INTRODUCTION TO LINKED DATA AND GRAPH DATABASES

HANDS-ON WORKSHOP

PHUSE 2017 ANNUAL CONFERENCE

Edinburgh, Scotland
11 October, 2017

INSTRUCTOR

Tim Williams	Statistical Systems Analyst UCB BioSciences Raleigh, NC, USA tim.williams@PhUSE.eu

Workshop Files, Presentation PDF:

https://github.com/phuse-org/LinkedDataWorkshop/Annual2017-EU

HELP!

WHO?

HOW?

OUTLINE

- Introduction
  - Server Login
  - Data as a Graph
- Exercises
  1. Neo4j Labeled Property Graph (LPG)
  2. Resource Description Framework (RDF)
- Demonstrations (time permitting)
  - SDTM as LGP
  - SDTM as RDF

OUTLINE

- **Introduction**
  - Server Login
  - Data as a Graph
- Exercises
  1. Neo4j Labeled Property Graph (LPG)
  2. Resource Description Framework (RDF)
- Demonstrations (time permitting)

"INTRODUCTION TO LINKED DATA AND GRAPH DATABASES"

"FROM WHITEBOARD TO QUERYABLE GRAPH: A *VERY BASIC* INTRODUCTION TO CONVERTING CLINICAL TRIALS CONCEP DATA."

MATERIALS

- Laptop - power up!
- Pencil + eraser, or pen
- Printed copies of:
  - Exercises
  - Neo4j Diagram
  - RDF Diagram
- Server IP Address

OUTLINE

- Introduction
  - **Server Login**
  - Data as a Graph
- Exercises
  1. Neo4j Labeled Property Graph (LPG)
  2. Resource Description Framework (RDF)
- Demonstrations (time permitting)

SERVER LOGIN

Instructions in Exercises (Page 3)

Computer:

User name: phuseldw

Password:

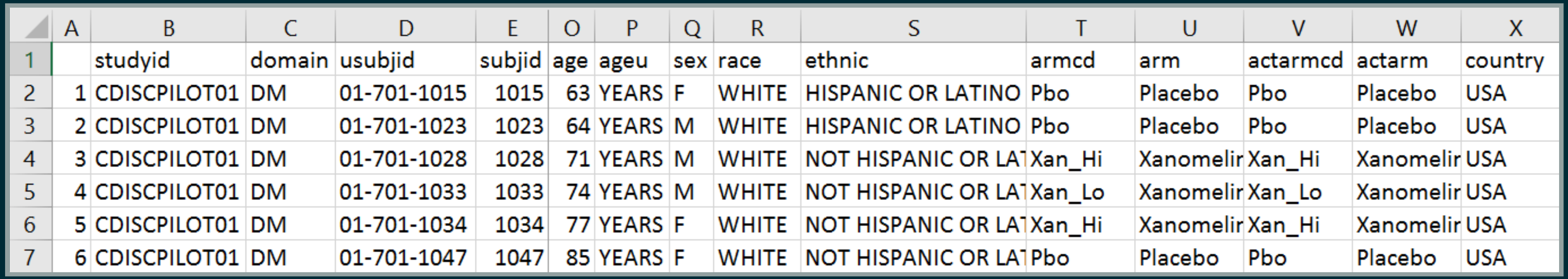
OUTLINE

- Introduction
  - Server Login
  - Data as a Graph
- Exercises
  1. Neo4j Labeled Property Graph (LPG)
  2. Resource Description Framework (RDF)
- Demonstrations (time permittting)

WHY DATA AS A GRAPH?

ONE EXAMPLE: SDTM DOMAINS

SDTM DM DOMAIN



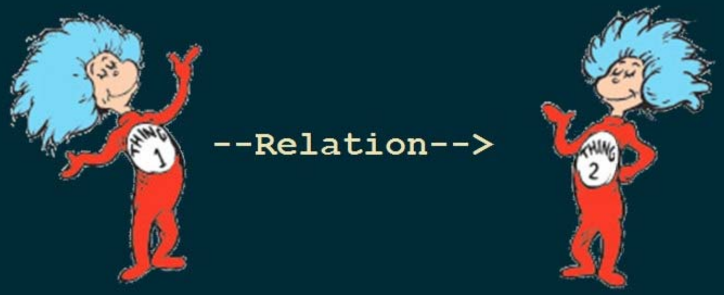
	A	B	C	D	E	O	P	Q	R	S	T	U	V	W	X
1		studyid	domain	usubjid	subjid	age	ageu	sex	race	ethnic	armcd	arm	actarmcd	actarm	country
2	1	CDISCPILOT01	DM	01-701-1015	1015	63	YEARS	F	WHITE	HISPANIC OR LATINO	Pbo	Placebo	Pbo	Placebo	USA
3	2	CDISCPILOT01	DM	01-701-1023	1023	64	YEARS	M	WHITE	HISPANIC OR LATINO	Pbo	Placebo	Pbo	Placebo	USA
4	3	CDISCPILOT01	DM	01-701-1028	1028	71	YEARS	M	WHITE	NOT HISPANIC OR LAT	Xan_Hi	Xanomelir	Xan_Hi	Xanomelir	USA
5	4	CDISCPILOT01	DM	01-701-1033	1033	74	YEARS	M	WHITE	NOT HISPANIC OR LAT	Xan_Lo	Xanomelir	Xan_Lo	Xanomelir	USA
6	5	CDISCPILOT01	DM	01-701-1034	1034	77	YEARS	F	WHITE	NOT HISPANIC OR LAT	Xan_Hi	Xanomelir	Xan_Hi	Xanomelir	USA
7	6	CDISCPILOT01	DM	01-701-1047	1047	85	YEARS	F	WHITE	NOT HISPANIC OR LAT	Pbo	Placebo	Pbo	Placebo	USA

What is wrong here?

- Inflexible, version specific row x column structure and format
- Mixture of concepts
- No integral metadata
- Data repetition

GRAPH DATA CAN FIX THESE PROBLEMS!

DATA AS A GRAPH?

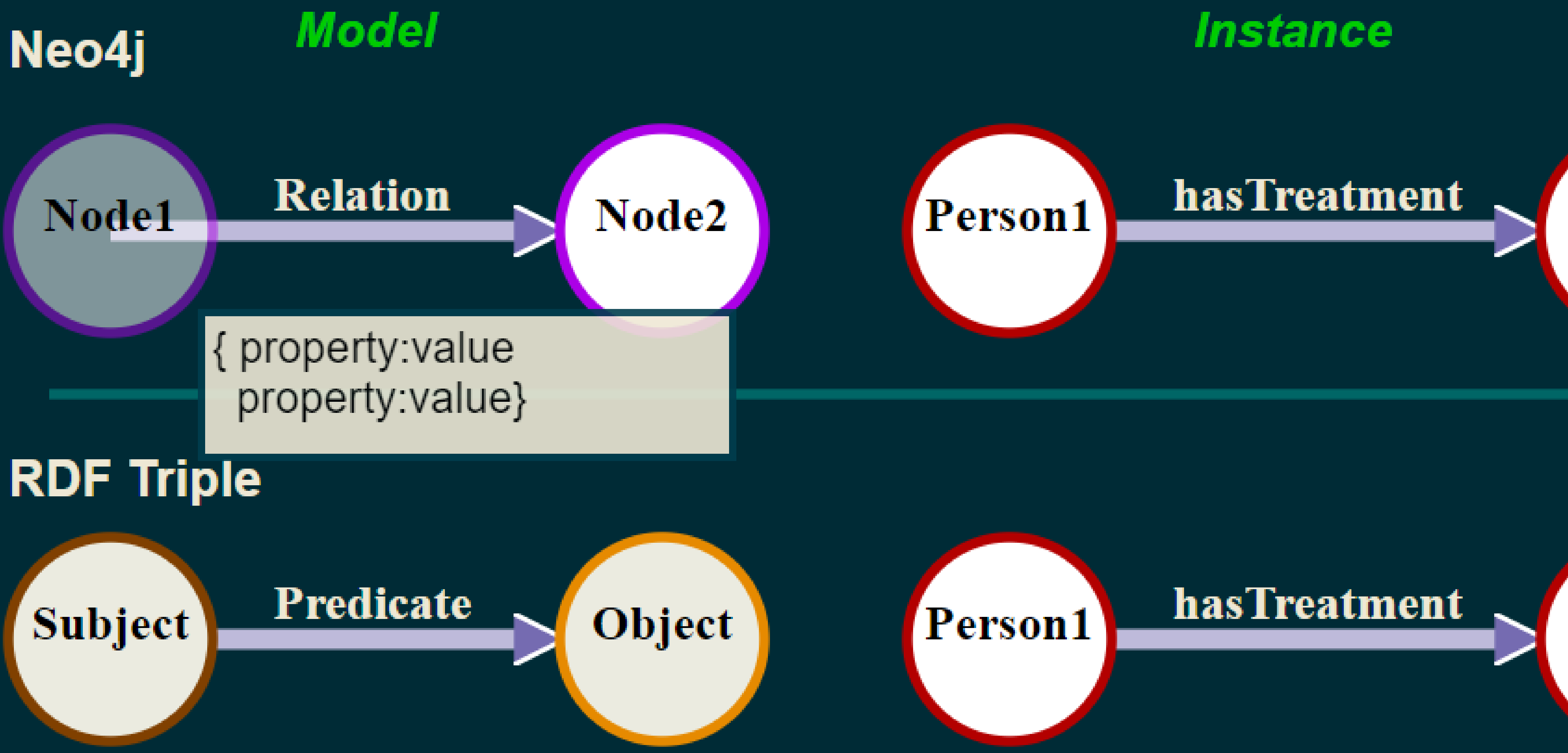


--Relation-->

Compare Neo4j with RDF

SCREEN SHOT OF INTERACTIVE GRAPH

NEO4J, RDF COMPARISON

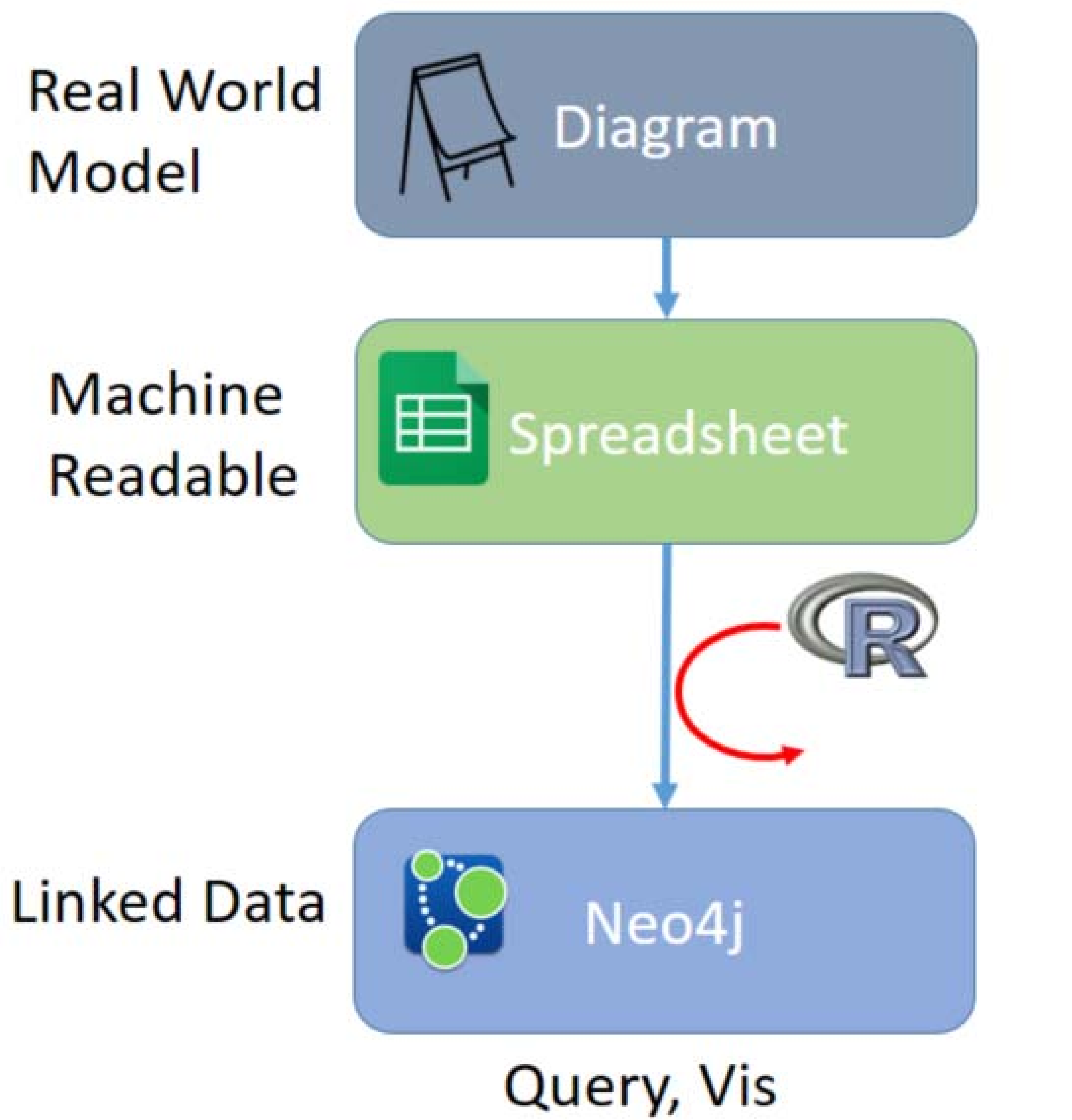


NEO4J, RDF: MORE CORE DIFFERENCES

	Neo4j	**RDF**
Query	Cypher	SPARQL
Traverse	Easier	Harder
Graph	Less complex, shallow	More complex, deep
Ontologies	Code them externally?	Many available & tools to make them.
Learning	Easier	Harder

OUTLINE

- Introduction
  - Server Login
  - Data as a Graph
- **Exercises**
  1. **Neo4j Labeled Property Graph (LPG)**
  2. Resource Description Framework (RDF)
- Demonstrations (time permitting)



Real World Model

Diagram

Machine Readable

Spreadsheet

Linked Data

Neo4j

Query, Vis

APPROACH

- Model instances, not ontology ; Real-life things, not a classification of *types* of things (onto
  - Example: PERSON1 *`enrolledin`* STUDY1 not "PATIENTS enrolledin STUDIES"
- See **"Guidelines for Adding Nodes and Relations"**

IDEAS FOR NEW NODES

Site

- SITE1 *`locatedin`* COUNTRY1
- SITE1 *`investigator`* INVESTIGATOR1

Person

- PERSON1 *`gender`* M

Study

- STUDY1 *`sponsor`* COMPANY1

1.1 DIAGRAM THE MODEL

"WHITE BOARD" THE THINGS

Handout: Neo4j Diagram

The Revised Neo4j Diagram

1.2 TRANSFER DIAGRAM TO SPREADSHEET

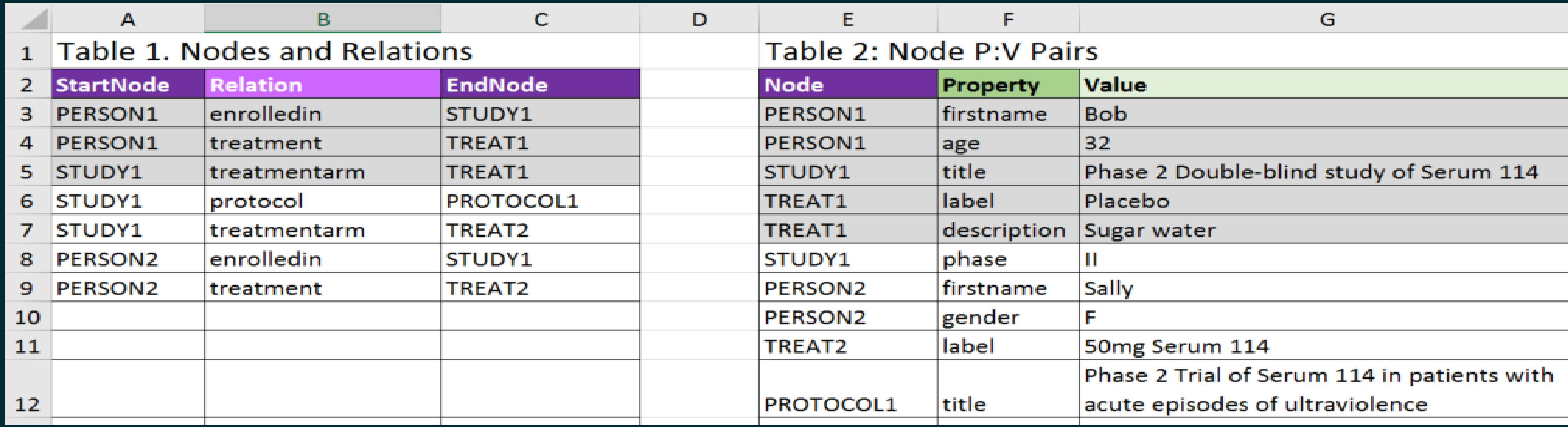


Table 1. Nodes and Relations

StartNode	Relation	EndNode
PERSON1	enrolledin	STUDY1
PERSON1	treatment	TREAT1
STUDY1	treatmentarm	TREAT1
STUDY1	protocol	PROTOCOL1
STUDY1	treatmentarm	TREAT2
PERSON2	enrolledin	STUDY1
PERSON2	treatment	TREAT2

Table 2: Node P:V Pairs

Node	Property	Value
PERSON1	firstname	Bob
PERSON1	age	32
STUDY1	title	Phase 2 Double-blind study of Serum 114
TREAT1	label	Placebo
TREAT1	description	Sugar water
STUDY1	phase	II
PERSON2	firstname	Sally
PERSON2	gender	F
TREAT2	label	50mg Serum 114
PROTOCOL1	title	Phase 2 Trial of Serum 114 in patients with acute episodes of ultraviolence

…WAIT FOR INSTRUCTOR WHEN DONE!

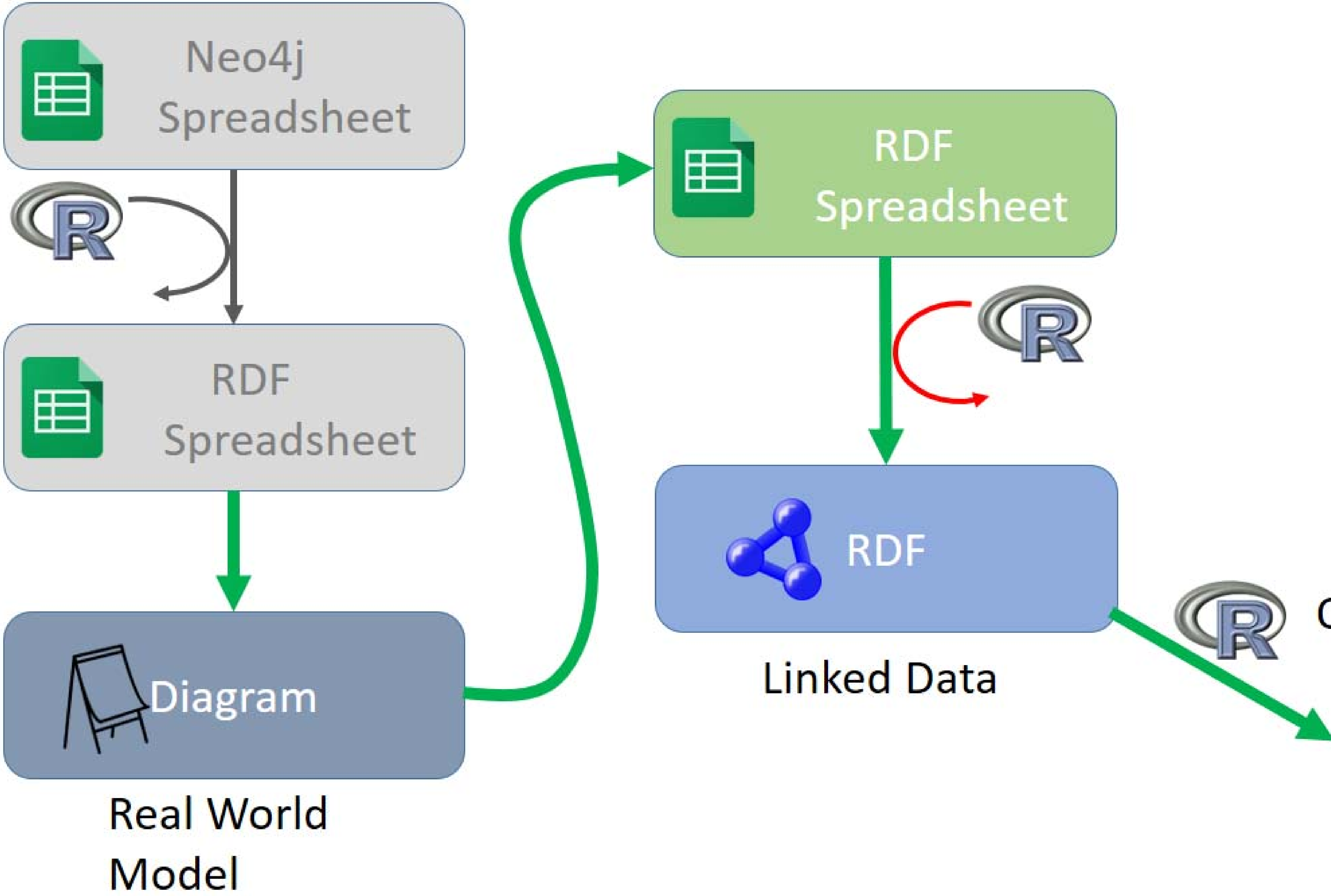
1.3 UPLOAD TO NEO4J

1.4 QUERY AND VISUALIZE

End Neo4 section

OUTLINE

- Introduction
  - Server Login
  - Data as a Graph
- Exercises
  1. Neo4j Labeled Property Graph (LPG)
  2. **Resource Description Framework (RDF)**
- Demonstrations (time permitting)



Neo4j Spreadsheet

RDF Spreadsheet

Diagram

Real World Model

RDF Spreadsheet

RDF

Linked Data

2.1 RDF SPREADSHEET TO DIAGRAM

NEO4J CONCEPTS TO THE RDF DIAGRAM

Handout: RDF Diagram

The Revised RDF Diagram

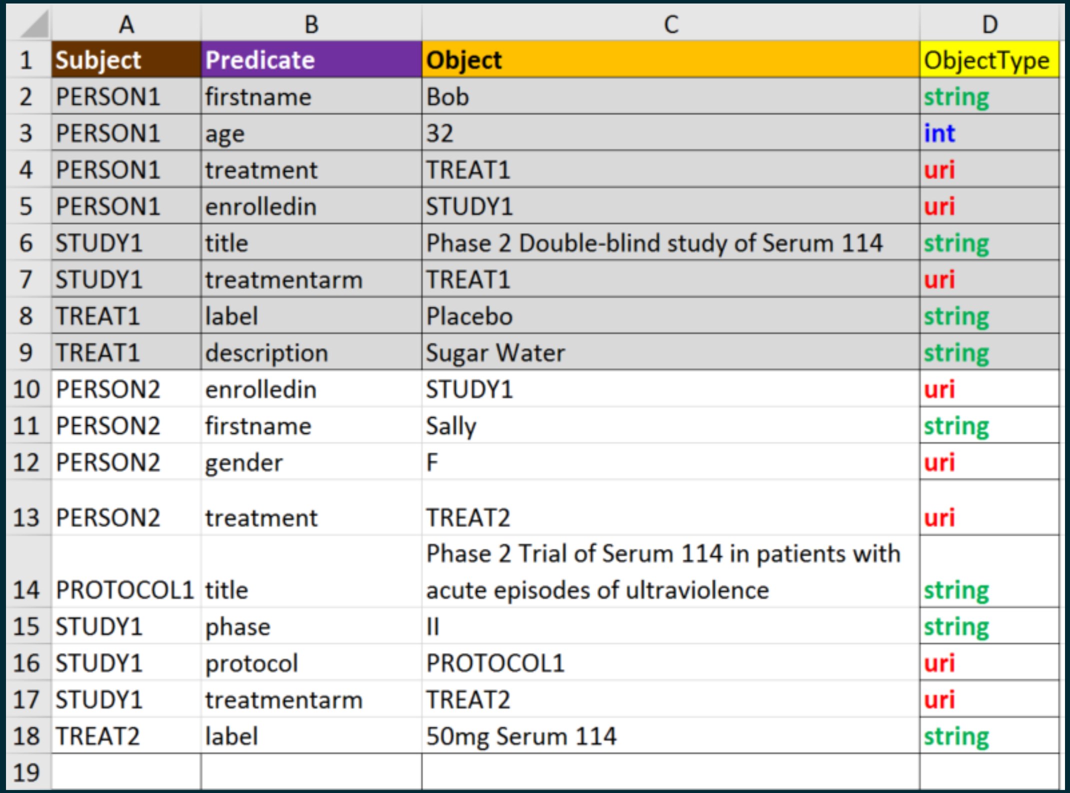
OBJECT TYPES

Type	Description
uri	Object links to another node or *could* link to another node
string	Character string/value that does not link to other nodes
int	Integer number. No link to other nodes

2.2 ASSIGN OBJECT TYPE

1. Add the Object Type in each row.

2.2 ASSIGN OBJECT TYPE



	A	B	C	D
1	**Subject**	**Predicate**	**Object**	ObjectType
2	PERSON1	firstname	Bob	string
3	PERSON1	age	32	int
4	PERSON1	treatment	TREAT1	uri
5	PERSON1	enrolledin	STUDY1	uri
6	STUDY1	title	Phase 2 Double-blind study of Serum 114	string
7	STUDY1	treatmentarm	TREAT1	uri
8	TREAT1	label	Placebo	string
9	TREAT1	description	Sugar Water	string
10	PERSON2	enrolledin	STUDY1	uri
11	PERSON2	firstname	Sally	string
12	PERSON2	gender	F	uri
13	PERSON2	treatment	TREAT2	uri
14	PROTOCOL1	title	Phase 2 Trial of Serum 114 in patients with acute episodes of ultraviolence	string
15	STUDY1	phase	II	string
16	STUDY1	protocol	PROTOCOL1	uri
17	STUDY1	treatmentarm	TREAT2	uri
18	TREAT2	label	50mg Serum 114	string
19				

2.3 CREATE RDF (TTL) FILE

2.4 QUERY AND VISUALIZE

NEO4J

WHICH TO CHOOSE: NEO4J OR RDF?

- Graph path traversal
- Process flow
- Changing (transactional) data
- Where connections/relations(links) are key

WHICH TO CHOOSE: NEO4J OR RDF?

RDF

- Classification (ontologies)
- Rules and Logic
- Datatyping, Time concepts
- Non-transactional data

OUTLINE

- Introduction
  - Server Login
  - Data as a Graph
- Exercises
  1. Neo4j Labeled Property Graph (LPG)
  2. Resource Description Framework (RDF)
- **Demonstrations (time permitting)**

BUT FIRST: ACKNOWLEDGEMENTS

- PhUSE - server costs
- **Lauren White, Wendy Dobson, Tora Whitworth** and the entire PhUSE admin team
- **Chris Decker** - server cloning
- **Johannes Ulander** - Neo4j SDTM Demo, exercises review, assistant
- **Scott Bahlavooni, Ian Fleming** - assistants
- **Mark Foxwell, Paula Finch** - Workshop coordinators
- ...and everyone else I forgot to mention
- ...and: **YOU**!

BUT SECOND: RESOURCES

- Workshop materials, including the SPARQL and CYPHER scripts, plus PDF of this present
  https://github.com/phuse-org/LinkedDataWorkshop/Annual2017-EU

NEO4J RESOURCES

Recommended Overview	https://neo4j.com/developer/graph-database/
Overview of graph db and Neo4j [optional]	https://youtu.be/U8ZGVx1NmQg [45min]
Intro to Cypher	https://www.youtube.com/watch?v=1TSBXZMv6tc [49min]
Graph Modeling	https://www.youtube.com/watch?v=AaJS-DGBQX4 [42min]

RDF AND PROTEGE RESOURCES

Introduction to Semantic Web	http://www.cambridgesemantics.com/semantic-university/introduction-semantic-web
What is Linked Data?	http://www.cambridgesemantics.com/semantic-university/what-linked-data
Introduction to Linked data	http://www.cambridgesemantics.com/semantic-university/introduction-linked-data
Protege Application	https://www.youtube.com/watch?v=8Nf2xf5akoM