Yandex

Data modeling and file formats

Storage:
HDFS

Application:
Real-time bidding
platform

Storage:
HDFS

Application:
Real-time bidding
platform

Data = Clicks, Impressions

Task = Compute CTR

Domain-specific terms

Storage:
HDFS

Data = Bytes, Files

Storage-specific terms

Application:
Real-time bidding
platform

Data = Clicks, Impressions

Task = Compute CTR

Domain-specific terms

Storage:
HDFS

Data = Bytes, Files

Storage-specific terms

Application:
Real-time bidding
platform

Data = Clicks, Impressions

Task = Compute CTR

Domain-specific terms

In-between

Data model,
File formats

Data modeling

- Data model – a way you think about your data elements, what they are, what domain they come from, how different elements relate to each other, what they are composed of

  - abstract model
  - explicitly defines the structure of data

Relational data model

C1	C2	C3	C4
v11	v12	v13	v14
v21	v22	v23	v24
v31	v32	v33	v34

Relational data model

Data set
(also: table, relation)

C1	C2	C3	C4
v11	v12	v13	v14
v21	v22	v23	v24
v31	v32	v33	v34

Relational data model

Data set
(also: table, relation)

Tuples
(also: rows)

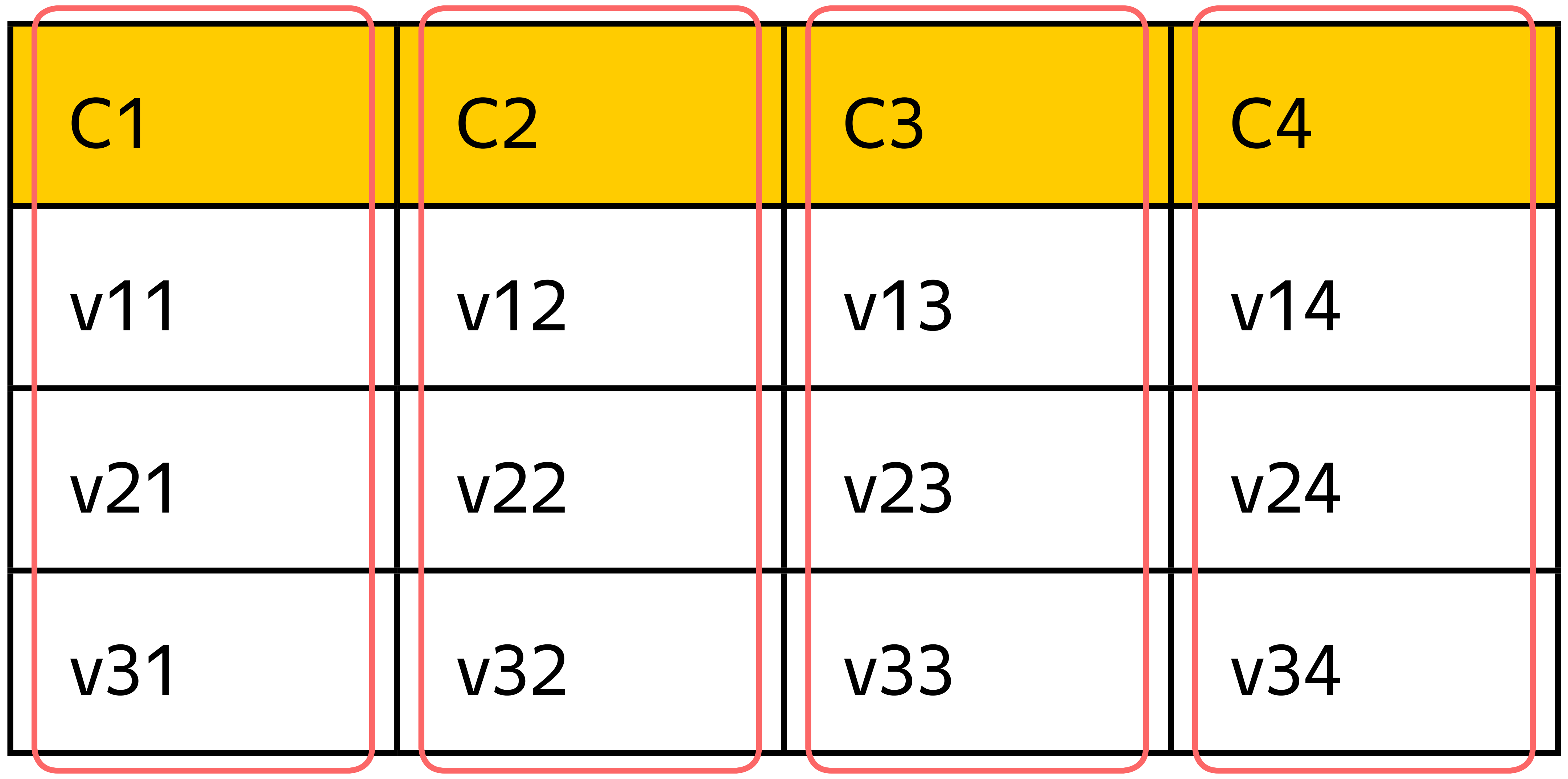
C1	C2	C3	C4
v11	v12	v13	v14
v21	v22	v23	v24
v31	v32	v33	v34

Relational data model

Data set
(also: table, relation)

Tuples
(also: rows)

Columns
(also: attributes)



C1	C2	C3	C4
v11	v12	v13	v14
v21	v22	v23	v24
v31	v32	v33	v34

Relational data model

Data set
(also: table, relation)

Tuples
(also: rows)

Columns
(also: attributes)

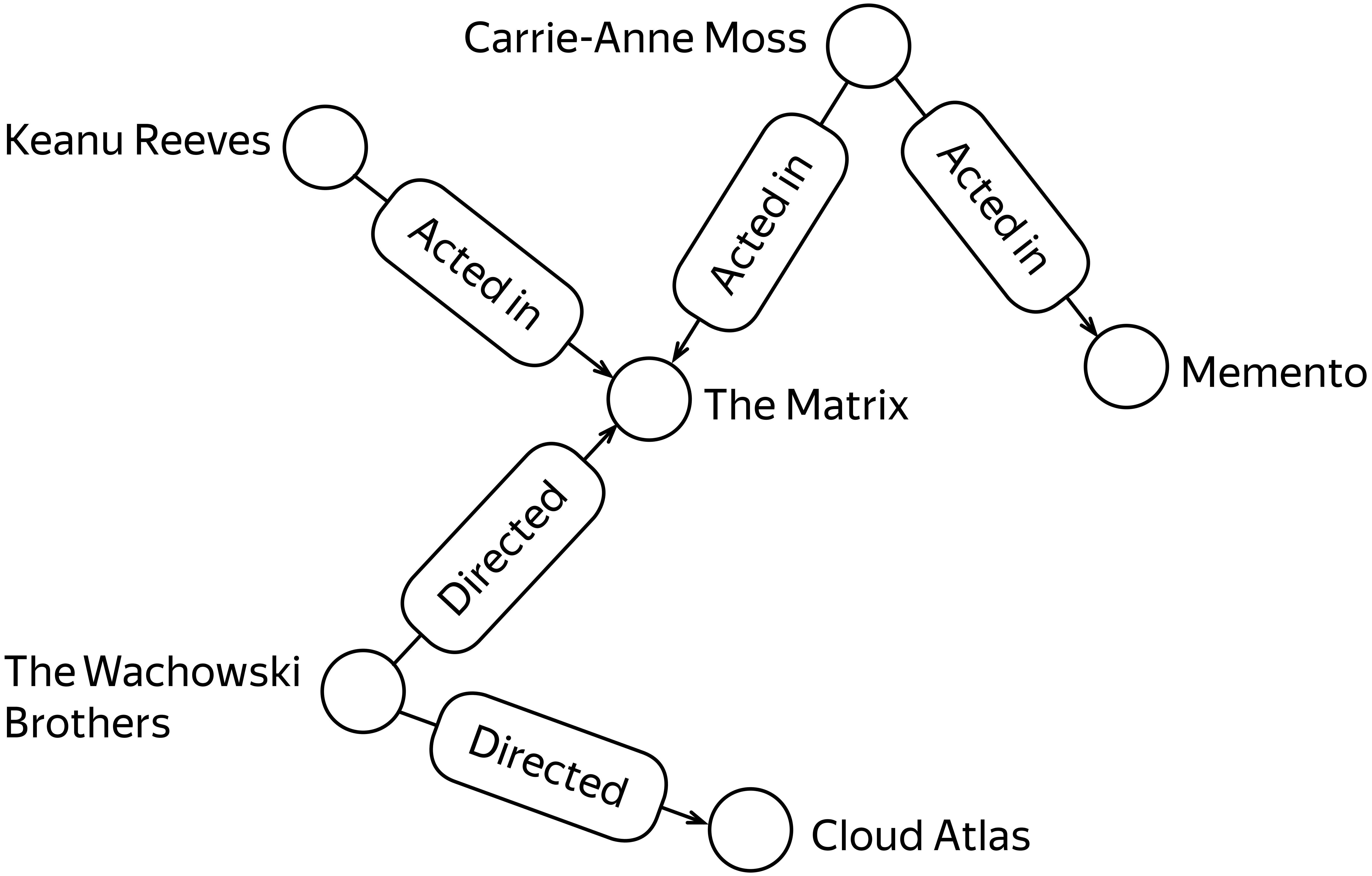
Values

C1	C2	C3	C4
v11	v12	v13	v14
v21	v22	v23	v24
v31	v32	v33	v34

Relational data model (example)

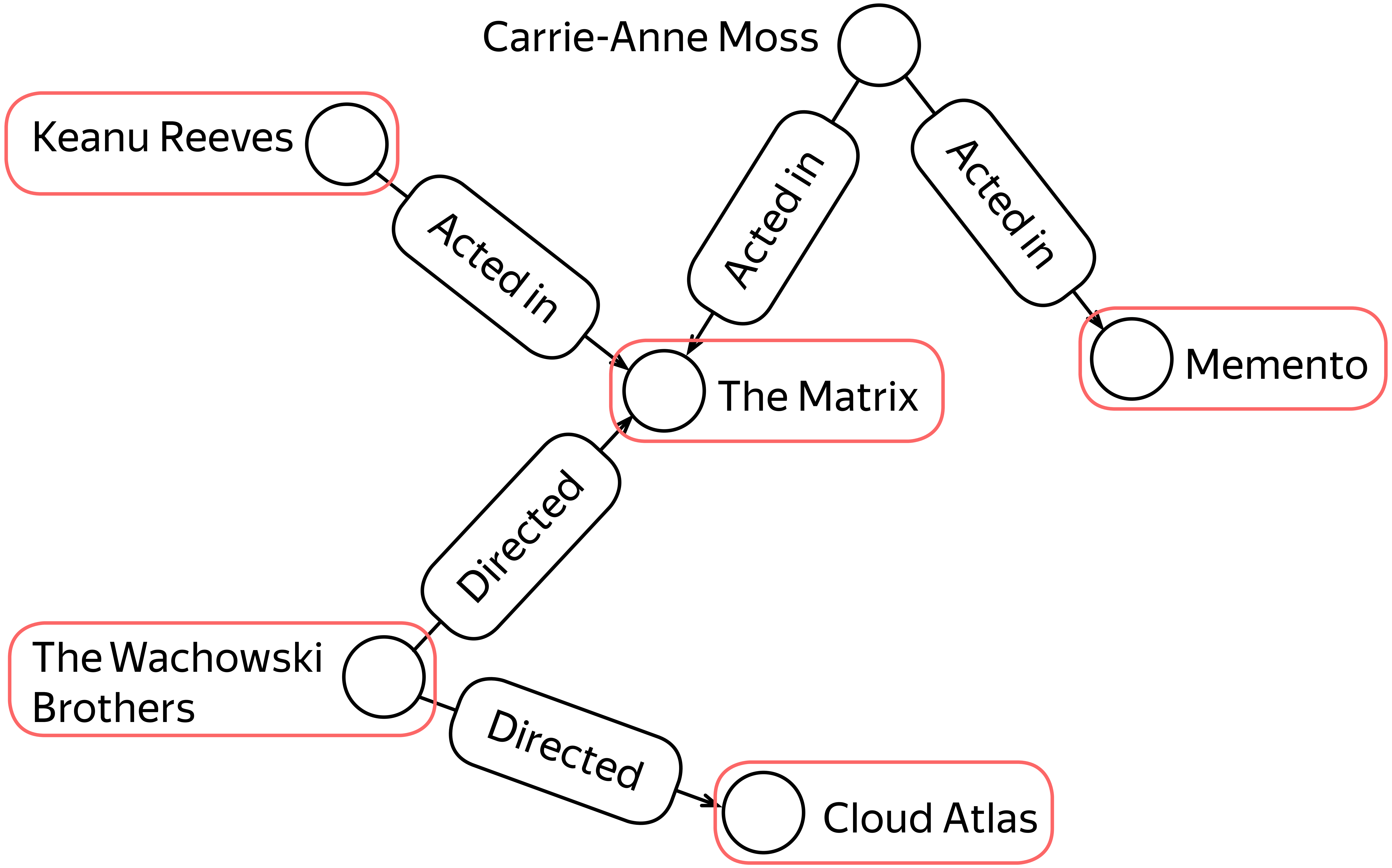
Event	Timestamp	User ID	Ad ID
IMPRESSION	T21:04:13	u1248	a864
IMPRESSION	T21:04:15	u3192	a711
CLICK	T21:04:20	u3192	a711

Graph data model



Graph data model

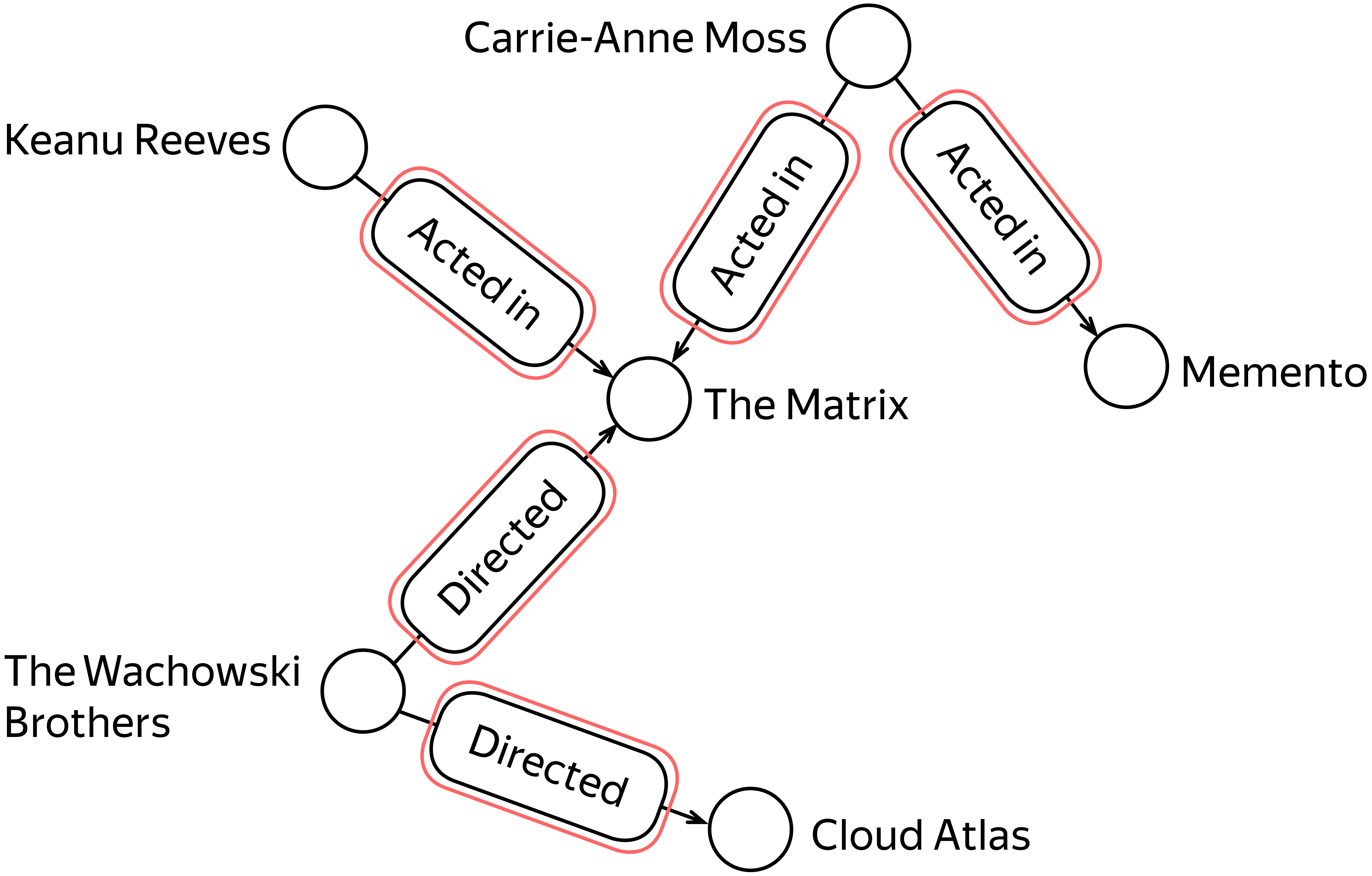
Vertices
(also: entities)



Graph data model

Vertices
(also: entities)

Edges
(also: relations)



Data model

- Defines the structure of data
- Makes some things easier to express than others
- Will use a *relational model*

Unstructured data?

Unstructured data?

- Technically, all data is structured at least as a byte sequence
- Usually, means “not structured enough for a task”

Unstructured data?

- Technically, all data is structured at least as a byte sequence
- Usually, means “not structured enough for a task”

- Ex. 1: Logs = Line per request with all related data
  - Easy to work with

Unstructured data?

- Technically, all data is structured at least as a byte sequence
- Usually, means “not structured enough for a task”

- Ex. 1: Logs = Line per request with all related data
  - Easy to work with

- Ex. 2: Video = Sequence of frames
  - Hard to work with

File format (also: storage format)

- Defines (physical) data layout
- Different design choices lead to different tradeoffs in complexity
  - affects performance, correctness

File format (also: storage format)

- Defines (physical) data layout
- Different design choices lead to different tradeoffs in complexity
  - affects performance, correctness

- Primary function: to transform between raw bytes and programmatical data structures (*serialization* & *deserialization*)

File formats

- Many!
- Differ in:
  - space efficiency
  - encoding & decoding speed
  - supported data types
  - splittable/monolithic structure
  - extensibility

Conclusion

- Deciding on a *data model* and *storage format* have far-reaching implications for your application performance, correctness, computation complexity, and resource usage

- Next videos
  - Text formats
  - Binary formats
  - Compression

BigDATAteam