计算分子生物学导论

中国医学科学院基础医学研究所

陈阳

yc@ibms.pumc.edu.cn

20250325

课程目标

- 了解计算生物学的发展历史与最新进展。
- 掌握常用生物信息学分析工具。
- 运用计算分子生物学方法推动自己的课题研究。

课程教学主要内容

序号	教学内容	授课教师	课时
1	计算分子生物学导论	陈阳	3
2	转录物组技术与数据分析	陈阳	3
3	单细胞转录物组技术与数据分析	陈阳	3（大作业）
4	空间转录组技术与数据分析	陈阳	3
5	表观基因组技术与数据分析	陈阳	3（大作业）
6	3D/4D基因组技术与数据分析	陈阳	3
7	蛋白质组与代谢物组技术与数据分析	李雪媛	3
8	生物分子网络	陈阳	3
9	人工基因线路与精准医学	谢震（清华大学）	3（大作业）

课程教学考核方式

考核方式：二次课后大作业

- ✓ 第一次，单细胞转录物组数据分析实践。30
- ✓ 第二次，表观基因组数据分析实践。30
- ✓ 第三次，医学人工基因线路设计。 30
- ✓ 课堂问答。 10

TA: 张玉祺

15610920026@163.com

大纲

- 1. 基本概念
- 2. 分子生物学发展简史
- 3. 计算科学发展简史
- 4. 课程常用计算软件工具介绍

分子生物学

- 分子生物学是研究生物体内分子结构、功能和相互作用的科学领域。
- 它探究生命现象背后的分子机制，涉及到DNA、RNA、蛋白质等生物大分子的结构、功能和调控。
- 分子生物学的研究范围包括基因的表达调控、细胞信号传导、蛋白质合成及其功能、细胞命运决定等。

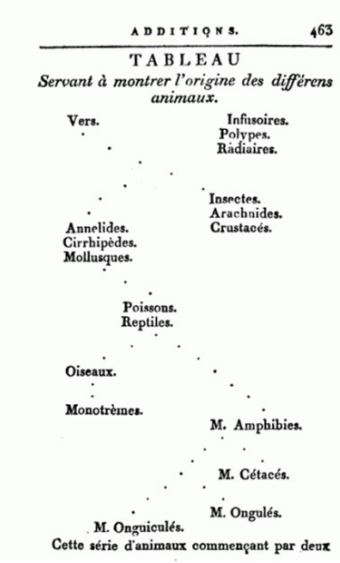
计算分子生物学

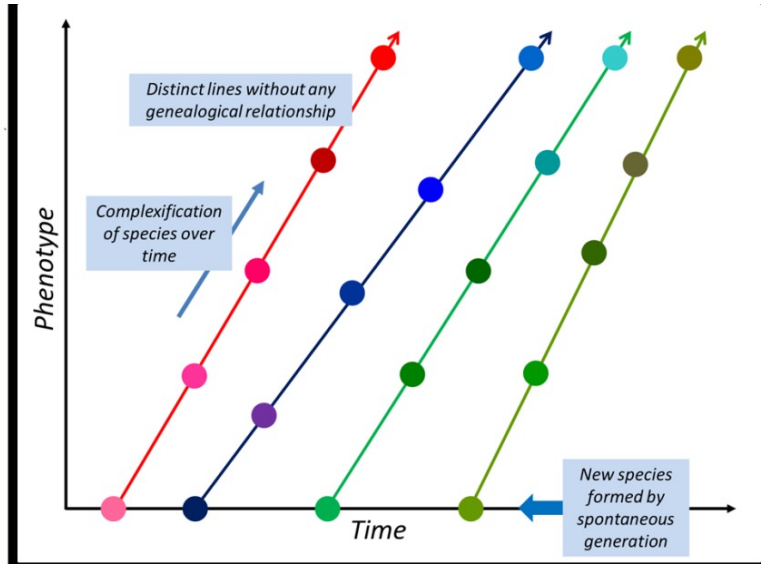
- 研究方式
  - 使用大数据分析、数学建模、计算机模拟等技术
- 研究内容
  - 生物分子的结构、功能和相互作用
  - 蛋白质、核酸、糖基以及分子间相互作用
- 发挥作用
  - 药物设计、疾病诊断和治疗、生物化学反应机理研究
- 新兴应用
  - 数字化生物技术与生物工程

发展历史简史

- 生物学
  - 拉马克“用进废退”和“获得性遗传”
  - 达尔文提出进化论
  - 孟德尔提出遗传定律
  - 约翰森首次提出基因一词
  - 摩尔根
    - 创立遗传的染色体学说
    - 提出基因的连锁互换定律
  - 艾佛里、麦克劳德、麦卡蒂验证DNA是生物的遗传物质
  - Chargaff提出了A=T,G=C的Chargeff规则
  - Sanger测定了胰岛素的氨基酸序列
  - 沃森和克拉克建立DNA双螺旋结构模型
  - 克里克提出中心法则
  - 科学家破译了遗传密码子

1809年获得性遗传学说

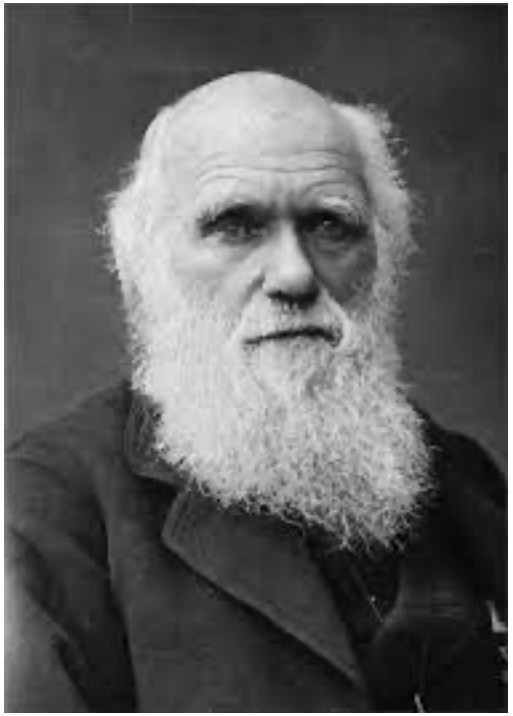


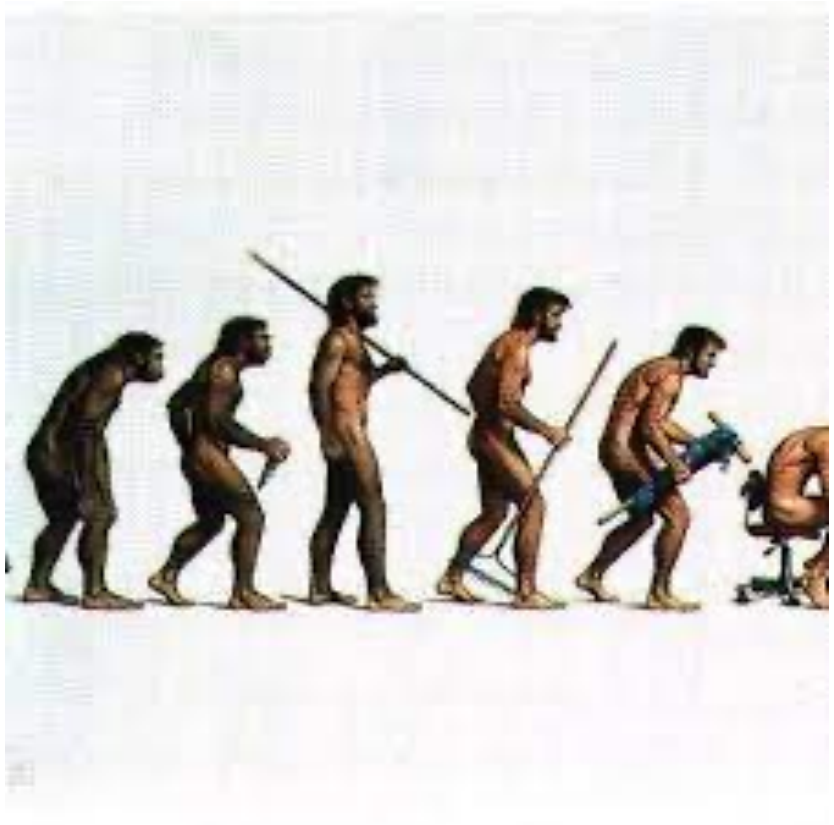


- 法国科学家拉马克
- 1809年“用进废退”、“获得性遗传”

1859年达尔文进化论

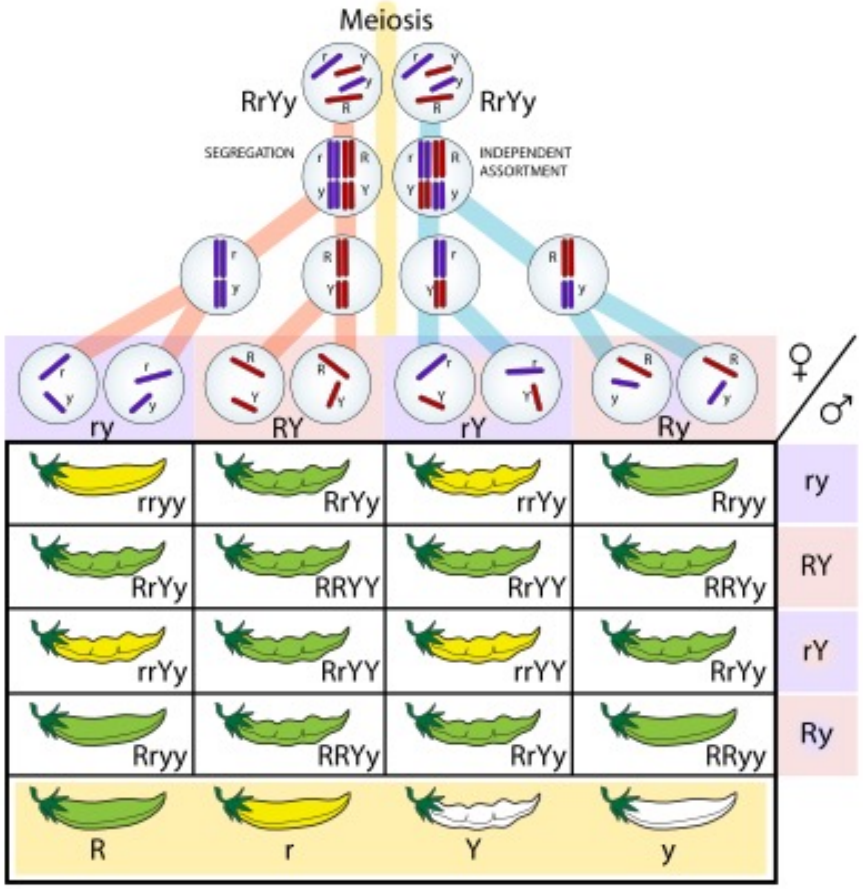
- 1859年达尔文提出进化论





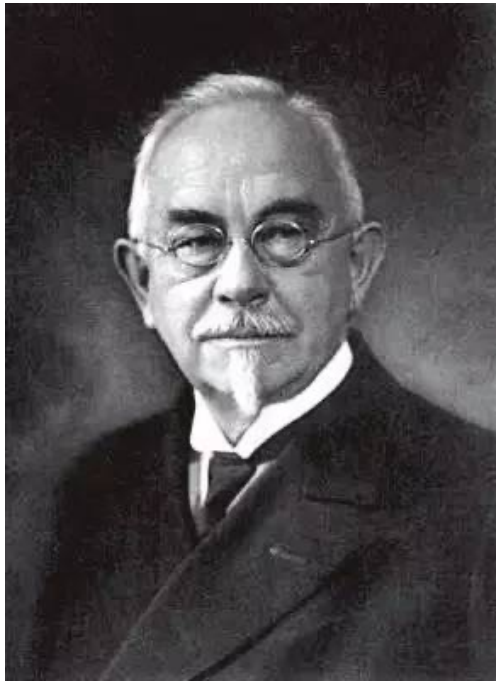
1865年经典遗传学

- 1865年孟德尔提出遗传定律
- 显性原则
- 分离定律
- 自由组合定律与独立分配率



1909年，基因概念提出

- 1909年约翰森首次提出基因一词



Abschnitt IV.

Die Hypothese der intracellularen Pangenesis.

Erstes Kapitel.

Pangene in Kern und Cytoplasma.

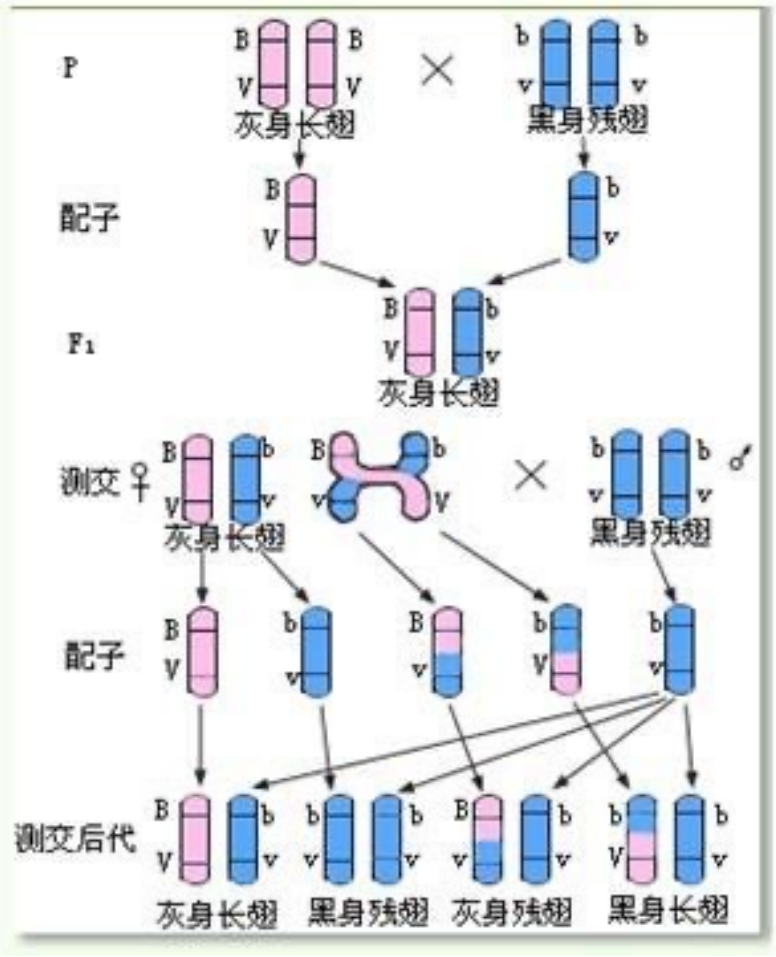
§ 1. Einleitung.

Die Schlussfolgerungen, zu denen uns im ersten Theile die kritische Betrachtung der bisherigen Theorien über die Erblichkeit, und im zweiten die Uebersicht über den jetzigen Stand der Zellenlehre geführt haben, wollen wir jetzt mit einander in Verbindung zu bringen suchen.

1910年，连锁互换定律

Thomas H. Morgan

- 托马斯·亨特·摩尔根
- 创立遗传的染色体学说
- 提出基因的“连锁-互换”定律



1928年，格里菲斯转化实验

弗雷德里克·格里菲斯（Frederick Griffith，1879年－1941年）英国细菌学家、传染病学家与肺炎病理学家。

1928年1月进行了一项“格里菲斯实验”，他透过肺炎链球菌（学名：Streptococcus pneumoniae）进行实验，发现细菌品系间可以互相转换，并改变其形态与功能，这项发现称为转型定律。

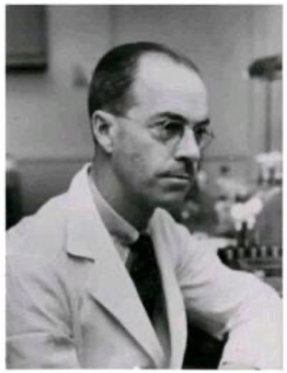
注入宿主的细菌类型	细菌特征	宿主感染后反应
II–R（粗糙型）	无荚膜	存活
III–S（平滑型）	有荚膜	死亡
死的III–S		存活
死的III–S + 活的II–R		死亡

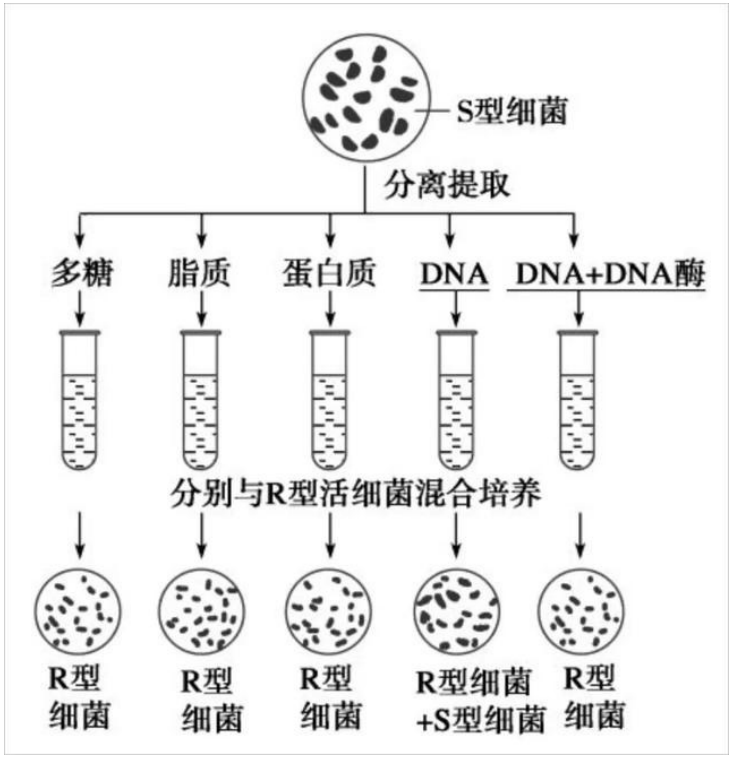
1930年，《自然选择的遗传理论》

- 罗纳德-费希尔，英国统计学家、演化生物学家与遗传学家。
- 为进化提供了数学理论基
- 群体遗传学和现代演化综论的奠基者。

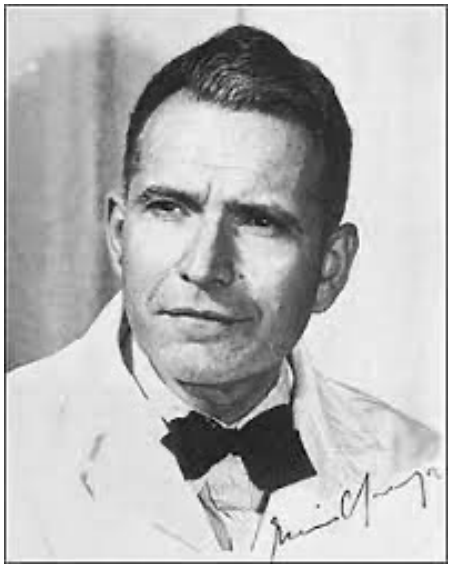
1944年发现DNA是生物的遗传物质

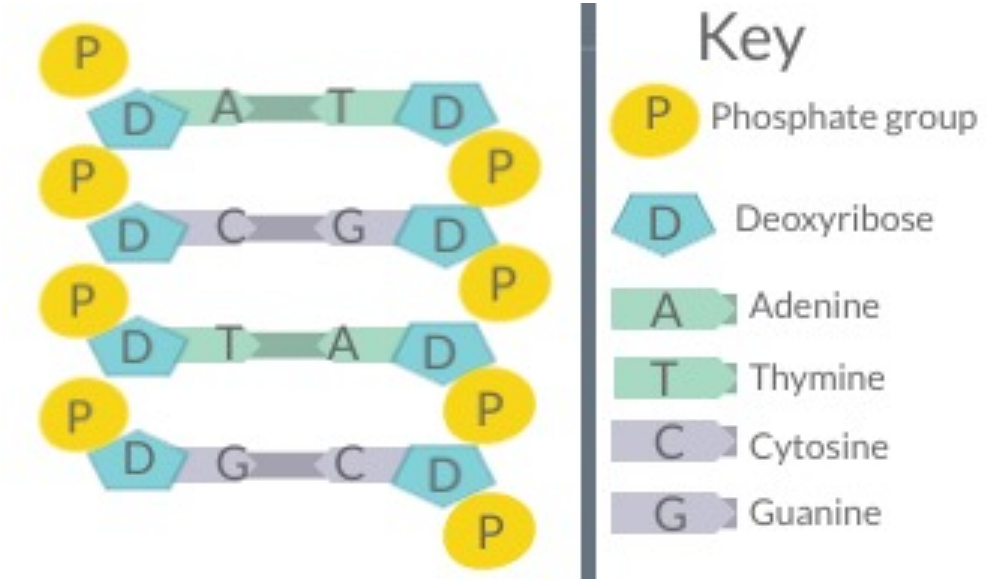
- 三位科学家艾佛里、麦克劳德、麦卡蒂的实验证实了遗传物质是DNA





1950年，查戈夫法则





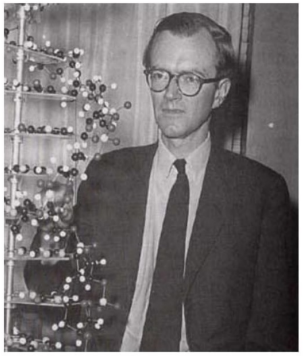
- 埃尔文-查戈夫，美国生物学家
- 1950年Chargaff提出了Chargeff规则
- 腺嘌呤与胸腺嘧啶的摩尔含量相等（A=T）

1953年DNA双螺旋模型的提出

■ 1953年沃森和克拉克建立DNA双螺旋结构模型

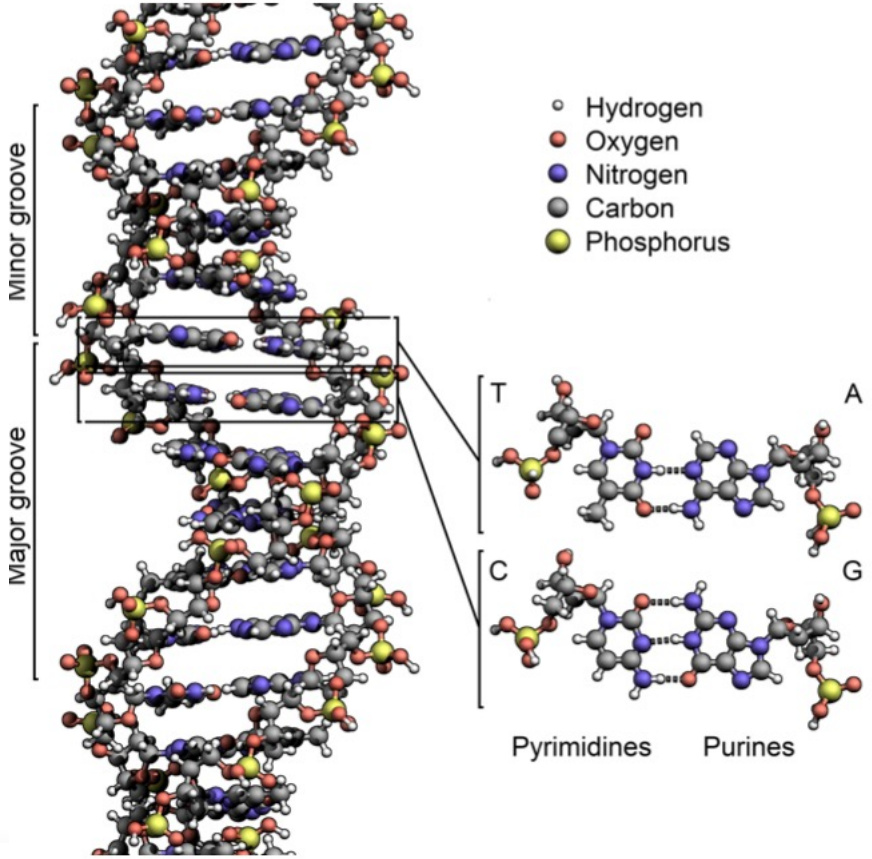
Francis Crick

James Watson



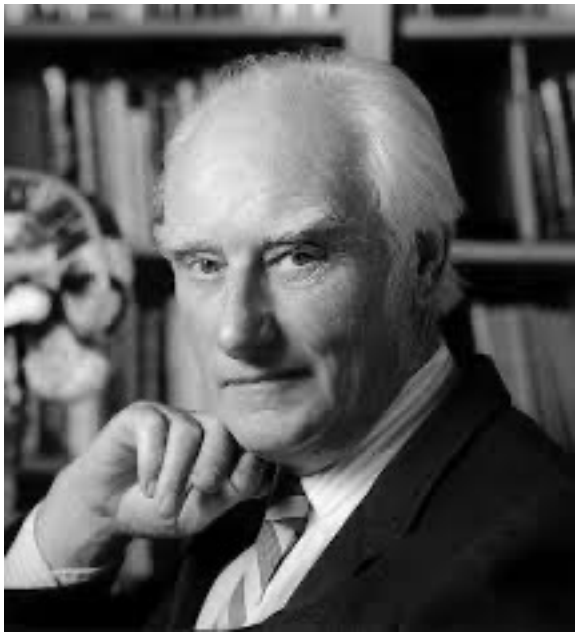
Maurice Wilkins

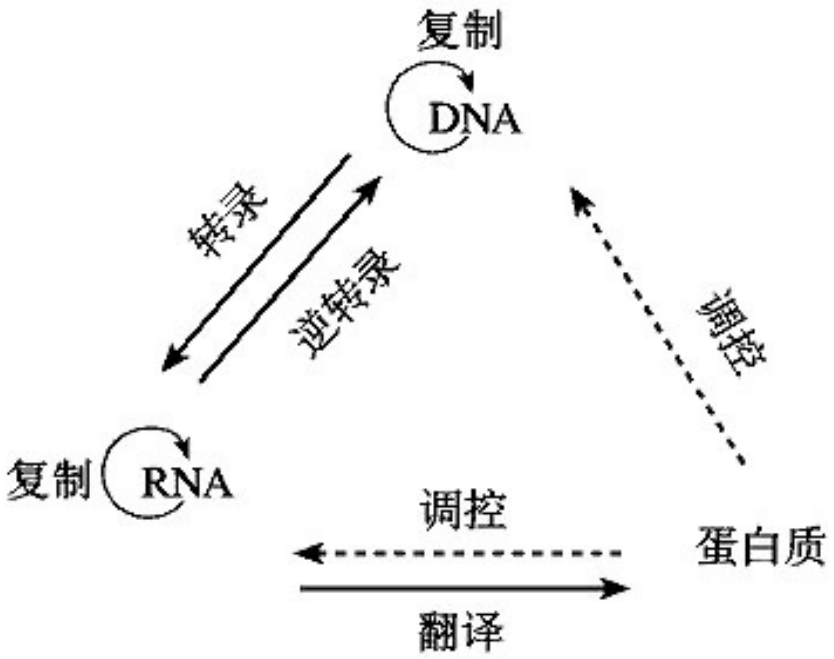
Rosalind Franklin



1957年生物中心法则

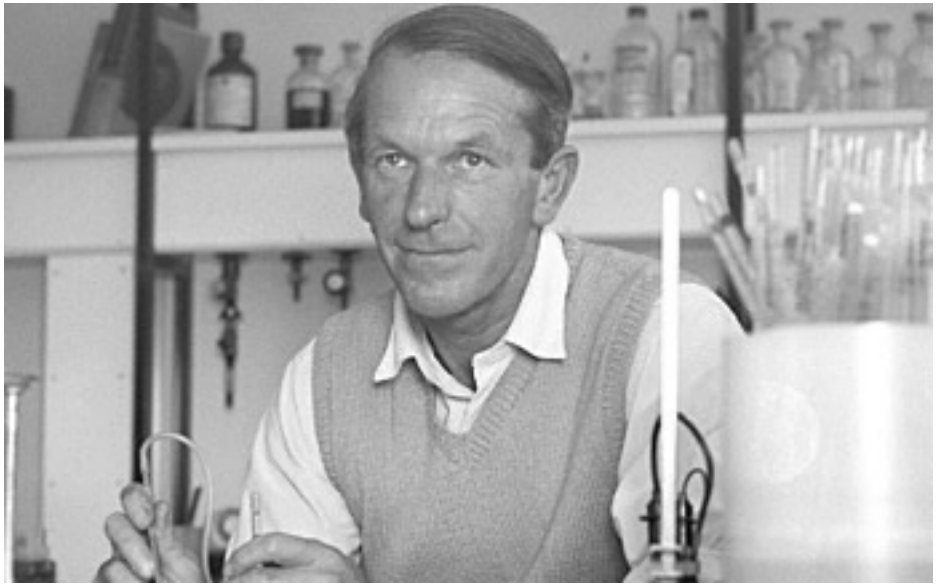
- 1957年克里克提出中心法则



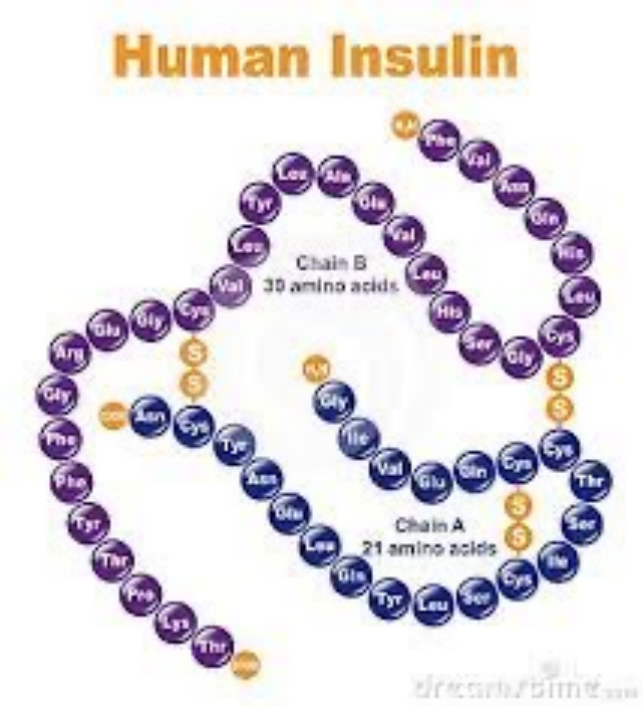


1951年-1952蛋白质氨基酸序列的测定

- 1958年Sanger测定了胰岛素的氨基酸序列获得诺奖



他利用自己发明的桑格试剂，也就是2,4-二硝基氟苯与胰岛素反应，使得2.4-二硝基苯基牢固的结合在胰岛素蛋白链N-端的氨基上，然后用盐酸将胰岛素彻底水解，进行纸层析



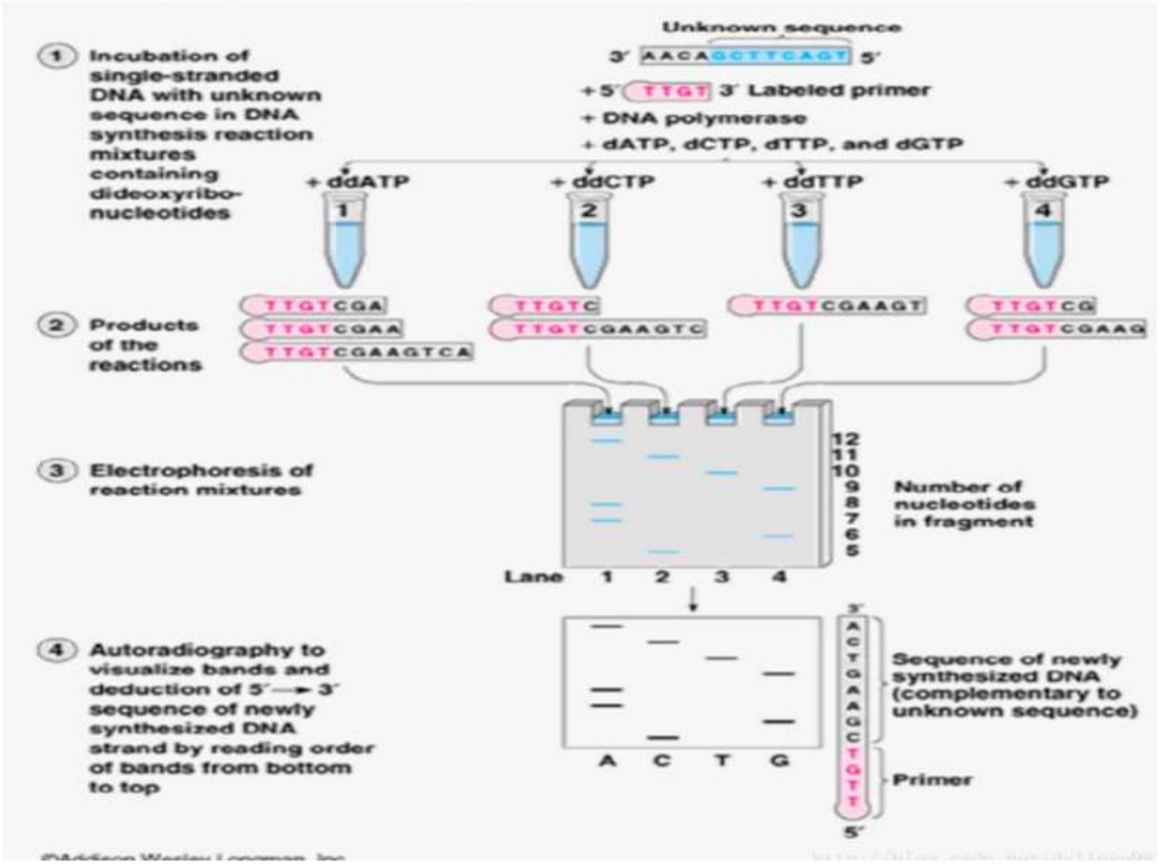
1977年Sanger测序技术

Dr. Fred Sanger

Frederick Sanger was awarded the prize in both 1958 and 1980. He is the fourth person in the world to have been awarded two Nobel Prizes and the only person to receive both in chemistry.

"dideoxy" sequencing technique (Sanger et al., 1977)

DNA双脱氧链终止法测序



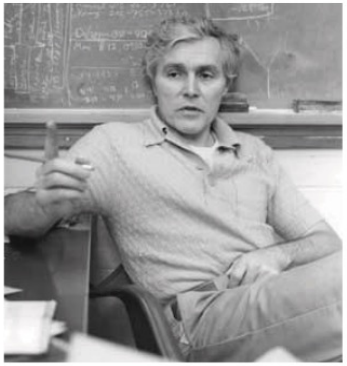
The 1970s and Earlier - Sequence Databases, Similarity Matrices and Molecular Evolution

Margaret Dayhoff

How do protein sequences evolve?

How should similarity between two proteins be scored to most accurately detect homology?

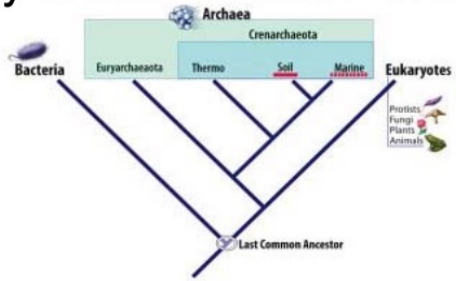
- First protein sequence databases / protein family classification
- PAM matrices for protein sequence comparisons (still used!)



Carl Woese

What can molecular sequences tell us about organismal evolution?

- Molecular classification of life
- Molecular clocks
- Use of ribosomal RNA to infer phylogeny
- Discovery of third 'domain' of life - Archaea



Russ Doolittle

The 1980s: Sequence Alignment/Search

Which specific residues/positions in a pair of proteins are homologous?

- Smith-Waterman alignment algorithm

Photographs of scientists removed due to copyright restrictions.

What RNA secondary structure has minimum folding free energy?

- Nussinov algorithm
- Zuker algorithm

How to rapidly and reliably find homologs to a query sequence in a sequence database?

- FastA and BLAST algorithms and associated statistics

The ‘90s: HMMs, Ab Initio Protein Structure Prediction, Genomics, Comparative Genomics

How to identify domains in a protein?

How to identify genes in a genome?

Hidden Markov Models as a framework for such problems

How to study gene expression globally, infer gene function from expression?

- Microarrays and clustering

How to predict protein function by comparing genomes?

- gene fusions, phylogenetic profiling, etc.

How to predict protein structure directly from primary sequence?

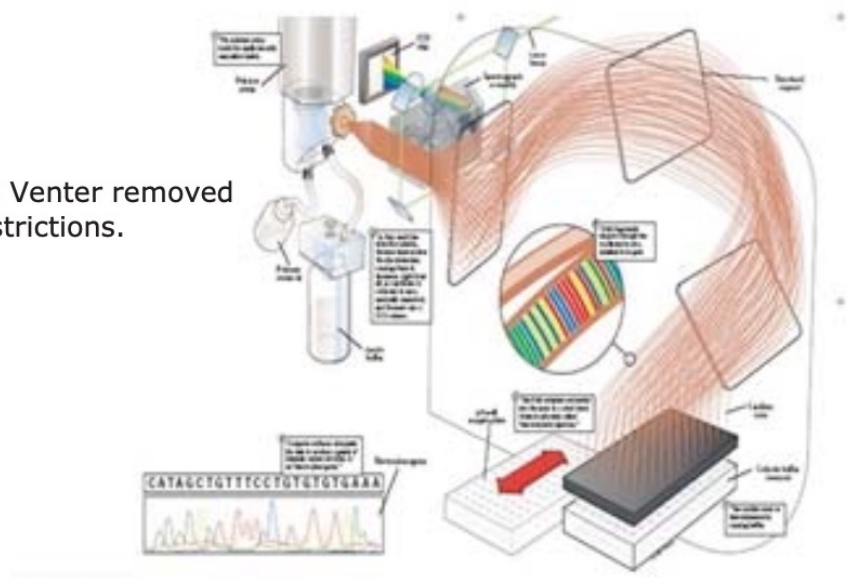
- Rosetta algorithm

The 2000s Part 1:

The human genome is sequenced, assembled, annotated

genomics becomes fashionable

Photograph of Craig Venter removed due to copyright restrictions.



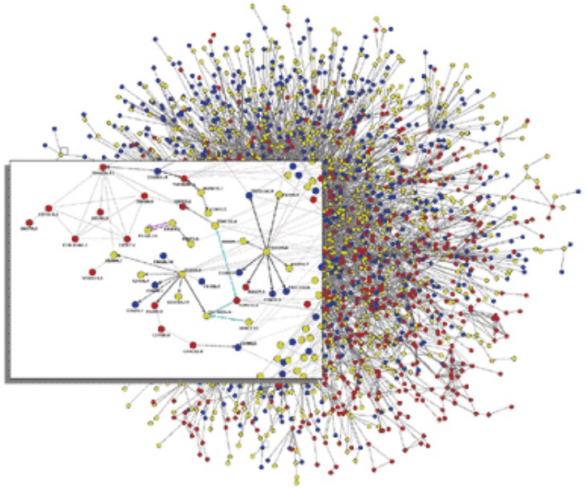
The 2000s Part 2:
Biological Experiments Become High-Throughput, Computational Biology Becomes more Biological

Massively parallel data collection - transcriptomics, proteomics, interactomics, metagenomics

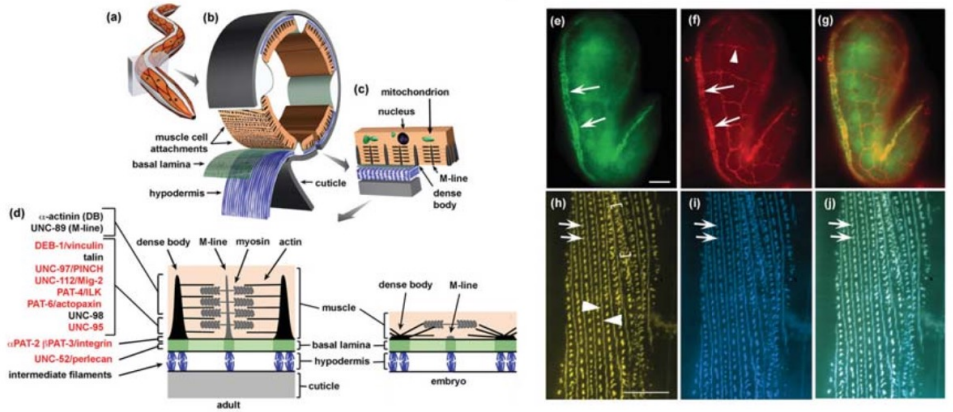
Using sequence and array data to address fundamental questions about transcription, splicing, microRNAs, translation, epigenetics, protein structure/function, development, evolution, disease, etc.

Integrated computational/experimental approaches

Rise of bioimage informatics

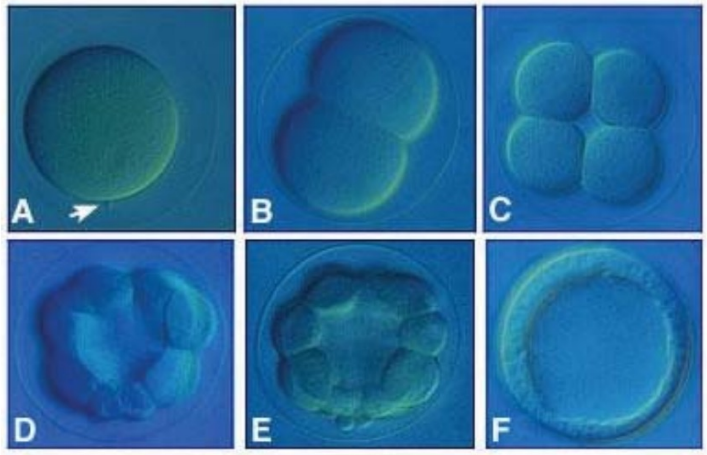


Courtesy of Marc Vidal. Used with permission.



Courtesy of Donald G. Moerman and Benjamin D. Williams. License: CC-BY.
Source: Moerman, D. G. and Williams, B. D. "Sarcomere Assembly in C. Elegans Muscle" (January 16, 2006), WormBook, ed. The C. elegans Research Community, WormBook.

Photograph of Eric Davidson removed due to copyright restrictions.

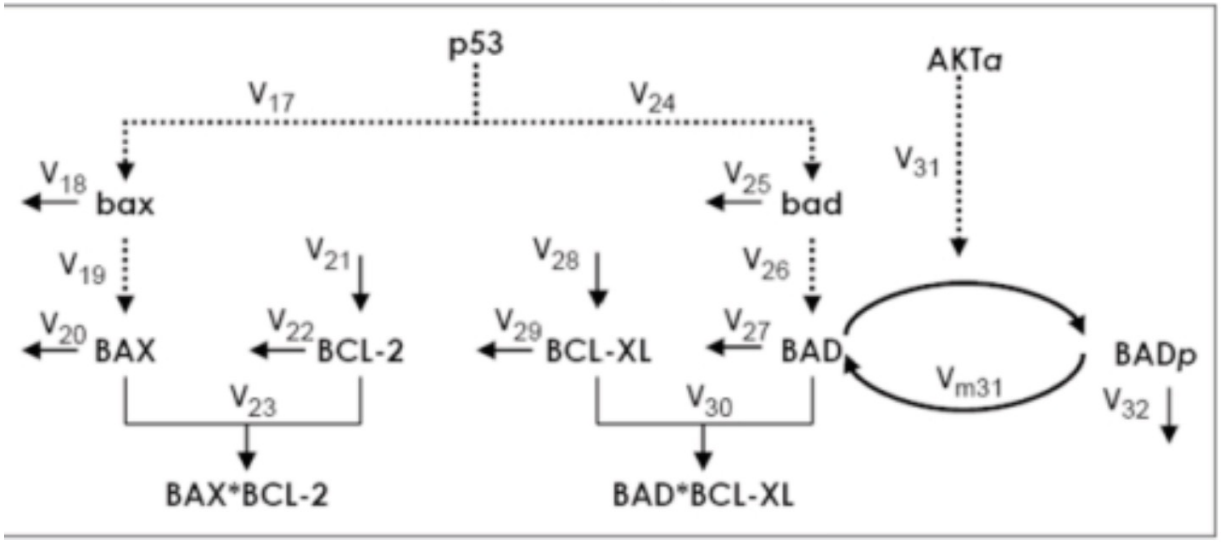


Courtesy of Charles Ettensohn. Used with permission.

The 2000s Part 3

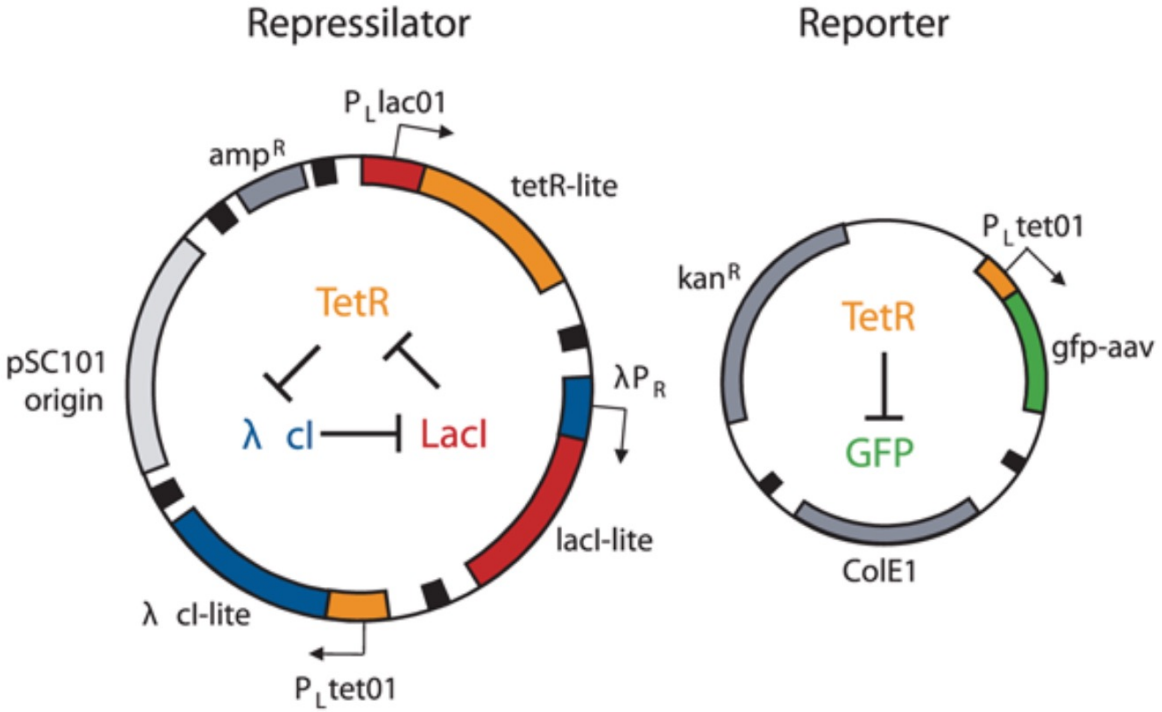
Systems Biology

Models of gene and protein networks in development, disease, etc.



The 2000s Part 4:
Synthetic Biology & Biological Engineering

Design of regulatory networks using biological components



Source: Elowitz, Michael B., and Stanislas Leibler. "A Synthetic Oscillatory Network of Transcriptional Regulators." *Nature* 403, no. 6767 (2000): 33S-8.

$$\frac{d[A]}{dt} = \frac{V_a}{1+\frac{[B]}{K_{iB}}} - k_a[A]$$

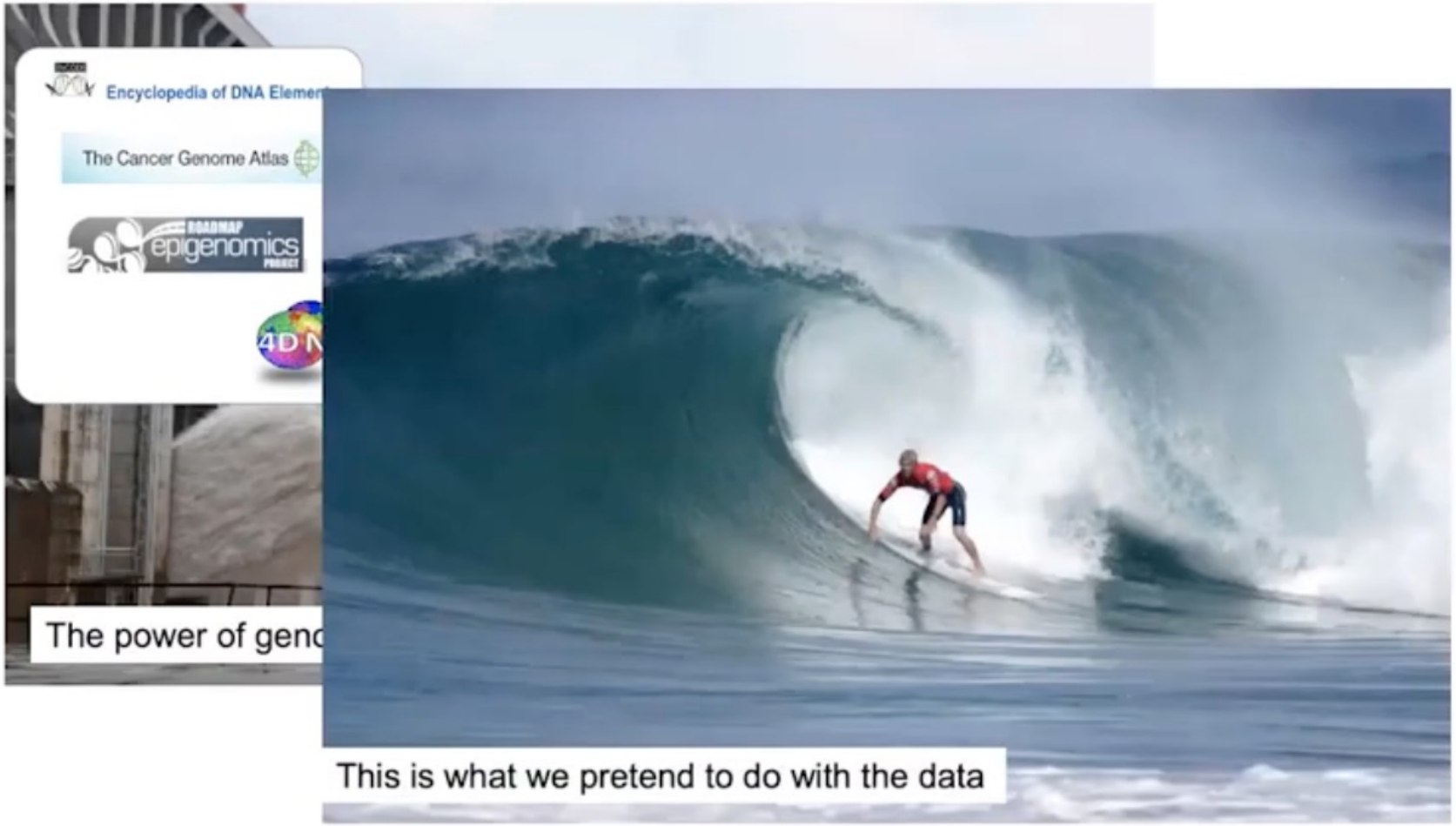
$$\frac{d[B]}{dt} = \frac{V_b}{1+\frac{[C]}{K_{iC}}} - k_b[B]$$

$$\frac{d[C]}{dt} = \frac{V_c}{1+\frac{[E]}{K_{iE}}} - k_c[C]$$

$$\frac{d[D]}{dt} = \frac{V_d}{1+\frac{K_{aB}}{[B]}} - k_d[D]$$

$$\frac{d[E]}{dt} = \frac{V_e}{\left(1+\frac{K_{aD}}{[D]}\right)\left(1+\frac{[C]}{K_{iC'}}\right)} - k_e[E]$$

In the post-genome era

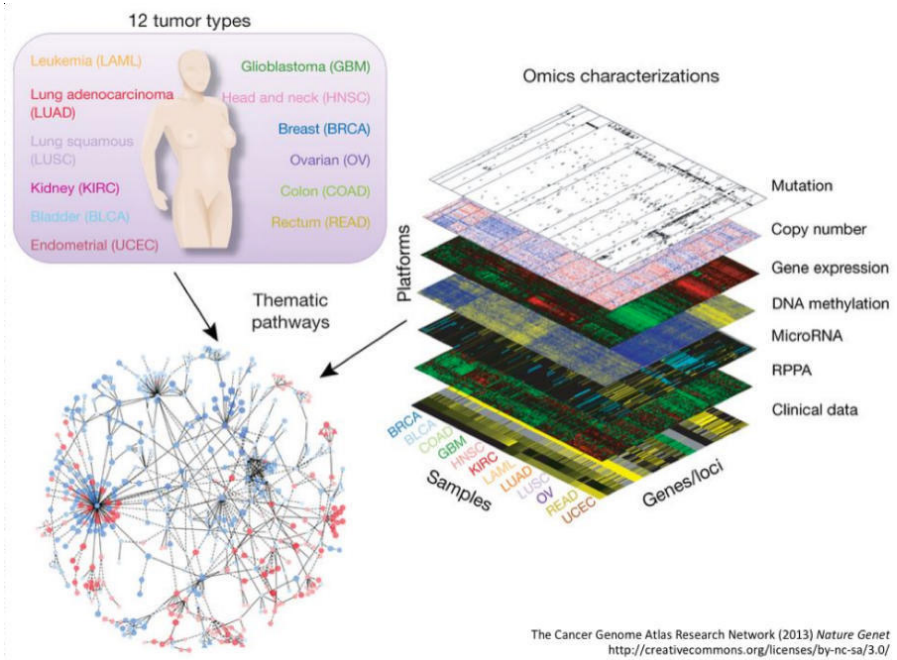


Slice from the workshop of WashU Epigenome Browser

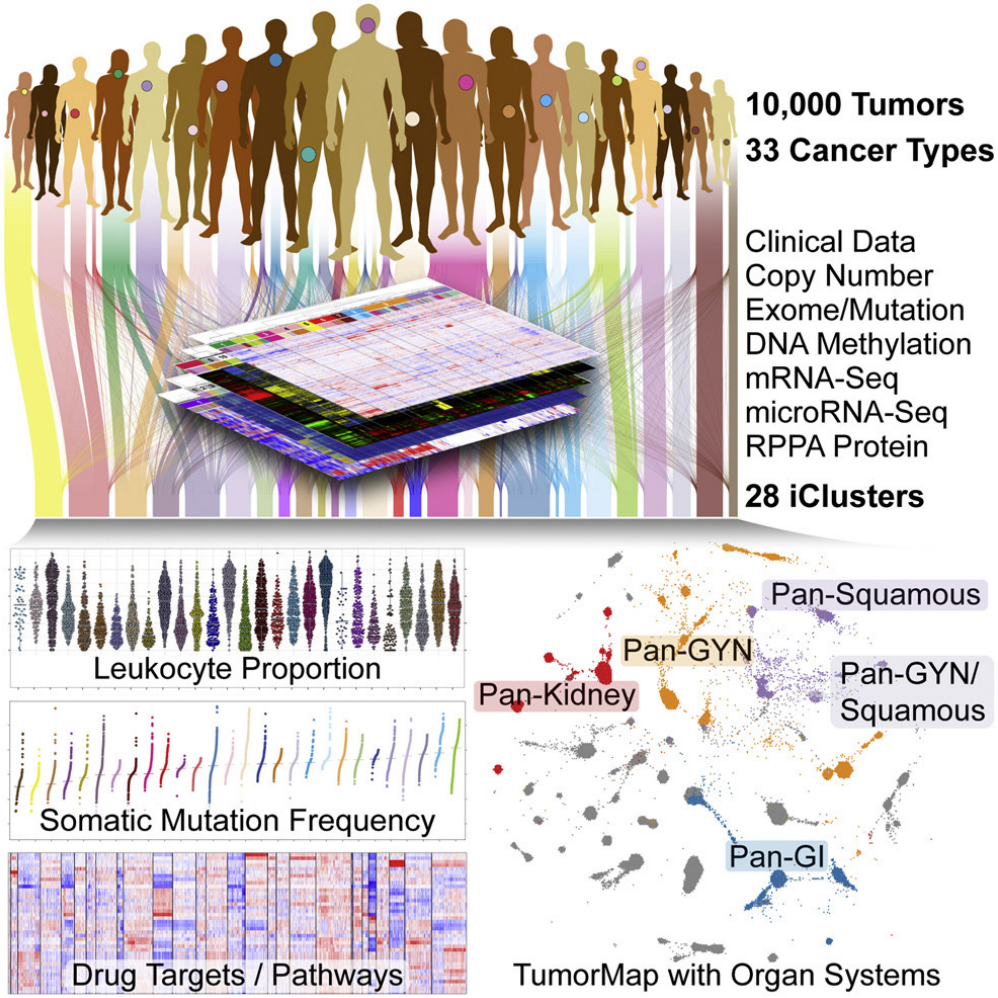
2005年：癌症基因组图谱（TCGA）启动

2008年第一个数据门户和数据分析工具

2013年成立泛癌症分析工作组



2018年：癌症基因组图谱（TCGA）进展

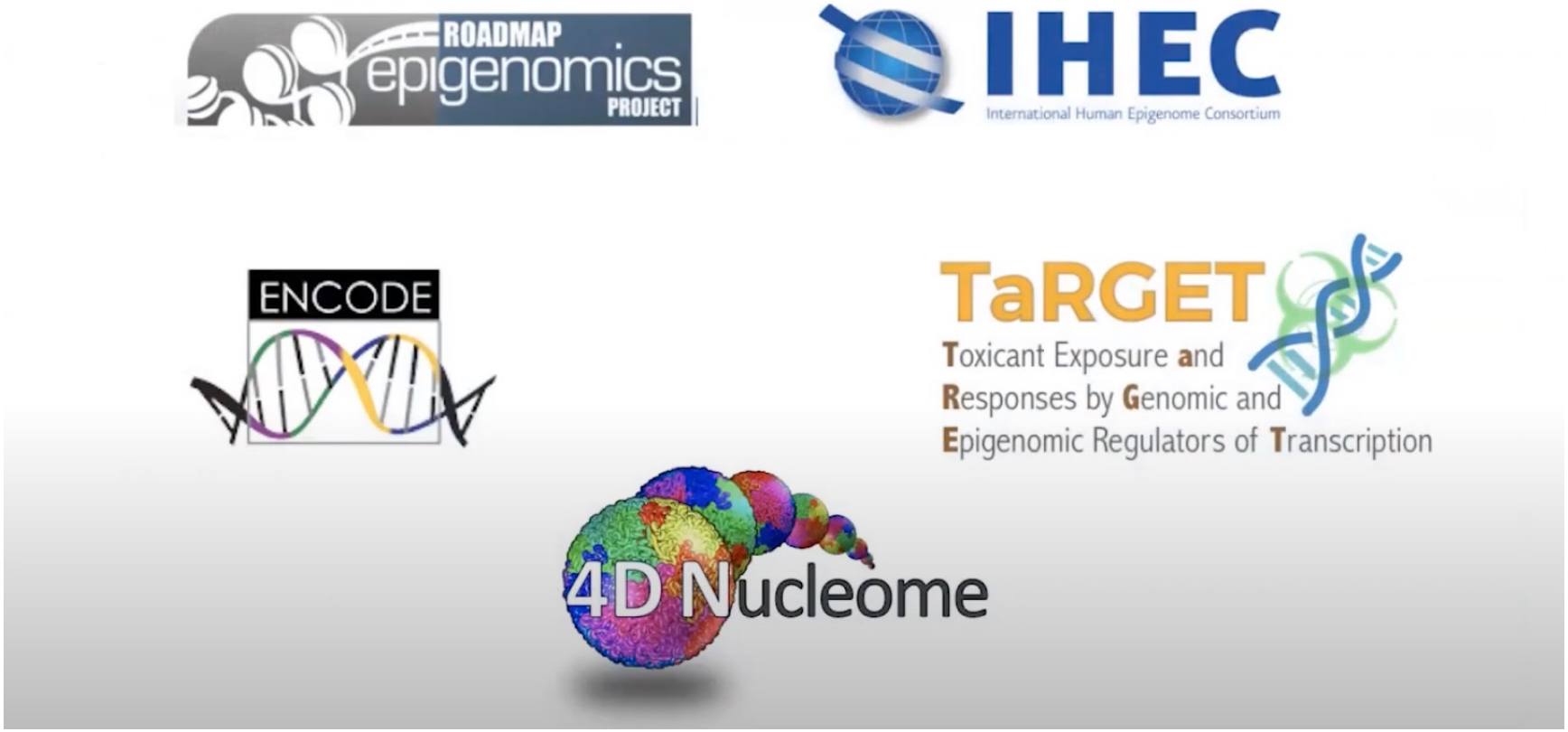


Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer

- Cell-of-origin influences, but does not fully determine, tumor classification
细胞起源影响肿瘤分类，但并不完全决定肿瘤分类
- Immune features and copy-number aberrations define the most mixed tumor groups
免疫特征和拷贝数异常定义了最混杂的肿瘤组

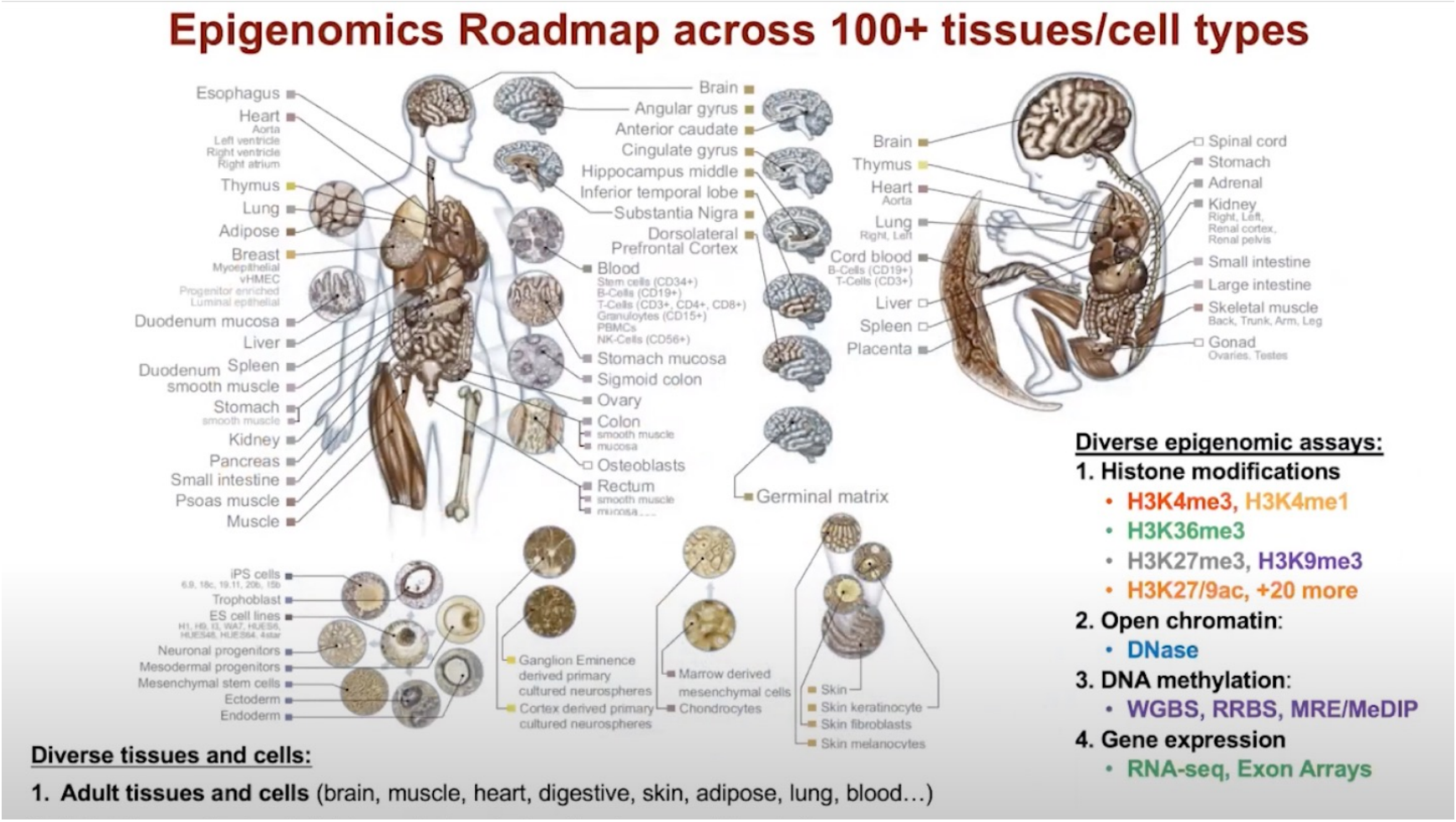
DOI: 10.1016/j.cell.2018.03.022

From 1D to 4D Genome

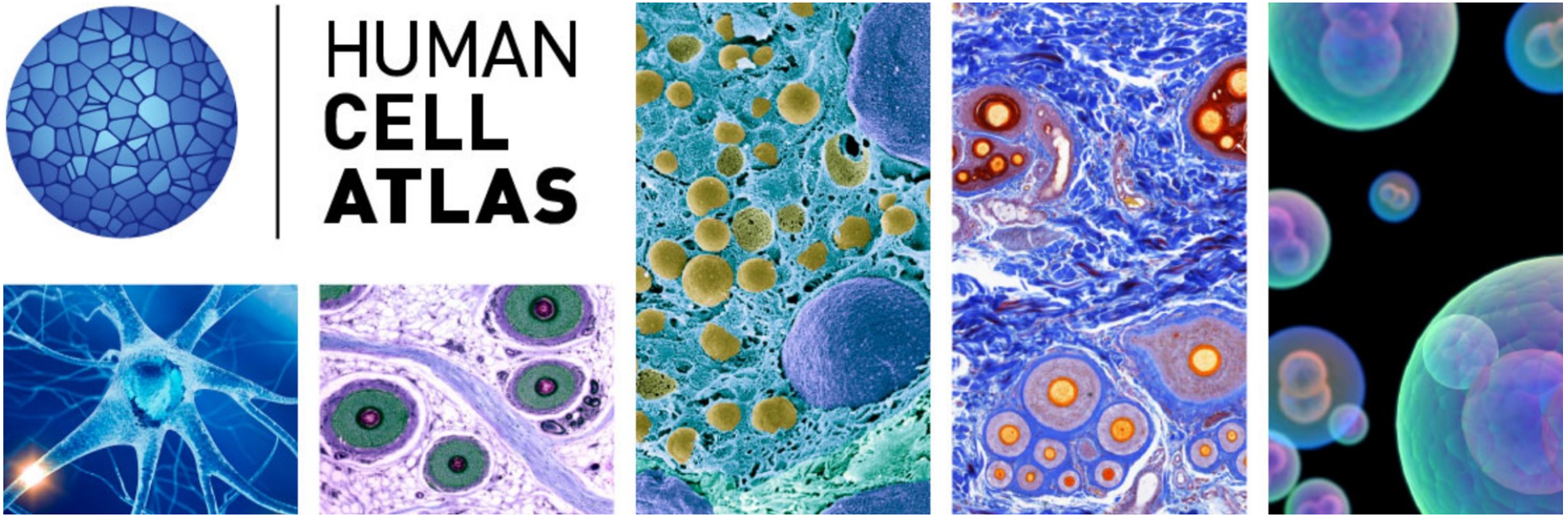


Slice from the workshop of WashU Epigenome Browser

Epigenomics Roadmap

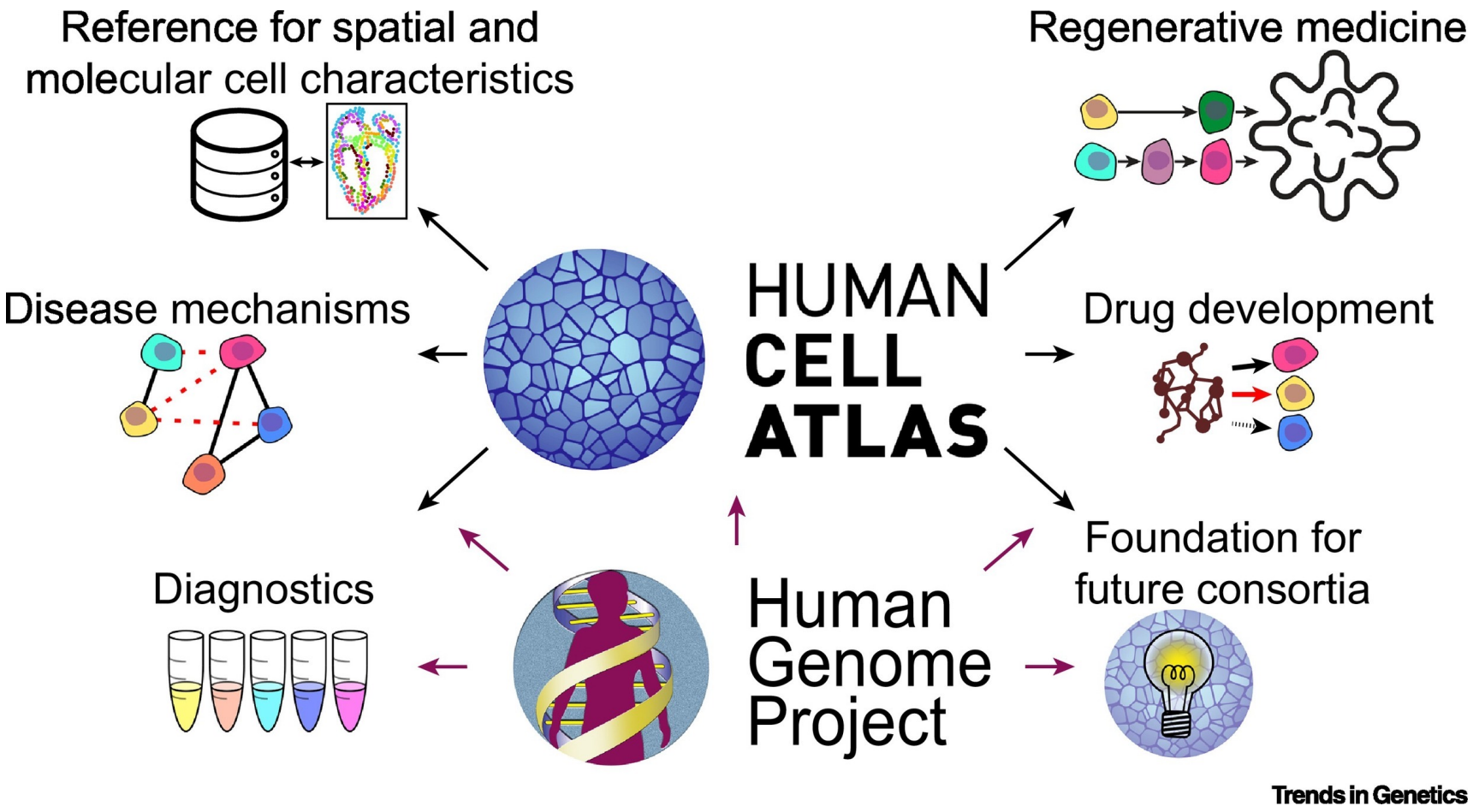


Human Cell Atlas



https://www.humancellatlas.org/

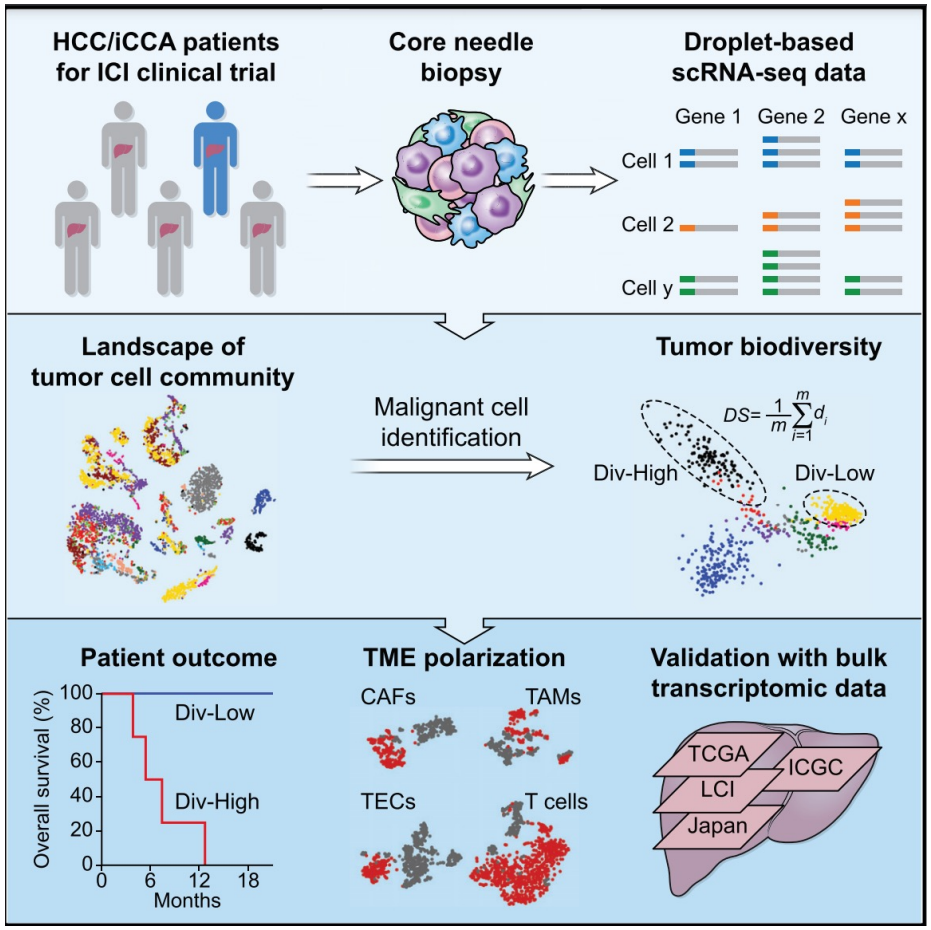
Human Cell Atlas

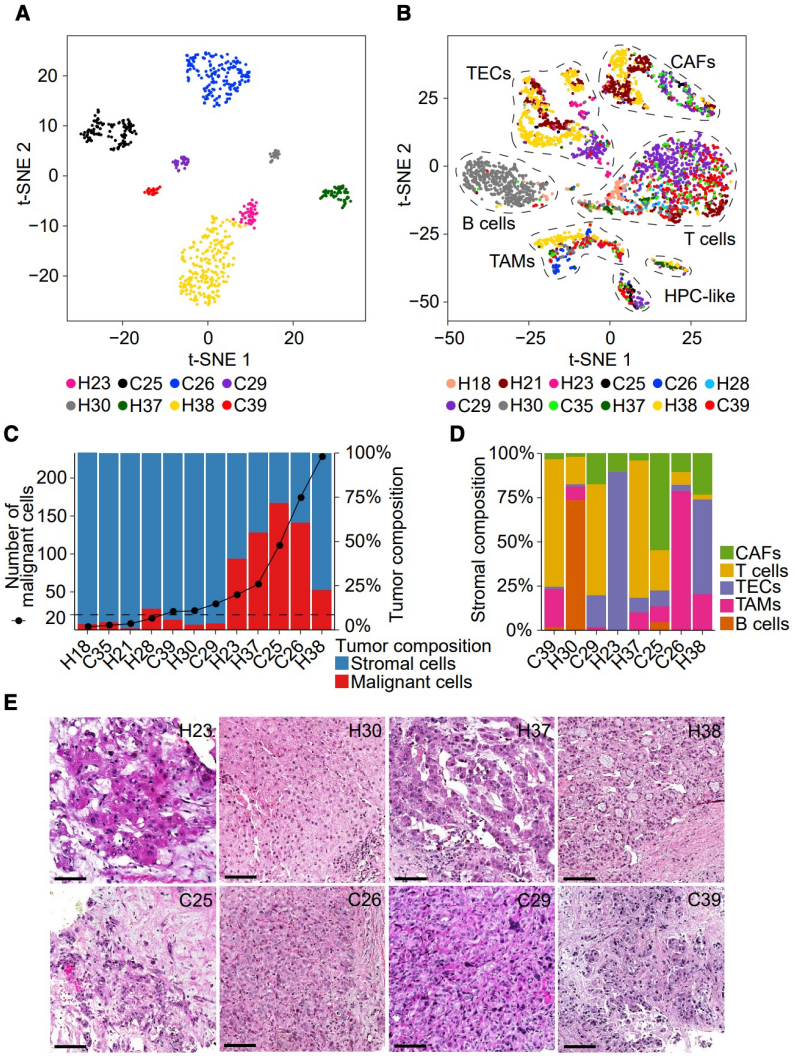


https://www.cell.com/trends/genetics/fulltext/S0168-9525%2821%2900076-7

计算分子生物学

■ 疾病–肿瘤异质性的揭示

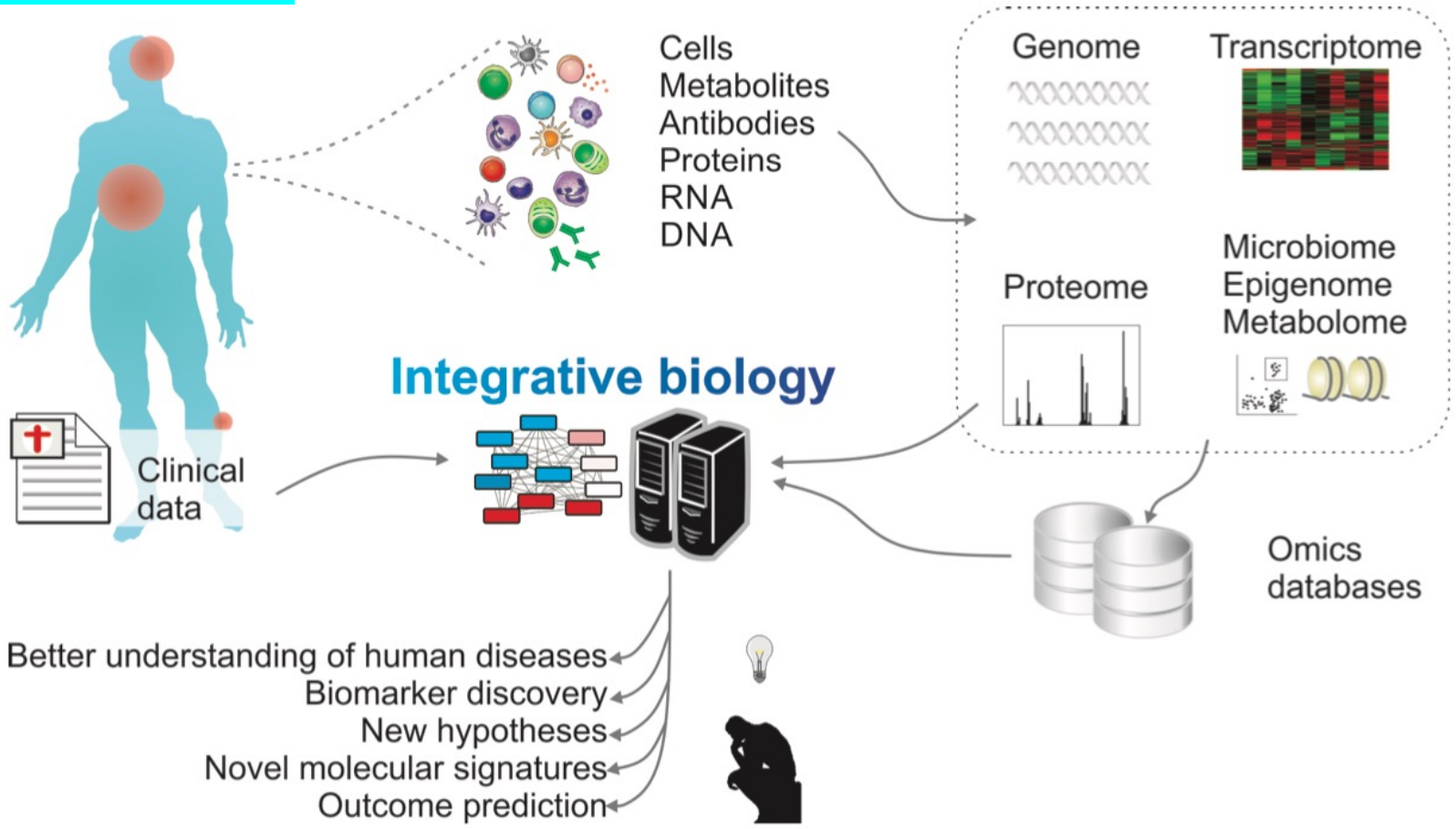




https://www.cell.com/cancer-cell/fulltext/S1535-6108(19)30375-7

计算分子生物学的医学应用

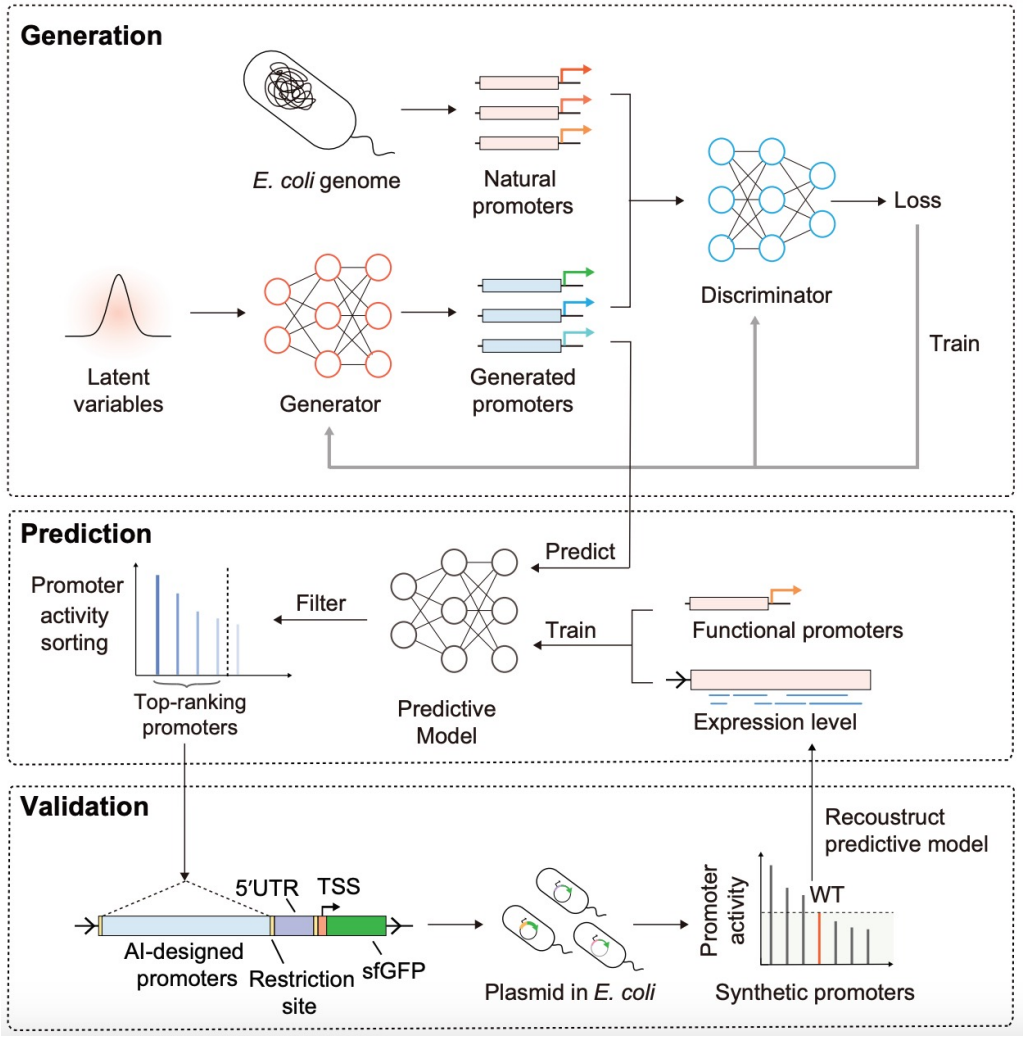
整合生物医学



https://www.ncbi.nlm.nih.gov/books/NBK550336/figure/Ch2-f0001/

计算分子生物学与生物工程

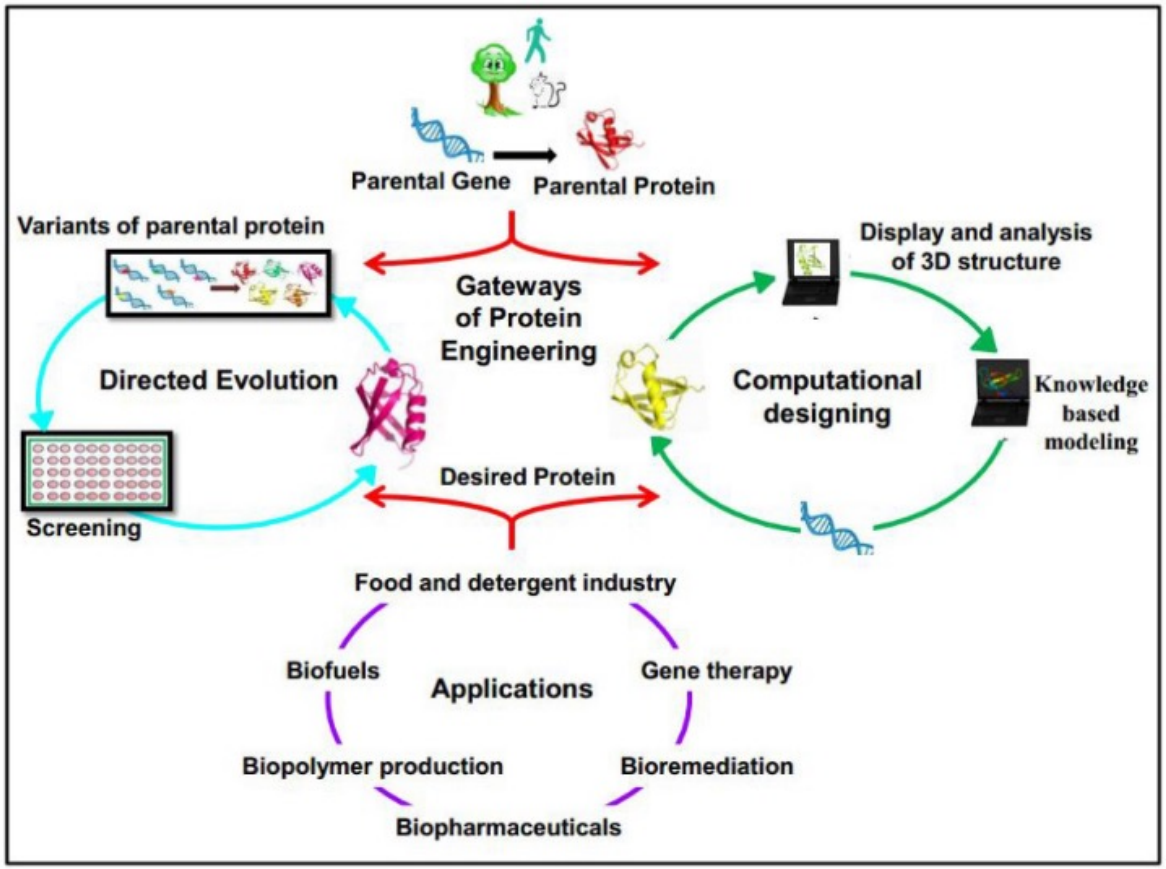
- **合成-基因调控**
  - 帮助研究人员设计和构建基因调控元件，以便在特定的时间点和条件下控制基因表达。
  - 这些基因调控元件可以用于调节代谢途径、合成代谢产物等。



https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkaa325/5837049

计算分子生物学与生物工程

- **合成–蛋白质工程**
  - 帮助研究人员设计和优化蛋白质序列，使其具有所需的性质。
  - 这些蛋白质可以用于制药、生物燃料和其他生物工程应用中。



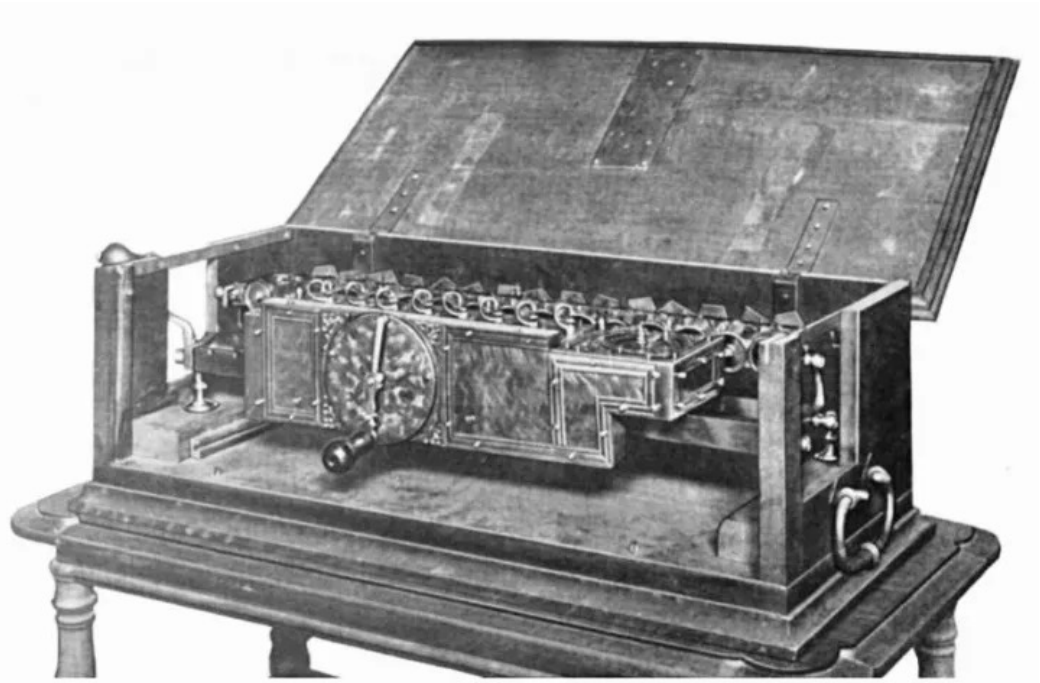
https://www.creativebiomart.net/protein-engineering-services.htm

计算科学发展简史

- **计算科学**
  - 帕斯卡发明了人类有史以来第一台机械计算机
  - 莱布尼茨
    - 发明二进制
    - 首次提出“计算机”的概念
  - 图灵提出了“图灵机”设想，是现代计算机的原型
  - 世界上第一台电子计算机“埃尼阿克”诞生
  - 比尔盖茨创立微软公司
  - 乔布斯创立苹果公司

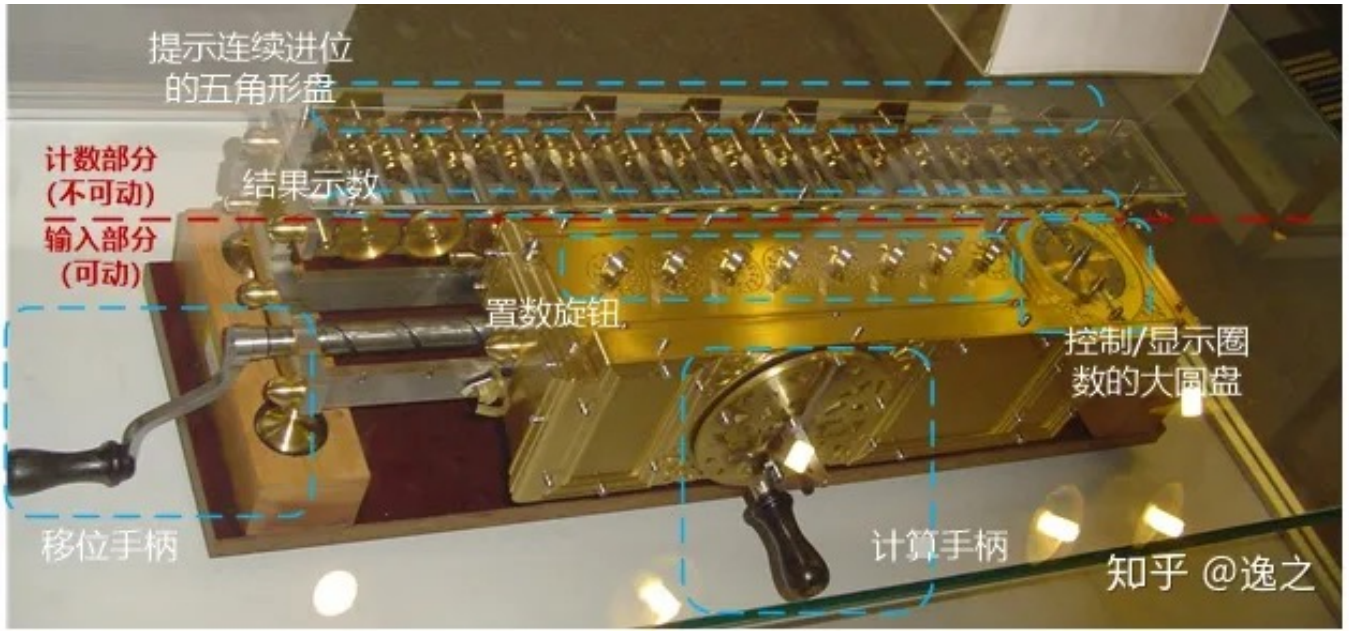
1642年人类第一台机械计算机

- 1642年布莱士.帕斯卡发明了人类的第一台机械计算机



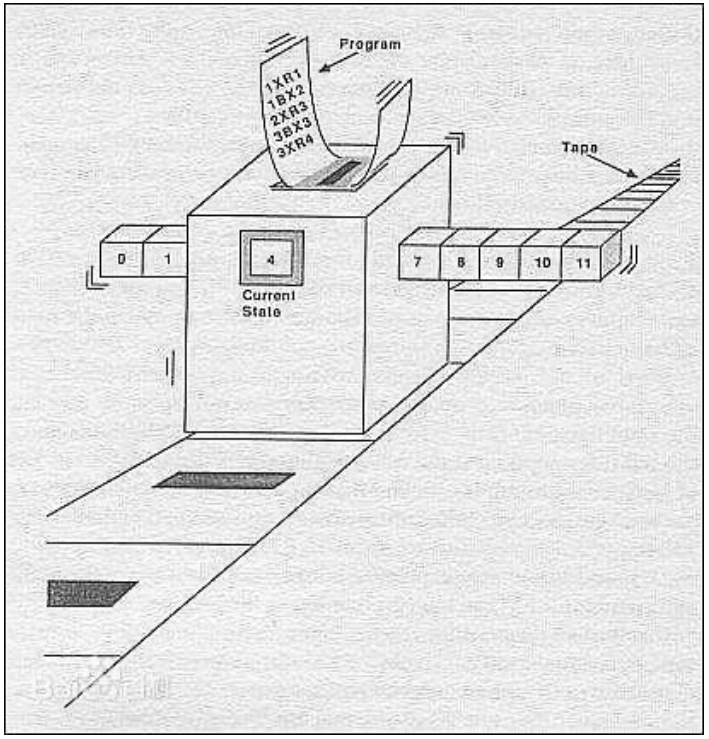
1673年人类第一台四则运算机械计算机

- 莱布尼茨（1646-1716）



1936年图灵计算机

- 图灵（1912-1954）提出了“图灵机”设想，是现代计算机的原型

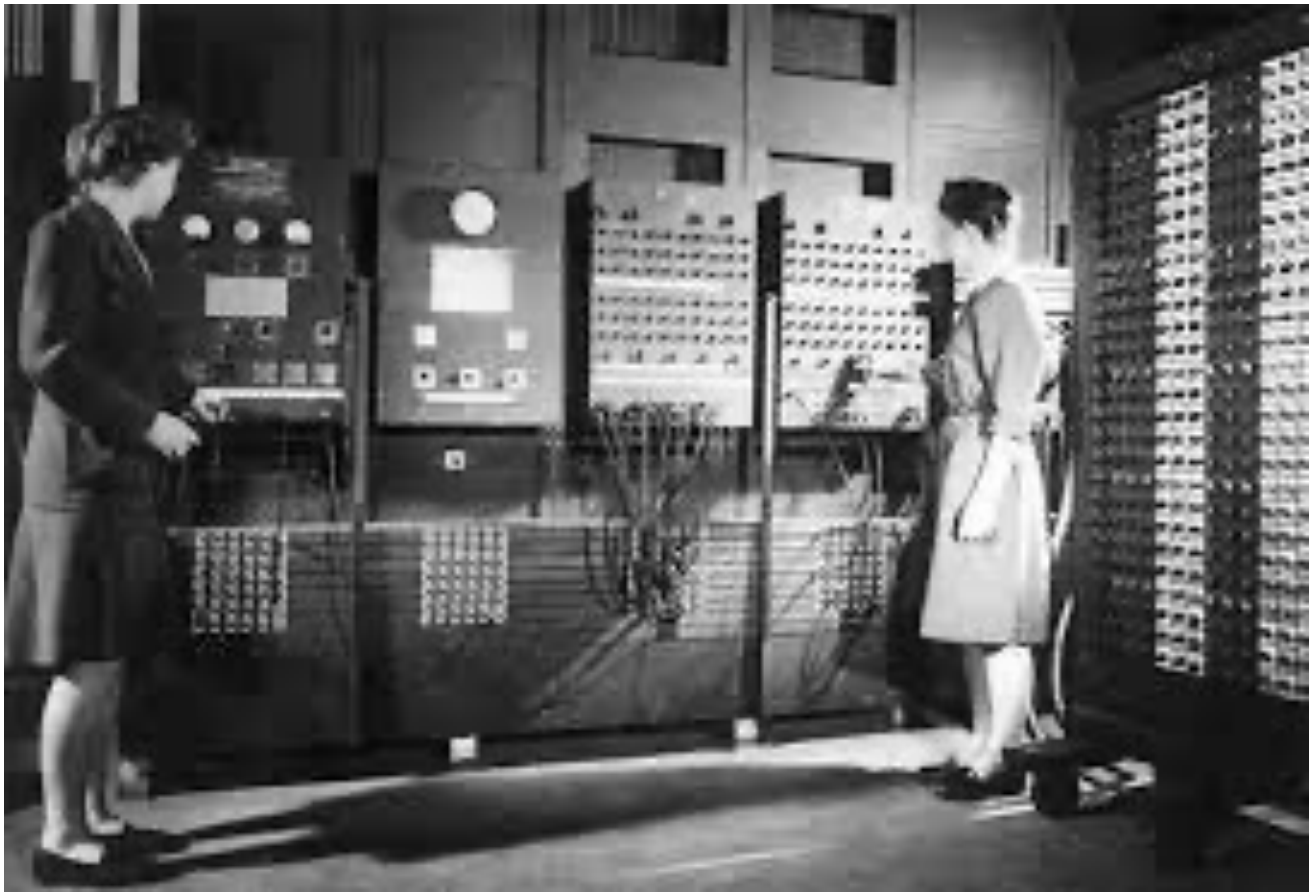


1936年，《论可计算数及其在判定问题上的应用》
1948年，《智能机器》
1950年，《计算机器与智能》
1952年，《形态发生的化学基础》

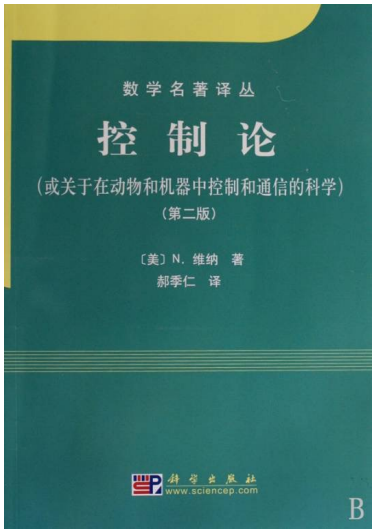
https://www.bilibili.com/video/BV1br4y1N762?t=12.7

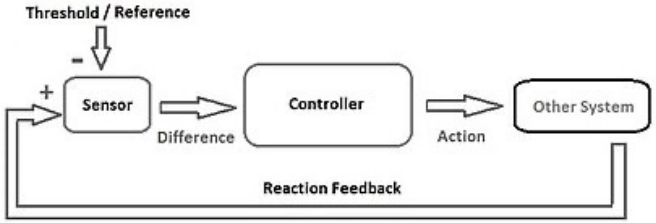
1946年人类第一台计算机

- 世界上第一台电子计算机“埃尼阿克”诞生



1948年诺伯特-维纳建立控制论

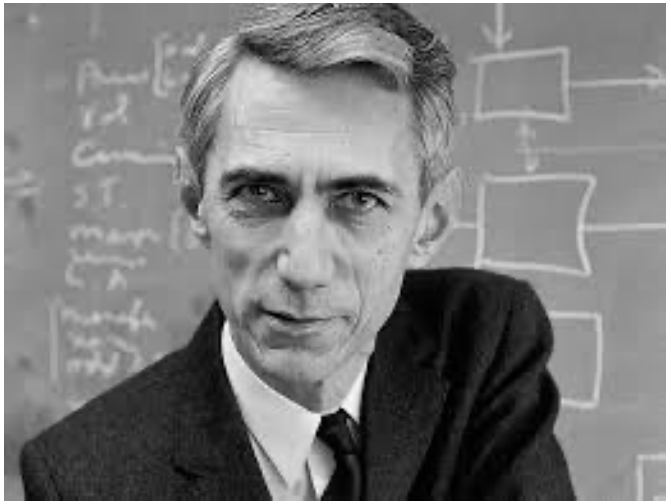




A Cybernetic Loop

- 研究机器和组织的内部或彼此之间的控制和通信的科学。
- 英文cybernetics。

1948年克劳德-香浓建立信息论



$$H(U) = E[-\log p_i] = -\sum_{i=1}^{n} p_i \log p_i$$

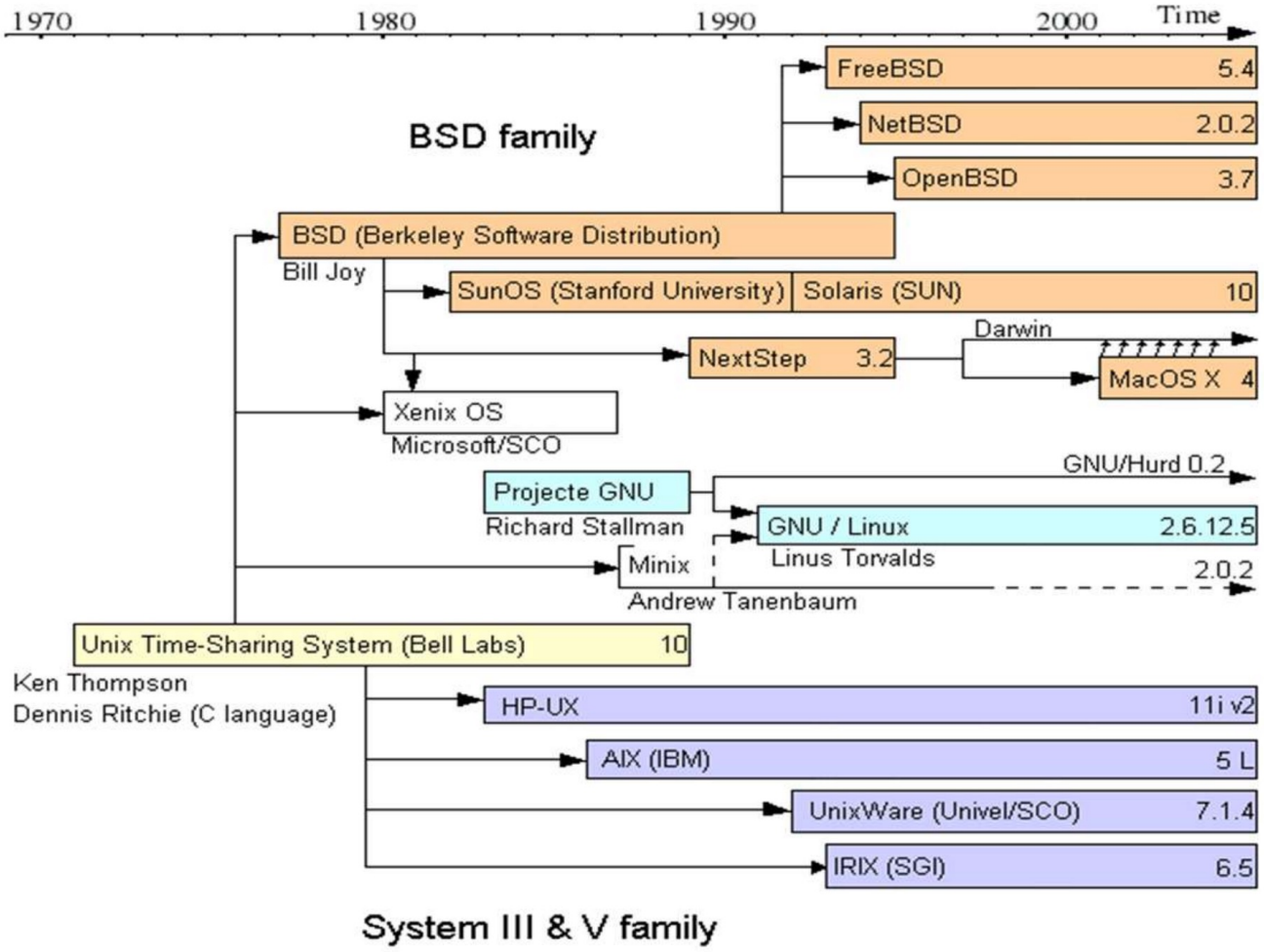
- 狭义信息论是应用统计方法研究通讯系统中信息传递和信息处理的共同规律的科学，即研究概率性语法信息的科学；
- 广义信息论是应用数学和其他有关科学方法研究一切现实系统中信息传递和处理、信息识别和利用的共同规律的科学，即研究语法信息、语义信息和语用信息的科学

1969年UNIX操作系统诞生

Ken Thompson and Dennis Ritchie (standing) at a PDP-11 in 1972.

Slide from Kong

类UNIX操作系统的演变



Slide from Kong

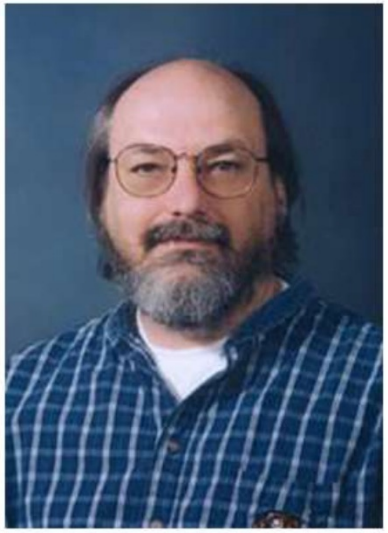
1990年Linux的诞生

- 1990年，芬兰赫尔辛基大学的学生Linus Torvalds用汇编语言写了一个在80386保护模式下处理多任务切换的程序。
- Linus Torvalds发布的运行在386机器上的内核程序。从版本0.0.1开始具有操作系统内核的雏形。
- Linux采用GNU的公共许可协议，可以免费使用、自由传播。
- Linux操作系统不仅包括Linux的内核，而且还包括shell、文本编辑器、高级语言编译器、办公等应用软件。具有UNIX操作系统的全部功能，包括多任务、多用户的能力，符合POSIX标准。

Slide from Kong

Linux和Unix的关系

Ken Thompson



Dennis Ritchie

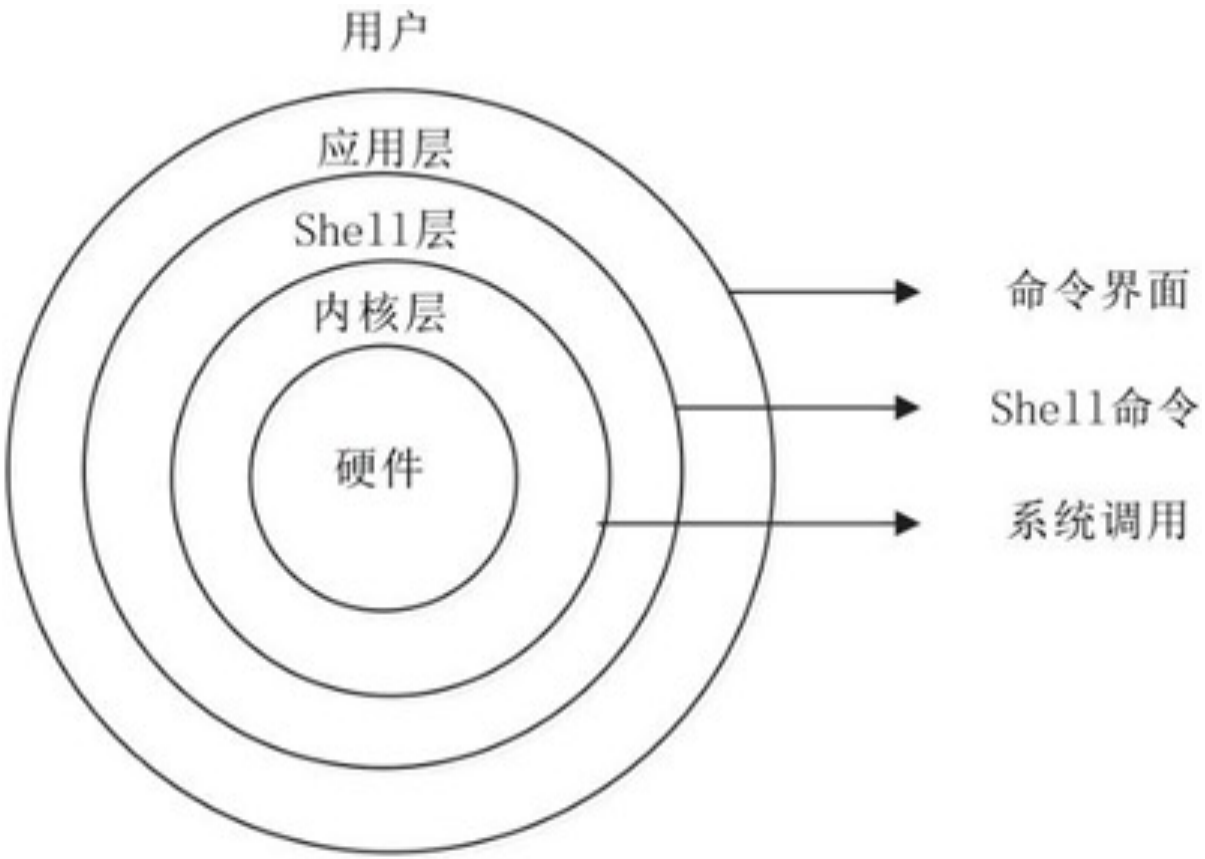
Richard Stallman

Linus Torvalds

1983年 图灵奖

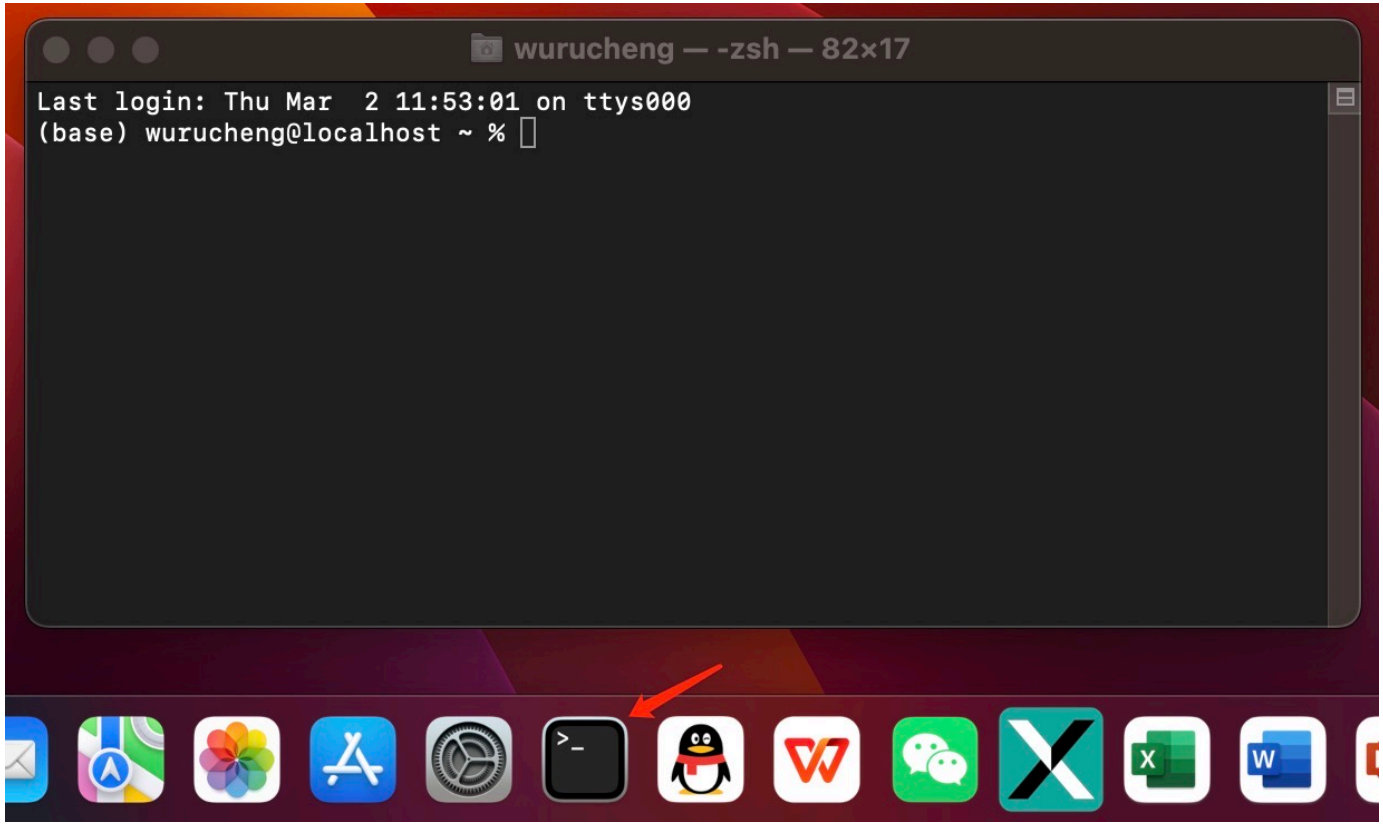
Slide from Kong

UNIX/Linux 系统结构层次概要



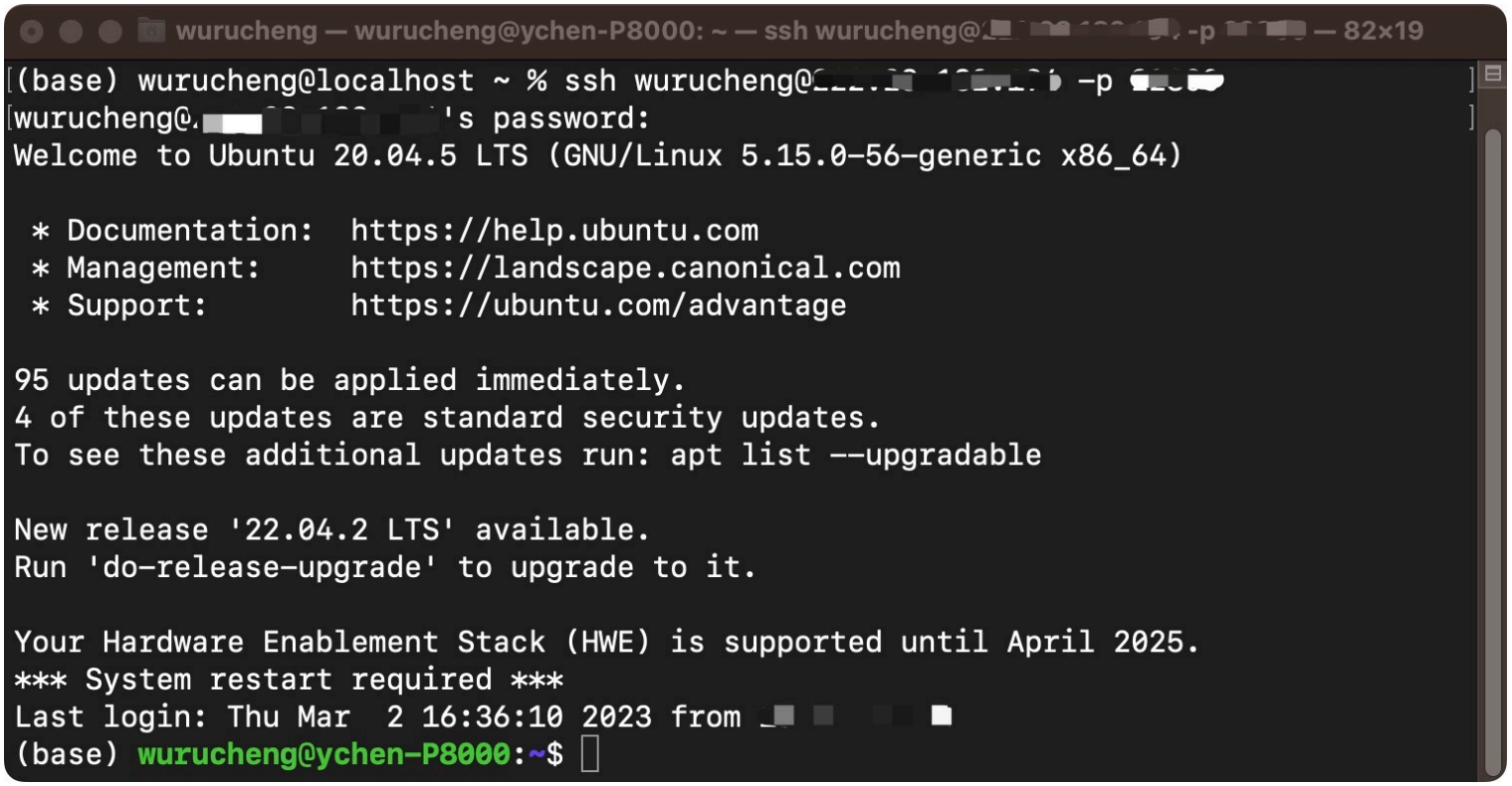
Linux练习环境

- **苹果电脑终端**
  - **苹果电脑使用的是Mac OS操作系统，也属于类Unix操作系统。**
  - **终端使用的是zsh，完全兼容Linux操作系统的bash。**



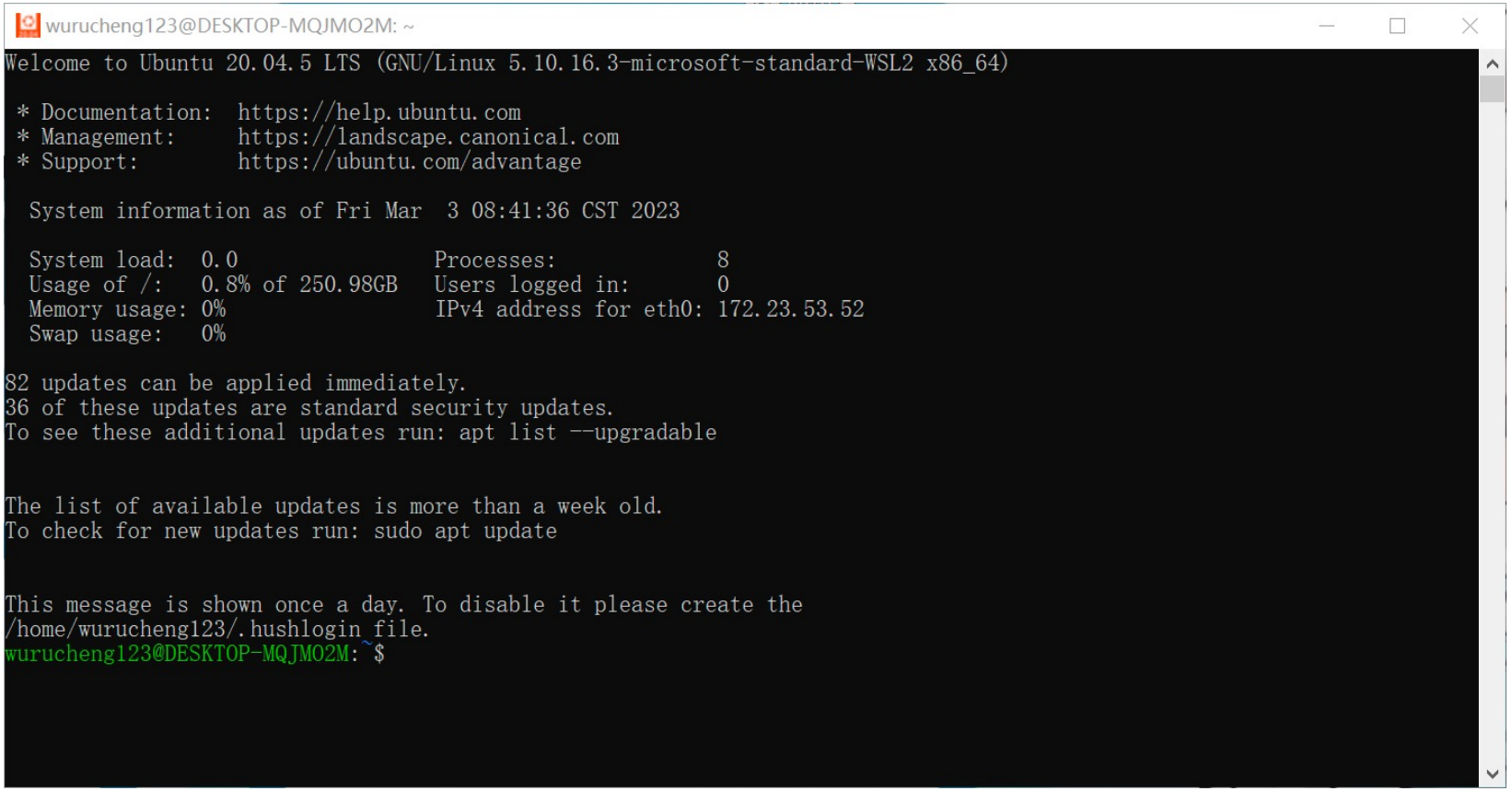
Linux练习环境

- **服务器**
  - **服务器是生物信息学分析的主战场。**
  - **第一种登陆服务器的方式是ssh直连，这种方式不需要额外安装其他应用，比较轻便，但是每次登陆和传输文件都需要自己输入ip和账号密码。**

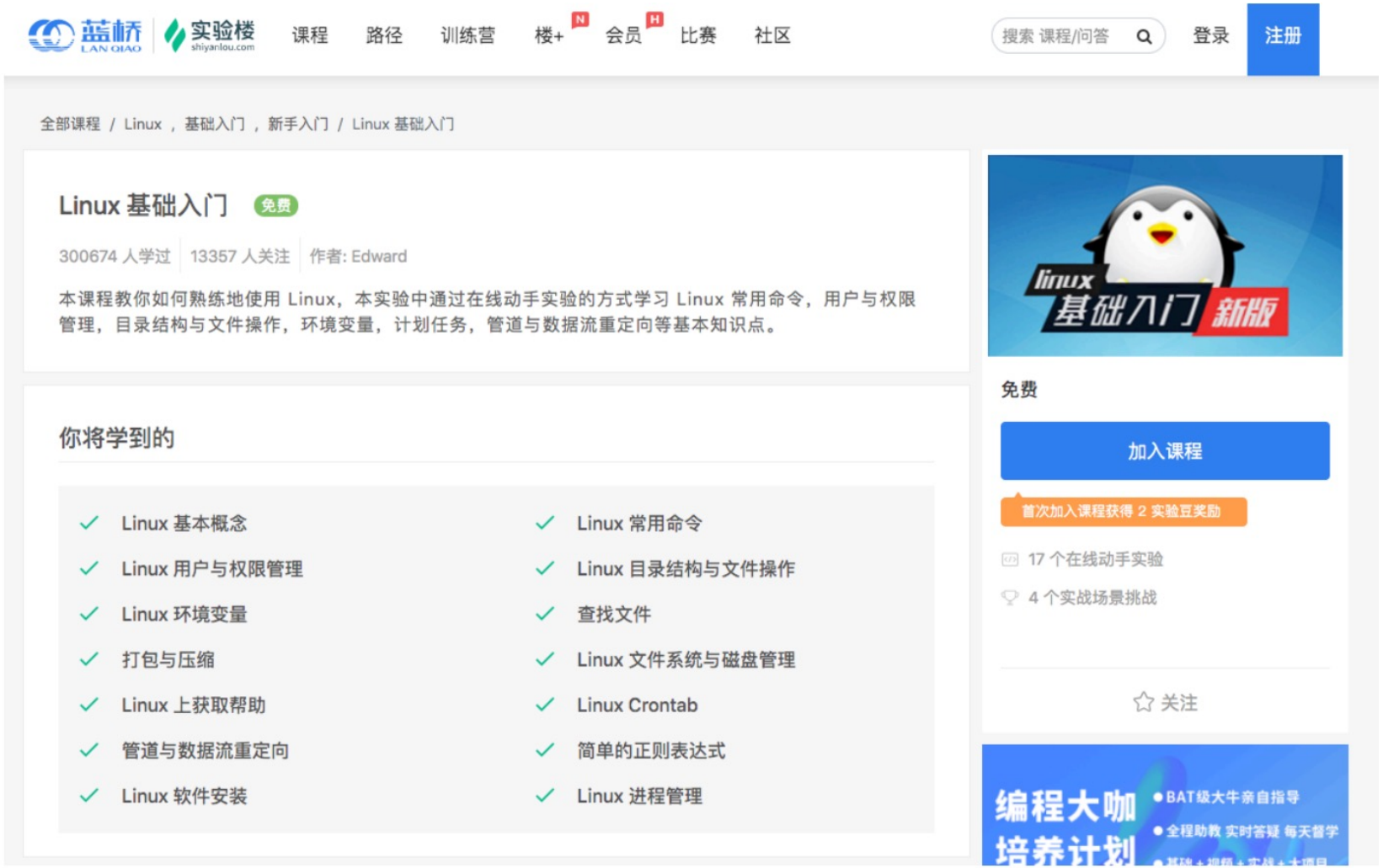


Linux练习

- **Windows子系统**



Linux学习推荐



Slide from Meng

常用编程语言——Python

- Guido van Rossum在1989年创造了Python，在1991年将Python首次公开发行
- 跨平台、开源、解释型语言、高级语言
- 源代码可见
- 第三方包种类多，可用底层语言（如 C、Rust）重写关键代码
- 开发效率高，执行效率低

Slide from Zhao

Python语言的特点

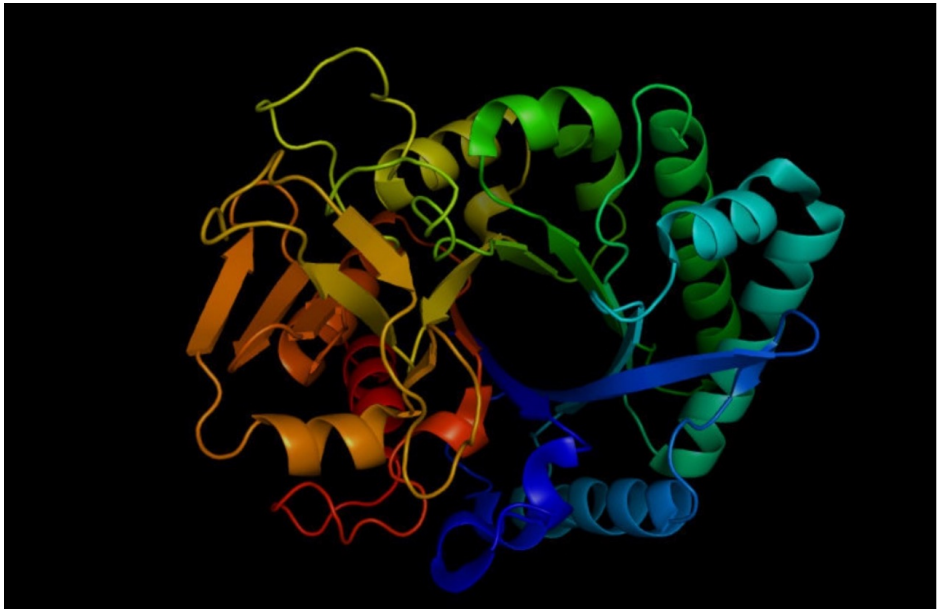
- Python 开发效率高，执行效率低 (科研, 数据分析)
- C 开发效率低, 执行效率高 (高频量化交易)

■ 执行效率对比

C	10000行	0.01s
Java	1000行	0.05s
Python	100行	0.1s

Python在生物信息学中的应用

- 数据爬取（爬取UniProt）
  - 爬虫
- Web前后端开发（AnnoLnc2）
  - Django, Flask, 数据库…
- 工作流程
  - Jupyterlab
  - Snakemake
- 字符串处理
  - FASTA, FASTQ, BED-like, BAM/SAM
  - Biopython, Pysam
- 数据分析机器学习（AlphaFold2）
  - 数据清洗，数据分析，可视化
  - 特征工程，机器学习

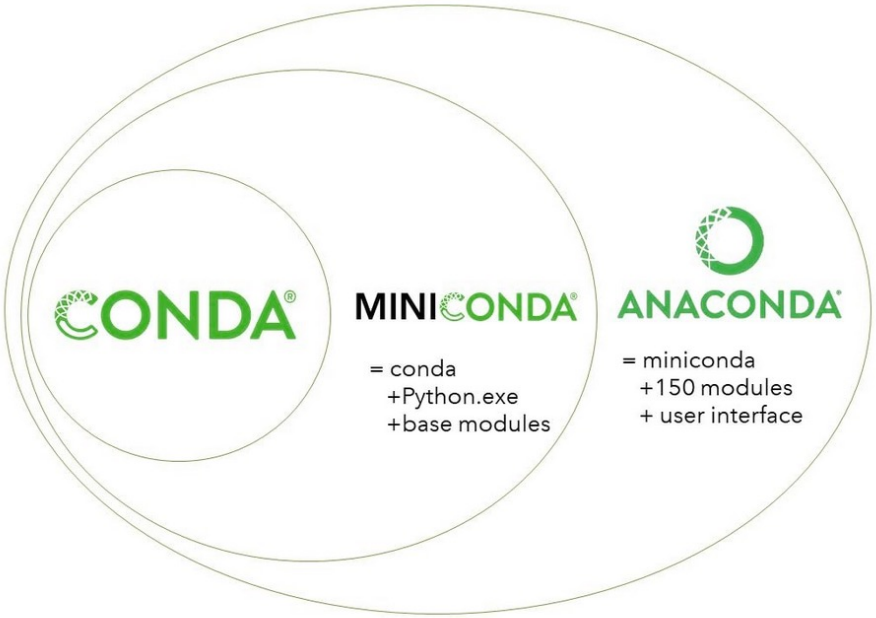


AlphaFold2蛋白结构预测

Slide from Zhao

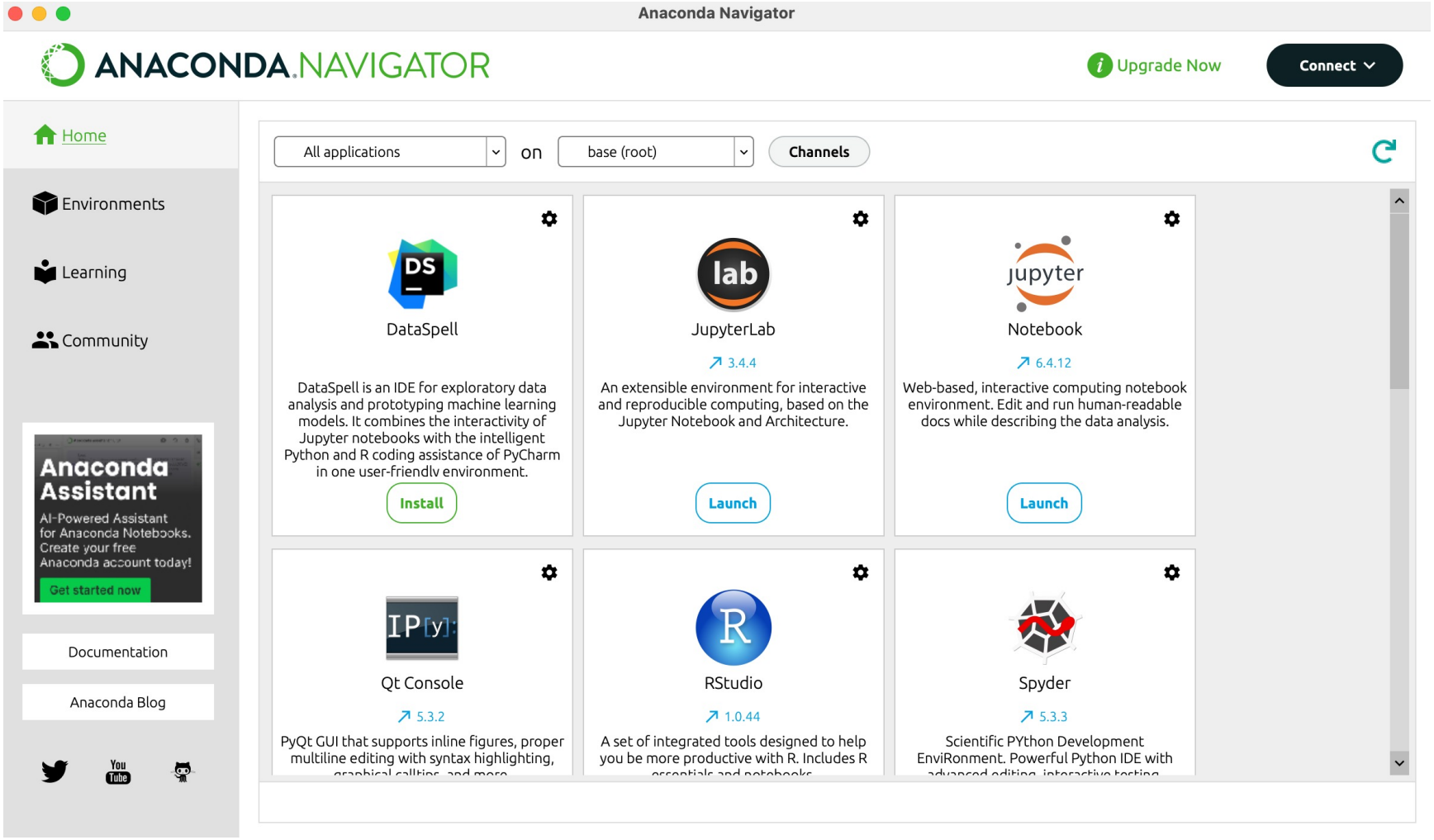
Conda简介

- Conda 是一款环境管理工具
  - 最流行的Python环境管理工具之一
  - 开源的软件包管理系统和环境管理系统，用于安装多个版本的软件包及其依赖关系，并在不同环境间切换。
  - Conda是为Python程序创建的，也可以打包和分发其他软件。
  - Linux，MacOS和Windows跨平台
- Conda
  - Anaconda
  - Miniconda



Slide from Zhao

ANACONDA


Anaconda Navigator
ANACONDA.NAVIGATOR
Upgrade Now
Connect
Home
Environments
Learning
Community
Anaconda Assistant
AI-Powered Assistant for Anaconda Notebooks. Create your free Anaconda account today!
Get started now
Documentation
Anaconda Blog
All applications
on
base (root)
Channels
DataSpell
DataSpell is an IDE for exploratory data analysis and prototyping machine learning models. It combines the interactivity of Jupyter notebooks with the intelligent Python and R coding assistance of PyCharm in one user-friendly environment.
Install
JupyterLab
3.4.4
An extensible environment for interactive and reproducible computing, based on the Jupyter Notebook and Architecture.
Launch
Notebook
6.4.12
Web-based, interactive computing notebook environment. Edit and run human-readable docs while describing the data analysis.
Launch
Qt Console
5.3.2
PyQt GUI that supports inline figures, proper multiline editing with syntax highlighting,
RStudio
1.0.44
A set of integrated tools designed to help you be more productive with R. Includes R
Spyder
5.3.3
Scientific PYthon Development EnviRonment. Powerful Python IDE with

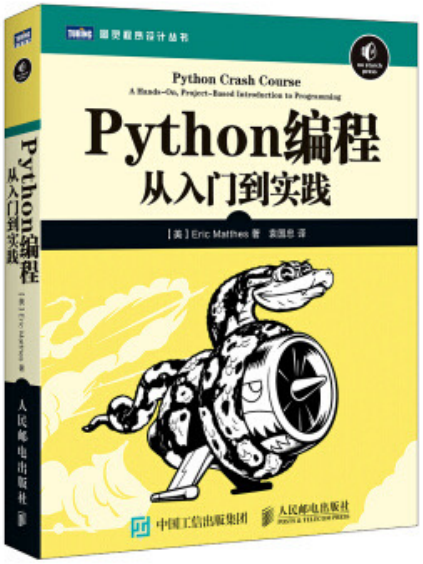
JupyterLab的使用

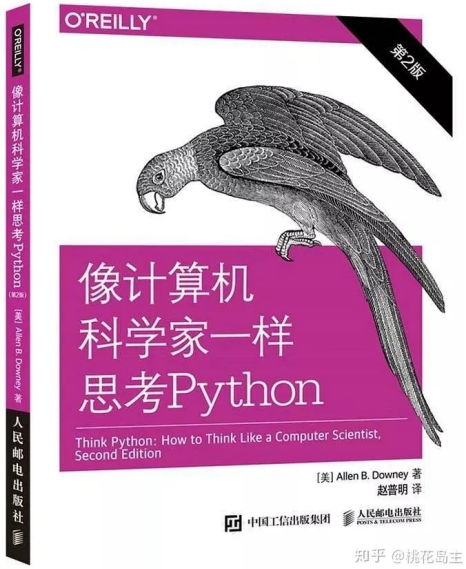
JupyterLab是一款交互式python IDE，现在可支持多种编程语言，写代码的同时可以进行markdown，还能保存代码输出结果，是非常好用的代码管理工具。

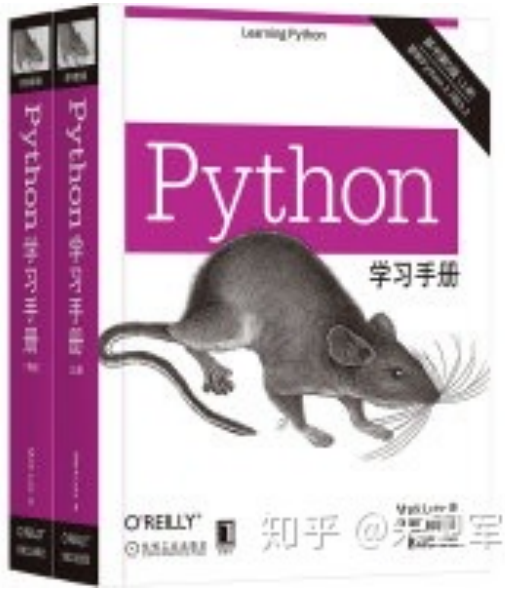
入门推荐

https://www.icourse163.org/course/0809BIT008-268001?outVendor=zw_mooc_pclszykctj_

入门推荐


Python Crash Course
A Hands-On, Project-Based Introduction to Programming
Python编程
从入门到实践
[美] Eric Matthes 著
人民邮电出版社


O'REILLY
第2版
像计算机
科学家一样
思考Python
Think Python: How to Think Like a Computer Scientist, Second Edition
[美] Allen B. Downey 著
赵普明 译
中国工信出版集团
人民邮电出版社
知乎 @桃花岛主


Learning Python
Python
学习手册
O'REILLY
知乎 @

常用编程语言——R

- R语言是从S语言演变而来的。
- S语言是二十世纪70年代诞生于贝尔实验室，由Rick Becker，John Chambers，Allan Wilks开发。
- 用S语言开发的商业软件Splus，取得了巨大成功。

Slide from Meng

R语言的发展历史

- 1995年由新西兰Auckland大学统计系的Robert Gentleman和Ross Ihaka，编写了一种能执行S语言的软件
- 并将该软件的源代码全部公开，这就是R软件，其命令统称为R语言。

Robert Gentleman

Ross Ihaka

Slide from Meng

R语言的优势

- 开源，免费
- 优秀的跨平台性
- 支持脚本和批处理
- 对统计问题支持良好
- 语法灵活
- 画图方便
- 丰富的扩展性（R包）
- ……

Slide from Meng

R语言的不足

- 慢的运算速度
- 慢的for循环
- 内存消耗大
- 非命令行下只支持单核计算
- ……

Slide from Meng

R语言的下载

- 下载地址：https://cran.r-project.org
- Windows、MacOS按普通软件安装方式安装即可。
- Linux需用命令安装，网站上有给出。

CRAN
Mirrors
What's new?
Search
CRAN Team

About R
R Homepage
The R Journal

Software
R Sources
R Binaries
Packages
Task Views
Other

Documentation
Manuals
FAQs
Contributed

The Comprehensive R Archive Network

Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- Download R for Linux (Debian, Fedora/Redhat, Ubuntu)
- Download R for macOS
- Download R for Windows

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

Source Code for all Platforms

Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!

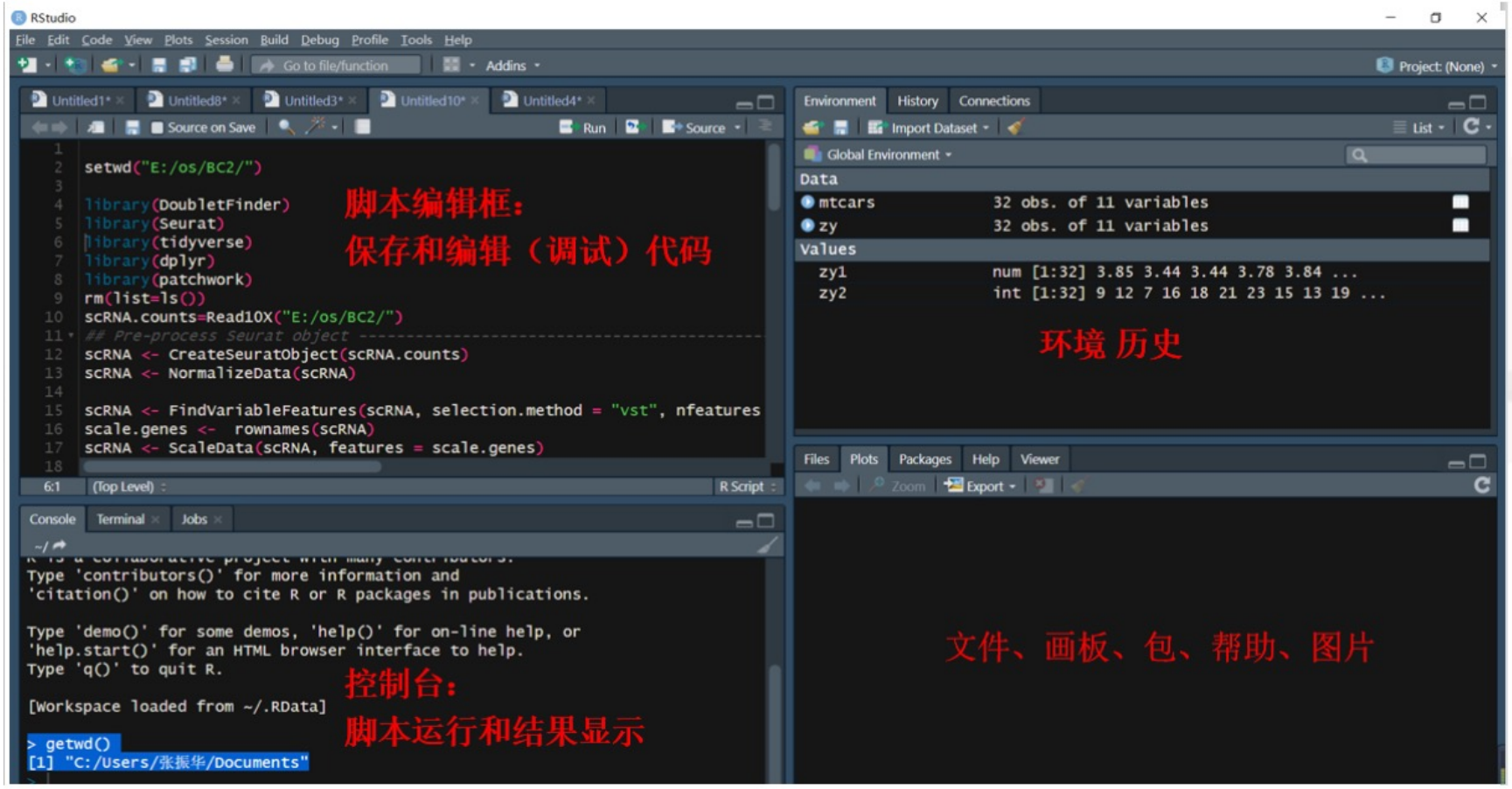
- The latest release (2022-10-31, Innocent and Trusting) R-4.2.2.tar.gz, read what's new in the latest version.
- Sources of R alpha and beta releases (daily snapshots, created only in time periods before a planned release).
- Daily snapshots of current patched and development versions are available here. Please read about new features and bug fixes before filing corresponding feature requests or bug reports.
- Source code of older versions of R is available here.
- Contributed extension packages

Questions About R

- If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.

Rstudio介绍

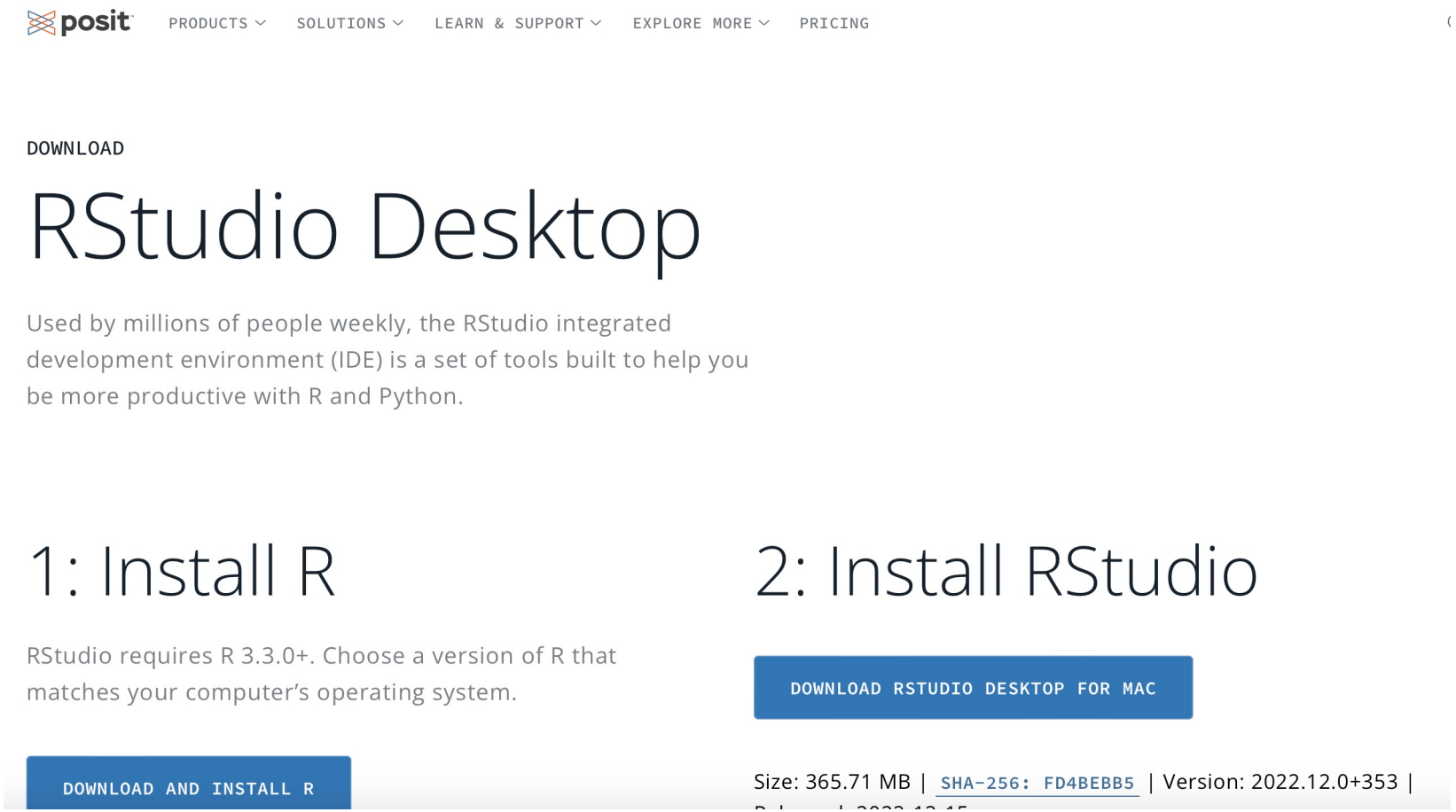
- Rstudio为一款R语言IDE，能够可视化当前环境、数据结构以及图，对新手入门R语言很友好。



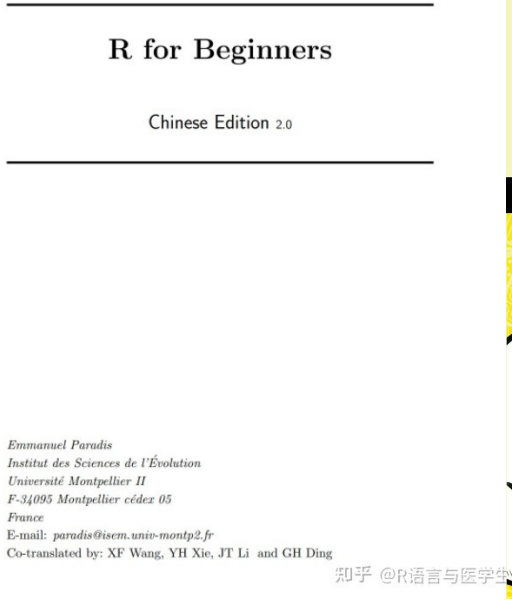
Slide from Hua

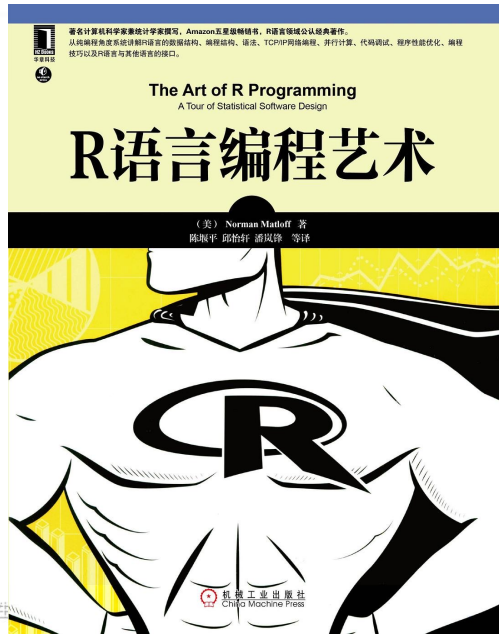
Rstudio的安装

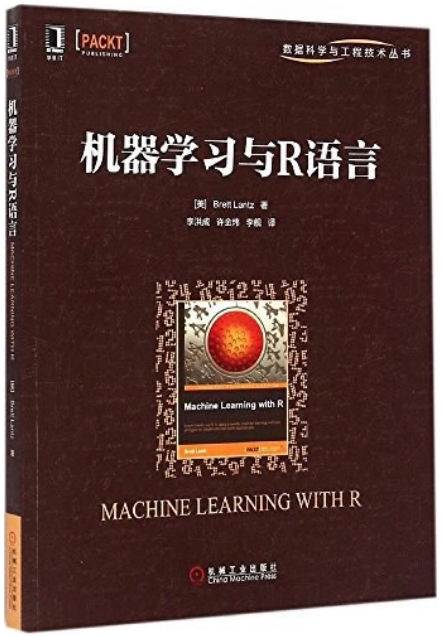
- 下载地址：https://posit.co/download/rstudio-desktop/



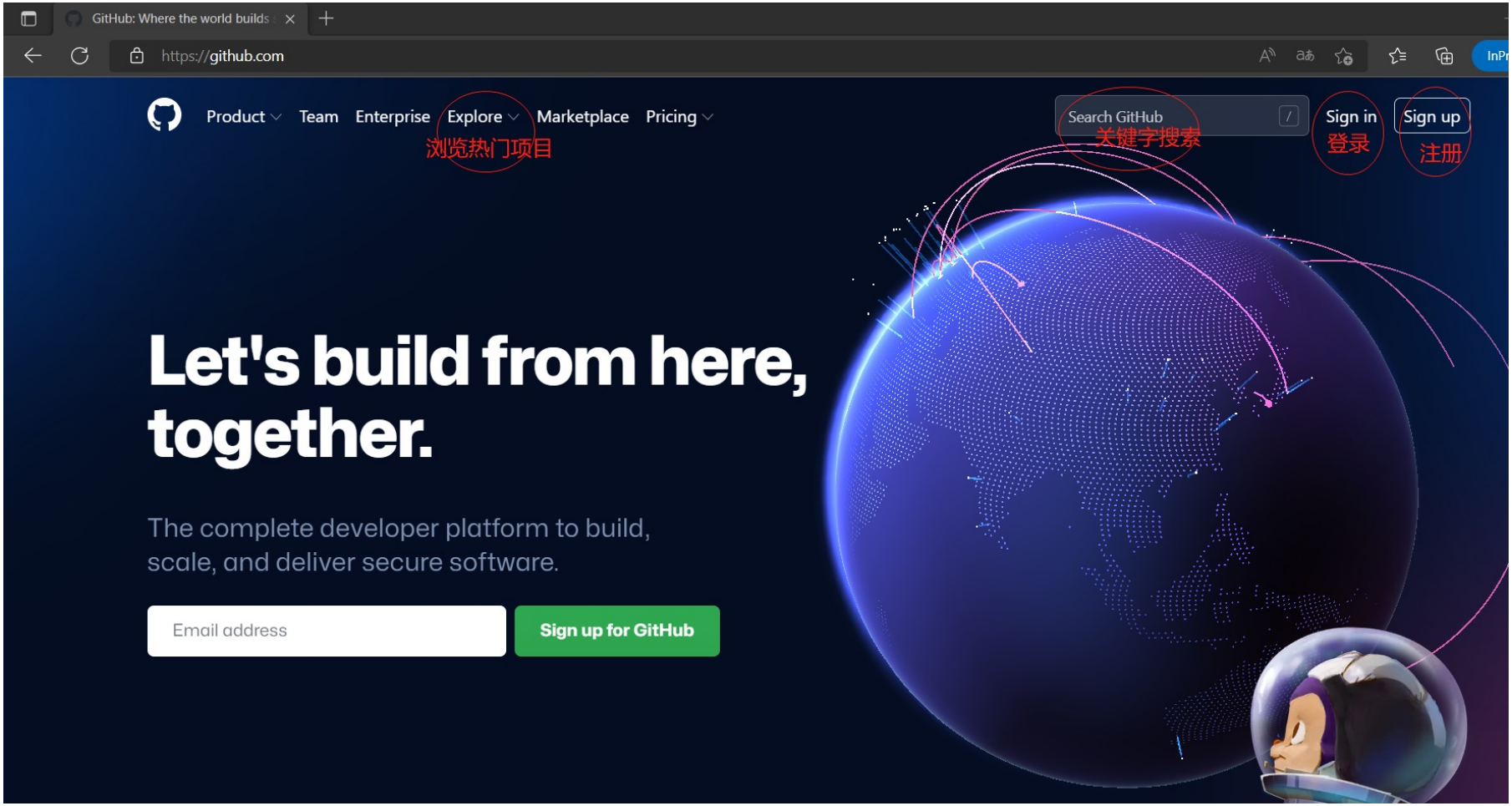
入门推荐


R for Beginners
Chinese Edition 2.0
Emmanuel Paradis
Institut des Sciences de l'Évolution
Université Montpellier II
F-34095 Montpellier cédex 05
France
E-mail: paradis@isem.univ-montp2.fr
Co-translated by: XF Wang, YH Xie, JT Li and GH Ding

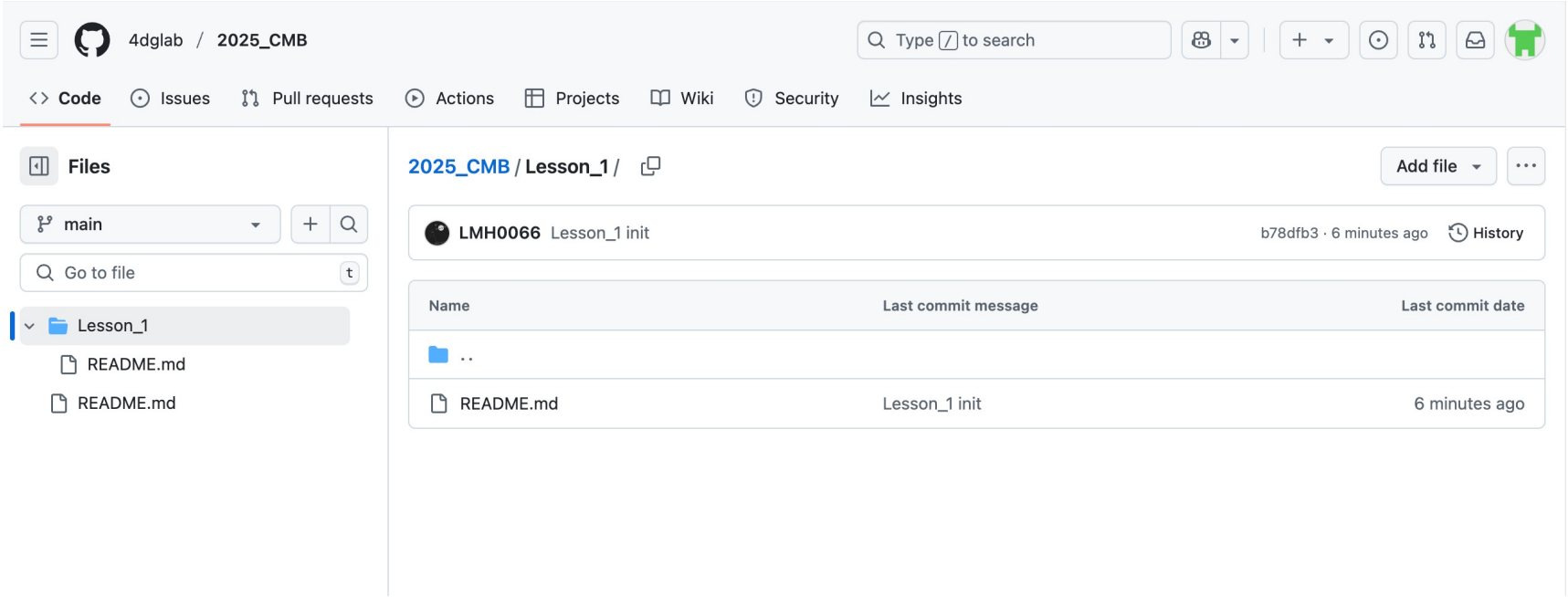

The Art of R Programming
A Tour of Statistical Software Design
R语言编程艺术
（美） Norman Matloff 著
陈振平 邱怡轩 潘岚锋 等译
机械工业出版社
China Machine Press


[PACKT]
数据科学与工程技术丛书
机器学习与R语言
[美] Brett Lantz 著
Machine Learning with R
MACHINE LEARNING WITH R
机械工业出版社
China Machine Press

GitHub介绍


GitHub: Where the world builds
https://github.com
Product
Team
Enterprise
Explore
浏览热门项目
Marketplace
Pricing
Search GitHub
关键字搜索
Sign in
登录
Sign up
注册
Let's build from here, together.
The complete developer platform to build, scale, and deliver secure software.
Email address
Sign up for GitHub

课程PPT下载



https://github.com/4dglab/2025_CMB/tree/main/Lesson_1

- **Thanks for your attention!**