浙江大学
ZHEJIANG UNIVERSITY

Mathematics in computer science

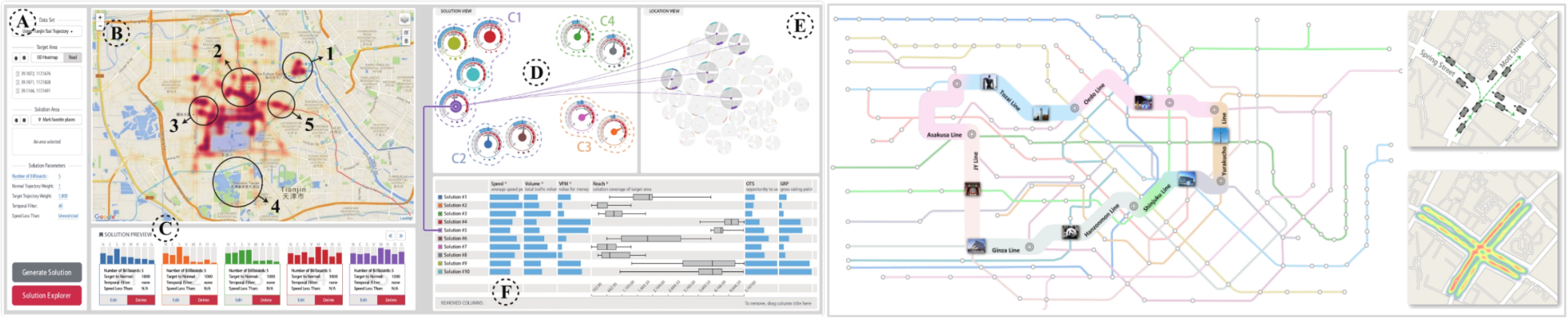
0. Why data-driven?

张宏鑫（Hongxin Zhang）

zhx@cad.zju.edu.cn

State Key Lab of CAD&CG, ZJU

2025-02-18

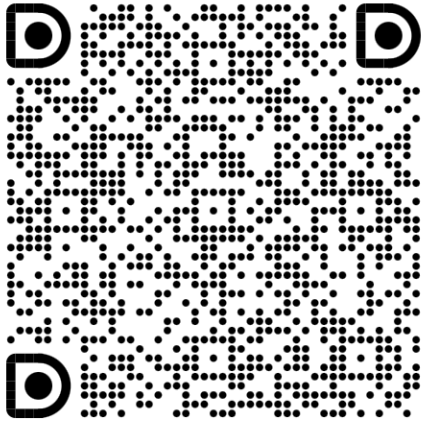


Outline

- Background
- What is data-driven about?
- Is it really useful for computer science and technology?

2025春春(每周)//星期二//11 -14 节/计算 内部

该群属于“浙江大学”内部群，仅组织内部成员可以加入，如果组织外部人员收到此分享，需要先申请加入该组织。



此二维码365天内有效（2026-02-18 前）

The largest challenge of Today's CS

- Big Data

- All big companies are collecting data!!!
  - Google, Apple, Facebook, IBM, Microsoft, Amazon, …

  - In China, Baidu, Alibaba, Tecent, 360, DiDi, Netease, Xiaomi, Sina, Huawei

The largest challenge of Today's CS

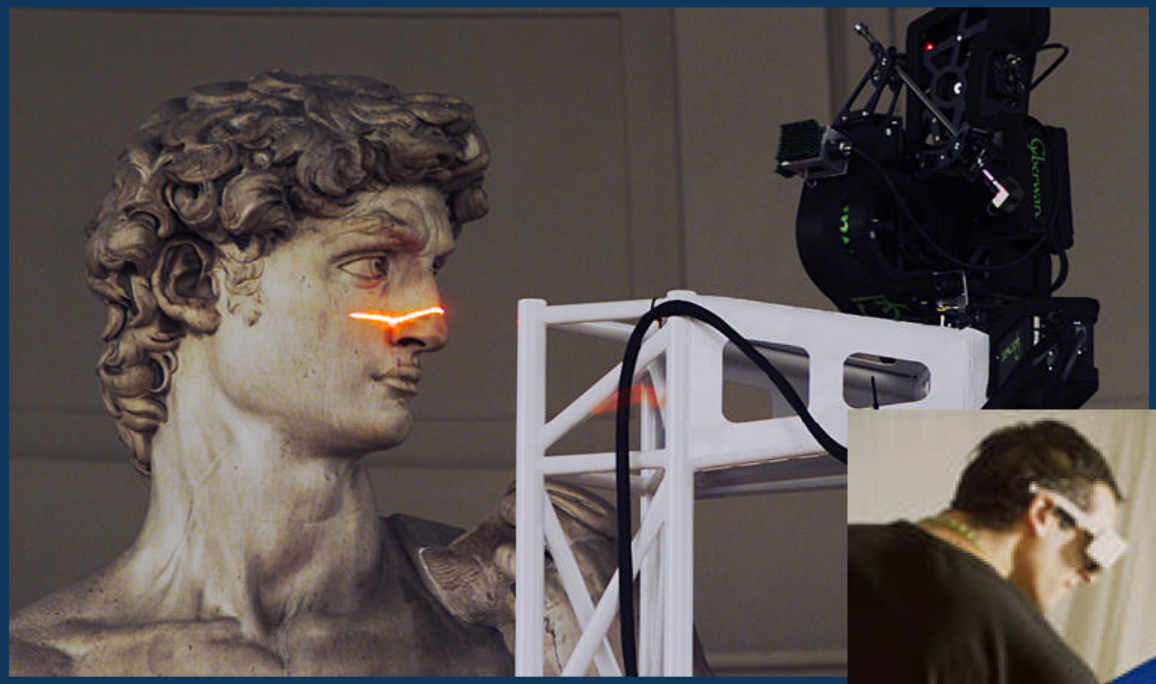
- Data, Data, Data …
  (in computer graphics)
  - The tedious effort required to create digital worlds and digital life.
    - Finding new ways to communicate and new kinds of media to create.
    - Experts are expensive: scientists, engineers, filmmakers, graphic designers, fine artists, and game designers.

- Process existing data and then create new ones from them.

Computers are really fast

- If you can create it, you can render it

How do you create it?



Digital Michaelangelo Project

Steven Schkolne

Pure procedural synthesis vs. Pure data

- Creating motions for a character in a movie
  - Pure procedural synthesis (model) (Geek, 64k)
    - compact, but very artificial, rarely used in practice.
  - “By hand” or “pure data” (data) (Artists and workers)
    - higher quality but lower flexibility.
  - the best of both worlds: hybrid methods?!?

Everything but Avatar

Make it easy and true

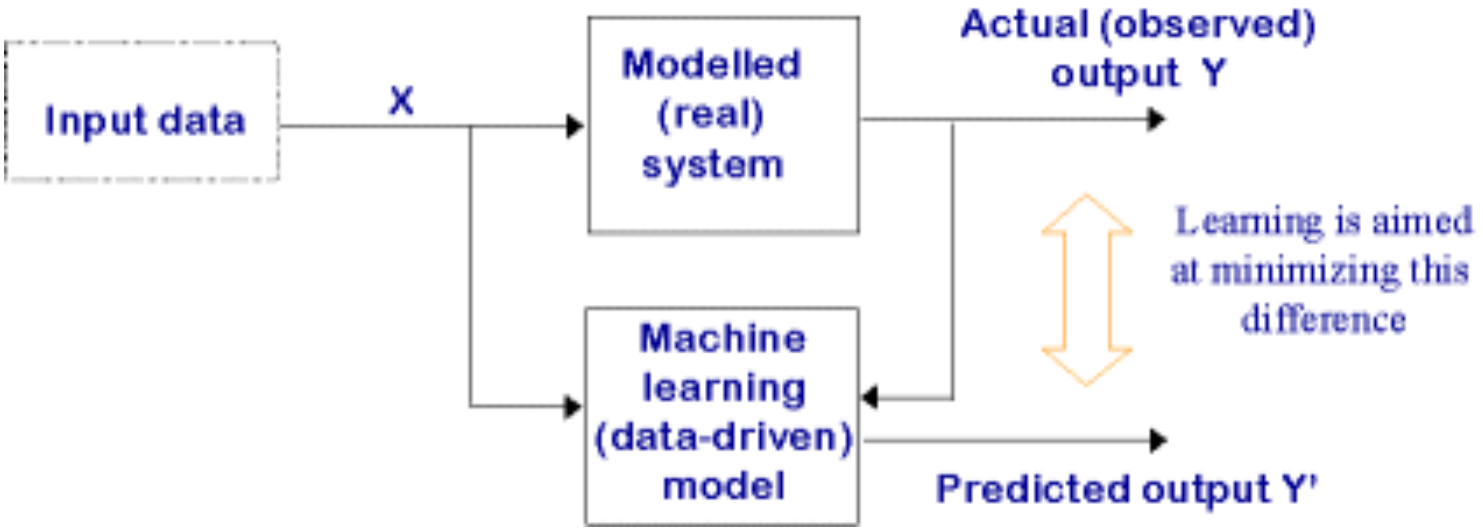
Bayesian Reasoning

- Principle modeling of uncertainty.
- General purpose models for unstructured data.
- Effective algorithm for data fitting and analysis under uncertainty.

➢ But currently it is always used as a black box.

Belief v.s. Probability

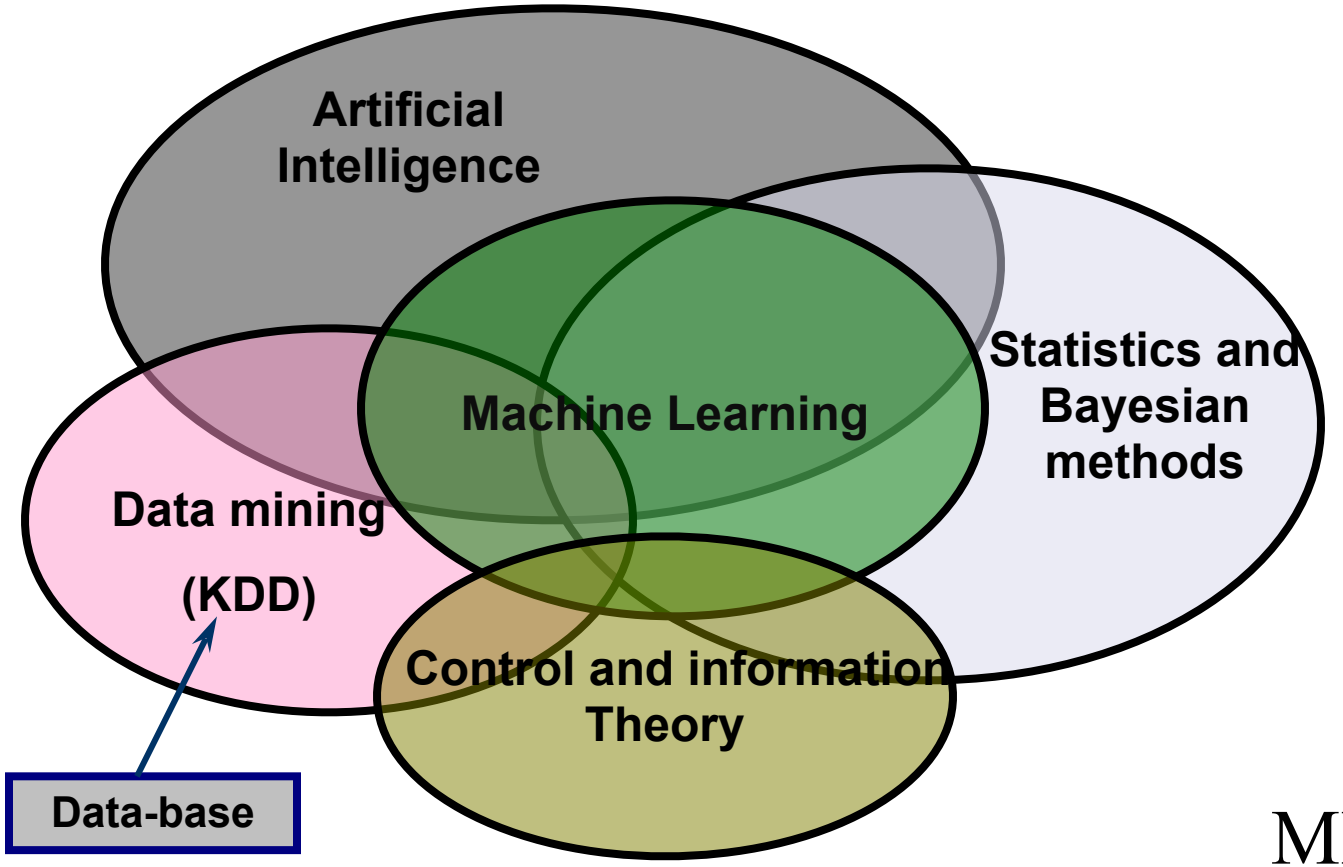
Data driven modeling



Data-driven vocabulary

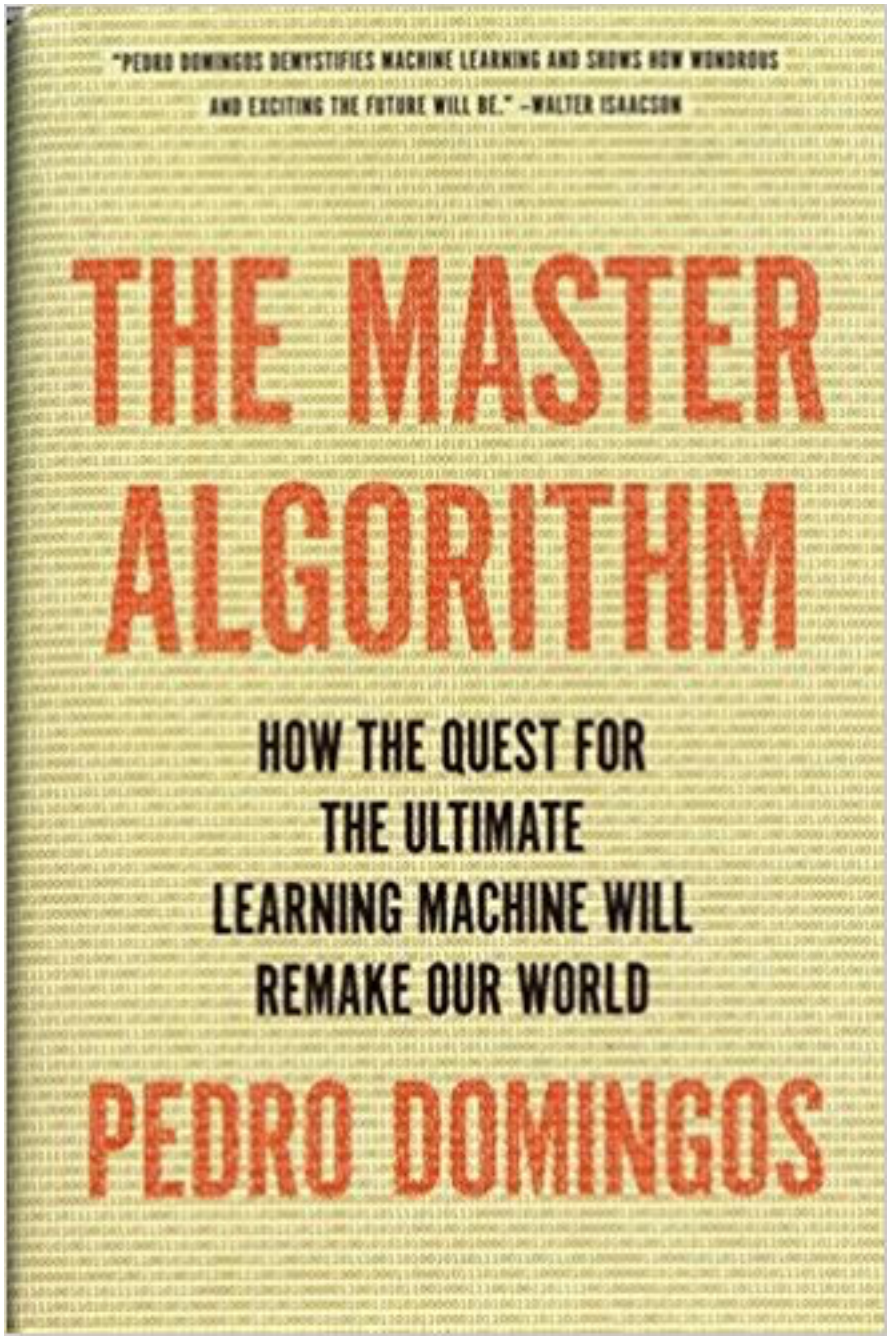
- Data
  - data-driven, data mining
- Learning
  - machine learning, statistical learning
- Uncertainty
  - probability, likelihood
- Intelligent
  - Inference, decision, detection, recognition

Data-driven related techniques



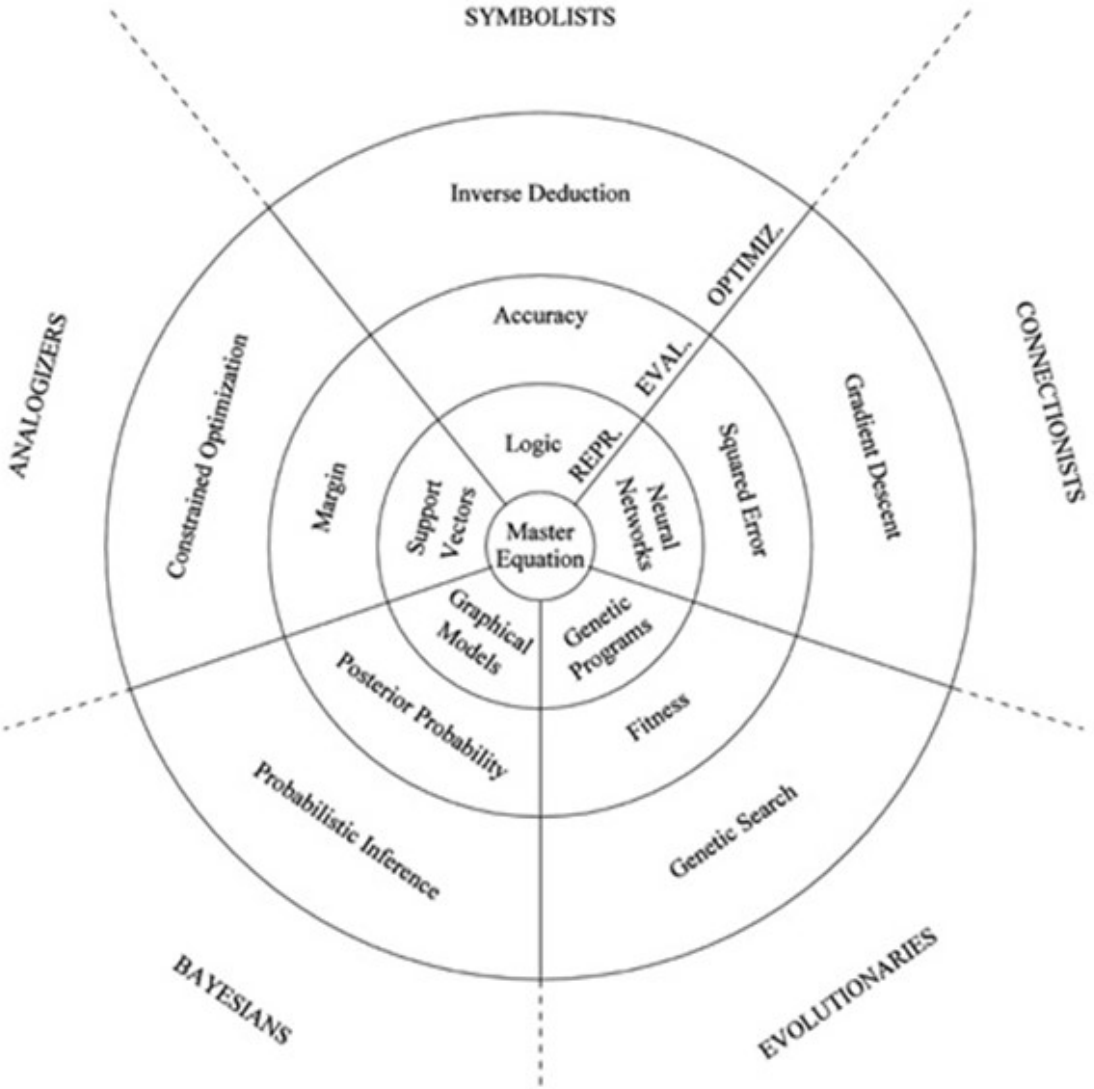
$ML \neq AI$

Computer Vision	Multi-media	Bio-informatics	Computer Graphics	Information retrieval


"PEDRO DOMINGOS DEMYSTIFIES MACHINE LEARNING AND SHOWS HOW WONDROUS AND EXCITING THE FUTURE WILL BE." –WALTER ISAACSON
THE MASTER ALGORITHM
HOW THE QUEST FOR THE ULTIMATE LEARNING MACHINE WILL REMAKE OUR WORLD
PEDRO DOMINGOS


中信出版集团
THE MASTER ALGORITHM
HOW THE QUEST FOR THE ULTIMATE LEARNING MACHINE WILL REMAKE OUR WORLD
终极算法
机器学习和人工智能如何重塑世界
Pedro Domingos
佩德罗·多明戈斯 著
近20年人工智能领域最具轰动性的著作!
揭秘机器学习的终极逻辑,
全景勾勒人工智能的商业未来
比尔·盖茨年度荐书!
《乔布斯传》作者 沃尔特·艾萨克森、图灵奖得主Judea Pearl
中国大数据领航人 车品觉、今日头条首席算法构架师 曹欢欢 倾力推荐!
Google X、微软研究院 密切关注!
中信出版集团

机器学习的5个学派



- **符号学派**

  将学习看作逆向演绎，并从哲学、心理学、逻辑学中寻求洞见

- **联结学派**

  对大脑进行逆向分析，灵感来源于神经科学和物理学

- **进化学派**

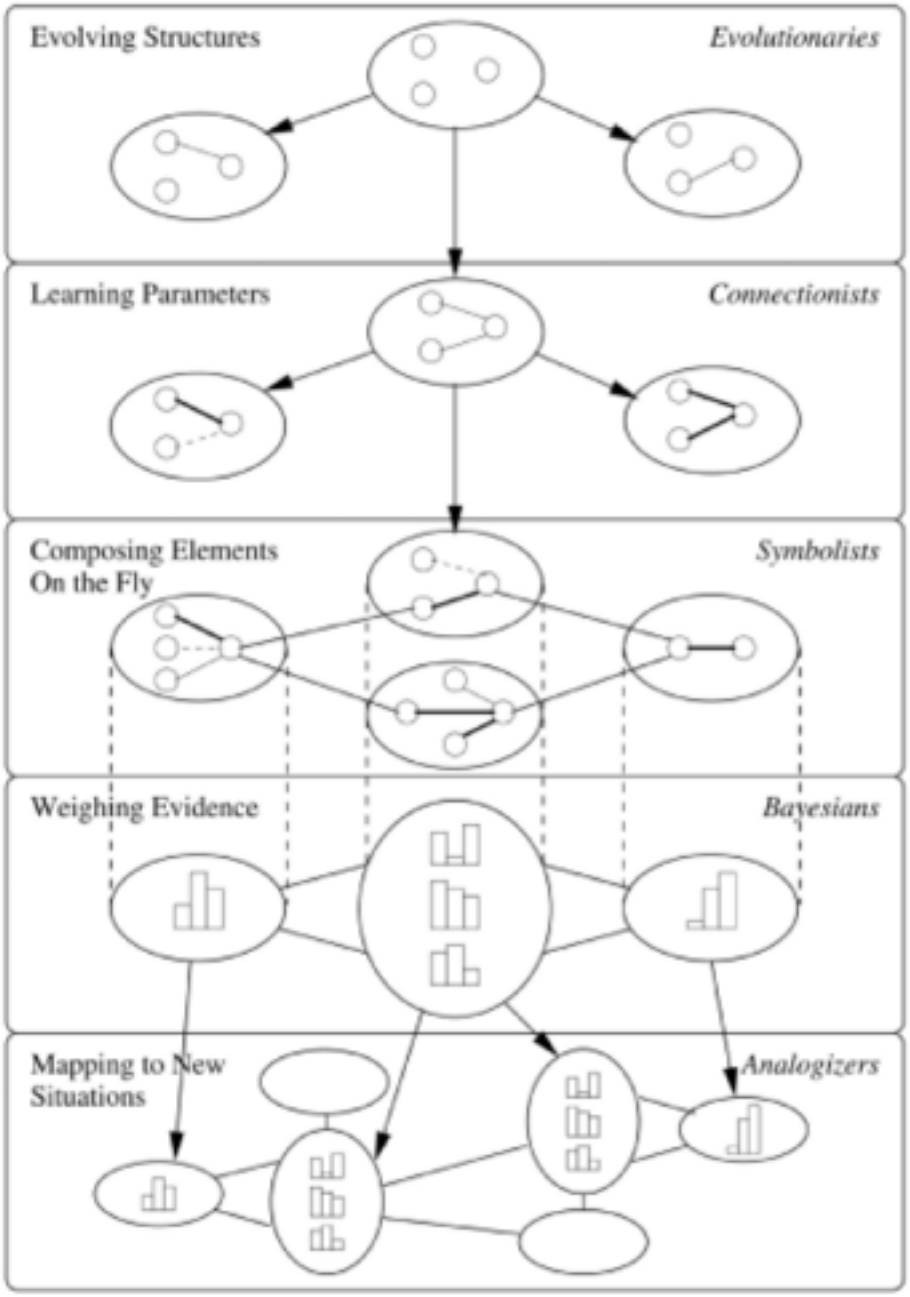
  在计算机上模拟进化，并利用遗传学和进化生物学知识

- **贝叶斯学派**

  认为学习是一种概率推理形式，理论根基在于统计学

- **类推学派**

  通过对相似性判断的外推来进行学习，并受心理学和数学最优化的影响



- **进化学派**

  在计算机上模拟进化，并利用遗传学和进化生物学知识

- **符号学派**

  将学习看作逆向演绎，并从哲学、心理学、逻辑学中寻求洞见

- **联结学派**

  对大脑进行逆向分析，灵感来源于神经科学和物理学

- **贝叶斯学派**

  认为学习是一种概率推理形式，理论根基在于统计学

- **类推学派**

  通过对相似性判断的外推来进行学习，并受心理学和数学最优化的影响

What is machine learning?

- Definition by Mitchell, 1997
  - *A program learns from experience E with respect to some class of tasks T and performance measure P, if its performance at task T, as measured by P, improves with experience E.*
  - 机器学习乃于某类任务兼性能度量的经验中学习之程序；若其作用于任务，可由度量知其于已知经验中获益。

- Comments from Hertzmann, 2003
  - For the purposes of computer graphics, machine learning should really be viewed as a set of techniques for leveraging data. Given some data, we can model the process that generated the data.

Data-driven system

- Learning systems are not directly programmed to solve a problem, instead develop own program based on:
  - examples of how they should behave
  - from trial-and-error experience trying to solve the problem

Different from standard CS: want to implement unknown function, only have access to sample input-output pairs (training examples)

Main categories of learning problems

Learning scenarios differ according to the available information in training examples

- **Supervised**: correct output available
  - **Classification**: 1-of-N output (speech recognition, object recognition,medical diagnosis)
  - **Regression**: real-valued output (predicting market prices, temperature)
- **Unsupervised**: no feedback, need to construct measure of good output
  - **Clustering** : Clustering refers to techniques to segmenting data into coherent “clusters.”
  - **Novelty-detection**: detecting new data points that deviate from the normal.
- **Reinforcement**: scalar feedback, possibly temporally delayed

Main class of learning problems

Learning scenarios differ according to the available information in training examples

- **Supervised**: correct output available
  - …
- **Semi-Supervised**: only a part of output available
  - **Ranking**:
- **Unsupervised**: no feedback, need to construct measure of good output
  - …
- *Reinforcement*: scalar feedback, possibly temporally delayed

And more …

- Time series analysis
- Dimension reduction
- Model selection
- Generic methods
- Graphical models

Why data driven methods?

- **Develop enhanced computer systems**
  - automatically adapt to user, customize
  - often difficult to acquire necessary knowledge
  - discover patterns offline in large databases (*data mining*)
- **Improve understanding of human, biological learning**
  - computational analysis provides concrete theory, predictions
  - explosion of methods to analyze brain activity during learning
- **Timing is good**
  - growing amounts of data available
  - cheap and powerful computers
  - suite of algorithms, theory already developed

Is it really useful for computer science and technology?

- **Con**: Everything is machine learning or everything is human tuning?
  - Sometimes, this may be true

- **Pro**: more understanding of learning, but yields much more powerful and effective algorithms.
  - Problem taxonomy
  - General-purpose models
  - Reasoning with probabilities

❖ I believe the mathematic magic

What will be a successful D-D algorithm?

- Computational efficiency
- Robustness
- Statistical stability

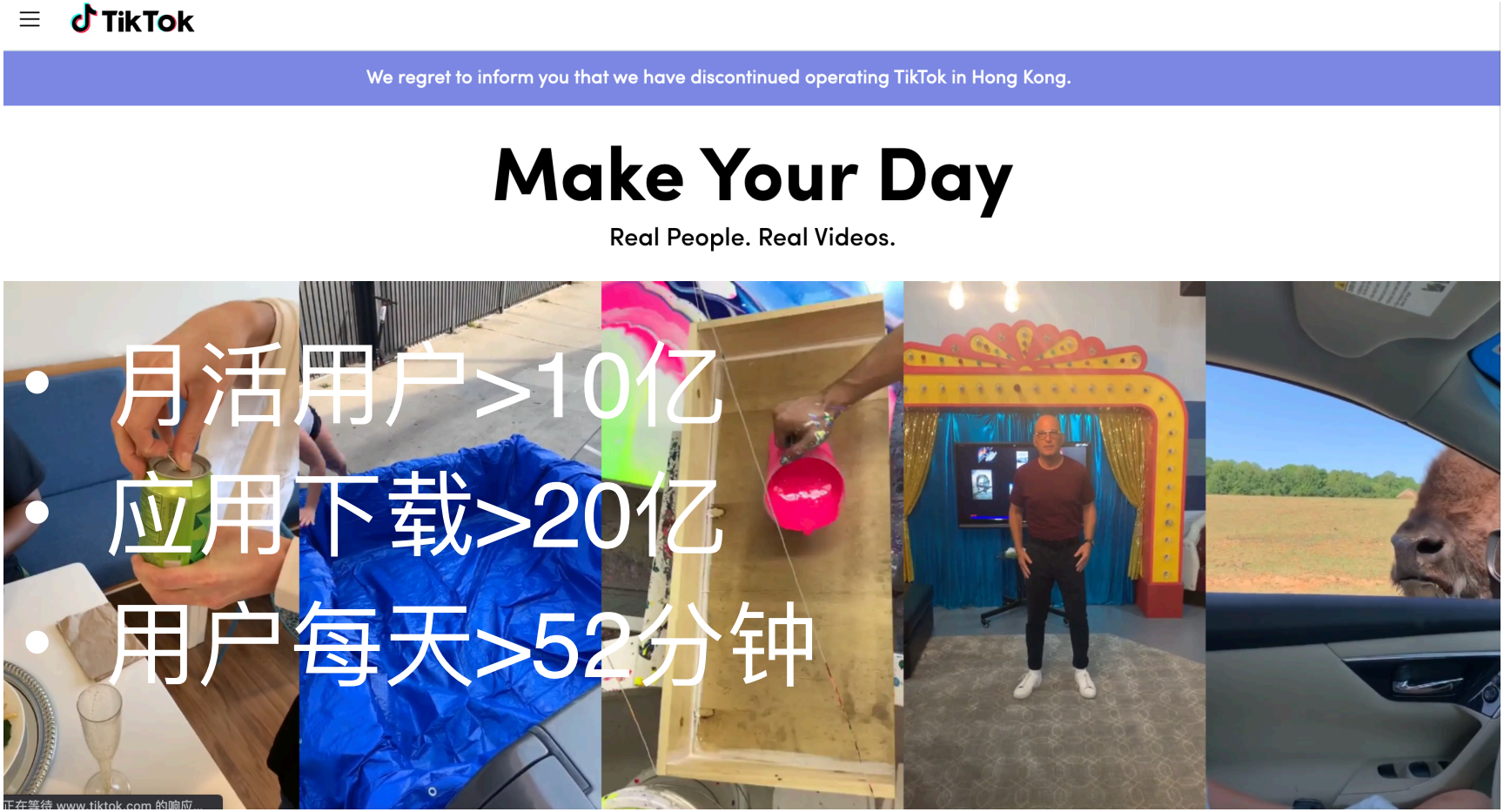
Applications of Data-driven

Old and New …

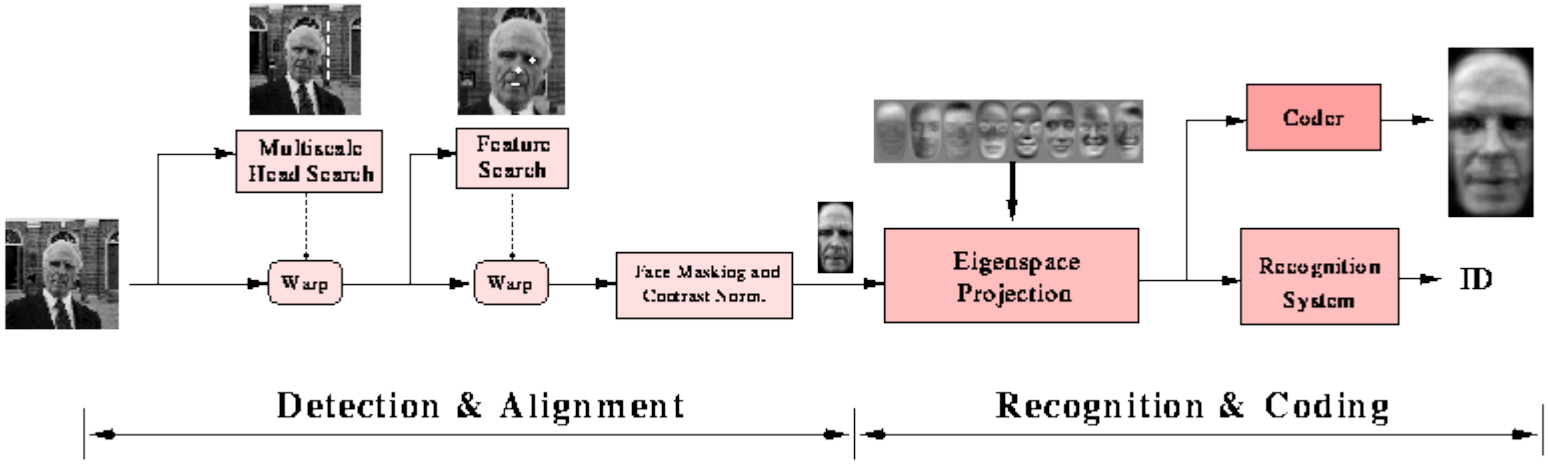
The First Example: Google!

- 每天过滤xxxx亿个网页
- 每天追踪xxxx亿个的独立URL
- 每月接受xxxx亿次搜索请求

The great example in China: TikTok/Douyin (ByteDance)



Object detection and recognition - the power of DD



The image is copied from

http://vismod.media.mit.edu/vismod/demos/facerec/

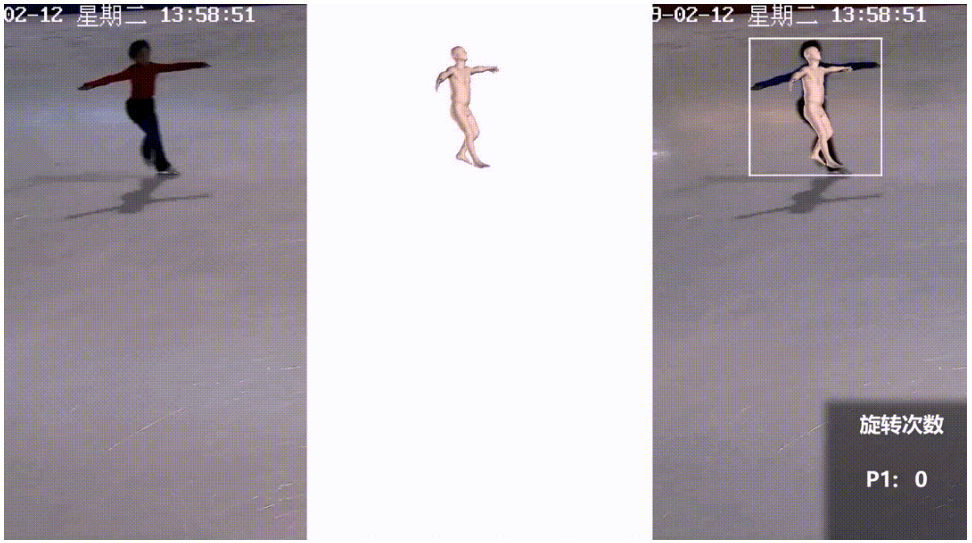
Object detection and recognition

Face [Vaillant et al IEE 1994] [Garcia et al PAMI 2005] [Osadchy et al JMLR 2007]
Pedestrian: [Kavukcuoglu et al. NIPS 2010] [Sermanet et al. CVPR 2013]

人脸识别-数据集

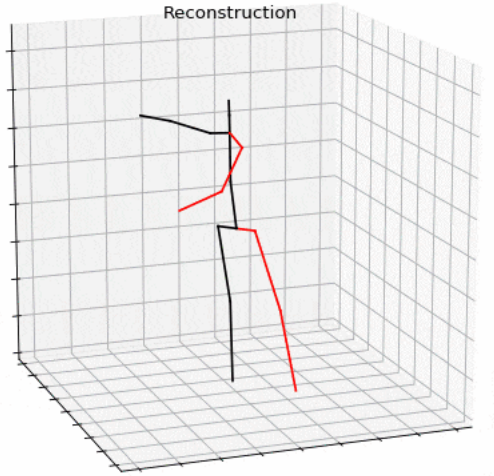
WebFace	10k+人，约500K张图片	非限制场景
FaceScrub	530人，约100k张图片	非限制场景
YouTube Face	1,595个人 3,425段视频	非限制场景、视频
LFW	5k+人脸，超过10K张图片	标准的人脸识别数据集
MultiPIE	337个人的不同姿态、表情、光照的人脸图像，共750k+人脸图像	限制场景人脸识别
MegaFace	690k不同的人的1000k人脸图像	新的人脸识别评测集合
IJB-A	24,327张图像 and 49,759 faces	人脸识别，人脸检测
CAS-PEAL	1040个人的30k+张人脸图像，主要包含姿态、表情、光照变化	限制场景下人脸识别
Pubfig	200个人的58k+人脸图像	非限制场景下的人脸识别
MS-Celeb-1M	100k人，10M人脸图像	非限制场景

In Olympic Games Beijing 2022



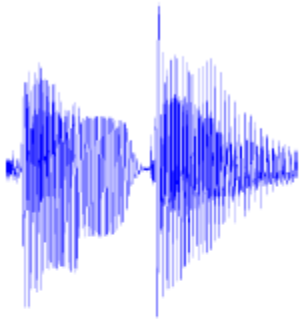
3D Modeling and Visualization

3D Skeleton Detection



Speech recognition

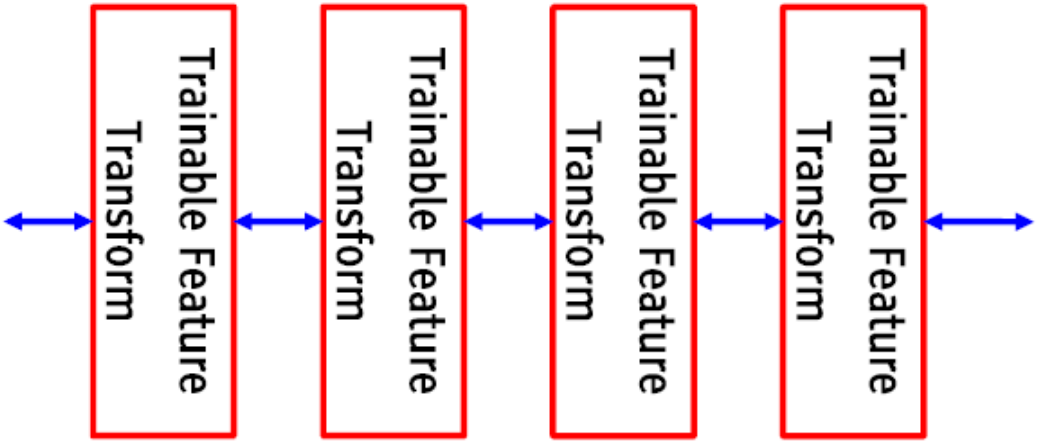
- Modern architecture for pattern recognition
  - Speech recognition: early 90's – 2011



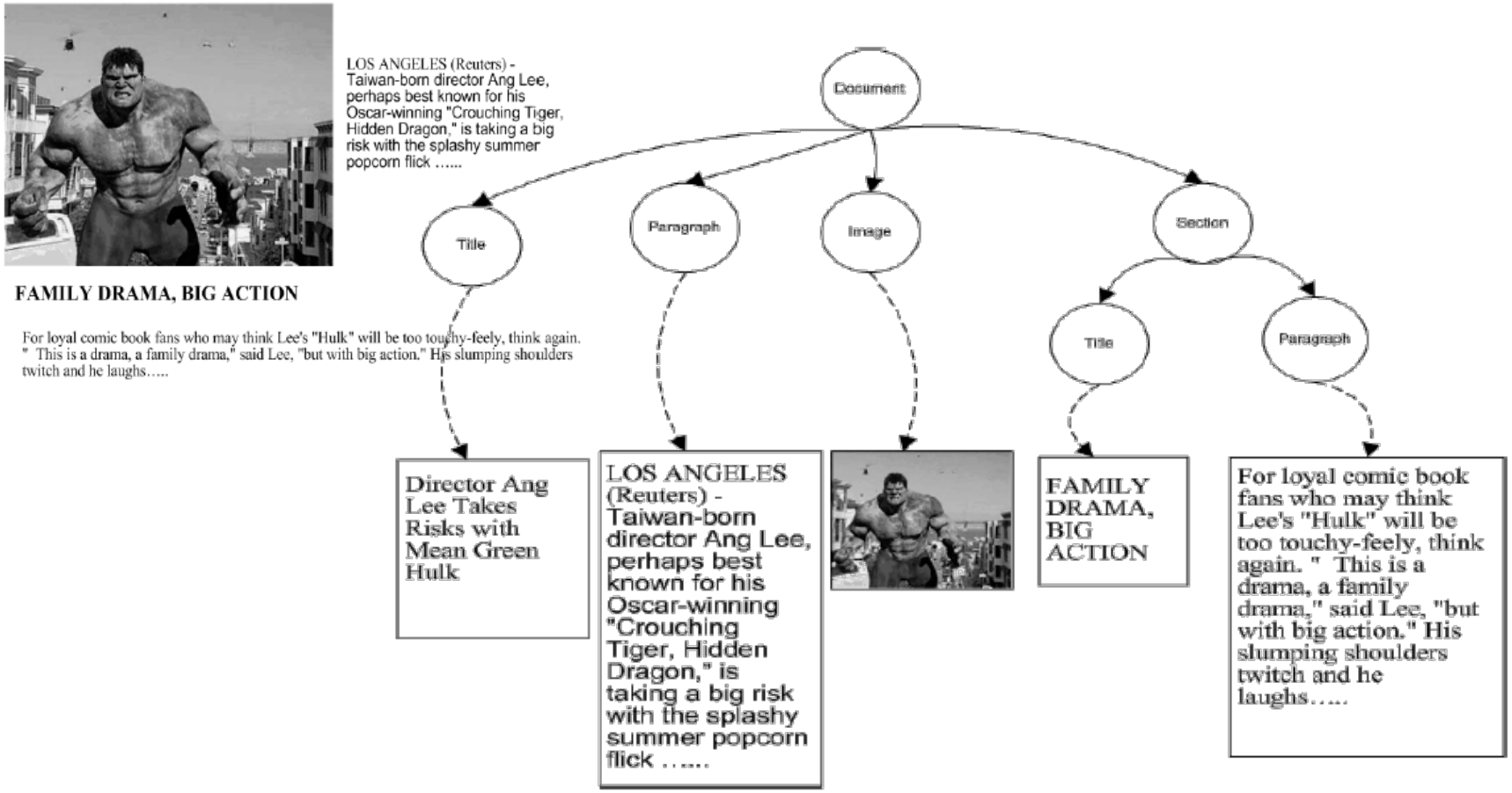
  - Object Recognition: 2006 - 2012

Speech recognition

- Hierarchy of representations with increasing level of abstraction
- Each stage is a kind of trainable feature transform
- Image recognition
  - Pixel → edge → texton → motif → part → object
- Text
  - Character → word → word group → clause → sentence → story
- Speech
  - Sample → spectral band → sound → … → phone → phoneme → word →

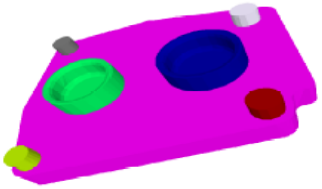


Document processing: Bayesian classification

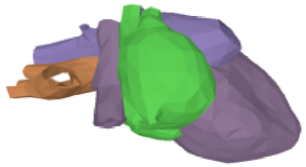


Mesh Processing – Data clustering/segmentation

- Hierarchical Mesh Decomposition using Fuzzy Clustering and Cuts. By Sagi Katz and Ayellet Tal, SIGGRAPH 2003

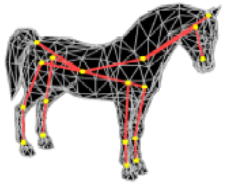


(c) mechanical part – 1270 faces
7 patches



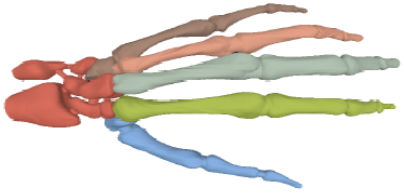
(d) heart – 1619 faces
4 patches

(a) object

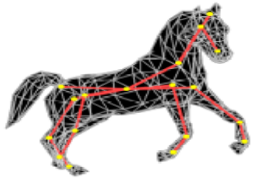


(b) skeleton

(e) Venus – 67,170 faces
3 patches



(f) skeleton hand – 654,666 faces
6 patches



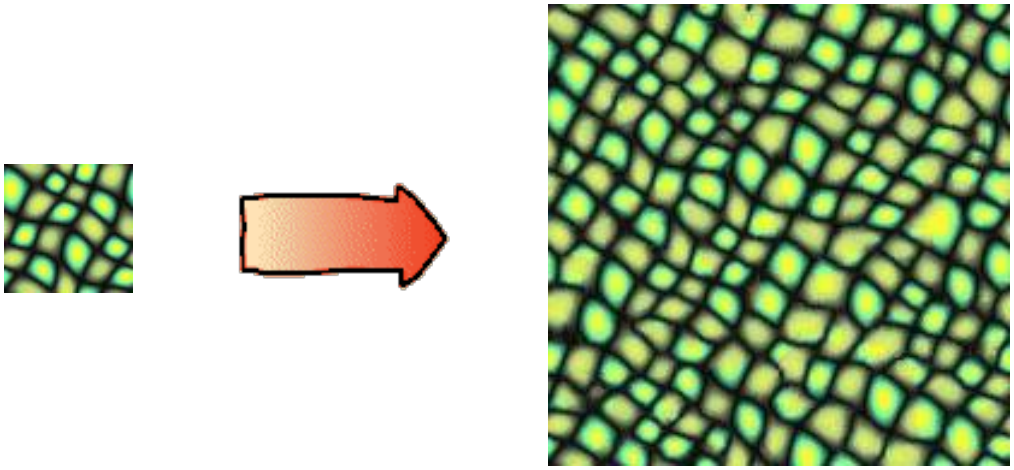
(c) deformed skeleton

(d) deformed object

Texture synthesis and analysis – Hidden Markov Model

- *Texture Synthesis over Arbitrary Manifold Surfaces.* Li-Yi Wei and Marc Levoy. SIGGRAPH 2001.

- *Fast Texture Synthesis using Tree-structured Vector Quantization.* Li-Yi Wei and Marc Levoy. SIGGRAPH 2000.



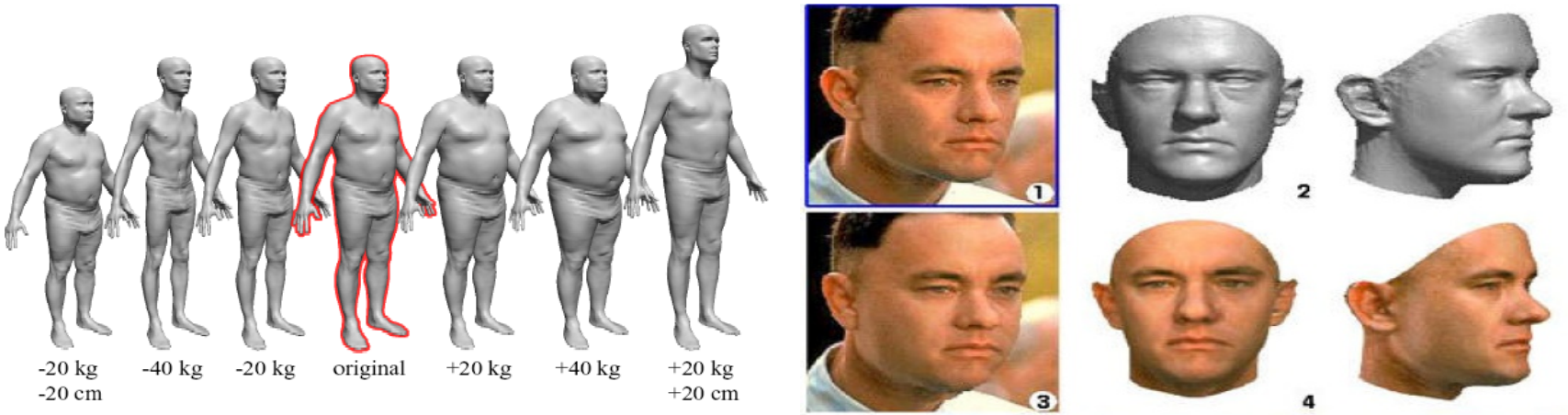
Reflectance texture synthesis – Dimension reduction

- *Synthesizing Bidirectional Texture Functions for Real-World Surfaces.* Xinguo Liu, Yizhou Yu and Heung-Yeung Shum. SIGGRAPH 2001.
- More recent papers…

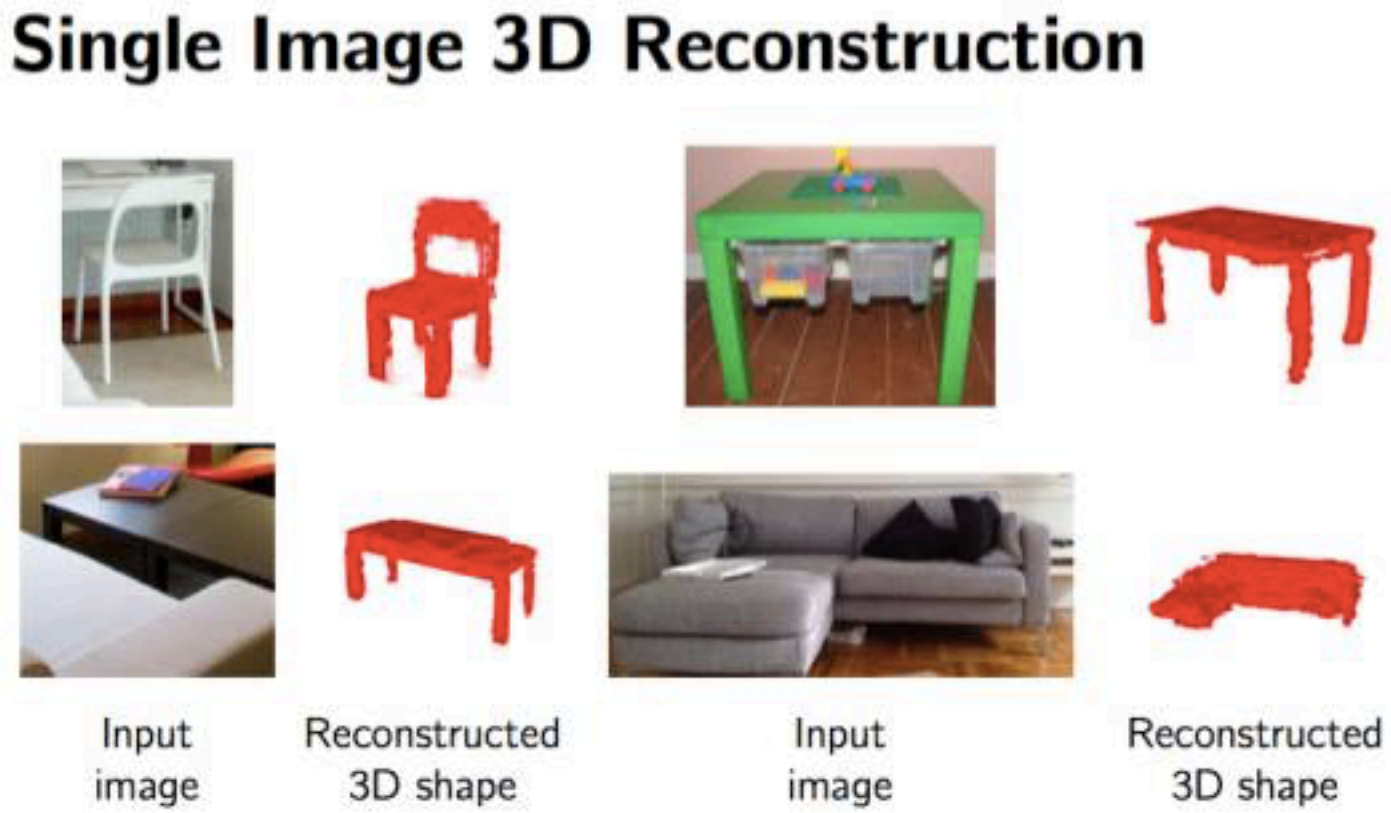
Human shapes - Dimension reduction

- *The Space of Human Body Shapes: Reconstruction and Parameterization From Range Scans.* Brett Allen, Brian Curless, Zoran Popovic. SIGGRAPH 2003.

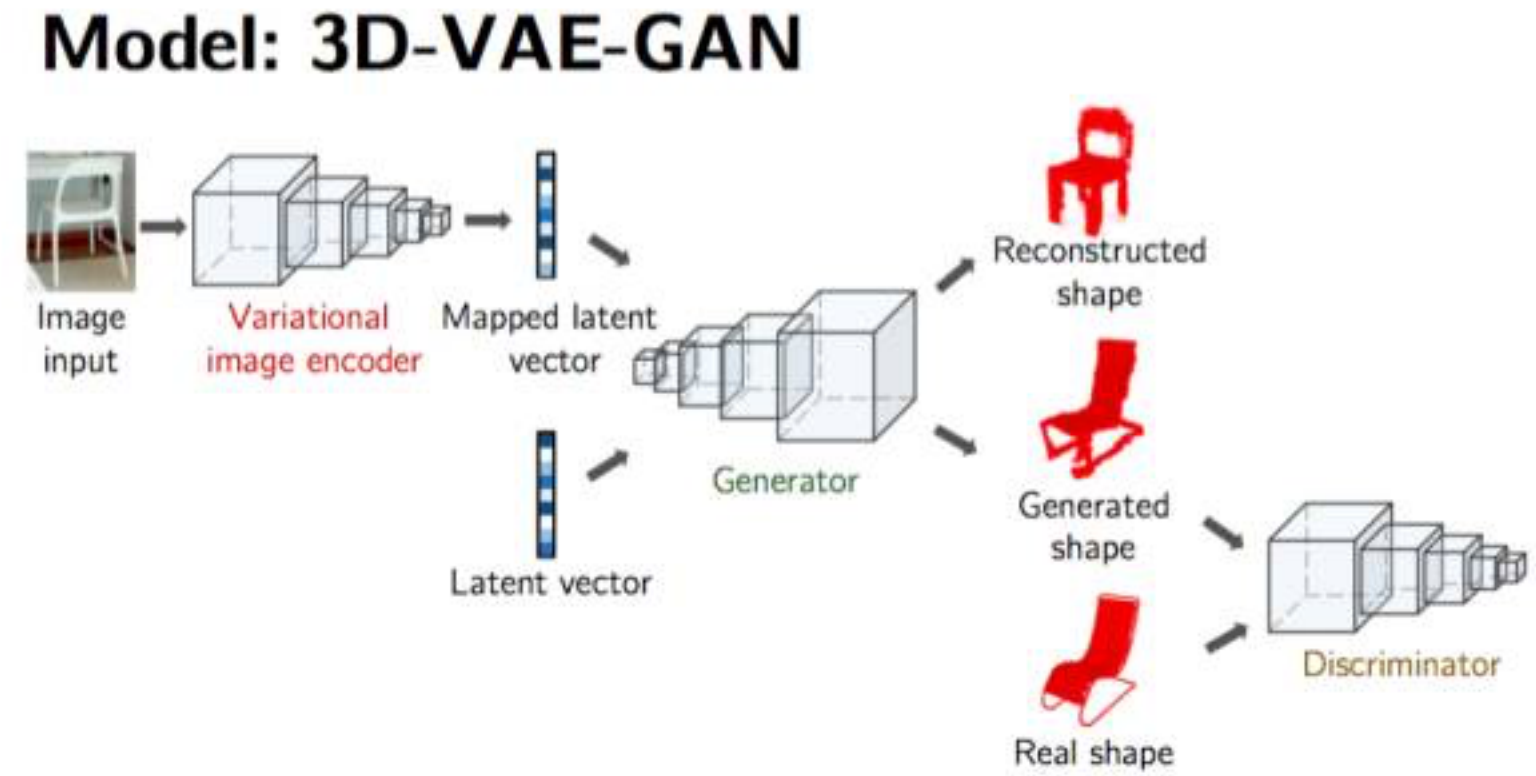
- *A Morphable Model for the Synthesis of 3D Faces.* Volker Blanz and Thomas Vetter. SIGGRAPH 1999.



Learning a Probabilistic Latent Space of Object Shapes – GANs (NIPS2016)



Learning a Probabilistic Latent Space of Object Shapes – GANs

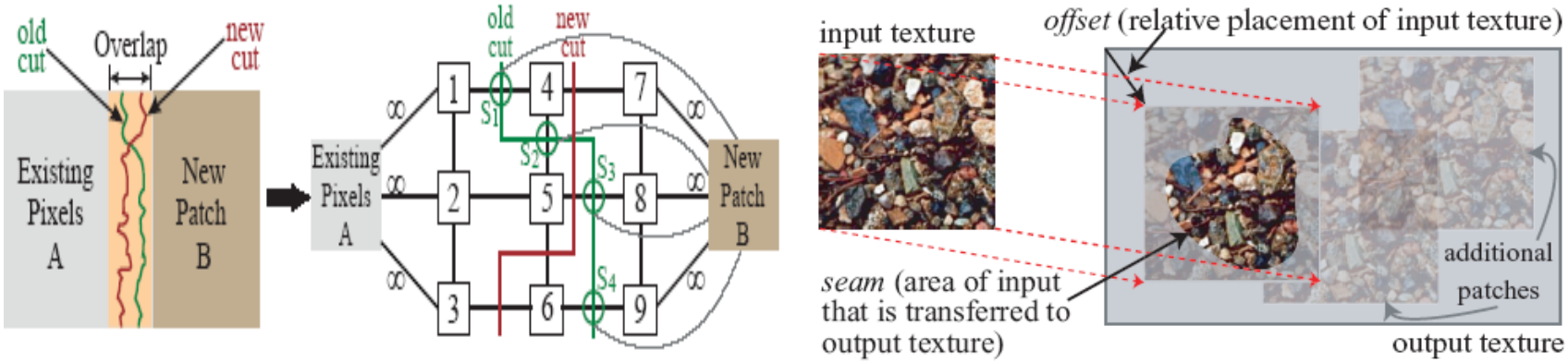


We combine the encoder with 3D-GAN for reconstruction and generation.

Image processing and synthesis - Graphical model

- *Image Quilting for Texture Synthesis and Transfer*. Alexei A. Efros and William T. Freeman. SIGGRAPH 2001.

- *Graphcut Textures: Image and Video Synthesis Using Graph Cuts*. V Kwatra, I. Essa, A. Schödl, G. Turk, and A. Bobick. SIGGRAPH 2003.



Data-driven based Rendering

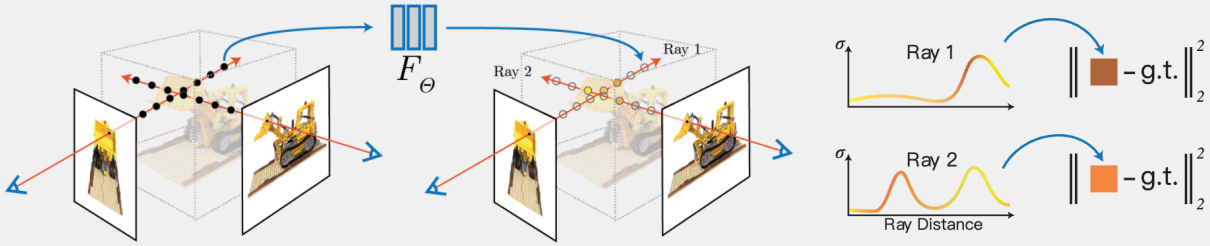
- [NeRF: Neural Radiance Fields (ECCV2020)]

Abstract & Method

We present a method that achieves state-of-the-art results for synthesizing novel views of complex scenes by optimizing an underlying continuous volumetric scene function using a sparse set of input views.

$$(x,y,z,\theta,\phi) \rightarrow F_\Theta \rightarrow (RGB\sigma)$$

Our algorithm represents a scene using a fully-connected (non-convolutional) deep network, whose input is a single continuous 5D coordinate (spatial location (x, y, z) and viewing direction (θ, φ)) and whose output is the volume density and view-dependent emitted radiance at that spatial location.

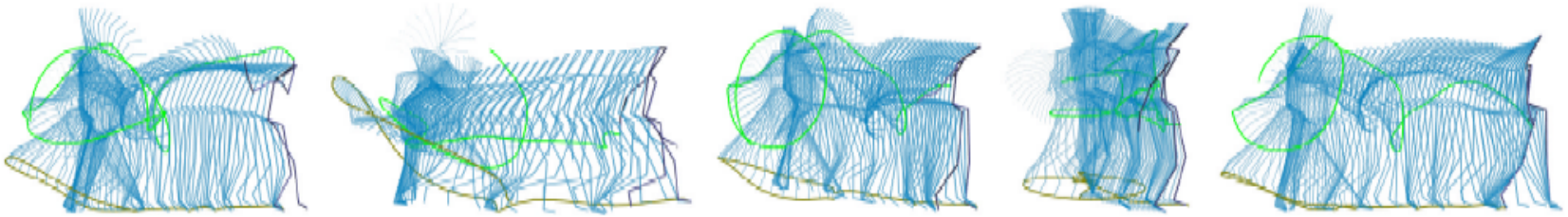


We synthesize views by querying 5D coordinates along camera rays and use classic volume rendering techniques to project the output colors and densities into an image. Because volume rendering is naturally differentiable, the only input required to optimize our representation is a set of images with known camera poses. We describe how to effectively optimize neural radiance fields to render photorealistic novel views of scenes with complicated geometry and appearance, and demonstrate results that outperform prior work on neural rendering and view synthesis.

Human Motion - Time series analysis

- *Style Machines.* M. Brand and A. Hertzmann. SIGGRAPH 2000.

- *A Data-Driven Approach to Quantifying Natural Human Motion*. L. Ren, A. Patrick, A. Efros, J. Hodgins, J. Rehg. SIGGRAPH 2005



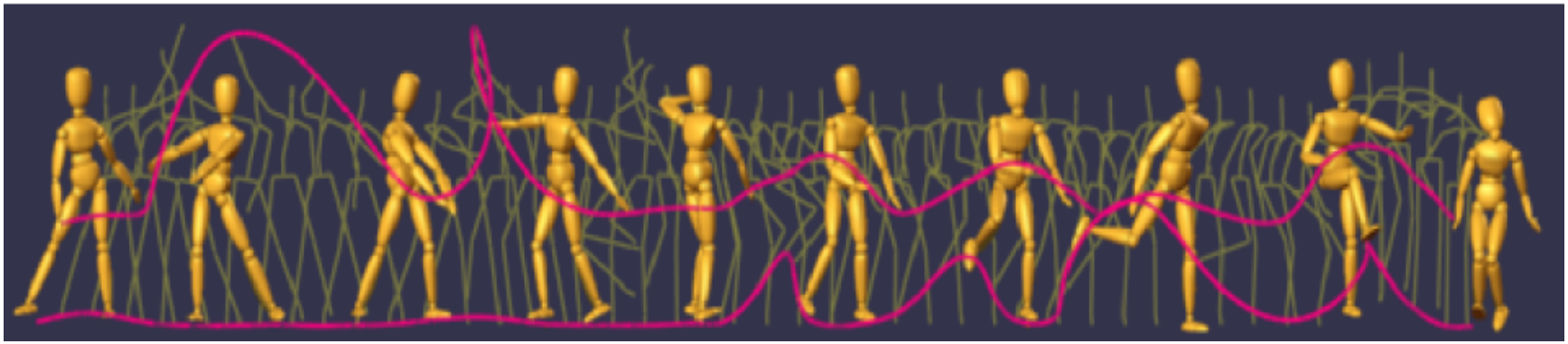
A pirouette and promenade in five synthetic styles drawn from a space that contains ballet, modern dance, and different body types. The choreography is also synthetic. Streamers show the trajectory of the left hand and foot.

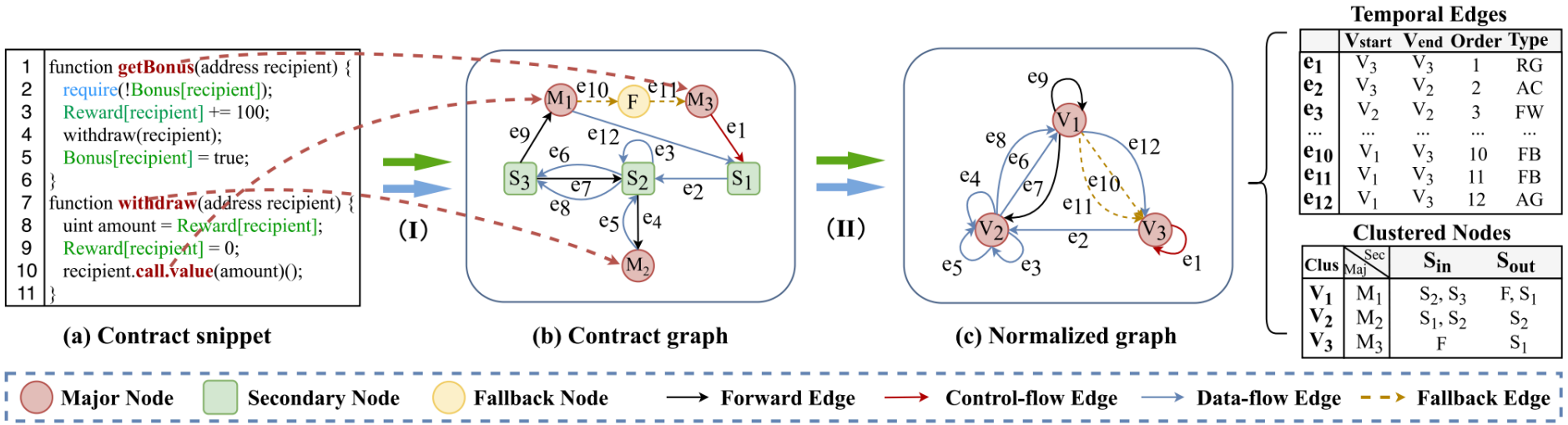
Video Textures - Reinforcement Learning

- *Video textures*. Arno Schödl, Richard Szeliski, David H. Salesin, and Irfan Essa. *SIGGRAPH 2000*.

Motion texture - Linear dynamic system

- *Motion Texture: A Two-Level Statistical Model for Character Motion Synthesis.* Yan Li, Tianshu Wang, and Heung-Yeung Shum. SIGGRAPH 2002.



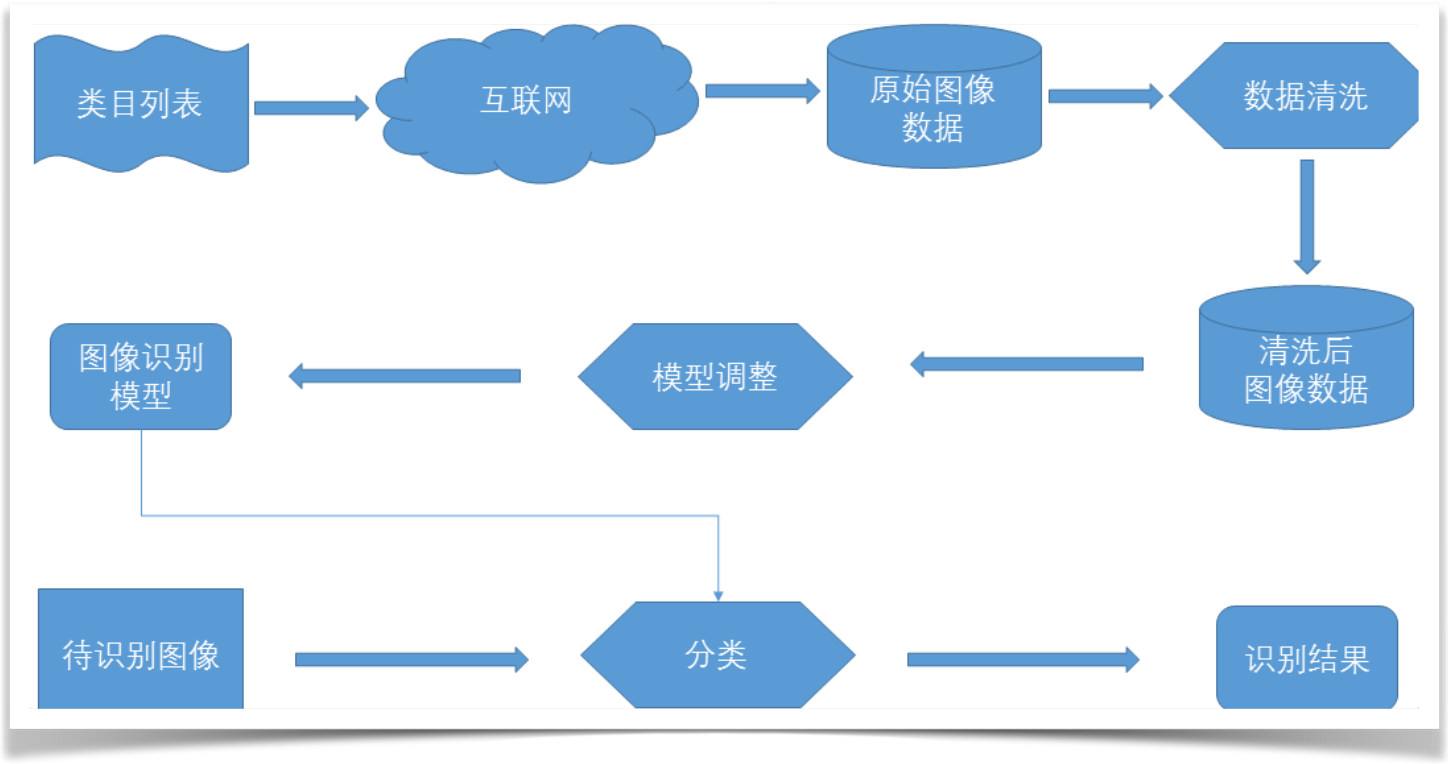


	Vstart	Vend	Order	Type
e_1	V_3	V_3	1	RG
e_2	V_3	V_2	2	AC
e_3	V_2	V_2	3	FW
...	...	...	...	...
e_{10}	V_1	V_3	10	FB
e_{11}	V_1	V_3	11	FB
e_{12}	V_1	V_3	12	AG

Clus	Maj \ Sec	S_{in}	S_{out}
V_1	M_1	S_2, S_3	F, S_1
V_2	M_2	S_1, S_2	S_2
V_3	M_3	F	S_1

Smart Contract Vulnerability Detection (IJCAI2020)

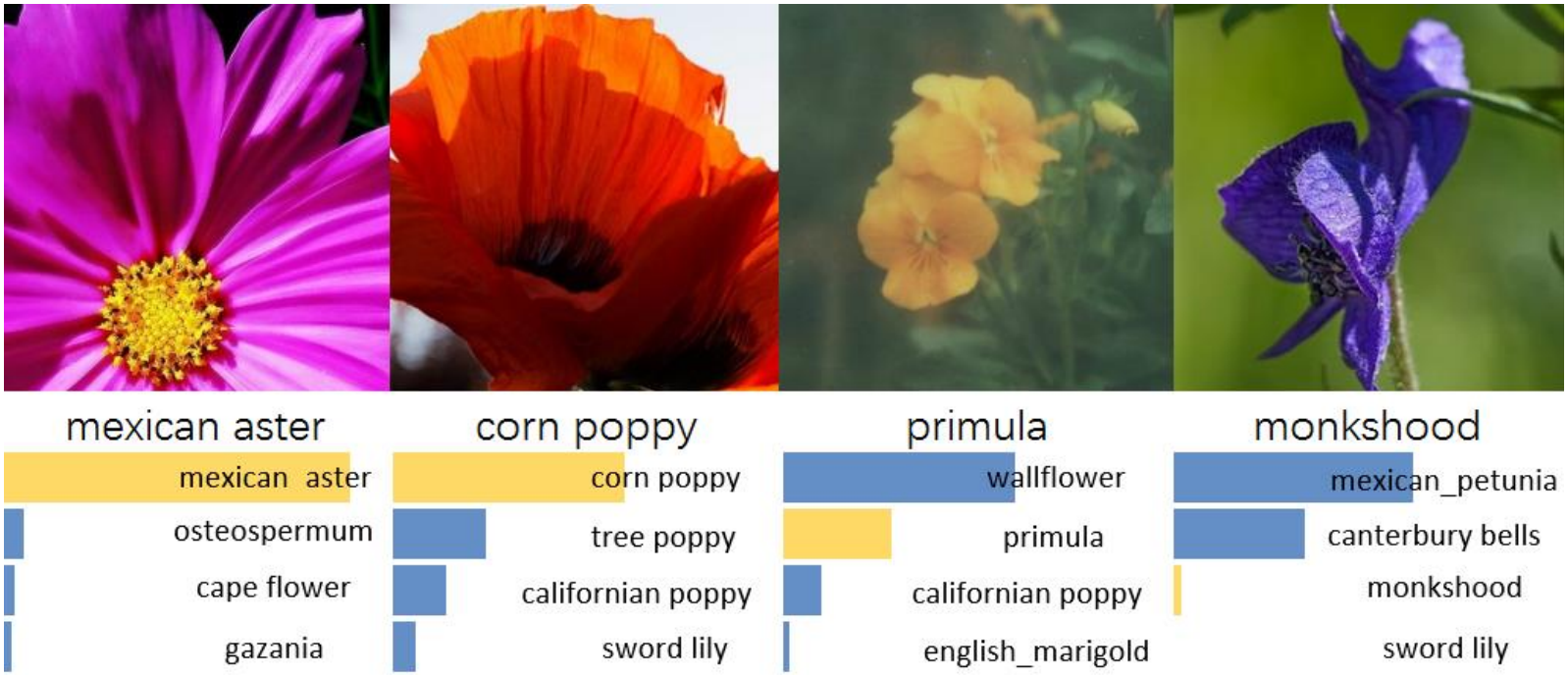
图像识别与深度学习



阿里云合作项目：互联网图像大数据的自动分析
(整体架构图)

来源：汪洋.大规模互联网图像自动识别技术研究[D].浙江大学,2017.

图像识别与深度学习

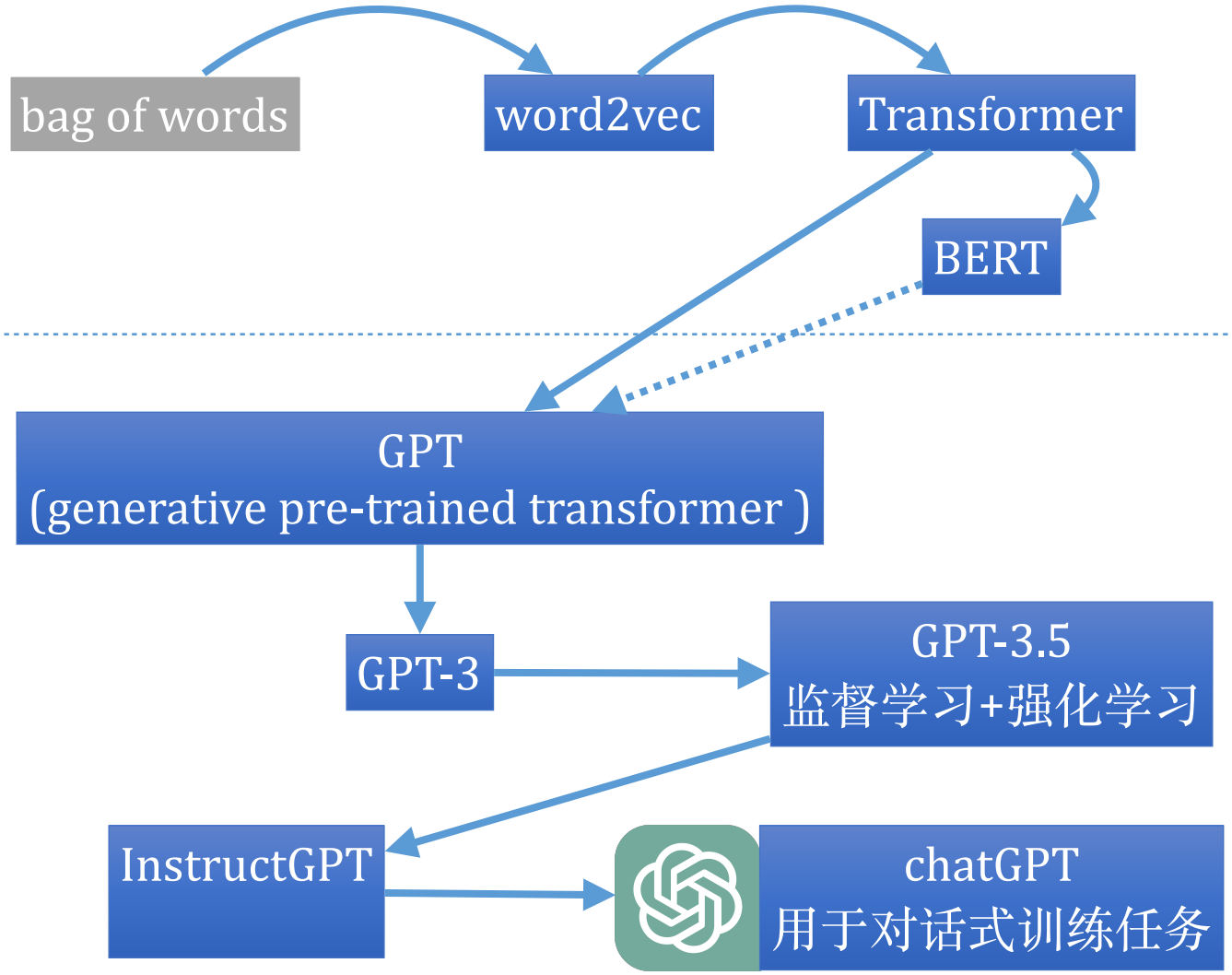


阿里云合作项目：互联网图像大数据的自动分析

应用：花的识别

来源：汪洋.大规模互联网图像自动识别技术研究[D].浙江大学,2017.

ChatGPT


bag of words
word2vec
Transformer
BERT
GPT
(generative pre-trained transformer)
GPT-3
GPT-3.5
监督学习+强化学习
InstructGPT
chatGPT
用于对话式训练任务

ChatGPT

Models referred to as "GPT 3.5"

GPT-3.5 series is a series of models that was trained on a blend of text and code from before Q4 2021. The following models are in the GPT-3.5 series:

1. `code-davinci-002` is a base model, so good for pure code-completion tasks
2. `text-davinci-002` is an InstructGPT model based on `code-davinci-002`
3. `text-davinci-003` is an improvement on `text-davinci-002`

InstructGPT models

We offer variants of InstructGPT models trained in 3 different ways:

TRAINING METHOD	MODELS
SFT Supervised fine-tuning on human demonstrations	`davinci-instruct-beta`[1]
FeedME Supervised fine-tuning on human-written demonstrations and on model samples rated 7/7 by human labelers on an overall quality score	`text-davinci-001`, `text-davinci-002`, `text-curie-001`, `text-babbage-001`
PPO Reinforcement learning with reward models trained from comparisons by humans	`text-davinci-003`

ChatGPT

GPT-3 Training Data

Dataset	# Tokens	Weight in Training Mix
Common Crawl	410 billion	60%
WebText2	19 billion	22%
Books1	12 billion	8%
Books2	55 billion	8%
Wikipedia	3 billion	3%

The model was trained on:

- Books1 - also known as BookCorpus. Here's a paper on BookCorpus, which maintains that it's free books scraped from smashwords.com.
- Books2 - No one knows exactly what this is, people suspect it's libgen
- Common Crawl
- WebText2 - an internet dataset created by scraping URLs extracted from Reddit submissions with a minimum score of 3 as a proxy for quality, deduplicated at the document level with MinHash
- What's in MyAI Paper, Source - Detailed dive into these datasets.

Large AI Models

Generated by Stable Diffusion AI

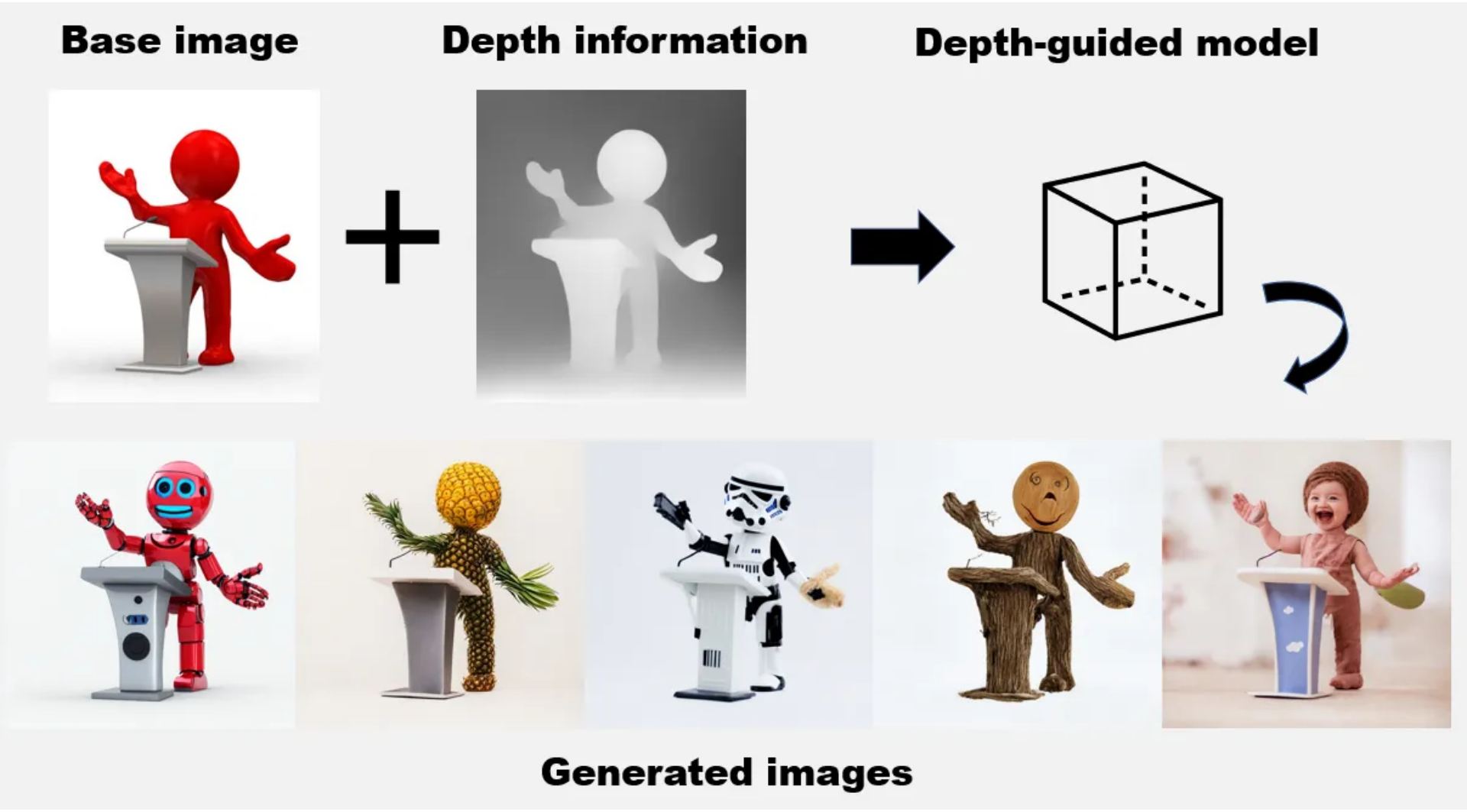
Large AI Models

Generated by Stable Diffusion V2

Large AI Models

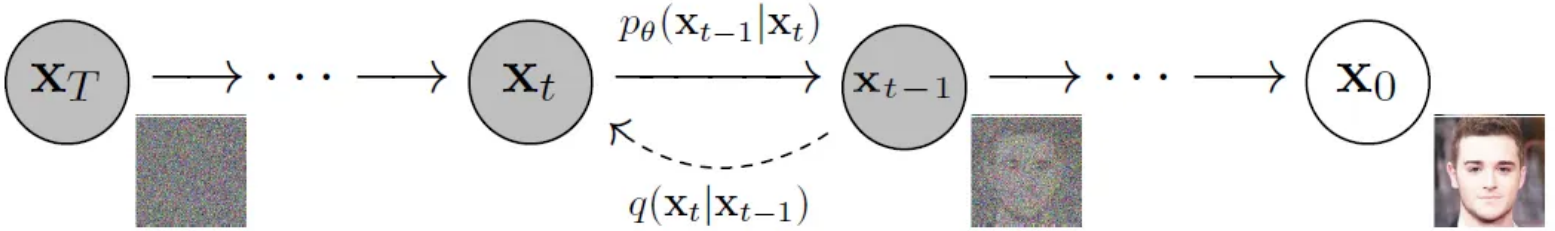
Generated by Stable Diffusion V2

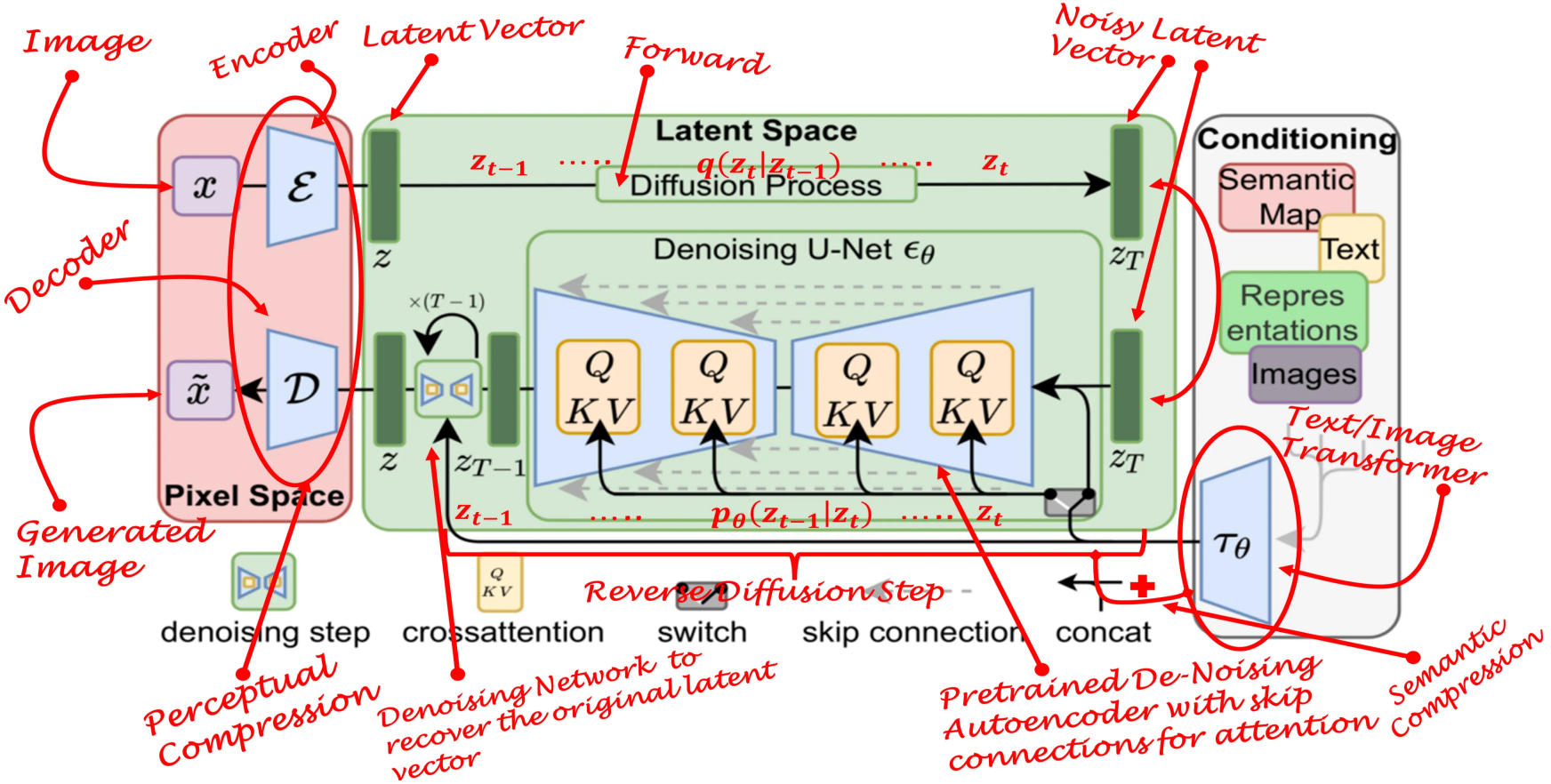
Big AI Models



Stable Diffusion v1 版本的模型单次训练需要使用A100 GPU总计150000小时，其背后公司Stability AI为了运营，拥有一个由4000块A100组成的GPU集群（估计5000万美元，即3.6亿元）

Large AI Models





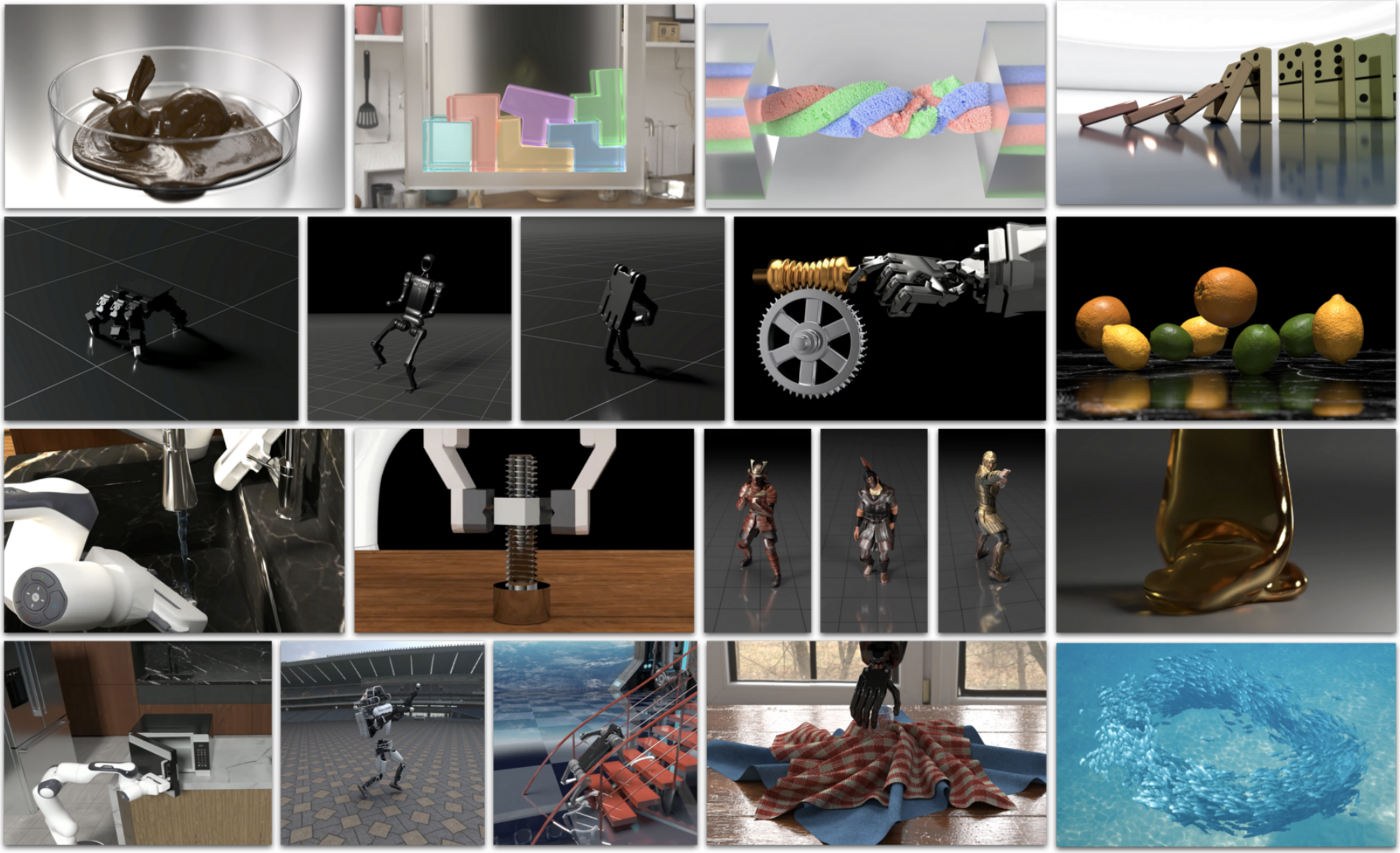
Large AI Models

Sora 生成的视频范例
提示词：「两艘海盗船在一个咖啡杯中航行、互相战斗的逼真特写视频。」
来源：*https://twitter.com/DrJimFan/status/1758210245799920123*

Large AI Models

模拟数字世界。Sora 还能模拟人工进程，视频游戏就是一个例子。Sora 可以通过基本策略同时控制 Minecraft 中的玩家，同时高保真地呈现世界及其动态。只需在 Sora 的提示字幕中提及「Minecraft」，就能零样本激发这些功能。

Large AI Models - Genesis



中国时刻已经到来

Marc Andreessen @pmarca · 1月24日

Deepseek R1 is one of the most amazing and impressive breakthroughs I've ever seen — and as open source, a profound gift to the world.

1,207 4,727 3.5万 1,061万

Sam Altman @sama · 1月28日

deepseek's r1 is an impressive model, particularly around what they're able to deliver for the price.

we will obviously deliver much better models and also it's legit invigorating to have a new competitor! we will pull up some releases.

6,934 1万 8.7万 1,413万

Yann LeCun
@ylecun

It's actually very good for Meta and the rest of the open source AI world. Since DeepSeek published the techniques and open sources the code, whatever innovation they came up with will make it to other models, including Llama.
Everyone wins.
The only parasites are the ones who profit from open innovation but are secretive themselves.

Andrej Karpathy
@karpathy

DeepSeek (Chinese AI co) making it look easy today with an open weights release of a frontier-grade LLM trained on a joke of a budget (2048 GPUs for 2 months, $6M).

For reference, this level of capability is supposed to require clusters of closer to 16K GPUs, the ones being brought up today are more around 100K GPUs. E.g. Llama 3 405B used 30.8M GPU-hours, while DeepSeek-V3 looks to be a stronger model at only 2.8M GPU-hours (~11X less compute). If the model also passes vibe checks (e.g. LLM arena rankings are ongoing, my few quick tests went well so far) it will be a highly impressive display of research and engineering under resource constraints.

Does this mean you don't need large GPU clusters for frontier LLMs? No but you have to ensure that you're not wasteful with what you have, and this looks like a nice demonstration that there's still a lot to get through with both data and algorithms.

Very nice & detailed tech report too, reading through.

翻译帖子

DeepSeek @deepseek_ai · 2024年12月26日

Introducing DeepSeek-V3!

Biggest leap forward yet:
60 tokens/second (3x faster than V2!)
Enhanced capabilities...

显示更多

Summary

- Learning (from Data) is a nut-shell, :-D
  - Keywords
    - Noun: data, models, patterns, features;
    - Adj.: probabilistic, statistical;
    - Verb: fitting, reasoning, mining

Human in the Loop

- 反思：是否要从数据驱动方法再回归到模型学习?
  - 可解释性
  - 算力需求
  - 系统、结果的可控性

中国传统中的统计（数据）思维 - 李舰

- https://cosx.org/2019/05/beauty-of-statistics/

英国学者李约瑟研究中国科技史时提出了一个问题：“尽管中国古代对人类科技发展做出了很多重要贡献，但为什么科学和工业革命没有在近代的中国发生？”这就是著名的李约瑟难题（Needham's Grand Question）。具体地说，是问“为什么近代科学没有产生在中国，而是在 17 世纪的西方，特别是文艺复兴之后的欧洲？”李约瑟通过对中国科学技术史的研究，在社会制度和地理环境中寻找答案。但这个问题一直被国人拿来反思自己的文化和传统，很多人都分析出了各种原因，大多数人认为中国的传统文化中缺少科学精神、甚至没有能够产生现代科学的基因，再结合现实生活中的各种乱象，无不痛心疾首，都想治病救人。

让我们把时间拉回到百年前的中国，轰轰烈烈的新文化运动已经开始，“德先生”和“赛先生”进了中国。国人深切地认识到了科学的威力，无数仁人志士立志向学，1923 年的“科玄之争”更是加速了科学在全民中的普及。当时“科学派”的观点不仅仅是科学在实业中的价值，更是要全面介入人们的生活。当然，当时的“玄学”也不是指魏晋那套老庄玄学和今天人们认为的旧中国玄学，而是指“在欧洲鬼混了二千多年的无赖鬼”，也就是形而上学。这次科玄之争可以说力度非常大，当时国人对科学的信仰程度超乎今天人们的想象。新中国成立后，对全民进行科学教育的成就更是有目共睹，中国的科技水平也是发展神速，但是如今国民科学素质的情况似乎仍然不容乐观，很多科普作者越科普越心焦，质疑中国科学精神的言论也仍然甚嚣尘上。

国民的科学素养真的这么差吗？科学素养的缺失真的是传统文化带来的吗？我看都不见得。梁启超在东南大学时，学生罗时实认为国粹将亡，因为读经的人太少了，梁启超闻声大怒，拍案道：“从古就是这么少”。当然，科学相比于经学更值得普及，但是对普通民众缺乏专业的科学知识不应苛责，这是正常现象。不同科学领域、不同知识内容的科普是一项漫长而有意义的事业，更需要普及的可能是科学思维。科学思维虽然与任何形式的玄学都水火不容，但也并不等于“死理性派”，也不是“死的机械论”，不能说演绎法是科学而归纳法就不是科学，也不能说理性主义是科学而经验主义就不是科学。不同的历史文化可能侧重不同，我们不能因为中国历史上三百年的特殊时期就质疑整个历史的科技成就，也不能因为中国传统公理体系的缺失就否认整个文化的科学精神，这是不科学的做法，也属于没有文化自信的表现。

中国传统中的统计（数据）思维 - 李舰

- https://cosx.org/2019/05/beauty-of-statistics/

也许是因为科学这个词听起来太高大上，也可能是科学比较接近真理，现在很多科普过于强调精确科学或者“硬”科学，有时候站在了普通人直觉或者经验的对立面，更侧重理性主义和演绎推理。这种精神放在一百年前的蒙昧期是合适的，放在今天全民教育水平不低的情形下可能有些矫枉过正，我觉得还是允执厥中比较好。能够在概念世界和知觉世界中达到和谐、能够在演绎法与归纳法中达到平衡，统计学可能是一个很好的桥梁。如今无论是自然科学还是社会科学都离不开统计学，尤其在应用领域，直接掀起了大数据的热潮，技术层面的威力已经深入人心，但是思维方面的普及还有所不足。实际上，对中国人来说，理解统计思维似乎是一件非常轻松的事，无论是上古伏羲观天法地的归纳精神，或者神农尝百草的试验精神，还是后世天人合一的整体思维、观过知仁的结果导向、未战而庙算的预测习惯，都是深合统计之道的。

很多人受到各种原因的误导之后对中国的文化不自信，易于走向“言必称希腊，对于自己的祖宗，则对不住，忘记了”的极端，这是不对的。即使是作为很多科学基础的数学，也不止一种思维方式。数学家吴文俊院士说过“我国古代数学并没有发展出一套演绎推理的形式系统，但却另有一套更有生命力的系统”，这个生命力就是“从实际中发现问题，提炼问题，进而分析问题和解决问题”， 完全不同于希腊几何学纯逻辑推理的形式主义道路，中国数学的经典著作大都是以问题集的形式出现的，对结果不是用定理来表达的，而是用“术”来表达的，用现代的话来讲就是程序，与近代计算机的使用融合无间。可见中国传统的数学思维是非常适合现在这个算法时代的。算法与统计的结合造就了机器学习、人工智能的大爆发，甚至可以说是主导了这个时代的科技应用方向。统计学家约翰·图基 1962 年的文章中指出，任何数理统计学工作都应该在纯数学或者数据分析的实践中二选一，两个标准都不符合的工作必然只是一时的过客。陈希孺院士也曾预测“新一轮的突破性进展正在孕育中，它也许就是数据分析？”如今大师们的论断都已言中，统计学与算法结合解决实际问题，已经渐成主流，甚至发展出了一门新的学科——数据科学。

卡瓦列里原理在西方数学史中被认为是微积分发明前的重要基础，而中国的祖暅原理与之等价。莱布尼茨在提出二进制的那篇著名文章里直接引用了伏羲八卦，他还认为“如果说我们（欧洲人）在手工技能上与他们（中国）不分上下、在理论科学方面超过他们的话，那么，在实践哲学方面…… 我不得不汗颜地承认他们远胜于我们”。在这里我们无意比较中西的优劣，也并不是为了说明中国有多厉害（如果是这个目的的话，可以举更多例子或者写另一本书），仅仅只是为了澄清一些误解，这些误解既是对中国传统的某种误读，同时也是科学思维上的某类误区。我们追求理性和完美的体系，也希望能止于至善，但我们也不应忽视经验主义和观察、试验、归纳、计算的力量，这些都是科学，不应偏颇。尤其对于普通人来说，多从观察身边的小事、解决实际问题的角度训练科学思维，可能效果更好，毕竟“刻鹄不成尚类鹜，画虎不成反类狗”。

Questions for review

- Try to find potential learning based (data driven) applications in your research area
- Is there any disadvantage / weakness?

Reference

- Reinforcement learning: A survey

微博：@浙大张宏鑫
邮件：zhx@cad.zju.edu.cn
主页：http://person.zju.edu.cn/zhx
手机：13958011790
微信：timothykull

谢谢

Thank You