数据库遇到深度学习

—面向智能应用的多模态数据库

崇志宏

东南大学数据与智能实验室（D&Intel Lab@SEU）

chongzhihong@seu.edu.cn

cse.seu.edu.cn/PersonalPage/zhchong/index.htm

东南大学数据与智能实验室 (D&Intel Lab)

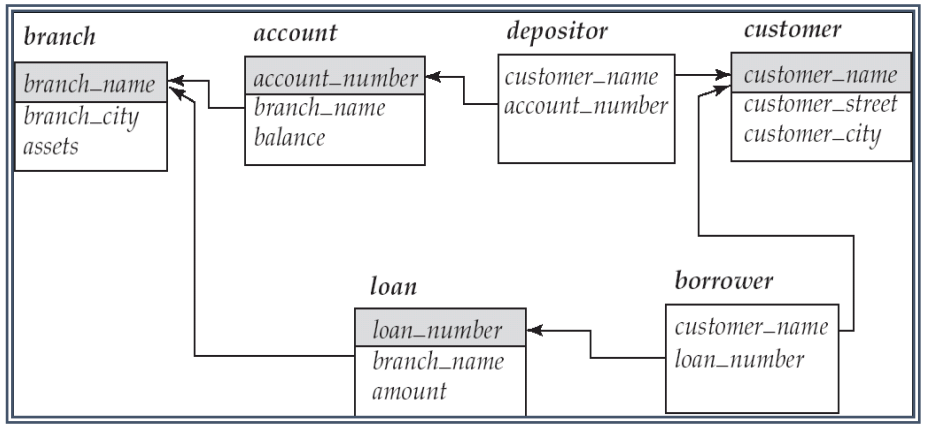
提纲

- 数据库和深度学习
  - 关系和流形结构
  - 关系表示和流形结构表示
- 多模态数据的语义关系
  - 多模态语义的层次组合结构表示
  - 相似和相关关系的表示
- 数据驱动的索引和查询策略优化
  - 索引学习
  - 查询策略优化

数据库和深度学习：关系和流形结构

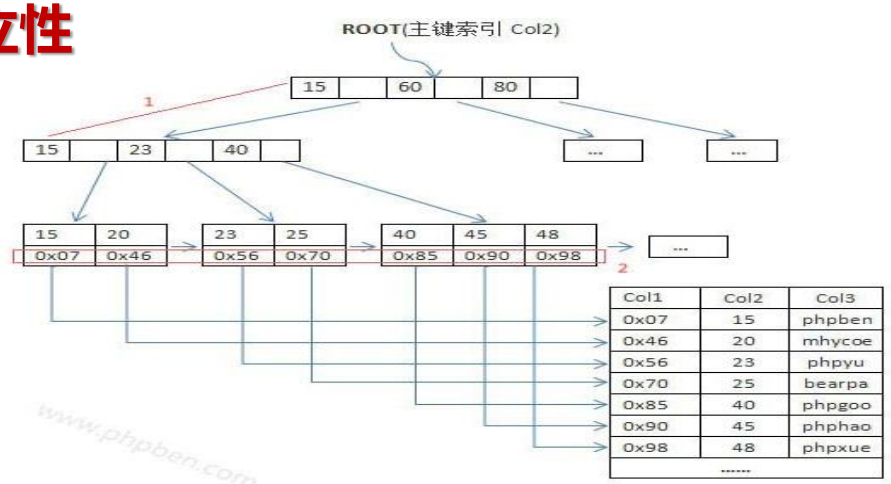
观察数据 → **隐藏结构**

逻辑模式

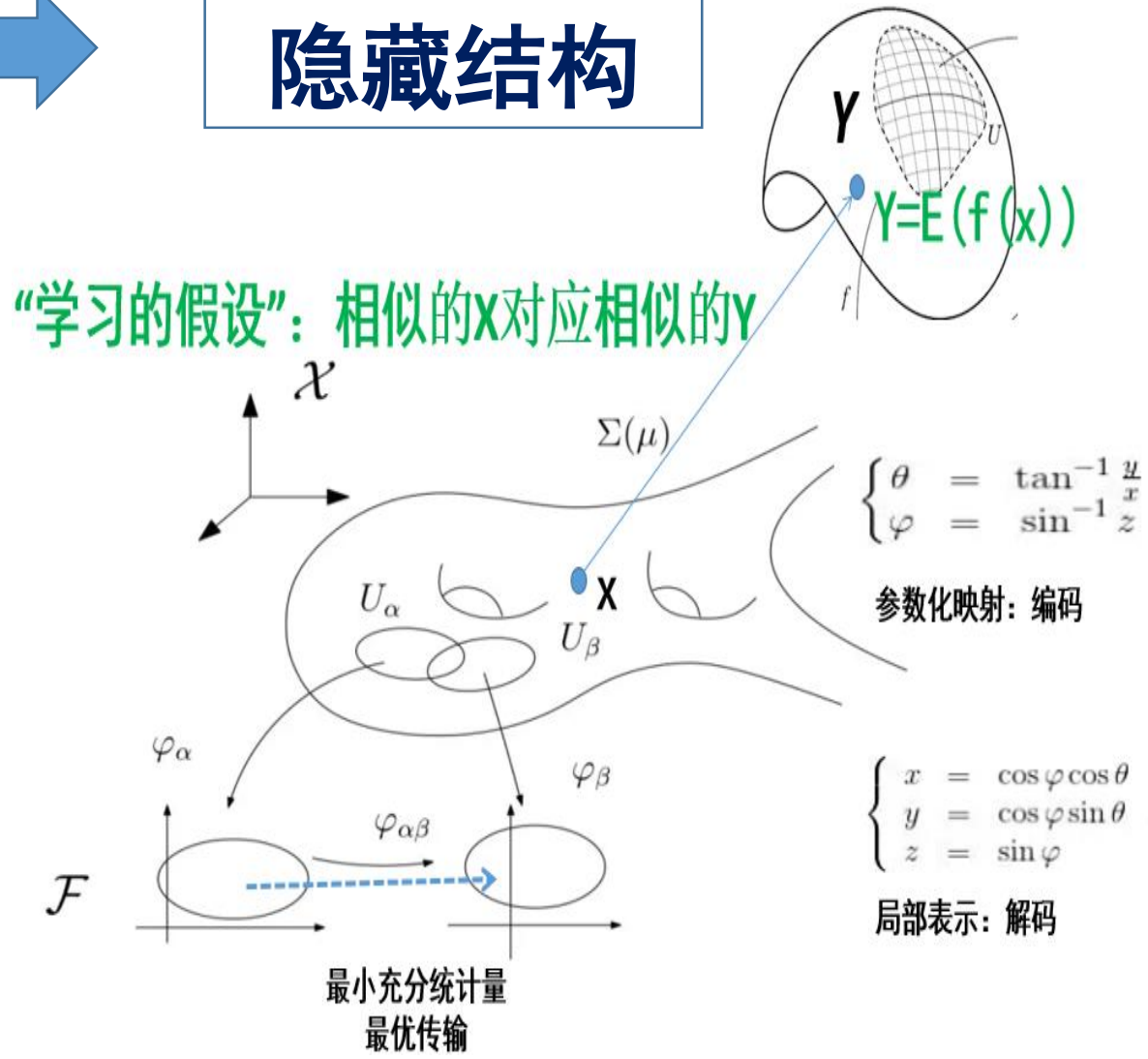


数据独立性

物理模式



工业级的强一致性协议：全球同步和保证高吞吐的一致性



Geometric deep learning: going beyond Euclidean data

流形结构的深度神经网络模型表示

1. 高维复杂函数的近似 2. 层次结构特征抽取

Deep Learning Refers to learning complicated concepts by building them from simpler ones in a hierarchical or multi-layer manner. Artificial neural networks are popular realizations of such deep multi-layer hierarchies.

$$y = f(x)$$

Using parallel computing techniques (e.g. GPU) to speed up matrix operation

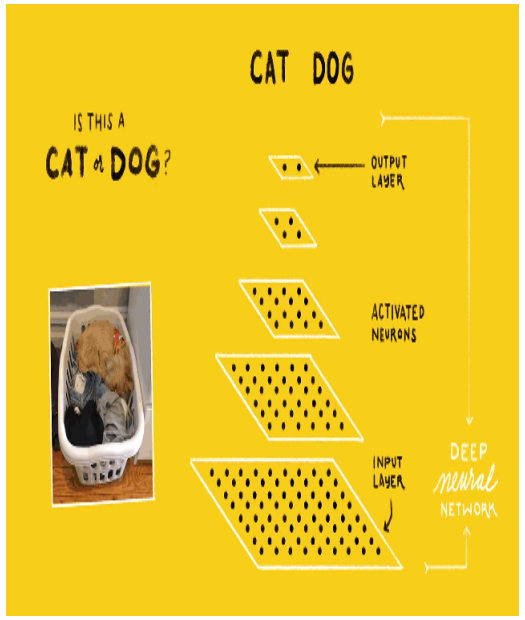
$$= \sigma(W^L \cdots \sigma(W^2 \sigma(W^1 x + b^1) + b^2) \cdots + b^L)$$

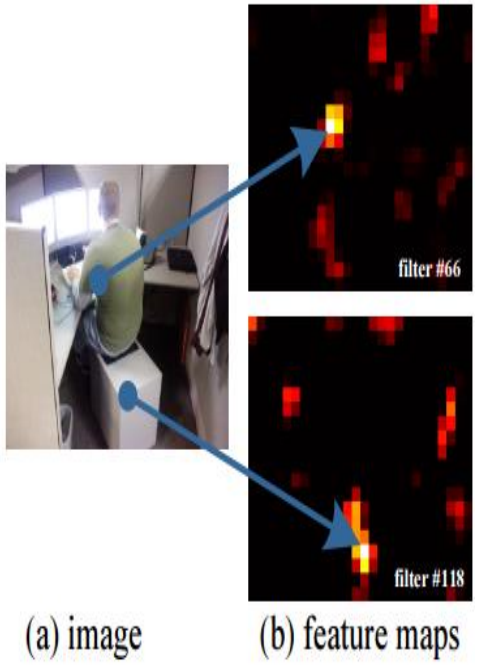
Geometric deep learning: going beyond Euclidean data

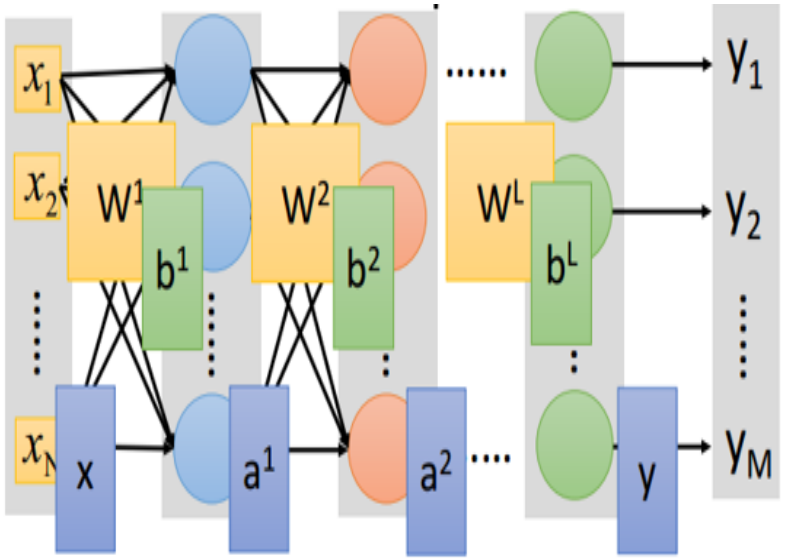
流形结构的深度神经网络模型表示

1. 高维复杂函数的近似 2. 层次结构特征抽取

投影算子、选择算子和组合算子！







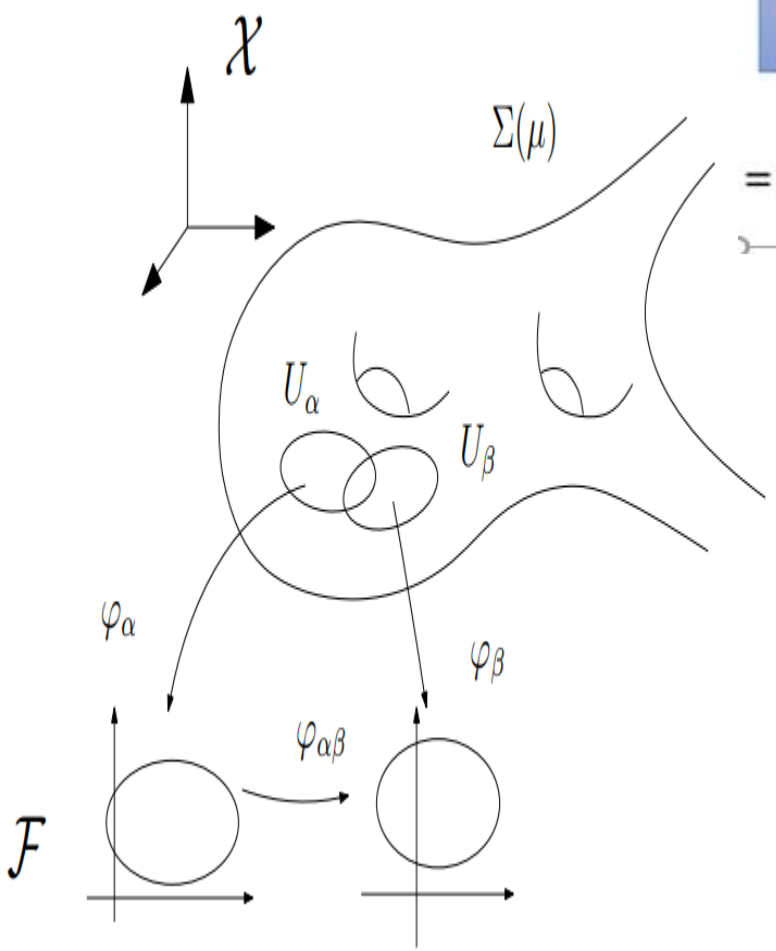
$$y = f(x)$$

Using parallel computing techniques (e.g. GPU) to speed up matrix operation

$$= \sigma(W^L \cdots \sigma(W^2 \sigma(W^1 x + b^1) + b^2) \cdots + b^L)$$

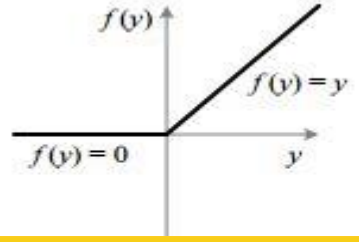
Deep Learning in Neural Networks: An Overview

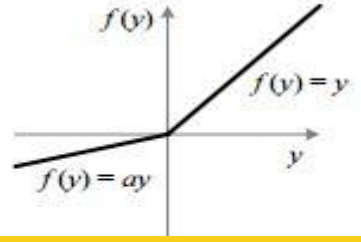
流形结构的深度神经网络模型表示

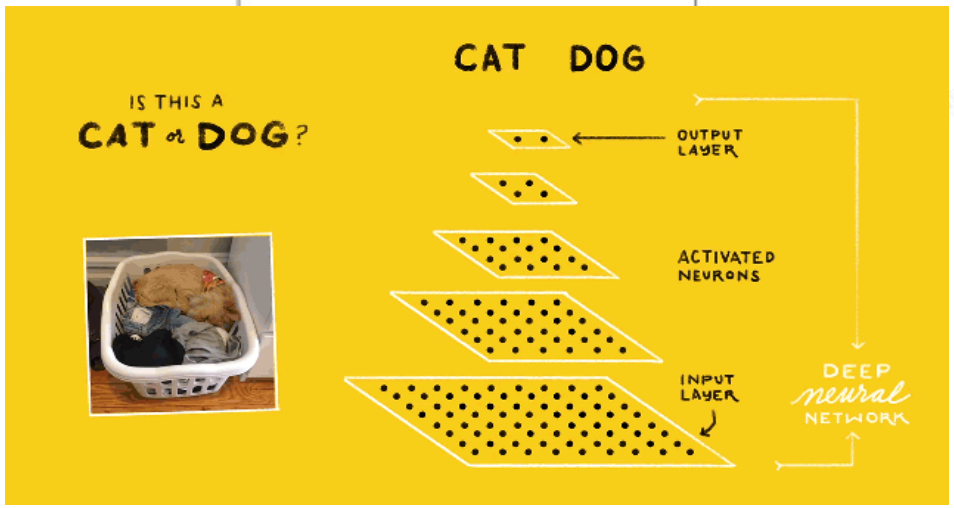


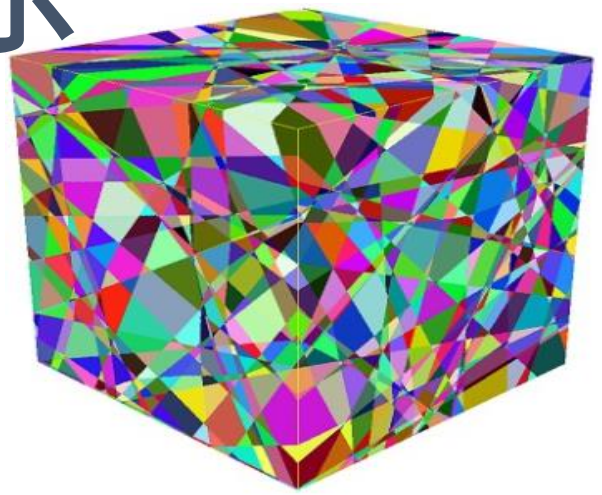
$$y = f(x)$$ Using parallel computing techniques (e.g. GPU) to speed up matrix operation

$$= \sigma(W^L \cdots \sigma(W^2 \sigma(W^1 x + b^1) + b^2) \cdots + b^L)$$

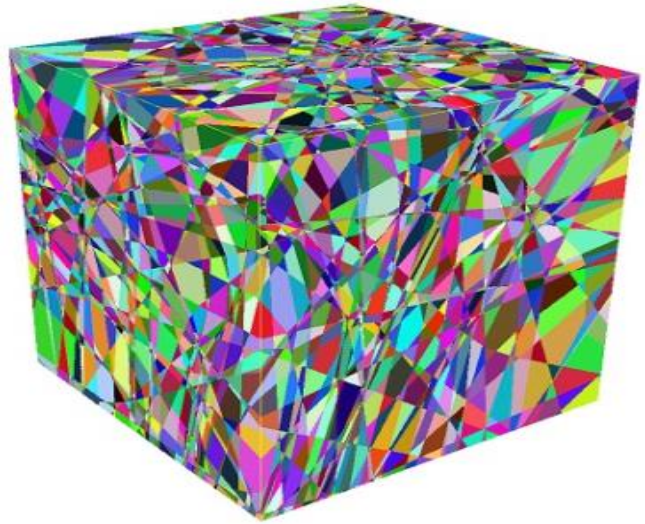








d. cell decomposition
$\mathcal{D}(\varphi_\theta)$



f. cell decomposition
$\mathcal{D}(\psi_\theta \circ \varphi_\theta)$

Geometric Understanding of Deep Learning

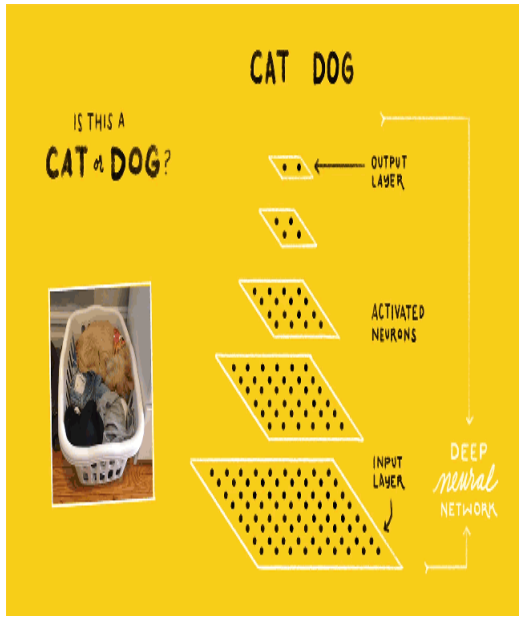
流形结构的深度神经网络模型表示

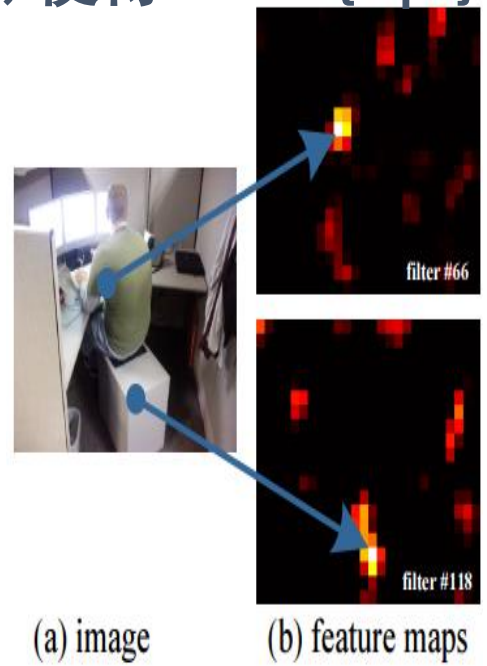
1. 高维复杂函数的近似 2. 层次结构特征抽取

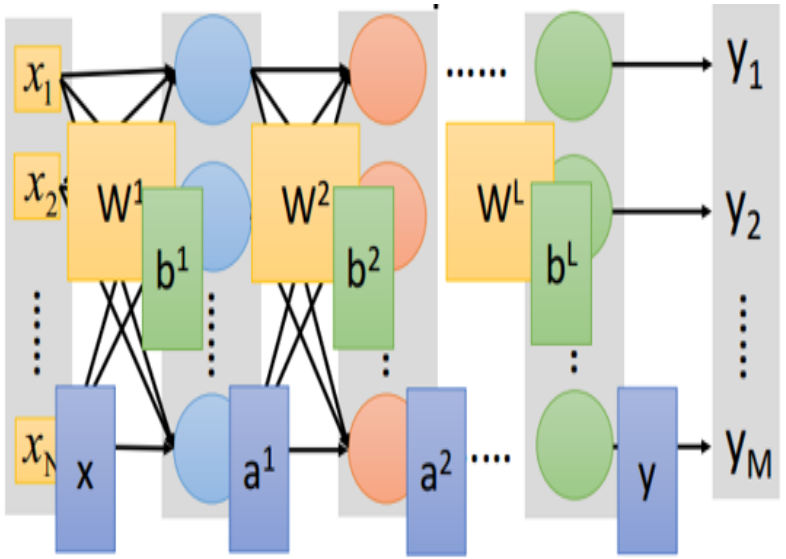
1、不确定关系的期望估计 $E(Y)=f_w(X)$

2、条件概率 $Pr\{Y|X\}=N(f_w(X), I)$

3、条件样本生成 $Y=f_w(X, e)$ 使得 $Y\sim Pr\{Y|X\}, e\sim N(0, 1)$







$$y = f(x)$$

Using parallel computing techniques (e.g. GPU) to speed up matrix operation

$$= \sigma(W^L \cdots \sigma(W^2 \sigma(W^1 x + b^1) + b^2) \cdots + b^L)$$

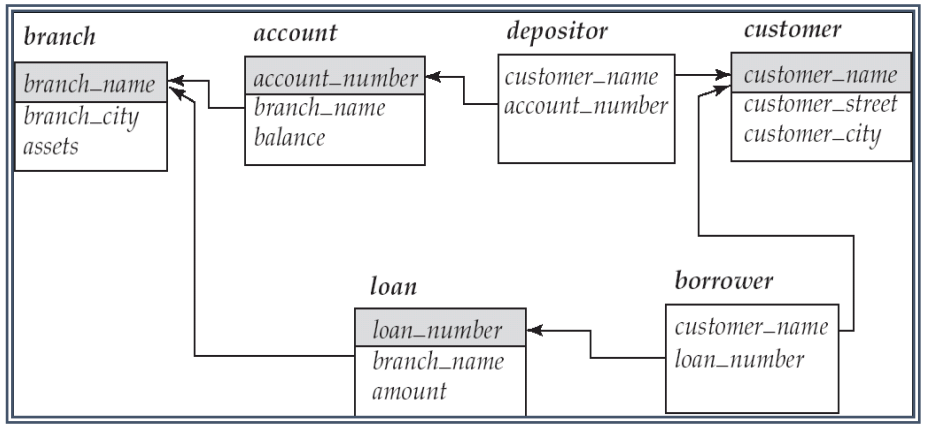
Geometric deep learning: going beyond Euclidean data

数据库和深度学习：关系和流形结构

观察数据

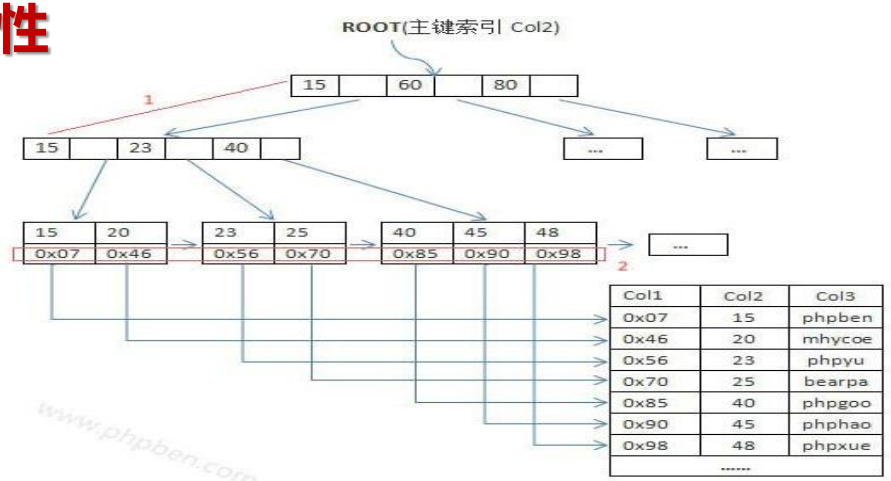
隐藏结构

逻辑模式

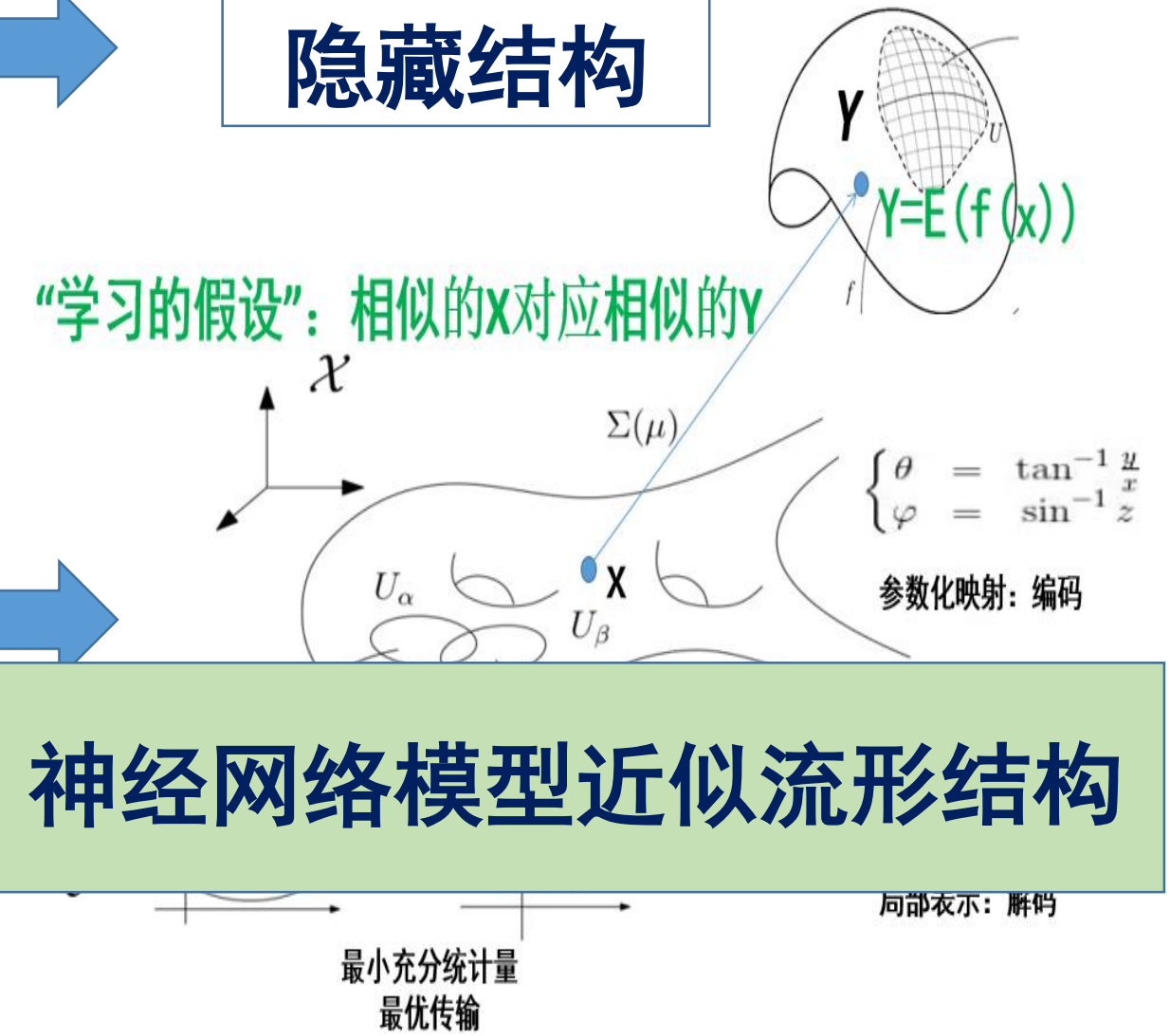


数据独立性

物理模式



工业级的强一致性协议：全球同步和保证高吞吐的一致性



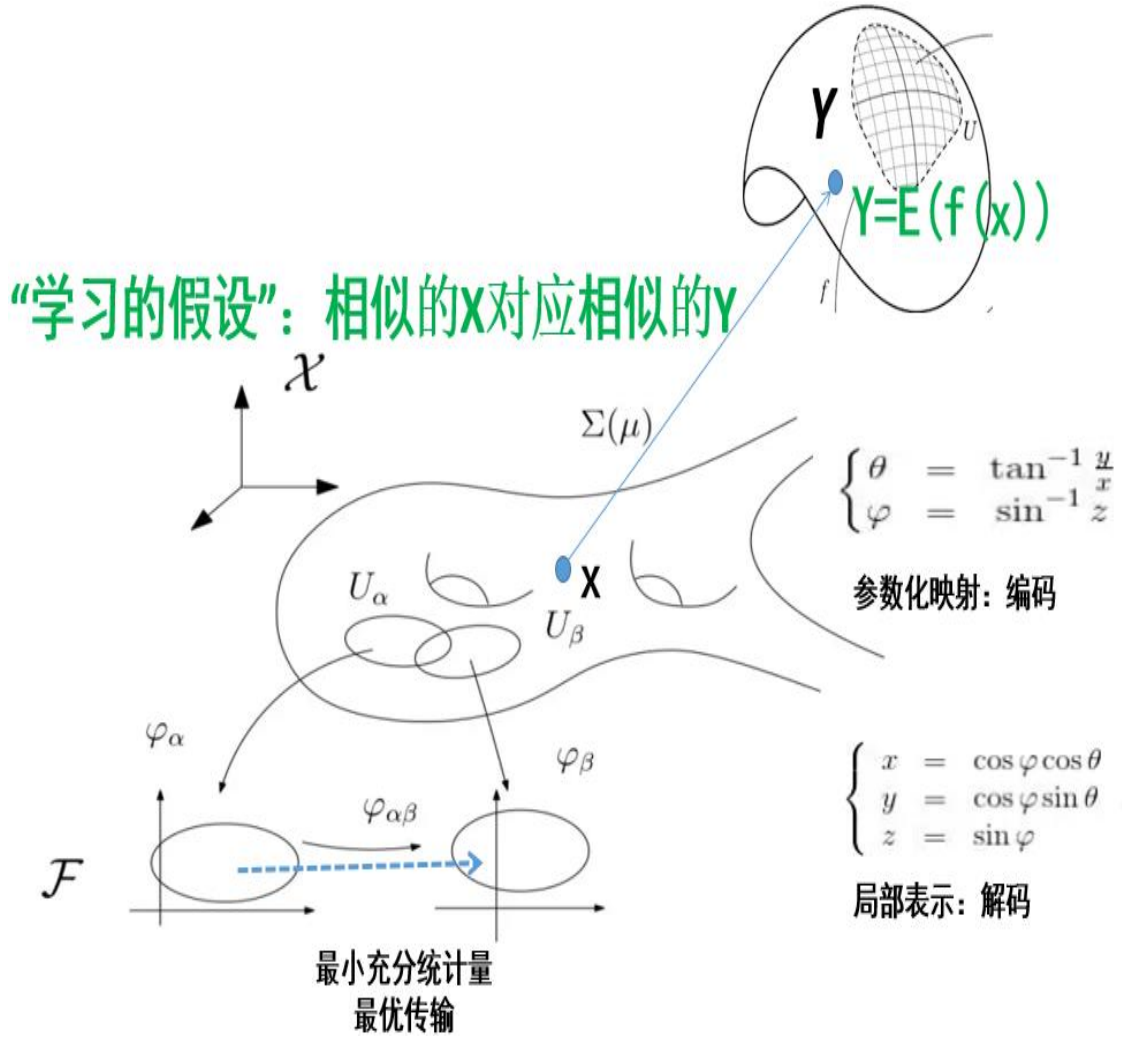
神经网络模型近似流形结构

Geometric Understanding of Deep Learning

提纲

- 多模态数据库和深度学习
  - 关系和流形结构
  - 关系表示和流形结构表示
- **数据的语义关系**
  - **多模态语义的层次组合结构表示**
  - **相似和相关关系的表示**
- 数据驱动的索引和查询策略优化
  - 索引学习
  - 查询策略优化

多模态数据的相似或相关关系表示



一、相似关系

相似的X对应相似的Y

二、相关关系：

$$Pr\{Y|X\} \approx Pr\{Y|X`\}\,Pr\{X` \approx X\}$$

$$\approx Pr\{Y|X`\} \mid \varphi_\alpha(X`) \approx \varphi_\alpha(X)\}$$

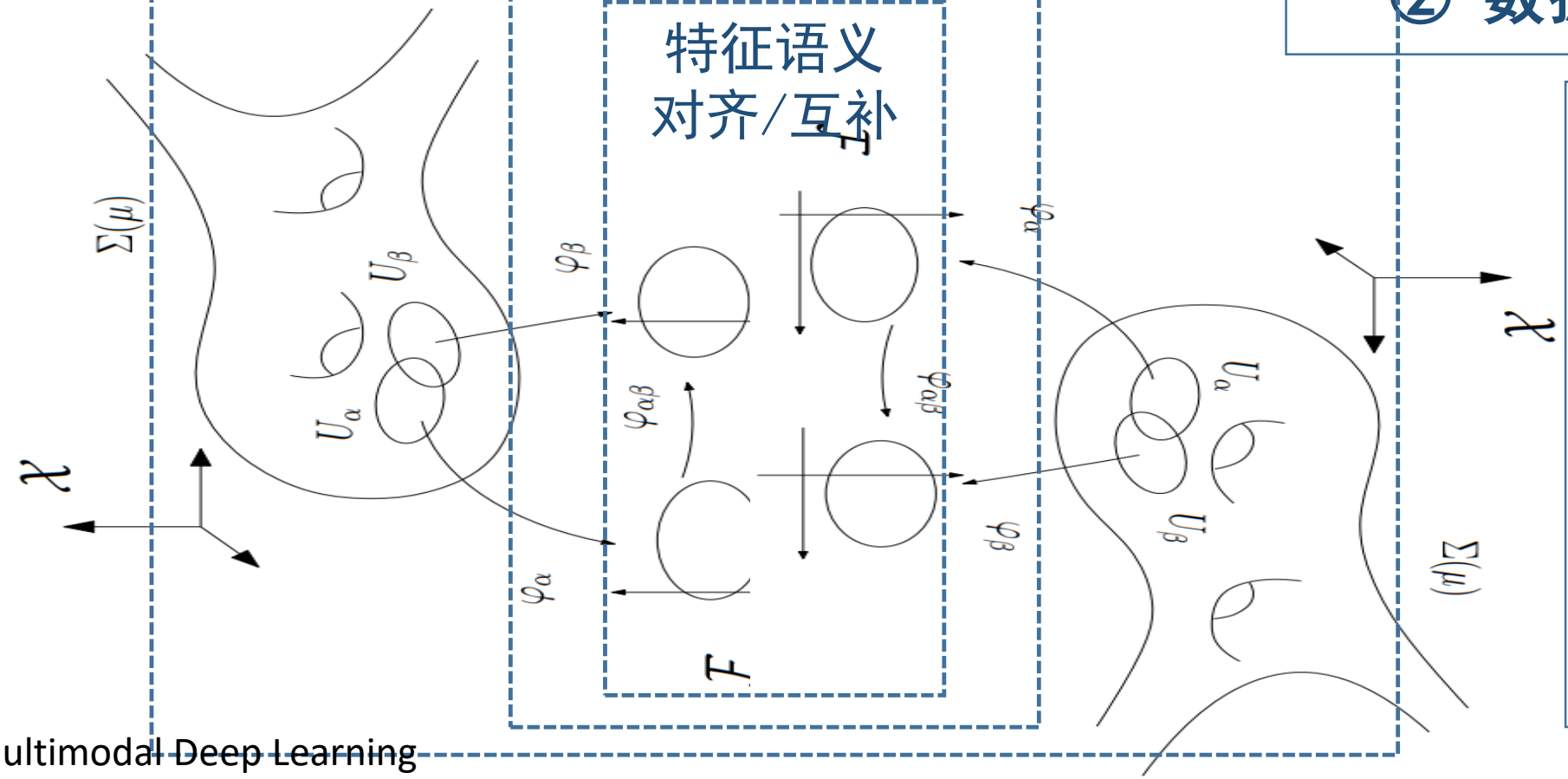
Geometric Understanding of Deep Learning

多模态的相似或相关关系

对齐/互补/冲突关系（1:1/1:n/n:1/n:n）

模型关系

特征语义
对齐/互补



① 模式匹配

② 数据合并

① 共同特征语义关系

② 多视图语义关系

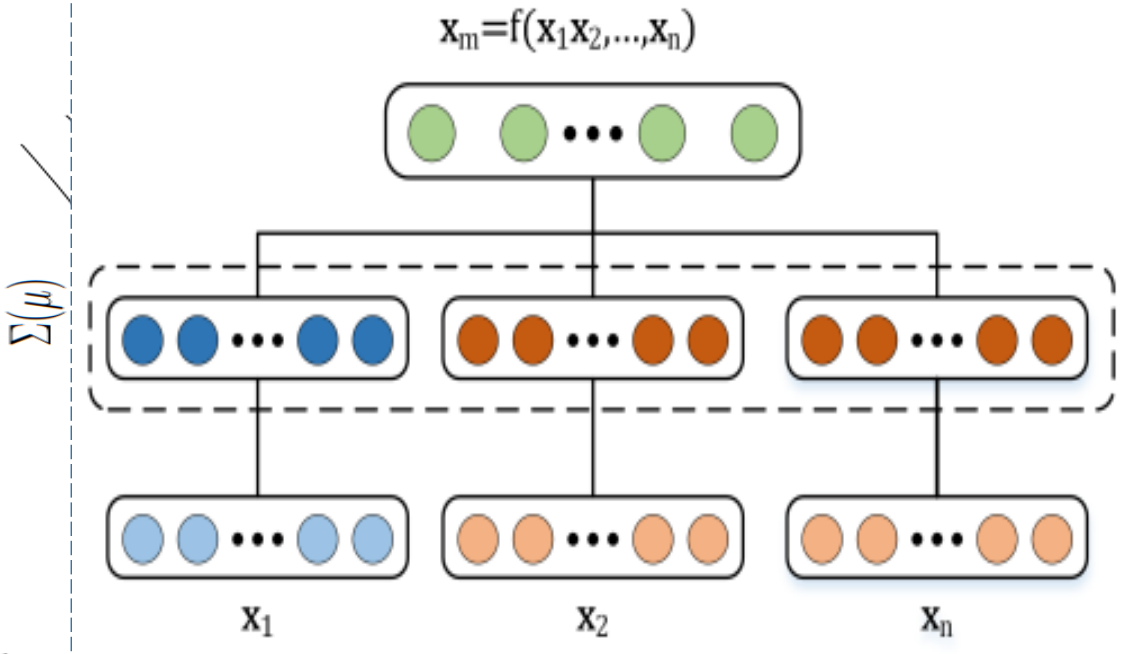
③ 概率依赖关系

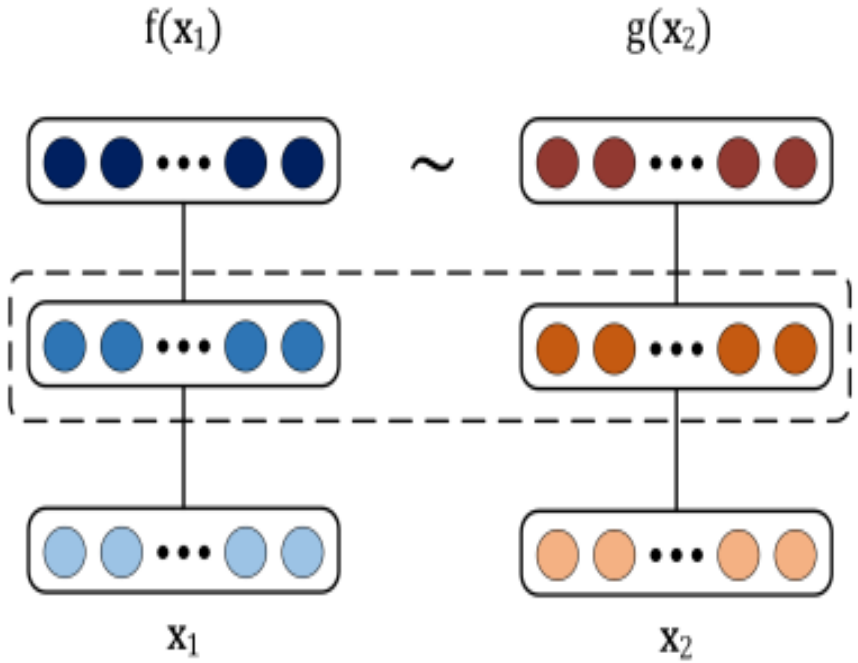
④ 迁移关系

多模态的相似或相关关系

对齐/互补/冲突关系（1:1/1:n/n:1/n:n）

① 模式匹配





义关系

关系

系

④ 迁移关系

多模态数据的相似或相关关系表示

$$f(x) = \phi_y \circ g \circ \phi_x^{-1}(x)$$

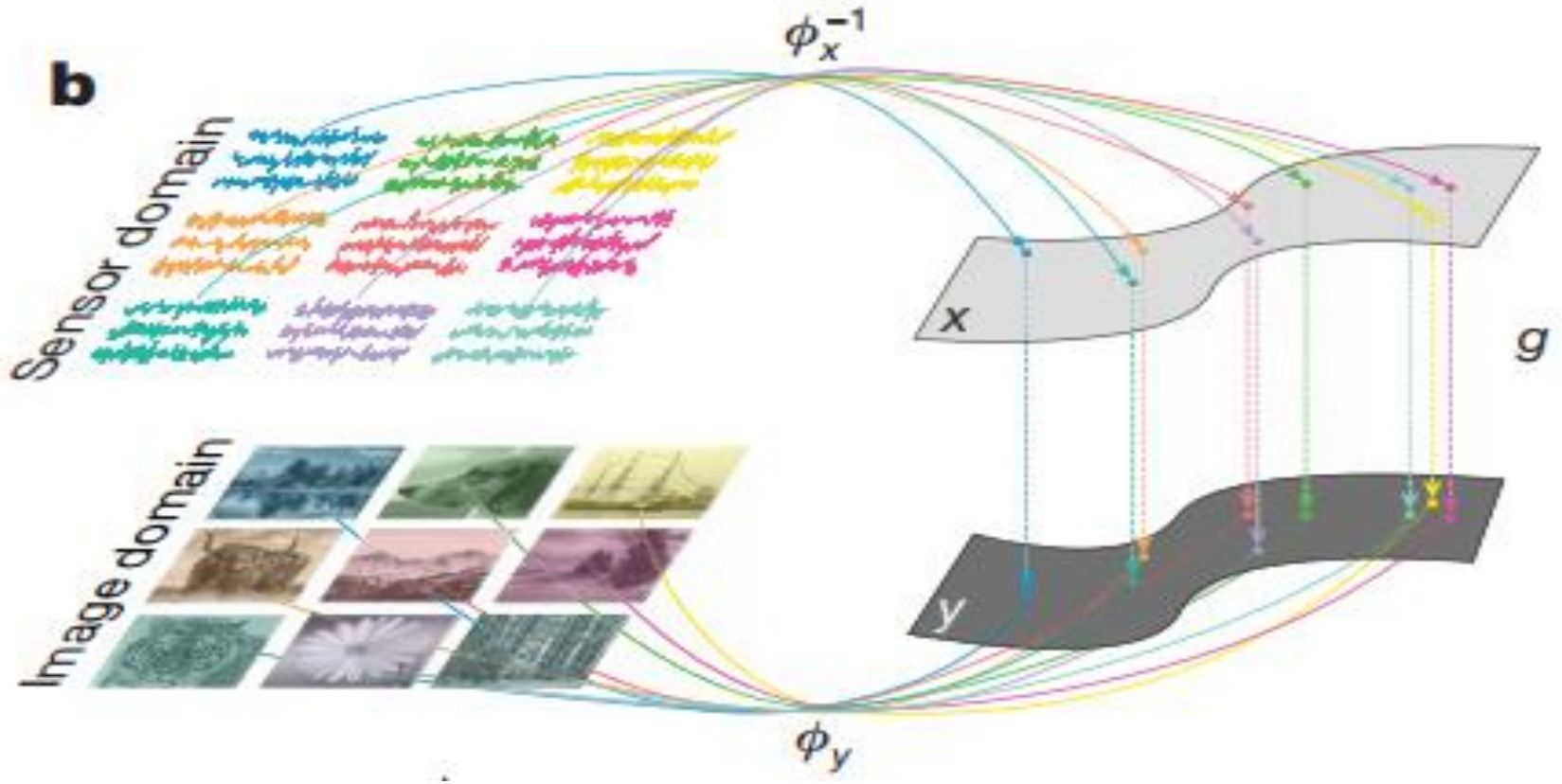


Image reconstruction by domain-transform manifold learning

数据库与深度学习

关系数据模型
（一阶结构化查询）

字符 匹配

结构化数据

结构化查询

R(I,Q)

非结构化检索

非结构化数据

语义 相似

语义关系模型
（相似、相关、距离）

数据库与深度学习

Mission Critic Application

效率

关系数据模型
（一阶查询）

字符 匹配

Mission Refined Application

效果

结构化数据

结构化查询

R(I,Q)

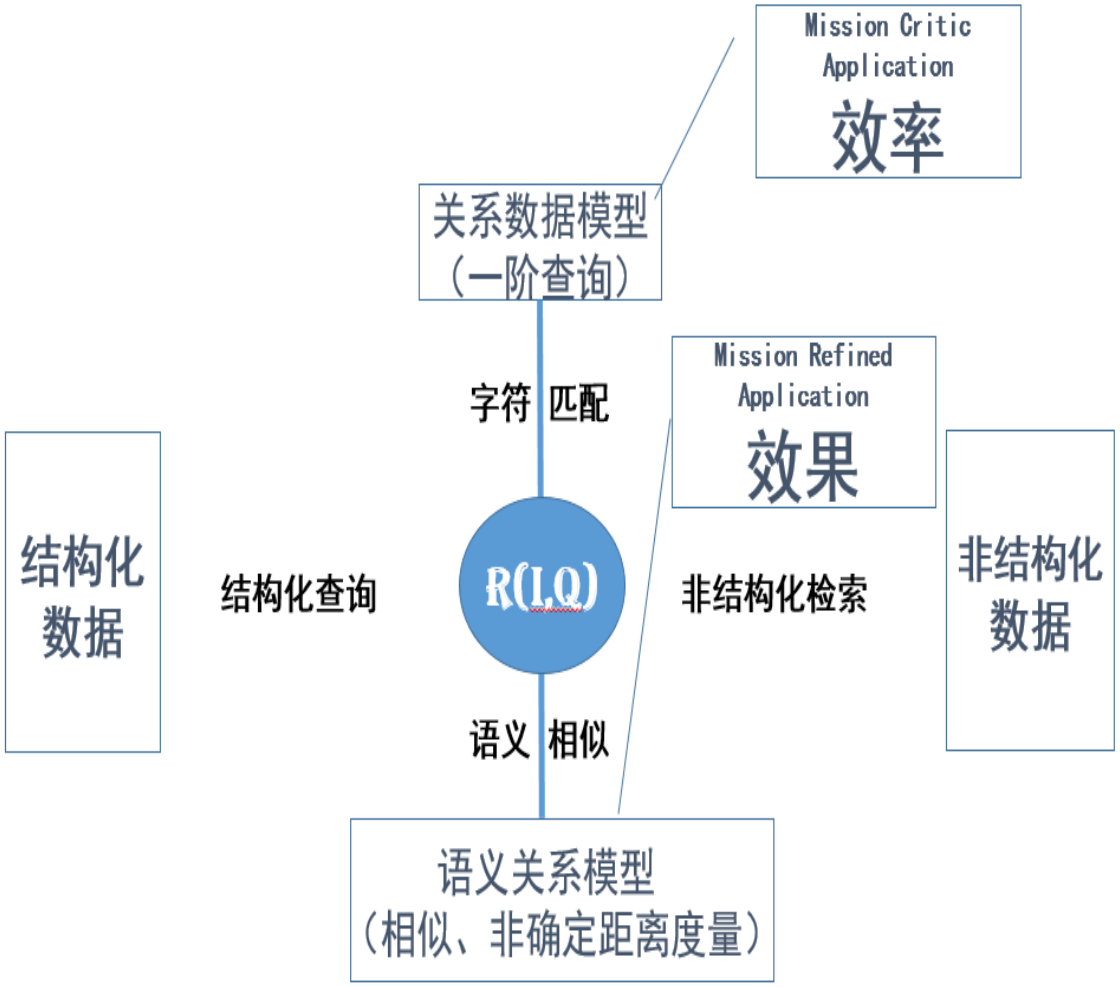
非结构化检索

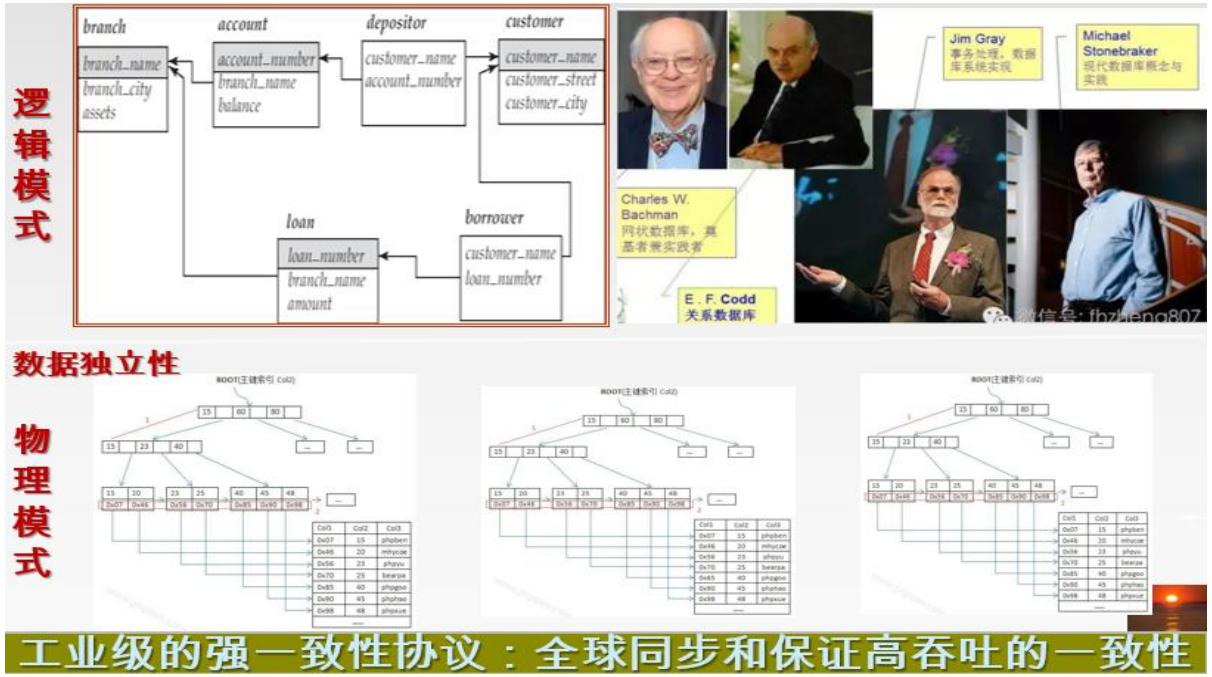
非结构化数据

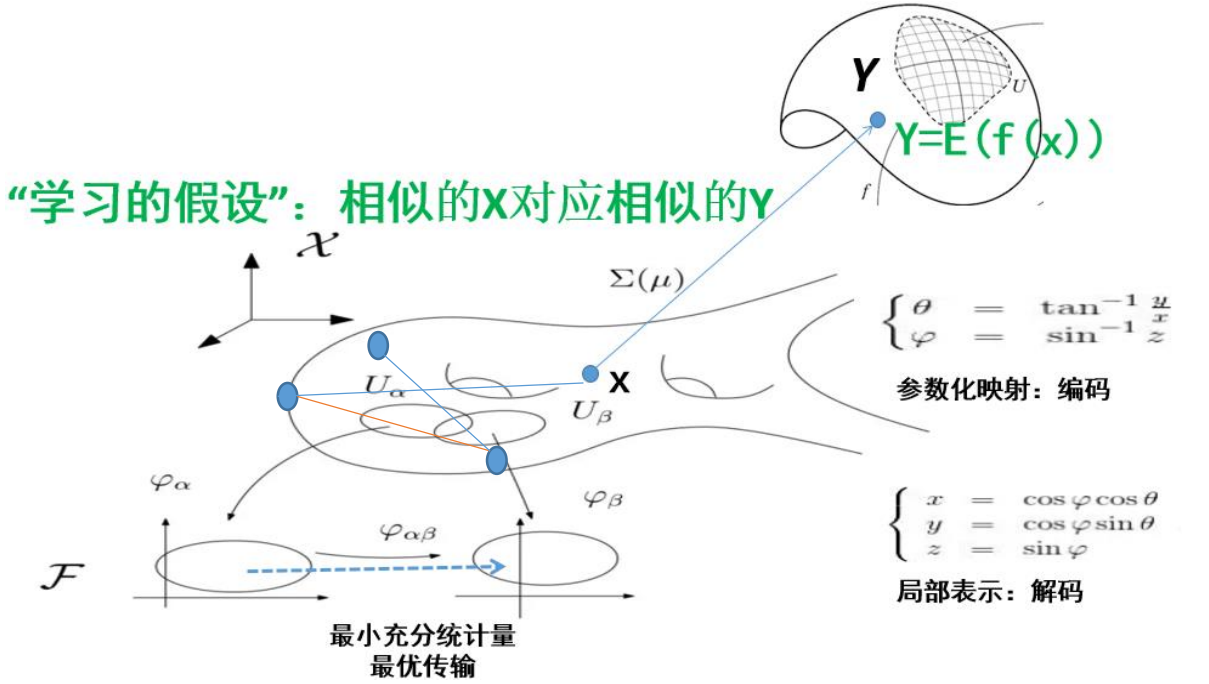
语义 相似

语义关系模型
（相似、非确定距离度量）

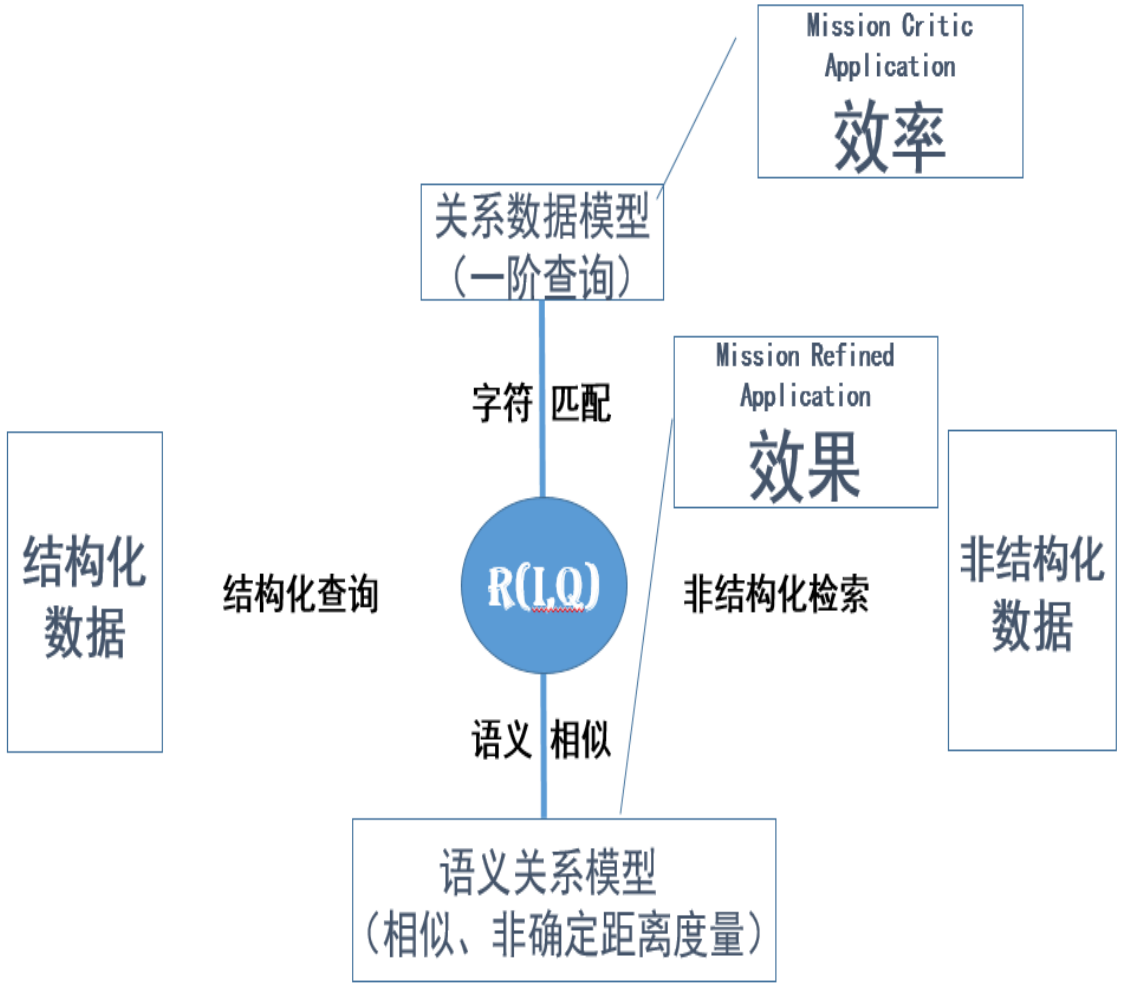
关系和流形结构


Mission Critic Application
效率
关系数据模型
（一阶查询）
字符 匹配
Mission Refined Application
效果
结构化数据
结构化查询
R(I,Q)
非结构化检索
非结构化数据
语义 相似
语义关系模型
（相似、非确定距离度量）


逻辑模式
branch
branch_name
branch_city
assets
account
account_number
branch_name
balance
depositor
customer_name
account_number
customer
customer_name
customer_street
customer_city
loan
loan_number
branch_name
amount
borrower
customer_name
loan_number
Jim Gray
事务处理：数据库系统实现
Michael Stonebraker
现代数据库概念与实践
Charles W. Bachman
网状数据库，真基者最实践者
E. F. Codd
关系数据库
数据独立性
物理模式
工业级的强一致性协议：全球同步和保证高吞吐的一致性


“学习的假设”：相似的X对应相似的Y
Y=E(f(x))
Y
U
f
X
$\mathcal{X}$
$\Sigma(\mu)$
U_α
U_β
φ_α
φ_β
$\varphi_{\alpha\beta}$
$\mathcal{F}$
$\begin{cases} \theta = \tan^{-1}\frac{y}{x} \\ \varphi = \sin^{-1}\frac{z}{z} \end{cases}$
参数化映射：编码
$\begin{cases} x = \cos\varphi\cos\theta \\ y = \cos\varphi\sin\theta \\ z = \sin\varphi \end{cases}$
局部表示：解码
最小充分统计量
最优传输

数据库与深度学习



关系模型

1、封闭假设 2、明确关系 3、精确匹配

效率

流形结构

1、开放假设 2、隐式关系 3、相似不确定

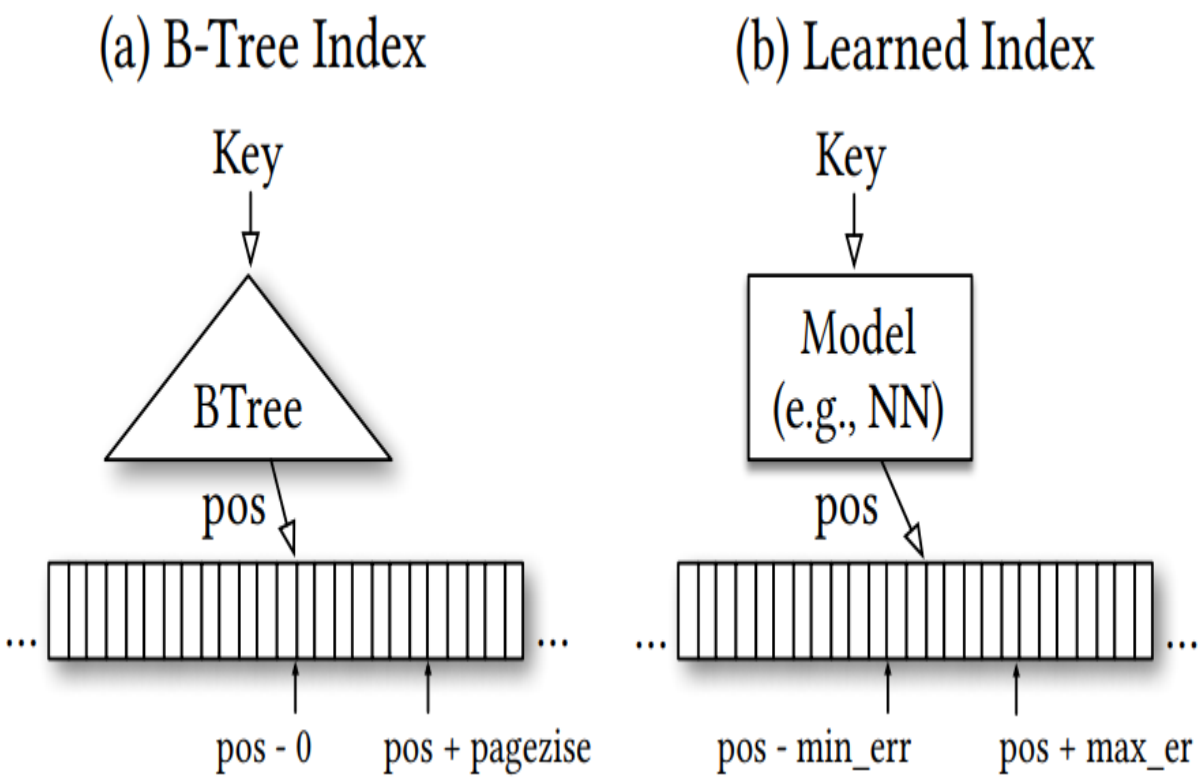
效果

提纲

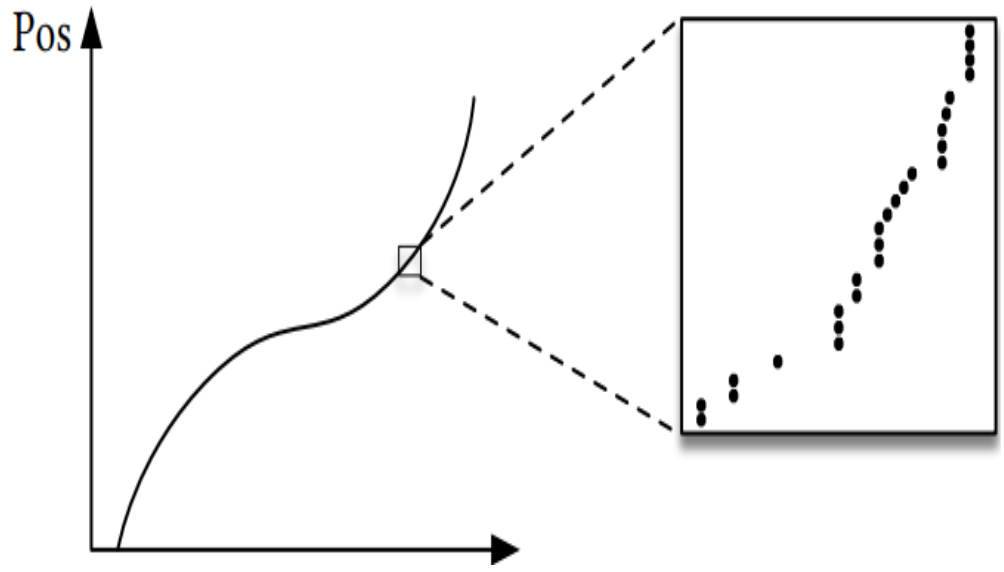
- 数据库和深度学习
  - 关系和流形结构
  - 关系表示和流形结构表示
- 多模态数据的语义关系
  - 多模态语义的层次组合结构表示
  - 相似和相关关系的表示
- 数据驱动的索引和查询策略优化
  - 索引学习
  - 查询策略优化

数据分布和索引

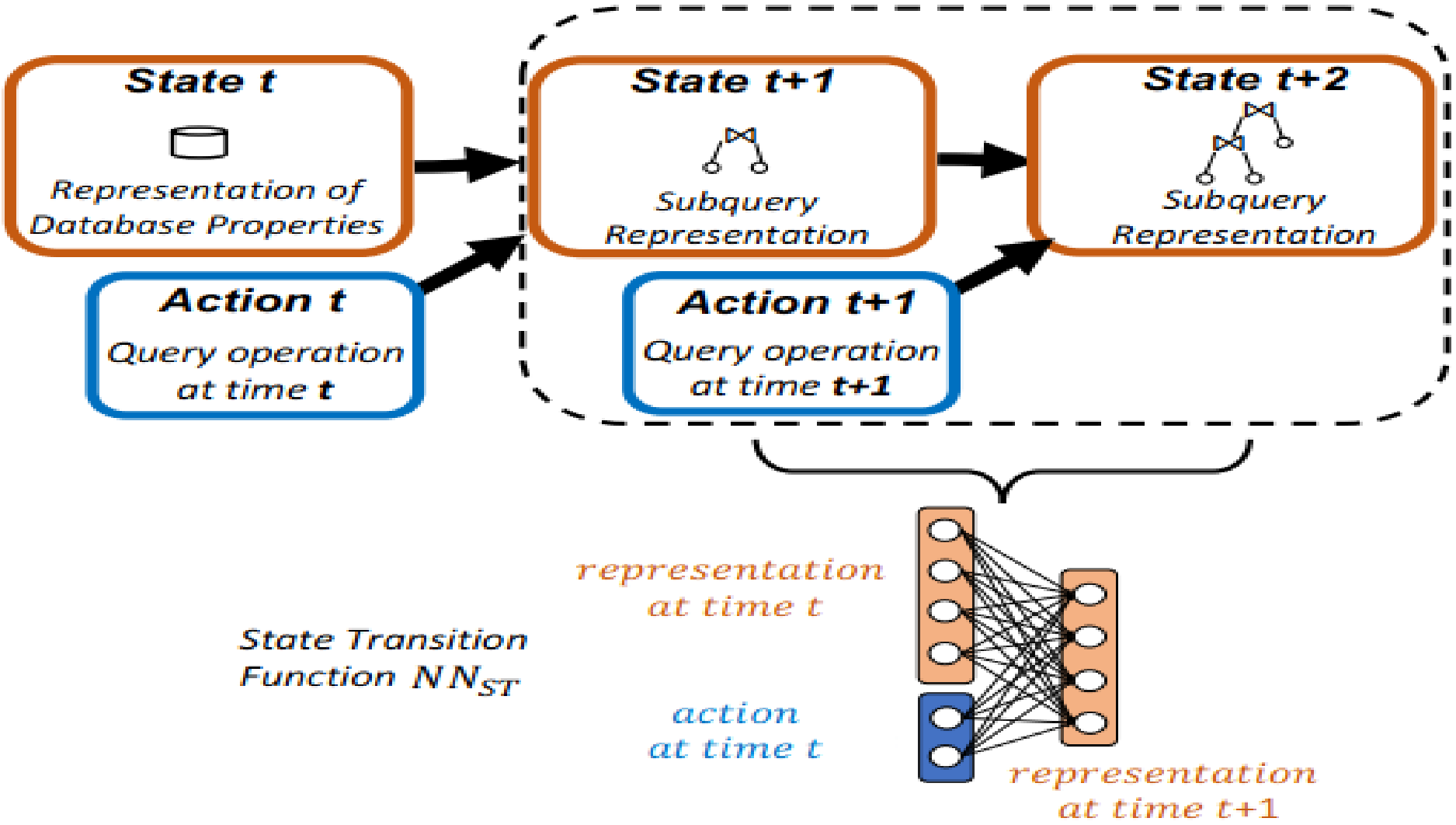
- 深度模型通过学习拟合数据分布



$$p = F(\text{Key}) * N$$



查询策略优化

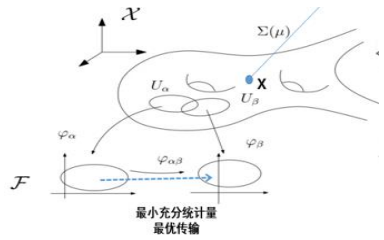


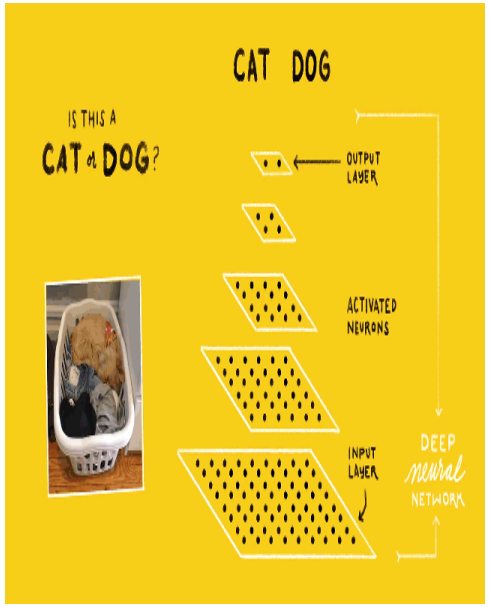
Learning State Representations for Query Optimization with Deep Reinforcement Learning

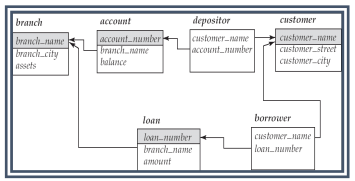
提纲

- 数据库和深度学习
  - 关系和流形结构
  - 关系表示和流形结构表示
- 多模态数据的语义关系
  - 多模态语义的层次组合结构表示
  - 相似和相关关系的表示
- 数据驱动的索引和查询策略优化
  - 索引学习
  - 查询策略优化

SQL-Like的多模态查询&检索







① 结构化特征表示和分解
② 相似和相关关系
③ 距离学习
④ 对象生成

SELECT O_1, ? O_2
WHERE ?O_2 LIKE O1 AND ?O_2 IS ABOUT CHINA

本体驱动的描述性算子定义与学习
（主动学习、元学习）

谢谢！