Infuse AI into Your Enterprise

IBM全球杰出工程师

IBM AI系统研究技术总监

林咏华 (IBM研究院)

AI时代的移动技术革新

Era of AI: Innovations in Mobile Technologies

Roads towards Artificial Intelligence

IBM Watson

Watson won humans in Jeopardy

Solution Supporting Technologies

Transportation | Cognitive Medical | Finance & Insurance | Smarter City

Media & Entertainment | Cognitive Retailer | Automobile | Manufacture

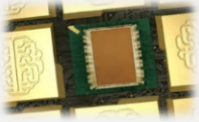
AI Cloud Computing

AI Vision, Acoustic, Language, Conversation

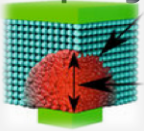
Deep Learning & Machine Learning

Quantum Computing

Neural Chips



In-memory Computing



Deep Learning Systems

When enterprises going into AI area ...

VALUE

TALENTS DATA ECONOMY

APICloud IT大咖说 知识共享平台

An Example – PowerAI Vision

An AI product which is powered by "AI for AI" innovation

AI时代的移动技术革新

Era of AI: Innovations in Mobile Technologies

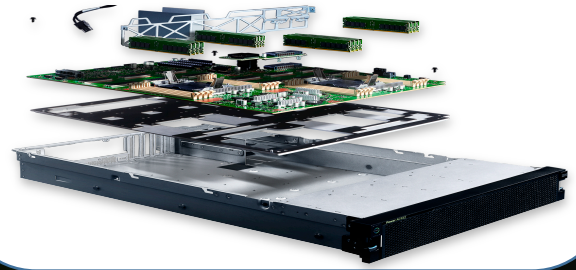
Very few enterprises have experienced teams in DL/AI

Deep Learning Experts

TensorFlow

Caffe

torch



Fixed API capability can not meet requirements in industries

Application Developers

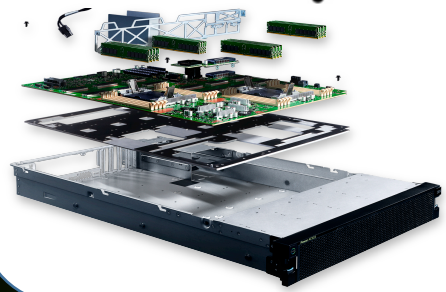
API

Cloud Service

TensorFlow

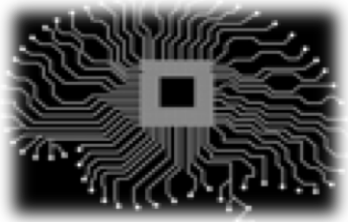
Caffe

torch



An AI Brain in Enterprise which will be used by application developer

- Learn high accurate models from enterprise data
- High productivity and efficiency

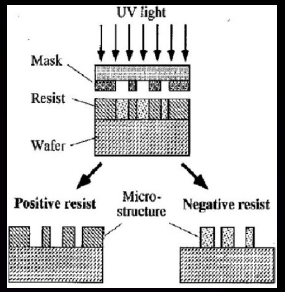


API

Cloud Service

Example 1: AI for Product Quality Inspection (Manufacture)

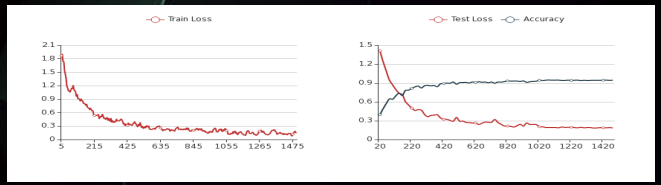
Inspect images of photoresist openings after having been exposed and developed (光刻是通过一系列生产步骤将晶圆表面薄膜的特定部分去除的工艺。被广泛用于集成电路的生产流程。**显影检查需要人工检验不合格的晶圆**，以便返工重新曝光、显影。)
显影检查：图形尺寸的偏差、光刻胶的污染、空洞、划伤，以及污点等。

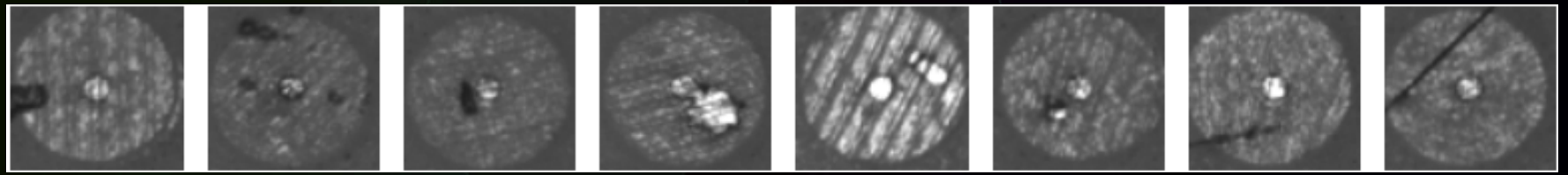


With **IBM AI Platform**, the manufacture could **quickly** build the auto defect inspection capability:

- Data import： 1 5 min.
- Data labeling： 5 min.
- AI Vision training： 1 0 min.

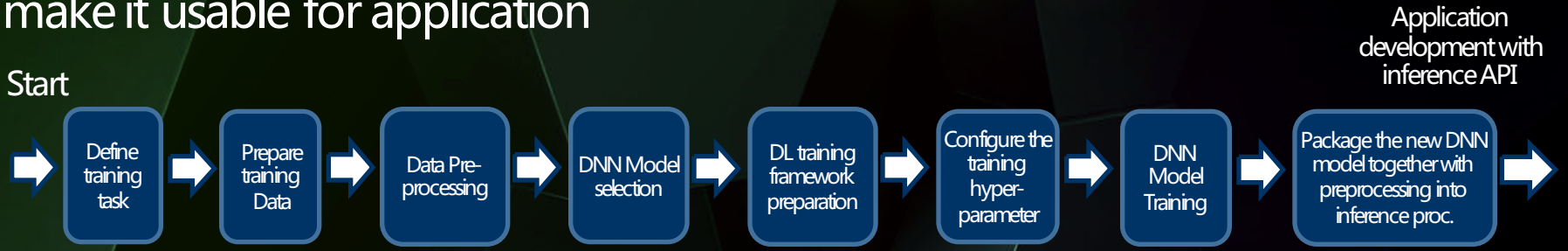
Accuracy: 94.5%





Steps for AI Deep Learning Development

Usually, developers need following steps to develop a DNN model and make it usable for application



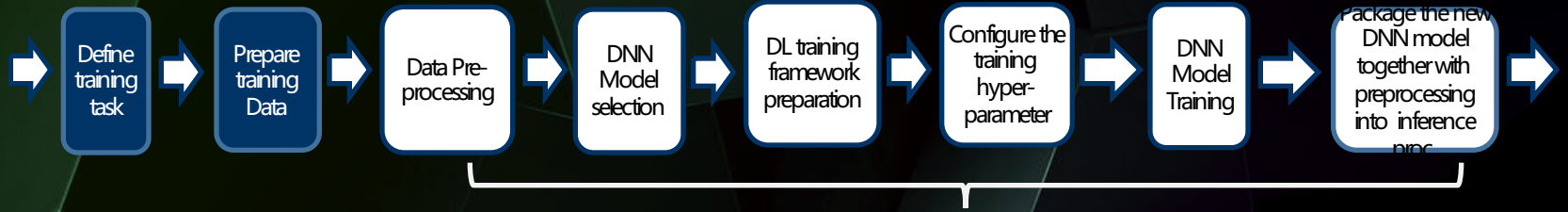
Most of enterprises are facing the challenges …

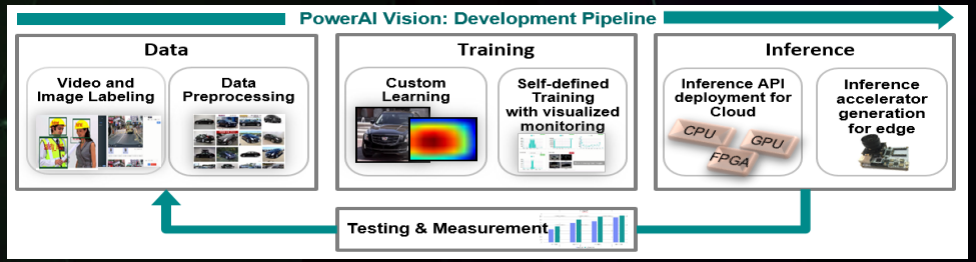
- No experience on DNN design and develop
- No experience on computer vision
- No experience on how to build a platform to support enterprise scale deep learning, including data preparation, training, and inference

We can help deep learning for Vision easier – PowerAI Vision

Deep knowledges of ML/DL and computer vision have been embedded into **PowerAI Vision**.

Start





APICloud IT大咖说 知识共享平台

How to ensure good accuracy without onsite deep learning experts?

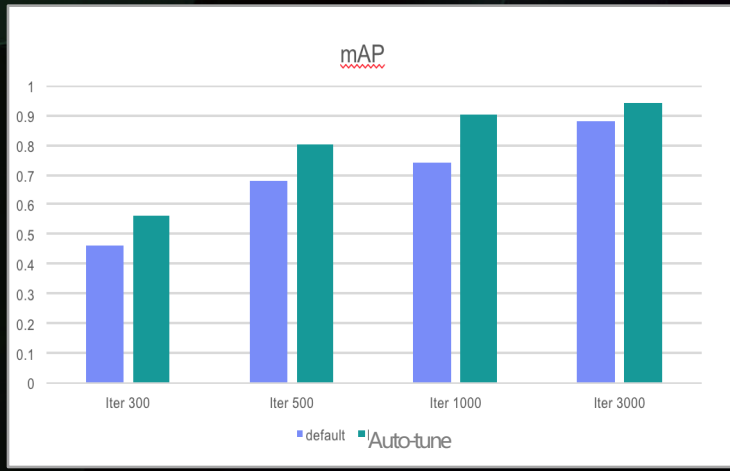
DL for DL: Learning to optimize parameters for visual analysis

- Through machine learning, PowerAI Vision will automatically tune parameters to achieve good accuracy for different training cases defined by users.
- In the following test case, our auto-tuning DL network could outperform the fix manual configuration (default) by >6%. And it could achieve the same accuracy (e.g. 90%) with much less training time (e.g <1/3).

18 parameters have been tuned, including
- Caffe training parameters
- Neural network parameters
- Object detection parameters.

Test data set: object detection for helmet and safety vest

Fig. 1 Performance comparison for object detection



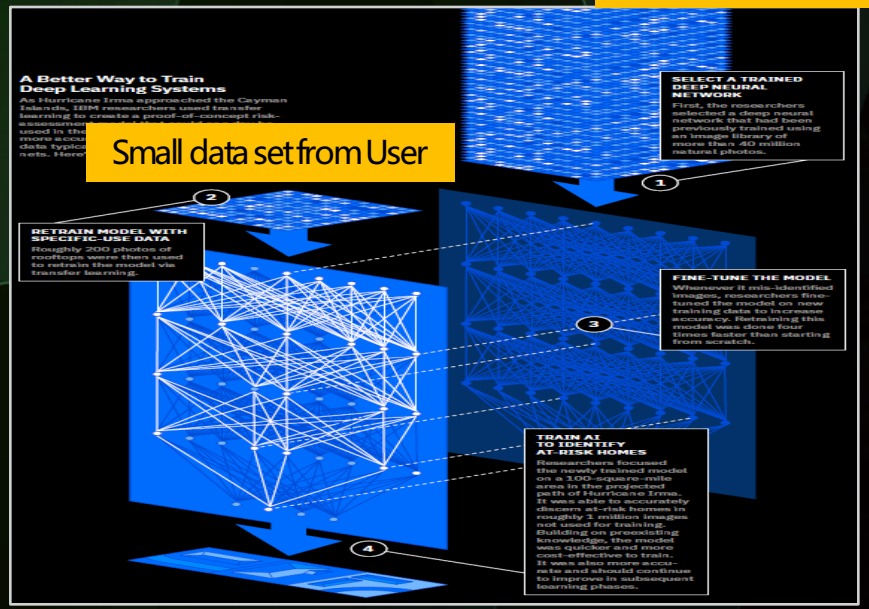
Machine	IBM P8
GPU	Tesla K80
GPU MEM	10G
CPU	ppc64le
CPU cores	160
MEM	55G
Frequency	3.5 GHz
OS	Ubuntu 14.04
caffe version	Py-faster-rcnn
Dataset	User-defined
Model	vgg16

Enterprise: I don’t have massive data

Transfer Learning for Learning from Small Data Set

- In lots of industry scenarios, we don't have huge data set.
- PowerAI Vision applied the optimized Transfer Learning technology for custom learning from small data set.

Good base model



Small data set from User

Base models supported by PowerAI Vision

- Flower (various flowers)
- Landscape (mountain, coast, forest, country side)
- Chinesefood (dumpling, rice, noodle, seafood, etc.)
- Action (fishing, reading, climbing, etc.)
- Scene (airport, street, building, campus, etc.)
- Face (human face)
- Vehicle (Jeep, Car, Sport Car, SUV, Van)
- Others (other scenarios)

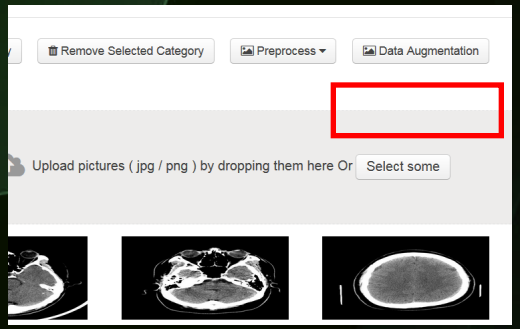
1. Select data scenario (base model) → 2. Add user's training data → 3. One-click to start training

Small data set, better accuracy, faster training

Data Augmentation for Learning from Small Data Set

- Data Augmentation can enhance the classification accuracy and reduce overfitting for small datasets
- Data Augmentation functions has been available on PowerAIVision

Fig.1 User could "one-click" and select different data augmentation methods



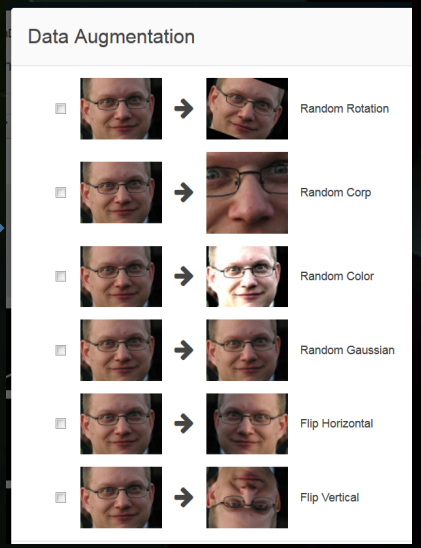
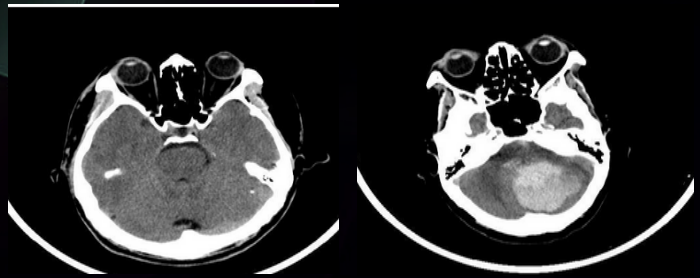


Fig.2 Data augmentation could improve the accuracy significantly

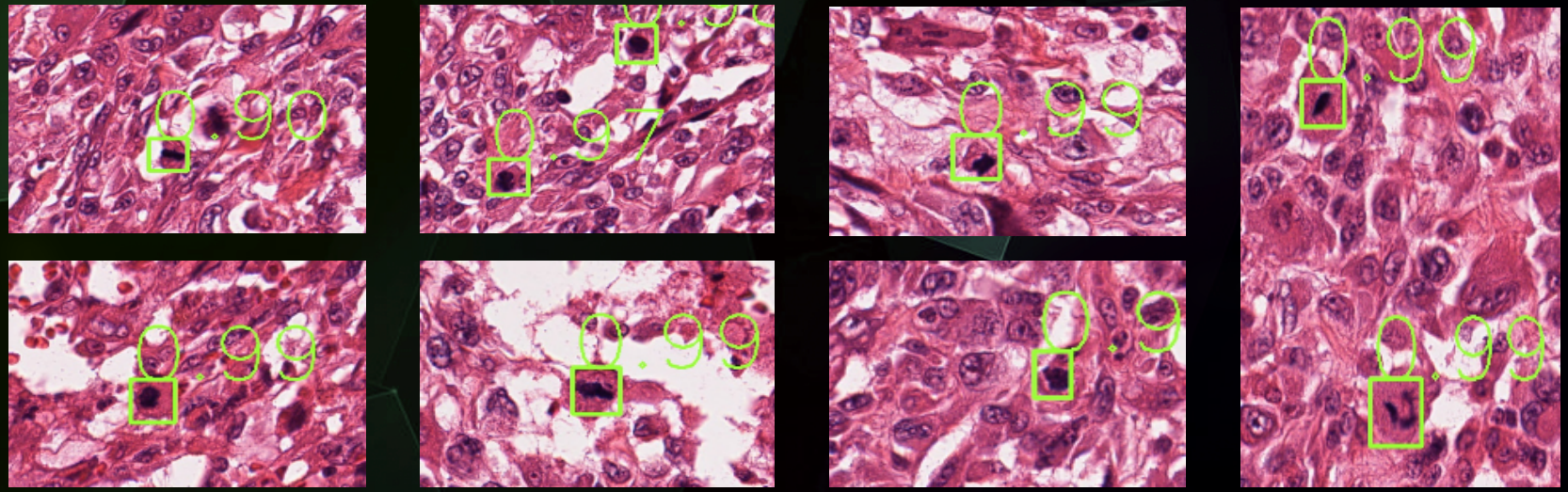
Medical image analysis for cerebral hemorrhage (脑出血)
(Original data: 157 pic.)



Accuracy: 97.9%

Example 2: Mitoses Detection for Medical Image

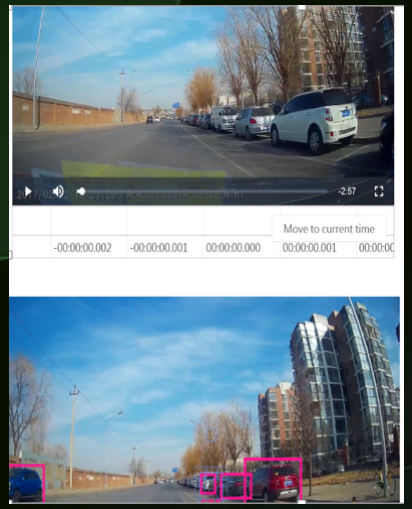
- Labeled data objects: 13
- Result: Detected 227 mitoses objects from 207 files.



Huge efforts on training data annotation

Semi auto-labeling : Reduce the time for data annotation

- **Semi – auto labeling** : To use AI technology for releasing most of human work for labeling (**10x ~ 50x**)



Manually label small data set

System will learn the objects for labeling

Auto – labeling by machines

Human review and adjust

Deep Learning vs. Communication

Image classification with AlexNet: ~1260 operations for each bit in image

Object detection with YOLO: ~900 operations for each bit in image

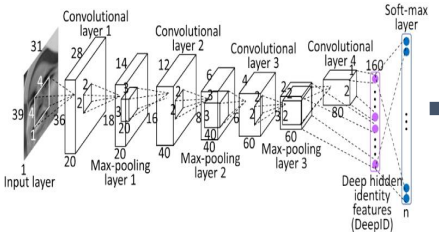
4G communication: ~2300 operations for each bit (transmit and receive)

AI should be deployed on edge

Enterprise: I don’t have experts knowing program for both deep learning and edge devices (e.g. FPGA)

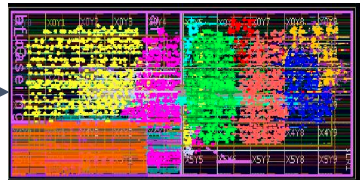
Automatically generate accelerator for deep learning

Automatically enable deep learning from cloud to edge – Enhance productivity



Trained Caffe CNN model in data center

AI Inference Engine tool



FPGA Accelerator bit-file for edge

translation

synthesis

download

Net Model File

```
name: "dummy-net"
layers { name: "data" ...}
layers { name: "conv" ...}
layers { name: "pool" ...}
... more layers ...
layers { name: "loss" ...}
```

Verilog File

```
--input module---
conv conv_instance(...)
pool pool_instance(...)
...more layers
loss loss_instance(...)
--output module---
```

FPGA Bit File

Net.bit

FPGA Execution

FPGA chip range from $20 to $1K

APICloud IT大咖说 知识共享平台

Infuse AI into Enterprises

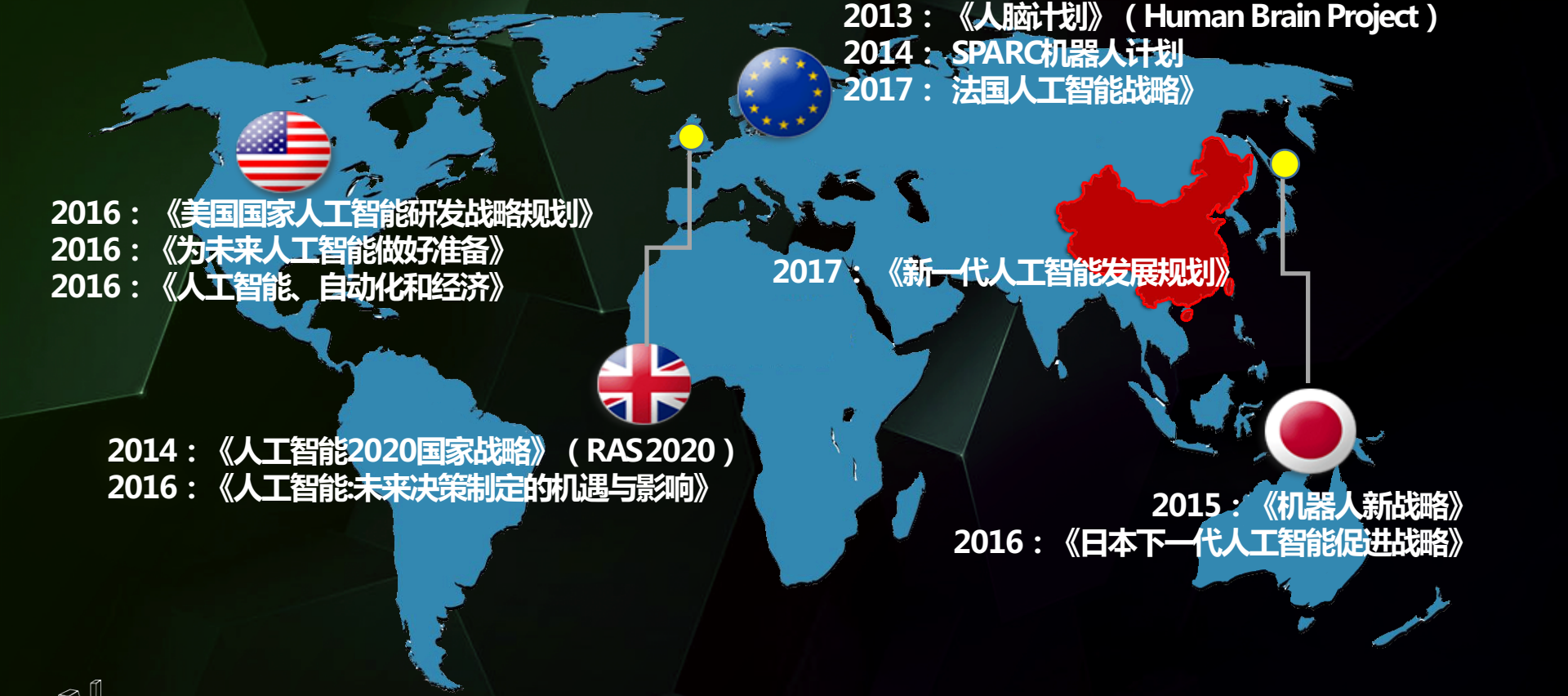
AI for AI

AI时代的移动技术革新

Era of AI: Innovations in Mobile Technologies

Backup

New Era of AI in the World



谢谢观看
THANKS