



# Will **knowledge graphs** save us from the mess of modern data practice?

**Dr. Ora Lassila**

Principal Technologist  
Amazon **Neptune**



Will **knowledge graphs** save us from  
the mess of modern data practice?

Will **knowledge graphs** save us from  
the mess of modern data practice?

**Yes**

Will **knowledge graphs** save us from  
the mess of modern data practice?

**Well, it depends...**

# Game plan for my talk

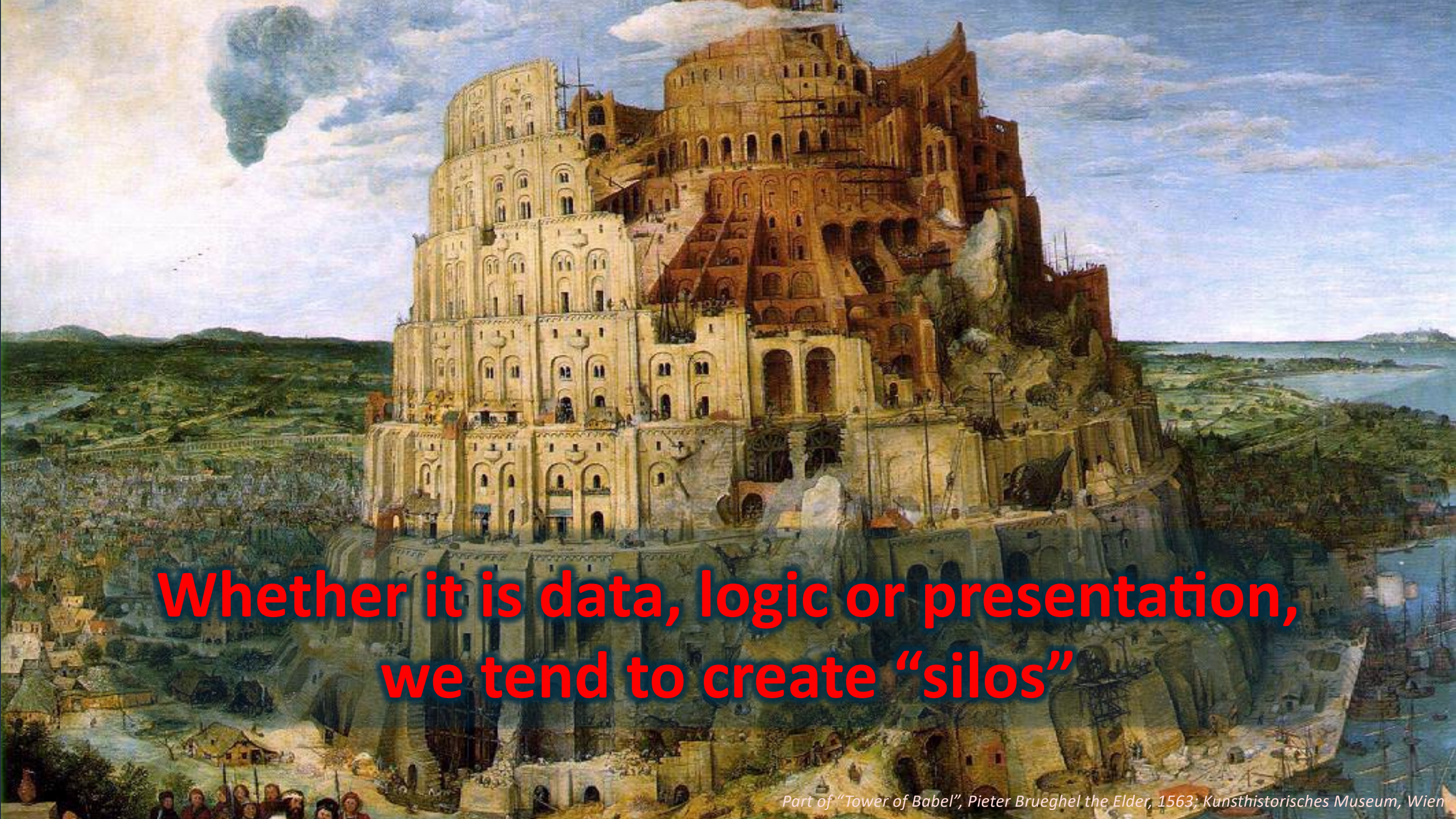
1. Problem (as I see it)
2. Solution (as I imagine it)

# Silos everywhere!

1. Each data model is a new “vocabulary”
2. Each API is also a new vocabulary (its own semantics, in a “black box”)
3. Each presentation (of data) speaks its own, new “language”

We spend a lot of time just **converting and “massaging” data...**

**Whether it is data, logic or presentation,  
we tend to create “silos”**



**Whether it is data, logic or presentation,  
we tend to create “silos”**

*Part of “Tower of Babel”, Pieter Bruegel the Elder, 1563; Kunsthistorisches Museum, Wien*

# Main challenges in modern data practice (as I see it)

Software

Silos

Semantics

These three are not unrelated...

# “Modern Data Stack”...?



*“**stack**” is often just a euphemism for “we have a lot of different tools and this is our attempt to make us look more organized”*

# Challenge: software

The whole industry is very “tool-oriented”

- constant stream of new tools being offered
- (the tools support a vibrant service economy...)
- most of these tools either fix some small problem (“band-aid”) or are yet another layer on top of something that is fundamentally broken

Also APIs...

# Challenge: software

Adding more software does not help with the already daunting complexity

More software is not the solution, we need to find ways to solve this with less

Insanity: doing the same thing over and over and expecting a different result

*(probably mis-attributed to Albert Einstein)*

# Challenge: silos

Data locked behind an API or (even worse) an “app”

- you have to **play by their rules** to get access

Data integration is difficult and leads to bespoke solutions

Elimination of silos would be nice, but is not only a technical problem

- attempting data integration across organizational boundaries needs to deal with issues of “turf”
- semantic alignment is ultimately a human problem
- also: regulatory compliance, etc.

# Challenge: semantics

Abused & overloaded term; most folks do not know what it means

In the context of information systems, semantics defines how data “behaves” and how machines can interpret data

Today, semantics is just hard-wired in software

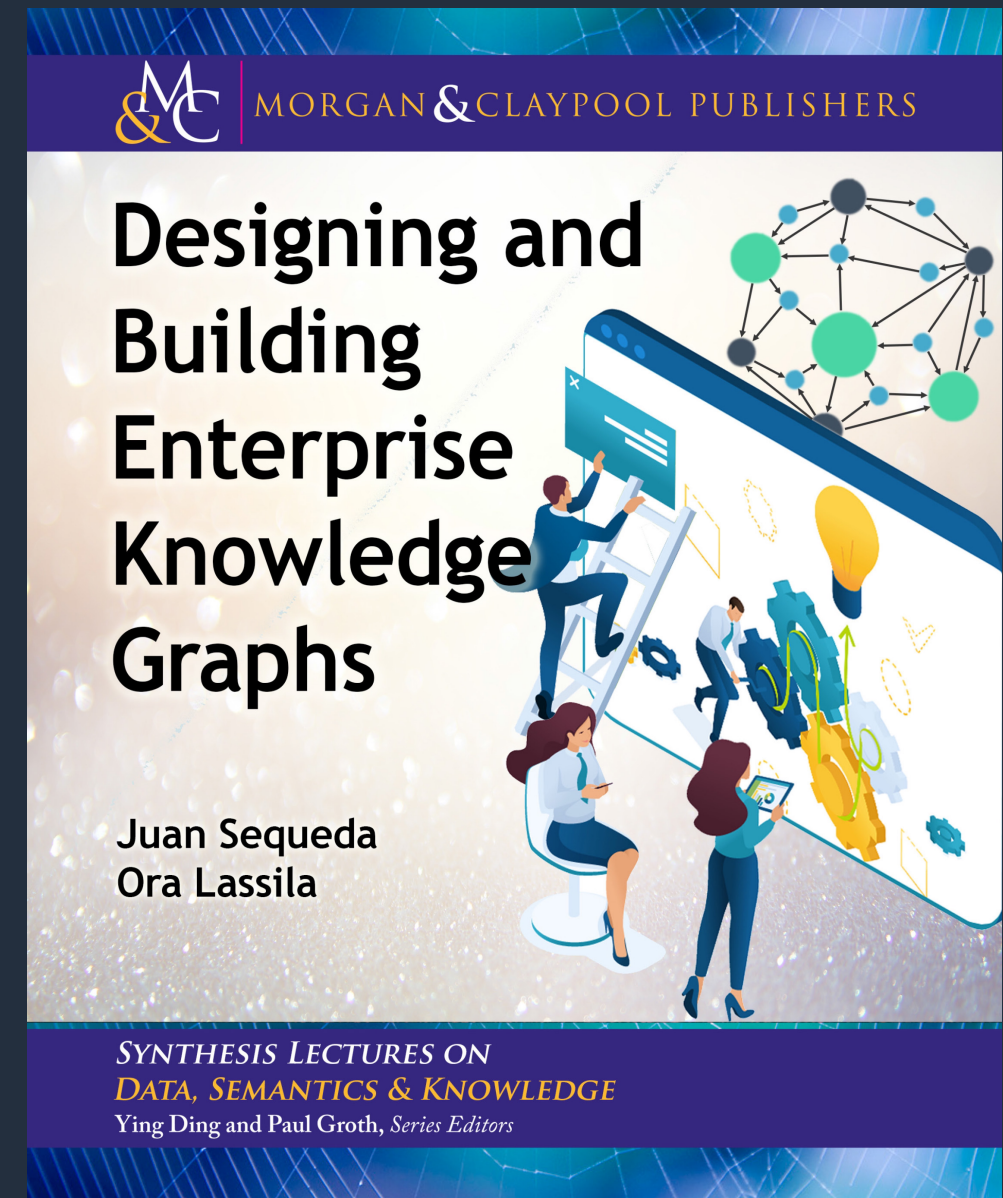
# Challenge: semantics

How do you define and communicate semantics?

How do you “inject” semantics into your data?

Data integration requires alignment of semantics

People tend to confuse their own interpretation with formal semantics



# Human interpretation vs. formal semantics (demonstrated in the context of JSON)

“My data model is JSON”

- JSON is **not a data model**, and mind you, JSON has **no semantics**

“My data is just JSON”

- your data is never “just JSON”, you **always impose external semantics**

“JSON is easy to understand”



# “JSON is easy to understand”

```
{  
  "first name": "Ora",  
  "family name": "Lassila",  
  "degree": "Ph.D",  
  "place of birth": "Helsinki",  
  "hobbies": [ "photography", "scale models" ]  
}
```

# “JSON is easy to understand”

```
{  
  "etunimi": "Ora",  
  "sukunimi": "Lassila",  
  "tutkinto": "TKT",  
  "syntymäpaikka": "Helsinki",  
  "harrastukset": [ "valokuvaus", "pienoismallit" ]  
}
```

# A brief history of graphs and ontologies ➡

3<sup>rd</sup> Century BCE: Categories & logic (Aristotle)

1730s: Graph theory (Euler)

1950s and onwards: Graphs as the essential underpinning of computer science

1960s: Social networks, "small-world experiment", Erdős number (Milgram et al)

1960s-1970s: Network databases (CODASYL), semantic networks (Quillian et al)

1997 and onwards: The Semantic Web, RDF, OWL, etc. (Lassila et al)

Today: Modern knowledge graphs and graph databases

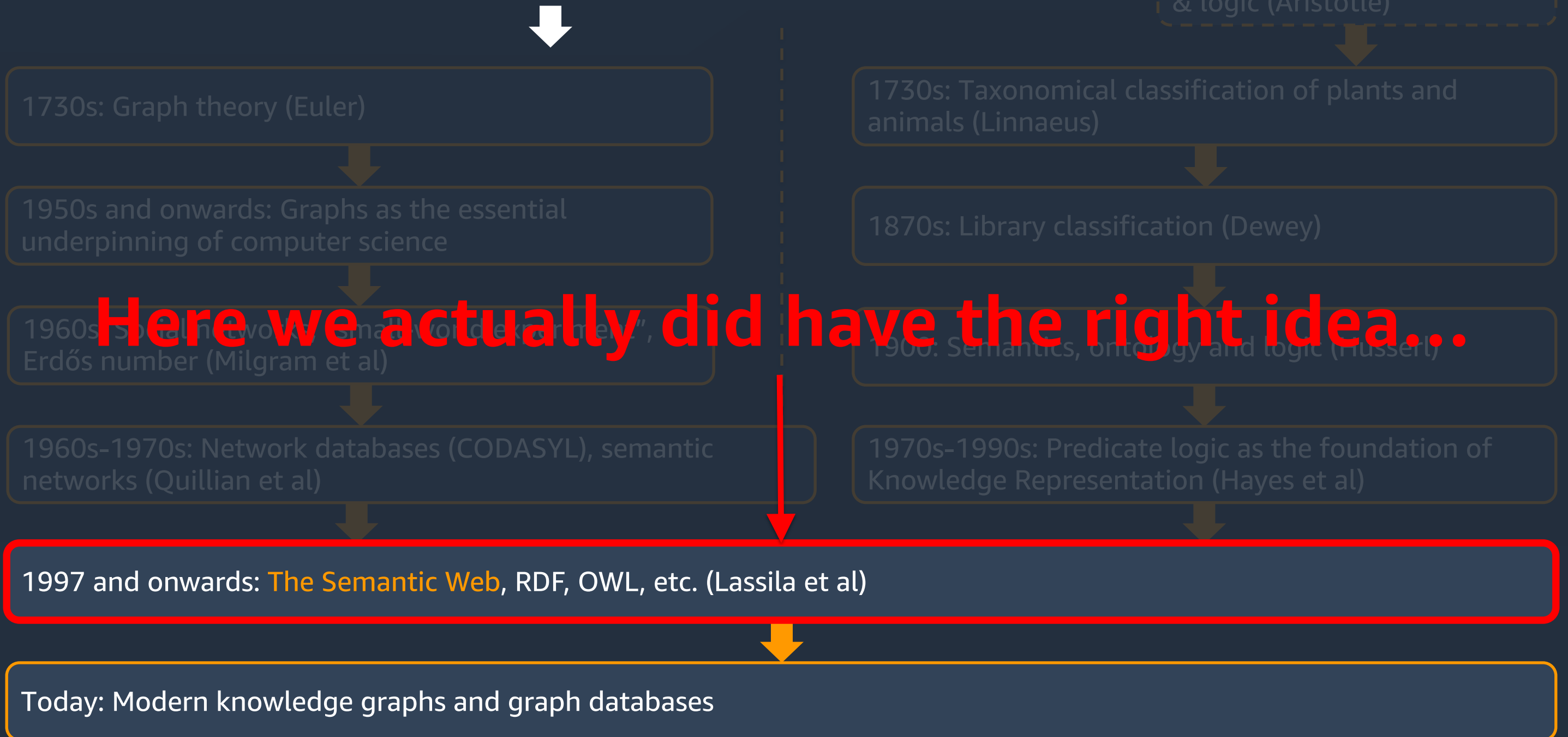
1730s: Taxonomical classification of plants and animals (Linnaeus)

1870s: Library classification (Dewey)

1900: Semantics, ontology and logic (Husserl)

1970s-1990s: Predicate logic as the foundation of Knowledge Representation (Hayes et al)

# A brief history of graphs and ontologies ➡



# Semantic Web, seriously...?



# Semantic Web, seriously...?

Yep, very seriously



We have been doing this for quite some time, you need to pay attention

# Why are Semantic Web technologies attractive?

1. Self-describing data with accessible semantics
  - remember: **accessible data = physical bits + semantics**
2. Data semantics not defined by code & applications
3. Procedural → declarative
4. Serendipity
5. Graph-based representation is intuitive
6. Based on well understood technologies and infrastructure

*NB: Outside popular non-symbolic AI methods, the Semantic Web technologies are the embodiment of what we wanted to do before the “AI winter”*

# Why are Semantic Web technologies attractive?

1. Self-describing data with **accessible semantics**
  - remember: **accessible data = physical bits + semantics**
2. Data semantics not defined by code & applications
3. Procedural → **unifying language for data!**
4. Serendipity
5. **Graph-based representation** is intuitive
6. Based on well-understood technologies and infrastructure

*NB: Outside popular non-symbolic AI methods, the Semantic Web technologies are the embodiment of what we wanted to do before the “AI winter”*

# I want a unifying **logical language** for data!

We need a language to

1. represent data
  2. talk about data
- } **thinking about your data as a graph is natural and intuitive**

We need to be able to define and communicate semantics

Syntax does not matter

(Also, in case you were wondering, this language is not SQL)

# I want a **unifying logical language** for data!

RDF & OWL are currently our **best candidate** for something like this

Graphs are nice, and so are graph databases, but if you want/need your data in **some other physical form**, go for it

# I want a **unifying logical language** for data!

We also need good, shared definitions of basic things:

- DC, FOAF, DOAP, PROV-O, FIBO, NIEM, ...
- schema.org, LOD cloud, ...

I really wish the open-source community would embrace ontologies with the same energy and enthusiasm they have done for software

# A graph is a graph is a graph... or is it?

RDF vs. labeled property graphs...?

- confuses users (when given the choice)
- should not matter
- the rift hurts the industry and community as a whole

**Stop reinforcing the rift!**

# A graph is a graph is a graph



The Amazon Neptune team is working to mitigate the rift: **Project OneGraph**

See this for more info:  
<https://arxiv.org/abs/2110.13348>

The solution is not easy

- RDF-star is a significant step in the right direction

The real issue may in fact be

Graph as a **logical representation** vs. graph as a **data structure**



Will **knowledge graphs** save us from  
the mess of modern data practice?

Will **knowledge graphs** save us from  
the mess of modern data practice?

**Yes**

Data → Knowledge

Focus on semantics: unifying language

Will **knowledge graphs** save us from the mess of modern data practice?

**Yes, but...**

The path to happiness is not an easy one, nor a quick one

There will be some pain in the short term

The technological solutions exist, but we need to work on people

# Thank you!

Thanks to my colleagues and collaborators:

- Charles Ivie, Brad Bebee, Piyush Mathur, Nicole Moldovan, Juan Sequeda

Contact:

- [ora@amazon.com](mailto:ora@amazon.com)
- Twitter: [@oralassila](https://twitter.com/oralassila)

