Emanuele Della Valle
Riccardo Tommasini
Emanuele Falzone

FROM THEORY TO PRACTICE

STREAM REASONING

RW 2020, 16th Reasoning Web Summer School - 25.06.2020

EMANUELE DELLA VALLE

- Associate Professor at DEIB Politecnico di Milano
- Expert in semantic technologies and stream computing
- Brander of **stream reasoning**
- 20+ years of experience in research and innovation projects
- Startupper

emanuele.dellavalle@polimi.it
@manudellavalle
http://emanueledellavalle.org
http://streamreasoning.org
http://fluxedo.com

RICCARDO TOMMASINI

- Assistant Professor of Computer Science, University of Tartu
- Expert in graph and streaming data processing, data integration and semantic technologies
- Main contributor of the RSP-QL stack Engine, author of VoCaLS ontology
- ~5 years experience in innovation and research projects

riccardo.tommasini@ut.ee
@rictomm
https://rictomm.me

EMANUELE FALZONE

- PhD student at DEIB
  Politecnico di Milano
- Investigating graph stream processing
- 2+ years of experience in research and innovation projects
- Open source contributor!

emanuele.falzone@polimi.it
http://emanuelefalzone.com

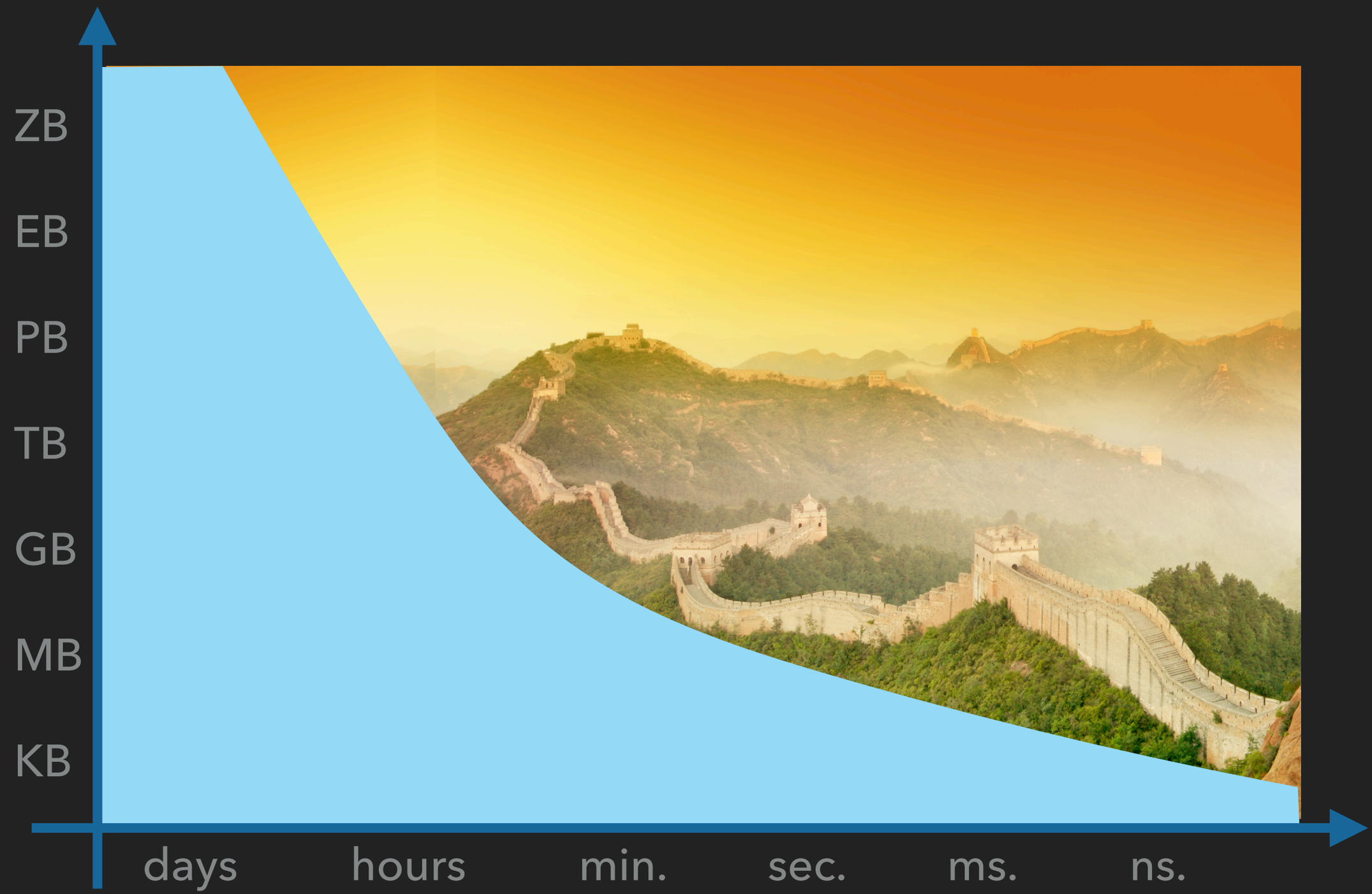
BIG DATA TECHS CAN TAME VOLUME

- Hadoop, MapReduce, HIVE
- “schema on read” methodology
- spark (x100 faster)
- “data lake” concept

BIG DATA TECHS CAN TAME VELOCITY

- Storm
- Kafka
- Spark Streaming
- Flink
- paradigmatic change
  - from persistent data and transient queries
  - **to persistent queries and transient data**

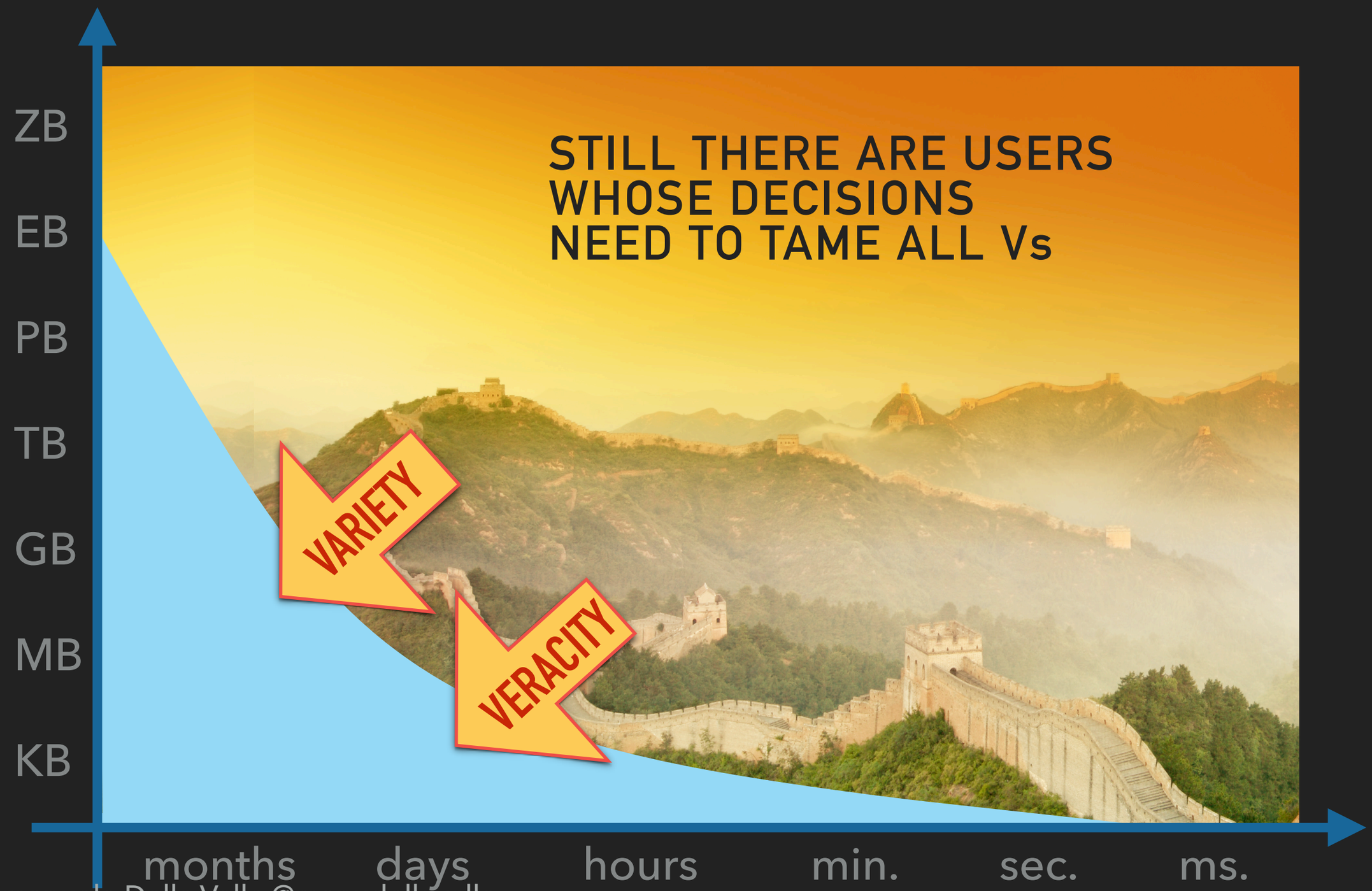
BIG DATA TECHS CANNOT TAME VOLUME AND VELOCITY SIMULTANEOUSLY



BIG DATA TECHS CAN TAME VARIETY USING SEMANTIC TECHNOLOGIES

- RDF data model
- SPARQL query language
- OWL ontological language
- R2RML mapping language
- Ontology Based Data Access methodology

VARIETY & VERACITY MAKES PROBLEMS HARDER



OFF-SHORE OIL OPERATIONS

- When sensors on a drilling pipe in an oil-rig indicate that it is about to get stuck, how long – according to historical records – **can I keep drilling?**

- **400,000 sensors** from 10s of differente producers
- **10,000 observations per second**, many out-of-operational-ranges

SMART CITIES

- Can you **suggest where to spend my next hours** given my interests, the presence of people and what their doing?
- **100,000s people** leaving **10,000s digital footprint per second** via Call Data Records, Bluetooth, WiFi, Social Media, …

REQUIREMENT ANALYSIS

A system able to answer those queries must be able to

- handle **massive** datasets
- process **data streams** on the fly
- cope with **heterogeneous** datasets
- cope with **incomplete** data
- cope with **noisy** data
- provide **reactive answers**
- support **fine-grained information access**
- integrate **complex** domain **models**

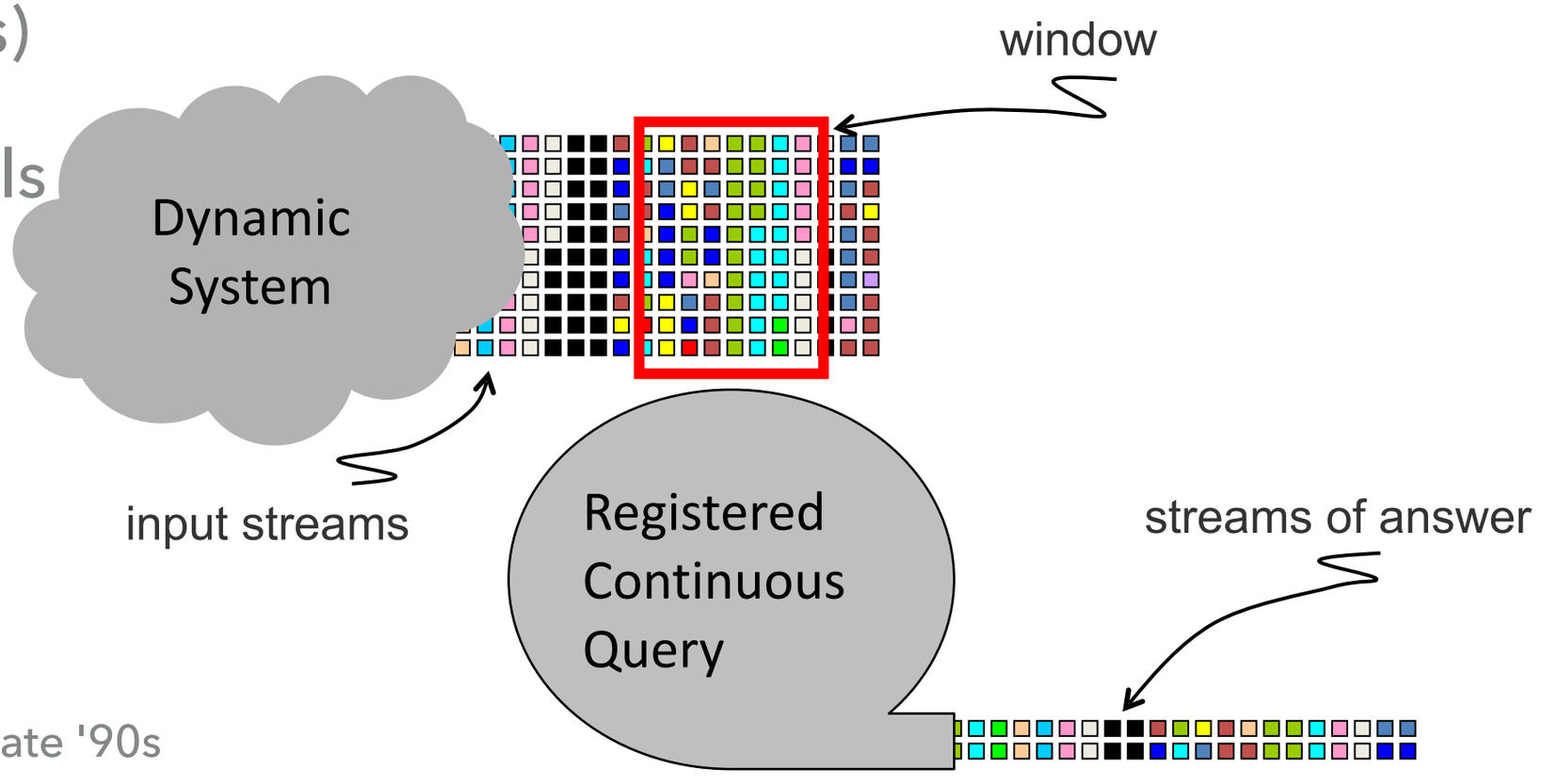
Volume	Velocity	Variety	Veracity
x			
	x		
		x	
		x	x
			x
		x	
		x	x
		x	

DATA STREAMS

- Data Streams are usually **unbounded**
- **No assumption** can be made **on** data arrival **order**
- Data items in streams often represent **observations not facts**
- Size and time constraints make it **difficult to store** and process data stream elements **after their arrival**
- **One-time processing** is the typical mechanism used to deal with streams

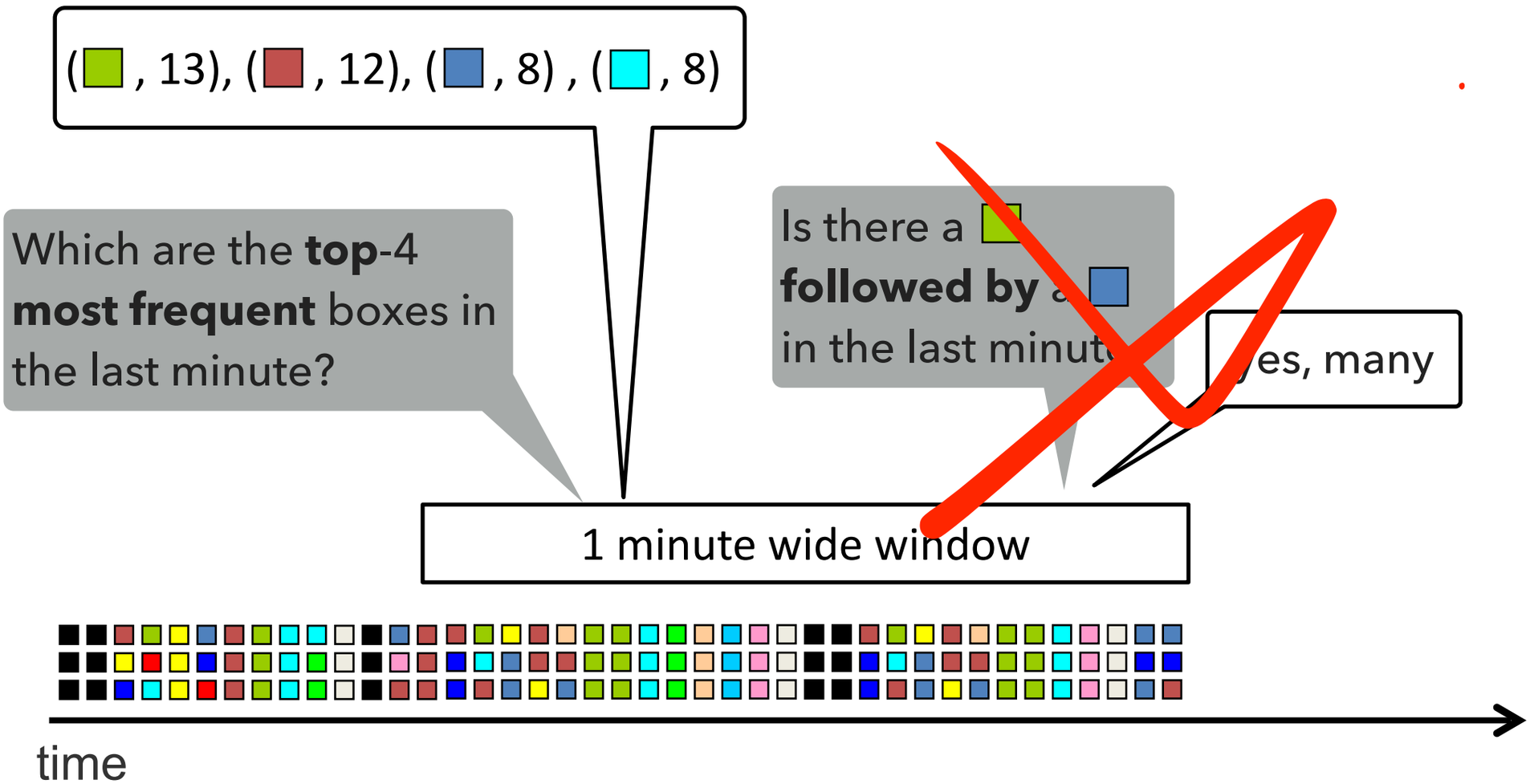
THE PARADIGMATIC CHANGE* OF STREAM PROCESSING

- From **persistent data** and **transient queries** (one time semantics)
- To **transient data** and **persistent queries** (continuous semantics)
- Two competing models
  - DSMS
  - CEP



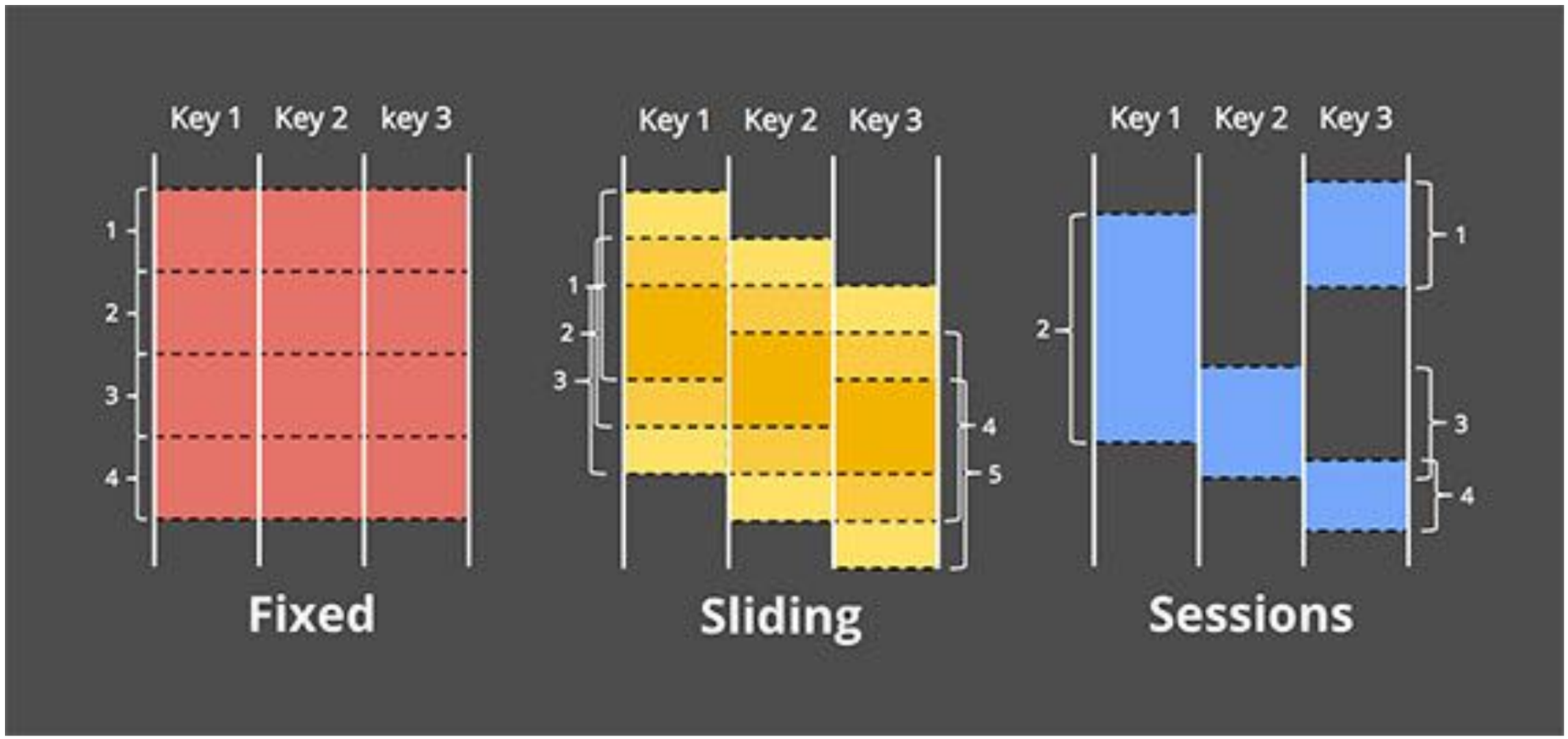
*first arose in DB community in the late '90s

STREAM PROCESSING A USER PERSPECTIVE



WINDOWS

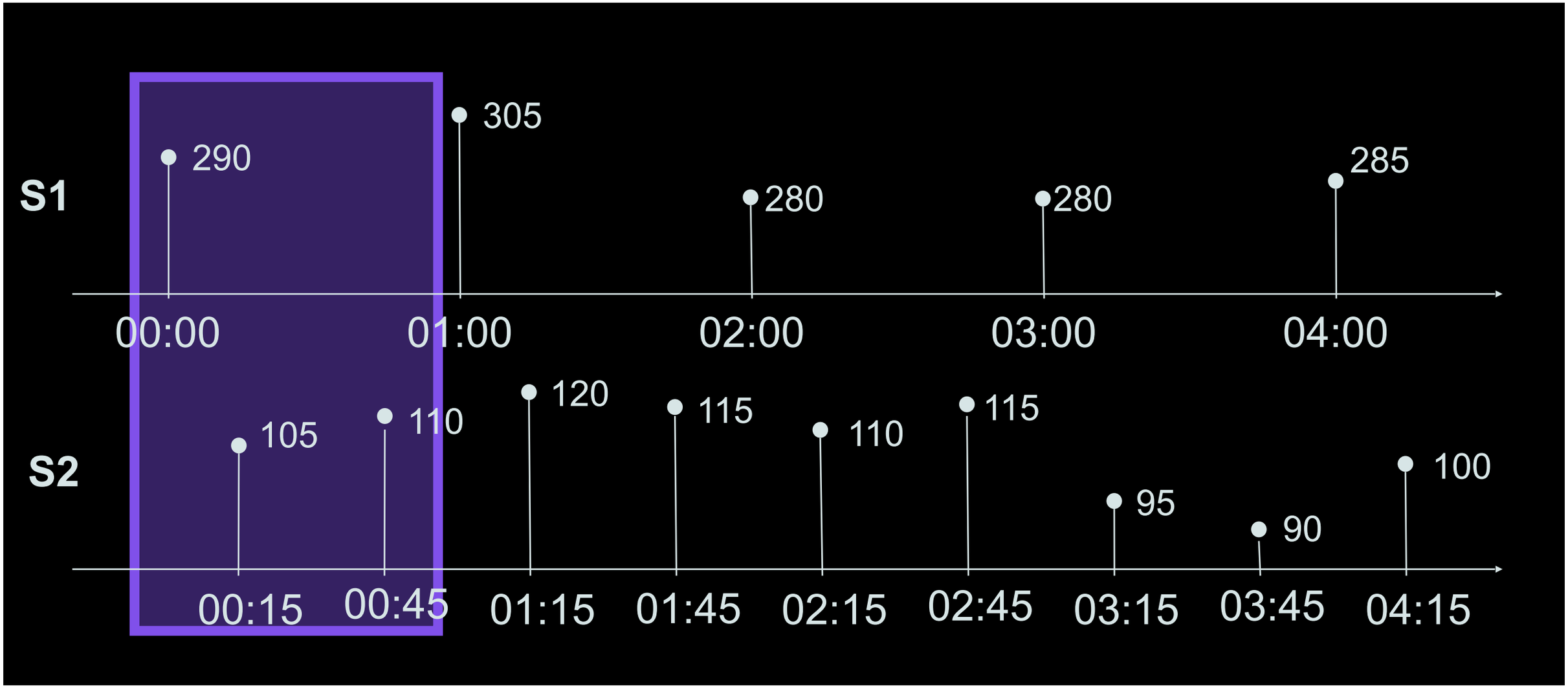
- Windows define a finite sub-streams of an unbounded stream



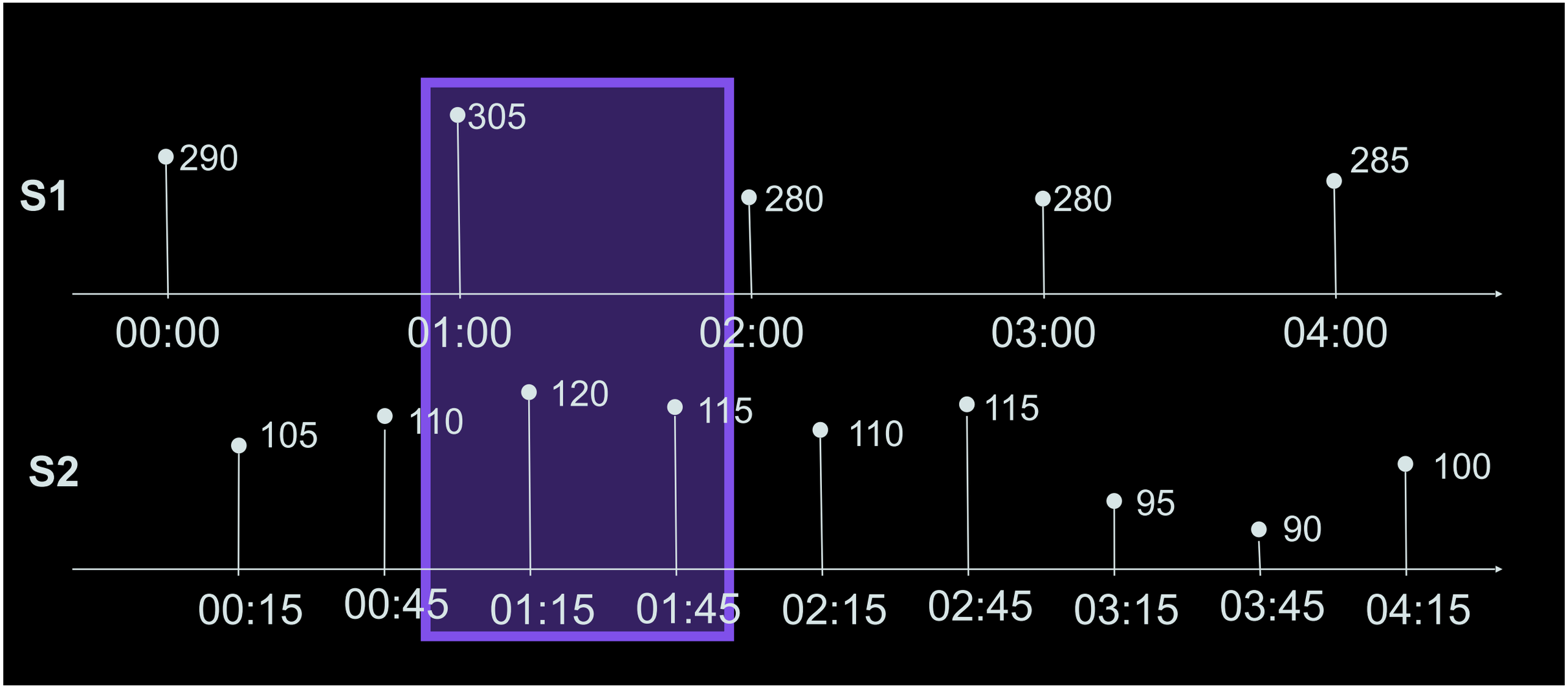
- They can be interpreted as locally closed worlds

JOINING STREAMS ON WINDOWS

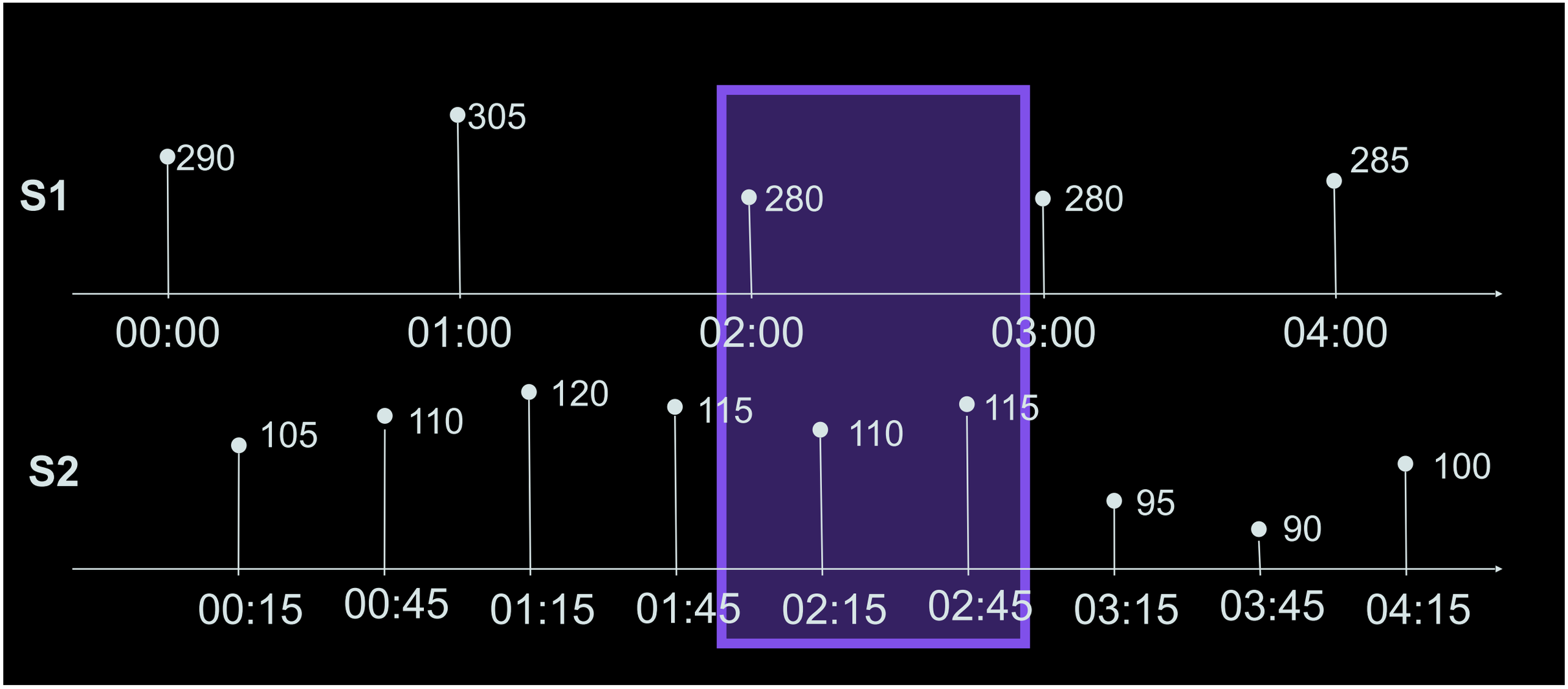
CAN WE SYNCHRONIZE THEM WITH A TUMBLING WINDOW?



CAN WE SYNCHRONIZE THEM WITH A TUMBLING WINDOW?

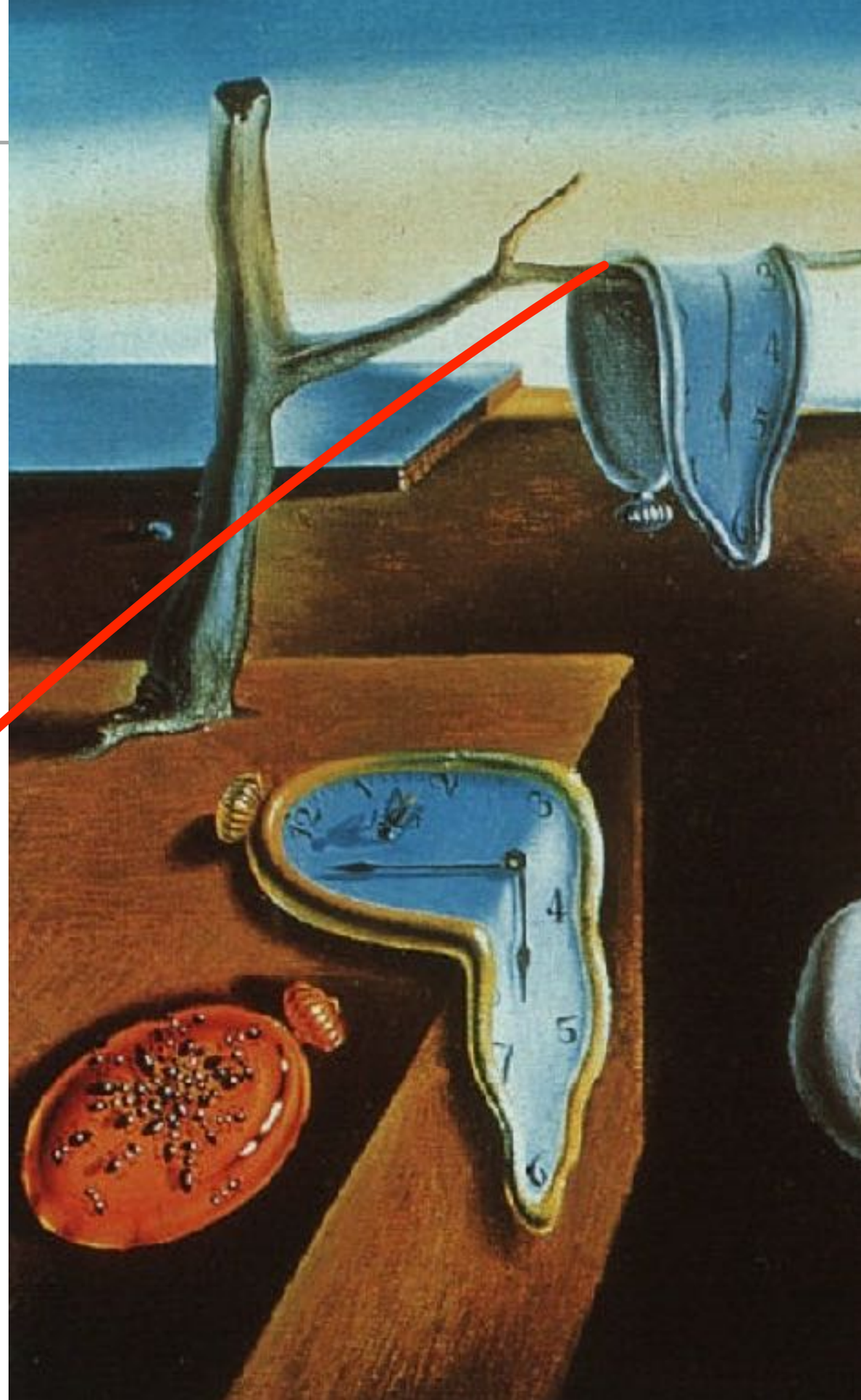


CAN WE SYNCHRONIZE THEM WITH A TUMBLING WINDOW?

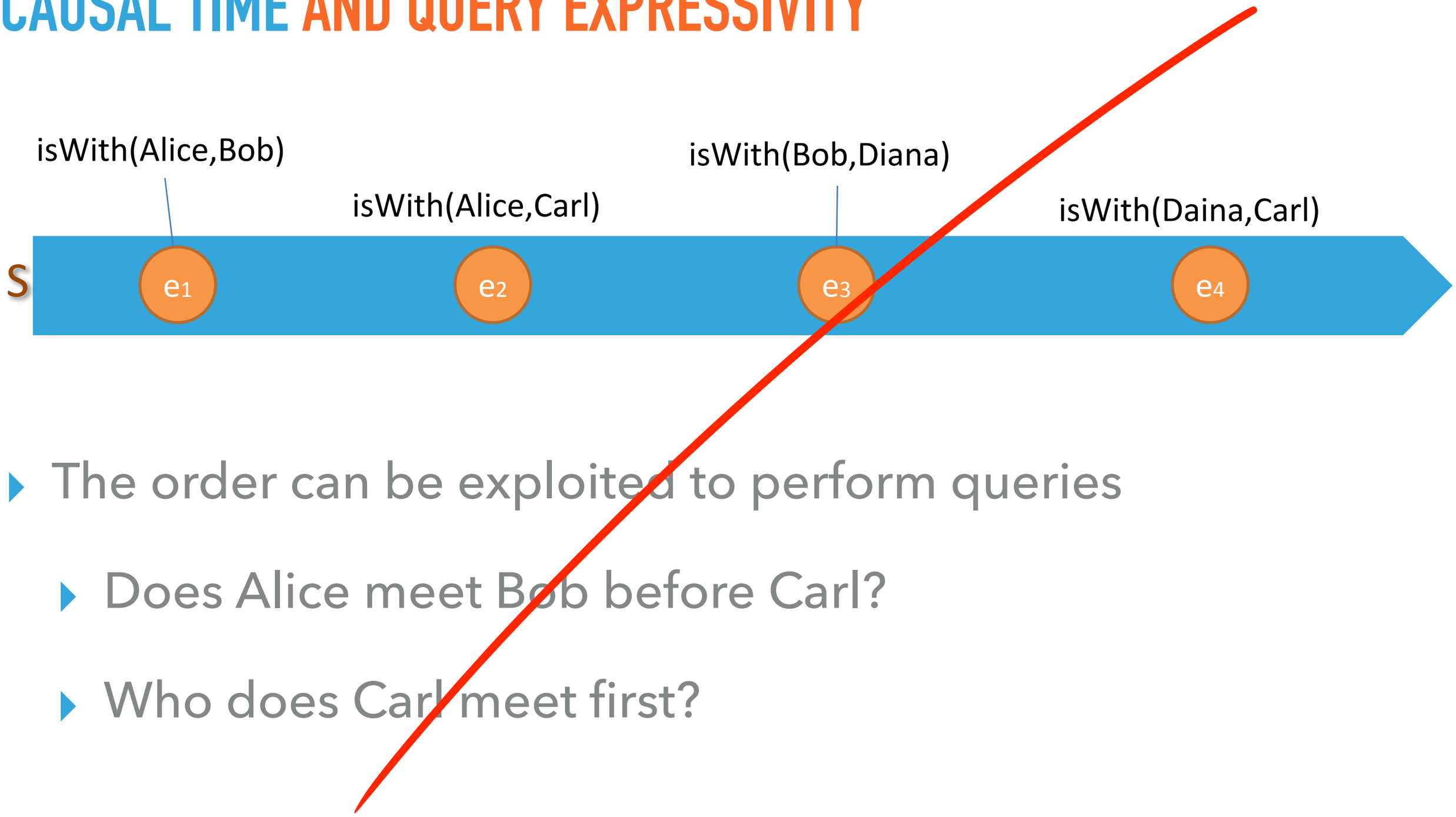


TIME MODELS AND . . .

- Causal
- Absolute
- Interval-based

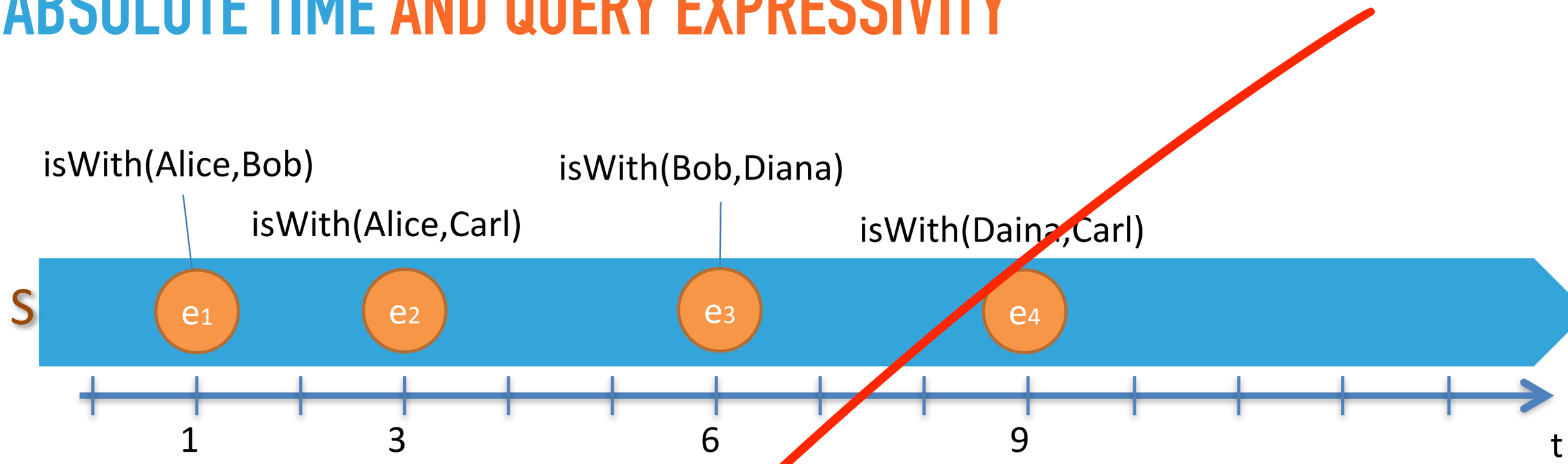


CAUSAL TIME AND QUERY EXPRESSIVITY



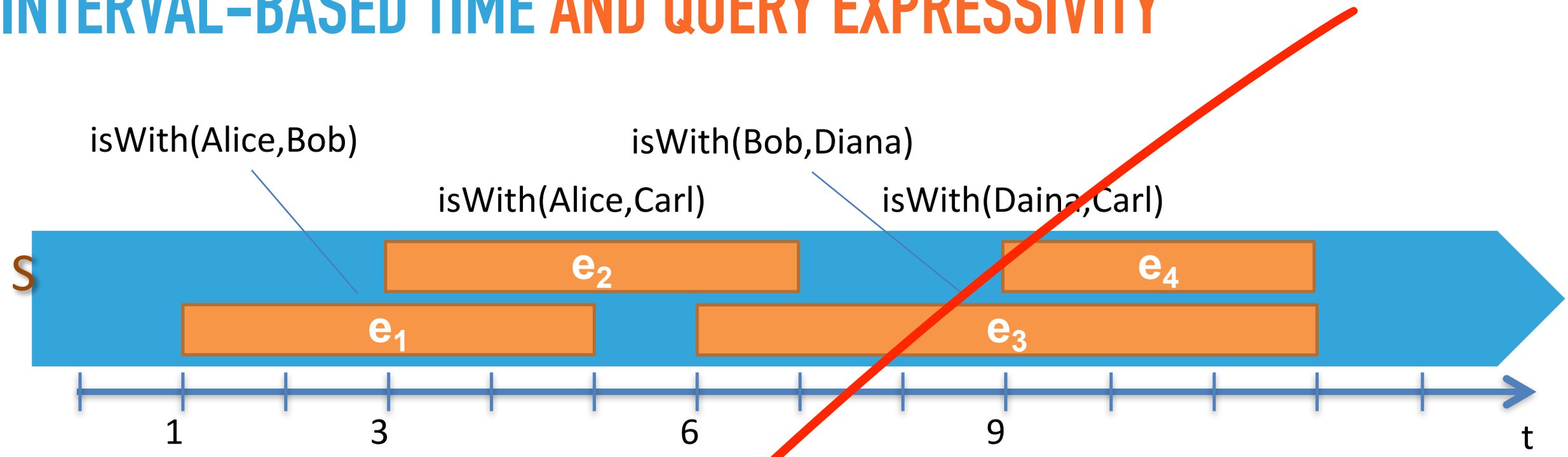
- The order can be exploited to perform queries
  - Does Alice meet Bob before Carl?
  - Who does Carl meet first?

ABSOLUTE TIME AND QUERY EXPRESSIVITY



- We can ask the queries in the previous slide
- We can start to compose queries taking into account the time
  - How many people has Alice met in the last 5m?
  - Does Diana meet Bob and then Carl within 5m?

INTERVAL-BASED TIME AND QUERY EXPRESSIVITY



- We can ask the queries in the previous two slides
- It is possible to write even more complex queries:
  - Which are the meetings the last less than 5m?
  - Which are the meetings with conflicts?

STREAM PROCESSING VS. REQUIREMENTS

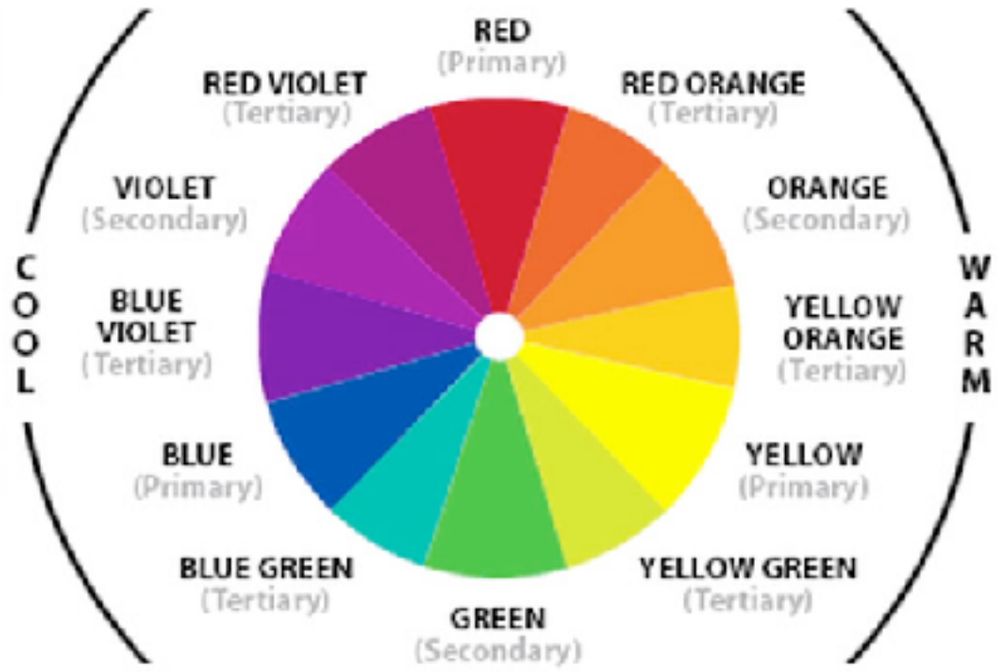
Requirement	SP
massive datasets	✓
data **streams**	✓
heterogeneous dataset	✗
incomplete data	✗
noisy data	✓
reactive answers	✓
fine-grained information **access**	✓
complex domain **models**	✗

SEMANTIC TECHS A USER PERSPECTIVE

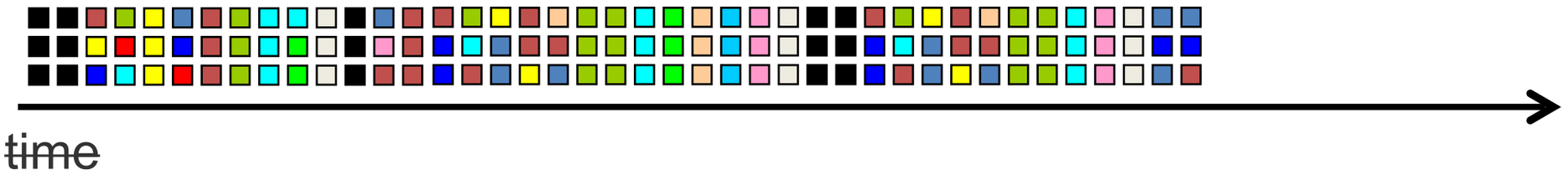
Are there any **cool colored** box?

yes, 7 , 13 , ...

An **ontology** of colors



~~1 minute wide window~~



~~time~~

STATE-OF-THE-ART: RDF MODEL

- **RDF**: Resource Description Framework
  - It allows to make **statements about resources** in the form of **subject-predicate-object** expressions
  - In RDF terminology **triples**
  - E.g. :Rodin (subject) :sculpts (predicate) :TheThinker (object)

- A collection of RDF statements represents a **labelled, directed graph**
  - In RDF terminology a **graph**
  - E.g., the triple above can be connected to millions of others telling information about Rodin and The Thinker

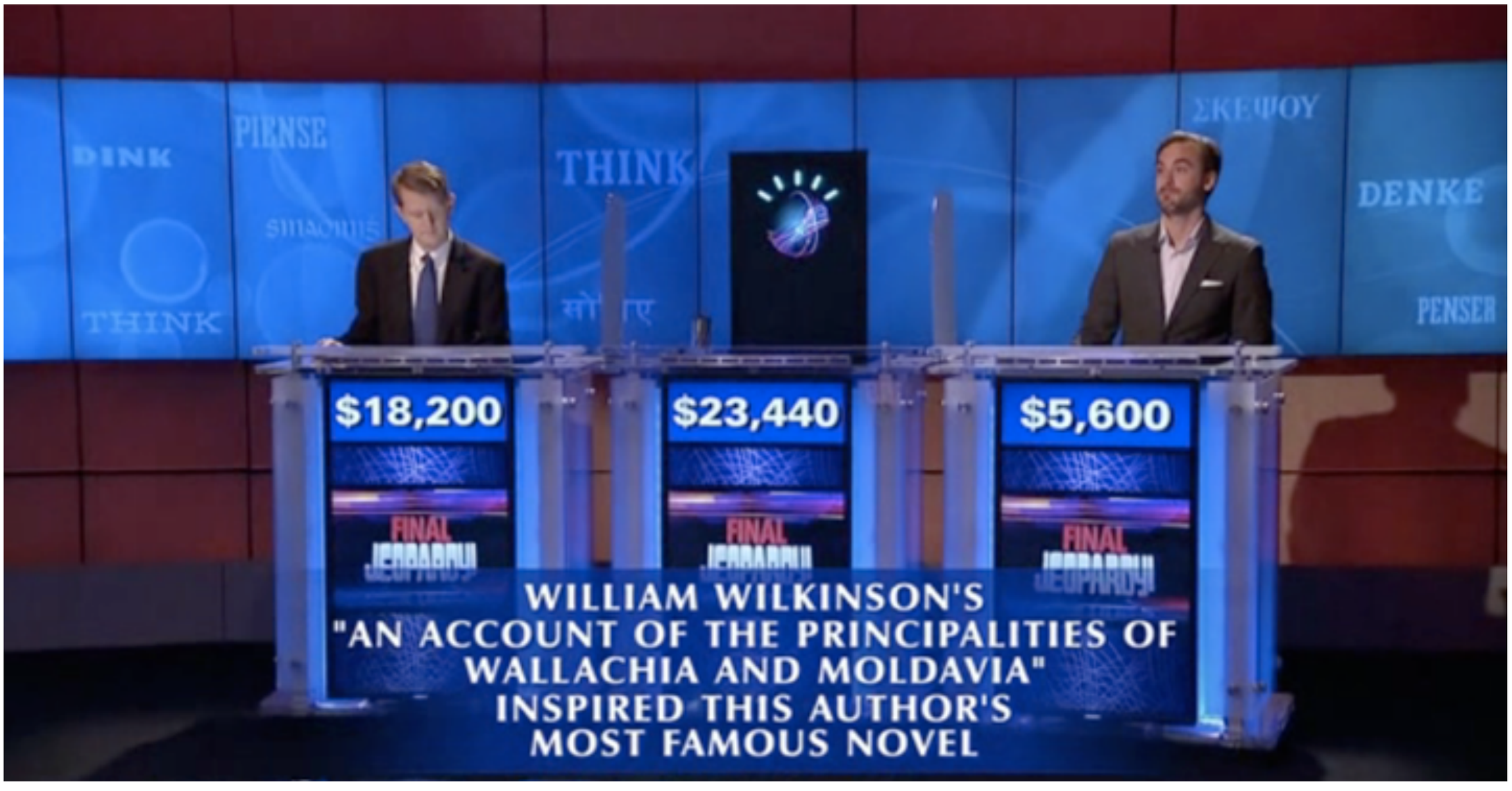
STATE-OF-THE-ART: OWL

- **OWL**: Web Ontology Language
  - It allows to give **well-defined meaning** to classes, properties, individuals, and data values
    - E.g.
      - sculpts is a property
      - sculpting is a special way of creating
      - those who create are artists
      - …
  - A collection of classes, properties, individuals, and data values forms a **vocabulary**
  - When using **OWL2DL**, **classes and properties** are isolated **in** a **T-box** while **individuals and values** are **in** an **A-box**

STATE-OF-THE-ART: SPARQL

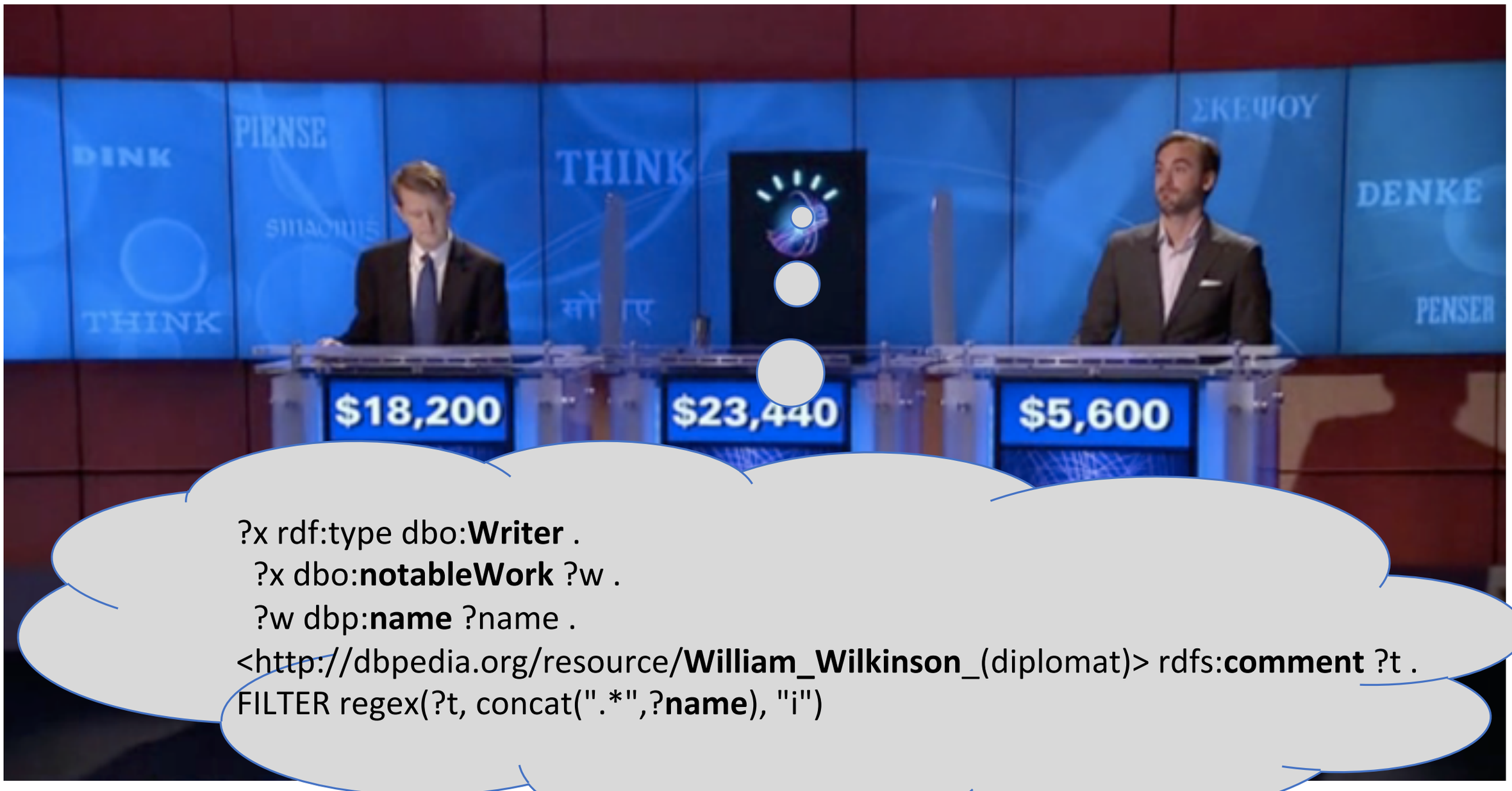
- **SPARQL**: Querying RDF under some entailment regime
  - It allows to make search for **statements about resources**
    - In SPARQL terminology a **triples pattern** is an RDF triple in which users can add variables
      - E.g. 1: what does Rodin sculpt? :Rodin :sculpts ?x
      - E.g. 2: what connects Rodin to The Thinker? :Rodin ?x :TheThinker
      - E.g. 3: what's Rodin? (*it requires inference*) :Rodin a ?x
  - A collection of triples patterns represents a **graph pattern**
  - Graph patterns can be combined with FILTER, UNION and other clauses to form an expressive query language

THE POWER OF SPARQL AND OPEN KNOWLEDGE GRAPHS

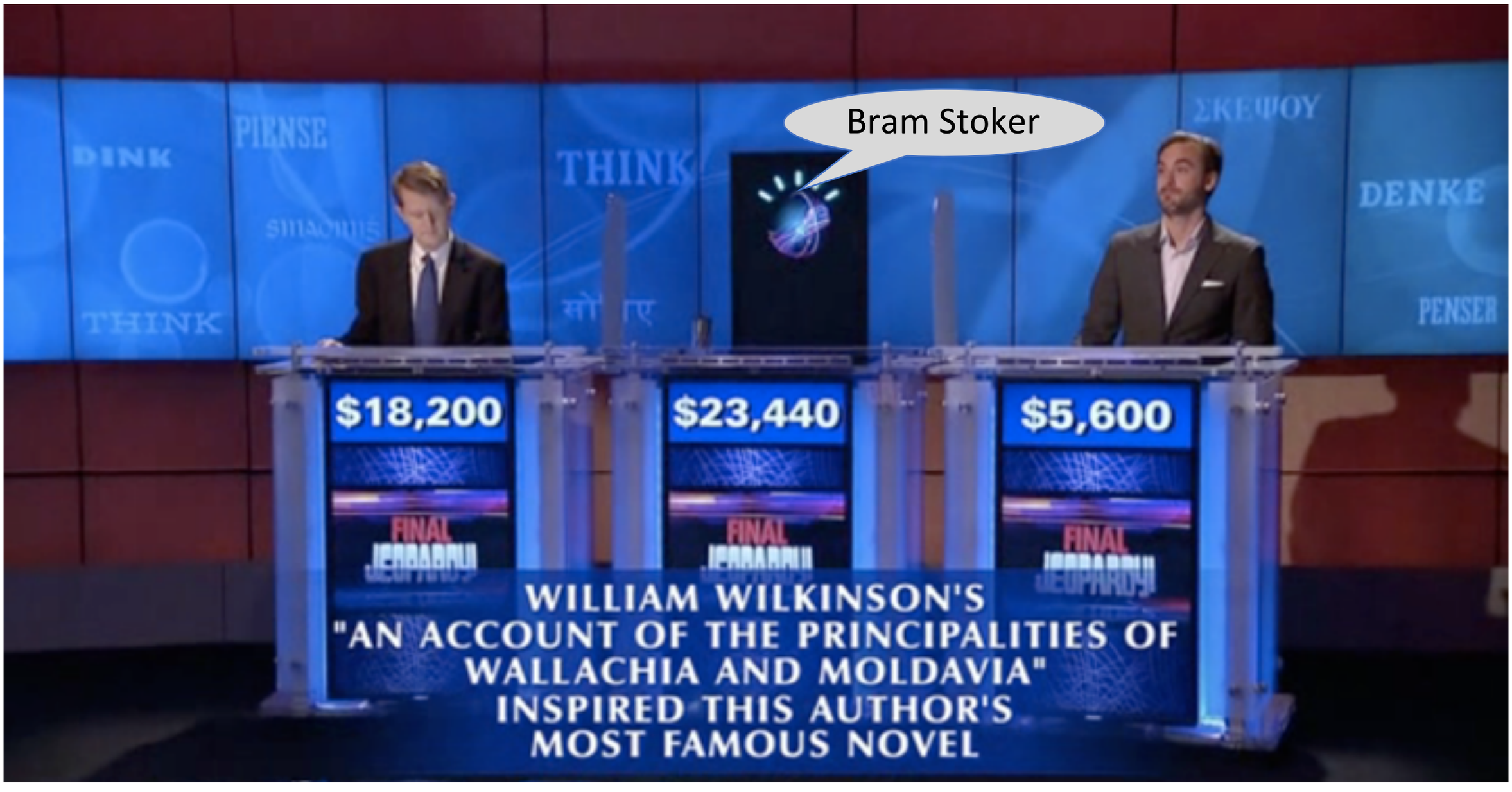


Emanuele Della Valle - http://emanueledellavalle.org - @manudellavalle

THE POWER OF SPARQL AND OPEN KNOWLEDGE GRAPHS



THE POWER OF SPARQL AND OPEN KNOWLEDGE GRAPHS



Emanuele Della Valle - http://emanueledellavalle.org - @manudellavalle

SEMANTIC TECHS VS. REQUIREMENTS

Requirement	SP	ST
massive datasets	✓	✓
data **streams**	✓	✗
heterogeneous dataset	✗	✓
incomplete data	✗	✓
noisy data	✓	✗
reactive answers	✓	✗
fine-grained information **access**	✓	✓
complex domain **models**	✗	✓

STREAM REASONING RESEARCH QUESTION

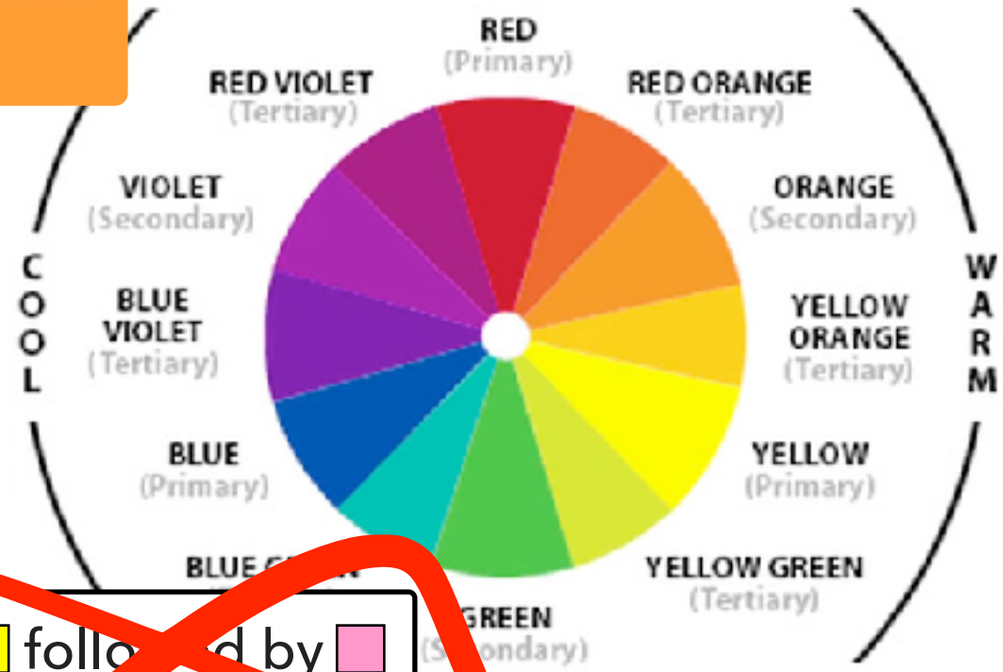
Is it possible to **make sense** in **real time** of **multiple**, **heterogeneous**, **gigantic** and inevitably **noisy** and **incomplete data streams** in order to support the **decision** processes of extremely large numbers of concurrent users?

E. Della Valle, S. Ceri, F. van Harmelen & H. Stuckenschmidt, 2010

STREAM REASONING A USER PERSPECTIVE

Is there a **primary cool** color followed by a **secondary warm** one in the last minute?

An **ontology** of colors

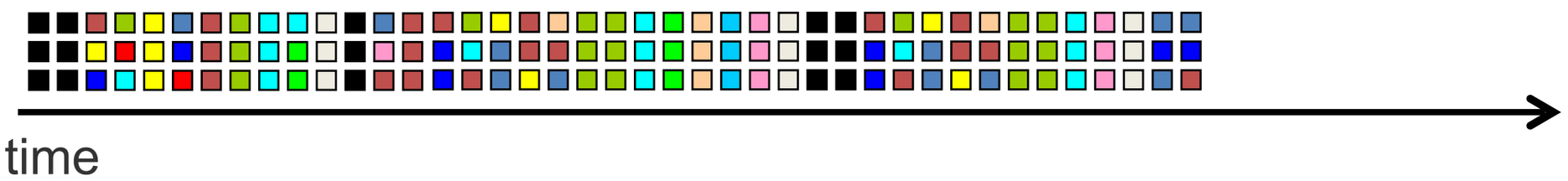


(■, 13),(■, 8) , (■, 8)

Which are the top-2 most frequent **cool** colors in the last minute?

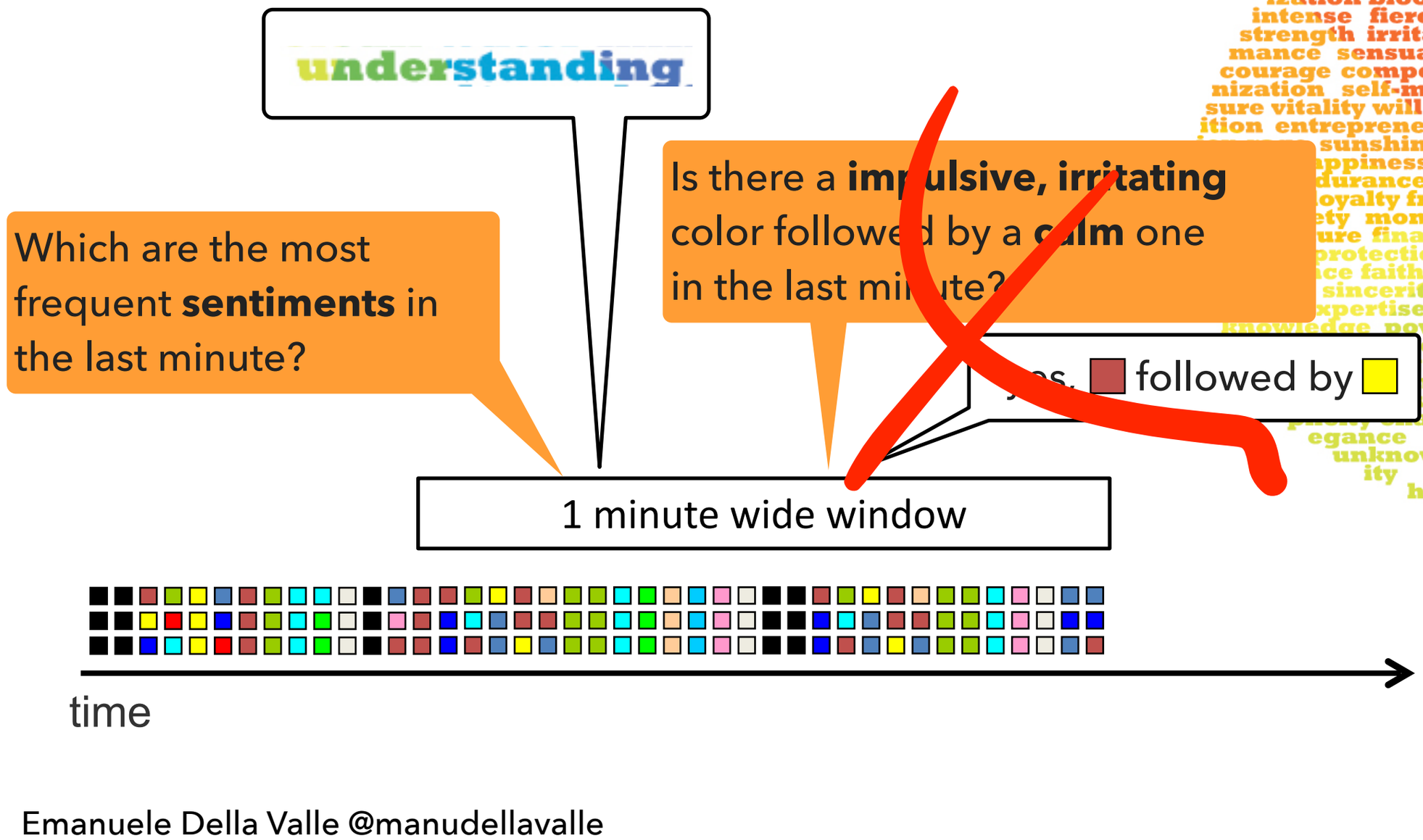
yes, ■ followed by ■

1 minute wide window



STREAM REASONING A USER PERSPECTIVE

The **better** is the **ontology** (of the colors) we are using
the **more expressive** are the **queries** we can register



A **better**
ontology (of colors)

Emanuele Della Valle
Riccardo Tommasini
Emanuele Falzone

add Matteo’s slides about VKG
positioning RSP vrt SR
present RSP-QL AND RDF streams
RSP Engine and Jasper

FROM THEORY TO PRACTICE

STREAM REASONING

RW 2020, 16th Reasoning Web Summer School - 25.06.2020