Automating Population Health Studies through Semantics and Statistics

Alexander New, Miao Qi, Shruthi Chari, Sabbir M. Rashid, Oshani Seneviratne, James P. McCusker, John S. Erickson, Deborah L. McGuinness, and Kristin P. Bennett

SemStats Talk – Oct 27, 2019

Rensselaer

why not change the world?®

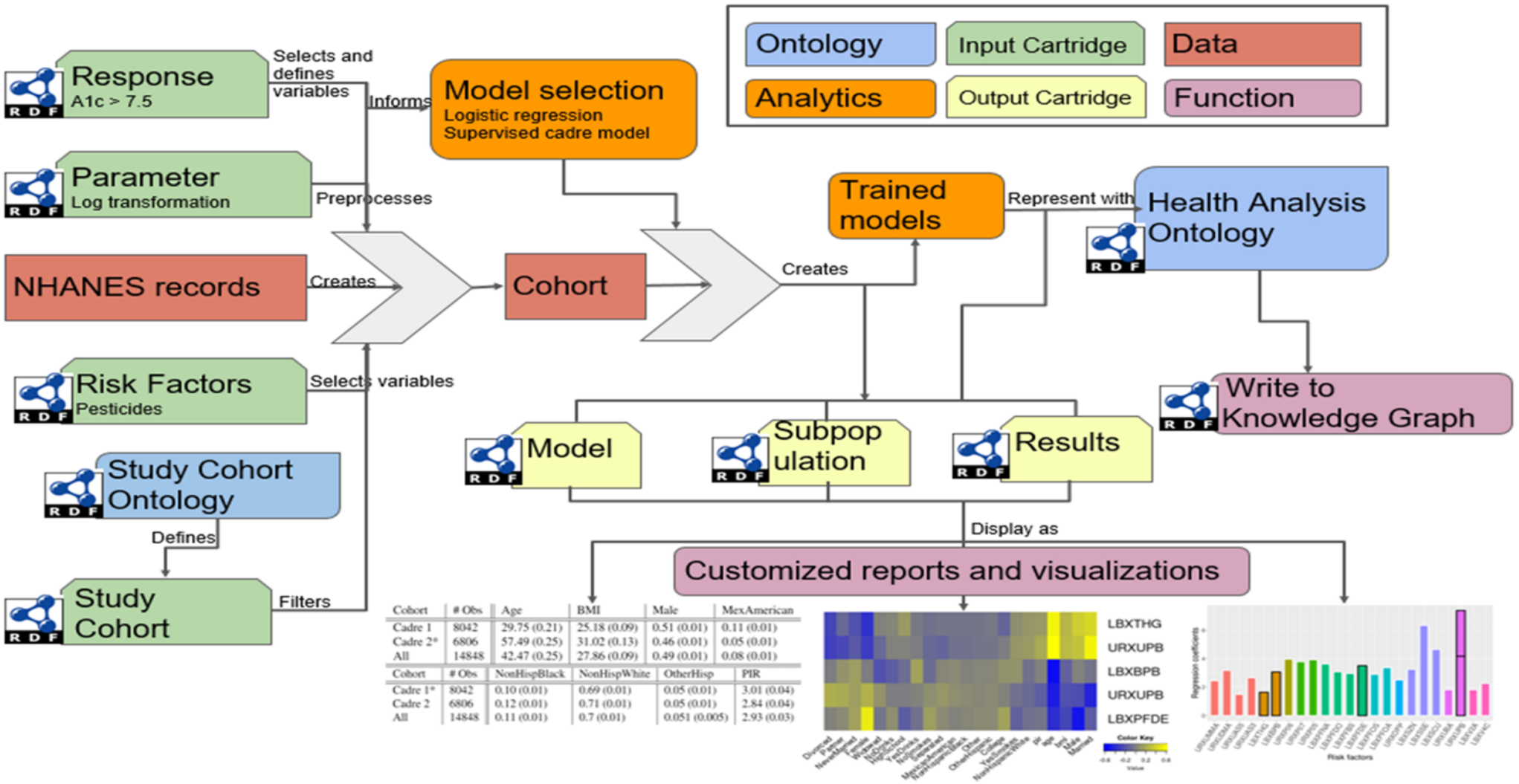
Project Summary

We use **ontologies** and **knowledge graphs** to represent data preparation and workflow modeling in a reusable and reproducible way using **Semantically-Targeted Analysis** with reusable modular knowledge called **cartridges**.

Use Case

For *[discovered subpopulation]* in *[study cohort]*, does *[risk factor]* have a significant association with *[chronic health condition]*?.

Semantically Targeted Analytics (STA) Framework



Health Analysis Ontology (HAO)

- It supports modeling of processes, components, models, variables and factors involved in a health analysis pipeline
- It provides a vocabulary necessary to model the reusable components of an analysis (sio:Analysis) implemented by an analysis workflow (hao:AnalysisWorkflow) that we store in cartridges (hao:Cartridge).
- Ontologies currently used in STA

Ontology
Health Analysis Ontology
Study Cohort Ontology
Children's Health Exposure Analysis Resource
The Statistical Methods Ontology
Semanticscience Integrated Ontology
National Cancer Institute Thesaurus
Ontology for Biomedical Investigations
The PROV Ontology
Ontology of Biological and Clinical Statistics
DC Terms
Simple Knowledge Organization System

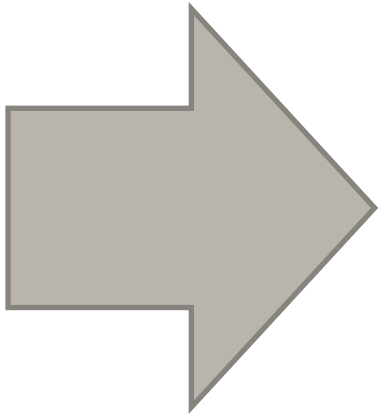
Cartridges: Application-specific Vocabularies That Extend A KG's Range Of Applicable Analyses

Response Variable
Analysis concepts and **background domain axioms** necessary to model a given health condition

Study cohort
Inclusion criteria used to determine if a given subject may be included in a study

Risk factor
Rules for modeling **semantically-similar risk factor categories** (e.g., pesticides)

Parameter
Rules to **complete chosen analysis workflow**, such as potential hyperparameter configurations to search over

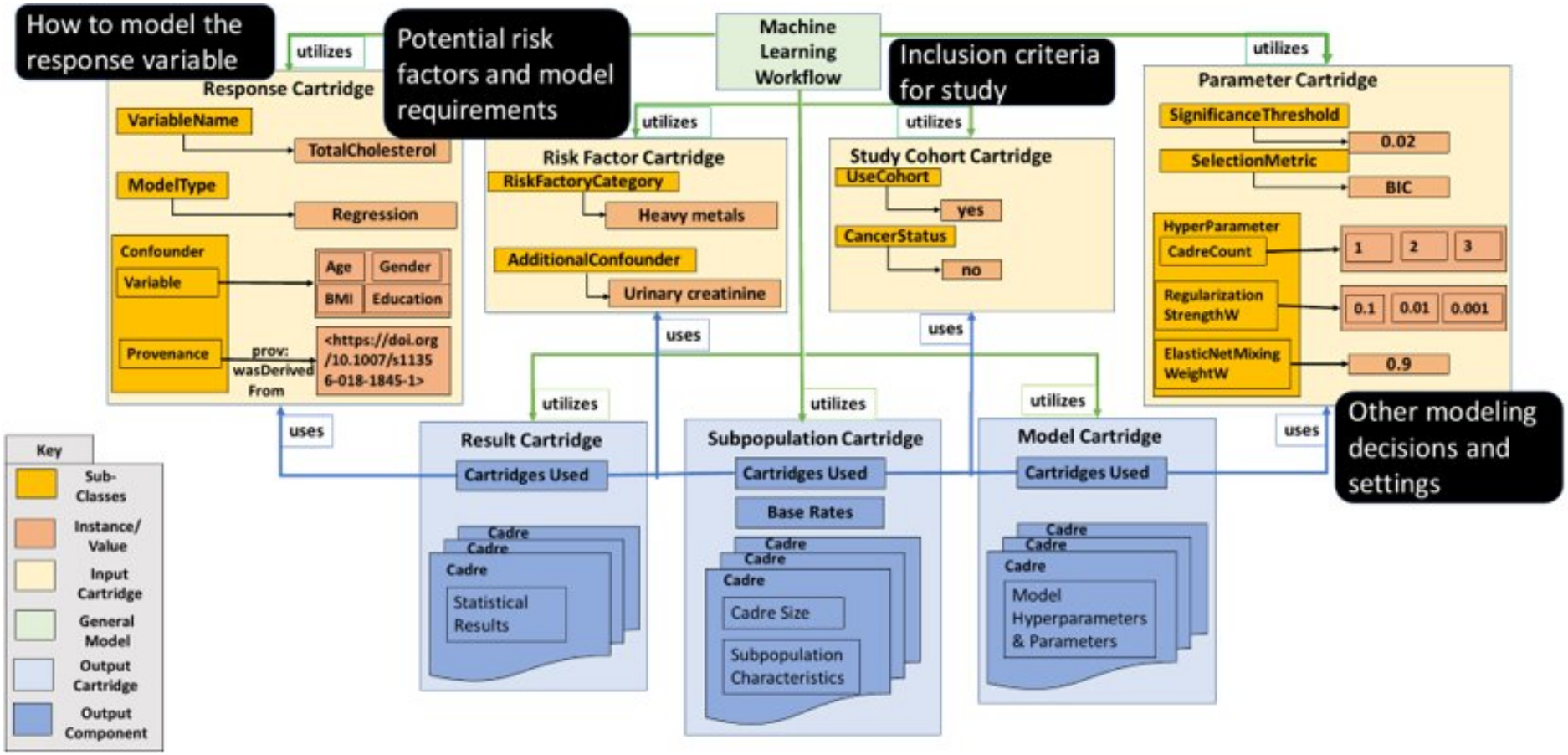


Model
Chosen **hyperparameters** and **optimal model**

Subpopulation
Summary statistics characterizing **discovered subpopulations**

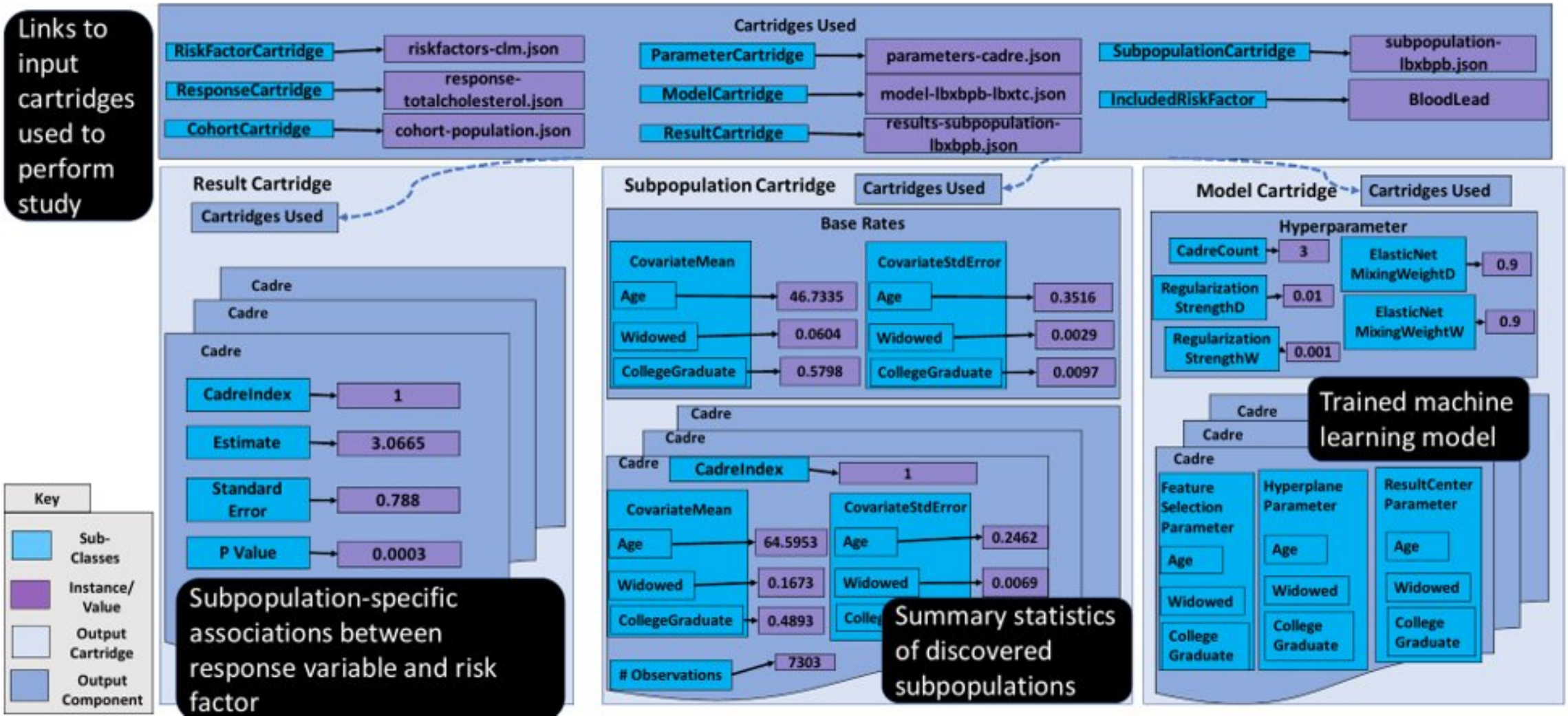
Results
Statistical quantification of **subpopulation-specific discovered associations** between the risk factor and the response variable

Input Cartridges (Yellow): Define Components Of A Risk Study



- Cartridges encode best practices for both analytics modeling and specific domains
- This allows rigorous studies to be constructed, represented, and interpreted by people with diverse background knowledge levels

Output Cartridges (Light Blue) Store Statistical Findings



Supervised Cadre Models For Subpopulation-discovery And Risk Analysis

- Supervised learning framework for heterogeneous data
  - Simultaneously divides observations into subpopulations (cadres) and learns subpopulation-specific risk models
  - E.g., subjects below a threshold based on age and BMI have a significant association between blood cadmium and systolic blood pressure

$$f(x) = g(x_{F_C})^T e(x_{F_T})$$

- Risk score function (e.g., for having hypertension)

$$e^m(x_{F_T}) = (W_m)^T x_{F_T}$$

- Risk score function for cadre *m*

$$g_m(x_{F_C}) = \frac{e^{-\gamma||x_{F_C} - c^m||_d^2}}{\sum_{m'} e^{-\gamma||x_{F_C} - c^{m'}||_d^2}}$$

- Probability that observation *x* belongs to cadre *m*

$$||z||_d = \left(\sum_p |d_p|(z_p)^2\right)^{1/2}$$

- Semimetric used for cadre-assignment

Example: Identify Risk Factors Associated With High Total Cholesterol

Response
Total cholesterol is a continuous response variable.

Response
Control for subjects' age, Body Mass Index (BMI), Poverty Income Ratio (PIR),smoking habits, drinking habits, gender, marital status, and education level.

Study cohort
All available NHANES subjects

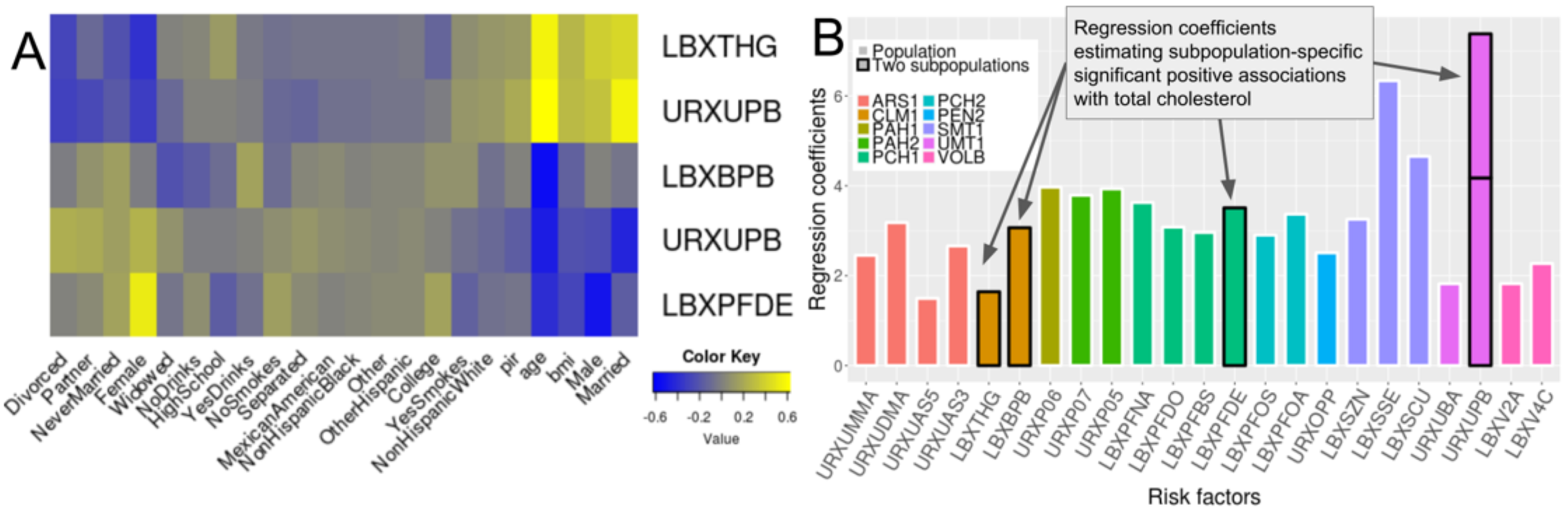
Risk Factor
201 environmental exposure risk factors divided into 17 categories

Parameter
Train models with M = 1, 2 and 3 cadres and choose best one using BIC for model selection

Parameter
Standardize risk factor measurements

Parameter
Significance threshold of $\alpha = 0.02$ for GLM hypothesis tests

Example: Identify Risk Factors Associated With High Total Cholesterol



- Heatmap of subpopulation means that have significant risk factor associated with high total cholesterol
- Significant positive regression coefficients associated with high total cholesterol

Conclusions

STA is a framework for performing end-to-end analyses on semantically-heterogeneous data

Via *cartridges*, novel statistical findings are written to a collective knowledge graph for future querying and reference.

Questions?

Thank You!

Points of contact:

Alexander New, newa2@rpi.edu

Kristin P. Bennett, bennek@rpi.edu

Deborah L. McGuinness, dlm@cs.rpi.edu