

# 深度学习框架的性能优化 及其在医药行业的应用实践

SPEAKER

朱智勇  
英特尔亚太研发中心

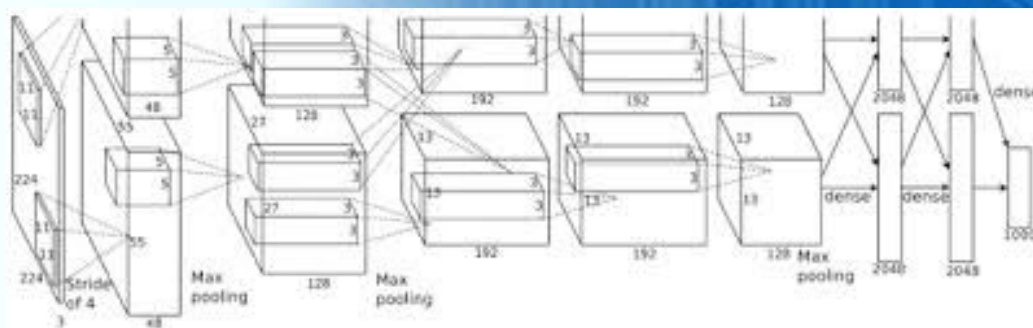
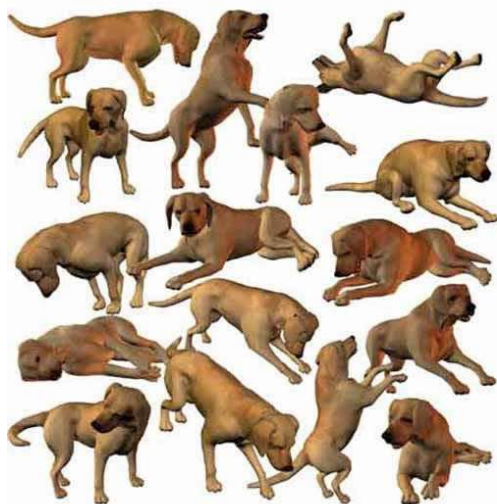


# 深度学习框架及性能优化

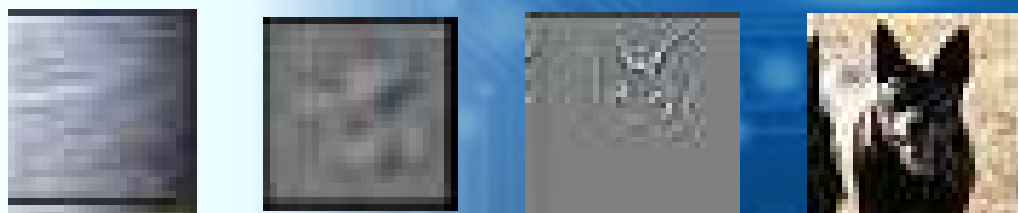
## 深度学习在医药行业案例分析

# 深度学习介绍

- 机器学习的一种
- 神经网络
- 深层线性和非线性
- CNN/RNN/DBN等模型
- 图形/图像/语音/文本等应用



→ “dog”



Deeper Layers of the Network → Higher Level Features

# 深度学习的性能优化

全面的软件优化

涵盖主要深度学习框架

浮点能力的优化

Cache和memory的优化

并行计算的支持

丰富的调优软件



theano

Caffe

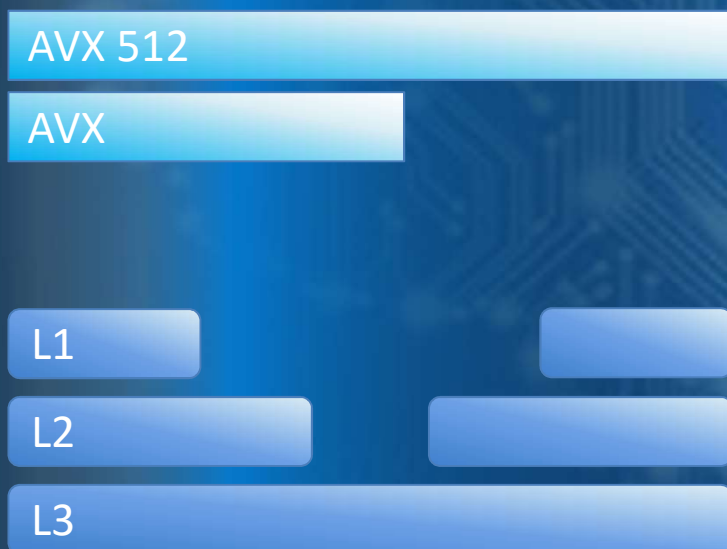


<https://software.intel.com/machine-learning/>

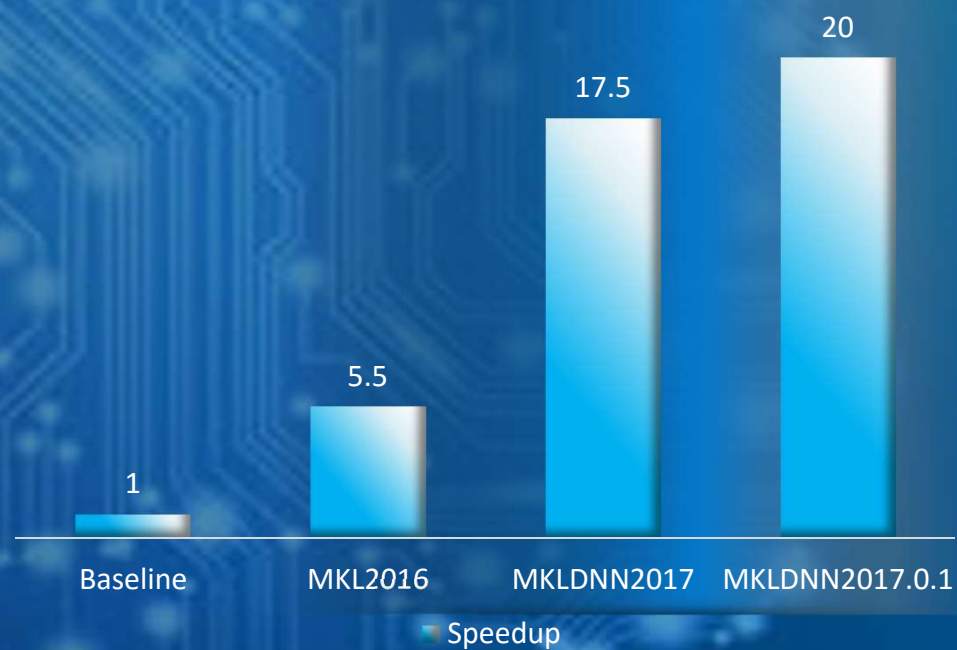


# 单节点优化

Single node optimization

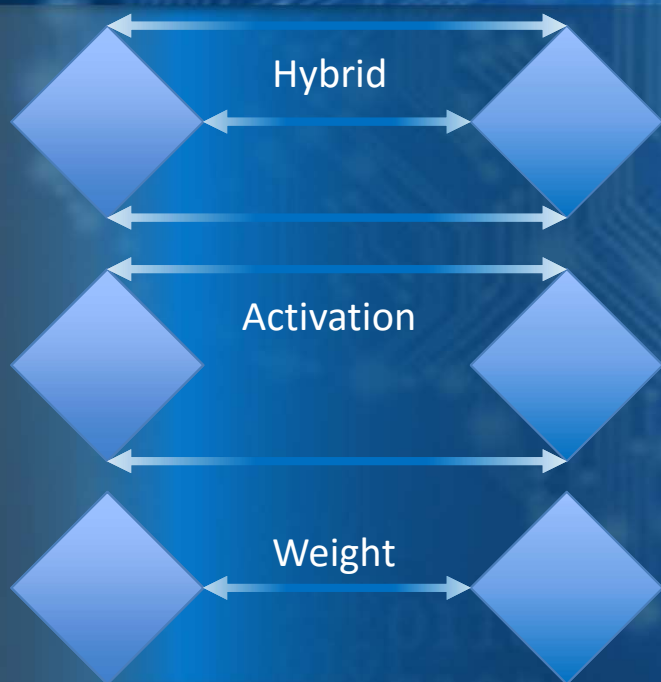


Alexnet training on Xeon BDW

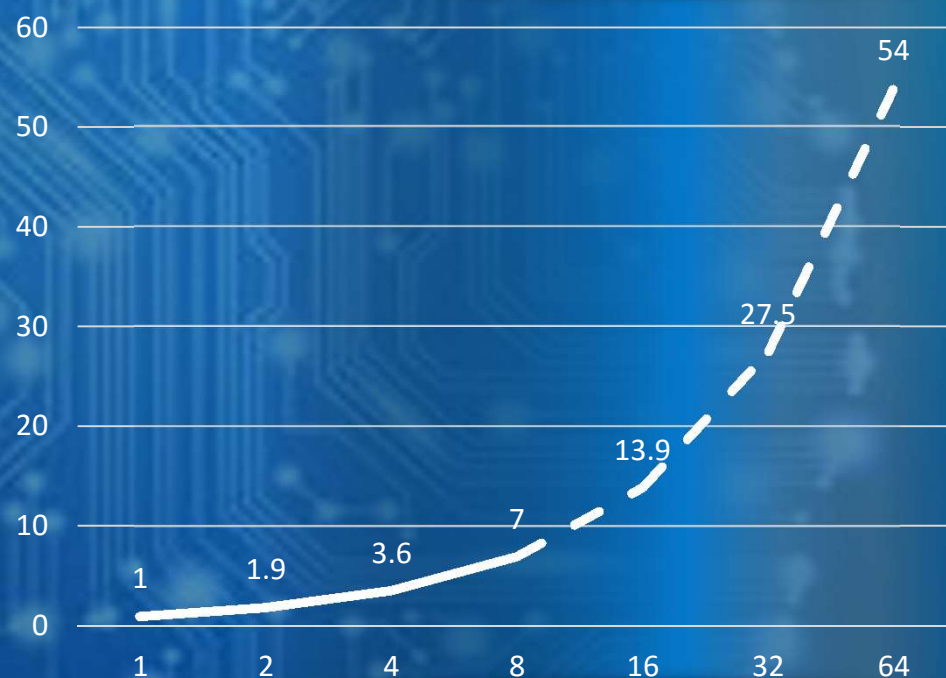


## 单节点性能不断飞跃

# 多节点支持



Training speedup on cluster



## 先进的并行技术

The background of the slide is a deep blue color. It features a faint, stylized graphic of a human brain. The brain's surface is covered with intricate white circuitry lines, resembling a printed circuit board. Interspersed among the circuitry are small, glowing white dots, some of which are arranged in a pattern that suggests binary code (0s and 1s). The overall effect is a high-tech, digital aesthetic.

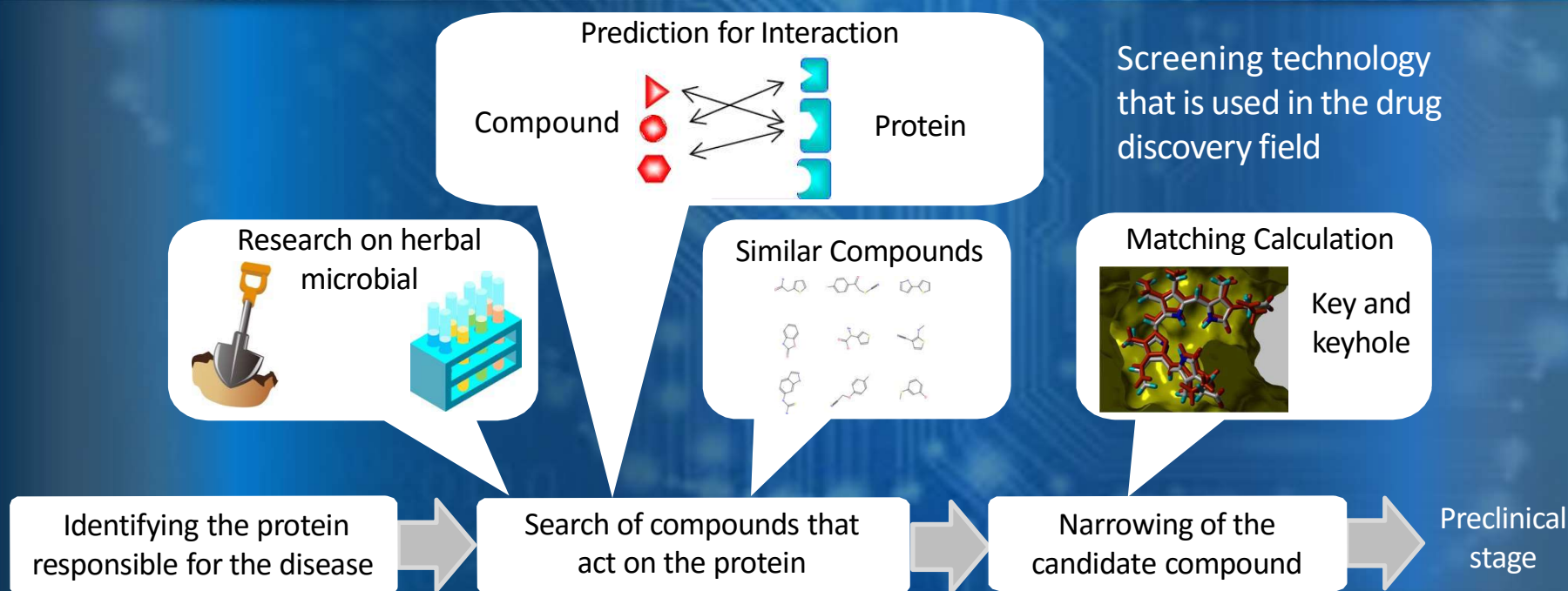
# 深度学习框架及性能优化

## 深度学习在医药行业案例分析

# 案例一：药物研发

医药研发中的问题：研发周期长，成功率低

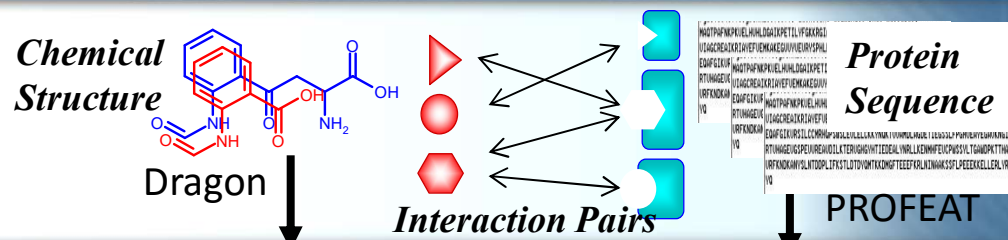
用深度模型：筛选候选化合物，缩小研究范围



## 深度学习在药物研发中的应用



# 化合物与蛋白质的匹配



将化合物表达为向量

将蛋白质表达为向量

**Vector**

**Compound Vector**: Mw, logP, #C, #OH, ...  
 (252, 7, 4, 5, ...)  
 (320, 1, 2, 1, ...)  
 (238, 6, 7, 4, ...)

**Protein Vector**: AA, AH, AS, ...  
 (72, 51, 47, ...)  
 (81, 53, 64, ...)  
 (60, 43, 48, ...)

**Interaction Vector**

(252, 7, 4, 5, ... 72, 51, 47, ...) **Bind**  
 (320, 1, 2, 1, ... 60, 43, 48, ...) **Bind**  
 (238, 6, 7, 4, ... 81, 53, 64, ...) **Bind**  
 (252, 7, 4, 5, ... 60, 43, 48, ...) **Non-bind**  
 (320, 1, 2, 1, ... 72, 51, 47, ...) **Non-bind**

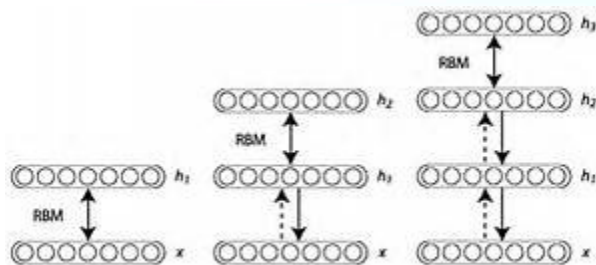
数据：向量的组合

**Query Pair**

(220, 3, 2, 3, ... 42, 31, 34, ...) **Bind or Non-Bind?**

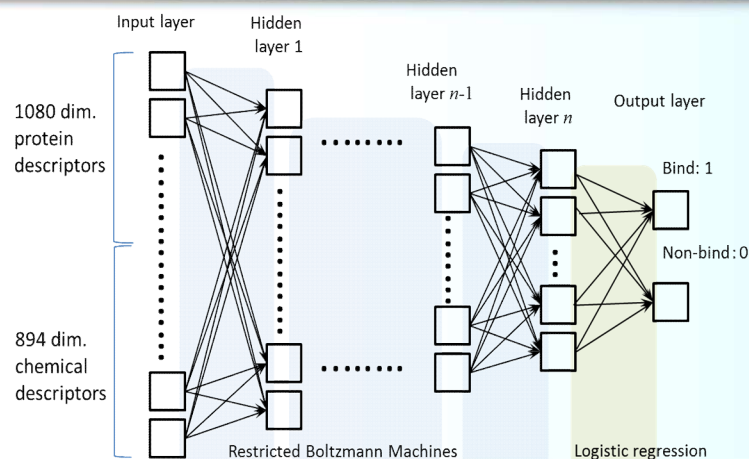
分类：是否有效组合

# 解决方案：Deep belief network



最终的方案：

- 基于Intel Xeon服务器
- 更多的内存
- 更快的训练速度
- 更多的数据
- 更高的预测精度

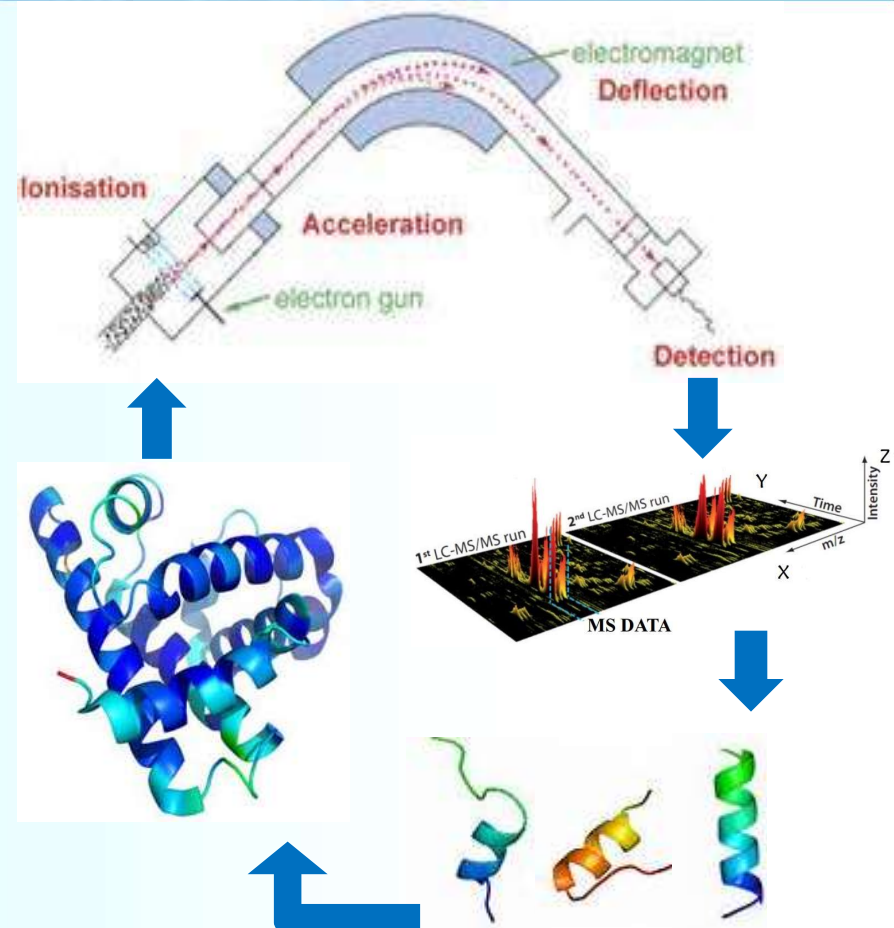


更多内存 更快速度 更高精度

## 案例二：蛋白质分析

### 蛋白质组学

- 蛋白质分解
- 双联质谱仪
- 肽链的分析
- 蛋白质的结构，表达和功能问题建模
- 预测肽链的出峰时间



蛋白质组学对于疾病研究有重要意义

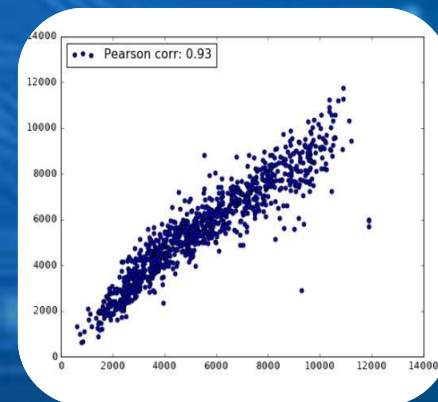
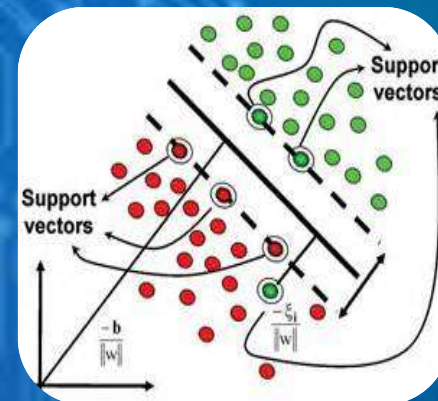


# 传统的预测方法

## 传统的预测方法

- 基于SVM
- 时间较长
- 准确率中等

Peptides	RT(min)
SALLALGLK	120.22
NALSSLWGK	122.51
YATLATVSR	53.88
YPMAGVLNK	71.74
NPITNALVR	83.94
QAYTQFGGK	42.86
VYGYVTNSK	??
FVYSLLGPR	??

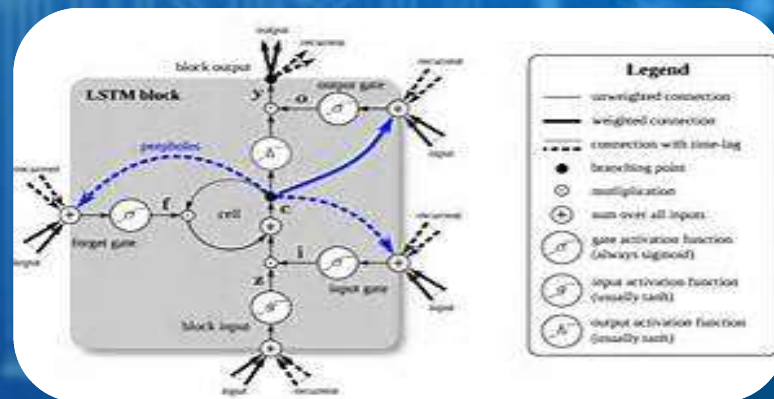
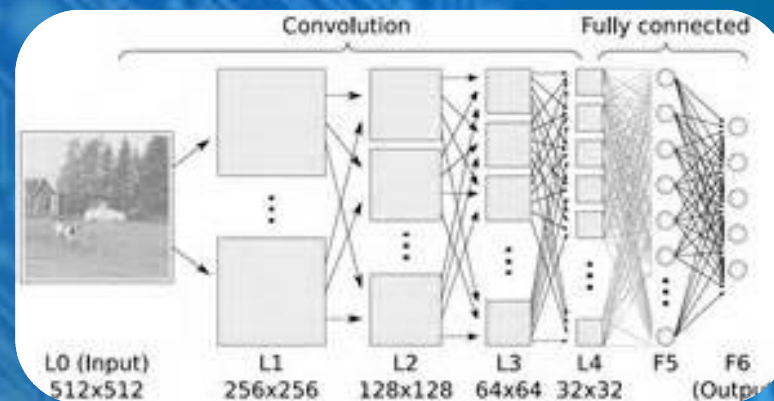




# 深度学习的方案

## 深度学习方案

- CNN模型
- LSTM模型
- 1/3训练时间
- 突破的预测准确率

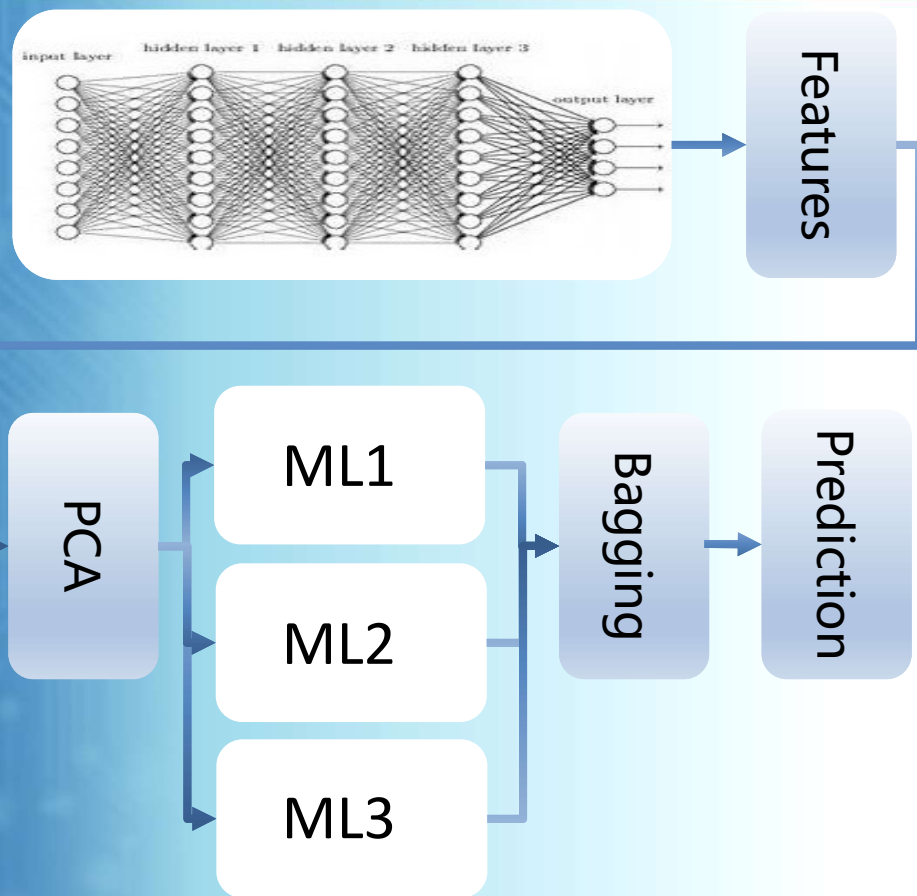


## 质谱仪 RT 预测的突破

# 深度学习 + 机器学习

## 组合方案

- 特征提取：CNN 和 LSTM
- 回归拟合：机器学习方法
- 预测结果：Bagging
- 最好的预测结果



深度学习 + 机器学习 = 更好结果

# 总结

全面的支持和优化，单节点及分布式

更多内存，更多适用性，更快速度，更高精度

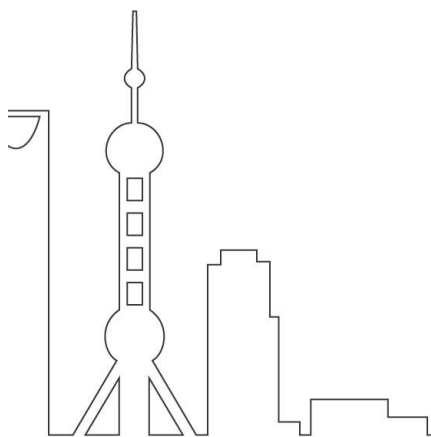
深度学习在医药领域有很好应用，其他行业也有机会

深度学习和机器学习结合，会有更好结果

# 联系方式

朱智勇(Steve)      邮件：[steve.zhu@intel.com](mailto:steve.zhu@intel.com)

微信：1256646377



# *Thanks!*

International Software Development Conference

主办方 **Geekbang**  **InfoQ**   
极客邦科技