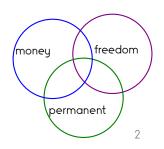
# A Hitchhiker's guide to Ontology

Fabian M. Suchanek Télécom Paris Tech University Paris, France

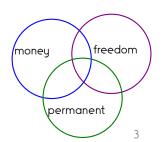




2003: BSc in Cognitive ScienceOsnabrück University/DE2005: MSc in Computer ScienceSaarland University/DE

2008: PhD in Computer Science Max Planck Institute/DE

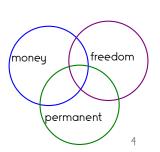




2003: BSc in Cognitive ScienceOsnabrück University/DE2005: MSc in Computer ScienceSaarland University/DE

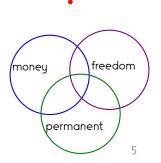
2008: PhD in Computer Science
Max Planck Institute/DE





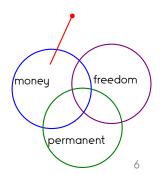
2009: PostDoc at Microsoft Research Silicon Valley/US





2009: PostDoc at Microsoft Research Silicon Valley/US



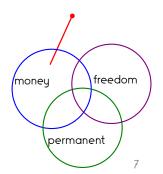


2009: PostDoc at Microsoft Research Silicon Valley/US

2010: PostDoc

INRIA Saclay/FR



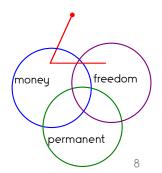


2009: PostDoc at Microsoft Research Silicon Valley/US

2010: PostDoc

INRIA Saclay/FR





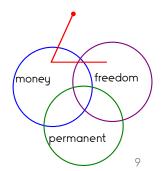
2009: PostDoc at Microsoft Research Silicon Valley/US

2010: PostDoc

INRIA Saclay/FR

2012: Research group leader
Max Planck Institute/DF





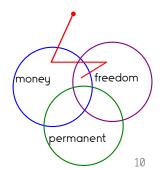
2009: PostDoc at Microsoft Research Silicon Valley/US

2010: PostDoc

INRIA Saclay/FR

2012: Research group leader
Max Planck Institute/DF





2009: PostDoc at Microsoft Research Silicon Valley/US

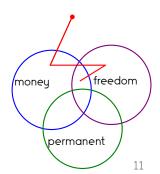
2010: PostDoc

INRIA Saclay/FR

2012: Research group leader
Max Planck Institute/DF

2013: Associate Professor Télécom Paris Tech/FR





2009: PostDoc at Microsoft Research Silicon Valley/US

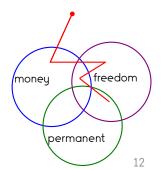
2010: PostDoc

INRIA Saclay/FR

2012: Research group leader
Max Planck Institute/DF

2013: Associate Professor Télécom Paris Tech/FR





2009: PostDoc at Microsoft Research Silicon Valley/US

2010: PostDoc

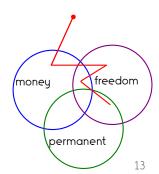
INRIA Saclay/FR

2012: Research group leader
Max Planck Institute/DF

2013: Associate Professor Télécom Paris Tech/FR

2016: Full Professor
Télécom Paris Tech/FR





2009: PostDoc at Microsoft Research Silicon Valley/US

2010: PostDoc

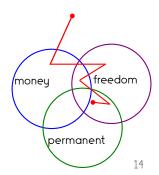
INRIA Saclay/FR

2012: Research group leader
Max Planck Institute/DF

2013: Associate Professor Télécom Paris Tech/FR

2016: Full Professor
Télécom Paris Tech/FR





# I am an Elvis Fan!



#### Recent News Article: Elvis died (?)

Friday September 5th, 2018

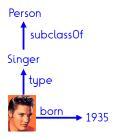
#### Rock legend Elvis Presley (83) died





Pago Pago / American Samoa (dpo) - The King is dead. Already on Sunday, the former singer, nusician and actor Elvis Presley (83) died peacefully in his adopted home Pago Pago on the Pacific Island Tutuli ain the circle of his family. Presley, who completed his active career in 1977, is considered the most successful solo artist in the world, with over one billion records sold.

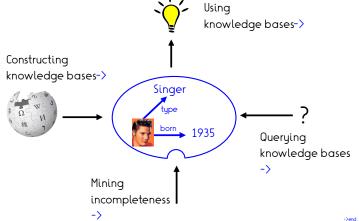
# Knowledge Bases



For us, a knowledge base (KB) is a graph, where the nodes are entities and the edges are relations.

(We do not distinguish T-Box and A-Box.)

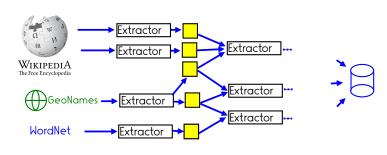
# Knowledge Base Life Cycle



# Extracting from Wikipedia



# Creating a large knowledge base

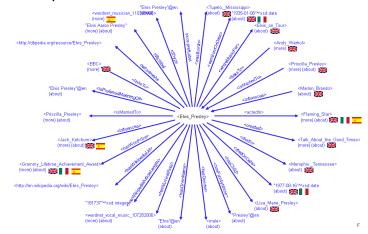


Intermediate extractors

- clean facts
- deduplicate facts and entities
- check consistency

ensuring high quality (95%)

# Example: YAGO about Elvis



### YAGO: a large knowledge base



http://yago-knowledge.org open code and open data

Wikipedia + WordNet time and space 10 languages 100 relations 100m facts 10m entities 95% accuracy used by DBpedia

and TBM Watson

Caveat: focus on precision!







[WWW'07, WS'08, WWW'11 demo, AI|'13, WWW'13 demo, CIDR'15, ISWC'16]













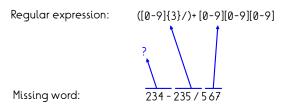


Regular expression:  $([0-9]{3}/)+[0-9][0-9][0-9]$ 

Missing word: 234 - 235 / 567

Regular expression: (([0-9]{3}/)+[0-9][0-9][0-9]) | 234-235/567

Missing word: 234 - 235 / 567



1. Find the optimal partial matching

Regular expression:

Missing word:

1. Find the optimal partial matching

234 - 235 / 567

Regular expression:

Missing word:

- 1. Find the optimal partial matching
- 2. Find the optimal repair

Regular expression:

Missing word:

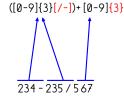
- 1. Find the optimal partial matching
- 2. Find the optimal repair

Regular expression:

Missing word:

- 1. Find the optimal partial matching
- 2. Find the optimal repair
- 3. Optimize and compress

Regular expression:



Missing word:

Caveat: Repair defaults to disjunction

- better recall
- ca. 5 times shorter than a disjunction on 8 datasets

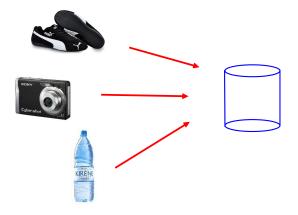




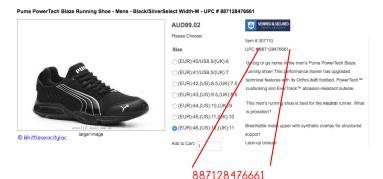


[ISWC 2017 demo. PAKDD 20181

#### Goal: Harvest entities from the Web



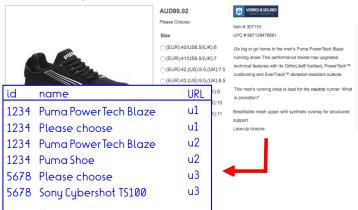
### IBEX: Collect unique ids



Unique identifiers can be found by a regular expression, and verified by a checksum.

### IBEX: Collect unique ids

#### Puma PowerTech Blaze Running Shoe - Mens - Black/SilverSelect Width-M - UPC # 887128476661



### IBEX: analyses

#### Found

- •13m email addresses with their name
- 235K chemical products
- •1.4m books
- •1.1m products

... with an accuracy of 73%-96%



#### Analyzed

- •Global trade flow
- frequent email providers
- frequent people names and more





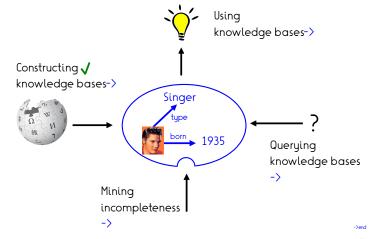




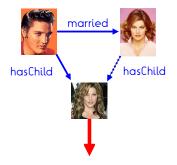
[WebDB 2015]

All data available online at http://resources.mpi-inf.mpg.de/d5/ibex/

# Knowledge Base Life Cycle

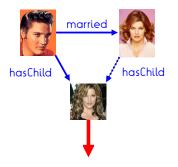


#### Incompleteness: Concrete facts



 $married(x,y) \wedge hasChild(x,z) \Rightarrow hasChild(y,z)$ 

#### Incompleteness: Concrete facts



 $married(x,y) \land hasChild(x,z) \Rightarrow hasChild(y,z)$ 

But: Rule mining needs counter examples and RDF ontologies are positive only

#### Partial Completeness Assumption



Assumption: If we know  $r(x,y_1),..., r(x,y_n)$ , then all other r(x,z) are false.

#### Partial Completeness Assumption



#### Assumption:

If we know  $r(x,y_1),..., r(x,y_n)$ , then all other r(x,z) are false.

#### Partial Completeness Assumption



#### Assumption:

If we know  $r(x,y_1),..., r(x,y_n)$ , then all other r(x,z) are false.

### AMIE finds rules in knowledge bases



AMIE is based on an efficient in-memory database implementation.

Caveat: rules cannot predict the unknown with high precision

# AMIE finds rules in knowledge bases





[WWW 2013, VLDB journal 2015]













the quality of YAGO w ls a precision of 95%, as lks to our brilliant algori

© Girt Happy 43



marriedTo





the quality of YAGO w ls a precision of 95%, as also to our brilliant algori



© Girt Happy 44



the quality of YAGO w ls a precision of 95%, as iks to our brilliant algori



Given a subject s and a relation r, do we know all o with r(s, o) ?

© Girt Happy 45

### Signals for Incompleteness



Closed World Assumption Partial Completeness Assumption Popularity oracle No-change oracle Star-pattern oracle Class-oracle

AMIE oracle: Learn rules such as  $moreThan_1(x, hasParent) \Rightarrow complete(x, hasParent)$ 



# Signals for Incompleteness (F1)

Relation	CWA	PCA	$\mathbf{card}_2$	Popularity	No change	Star	Class	AMIE	
diedIn	60%	22%	_	4%	15%	50%	99%	96%	
directed	40%	96%	19%	7%	71%	0%	0%	100%	
graduatedFrom	89%	4%	2%	2%	10%	89%	92%	87%	
hasChild	71%	1%	1%	2%	13%	40%	78%	78%	
hasGender	78%	100%	_	2%	_	86%	95%	100%	`
hasParent*	1%	54%	100%	_	_	0%	0%	100%	)
isCitizenOf*	4%	98%	11%	1%	4%	10%	5%	100%	
isConnectedTo	87%	34%	19%	_	_	68%	88%	89%	
isMarriedTo*	55%	7%	0%	3%	12%	37%	57%	46%	
wasBornIn	28%	100%	_	5%	8%	0%	0%	100%	_



Relation	CWA	PCA	$\mathbf{card}_2$	Popularity	Star	Class	AMIE
alma_mater	90%	14%	5%	1%	87%	87%	87%
brother	93%	1%	_	1%	94%	96%	96%
child	70%	1%	_	1%	79%	72%	73%
country_of_citizenship*	42%	97%	10%	3%	0%	0%	98%
director	81%	100%	_	3%	94%	89%	100%
father*	5%	100%	6%	9%	89%	8%	100%
mother*	3%	100%	3%	10%	67%*	5%	100%
place_of_birth	53%	100%	7%	5%	55%	0%	100%
place_of_death	89%	35%	1%	2%	81%	81%	96%
sex_or_gender	81%	100%	6%	3%	92%	91%	100%
spouse*	57%	7%	_	1%	54%	54%	55%



•= bi ased training sample



AMIE can predict incompleteness

•bornIn: 100% F1-measure

•diedIn: 96%

•directed: 100%

• graduatedFrom: 87%

•hasChild: 78%

isMarriedTo: 46%

... and more.





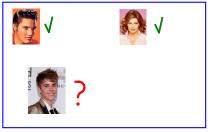


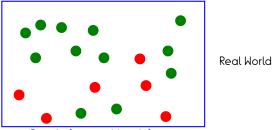


[WSDM 2017]

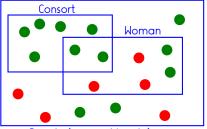
>rep&married >married





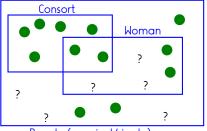


People (married/single)



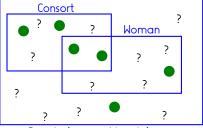
Real World

People (married/single)



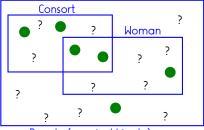
People (married/single)

Knowledge base under the Open World Assumption



People (married/single)

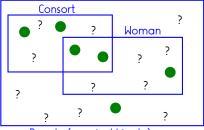
Knowledge base under the Open World Assumption and incompleteness



People (married/single)

Knowledge base under the Open World Assumption and incompleteness

Baseline 1: Obligatory if all instances have it

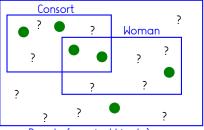


Open World Assumption and incompleteness

Knowledge base under the

People (married/single)

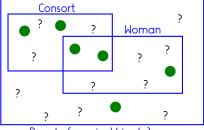
Baseline 1: Obligatory if all instances have it XBaseline 2: Obligatory if at least n% of instances have it



Knowledge base under the Open World Assumption and incompleteness

People (married/single)

Baseline 1: Obligatory if all instances have it X
Baseline 2: Obligatory if at least n% of instances have ₦₩oman
Baseline 3: Obligatory if all instances that have it fall in the class



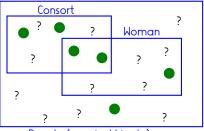
Knowledge base under the Open World Assumption and incompleteness

People (married/single)

Baseline 1: Obligatory if all instances have it X

Baseline 2: Obligatory if at least n% of instances have ₦₩oman X

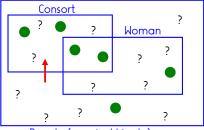
Baseline 3: Obligatory if all instances that have it fall in the class



People (married/single)

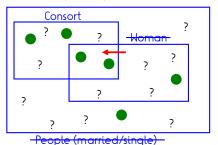
Knowledge base under the Open World Assumption and incompleteness

Theorem: If the KB is sampled randomly uniformly from the real world, and if the density of an attribute changes when we go into an intersecting class, then the attribute cannot be obligatory.



Knowledge base under the Open World Assumption and incompleteness

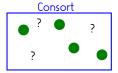
Theorem: If the KB is sampled randomly uniformly from the real world, and if the density of an attribute changes when we go into an intersecting class, then the attribute cannot be obligatory.



Knowledge base under the Open World Assumption and incompleteness

Theorem: If the KB is sampled randomly uniformly from the real world, and if the density of an attribute changes when we go into an intersecting class, then the attribute cannot be obligatory.

# Determining obligatory attributes



We can predict obligatory attributes of classes with up to 80% precision (at 40% recall).





Caveat: We do not actually predict, but exclude.





[WWW 2018]

#### Incompleteness: Missing entities

We have the following cities in our knowledge base:



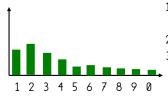
Are there any cities missing?

### Incompleteness: Missing entities

We have the following cities in our knowledge base:



Are there any cities missing?



- 1) Take the number of inhabitants of each city
- 2) Take the first digit
- Plot the number of cities per first digit

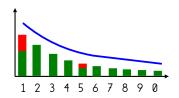
#### Incompleteness: Missing entities

Benford's law says that the first digit! appears with probability

$$log_{10}(1+\frac{1}{d})$$

=> We can give a minimum numbers of cities that are missing to make the distribution representative of the real world.

(For other classes, we can learn a parameter for a variant of the law.)







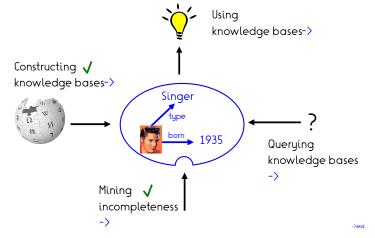




[ISWC 2018]

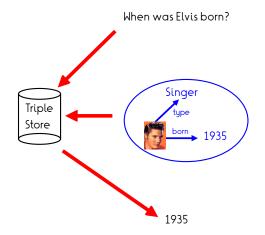


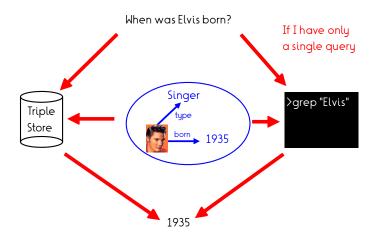
### Knowledge Base Life Cycle

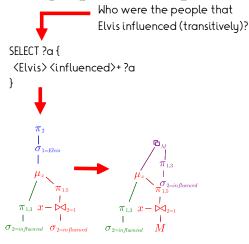


When was Elvis born?









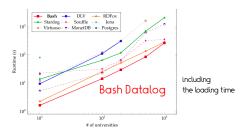
```
Who were the people that
                               Elvis influenced (transitively)?
 SFLECT ?a {
  <Elvis> <influenced>+?a
          \pi_2
         \sigma_{1=Elvis}
                                                     \sigma_{2=in\,fluenced}
    \pi_{1.3} x - \bowtie_{2=1}
                                       \pi_{1,3} x - \bowtie_{2=1}
\sigma_{2=in\,fluenced} \sigma_{2=in\,fluenced}
                                 \sigma_{2=in\,fluenced}
```

while sort -t\$'\t' -k1 -k<(cat tmp/lock\_mat1</pre> join -t \$'\t' -12-21-o  $\langle (sort - t ) | t' - k 1 tmp \rangle$ Icomm -23 - tmp/full2 mv tmp/new2 tmp/de sort -u -merge -o tmp [-s tmp/delta2]; do continue; done

VLDB reviewer: Unix Shell commands to run queries is not very innovative. The DB community exists to overcome the shortcomings of this approach

### Querying Knowledge Bases

VLDB reviewer: Unix Shell commands to run queries is not very innovative. The DB community exists to overcome the shortcomings of this approach



	Bash	RDFox	BigDatalog	Stardog
LiveJournal (69M edges)	2min	1min	9min	15min
Orkut (117M e dges)	2.5min	2min	30min	18min
Friendster (2586M edges)	4:30h	OOM	005	>10h
WikiData (2100M edges)	1h	OOM		73

### Convert SPARQL to Bash online

#### Bash Datalog

Answering Datalog Queries with Unix Shell Commands

About Datalog mode

SPARQL/OWL mode

API

#### SPARQL query

Convert to bash script

Download script

#### Bash script

```
read_ntriples() { Sawk -F' " { sub(" , "\t"); sub(" ", "\t"); sub(" ), "\t"); print 50 ) " S8" ; " ; " ; sub(" ", "\t"); sub(" ), " ; print 51 " $2 " " $3 " ."); ; }

touch tup/mat0 tup/mat1
touch tup/mat0 tup/mat1
sawk -v F9-6'\t" (31 = "chttp://yago-knowledge.org
knowledge.org/rsource/influenced>" { brint 53 >> knowledge.org/rsource/influenced>" } { print 54 > knowledge.org/rsource/influenced>" } { print 54 > knowledge.org/rsource/influenced>" } { print 54 >
```



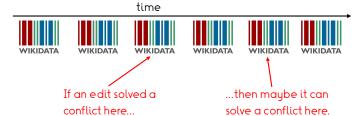




[ISWC 2018]

>wikihist

## Correcting Knowledge Bases



On January 8th, user Bob changed the gender from "m" to "man" for an entity that

- is a person
- is an American rock singer
- was born in 1935
- is a living being

can we do the same for all people? or just for American rock singers? or only for people born in 1935? or in general for all living beings?

### Correcting Knowledge Bases

time













We mine correction rules of the form

If x is an instance of "person", and the gender is "m"

Constraint: The gender has to be "male", "female", "man", "woman",...

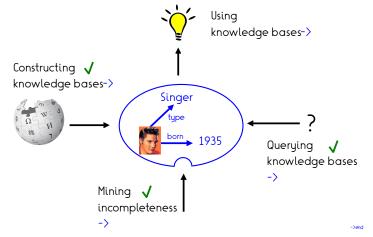






[WWW 2019]

## Knowledge Base Life Cycle



# Combinatorial Creativity







# Description Logics do not work

 $Mop \equiv Tool \sqcap \exists has.Stick \sqcap \exists has.Strings$ 

$$\begin{array}{lll} \textit{BabyMop} & \equiv \\ \textit{Romper} & \sqcap \; \exists \; has.(Mop \; \sqcap \; \neg \exists \; has.Stick) \; \sqcap \; \exists \; has.Baby \\ & \equiv \; \dots \sqcap \; \exists \; has.\bot \quad \sqcap \dots \end{array}$$



© Stone Mens Wear © Vileda © Avsar Aras 79

## Language for Combinatorial Creativity

 $Mop \equiv Tool \sqcap \exists has.Stick \sqcap \exists has.Strings$ 

Subtraction:  $Mop - \exists has. \top \equiv Tool \sqcap \exists has. Strings$ 

Addition:  $Mop + \exists has. \top \equiv Mop$ Succession:  $Mop \rightarrow \exists u. \top \equiv Stick$ 

Selection\*:  $Mop \uparrow \exists has. \top \equiv \exists has. Stick$ 

$$Romper + \exists has.(Mop - \exists has.Stick) + \exists has.Baby$$
  
 $\equiv BabyMop$ 











# Language for Combinatorial Creativity

 $Mop \equiv Tool \sqcap \exists has.Stick \sqcap \exists has.Strings$ 

Subtraction:  $Mop - \exists has. \top \equiv Tool \sqcap \exists Strings$ 

Addition:  $Mop + \exists has. \top \equiv Mop$ Succession:  $Mop \rightarrow \exists r. \top \equiv Stick$ 

Selection\*:  $Mop \uparrow \exists has. \top \equiv \exists has. Stick$ 

1) Descriptive experiments

TECH

The 25 Best Inventions

TIME Staff | Nov. 19, 2015

2) Generative experiments 1/3 nonsense, 1/3 exists, 1/3 "imaginable"



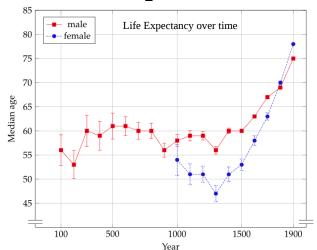




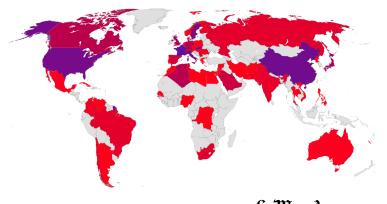


[ISWC 2016 paper & demo]

### YAGO for the Digital Humanities



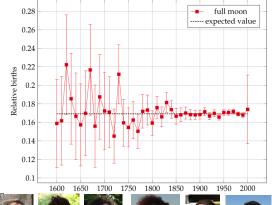
# Presence of foreign companies



red: many foreign companies mentioned blue: few foreign companies mentioned



# People born on a full moon day









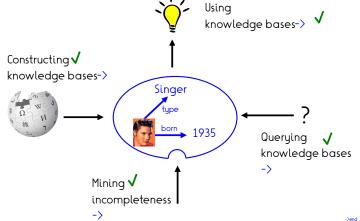






[AKBC 2013, VLDB 2014 vision. LDOW 2018]

## Knowledge Base Life Cycle



### Is Elvis dead?

Friday September 5th, 2018

#### Rock legend Elvis Presley (83) died



Page Pago / American Samoa (dpo) - The King is dead. Already on Sunday, the former singer, musician and actor Evike Presley (83) died peacefully in his adopted home Pago Pago on the Pacific island Tutuila in the circle of his family. Presley, who completed his active career in 1977, is considered the most successful solo artist in the world, with over one billion records sold.

#### Is Elvis dead?

Friday September 5th, 2018

#### Rock legend Elvis Presley (83) died



Pago Pago / American Samoa (dpo) - The King is dead. Already on Sunday, the former singer, musician and actor Elvis Presley (83) died peacefully in his adopted home Pago Pago on the Pacific Island Tutulla in the circle of his family. Presley, who completed his active career in 1977, is considered the most successful solo artist in the world, with over one billion records sold.



#### Is Elvis dead?

Friday September 5th, 2018

#### Rock legend Elvis Presley (83) died



Pago Pago / American Samoa (dpo) - The King is dead. Already on Sunday, the former singer, musician and actor Ekiv Bresley (83) died peacefully in his adopted home Pago Pago on the Pacific island Tutuila in the circle of his family. Presley, who completed his active career in 1977, is considered the most successful solo artist in the world, with over one billion records sold.



Fake news!!

## Knowledge Base Life Cycle

