Published online: July 29, 2016

Review

OPEN ACCESS

molecular systems biology

Deep learning for computational biology

Christof Angermueller[1,†], Tanel Pärnamaa[2,3,†], Leopold Parts[2,3,*] & Oliver Stegle[1,**]

信息学院　马再兴

目录

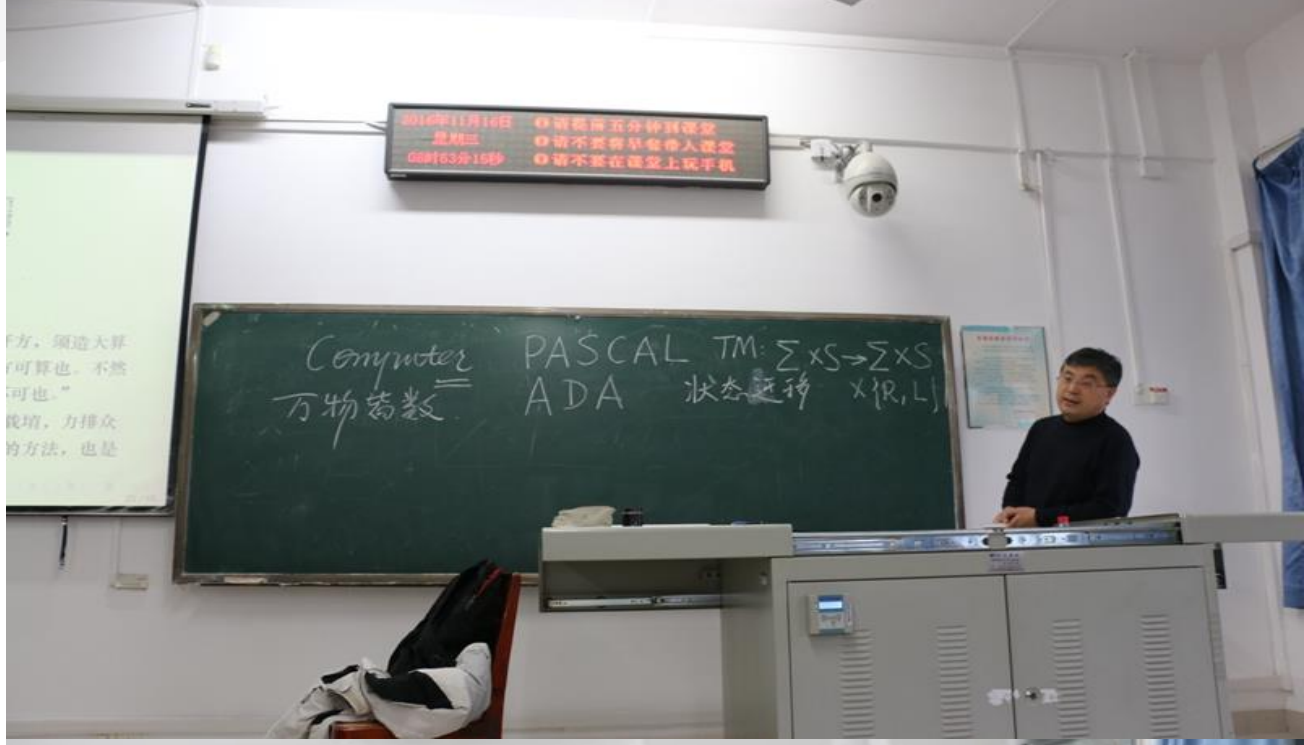
- 背景

基本概念

计算生物学

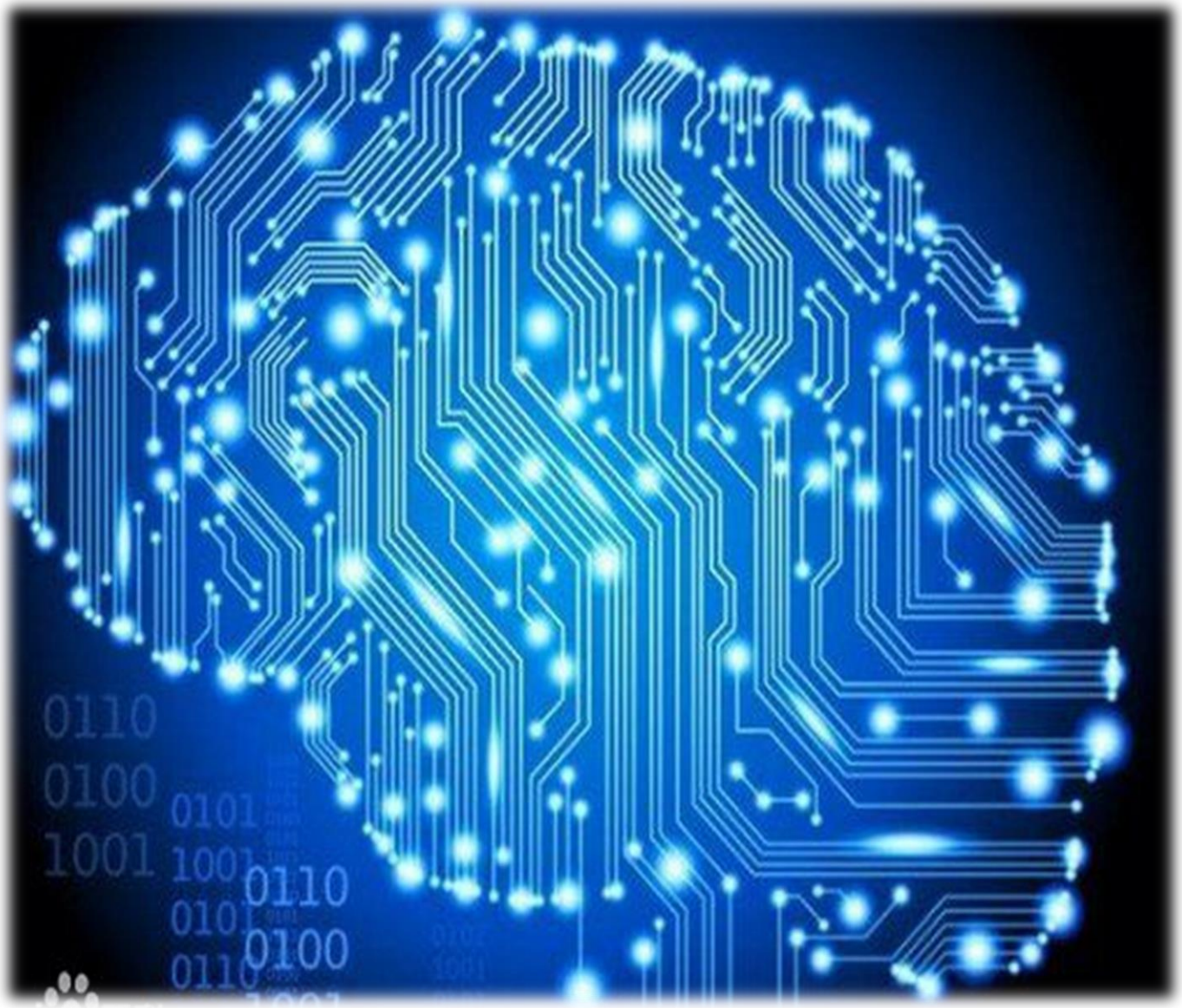
一句话解释：

关于数据分析的理论与技术，包括软件开发、数学建模、数字仿真



- 背景

基本概念



深度学习

一句话解释：

建立、模拟人脑进行分析学习的神经网络，它模仿人脑的机制来解释数据，例如图像，声音和文本

- 背景

基本概念

人工神经网络

一句话解释：

模仿动物神经网络行为特征，进行分布式并行信息处理的算法数学模型

计算生物学

深度学习的应用背景

生物数据增长

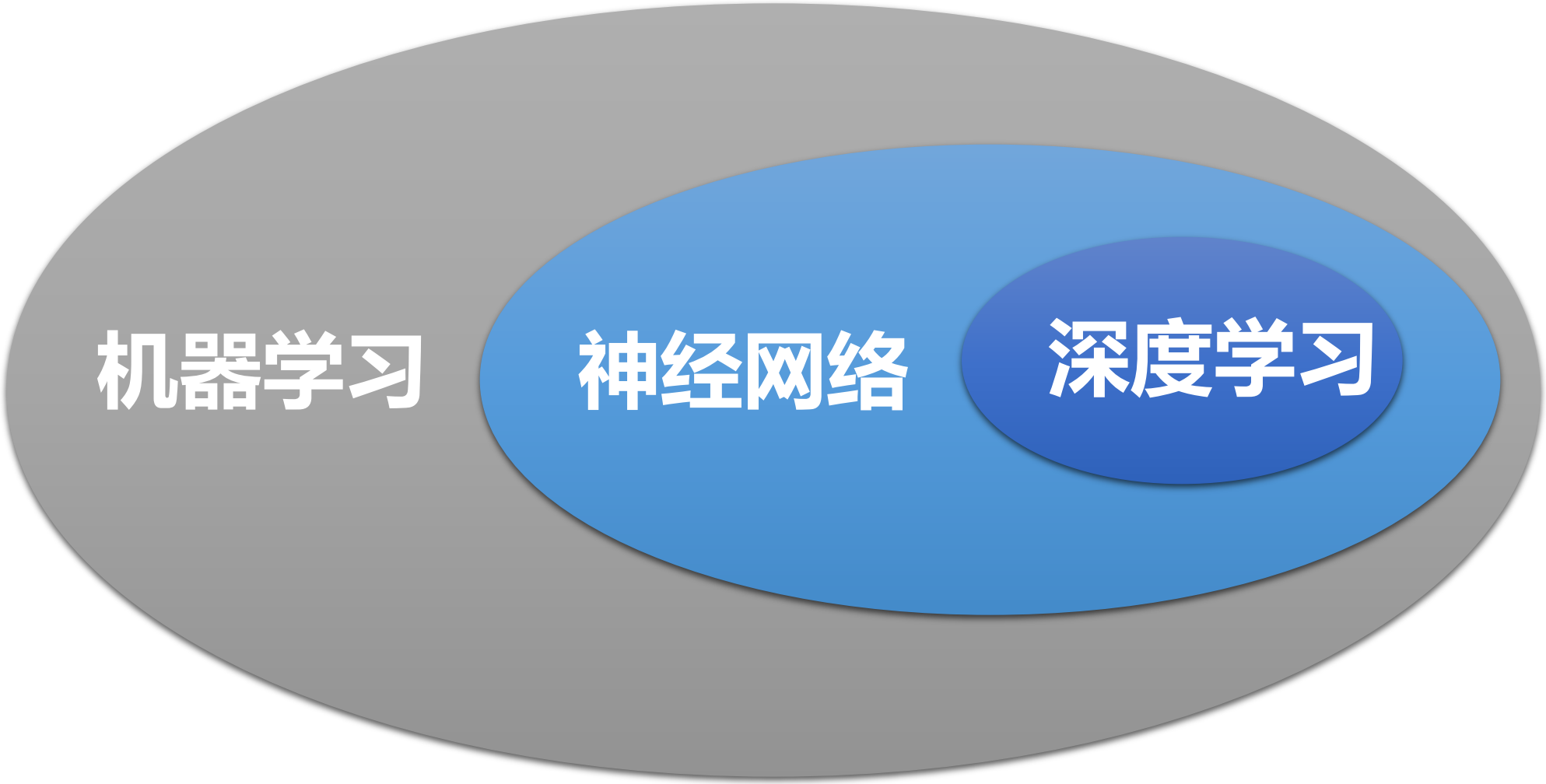
Illumina仪器测序所得的数据，每12个月就能翻一番;如果仅以摩尔定律来看，每18个月数据量就能翻一番

◆深度学习

◆利用大量数据寻找隐藏的内部结构

◆做出准确的预测

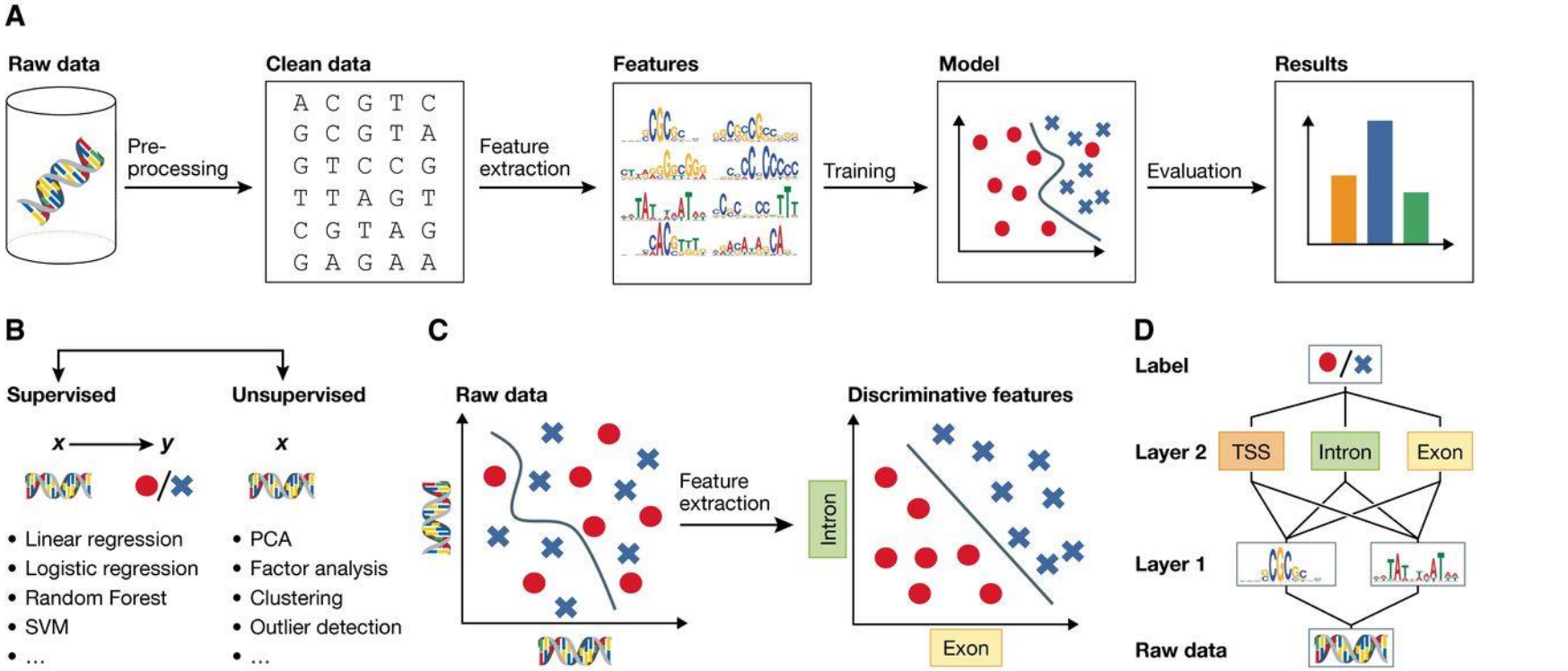
关系



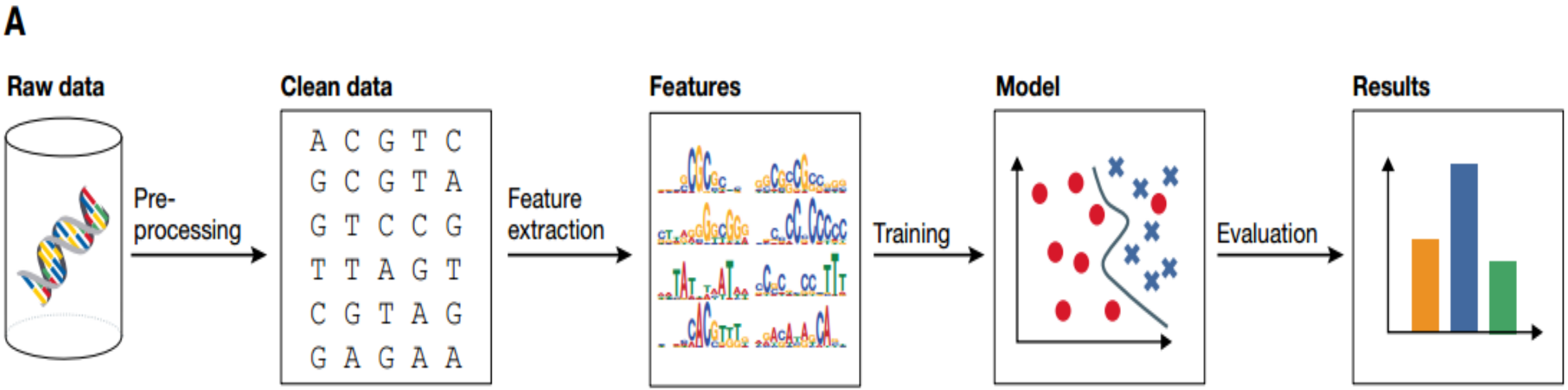
深度学习（Deep learning）

深度学习的基本知识与方法

Machine learning and representation learning



Christof Angermueller et al. Mol Syst Biol 2016;12:878



经典机器学习的四个步骤

（A）预处理-特征提取-模型训练-模型评估

监督机器学习

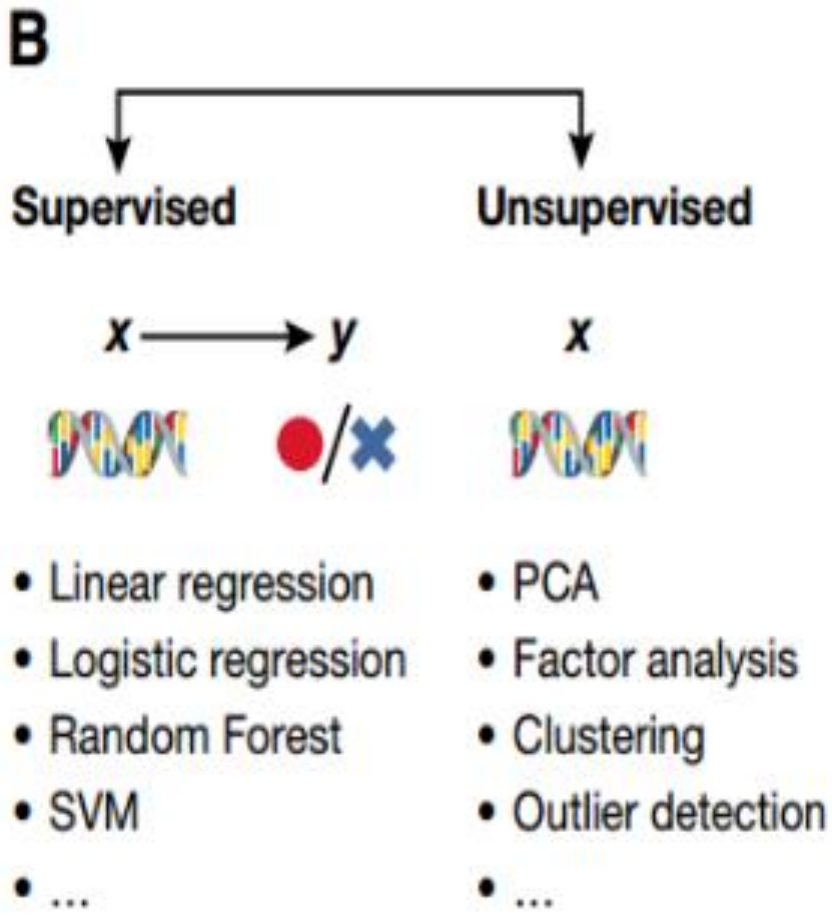
Supervised Machine learning

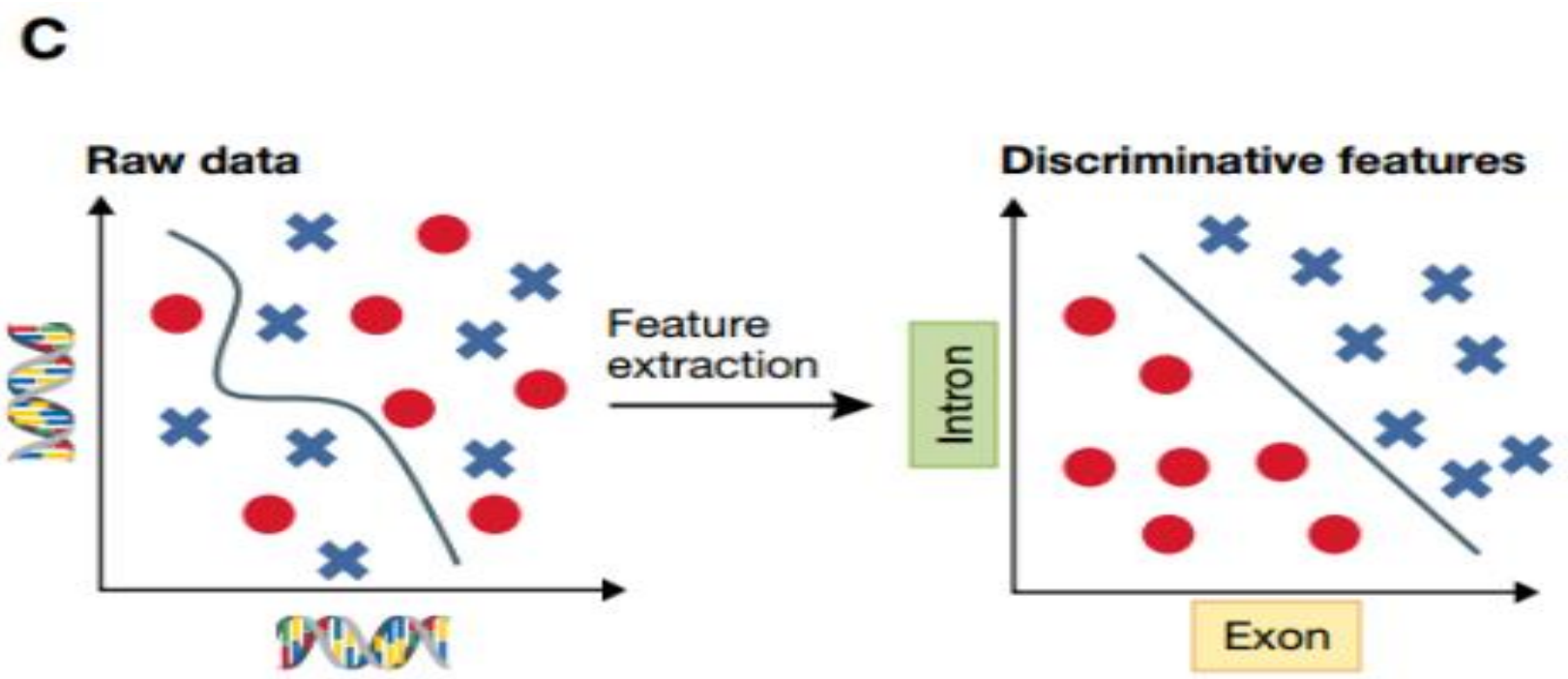
监督机器学习方法

输入特征x输出标签y

而非监督学习方法

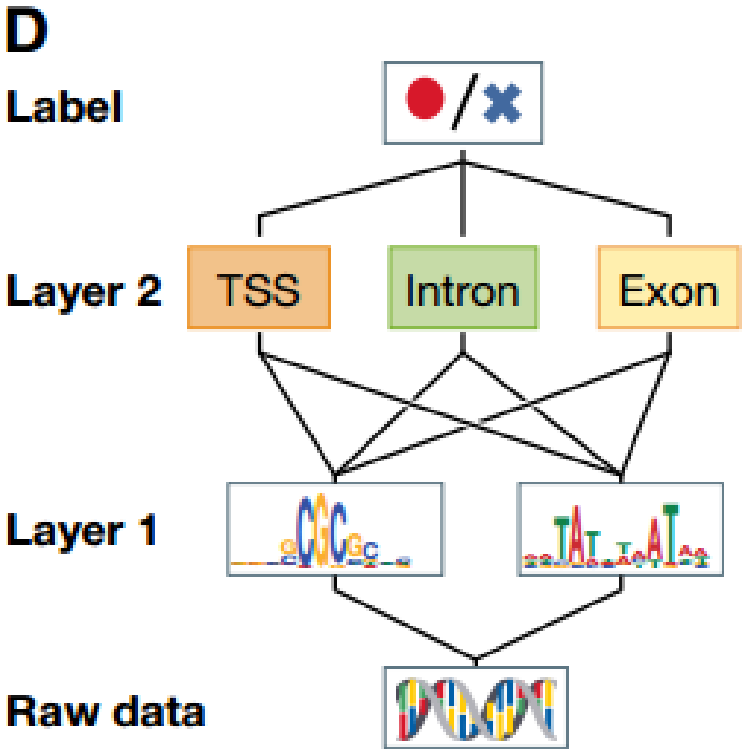
关于因素x,没有标签y





原始输入数据经常通过一个复杂方式高维及相关对应的标签挑战了许多经典的机器学习算法

深度网络通过一个层次结构从原始数据学习抽象的特性表征



DNA 测序

RNA测量

流式细胞术

自动化显微镜

深度学习（Deep learning）

深度学习应用：调控基因组学

生物图像分析

深度学习-调控基因组学

Deep learning-regulatory genomics

传统方法

使用(QTL)定位（需要大量样本）

使用有差异的训练模型内的基因组（不能直接作用于序列）

◆**深度学习**

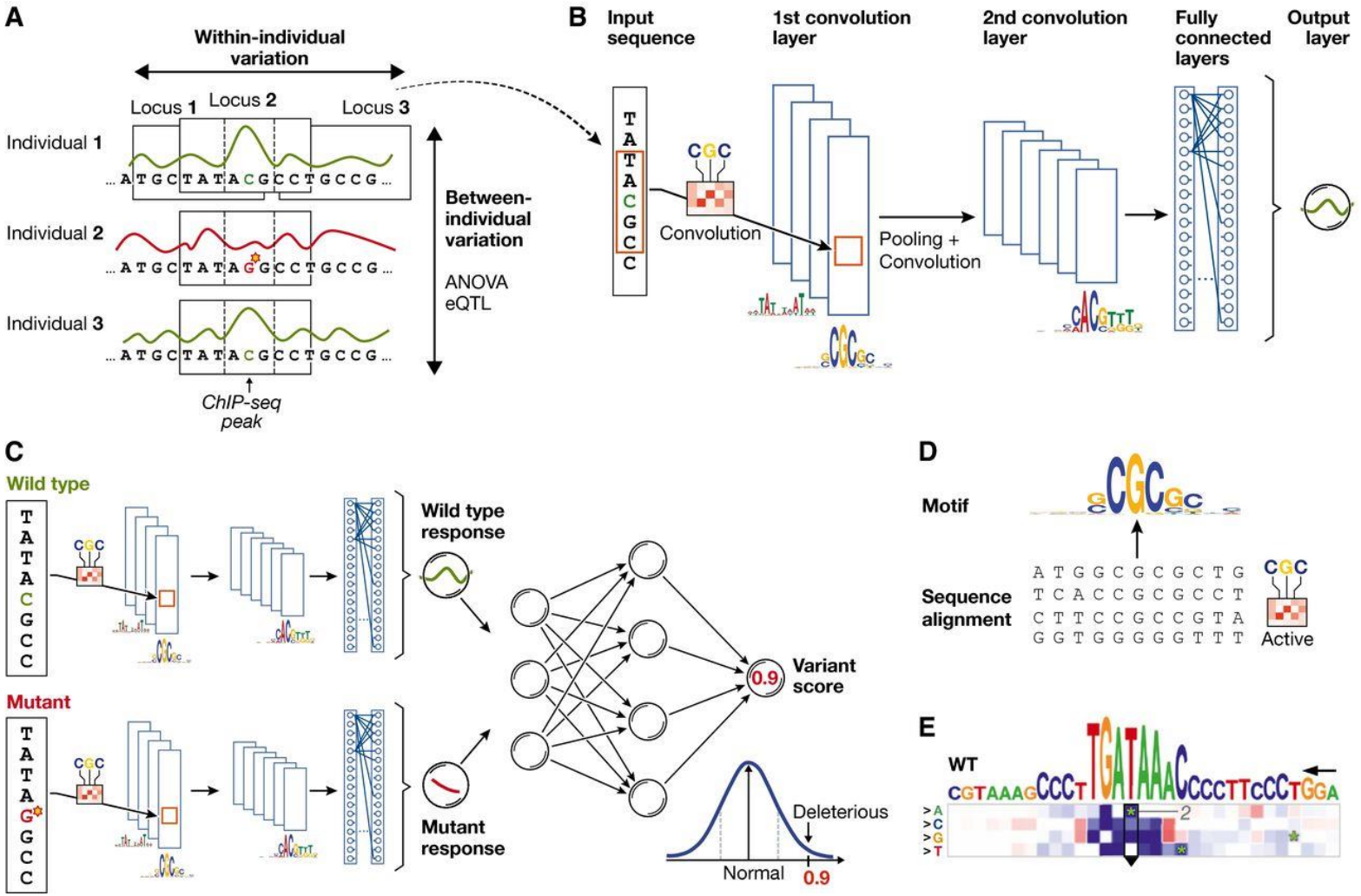
◆绕过手动提取数据

◆表征性丰富

（捕获序列间的非线性依赖关系，跨序列作用）

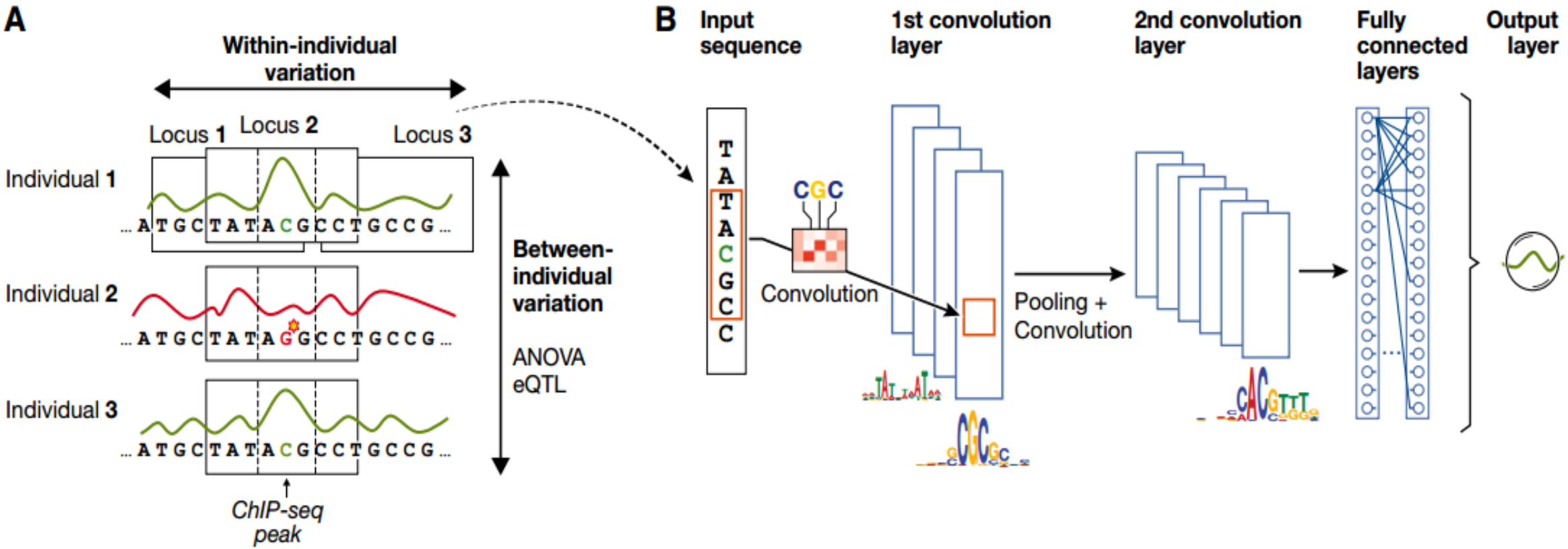
◆用于预测剪接，表观遗传标记

Principles of using neural networks for predicting molecular traits from DNA sequence



Christof Angermueller et al. Mol Syst Biol 2016;12:878

molecular systems biology



使用神经网络原理从DNA序列预测分子的特征

深度学习应用

深度学习的前期应用

前馈神经网络

预测个体的外显子剪接

精度更高，能识别罕见的错误拼接调节引起的突变

卷积设计

使用CNN 允许直接在DNA序列训练,而不需要定义功能

突变效果的芯片预测

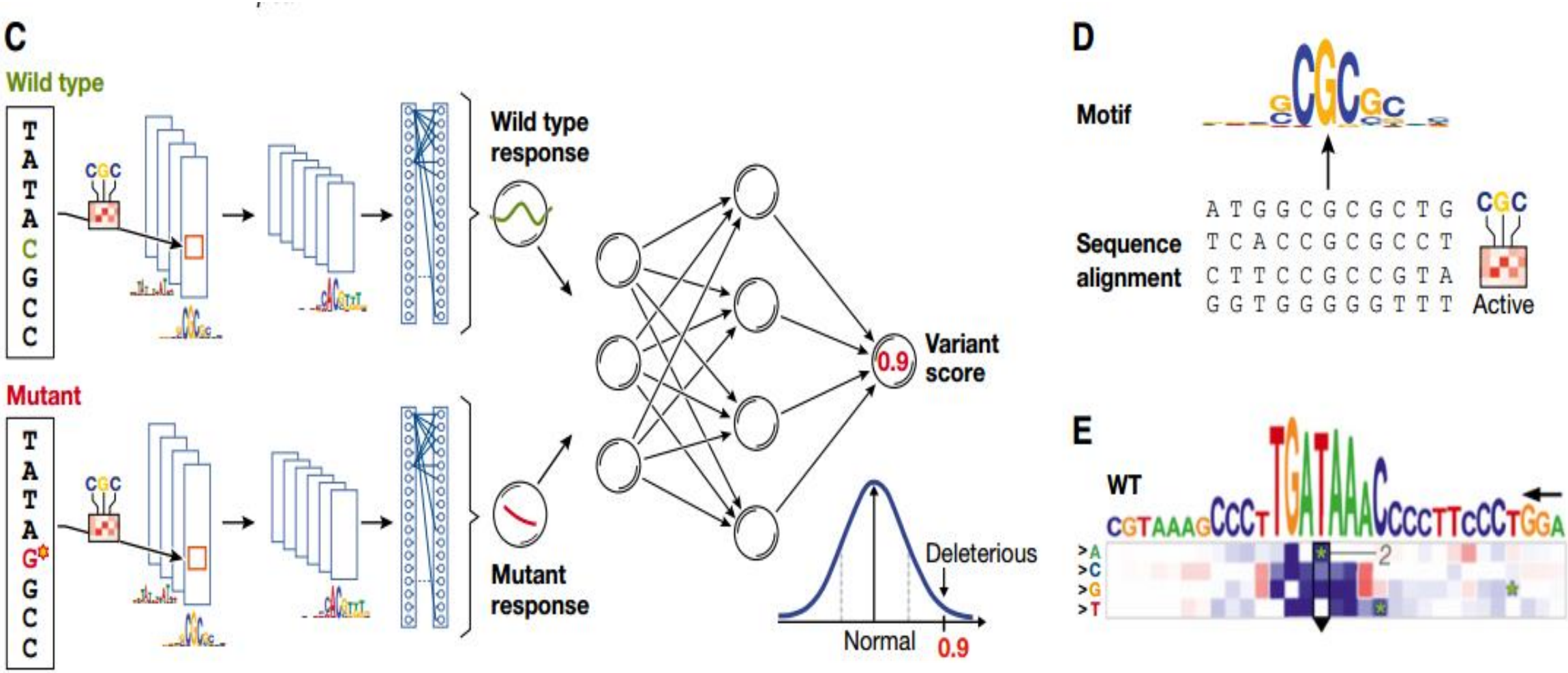
训练原始DNA序列来在硅片预测中突变的效果

多特征的联合预测及进展

从DNA序列预测染色质标记

预测并行多个染色质状态

无监督深度学习的架构



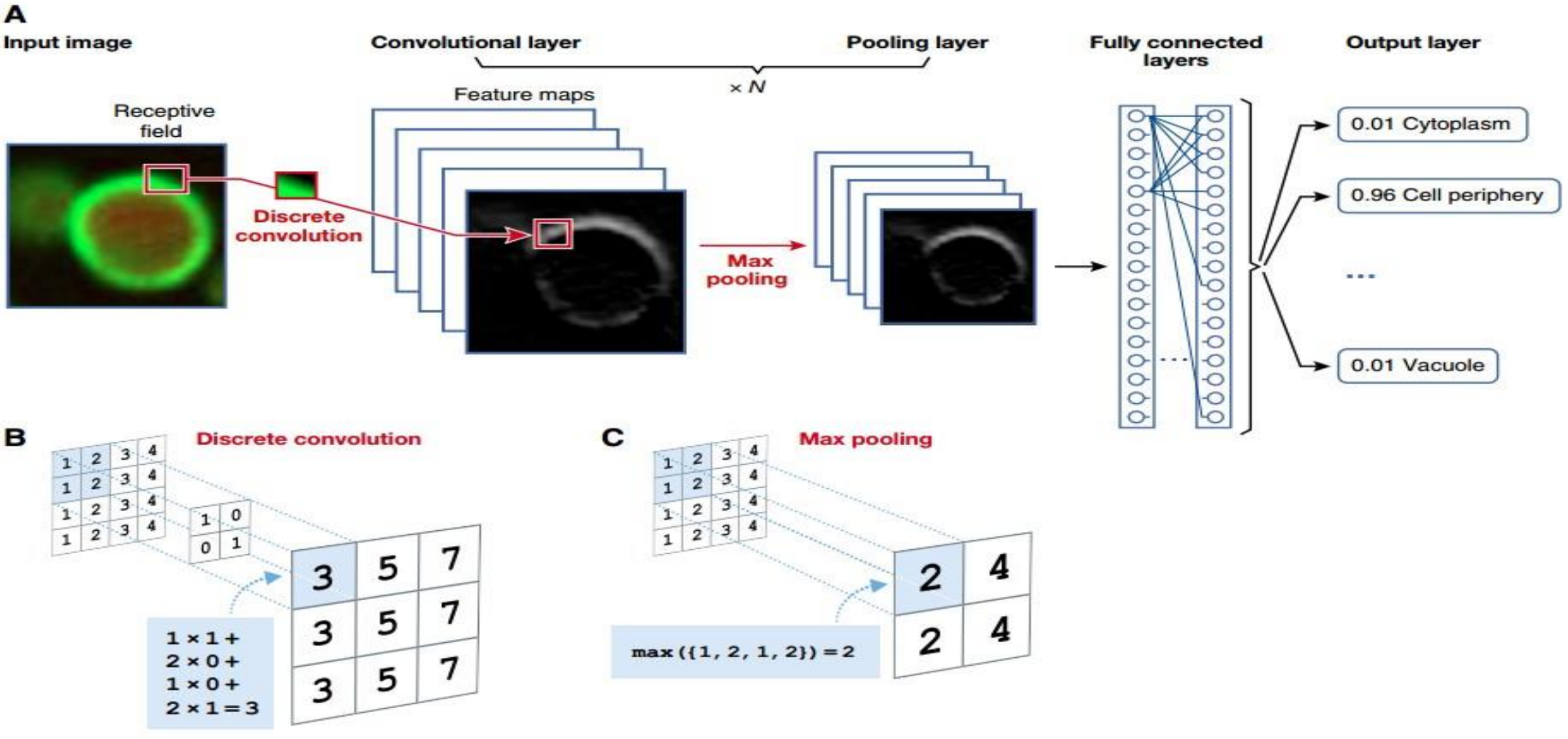
使用神经网络原理从DNA序列预测分子的特征

生物图像分析

最成功的的应用：图像分析

当前最先进的模型在利用卷积网络：

图像分类
目标检测、
图像检索语义分割



深度学习—生物图像分析

Deep learning—biological imagae analysis

深度学习-生物图像分析

Deep learning —biological imagae analysis

01 进行像素级别分类

02 细胞，细胞群，组织的分析 、重用经过训练的模型

03 解释、可视化卷积网络

04 隐藏重要图像部分、二维中可视化输入

Table 1. Overview of existing deep learning frameworks, comparing four widely used software solutions.

	Caffe	Theano	Torch7	TensorFlow
Core language	C++	Python, C++	LuaJIT	C++
Interfaces	Python, Matlab	Python	C	Python
Wrappers		Lasagne, Keras, sklearn-theano		Keras, Pretty Tensor, Scikit Flow
Programming paradigm	Imperative	Declarative	Imperative	Declarative
Well suited for	CNNs, Reusing existing models, Computer vision	Custom models, RNNs	Custom models, CNNs, Reusing existing models	Custom models, Parallelization, RNNs

现有深度学习框架

深度学习框架已经很容易从现有模块基础上构建高水平的神经网络.

软件处于开发阶段，目前只有几个预先训练模型可以使用

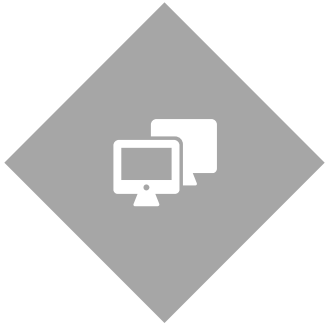
数据流程图

数据准备

所需数据大小

将数据分成训练集和测试集

规范原始数据



建立模型

模型结构选择

确定网络中神经元数量

模型训练

随机梯度下降

参数初始化

学习率和样本大小

学习衰减速率

动量

模型训练

先验参数学习自适应学习速率方法

批量常规化

分析学习曲线

监控培训和验证性能

注意避免过渡拟合 超参数优化 在GPU上训练

个人总结

深度学习的作为经典机器学习的工具和策略补充。

但没有任何一种的方法是普遍适用的。深度学习也是

传统分析方法仍然是有效的：当数据稀缺或如果目标评估统计学意义。

前景可观，但当一种新技术出现我们应该是冷静的对待。

需要改进

- 深度学习的发展还不成熟，相关软件还在开发
- 模型设计复杂、计算环境复杂
- 选择使用是否使用本身就是个问题

THANK YOU.