

# The WPU YRBS (Youth Risk Behavior Survey) Data Mart

## EXPLORATIONS: Research, Scholarship and Creative Expression at WPU

Ana K. Ocampo and Cyril S. Ku, Department of Computer Science, William Paterson University

### Outline

- Overview
- Project Description
- Server Architecture
- WPU Data Warehousing Process
- YRBSS (CDC)
- WPU YRBS Data Mart
- Project Phases
- Acknowledgement
- Bibliography

### Overview

The WPU YRBS (Youth Risk Behavior Survey) Data Mart is a multi-user and parallel access platform in a client-server environment for behavioral research in the Department of Public Health and for data analytic and data mining research in the Department of Computer Science at William Paterson University. This presentation outlines our effort in data extraction, data transformation, and data loading processes from the YRBSS (Youth Risk Behavior Surveillance System) at CDC (Centers for Disease Control and Prevention). The architecture of our data mart as well as the future plan of data mining activities will be presented.

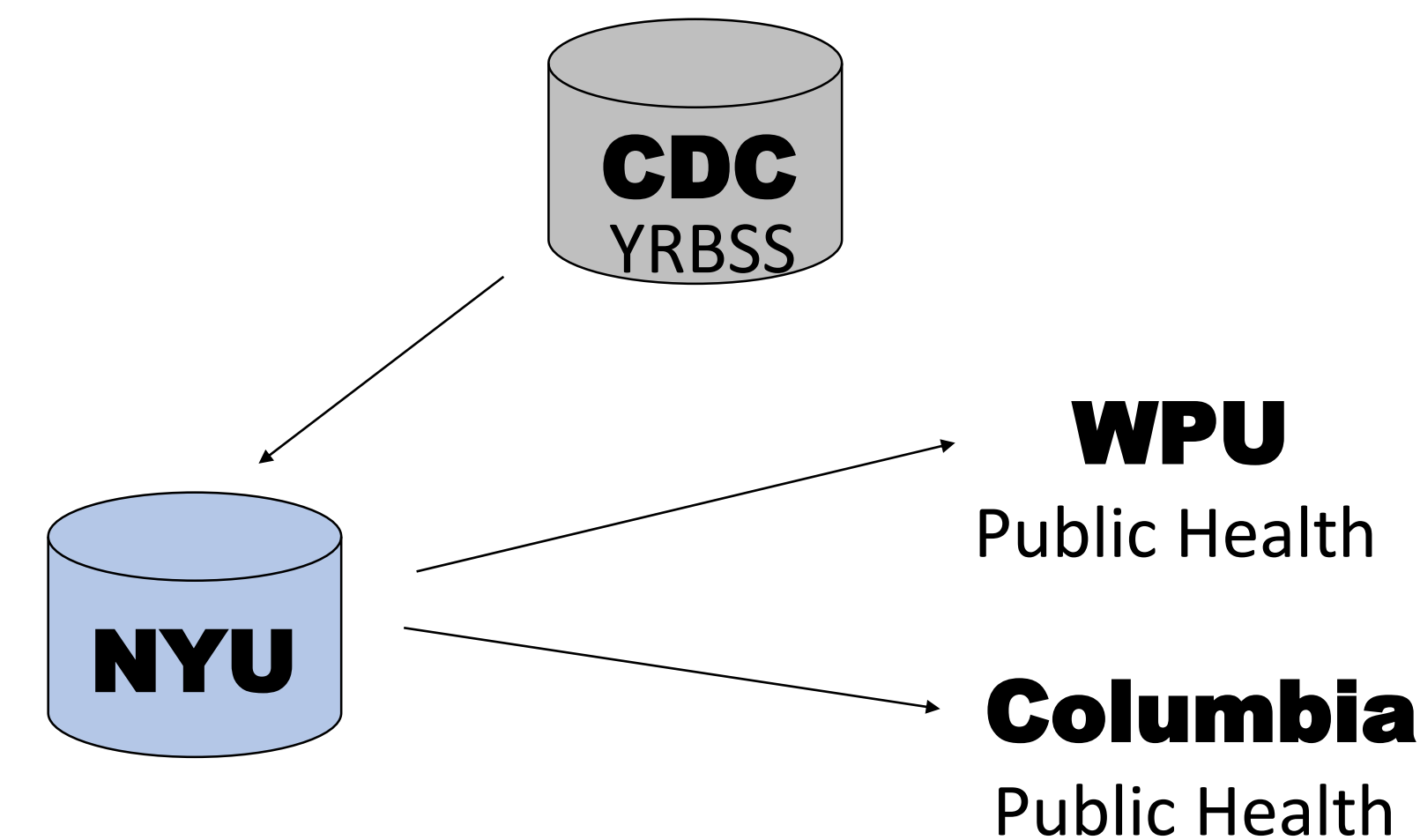
### Project Description

- An inter-disciplinary research project:
  - Department of Public Health – Dr. Corey H. Basch, Dr. Alex Kecojevic
  - Department of Computer Science – Dr. Cyril S. Ku, Dr. Ali Moghani, Ms. Ana K. Ocampo (Research Assistant)
- Data Warehouse
  - YRBSS (Youth Risk Behavior Surveillance System) from CDC (Centers for Disease Control and Prevention)
- Data Mart at New York University
  - Accessed by Columbia University (Dr. Charles E. Basch)
  - Accessed by WPU (Dr. Corey H. Basch)
  - Data feed no longer exists

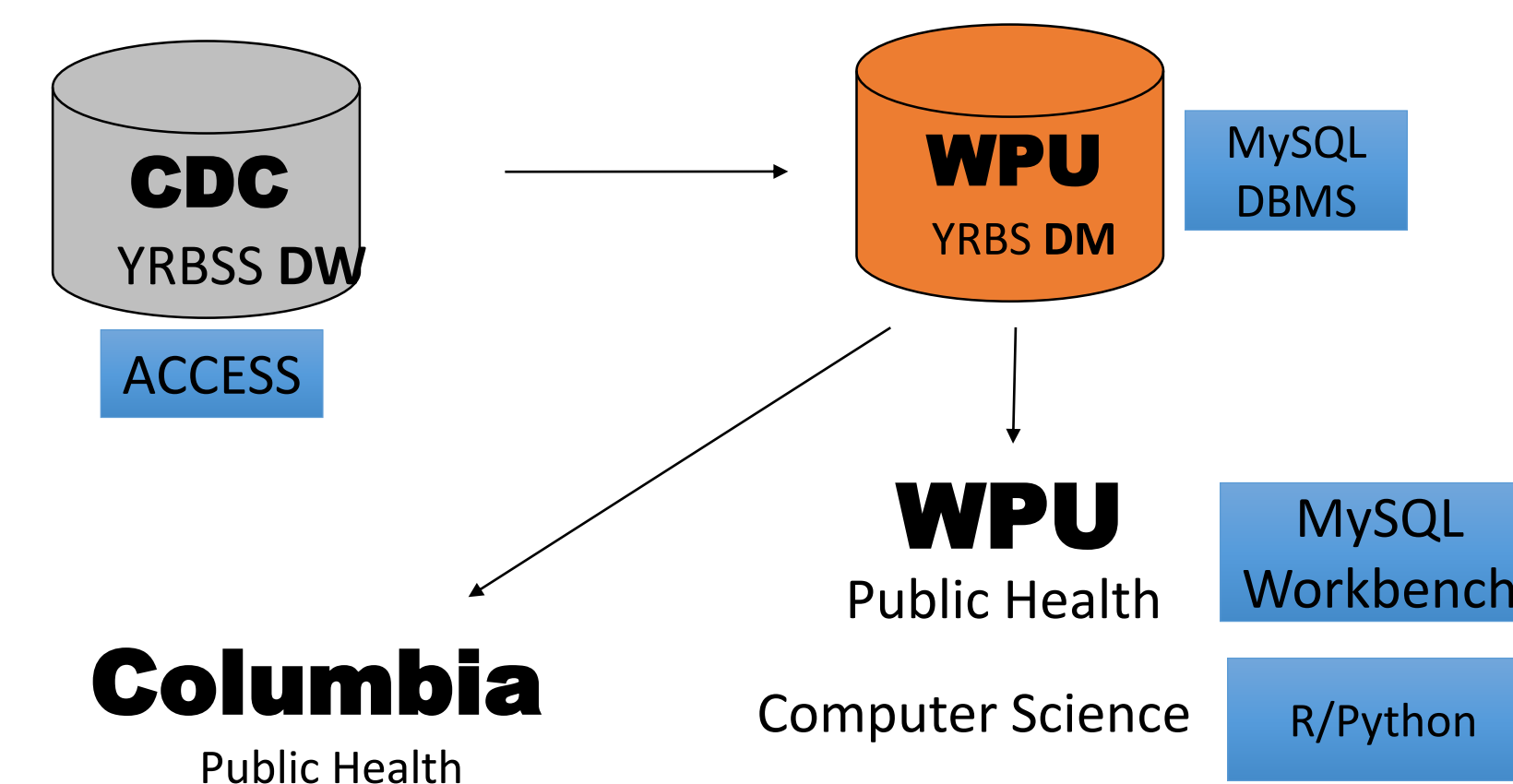
### Project Description

- Create a Data Mart at William Paterson University
  - The WPU YRBS Data Mart consists of Youth Risk Behavior Survey data
  - Replaced the data feed from NYU
  - Accessed by Dr. Corey Basch and Dr. Alex Kecojevic for their behavioral research on American youth in the Department of Public Health
  - Maintained by the Department of Computer Science. The data mart will be used for data warehousing (big data), data analytics, and data mining research (Dr. Cyril Ku and Dr. Ali Moghani)
  - Will be accessed by WPU and Columbia University

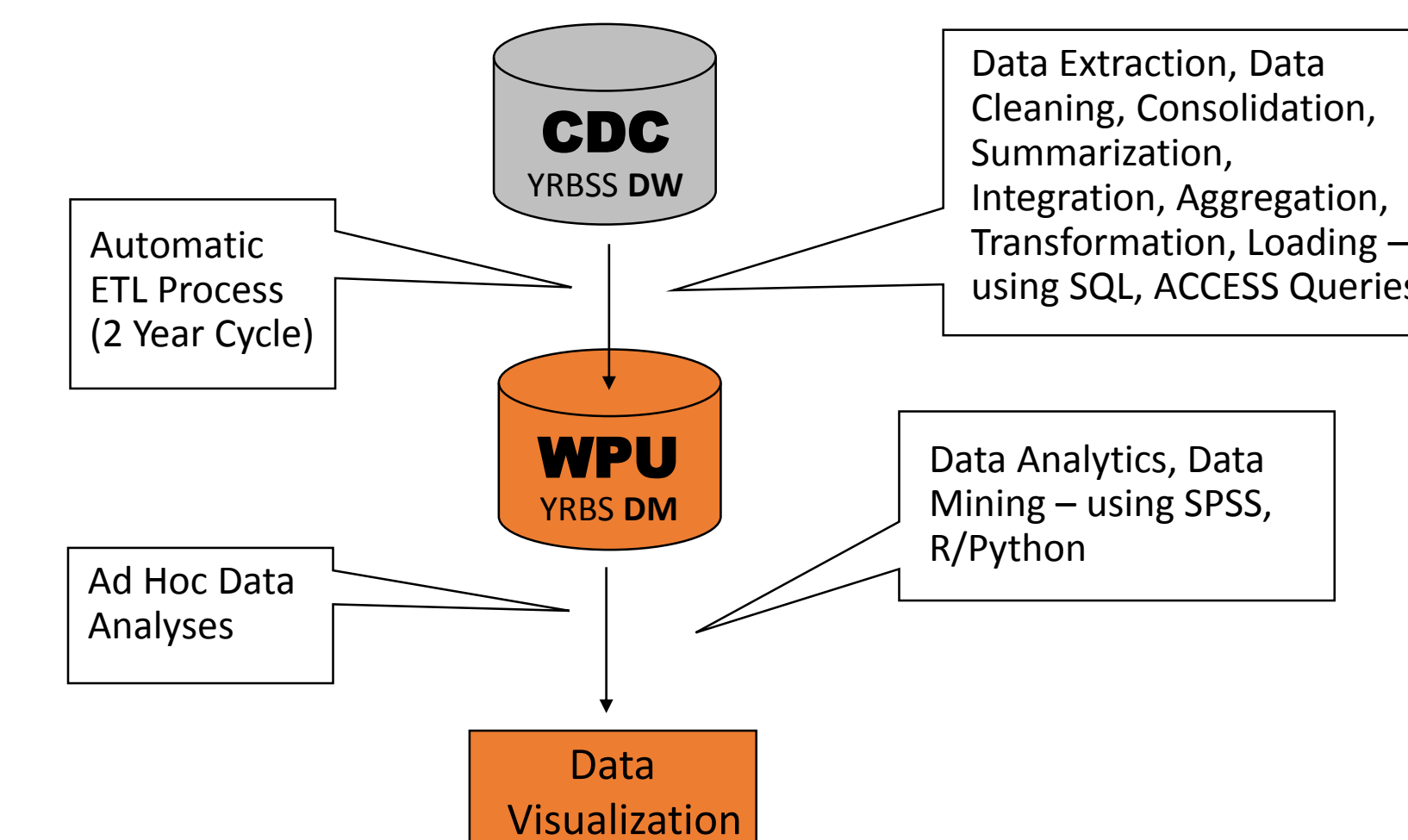
### Server Architecture (Original Setting)



### Server Architecture (New Setting)



### WPU Data Warehousing Process



### YRBSS (CDC)

- Developed in 1990 to monitor priority health risk behaviors that contribute to the leading causes of death, disability, and social problems among youth and young adults in the U.S.
  - Behaviors that contribute to unintentional injuries and violence
  - Sexual behaviors that contribute to unintended pregnancy and sexually transmitted infections, including HIV infection
  - Alcohol and other drug use
  - Tobacco use
  - Unhealthy dietary behaviors
  - Inadequate physical activity
  - Monitors the prevalence of obesity and asthma
- From 1991 to 2013, YRBSS has collected data from more than 2.6 million high school students in more than 1,100 separate surveys (national, state, territorial, tribal government, and local school-based surveys – 9<sup>th</sup> to 12<sup>th</sup> grade students)

### WPU YRBS Data Mart

- Subsets of YRBSS from CDC
- Data Mart (data subset 1)
  - Time frame: the latest 12 months
  - Data location: national
  - Demographics: BMI, sexual minority
  - Data categories: mental health, bullying, sleep patterns
- Data Mart (data subset 2)
  - Time frame: 2010 to present (2015)
  - Data locations: national, state (NJ, NY, PA), city (NYC, Philadelphia)
  - Demographics: Sexual minority
  - Data points: tobacco use, cigarette use, current asthma, electronic vapor product use, alcohol and other drug use, sexual behavior, HIV, sex education, STD, benefits of abstinence

### Project Phases

- First phase (2015-2016):
  - Establish server architecture
  - Program ETL (Extraction, Transformation, Loading) process and 2-year cycle procedure
  - Create data mart using MySQL DBMS
- Second phase (2016-2017):
  - OLAP (On-Line Analytical Processing) architecture – from relational to multi-dimensional database
  - Data analytics and data mining – research and tool selections
  - Data visualization – data presentation techniques

### Acknowledgement

- This research was supported in part by the ART (Assigned Released Time for Research) program, Office of the Provost; and in part by the Student Research Funds of the College of Science and Health, William Paterson University.

### Bibliography

- Brener, N. D., Kann, L., Shanklin, S., Kinchen, S., Eaton, D. K., Hawkins, J., and Flint, K. H., "Methodology of the Youth Risk Behavior Surveillance System – 2013," *CDC MMWR Recommendations and Reports*, Vol. 62, No. 1, March 1, 2013
- <http://www.cdc.gov/healthyyouth/data/yrbs/index.htm>