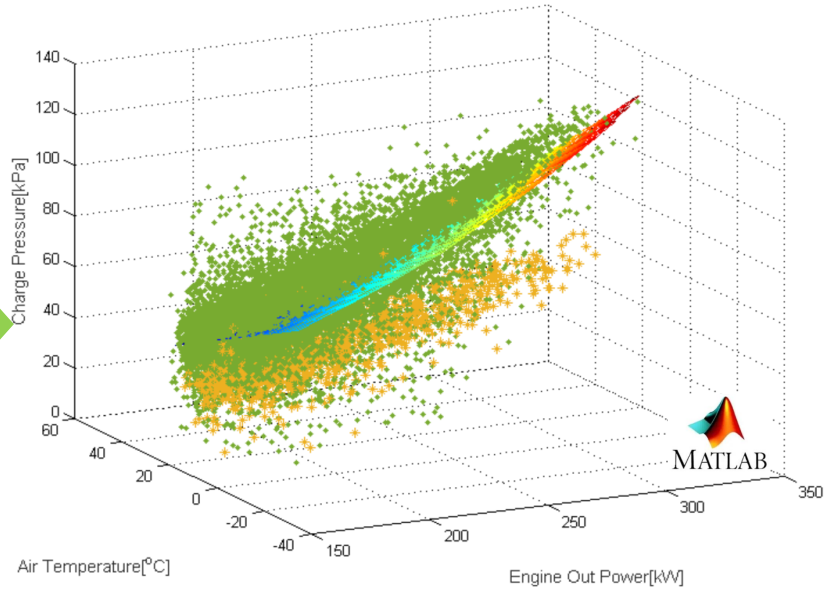
- 工程专家：领域的技术知识帮助实验设计的落实；
- 数据科学家：统计或数学知识确保折衷的合理性。

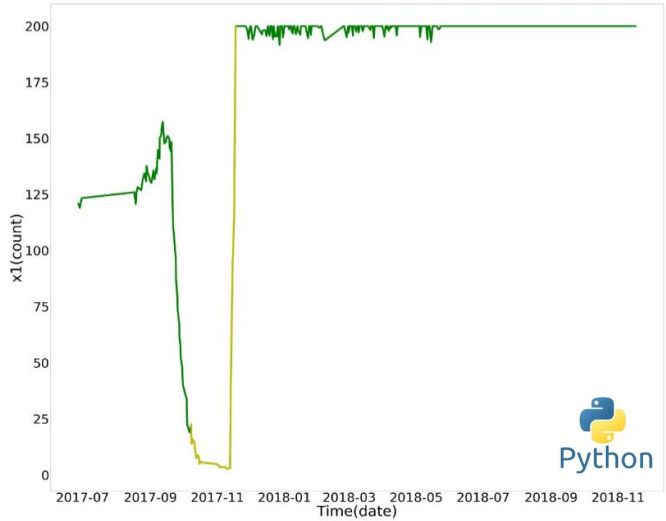
工作坊式合作模式的案例

柴油机增压器故障预测

- 面临的挑战
  - ✓ 只有大约几十个失效案例
  - ✓ 只有少数参数和增压器有关系，如增压压力
  - ✓ 分钟级的采样频率



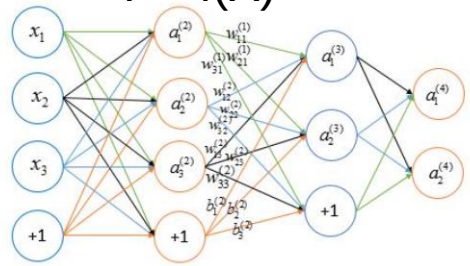




流水线式的合作模式

传统方式： 通过物联网把信号数据都采集到云端后再开发数据模型

Y = f(X)

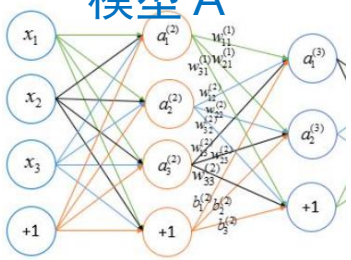


车辆控制模块

车联网终端

云平台

模型 A



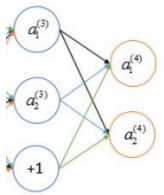
U = f(X)

MATLAB SIMULINK

新的机会:

- 完整的模型 = 模型 A + 模型 B
- 模型A放置在控制器或车联网终端中，以较低的成本使用高频和多维数据
- 模型B在云端使用模型A处理后的低频高信息含量数据以获得较好的性能
- 技术专家在边缘端基于原理主导模型A的开发
- 数据科学家后续在云端基于机器学习主导模型B的开发

模型 B



Y = f(U)

MATLAB or Python

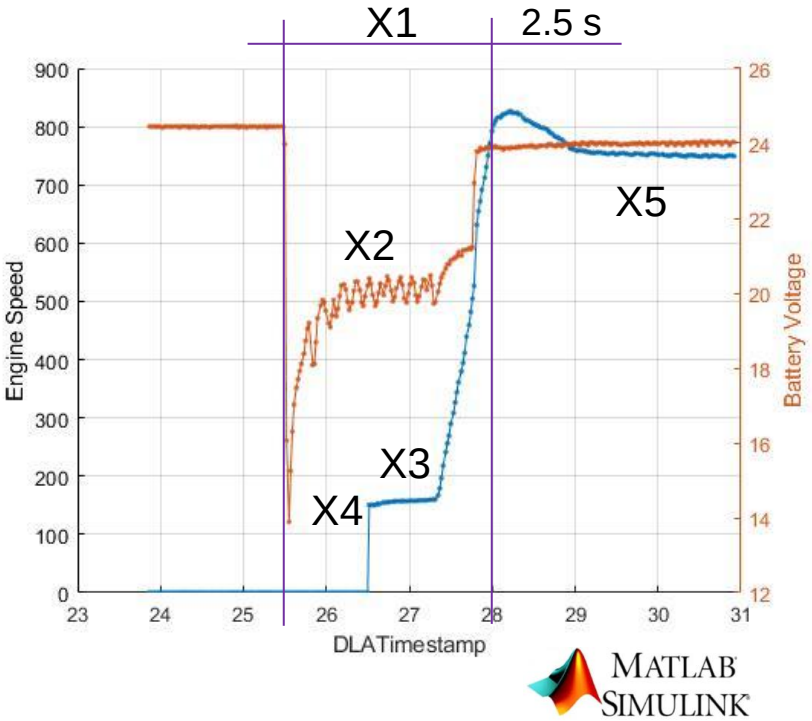
流水线式合作模式的案例

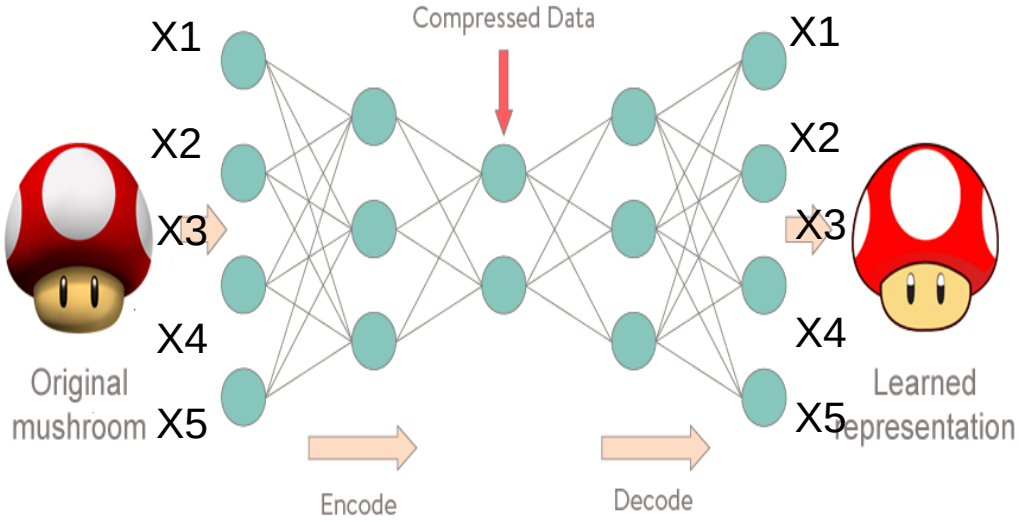
- 以启动系统性能监测为例

ECU

车联网终端

云平台

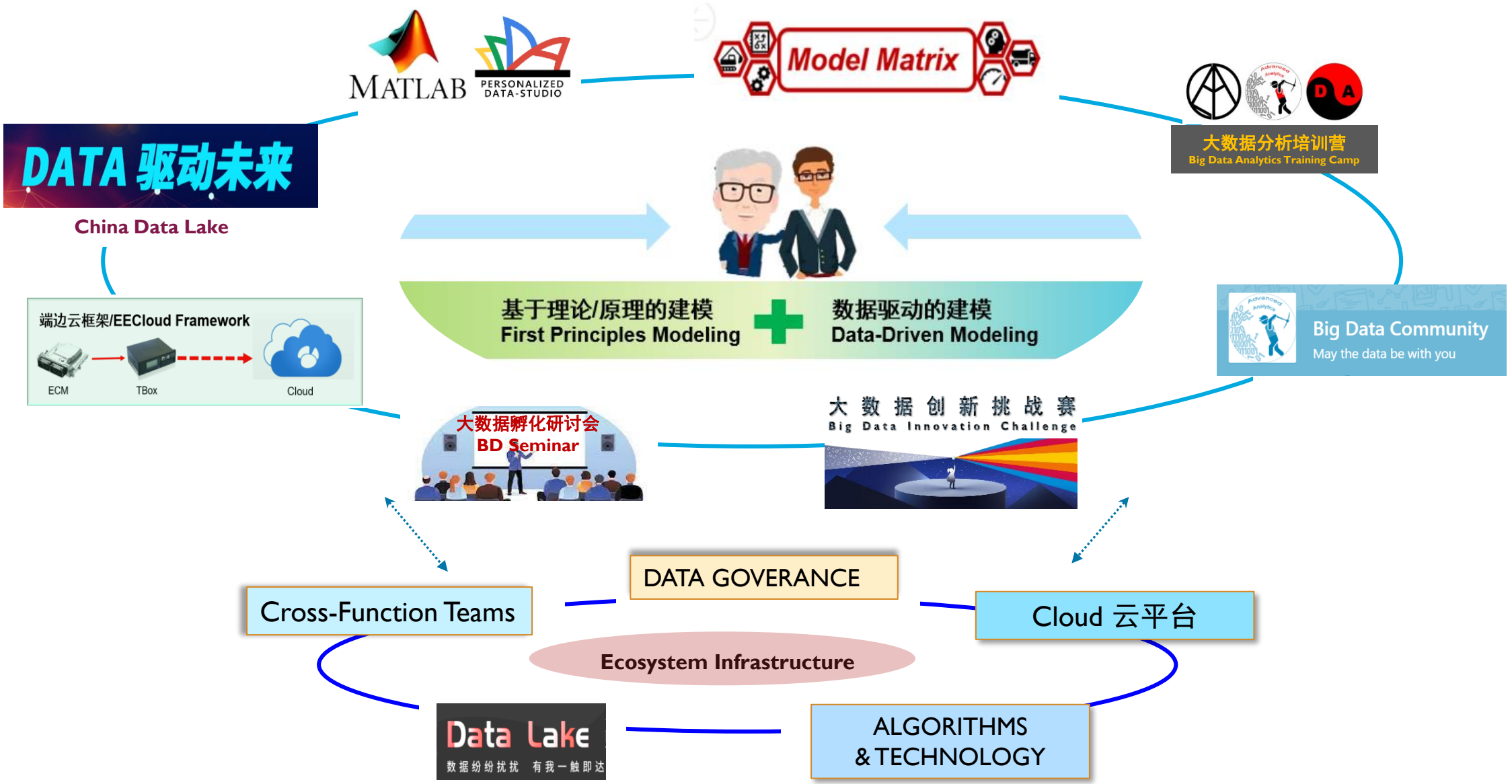




Auto-Encoder 自编码器

MATLAB or Python

大数据分析文化的建设


MATLAB
PERSONALIZED DATA-STUDIO
Model Matrix
DATA 驱动未来
China Data Lake
大数据分析培训营
Big Data Analytics Training Camp
基于理论/原理的建模
First Principles Modeling
数据驱动的建模
Data-Driven Modeling
端边云框架/EECloud Framework
ECM
TBox
Cloud
Big Data Community
May the data be with you
大数据孵化研讨会
BD Seminar
大 数 据 创 新 挑 战 赛
Big Data Innovation Challenge
DATA GOVERANCE
Cross-Function Teams
Cloud 云平台
Ecosystem Infrastructure
Data Lake
数据纷纷扰扰 有我一触即达
ALGORITHMS
& TECHNOLOGY

2022 MathWorks
中国汽车年会

Thank you