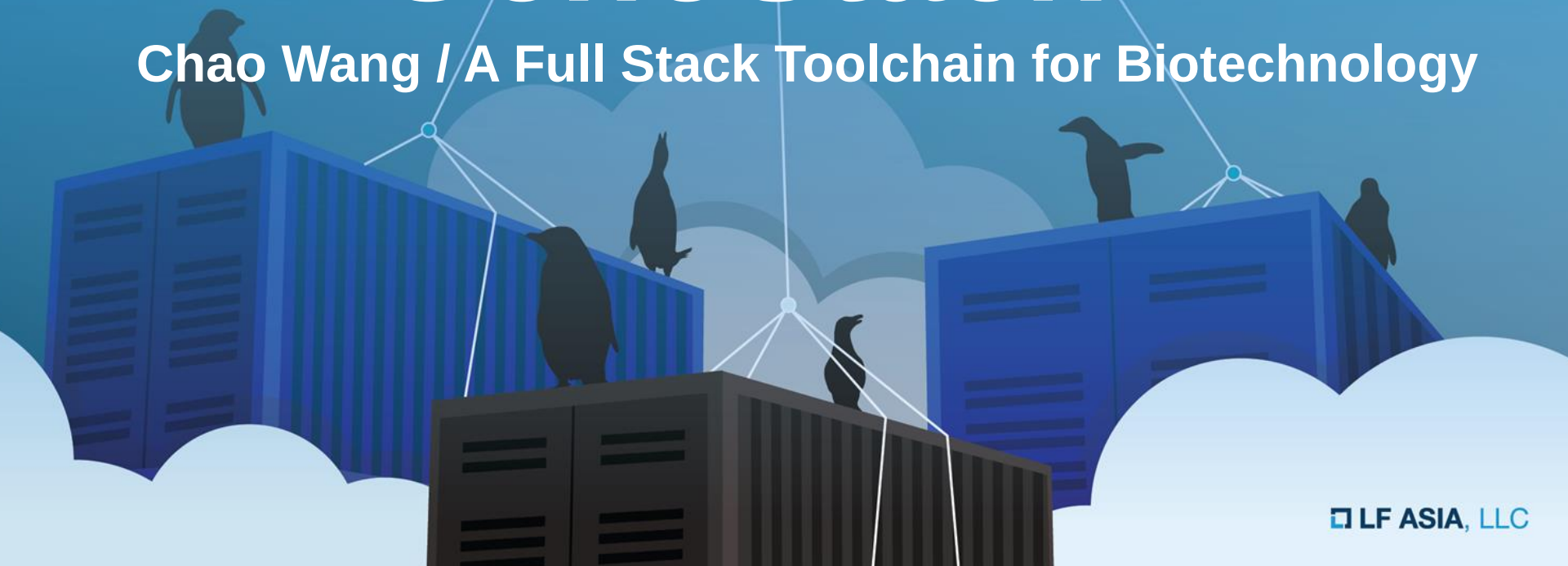
LINUXCON
containercon
CHINA 中国
CLOUDOPEN
THINK OPEN
开放性思维

# GenoStack™

## Chao Wang / A Full Stack Toolchain for Biotechnology



LF ASIA, LLC

# 目录(Contents)

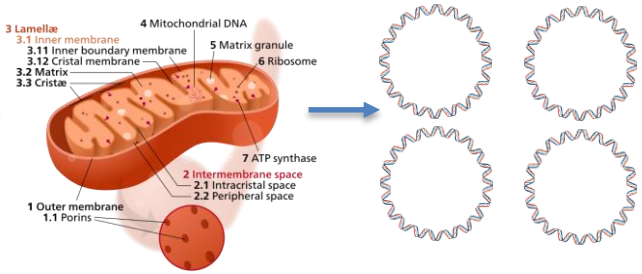


- 背景知识：生物信息处理基本过程
  Background：Workflow of bioinformatics
- GenoStack概览(Overview)
- GenoStack生物信息工具链
  Full Stack Toolchain for biotechnology
- GenoStack推荐系统(Recommend system)
- GenoStack场景化服务(Context service)

# 背景知识(Background)



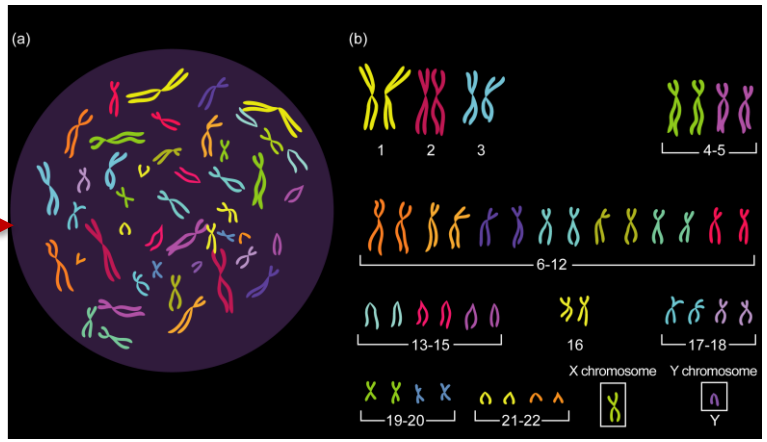
先补充点初中的生物知识（biology from high school）



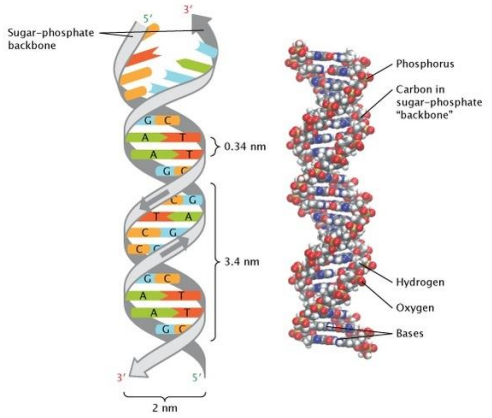


线粒体作为细胞的能量工厂，也有自己的遗传物质，其DNA为环形，只来自于母亲，可用于追溯母系祖源

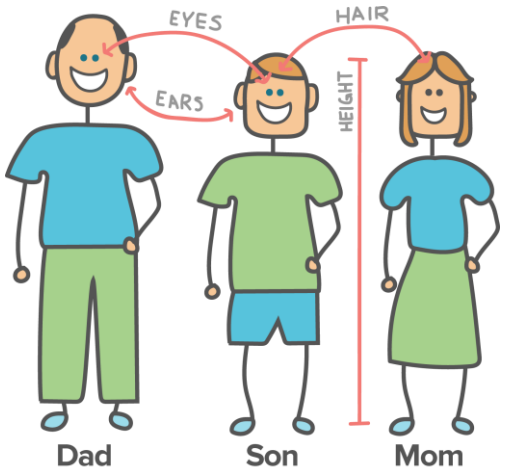
人类是真核生物




人类有23对染色体(22对常染色体，1对性染色体)其中一组来自父亲，一组来自母亲




约30亿个碱基对 2万到2.5万个基因，决定了我们的各种特征

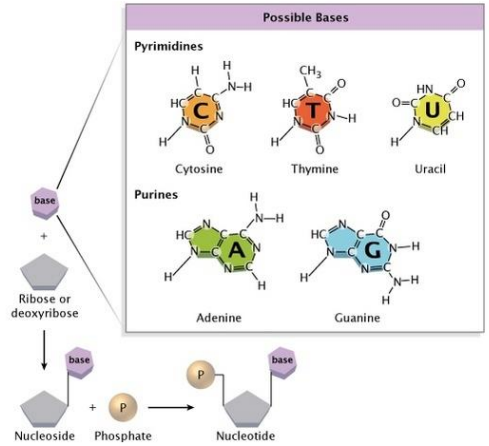



https://www.slideserve.com/cynthia-pittman/two-basic-cell-types-prokaryotic-vs-eukaryotic-cells
https://pmgbiology.files.wordpress.com/2015/10/chromosomes.png
https://www.semanticscholar.org/paper/Circular-DNA-Vologodskii/4942d0246f18be904a25db942d304e52ca7147be
https://www.khanacademy.org/test-prep/mcat/behavior/behavior-and-genetics/a/genes-environment-and-behavior

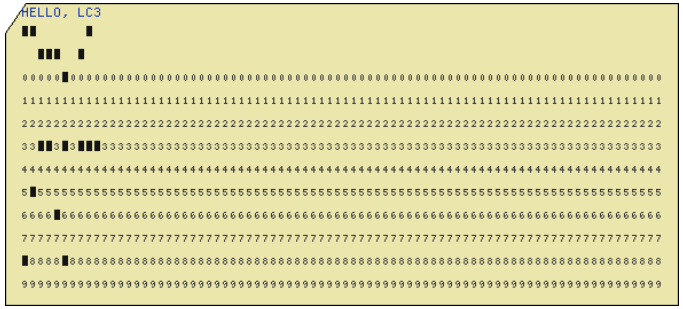
# 背景知识(Background)



DNA is the code of life




Vs.




IBM029 and punch cards

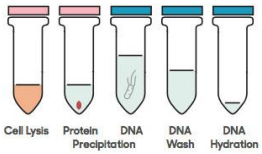
http://www.kloth.net/services/cardpunch.php

# 背景知识(Background)

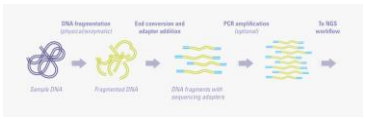


基因数据处理的关键过程，数据分析所需要的软件平台会越来越重要

唾液样本
(saliva)
无创非接触式




核酸提取(dna extraction)



文库制备(library construction)

准备阶段(preparation)

二代测序(NGS)

芯片(chip array)

上机测序(sequencing)

NGS

原始文件
(raw data)

第一阶段(primary)
质量控制 通常和测序仪类型强相关

Chip array

Fastq

第二阶段(Secondary)
生物学流水线，GATK bio-pipeline

基因型

第三阶段(Tertiary)
解读 综合分析 大数据 机器学习

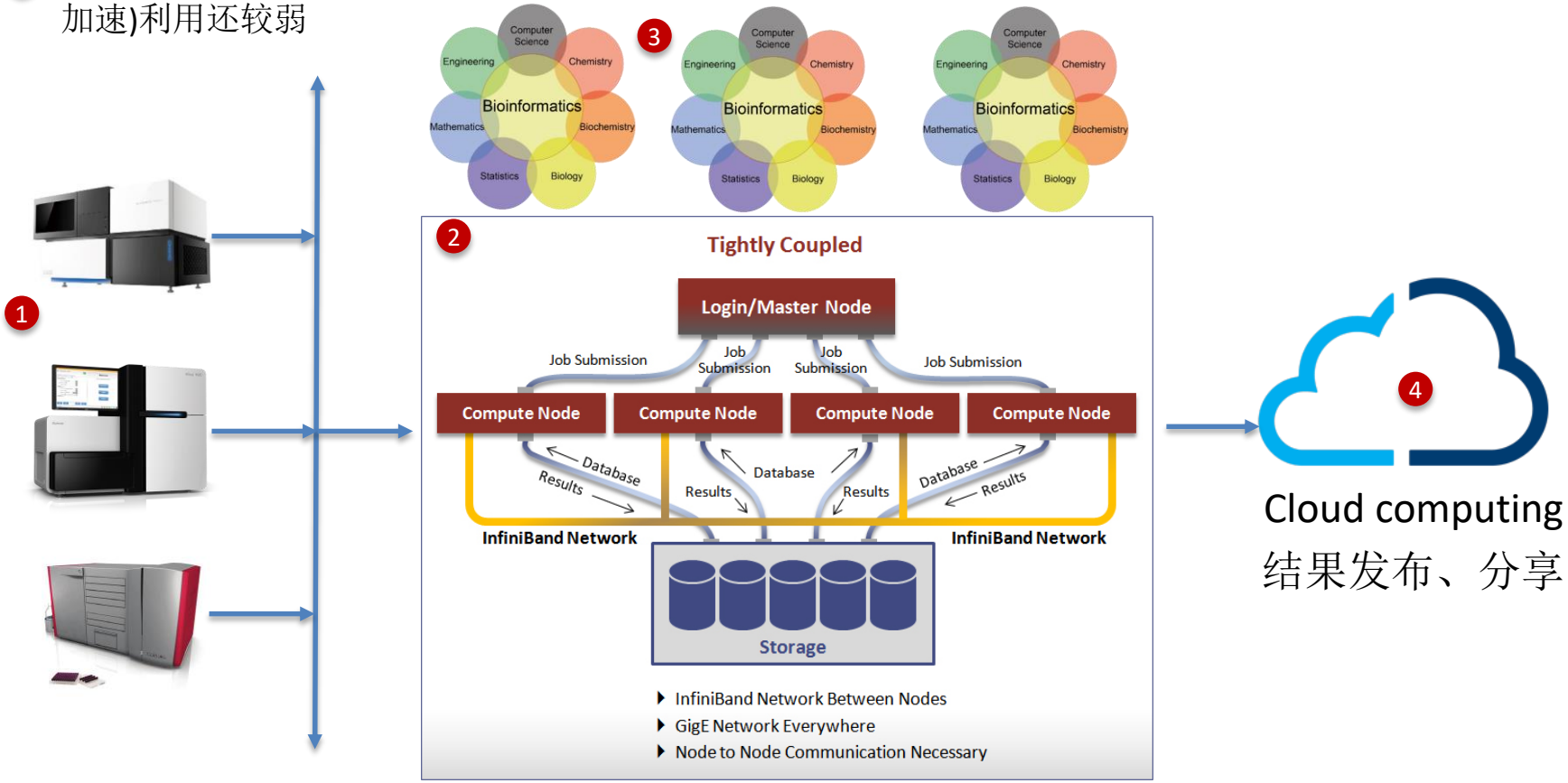
数据分析(data analysis)

http://bioted.es/1-1-extraction-and-purification-of-nucleics-acids/?lang=en
https://www.thermofisher.com/cn/zh/home/brands/thermo-scientific/molecular-biology/thermo-scientific-specialized-molecular-biology-applications/sequencing-thermo-scientific/next-generation-sequencing-library-preparation-thermo-scientific.html

# 生命科学领域IT设施面临的挑战



1. 针对不同场景 存在多种设备类型 产生的数据格式有差异
2. 传统HPC模式的扩展性、可维护性带来巨大的瓶颈
3. 多样的细分领域（基因组学、代谢组、蛋白质组，动植物）巨大的工具集 碎片化导致代码可维护性低
4. 向云端迁移 需要专业的技术支撑来平滑迁移;当前对云计算的成熟技术(大数据、人工智能、GPU/FPGA加速)利用还较弱



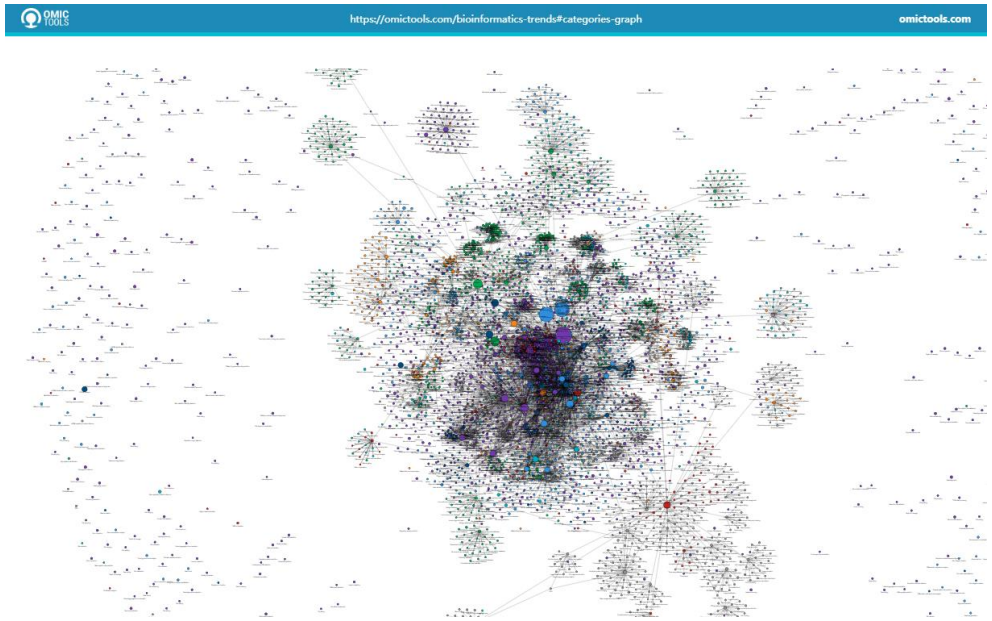

https://insidehpc.com/white-paper/cloud-computing-guide/

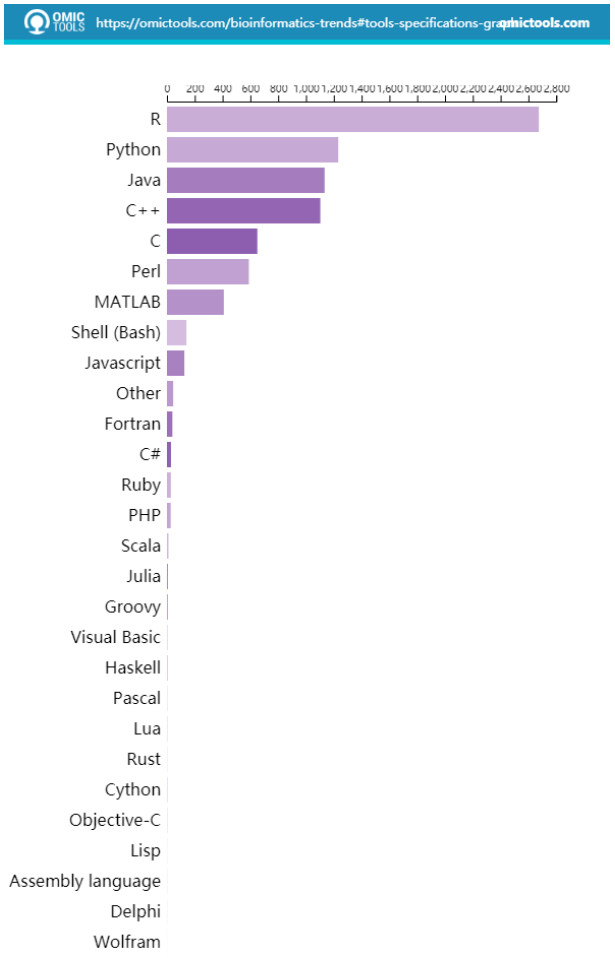
http://www.nust.edu.pk/INSTITUTIONS/Centers/RCMS/ap/pg/MSBioinformatics/Pages/default.aspx

# 基因领域IT设施面临的挑战

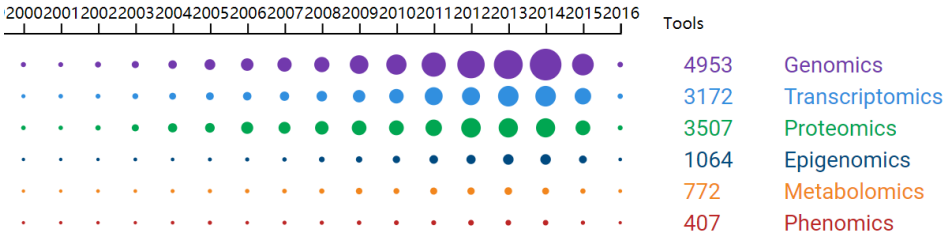


生物信息处理的工具碎片化




多达5000种不同的工具 而且还在增加

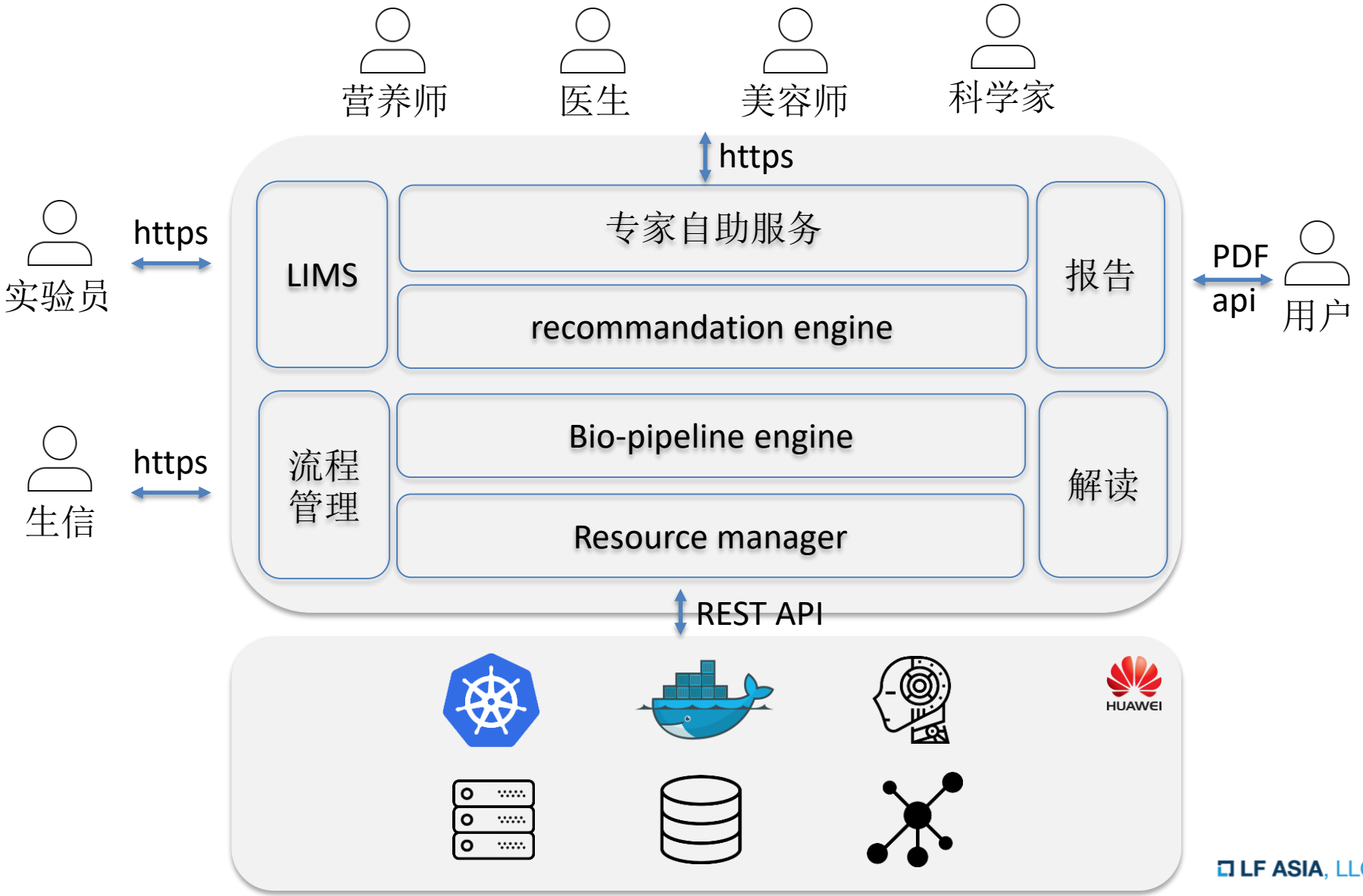



不同的细分领域有自己的工具集

不同的语言环境

# GenoStack架构概览(Overview)

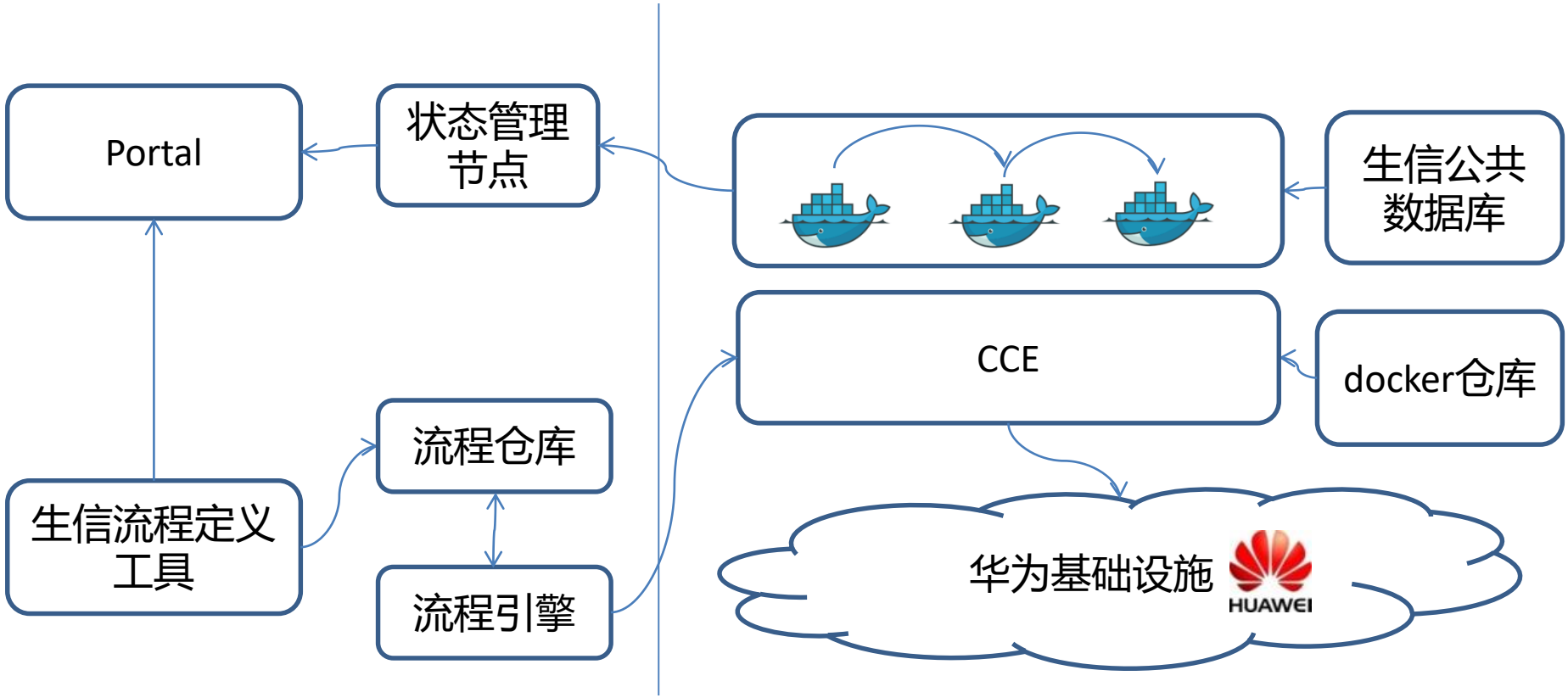




# GenoStack基于容器的Pipeline



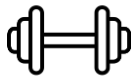
1.可复用的流程（WDL/CWL语言描述）
2.工具进行容器化管理 方便复用



# GenoStack基于基因的推荐系统



饮食 运动 提醒

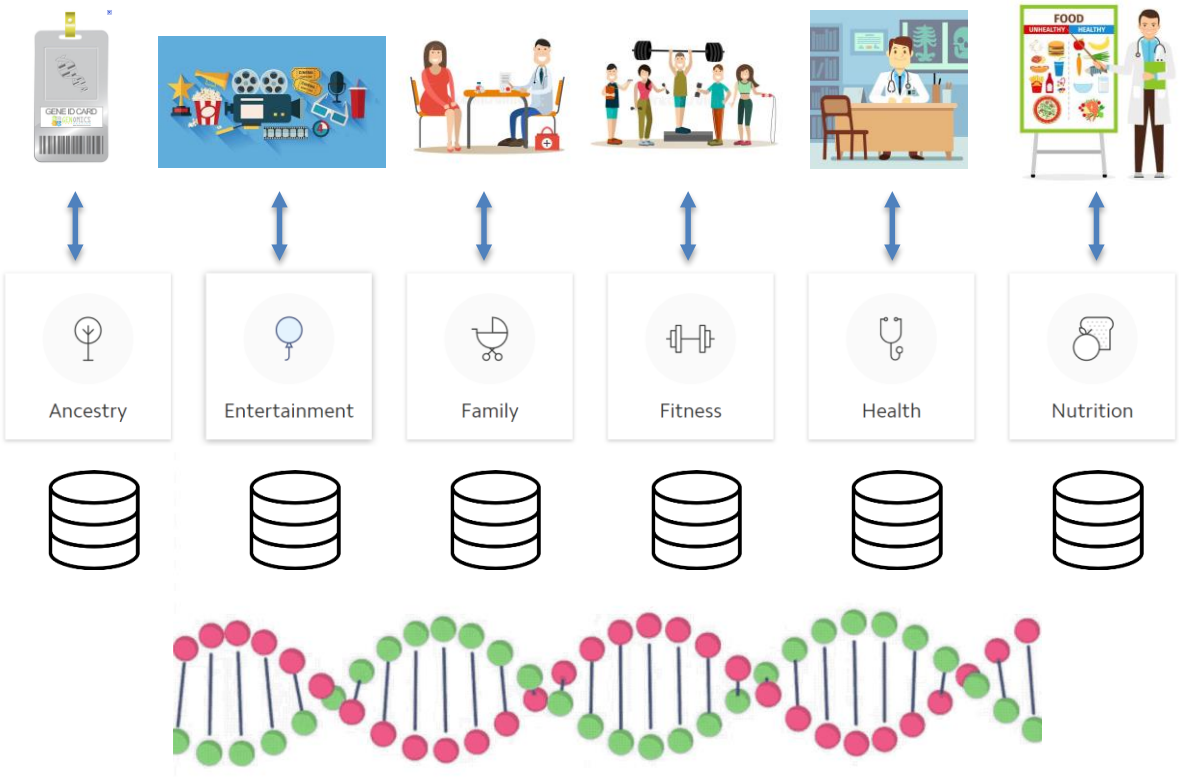





乳糖不耐受
咖啡因代谢快
最大吸氧量高

# GenoStack基因场景化服务






30亿碱基 2万多个基因

Like 5G network slicing

https://www.helix.com/blog/decides-dna-read-cells/

# 愿景



- 充分利用云原生的能力，紧密集成
- 让每个人都用好（简单、快捷、安全）自己的基因数据

LINUXCON
containercon
CLOUDOPEN
CHINA 中国
THINK OPEN
开放性思维

# “Quote Placeholder”


LINUXCON
containercon
CLOUDOPEN
CHINA 中国
THINK OPEN
开放性思维
LF ASIA, LLC