CAICT 中国信通院

人工智能数据集工作介绍

中国信息通信研究院人工智能研究所

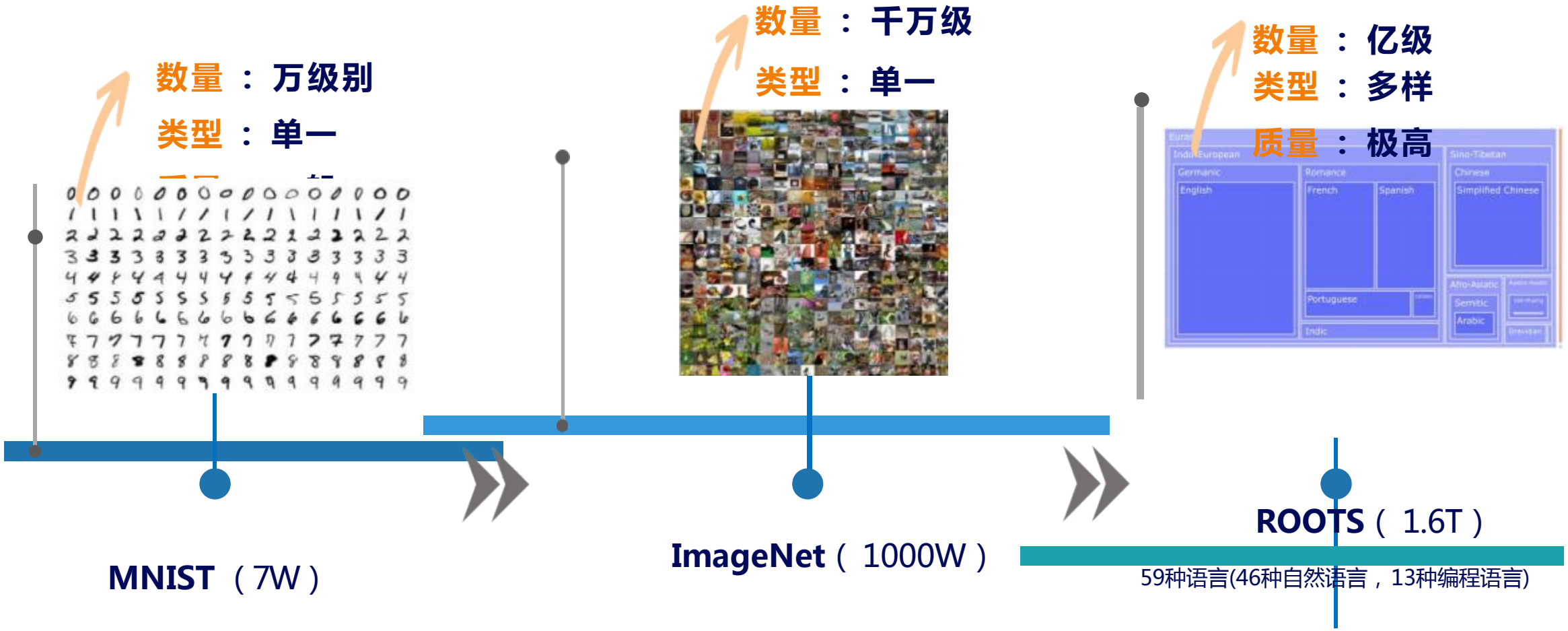
人工智能关键技术和应用评测工业和信息化部重点实验室

2024年4月

大模型时代，数据成为新一轮人工智能竞争的壁垒与制高点

CAICT 中国信通院

- 人工智能每次阶段性的进步，数据都扮演着重要角色，尤其在大模型时代，**海量、高质量、多样化**的训练数据集，成为拉开能力差距的关键要素。



浅层学习时期

（ ~2012 ）

深度学习时期

（ 2012~2018 ）

预训练模型时期

（ 2019~ ）

2022年起，人工智能正在由“模型中心”转向“数据为中心”

CAICT 中国信通院

- 2022年产学研提出**“以数据为中心的人工智能”（ Data-centric AI ）**，高质量的训练数据集、完备的数据应用策略将会更好的服务于模型的开发与应用。
- 人工智能领域的权威学者**吴恩达，**发起了**“以数据为中心的 AI”** ，即在模型相 对固定的前提下，通过提升数据的质量和数量来提升整个模型的训练效果。
- 通过**添加数据标记、清洗和转换数据、数据缩减、增加数据多样性、持续监 测和维护数据**等手段，形成优质的标准化数据集和完备的数据全生命周期管理体系。

李飞飞团队：实现可信AI，数据的设计、完善、质量评估是关键

吴恩达：80%的高质量数据与20%的模型训练构成了更好的AI模型。

- 2021年举办了首届“以数据为中心的人工智能竞赛”，比赛仅允许通过改进数据来提升模型的性能。

80%的高质量数据与20%的模型训练构成了更好的AI模型。

数据集贯穿于大模型全生命周期，输出高质量数据集

CAICT 中国信通院