Leveraging *Local* Data Sources

May 20, 2016

Sallie Keller, Director

VirginiaTech
Biocomplexity Institute

SDAL SOCIAL & DECISION ANALYTICS LABORATORY

Data characterizes the world and the ways in which communication occurs between its individuals

Infrastructure

- Condition
- Operations
- Resilience
- Sustainability

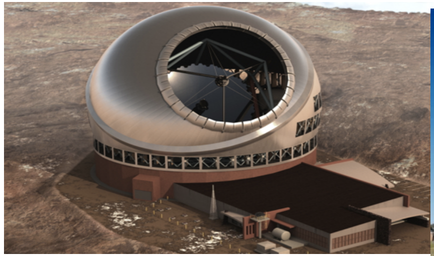
Environment

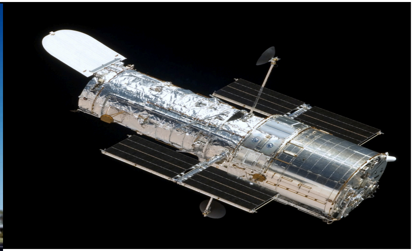
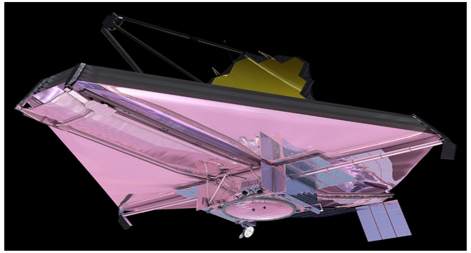
- Climate
- Pollution
- Noise
- Flora/ Fauna

People

- Relationships
- Location
- Economic Condition
- Communication
- Activities
- Health

The New Lens for Social Behavioral Observing



- Data collected faster, while individuals are in the act of behaving in real life situation
- Adapt methods to make the best use of these data
- New data streams produce new discoveries but should not be allowed to degrade the scientific approach

We are in the middle of an *ALL data* revolution

Designed Data Collections

Statistically-designed and intentional observational data collections

Administrative Data

Data collected for the administration of an organization or program

Opportunity Data

Data generated as we move through our daily paces

Procedural Data

Data derived from policies and procedures

What about data quality?

Traditional Approach:

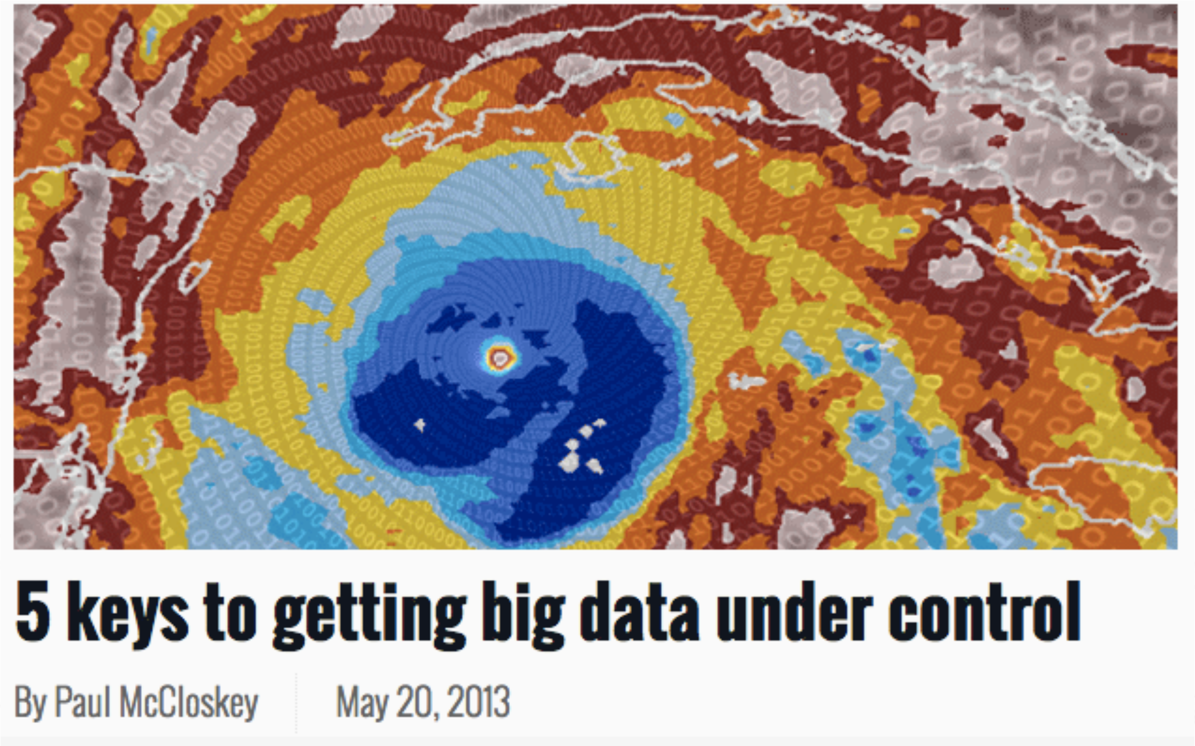
- Control over measurement processes
- Control over collection processes
  - Optimization
  - information maximization
- Clear and controlled ownership

Reliability
Accuracy
Timeliness
Completeness
Relevance
Consistency
Currency
Precision
Format
Sufficiency
Flexibility
Conciseness
Comparability
Scope
Level-of-detail
Effeciency
Quantitativeness
Interpretability
Understandability
Usefulness
Usableness
Informativeness
Clarity
Content
Importance
Freedom from bias

Evolution of data quality by discipline

Discipline	Contributions to Data Quality
Physical and Biological Sciences	• Experimental methods • Data repositories/portals • Reproducibility and replication
Engineering, IT, Business	• Pareto Principle (80 - 20 rule) • Fitness-for-use • Total Data Quality Management (TDQM) • Data management • Standards
Social and Behavioral Sciences	• Total Survey Error • Randomized Control Trials, Observational Studies, and Natural Experiments
Statistics	• Decision theoretic approach • Statistical methods • Adoption of definitions, methods, and approaches by official statistics

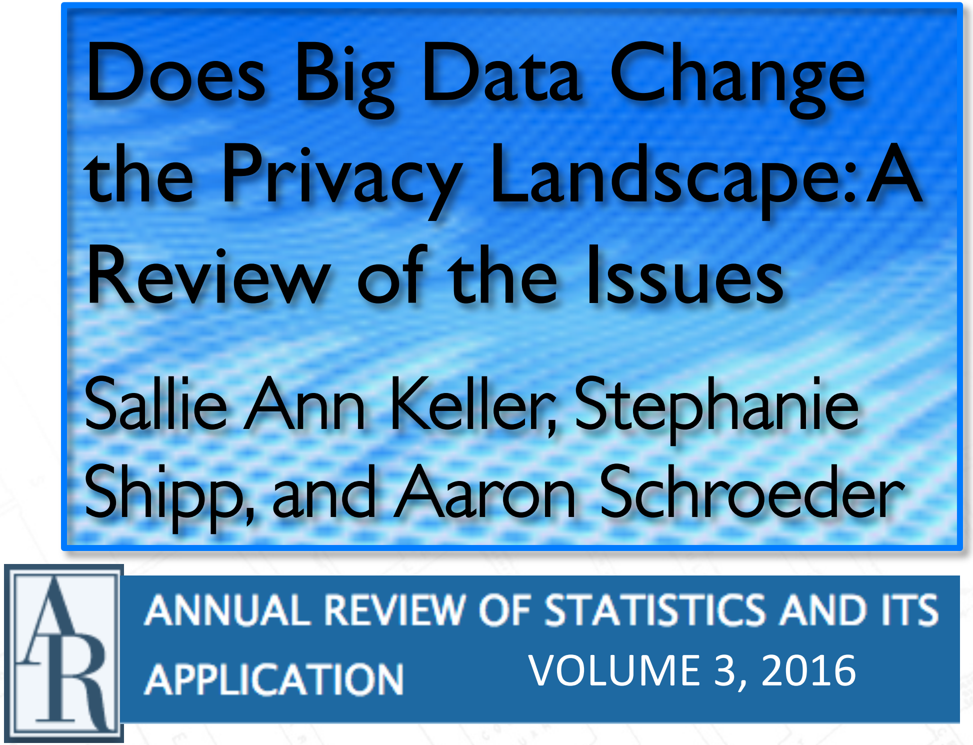
In this data revolution, we have lost control!



5 keys to getting big data under control

By Paul McCloskey May 20, 2013

And then there is privacy!



Does Big Data Change the Privacy Landscape: A Review of the Issues

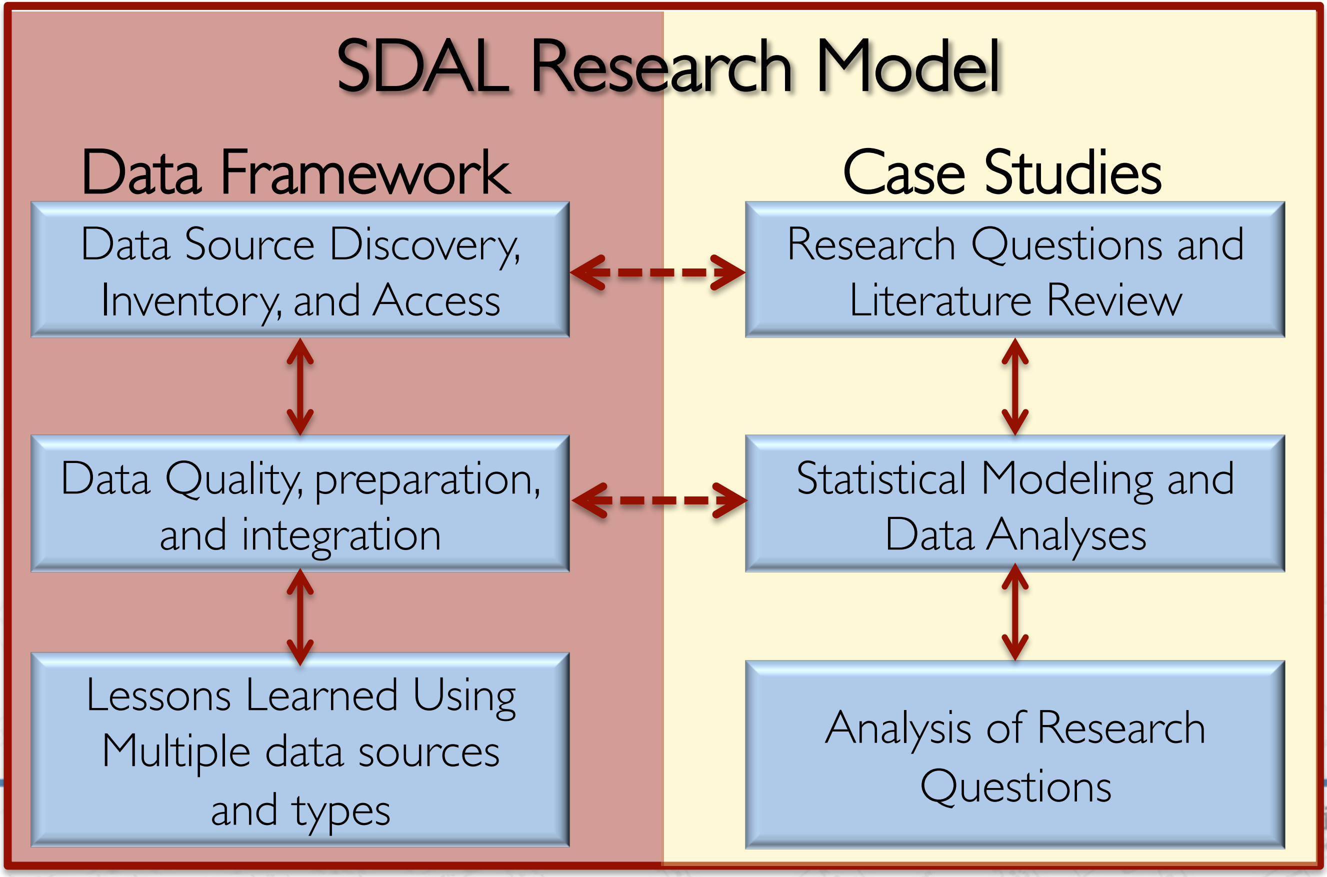
Sallie Ann Keller, Stephanie Shipp, and Aaron Schroeder

ANNUAL REVIEW OF STATISTICS AND ITS APPLICATION VOLUME 3, 2016

Innovation data and measurement problems

- For what purpose?
  - Research, researcher access, policy, program evaluation, …
- What dimension or aspect of innovaion?
  - entrepreneurship, R&D, advances, productivity, productivity growth, technology, workforce, household innovators, automation, robotics, non-market diffusion, collective knowledge, …
- Can we find local data surrogates?

A disciplined approach is needed to develop the theory and methods for using All the data



SDAL Research Model

Data Framework

Data Source Discovery, Inventory, and Access

Data Quality, preparation, and integration

Lessons Learned Using Multiple data sources and types

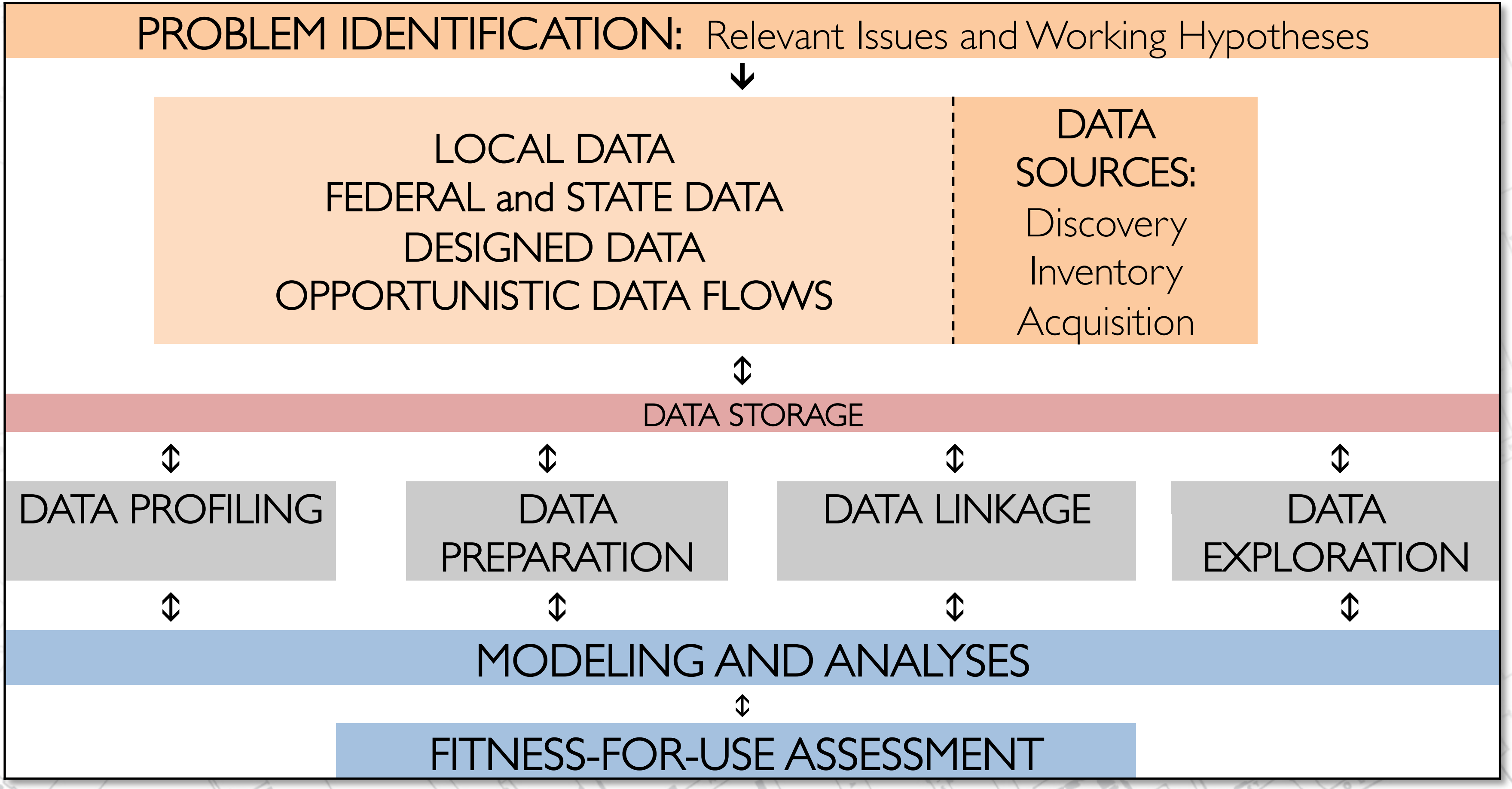
Case Studies

Research Questions and Literature Review

Statistical Modeling and Data Analyses

Analysis of Research Questions

Data Framework



PROBLEM IDENTIFICATION: Relevant Issues and Working Hypotheses

LOCAL DATA
FEDERAL and STATE DATA
DESIGNED DATA
OPPORTUNISTIC DATA FLOWS

DATA SOURCES:
Discovery
Inventory
Acquisition

DATA STORAGE

DATA PROFILING

DATA PREPARATION

DATA LINKAGE

DATA EXPLORATION

MODELING AND ANALYSES

FITNESS-FOR-USE ASSESSMENT

Data inventory process in-action

- Purpose of organization collecting the data
- Description of the data
  - content
  - Unit represented
  - Longitudinal or cross-sectional
  - Geographic coverage
  - Timing of collection and release

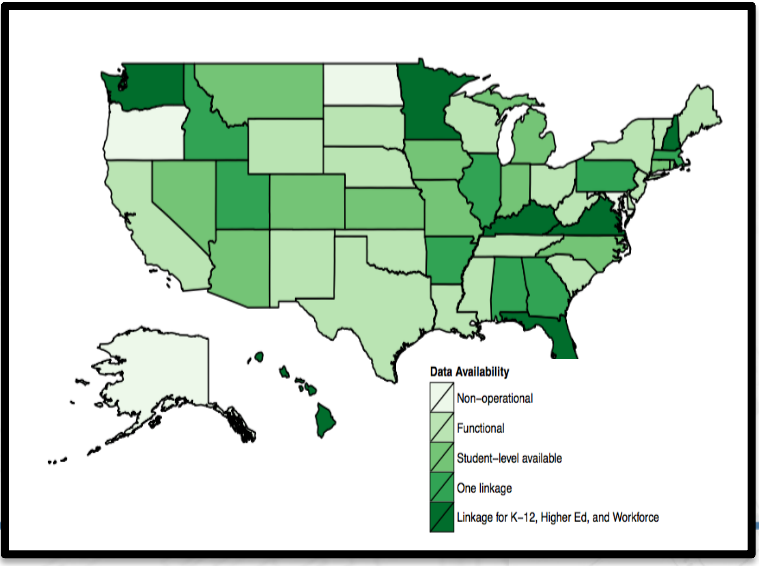
- Data collection method
- Metadata and provenance
- Other issues
  - Selectivity
  - Stability
  - Accessibility
  - Privacy and security
  - Research using data

Collaborative wiki – dynamic report appendix

Case study data sources that moved onto acquisition

Education – State Longitudinal Data Systems

- Washington
- North Carolina
- Virginia
- Kentucky
- Texas



Housing

- Local Count Real Estate Assessment Data
- Black Knight Financial Services
- CoreLogic
- MLS Data
- Local Crime Reports
- Location Inc.

Focus Briefly on Data Profiling

- Determination of quality of data and utility to project at hand
  - Data Quality
  - Data Structure
  - Metadata and Provenance
- Issues are discovered but not fixed, fix depends on research need

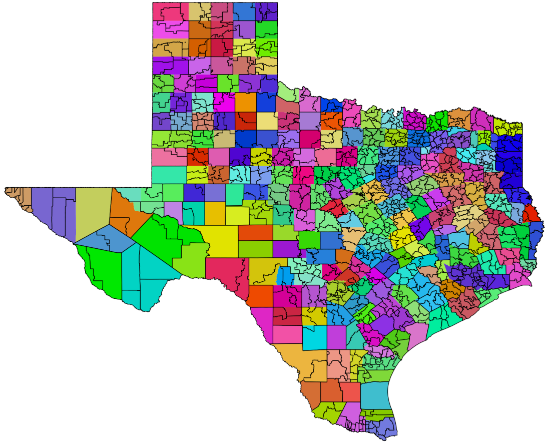
Local Data: Education Example

Exercising the data framework for the education data

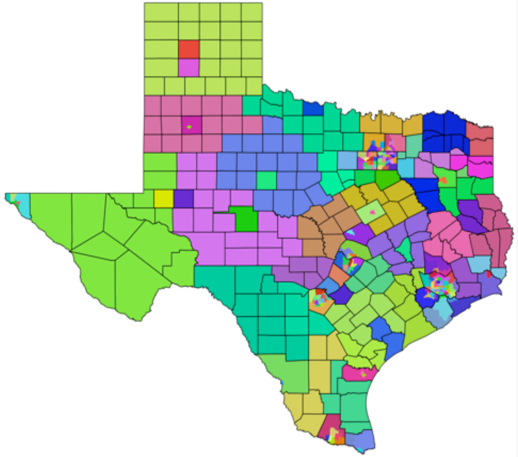
- Profiled variables
  - Student ID, district code, year, gender, race/ethnicity, grade, age, and other variables (e.g., limited English proficiency)
  - Most variables were valid and consistent requiring very little cleaning
- Transformed variables
  - Matched school districts with counties
  - Calculated ages from birthdates (NC and KY)
  - Texas enrollment estimates weighted to match the state level counts
- Restructured variables
  - Virginia data were restructured to create three main tables for race/ethnicity, by grade, gender by grade, and disadvantaged status by grade

Local data challenges with geographic alignment

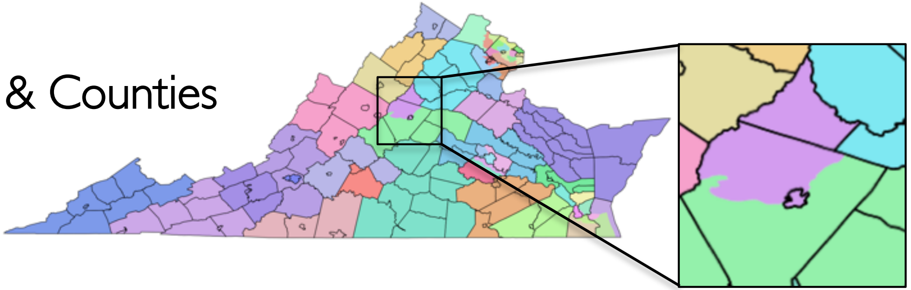
Counties & School districts



PUMAs & Counties



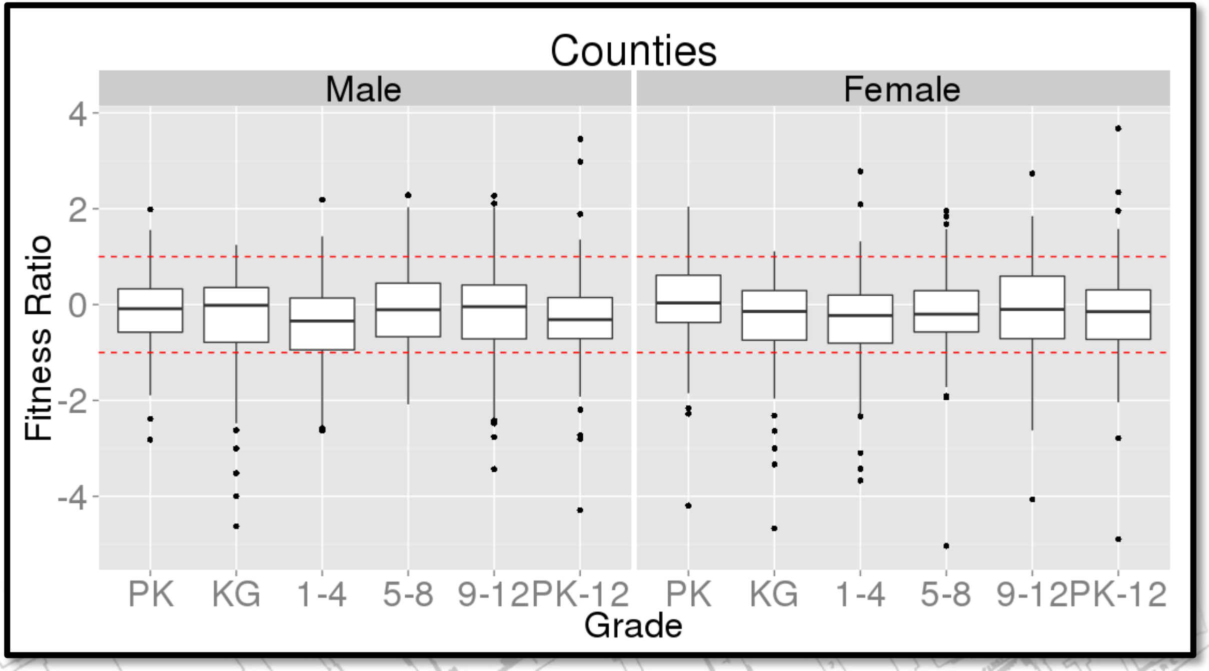
PUMAs & Counties



Education SLSD Alignment with ACS – Kentucky

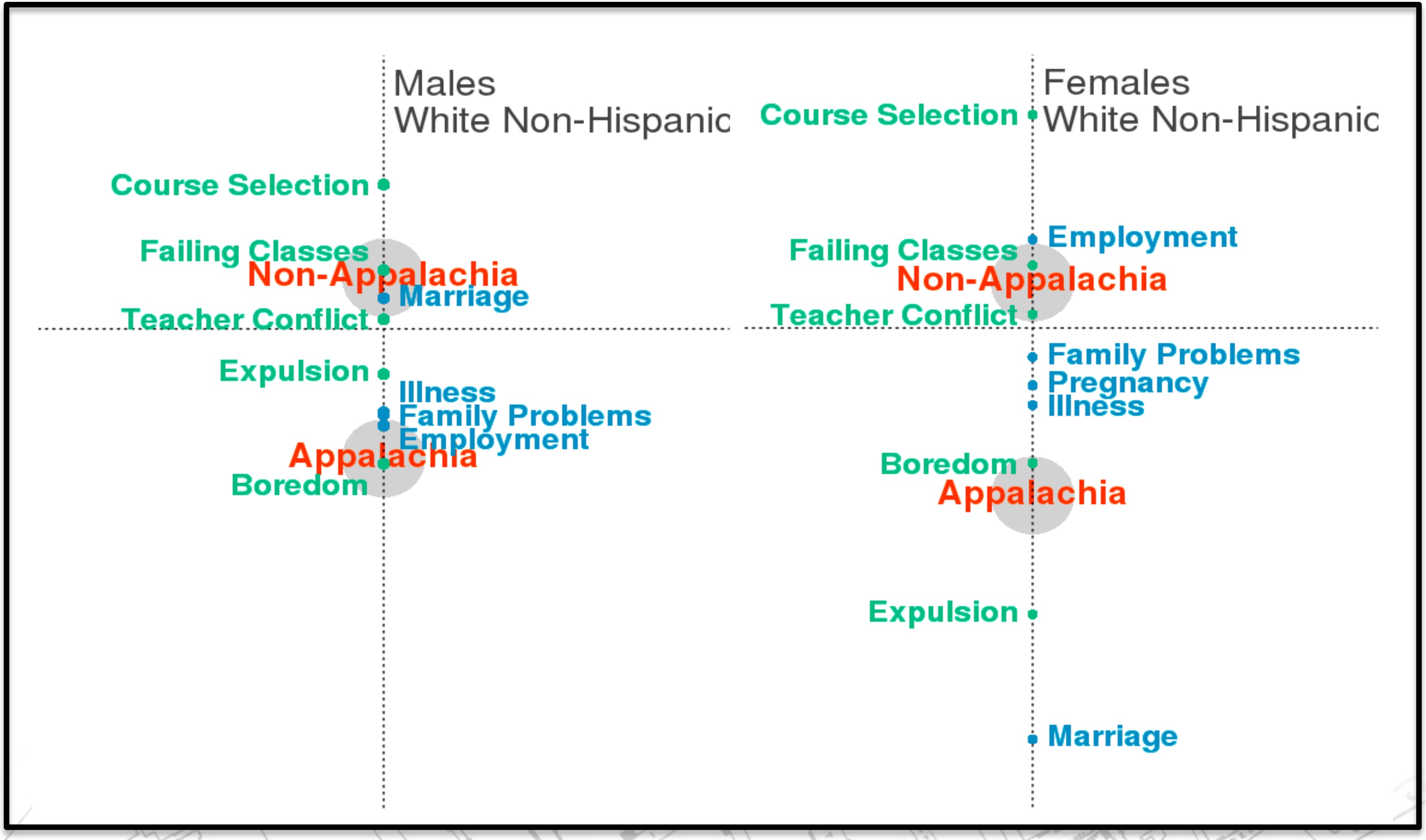
$$Fitness\ Ratio = \frac{ACS\ Estimate - Local\ Data\ Estimate}{ACS\ 90\%\ Margin\ of\ Error}$$

Public Enrollment at Kentucky State and County Levels, 2009-2013



Kentucky Public School Dropouts

Grades 9-12, KLDS 2009-2013



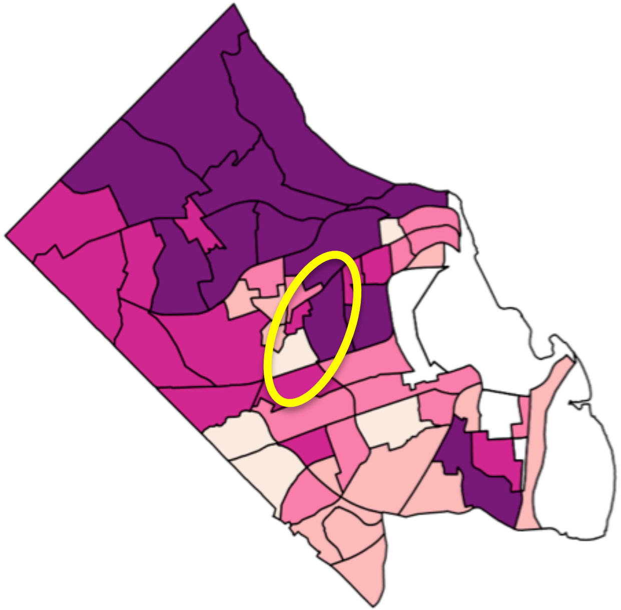
Local Data: Housing Example

Exercising the data framework for the housing data

- Profiled variables
  - Need to align units of analysis, parcels versus housing units
  - Reconcile land use codes through discussion with experts
  - Remove non-residential parcels
- Transformed variables
  - Weight parcels by (estimated) number of units
  - Adjusted inconsistent longitudinal data, e.g., number of units or year built
- Restructured variables
  - Create a consistent set of geocodes per parcel

Characterizing neighborhoods

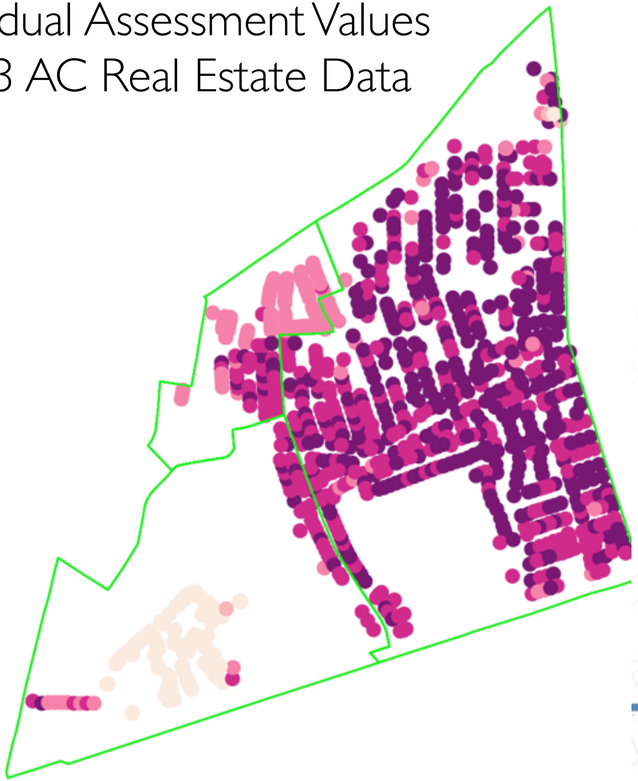
Medan House Value
Owner Occupied Units
ACS 2009-2013 Estimate



Scale

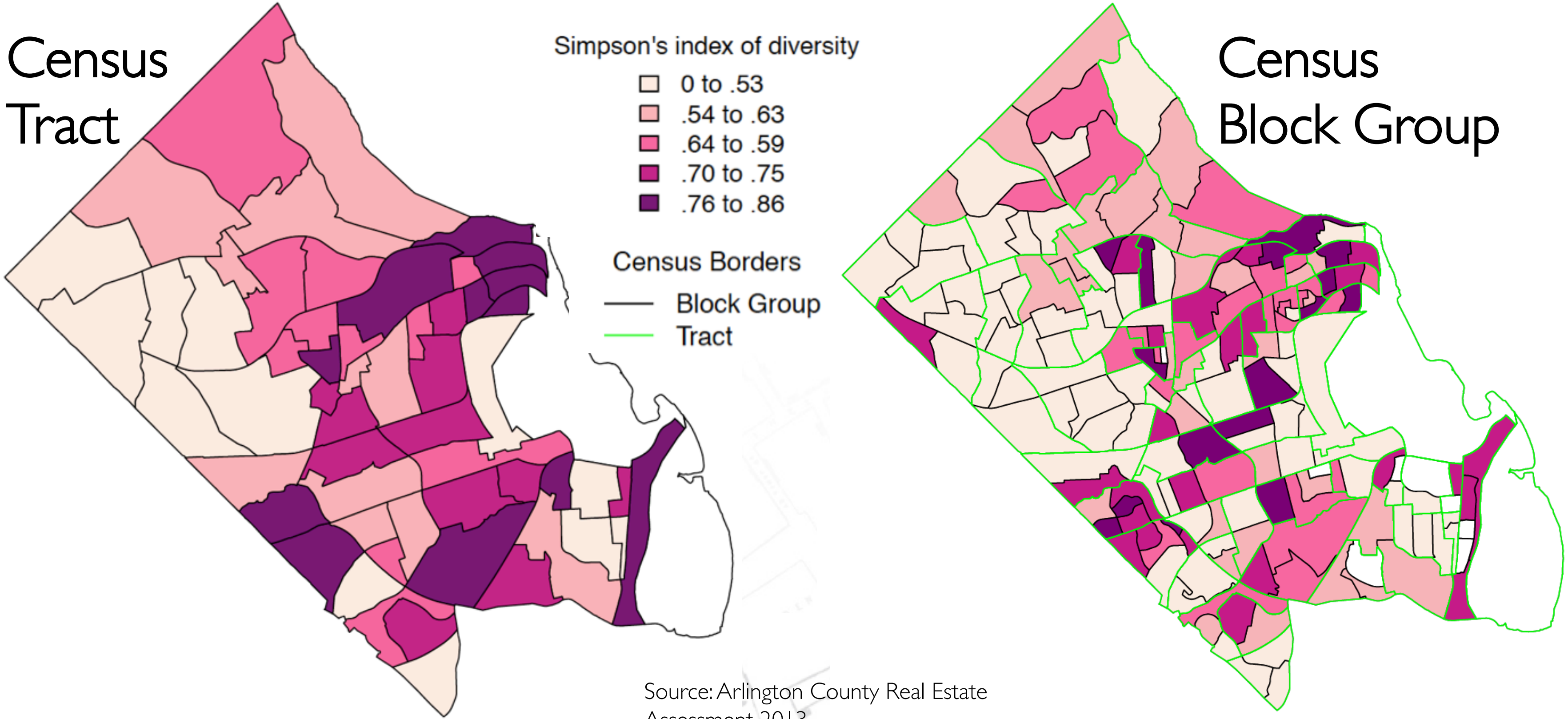
- □ Below $350,000
- □ $350,000 to $449,999
- □ $450,000 to $599,999
- □ $600,000 to $699,999
- ■ $700,000 or more

Individual Assessment Values
2013 AC Real Estate Data



Source: Arlington County Real Estate Assessment, 2013; ACS 2009-2013

Simpson Index of Home Value Diversity



Source: Arlington County Real Estate Assessment 2013

Concluding Remarks

- Use of external data – no control over collection, unlike federal statistics
- Through lens of case studies develop and test data framework
- Need a disciplined, yet flexible and adaptable, data framework to assess data quality and fitness-for-use that is dependent on use, e.g., Innovation measures

Proposed Criteria for Future Research

- Speed and regularity of acquiring data
- Flexibility in negotiating data agreements
- Capacity to implement Data Framework
  - e.g., Organizational structure and constraints, GIS, longitudinal discrepancies, statistical data integration, etc.
- Research needed to align and harmonize data sources
- Value proposition to sponsor