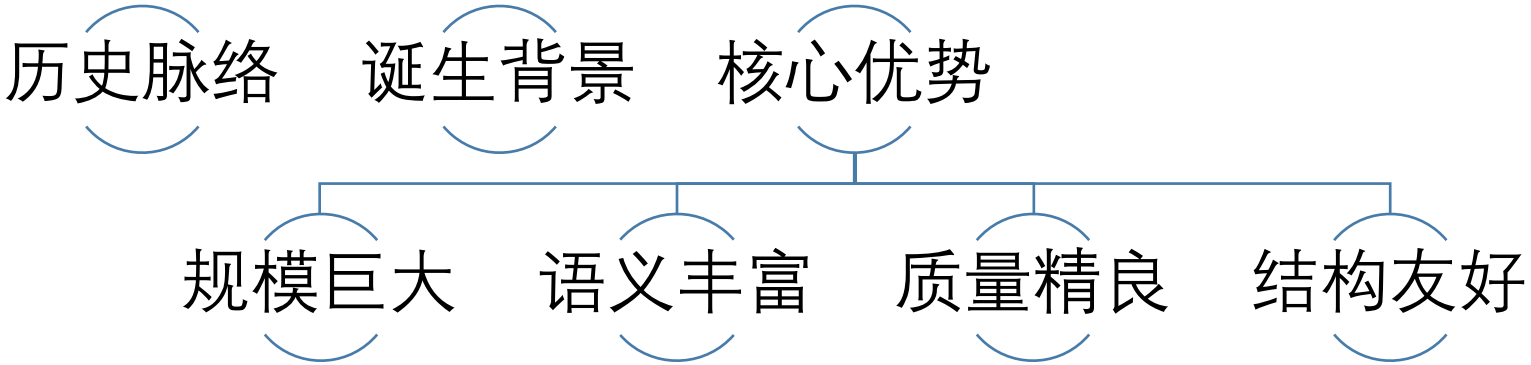
Introduction to Knowledge Graphs

肖仰华
复旦大学知识工场实验室
shawyh@fudan.edu.cn

2017-7-13

Outline

源起



历史脉络

知識工場

人工智能 → 知识工程 → 知识表示 → 知识图谱

AI(Artificial Intelligence): Think, act, humanly or rationally

"The exciting new effort to make computers ***think*** … *machines with minds*, in the full and literal sense." (Haugeland, 1985)
"AI … is concerned with ***intelligent behavior*** in artifacts." (Nilsson, 1998)

KE (Knowledge engineering) is an engineering discipline that involves ***integrating knowledge into computer systems*** in order to solve complex problems normally requiring a high level of human expertise

KR (Knowledge representation) is dedicated to ***representing information about the world*** in a form that a computer system can utilize to solve complex tasks such as diagnosing a medical condition or having a dialog in a natural language.

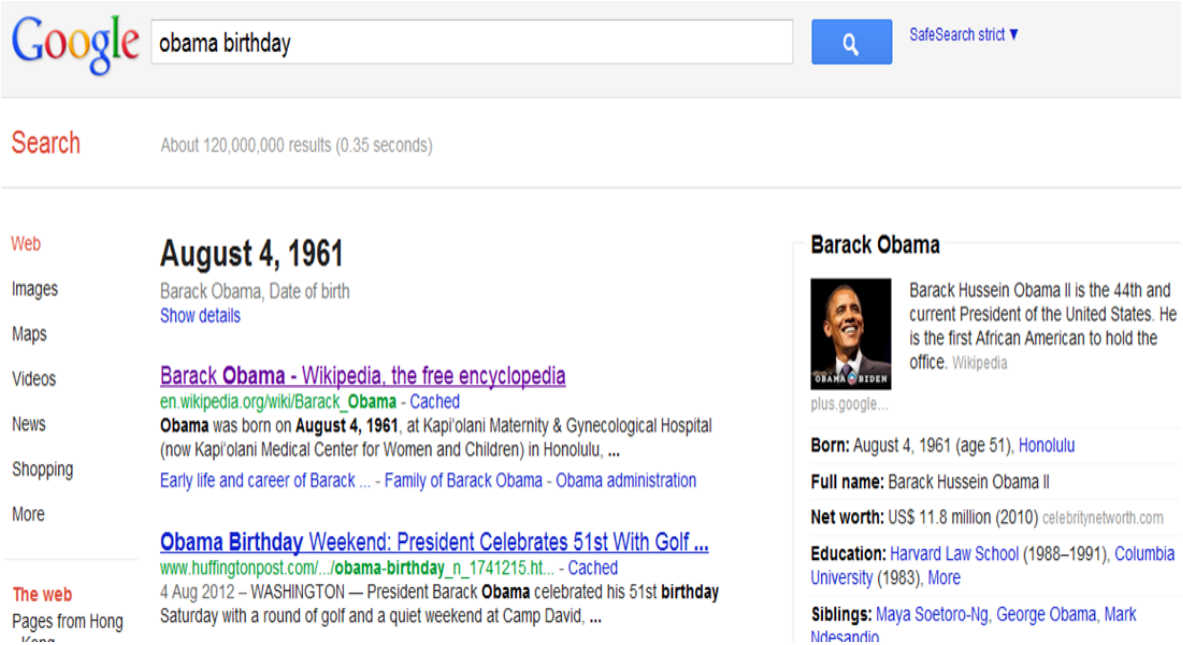
KG (Knowledge graph) is a large scale ***semantic network*** consisting of entities/concepts as well as the semantic relationships among them

诞生背景

- 2012年5月， Goolge正式发表自己的知识图谱

- 搜索核心需求： 让搜索通往答案
  - 无法理解搜索关键词
  - 无法精准回答

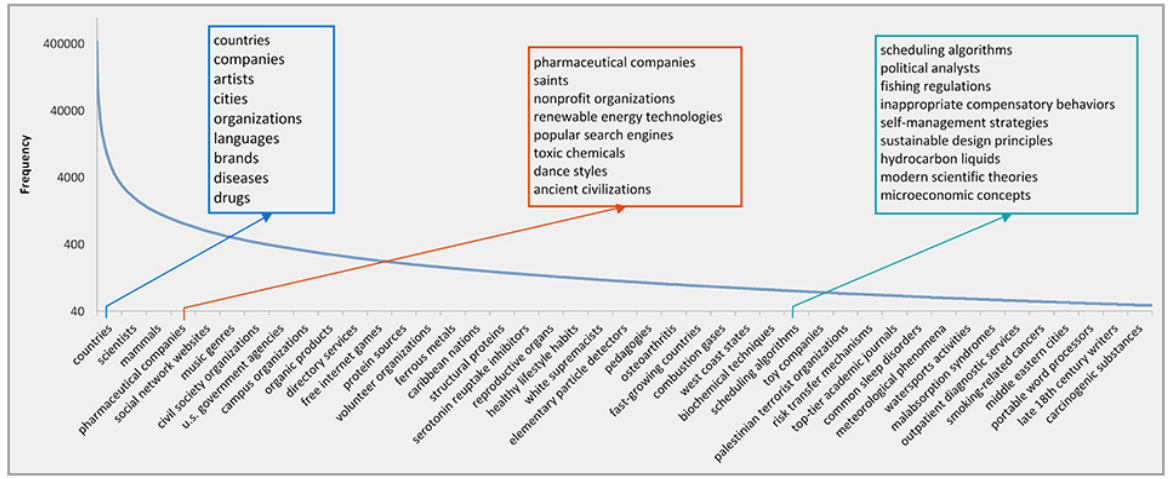
- 根本问题
  - 缺乏大规模背景知识
  - 传统知识表示难以满足需求



KG优势1： large scale

- Higher coverage over entities and concepts

KGs	# of Entities/Concepts	# of Relations
YAGO	10 Million	120 Million
DBpedia	28 Million	9.5 **Billion**
Probase	2.7 Million	70 **Billion**
BabelNet	14 Million	**5 Billion**
CN-DBpedia	17 Million	200 Million

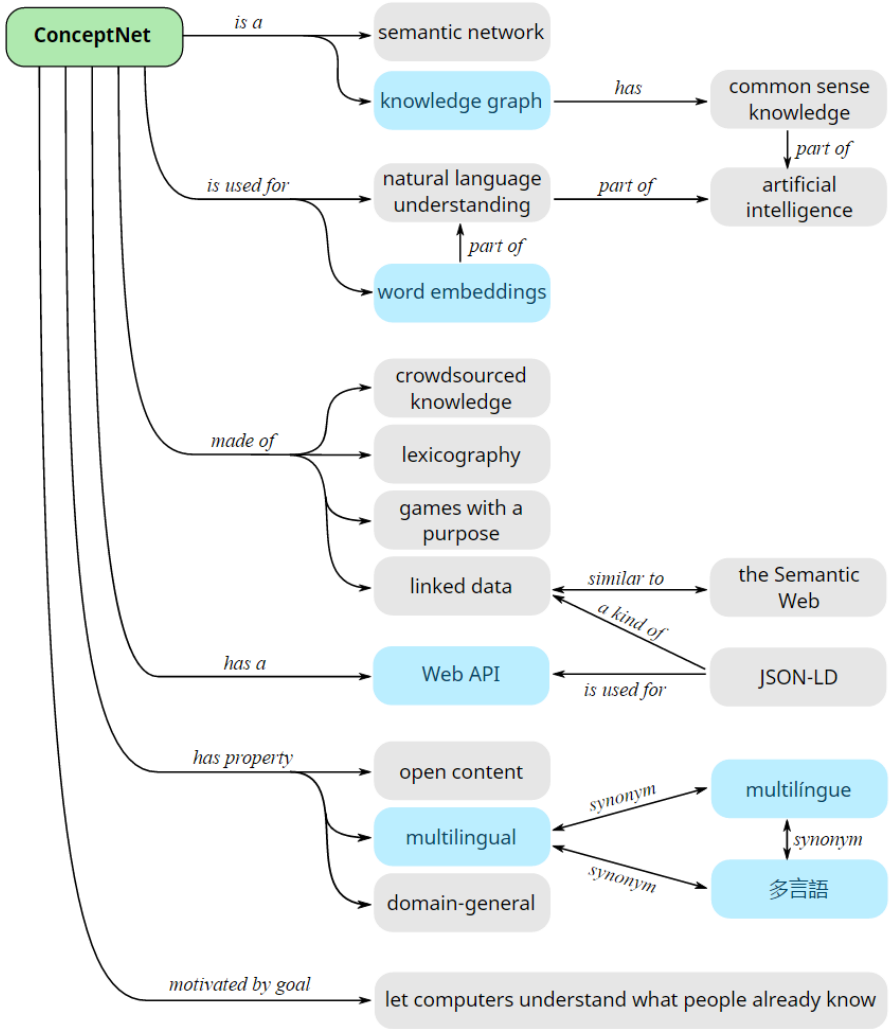


Existing Taxonomies	Number of Concepts
Freebase [5]	1,450
WordNet [13]	25,229
WikiTaxonomy [26]	111,654
YAGO [35]	352,297
DBPedia [1]	259
ResearchCyc [18]	≈ 120,000
KnowItAll [12]	N/A
TextRunner [2]	N/A
OMCS [31]	N/A
NELL [7]	123
Probase	**2,653,872**

KG优势2: semantically rich

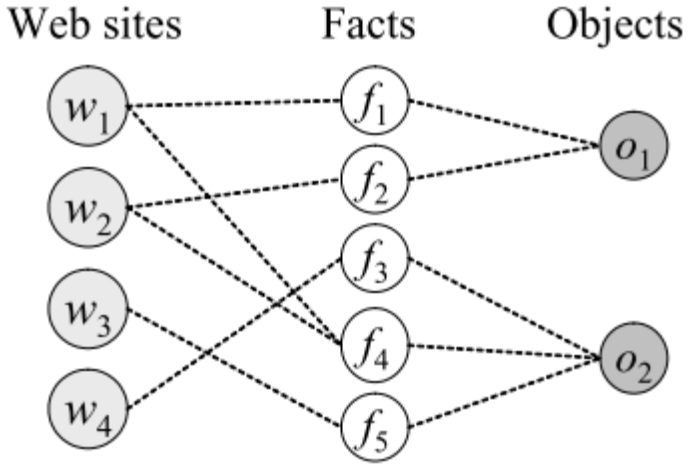
- Higher coverage over numerous semantic relationships

KGs	# of Relations
DBpedia	1,650
YAGO1	14
YAGO3	74
CN-DBpedia	100 Thousands



KG优势3： high quality

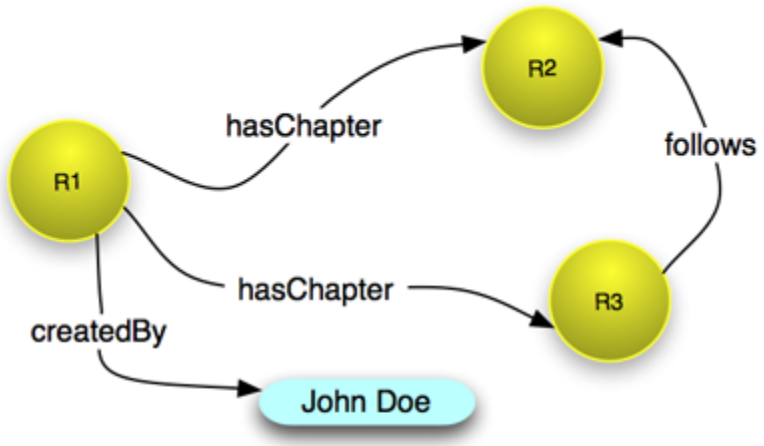
- High quality
  - Big data: Cross validation by multiple sources
  - Crowd sourcing: quality guarantee



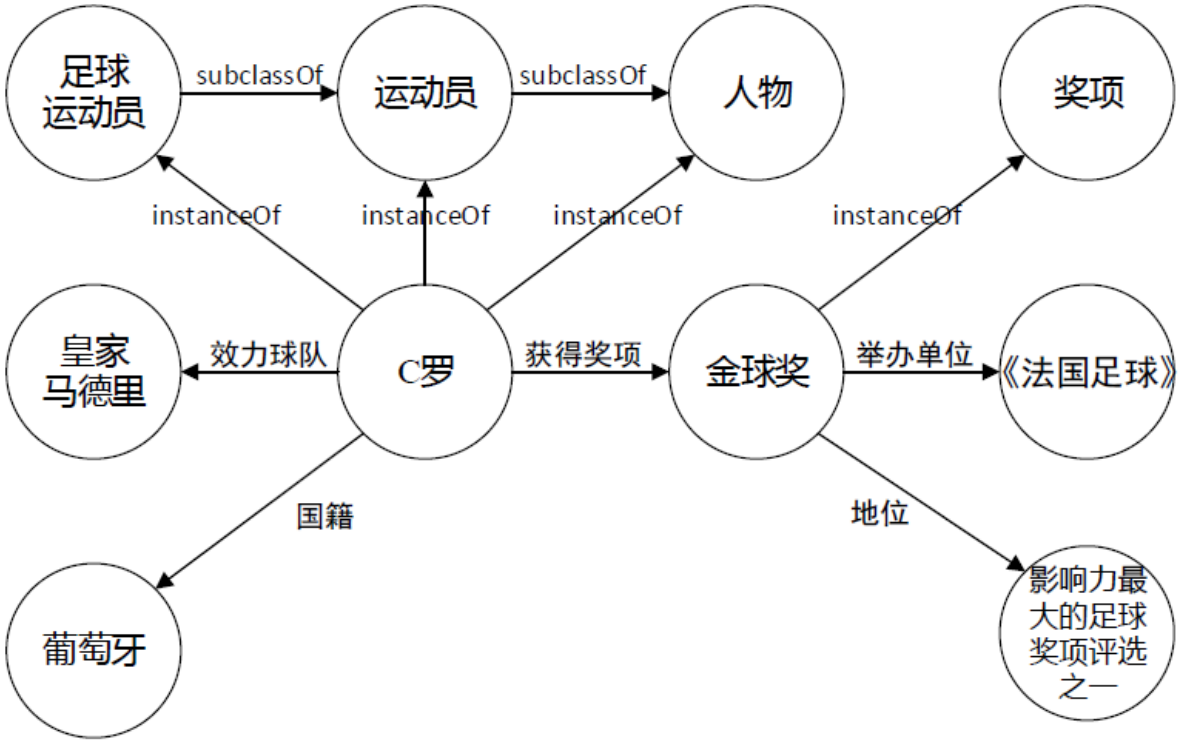
[Yin, etc., Truth Discovery with Multiple Conflicting Information Providers on the Web, kdd07]

KG优势4: friendly structure

- Structured organization
  - By RDF
  - By graph



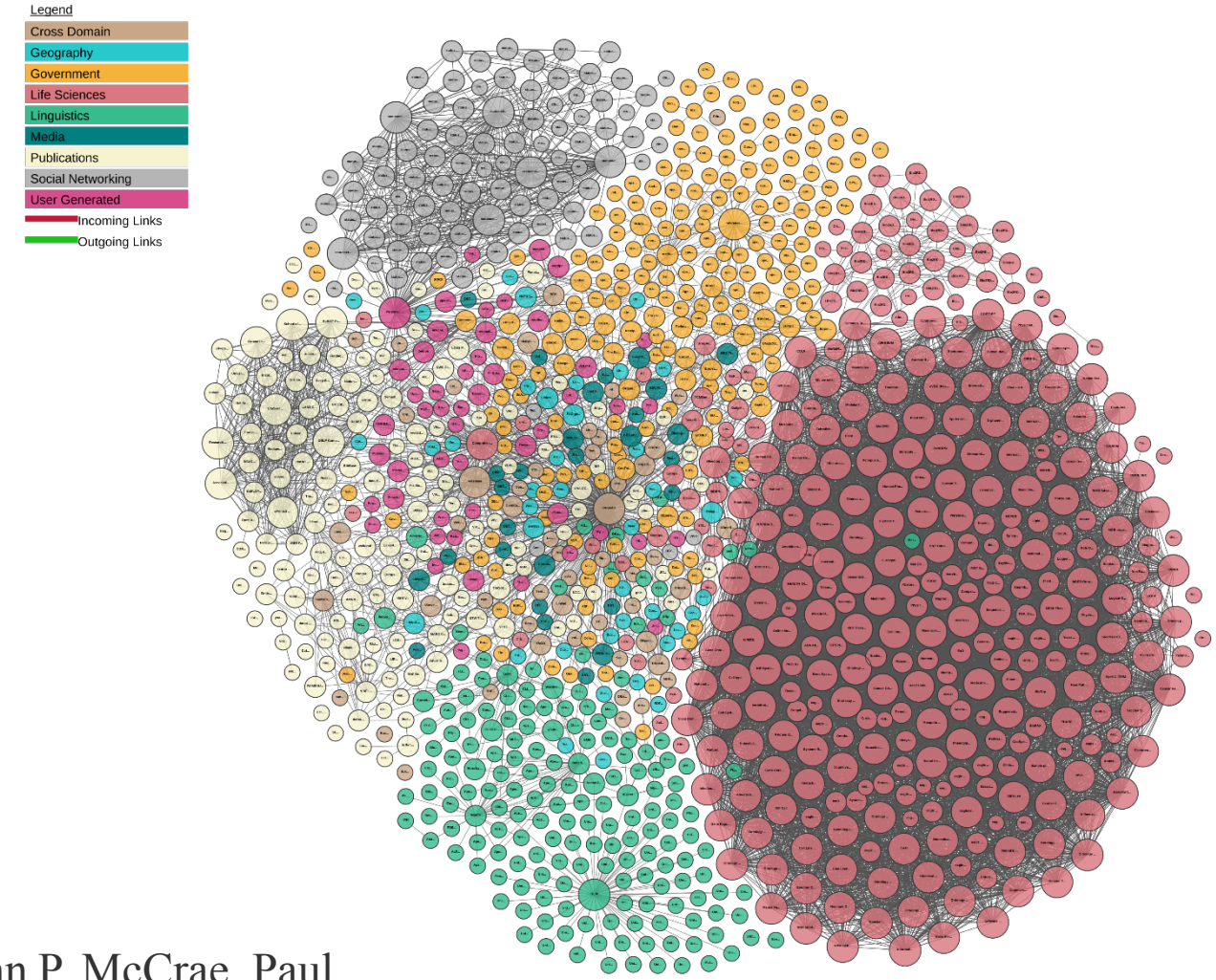
Subject	Predicate	Object
R1	hasChapter	R2
R1	hasChapter	R3
R3	follows	R2
R1	createdBy	"John Doe"



越来越多的知识图谱应运而生

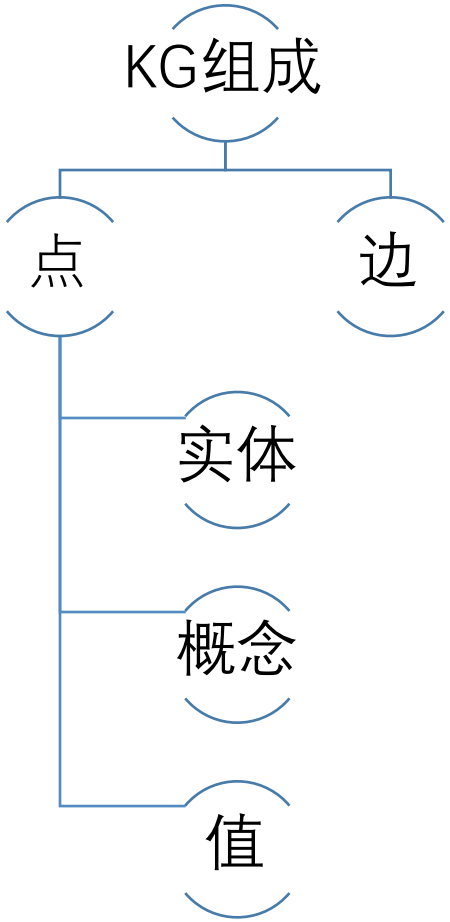
Yago， WordNet, FreeBase, Probase, NELL, CYC, DBpedia….

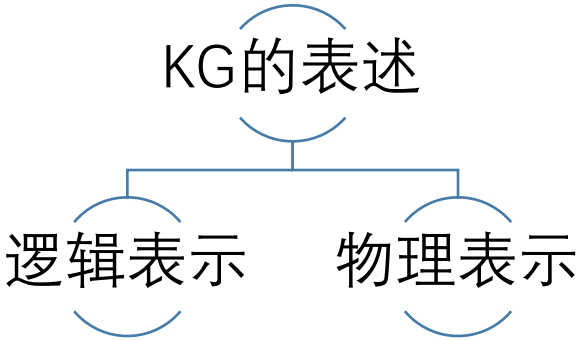
时间	知识图谱数量
2017-03-16	1,139
2014-08-30	570
2011-09-19	295
2010-09-22	203
2009-07-14	95
2008-09-18	45
2007-11-07	28
2007-05-01	12



"Linking Open Data cloud diagram 2017, by Andrejs Abele, John P. McCrae, Paul Buitelaar, Anja Jentzsch and Richard Cyganiak. http://lod-cloud.net/"

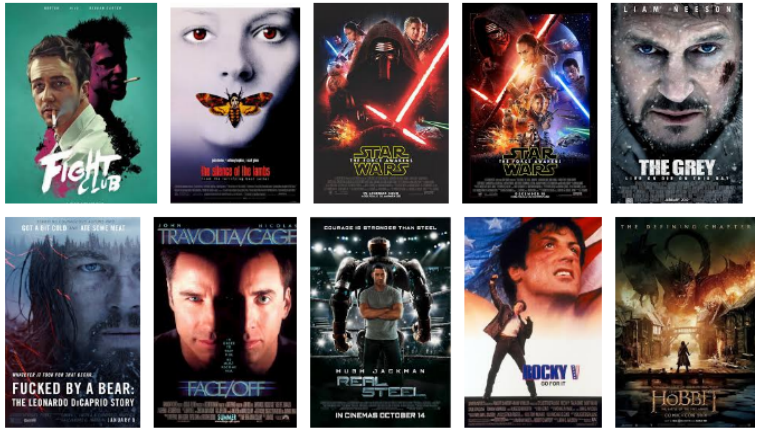
内涵





KG组成- Node-Entity

- Entity/Objects/Instances
  - Wikipedia: An **entity** is something that exists as itself, as a subject or as an object, actually or potentially, concretely or abstractly, physically or not.
  - 黑格尔《小逻辑》：能够独立存在的，作为一切属性的基础和万物本原的东西



KG组成- Node-Concept

- Concept
  - In metaphysics, and especially ontology, a concept is a fundamental category of existence.
  - (mental) representations of categories
- Category
  - Groups of entities which have something in common;
- Type/class
  - WIKITIONARY: A grouping based on shared characteristics; a class.

CATEGORIZATION:

1、the process of formation of categories;

2、the process of identifying X as a member of a particular category Y;

owl:Thing
- Activity (edit)
  - Game (edit)
    - BoardGame (edit)
    - CardGame (edit)
  - Sales (edit)
  - Sport (edit)
    - Athletics (edit)
    - Boxing (edit)

DBpedia Types

microsoft
- company
- software company (Basic-level concept)
- largest OS vendor

Probase Categories

KG组成- Node-Value

- Date
  - 特朗普 出生日期 1946年6月14日
- String
  - 特朗普 简介 “唐纳德·特朗普（Donald Trump），第45任美国总统，1946年6月14日生于纽约，美国共和党籍政治家”
- Numeric
  - 特朗普 年龄 71

KG组成- 边

- Relation
  - 侧重实体(individual)之间的关系
  - Examples:
    - Sitting-On: An apple sitting on a table
    - Taller-than: Washington Monument is taller than the White House

- Property/Attribute/Quality
  - A characteristic/quality that describes an object
  - Examples:
    - size, color, weight, composition, and so forth, of an object

Models of Knowledge Graph

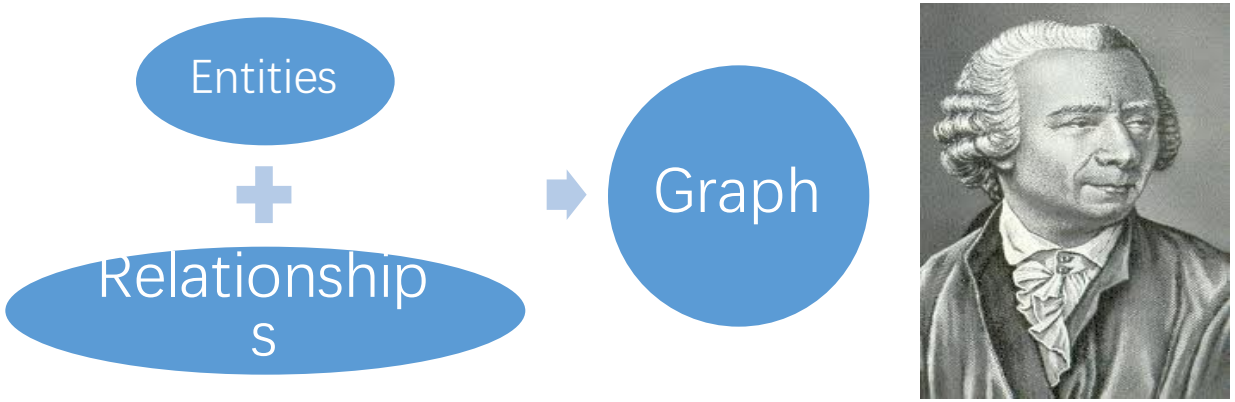
Entities
- Concepts
- Instance
- Value

Relationships
- IsA
- Co-occurrence
- Synonyms
- Others….

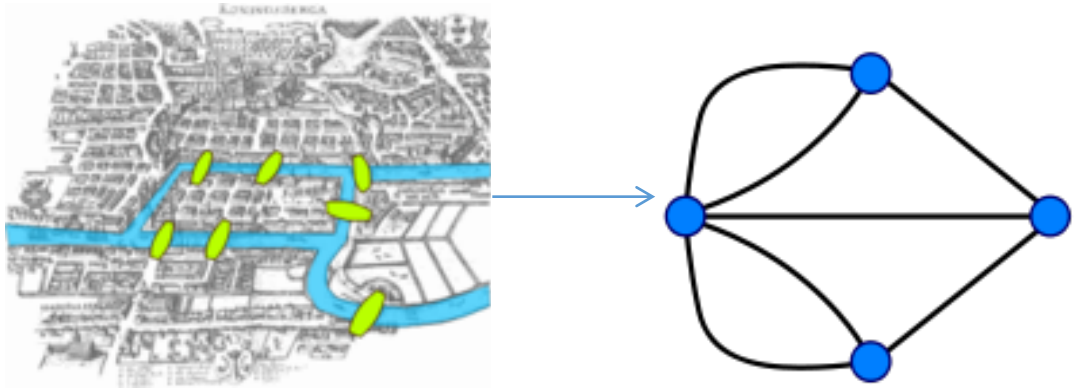
Knowledge Graph

What is a graph?

- A collections of entities and relationship between them
  - Entity
  - Relationships

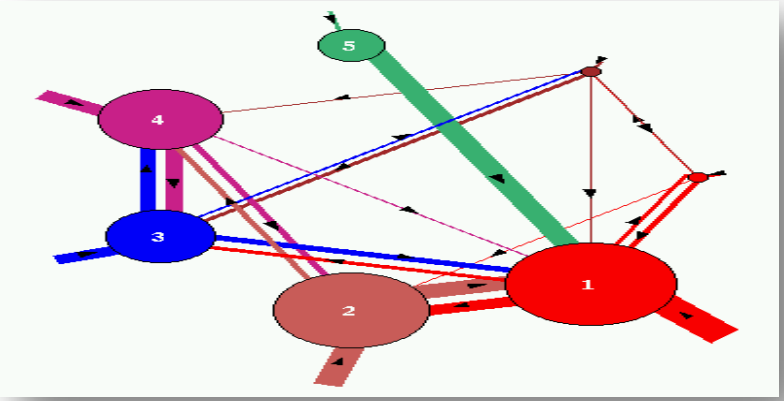


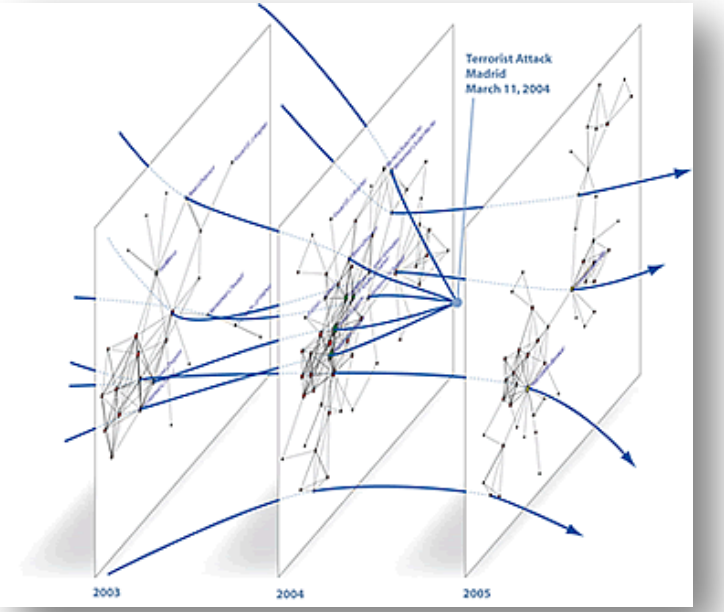
- Euler
  - Seven Bridges of Königsberg

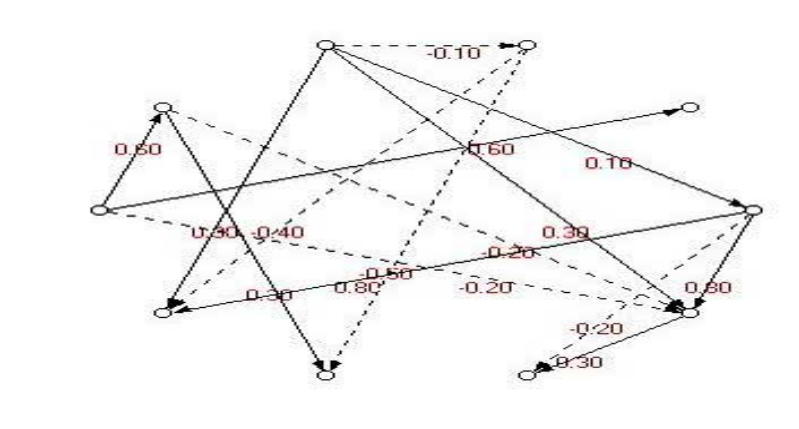


Models of graphs

- Weighted graphs
- Directed graphs
- Probalistic graphs
- Evolving graphs





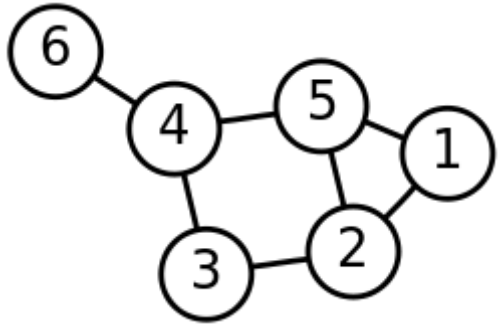


Notations

- Vertices/Nodes
- Edges/arcs

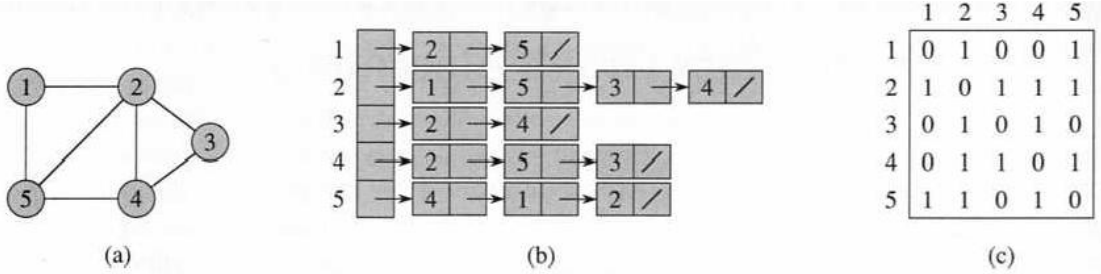
- Neighbors of a vertex
- Degree of a vertex
- Subgraph
- Shortest path

- Example graph



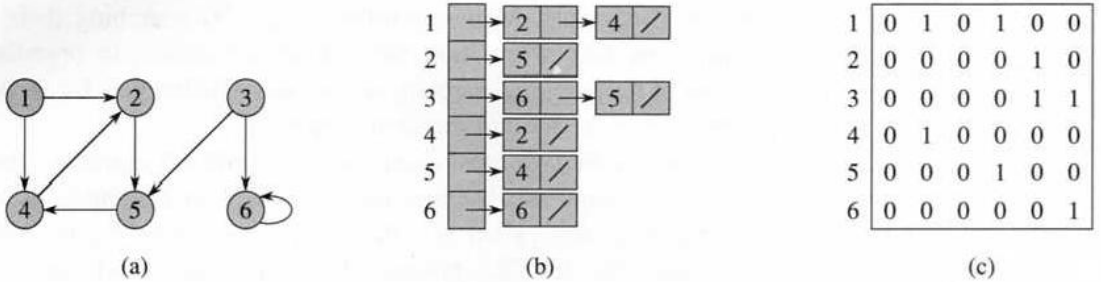
Representation of a graph

- Adjacent list
  - Space efficient on sparse graph
- Matrix



	1	2	3	4	5
1	0	1	0	0	1
2	1	0	1	1	1
3	0	1	0	1	0
4	0	1	1	0	1
5	1	1	0	1	0

Figure 22.1 Two representations of an undirected graph. **(a)** An undirected graph G having five vertices and seven edges. **(b)** An adjacency-list representation of G. **(c)** The adjacency-matrix representation of G.



	1	2	3	4	5	6
1	0	1	0	1	0	0
2	0	0	0	0	1	0
3	0	0	0	0	1	1
4	0	1	0	0	0	0
5	0	0	0	1	0	0
6	0	0	0	0	0	1

Figure 22.2 Two representations of a directed graph. **(a)** A directed graph G having six vertices and eight edges. **(b)** An adjacency-list representation of G. **(c)** The adjacency-matrix representation of G.

RDF: Resource Description Framework

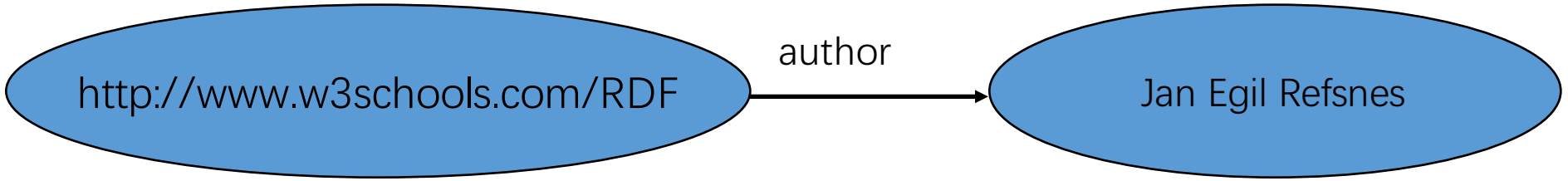
- A framework (not a language) for describing resources, recommended by W3C
- Facilitating reading and correct use of information by *computers*, not necessarily by people

- Resource, Property, Property Value =Subject, Predicate, Object of a statement
- RDF identifies resources with URIs

RDF representations

- RDF offers only binary predicates.
- Think of them as P(x,y) where P is the relationship between the objects x and y.

- From the example,
- X = http://www.w3schools.com/RDF
- Y = Jan Egil Refsnes
- P = author



```
<?xml version="1.0"?>
<rdf:RDF
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:cd="http://www.recshop.fake/cd#">
<rdf:Description
 rdf:about="http://www.recshop.fake/cd/Empire Burlesque">
  <cd:artist>Bob Dylan</cd:artist>
  <cd:country>USA</cd:country>
  <cd:company>Columbia</cd:company>
  <cd:price>10.90</cd:price>
  <cd:year>1985</cd:year>
</rdf:Description>
<rdf:Description
 rdf:about="http://www.recshop.fake/cd/Hide your heart">
  <cd:artist>Bonnie Tyler</cd:artist>
  <cd:country>UK</cd:country>
  <cd:company>CBS Records</cd:company>
  <cd:price>9.90</cd:price>
  <cd:year>1988</cd:year>
</rdf:Description>
… </rdf:RDF>.
```

Root element of RDF documents

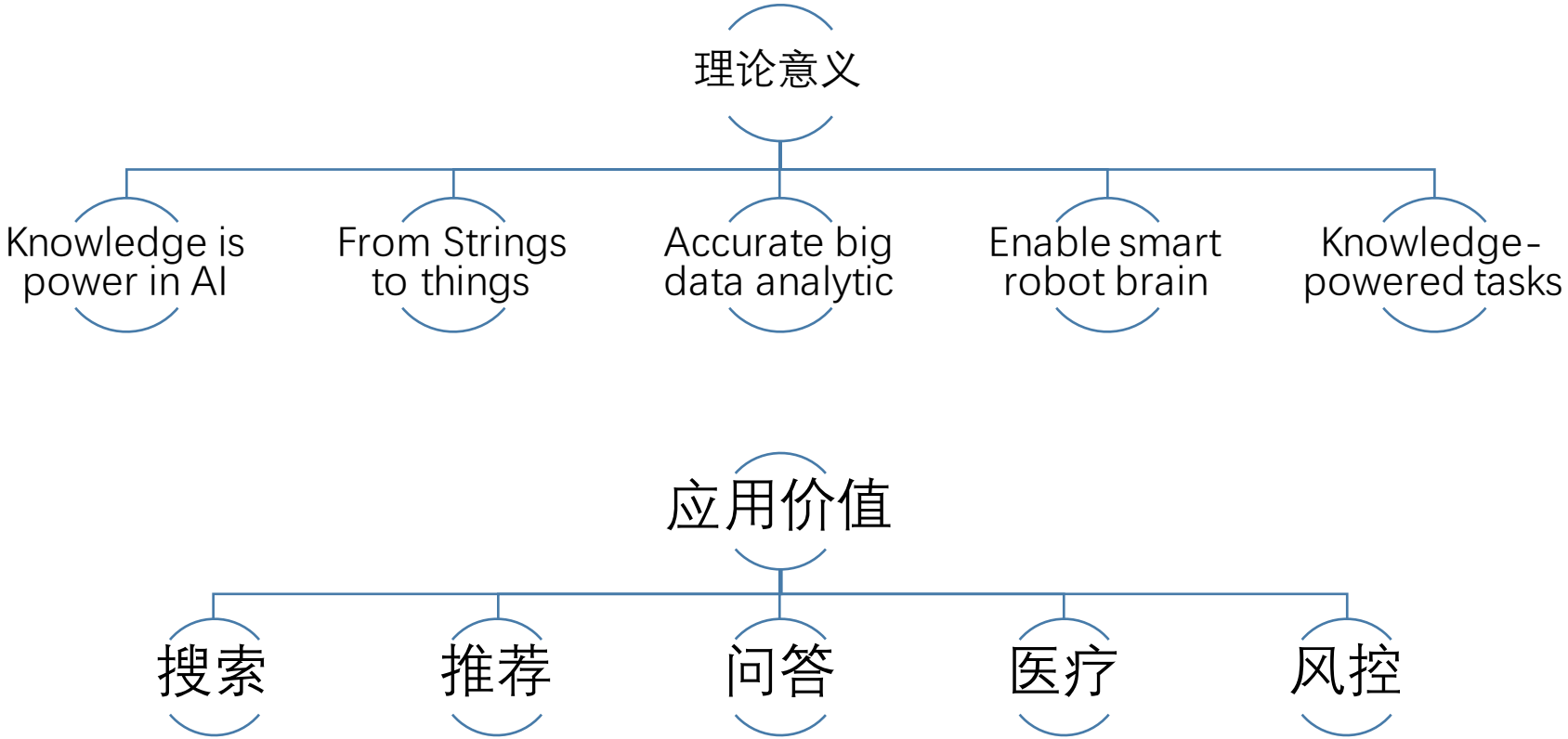
Source of namespace for elements with rdf prefix

Source of namespace for elements with cd prefix

Description element describes the resource identified by the rdf:about attribute.

Cd:country etc are properties of the resource.

价值


理论意义
Knowledge is power in AI
From Strings to things
Accurate big data analytic
Enable smart robot brain
Knowledge-powered tasks
应用价值
搜索
推荐
问答
医疗
风控

Knowledge is power in AI

- AI system=knowledge + reasoning

- Edward Feigenbaum: *"father of expert systems"*
  - *Knowledge is power, and the computer is an amplifier of that power. We are now at the dawn of a new computer revolution… Knowledge itself is to become the new wealth of nations.*

http://www.computerhistory.org/fellowawards/hall/bios/Edward,Feigenbaum/

Big Knowledge is Big Power in AI

From Strings to Things

- Mapping text to a mental world consisting of entities and concepts
- Enables
  - From Search keywords to answers
    - Ex: obama birthday
  - Text understanding

Big data analytic

- Big data analytic needs more background knowledge for
  - A better accuracy
  - More intelligence
- Example
  - How hot is baoqiang's divorce on web?
  - Why baoqiang select Qizhun Zhang as his lawyer?

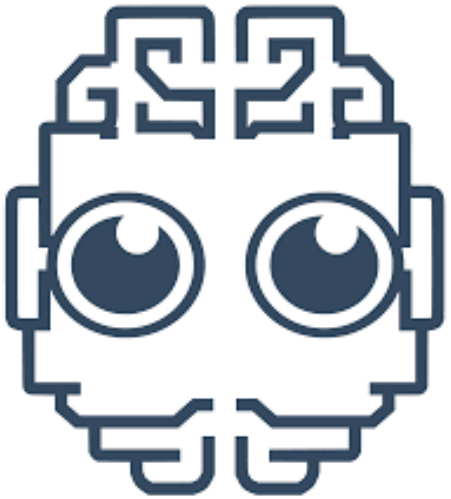
[深扒王宝强离婚内幕 最大祸根源于谁_百山探索](#)
[深度解析宝宝离婚闹剧事件 细说婚姻幸福真谛!_央广网](#)
[宝强离婚最新动态,DNA结果公布马蓉原形毕露_新闻频道_中华网](#)

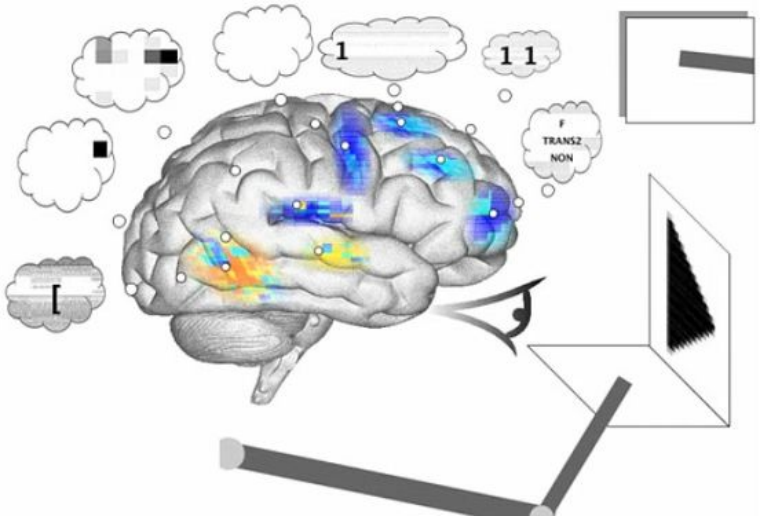
.....宝宝不知道宝宝的宝宝是不是宝宝亲生的宝宝，宝宝现在担心的是宝宝的宝宝不是宝宝的宝宝如果宝宝的宝宝真的不是宝宝的宝宝那就吓死宝宝了宝宝的宝宝为什么要这样对待宝宝，宝宝很难过，如果宝宝和宝宝的宝宝因为宝宝的宝宝打起来了，你们到底支持宝宝还是宝宝的宝宝！【宝宝心里苦，但是宝宝不说】

Smart Robot Brain

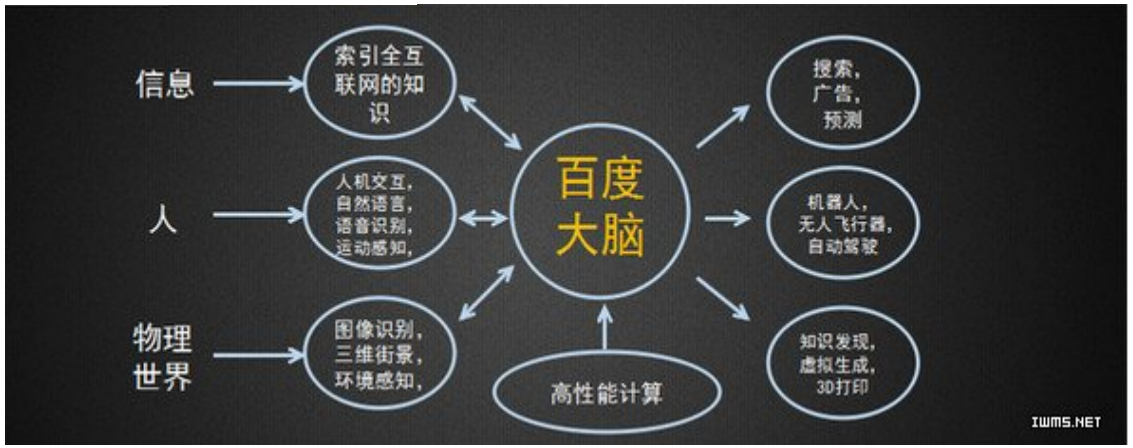
- Smart robot brain needs background knowledge to
  - Interact with the world
  - Understand the world

http://robobrain.me





- Robot Brain
  - 美国脑计划
  - 欧盟、IBM蓝脑计划
  - 中国脑计划
  - 百度大脑
  - 讯飞超脑



Knowledge-powered tasks

- Human can learn concepts from few samples because we are knowledgeable

- ***Knowledge+***
  - Search
  - Recommendation
  - Machine learning
    - Must or must-not constrain in semi-supervised learning
    - Regulizers with knowledge as constraint
  - ……

"People learning new concepts can often generalize successfully from just a single example, yet machine learning algorithms typically require tens or hundreds of examples to perform with similar accuracy.."

RESEARCH ARTICLES

COGNITIVE SCIENCE

Human-level concept learning through probabilistic program induction

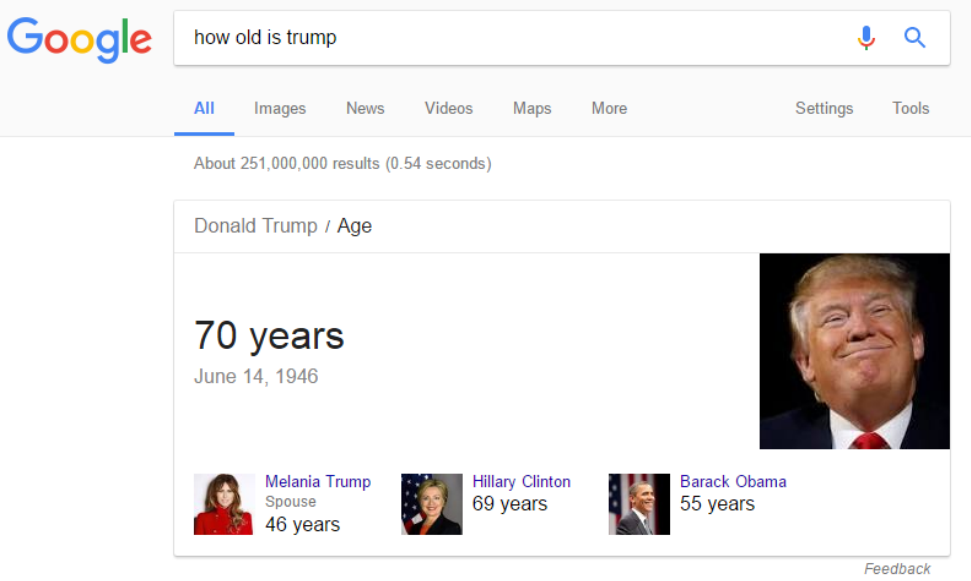
Brenden M. Lake,[1*] Ruslan Salakhutdinov,[2] Joshua B. Tenenbaum[3]

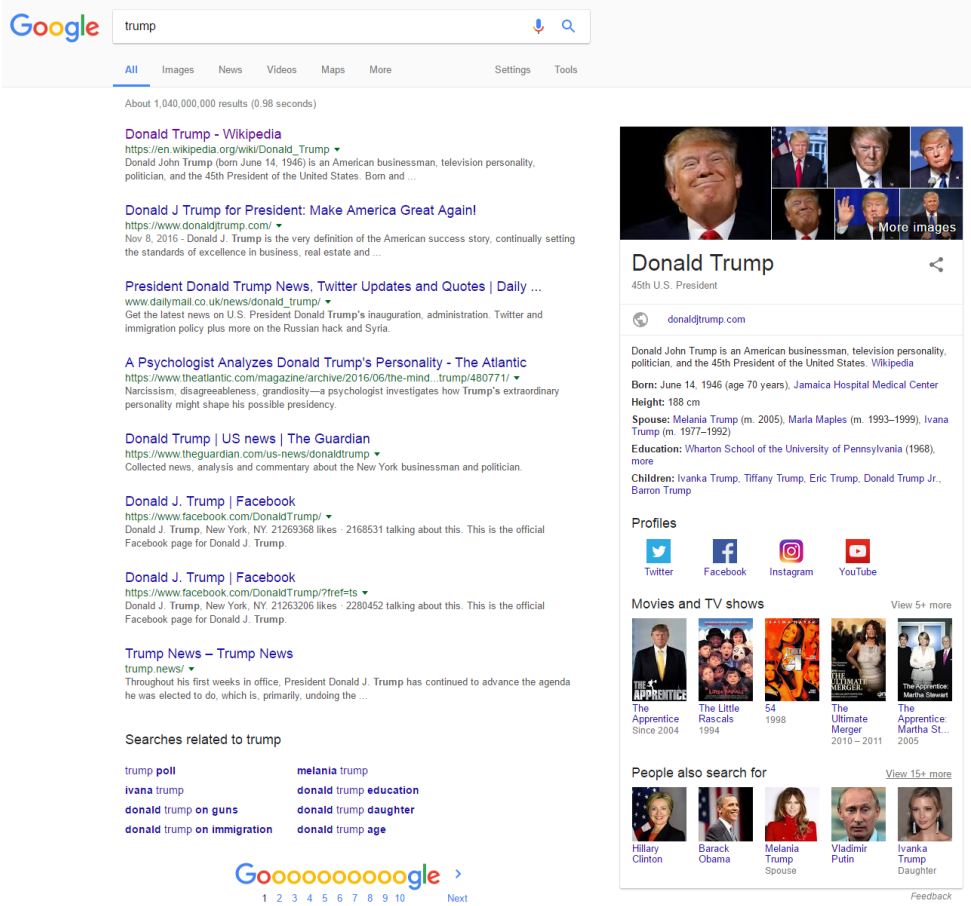
People learning new concepts can often generalize successfully from just a single example, yet machine learning algorithms typically require tens or hundreds of examples to perform with similar accuracy. People can also use learned concepts in richer ways than conventional algorithms—for action, imagination, and explanation. We present a computational model that captures these human learning abilities for a large class of simple visual concepts: handwritten characters from the world's alphabets. The model represents concepts as simple programs that best explain observed examples under a Bayesian criterion. On a challenging one-shot classification task, the model achieves human-level performance while outperforming recent deep learning approaches. We also present several "visual Turing tests" probing the model's creative generalization abilities, which in many cases are indistinguishable from human behavior.

SCIENCE, 11 DECEMBER 2015 • VOL 350 ISSUE 6266

User case: search engine

- Leads to answers instead of web pages
  - Knowledge card provides rich background
  - Identify the entity from the search key word
  - Give answers directly





User case: recommendation

- Recommend semantically related key words, or items

- Recommend conceptually consistent entities
  - What amazon should recommend if users search samsumg s6 and iphone 6?
  - What should taobao recommend if users search ‘康师傅’?

Use Case: Question Answering

This town is known as "Sin City" & its downtown is "Glitter Gulch"

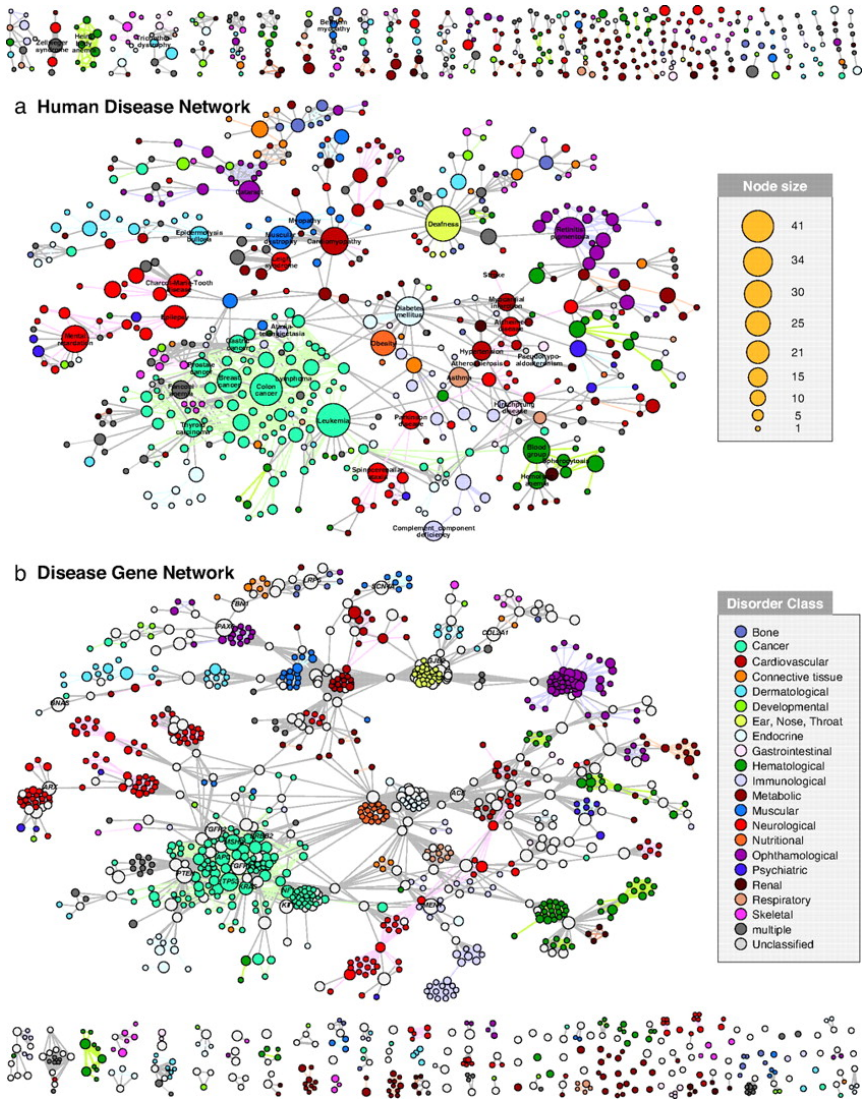
Q: Sin City ?
→ movie, graphical novel, nickname for city, …
A: Vegas ? Strip ?
→ Vega (star), Suzanne Vega, Vincent Vega, Las Vegas, …
→ comic strip, striptease, Las Vegas Strip, …

This American city has two airports named after a war hero and a WW II battle

question classification & decomposition → **knowledge back-ends**

D. Ferrucci et al.: Building Watson. AI Magazine, Fall 2010.
IBM Journal of R&D 56(3/4), 2012: This is Watson.

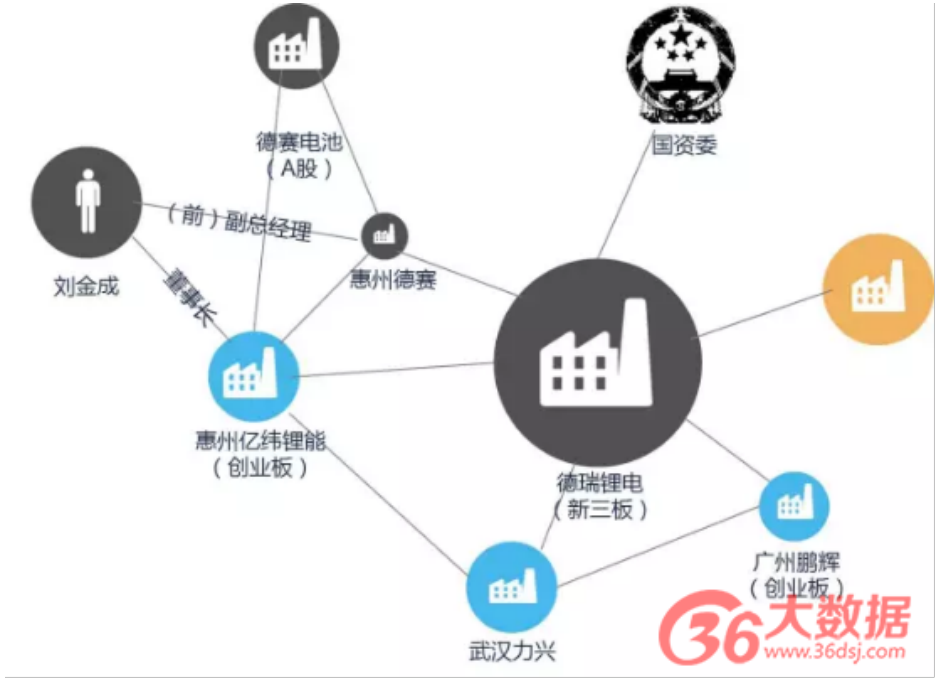
Use Case: Medical Text Analytics



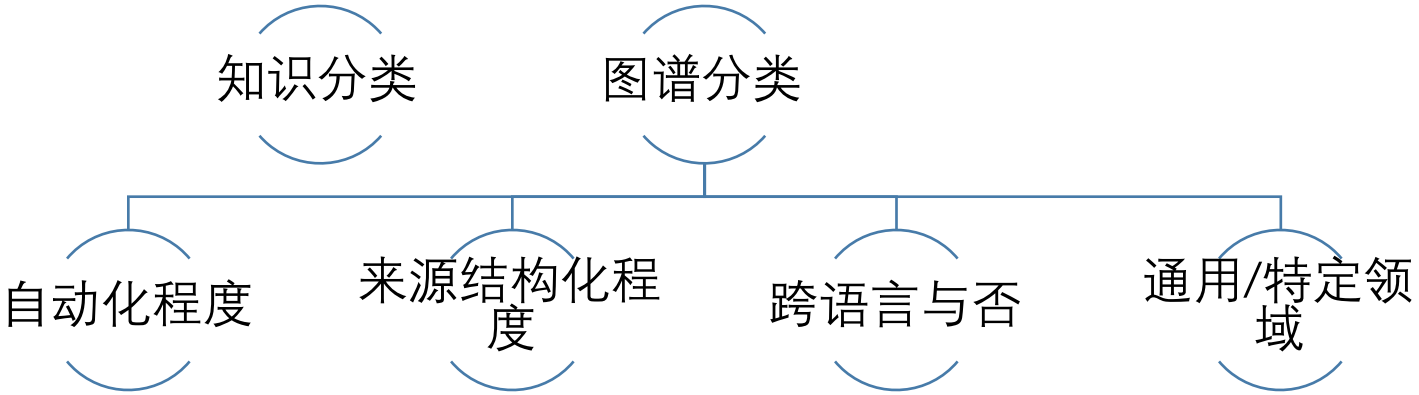
K.Goh,M.Kusick,D.Valle,B.Childs,M.Vidal,A.Barabasi: The Human Disease Network, PNAS, May 2007

Use Case：Risk analytic

- 投资关系分析
- 征信



分类



知识类别

- factual knowledge
  - bornIn (SteveJobs, SanFrancisco), hasFounded (SteveJobs, Pixar),
  - hasWon (SteveJobs, NationalMedalOfTechnology), livedIn (SteveJobs, PaloAlto)

- taxonomic knowledge (ontology):
  - instanceOf (SteveJobs, computerArchitects), instanceOf(SteveJobs, CEOs)
  - subclassOf (computerArchitects, engineers), subclassOf(CEOs, businesspeople)

知识类别

- lexical knowledge (terminology):
  - means (“Big Apple“, NewYorkCity), means (“Apple“, AppleComputerCorp)
  - means (“MS“, Microsoft) , means (“MS“, MultipleSclerosis)
- contextual knowledge (entity occurrences, entity-name disambiguation)
  - maps (“Gates and Allen founded the Evil Empire“, BillGates, PaulAllen, MicrosoftCorp)
- linked knowledge (entity equivalence, entity resolution):
  - sameAs (Apple, AppleCorp), sameAs (hasFounded, isFounderOf)

知识类别

- multi-lingual knowledge:
  - meansInChinese („乔戈里峰“, K2), meansInUrdu („وٹےک“, K2)
  - meansInFr („école“, school (institution)), meansInFr („banc“, school (of fish))
- temporal knowledge (fluents):
  - hasWon (SteveJobs, NationalMedalOfTechnology)@1985
  - marriedTo (AlbertEinstein, MilevaMaric)@[6-Jan-1903, 14-Feb-1919]
  - presidentOf (NicolasSarkozy, France)@[16-May-2007, 15-May-2012]
- spatial knowledge:
  - locatedIn (YumbillaFalls, Peru), instanceOf (YumbillaFalls, TieredWaterfalls)
  - hasCoordinates (YumbillaFalls, 5°55‘11.64‘‘S 77°54‘04.32‘‘W),
  - closestTown (YumbillaFalls, Cuispes), reachedBy (YumbillaFalls, RentALama)

知识类别

- common-sense knowledge (properties):
  - hasAbility (Fish, swim), hasAbility (Human, write),
  - hasShape (Apple, round), hasProperty (Apple, juicy),
  - hasMaxHeight (Human, 2.5 m)

- common-sense knowledge (rules):
  - $\forall$ x: human(x) $\Rightarrow$ male(x) $\vee$ female(x)
  - $\forall$ x: (male(x) $\Rightarrow$ $\neg$ female(x)) $\wedge$ (female(x)) $\Rightarrow$ $\neg$ male(x))
  - $\forall$ x: human(x) $\Rightarrow$ ($\exists$ y: mother(x,y) $\wedge$ $\exists$ z: father(x,z))
  - $\forall$ x: animal(x) $\Rightarrow$ (hasLegs(x) $\Rightarrow$ isEven(numberOfLegs(x))

知识类别

- emerging knowledge (open IE):
  - hasWon (MerylStreep, AcademyAward)
  - occurs („Meryl Streep“, „celebrated for“, „Oscar for Best Actress“)
  - occurs („Quentin“, „nominated for“, „Oscar“)

- multimodal knowledge (photos, videos):
  - JimGray
  - JamesBruceFalls

- social knowledge (opinions):
  - admires (maleTeen, LadyGaga), supports (AngelaMerkel, HelpForGreece)

- epistemic knowledge ((un-)trusted beliefs):
  - believe(Ptolemy,hasCenter(world,earth)), believe(Copernicus,hasCenter(world,sun))
  - believe (peopleFromTexas, bornIn(BarackObama,Kenya))

知识图谱分类

- 自动化程度
- 数据来源结构化程度
- 跨语言
- 通用/specific

ID	知识图谱	构建方式	数据来源	语言	范围
1	Cyc	人工	——	英文	通用
2	WordNet	人工	——	英文	通用
3	ConceptNet	自动	知识图谱	多语言	通用
4	GeoNames	半自动	百科	多语言	领域
5	Freebase	半自动	百科	英文	通用
6	YAGO	自动	百科	多语言	通用
7	DBpedia	半自动	百科	多语言	通用
8	Open IE	自动	纯文本	英文	通用
9	BabelNet	自动	知识图谱	多语言	通用
10	Google KG	自动	混合	多语言	通用
11	Probase	自动	纯文本	英文	通用
12	搜狗知立方	自动	百科	中文	通用
13	百度知心	自动	百科	中文	通用
14	CN-DBpedia	自动	百科	中文	通用

Thank YOU !

Our LAB: Knowledge Works at Fudan University

http://kw.fudan.edu.cn