Data Extraction, Transformation, and Integration for Loading Data into a Psychiatric Research Data Warehouse

Data Analytics: Biomedical and Health Informatics 509

Presentation by Joseph Miles

Introduction

"The decentralized nature of our scientific communities and healthcare systems has created a sea of valuable but incompatible electronic databases." (Sujansky, 2001)

Data Warehouse: (Schmidt and Prado, 2014)

1. Accuracy
2. Reliability
3. Consistency
4. Accessibility

Also: completeness,

and interpretability

Sujansky, W. (2001). Heterogeneous database integration in biomedicine. *Journal Of Biomedical Informatics*. 34(4): 285-298. Doi:10.1006/jbin.2001.1024

Schmidt, SO; Prado, EPV. (2014) IT Architecture and Information Quality in Data Warehouse and Business Intelligence Environments. *IGI Global*. 6: 121-127. DOI: 10.4018/978-1-4666-4892-0.ch06

Integration

Heterogeneous methods of labeling and storing data confound data integration

- Different labels, same meaning
- Same labels, different meaning
- Same label, same meaning but different semantics

ETL: Extraction, Transformation, and Loading

- To maintain the integrity of the data, accurate ETL is of highest importance (Schmidt and Prado, 2014)

Schmidt, SO; Prado, EPV. (2014) IT Architecture and Information Quality in Data Warehouse and Business Intelligence Environments. *IGI Global*. 6: 121-127. DOI: 10.4018/978-1-4666-4892-0.ch06

Heterogeneous Grouping

Patients with the same diagnosis do not always have the same problems

- Depression:
  - Sad
  - Anxious
  - Empty
  - Hopeless
  - Irritable
  - Feelings of guilt, worthlessness, or helplessness
  - Decreased energy or fatigue
  - Feeling restless or having trouble sitting still
  - Difficulty sleeping
  - Oversleeping

NIMH: Depression. https://www.nimh.nih.gov/health/topics/depression/index.shtml

RDoC (Research Domain Criteria)

We tend to treat patients by diagnosis rather than treating the problem(s) that established the original diagnosis.

RDoC

- RDoC "*integrates many levels of information (from genomics to self-report) to better understand basic dimensions of functioning underlying the full range of human behavior from normal to abnormal.*" (NIMH, RDoC)
- Beyond categorization:
  - "*The best time to address a mental illness is before the appearance of symptoms that disrupt daily life*." https://www.nimh.nih.gov/about/strategic-planning-reports/strategic-objective-2.shtml

NIMH. Research Domain Criteria (RDoC). https://www.nimh.nih.gov/research-priorities/rdoc/index.shtml

RDoC Studies

Accumulate data:

1. Genomics and inheritance

   Currently known candidate genes in the psychiatric "Positive Valence System" are only loosely correlated with observed behaviors (phenotype) [r^2 values ranging from 0.181 to 0.233, Hess et al, 2016].

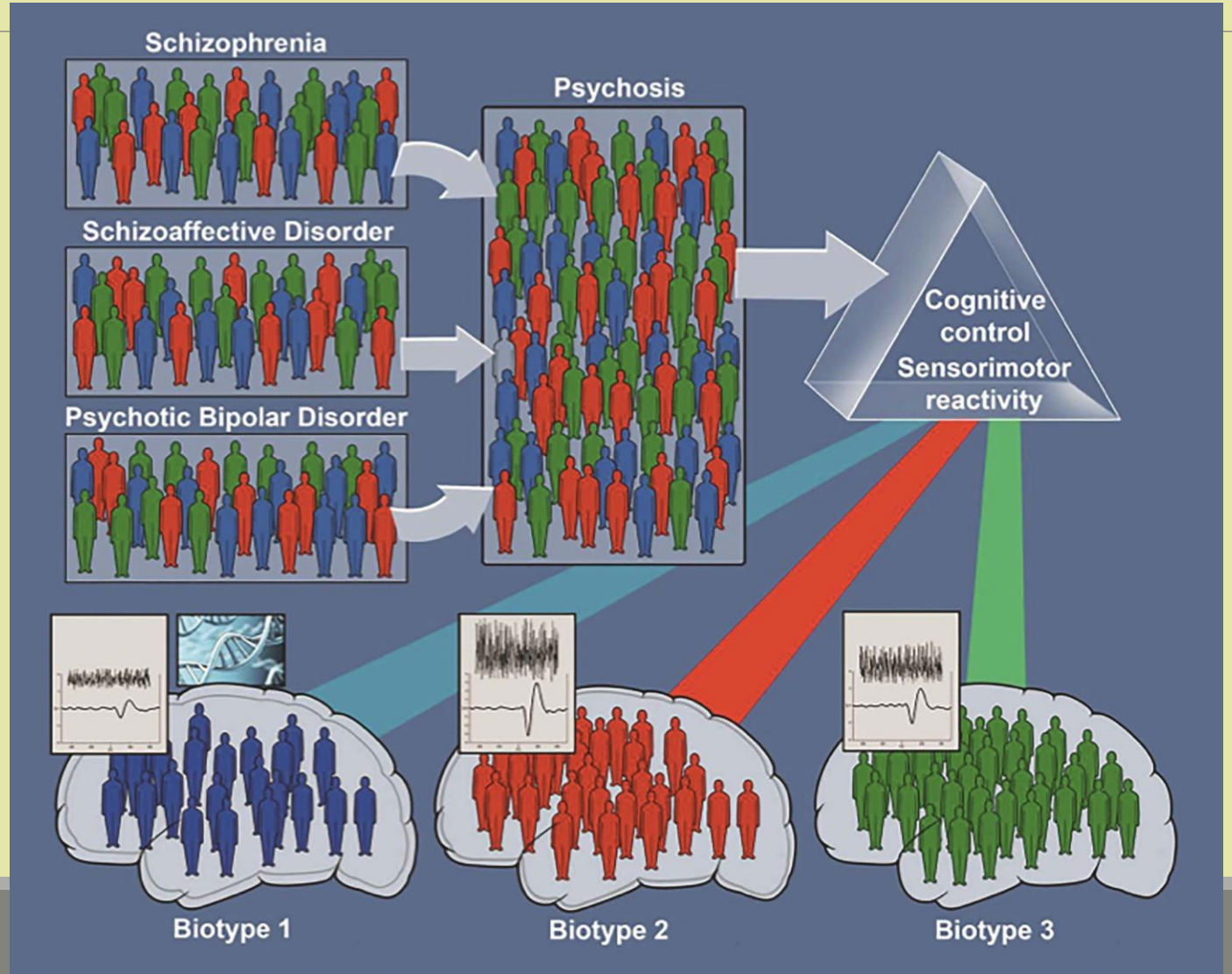
2. Environmental Considerations
   - Highest level of education
   - Employment status
   - Household income

3. Self-report of symptomology
   - Discover phenotype – genotype linkage
   - Assess differences in clinical presentation at different patient ages
     - Longitudinal study

Hess, JL; Kawaguchi, DM; Wagner, KE; Faraone, SV; Glatt, SJ. (2016). The influence of genes on "positive valence systems" constructs: A systematic review. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, *171*(1), 92-110.

RDoC Studies

RDoC
Research Domain Criteria Initiative



Upstate RDoC Family Study

UPSTATE MEDICAL UNIVERSITY
Knowing changes everything™

SUNY Upstate Medical University has been evaluating constructs from the Positive Valence System

- Research subject (proband) is a child with a “Mental Health Issue.”
- Family Study:
  - Also examine biological parent(s)
  - Also examine biological siblings (up to 4 siblings)
- Each research subject (proband and family members) is documented in a spreadsheet with 96 variables
  - As of 10/28/2016, there were 465 families and 1362 total research subjects
  - 1334 subjects either donated whole blood or saliva for genetic testing

What are we creating, short term?

NIMH requires RDoC research sites to submit cumulative data to the data warehouse every 6 months (> 80 hours?!?)

Upstate is submitting 2 files:

1. Family Studies Demographics (53 variables per record)
   - A comma separated variables (csv) file with each record representing a family
2. Research Subjects Demographics (51 variables per record)
   - A csv file with each records representing a research subject
     - Research subject data also requests parent/sibling ID data

ElementName	DataType	Size	Required	ElementDescription	ValueRange
subjectkey	GUID		Required	The NDAR Global Unique Identifier (GUID) for research subject	NDAR*
src_subject_id	String	20	Required	Subject ID how it's defined in lab/project	
interview_date	Date		Required	Date on which the interview/genetic test/sampling/imaging was completed. MM/DD/YYYY	
interview_age	Integer		Required	Age in months at the time of the interview/test/sampling/imaging.	0 :: 1260
gender	String	20	Required	Sex of the subject	M;F
race	String	30	Recommended	Race of study subject	American Indian/Alaska Native; Asian; Hawaiian or Pacific Islander; Black or African American; White; More than one race; Unknown or not reported
ethnicity	String	30	Recommended	Ethnicity of participant	Hispanic or Latino; Not Hispanic or Latino; Unknown
subjectkey_father	GUID		Recommended	The NDAR Global Unique Identifier (GUID) for subject's biological father	NDAR*
src_father_id	String	100	Recommended	site specific father ID	
ques_agebf	Integer		Recommended	Age, Biological Father(in months)	0 :: 1200; 999

Data dictionary, Family Studies Demographics

0. subjectkey
1. src_subject_id
2. interview_date
3. interview_age
4. gender
5. race
6. ethnicity
7. subjectkey_father
8. src_father_id
9. ques_agebf
10. ques_genderbf
11. bio_father_race
12. bio_father_ethnicity
13. et2b
14. dem_08b
15. fa_maritalstatus
16. fa_householdincome
17. subjectkey_mother
18. src_mother_id
19. ques_agebm
20. ques_genderbm
21. bio_mother_race
22. bio_mother_ethnicity
23. et2a
24. dem_08a
25. mo_maritalstatus
26. mo_householdincome
27. subjectkey_sibling1
28. src_sibling1_id
29. ques_age1
30. ques_gender1
31. sib1_race
32. sib1_ethnicity
33. subjectkey_sibling2
34. src_sibling2_id
35. ques_age2
36. ques_gender2
37. sib2_race
38. sib2_ethnicity
39. subjectkey_sibling3
40. src_sibling3_id
41. ques_age3
42. ques_gender3
43. sib3_race
44. sib3_ethnicity
45. subjectkey_sibling4
46. src_sibling4_id
47. ques_age4
48. ques_gender4
49. sib4_race
50. sib4_ethnicity
51. mother_other_race
52. father_other_race

Data dictionary, Research Subject Demographics

0. ndar_subjectkey
1. src_subject_id
2. interview_date
3. interview_age
4. gender
5. race
6. ethnic_group
7. phenotype
8. phenotype_description
9. twins_study
10. sibling_study
11. family_study
12. family_user_def_id
13. subjectkey_mother
14. src_mother_id
15. subjectkey_father
16. src_father_id
17. subjectkey_sibling1
18. src_sibling1_id
19. sibling_type1
20. subjectkey_sibling2
21. src_sibling2_id
22. sibling_type2
23. subjectkey_sibling3
24. src_sibling3_id
25. sibling_type3
26. subjectkey_sibling4
27. src_sibling4_id
28. sibling_type4
29. zygosity
30. sample_taken
31. sample_id_original
32. sample_description
33. biorepository_name
34. sample_id_biorepository
35. patient_id_biorepository
36. cell_id_original
37. cell_id_biorepository
38. agre_subject_id
39. sfari_subject_id
40. cpea_site
41. cpea_id
42. blood_id
43. adi_dx
44. ados_dx
45. agp_family_id
46. agp_subject_id
47. site
48. comments_misc
49. week
50. study

What to Integrate?

Each family has up to 7 people = up to 672 variables to extract from for each family!

- 2 variables are “family ID” and “individual ID” which were used as a primary key and for generating the Upstate ID
- 12 other variables for each individual were also needed.

A second file contains a deidentified NIMH ID code for each subject.

- NIMH ID is called a GUID [Global Unique Identifier] and is required for every research subject in both upload files.

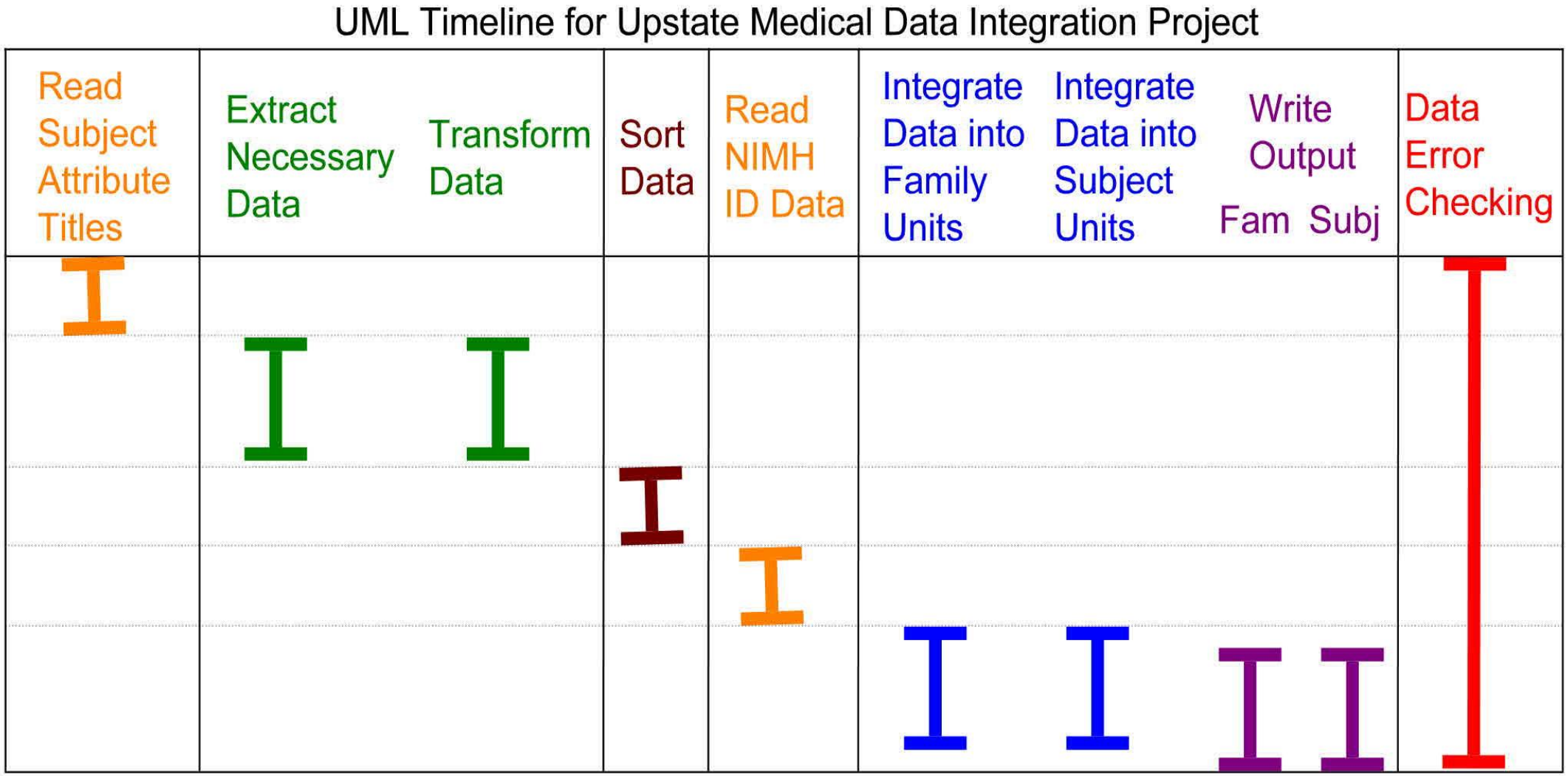
Simplified Data Dictionary

Table 1

Table of input variables needed for the data dictionary

Variable Name	Transformation necessary: Convert data to NIMH format (*Yes* or *No*)
Family ID *(Part of composite key)	No
Individual ID *(Part of composite key)	No
Study Visit Date	No
Mental Health	No
Date of Birth	Yes
Gender	Yes
Racial Category	Yes
Ethnic Category	Yes
Highest Level of Education	Yes
Employment Status	Yes
Marital Status	No
Household Annual Income	No
Blood Sample ID	Yes
Saliva Sample ID	Yes

Planning the application

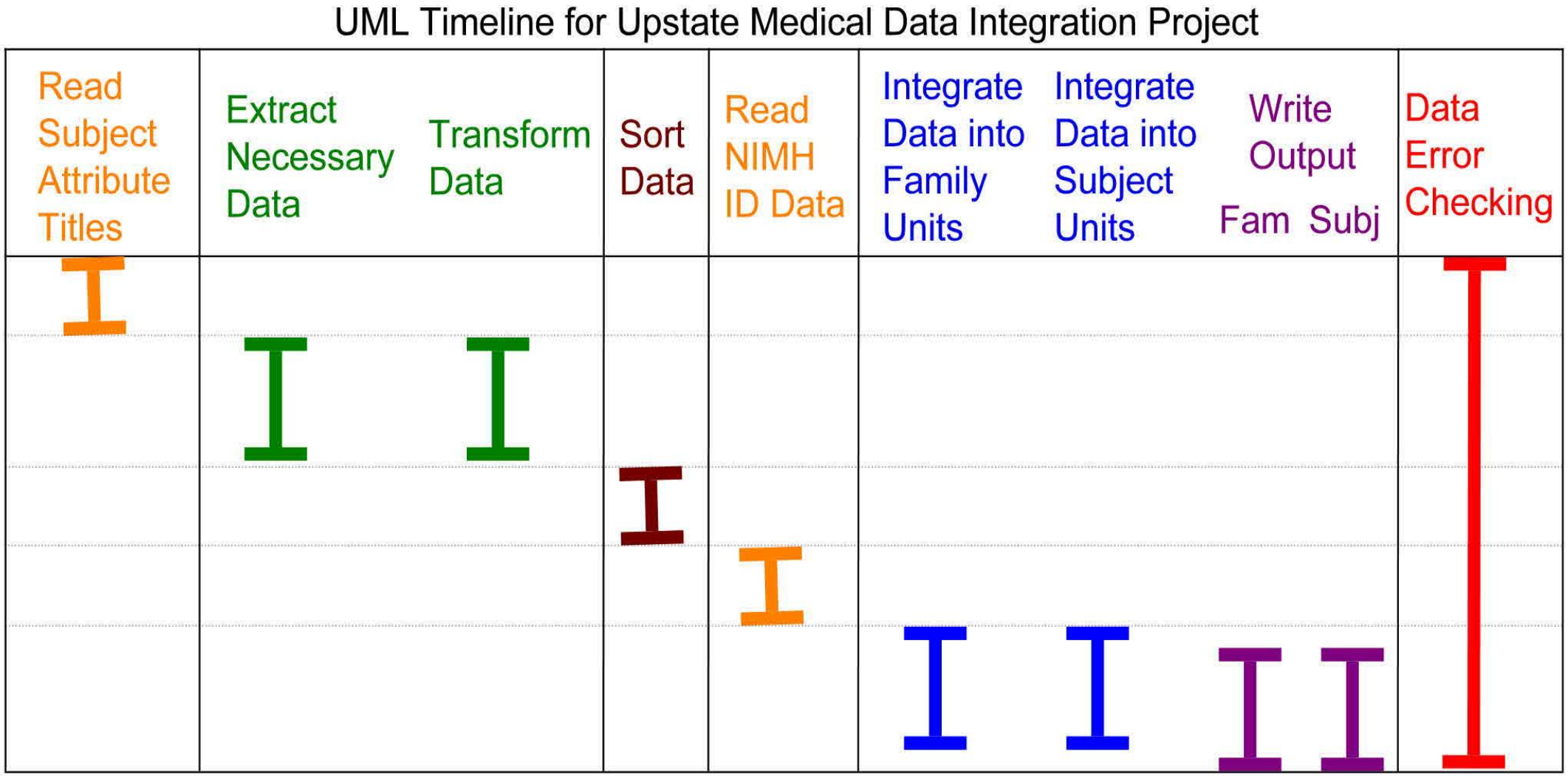


Read Variable Titles

FamID	Ind. ID	Date of Birth	Gender 1=Male 2=Female	Racial Category 1=Am Indian or Alaska Native 2=Asian 3=Black/African American 4=Nat Hawaiian or other Pacific Islander 5=White 6=Two or	Racial Category (Other specify)	Ethnic Category 1=Hispanic or Latino; 2=Not Hispanic or Latino 3=Does not wish to say	Highest Level of Education 1=Less than HS 2=HS diploma or equivalent 3=Some college - no degree 4=Postsecondary non-degree award	Employment Status 1=Unemployed 2=Employed part-time 3=Employed full-time 4=Homemaker 5=Retired

Marital Status 1=Married 2=Never married 3=Not married - living tog 4=Divorced	Household Annual Income 0=NA (child)	Parent agrees to repositories 1=yes both 2=Local Study only	Parent agrees to be contacted and child after 18 1=Yes 2=No	Blood Draw Completed 1=Yes 2=No	Date Blood Draw Completed 0=No blood draw	Any Issues (Nurse) 1=Yes 2=No	Blood Sample sent to RUCDR 1=Yes 2=No 0=NA (no blood draw)

Planning the application: continued



Discussion: Data integration for bioinformatics

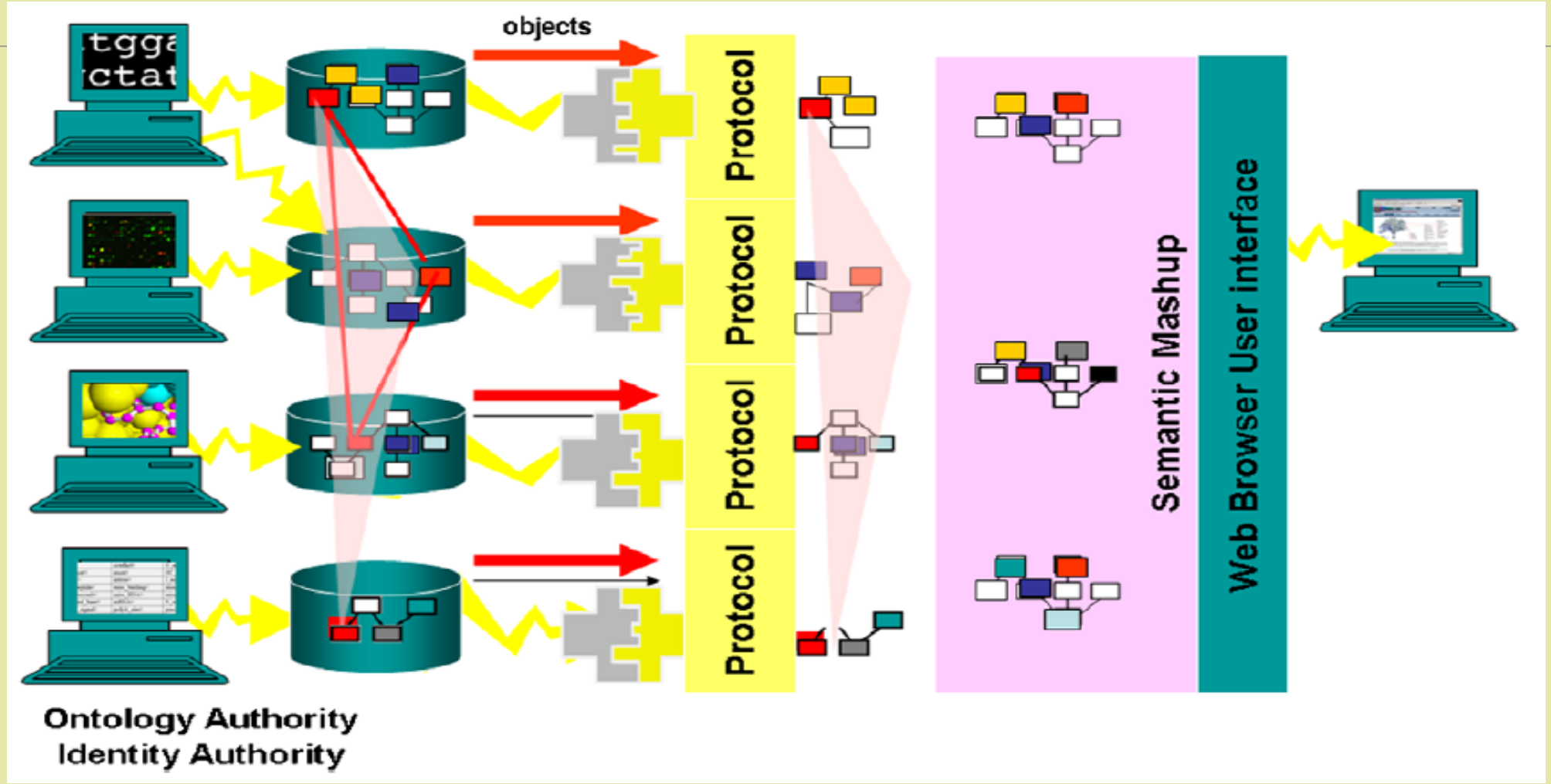
Goble and Stevens, 2008

- Need common, shared identities and names
- Need shared semantics
- Need shared access mechanisms
- Need to adhere to standards
- Need to explicitly state collection policies and governance
- Need to balance curation with data usability

Goble, C; Stevens, R. (2008) State of the nation in data integration for bioinformatics. Journal of Biomedical Informatics. 41: 687-693. doi:10.1016/j.jbi.2008.01.008

Discussion: Resource Description Framework (RDF)

And the future of data integration. Goble and Stevens, 2008



Goble, C; Stevens, R. (2008) State of the nation in data integration for bioinformatics. Journal of Biomedical Informatics. 41: 687-693. doi:10.1016/j.jbi.2008.01.008

Conclusion

- ✓Precision in extraction and transformation is essential for data to be used in data analysis and querying.
- ✓Python programming is ideal for ETL due to consistent and quick manipulation of data.
  - There are many possibilities for human error with manual ETL
- ✓Data Warehouses provide a mechanism for accuracy, reliability, consistency as well as remote access for analysis.

References:

Goble, C; Stevens, R. (2008) State of the nation in data integration for bioinformatics. *Journal of Biomedical Informatics*. 41: 687-693. doi:10.1016/j.jbi.2008.01.008

Hess, JL; Kawaguchi, DM; Wagner, KE; Faraone, SV; Glatt, SJ. (2016). The influence of genes on "positive valence systems" constructs: A systematic review. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, *171*(1), 92-110.

NIMH Websites Retrieved 4/30/2017: Depression. https://www.nimh.nih.gov/health/topics/depression/index.shtml; Research Domain Criteria (RDoC). https://www.nimh.nih.gov/research-priorities/rdoc/index.shtml; Strategic Objective 2. https://www.nimh.nih.gov/about/strategic-planning-reports/strategic-objective-2.shtml

Schmidt, SO; Prado, EPV. (2014) IT Architecture and Information Quality in Data Warehouse and Business Intelligence Environments. *IGI Global*. 6: 121-127. DOI: 10.4018/978-1-4666-4892-0.ch06

Sujansky, W. (2001). Heterogeneous database integration in biomedicine. *Journal Of Biomedical Informatics*. 34(4): 285-298. Doi:10.1006/jbin.2001.1024

Any questions?: Joseph Miles <jmiles3@oswego.edu>