

第十三章 生物信息学实验基础

陈迪俊

南京大学

目录

第一节 生物信息学编程概述（生物技术与信息技术融合）

第二节 Linux系统及编程语言简介（生物问题 → 信息表示）

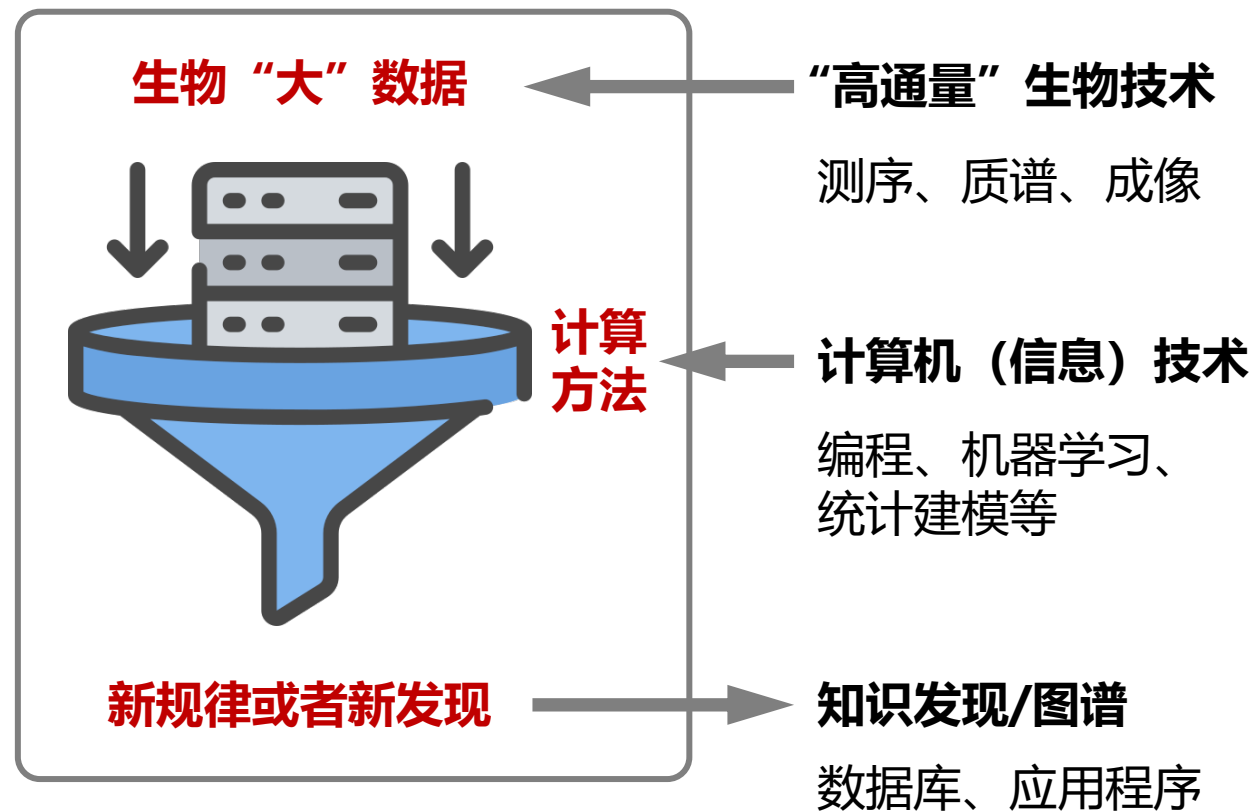
第三节 数据分析流程搭建（生物数据 → 知识发现）

第四节 数据库开发基础（生物数据/知识图谱 → 数据再利用）



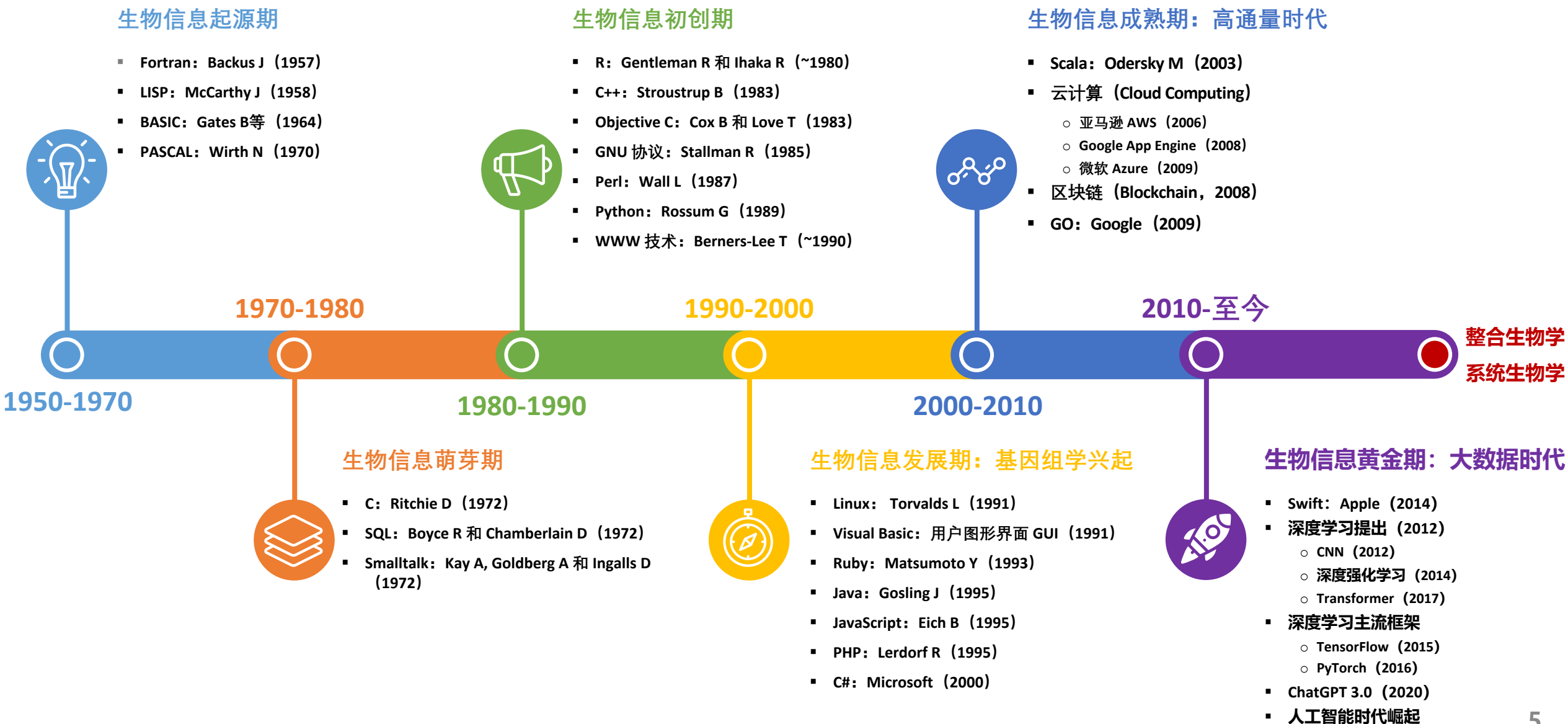
• 生物信息学编程

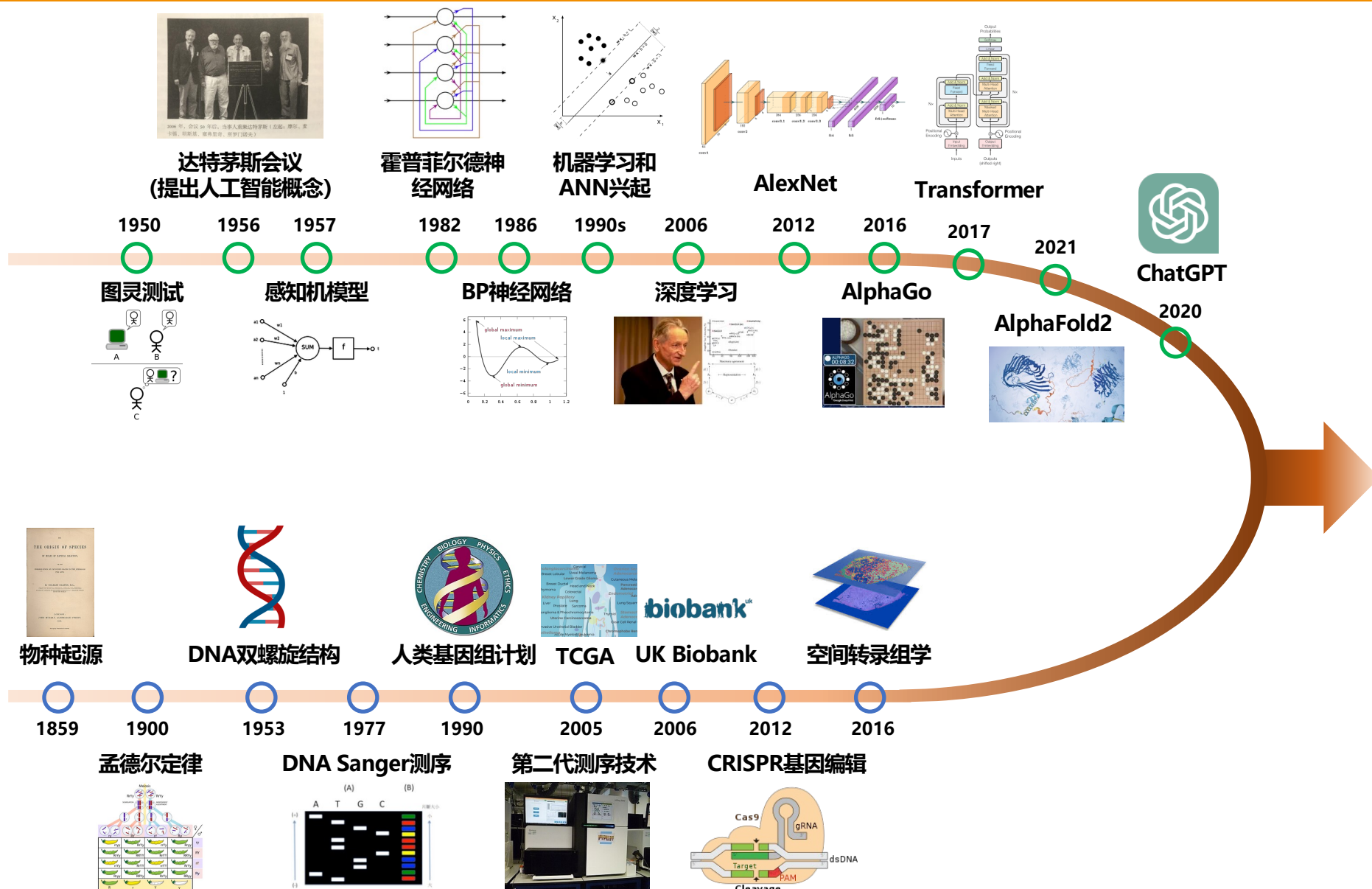
- 利用**计算工具和算法**来处理和分析**生物学数据**，以解答生物学**问题**和推动生物科学**新发现**。
- **目的**是将复杂的生物问题转化为可以用信息化方法（包括代码和模型）表示并进行计算和分析从而发现新规律的过程。



- 生物问题 → 信息表示 → 新规律/发现

生物信息起源期





- 第一节：生物信息学编程概述
- 第二节：Linux系统及编程语言简介 (Linux、Python、R)
- 第三节：数据分析流程搭建 (组学“大”数据 → 知识发现)
- 第四节：数据库开发基础 (生物数据/知识图谱 → 数据再利用)



目录

第一节 生物信息学编程概述（生物技术与信息技术融合）

第二节 Linux系统及编程语言简介（生物问题 → 信息表示）

第三节 数据分析流程搭建（生物数据 → 知识发现）

第四节 数据库开发基础（生物数据/知识图谱 → 数据再利用）



Linux 是一个开源的类 Unix 操作系统

<

Linux是生物信息学实验的首选操作系统之一。由于开源性、稳定性和强大的命令行工具，Linux成为了

Linux系统的组成部分包括操作系统的各个层次和组件，它们共同协作以提供一个功能齐全、稳定和高效的计算环境。

- **内核 (Kernel)**：操作系统的核心部分，它负责管理硬件资源，包括处理器、内存、设备驱动程序、文件系统等

- **服务和守护进程**：如Web服务器（如Apache、Nginx）、数据库服务器（如MySQL、PostgreSQL）、邮件服务器（如Postfix、Exim）、日志服务（如sys

Linux文件系统是Linux操作系统用于管理和存储数据的结构化方法。它负责组织文件、目录，以及数据在硬盘或其他存储设备上的存储和检索。

Linux 文件系统采用三种基本权限来管理对文件和目录的访问：

- <

Shell是一个应用程序，它连接了用户和Linux内核，让用户能够更加高效、安全、低成本地使用Linux内核。**Shell脚本语言**是一种在Linux操作系统中由Shell（命令解释器）执行的

管道技术将多个命令连接在一起，将一个命令的输出

Shell输入/输出重定向是一种强大的技术，允许控制命令如何处理输入和输出。在生物信息学中，它可以用于从文件中读取数据、将输出保存到文件、以及管道操作的数据流重定向。

script.sh

```
#!/bin/bash
```

```
#
```

R是一种专为数据分析和可视化设计的统计编程语言和环境，广泛用于生物信息学等数据科学、商业分析等领域。

安装 R 包

-

- ❑ dplyr 包的 “%>%” 运算符实现了

R提供了广泛的统计方法：

-

ggplot2 是 R 语言中最流行和强大的数据可视化包之一，由 Hadley Wickham 开发。它基于“语法图形”（Grammar of Graphics）的理念，允许用户通过定义

- ❑ 高效数据处理: tidyverse、dply、...r
- ❑ 数据可视化: ggplot2、...
- ❑ 机器学习与统计建模:

- Python是一种通用编程语言，以其简洁易用的语法和丰富的库生态系统而闻名，广泛应用于数据科学、机器学习（包括深度学习）、

- ❑ 易于学习和使用
- ❑ 广泛的应用领域
-

□ 通过Anaconda来安装并管理Python环境

□ 开发工具：

- ❑ 赋值语法: 变量名 = 值

目录

第一节 生物信息学编程概述（生物技术与信息技术融合）

第二节 Linux系统及编程语言简介（生物问题 → 信息表示）

第三节 数据分析流程搭建（生物

- **数据分析流程 (data analysis pipeline)** 是指一系列有序的步骤和方法，用于处理和分析生物学数据，以便从中提取有用的信息和



1. **明确任务和目标：** 确定要执行的任务和分析的目标。
2. **准备**

数据分析流

- **workflow 管理软件通常设计用于处理大规模数据分析**

➤ 第1步: 读数比对 (mapping reads)


```
SAMPLES = ["A", "B"]
```