HEALTH DATAPALOOZA AcademyHealth

AcademyHealth NATIONAL HEALTH POLICY CONFERENCE

EVALUATING HEALTHCARE DATASETS:

A Framework to Select Datasets and to Standardize, Classify and Link Variables

FEBRUARY 18, 2021

Brian Williams & David M.C. Stern

Stern Consulting LLC

The Challenge

Selecting the few based on suitability rather than familiarity

Many are Called

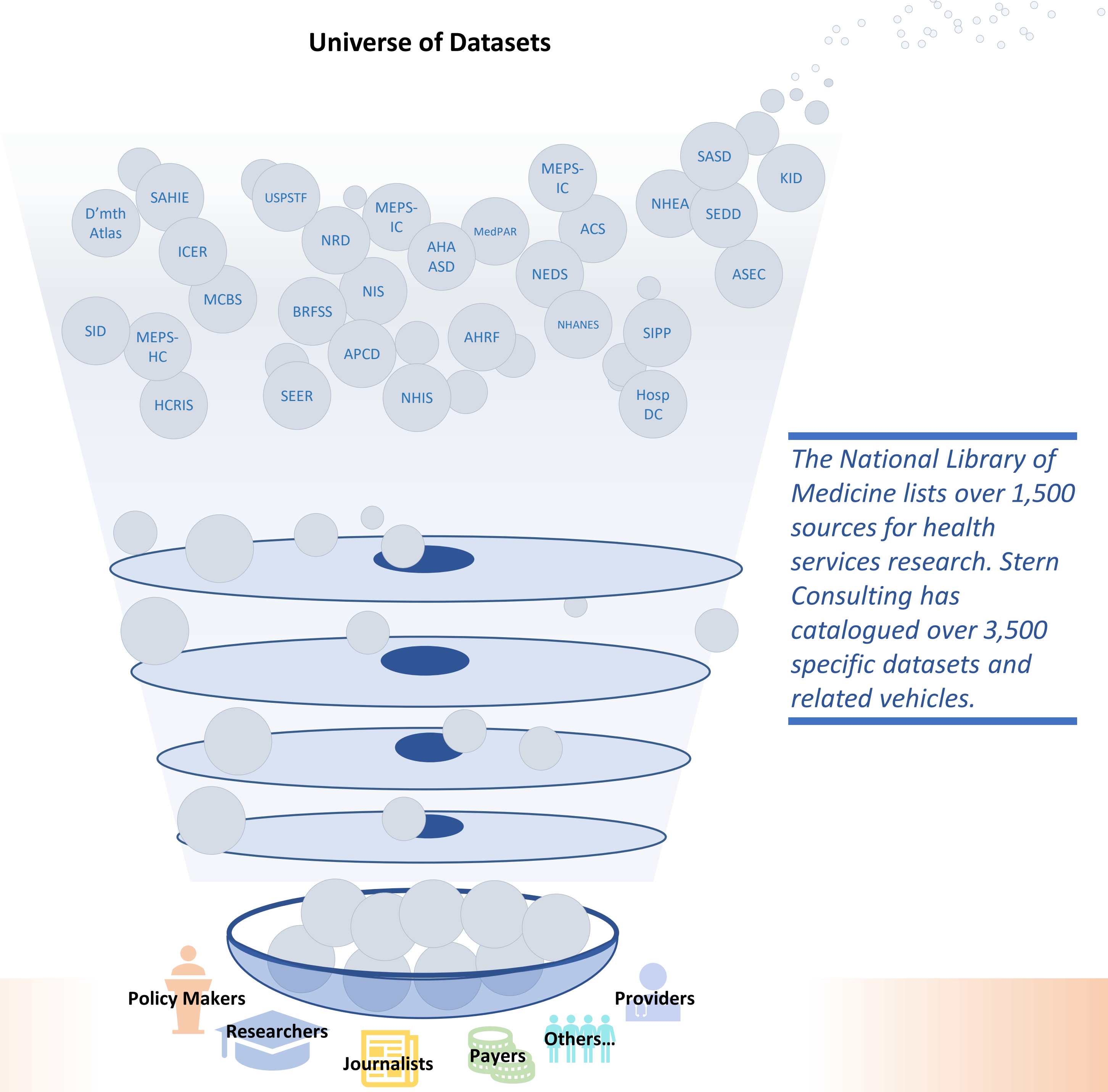
There are thousands of healthcare datasets.

Few are Chosen

A handful may be suitable for a particular need.

Stern Consulting LLC

Universe of Datasets



The National Library of Medicine lists over 1,500 sources for health services research. Stern Consulting has catalogued over 3,500 specific datasets and related vehicles.

Framework for Sorting Through the Data Universe

(version 2.0)

To sort through it all, we developed a framework that

- ***Defines*** *the characteristics of datasets* (Data Dimensions),
- ***Identifies*** *user requirements* (User Considerations), *and*
- ***Matches*** *those considerations to find the datasets best suited for a particular project* (Selected Datasets).

Stern Consulting LLC

Universe of Datasets

Datasets are defined by eight key dimensions.

Data Dimensions

Mechanism of Data Generation
- Administrative
- Survey
- Disease Surveillance
- Evidence Based Healthcare
- Regulatory (e.g. Cost Reports)
- Medical Record Abstracts
- Vital Records
- Peer Reviewed Literature
- Gray Literature
- Directories/Code Books/Lists
- Other

Sponsor
- Federal Government
- State Government
- Foundations
- Industry Associations
- Disease-Specific Organizations
- Academia
- Accreditation Bodies
- Commercial Entities
- Other

Unit
- Person
- Family
- Household
- Employer
- Encounter/Claim
- Diagnosis
- Procedure
- Provider
- Location
- Other

Data Dimensions

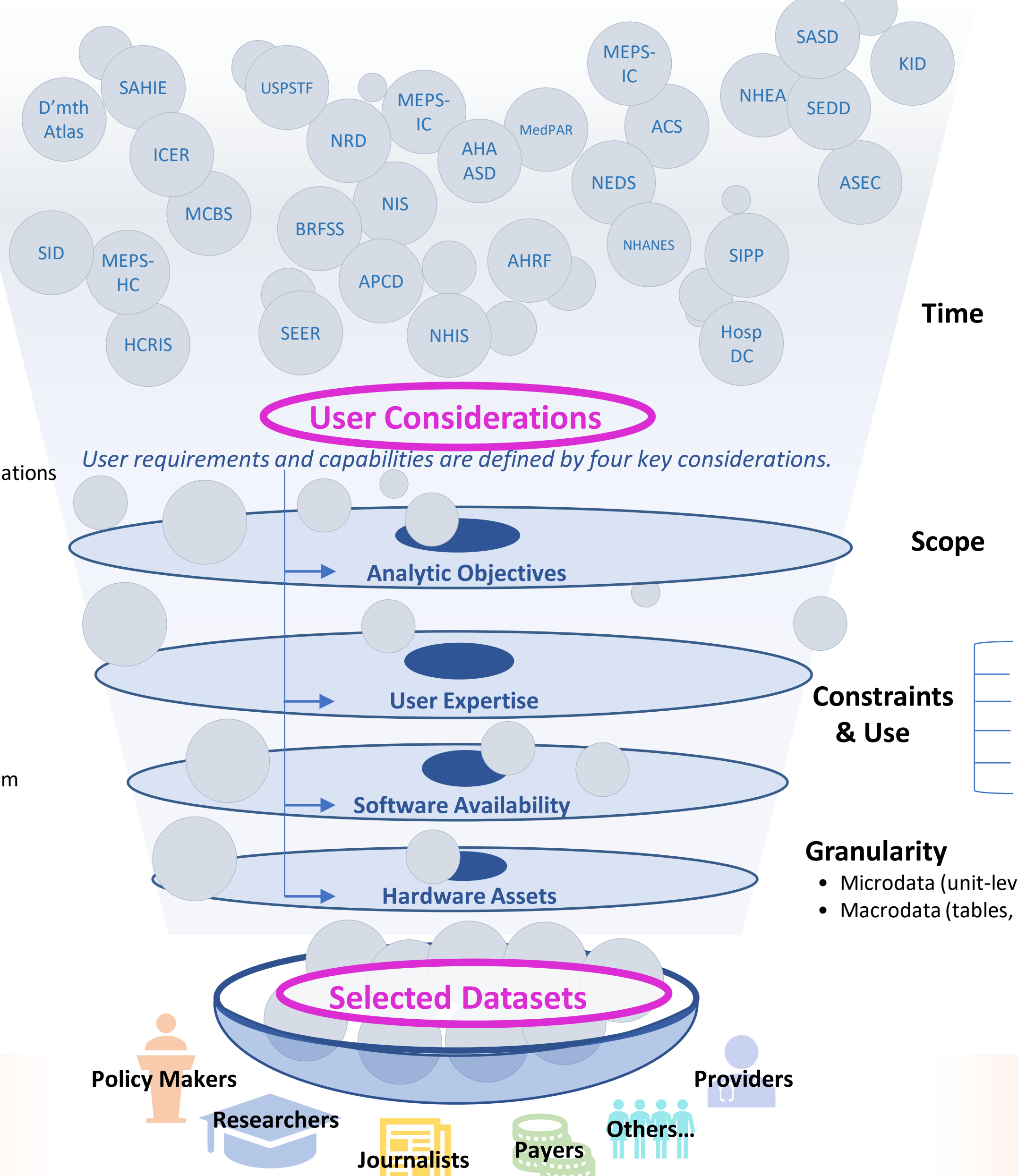
Content
- Access
- Cost
- Healthiness
- Demographics
- Geographic Levels
- Economics
- Social
- Other

Time
- **Span**
- **Periodicity**
- **"Longitudinality"**
  - Longitudinal
    - –Panel
    - –Cohort
    - –Retrospective
  - Cross-sectional
- **Other**

Scope
- **Population**
- **Geography**
- **Other**

Constraints & Use
- **Acquisition Requirements**
- **Required Expertise**
- **Cost**
- **Hard/Software Needed**
- **Restrictions on Use**
- **Other**

Granularity
- Microdata (unit-level source data)
- Macrodata (tables, online queries, other)



User Considerations

User requirements and capabilities are defined by four key considerations.

- Analytic Objectives
- User Expertise
- Software Availability
- Hardware Assets

Selected Datasets

Sample Users

Case Study:

"Datasets to Evaluate the Impact of National Healthcare Policy"

User Considerations

***Analytic Objectives**: The selected datasets had to address access, cost and/or healthiness; be available to the public pre- and post-ACA; and reflect the U.S. population.*

***User Capabilities:** Selections had to account for differing user capabilities concerning database management, data analysis, and statistics.*

Universe of Datasets

Datasets are defined by eight key dimensions.

Dimensions

Mechanism of Data Generation

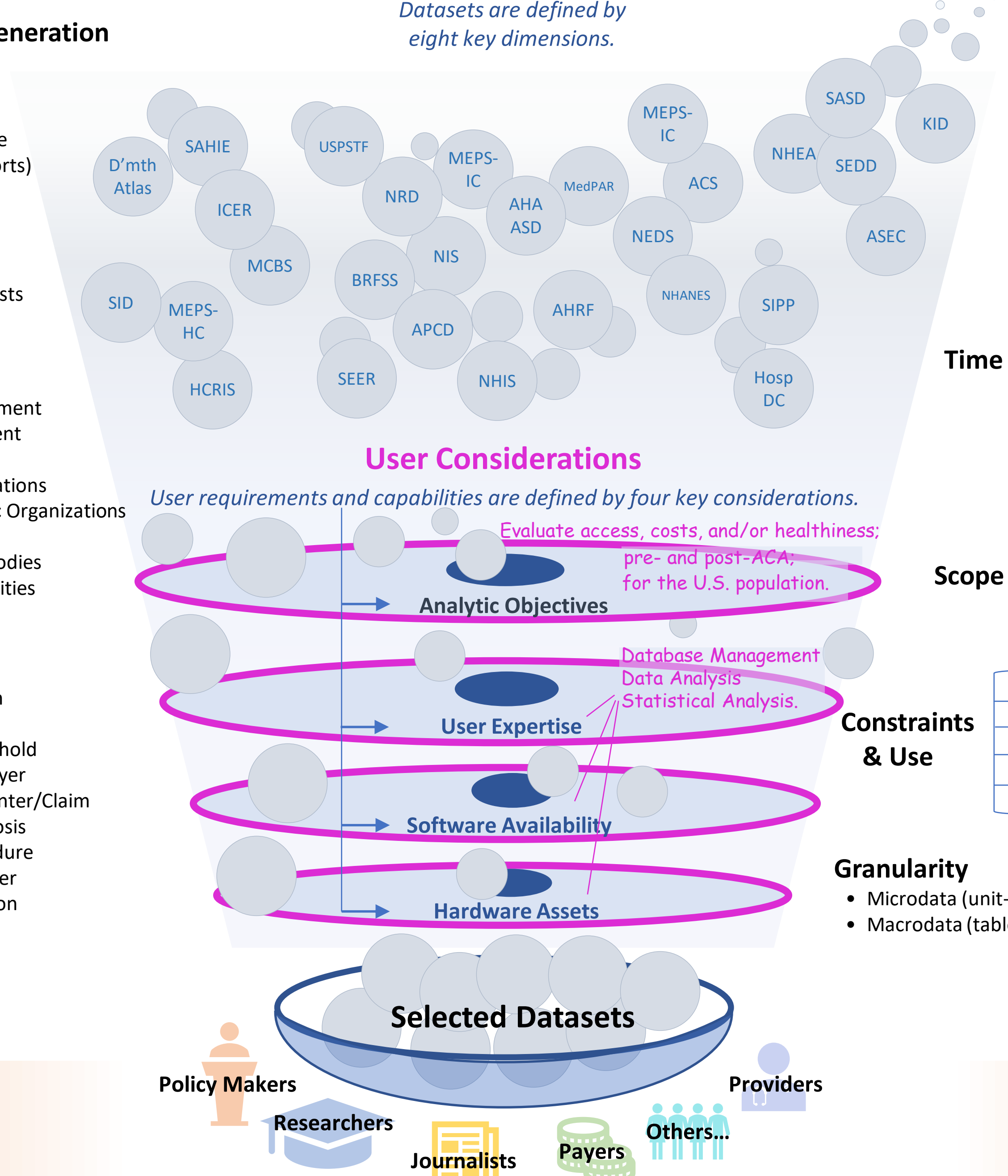
- Administrative
- Survey
- Disease Surveillance
- Evidence Based Healthcare
- Regulatory (e.g. Cost Reports)
- Medical Record Abstracts
- Vital Records
- Peer Reviewed Literature
- Gray Literature
- Directories/Code Books/Lists
- Other

Sponsor

- Federal Government
- State Government
- Foundations
- Industry Associations
- Disease-Specific Organizations
- Academia
- Accreditation Bodies
- Commercial Entities
- Other

Unit

- Person
- Family
- Household
- Employer
- Encounter/Claim
- Diagnosis
- Procedure
- Provider
- Location
- Other



User Considerations

User requirements and capabilities are defined by four key considerations.

Analytic Objectives

Evaluate access, costs, and/or healthiness; pre- and post-ACA; for the U.S. population.

User Expertise

Database Management
Data Analysis
Statistical Analysis.

Software Availability

Hardware Assets

Selected Datasets

Dimensions

Content

- Access
- Cost
- Healthiness
- Demographics
- Geographic Levels
- Economics
- Social
- Other

Time

Span
Periodicity
"Longitudinality"

- Longitudinal
  - –Panel
  - –Cohort
  - –Retrospective
- Cross-sectional

Other

Scope

Population
Geography
Other

Constraints & Use

Acquisition Requirements
Required Expertise
Cost
Hard/Software Needed
Restrictions on Use
Other

Granularity

- Microdata (unit-level source data)
- Macrodata (tables, online queries, other)

Sample Users

Policy Makers
Researchers
Journalists
Payers
Others...
Providers

Stern Consulting LLC

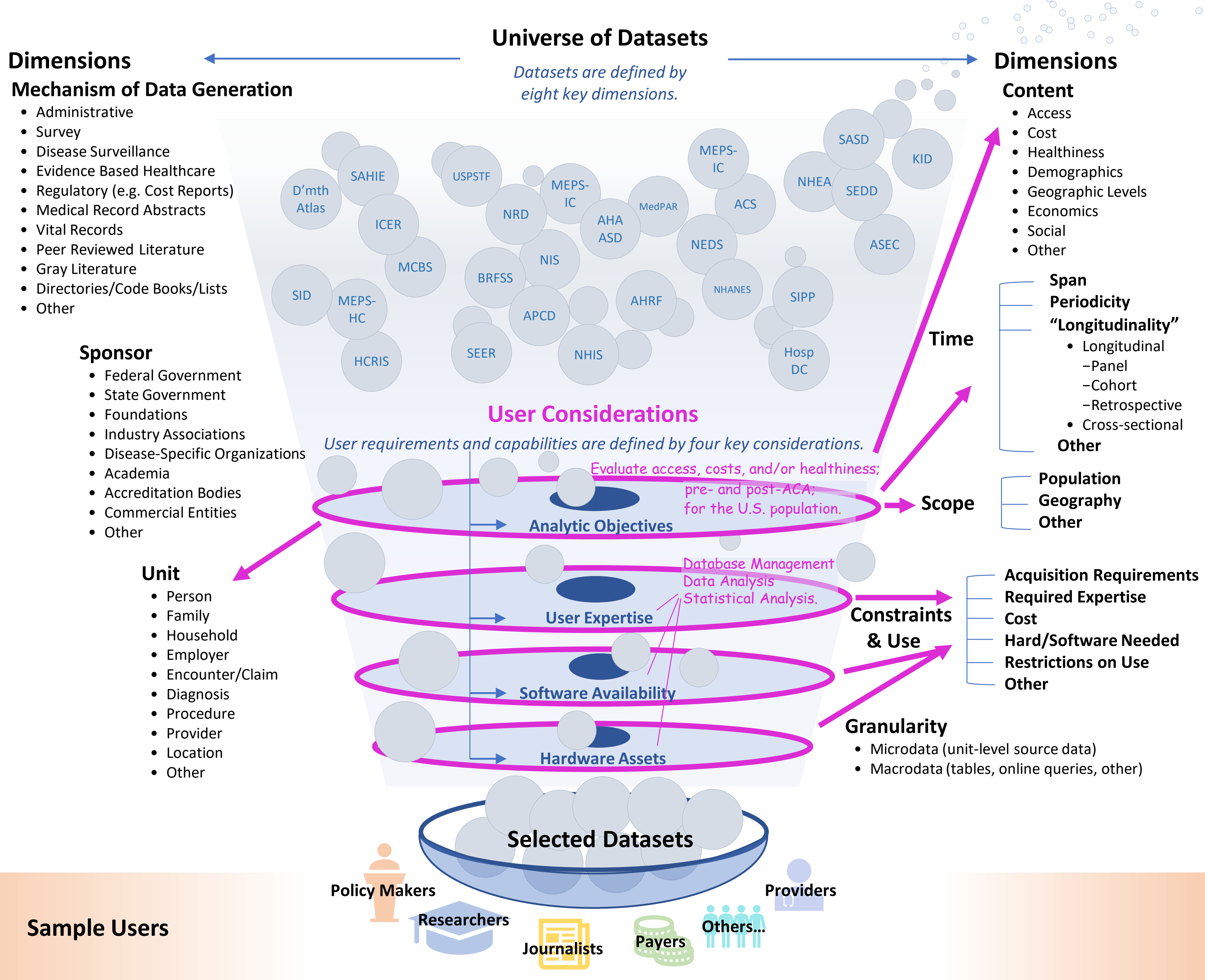
Case Study:

"Datasets to Evaluate the Impact of National Healthcare Policy"

Key Dimensions

***Analytic objectives** were addressed primarily by the Content, Time, Scope and Unit dimensions.*

***User Capabilities** were addressed by the "Constraints & Use" dimension.*



Universe of Datasets

Datasets are defined by eight key dimensions.

Dimensions

Mechanism of Data Generation
- Administrative
- Survey
- Disease Surveillance
- Evidence Based Healthcare
- Regulatory (e.g. Cost Reports)
- Medical Record Abstracts
- Vital Records
- Peer Reviewed Literature
- Gray Literature
- Directories/Code Books/Lists
- Other

Sponsor
- Federal Government
- State Government
- Foundations
- Industry Associations
- Disease-Specific Organizations
- Academia
- Accreditation Bodies
- Commercial Entities
- Other

Unit
- Person
- Family
- Household
- Employer
- Encounter/Claim
- Diagnosis
- Procedure
- Provider
- Location
- Other

Dimensions

Content
- Access
- Cost
- Healthiness
- Demographics
- Geographic Levels
- Economics
- Social
- Other

Time

Span
Periodicity
"Longitudinality"
- Longitudinal
  - −Panel
  - −Cohort
  - −Retrospective
- Cross-sectional

Other

Scope

Population
Geography
Other

Constraints & Use

Acquisition Requirements
Required Expertise
Cost
Hard/Software Needed
Restrictions on Use
Other

Granularity
- Microdata (unit-level source data)
- Macrodata (tables, online queries, other)

User Considerations

User requirements and capabilities are defined by four key considerations.

Evaluate access, costs, and/or healthiness; pre- and post-ACA; for the U.S. population.

Database Management
Data Analysis
Statistical Analysis.

Stern Consulting LLC

Case Study:

"Datasets to Evaluate the Impact of National Healthcare Policy"

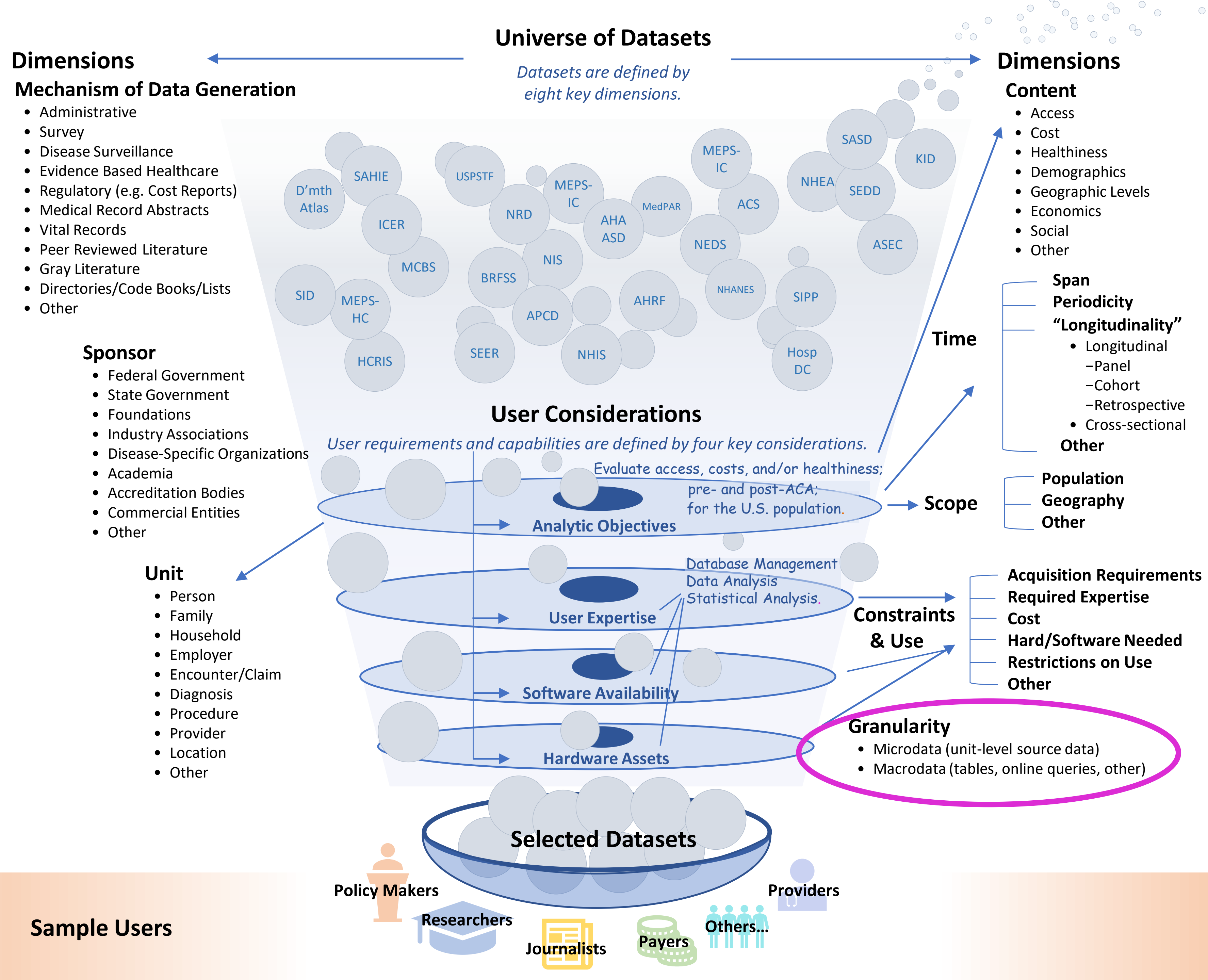
The "Granularity" Dimension

***Microdata:** most flexible, but high degree of user expertise required.*

***Macrodata:** less flexible, but easier to use and can be quite robust.*

Both were included in the data selection process.

Stern Consulting LLC



Universe of Datasets

Datasets are defined by eight key dimensions.

Dimensions

Mechanism of Data Generation

- Administrative
- Survey
- Disease Surveillance
- Evidence Based Healthcare
- Regulatory (e.g. Cost Reports)
- Medical Record Abstracts
- Vital Records
- Peer Reviewed Literature
- Gray Literature
- Directories/Code Books/Lists
- Other

Sponsor

- Federal Government
- State Government
- Foundations
- Industry Associations
- Disease-Specific Organizations
- Academia
- Accreditation Bodies
- Commercial Entities
- Other

Unit

- Person
- Family
- Household
- Employer
- Encounter/Claim
- Diagnosis
- Procedure
- Provider
- Location
- Other

Dimensions

Content

- Access
- Cost
- Healthiness
- Demographics
- Geographic Levels
- Economics
- Social
- Other

Time

- Span
- Periodicity
- "Longitudinality"
  - Longitudinal
    - −Panel
    - −Cohort
    - −Retrospective
  - Cross-sectional
- Other

Scope

- Population
- Geography
- Other

Constraints & Use

- Acquisition Requirements
- Required Expertise
- Cost
- Hard/Software Needed
- Restrictions on Use
- Other

Granularity

- Microdata (unit-level source data)
- Macrodata (tables, online queries, other)

User Considerations

User requirements and capabilities are defined by four key considerations.

Analytic Objectives – Evaluate access, costs, and/or healthiness; pre- and post-ACA; for the U.S. population.

User Expertise – Database Management, Data Analysis, Statistical Analysis.

Software Availability

Hardware Assets

Selected Datasets

Sample Users: Policy Makers, Researchers, Journalists, Payers, Others…, Providers

Case Study:

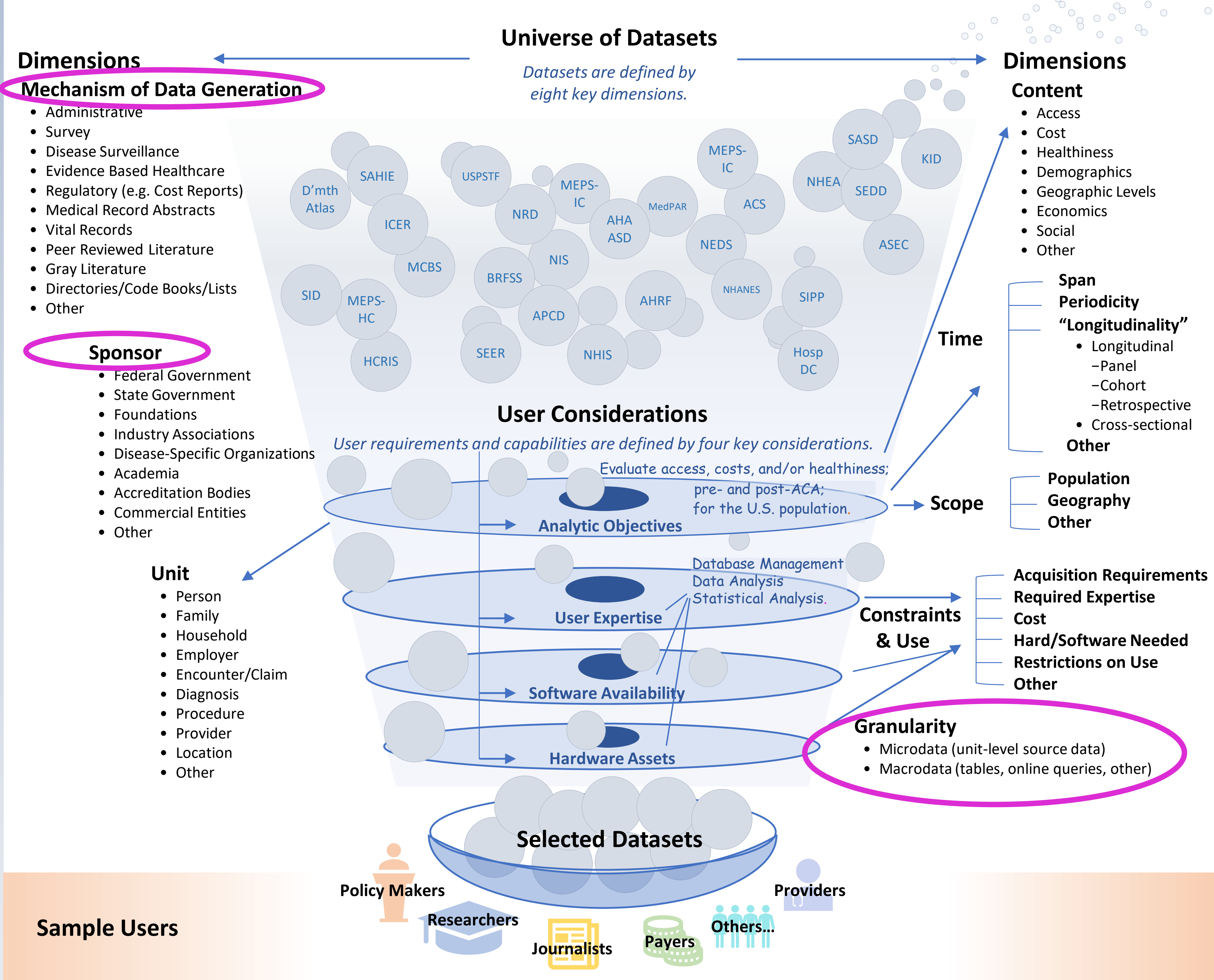
“Datasets to Evaluate the Impact of National Healthcare Policy”

Dimensions Not Linked to User Considerations

Some data dimensions are not constrained by user requirements.

They offer the most opportunity for thinking broadly about available datasets and looking beyond the familiar.

Stern Consulting LLC



Universe of Datasets

Datasets are defined by eight key dimensions.

Dimensions

Mechanism of Data Generation

- Administrative
- Survey
- Disease Surveillance
- Evidence Based Healthcare
- Regulatory (e.g. Cost Reports)
- Medical Record Abstracts
- Vital Records
- Peer Reviewed Literature
- Gray Literature
- Directories/Code Books/Lists
- Other

Sponsor

- Federal Government
- State Government
- Foundations
- Industry Associations
- Disease-Specific Organizations
- Academia
- Accreditation Bodies
- Commercial Entities
- Other

Unit

- Person
- Family
- Household
- Employer
- Encounter/Claim
- Diagnosis
- Procedure
- Provider
- Location
- Other

Dimensions

Content

- Access
- Cost
- Healthiness
- Demographics
- Geographic Levels
- Economics
- Social
- Other

Time

Span
Periodicity
“Longitudinality”

- Longitudinal
  - –Panel
  - –Cohort
  - –Retrospective
- Cross-sectional

Other

Scope

Population
Geography
Other

Constraints & Use

Acquisition Requirements
Required Expertise
Cost
Hard/Software Needed
Restrictions on Use
Other

Granularity

- Microdata (unit-level source data)
- Macrodata (tables, online queries, other)

User Considerations

User requirements and capabilities are defined by four key considerations.

Analytic Objectives — Evaluate access, costs, and/or healthiness; pre- and post-ACA; for the U.S. population.

User Expertise — Database Management; Data Analysis; Statistical Analysis.

Software Availability

Hardware Assets

Selected Datasets

Sample Users

Case Study:

“Datasets to Evaluate the Impact of National Healthcare Policy”

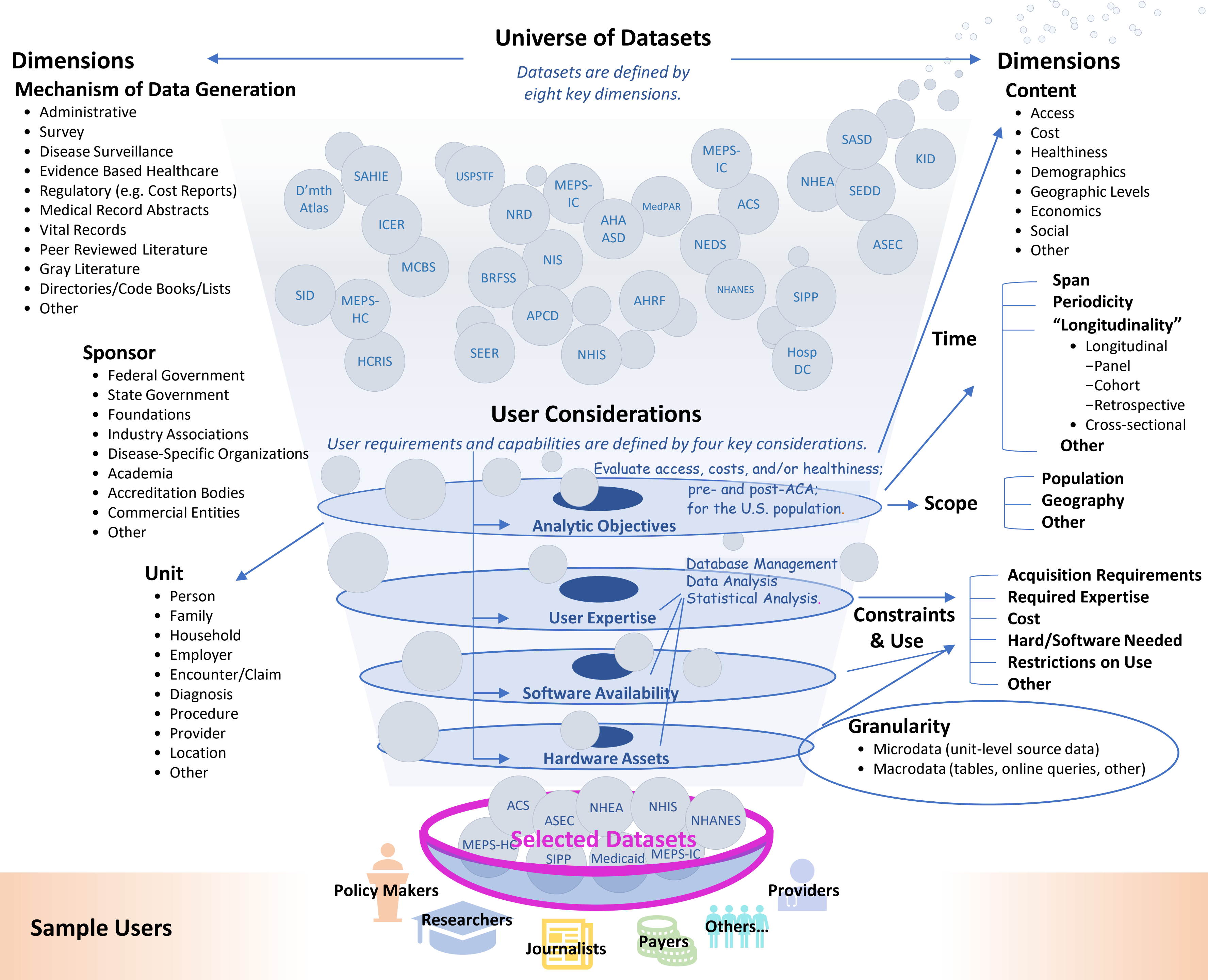
From Many to Few

***Selected Datasets**: Nine datasets “made the cut” after matching user needs to data dimensions.*

Beyond the Familiar

The best-suited datasets for our case study were a mixture of microdata and macrodata, surveys and admin data, healthcare- and nonhealthcare-focused data sets.

The framework not only narrowed our field of datasets to the most appropriate, but broadened our thinking beyond the familiar.



Universe of Datasets

Datasets are defined by eight key dimensions.

Dimensions

Mechanism of Data Generation

- Administrative
- Survey
- Disease Surveillance
- Evidence Based Healthcare
- Regulatory (e.g. Cost Reports)
- Medical Record Abstracts
- Vital Records
- Peer Reviewed Literature
- Gray Literature
- Directories/Code Books/Lists
- Other

Sponsor

- Federal Government
- State Government
- Foundations
- Industry Associations
- Disease-Specific Organizations
- Academia
- Accreditation Bodies
- Commercial Entities
- Other

Unit

- Person
- Family
- Household
- Employer
- Encounter/Claim
- Diagnosis
- Procedure
- Provider
- Location
- Other

Dimensions

Content

- Access
- Cost
- Healthiness
- Demographics
- Geographic Levels
- Economics
- Social
- Other

Time

Span

Periodicity

“Longitudinality”

- Longitudinal
  - −Panel
  - −Cohort
  - −Retrospective
- Cross-sectional

Other

Scope

Population

Geography

Other

Constraints & Use

Acquisition Requirements

Required Expertise

Cost

Hard/Software Needed

Restrictions on Use

Other

Granularity

- Microdata (unit-level source data)
- Macrodata (tables, online queries, other)

User Considerations

User requirements and capabilities are defined by four key considerations.

Analytic Objectives: Evaluate access, costs, and/or healthiness; pre- and post-ACA; for the U.S. population.

User Expertise: Database Management; Data Analysis; Statistical Analysis.

Software Availability

Hardware Assets

Selected Datasets

Sample Users

Policy Makers

Researchers

Journalists

Payers

Others…

Providers

Stern Consulting LLC

Case Study:

“Datasets to Evaluate the Impact of National Healthcare Policy”

Selected Datasets by Dimension

Each of the nine selected datasets is characterized by each of the eight data dimensions.

		Dimensions of Datasets							
Datasets	**Notes**	**Mechanism of Data Generation**	**Sponsor** (Data Collector)	**Content** *Variable Counts*:* Health Care Non-Healthcare	**Unit**	**Granularity**	**Constraints & Use**** Requirements	**Time** Span/ Longitudinality	**Scope**
ACS American Community Survey	Extensive geographic and demographic drill downs on disability and health insurance.	Survey	Census Bureau	Healthcare 17 Non-Healthcare 199	Person	Macro/ Microdata	Ready-to-Use/ Requirements	2005-present Cross-sectional	National
ASEC Annual Social and Economic Supplement to the CPS	Labor force data with health insurance, out-of-pocket $ and health status fields.	Survey	BLS (Census Bureau)	Healthcare 182 Non-Healthcare 479	Person	Macro/ Microdata	Ready-to-Use/ Requirements	1998-present Cross-sectional	National
SIPP Survey of Income and Program Participation	Premier source of information on income and program participation. Addresses health insurance.	Survey	Census Bureau	Healthcare 253 Non-Healthcare 2,406	Person	Microdata	Requirements	1984-present Longitudinal	National
MEPS-HC Medical Expenditure Panel Survey, Household Component	Person-level health expenditures with longitudinal capabilities.	Survey	AHRQ (Westat)	Healthcare 1,252 Non-Healthcare 330	Person	Macro/ Microdata	Ready-to-Use/ Requirements	1996-present Cross-sectional Longitudinal	National
NHANES National Health and Nutrition Examination Survey	Survey combines interviews and physical examination, including lab tests.	Survey	NCHS/CDC	Healthcare 1,733 Non-Healthcare 147	Person	Microdata	Requirements	1999-present Cross-sectional	National
NHIS National Health Interview Survey	Principal source of information on health of U.S. population. Robust demographic, socioeconomic data.	Survey	CDC (NCHS)	Healthcare 1,388 Non-Healthcare 212	Person	Macro/ Microdata	Ready-to-Use/ Requirements	1963-present Cross-sectional	National
MEPS-IC Medical Expenditure Panel Survey, Insurance Component	Factors contributing to use of employer sponsored insurance. Premiums and cost sharing.	Survey	AHRQ (Census Bureau)	Healthcare 153 Non-Healthcare 22	Employers	Macrodata	Ready-to-use	1996-present Cross-sectional	National
Medicaid (various program data)	Actual enrollment data. Breakouts of new eligibility categories created by ACA.	Admin	CMS	Healthcare 30 Non-Healthcare 8	Person	Macrodata	Ready-to-use	Span[†] Cross-sectional	National
NHEA National Health Expenditure Accounts	Official estimates of healthcare spending in U.S. Includes care, admin, research and infrastructure.	Multiple Sources	HHS	Healthcare 640 Non-Healthcare 14	Services, Payers, Sponsors	Macrodata	Ready-to-use	1960-present Cross-sectional	National

BLS: Bureau of Labor Statistics
AHRQ: Agency for Healthcare Research and Quality
NCHS: National Center for Health Statistics
CDC: Centers for Disease Control and Prevention
CMS: Centers for Medicare and Medicaid Services
HHS: Department of Health and Human Services

[*] “Counts of variables” by topic is a reasonable method of determining a dataset’s areas of focus. Each variable from the nine selected datasets has been categorized by subject matter. All ultimately roll up to either “non-healthcare” or “healthcare.” Additional detail on content is provided below. (Counts exclude sample weights and variables related to survey administration.)

[**] All selected datasets are free. Macrodata versions are typically easy- or ready-to-use. For the microdata versions, users must have database management skills and ability to generate population estimates (relatively easy) and margins of error (more complicated) from raw survey data. Microdata are typically too large for MS Excel or MS Access and require database management/statistical analysis software and the corresponding hardware.

[†] Varies by program. Performance Indicator Project (PIP): Oct 2013-present. Medicaid Budget & Expenditure System (MBES): Jan 2014-present. Statistical Enrollment Data System (SEDS): Oct 2013-present.

Stern Consulting LLC

Case Study:

"Datasets to Evaluate the Impact of National Healthcare Policy"

Selected Datasets by Dimension: Content

"Variable count" by topic reveals each dataset's areas of focus.

Note that even the datasets not focused on healthcare contain extensive economic, social and demographic data, by which the limited health data they do contain, can be analyzed.

Stern Consulting LLC

Dimensions of Datasets

Datasets	Notes	Mechanism of Data Generation	Sponsor (Data Collector)	Content *Variable Counts*:* Health Care Non-Healthcare	Unit	Granularity	Constraints & Use** Requirements	Time Span/ Longitudinality	Scope
ACS American Community Survey	Extensive geographic and demographic drill downs on disability and health insurance.	Survey	Census Bureau	Healthcare 17 Non-Healthcare 199	Person	Macro/ Microdata	Ready-to-Use/ Requirements	2005-present Cross-sectional	National
ASEC Annual Social and Economic Supplement to the CPS	Labor force data with health insurance, out-of-pocket $ and health status fields.	Survey	BLS (Census Bureau)	Healthcare 182 Non-Healthcare 479	Person	Macro/ Microdata	Ready-to-Use/ Requirements	1998-present Cross-sectional	National
SIPP Survey of Income and Program Participation	Premier source of information on income and program participation. Addresses health insurance.	Survey	Census Bureau	Healthcare 253 Non-Healthcare 2,406	Person	Microdata	Requirements	1984-present Longitudinal	National
MEPS-HC Medical Expenditure Panel Survey, Household Component	Person-level health expenditures with longitudinal capabilities.	Survey	AHRQ (Westat)	Healthcare 1,252 Non-Healthcare 330	Person	Macro/ Microdata	Ready-to-Use/ Requirements	1996-present Cross-sectional Longitudinal	National
NHANES National Health and Nutrition Examination Survey	Survey combines interviews and physical examination, including lab tests.	Survey	NCHS/CDC	Healthcare 1,733 Non-Healthcare 147	Person	Microdata	Requirements	1999-present Cross-sectional	National
NHIS National Health Interview Survey	Principal source of information on health of U.S. population. Robust demographic, socioeconomic data.	Survey	CDC (NCHS)	Healthcare 1,388 Non-Healthcare 212	Person	Macro/ Microdata	Ready-to-Use/ Requirements	1963-present Cross-sectional	National
MEPS-IC Medical Expenditure Panel Survey, Insurance Component	Factors contributing to use of employer sponsored insurance. Premiums and cost sharing.	Survey	AHRQ (Census Bureau)	Healthcare 153 Non-Healthcare 22	Employers	Macrodata	Ready-to-use	1996-present Cross-sectional	National
Medicaid (various program data)	Actual enrollment data. Breakouts of new eligibility categories created by ACA.	Admin	CMS	Healthcare 30 Non-Healthcare 8	Person	Macrodata	Ready-to-use	Span[†] Cross-sectional	National
NHEA National Health Expenditure Accounts	Official estimates of healthcare spending in U.S. Includes care, admin, research and infrastructure.	Multiple Sources	HHS	Healthcare 640 Non-Healthcare 14	Services, Payers, Sponsors	Macrodata	Ready-to-use	1960-present Cross-sectional	National

BLS: Bureau of Labor Statistics
AHRQ: Agency for Healthcare Research and Quality

NCHS: National Center for Health Statistics
CDC: Centers for Disease Control and Prevention

CMS: Centers for Medicare and Medicaid Services
HHS: Department of Health and Human Services

[*] "Counts of variables" by topic is a reasonable method of determining a dataset's areas of focus. Each variable from the nine selected datasets has been categorized by subject matter. All ultimately roll up to either "non-healthcare" or "healthcare." Additional detail on content is provided below. (Counts exclude sample weights and variables related to survey administration.)

[**] All selected datasets are free. Macrodata versions are typically easy- or ready-to-use. For the microdata versions, users must have database management skills and ability to generate population estimates (relatively easy) and margins of error (more complicated) from raw survey data. Microdata are typically too large for MS Excel or MS Access and require database management/statistical analysis software and the corresponding hardware.

[†] Varies by program. Performance Indicator Project (PIP): Oct 2013-present. Medicaid Budget & Expenditure System (MBES): Jan 2014-present. Statistical Enrollment Data System (SEDS): Oct 2013-present.

Case Study:

“Datasets to Evaluate the Impact of National Healthcare Policy”

Content Detail

The main healthcare and non-healthcare categories are broken out into increasingly finer levels of detail and standardized across datasets.

This allows the comparison of datasets by specific areas of strength. Here we see, for example, that the best resources for “Ability to Get Care” data are NHIS, MEPS-HC and NHANES (with SIPP also “on the board.”)

Stern Consulting LLC

🟥	Present
🟨	Better
🟩	Best

Variable Counts by Content Category Dataset

Content Categories			Datasets								
			Non-Healthcare Focused			Healthcare Focused					
Main	Sub-Category 1	Sub-Category 2	ACS	ASEC	SIPP	MEPS-HC	NHANES	NHIS	Medicaid	MEPS-IC	NHEA
Healthcare	Access	Ability to Get Care			19	178	171	263			
		Ability to Pay for Care	10	157	153	346	17	203	22	85	11
	Cost	Charges				19					
		Encounters			12	21	4	17			
		Expenditures				300			8		586
		Expenditures by Sponsor		15	20			4		66	43
	Healthiness	Behavior/Attitude				4	224	62			
		Body Composition				1	151	8			
		Child-Specific Problems			29	28		10			
		Clinical Results					381				
		Condition			2	81	306	702			
		Days Lost Due to Illness			3	3		5			
		Diet					406				
		Functional Limitation	7	9	14	39	55	79			
		Status		1	1	32	18	35			
Non-Healthcare	Demographics	Age	2	3	14	7	5	3			1
		Population									1
		Race/Ethnicity	12	7	3	8	2	15			
		Sex	1	1	11	1	2	3			1
	Social	Child Care	🟥	9	47	🟨	🟥	3			
		Education	7	4	24	6	4	4			
		Household Composition	33	77	314	65	88	42			
		Heritage	17	5	14	7	19	7			
		Internet/Computer Use	11	🟩	🟨	🟨	1	7			
		Marital Status	6	2	8	8	2	5			
		Migration	3	10	3	🟩	🟥	🟨		🟥	
		Military Status	13	5	13	6	2	11		🟨	
		Neighborhood			4			5			
	Geographic Level	Specified Levels	8	13	7	4		4	8	4	4
	Economics	Income	13	211	428	25	3	51			
		Other Benefits	1	19	41		10	8			
		Assets	1	1	267						
		Debt			108						
		Taxes		20	5	6					
		Paid Support			17						
		Labor Force	13	37	221	48	5	15		5	
		Job Characteristics	15	47	800	139	4	12		14	
		Problems Paying Bills			2						
		Food Security			8			10			
		Indicators									6
	Housing	Financial	26	3	34			1			
		Physical	17	5	13			6			

Case Study:

“Datasets to Evaluate the Impact of National Healthcare Policy”

Content Comparison

Example: Ability to Get Care

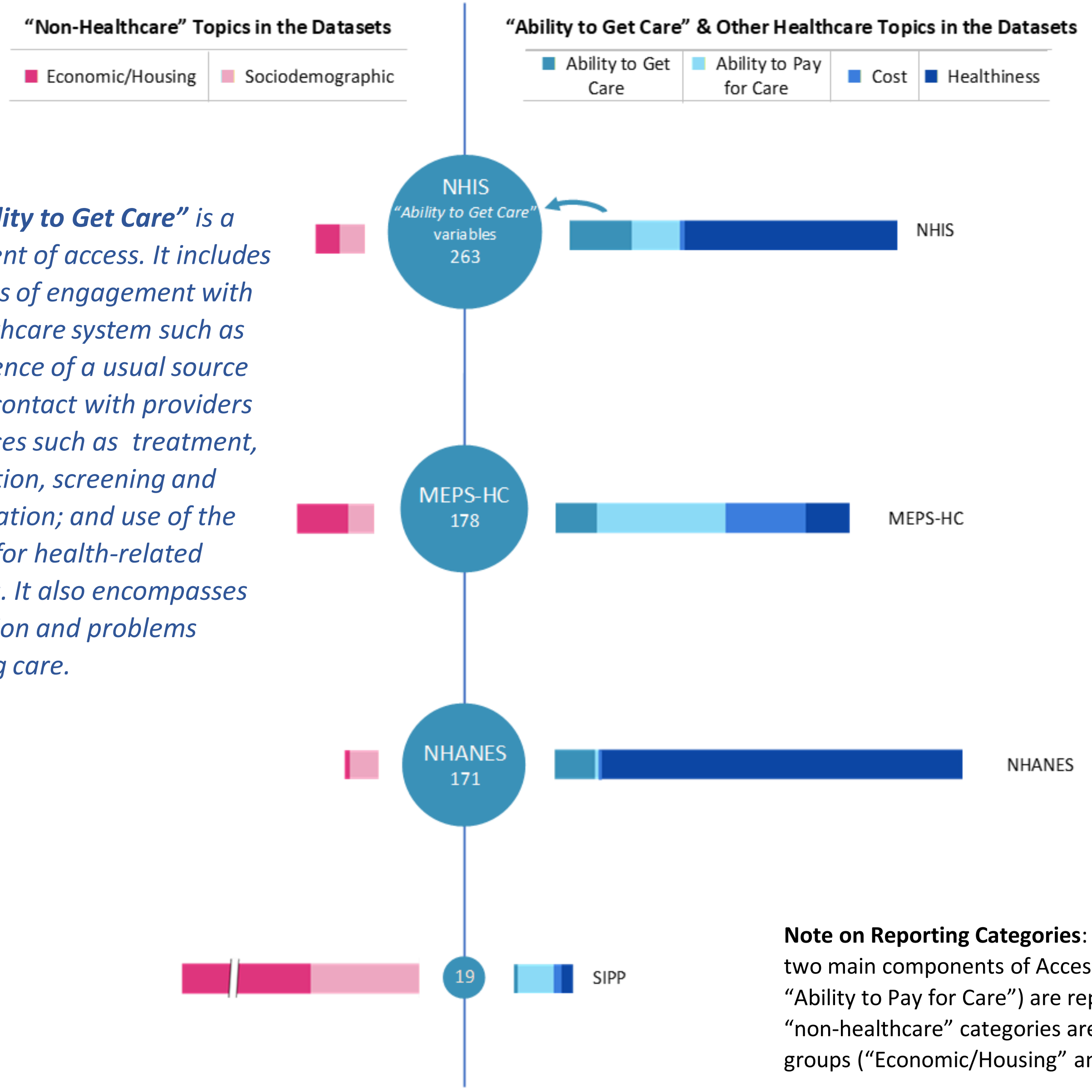
The framework enables comparison of datasets in multiple ways. This example shows

- *A ranking of the selected datasets according to their treatment of the “Ability to Get Care” (based on counts of relevant variables), and*
- *The availability of other variables for cross classification.*

Stern Consulting LLC

Datasets with “Ability to Get Care” Data

(Bars show the relative variable counts for content categories within each dataset. Bubbles compare the “Ability to Get Care” category across datasets.)



*The **“Ability to Get Care”** is a component of access. It includes indicators of engagement with the healthcare system such as the presence of a usual source of care; contact with providers for services such as treatment, consultation, screening and immunization; and use of the internet for health-related purposes. It also encompasses satisfaction and problems obtaining care.*

Note on Reporting Categories: For purposes of this chart, the two main components of Access (“Ability to Get Care” and “Ability to Pay for Care”) are reported separately, while the “non-healthcare” categories are collapsed into two roll-up groups (“Economic/Housing” and “Sociodemographic.”)

ACS, ASEC, Medicaid (eligibility data), MEPS-IC and NHEA contain no “ability to get care” data.

Framework for Sorting Through the Data Universe

(version 2.0)

Creating a Variable Classification System

One of our most important efforts to make this framework possible, was to standardize and classify variables across the profiled datasets. Each variable has up to nine levels of classification.

Once categorized, variables may be

- ***Counted** to determine the focus of datasets, and*
- ***"Searched on"** to find variables relevant to specific topics.*

This process is reflected through the "Content" dimension of the framework.

Variable Classification: Example

Main	Sub-cat. 1	Sub-cat. 2	Sub-cat. 3	Sub-cat 4	*Additional sub-categories* →
Healthcare	**Access**	**Ability to Get Care**	**Source of Care**	(Y/N)	
				Type	
				Purpose	
				Provider Race	
				Provider Sex	
				Time to Get to	
				Reason Without	
				Caregiver Assistance	
				Doctor Treats Adults and Children	
				Impact of Health Insurance	
			Contact with System	Contact by Type of Service	
				Immunization	
				Instruction	
				Internet/Computer Use for Healthcare	
				Prophylactic Medications	
				Screening	
				Seen/Talked to Health Professional	
				Treatment	
			Problems Obtaining Care		
			Satisfaction		
		Ability to Pay for Care	**Health Insurance**		

We built on the existing sponsor categories where indicated and filled in the gaps as needed. This work in progress currently includes hundreds of categories for thousands of variables, and we are continually refining it.

Stern Consulting LLC

Framework for Sorting Through the Data Universe

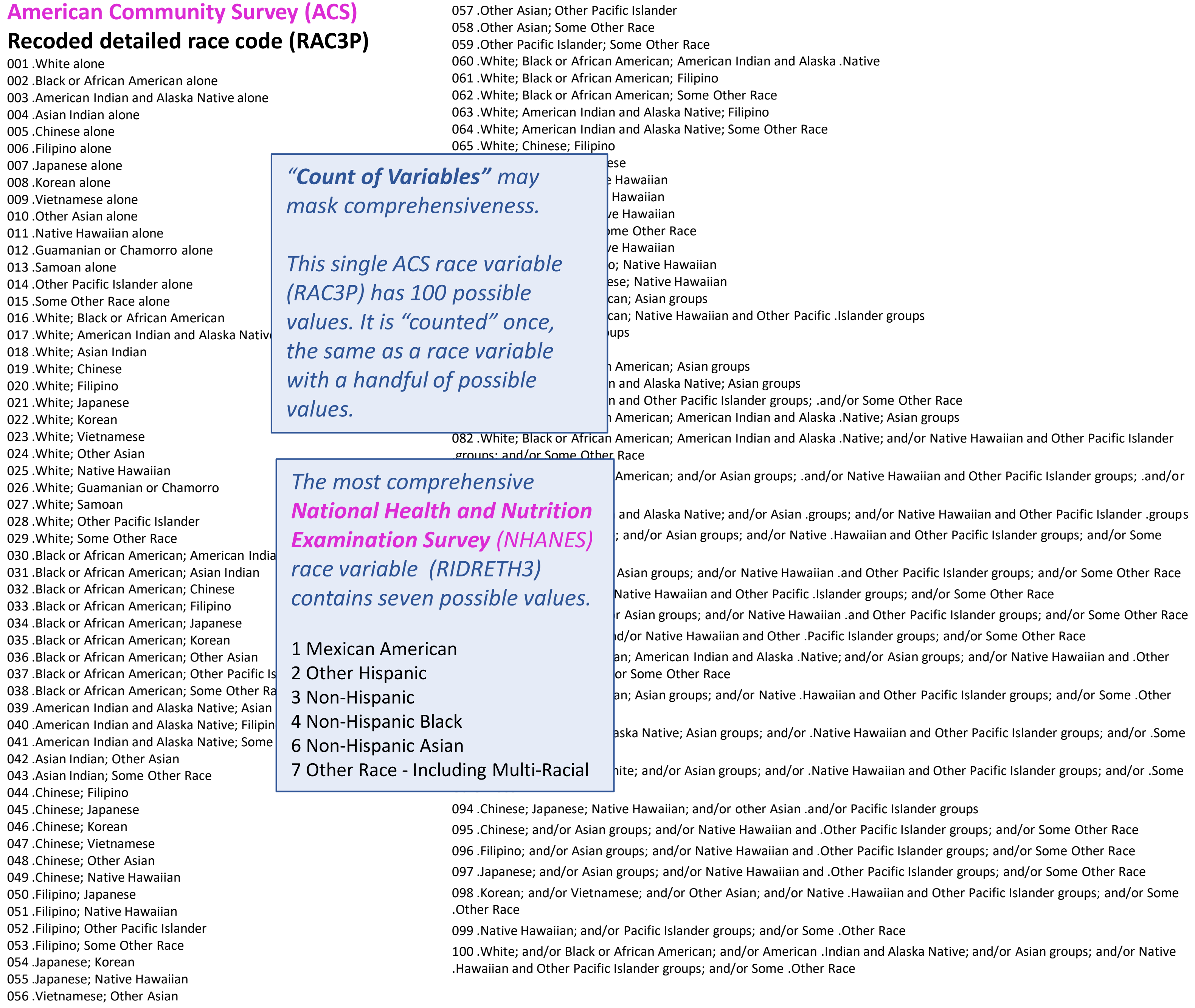
(version 2.0)

Additional Applications of the Framework

Variable Composition

Analyzing the possible values for a given variable provides an additional opportunity for comparing datasets.

In this example, note that the most comprehensive “race” variable in the American Community Survey (ACS) contains 100 possible values. In the National Health and Nutrition Examination Survey (NHANES) the most comprehensive race variable contains seven.



American Community Survey (ACS)

Recoded detailed race code (RAC3P)

001 .White alone
002 .Black or African American alone
003 .American Indian and Alaska Native alone
004 .Asian Indian alone
005 .Chinese alone
006 .Filipino alone
007 .Japanese alone
008 .Korean alone
009 .Vietnamese alone
010 .Other Asian alone
011 .Native Hawaiian alone
012 .Guamanian or Chamorro alone
013 .Samoan alone
014 .Other Pacific Islander alone
015 .Some Other Race alone
016 .White; Black or African American
017 .White; American Indian and Alaska Nativ
018 .White; Asian Indian
019 .White; Chinese
020 .White; Filipino
021 .White; Japanese
022 .White; Korean
023 .White; Vietnamese
024 .White; Other Asian
025 .White; Native Hawaiian
026 .White; Guamanian or Chamorro
027 .White; Samoan
028 .White; Other Pacific Islander
029 .White; Some Other Race
030 .Black or African American; American India
031 .Black or African American; Asian Indian
032 .Black or African American; Chinese
033 .Black or African American; Filipino
034 .Black or African American; Japanese
035 .Black or African American; Korean
036 .Black or African American; Other Asian
037 .Black or African American; Other Pacific Is
038 .Black or African American; Some Other Ra
039 .American Indian and Alaska Native; Asian
040 .American Indian and Alaska Native; Filipin
041 .American Indian and Alaska Native; Some
042 .Asian Indian; Other Asian
043 .Asian Indian; Some Other Race
044 .Chinese; Filipino
045 .Chinese; Japanese
046 .Chinese; Korean
047 .Chinese; Vietnamese
048 .Chinese; Other Asian
049 .Chinese; Native Hawaiian
050 .Filipino; Japanese
051 .Filipino; Native Hawaiian
052 .Filipino; Other Pacific Islander
053 .Filipino; Some Other Race
054 .Japanese; Korean
055 .Japanese; Native Hawaiian
056 .Vietnamese; Other Asian
057 .Other Asian; Other Pacific Islander
058 .Other Asian; Some Other Race
059 .Other Pacific Islander; Some Other Race
060 .White; Black or African American; American Indian and Alaska .Native
061 .White; Black or African American; Filipino
062 .White; Black or African American; Some Other Race
063 .White; American Indian and Alaska Native; Filipino
064 .White; American Indian and Alaska Native; Some Other Race
065 .White; Chinese; Filipino
ese
e Hawaiian
Hawaiian
ve Hawaiian
ome Other Race
ve Hawaiian
o; Native Hawaiian
ese; Native Hawaiian
can; Asian groups
can; Native Hawaiian and Other Pacific .Islander groups
ups
n American; Asian groups
n and Alaska Native; Asian groups
n and Other Pacific Islander groups; .and/or Some Other Race
n American; American Indian and Alaska .Native; Asian groups
082 .White; Black or African American; American Indian and Alaska .Native; and/or Native Hawaiian and Other Pacific Islander .groups; and/or Some Other Race
American; and/or Asian groups; .and/or Native Hawaiian and Other Pacific Islander groups; .and/or
and Alaska Native; and/or Asian .groups; and/or Native Hawaiian and Other Pacific Islander .groups
; and/or Asian groups; and/or Native .Hawaiian and Other Pacific Islander groups; and/or Some
Asian groups; and/or Native Hawaiian .and Other Pacific Islander groups; and/or Some Other Race
Native Hawaiian and Other Pacific .Islander groups; and/or Some Other Race
r Asian groups; and/or Native Hawaiian .and Other Pacific Islander groups; and/or Some Other Race
d/or Native Hawaiian and Other .Pacific Islander groups; and/or Some Other Race
an; American Indian and Alaska .Native; and/or Asian groups; and/or Native Hawaiian and .Other
or Some Other Race
an; Asian groups; and/or Native .Hawaiian and Other Pacific Islander groups; and/or Some .Other
aska Native; Asian groups; and/or .Native Hawaiian and Other Pacific Islander groups; and/or .Some
hite; and/or Asian groups; and/or .Native Hawaiian and Other Pacific Islander groups; and/or .Some
094 .Chinese; Japanese; Native Hawaiian; and/or other Asian .and/or Pacific Islander groups
095 .Chinese; and/or Asian groups; and/or Native Hawaiian and .Other Pacific Islander groups; and/or Some Other Race
096 .Filipino; and/or Asian groups; and/or Native Hawaiian and .Other Pacific Islander groups; and/or Some Other Race
097 .Japanese; and/or Asian groups; and/or Native Hawaiian and .Other Pacific Islander groups; and/or Some Other Race
098 .Korean; and/or Vietnamese; and/or Other Asian; and/or Native .Hawaiian and Other Pacific Islander groups; and/or Some .Other Race
099 .Native Hawaiian; and/or Pacific Islander groups; and/or Some .Other Race
100 .White; and/or Black or African American; and/or American .Indian and Alaska Native; and/or Asian groups; and/or Native .Hawaiian and Other Pacific Islander groups; and/or Some .Other Race

*“**Count of Variables**” may mask comprehensiveness.*

This single ACS race variable (RAC3P) has 100 possible values. It is “counted” once, the same as a race variable with a handful of possible values.

*The most comprehensive **National Health and Nutrition Examination Survey** (NHANES) race variable (RIDRETH3) contains seven possible values.*

1 Mexican American
2 Other Hispanic
3 Non-Hispanic
4 Non-Hispanic Black
6 Non-Hispanic Asian
7 Other Race - Including Multi-Racial

Stern Consulting LLC

Framework for Sorting Through the Data Universe

(version 2.0)

Additional Applications of the Framework

Opportunities for Linkage

The framework provides a mechanism for identifying variables where linkage among datasets is possible.

For example, several of the datasets include "Region" and "State" enabling linkage at those levels.

ACS provides, by far, the most geographic levels. Its granularity permits linkage to a wide range of datasets beyond those profiled herein.

Stern Consulting LLC

Opportunities for Linkage: Example

Geographic Levels by Data Source

Geographic Level (All are national in scope.)	Data Sources									
	Microdata & Macrodata							Macrodata Only		
	ACS		ASEC	MEPS-HC	NHANES	NHIS	SIPP	Medicaid	MEPS-IC	NHEA
	Microdata	Macrodata								
Nation	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Region	✓	✓	✓	✓		✓	✓			✓*
Census Division	✓	✓	✓						✓	
State	✓	✓	✓				✓	✓	✓	✓
Consolidated Statistical Area			✓							
Core Based Statistical Area			✓							
County		✓	✓							
Metropolitan Status		✓	✓							
Principal City		✓	✓							
Public Use Microdata Area	✓	✓								
Additional 160 Levels		✓								

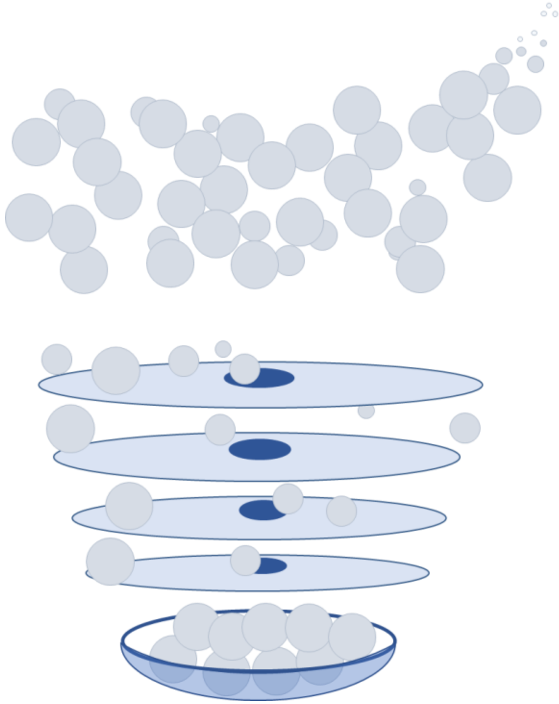
* NHEA uses different "region" categories than the other datasets and is therefore not linkable on that variable.

Note: Linkage at the person level is typically not possible with public use files due to confidentiality. However, for certain datasets and subject to approval, person-linkable versions are available onsite at designated research data centers.

Summary

In Conclusion, We Have:

- Developed a framework for selecting among 1000's of datasets based on
  - Eight data dimensions
  - Four user considerations
- Created detailed standardization and mapping of variables across datasets
- Applied this framework and mapping to identify nine datasets most apt for evaluation of the impact of national healthcare policy on the U.S. population.



Stern Consulting LLC

Stern Consulting LLC
Data-driven support for the healthcare community

AcademyHealth

About

Stern Consulting

Stern Consulting LLC provides specialty analytic and consulting services to healthcare leaders, hospital systems, healthcare companies, and investors. For more information, see www.sternconsulting.com.

Brian Williams oversees Stern Consulting's healthcare database and analysis functions. He has over 30 years of experience in healthcare decision support. Mr. Williams earned an MBA with concentration in healthcare management from Boston University and a BA in political science from Holy Cross College.
bwilliams@sternconsulting.com

David M. C. Stern is the founder and president of Stern Consulting. He has advised some of the pre-eminent organizations in the healthcare industry. He earned an MBA from the Yale School of Management and a BA in economics from Yale College.
dstern@sternconsulting.com

AcademyHealth

AcademyHealth is a leading national organization for health services researchers, policymakers, and health care practitioners and stakeholders. AcademyHealth – together with its members – increases the understanding of methods and data used in the field, enhances the professional skills of researchers and research users, and expands awareness.

The current slides were prepared for presentation at AcademyHealth's 2021 Health Datapalooza and National Health Policy Conference.

Stern Consulting LLC