Learning Semantic Parsers

Percy Liang

2014 JHU/CLSP Fred Jelinek Memorial PIRE Workshop

Stanford University

Paleobiology

Paleobiology

The transition between the Proterozoic and Phanerozoic eons, beginning 542 million years (Myr) ago, is distinguished by the diversification of multicellular animals and by their acquisition of mineralized skeletons during the Cambrian period[1]. Considerable progress has been made in documenting and more precisely correlating biotic patterns in the Neoproterozoic–Cambrian fossil record with geochemical and physical environmental perturbations[2–5], but the mechanisms responsible for those perturbations remain uncertain[1,2]. Here we use new stratigraphic and geochemical data to show that early Palaeozoic marine sediments deposited approximately 540–480 Myr ago record both an expansion in the area of shallow epicontinental seas and anomalous patterns of chemical sedimentation that are indicative of increased oceanic alkalinity and enhanced chemical weathering of continental crust. These geochemical conditions were caused by a protracted period of widespread continental denudation during the Neoproterozoic followed by extensive physical reworking of soil, regolith and basement rock during the first continental-scale marine transgression of the Phanerozoic. The resultant globally occurring stratigraphic surface, which in most regions separates continental crystalline basement rock from much younger Cambrian shallow marine sedimentary deposits, is known as the Great Unconformity[6]. Although Darwin and others have interpreted this widespread hiatus in sedimentation on the continents as a failure of the geologic record, this palaeogeomorphic surface represents a unique physical environmental boundary condition that affected seawater chemistry during a time of profound expansion of shallow marine habitats. Thus, the formation of the Great Unconformity may have been an environmental trigger for the evolution of biomineralization and the 'Cambrian explosion' of ecologic and taxonomic diversity following the Neoproterozoic emergence of animals.

The term Great Unconformity was first used in the year 1869 to describe the prominent stratigraphic surface in the Grand Canyon that separates the shallow marine, ~525-Myr-old Cambrian Tapeats Sandstone from the underlying metamorphosed, 1,740-Myr-old Vishnu Schist and structurally tilted sedimentary rocks of the 1,200–740 Myr-old Grand Canyon Supergroup[6]. The Great Unconformity is well exposed in the Grand Canyon, but this geomorphic surface, which records the erosion and weathering of continental crust followed by sediment accumulation, can be traced across Laurentia and globally, including Gondwana[7,8], Baltica[9], Avalonia[9] and Siberia[10], making it the most widely recognized and distinctive stratigraphic surface in the rock record. It is also notable because the Cambrian sediments that overlie it in many regions preserve the first skeletonized crown-group animals, a fact that some palaeontologists have interpreted as evidence for stratigraphic bias and an incomplete record of early animal evolution[1,6].

Here we use stratigraphic and lithologic data for 21,521 rock units from 830 geographic locations in North America, in conjunction with petrologic and geochemical data (Methods; see also Supplementary Information), to explore the hypothesis that the formation of the Great Unconformity is causally linked to the evolution of biomineralization; this linkage is proposed to occur by means of the geochemical effects of prolonged continental denudation followed by enhanced physical and chemical weathering of continental crust during terminal Ediacaran and Cambrian time.

The Cambrian- to Early Ordovician-aged sediments of the Sauk Sequence[11,12] that overlie the Great Unconformity are time-transgressive, such that Early Cambrian sediments occur on the margins of the palaeocontinents and Late Cambrian sediments overlie the Great Unconformity in continental interiors (Fig. 1). The spatial extent of the Sauk Sequence is comparable to other Phanerozoic continent-scale sedimentary sequences[11,12], but its geological characteristics are unique. In most places, undeformed Cambrian sedimentary rocks deposited on Earth's surface rest non-conformably on much older continental crystalline basement rocks, many of which were formed and/or metamorphosed within the Earth's crust (Fig. 2a). Thus, the Great Unconformity marks the termination of an extended period of continental denudation that exhumed and exposed large areas of igneous and metamorphic rocks to subaerial weathering before marine transgression and subsequent sedimentation.

Continental-scale marine transgression during the Cambrian–Early Ordovician accentuated rates of weathering on the Great Unconformity by shifting landward the position of the erosive transgressive shoreface system, often called the 'wave-base razor'[13], as well as adjacent transitional backshore, aeolian and fluvial systems. As a result, much of the soil and weathered basement rock (regolith) that covered low-relief continental interiors[7] was eroded and mobilized during the transgression, thereby exposing silicate mineral surfaces to weathering over an area that is unprecedented in the rock record (Fig. 2a). This is important because freshly exposed rock weathers chemically at rates more than three times faster than undisturbed soils and regolith[14,15], and

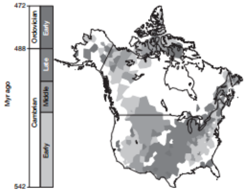


Figure 1 | Sauk Sequence in North America. Distribution and age of the oldest Phanerozoic sedimentary rocks in North America.

[1]Department of Geoscience, University of Wisconsin, Madison, Wisconsin 53706, USA. [2]Geology Department, Pomona College, Claremont, California 91711, USA.

Paleobiology

The transition between the Proterozoic and Phanerozoic eons, beginning 542 million years (Myr) ago, is distinguished by the diversification of multicellular animals and by their acquisition of mineralized skeletons during the Cambrian period[1]. Considerable progress has been made in documenting and more precisely correlating biotic patterns in the Neoproterozoic–Cambrian fossil record with geochemical and physical environmental perturbations[2–5], but the mechanisms responsible for those perturbations remain uncertain[1,2]. Here we use new stratigraphic and geochemical data to show that early Palaeozoic marine sediments deposited approximately 540–480 Myr ago record both an expansion in the area of shallow epicontinental seas and anomalous patterns of chemical sedimentation that are indicative of increased oceanic alkalinity and enhanced chemical weathering of continental crust. These geochemical conditions were caused by a protracted period of widespread continental denudation during the Neoproterozoic followed by extensive physical reworking of soil, regolith and basement rock during the first continental-scale marine transgression of the Phanerozoic. The resultant globally occurring stratigraphic surface, which in most regions separates continental crystalline basement rock from much younger Cambrian shallow marine sedimentary deposits, is known as the Great Unconformity[6]. Although Darwin and others have interpreted this widespread hiatus in sedimentation on the continents as a failure of the geologic record, this palaeogeomorphic surface represents a unique physical environmental boundary condition that affected seawater chemistry during a time of profound expansion of shallow marine habitats. Thus, the formation of the Great Unconformity may have been an environmental trigger for the evolution of biomineralization and the 'Cambrian explosion' of ecologic and taxonomic diversity following the Neoproterozoic emergence of animals.

The term Great Unconformity was first used in the year 1869 to describe the prominent stratigraphic surface in the Grand Canyon that separates the shallow marine, ~525-Myr-old Cambrian Tapeats Sandstone from the underlying metamorphosed, 1,740-Myr-old Vishnu Schist and structurally tilted sedimentary rocks of the 1,200–740 Myr-old Grand Canyon Supergroup[6]. The Great Unconformity is well exposed in the Grand Canyon, but this geomorphic surface, which records the erosion and weathering of continental crust followed by sediment accumulation, can be traced across Laurentia and globally, including Gondwana[7,8], Baltica[8], Avalonia[9] and Siberia[10], making it the most widely recognized and distinctive stratigraphic surface in the rock record. It is also notable because the Cambrian sediments that overlie it in many regions preserve the first skeletonized crown-group animals, a fact that some palaeontologists have interpreted as evidence for stratigraphic bias and an incomplete record of early animal evolution[1,6].

Here we use stratigraphic and lithologic data for 21,521 rock units from 830 geographic locations in North America, in conjunction with petrologic and geochemical data (Methods; see also Supplementary Information), to explore the hypothesis that the formation of the Great Unconformity is causally linked to the evolution of biomineralization; this linkage is proposed to occur by means of the geochemical effects of prolonged continental denudation followed by enhanced physical and chemical weathering of continental crust during terminal Ediacaran and Cambrian time.

The Cambrian- to Early Ordovician-aged sediments of the Sauk Sequence[11,12] that overlie the Great Unconformity are time-transgressive, such that Early Cambrian sediments occur on the margins of the palaeocontinents and Late Cambrian sediments overlie the Great Unconformity in continental interiors (Fig. 1). The spatial extent of the Sauk Sequence is comparable to other Phanerozoic continent-scale sedimentary sequences[11,12], but its geological characteristics are unique. In most places, undeformed Cambrian sedimentary rocks deposited on Earth's surface rest non-conformably on much older continental crystalline basement rocks, many of which were formed and/or metamorphosed within the Earth's crust (Fig. 2a). Thus, the Great Unconformity marks the termination of an extended period of continental denudation that exhumed and exposed large areas of igneous and metamorphic rocks to subaerial weathering before marine transgression and subsequent sedimentation.

Continental-scale marine transgression during the Cambrian–Early Ordovician accentuated rates of weathering on the Great Unconformity by shifting landward the position of the erosive transgressive shoreface system, often called the 'wave-base razor'[13], as well as adjacent transitional backshore, aeolian and fluvial systems. As a result, much of the soil and weathered basement rock (regolith) that covered low-relief continental interiors[7] was eroded and mobilized during the transgression, thereby exposing silicate mineral surfaces to weathering over an area that is unprecedented in the rock record (Fig. 2a). This is important because freshly exposed rock weathers chemically at rates more than three times faster than undisturbed soils and regolith[14,15], and

Figure 1 | Sauk Sequence in North America. Distribution and age of the oldest Phanerozoic sedimentary rocks in North America.

[1]Department of Geoscience, University of Wisconsin, Madison, Wisconsin 53706, USA. [2]Geology Department, Pomona College, Claremont, California 91711, USA.

paleobiodb.org

[PaleoDeepDive (Shanan Peters, Chris Ré)]

Paleobiology

The transition between the Proterozoic and Phanerozoic eons, beginning 542 million years (Myr) ago, is distinguished by the diversification of multicellular animals and by their acquisition of mineralized skeletons during the Cambrian period[1]. Considerable progress has been made in documenting and more precisely correlating biotic patterns in the Neoproterozoic–Cambrian fossil record with geochemical and physical environmental perturbations[2–5], but the mechanisms responsible for those perturbations remain uncertain[1,2]. Here we use new stratigraphic and geochemical data to show that early Palaeozoic marine sediments deposited approximately 540–480 Myr ago record both an expansion in the area of shallow epicontinental seas and anomalous patterns of chemical sedimentation that are indicative of increased oceanic alkalinity and enhanced chemical weathering of continental crust. These geochemical conditions were caused by a protracted period of widespread continental denudation during the Neoproterozoic followed by extensive physical reworking of soil, regolith and basement rock during the first continental-scale marine transgression of the Phanerozoic. The resultant globally occurring stratigraphic surface, which in most regions separates continental crystalline basement rock from much younger Cambrian shallow marine sedimentary deposits, is known as the Great Unconformity[6]. Although Darwin and others have interpreted this widespread hiatus in sedimentation on the continents as a failure of the geologic record, this palaeogeomorphic surface represents a unique physical environmental boundary condition that affected seawater chemistry during a time of profound expansion of shallow marine habitats. Thus, the formation of the Great Unconformity may have been an environmental trigger for the evolution of biomineralization and the 'Cambrian explosion' of ecologic and taxonomic diversity following the Neoproterozoic emergence of animals.

The term Great Unconformity was first used in the year 1869 to describe the prominent stratigraphic surface in the Grand Canyon that separates the shallow marine, ~525-Myr-old Cambrian Tapeats Sandstone from the underlying metamorphosed, 1,740-Myr-old Vishnu Schist and structurally tilted sedimentary rocks of the 1,200–740 Myr-old Grand Canyon Supergroup[6]. The Great Unconformity is well exposed in the Grand Canyon, but this geomorphic surface, which records the erosion and weathering of continental crust followed by sediment accumulation, can be traced across Laurentia and globally, including Gondwana[7,8], Baltica[8], Avalonia[9] and Siberia[10], making it the most widely recognized and distinctive stratigraphic surface in the rock record. It is also notable because the Cambrian sediments that overlie it in many regions preserve the first skeletonized crown-group animals, a fact that some palaeontologists have interpreted as evidence for stratigraphic bias and an incomplete record of early animal evolution[1,6].

Here we use stratigraphic and lithologic data for 21,521 rock units from 830 geographic locations in North America, in conjunction with petrologic and geochemical data (Methods; see also Supplementary Information), to explore the hypothesis that the formation of the Great Unconformity is causally linked to the evolution of biomineralization; this linkage is proposed to occur by means of the geochemical effects of prolonged continental denudation followed by enhanced physical and chemical weathering of continental crust during terminal Ediacaran and Cambrian time.

The Cambrian- to Early Ordovician-aged sediments of the Sauk Sequence[11,12] that overlie the Great Unconformity are time-transgressive, such that Early Cambrian sediments occur on the margins of the palaeocontinents and Late Cambrian sediments overlie the Great Unconformity in continental interiors (Fig. 1). The spatial extent of the Sauk Sequence is comparable to other Phanerozoic continent-scale sedimentary sequences[11,12], but its geological characteristics are unique. In most places, undeformed Cambrian sedimentary rocks deposited on Earth's surface rest non-conformably on much older continental crystalline basement rocks, many of which were formed and/or metamorphosed within the Earth's crust (Fig. 2a). Thus, the Great Unconformity marks the termination of an extended period of continental denudation that exhumed and exposed large areas of igneous and metamorphic rocks to subaerial weathering before marine transgression and subsequent sedimentation.

Continental-scale marine transgression during the Cambrian–Early Ordovician accentuated rates of weathering on the Great Unconformity by shifting landward the position of the erosive transgressive shoreface system, often called the 'wave-base razor'[13], as well as adjacent transitional backshore, aeolian and fluvial systems. As a result, much of the soil and weathered basement rock (regolith) that covered low-relief continental interiors[7] was eroded and mobilized during the transgression, thereby exposing silicate mineral surfaces to weathering over an area that is unprecedented in the rock record (Fig. 2a). This is important because freshly exposed rock weathers chemically at rates more than three times faster than undisturbed soils and regolith[14,15], and

472, 488, 542, Myr ago, Ordovician, Cambrian, Early, Late, Middle, Early

Figure 1 | Sauk Sequence in North America. Distribution and age of the oldest Phanerozoic sedimentary rocks in North America.

[1]Department of Geoscience, University of Wisconsin, Madison, Wisconsin 53706, USA. [2]Geology Department, Pomona College, Claremont, California 91711, USA.

paleobiodb.org

Where was the last American Mastadon found?

Paleobiology

The transition between the Proterozoic and Phanerozoic eons, beginning 542 million years (Myr) ago, is distinguished by the diversification of multicellular animals and by their acquisition of mineralized skeletons during the Cambrian period[1]. Considerable progress has been made in documenting and more precisely correlating biotic patterns in the Neoproterozoic–Cambrian fossil record with geochemical and physical environmental perturbations[2–5], but the mechanisms responsible for those perturbations remain uncertain[1,2]. Here we use new stratigraphic and geochemical data to show that early Palaeozoic marine sediments deposited approximately 540–480 Myr ago record both an expansion in the area of shallow epicontinental seas and anomalous patterns of chemical sedimentation that are indicative of increased oceanic alkalinity and enhanced chemical weathering of continental crust. These geochemical conditions were caused by a protracted period of widespread continental denudation during the Neoproterozoic followed by extensive physical reworking of soil, regolith and basement rock during the first continental-scale marine transgression of the Phanerozoic. The resultant globally occurring stratigraphic surface, which in most regions separates continental crystalline basement rock from much younger Cambrian shallow marine sedimentary deposits, is known as the Great Unconformity[6]. Although Darwin and others have interpreted this widespread hiatus in sedimentation on the continents as a failure of the geologic record, this palaeogeomorphic surface represents a unique physical environmental boundary condition that affected seawater chemistry during a time of profound expansion of shallow marine habitats. Thus, the formation of the Great Unconformity may have been an environmental trigger for the evolution of biomineralization and the 'Cambrian explosion' of ecologic and taxonomic diversity following the Neoproterozoic emergence of animals.

The term Great Unconformity was first used in the year 1869 to describe the prominent stratigraphic surface in the Grand Canyon that separates the shallow marine, ~525-Myr-old Cambrian Tapeats Sandstone from the underlying metamorphosed, 1,740-Myr-old Vishnu Schist and structurally tilted sedimentary rocks of the 1,200–740 Myr-old Grand Canyon Supergroup[6]. The Great Unconformity is well exposed in the Grand Canyon, but this geomorphic surface, which records the erosion and weathering of continental crust followed by sediment accumulation, can be traced across Laurentia and globally, including Gondwana[7,8], Baltica[8], Avalonia[9] and Siberia[10], making it the most widely recognized and distinctive stratigraphic surface in the rock record. It is also notable because the Cambrian sediments that overlie it in many regions preserve the first skeletonized crown-group animals, a fact that some palaeontologists have interpreted as evidence for stratigraphic bias and an incomplete record of early animal evolution[1,6].

Here we use stratigraphic and lithologic data for 21,521 rock units from 830 geographic locations in North America, in conjunction with petrologic and geochemical data (Methods; see also Supplementary Information), to explore the hypothesis that the formation of the Great Unconformity is causally linked to the evolution of biomineralization; this linkage is proposed to occur by means of the geochemical effects of prolonged continental denudation followed by enhanced physical and chemical weathering of continental crust during terminal Ediacaran and Cambrian time.

The Cambrian- to Early Ordovician-aged sediments of the Sauk Sequence[11,12] that overlie the Great Unconformity are time-transgressive, such that Early Cambrian sediments occur on the margins of the palaeocontinents and Late Cambrian sediments overlie the Great Unconformity in continental interiors (Fig. 1). The spatial extent of the Sauk Sequence is comparable to other Phanerozoic continent-scale sedimentary sequences[11,12], but its geological characteristics are unique. In most places, undeformed Cambrian sedimentary rocks deposited on Earth's surface rest non-conformably on much older continental crystalline basement rocks, many of which were formed and/or metamorphosed within the Earth's crust (Fig. 2a). Thus, the Great Unconformity marks the termination of an extended period of continental denudation that exhumed and exposed large areas of igneous and metamorphic rocks to subaerial weathering before marine transgression and subsequent sedimentation.

Continental-scale marine transgression during the Cambrian–Early Ordovician accentuated rates of weathering on the Great Unconformity by shifting landward the position of the erosive transgressive shoreface system, often called the 'wave-base razor'[13], as well as adjacent transitional backshore, aeolian and fluvial systems. As a result, much of the soil and weathered basement rock (regolith) that covered low-relief continental interiors[7] was eroded and mobilized during the transgression, thereby exposing silicate mineral surfaces to weathering over an area that is unprecedented in the rock record (Fig. 2a). This is important because freshly exposed rock weathers chemically at rates more than three times faster than undisturbed soils and regolith[14,15], and

Figure 1 | Sauk Sequence in North America. Distribution and age of the oldest Phanerozoic sedimentary rocks in North America.

[1]Department of Geoscience, University of Wisconsin, Madison, Wisconsin 53706, USA. [2]Geology Department, Pomona College, Claremont, California 91711, USA.

paleobiodb.org

Where was the last American Mastadon found?

How long do species exist on average before going extinct?

Paleobiology

The transition between the Proterozoic and Phanerozoic eons, beginning 542 million years (Myr) ago, is distinguished by the diversification of multicellular animals and by their acquisition of mineralized skeletons during the Cambrian period[1]. Considerable progress has been made in documenting and more precisely correlating biotic patterns in the Neoproterozoic–Cambrian fossil record with geochemical and physical environmental perturbations[2–5], but the mechanisms responsible for those perturbations remain uncertain[1,2]. Here we use new stratigraphic and geochemical data to show that early Palaeozoic marine sediments deposited approximately 540–480 Myr ago record both an expansion in the area of shallow epicontinental seas and anomalous patterns of chemical sedimentation that are indicative of increased oceanic alkalinity and enhanced chemical weathering of continental crust. These geochemical conditions were caused by a protracted period of widespread continental denudation during the Neoproterozoic followed by extensive physical reworking of soil, regolith and basement rock during the first continental-scale marine transgression of the Phanerozoic. The resultant globally occurring stratigraphic surface, which in most regions separates continental crystalline basement rock from much younger Cambrian shallow marine sedimentary deposits, is known as the Great Unconformity[6]. Although Darwin and others have interpreted this widespread hiatus in sedimentation on the continents as a failure of the geologic record, this palaeogeomorphic surface represents a unique physical environmental boundary condition that affected seawater chemistry during a time of profound expansion of shallow marine habitats. Thus, the formation of the Great Unconformity may have been an environmental trigger for the evolution of biomineralization and the 'Cambrian explosion' of ecologic and taxonomic diversity following the Neoproterozoic emergence of animals.

The term Great Unconformity was first used in the year 1869 to describe the prominent stratigraphic surface in the Grand Canyon that separates the shallow marine, ~525-Myr-old Cambrian Tapeats Sandstone from the underlying metamorphosed, 1,740-Myr-old Vishnu Schist and structurally tilted sedimentary rocks of the 1,200–740 Myr-old Grand Canyon Supergroup[6]. The Great Unconformity is well exposed in the Grand Canyon, but this geomorphic surface, which records the erosion and weathering of continental crust followed by sediment accumulation, can be traced across Laurentia and globally, including Gondwana[7,8], Baltica[9], Avalonia[9] and Siberia[10], making it the most widely recognized and distinctive stratigraphic surface in the rock record. It is also notable because the Cambrian sediments that overlie it in many regions preserve the first skeletonized crown-group animals, a fact that some palaeontologists have interpreted as evidence for stratigraphic bias and an incomplete record of early animal evolution[1,6].

Here we use stratigraphic and lithologic data for 21,521 rock units from 830 geographic locations in North America, in conjunction with petrologic and geochemical data (Methods; see also Supplementary Information), to explore the hypothesis that the formation of the Great Unconformity is causally linked to the evolution of biomineralization; this linkage is proposed to occur by means of the geochemical effects of prolonged continental denudation followed by enhanced physical and chemical weathering of continental crust during terminal Ediacaran and Cambrian time.

The Cambrian- to Early Ordovician-aged sediments of the Sauk Sequence[11,12] that overlie the Great Unconformity are time-transgressive, such that Early Cambrian sediments occur on the margins of the palaeocontinents and Late Cambrian sediments overlie the Great Unconformity in continental interiors (Fig. 1). The spatial extent of the Sauk Sequence is comparable to other Phanerozoic continent-scale sedimentary sequences[11,12], but its geological characteristics are unique. In most places, undeformed Cambrian sedimentary rocks deposited on Earth's surface rest non-conformably on much older continental crystalline basement rocks, many of which were formed and/or metamorphosed within the Earth's crust (Fig. 2a). Thus, the Great Unconformity marks the termination of an extended period of continental denudation that exhumed and exposed large areas of igneous and metamorphic rocks to subaerial weathering before marine transgression and subsequent sedimentation.

Continental-scale marine transgression during the Cambrian–Early Ordovician accentuated rates of weathering on the Great Unconformity by shifting landward the position of the erosive transgressive shoreface system, often called the 'wave-base razor'[13], as well as adjacent transitional backshore, aeolian and fluvial systems. As a result, much of the soil and weathered basement rock (regolith) that covered low-relief continental interiors[7] was eroded and mobilized during the transgression, thereby exposing silicate mineral surfaces to weathering over an area that is unprecedented in the rock record (Fig. 2a). This is important because freshly exposed rock weathers chemically at rates more than three times faster than undisturbed soils and regolith[14,15], and

Figure 1 | Sauk Sequence in North America. Distribution and age of the oldest Phanerozoic sedimentary rocks in North America.

[1]Department of Geoscience, University of Wisconsin, Madison, Wisconsin 53706, USA. [2]Geology Department, Pomona College, Claremont, California 91711, USA.

paleobiodb.org

Where was the last American Mastadon found?

How long do species exist on average before going extinct?

Goal: help scientists answer macro-questions

Challenge: requires computation / aggregation

Question answering via semantic parsing

Where was the last American Mastadon found?

Question answering via semantic parsing

Where was the last American Mastadon found?

semantic parsing

$\text{LocationOf.argmax}(\text{Type.Occurrence} \sqcap \text{Genus.Mammut}, \text{Period})$

Question answering via semantic parsing

Where was the last American Mastadon found?

semantic parsing

$\text{LocationOf.argmax}(\text{Type.Occurrence} \sqcap \text{Genus.Mammut}, \text{Period})$

execute

`New Mexico`

Question answering via semantic parsing

Where was the last American Mastadon found?

semantic parsing

execute

`New Mexico`

Email assistant via semantic parsing

Send a reminder to all authors who haven't sent an abstract.

Email assistant via semantic parsing

Send a reminder to all authors who haven't sent an abstract.

$\Downarrow$ semantic parsing

$\forall x \in (\text{Author} \sqcap \neg\text{Sent.Subject.Abstract}) : \text{Remind}(x)$

Email assistant via semantic parsing

Send a reminder to all authors who haven't sent an abstract.

$\Big\downarrow$ semantic parsing

$\forall x \in (\text{Author} \sqcap \neg\text{Sent.Subject.Abstract}) : \text{Remind}(x)$

$\Big\downarrow$ execute

`[5 emails sent]`

Email assistant via semantic parsing

Send a reminder to all authors who haven't sent an abstract.

semantic parsing

execute

`[5 emails sent]`

Semantic parsing

[utterance: user input]

semantic parsing

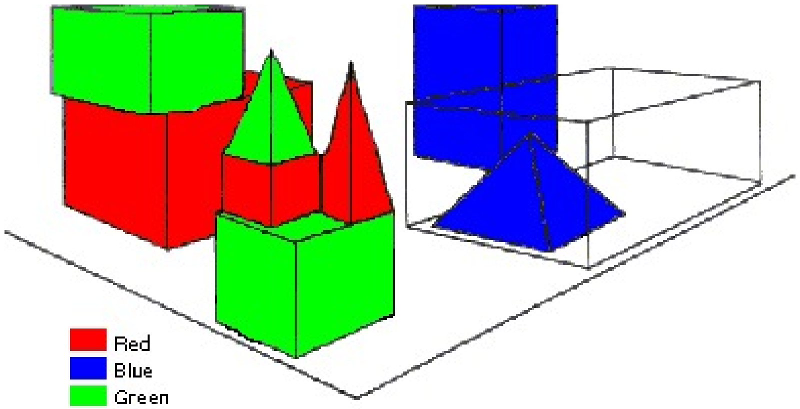
[program]

execute

`[behavior: user output]`

Semantics = how language connects with the world

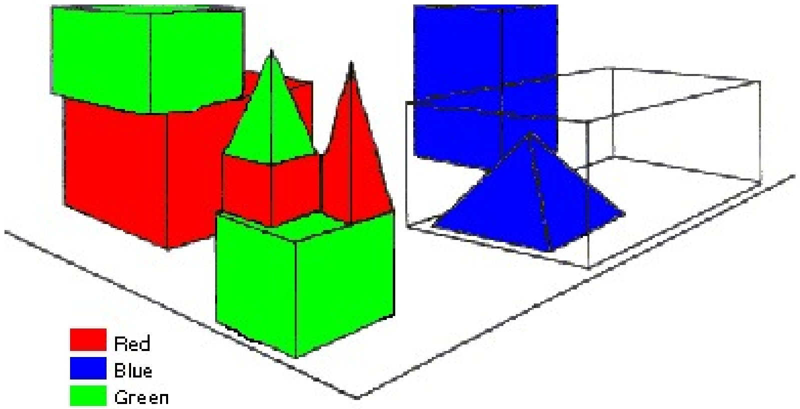
SHRDLU [1971]



SHRDLU [1971]

Person: Pick up a big red block.

Computer: OK.



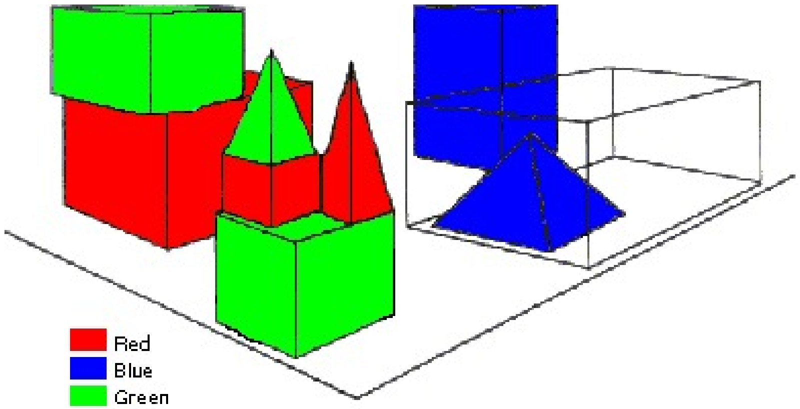
SHRDLU [1971]

Person: Pick up a big red block.

Computer: OK.

Person: Grasp the pyramid.

Computer: I don't understand which pyramid you mean.



SHRDLU [1971]

Person: Pick up a big red block.

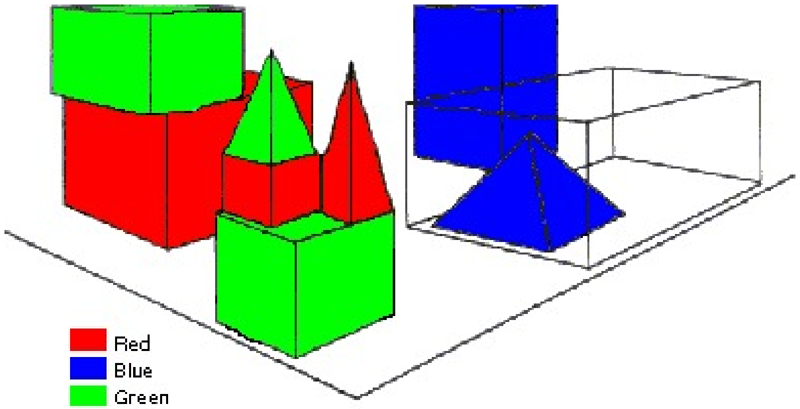
Computer: OK.

Person: Grasp the pyramid.

Computer: I don't understand which pyramid you mean.

Person (changing their mind): Find a block which is taller than the one you are holding and put it into the box.

Computer: By "it", I assume you mean the block which is taller than the one I am holding.



SHRDLU [1971]

Person: Pick up a big red block.

Computer: OK.

Person: Grasp the pyramid.

Computer: I don't understand which pyramid you mean.

Person (changing their mind): Find a block which is taller than the one you are holding and put it into the box.

Computer: By "it", I assume you mean the block which is taller than the one I am holding.

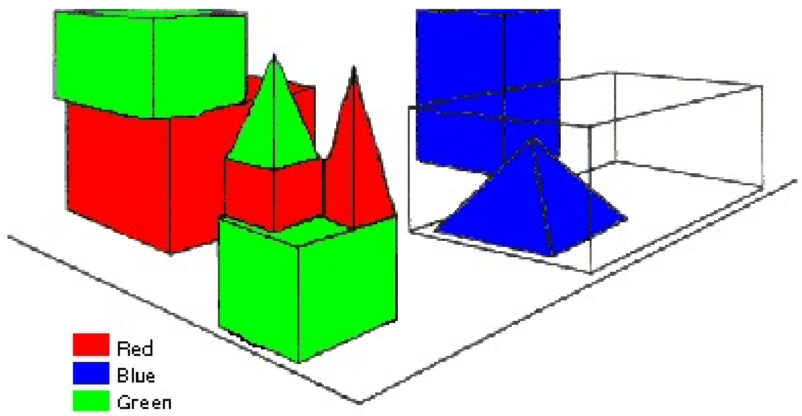
Computer: OK.

Person: What does the box contain?

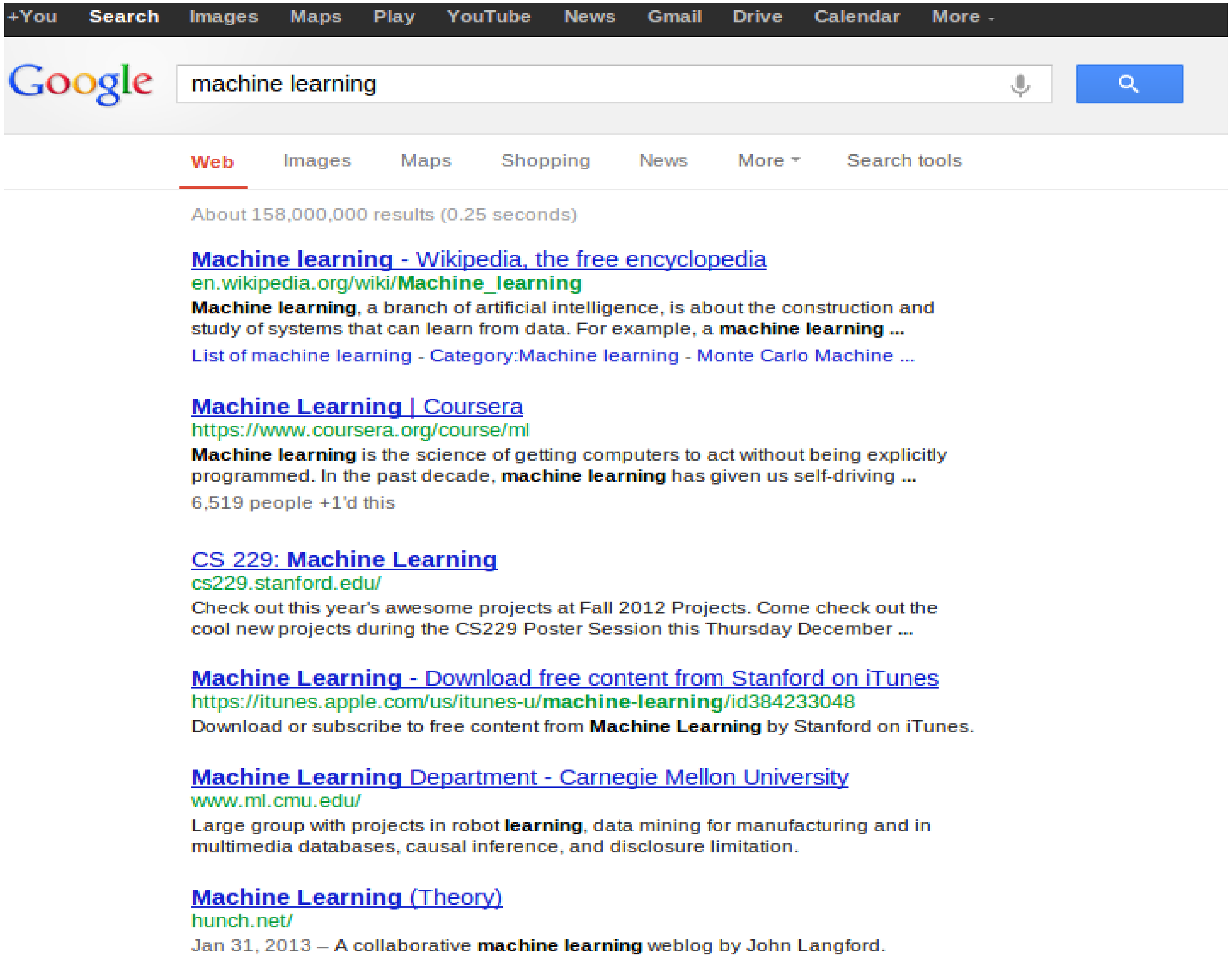
Computer: The blue pyramid and the blue block.

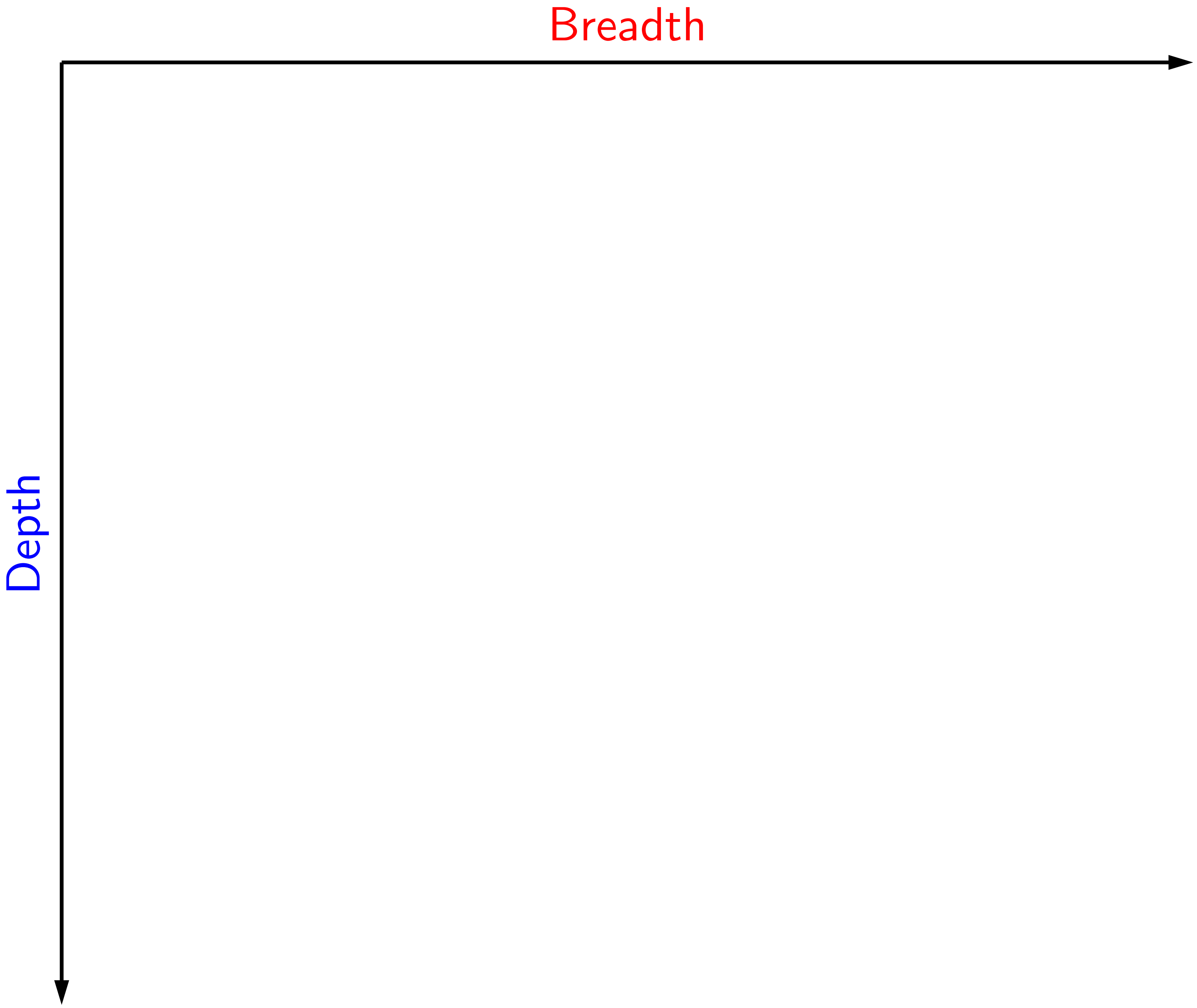
Person: What is the pyramid supported by?

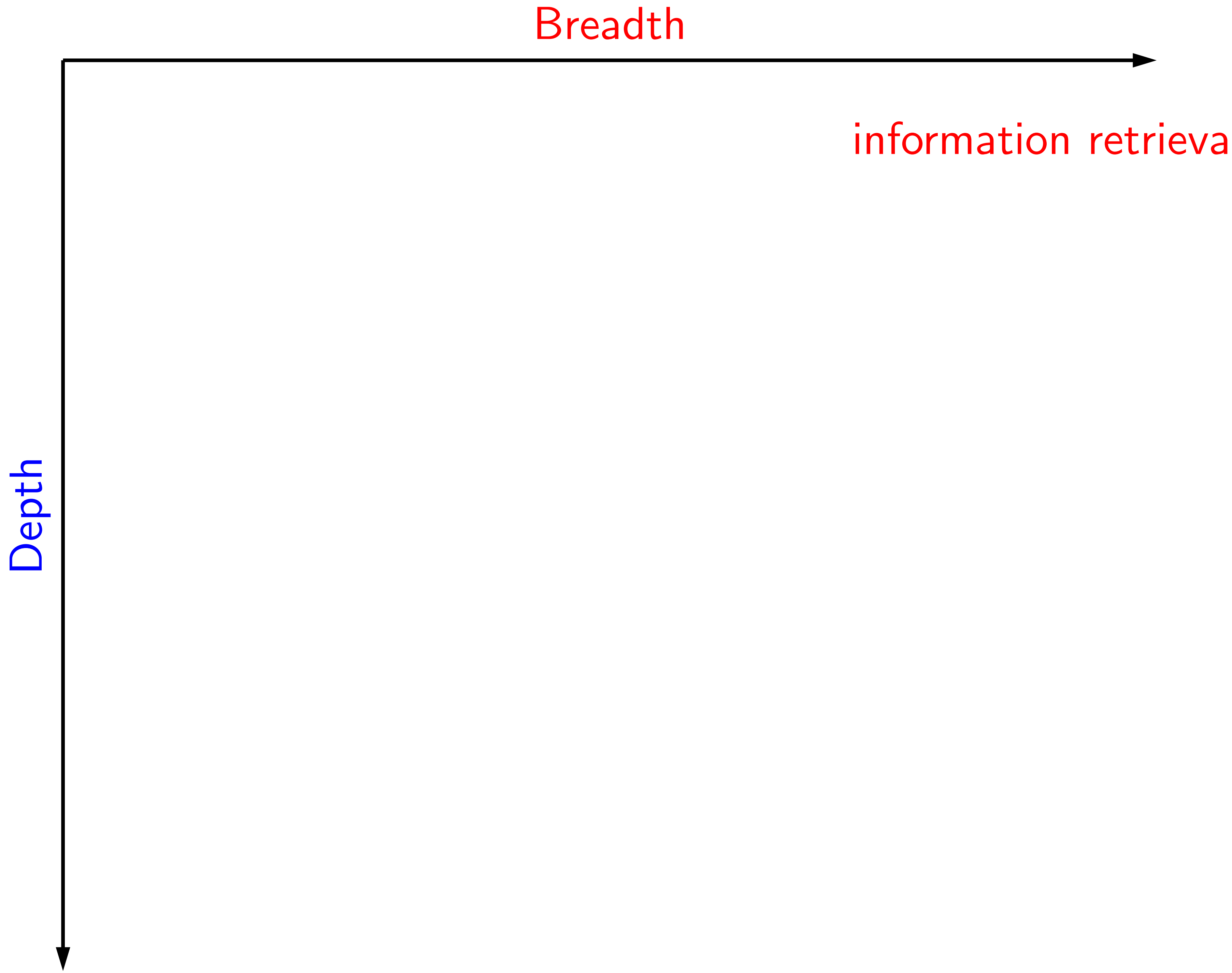
Computer: The box.

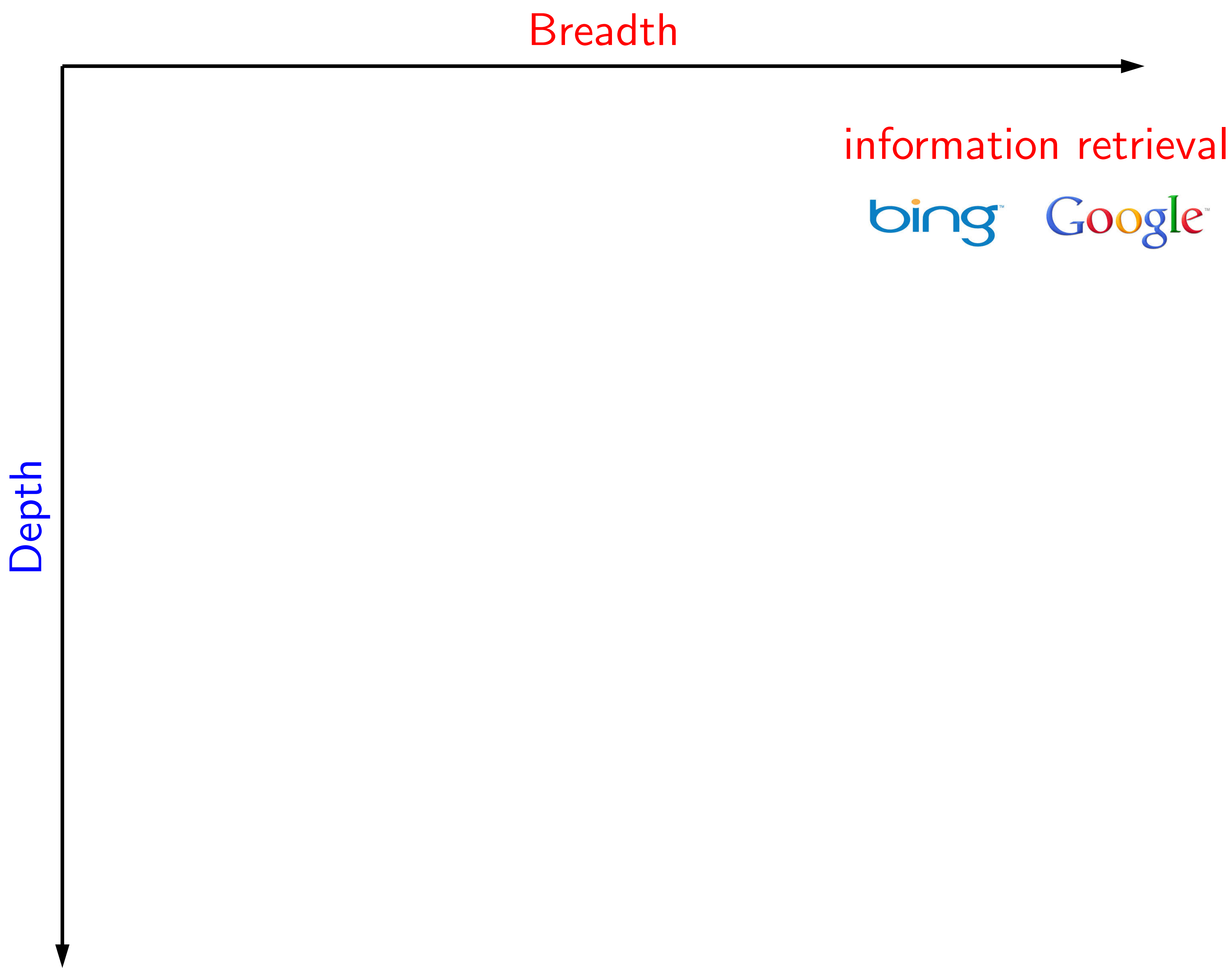


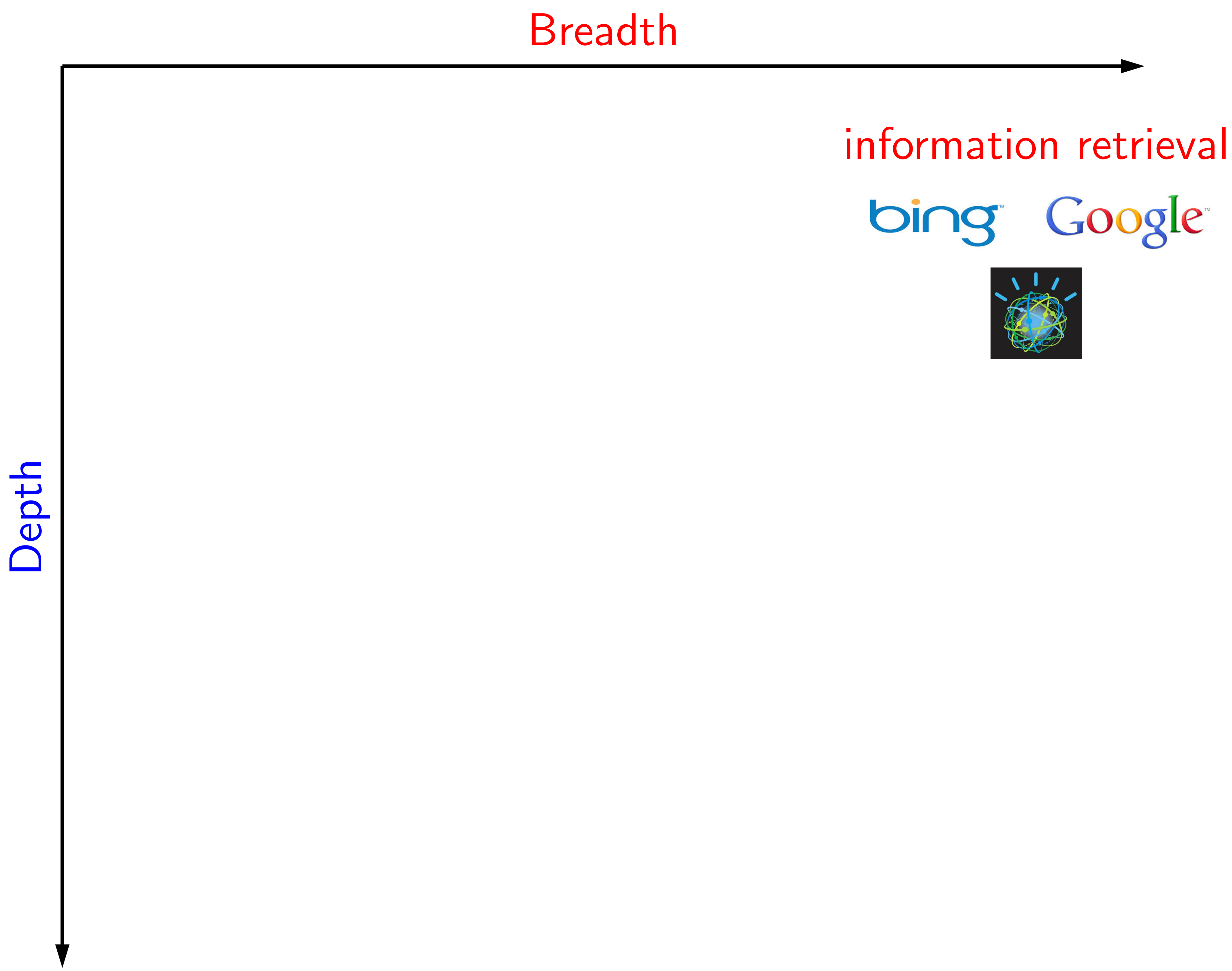
Web search [1990s-]

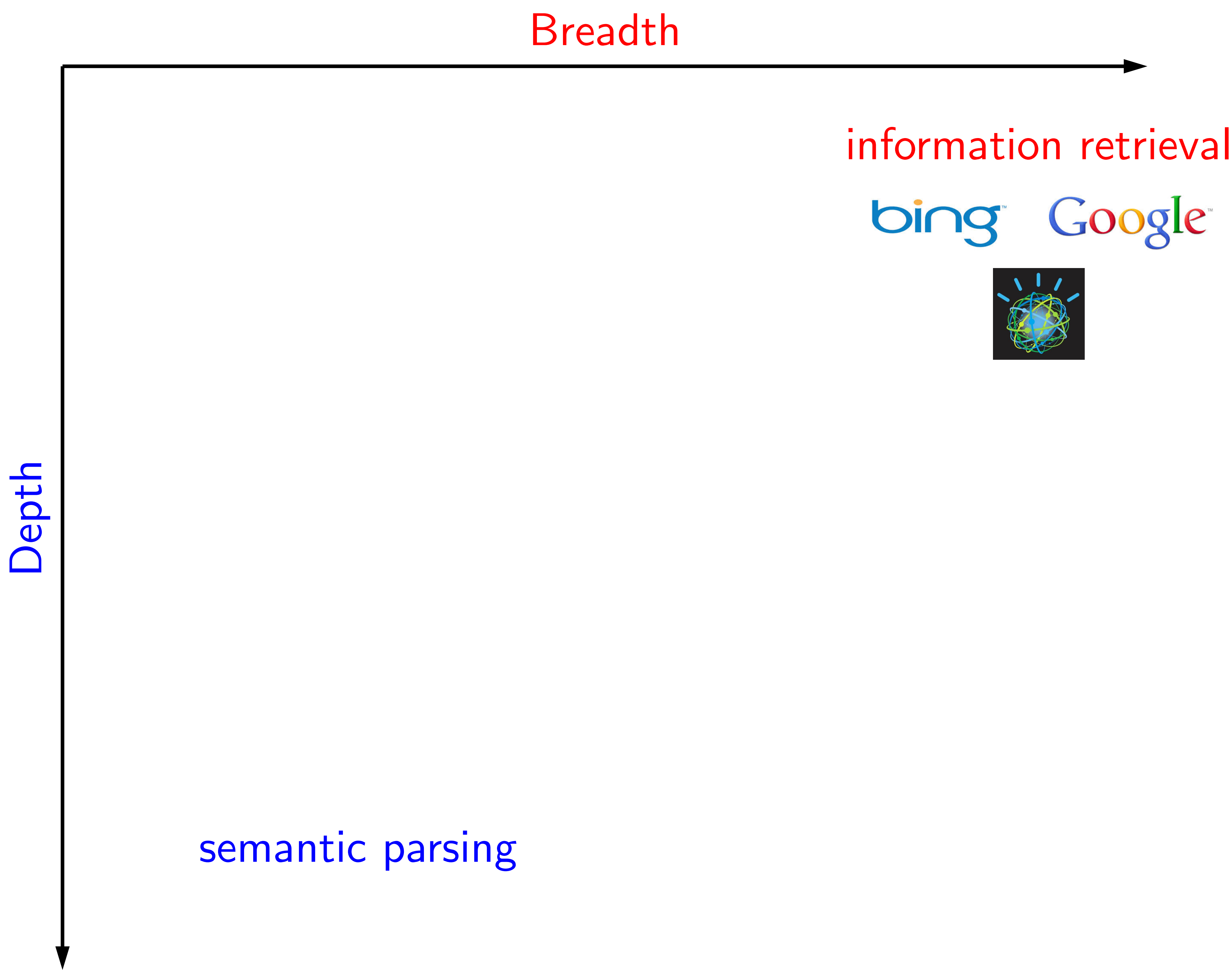



Breadth
Depth


Breadth
information retrieval
Depth


Breadth
Depth
information retrieval

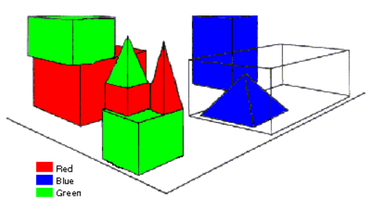

Breadth
information retrieval
Depth


Breadth
information retrieval
Depth
semantic parsing

Breadth

information retrieval

Depth

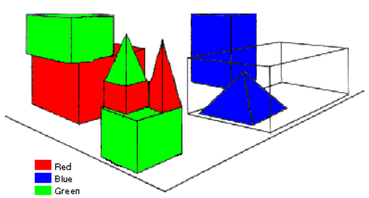


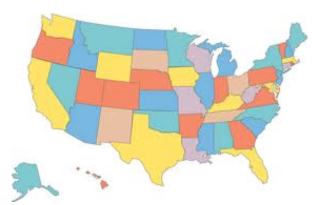
semantic parsing

Breadth

Depth

information retrieval





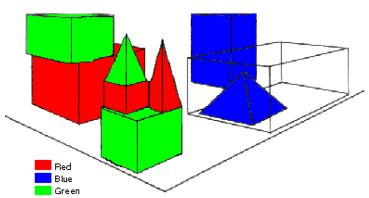
semantic parsing

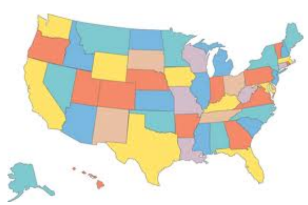
Breadth

information retrieval

Depth

Objective: to develop semantic parsers with the modern sensibilities of web search.





semantic parsing

[Zelle & Mooney, 1996; Zettlemoyer & Collins, 2005; Clarke et al. 2010; Liang et al., 2011]

Training data for semantic parsing

Heavy supervision

What's Bulgaria's capital?

Capital.Bulgaria

When was Walmart started?

DateFounded.Walmart

What movies has Tom Cruise been in?

Type.Movie $\sqcap$ Starring.TomCruise

...

[Zelle & Mooney, 1996; Zettlemoyer & Collins, 2005; Clarke et al. 2010; Liang et al., 2011]

Training data for semantic parsing

Heavy supervision

What's Bulgaria's capital?

Capital.Bulgaria

When was Walmart started?

DateFounded.Walmart

What movies has Tom Cruise been in?

Type.Movie $\sqcap$ Starring.TomCruise

...

[Zelle & Mooney, 1996; Zettlemoyer & Collins, 2005; Clarke et al. 2010; Liang et al., 2011]

Training data for semantic parsing

Heavy supervision	**Light supervision**
What's Bulgaria's capital?	*What's Bulgaria's capital?*
Capital.Bulgaria	`Sofia`
When was Walmart started?	*When was Walmart started?*
DateFounded.Walmart	`1962`
What movies has Tom Cruise been in?	*What movies has Tom Cruise been in?*
Type.Movie $\sqcap$ Starring.TomCruise	`TopGun,VanillaSky,...`
...	...

[Zelle & Mooney, 1996; Zettlemoyer & Collins, 2005; Clarke et al. 2010; Liang et al., 2011]

Training data for semantic parsing

Heavy supervision	**Light supervision**
What's Bulgaria's capital?	*What's Bulgaria's capital?*
Capital.Bulgaria	`Sofia`
When was Walmart started?	*When was Walmart started?*
DateFounded.Walmart	`1962`
What movies has Tom Cruise been in?	*What movies has Tom Cruise been in?*
Type.Movie $\sqcap$ Starring.TomCruise	`TopGun,VanillaSky,...`
...	...

[Zelle & Mooney, 1996; Zettlemoyer & Collins, 2005; Clarke et al. 2010; Liang et al., 2011]

Training data for semantic parsing

Heavy supervision

What's Bulgaria's capital?

Capital.Bulgaria

When was Walmart started?

DateFounded.Walmart

What movies has Tom Cruise been in?

Type.Movie $\sqcap$ Starring.TomCruise

...

Light supervision

What's Bulgaria's capital?

`Sofia`

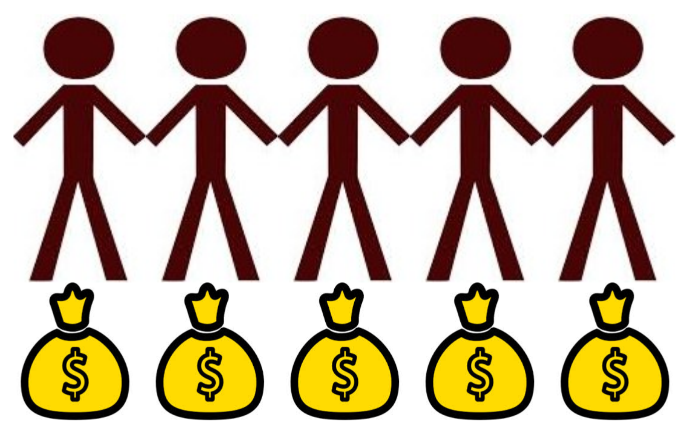
When was Walmart started?

`1962`

What movies has Tom Cruise been in?

`TopGun,VanillaSky,...`

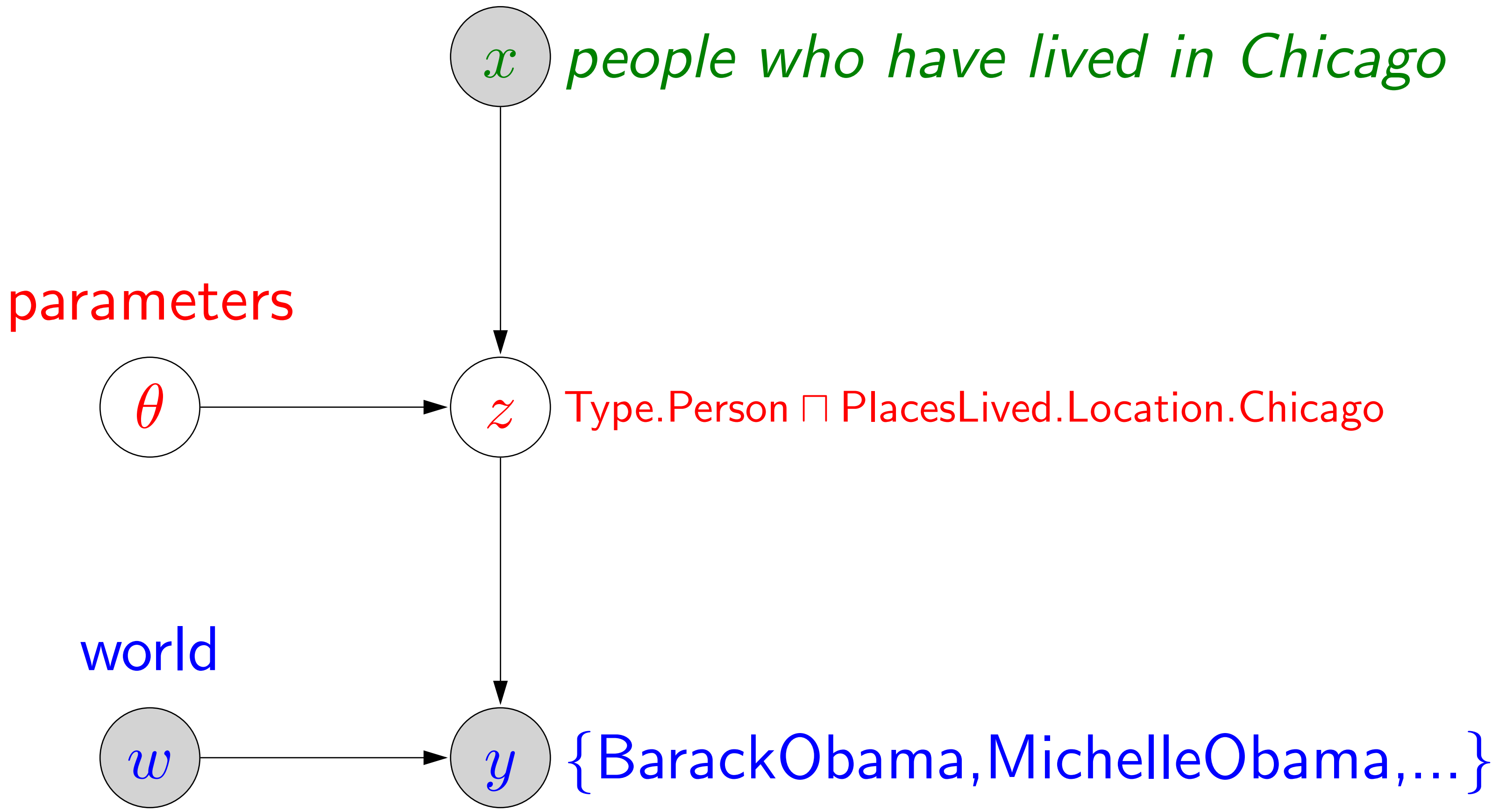
...



Outline

- **A semantic parsing framework**
- A closer look at the elements
  - Logical forms: lambda DCS
  - Lexical coverage
  - Grammar: building logical forms
  - Learning via bootstrapping
  - Leveraging denotations
  - Datasets/results
- Beyond Freebase
- Final remarks

Probabilistic framework



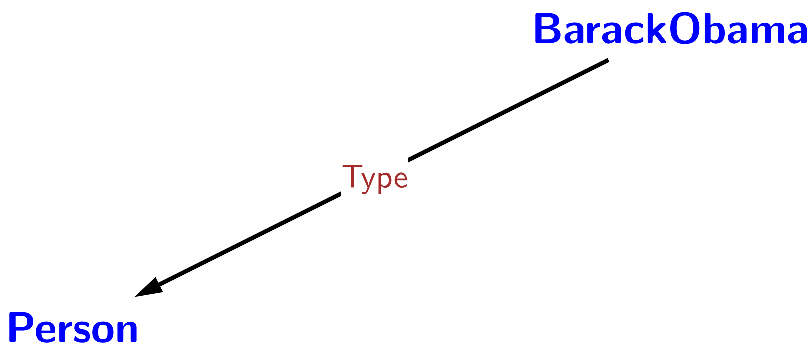
[Bollacker, 2008; Google, 2013]

World: Freebase

100M entities (nodes) 1B assertions (edges)

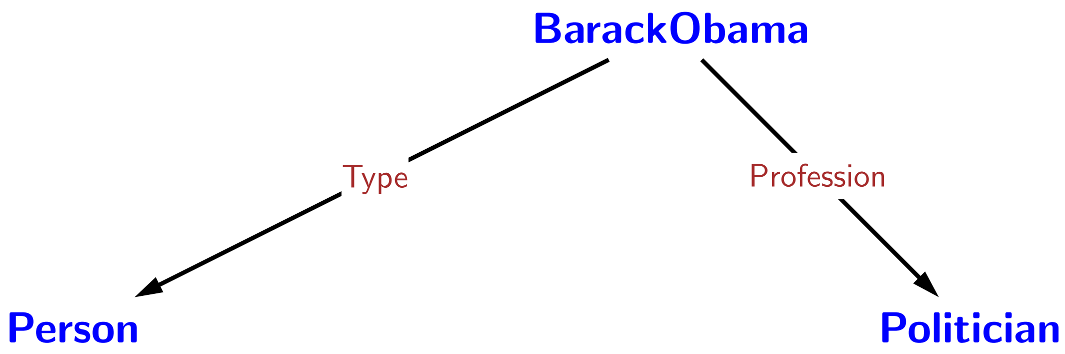
World: Freebase

100M entities (nodes)   1B assertions (edges)



World: Freebase

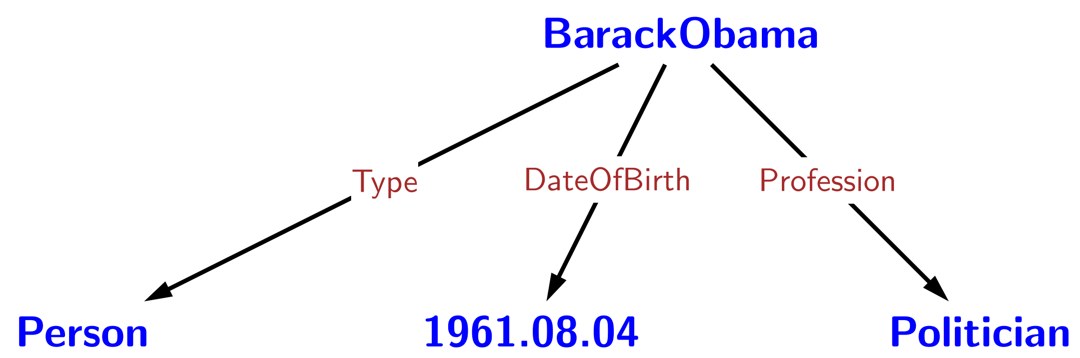
100M entities (nodes)   1B assertions (edges)



[Bollacker, 2008; Google, 2013]

World: Freebase

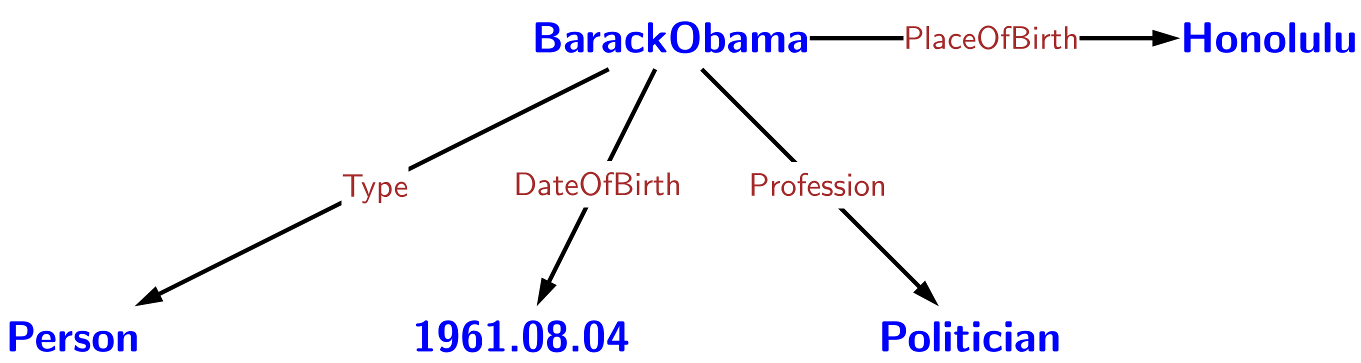
100M entities (nodes)   1B assertions (edges)



[Bollacker, 2008; Google, 2013]

World: Freebase

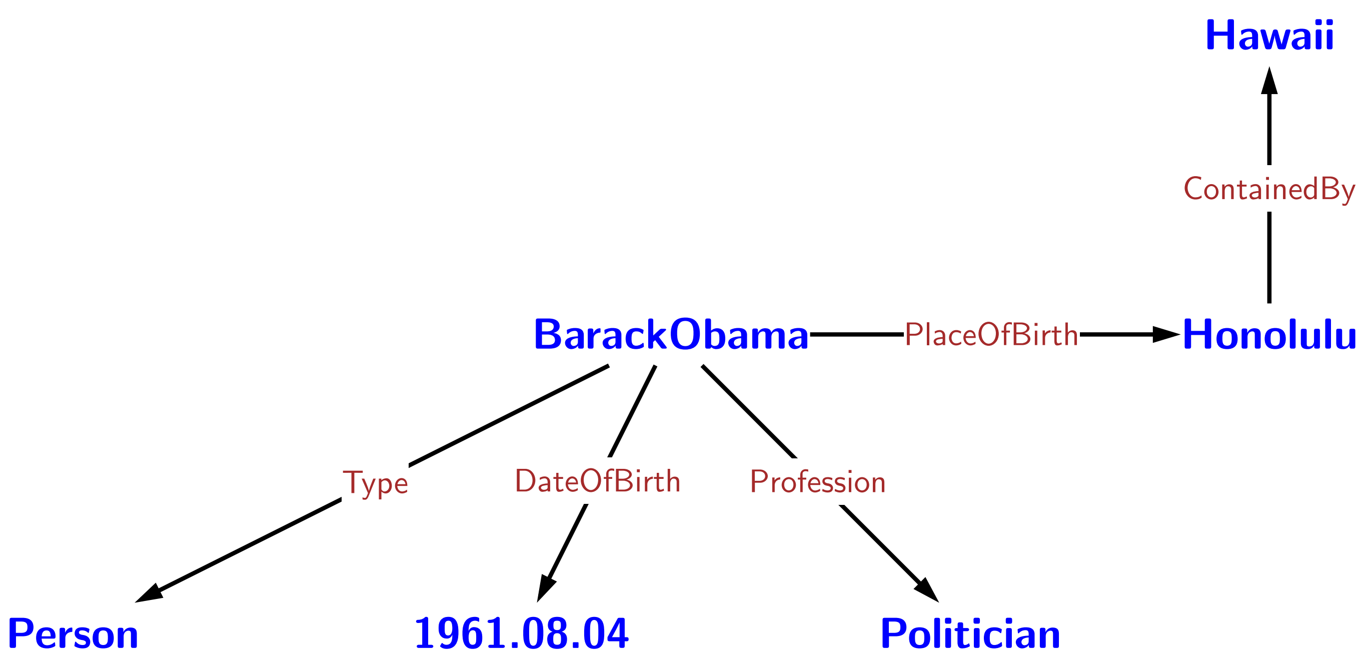
100M entities (nodes)   1B assertions (edges)



[Bollacker, 2008; Google, 2013]

World: Freebase

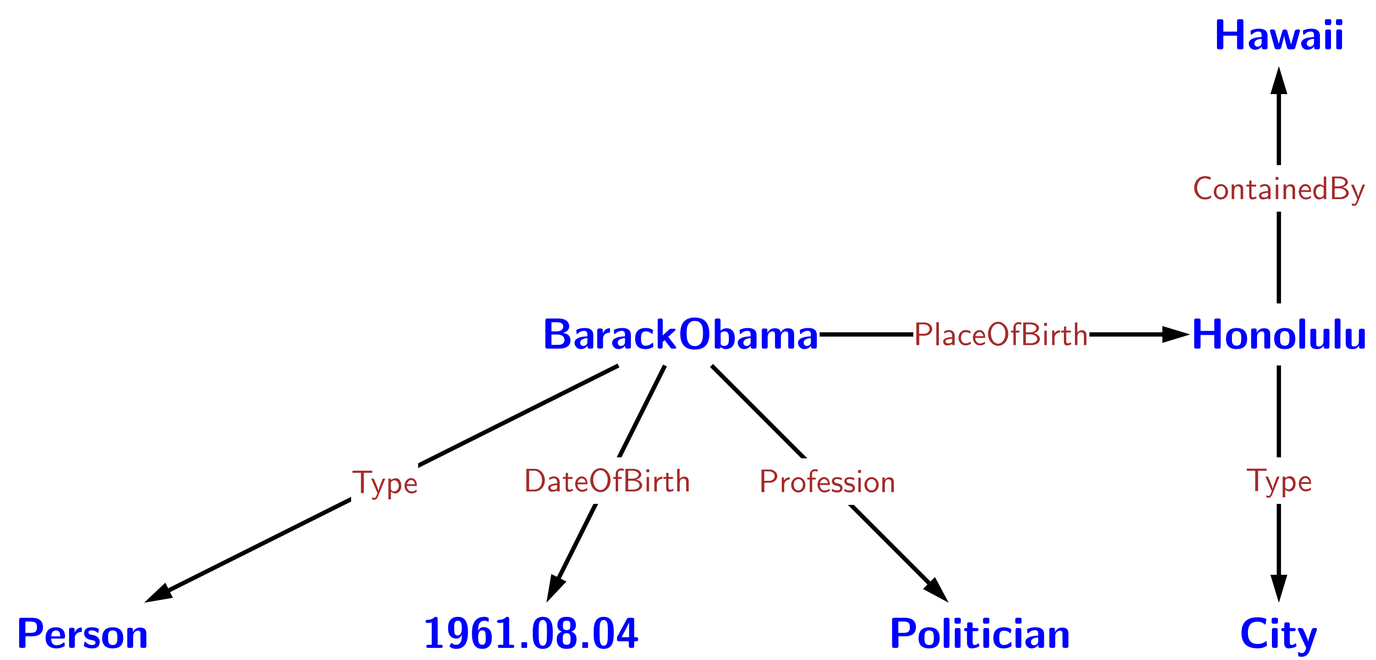
100M entities (nodes)   1B assertions (edges)



[Bollacker, 2008; Google, 2013]

World: Freebase

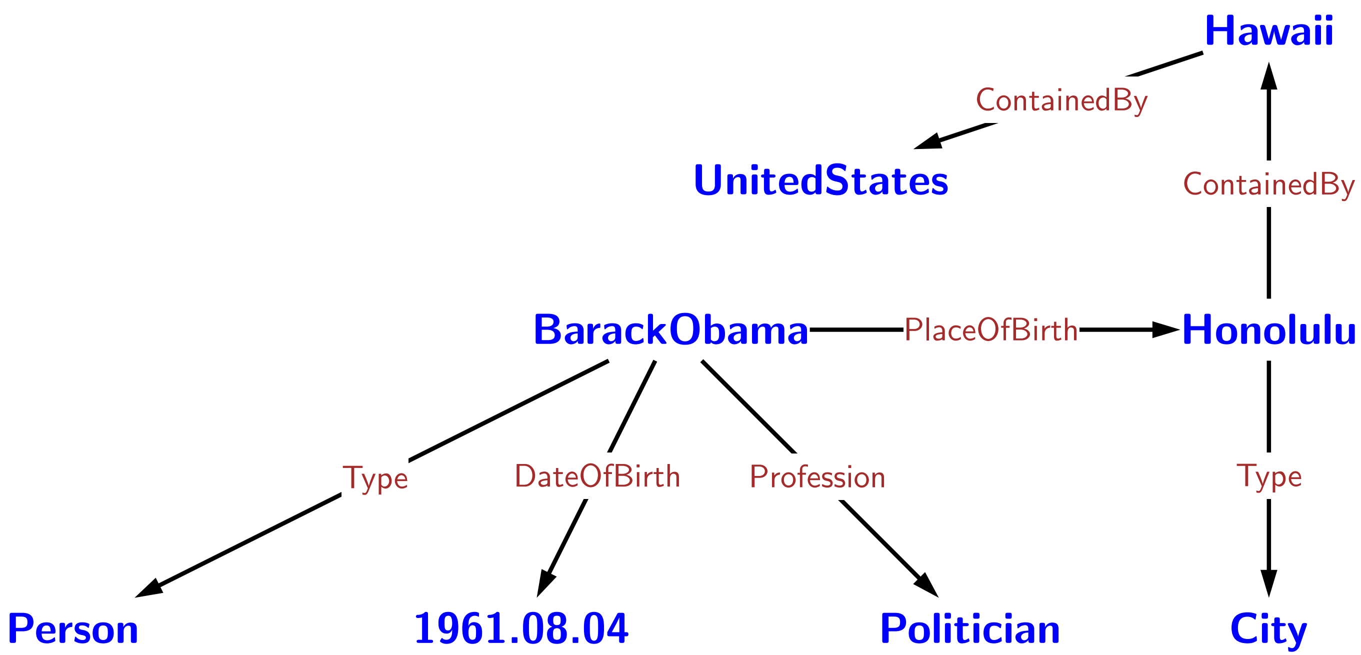
100M entities (nodes)   1B assertions (edges)



[Bollacker, 2008; Google, 2013]

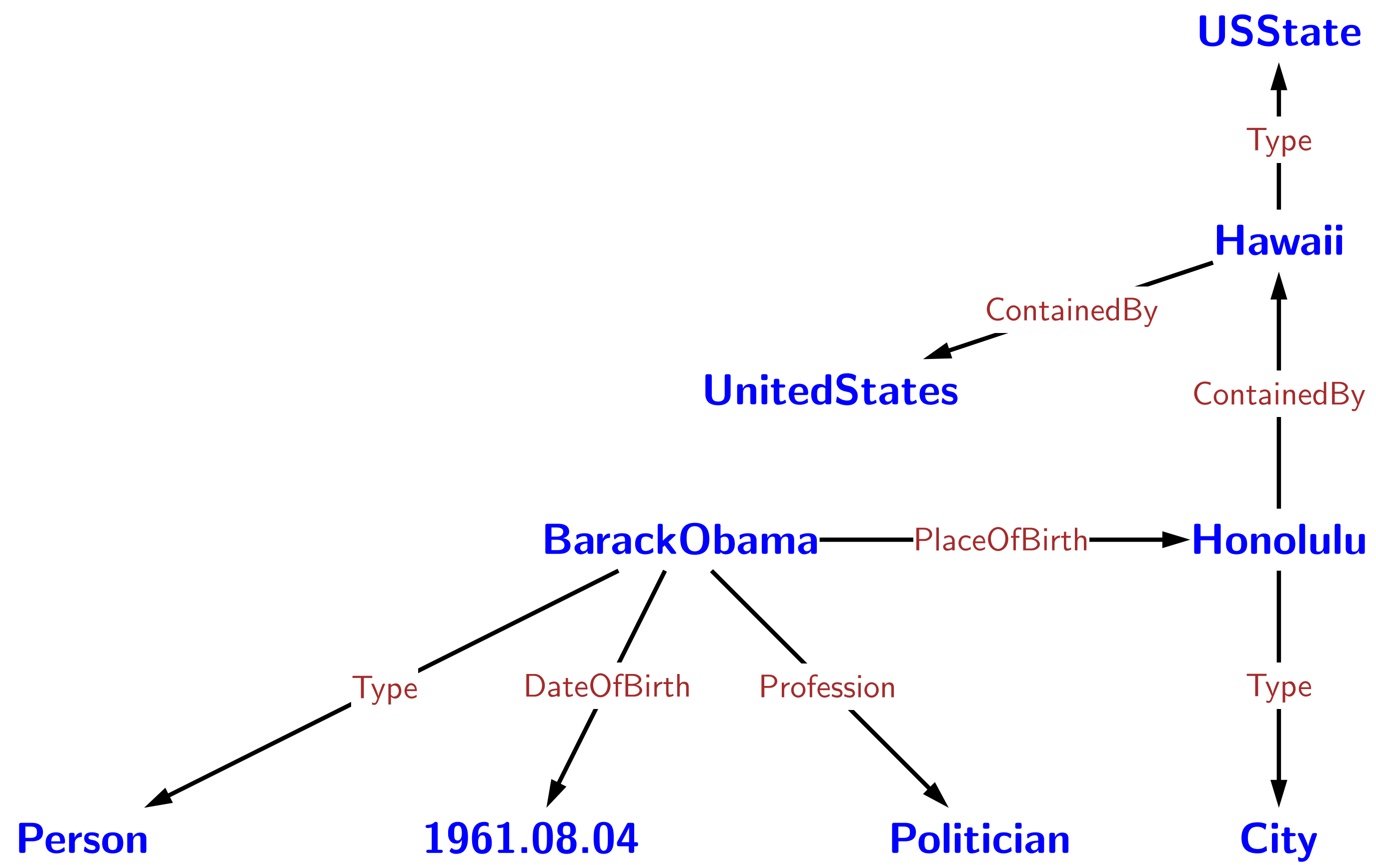
World: Freebase

100M entities (nodes)   1B assertions (edges)



World: Freebase

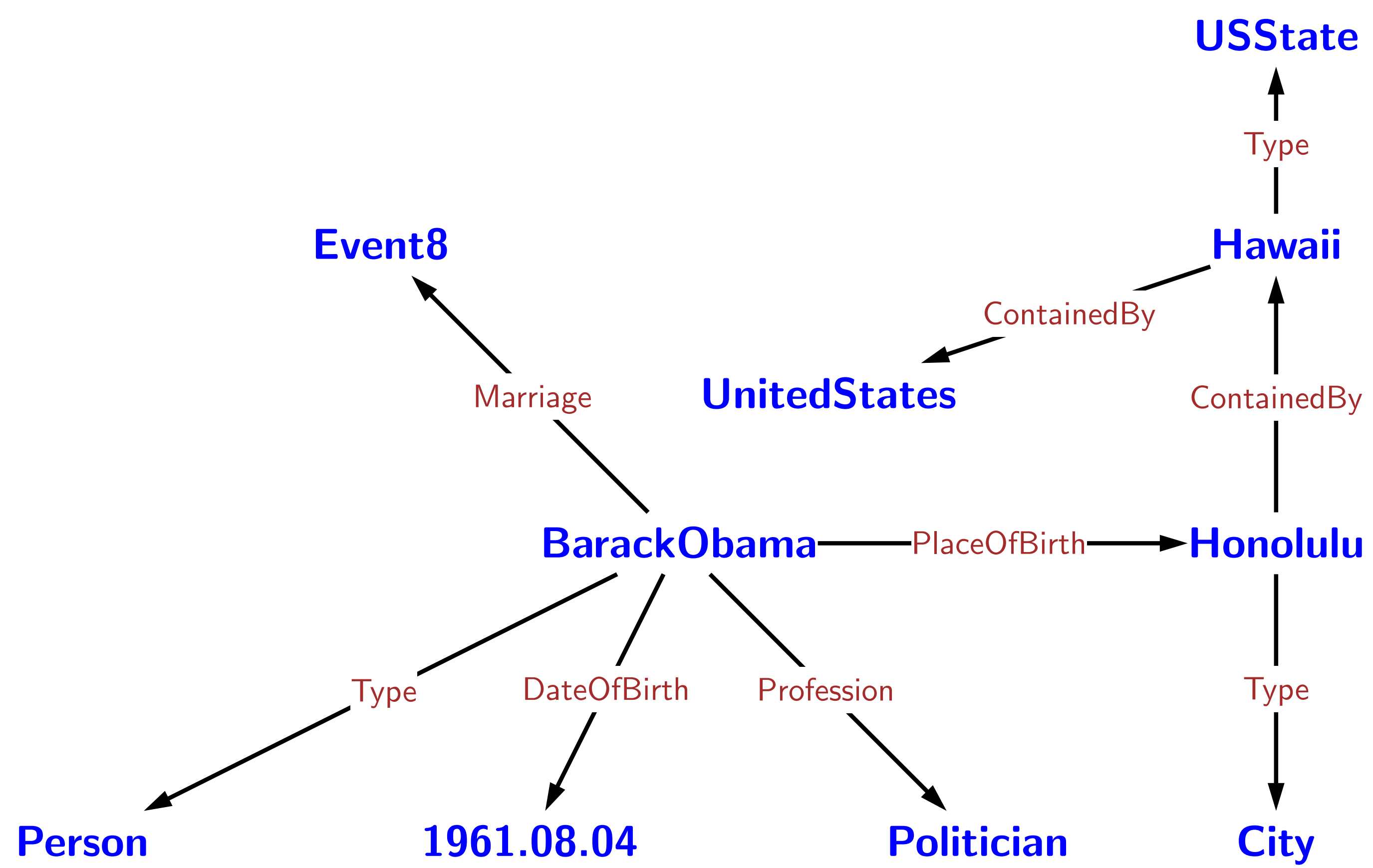
100M entities (nodes)   1B assertions (edges)



[Bollacker, 2008; Google, 2013]

World: Freebase

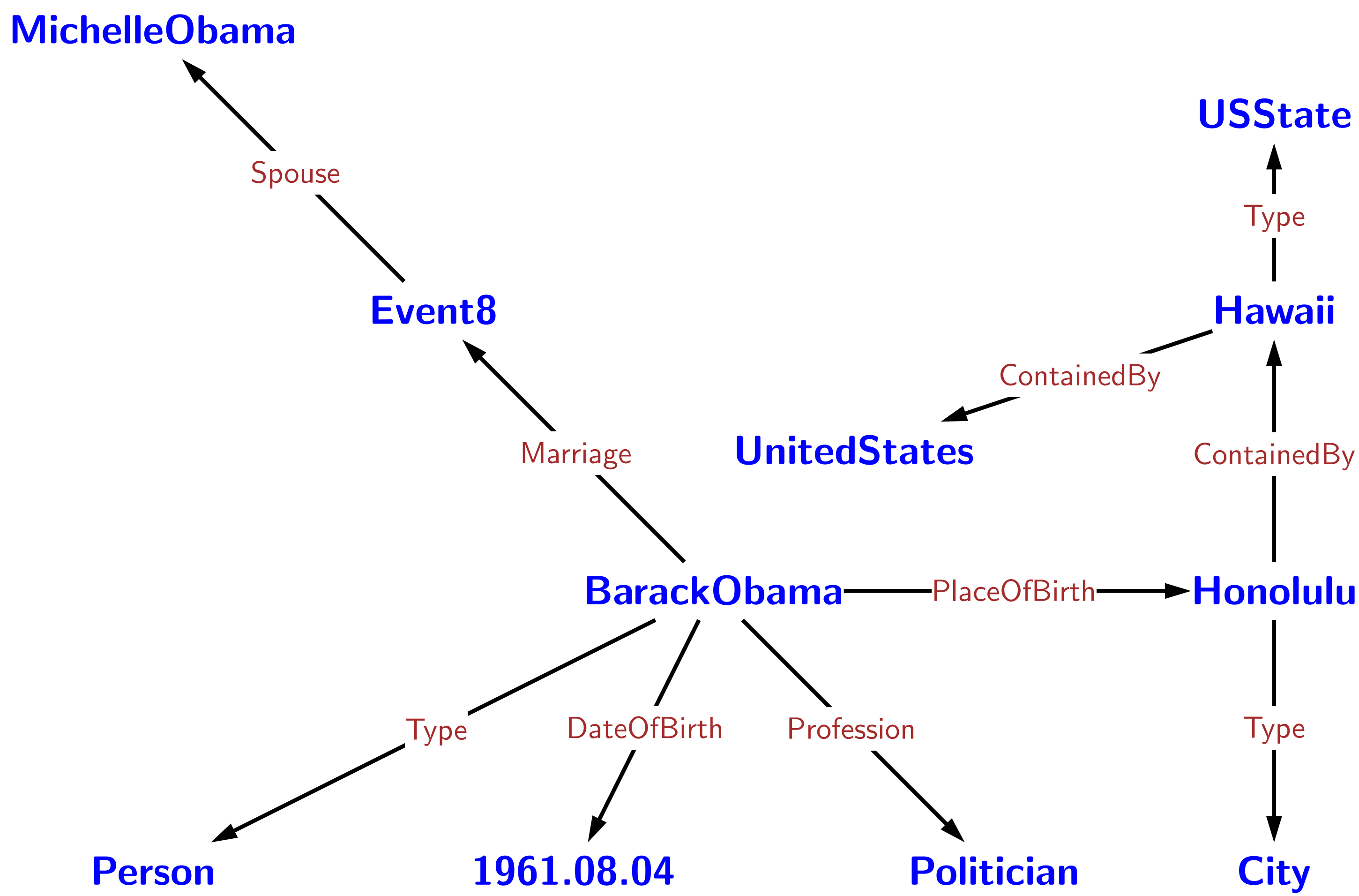
100M entities (nodes)   1B assertions (edges)



[Bollacker, 2008; Google, 2013]

World: Freebase

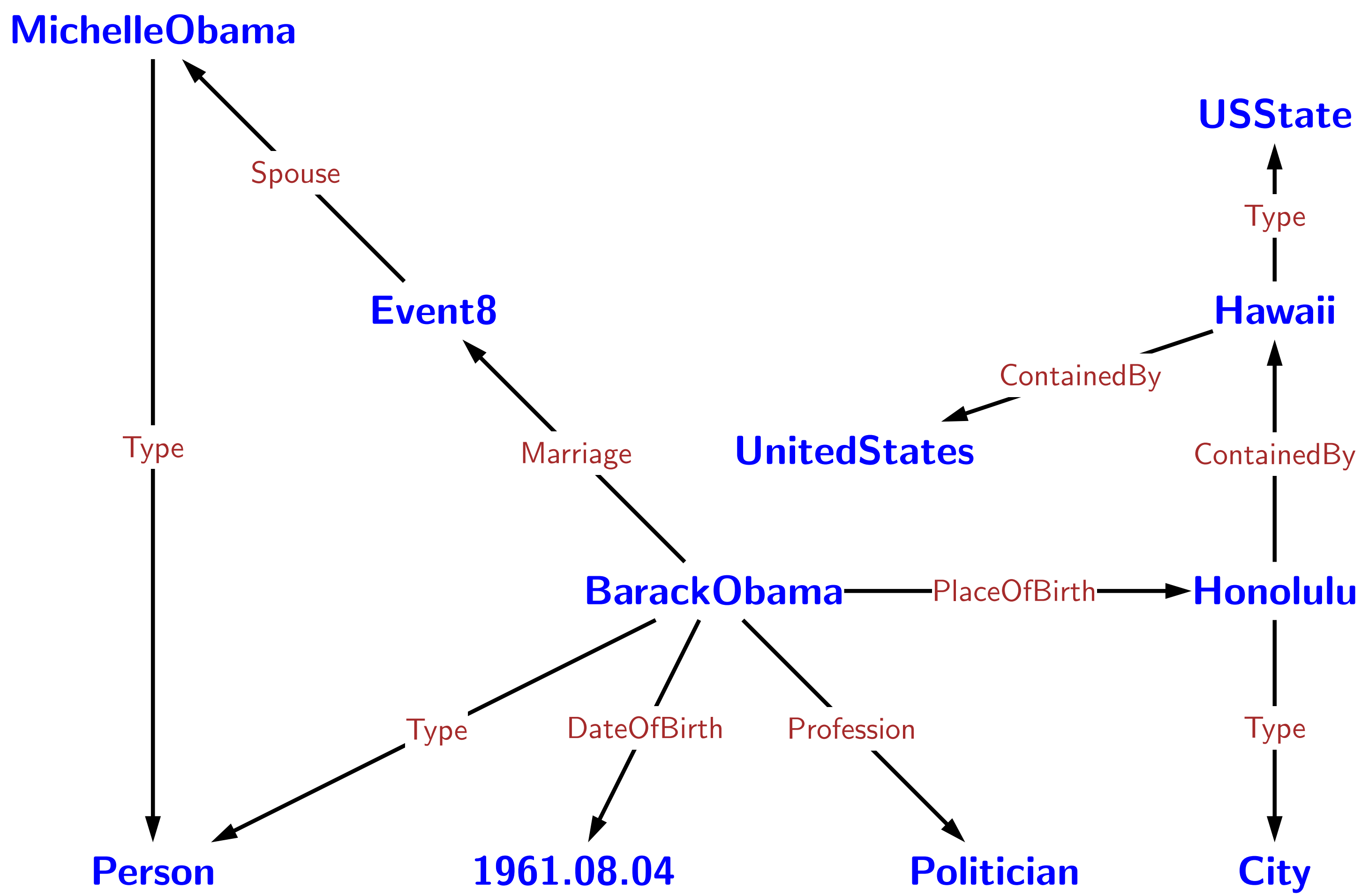
100M entities (nodes) 1B assertions (edges)



[Bollacker, 2008; Google, 2013]

World: Freebase

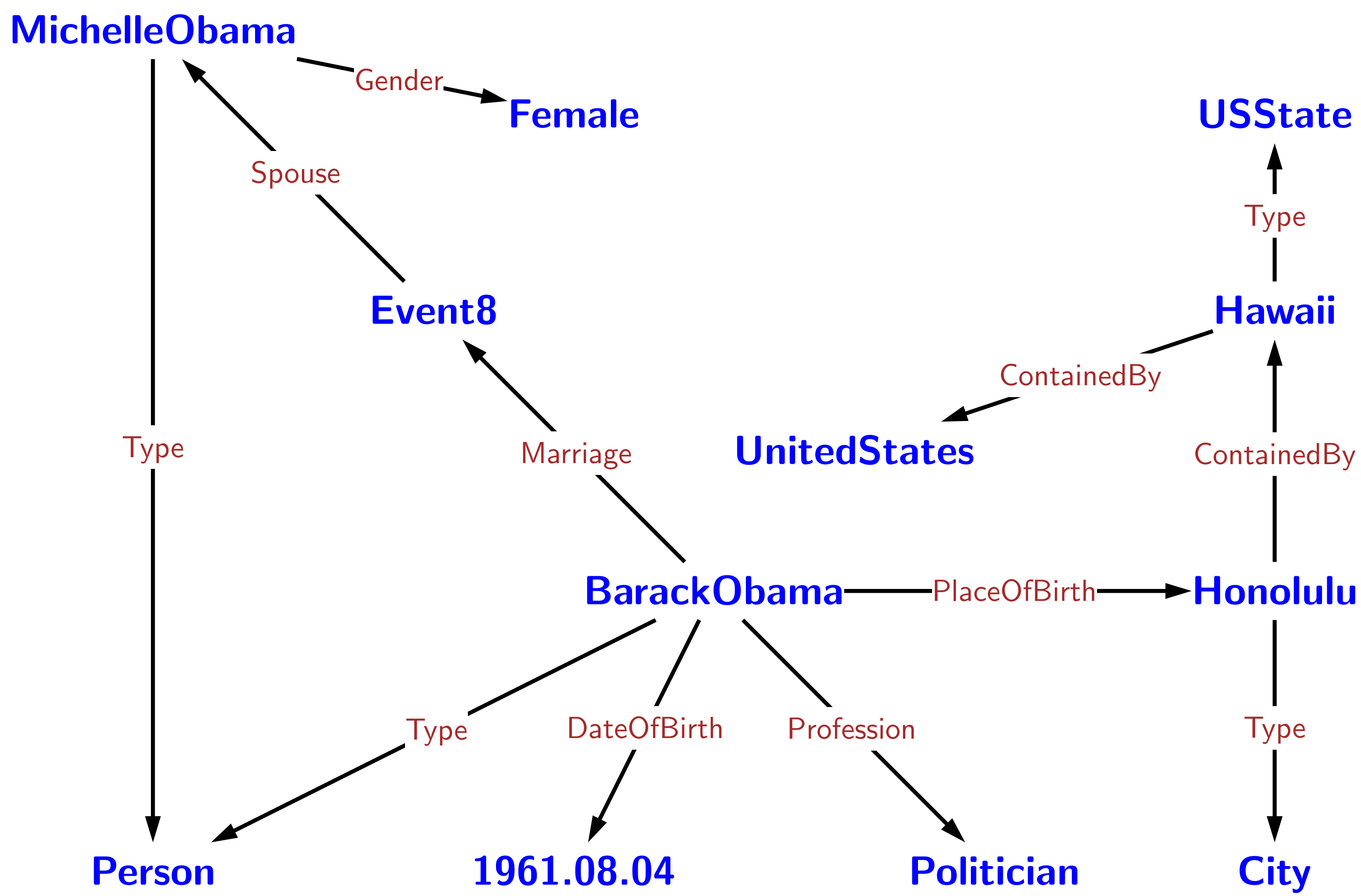
100M entities (nodes)   1B assertions (edges)



[Bollacker, 2008; Google, 2013]

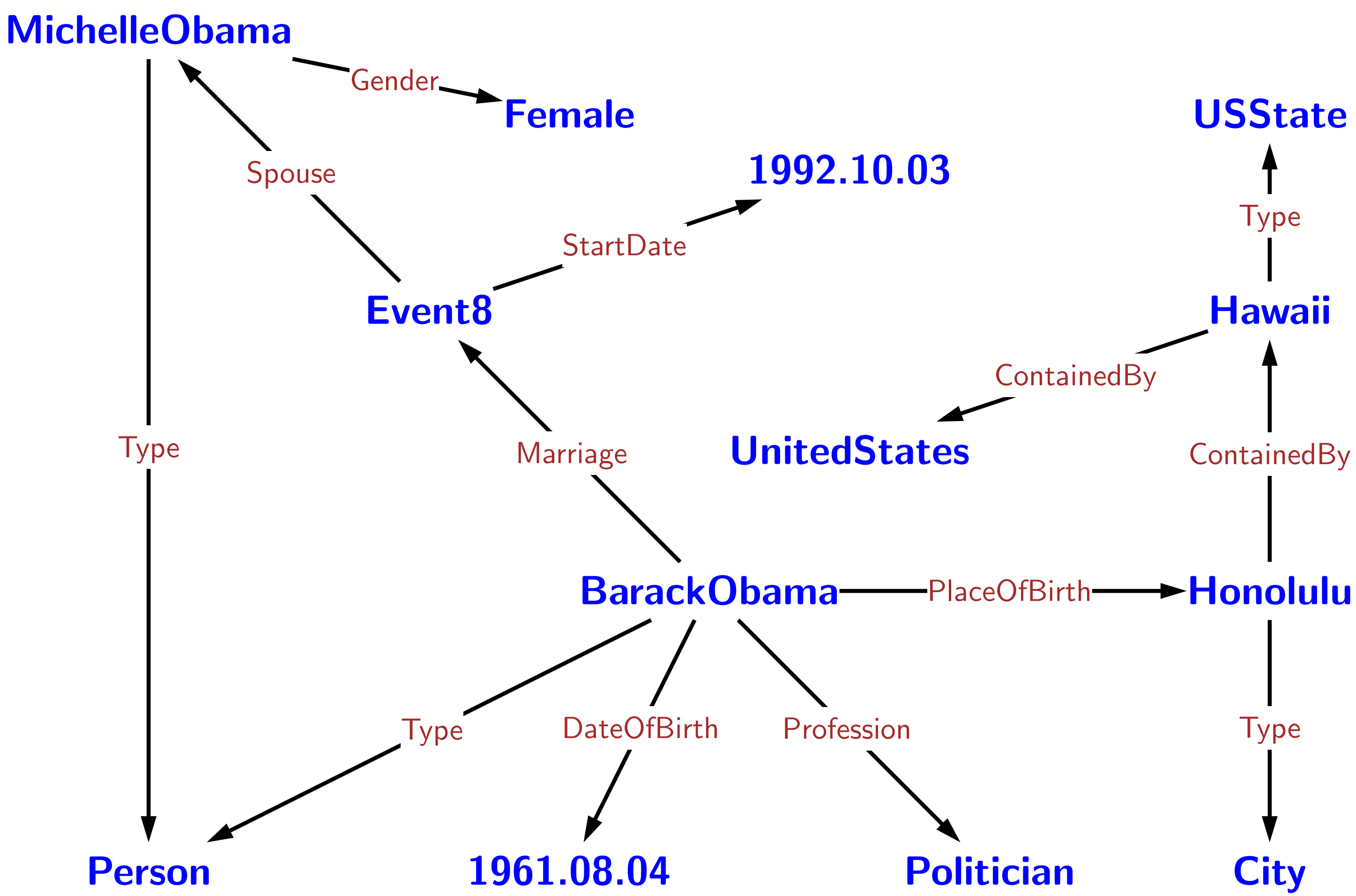
World: Freebase

100M entities (nodes)   1B assertions (edges)



World: Freebase

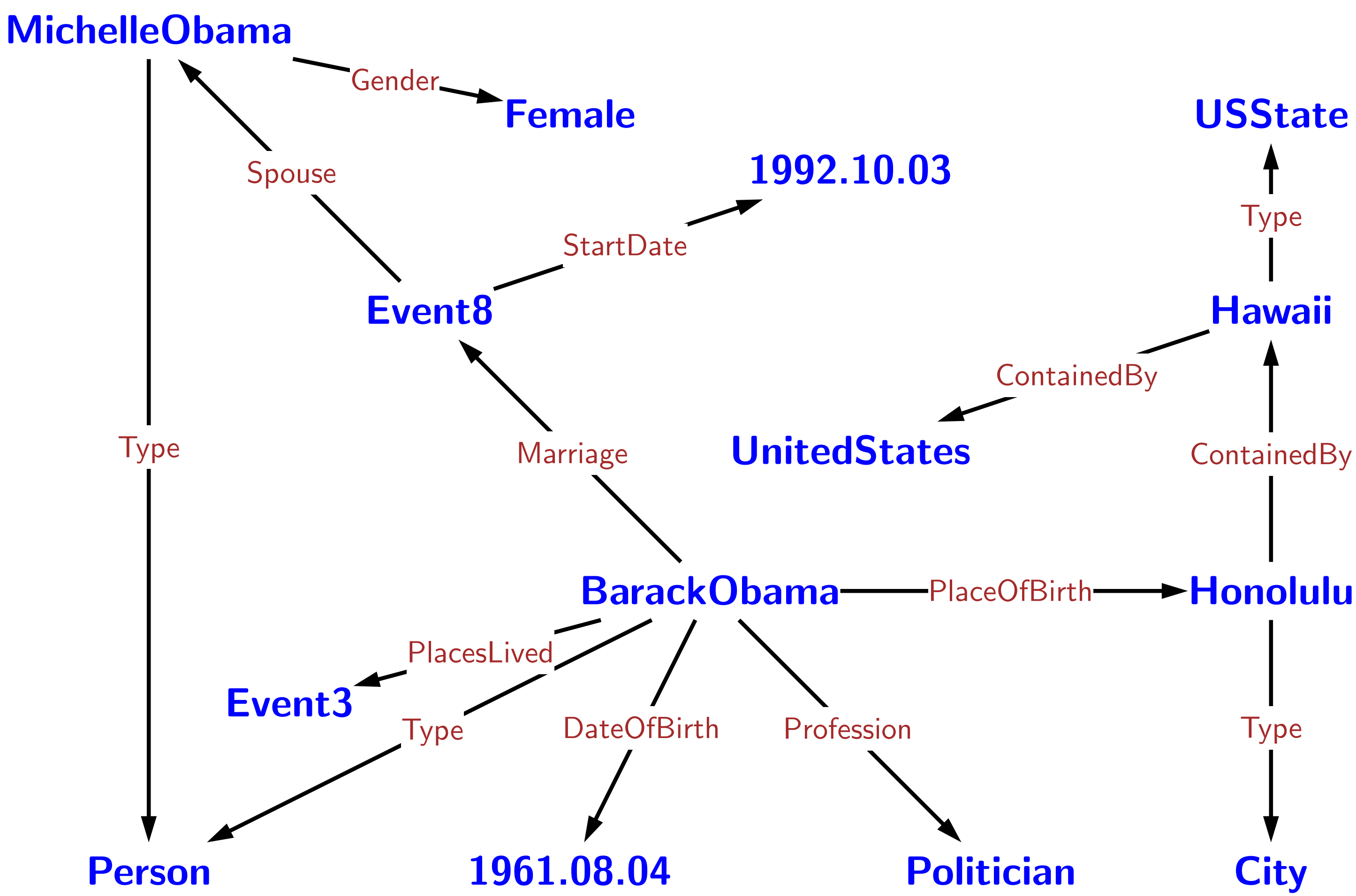
100M entities (nodes) 1B assertions (edges)



[Bollacker, 2008; Google, 2013]

World: Freebase

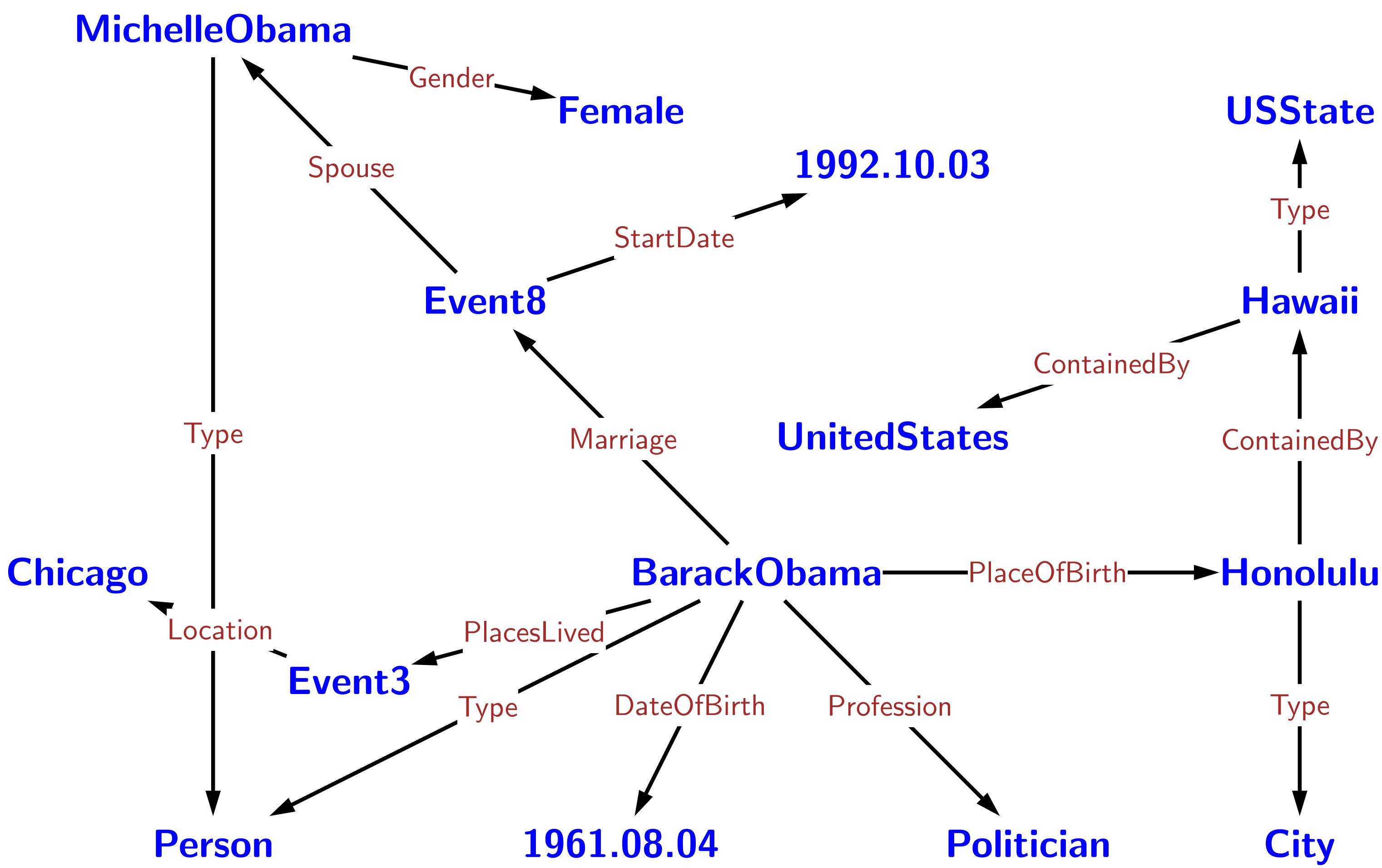
100M entities (nodes) 1B assertions (edges)



[Bollacker, 2008; Google, 2013]

World: Freebase

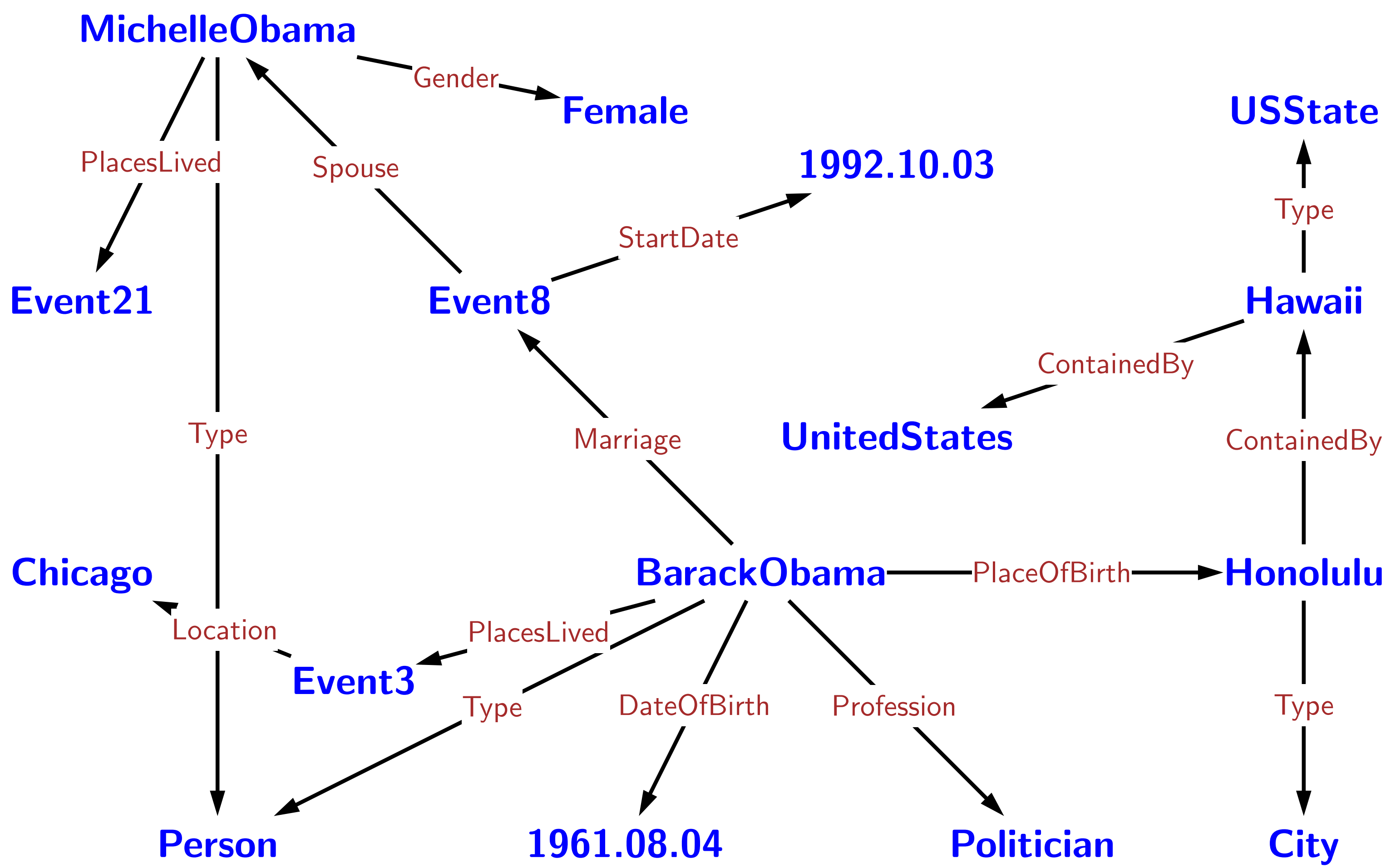
100M entities (nodes) 1B assertions (edges)



[Bollacker, 2008; Google, 2013]

World: Freebase

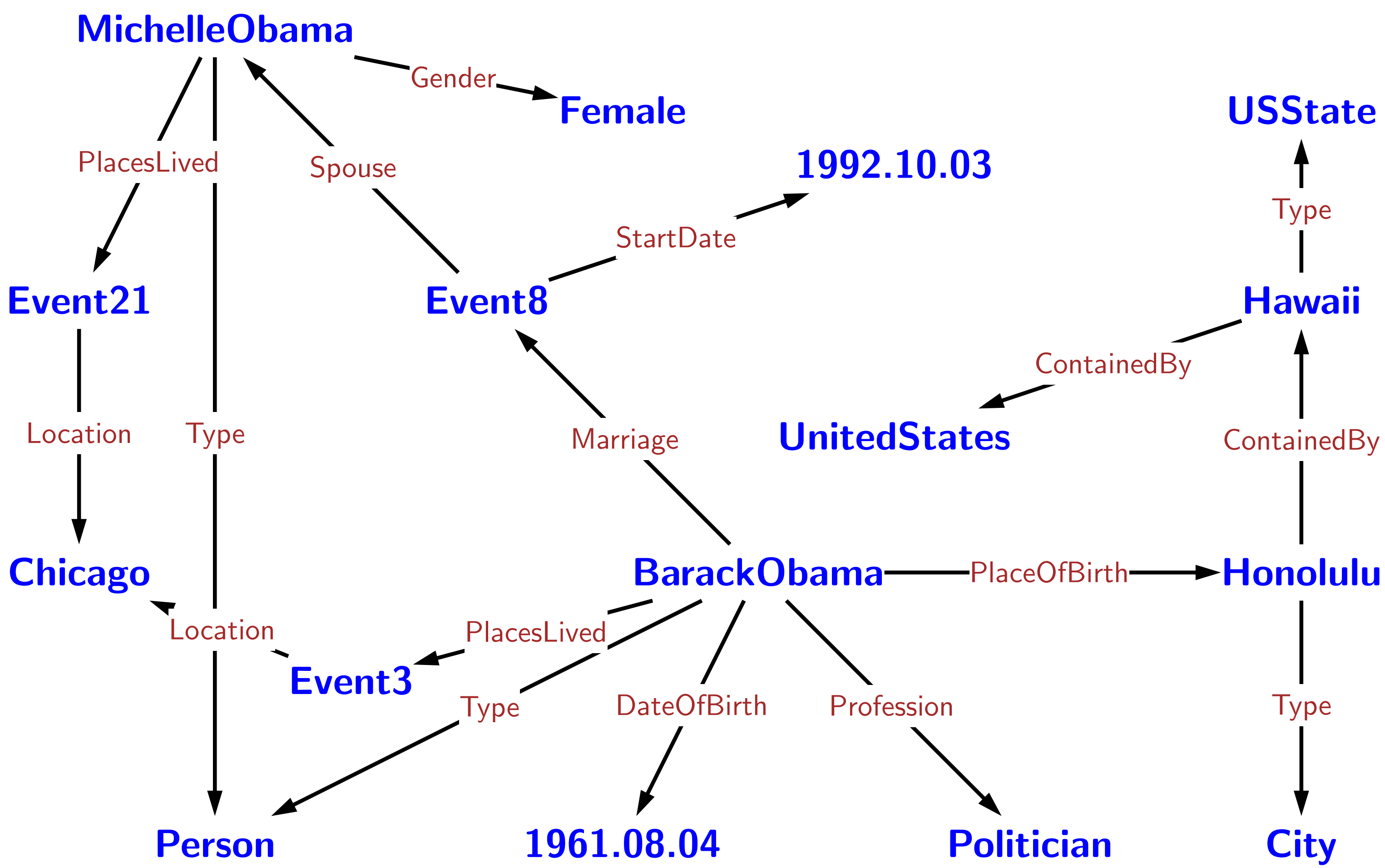
100M entities (nodes) 1B assertions (edges)



[Bollacker, 2008; Google, 2013]

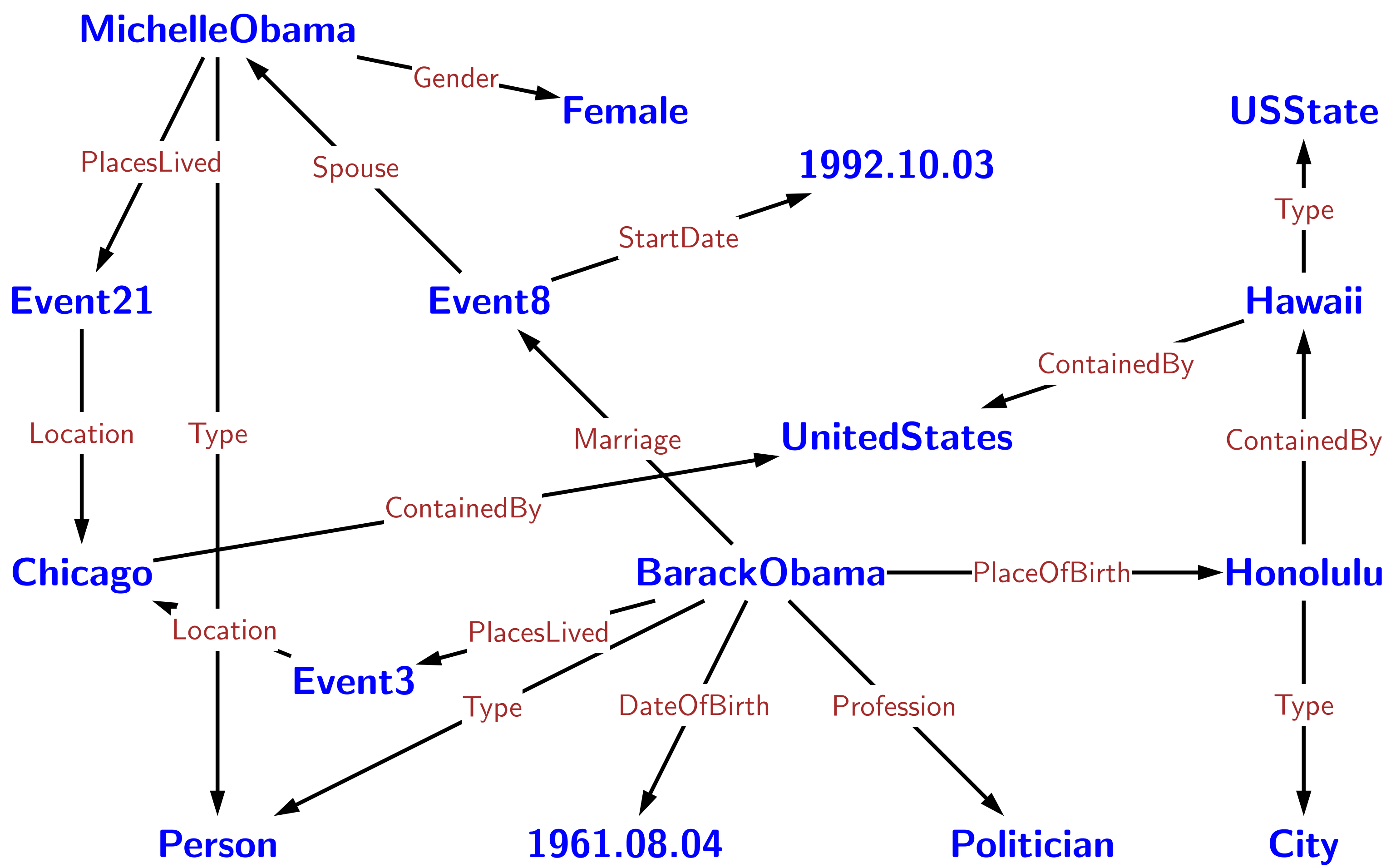
World: Freebase

100M entities (nodes) 1B assertions (edges)



World: Freebase

100M entities (nodes) 1B assertions (edges)

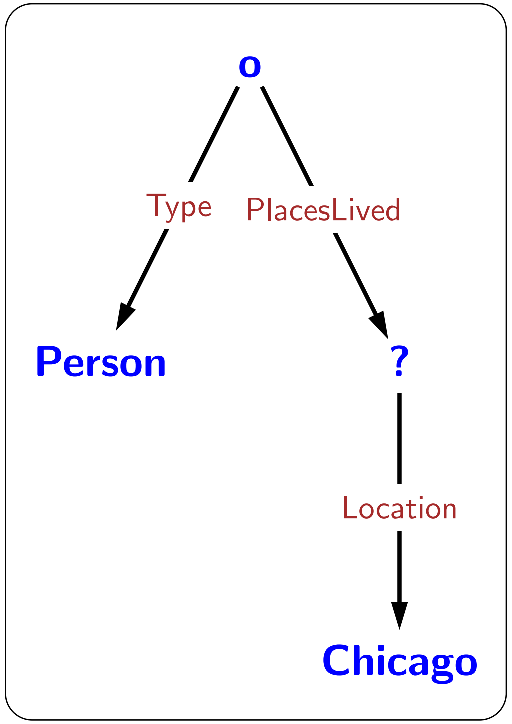


Logical forms: lambda DCS

Type.Person $\sqcap$ PlacesLived.Location.Chicago

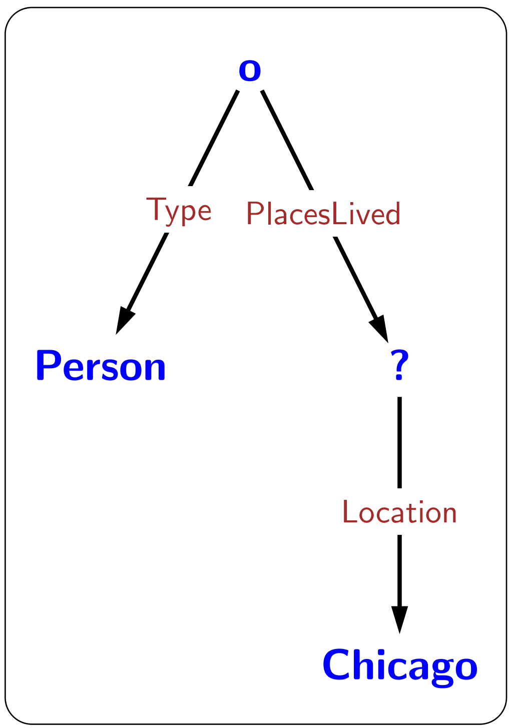
Logical forms: lambda DCS

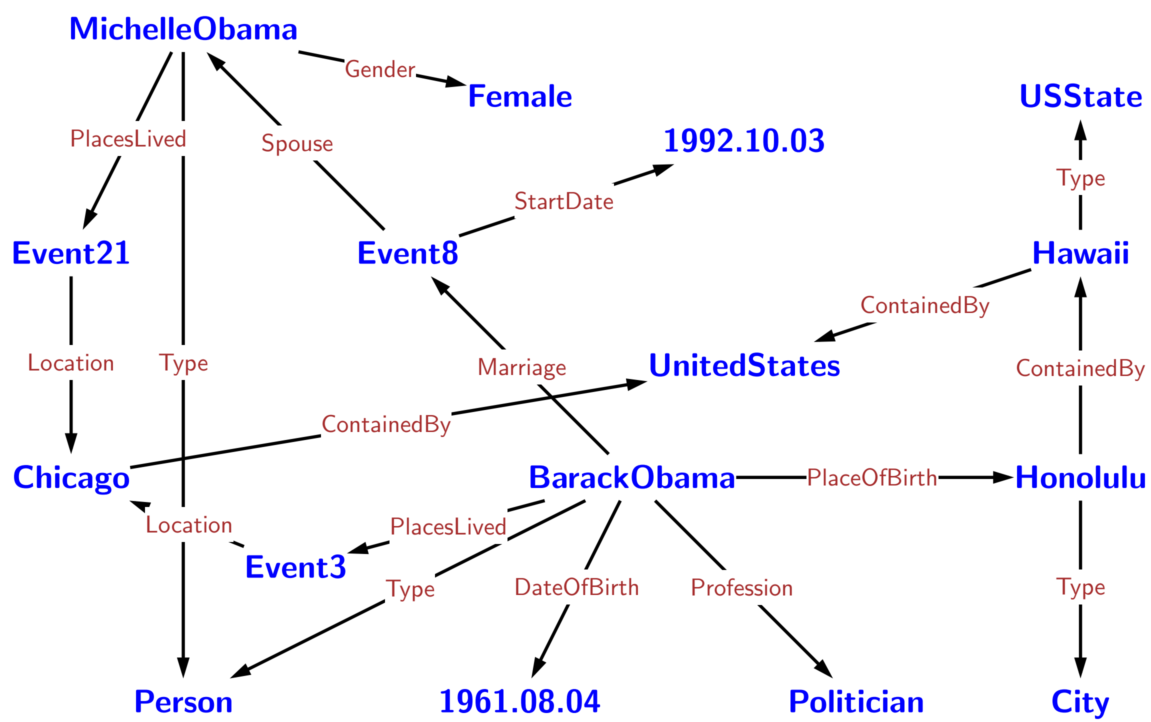
Type.Person $\sqcap$ PlacesLived.Location.Chicago



Logical forms: lambda DCS

$\text{Type.Person} \sqcap \text{PlacesLived.Location.Chicago}$

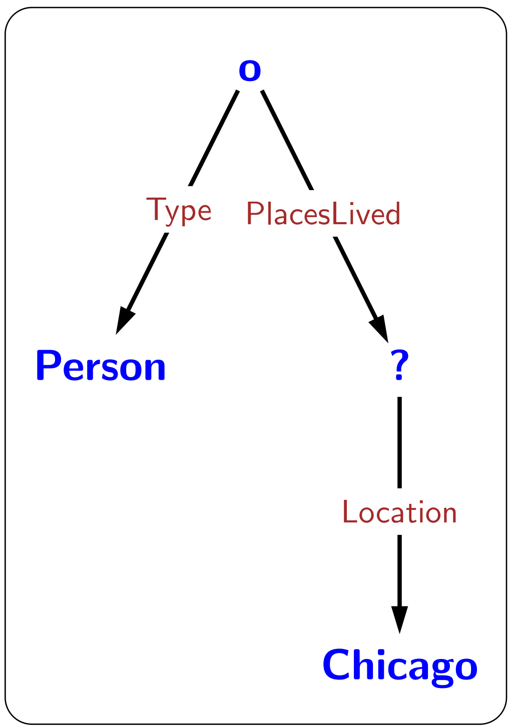


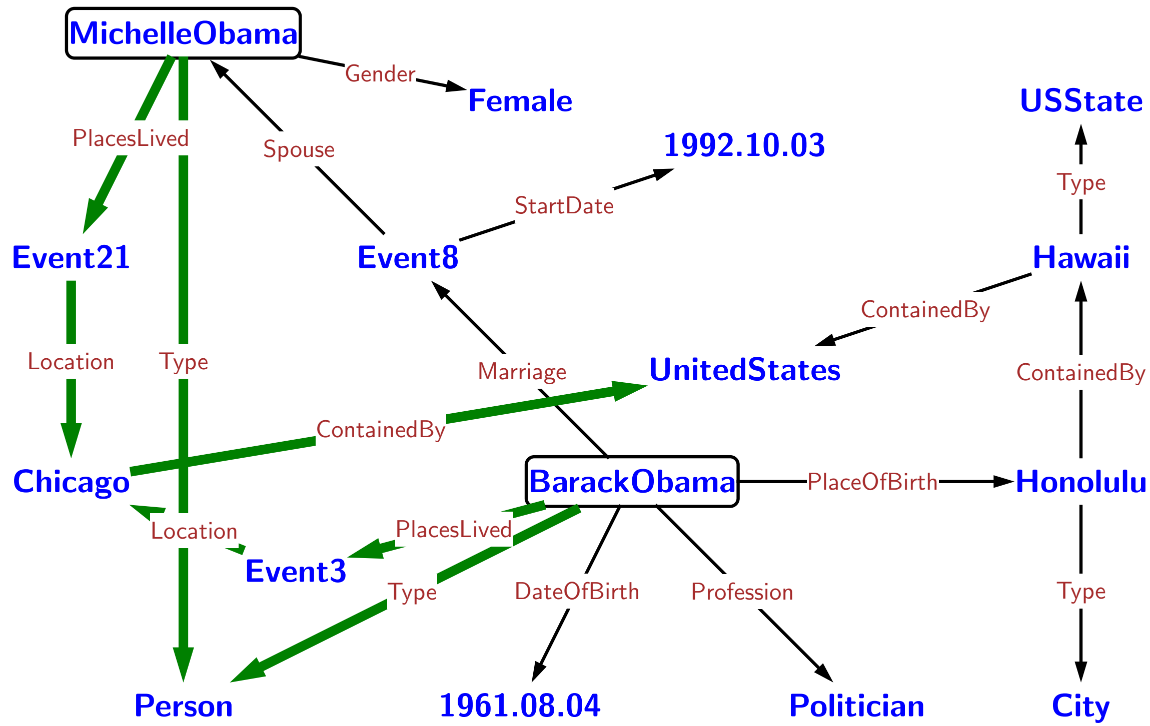


[Liang, 2013]

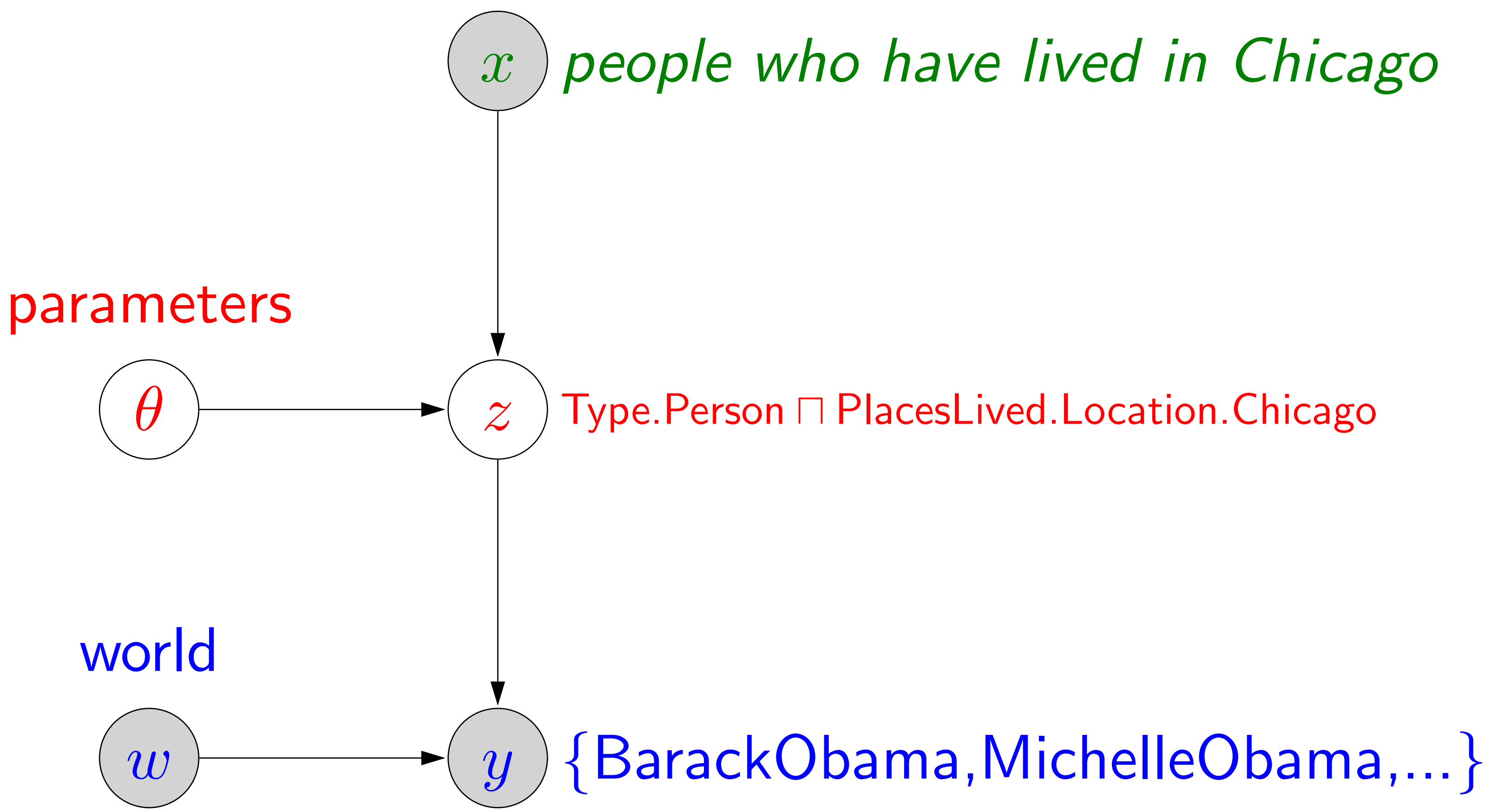
Logical forms: lambda DCS

Type.Person ⊓ PlacesLived.Location.Chicago



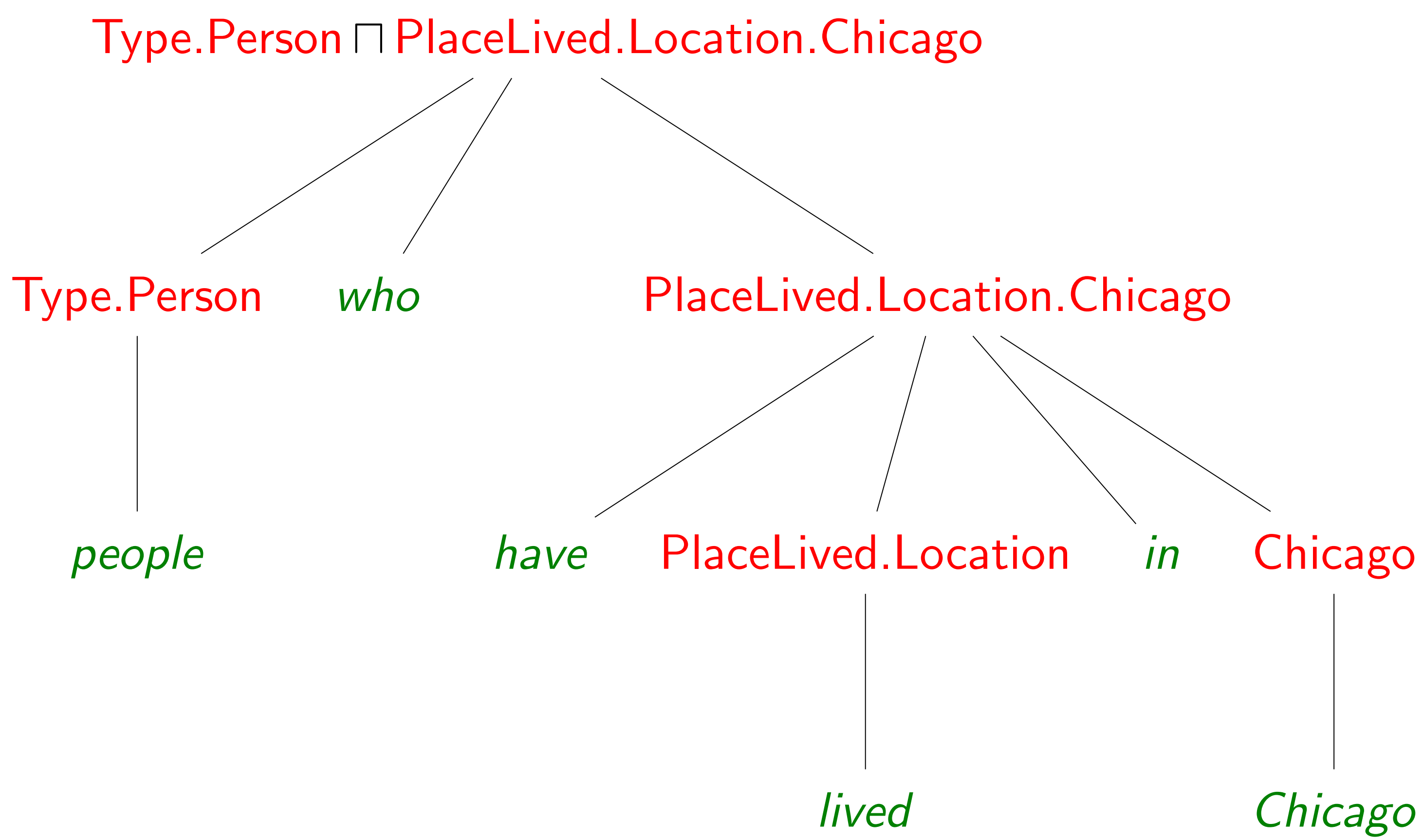


Probabilistic framework



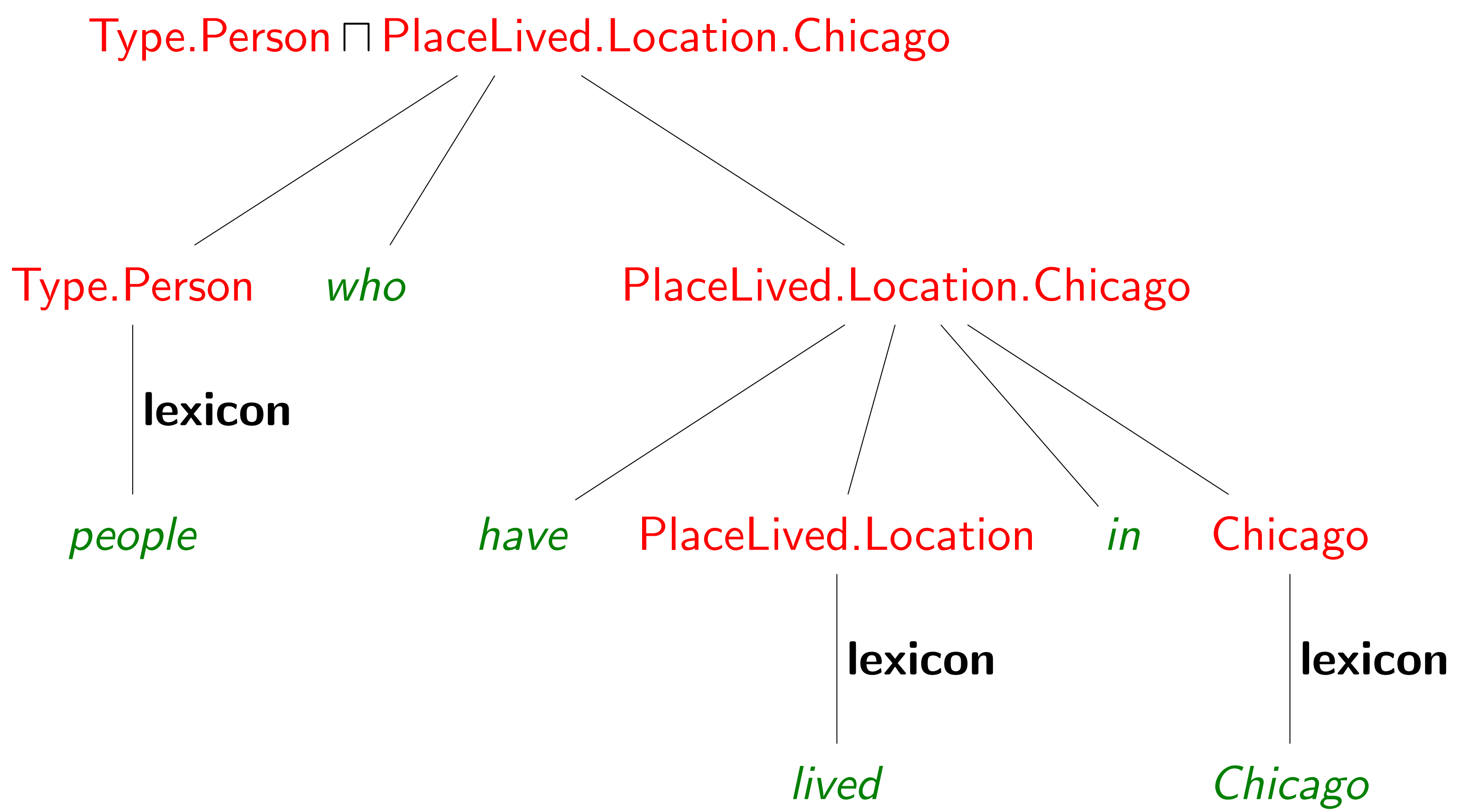
Derivations

Derivation: construction of logical form given utterance



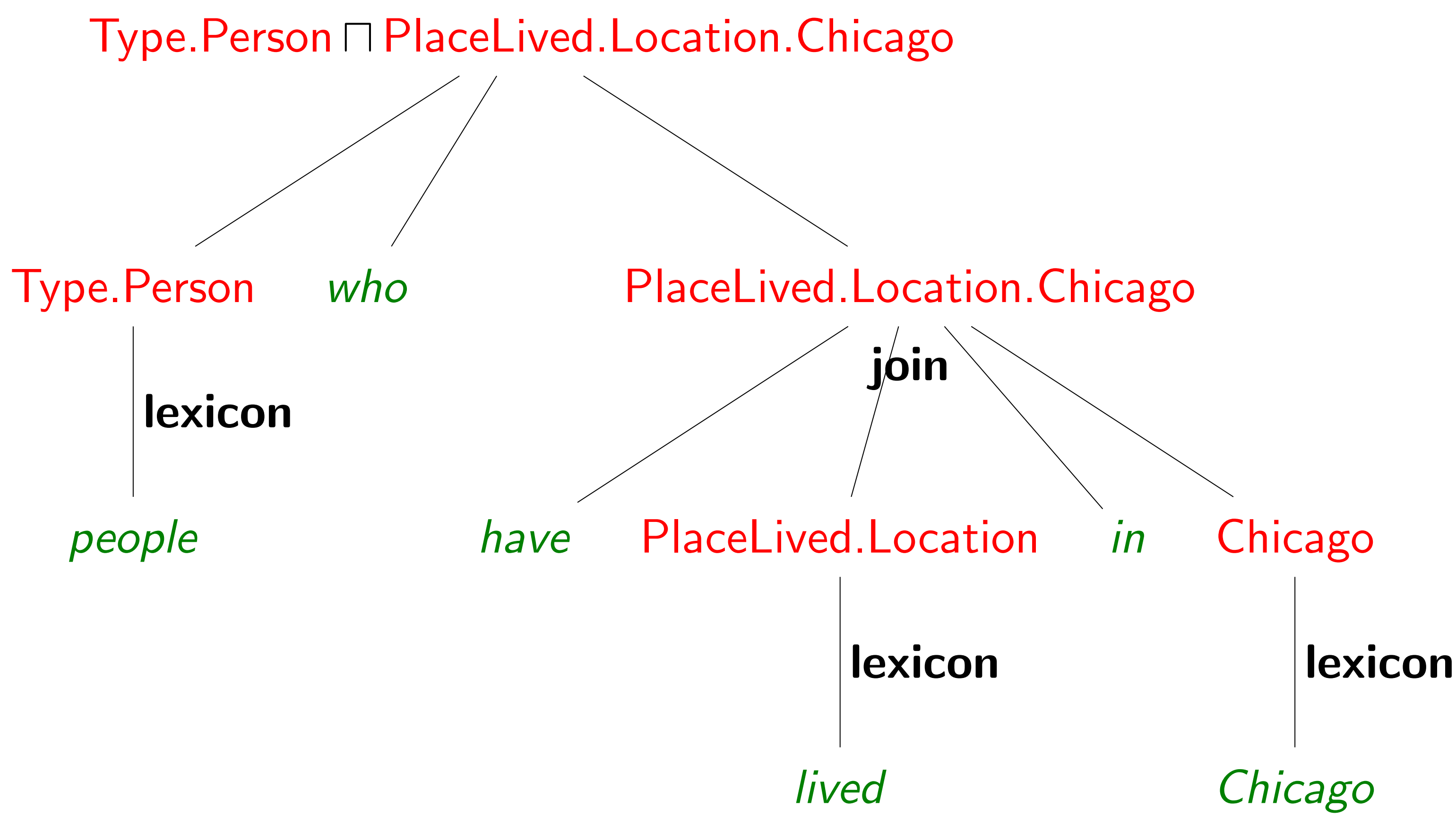
Derivations

Derivation: construction of logical form given utterance



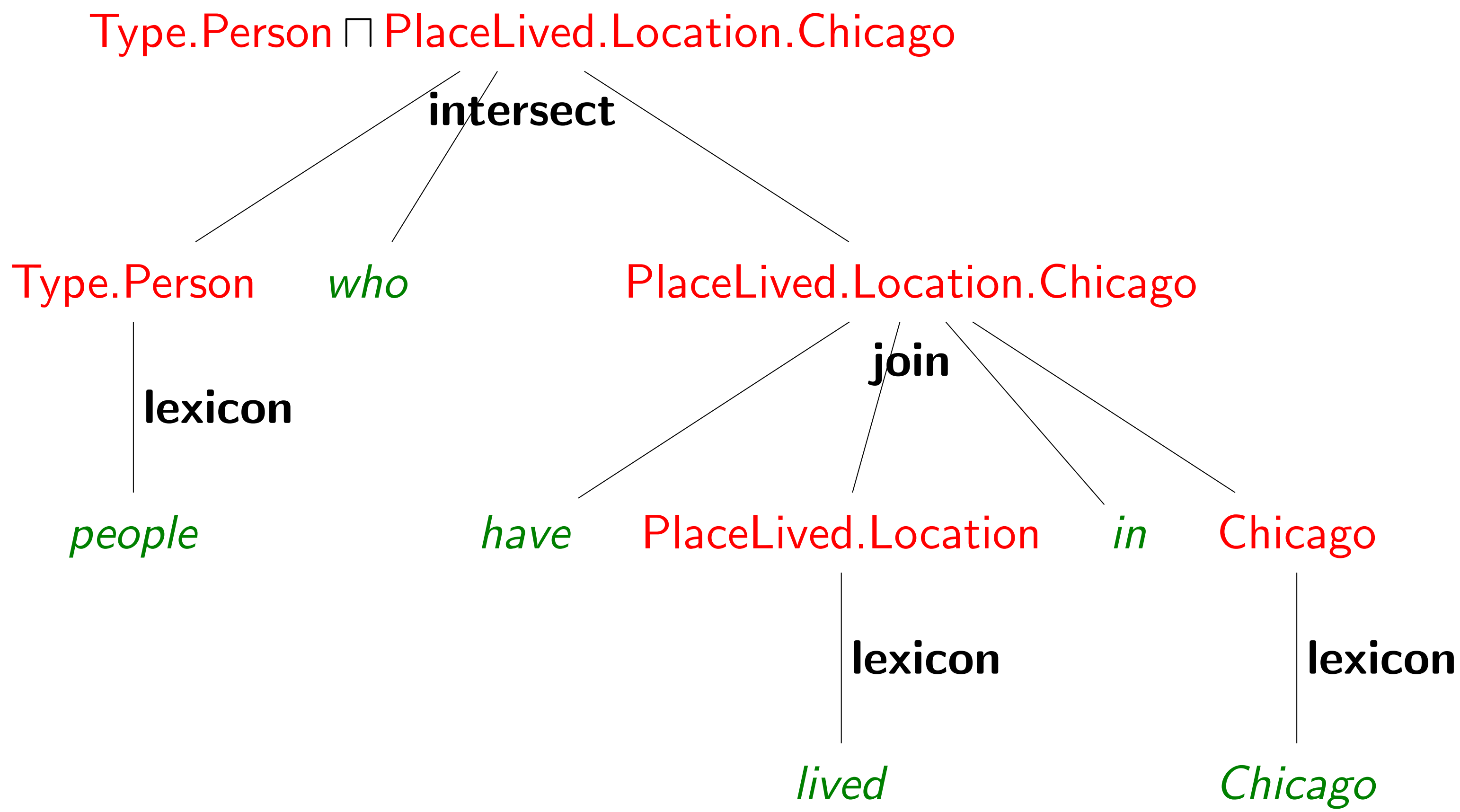
Derivations

Derivation: construction of logical form given utterance



Derivations

Derivation: construction of logical form given utterance



(Over)-generating derivations

utterance → Grammar → *derivation 1*
derivation 2
...

(Over)-generating derivations

utterance → Grammar → *derivation 1*
derivation 2
...

A Real Dumb Grammar

(lexicon) *Chicago* $\Rightarrow$ E : Chicago

(lexicon) *people* $\Rightarrow$ E : Type.Person

(lexicon) *live* $\Rightarrow$ E $\times$ E : PlacesLived

...

(join) E $\times$ E : b E : u $\Rightarrow$ E : $b.u$

(intersect) E : u E : v $\Rightarrow$ E : $u \sqcap v$

Many possible derivations!

people who have lived in Chicago

Many possible derivations!

people who have lived in Chicago

$\Downarrow$?

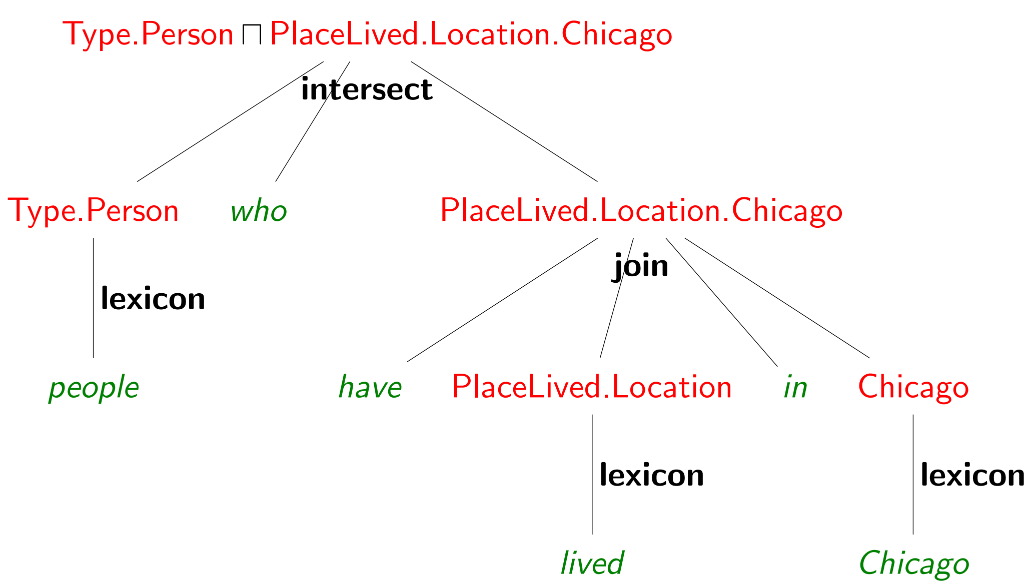
set of candidate derivations $\mathcal{D}(x)$

Many possible derivations!

people who have lived in Chicago

?

set of candidate derivations $\mathcal{D}(x)$

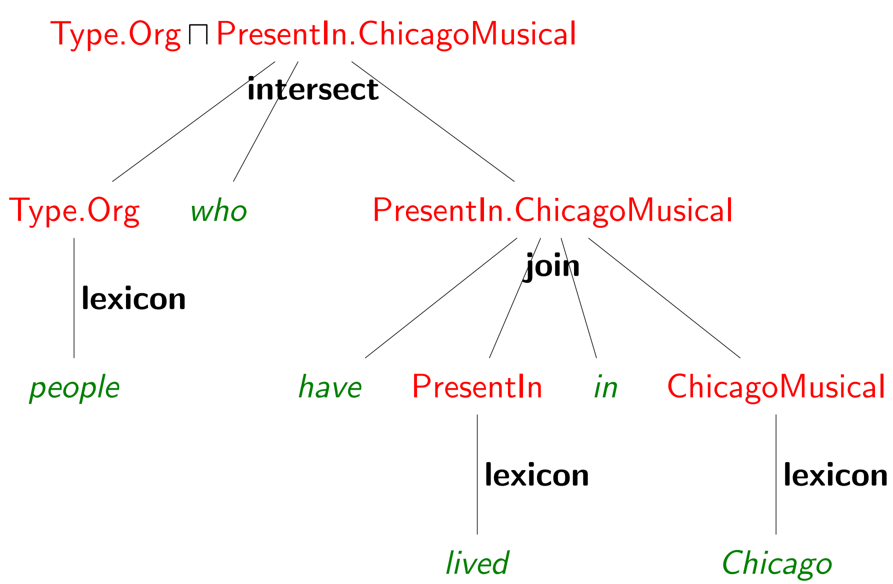


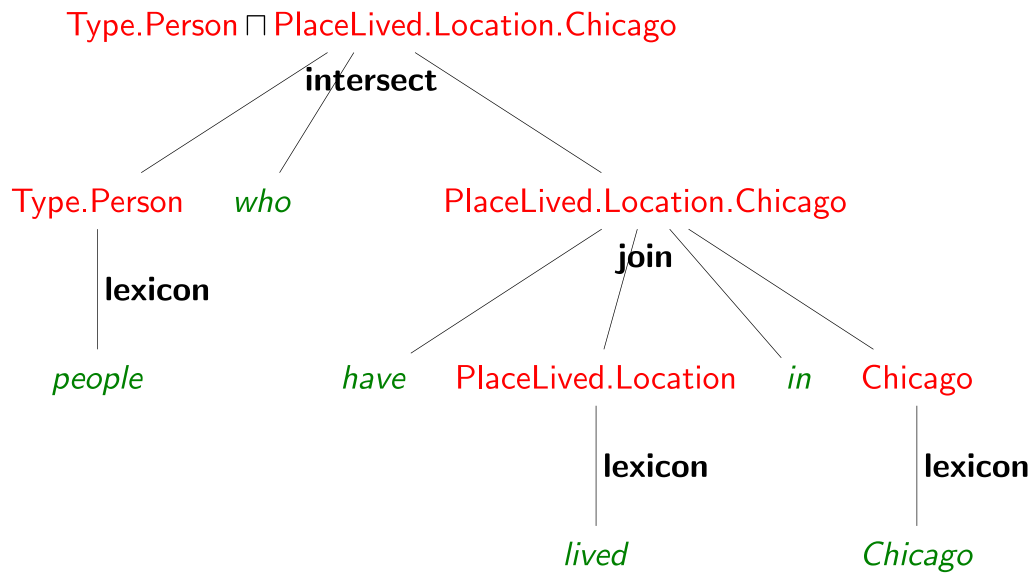
Many possible derivations!

people who have lived in Chicago

?

set of candidate derivations $\mathcal{D}(x)$





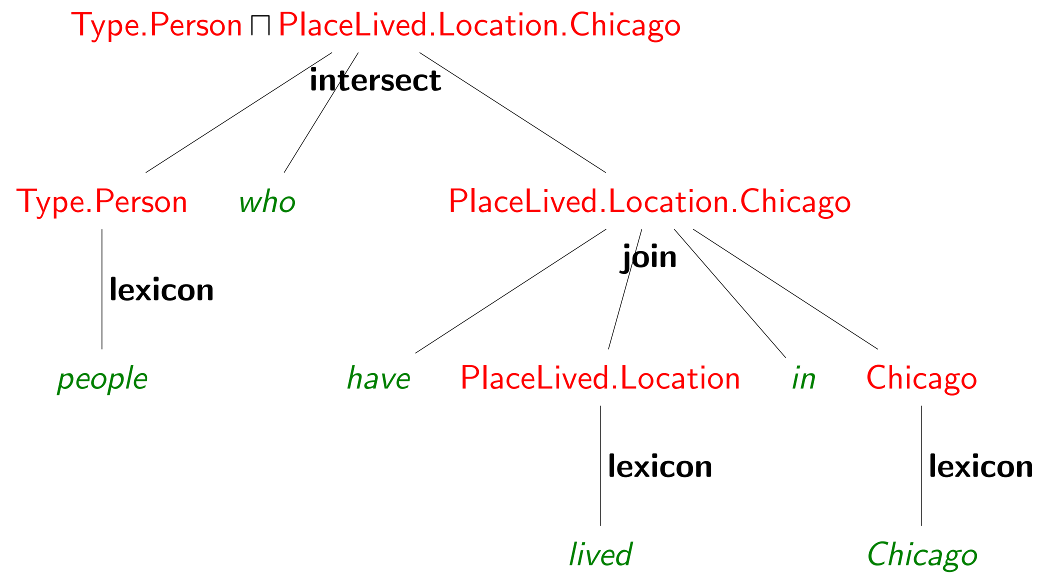
x: utterance

d: derivation

Feature vector $\phi(x, d) \in \mathbb{R}^F$:

x: utterance

d: derivation



Feature vector $\phi(x, d) \in \mathbb{R}^F$:

apply join	1
apply intersect	1
apply lexicon	3
skipped IN	1
skipped NN	0
lived maps to PlacesLived.Location	1
lived maps to PlaceOfBirth	0
alignmentScore	1.52
denotation-size=1	1
...	...

Scoring derivations

Feature vector: $\phi(x, d) = [1.3, 2, 0, 1, 0, 0, \dots] \in \mathbb{R}^F$

Scoring derivations

Feature vector: $\phi(x, d) = [1.3, 2, 0, 1, 0, 0, \dots] \in \mathbb{R}^F$

Parameter vector: $\theta = [1.2, -2.7, 3.4, \dots] \in \mathbb{R}^F$

Scoring derivations

Feature vector: $\phi(x, d) = [1.3, 2, 0, 1, 0, 0, \dots] \in \mathbb{R}^F$

Parameter vector: $\theta = [1.2, -2.7, 3.4, \dots] \in \mathbb{R}^F$

Scoring function:

$$\mathsf{Score}_\theta(x, d) = \phi(x, d) \cdot \theta = \sum_{j=1}^{F} \theta_j \phi_j(x, d)$$

Log-linear model

Candidate derivations: $\mathcal{D}(x)$

Log-linear model

Candidate derivations: $\mathcal{D}(x)$

Model: distribution over derivations d given utterance x

$$p(d \mid x, \theta) = \frac{\exp(\mathsf{Score}_\theta(x, d))}{\sum_{d' \in \mathcal{D}(x)} \exp(\mathsf{Score}_\theta(x, d'))}$$

Learning

Training data:

What's Bulgaria's capital?

`Sofia`

What movies has Tom Cruise been in?

`TopGun,VanillaSky,...`

...

+grammar, +features

Learning

Training data:

What's Bulgaria's capital?
`Sofia`
What movies has Tom Cruise been in?
`TopGun,VanillaSky,...`
...

+grammar, +features

Objective: Maximum likelihood

$$\arg\max_\theta \sum_{i=1}^{n} \log p_\theta(y^{(i)} \mid x^{(i)})$$

Learning

Training data:

What's Bulgaria's capital?
`Sofia`
What movies has Tom Cruise been in?
`TopGun,VanillaSky,...`
...

+grammar, +features

Objective: Maximum likelihood

$$\arg\max_\theta \sum_{i=1}^n \log p_\theta(y^{(i)} \mid x^{(i)})$$

Algorithm:

AdaGrad (stochastic gradient with per-feature step size)

Training intuition

Where did Mozart tupress?

Vienna

Training intuition

Where did Mozart tupress?

`PlaceOfBirth.Mozart`

`PlaceOfDeath.Mozart`

`PlaceOfMarriage.Mozart`

Vienna

Training intuition

Where did Mozart tupress?

`PlaceOfBirth.Mozart`	$\Rightarrow$ Salzburg
`PlaceOfDeath.Mozart`	$\Rightarrow$ Vienna
`PlaceOfMarriage.Mozart`	$\Rightarrow$ Vienna

Vienna

Training intuition

Where did Mozart tupress?

$\cancel{\texttt{PlaceOfBirth.Mozart} \Rightarrow \text{Salzburg}}$

$\texttt{PlaceOfDeath.Mozart} \Rightarrow \text{Vienna}$

$\texttt{PlaceOfMarriage.Mozart} \Rightarrow \text{Vienna}$

Vienna

Training intuition

Where did Mozart tupress?

~~`PlaceOfBirth.Mozart` $\Rightarrow$ Salzburg~~

`PlaceOfDeath.Mozart` $\Rightarrow$ Vienna

`PlaceOfMarriage.Mozart` $\Rightarrow$ Vienna

Vienna

Where did William Hogarth tupress?

Training intuition

Where did Mozart tupress?

~~`PlaceOfBirth.Mozart` ⇒ Salzburg~~

`PlaceOfDeath.Mozart` ⇒ Vienna

`PlaceOfMarriage.Mozart` ⇒ Vienna

Vienna

Where did William Hogarth tupress?

`PlaceOfBirth.WilliamHogarth`

`PlaceOfDeath.WilliamHogarth`

`PlaceOfMarriage.WilliamHogarth`

London

Training intuition

Where did Mozart tupress?

$\cancel{\texttt{PlaceOfBirth.Mozart} \Rightarrow \text{Salzburg}}$

$\texttt{PlaceOfDeath.Mozart} \Rightarrow \text{Vienna}$

$\texttt{PlaceOfMarriage.Mozart} \Rightarrow \text{Vienna}$

Vienna

Where did William Hogarth tupress?

$\texttt{PlaceOfBirth.WilliamHogarth} \Rightarrow \text{London}$

$\texttt{PlaceOfDeath.WilliamHogarth} \Rightarrow \text{London}$

$\texttt{PlaceOfMarriage.WilliamHogarth} \Rightarrow \text{Paddington}$

London

Training intuition

Where did Mozart tupress?

~~`PlaceOfBirth.Mozart`~~	~~$\Rightarrow$~~	~~Salzburg~~
`PlaceOfDeath.Mozart`	$\Rightarrow$	Vienna
`PlaceOfMarriage.Mozart`	$\Rightarrow$	Vienna

Vienna

Where did William Hogarth tupress?

`PlaceOfBirth.WilliamHogarth`	$\Rightarrow$	London
`PlaceOfDeath.WilliamHogarth`	$\Rightarrow$	London
~~`PlaceOfMarriage.WilliamHogarth`~~	~~$\Rightarrow$~~	~~Paddington~~

London

Training intuition

Where did Mozart tupress?

~~`PlaceOfBirth.Mozart` ⇒ Salzburg~~

$\boxed{\texttt{PlaceOfDeath.Mozart} \Rightarrow \text{Vienna}}$

`PlaceOfMarriage.Mozart` ⇒ Vienna

Vienna

Where did William Hogarth tupress?

`PlaceOfBirth.WilliamHogarth` ⇒ London

$\boxed{\texttt{PlaceOfDeath.WilliamHogarth} \Rightarrow \text{London}}$

~~`PlaceOfMarriage.WilliamHogarth` ⇒ Paddington~~

London

Outline

- A semantic parsing framework
- **A closer look at the elements**
  - Logical forms: lambda DCS
  - Lexical coverage
  - Grammar: building logical forms
  - Learning via bootstrapping
  - Leveraging denotations
  - Datasets/results
- Beyond Freebase
- Final remarks

Outline

- A semantic parsing framework
- A closer look at the elements
  - **Logical forms: lambda DCS**
  - Lexical coverage
  - Grammar: building logical forms
  - Learning via bootstrapping
  - Leveraging denotations
  - Datasets/results
- Beyond Freebase
- Final remarks

Lambda DCS

Entity

Chicago

Lambda DCS

Entity

Chicago

Join

PlaceOfBirth.Chicago

Lambda DCS

Entity

Chicago

Join

PlaceOfBirth.Chicago

Intersect

Type.Person $\sqcap$ PlaceOfBirth.Chicago

Lambda DCS

Entity

Chicago

Join

PlaceOfBirth.Chicago

Intersect

Type.Person$\sqcap$PlaceOfBirth.Chicago

Aggregation

count(Type.Person $\sqcap$ PlaceOfBirth.Chicago)

Lambda DCS

Entity

Chicago

Join

PlaceOfBirth.Chicago

Intersect

Type.Person $\sqcap$ PlaceOfBirth.Chicago

Aggregation

count(Type.Person $\sqcap$ PlaceOfBirth.Chicago)

Superlative

argmin(Type.Person $\sqcap$ PlaceOfBirth.Chicago, DateOfBirth)

Lambda DCS

Entity

Chicago

Join

PlaceOfBirth.Chicago

Intersect

Type.Person$\sqcap$PlaceOfBirth.Chicago

Aggregation

count(Type.Person $\sqcap$ PlaceOfBirth.Chicago)

Superlative

argmin(Type.Person $\sqcap$ PlaceOfBirth.Chicago, DateOfBirth)

Anaphora

$\mu x.$Type.Person $\sqcap$ Children.Influence.x

Lambda DCS

Entity

Chicago

Join

PlaceOfBirth.Chicago

Intersect

Type.Person$\sqcap$PlaceOfBirth.Chicago

Aggregation

count(Type.Person $\sqcap$ PlaceOfBirth.Chicago)

Superlative

argmin(Type.Person $\sqcap$ PlaceOfBirth.Chicago, DateOfBirth)

Anaphora

μx.Type.Person $\sqcap$ Children.Influence.x

Variable

λx.Gender.Female $\sqcap$ Children.Children.x

Comparison to lambda calculus

Lambda calculus

$$\lambda x.\mathsf{Type}(x, \mathsf{Person}) \wedge \exists e.\mathsf{PlacesLived}(x, e) \wedge \mathsf{Location}(e, \mathsf{Chicago})$$

Comparison to lambda calculus

Lambda calculus

$$\lambda x.\mathsf{Type}(x, \mathsf{Person}) \wedge \exists e.\mathsf{PlacesLived}(x, e) \wedge \mathsf{Location}(e, \mathsf{Chicago})$$

Lambda dependency-based compositional semantics (DCS)

$$\mathsf{Type.Person} \sqcap \mathsf{PlacesLived.Location.Chicago}$$

Comparison to lambda calculus

Lambda calculus

$$\lambda x.\mathsf{Type}(x, \mathsf{Person}) \wedge \exists e.\mathsf{PlacesLived}(x, e) \wedge \mathsf{Location}(e, \mathsf{Chicago})$$

Lambda dependency-based compositional semantics (DCS)

$$\mathsf{Type.Person} \sqcap \mathsf{PlacesLived.Location.Chicago}$$

- Eliminate variables to simplify

Comparison to lambda calculus

Lambda calculus

$$\lambda x.\mathsf{Type}(x, \mathsf{Person}) \wedge \exists e.\mathsf{PlacesLived}(x, e) \wedge \mathsf{Location}(e, \mathsf{Chicago})$$

Lambda dependency-based compositional semantics (DCS)

$$\mathsf{Type.Person} \sqcap \mathsf{PlacesLived.Location.Chicago}$$

- Eliminate variables to simplify
- Existential quantification by default

Comparison to lambda calculus

Lambda calculus

$$\lambda x.\text{Type}(x, \text{Person}) \wedge \exists e.\text{PlacesLived}(x, e) \wedge \text{Location}(e, \text{Chicago})$$

Lambda dependency-based compositional semantics (DCS)

$$\text{Type.Person} \sqcap \text{PlacesLived.Location.Chicago}$$

- Eliminate variables to simplify
- Existential quantification by default
- Replace entity sets ($e \to t$) rather than truth values (t)

Comparison to lambda calculus

Lambda calculus

$$\lambda x.\textsf{Type}(x, \textsf{Person}) \wedge \exists e.\textsf{PlacesLived}(x, e) \wedge \textsf{Location}(e, \textsf{Chicago})$$

Lambda dependency-based compositional semantics (DCS)

$$\textsf{Type.Person} \sqcap \textsf{PlacesLived.Location.Chicago}$$

- Eliminate variables to simplify
- Existential quantification by default
- Replace entity sets ($e \to t$) rather than truth values (t)
- Use tree/graph structures (analogy to dependency syntax)

Comparison to lambda calculus

Lambda calculus

$$\lambda x.\text{Type}(x, \text{Person}) \wedge \exists e.\text{PlacesLived}(x, e) \wedge \text{Location}(e, \text{Chicago})$$

Lambda dependency-based compositional semantics (DCS)

$$\text{Type.Person} \sqcap \text{PlacesLived.Location.Chicago}$$

- Eliminate variables to simplify
- Existential quantification by default
- Replace entity sets ($e \to t$) rather than truth values (t)
- Use tree/graph structures (analogy to dependency syntax)
- Inspired by Discourse Representation Theory, description logic, modal logic, similarities to AMR

An example

states west of Texas whose capitals are their largest cities

Outline

- A semantic parsing framework
- A closer look at the elements
  - Logical forms: lambda DCS
  - **Lexical coverage**
  - Grammar: building logical forms
  - Learning via bootstrapping
  - Leveraging denotations
  - Datasets/results
- Beyond Freebase
- Final remarks

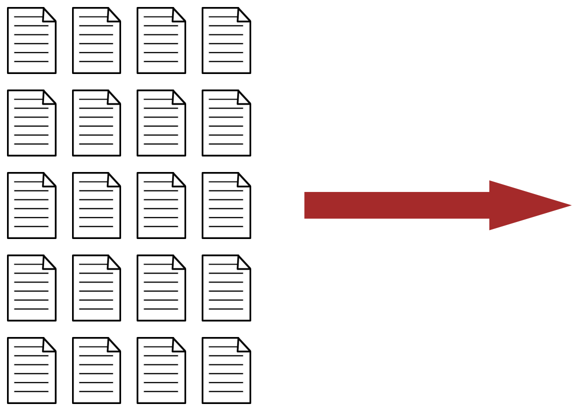
Challenge: lexical coverage

born ⇒ Type.City, PlaceOfBirth, Profession.Lawyer, ...

?

Solution: alignment

Open information extraction on ClueWeb09:



(*Barack Obama*, *was born in*, *Honolulu*)

(*Albert Einstein*, *was born in*, *Ulm*)

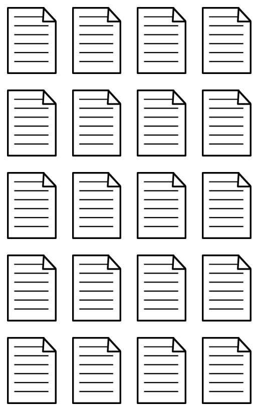
(*Barack Obama*, *lived in*, *Chicago*)

... 15M triples ...

[Fader et al. 2011]

Solution: alignment

Open information extraction on ClueWeb09:



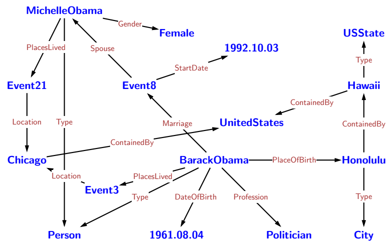
(*Barack Obama*, *was born in*, *Honolulu*)

(*Albert Einstein*, *was born in*, *Ulm*)

(*Barack Obama*, *lived in*, *Chicago*)

... 15M triples ...

Freebase:



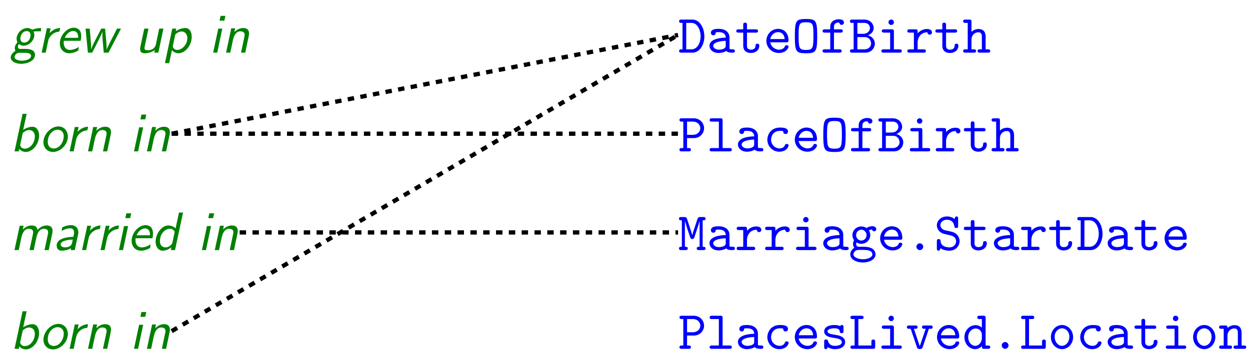
(`BarackObama`, `PlaceOfBirth`, `Honolulu`)

(`Albert Einstein`, `PlaceOfBirth`, `Ulm`)

(`BarackObama`, `PlacesLived.Location`, `Chicago`)

... 400M triples ...

Match text and Freebase predicates



Similar schema matching / alignment ideas [Cai & Yates, 2013, Fader et. al, 2013, Yao & van Durme, 2014; etc.]

Challenge: variability in language

What is the currency in the US?

Challenge: variability in language

What is the currency in the US?

What money do they use in the states?

How do you pay in America?

What's the currency of the US?

What money is accepted in the United States?

What money to take to the US?

...

[Berant & Liang, 2014]

A solution: paraphrasing

How many people live in Seattle?

↓

paraphrase

↓

What is the population of Seattle?

↓

PopulationOf(Seattle)

↓

`850,000`

Convert to a text-only problem

Challenge: "sub-lexical compositionality"

grandmother

$$\lambda x.\text{Gender.Female} \sqcap \text{Children.Children}.x$$

mayor

$$\lambda x.\text{GovtPositionsHeld.}(\text{Title.Mayor} \sqcap \text{OfficeOfJurisdiction}.x)$$

Challenge: "sub-lexical compositionality"

grandmother

$$\lambda x.\text{Gender.Female} \sqcap \text{Children.Children}.x$$

mayor

$$\lambda x.\text{GovtPositionsHeld.}(\text{Title.Mayor} \sqcap \text{OfficeOfJurisdiction}.x)$$

This is "grounding".

Challenge: ”sub-lexical compositionality”

presidents who have served two **non-consecutive** *terms*

[requires higher-order quantification]

presidents who were **previously** *vice-presidents*

[anaphora]

every other *president*

[weird quantification]

Outline

- A semantic parsing framework
- A closer look at the elements
  - Logical forms: lambda DCS
  - Lexical coverage
  - **Grammar: building logical forms**
  - Learning via bootstrapping
  - Leveraging denotations
  - Datasets/results
- Beyond Freebase
- Final remarks

Over-simplifying the grammar

A Real Dumb Grammar

(lexicon) *Chicago* $\Rightarrow$ E : Chicago

(lexicon) *people* $\Rightarrow$ E : Type.Person

(lexicon) *live* $\Rightarrow$ E $\times$ E : PlacesLived

...

(join) E $\times$ E : b E : u $\Rightarrow$ E : $b.u$

(intersect) E : u E : v $\Rightarrow$ E : $u \sqcap v$

Over-simplifying the grammar

A Real Dumb Grammar

(lexicon) *Chicago* $\Rightarrow$ E : Chicago

(lexicon) *people* $\Rightarrow$ E : Type.Person

(lexicon) *live* $\Rightarrow$ E $\times$ E : PlacesLived

...

(join) E $\times$ E : b E : u $\Rightarrow$ E : $b.u$

(intersect) E : u E : v $\Rightarrow$ E : $u \sqcap v$

A CCG Grammar

(lexicon) *Chicago* $\Rightarrow$ NP : PlaceOfBirth

(lexicon) *people* $\Rightarrow$ N : $\lambda x.\text{Type}(x, \text{Person})$

(lexicon) *live* $\Rightarrow$ (N\N)/NP : $\lambda y.\lambda f.\lambda x.f(x) \wedge \text{PlacesLived}(x, y)$

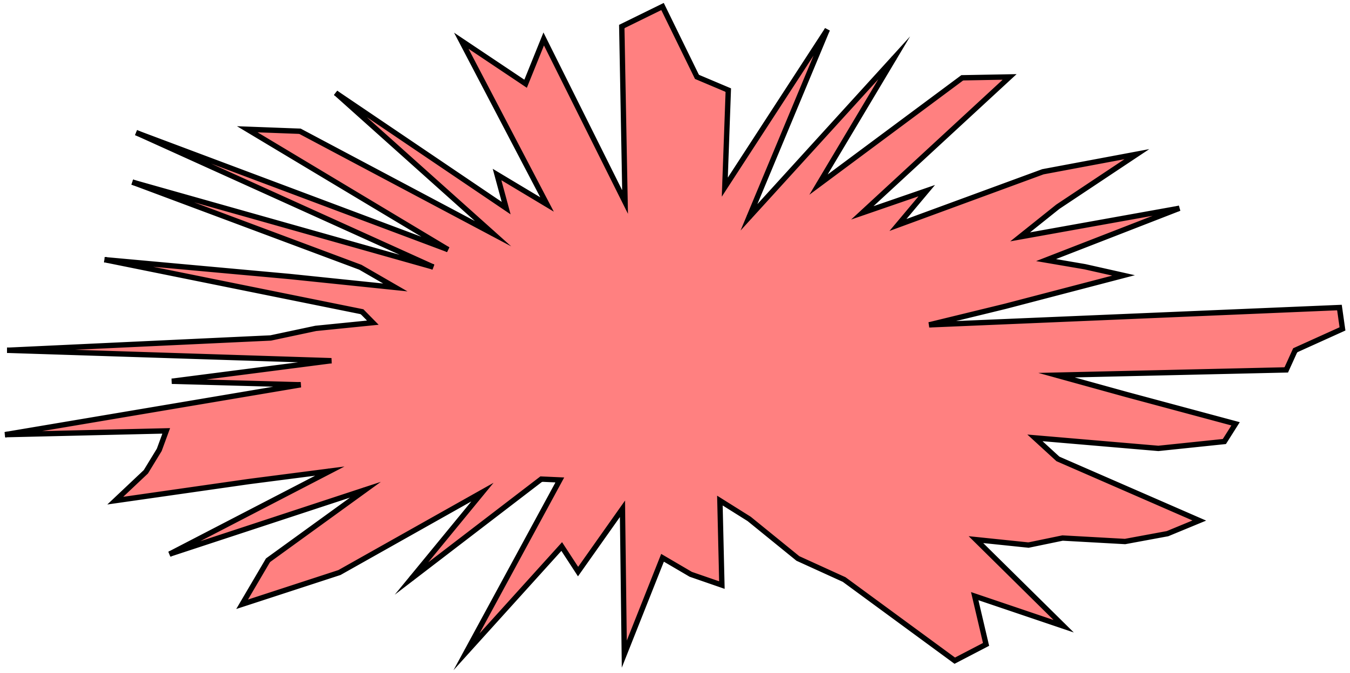
...

(backward) X : x $Y\backslash$X : f $\Rightarrow$ Y : $f(x)$

(forward) $Y/$X : b X : u $\Rightarrow$ Y : $f(x)$

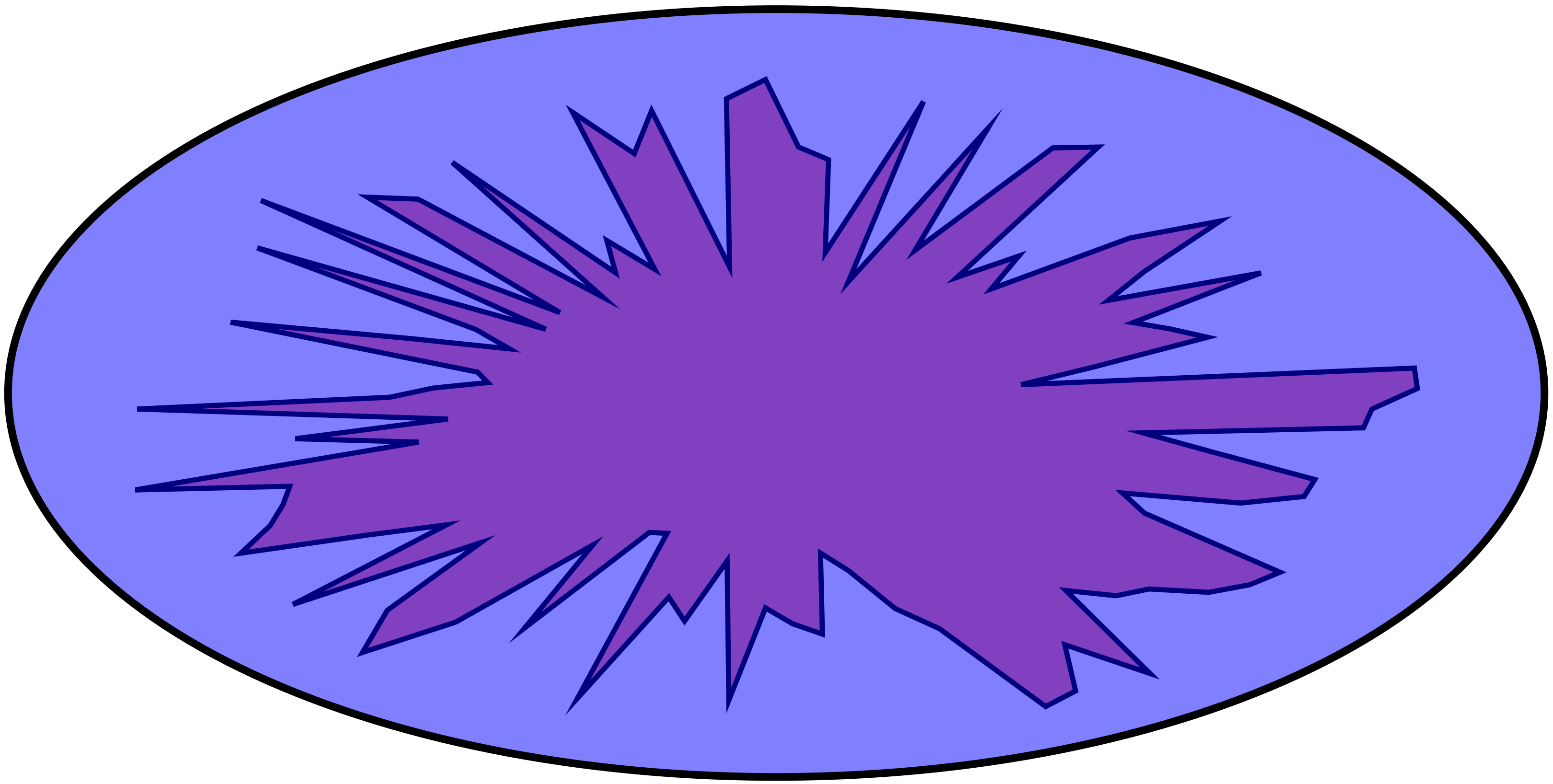
Overapproximation via simple grammars

- Modeling correct derivations requires complex rules



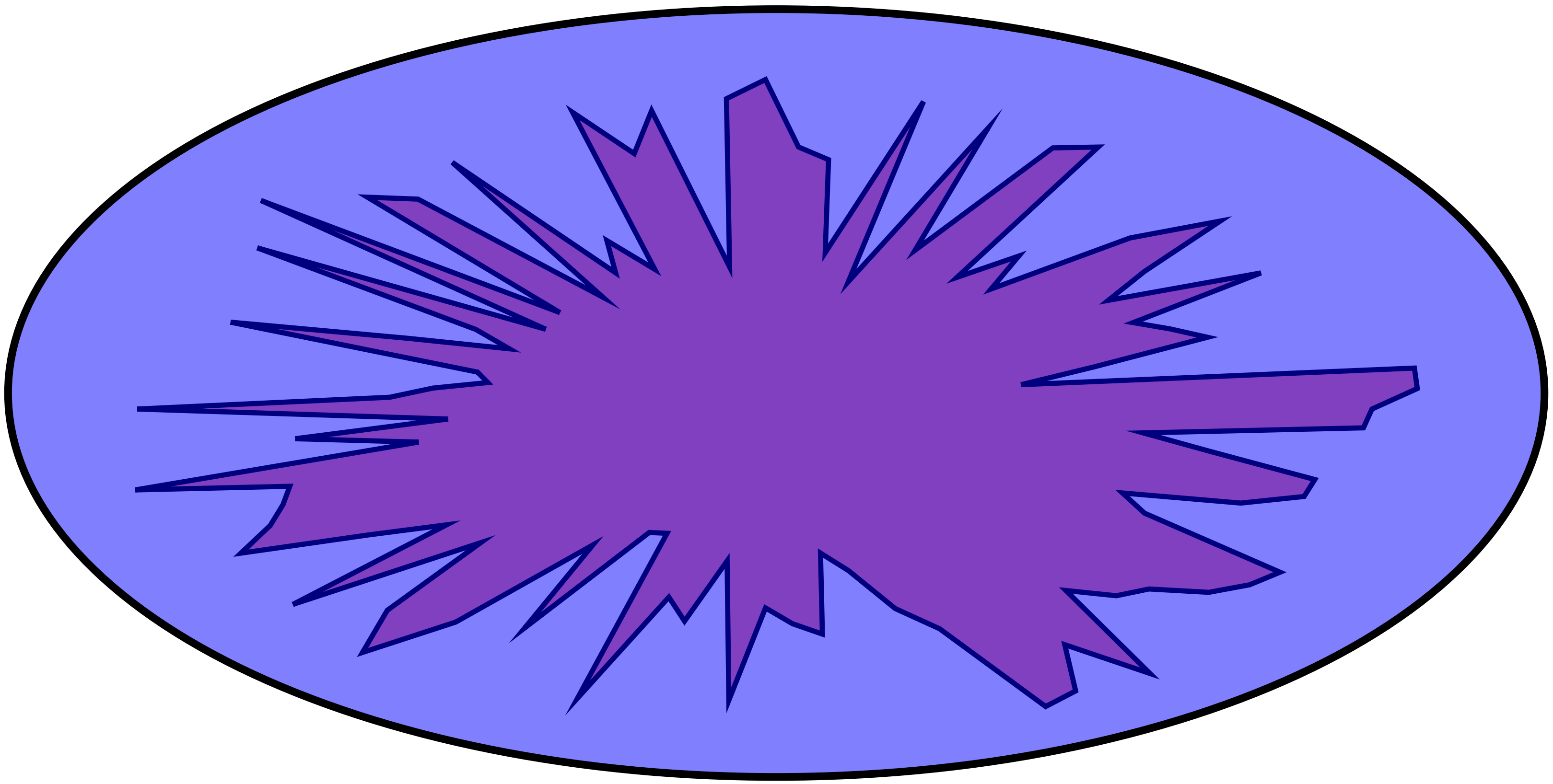
Overapproximation via simple grammars

- Modeling correct derivations requires complex rules
- Simple rules generate overapproximation of good derivations



Overapproximation via simple grammars

- Modeling correct derivations requires complex rules
- Simple rules generate overapproximation of good derivations



- Hard grammar rules $\Rightarrow$ soft/overlapping features

Bridging

$\text{Type.University} \sqcap \text{Education.BarackObama}$

`Type.University`

alignment

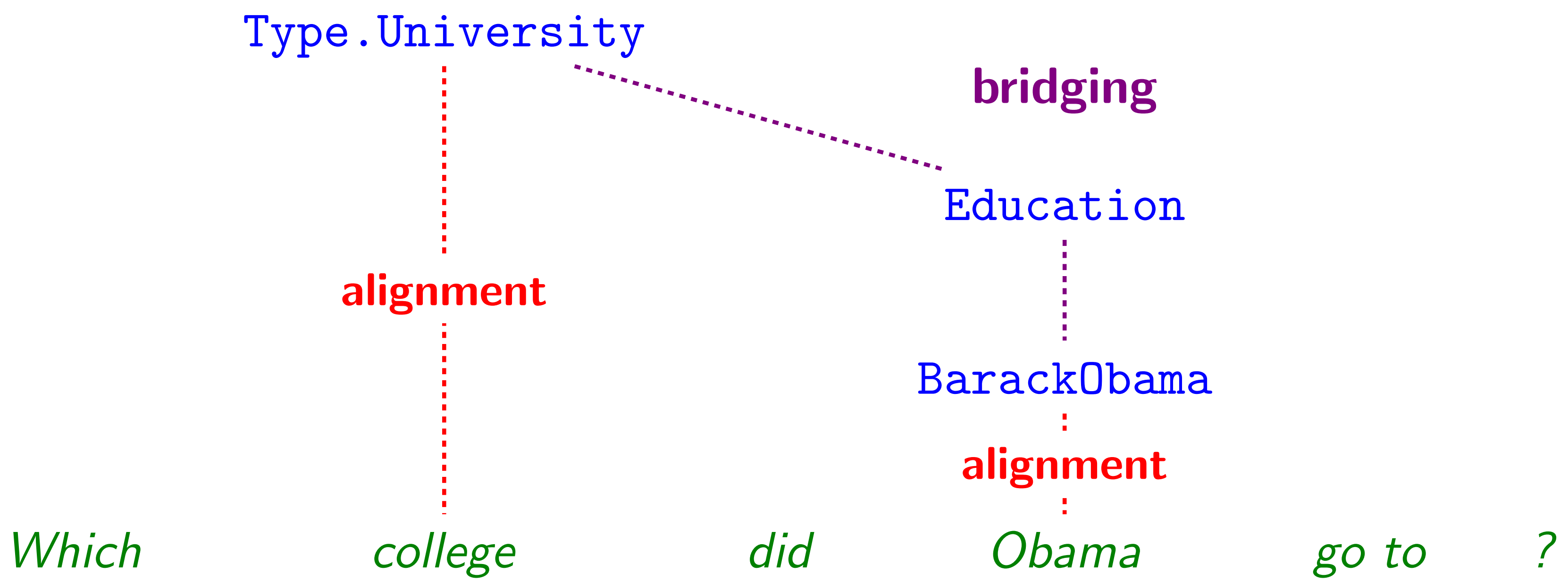
`BarackObama`

alignment

Which college did Obama go to ?

Bridging

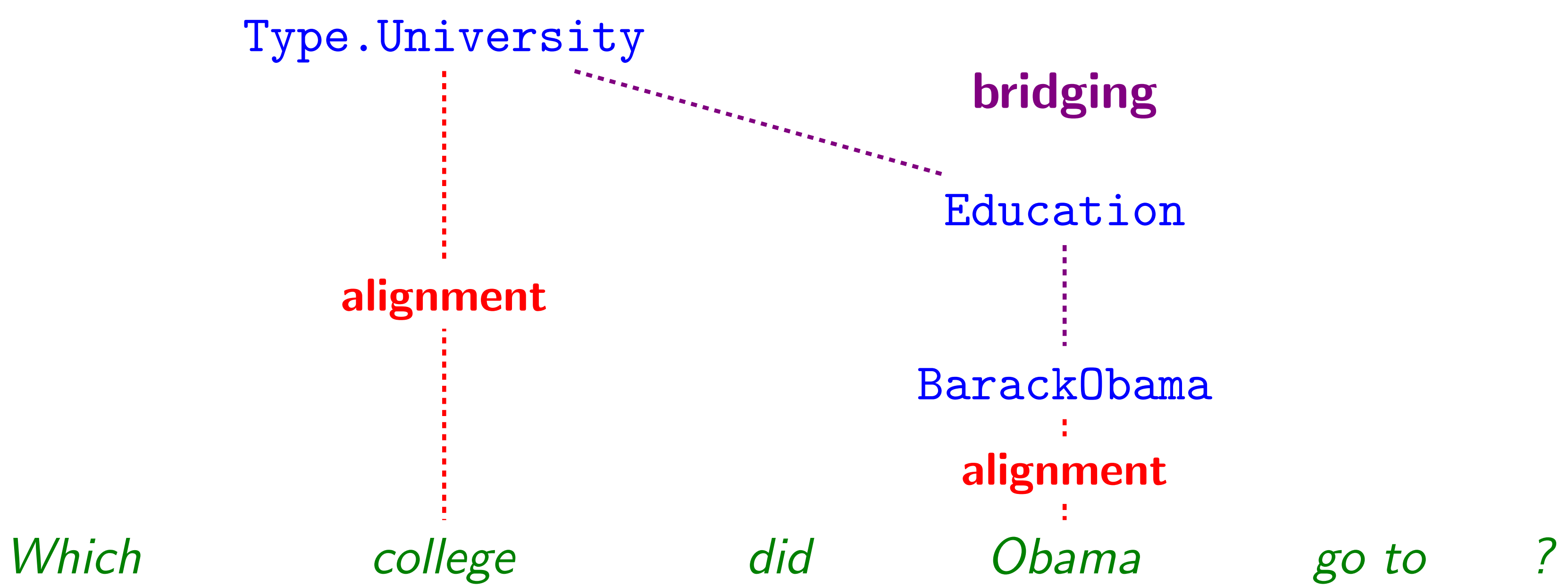
$\text{Type.University} \sqcap \text{Education.BarackObama}$



Bridging: use neighboring predicates / type constraints

Bridging

$\text{Type.University} \sqcap \text{Education.BarackObama}$



Bridging: use neighboring predicates / type constraints

Start building from parts with more certainty

Bridging to nowhere

Search logical forms based on "prior":

What countries in the world speak Arabic?

[Berant & Liang, 2014]

Bridging to nowhere

Search logical forms based on "prior":

What countries in the world speak Arabic?

ArabicAlphabet ArabicLang

Bridging to nowhere

Search logical forms based on "prior":

What countries in the world speak Arabic?

ArabicAlphabet

ArabicLang

`LangSpoken.ArabicLang`

`LangFamily.Arabic`

[Berant & Liang, 2014]

Bridging to nowhere

Search logical forms based on "prior":

What countries in the world speak Arabic?

- ArabicAlphabet
- ArabicLang
  - `LangSpoken.ArabicLang`
    - `Type.Country` $\sqcap$ `LangSpoken.ArabicLang`
      - `Count(Type.Country` $\sqcap$ `LangSpoken.ArabicLang)`
  - `LangFamily.Arabic`

Bridging to nowhere

Search logical forms based on "prior":

What countries in the world speak Arabic?

ArabicAlphabet

ArabicLang

`LangSpoken.ArabicLang`

`LangFamily.Arabic`

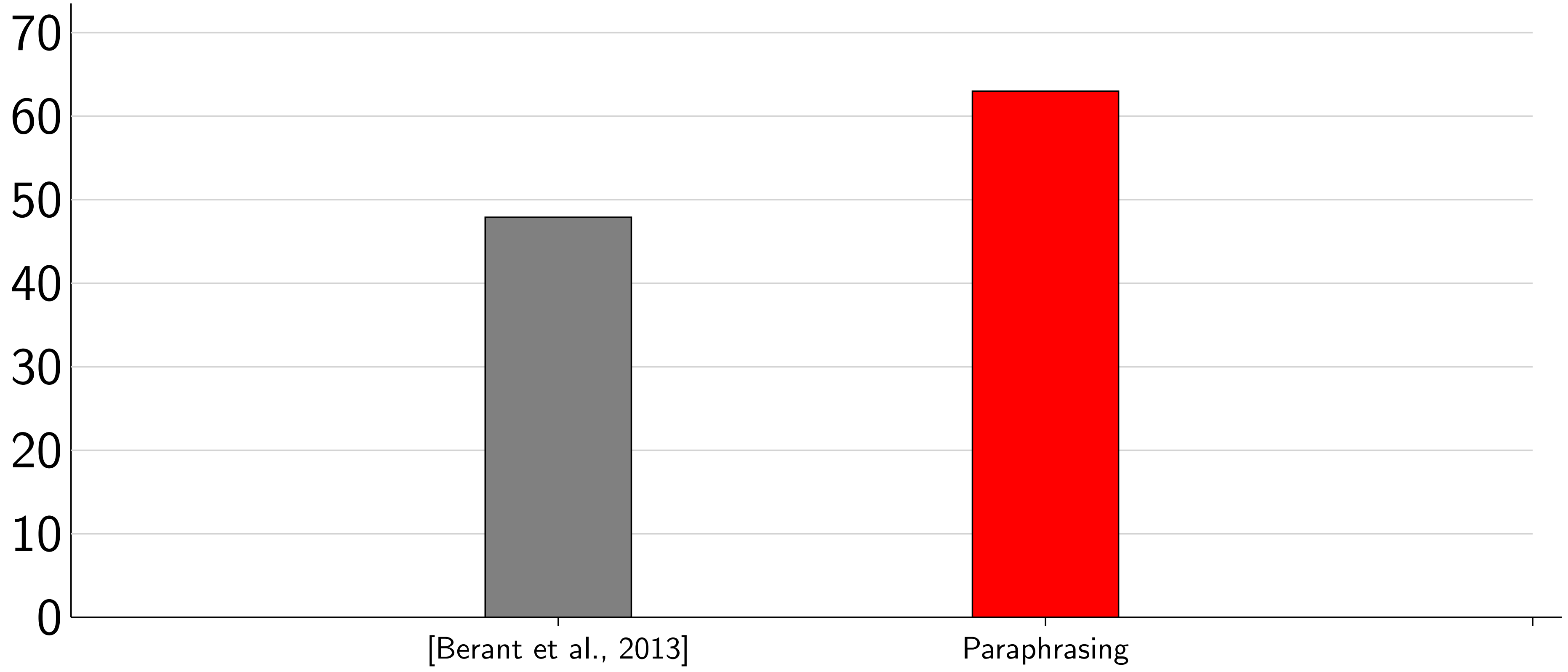
$\texttt{Type.Country} \sqcap \texttt{LangSpoken.ArabicLang}$

$\texttt{Count(Type.Country} \sqcap \texttt{LangSpoken.ArabicLang)}$

$\boxed{\textbf{Start building from parts with more certainty}}$

Oracle on WebQuestions

For what fraction of utterances was a candidate logical form correct?



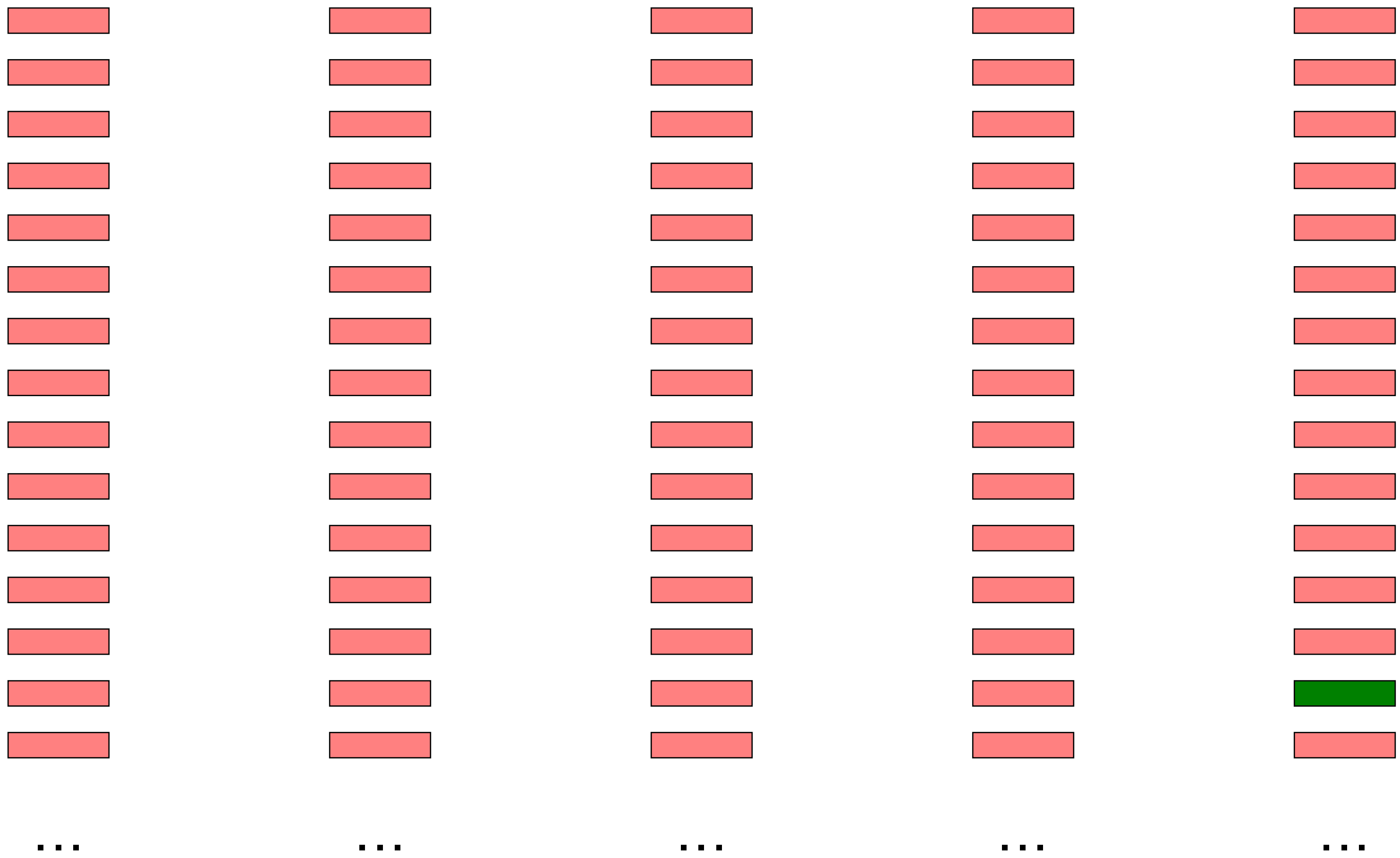
Outline

- A semantic parsing framework
- A closer look at the elements
  - Logical forms: lambda DCS
  - Lexical coverage
  - Grammar: building logical forms
  - **Learning via bootstrapping**
  - Leveraging denotations
  - Datasets/results
- Beyond Freebase
- Final remarks

Bootstrapping from easy examples

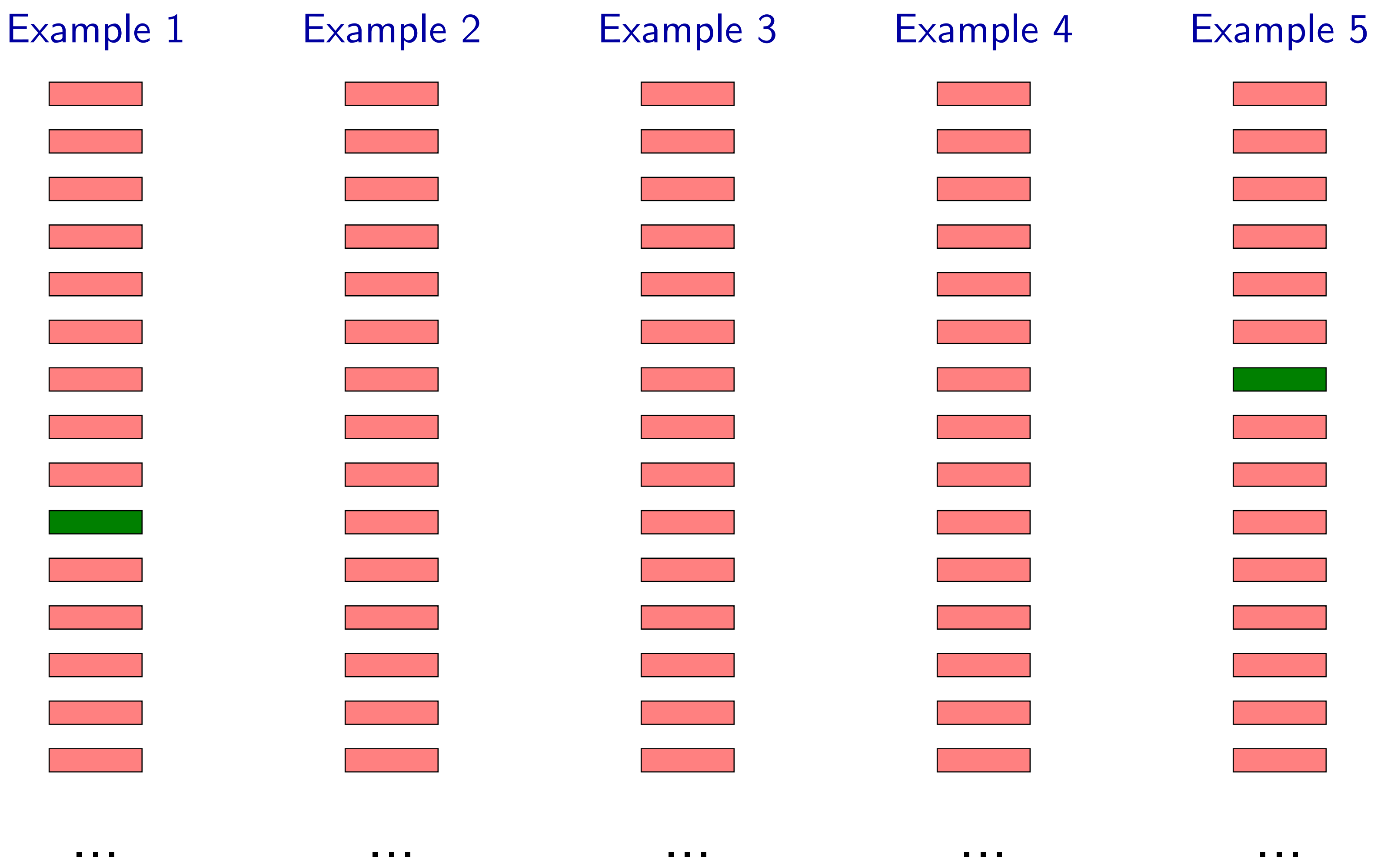
Iteration 1

Example 1 Example 2 Example 3 Example 4 Example 5



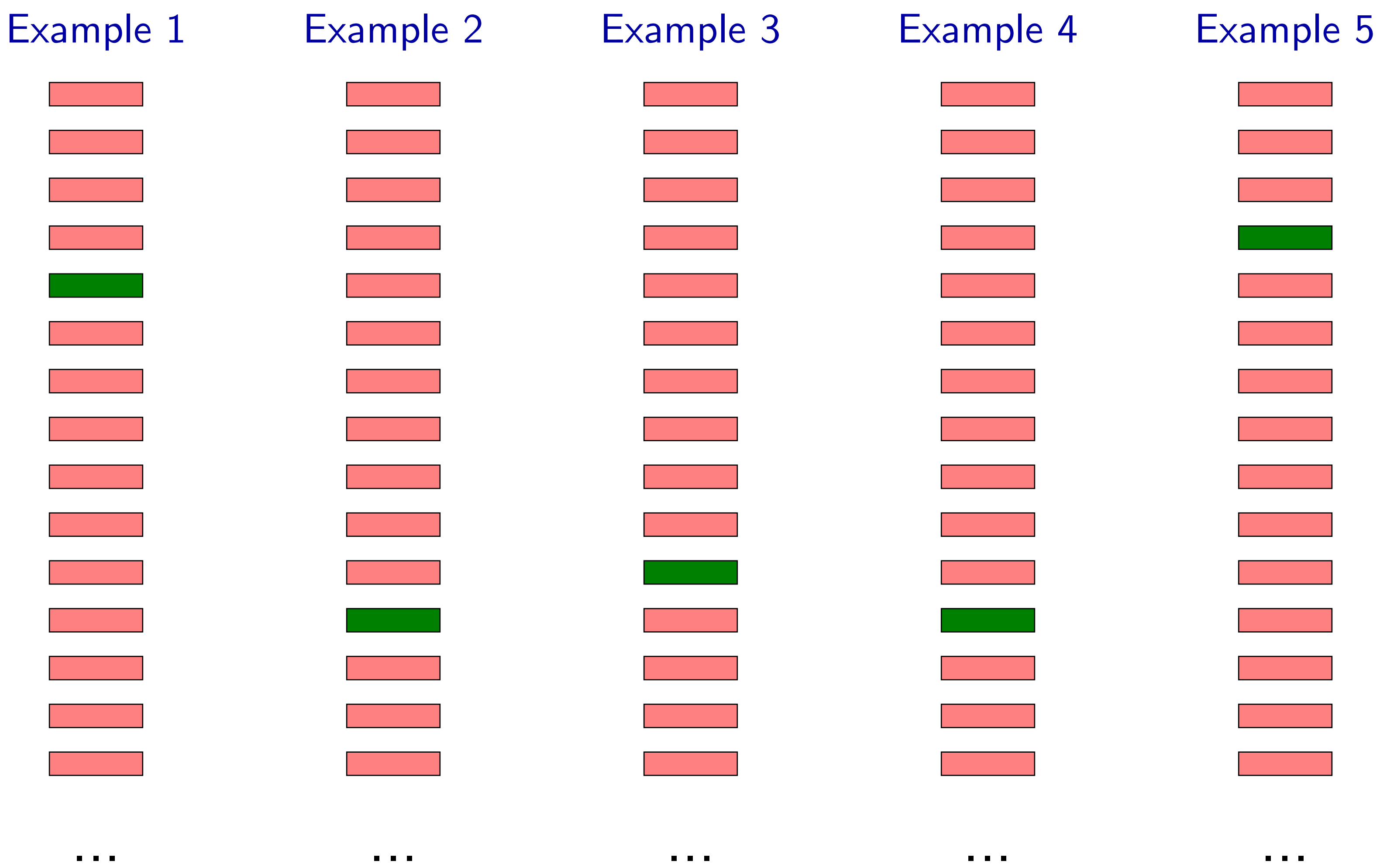
Bootstrapping from easy examples

Iteration 2



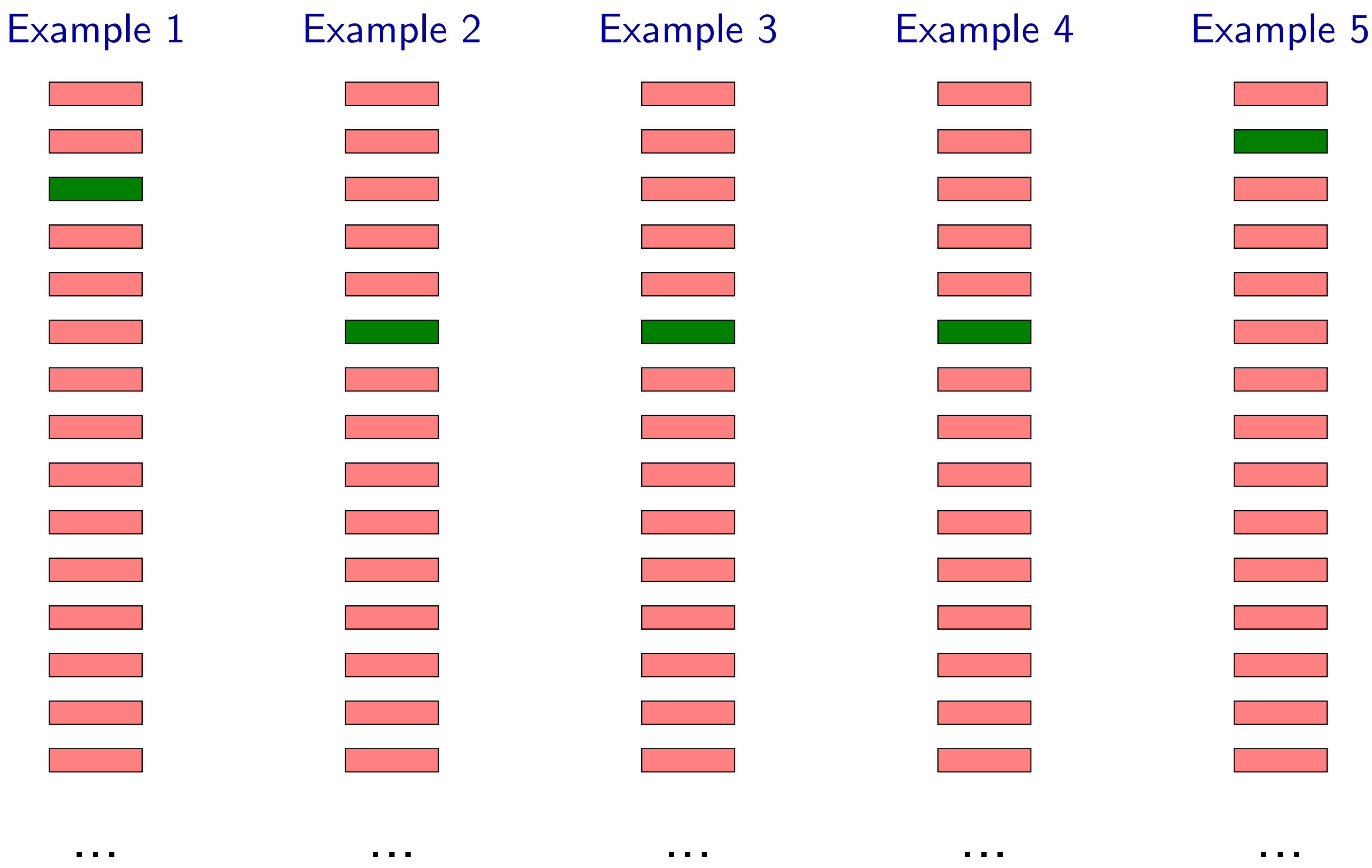
Bootstrapping from easy examples

Iteration 3



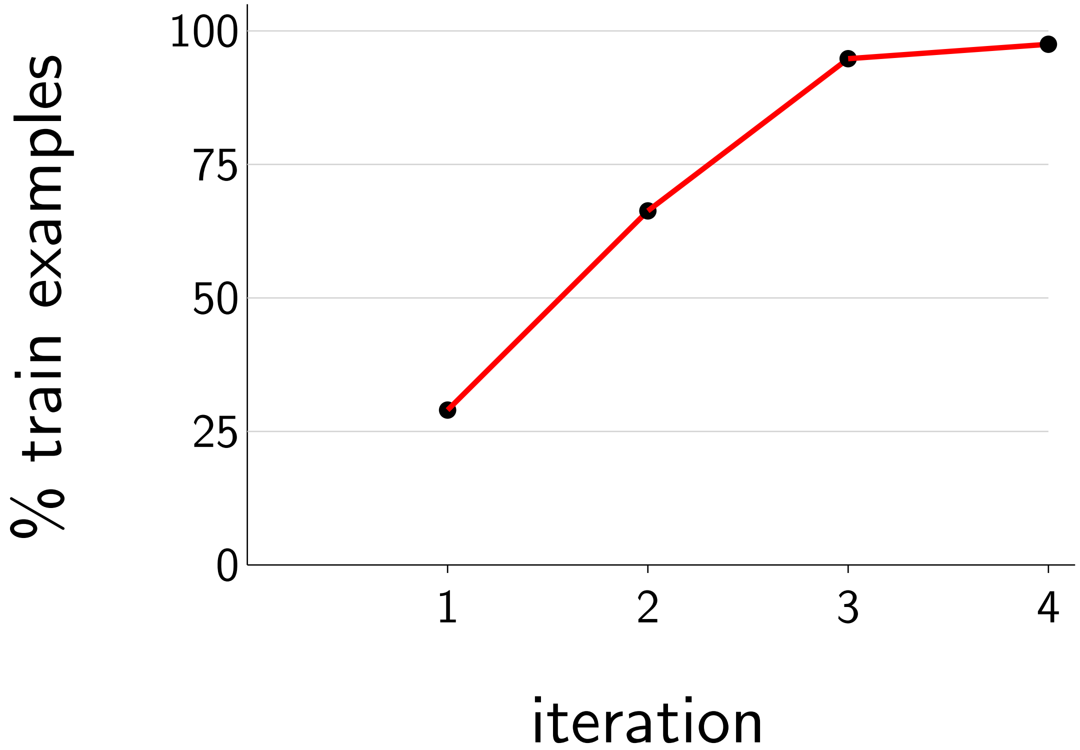
Bootstrapping from easy examples

Iteration 4



Bootstrapping from easy examples

On GeoQuery [Liang et al., 2011]:

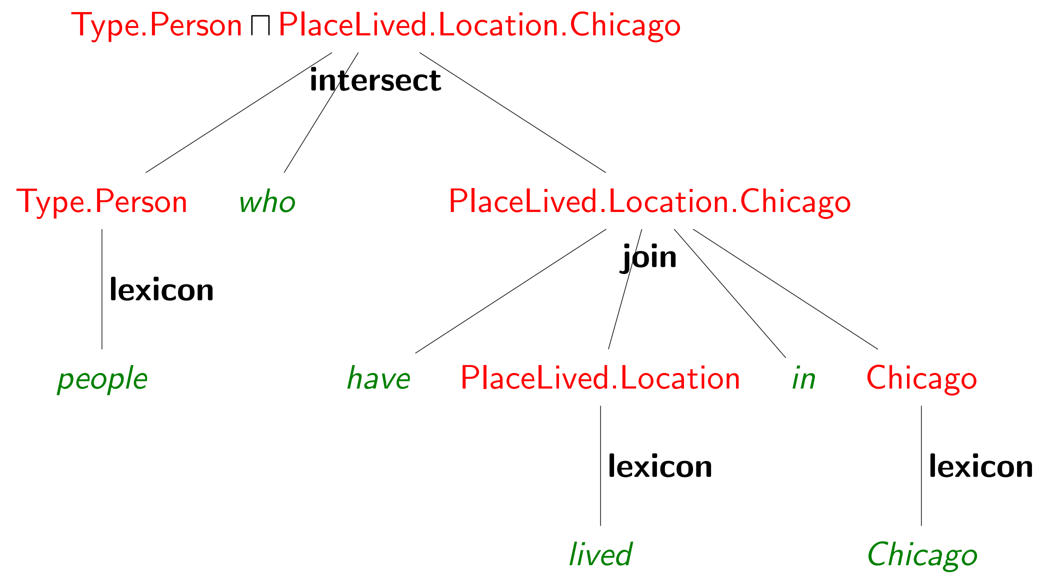


Outline

- A semantic parsing framework
- A closer look at the elements
  - Logical forms: lambda DCS
  - Lexical coverage
  - Grammar: building logical forms
  - Learning via bootstrapping
  - **Leveraging denotations**
  - Datasets/results
- Beyond Freebase
- Final remarks

x: utterance

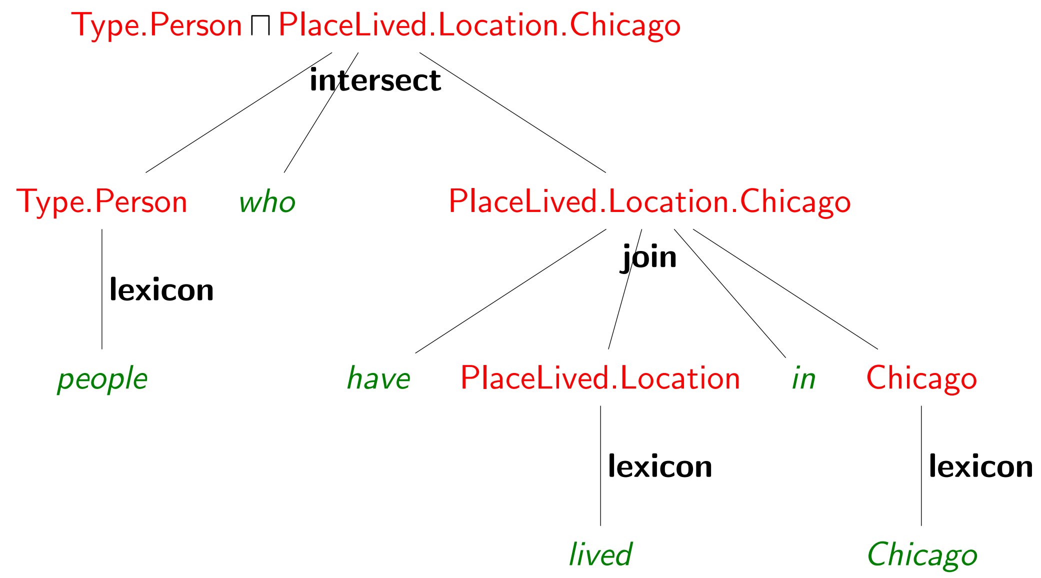
d: derivation



Feature vector $\phi(x, d) \in \mathbb{R}^F$:

x: utterance

d: derivation

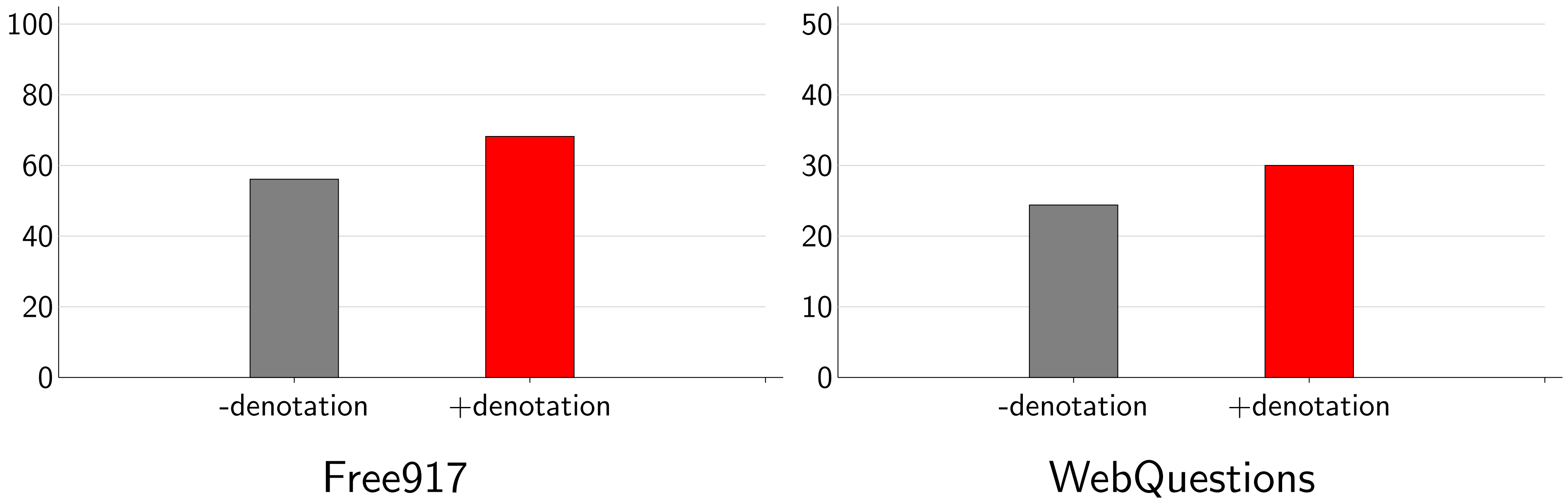


Feature vector $\phi(x, d) \in \mathbb{R}^F$:

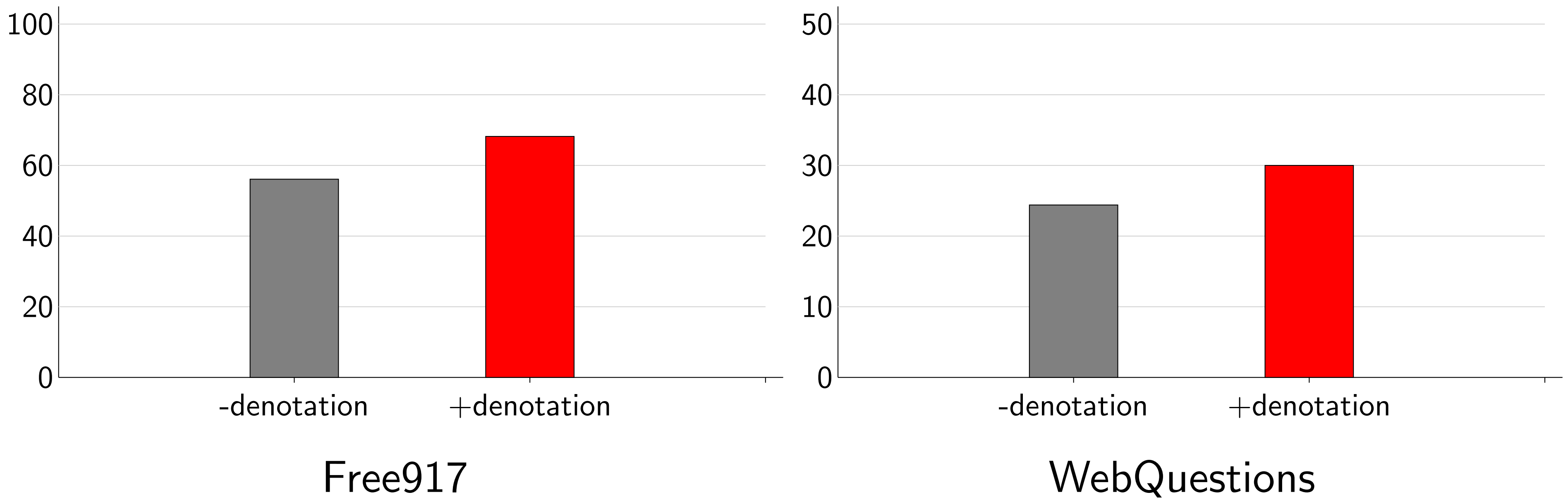
apply join	1
apply intersect	1
apply lexicon	3
skipped IN	1
skipped NN	0
lived maps to PlacesLived.Location	1
lived maps to PlaceOfBirth	0
alignmentScore	1.52
denotation-size=1	1
...	...

Impact of denotation features

Impact of denotation features



Impact of denotation features



Working with denotations actually provides more information than just logical forms

Outline

- A semantic parsing framework
- A closer look at the elements
  - Logical forms: lambda DCS
  - Lexical coverage
  - Grammar: building logical forms
  - Learning via bootstrapping
  - Leveraging denotations
  - **Datasets/results**
- Beyond Freebase
- Final remarks

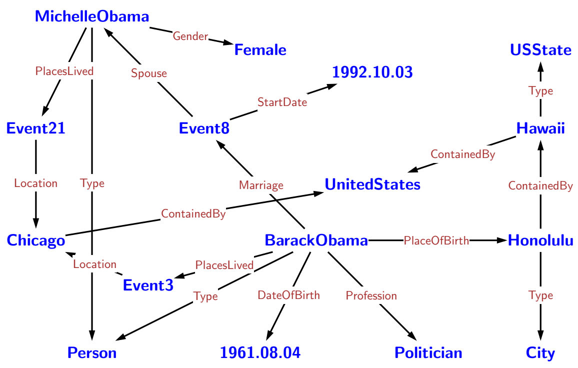
Question answering on Freebase

Free917 (1K questions) [Cai & Yates, 2013]

At what institutions was Marshall Hall a professor?

How many companies are traded by the NYSE?

Who was the newscaster in 1948 on CBS evening news?



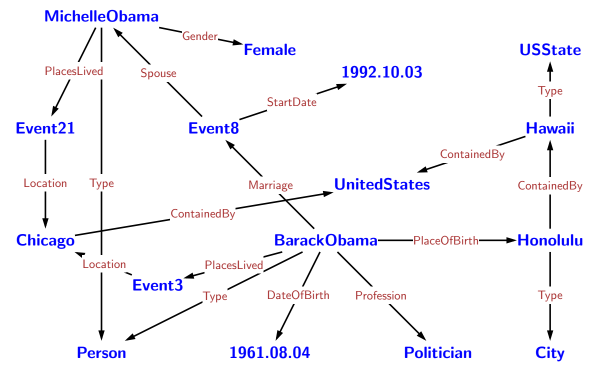
Question answering on Freebase

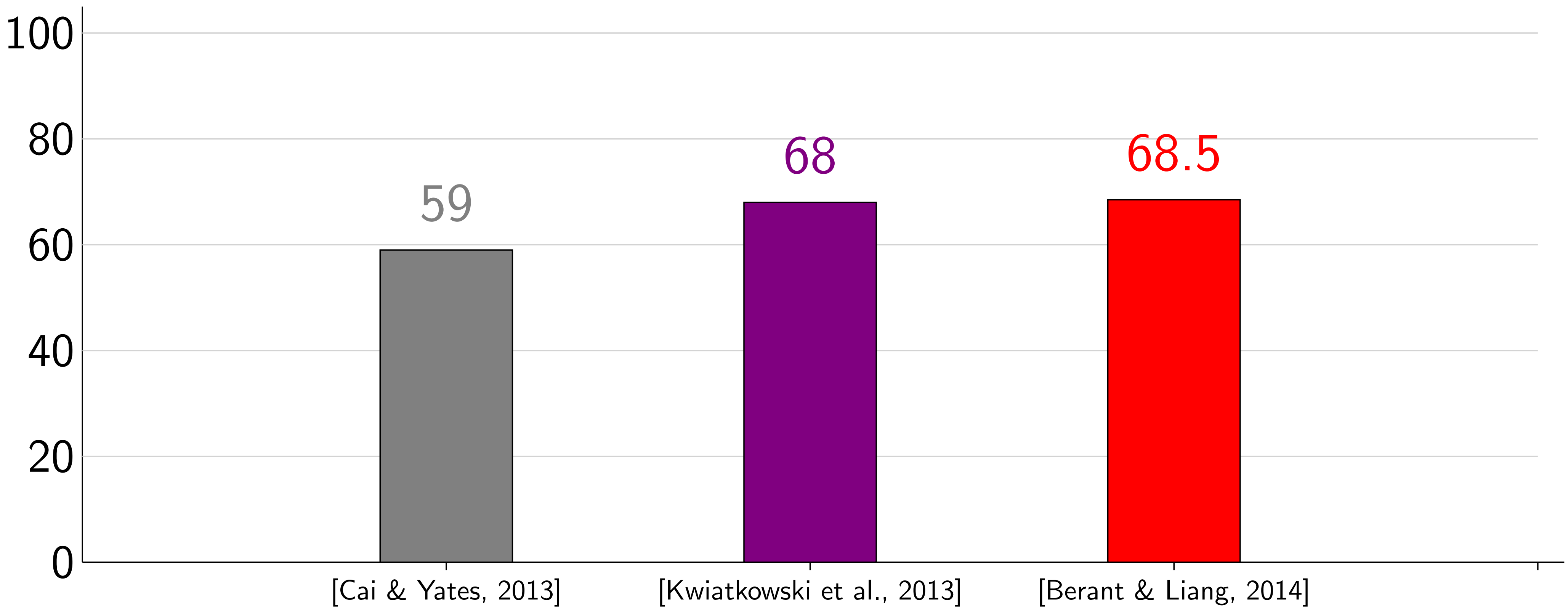
Free917 (1K questions) [Cai & Yates, 2013]

At what institutions was Marshall Hall a professor?

How many companies are traded by the NYSE?

Who was the newscaster in 1948 on CBS evening news?





Dataset collection

How to get natural questions (inputs)?

Dataset collection

How to get natural questions (inputs)?

Strategy: breadth-first search over Google Suggest graph

Dataset collection

How to get natural questions (inputs)?

Strategy: breadth-first search over Google Suggest graph

Where was Barack Obama born?

Dataset collection

How to get natural questions (inputs)?

Strategy: breadth-first search over Google Suggest graph

Where was Barack Obama born?

Where was _ born? $\xrightarrow{\textbf{Google Suggest}}$ Barack Obama
Lady Gaga
Steve Jobs

Dataset collection

How to get natural questions (inputs)?

Strategy: breadth-first search over Google Suggest graph

Where was Barack Obama born?

Where was _ born? $\xrightarrow{\textbf{Google Suggest}}$ Barack Obama
Lady Gaga
Steve Jobs

Where was Steve Jobs born?

Dataset collection

How to get natural questions (inputs)?

Strategy: breadth-first search over Google Suggest graph

Where was Barack Obama born?

Where was _ born? $\xrightarrow{\textbf{Google Suggest}}$ Barack Obama / Lady Gaga / Steve Jobs

Where was Steve Jobs born?

Where was Steve Jobs _? $\xrightarrow{\textbf{Google Suggest}}$ born / raised / on the Forbes list

Dataset collection

How to get natural questions (inputs)?

Strategy: breadth-first search over Google Suggest graph

Where was Barack Obama born?

Where was _ born? $\xrightarrow{\textbf{Google Suggest}}$ Barack Obama / Lady Gaga / Steve Jobs

Where was Steve Jobs born?

Where was Steve Jobs _? $\xrightarrow{\textbf{Google Suggest}}$ born / raised / on the Forbes list

Where was Steve Jobs raised?

Dataset collection

How to get natural questions (inputs)?

Strategy: breadth-first search over Google Suggest graph

Where was Barack Obama born?

Where was _ born? $\xrightarrow{\textbf{Google Suggest}}$ Barack Obama / Lady Gaga / Steve Jobs

Where was Steve Jobs born?

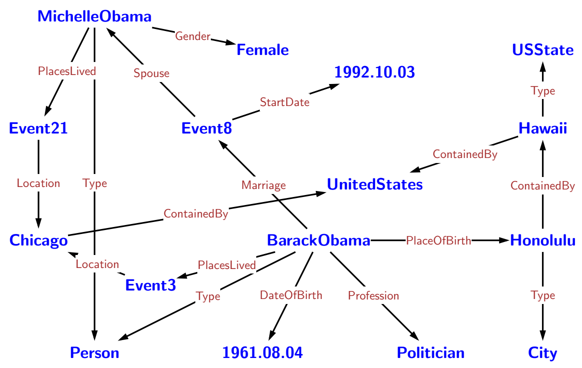
Where was Steve Jobs _? $\xrightarrow{\textbf{Google Suggest}}$ born / raised / on the Forbes list

Where was Steve Jobs raised?

...

AMT annotation ⇒ 6.6K question/answer pairs

Question answering on Freebase



WebQuestions dataset (6K questions) [Berant et al., 2013]

what did obama study in school

where to fly into bali

what was tupac name in juice

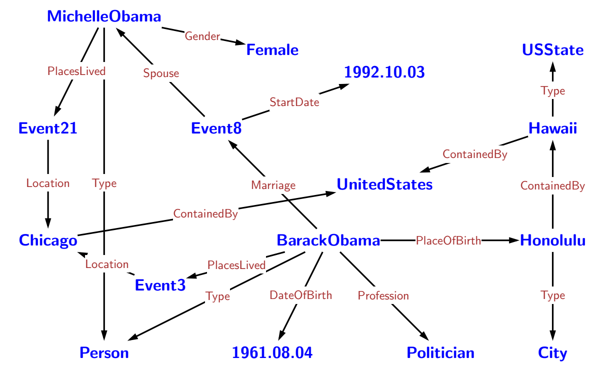
Question answering on Freebase

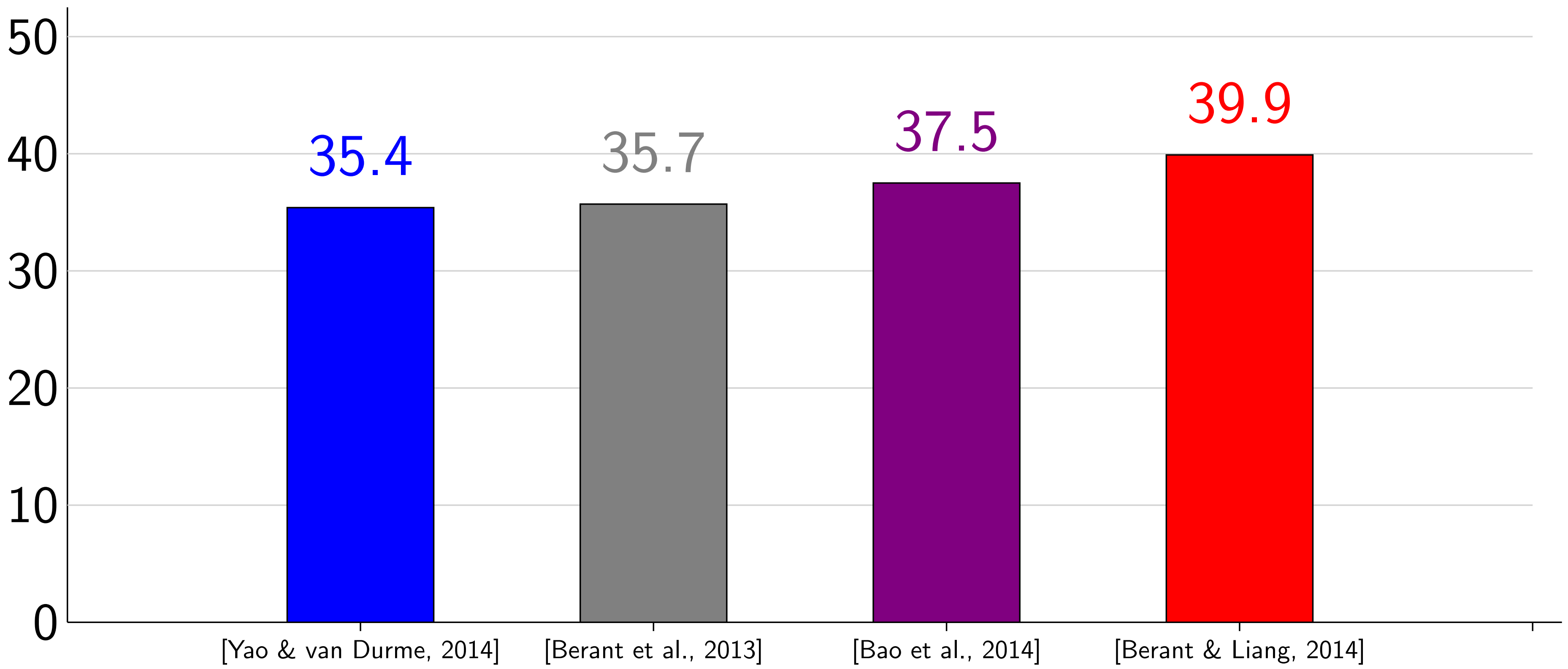
WebQuestions dataset (6K questions) [Berant et al., 2013]

what did obama study in school

where to fly into bali

what was tupac name in juice





Other semantic parsing tasks

Playing computer games [Branavan et al., 2010, 2011]

Following navigational instructions [Tellex et. al 2011; Chen et. al 2012; Artzi et. al 2013]

Understanding visual world [Matuszek et. al; 2012, Krishnamurthy & Kollar, 2013]

Solving algebra word problems [Kushman et. al, 2013]

Outline

- A semantic parsing framework
- A closer look at the elements
  - Logical forms: lambda DCS
  - Lexical coverage
  - Grammar: building logical forms
  - Learning via bootstrapping
  - Leveraging denotations
  - Datasets/results
- **Beyond Freebase**
- Final remarks

Challenge: incomplete knowledge base

What are the longest **hiking trails in Baltimore***?*

Data Source

hiking trails in Baltimore

Avalon Super Loop

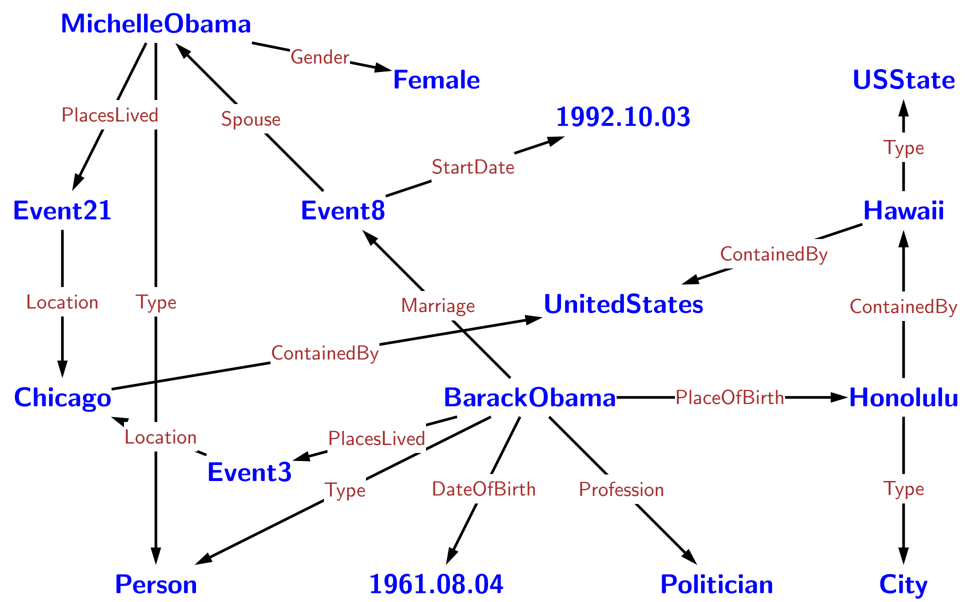
Patapsco Valley State Park

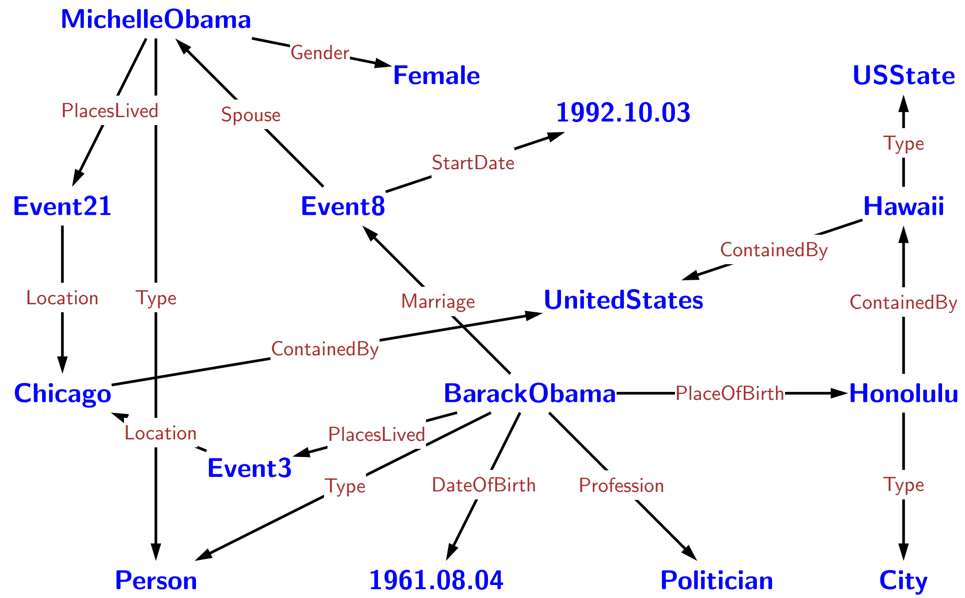
Gunpowder Falls State Park

Union Mills Hike

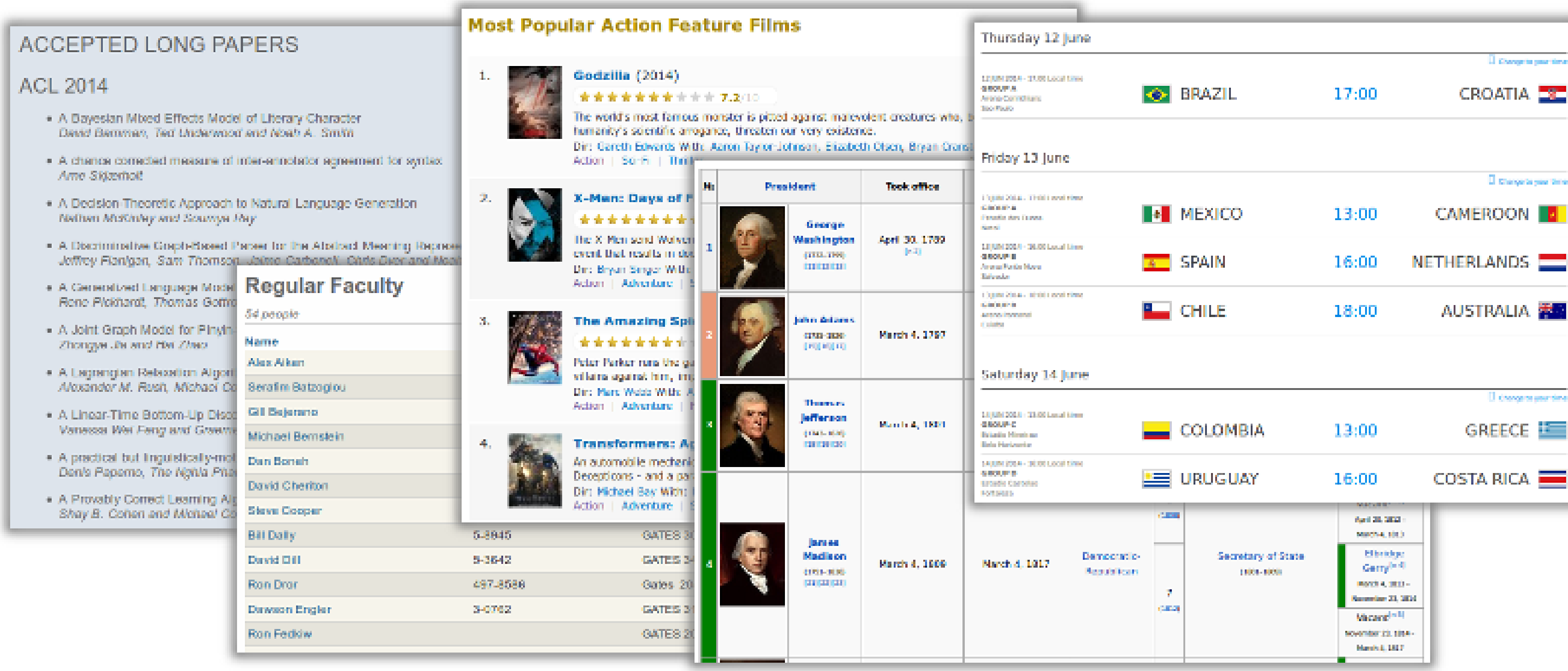
Greenbury Point

...


MichelleObama
Gender
Female
USState
PlacesLived
Spouse
1992.10.03
Type
StartDate
Event21
Event8
Hawaii
ContainedBy
Location
Type
Marriage
UnitedStates
ContainedBy
ContainedBy
Chicago
BarackObama
PlaceOfBirth
Honolulu
Location
PlacesLived
Event3
Type
DateOfBirth
Profession
Type
Person
1961.08.04
Politician
City



Fewer than 10% general web questions can be answered via Freebase



ACCEPTED LONG PAPERS
ACL 2014
Regular Faculty
Most Popular Action Feature Films
Godzilla (2014)
Thursday 12 June
BRAZIL 17:00 CROATIA
Friday 13 June
MEXICO 13:00 CAMEROON
SPAIN 16:00 NETHERLANDS
CHILE 18:00 AUSTRALIA
Saturday 14 June
COLOMBIA 13:00 GREECE
URUGUAY 16:00 COSTA RICA

[Pasupat & Liang, 2014]

Semantic parsing on the web

Input:

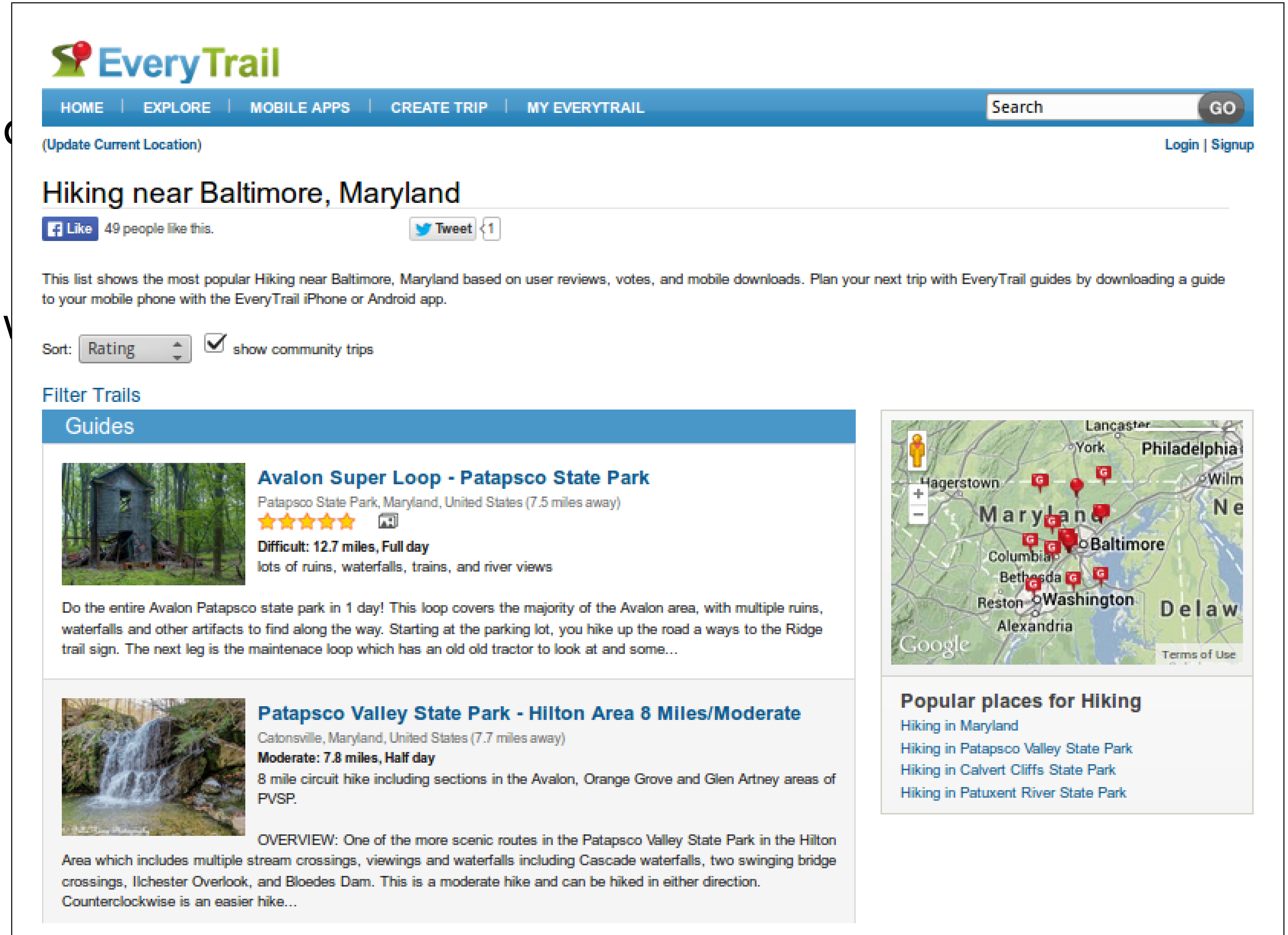
- query x

hiking trails near Baltimore

- web page w

Semantic parsing on the web

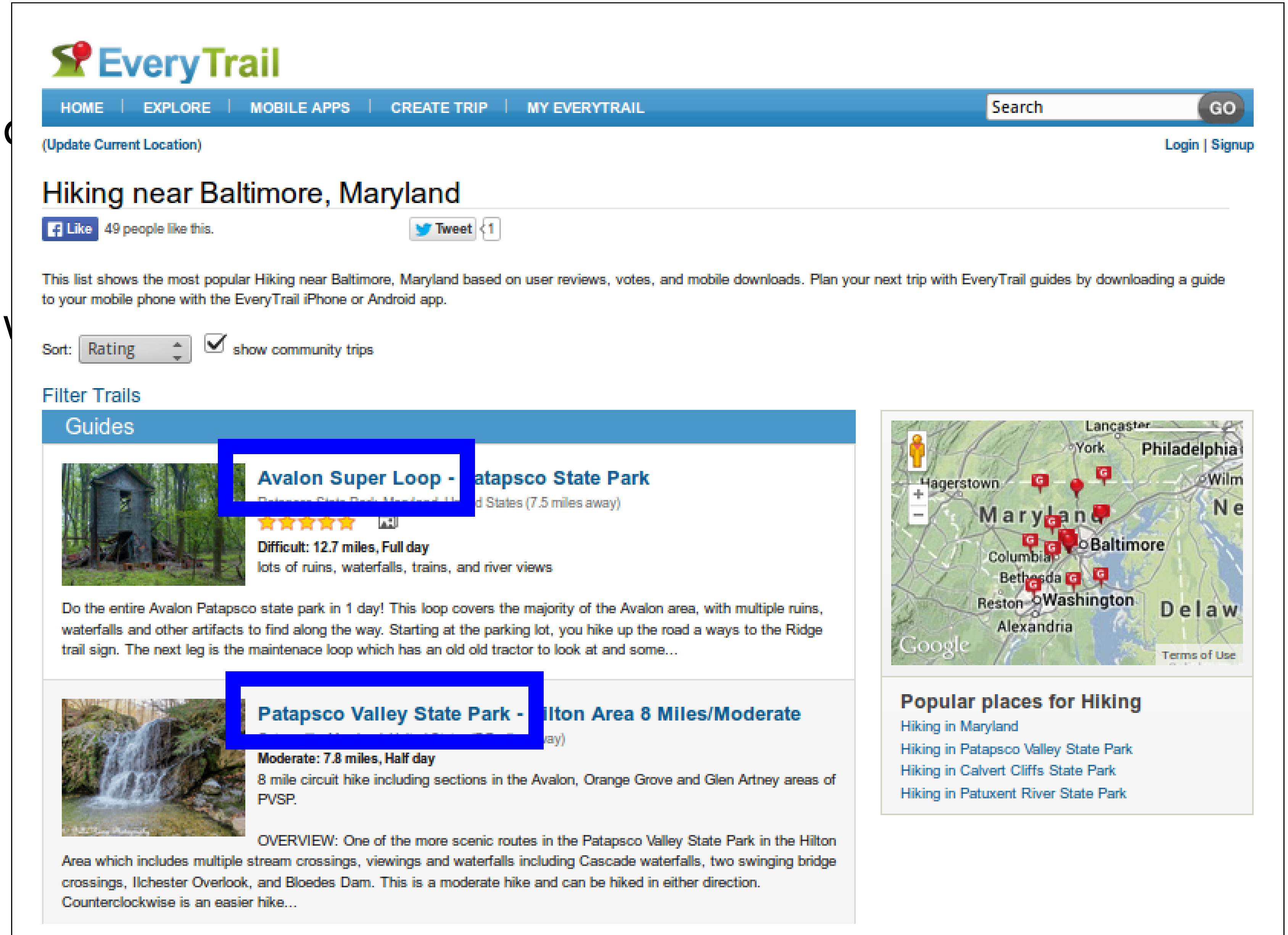
Input:



[Pasupat & Liang, 2014]

Semantic parsing on the web

Input:



[Pasupat & Liang, 2014]

Semantic parsing on the web

Input:

- query x

hiking trails near Baltimore

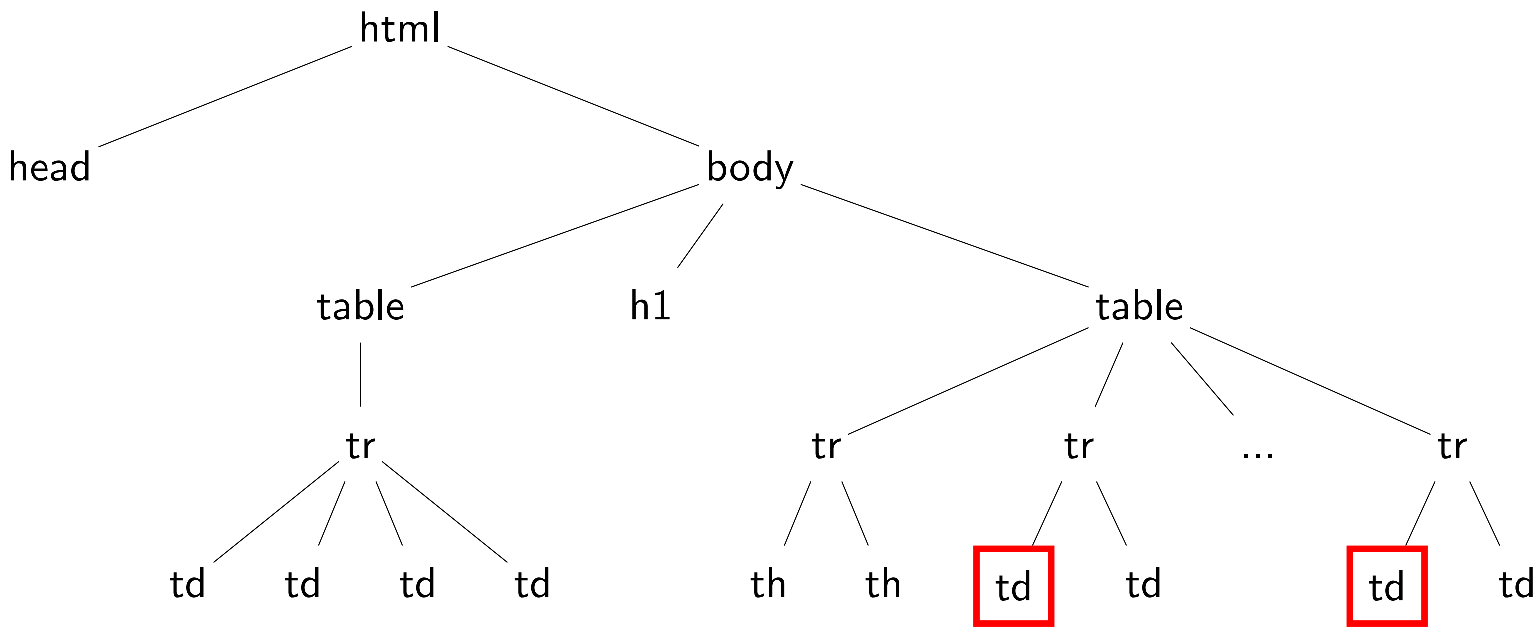
- web page w

Output:

- list of entities y

[Avalon Super Loop, Patapsco Valley State Park, ...]

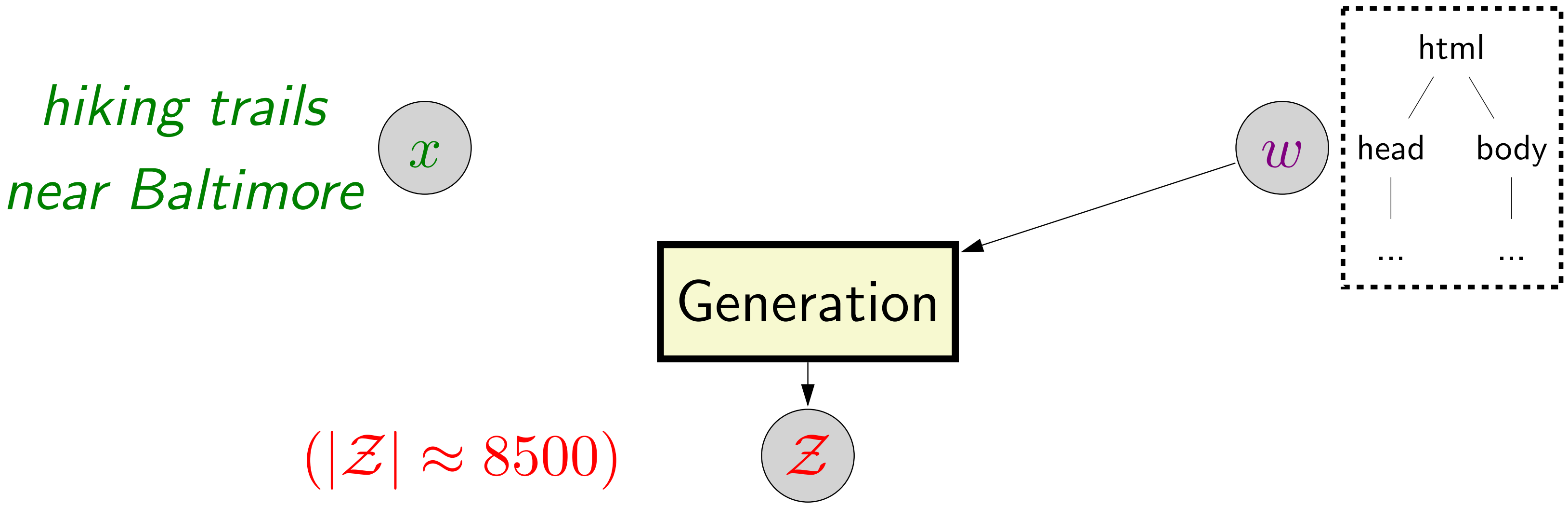
Logical forms: XPath expressions



$$z = \texttt{/html[1]/body[1]/table[2]/tr/td[1]}$$

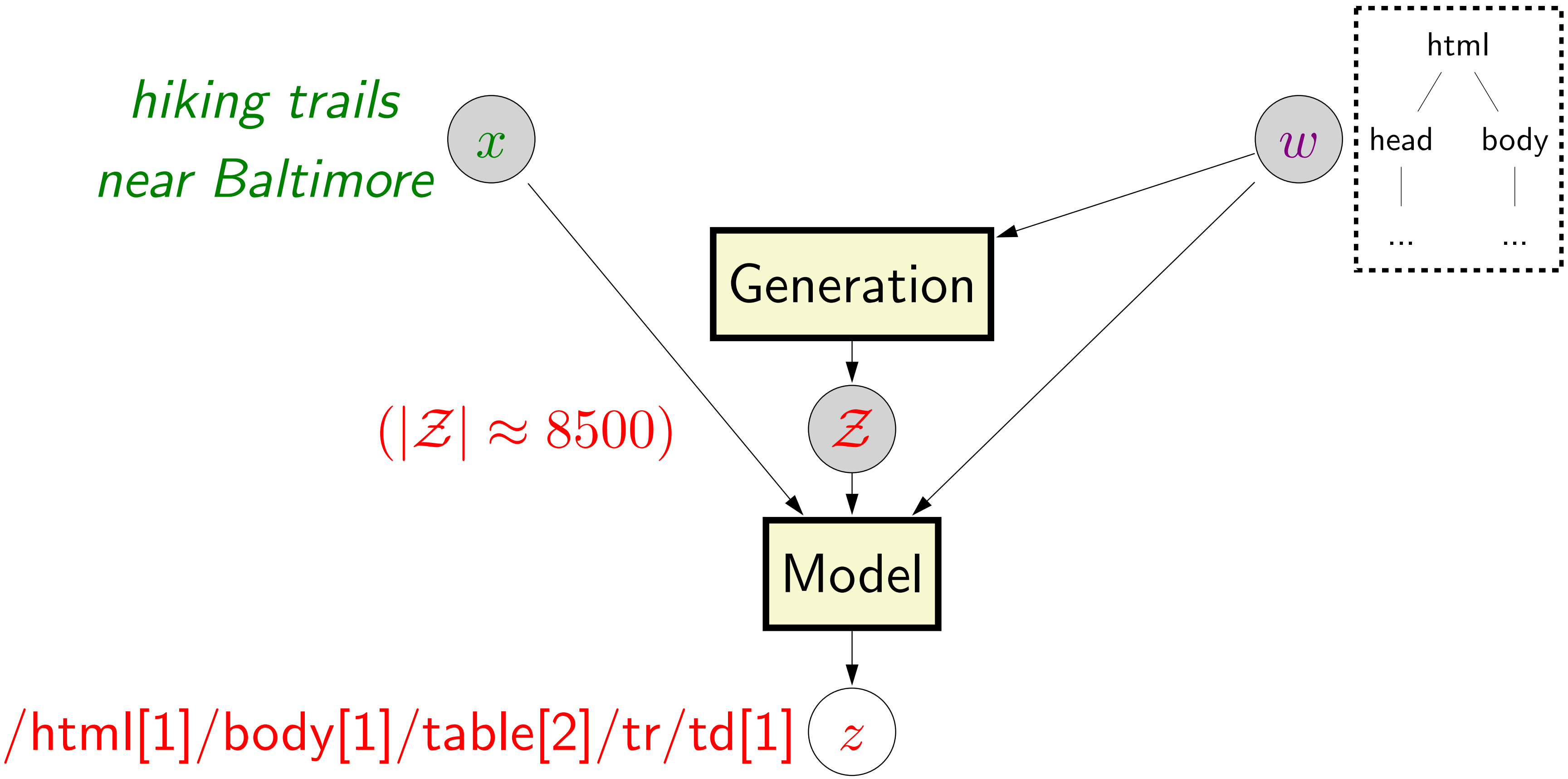
Framework

Framework



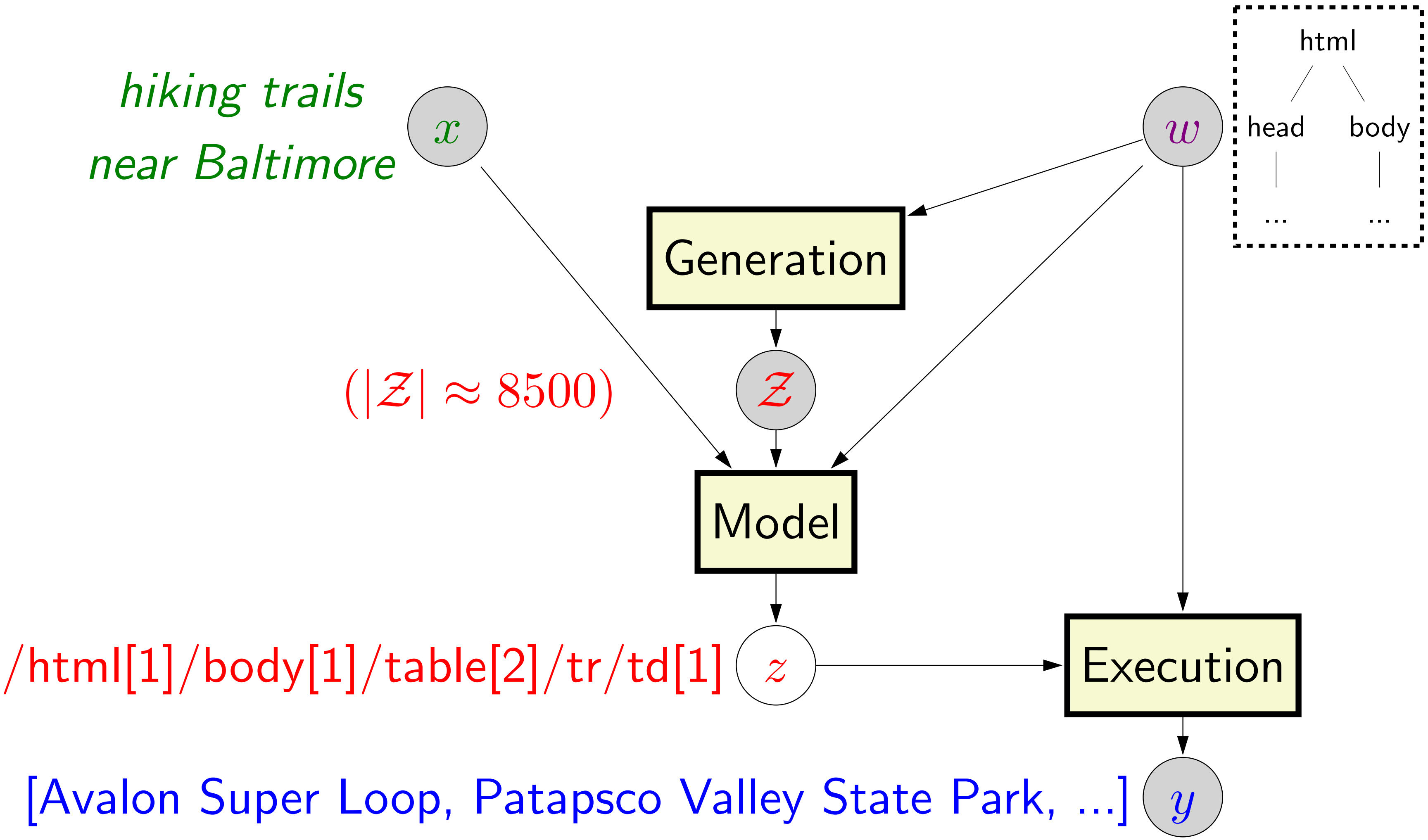
hiking trails near Baltimore
x
w
html
head
body
...
...
Generation
$\mathcal{Z}$
$(|\mathcal{Z}| \approx 8500)$

Framework



hiking trails
near Baltimore
x
w
html
head
body
...
...
Generation
(|𝒵| ≈ 8500)
𝒵
Model
/html[1]/body[1]/table[2]/tr/td[1]
z

Framework



hiking trails
near Baltimore
x
w
html
head
body
...
...
Generation
$(|\mathcal{Z}| \approx 8500)$
$\mathcal{Z}$
Model
/html[1]/body[1]/table[2]/tr/td[1]
z
Execution
[Avalon Super Loop, Patapsco Valley State Park, ...]
y

Features

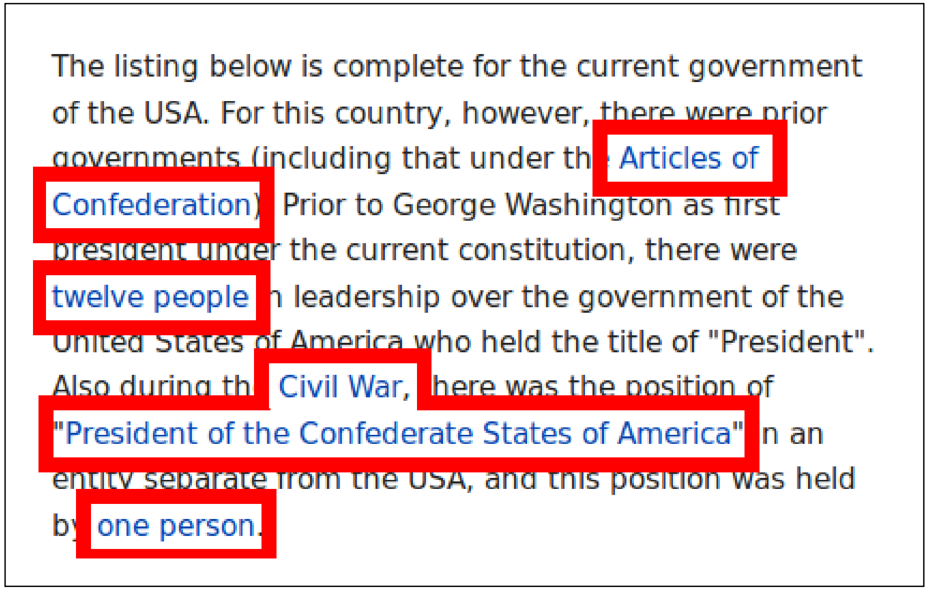
$$p_\theta(z \mid x, w) \propto \exp\{\theta^\top \phi(x, w, z)\}$$

Features

$$p_\theta(z \mid x, w) \propto \exp\{\theta^\top \phi(x, w, z)\}$$

Structural Features: captures context

$>$


The listing below is complete for the current government of the USA. For this country, however, there were prior governments (including that under the Articles of Confederation). Prior to George Washington as first president under the current constitution, there were twelve people in leadership over the government of the United States of America who held the title of "President". Also during the Civil War, there was the position of "President of the Confederate States of America" in an entity separate from the USA, and this position was held by one person.

Features

$$p_\theta(z \mid x, w) \propto \exp\{\theta^\top \phi(x, w, z)\}$$

Denotation Features: captures content

hiking trails near Baltimore		*hiking trails near Baltimore*
Avalon Super Loop		Home
Patapsco Valley State Park		About Baltimore Tour
Gunpowder Falls State Park	$>$	Pricing
Rachel Carson Conservation Park		Contact
Union Mills Hike		Online Support
...		...

Dataset

We created the OpenWeb dataset with diverse queries and web pages.

airlines of italy

natural causes of global warming

lsu football coaches

bf3 submachine guns

badminton tournaments

foods high in dha

technical colleges in south carolina

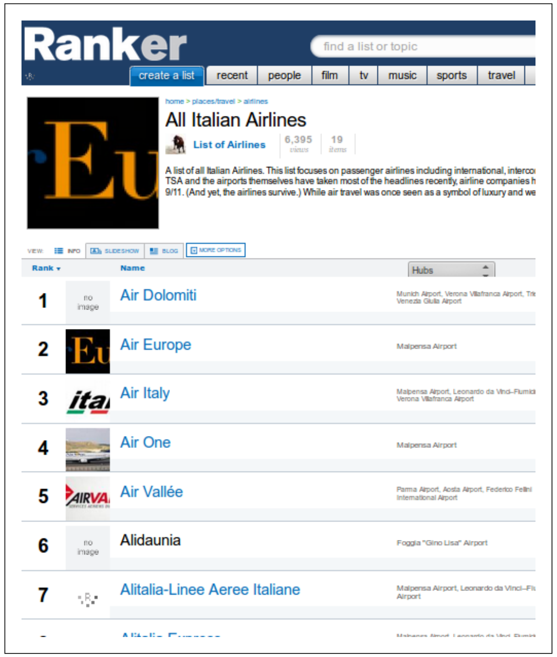
songs on glee season 5

singers who use auto tune

san francisco radio stations

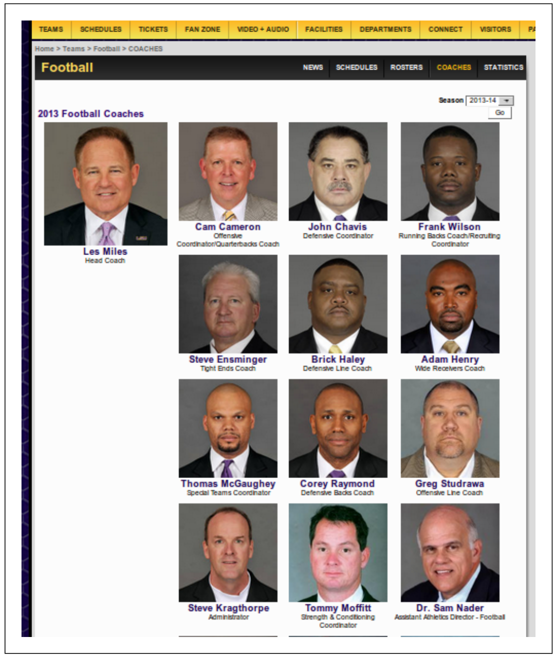
Dataset

We created the OpenWeb dataset with diverse queries and web pages.



airlines of italy

natural causes of global warming



lsu football coaches

Dataset Statistics

2773 examples

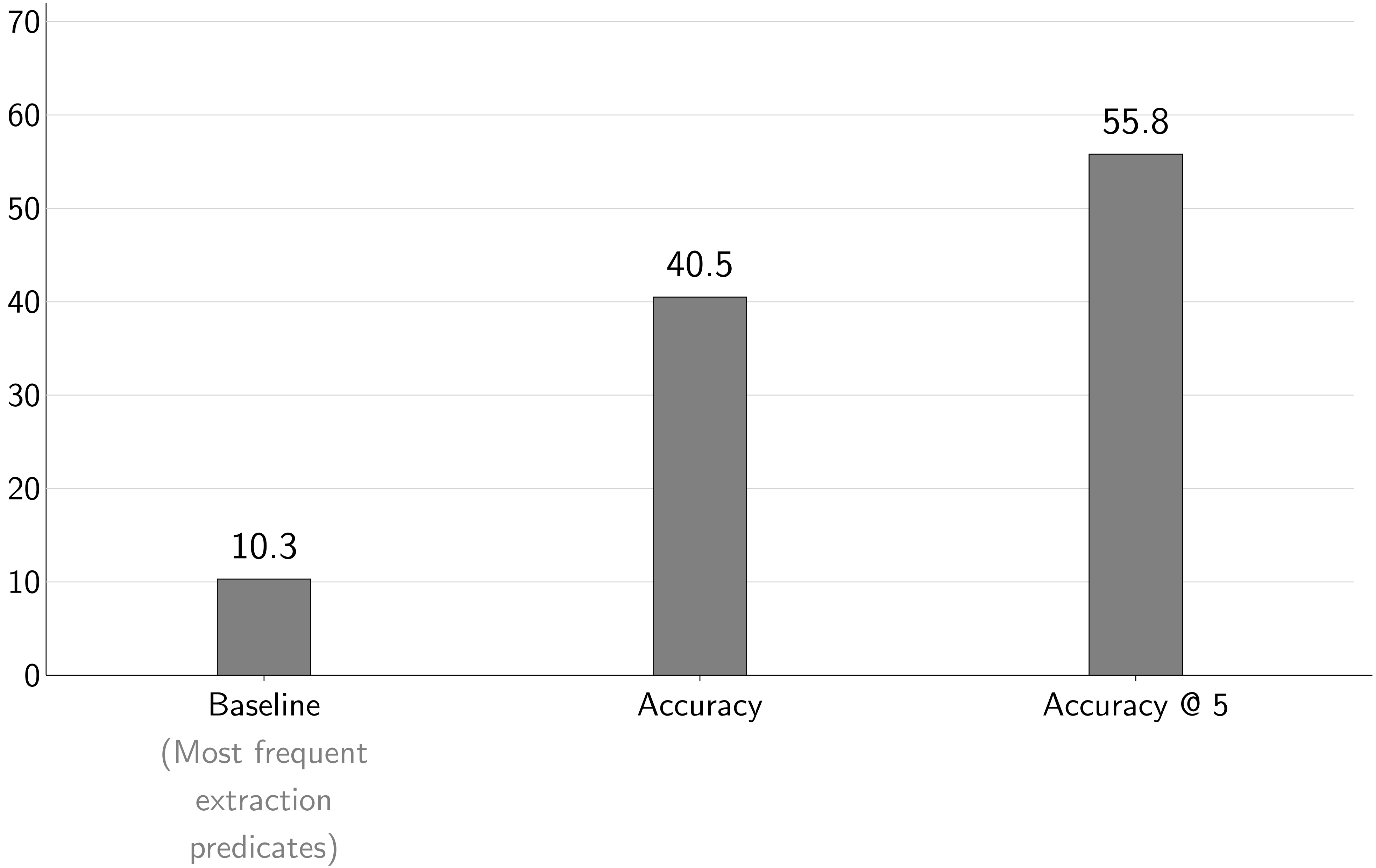
2269 unique queries

894 unique headwords $\leftarrow$ long tail!

1483 unique web domains $\leftarrow$ long tail!

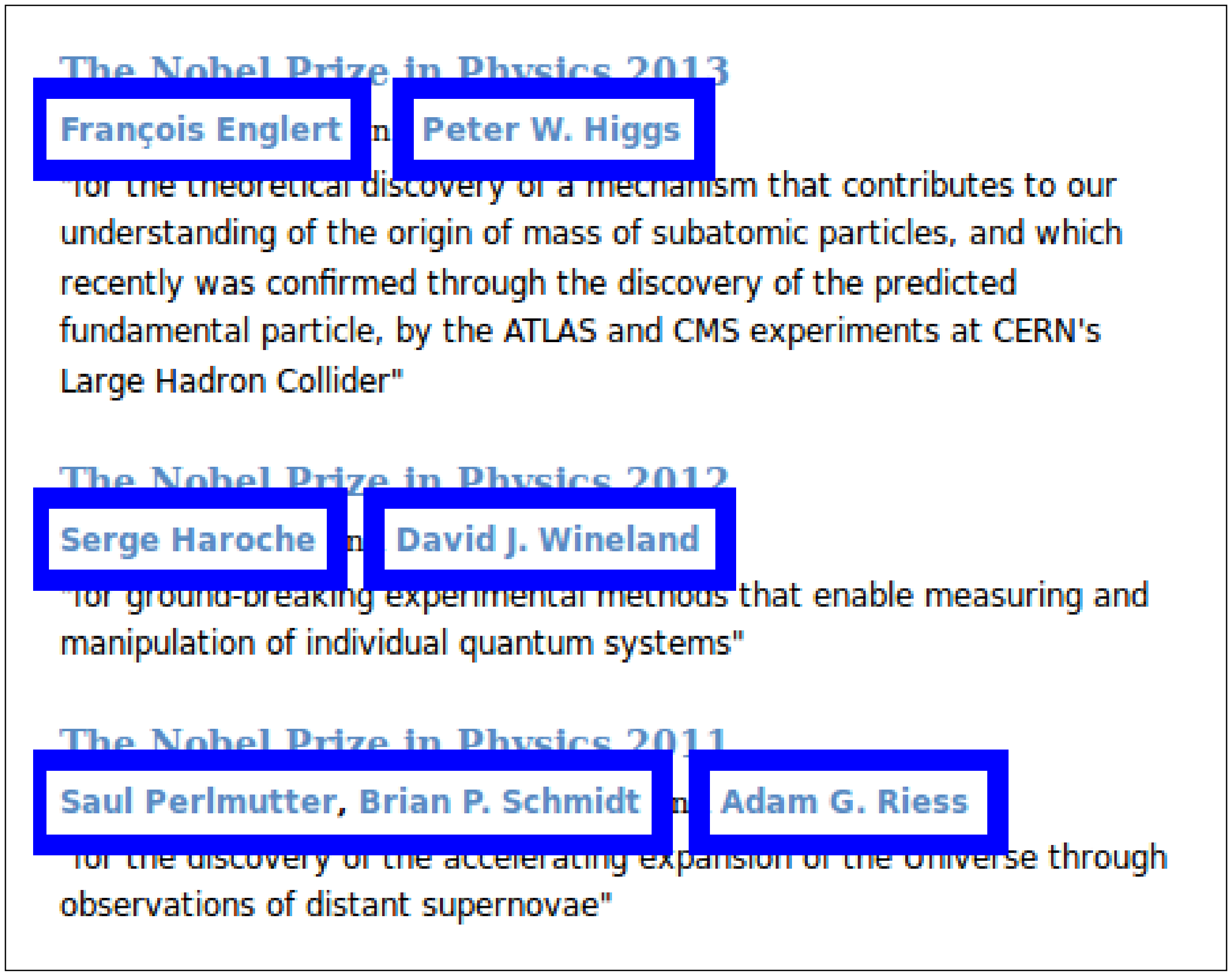
($\neq$ wrapper induction)

Main Results



Examples of Correct Predictions

Query: *nobel prize winners*



/html[1]/body/div/div[2]/div/div/div/h6/a/text

Examples of Coverage Errors

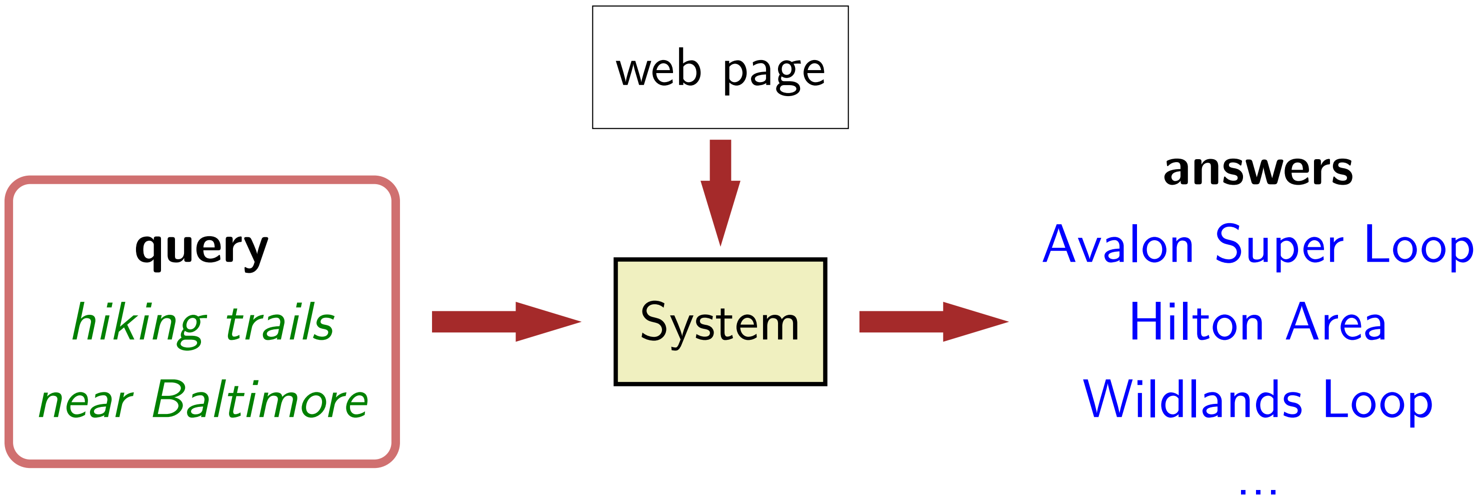
Query: *hedge funds in new york*

Rank	Firm	Headquarters
1	Bridgewater Associates	Westport, CT
2	Man Group	London
3	J.P. Morgan Asset Management	New York
4	Brevan Howard Asset Management	London
5	Och-Ziff Capital Management Group	New York
6	Paulson & Co.	New York
7	BlackRock Advisors	New York

`/html/body/div[3]/div[3]/div[4]/.../table/tbody/tr/td[2]/a`

Need compositionality!

Summary



A framework for extracting entities from a natural language query and a single web page

Outline

- A semantic parsing framework
- A closer look at the elements
  - Logical forms: lambda DCS
  - Lexical coverage
  - Grammar: building logical forms
  - Learning via bootstrapping
  - Leveraging denotations
  - Datasets/results
- Beyond Freebase
- **Final remarks**

Semantics

[utterance: user input]

semantic parsing

[intermediate semantic representations (text? logical forms? vectors?)]

execute

`[denotation: user output]`

Semantic representations as a means to an end

Long-term engineering vision

hiking trails near Baltimore

Long-term engineering vision

hiking trails near Baltimore under 5 miles

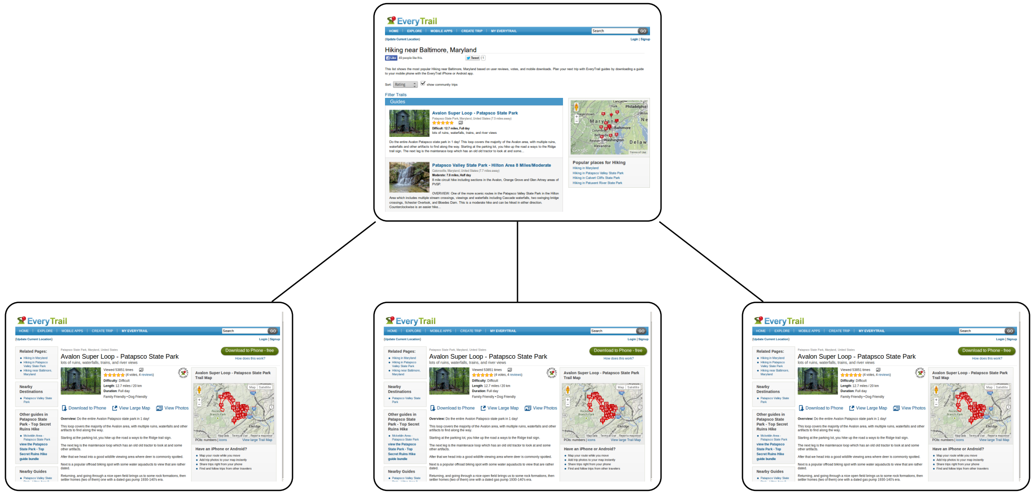
Long-term engineering vision

dog-friendly hiking trails near Baltimore under 5 miles

Long-term engineering vision

dog-friendly hiking trails near Baltimore under 5 miles

/body/div[2]/(CLICK)/body/div/(SEARCH)/body/div[3]



Conclusions

- **Learning and computation** are important to constrain representation

Conclusions

- **Learning and computation** are important to constrain representation

- Learning from denotations framework in which logical forms are **intermediate**

Conclusions

- **Learning and computation** are important to constrain representation

- Learning from denotations framework in which logical forms are **intermediate**

- Need **richer representations** that connect to real world with richer language

Code and data online

http://www-nlp.stanford.edu/software/sempre/

http://www-nlp.stanford.edu/software/web-entity-extractor-ACL2014/

Code and data online

http://www-nlp.stanford.edu/software/sempre/

http://www-nlp.stanford.edu/software/web-entity-extractor-ACL2014/

Collaborators

Jonathan Berant (post-doc)

Andrew Chou (masters)

Roy Frostig (Ph.D.)

Panupong Pasupat (Ph.D.)

Code and data online

http://www-nlp.stanford.edu/software/sempre/

http://www-nlp.stanford.edu/software/web-entity-extractor-ACL2014/

Collaborators

Jonathan Berant (post-doc)

Andrew Chou (masters)

Roy Frostig (Ph.D.)

Panupong Pasupat (Ph.D.)

Thank you!