Big Data,
AI,
TT (Thinking Technology)

Eric Zhao

JD.京东 JD.COM

JD.COM 京东

Big data

Why big data?

Big value

big data + machine learning

Big sciences

Thinking technology is the future

Big Data

Big Volume

There is no real big data until web search

Fast Velocity

Fast processing data offline and online

Big Scale

Growing number of nodes and jobs

Open source

Open source makes big data popular

Hadoop Core Idea

JD.COM 京东

Hadoop

A large scale distributed system built for processing large volume of data

Partition data

Partition on key of key/value pairs

Divide computation on thousands of nodes

Coordinate and schedule tasks automatically

All these are old technology back in 90's parallel database system, but there was no big data

Based on application usage pattern, optimize for CPU, I/O, and memory

Column store

- **Read the data that you need instead of fetching every column**

Reduce disk I/O and utilize memory more aggressively

- **Spark**
  - Read once - compute everything - write once VS. Hadoop read-compute-write-read-compute-write
  - HDFS vs. RDD
- **Presto**

Challenges in “Big” Scale

JD.COM 京东

- **Handle huge volume of data with reliability**
- **Design flexible data model and schema to support business use cases and to minimize multiple copies of same data**
- **Robust job management system to support 10K+ nodes cluster**
- **High performance and throughput**
  - Support 100K+ offline jobs per day
- **Low latency for real-time data**

Big Data @JD

JD.COM 京东

Offline and Online Data Clusters

10000+

Cluster Size

Daily Jobs 200K+ Daily online data 15 Billion

200000+

Computing Load

Total Volume 100PB+ Daily Increase 15PB

100PB+

Data Size

Big data along machine learning generates real big business value

Two main learning approaches

- **Deep learning**
  - SL(Supervised learning) learns hidden features from **large data** and uses more layers NNs without feature engineering
  - RL (Reinforced learning) is a MDP (Markov decision process) for self-learning and self-improving
- **Bayesian program learning**
  - Use parameters prior distribution and **small data** to infer the whole distribution
  - LDA is the best known model

Supervised learning (SL)

- **SL with feature engineering**

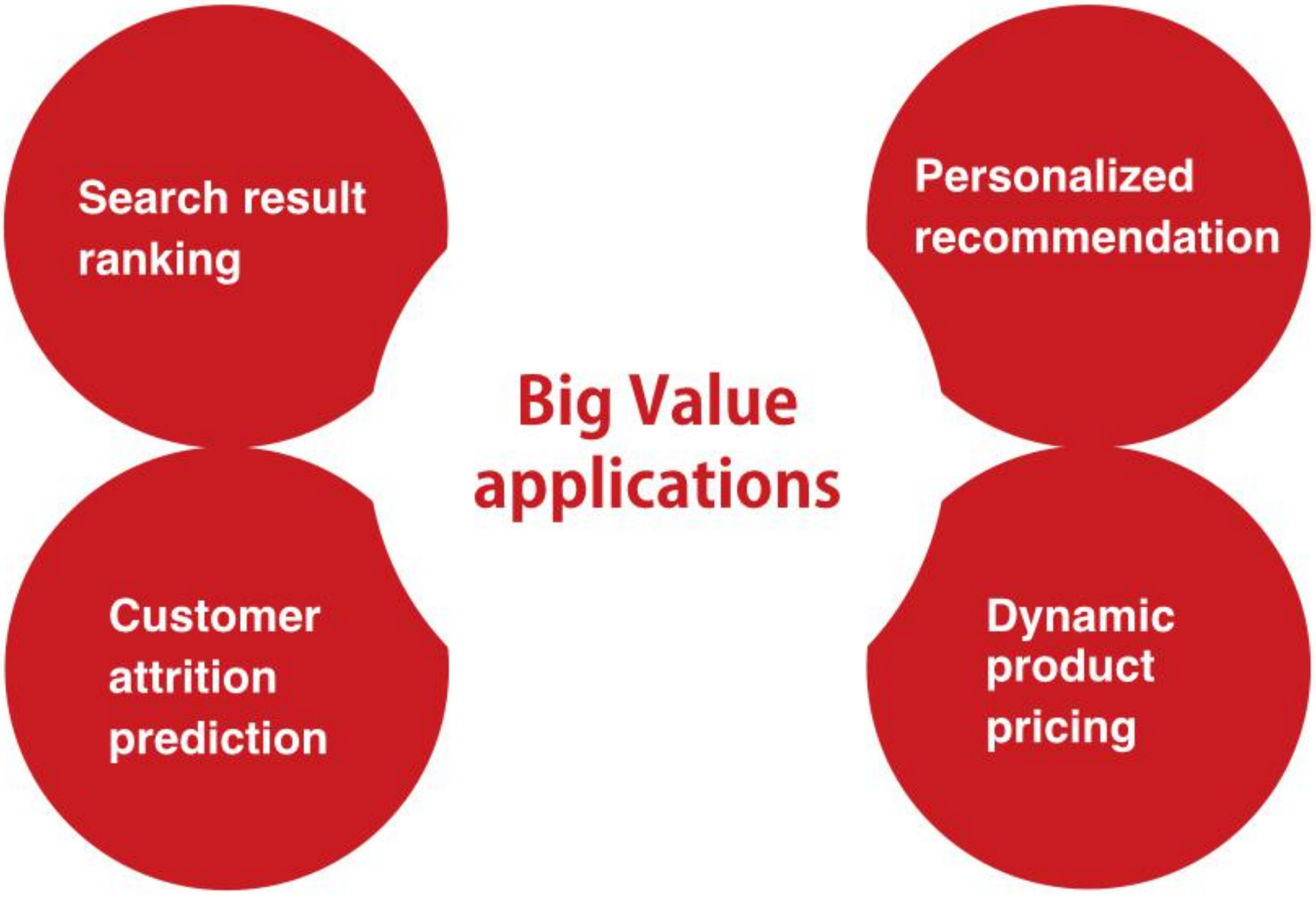
  - Feature engineering means handcrafted feature extraction
  - Regression, GBDT, and CRF

- **Supervised Deep learning**

  - Learn on **large data** without handcrafted feature extraction and use multi-stage non-linear feature transformation
  - Produces high quality results on learning large number of image, voice, and video data

Big Value applications

JD.COM 京东



Classic SL on Big Data

JD.COM 京东

- Accurate and high quality data is more important than big data.

- Big data is not equal to high coverage of the important features. The low coverage of key features sometimes means low model quality.

- High quality feature engineering is more important than the better science models.

- Real-time user feedback and behavior is the last mile of runtime model improvement.

- Self-learning and self-improving AI algorithms, which make better decisions than human does.

— AlphaGo

- Combine several SL deep learning models and RL model to make better strategy decision than human

- Reinforcement learning

  It allows machines and software agents to automatically determine the ideal behavior within a specific context, in order to maximize its performance

- JD is working on RL solutions to improve our core business decision making, such as product life cycle management.

- **Elon Musk's Open AI**
  - Open best AI algorithms to everyone

- Human needs drive technology innovation and advance
  - **Searching large web created Hadoop**
  - **Large social media data and proliferation of smart phone gives birth to deep learning**
  - **People want to make smarter and higher quality decision**
- Open source is a great movement

JD.COM 京东

Thanks!