计算系统生物学

王 勇

中国科学院数学与系统科学研究院

个人介绍

- 专业： 运筹学与控制论
- 2002年开始从事数学与生物学交叉研究

- 问题： 最优化算法→蛋白质结构预测→蛋白质比对→生物分子网络→生物数据集成→干细胞

- 方法： 最优化模型→统计学习模型→机器学习模型→统计

联系信息

- 答疑&讨论: 课后或思源楼 1003

* 对课程的意见和建议

邮件： ywang@amss.ac.cn

Personal viewpoint

计算系统生物学

Computational Systems Biology

生物学

- 生物学是研究生命现象和生物活动规律的科学。
- 生物学的一些基本研究方法——观察描述的方法、比较的方法和实验的方法等是在生物学发展进程中逐步形成的。在生物学的发展史上，这些方法依次兴起，成为一定时期的主要研究手段。现在，这些方法综合而成现代生物学研究方法体系。

From Wiki

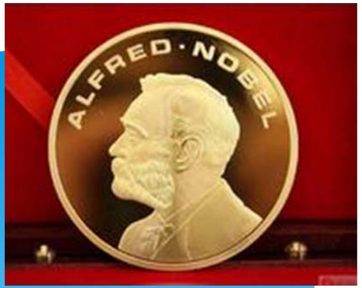
One thing important

- **"Nothing in Biology Makes Sense Except in the Light of Evolution"** is a 1973 essay by the evolutionary biologist Theodosius Dobzhansky, criticising anti-evolution creationism and espousing theistic evolution.
- The essay was first published in the *American Biology Teacher*, volume 35, pages 125-129.

系统科学

- 系统是由相互联系、相互作用的要素(部分)组成的具有一定结构和功能的有机整体。
- **系统科学**(系统理论)是以系统为研究和应用对象的一门科学。研究系统的结构、动力学、调控、信息与组织化等，系统科学是分析与综合渗透的研究方法，包括物理学、化学、生物学、心理学、社会学和医学、工程等领域的系统方法、原理和技术等研究。
- 研究系统与环境的关系
- “整体大于部分之和”

系统



- “Physics is the only real science. The rest are just stamp collecting.” or “All science is either physics or stamp collecting”

-- Ernest Rutherford

- 卢瑟福是20世纪初最伟大的实验物理学家，1908年诺贝尔化学奖得主，一生发表论文约215篇，著作6部，**培养了10位诺贝尔奖获得者**。卢瑟福的实验室被后人称为“诺贝尔奖得主的幼儿园”（剑桥大学卡文迪许实验室（Cavendish Laboratory））。

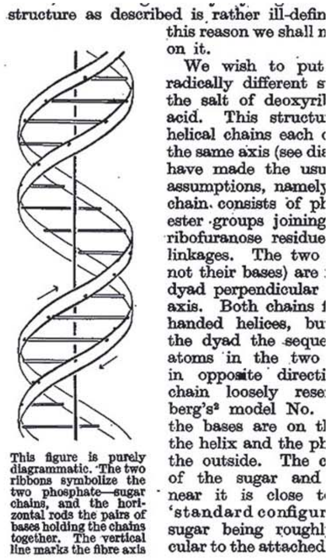
由卡文迪许衍生的LMB（Laboratory of Molecular Biology）更创造了一个机构高峰纪录：迄今55年历史有12个诺贝尔奖获得者

计算系统生物学的本质？？？

生物信息学

系统生物学

《Nature》1953



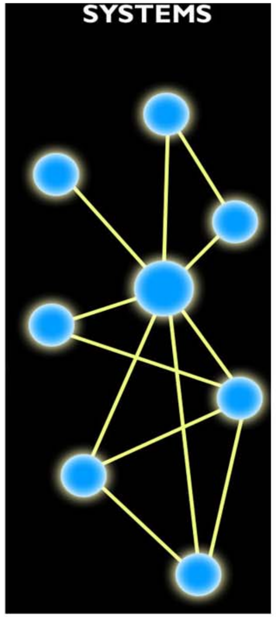
structure as described is rather ill-defin
this reason we shall n
on it.
We wish to put
radically different s
the salt of deoxyrib
acid. This structu
helical chains each
the same axis (see dia
have made the usu
assumptions, namely
chain consists of ph
ester groups joining
ribofuranose residue
linkages. The two
not their bases) are
dyad perpendicular
axis. Both chains f
handed helices, bu
the dyad the seque
atoms in the two
in opposite directi
chain loosely rese
berg's² model No.
the bases are on th
the helix and the ph
the outside. The c
of the sugar and
near it is close t
'standard configur
sugar being roughl
cular to the attached

This figure is purely diagrammatic. The two ribbons symbolize the two phosphate—sugar chains, and the horizontal rods the pairs of bases holding the chains together. The vertical line marks the fibre axis

DNA双螺旋结构发现，
开启分子生物学研究

《Science》2001

人类基因组计划完成



系统生物学

1950
低通量观测数据

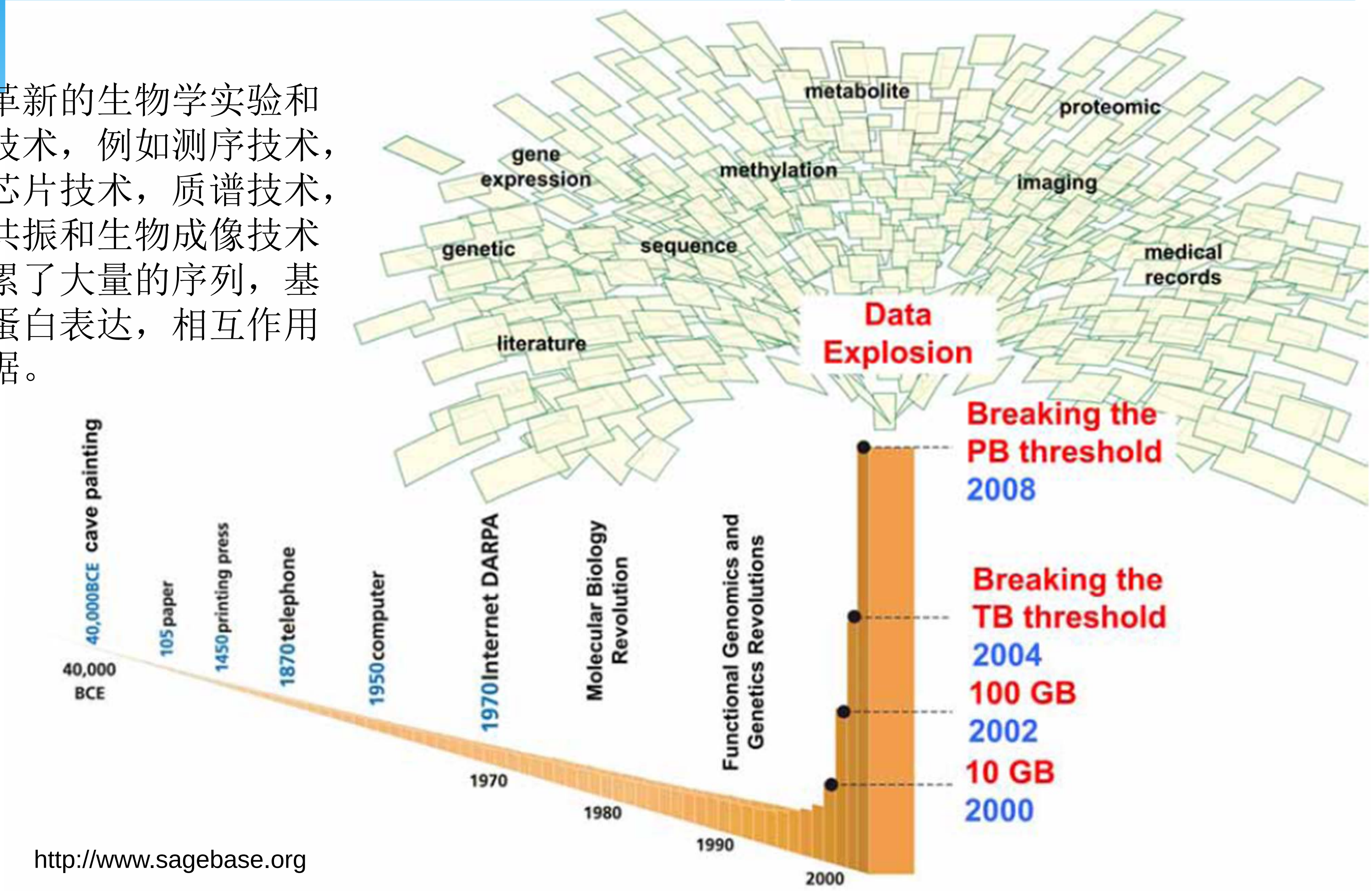
1990
基因组数据

2002
组学数据

本质上是由生物数据海量积累驱动

生物数据井喷

不断革新的生物学实验和观测技术，例如测序技术，基因芯片技术，质谱技术，核磁共振和生物成像技术等积累了大量的序列，基因、蛋白表达，相互作用等数据。



http://www.sagebase.org

BIG DATA

A REVOLUTION THAT WILL TRANSFORM HOW WE LIVE, WORK, AND THINK

大数据时代

生活、工作与思维的大变革

[英]维克托・迈尔–舍恩伯格（Viktor Mayer-Schönberger） 肯尼思・库克耶（Kenneth Cukier）◎著 盛杨燕 周涛◎译

浙江人民出版社

孜孜不倦的数据科学家

“维克托・迈尔－舍恩伯格二十多年来一直致力于网络经济、信息与创新、信息监管、网络规范与战略管理的研究。

从维也纳大学到哈佛大学，从新加坡国立大学到牛津大学，世界上最著名的互联网研究学府都留下了他的足迹。”

开大数据系统研究之先河

“他说，世界的本质就是数据，大数据将开启一次重大的时代转型；

他说，大数据发展的核心动力来源于人类测量、记录和分析世界的渴望。

他说，从因果关系到相关关系的思维变革才是大数据的关键，建立在相关关系分析法基础上的预测才是大数据的核心。”

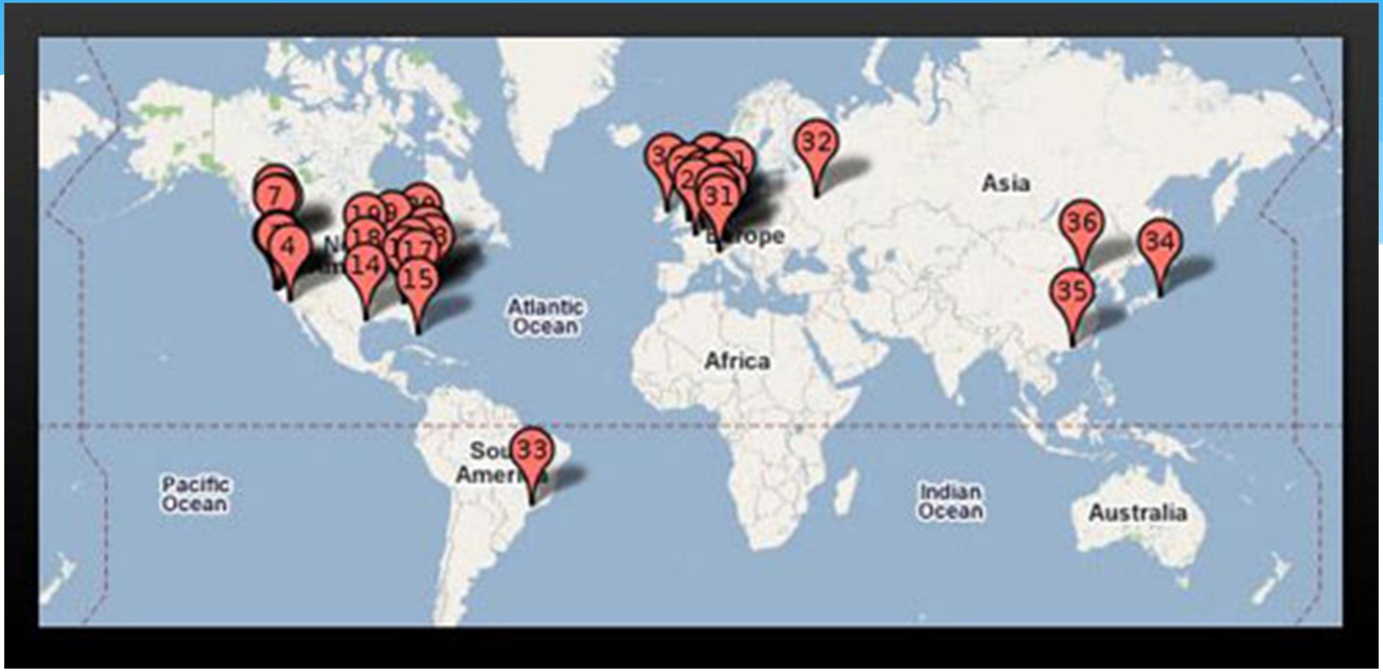
大数据的特点

- Big Data大数据，谈的不仅仅是数据量，其实包含了数据量(Volume)、时效性(Velocity)、多变性(Variety)、可疑性(Veracity)：
- Volume: 数据量大量数据的产生、处理、保存，海量数据
- Velocity:时效性：处理的时效，500万笔数据的深入分析,可能只能花5分钟的时间
- Variety: 多变性指的是数据的形态，包含文字、影音、网页、串流等等结构性、非结构性的数据
- Veracity: 可疑性指的是当数据的来源变得更多元时,这些数据本身的可靠度、质量是否足够，若数据本身就是有问题的,那分析后的结果也不会是正确的。

From 百度百科

Taming big data， Lisa K. Stapleton, Senior Editor, IBM Data Management magazine

Google data center



- Google共有36个数据中心。其中美国有19个、欧洲12个、俄罗斯1个、南美1个,亚洲3个（北京-Google.cn、香港-Google.com.hk和东京各1个）

服务器数量

- 服务器总量外界不清楚。据猜测，应远远不止20万台（2005年35万台），并且每天都在增长。

- 数据中心的专利

  1、服务器内置电池。每台服务器都有一颗12伏特电池，确保万一主断源断电时还可持续供电。最终目的，节约成本。

  2、可移动的数据中心集装箱。2008年10月获得该项专利，每个集装箱中最多可容纳1160台服务器。

定制型数据中心

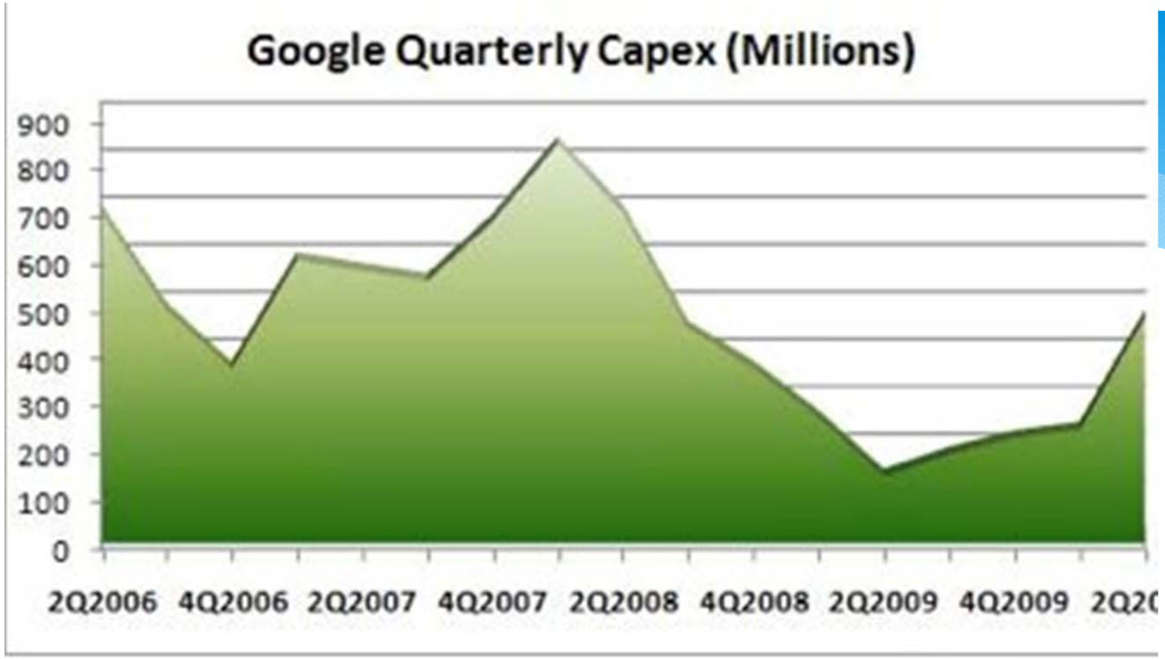
比利时水侧自然冷数据中心、爱尔兰空气侧自然冷数据中心，到芬兰海水直接冷却数据中心

Google比利时数据中心

1、大量的廉价电力；
2、绿色能源，更注重可再生能源；
3、靠近河流或湖泊；（设备冷却需要大量水源）
4、用地广阔；（隐秘性和安全性）
5、和其他数据中心的距离；（数据中心之间的快速链接）
6、税收优惠。

用电量

- 根据Google发布的数据，每位使用Google服务的用户平均每月将消耗Google服务器180瓦的电能，这相当于让一支60瓦的普通灯泡持续工作三小时，该数据来源于Google公司的官方博客。
- 此前有环保组织宣称，Google数据中心的耗电量已经超过了全球总用电量的1%。而根据Google所公布的官方数据，其数据中心在2010年的耗电量为22.6亿千瓦，略微超过老挝的用电量。
- 对比：据国家电网公司预测，在电力供需总体紧张的情况下，今年内和明年中国电力需求依然高增长，预计2005年全国用电量需求达到23,910亿千瓦时。


Google Quarterly Capex (Millions)

- 1Q 2006: $345 million
- 2Q 2006: $699 million
- 3Q 2006: $492 million
- 4Q 2006: $367 million
- 1Q 2007: $597 million
- 2Q 2007: $575 million
- 3Q 2007: $553 million
- 4Q 2007: $678 million
- 1Q 2008: $842 million
- 2Q 2008: $698 million
- 3Q 2008: $452 million
- 4Q 2008:$368 million
- 1Q 2009: $263 million
- 2Q 2009: $139 million
- 3Q 2009: $186 million
- 4Q2009: $221 million
- 1Q2010: $239 million
- 2Q2010: $476 million

根据Google的盈利报告，2006年Google在数据中心上的开销是19亿美元，2007年是24 亿，2008年23.6亿。

Google的最初宗旨：“使世界上所有的信息，能被普遍和有用的被搜寻到。”

《撬动地球的google》

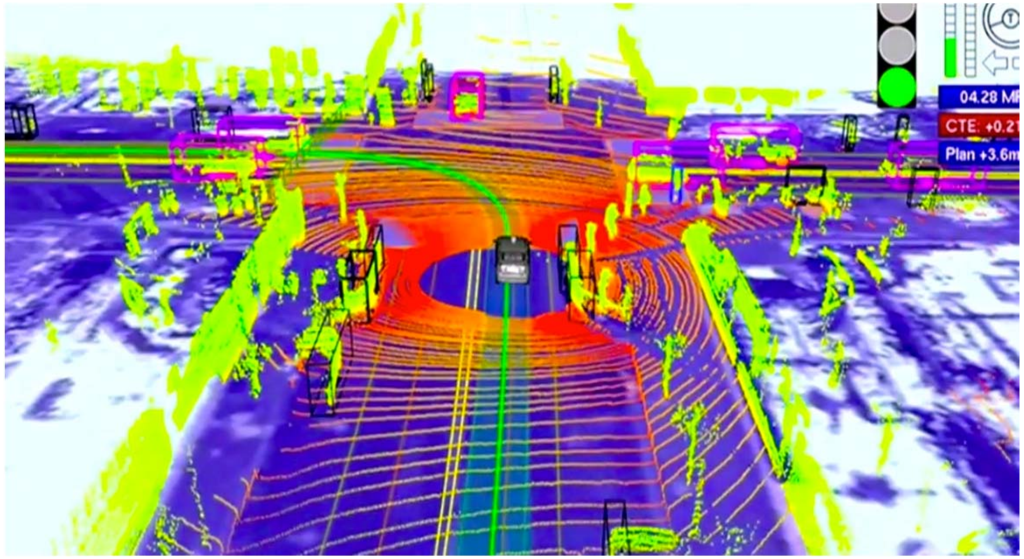
Google X Lab

- **Google X Lab**, sometimes known as **Google X**, is a secret facility run by Google thought to be located somewhere in the Bay Area of Northern California. Work at the lab is overseen by Sergey Brin, one of Google's co-founders.
- Reportedly worked on at the lab is a list of 100 projects pertaining to future technologies such as a space elevator, self-driving car, augmented reality glasses, a neural network that uses semi-supervised learning, enabling speech recognition and extraction of objects from video - for instance detecting if a cat is in a frame of video, and the Web of Things.

From Wiki

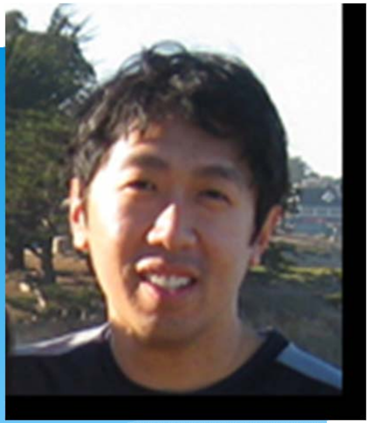
Self-driving car

- **Sebastian Thrun** is a Research Professor of Computer Science at Stanford University, a Google Fellow, a member of the National Academy of Engineering and the German Academy of Sciences. Thrun is best known for his research in robotics and machine learning, specifically his work with self-driving cars.
- How Google's Self-Driving Car Works. *IEEE Spectrum*



http://robots.stanford.edu/

Computer vision



- **Andrew Ng**, Director, Stanford Artificial Intelligence Lab, Computer Science Department Stanford University
- Google研究人员利用一千台电脑的1.6万核处理器组建了一个机器学习神经网络，网络包含超过10亿条链接。他们用从YouTube视频中提取出的1000万幅200x200缩略图训练神经网络，目的是寻找和识别猫”。神经网络利用无指导的机器学习去识别图像特征。
- **Building High-Level Features using Large Scale Unsupervised Learning**. Quoc V. Le, Marc'Aurelio Ranzato, Rajat Monga, Matthieu Devin, Kai Chen, Greg S. Corrado, Jeffrey Dean and Andrew Y. Ng. In*Proceedings of the Twenty-Ninth International Conference on Machine Learning*, 2012.

http://ai.stanford.edu/~ang/

Google的猫脸识别：人工智能的新突破

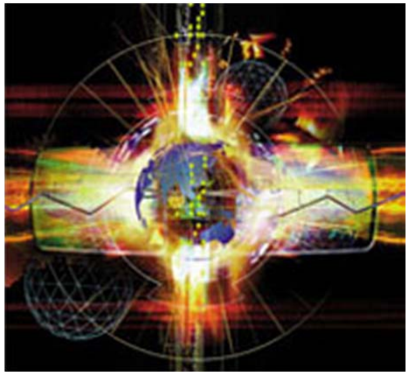
Machine translation

- Even Google Translate is based on Machine Learning techniques, which might sound surprising.
- Intuitively one might think, that language experts crafted tons of rules and translated millions of words. This work would then serve as a basis for the translation service. Far from it!
- In reality, Google uses publicly available data on the Internet which already exists in a translated version. For example all the European Union documents need to be translated by experts into the several languages of the European countries. Google grabs those documents and many many more and allows a Machine Learning algorithm to crunch this massive amount of data to detect patterns and learn translations. This approach makes it possible to translate texts to **58 different languages** in a decent quality.

Others

- 大数据的挖掘越来越多的渗透到生活的方方面面:奥巴马竞选团队利用数据分析筹款;成功预测2012年50个州选举结果的Nate Silver; 微软研究院也称成功预测了大部分奥斯卡奖项
- **Netflix**是一家美国公司，在美国、加拿大提供互联网随选流媒体播放.在美国有2700万订阅用户，在全世界则有3300万.
- 2006年10月，Netflix建立了**Netflix Prize竞赛**，并对外发布了一个电影评分（评分为**1, …, 5**的整数）数据集, **480,189**个用户对**17,770**部电影的评分----→Matrix completion
- Netflix 几乎比所有人都清楚大家喜欢看什么。它已经知道用户很喜欢 Fincher（社交网络、七宗罪的导演），也知道 Spacey 主演的片子表现都不错，还知道英剧版的《纸牌屋》很受欢迎
- 在不久的将来???

SEARCH ENGINES
SOCIAL NETWORKS
BLOGGING
CONTENT SITES
LOCATION BASED SERVICES
REVIEWS & RECOMMENDS
DIRECTORIES
SOCIAL COMMERCE
Google
AOL
Ask.com
YAHOO
bing
facebook
twitter
Zvents
twitpic
my____
tumblr
Ning
vimeo
YouTube
USTREAM
tumblr
Google hotpot
TypePad
Blogger
WordPress
slideshare
metacafe
flickr
Picasa
shopkick
foursquare
Bizzy
scvngr
Gowalla
Facebook Places
BBB
TrustLink
Citysearch
Google hotpot
insiderpages
Bizzy
tripadvisor
yelp
Angie's list
MagicYellow
superpages.com
YAHOO! LOCAL
Google places
Local.com
yp.com
GROUPON
ZAGAT
yipit
TIPPR
BuyWithMe
DealOn
livingsocial
DATA EXPLOSION
©2011 Miles BrandDNA

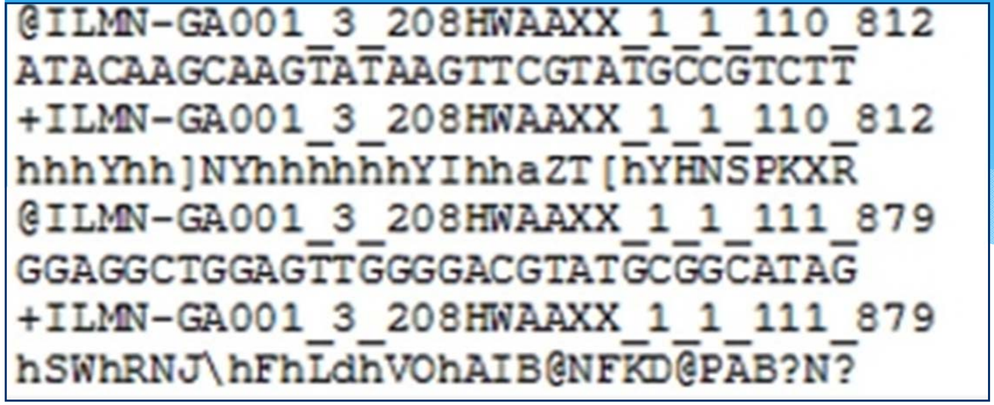


Data Exploration is everywhere
What’s we should do for biological data?

生物医学大数据？

- Sergey Brin的妻子是一位生物技术博士，且于1996年于耶鲁大学荣获生物学理学学士;之后她将重点放在医学信息学上，并与Sergey Brin共同希望能找出新方法来改善这方面的课题。他们俩亦深深影响医学信息学，并带领着人类基因组计划。在一次Google的时代精神会议上，Sergey Brin提出他希望总有一天人们都能了解自己的基因排序，借此来帮助医生、患者、研究人员分析这些数据，来解决身体上的问题。
- 开创一种完全不同的医学科学研究方法。与大多数医学研究相同的帕金森病研究依赖于经典的研究方法，即提出假设、进行分析、同行/同级评审(peerreview)和发表研究结果。但有着强大计算处理能力和多的令人震惊的数据的支撑，Sergey Brin开创了一种不同的研究方式。
- Sergey Brin：美国式的成功，《领袖人物》2012年3月

NGS: Next generation Sequencing data: RNA-seq & Chip-seq



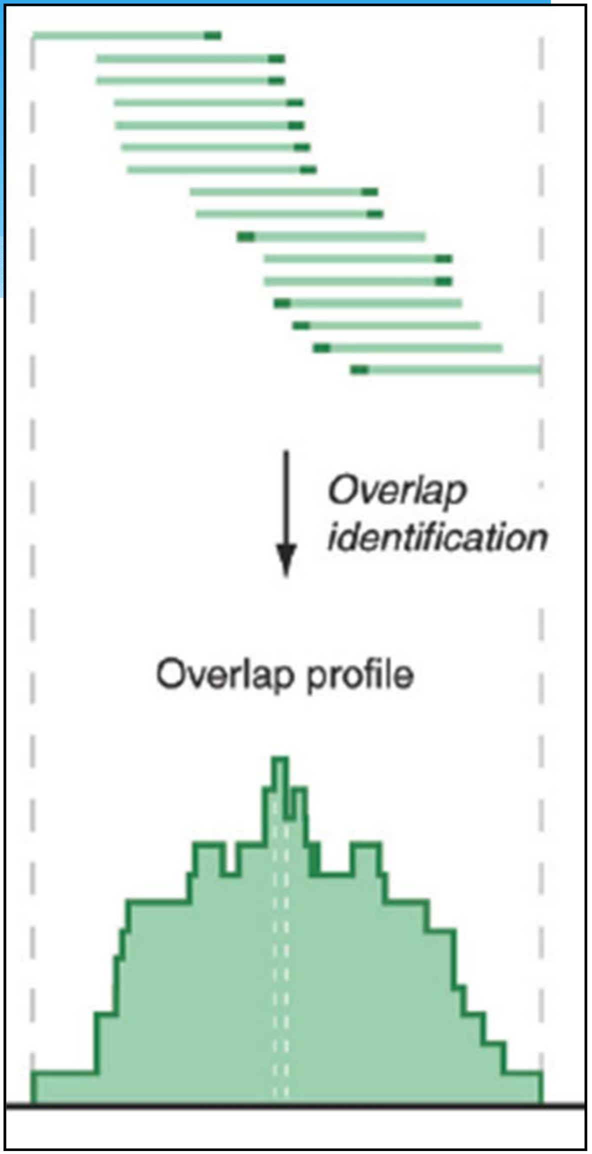
```
@ILMN-GA001_3_208HWAAXX_1_1_110_812
ATACAAGCAAGTATAAGTTCGTATGCCGTCTT
+ILMN-GA001_3_208HWAAXX_1_1_110_812
hhhYhh]NYhhhhhhYIhhaZT[hYHNSPKXR
@ILMN-GA001_3_208HWAAXX_1_1_111_879
GGAGGCTGGAGTTGGGGACGTATGCGGCATAG
+ILMN-GA001_3_208HWAAXX_1_1_111_879
hSWhRNJ\hFhLdhVOhAIB@NFKD@PAB?N?
```

Reads (fasta)
+ quality scores (fastq)
+ mapping (BAM)

Reads => Signal (Intermediate file)

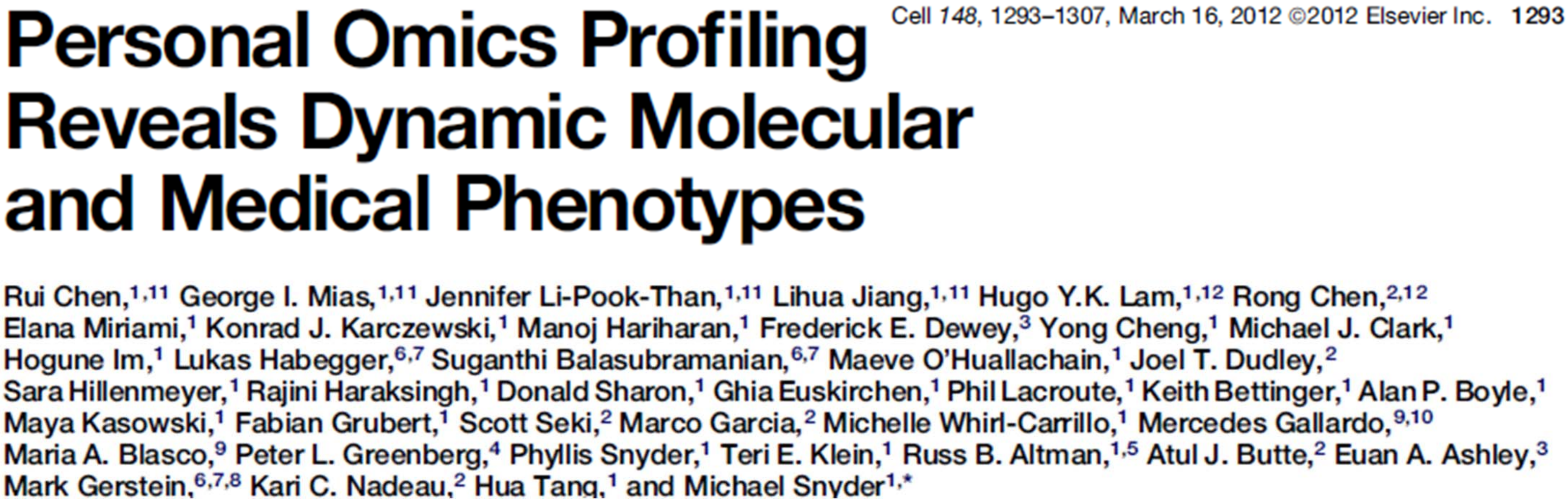
Accumulating @ >1 Pbp/yr (currently),
~20% of tot. HiSeq output

Should we store & share high or low level data ?



[PLOS CB 4:e1000158]

Personal omics is coming



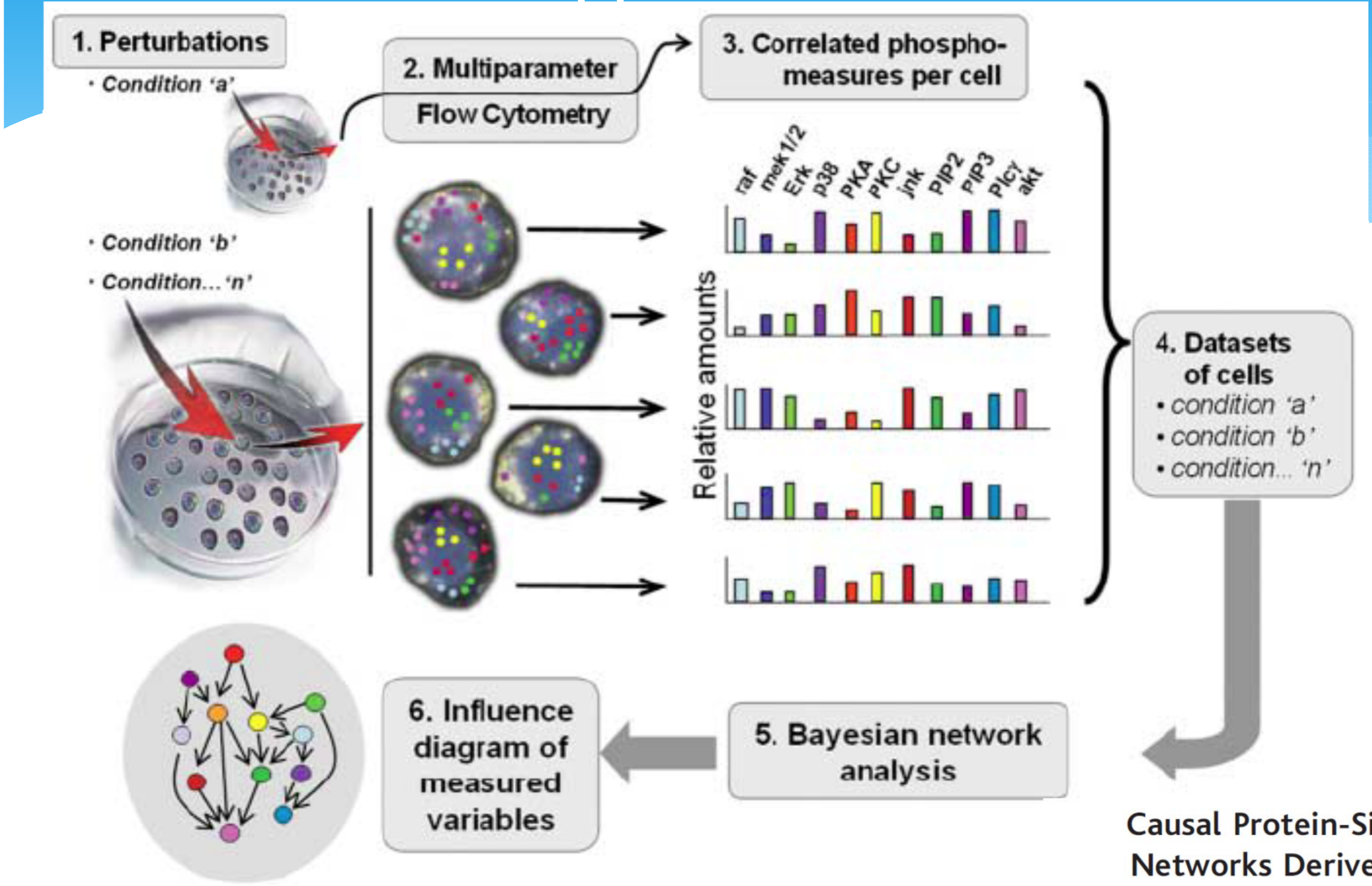
Cell *148*, 1293–1307, March 16, 2012 ©2012 Elsevier Inc. 1293

Personal Omics Profiling Reveals Dynamic Molecular and Medical Phenotypes

Rui Chen,[1,11] George I. Mias,[1,11] Jennifer Li-Pook-Than,[1,11] Lihua Jiang,[1,11] Hugo Y.K. Lam,[1,12] Rong Chen,[2,12] Elana Miriami,[1] Konrad J. Karczewski,[1] Manoj Hariharan,[1] Frederick E. Dewey,[3] Yong Cheng,[1] Michael J. Clark,[1] Hogune Im,[1] Lukas Habegger,[6,7] Suganthi Balasubramanian,[6,7] Maeve O'Huallachain,[1] Joel T. Dudley,[2] Sara Hillenmeyer,[1] Rajini Haraksingh,[1] Donald Sharon,[1] Ghia Euskirchen,[1] Phil Lacroute,[1] Keith Bettinger,[1] Alan P. Boyle,[1] Maya Kasowski,[1] Fabian Grubert,[1] Scott Seki,[2] Marco Garcia,[2] Michelle Whirl-Carrillo,[1] Mercedes Gallardo,[9,10] Maria A. Blasco,[9] Peter L. Greenberg,[4] Phyllis Snyder,[1] Teri E. Klein,[1] Russ B. Altman,[1,5] Atul J. Butte,[2] Euan A. Ashley,[3] Mark Gerstein,[6,7,8] Kari C. Nadeau,[2] Hua Tang,[1] and Michael Snyder[1,*]

Personalized medicine is expected to benefit from combining genomic information with regular monitoring of physiological states by multiple highthroughput methods. Here, we present an integrative personal omics profile (iPOP), an analysis that combines genomic, transcriptomic, proteomic, metabolomic,and autoantibody profiles from a single individual over a 14 month period. Our iPOP analysis revealed various medical risks, including type 2 diabetes.

Single cell data

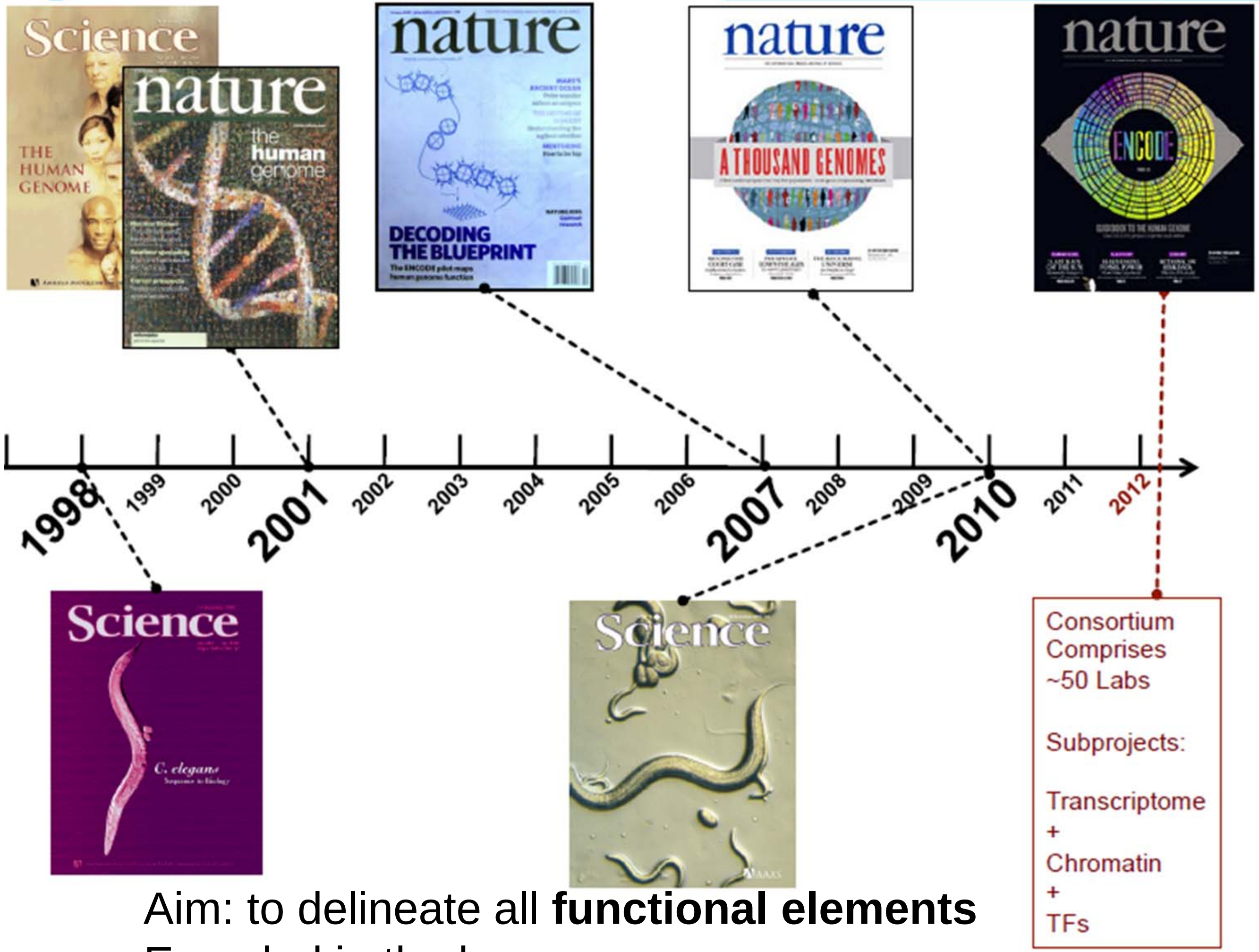


22 APRIL 2005 VOL 308 SCIENCE

Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data

Karen Sachs,[1]* Omar Perez,[2]* Dana Pe'er,[3]*
Douglas A. Lauffenburger,[1]† Garry P. Nolan[2]†

The Encyclopedia of DNA Elements (ENCODE)



Consortium Comprises ~50 Labs

Subprojects:

Transcriptome
+
Chromatin
+
TFs

Aim: to delineate all **functional elements**
Encoded in the human genome.

Gene expression data
CAGE data
DNA methylation data
TF binding data
Histone modification data
CHIA - PET data
Dnase hypersensivity data
Protein expression data
…..

A total of >40 papers
In 2012

The Cancer Genome Atlas (TCGA)

- 针对复杂疾病这一危害人类健康的主要杀手, 由美国国家癌症和肿瘤研究所(NCI)和国家人类基因组研究所(NHGRI)联合进行了癌症和肿瘤基因图谱(The Cancer Genome Atlas，TCGA)计划.
- 采用大规模的基因组测序，从转录、序列、表观修饰等不同层次采集数据，将50种肿瘤的基因组变异图谱绘制出来，进行计算分析，旨在找到所有致癌和抑癌基因的微小变异，了解癌细胞发生、发展的机制。
- 除了分子层面数据，一些表型层面的数据包括临床诊断、药物干扰、疾病间关系也可以公开获取。
- 对这些数据进行数学建模研究，探索复杂疾病早期诊断及治疗方法，是本世纪生物医学研究的重大挑战，同时海量、高维、复杂生物数据对数学工具也提出了更高的要求!

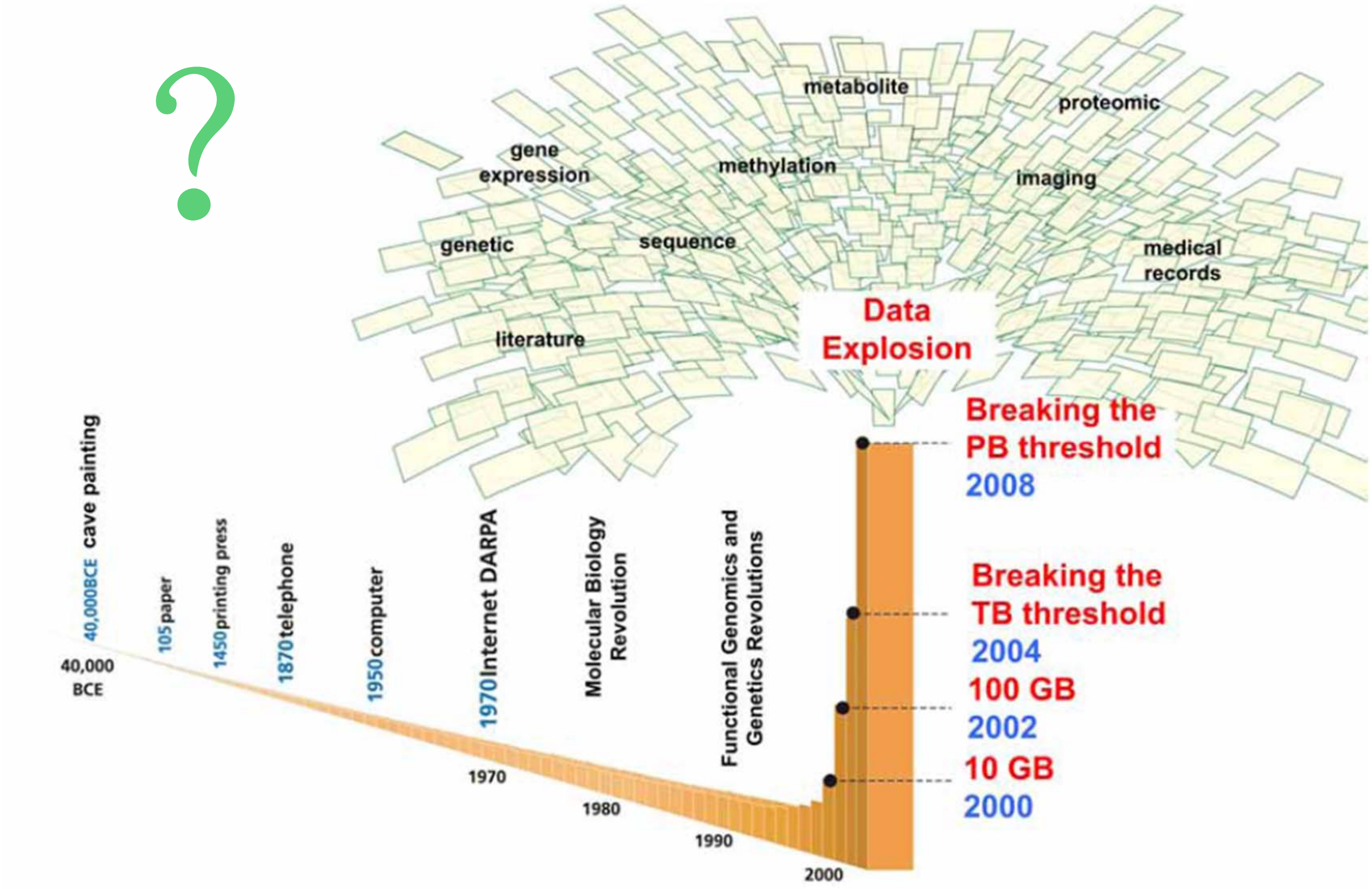
本质困难

- **生物数据的复杂性：** 维数高，高达数千、上万维；结构复杂，具有高度的非线性性和非一致性；噪声强而信号弱；来源不同，在时间和空间上呈现多尺度。分析集成这样的数据，需要探索和发展新的数学建模和计算方法。
- **生物系统的复杂性：** 数学建模中要充分考虑生物系统中的相互作用、非线性、调控、动态行为等特征。

Curse of dimensionality（维度灾难）

- The **curse of dimensionality** refers to various phenomena that arise when analyzing and organizing high-dimensional spaces (often with hundreds or thousands of dimensions) that do not occur in low-dimensional settings such as the physical space commonly modeled with just three dimensions.
- There are multiple phenomena referred to by this name in domains such as sampling, combinatorics, machine learning and data mining. The common theme of these problems is that when the dimensionality increases, the volume of the space increases so fast that the available data becomes sparse.
- The term *curse of dimensionality* was coined by Richard E. Bellman when considering problems in dynamic optimization

?


metabolite
proteomic
gene expression
methylation
imaging
genetic
sequence
medical records
literature
Data Explosion
Breaking the PB threshold
2008
40,000BCE cave painting
105 paper
1450 printing press
1870 telephone
1950 computer
1970 Internet DARPA
Molecular Biology Revolution
Functional Genomics and Genetics Revolutions
Breaking the TB threshold
2004
100 GB
2002
10 GB
2000
40,000 BCE
1970
1980
1990
2000

生物数据分析（挖掘）的三个层次

- □ 计数(Counting)
- □ 相关性分析(Correlation)
- □ 集成(Integration)

由易到难,由浅到深

从无模型到基于模型的分析

Counting is important

- The first step to deal with first-hand data
- The way to count really matters (ask a good question and count in a smart way)

仰望星空

“一个民族有一些关注天空的人，他们才有希望；一个民族只是关心脚下的事情，那是没有未来的。”　——温家宝

一个仰望星空的人第谷：在天文历史上以观测精密而著称，是一个善于“看”的人。清醒地知道要认识行星运动的规律，积累高度精确测量数据的重要性，并身体力行地测出了大量的原始精确的数据。

第谷给开普勒留下的，是他**20**年来观测的大量星空变化数据，还有一句忠告：“一定要尊重观测事实。”第谷便是开普勒仰望星空的第一双眼睛。（这位勤奋的天文学家也是物理学家，却无法仰望星空。因为他从小就损坏视力，在靠肉眼观测的年代，他无法成为一个天文观测学家。）

Good examples for counting

- 元素周期表 (Dmitri Mendeleev 1869),

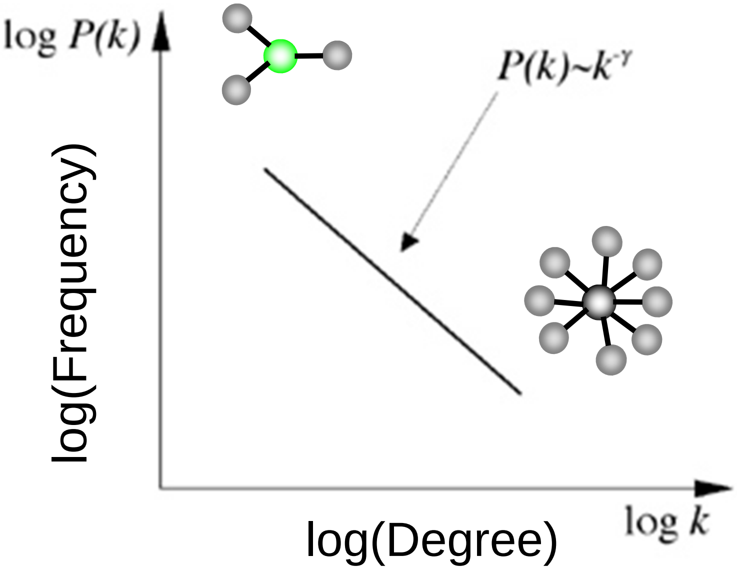
  他将当时已知的63种元素依原子量大小并以表的形式排列，把有相似化学性质的元素放在同一行，就是元素周期表的雏形。利用周期表，门捷列夫成功的预测当时尚未发现的元素的特性(镓、钪、锗)。直至2010年4月，周期表中共有118种已经发现的元素

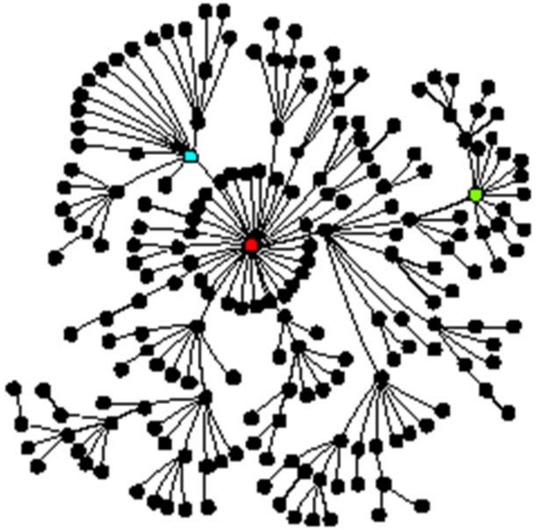
- 进化论 (Charles Robert Darwin 1809),

  查尔斯·罗伯特·达尔文，英国生物学家，进化论的奠基人。曾乘贝格尔号舰作了历时5 年的环球航行， 对动植物和地质结构等进行了大量的观察和采集。达尔文因此领导了人类历史上最为伟大、影响最为深远的一场理性革命。

Recent example: Scale-free networks

Power-law distribution





Hubs dictate the structure of the network

[Barabasi]

少数节点连接数超乎异常的事实。

Correlation

- Low level correlation

Two variables (correlation and causal relationship)

- Middle level correlation

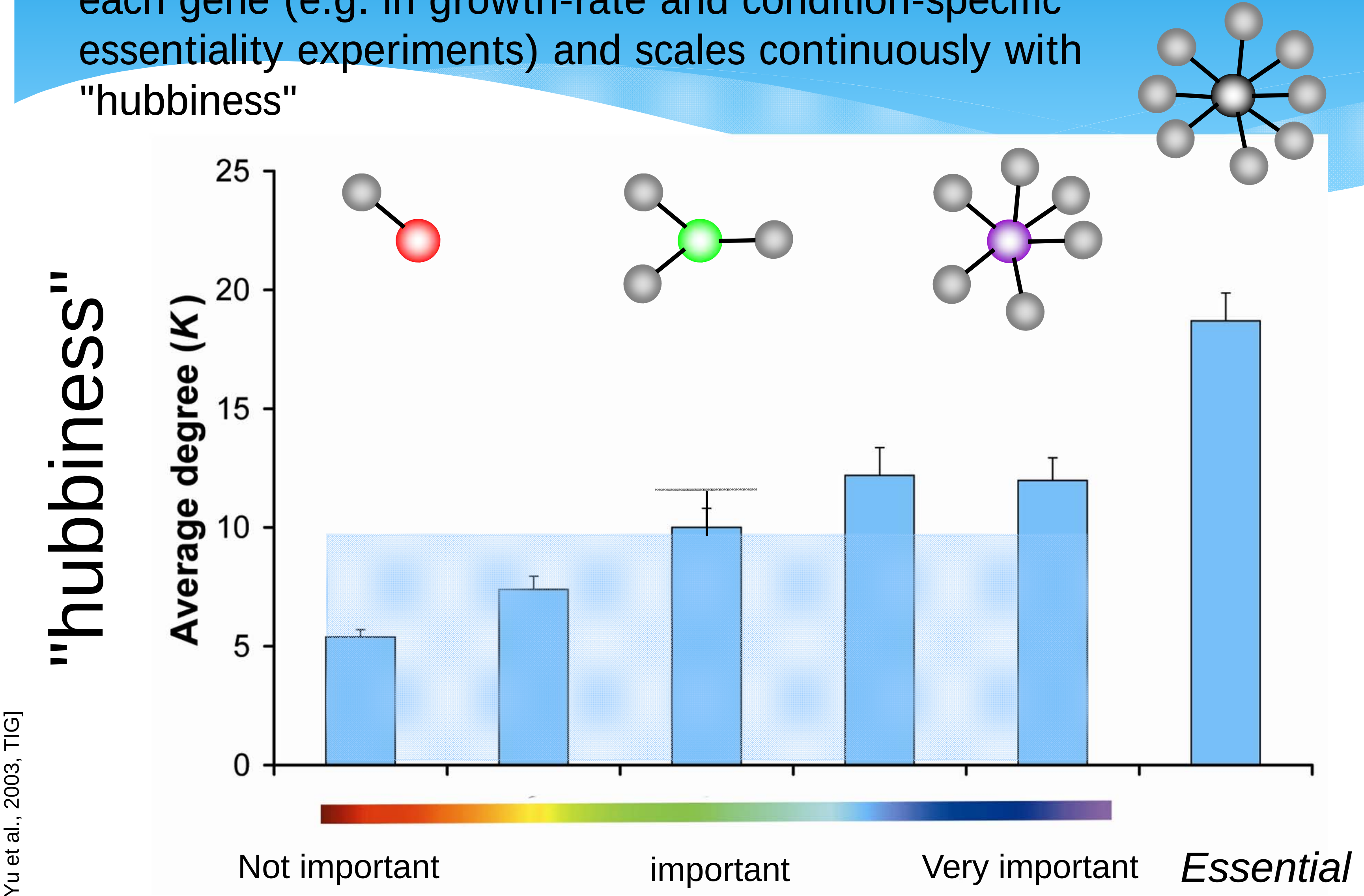
Two data sources

- High level correlation

Two disciplines 放之四海而皆准

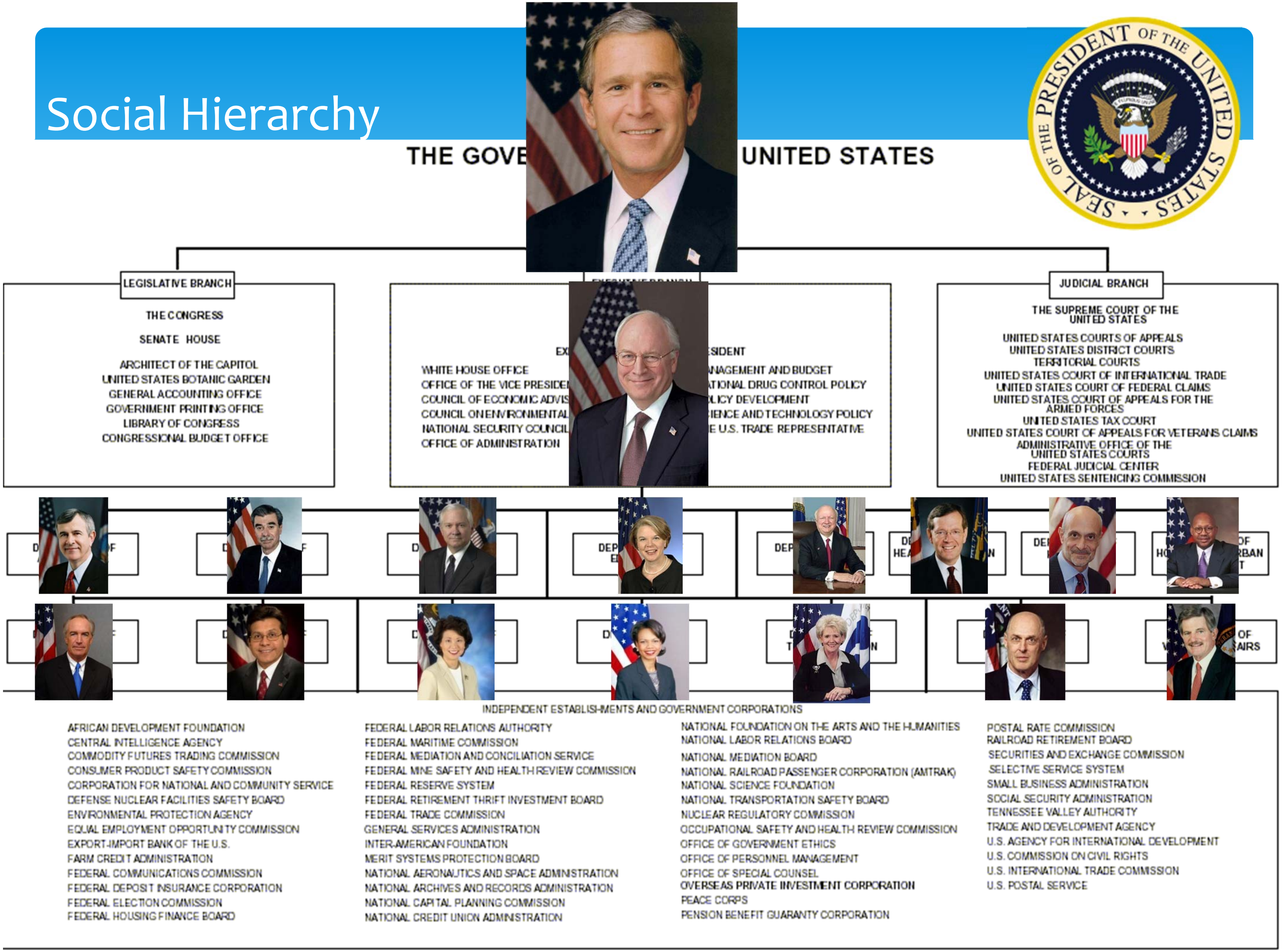
Degree correlates with essentiality

Marginal essentiality measures relative importance of each gene (e.g. in growth-rate and condition-specific essentiality experiments) and scales continuously with "hubbiness"



[Yu et al., 2003, TIG]

Social Hierarchy



THE GOVE UNITED STATES

LEGISLATIVE BRANCH

THE CONGRESS

SENATE HOUSE

ARCHITECT OF THE CAPITOL
UNITED STATES BOTANIC GARDEN
GENERAL ACCOUNTING OFFICE
GOVERNMENT PRINTING OFFICE
LIBRARY OF CONGRESS
CONGRESSIONAL BUDGET OFFICE

EXECUTIVE BRANCH

EX SIDENT

WHITE HOUSE OFFICE
OFFICE OF THE VICE PRESIDE
COUNCIL OF ECONOMIC ADVIS
COUNCIL ON ENVIRONMENTAL
NATIONAL SECURITY COUNCIL
OFFICE OF ADMINISTRATION

NAGEMENT AND BUDGET
TIONAL DRUG CONTROL POLICY
LICY DEVELOPMENT
IENCE AND TECHNOLOGY POLICY
E U.S. TRADE REPRESENTATIVE

JUDICIAL BRANCH

THE SUPREME COURT OF THE UNITED STATES

UNITED STATES COURTS OF APPEALS
UNITED STATES DISTRICT COURTS
TERRITORIAL COURTS
UNITED STATES COURT OF INTERNATIONAL TRADE
UNITED STATES COURT OF FEDERAL CLAIMS
UNITED STATES COURT OF APPEALS FOR THE ARMED FORCES
UNITED STATES TAX COURT
UNITED STATES COURT OF APPEALS FOR VETERANS CLAIMS
ADMINISTRATIVE OFFICE OF THE UNITED STATES COURTS
FEDERAL JUDICIAL CENTER
UNITED STATES SENTENCING COMMISSION

D OF
D F
D
DEP E
DEP
DE HEA N
DEI
HO OF RBAN T

D F
D F
D F
D
D T N
D
V OF AIRS

INDEPENDENT ESTABLISHMENTS AND GOVERNMENT CORPORATIONS

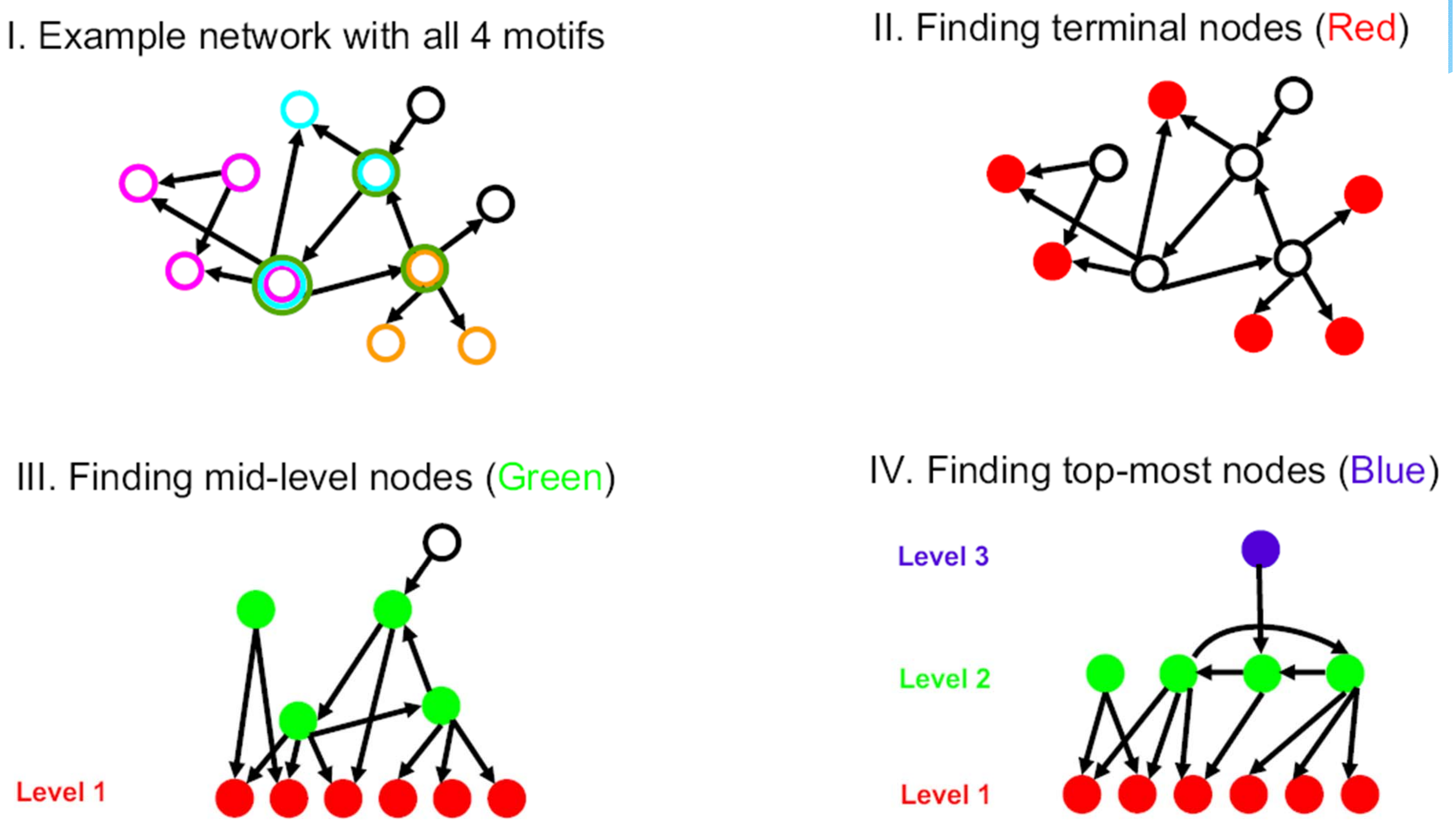
AFRICAN DEVELOPMENT FOUNDATION
CENTRAL INTELLIGENCE AGENCY
COMMODITY FUTURES TRADING COMMISSION
CONSUMER PRODUCT SAFETY COMMISSION
CORPORATION FOR NATIONAL AND COMMUNITY SERVICE
DEFENSE NUCLEAR FACILITIES SAFETY BOARD
ENVIRONMENTAL PROTECTION AGENCY
EQUAL EMPLOYMENT OPPORTUNITY COMMISSION
EXPORT-IMPORT BANK OF THE U.S.
FARM CREDIT ADMINISTRATION
FEDERAL COMMUNICATIONS COMMISSION
FEDERAL DEPOSIT INSURANCE CORPORATION
FEDERAL ELECTION COMMISSION
FEDERAL HOUSING FINANCE BOARD

FEDERAL LABOR RELATIONS AUTHORITY
FEDERAL MARITIME COMMISSION
FEDERAL MEDIATION AND CONCILIATION SERVICE
FEDERAL MINE SAFETY AND HEALTH REVIEW COMMISSION
FEDERAL RESERVE SYSTEM
FEDERAL RETIREMENT THRIFT INVESTMENT BOARD
FEDERAL TRADE COMMISSION
GENERAL SERVICES ADMINISTRATION
INTER-AMERICAN FOUNDATION
MERIT SYSTEMS PROTECTION BOARD
NATIONAL AERONAUTICS AND SPACE ADMINISTRATION
NATIONAL ARCHIVES AND RECORDS ADMINISTRATION
NATIONAL CAPITAL PLANNING COMMISSION
NATIONAL CREDIT UNION ADMINISTRATION

NATIONAL FOUNDATION ON THE ARTS AND THE HUMANITIES
NATIONAL LABOR RELATIONS BOARD
NATIONAL MEDIATION BOARD
NATIONAL RAILROAD PASSENGER CORPORATION (AMTRAK)
NATIONAL SCIENCE FOUNDATION
NATIONAL TRANSPORTATION SAFETY BOARD
NUCLEAR REGULATORY COMMISSION
OCCUPATIONAL SAFETY AND HEALTH REVIEW COMMISSION
OFFICE OF GOVERNMENT ETHICS
OFFICE OF PERSONNEL MANAGEMENT
OFFICE OF SPECIAL COUNSEL
OVERSEAS PRIVATE INVESTMENT CORPORATION
PEACE CORPS
PENSION BENEFIT GUARANTY CORPORATION

POSTAL RATE COMMISSION
RAILROAD RETIREMENT BOARD
SECURITIES AND EXCHANGE COMMISSION
SELECTIVE SERVICE SYSTEM
SMALL BUSINESS ADMINISTRATION
SOCIAL SECURITY ADMINISTRATION
TENNESSEE VALLEY AUTHORITY
TRADE AND DEVELOPMENT AGENCY
U.S. AGENCY FOR INTERNATIONAL DEVELOPMENT
U.S. COMMISSION ON CIVIL RIGHTS
U.S. INTERNATIONAL TRADE COMMISSION
U.S. POSTAL SERVICE

Determination of "Level" in Regulatory Network Hierarchy with Breadth-first Search



I. Example network with all 4 motifs

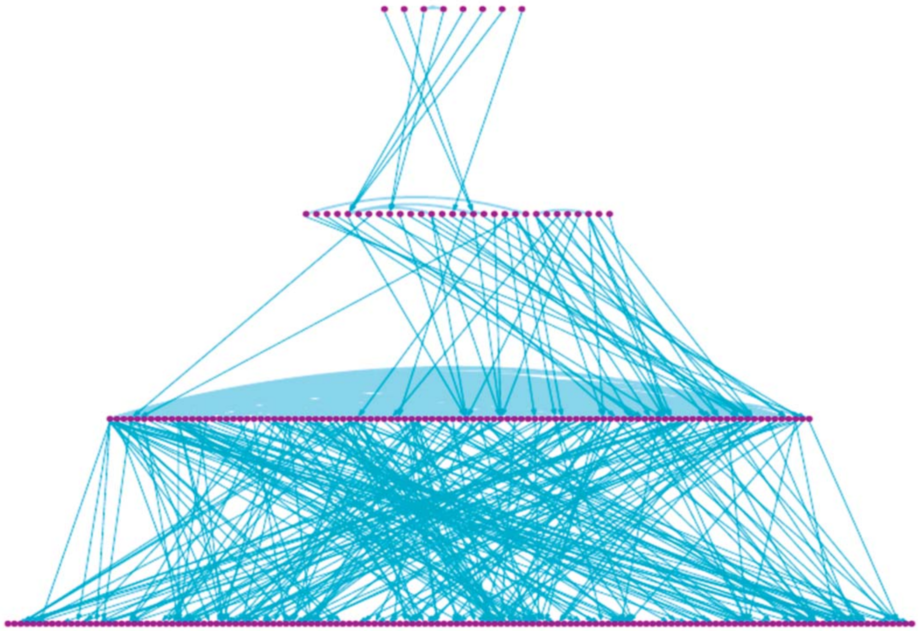
II. Finding terminal nodes (Red)

III. Finding mid-level nodes (Green)

IV. Finding top-most nodes (Blue)

[Yu et al., PNAS (2006)]

Regulatory Networks have similar hierarchical structures

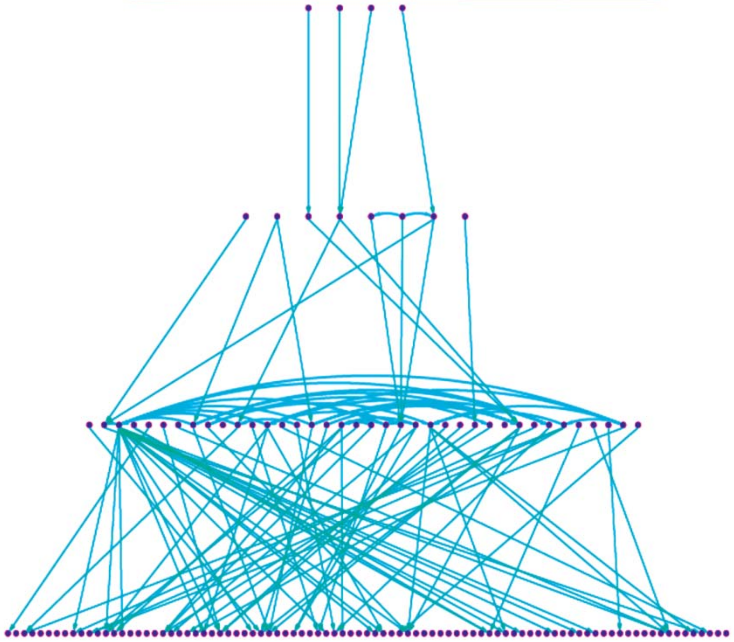


1

2

3

4

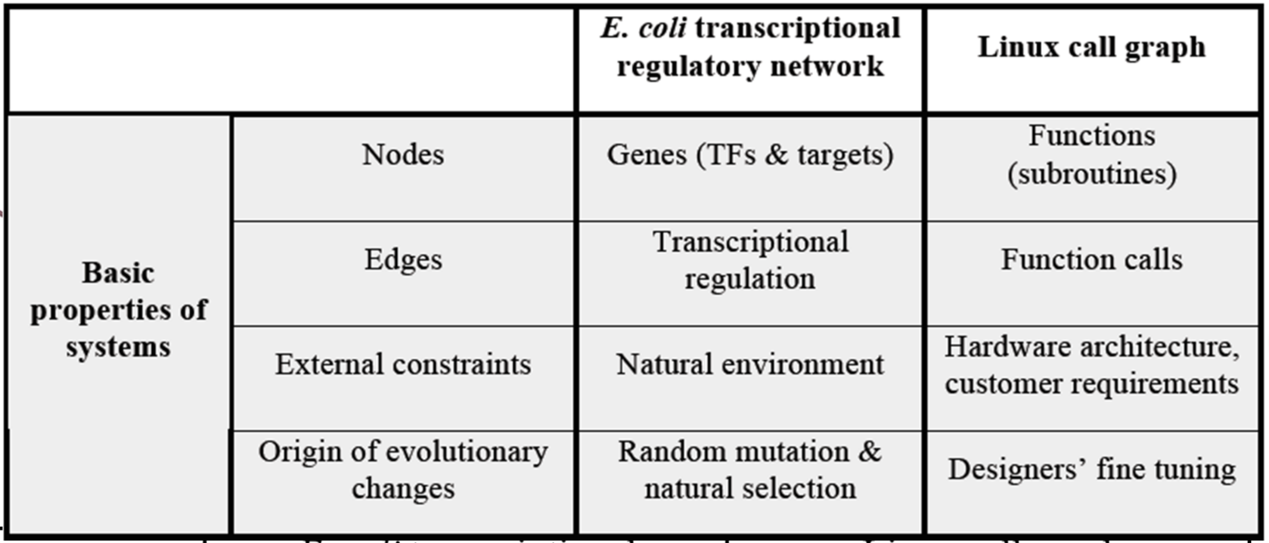


S. cerevisiae

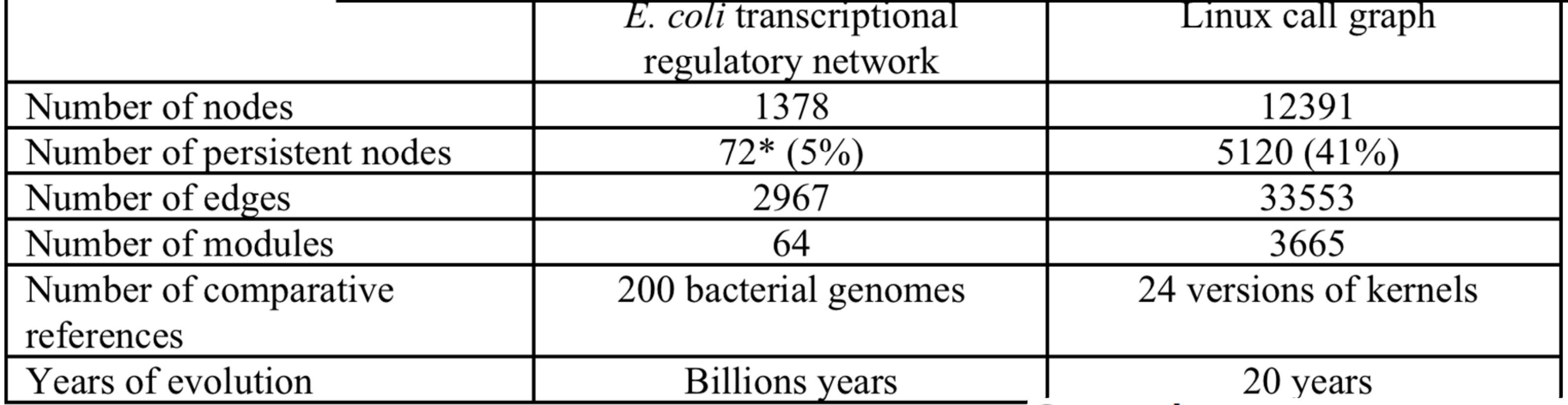
E. coli

[Yu *et al.*, *Proc Natl Acad Sci U S A* (2006)]

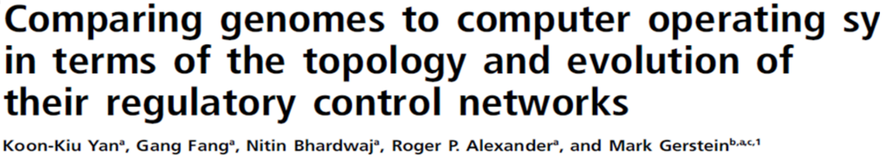
E. Coli Transcriptional regulatory networks vs Linux kernel call graph



		E. coli transcriptional regulatory network	Linux call graph
Basic properties of systems	Nodes	Genes (TFs & targets)	Functions (subroutines)
	Edges	Transcriptional regulation	Function calls
	External constraints	Natural environment	Hardware architecture, customer requirements
	Origin of evolutionary changes	Random mutation & natural selection	Designers' fine tuning



	E. coli transcriptional regulatory network	Linux call graph
Number of nodes	1378	12391
Number of persistent nodes	72* (5%)	5120 (41%)
Number of edges	2967	33553
Number of modules	64	3665
Number of comparative references	200 bacterial genomes	24 versions of kernels
Years of evolution	Billions years	20 years



Comparing genomes to computer operating sy in terms of the topology and evolution of their regulatory control networks

Koon-Kiu Yan[a], Gang Fang[a], Nitin Bhardwaj[a], Roger P. Alexander[a], and Mark Gerstein[b,a,c,1]

[Yan et al., PNAS (2010), in press]

集成=数学建模

- □ 计数(Counting)
- □ 相关性分析(Correlation)
- □ 集成(Integration)

由易到难,由浅到深

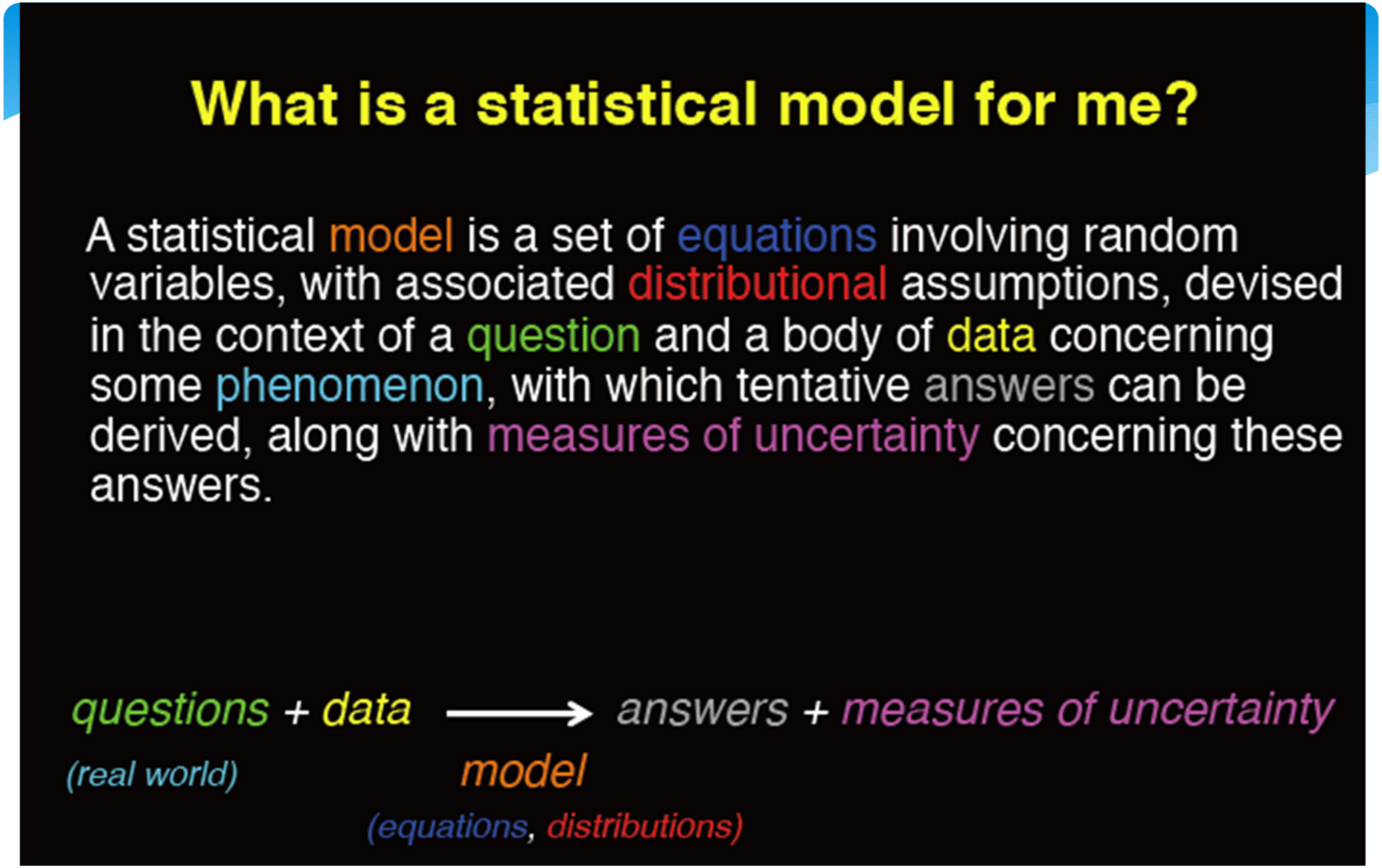
基于模型的分析

什么是建模？

- 模型的定义：科学研究中对事物的合理简化，例如牛顿力学模型、氢原子的玻尔模型（Wiki）

- 第一步是对观测数据建立一个模型。
- 第二步则是使用这个模型来推测未知现象发生的概率。

- Data → Model
- Model → Prediction



What is a statistical model for me?

A statistical model is a set of equations involving random variables, with associated distributional assumptions, devised in the context of a question and a body of data concerning some phenomenon, with which tentative answers can be derived, along with measures of uncertainty concerning these answers.

questions + *data* ⟶ *answers* + *measures of uncertainty*

(real world)

model

(equations, distributions)

Department of Statistics，University of California at Berkeley

建模本质上是找出最能解释数据的模型(最优化)

- Max P(D|M) maximum likelihood estimation（MLE）

- 贝叶斯（**Bayesian**）公式

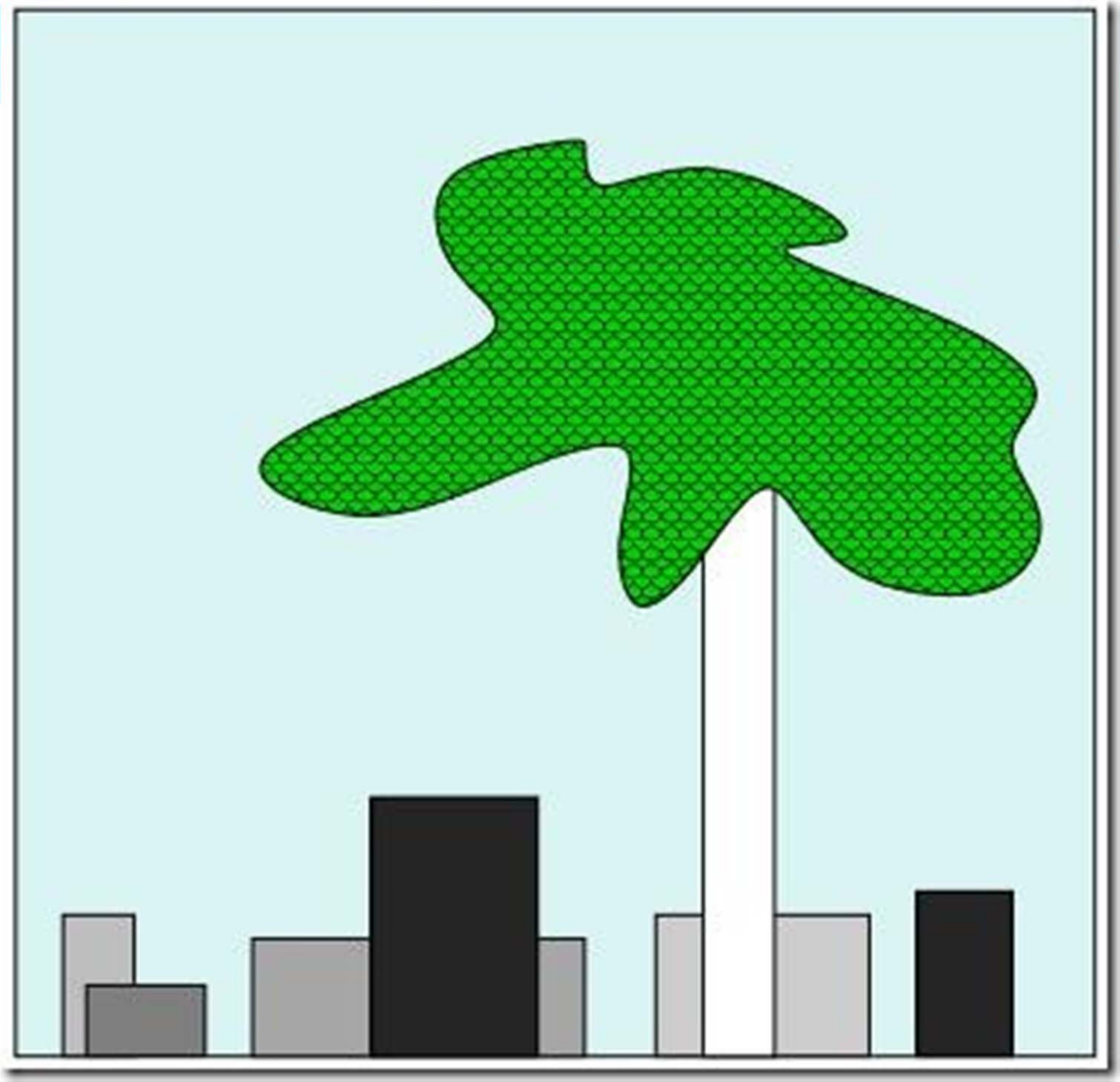
P(M|D)= P(D|M)P(M)/P(D)

Maximmum a posterior （MAP）

- 揭示了数据与模型的复杂关系

Overfitting

- When estimating parameters for a model from a limited amount of data, there is a danger of overfitting, which means that the model becomes very well adapted to the training data, but it will not generalise well to new data.
- Observing for instance that three flips of a coin[tail, tail, tail] would lead to the maximum likelihood estimate that the propobality of head is 0 and that of tail is 1.
- We can use prior knowledge to constrain the estimates. For example pseudocounts can be introduced



Observed Data

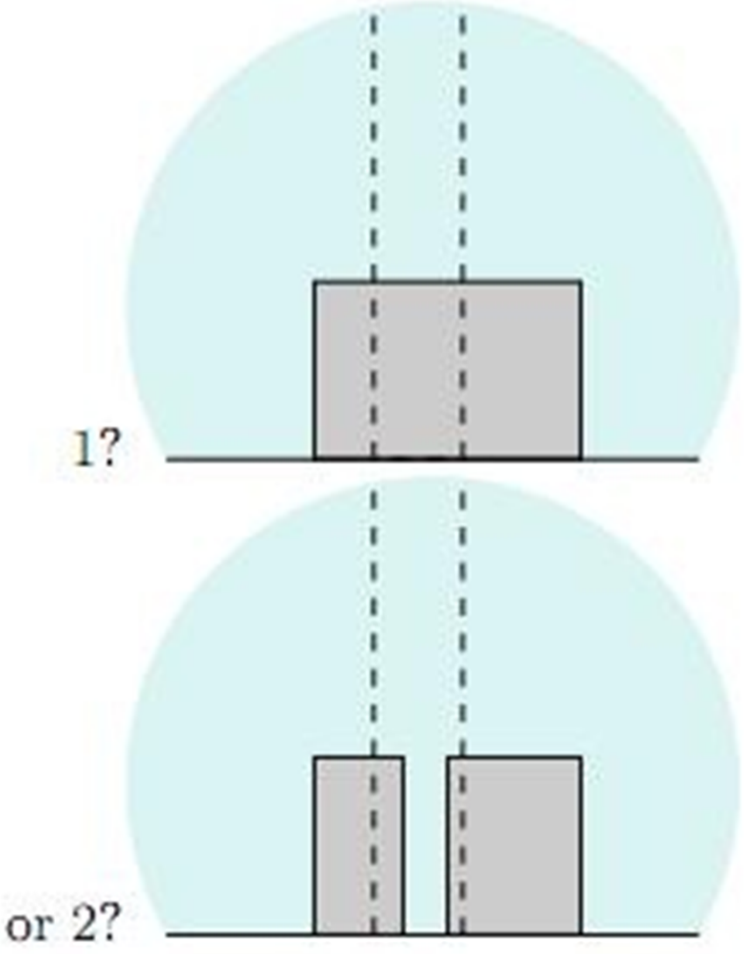


Figure 28.2. How many boxes are behind the tree?

Model #1

Model #2

MacKay 的著作 《Information Theory : Inference and Learning Algorithms》

奥卡姆剃刀

- 贝叶斯模型比较理论：
- $P(M \mid D) \propto P(M) * P(D \mid M)$
- 与信息论的关联：**最小描述长度原则**
- $\ln P(M \mid D) \propto \ln P(M) + \ln P(D \mid M)$

Also known as Occam’s Razor

When the solution is simple, God is answering.
Albert Einstein

Make everything as simple as possible, but not simpler.
Albert Einstein

One important note

- All models are wrong, but some are useful
  George E.P. Box
- **George Edward Pelham Box** (born 18 October 1919) is a statistician, who has made important contributions in the areas of quality control, time-series analysis, design of experiments, andBayesian inference.
- Box famously wrote that "essentially, all models are wrong, but some are useful" in his book on response surface methodology with Norman R. Draper.
- 美国威斯康星大学麦迪逊分校R. A. Fisher统计名誉教授。他是美国人文和自然科学研究院院士，美国统计学会S. S. Wilks纪念奖章、美国质量协会Shewhart奖章和英国皇家统计协会Guy银奖的获得者

The End of Theory: The Data Deluge Makes the Scientific Method Obsolete
By Chris Anderson

WIRED MAGAZINE: 16.07

Example

- Tennis playing
- The trace?
- Theory or data?

Data driven

- 谷歌的奠基哲学就是“我们不知道为什么这张网页比那张网页好”：只要引入链接的统计数据说明它好就行了，并不需要语义上或者是因果关系的分析。

Google changed the situation

- 2011年三月的O‘Reilly 前沿技术会议（ O’Reilly Emerging Technology Conference）上，Peter Norvig（谷歌的研究指导）对 George Box的座右铭进行了更新：所有模型都是错误的，愈加地，你能在没有模型的情况下成功。（"All models are wrong, and increasingly you can succeed without them."）
- O’Reilly Media 通过图书、杂志、在线服务、调查研究和会议等方式传播创新知识。自1978年开始，O’Reilly 一直都是前沿发展的见证者和推动者。

小结 Take home message

- 生物信息学与计算系统生物学本质上是由数据驱动的
- 复杂生物数据的建模
- 模型与数据的关系?

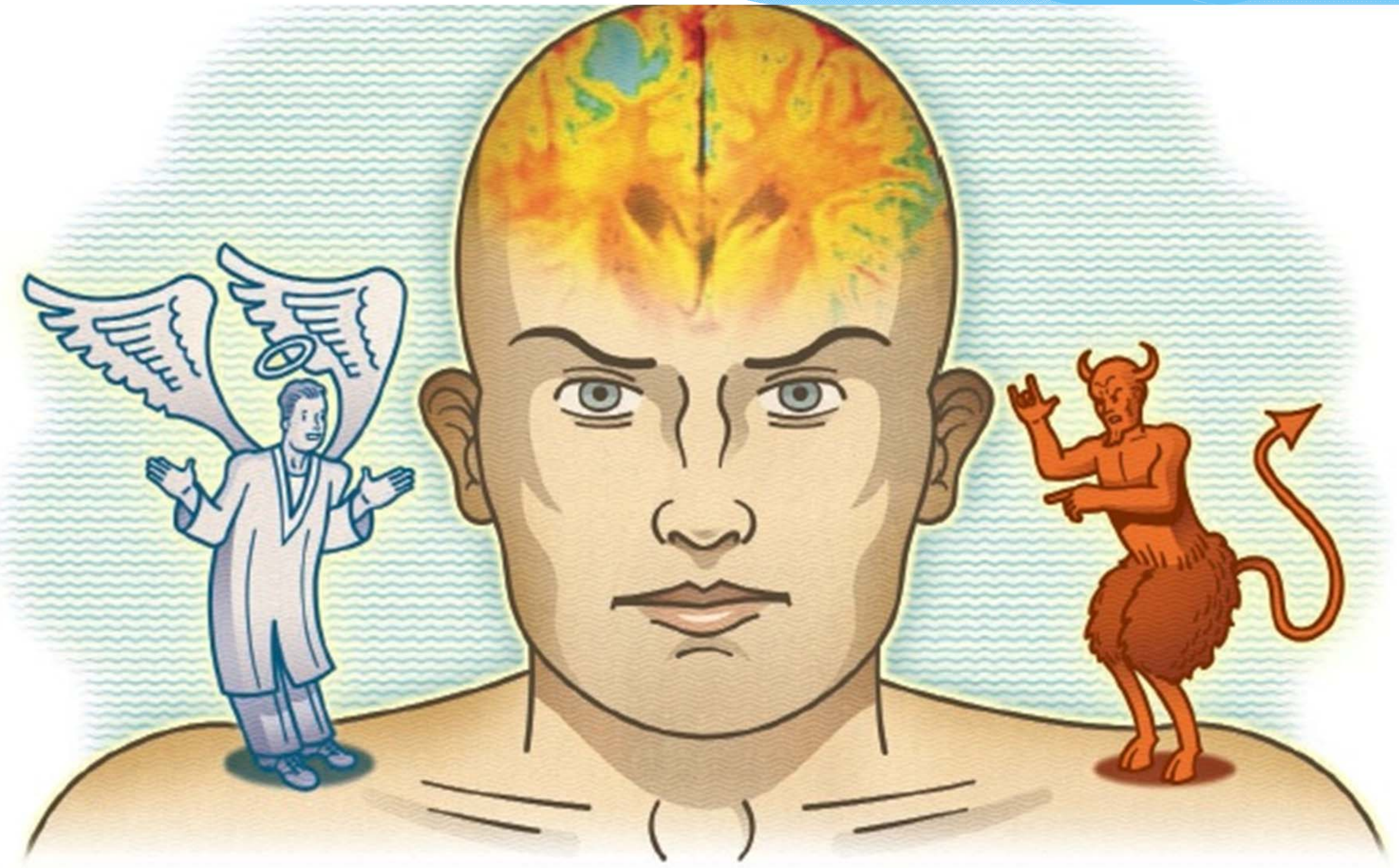
Procedure---Systems Biology

- System Perturbation
- Generating of comprehensive global data
- Identification of key molecules
- Network modelling
- Generation of hypotheses
- Validation of hypotheses

Taste the science

- 科学家“假设、模型、检验”的方法

- Several examples to taste the style



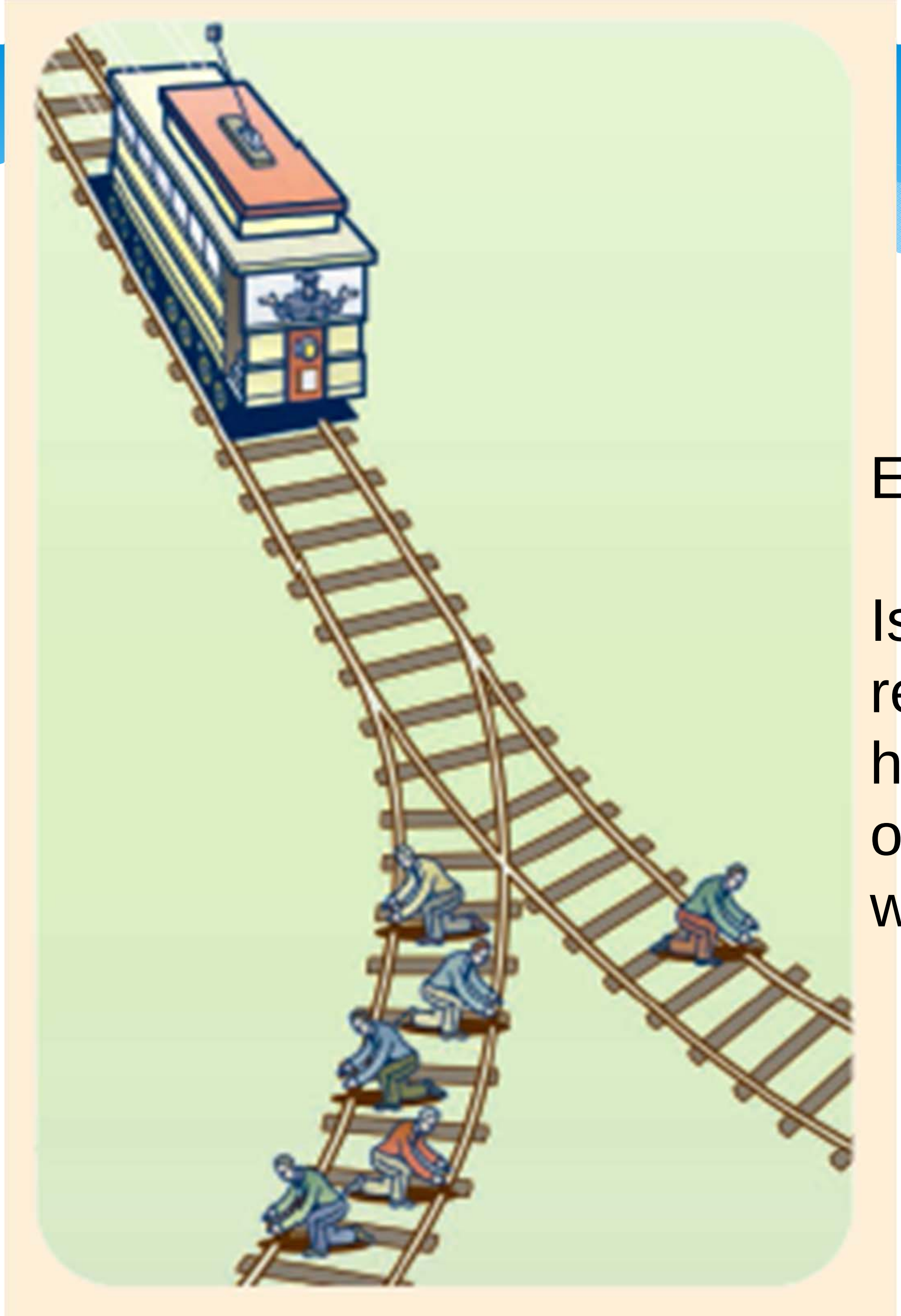
The Roots of Morality

An fMRI Investigation of Emotional Engagement in Moral Judgment

Joshua D. Greene,[1,2*] R. Brian Sommerville,[1] Leigh E. Nystrom,[1,3]
John M. Darley,[3] Jonathan D. Cohen[1,3,4]

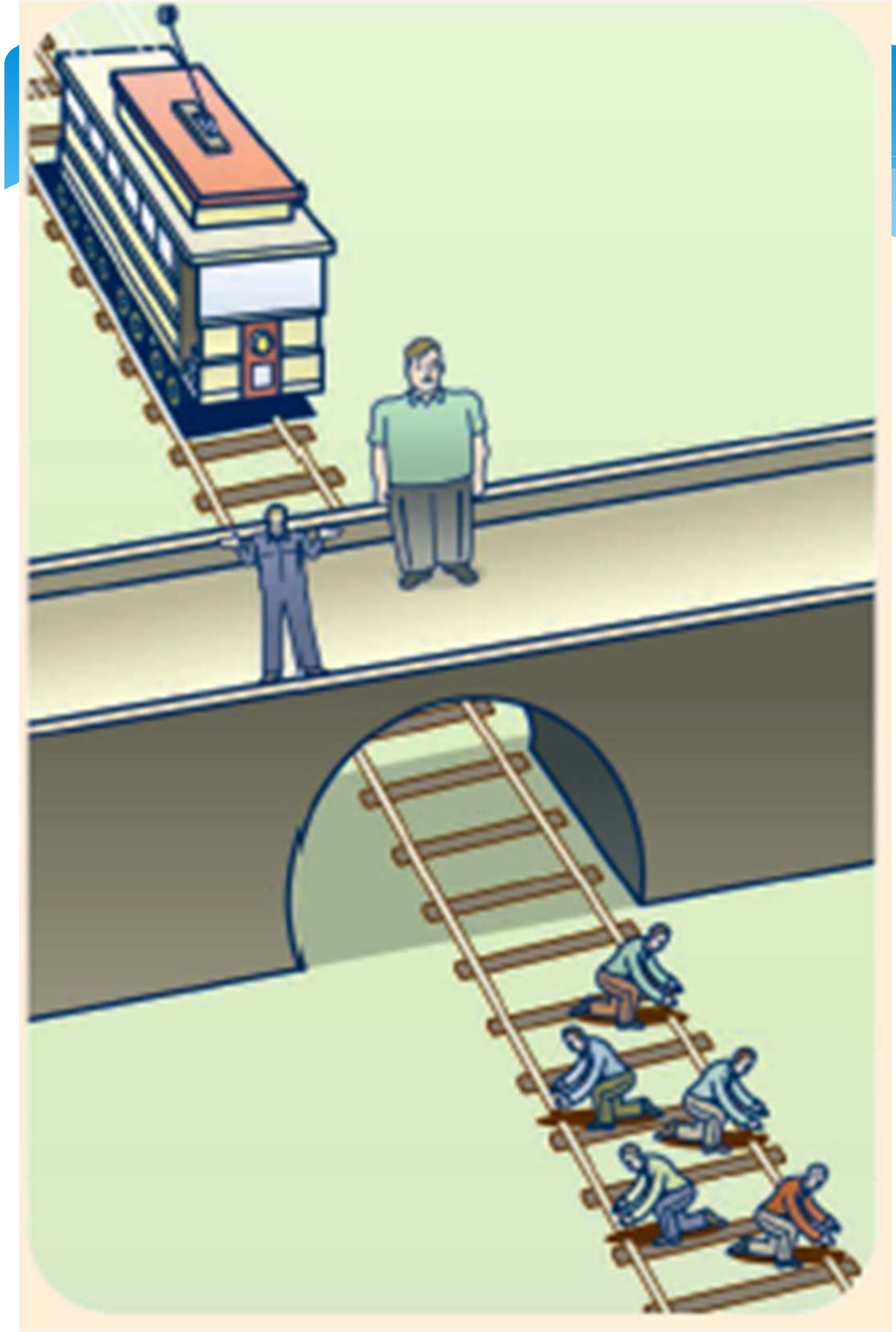
SCIENCE VOL 293 14 SEPTEMBER 2001

In two functional magnetic resonance imaging (fMRI) studies using moral dilemmas as probes, we apply the methods of cognitive neuroscience to the study of moral judgment.



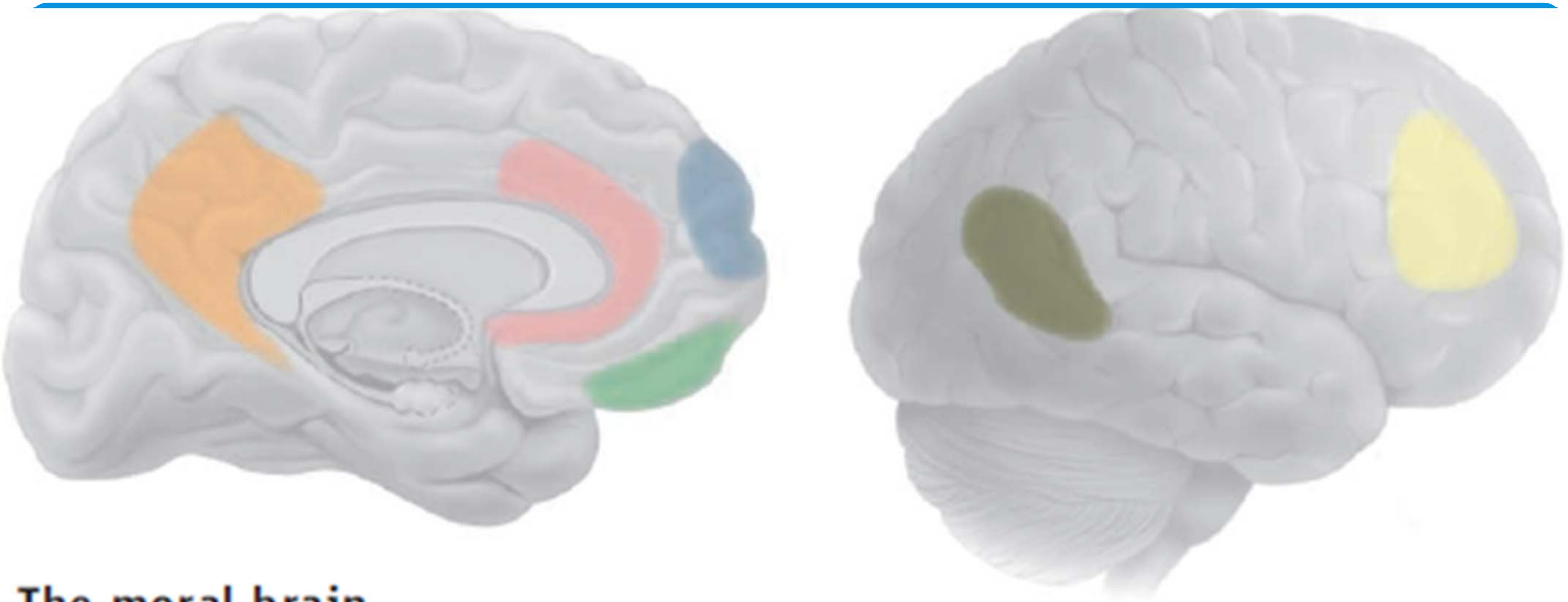
Experiment #1

Is it morally acceptable to redirect a runaway trolley car hurtling toward five workers onto a track with just one worker?



Experiment #2

How about pushing a man off a footbridge into the path of the trolley to stop it before it hits the hapless workers?



The moral brain.

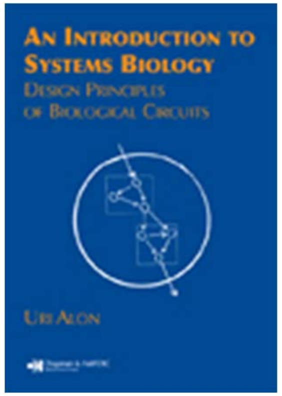
Neuroimaging studies have linked several brain regions to moral cognition. Disruptions to the right temporoparietal junction (brown), which is involved in understanding intentions, or the ventromedial prefrontal cortex (green), which processes emotion, have been found to alter moral judgments. Greene and colleagues have suggested that activity in the anterior cingulate cortex (pink) signals conflict between emotion, reflected by activity in the medial frontal gyrus (blue) and other areas (orange, brown), and "cold" cognition, reflected by activity in dorsolateral prefrontal cortex (yellow).

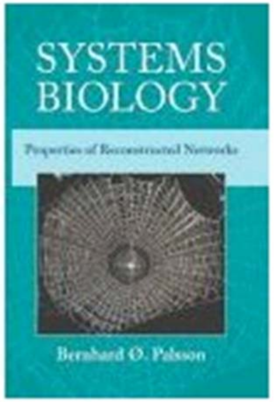
In a 2001 *Science paper, Greene, then a* postdoc with Jonathan Cohen at Princeton University, and colleagues reported that the medial frontal gyrus and other brain regions linked to emotion become more active when people contemplate “personal” moral dilemmas These impersonal dilemmas preferentially activate a different set of brain regions thought to contribute to abstract reasoning and problem solving.

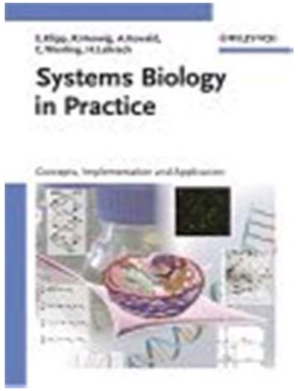
Greene envisions a tug of war between emotion and cognition in the brain: Emotions tell us we’ll feel terrible if we push the man; cognition says: Push him! Five is greater than one. Greene suspects that the arbiter in this conflict may be a brain region called the anterior cingulate cortex. Previous studies have found that this region fires up when people wrestle with many types of internal conflicts, and it did so when subjects in Greene’s study faced particularly difficult moral dilemmas.

推荐书目

- An introduction to Systems Biology: Design Principles of Biological Circuits
  by Uri Alon
  June 2006, Chapman&Hall/CRC, Taylor and Francis Group
- Systems Biology : Properties of Reconstructed Networks
  by Bernard Palsson
  January 2006, published by Cambridge Univ. Press
- Systems Biology in Practice: Concepts, Implementation And Application
  Klipp, E et al.
  John Wiley & Sons Inc. 2005
- Systems Biology: A Textbook Edda Klipp, et al. 2009


An Introduction to Systems Biology
Design Principles of Biological Circuits
Uri Alon




Systems Biology in Practice
Concepts, Implementation and Application

http://www.systems-biology.org

科普小书

- The music of life: Biology beyond the genome
- Denis Noble
- He is one of the pioneers of Systems Biology and developed the first viable mathematical model of the working heart in 1960. He is also a philosopher of biology, and his book *The Music of Life* challenges the foundations of current biological sciences, questions the central dogma, its unidirectional view of information flow, and its imposition of a bottom-up methodology for research in the life sciences

生命的乐章——后基因组时代的生物学(科学出版社)

State-of-Arts

- Alerts from Science, Nature, Cell, PNAS

- Nature Molecular Systems Biology
- BMC Systems Biology
- IET Systems Biology
- Other related journals

- Google, Wiki

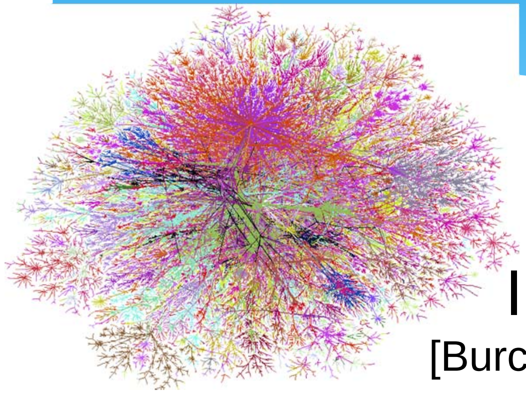
Model Vs Data

- Networking the whole biological system, rather than studying its isolated parts.
- Integrating large amounts of data in the context of biological network (Sequence, structure, function, gene expression, protein expression, protein interaction, protein-DNA interaction, and literature data).

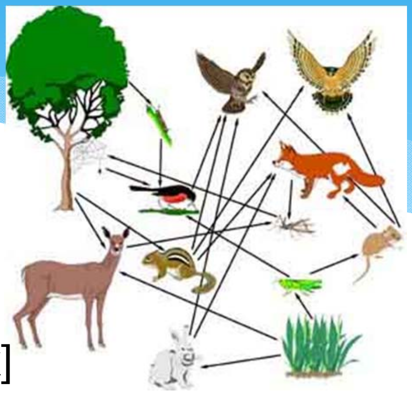
我们主要针对一类模型：

生物分子网络

Networks as a universal model

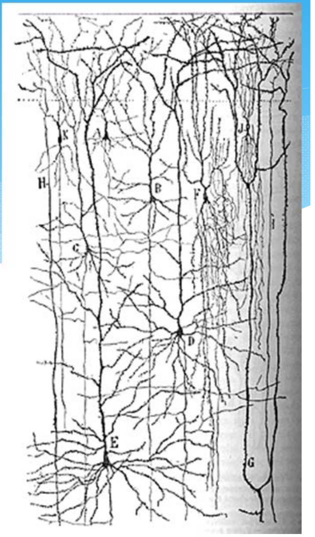


Internet
[Burch & Cheswick]

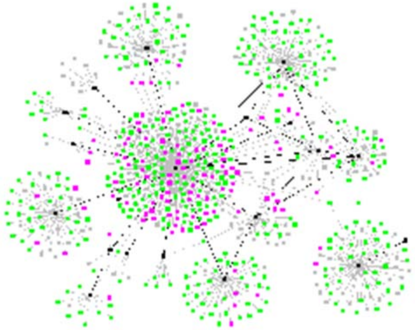


Food Web

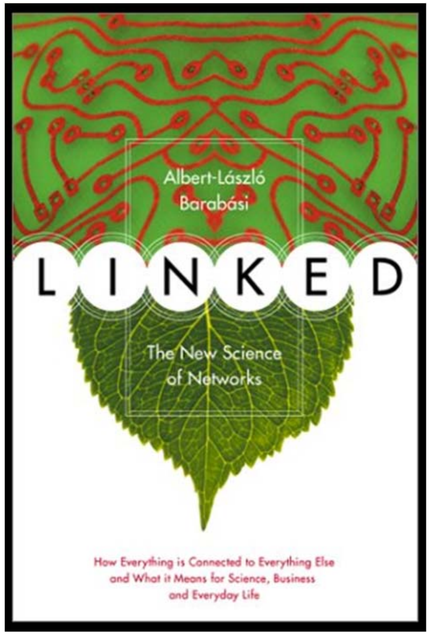
Electronic Circuit



Neural Network
[Cajal]



Disease Spread
[Krebs]



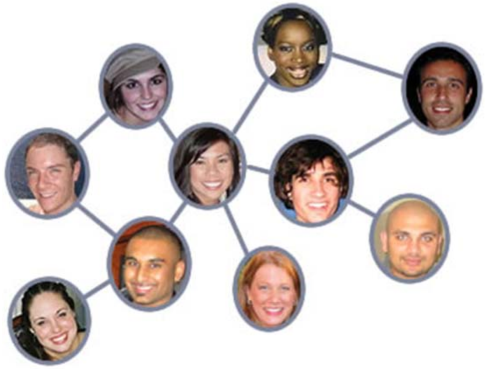
Albert-László Barabási

LINKED

The New Science of Networks

How Everything is Connected to Everything Else and What it Means for Science, Business and Everyday Life

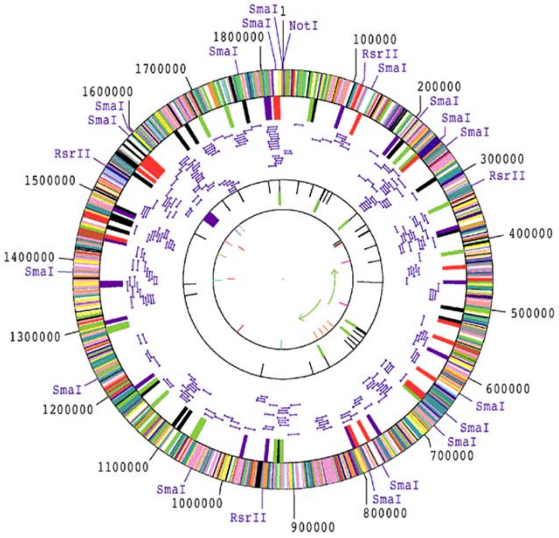
Protein Interactions
[Barabasi]



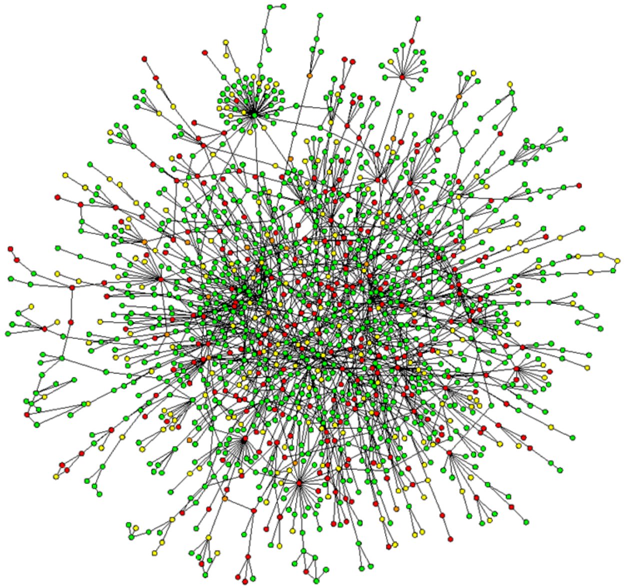
Social Network

生物分子网络

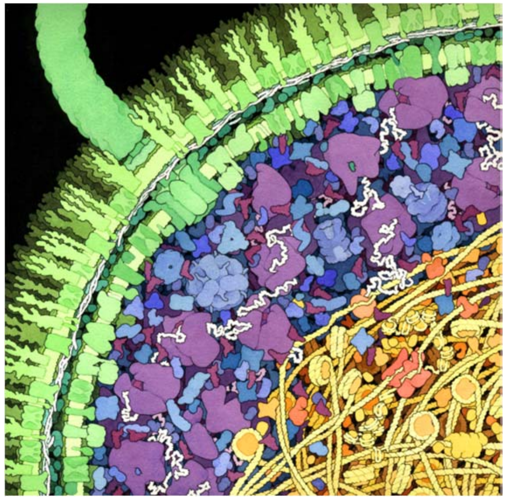
- 图作为基本工具用来强调相互作用并直观表示复杂的生物系统
- 节点代表生物分子，边代表他们之间在生命过程中的某种关系



1D: Complete Genetic Partslist



~2D: Bio-molecular Network



3D: Detailed structural understanding of cellular machinery

生物分子网络研究的科学问题

- **如何构建网络?**

建模，数据处理、集成

- **如何分析网络？网络与其它数据的集成**

静态：结构，功能等

动态：不同条件，进化

大纲

1. 异源同质数据集成
2. 异源异质数据集成
3. 动态与静态数据集成

1. 基因调控网络重建
2. 转录调控网络重建
3. 转录因子合作网络预测
4. 生物活性通路与网络标记物识别