Linked Data Query Processing

Tutorial at the 22nd International World Wide Web Conference (WWW 2013)

May 14, 2013

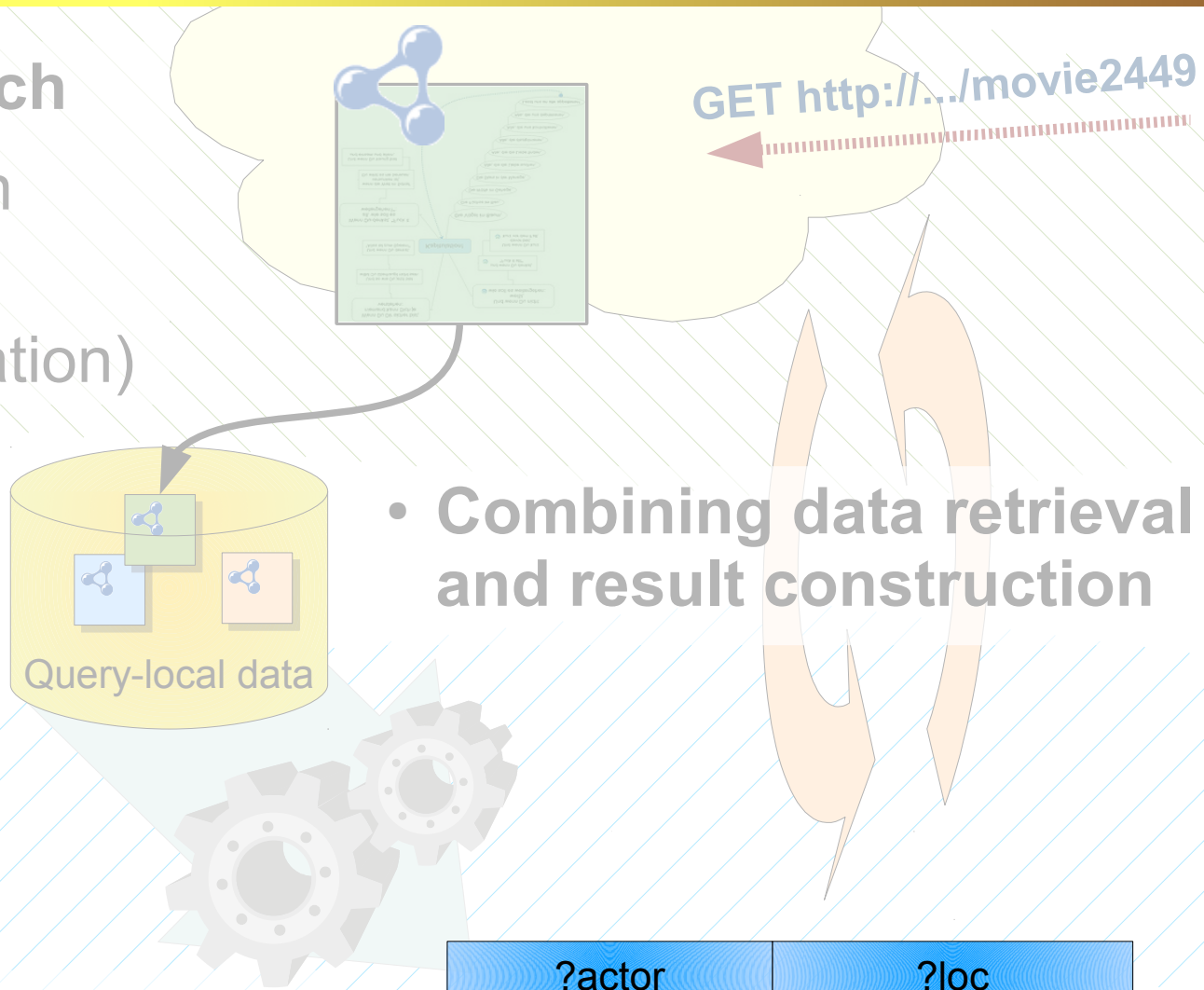
http://db.uwaterloo.ca/LDQTut2013/

2.
Theoretical Foundations

Olaf Hartig
University of Waterloo

“Ingredients” for LD Query Execution

- **Data retrieval approach**
  - Data source selection
  - Data source ranking (optional, for optimization)



- **Combining data retrieval and result construction**

- **Result construction approach**
  - i.e., query-local data processing

?actor	?loc
http://mdb.../Paul http://mdb.../Ric	http://geo.../Berlin http://geo.../Rome

Outline

- **Basics of the SPARQL Query Language**
- **Data Model for Linked Data Queries**
- **Full-Web Query Semantics**
  - Definition
  - Theoretical Properties
- **Reachability-Based Query Semantics**
  - Definition
  - Theoretical Properties

General Idea of SPARQL

- **SPARQL: Declarative query language for RDF data**
- **Main idea: pattern matching**
  - Query describes a pattern to be found in RDF data
  - Basically, query patterns consist of RDF triples with variables

Basic Query Pattern:

(*?p* , affiliated with , orgaX) ,

(*?p* , interested in , *?i*)

General Idea of SPARQL

- **SPARQL: Declarative query language for RDF data**
- **Main idea: pattern matching**
  - Query describes a pattern to be found in RDF data
  - Basically, query patterns consist of RDF triples with variables
  - Any part of the queried data that matches the query pattern contributes a solution to the query result

Basic Query Pattern:

(*?p* , affiliated with , orgaX) ,

(*?p* , interested in , *?i*)

Query Result:

?p	*?i*
alice	yoga
alice	tea

Data:

(alice , affiliated with , orgaX) ,

(bob , affiliated with , acme) ,

(alice , interested in , yoga) ,

(alice , interested in , tea) ,

(bob , interested in , tea) ,

Some Terminology

- **Any query result is represented by a set of valuations**

$$\Omega = \{ \mu_1 , \ldots , \mu_n \}$$

  - Valuation: mapping of query variables to RDF terms

$$\mu = \{ ?p \rightarrow \text{alice} , ?i \rightarrow \text{yoga} \}$$

  - Any valuation in a query result is a solution

?p	?i
alice	yoga
alice	tea

- **Two valuations μ_1 and μ_2 are compatible if they agree on the binding for any shared variable**
  - i.e., $\mu_1(?v) = \mu_2(?v)$ for all $?v \in \text{dom}(\mu_1) \cap \text{dom}(\mu_2)$

Query Patterns and their Semantics

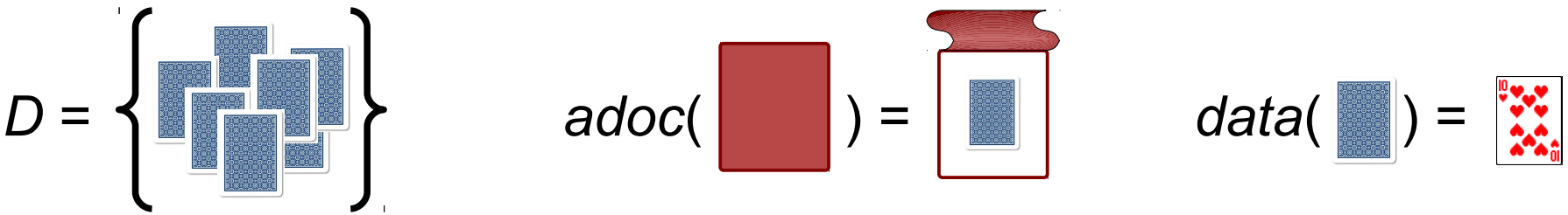
- **Triple pattern: An RDF triple that may have a variable as subject, predicate, or object, respectively**
  - $\text{eval}(\, tp,\, G\,) := \{\, \mu \mid \mu[tp] \in G \text{ and } \text{dom}(\mu) = \text{vars}(tp) \,\}$
- **Group graph pattern: Conjunction of two query patterns**
  - $\text{eval}(\, (P_1 \text{ AND } P_2),\, G\,) := \{\, \mu_1 \cup \mu_2 \mid \mu_1 \in \text{eval}(P_1, G) \text{ and } \mu_2 \in \text{eval}(P_2, G) \text{ and } \mu_1 \text{ and } \mu_2 \text{ are compatible} \,\}$
  - Associative and commutative
- **Basic graph pattern (BGP): A set of triple patterns**
  - $B = \{\, tp_1, \ldots, tp_n \,\}$ is equivalent to $(\, tp_1 \text{ AND } \ldots \text{ AND } tp_n \,)$
- **Filter, union, optional, etc.**

Outline

- **Basics of the SPARQL Query Language** √
- **Data Model for Linked Data Queries**
- **Full-Web Query Semantics**
  - Definition
  - Theoretical Properties
- **Reachability-Based Query Semantics**
  - Definition
  - Theoretical Properties

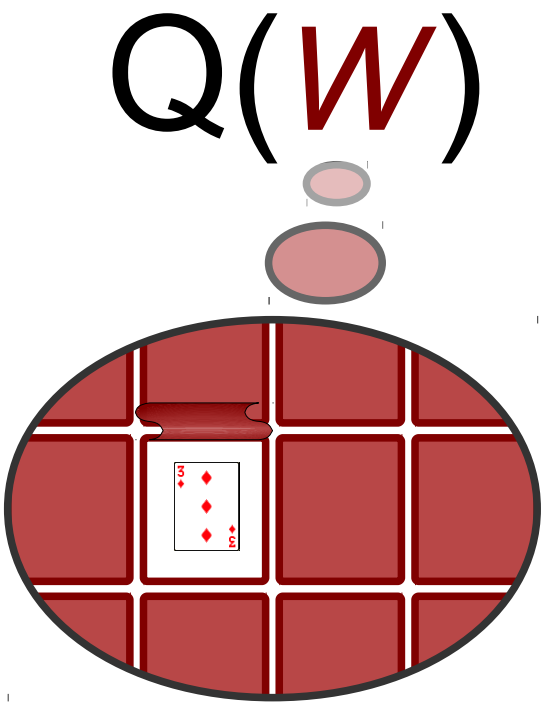
Data Model

- **Static view (i.e., no changes during computation)**
- **Web of Linked Data:** $W = (D, adoc, data)$
  - D – set of symbols that represent Web documents from which Linked Data can be extracted
  - $adoc$ – partial mapping from URIs to D
  - $data$ – total mapping from D to finite sets of RDF triples



- **All data in *W*:** $\text{AllData}(W) := \bigcup_{d \in D} data(d)$

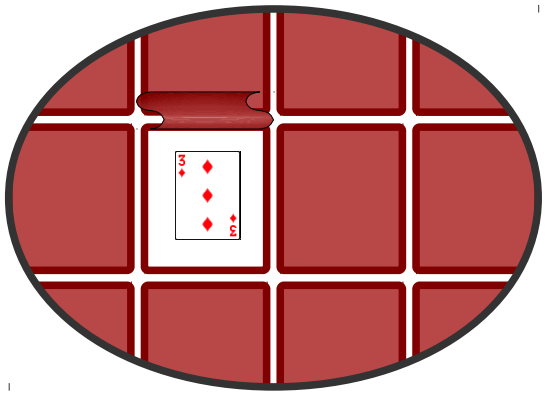
Linked Data Query



SPARQL-Based Linked Data Query

(*?p* , affiliated with , orgaX) ,
(*?p* , interested in , *?i*)

$$Q^{P}(W) = \{ \mu_1 , \mu_2 , \ldots \}$$



?p	*?i*
alice	yoga
alice	tea

Outline

- **Basics of the SPARQL Query Language** √
- **Data Model for Linked Data Queries** √
- **Full-Web Query Semantics**
  - Definition
  - Theoretical Properties
- **Reachability-Based Query Semantics**
  - Definition
  - Theoretical Properties

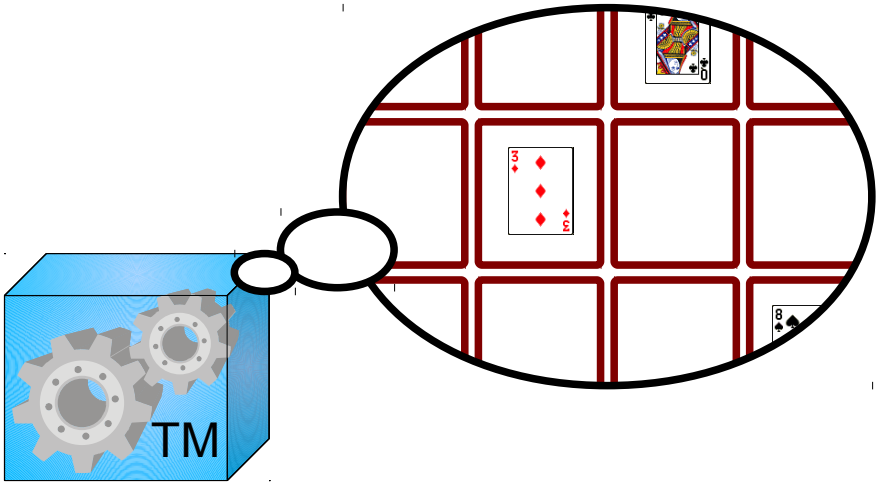
Full-Web Semantics

$$Q^{P}(W) := \text{eval}(P, \text{AllData}(W))$$

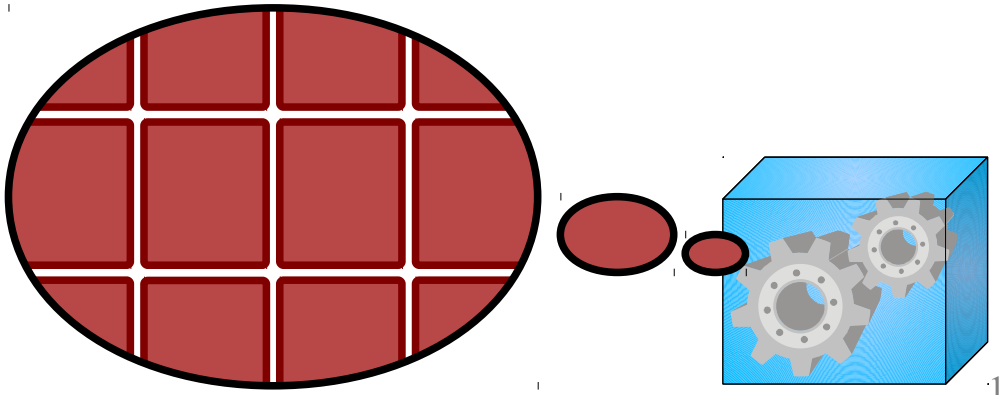
Computability

$$Q^{P}(W) := \mathrm{eval}(P, \mathrm{AllData}(W))$$

- **(Ordinary) Turing machines are unsuitable:**
  - Limited data access capabilities not properly captured

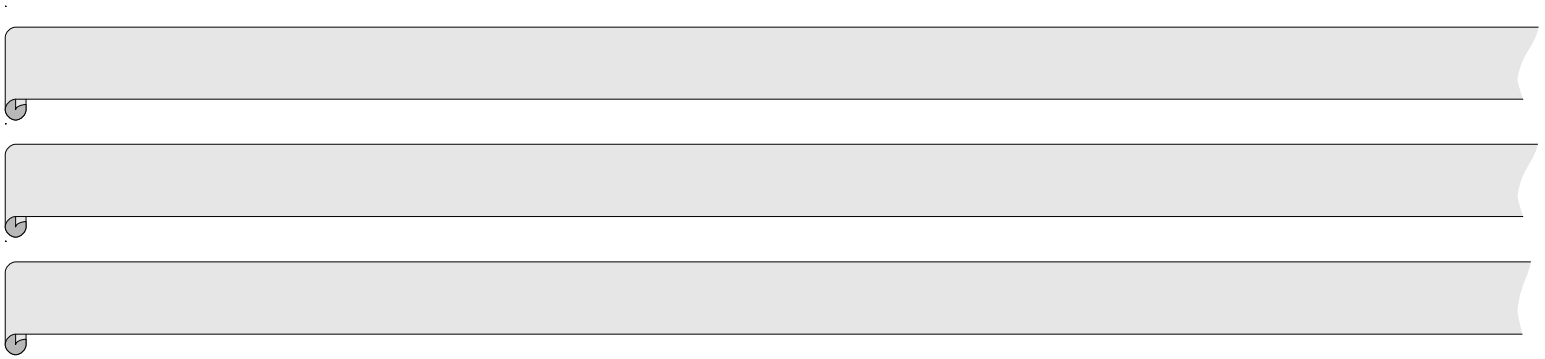


- **Web machines**
  - Abiteboul and Vianu, 1997
  - Mendelzon and Milo, 1997



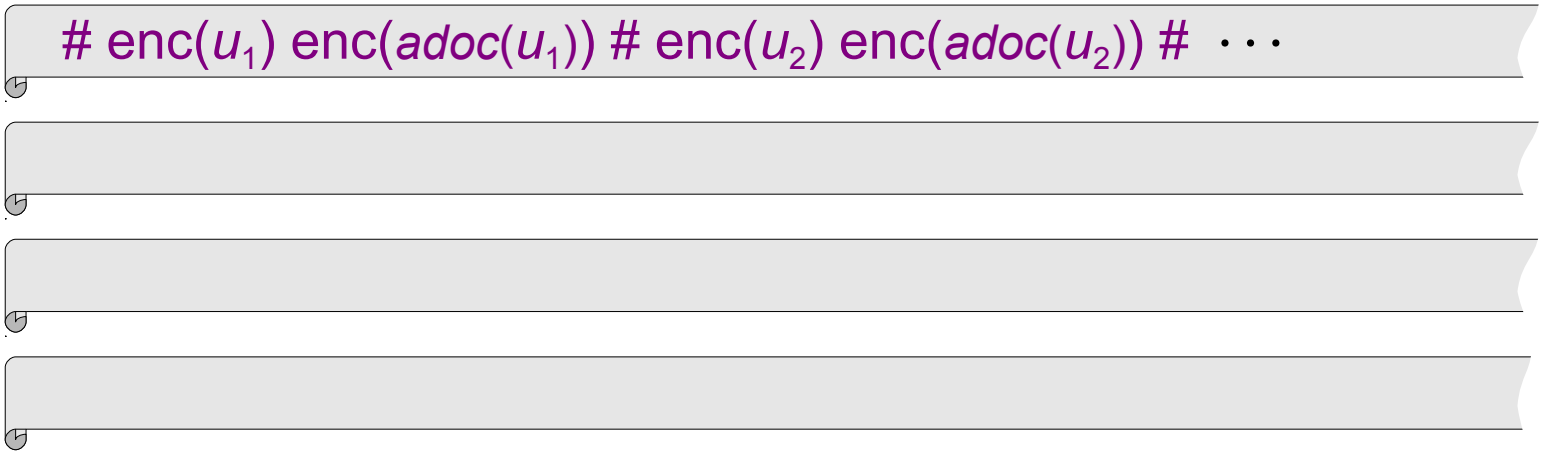
LD Machine

- **Multi-tape Turing machine**
  - Input
  - Work
  - Output



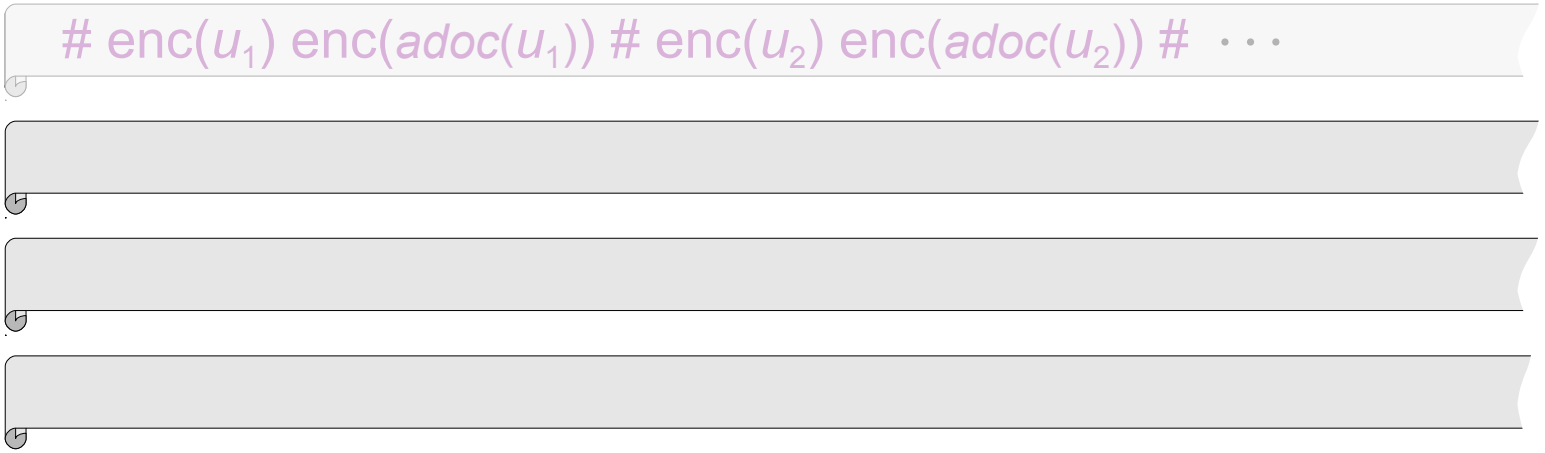
LD Machine

- **Multi-tape Turing machine**
  - Web Input
  - Input
  - Work
  - Output



LD Machine

- **Multi-tape Turing machine**
  - Web Input
  - Input
  - Work
  - Output

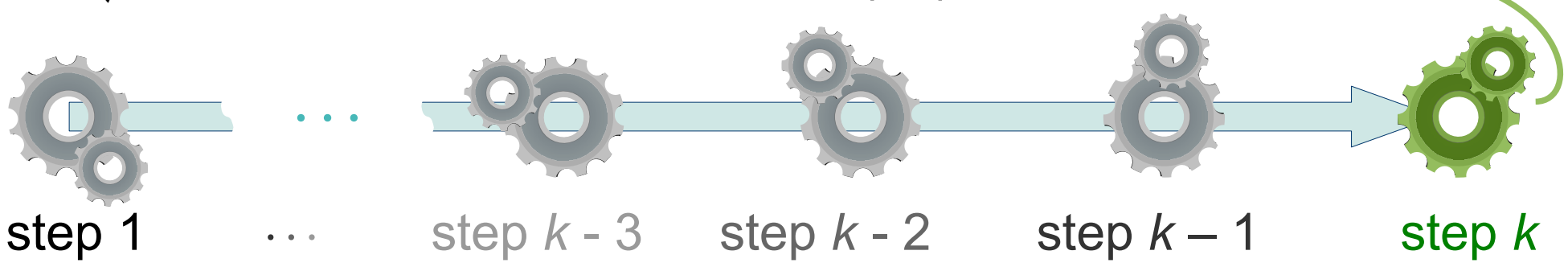


- **Access to** Web Input **is restricted**
  - Only by performing
    a particular procedure
    in a particular state

Finitely Computable LD Queries

- Web Input: # enc(u_1) enc($adoc(u_1)$) # enc(u_2) enc($adoc(u_2)$) # · · ·
- Input
- Work
- Output: # enc(μ_1) # enc(μ_2) # · · · # enc(μ_n) #

- **For Q exists an LD machine M_Q such that for any W holds:**
  - M_Q halts after a finite number of computation steps, and
  - M_Q outputs the complete result $Q(W)$



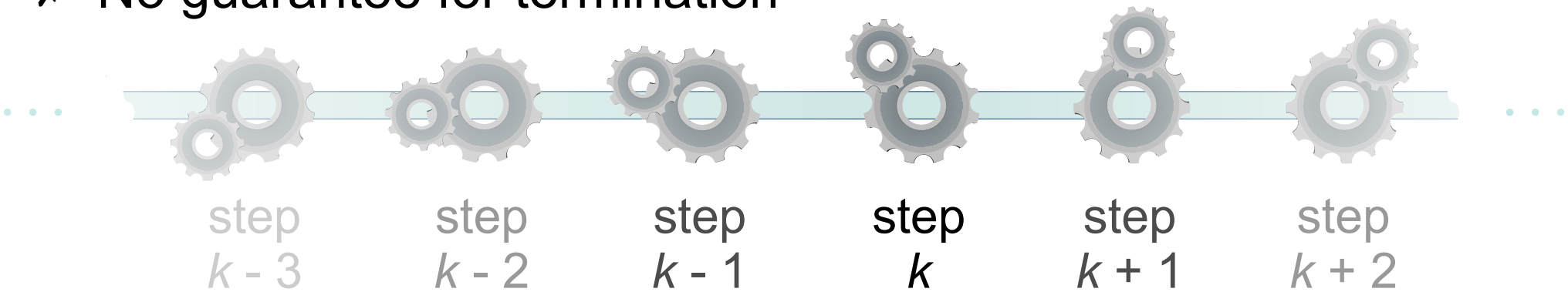
Eventually Computable LD Queries

- Web Input: # enc(u_1) enc($adoc(u_1)$) # enc(u_2) enc($adoc(u_2)$) # · · ·
- Input
- Work
- Output: # enc(μ_1) # enc(μ_2)

- **For Q exists an LD machine M_Q such that for any W holds:**

1. Output always encodes a subset of query result $Q(W)$, and
2. Each $\mu \in Q(W)$ eventually appears on the output

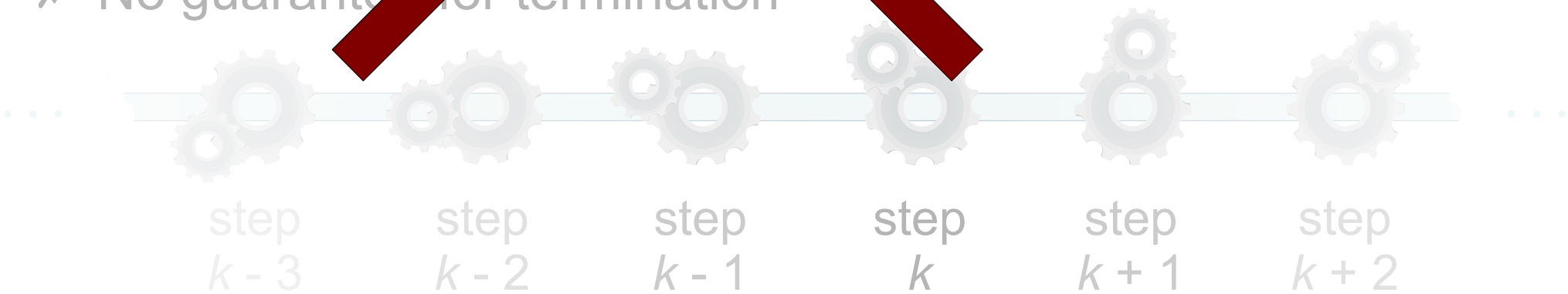
✗ No guarantee for termination



Not (LD Machine) Computable

- Web Input: # enc(u_1) enc($adoc(u_1)$) # enc(u_2) enc($adoc(u_2)$) # · · ·
- Input
- Work
- Output: # enc(μ_1) # enc(μ_2)

- **For Q exists an LD machine M_Q such that for any W holds:**
  1. Output always encodes a subset of query result $Q(W)$, and
  2. Each $\mu \in Q(W)$ eventually appears on the output
  - ✗ No guarantee for termination



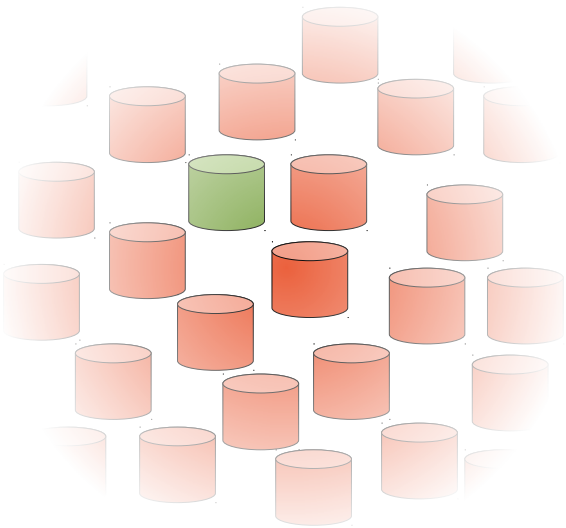
Computability (Full-Web Semantics)

$$Q^{P}(W) := \text{eval}(P, \text{AllData}(W))$$

Theorem [Har12]: If a satisfiable $\text{SPARQL}_{\text{LD}}$ query Q^{P} under full-Web semantics is *monotonic*, then it is eventually computable (but not finitely computable); otherwise, it is not even eventually computable.

- **Main reason: Set of all URIs is infinite (but countable)**

Satisfiability and Monotonicity

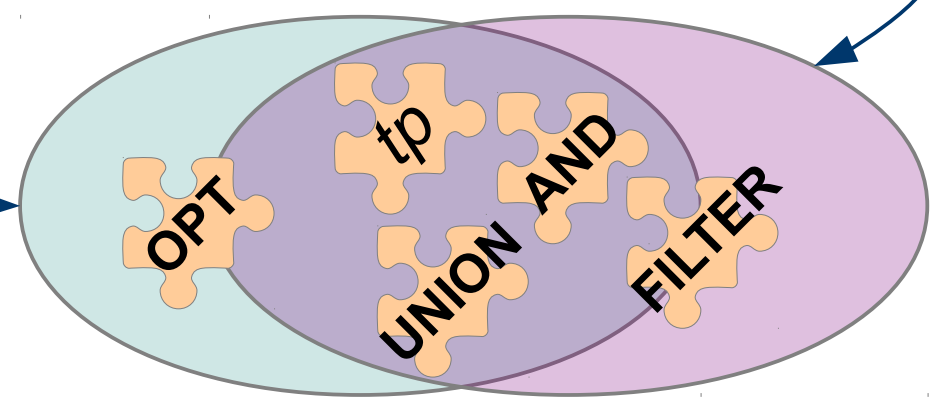


Satisfiability:

$Q(\text{▮}) = \{ \ldots, r_i, \ldots \} \neq \emptyset$

Monotonicity:

$\text{▮} \implies Q(\text{▮}) \subseteq Q(\text{▮})$



Unfortunately:
Satisfiability and monotonicity of SPARQL query patterns is undecidable

Proposition [Har12]: Let Q^P be a $\text{SPARQL}_{\text{LD}}$ query under full-Web semantics.

- Q^P is satisfiable $\iff$ its query pattern P is satisfiable
- Q^P is monotonic $\iff$ its query pattern P is monotonic

Computability (Full-Web Semantics)

$$Q^{P}(W) := \text{eval}(P, \text{AllData}(W))$$

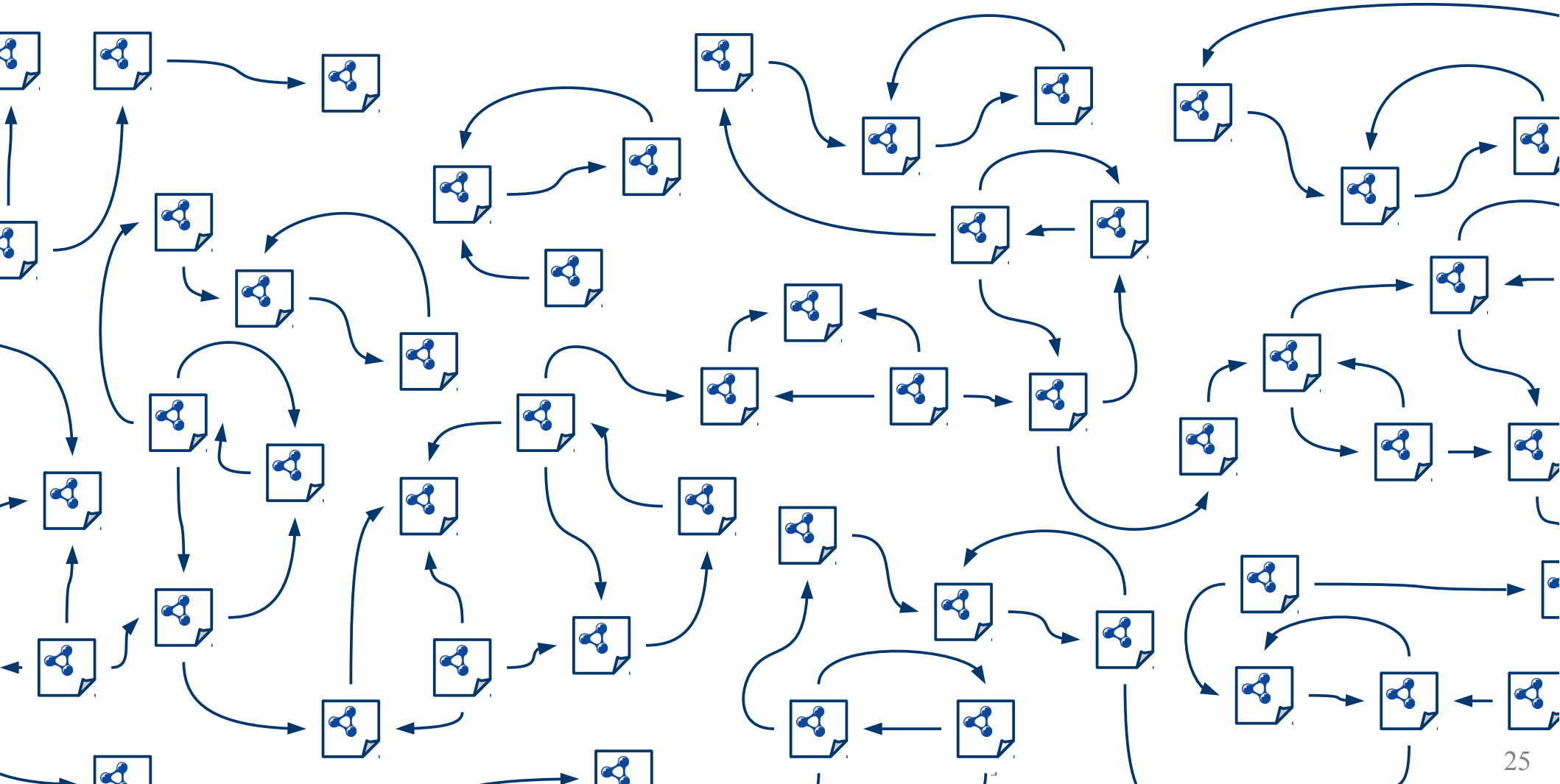
Theorem [Har12]: If a satisfiable $\text{SPARQL}_{\text{LD}}$ query Q^{P} under full-Web semantics is *monotonic*, then it is eventually computable (but not finitely computable); otherwise, it is not even eventually computable.

- **Main reason: Set of all URIs is infinite (but countable)**

Outline

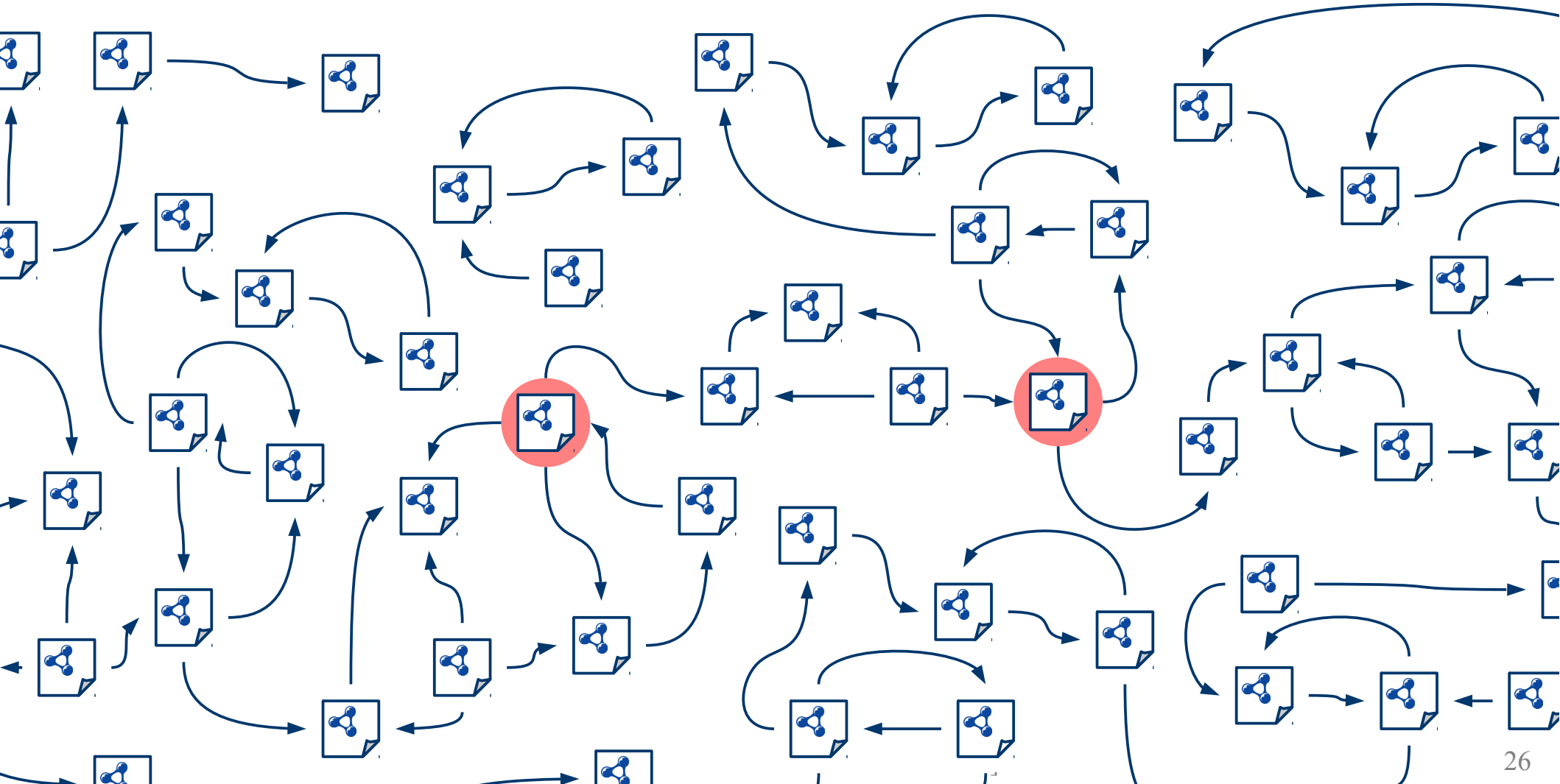
- **Basics of the SPARQL Query Language** √
- **Data Model for Linked Data Queries** √
- **Full-Web Query Semantics**
  - Definition
  - Theoretical Properties √
- **Reachability-Based Query Semantics**
  - Definition
  - Theoretical Properties

Reachability-Based Query Semantics



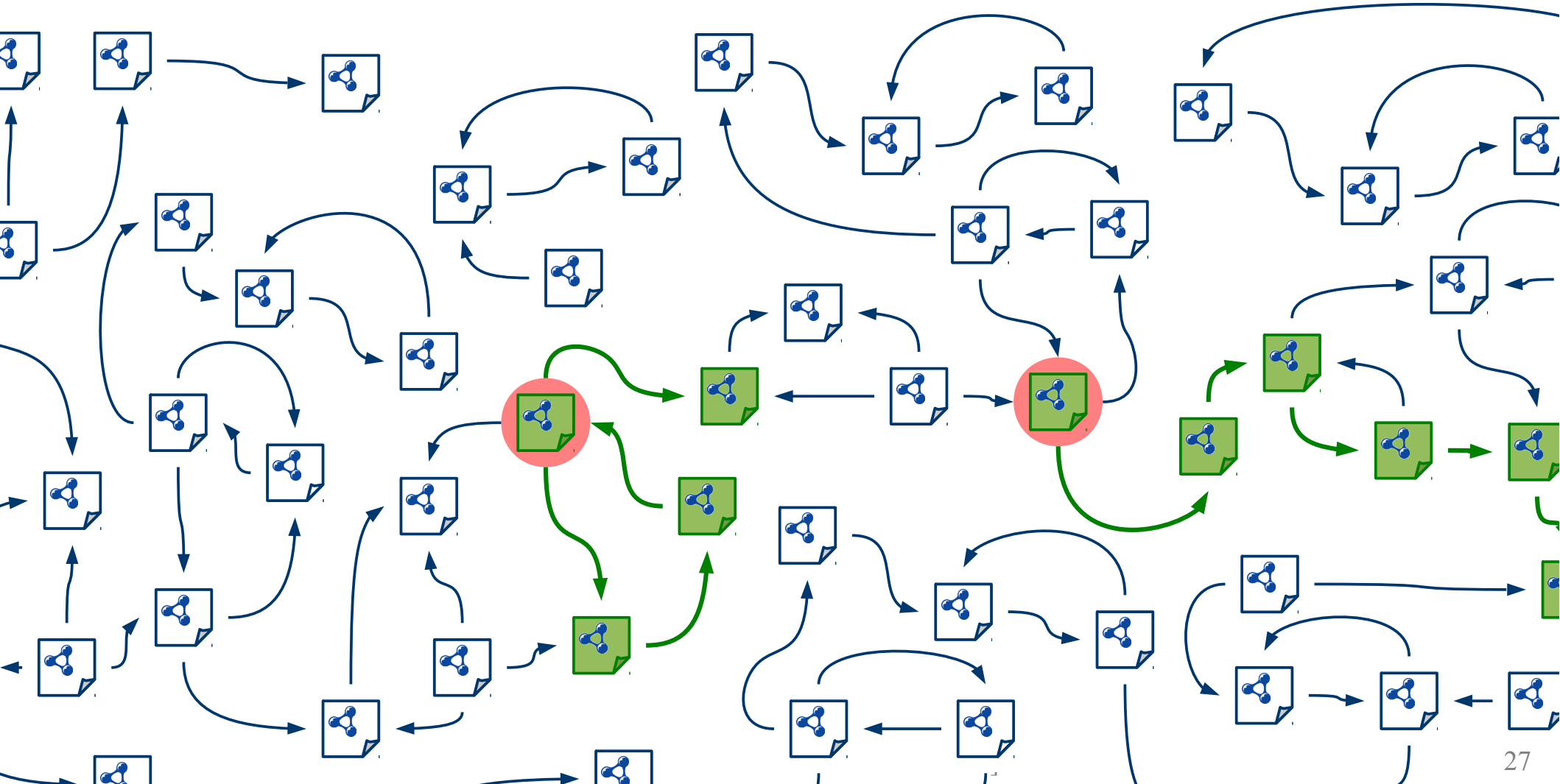
Reachability-Based Query Semantics

- **Seed URIs** $S \subset \mathcal{U}$



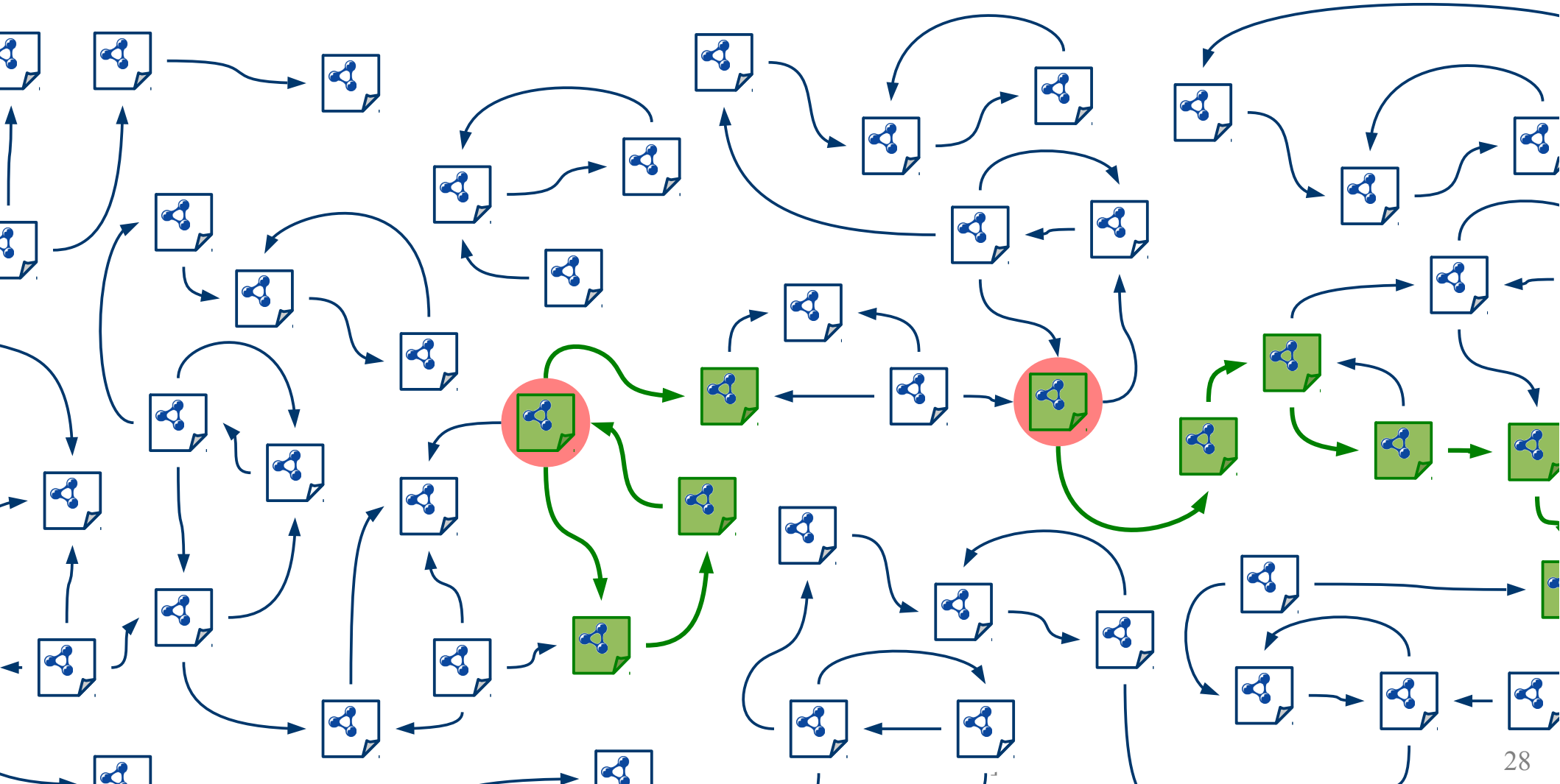
Reachability-Based Query Semantics

- **Reachability criterion *c***
  - (Turing) computable function *c*: $\mathcal{T} \times \mathcal{U} \times \mathcal{P} \rightarrow$ {true,false}
- **Seed URIs *S*** $\subset \mathcal{U}$



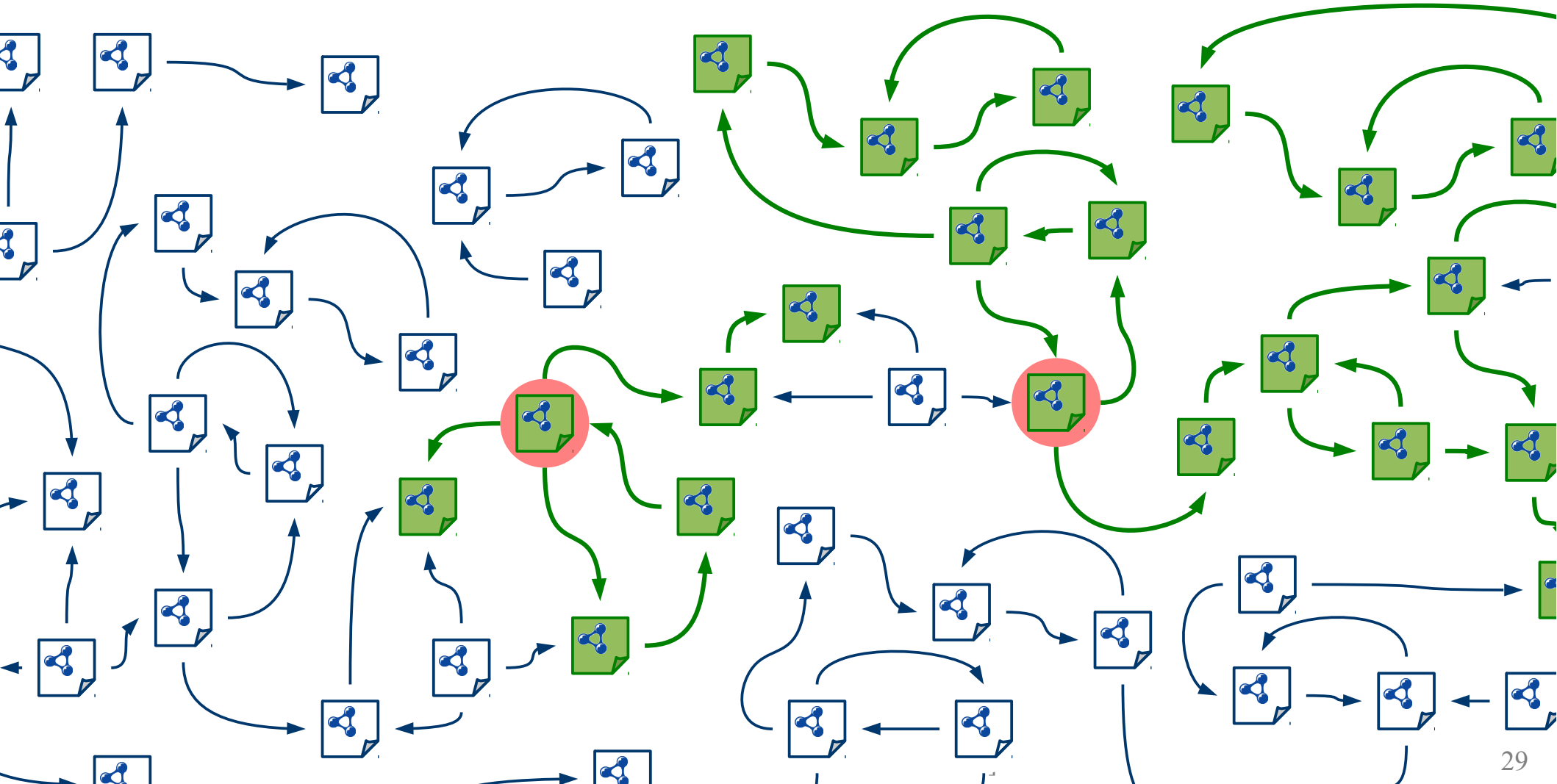
Reachability-Based Query Semantics

$$Q_c^{P,S}(W) := \mathrm{eval}(P, \mathrm{AllData}(W^*))$$



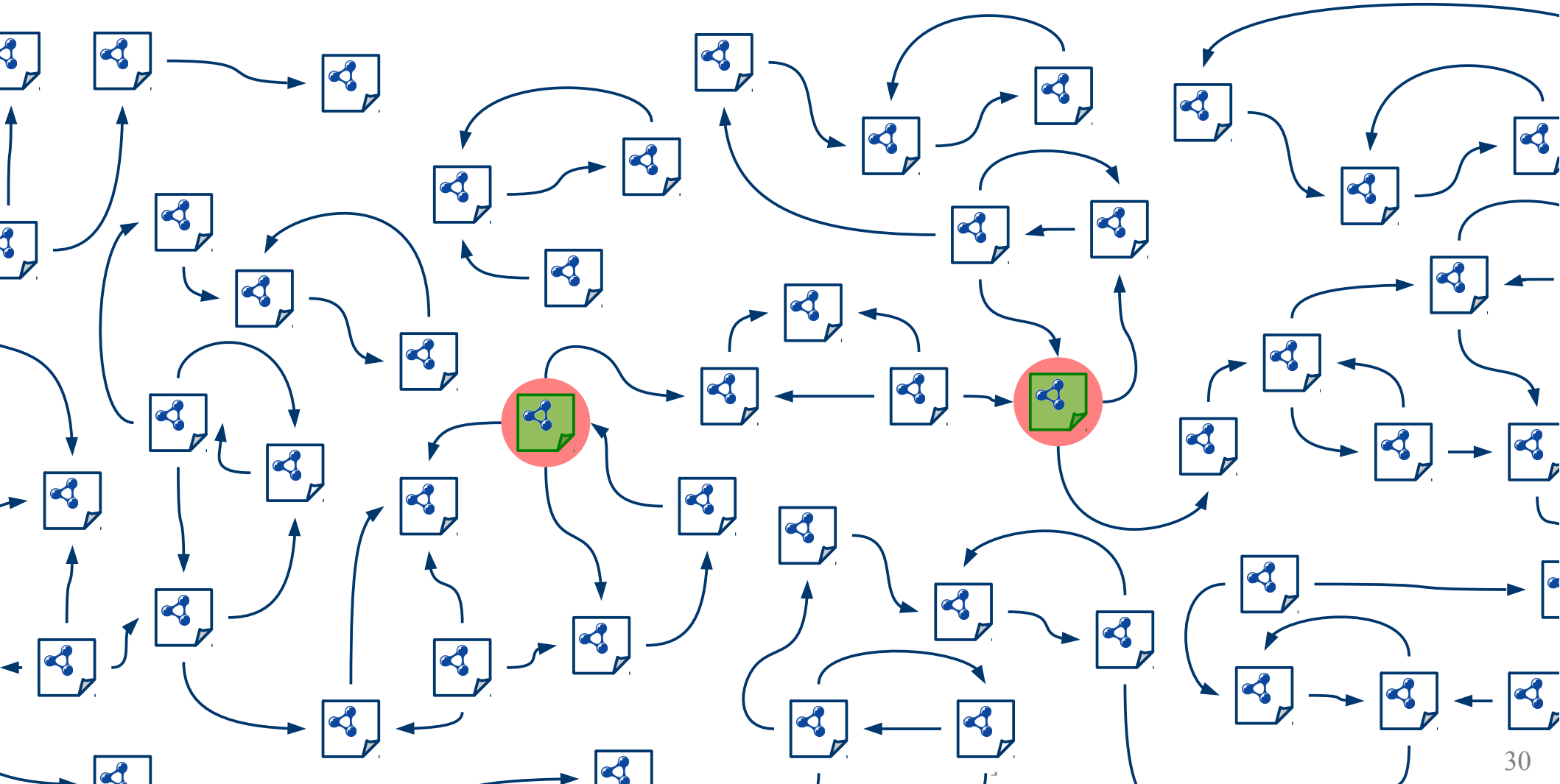
(Reachability-Based) c_{All}-Semantics

$$Q^{P,S}_{c_{\mathsf{All}}}(W) := \mathrm{eval}(P, \mathrm{AllData}(W^*))$$



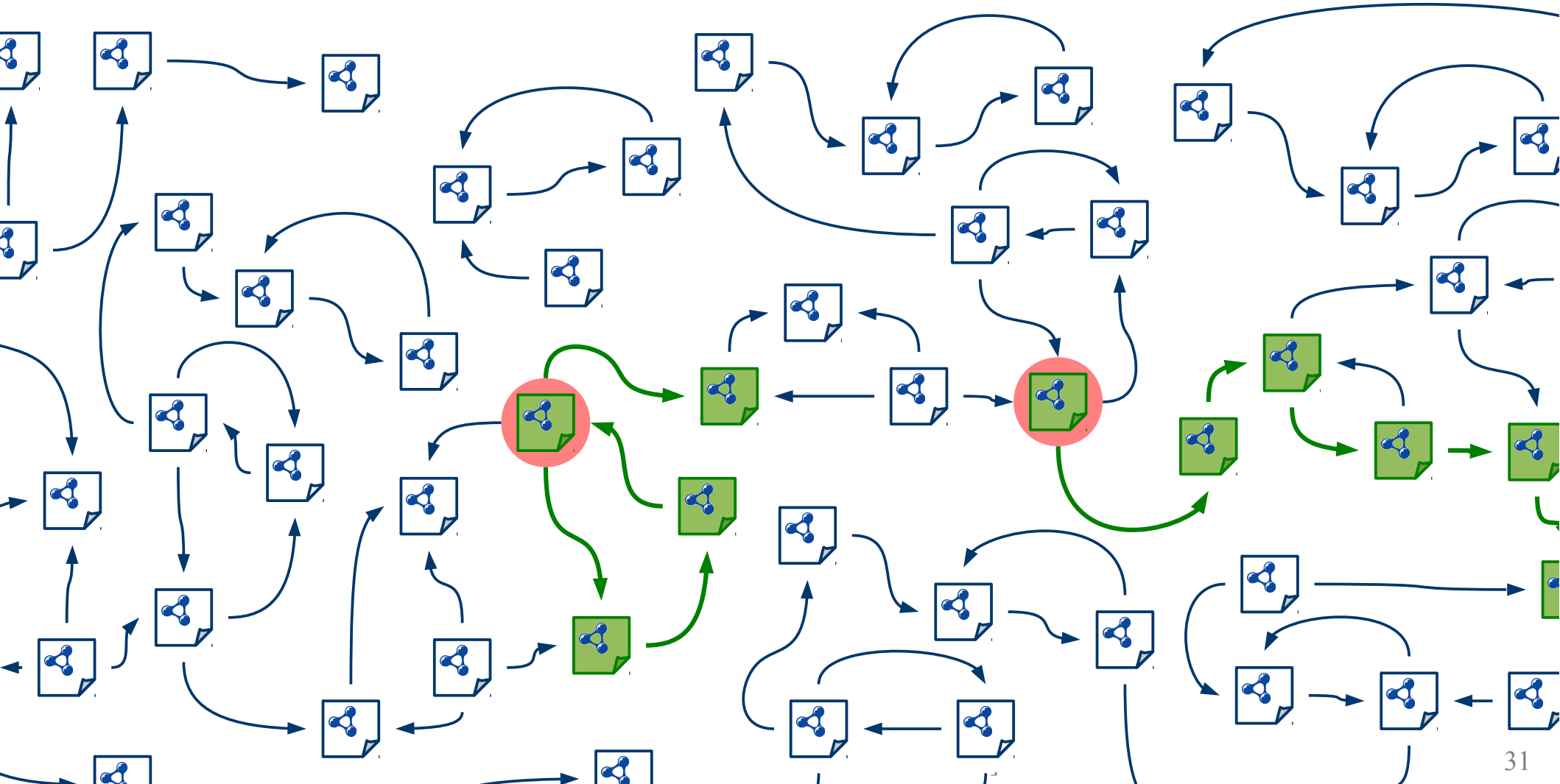
(Reachability-Based) c_{None}-Semantics

$$Q^{P,S}_{c_{\text{None}}}(W) := \text{eval}(P, \text{AllData}(W^*))$$

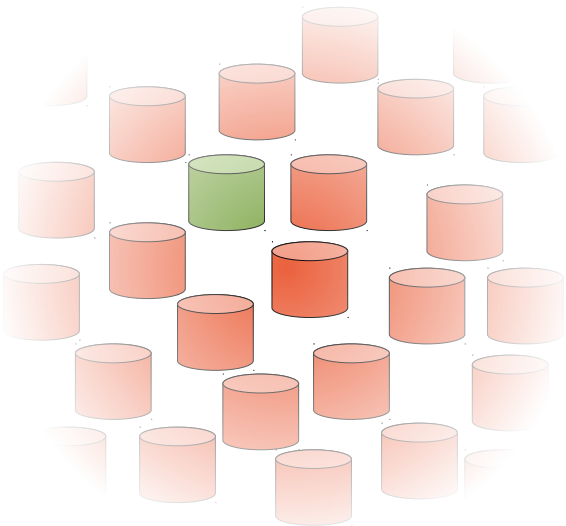


(Reachability-Based) c_{Match}-Semantics

$$Q^{P,S}_{c_{\text{Match}}}(W) := \text{eval}(P, \text{AllData}(W^*))$$



Satisfiability and Monotonicity



Satisfiability:

$Q(\;) = \{ \ldots, r_i, \ldots \} \neq \emptyset$

Monotonicity:

$\;\Longrightarrow\; Q(\;) \subseteq Q(\;)$

Proposition [Har12]: Let c be a reachability criterion and $Q^{P}_{c}{}^{S}$ be a $\text{SPARQL}_{\text{LD}}$ query under c-semantics.

- $Q^{P}_{c}{}^{S}$ is satisfiable $\Longleftrightarrow$ its SPARQL expression P is satisfiable
- $Q^{P}_{c}{}^{S}$ is monotonic $\Rightarrow$ its SPARQL expression P is monotonic

Counterexample: Let $Q^{P}_{c}{}^{S}$ be a $\text{SPARQL}_{\text{LD}}$ query under c-semantics. If $c = c_{\text{None}}$ and $|S| = 1$, then $Q^{P}_{c}{}^{S}$ is monotonic.

LD Machine-Based Computability

- **If we restrict our data model such that Webs of Linked Data are finite (i.e., consist of a finite number of documents), then:**

> **Theorem [Har12]:** For any reachability criterion *c*, any SPARQL based Linked Data query under *c*-semantics is finitely computable.

For results w.r.t. the unrestricted data model (i.e., Webs of Linked Data that may be infinitely large), refer to [Har12].

Outline

- **Basics of the SPARQL Query Language** √
- **Data Model for Linked Data Queries** √
- **Full-Web Query Semantics**
  - Definition
  - Theoretical Properties √
- **Reachability-Based Query Semantics**
  - Definition
  - Theoretical Properties √

Next part: 3. Source Selection Strategies ...

These slides have been created by
Olaf Hartig
for the
WWW 2013 tutorial on
Link Data Query Processing

Tutorial Website: http://db.uwaterloo.ca/LDQTut2013/