Hewlett Packard
Enterprise

Adopting an Effective LLM Platform

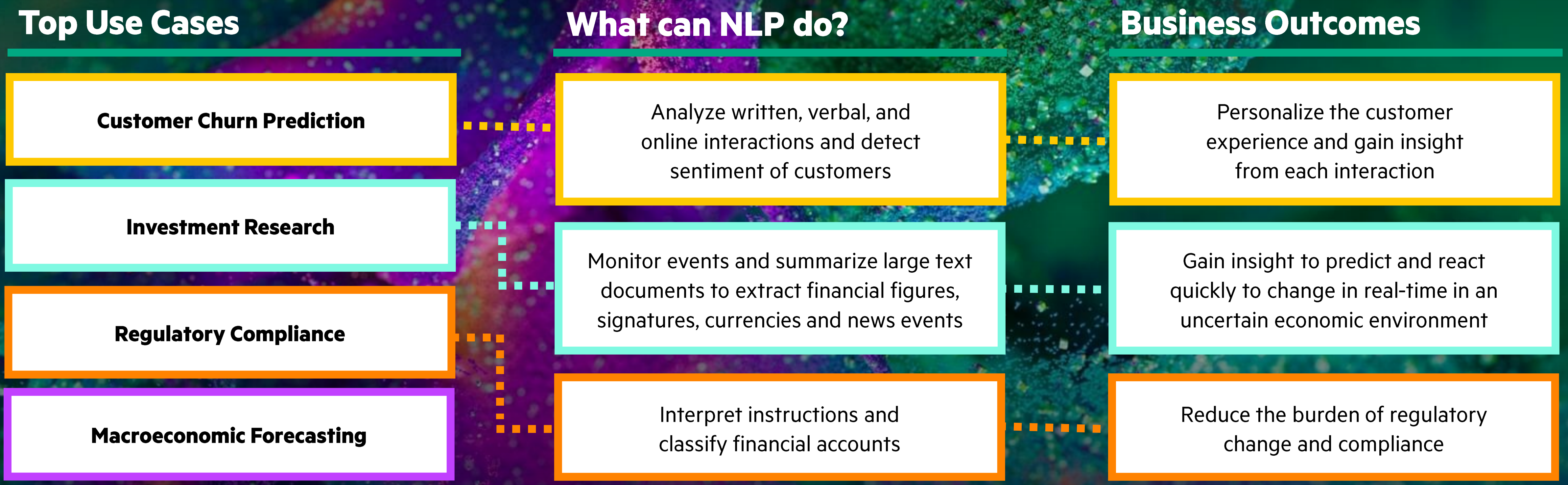
Hunter Almgren, Distinguished Technologist

May 31st, 2023

NLP in Financial Services

Customer Challenges

- Ongoing regulatory change
- Ability to summarize vast amounts of qualitative data
- Manual processes requiring specialized expertise



Top Use Cases

- Customer Churn Prediction
- Investment Research
- Regulatory Compliance
- Macroeconomic Forecasting

What can NLP do?

- Analyze written, verbal, and online interactions and detect sentiment of customers
- Monitor events and summarize large text documents to extract financial figures, signatures, currencies and news events
- Interpret instructions and classify financial accounts

Business Outcomes

- Personalize the customer experience and gain insight from each interaction
- Gain insight to predict and react quickly to change in real-time in an uncertain economic environment
- Reduce the burden of regulatory change and compliance

Do you have what NLP takes?

Data

- NLP needs lots of data
- The best data are often specific, proprietary, or sensitive

Compute

- NLP needs specialized infrastructure
- …and large numbers of GPUs

Time *and* Expertise

- NLP infrastructure is complex and time-consuming to manage
- Models can take months to train

Investment

- Training a model can cost $ millions
- Running and evaluating models adds to the bill

Challenges with training Large-scale language models

Why is it important for HPE to address these challenges?

Massive GPU clusters with optimized networking and storage	Resource & experiment management, distributed training, centralized UI	Adaptable infrastructure for future models and hardware
Hardware	**Software**	**Flexibility**
How do we design & build hardware platforms for the use case at hand and optimized on day one?	How do we provide all required and productivity enhancing capabilities in an end-to-end software platform?	How do we best prepare for a future where there may be novel models and architectures?

Expanding capabilities

Language models → Multimodal models

→ Significant applications for many industries

Increasing model size

Parameters: 100s million→ 100s billion → Trillions

→ Complex to train and optimize successfully

Emphasis on Alignment

Driving toward human-centric decision making

→ Ensure trustworthiness within model decisions

Solve your NLP challenges

Compute?

Access everything you need in one solution

- Develop and train models from day one
- Choose optimal infrastructure for any workload at scale
- Work across on-premises, private cloud, and public cloud
- Access emerging tech
- Get more from your GPUs
- Easily share on-premise or cloud GPUs with your team

Time and expertise?

Build models, not infrastructure

- Find and train more accurate models faster
- Save time with seamless distributed training and easy-to-use interface
- No need to rewrite code or manage infrastructure
- Easily interpret and reproduce your experiments
- Access solution-level support and deep expertise

Investment?

Spend less time and money

- Optimize all the GPUs you need, when you need them
- Fine-tune models faster
- Reduce headcount by focusing teams on delivering value
- Avoid hardware vendor lock-in and reduce cloud fees
- Pay your own way with a range of license, SaaS, and PaaS Options

COMPLEXITY WITH LOTS OF CHOICES: THREE ASPECTS OF AN AI PLATFORM

Data

EDA

pandas HIVE dask trino RAPIDS APACHE Spark

Pipelines

Versioning, Labelling

DVC LFS Pachyderm scale DELTA LAKE Labelbox annotell

Data Sources

LOG CSV aws Azure SQL snowflake CLOUDERA databricks lustre MAPR S3

Development

Collaboration

GitHub JFrog

Experiments

Determined AI

Scheduling

kubernetes slurm workload manager

Environment

jupyter PC

Evaluation

TensorBoard gradio

Compute

intel NVIDIA AMD habana An Intel Company cerebras Qualcomm

Deployment

Monitoring

Prometheus Grafana

Optimization

tvm deci

Testing

great_expectations Buildkite

Observability

ALIBI EXPLAIN truera modzy fiddler

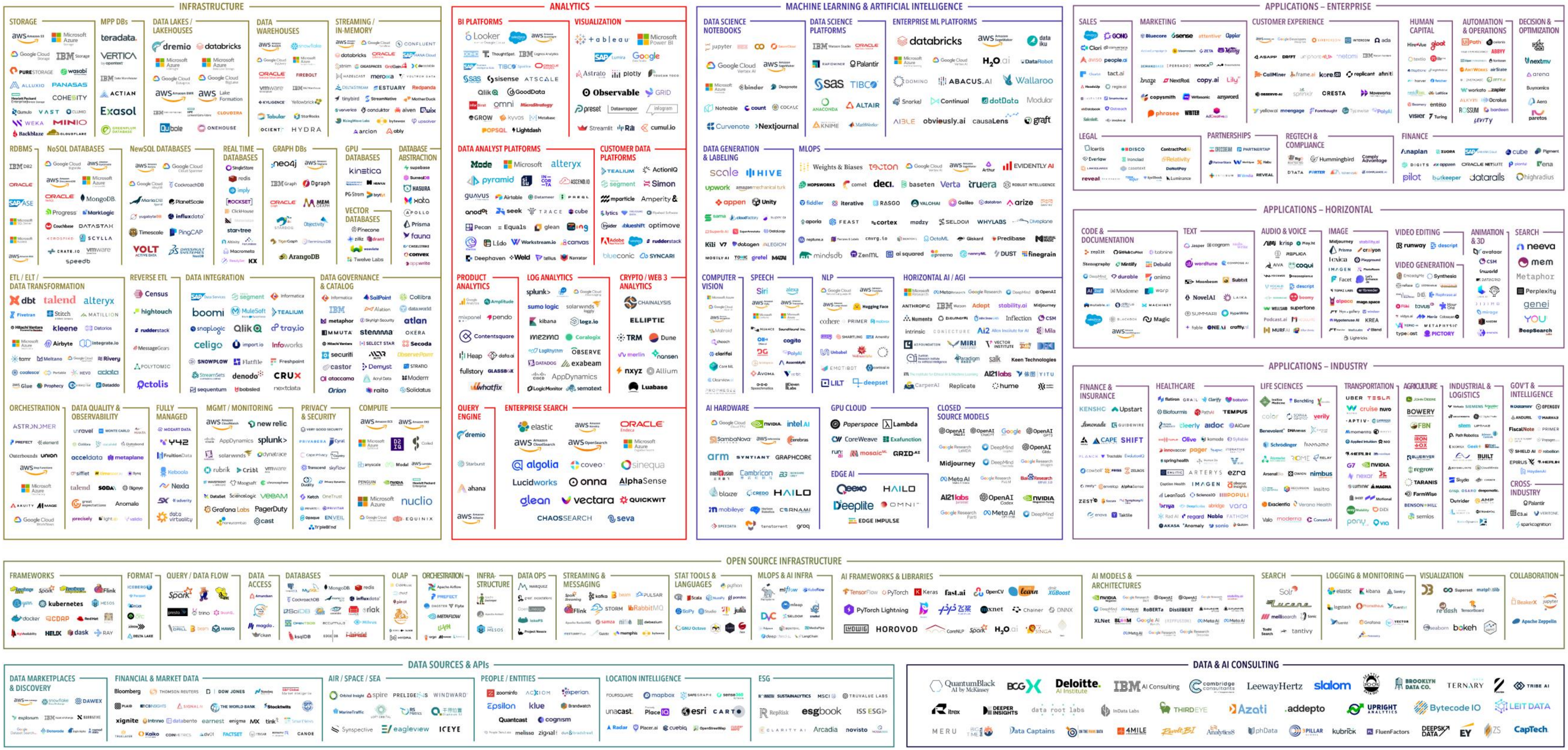
Bias, Robustness

ALIBI DETECT TROJ.AI

Serving, Rollout

TensorFlow Extended ONNX TORCHSERVE BENTOML SELDON KServe TensorRT

THE 2023 MAD (MACHINE LEARNING, ARTIFICIAL INTELLIGENCE & DATA) LANDSCAPE



INFRASTRUCTURE

STORAGE · MPP DBs · DATA LAKES / LAKEHOUSES · DATA WAREHOUSES · STREAMING / IN-MEMORY · RDBMS · NoSQL DATABASES · NewSQL DATABASES · REAL TIME DATABASES · GRAPH DBs · GPU DATABASES · VECTOR DATABASES · DATABASE ABSTRACTION · ETL / ELT / DATA TRANSFORMATION · REVERSE ETL · DATA INTEGRATION · DATA GOVERNANCE & CATALOG · ORCHESTRATION · DATA QUALITY & OBSERVABILITY · FULLY MANAGED · MGMT / MONITORING · PRIVACY & SECURITY · COMPUTE

ANALYTICS

BI PLATFORMS · VISUALIZATION · DATA ANALYST PLATFORMS · CUSTOMER DATA PLATFORMS · PRODUCT ANALYTICS · LOG ANALYTICS · CRYPTO / WEB 3 ANALYTICS · QUERY ENGINE · ENTERPRISE SEARCH

MACHINE LEARNING & ARTIFICIAL INTELLIGENCE

DATA SCIENCE NOTEBOOKS · DATA SCIENCE PLATFORMS · ENTERPRISE ML PLATFORMS · DATA GENERATION & LABELING · MLOPS · COMPUTER VISION · SPEECH · NLP · HORIZONTAL AI / AGI · AI HARDWARE · GPU CLOUD · EDGE AI · CLOSED SOURCE MODELS

APPLICATIONS – ENTERPRISE

SALES · MARKETING · CUSTOMER EXPERIENCE · HUMAN CAPITAL · AUTOMATION & OPERATIONS · DECISION & OPTIMIZATION · LEGAL · PARTNERSHIPS · REGTECH & COMPLIANCE · FINANCE

APPLICATIONS – HORIZONTAL

CODE & DOCUMENTATION · TEXT · AUDIO & VOICE · IMAGE · VIDEO EDITING · VIDEO GENERATION · ANIMATION & 3D · SEARCH

APPLICATIONS – INDUSTRY

FINANCE & INSURANCE · HEALTHCARE · LIFE SCIENCES · TRANSPORTATION · AGRICULTURE · INDUSTRIAL & LOGISTICS · GOV'T & INTELLIGENCE · CROSS-INDUSTRY

OPEN SOURCE INFRASTRUCTURE

FRAMEWORKS · FORMAT · QUERY / DATA FLOW · DATA ACCESS · DATABASES · OLAP · ORCHESTRATION · INFRASTRUCTURE · DATA OPS · STREAMING & MESSAGING · STAT TOOLS & LANGUAGES · MLOPS & AI INFRA · AI FRAMEWORKS & LIBRARIES · AI MODELS & ARCHITECTURES · SEARCH · LOGGING & MONITORING · VISUALIZATION · COLLABORATION

DATA SOURCES & APIs

DATA MARKETPLACES & DISCOVERY · FINANCIAL & MARKET DATA · AIR / SPACE / SEA · PEOPLE / ENTITIES · LOCATION INTELLIGENCE · ESG

DATA & AI CONSULTING

Version 1.0 - Feb 2023 © Matt Turck (@mattturck), Kevin Zhang (@ykevinzhang) & FirstMark (@firstmarkcap) Blog post: mattturck.com/MAD2023 Interactive version: MAD.firstmarkcap.com Comments? Email MAD2023@firstmarkcap.com

FIRSTMARK
EARLY STAGE VENTURE CAPITAL

MACHINE LEARNING DATA MANAGEMENT FOR THE ENTERPRISE

Flexibility

Diverse set of users

Diverse environments & infra

Diverse types use cases

Scalability

Scale manual user tasks through automation

Scale to massive data volumes

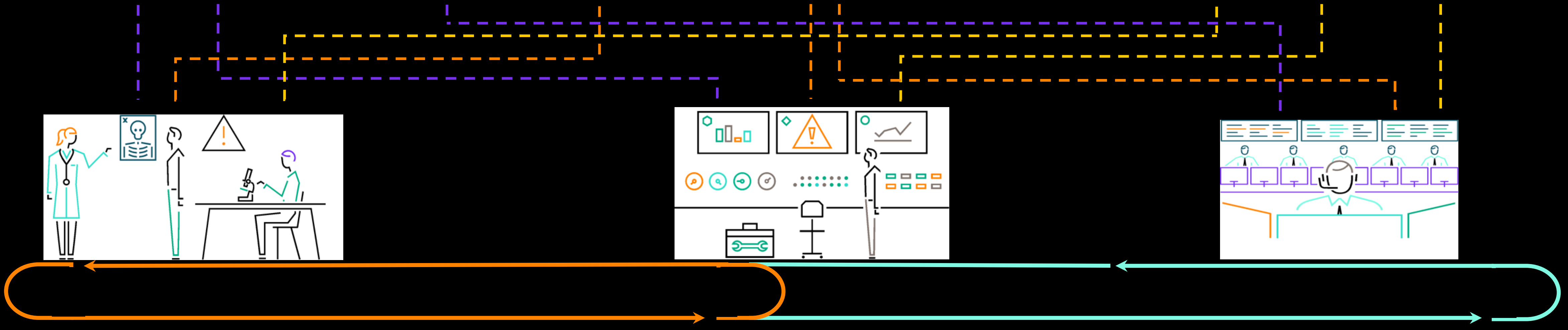
Scale across teams

Reproducibility

Developers can recreate and debug workflows

Teams can reuse and build upon each others' work

Organizations must meet compliance and regulatory requirements

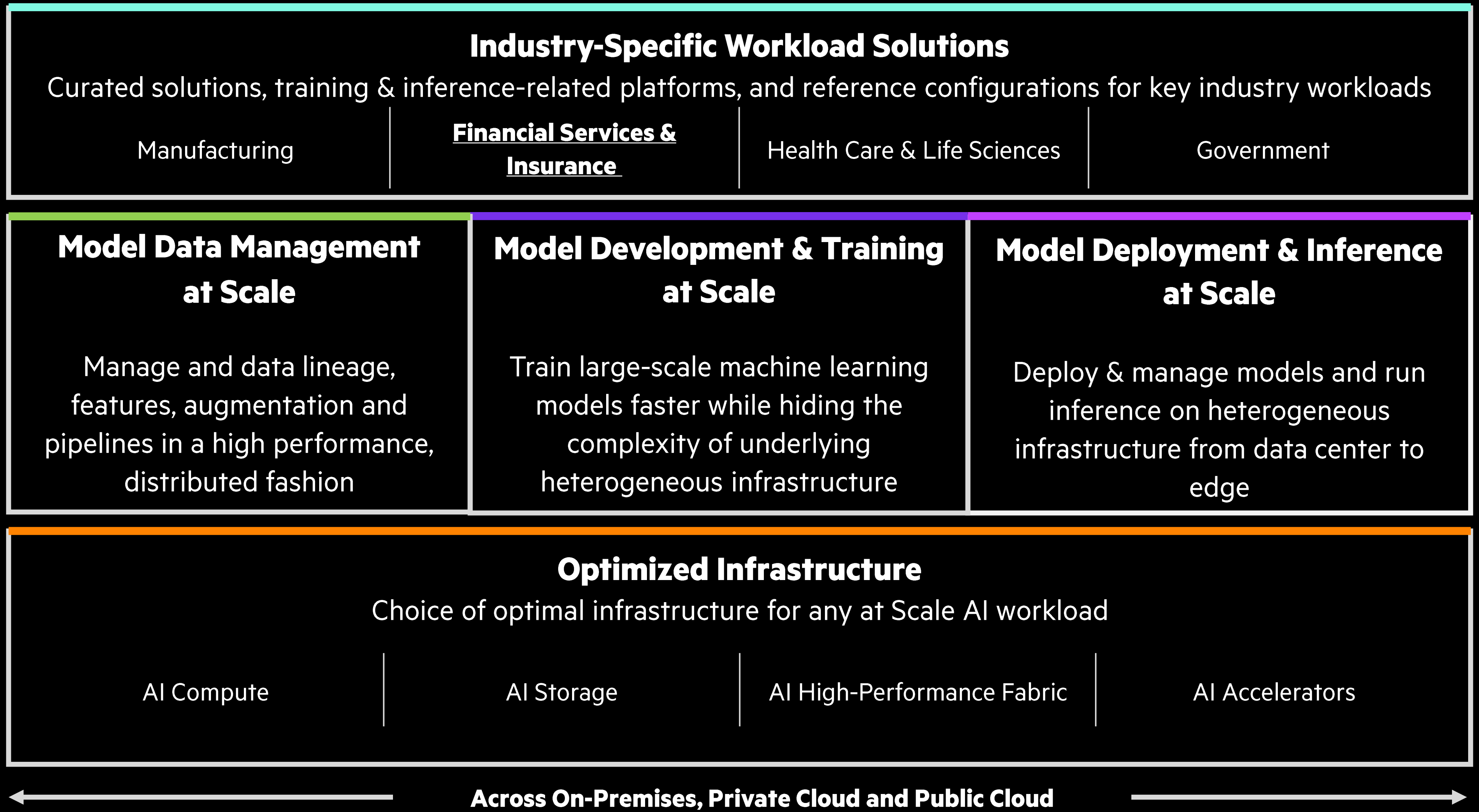


Data Processing & Pipelining

Model Development & Optimization

Model Deployment & Monitoring

AI at Scale Platform



Industry-Specific Workload Solutions

Curated solutions, training & inference-related platforms, and reference configurations for key industry workloads

Manufacturing	**Financial Services & Insurance**	Health Care & Life Sciences	Government

Model Data Management at Scale

Manage and data lineage, features, augmentation and pipelines in a high performance, distributed fashion

Model Development & Training at Scale

Train large-scale machine learning models faster while hiding the complexity of underlying heterogeneous infrastructure

Model Deployment & Inference at Scale

Deploy & manage models and run inference on heterogeneous infrastructure from data center to edge

Optimized Infrastructure

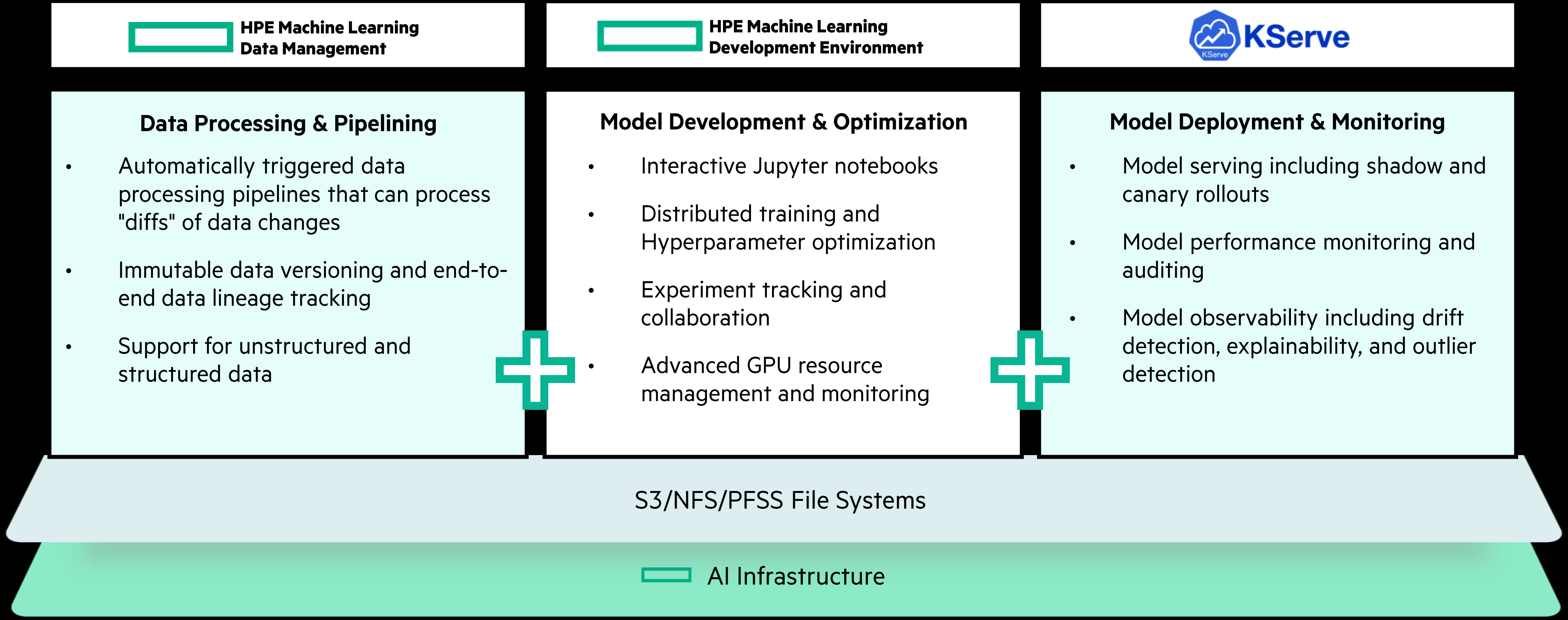
Choice of optimal infrastructure for any at Scale AI workload

AI Compute	AI Storage	AI High-Performance Fabric	AI Accelerators

Across On-Premises, Private Cloud and Public Cloud

THE HPE AI-AT-SCALE PLATFORM

Hosted on Kubernetes



HPE Machine Learning Data Management

HPE Machine Learning Development Environment

KServe

Data Processing & Pipelining

- Automatically triggered data processing pipelines that can process "diffs" of data changes
- Immutable data versioning and end-to-end data lineage tracking
- Support for unstructured and structured data

Model Development & Optimization

- Interactive Jupyter notebooks
- Distributed training and Hyperparameter optimization
- Experiment tracking and collaboration
- Advanced GPU resource management and monitoring

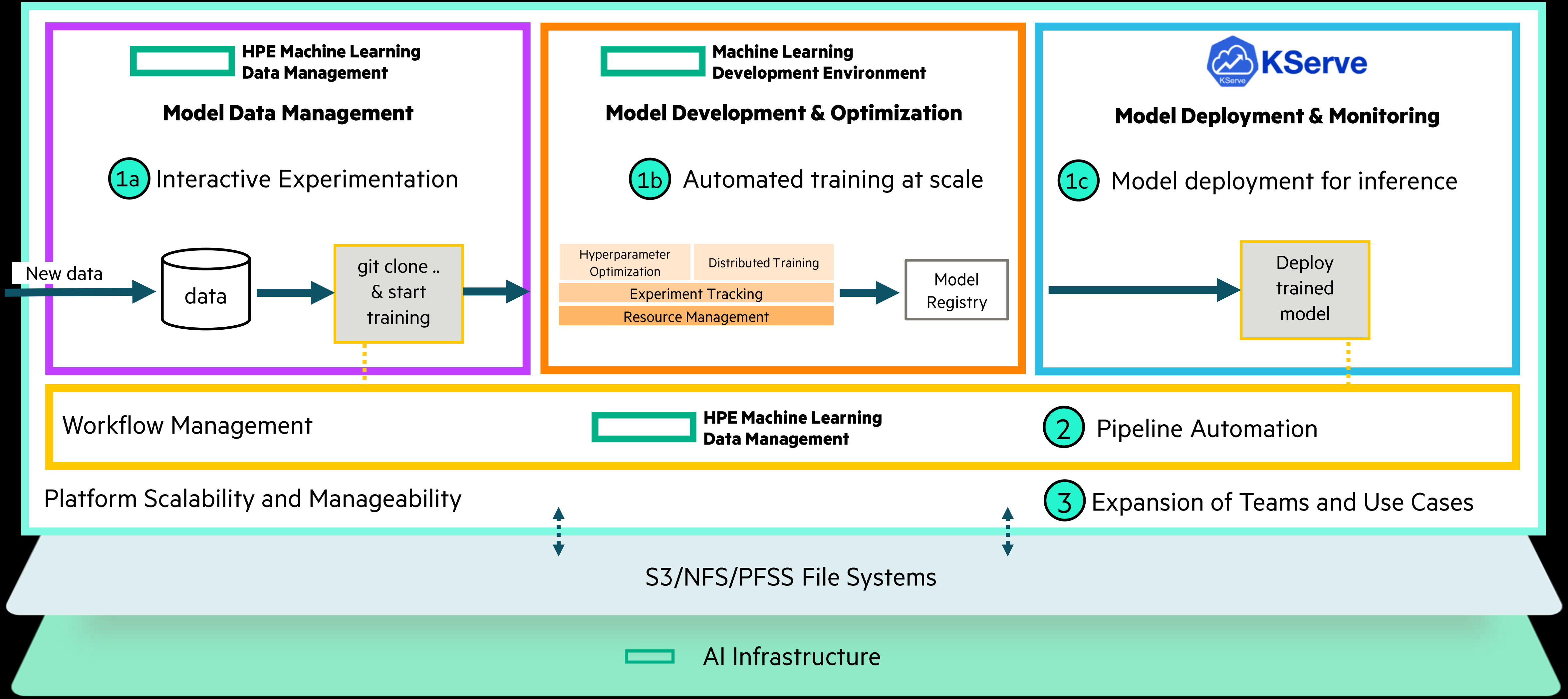
Model Deployment & Monitoring

- Model serving including shadow and canary rollouts
- Model performance monitoring and auditing
- Model observability including drift detection, explainability, and outlier detection

S3/NFS/PFSS File Systems

AI Infrastructure

EXAMPLE WORKFLOW WITH ML PLATFORM



HPE Machine Learning Data Management

Model Data Management

1a Interactive Experimentation

New data

data

git clone .. & start training

Machine Learning Development Environment

Model Development & Optimization

1b Automated training at scale

Hyperparameter Optimization

Distributed Training

Experiment Tracking

Resource Management

Model Registry

KServe

Model Deployment & Monitoring

1c Model deployment for inference

Deploy trained model

Workflow Management

HPE Machine Learning Data Management

2 Pipeline Automation

Platform Scalability and Manageability

3 Expansion of Teams and Use Cases

S3/NFS/PFSS File Systems

AI Infrastructure

Thank You

hunter.almgren@hpe.com
Distinguished Technologist