PRACTICAL MACHINE LEARNING PIPELINE USING STREAMING IOT SENSOR DATA

Mathieu Dumoulin - MapR, Mateusz Dymczyk - H2O.ai

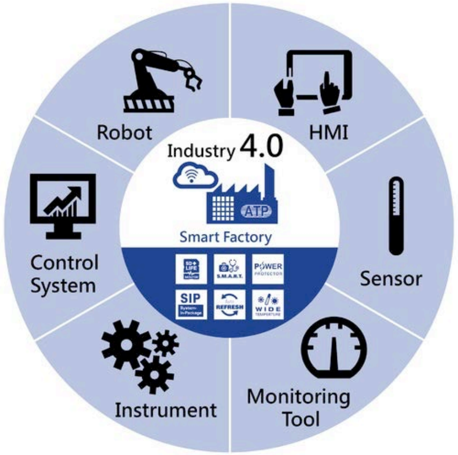
February 7, 2017 @ Tokyo Big Data Analytics 2017

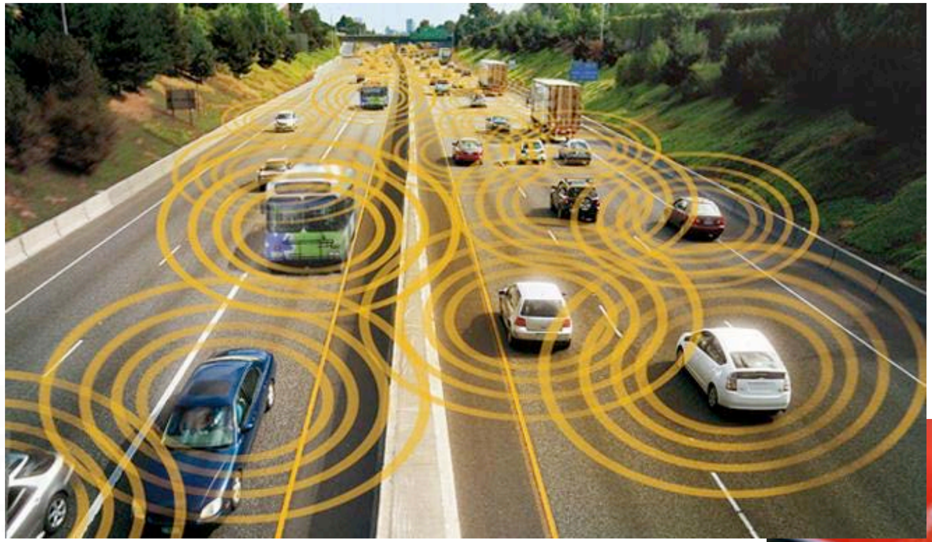
Mathieu Dumoulin and Mateusz Dymczyk

- Data Engineer @ MapR Technologies
- Previously data scientist and DS team manager, search, NLP and ML engineer

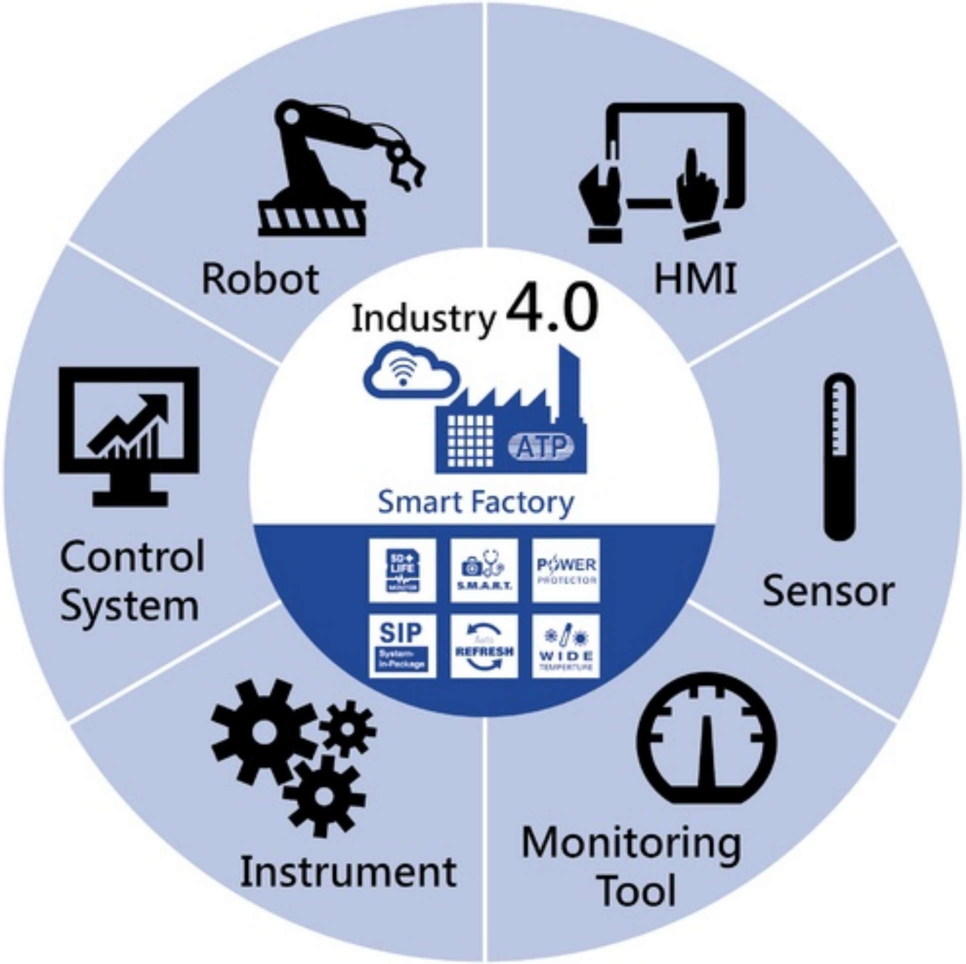
- Software Engineer @ H2O.ai
- Previously ML/NLP @ Fujitsu Laboratories and en-japan inc
- Sommelier in training

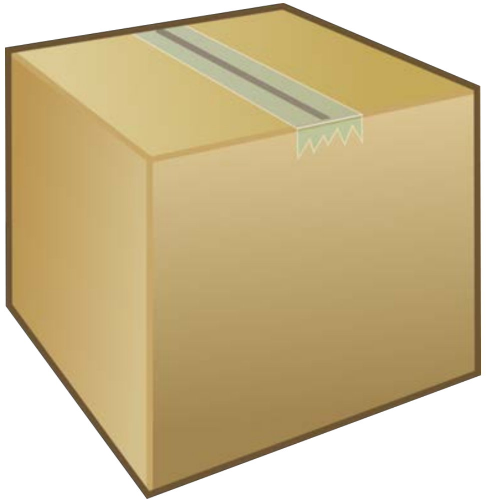
The Time for IoT is NOW





IoT and Industry 4.0: Predictive Maintenance





Problem Statement and Business Value

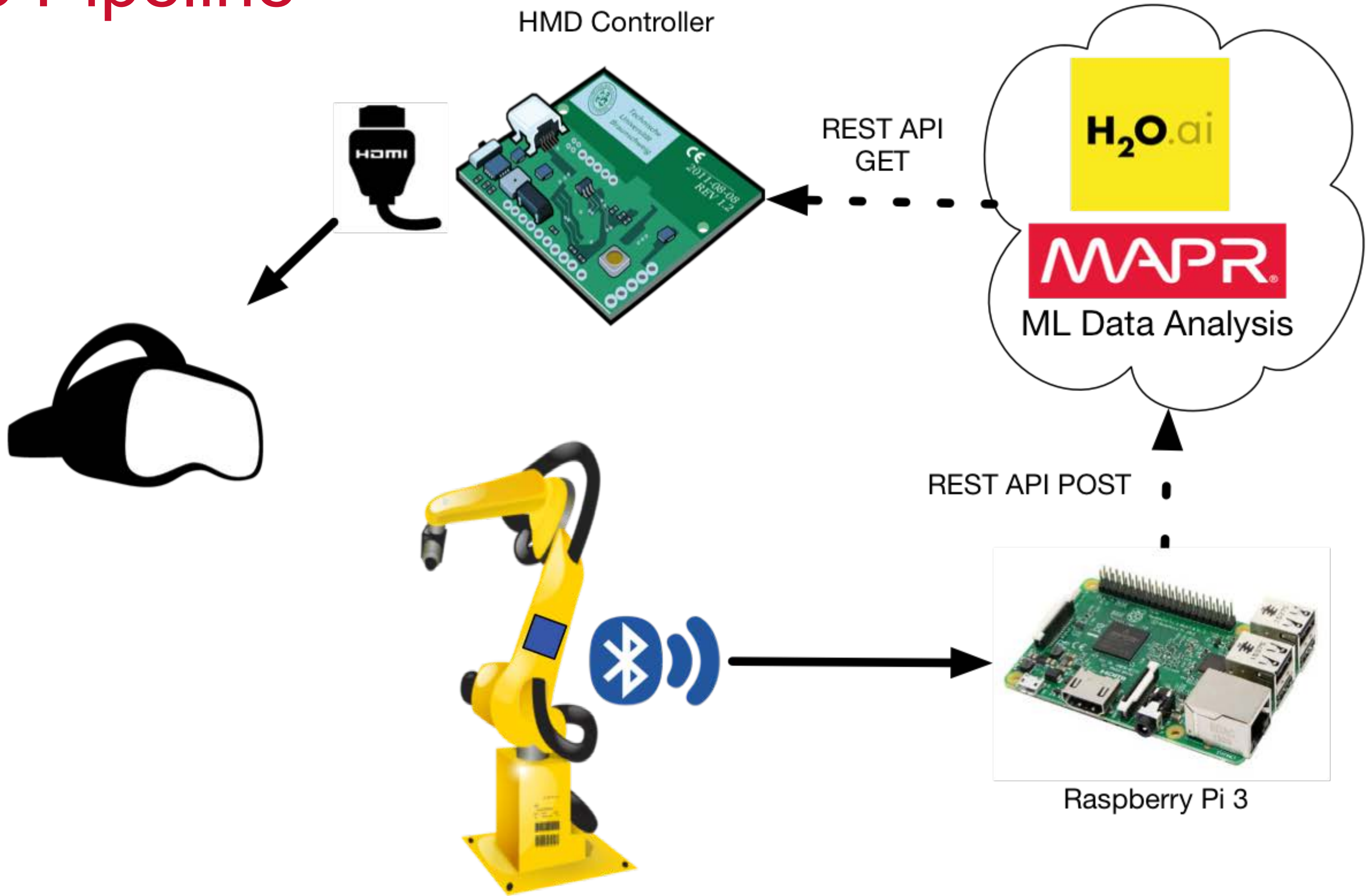
Situation: 生産能力を2倍以上にすることがゴール

We want a **real-time** view to allow factory staff to **see** which robots will *probably* fail, before they actually fail.

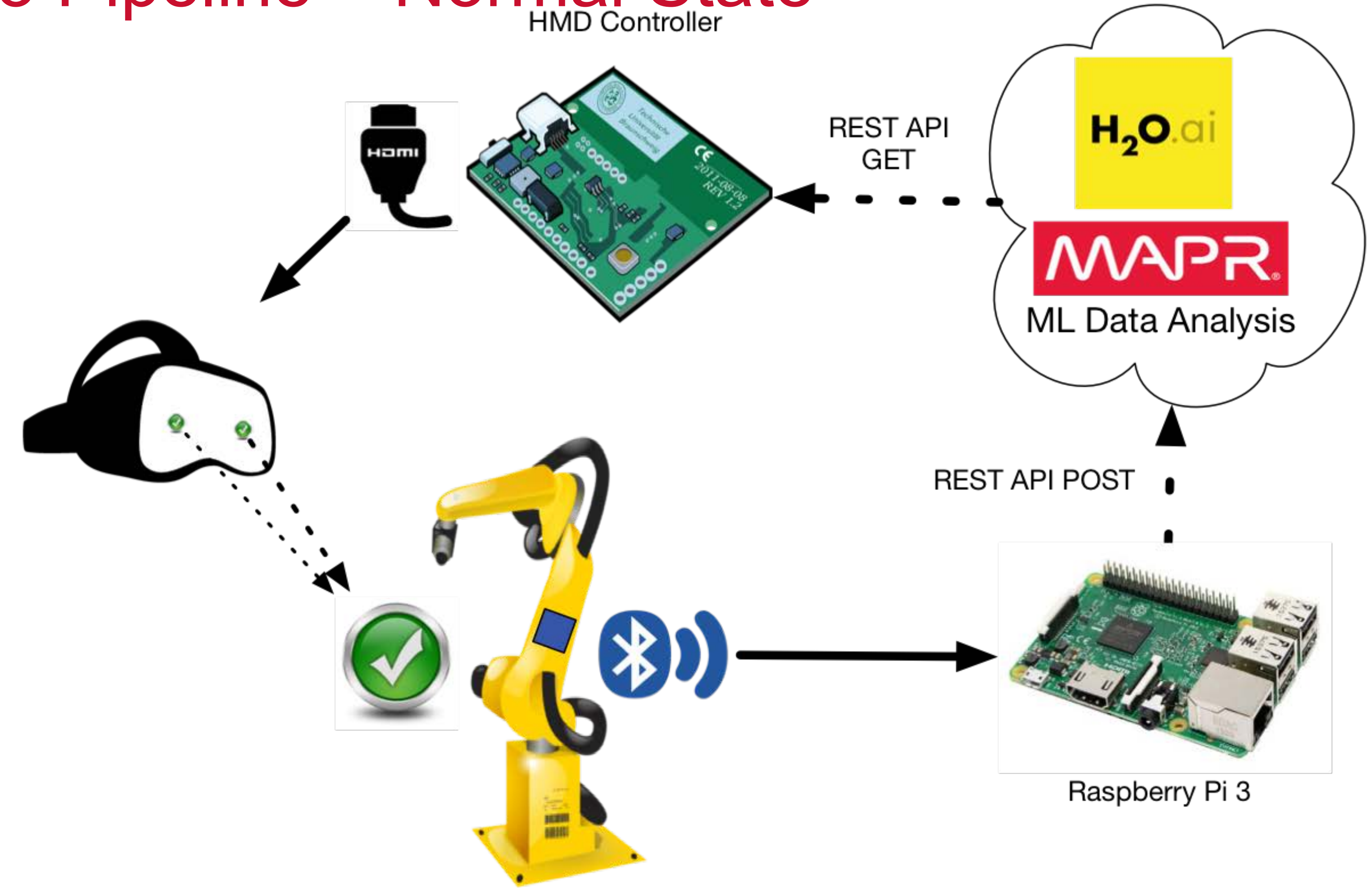
More Requirements:

- Deal with **variety** of robots (age, maker, function)
- **Scale** to [100-10,000] robots in real-time and **multiple factories**
- Ensure data **reliability**
- Factory staff has **low level of IT knowledge**

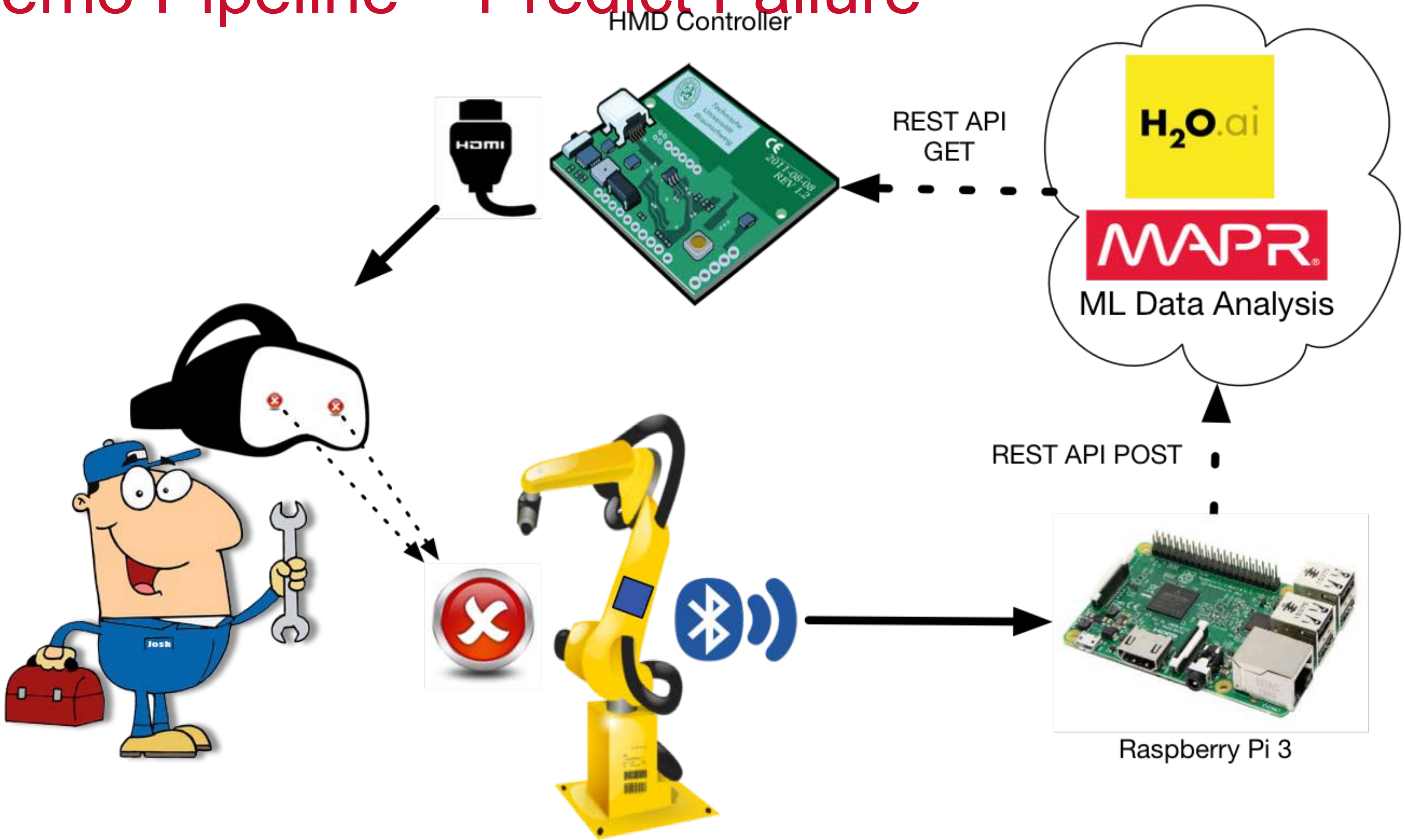
Demo Pipeline



Demo Pipeline – Normal State

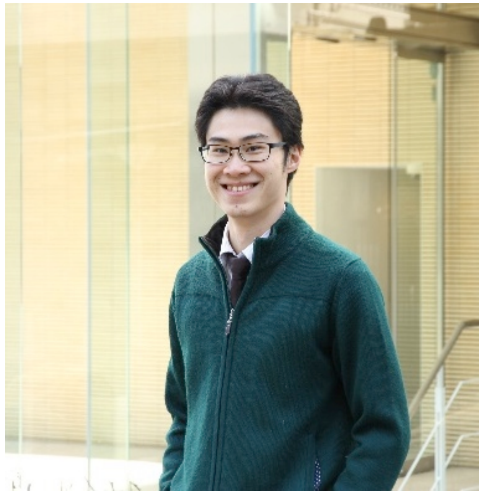


Demo Pipeline – Predict Failure



LP-RESEARCH Motion Sensor

- Tokyo-based startup
- Hardware R&D for Industry 4.0 applications
- Founded by Waseda University Ph.D. grads
- www.lp-research.com



LIFE PERFORMANCE RESEARCH

Our Demo – Real Time Robot Failure Prediction… with AR Visualization

How we Made our Demo

github.com/mdymczyk/iot-pipeline

1. Machine learning modeling
2. Data input
    1. Sensor to backend analysis
3. Backend data analysis
    1. MapR Converged Data Platform
    2. Streaming Architecture, MapR Streams (Apache Kafka)
4. Data output: visualizing predictions
    1. Augmented Reality Headset

Machine Learning Modeling

1. Set the Machine Learning goal
   1. Detect abnormal events > 90% accuracy
   2. Avoid false positives
   3. Decide output
2. How to reach the goal
   1. Supervised vs. unsupervised
   2. Choose algorithm
   3. Initial dataset exploration
   4. Data cleaning and feature extraction
   5. Deal with real-time and large scale
3. Deploy to production
   1. Use MapR CDP and custom software
   2. H2O’s export to POJO function

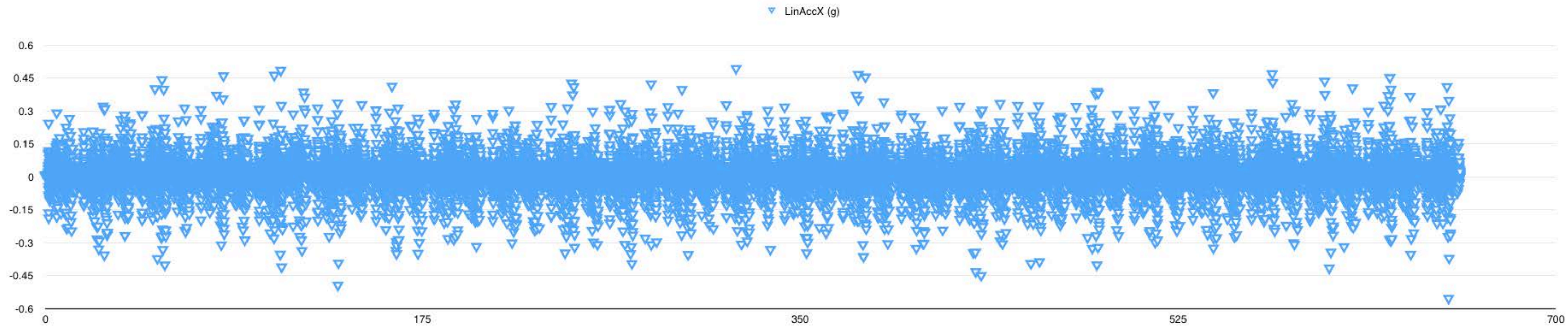
Normal State (OK!)

PREDICT FAILURE

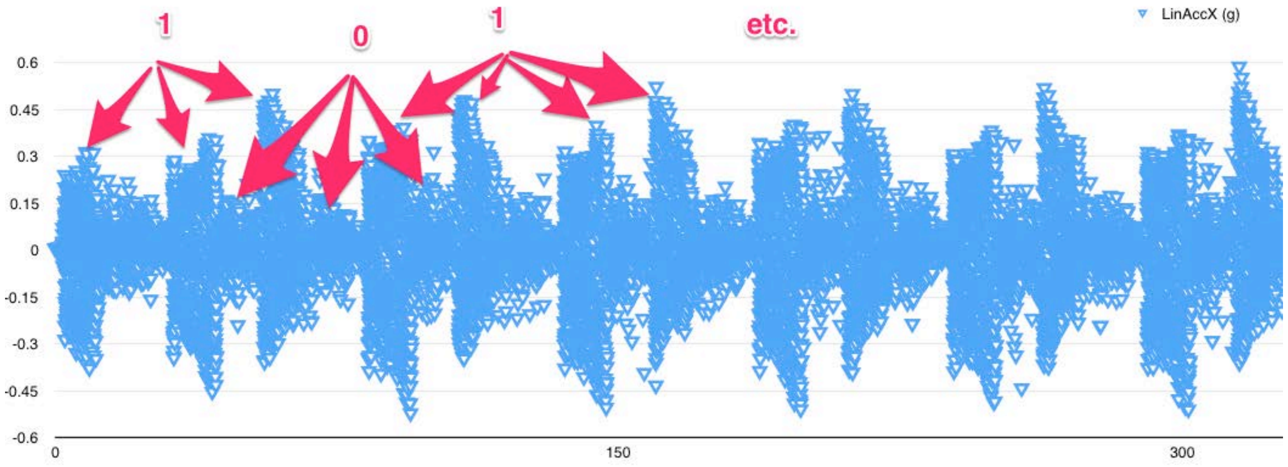
H2O.ai

ML – Looking at the data

Sensor Id	TimeStamp (s)	FrameNumber	AccX (g)	AccY (g)	AccZ (g)	GyroX (deg/s)	GyroY (deg/s)	GyroZ (deg/s)	MagX (uT)	MagY (uT)	MagZ (uT)	EulerX (deg)	EulerY (deg)	EulerZ (deg)	Quat W	QuatX	QuatY	QuatZ	LinAccX (g)	LinAccY (g)	LinAccZ (g)	Pressure (kPa)	Altitude (m)	Temp (degC)	HeaveMotion (m)
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	-0.0103	0.0137	0.0143	0	0	0	0
1	0.01	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0.0003	-0.0002	0.0188	0	0	0	0
1	0.02	2	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0.0023	0.0031	0.0227	0	0	0	0
1	0.03	3	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0.0015	0.0111	0.0182	0	0	0	0
1	0.04	4	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0.0048	0.0228	0.0042	0	0	0	0
1	0.05	5	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0.006	0.0311	-0.0008	0	0	0	0
1	0.06	6	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0.013	0.0205	-0.0205	0	0	0	0
1	0.07	7	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0.0191	0.0067	-0.0486	0	0	0	0
1	0.08	8	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0.0206	0.0022	-0.0653	0	0	0	0
1	0.09	9	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0.0156	0.002	-0.0761	0	0	0	0
1	0.1	10	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0.0126	-0.0163	-0.083	0	0	0	0
1	0.11	11	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0.0144	-0.0327	-0.0807	0	0	0	0
1	0.12	12	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0.0135	-0.0411	-0.0815	0	0	0	0
1	0.13	13	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0.0106	-0.0437	-0.0734	0	0	0	0

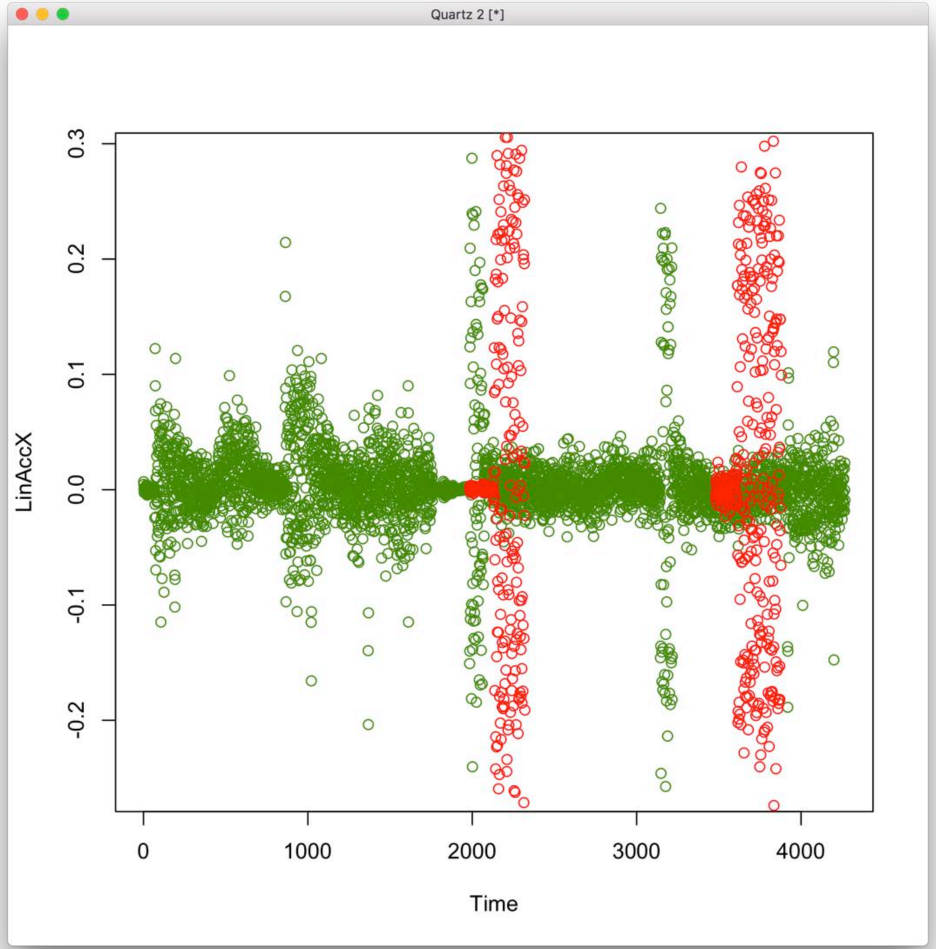


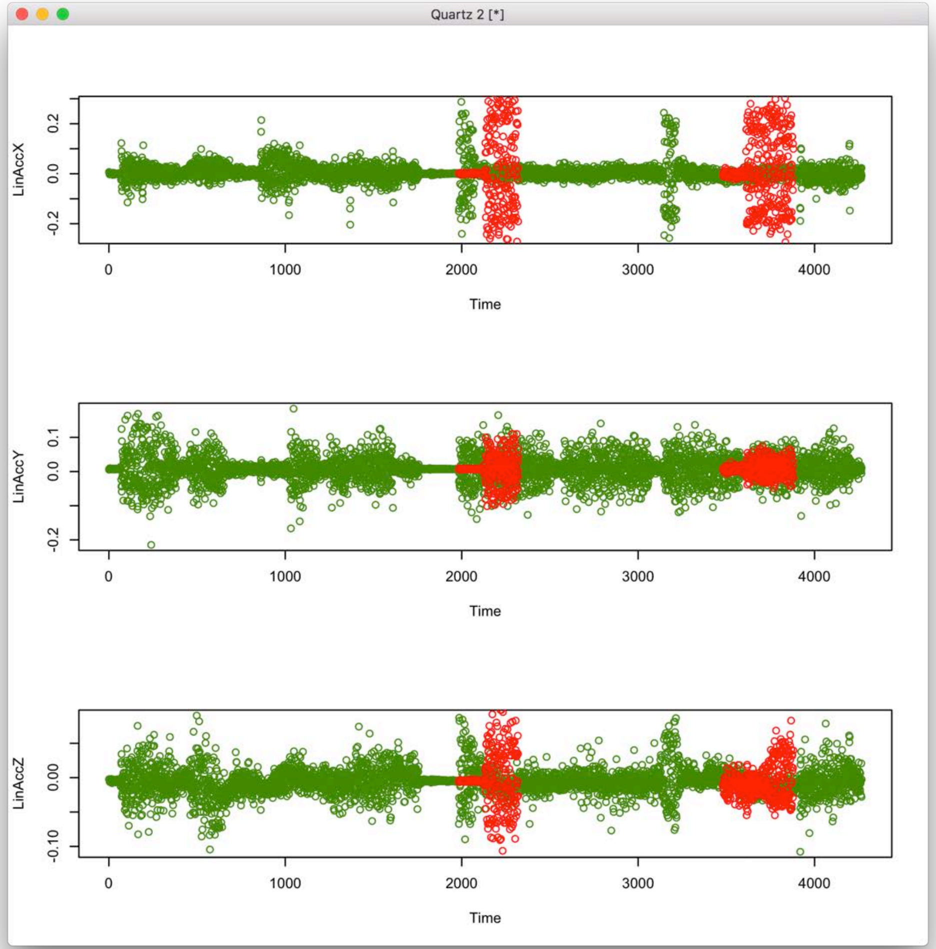
ML – Anomaly Detection



- Unsupervised: 教師なし学習
- Anomaly detection: 異常認識
- H2O uses autoencoder algorithm (deep learning)
- H2O's R API for modeling
  - Very productive API
  - Good graphs
- Parameter tuning of models
- See [H2O's training-book on GitHub]

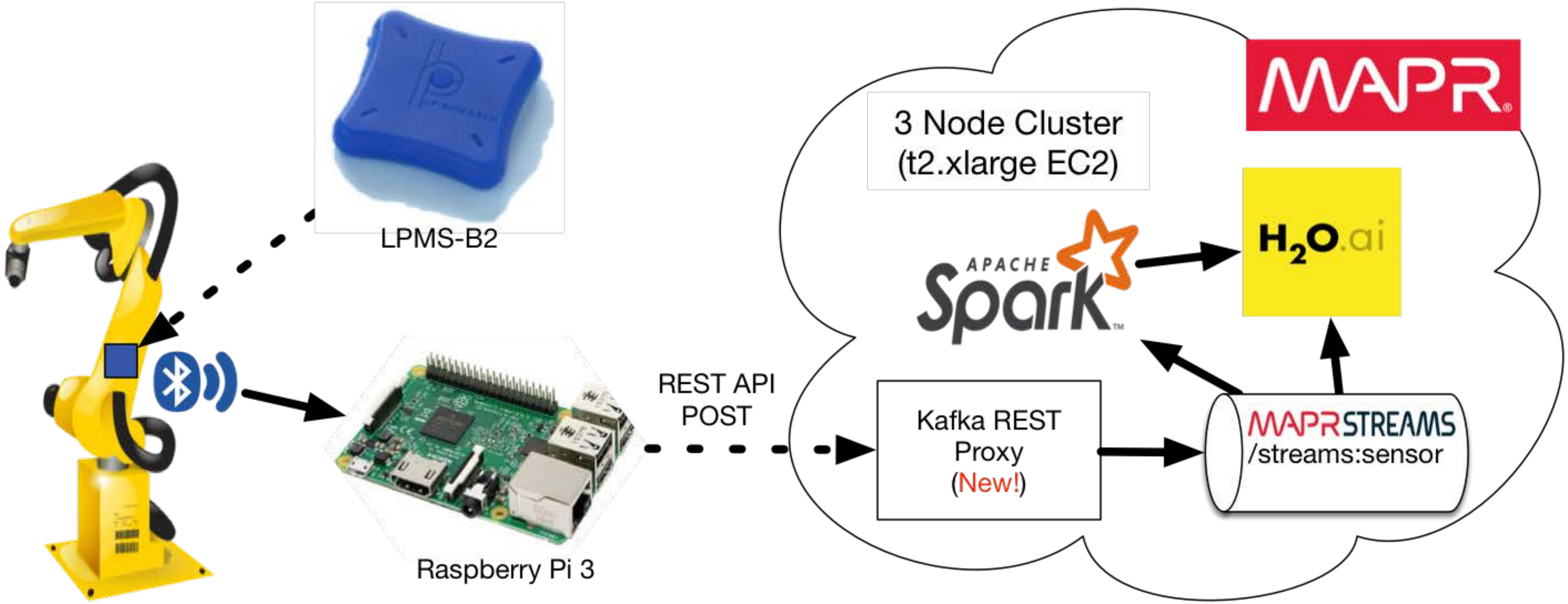
ML – Results

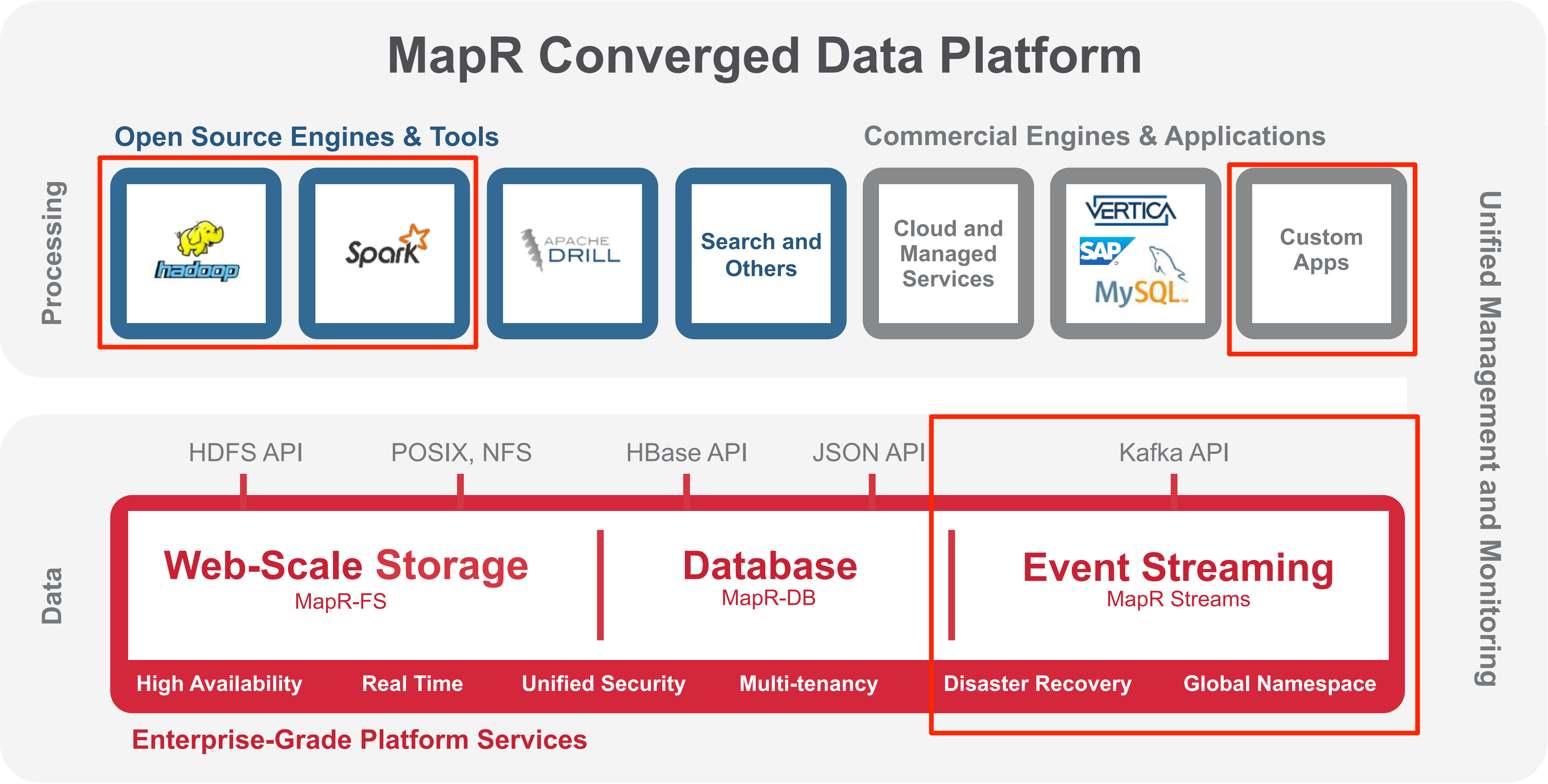




Note: Time window: 200ms, Threshold: 1SD (標準偏差)

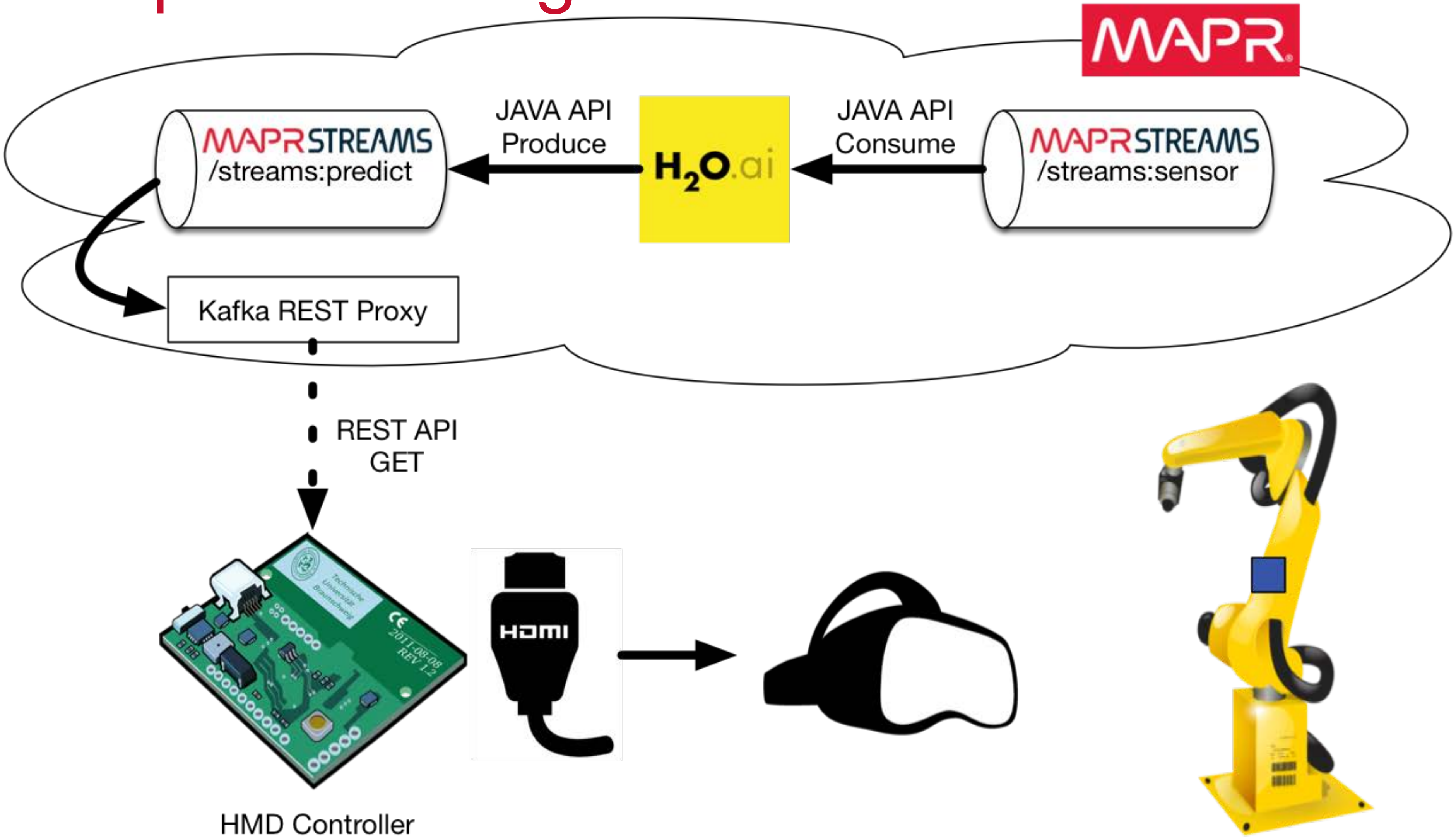
ML – Deploy to Production – Real-time Data



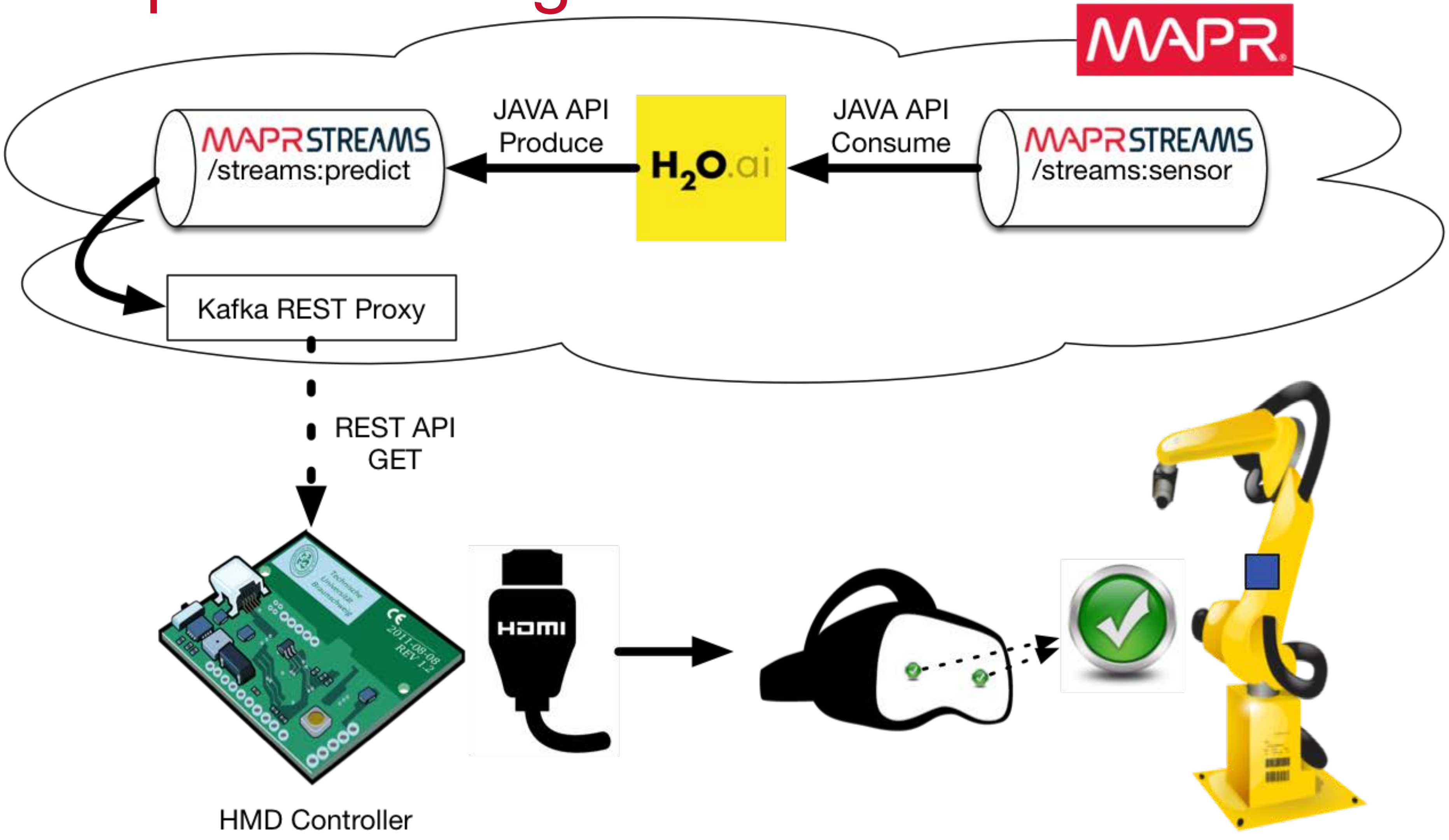


MapR Converged Data Platform
Open Source Engines & Tools
Commercial Engines & Applications
Processing
hadoop
Spark
APACHE DRILL
Search and Others
Cloud and Managed Services
VERTICA
SAP
MySQL
Custom Apps
Unified Management and Monitoring
Data
HDFS API
POSIX, NFS
HBase API
JSON API
Kafka API
Web-Scale Storage
MapR-FS
Database
MapR-DB
Event Streaming
MapR Streams
High Availability
Real Time
Unified Security
Multi-tenancy
Disaster Recovery
Global Namespace
Enterprise-Grade Platform Services

Data Output – Making Predictions



Data Output – Making Predictions



Conclusion

- Getting a good enough model on some data was less than 10% of the total work.
- Team members need to have ALL expertise for this kind of project. Hardware, software, big data, ML.

- MapR, H2O and LP-RESEARCH's sensor were all essential parts of the project success.
  - The MapR platform worked perfectly, H2O model is high quality and fast.
- The hardware expertise of LP-RESEARCH was critical

Q&A

Engage with us!

PROJECT GITHUB: github.com/mdymczyk/iot-pipeline

Mathieu Dumoulin, mdumoulin@mapr.com @Lordxar

- Blog: https://www.mapr.com/blog/author/mathieu-dumoulin

Mateusz Dymczyk, H2O.ai, mateusz@h2o.ai @mdymczyk

Klaus Petersen, klaus@lp-research.com

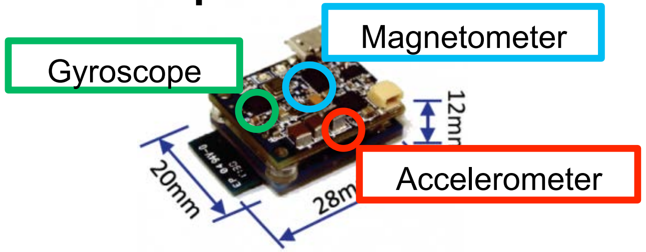
- LP-RESEARCH: www.lp-research.com

Thank you to LP-RESEARCH!

Hardware design and production
Expertise in Motion sensors

LIFE PERFORMANCE RESEARCH

Sensor fusion algorithm development



Multi-platform application development

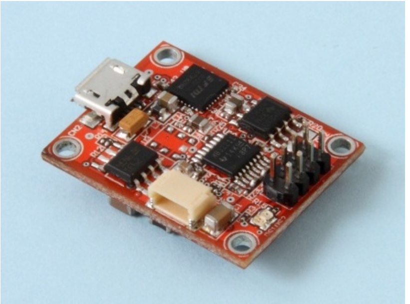
android
iOS

LPMS-B2

LPMS-CU2

LPMS-CANAL2

LPMS-USBAL2



OEM also available!

See all our products: https://www.lp-research.com/products/