



Tidy Data

Principles for Standardized Scientific
Data Submission and Reuse

Tyler Aguilar, MS

LSDA Data Scientist | Leidos

Agenda

- LSDA Data Landscape
- Data Preparation for Modern Analysis
- Tidy Data Framework
- Modernizing Legacy Research Datasets
- Benefits of Tidy Data for Research
- Modern Workflow Integration
- Future Work





LSDA Data Landscape

- LSDA houses human, animal, and plant research data across 6 decades
- Highly varied datasets, legacy spreadsheets were structured for human readability and the Excel toolkit
- Modern analysis requires machine-readable datasets for commonly used tools and programming languages (PowerBI, Python, R)
- Going from data to decision first requires restructuring and cleaning





Legacy Data Spreadsheets

Sample Spine Data

Site:	JSC			Proctor:	Smith		Date Analyzed		5/3/17			
UTMB	C1			Doctor:	Alice							
				Area:	Spine							
Subject	Sex	Age	BR	L1 BMD	L1 Area	L1 BMC	L2 BMD	L2 Area	L2 BMC	L3 BMC	L3 Area	L3 BMC
S0001	M	40	PRE	0.91166	23.7882	48.729	0.91884	46.6289	50.3922	0.82691	21.0317	81.437
S0003	M	35	PRE	0.90211	24.4419	46.1183	0.91055	47.1031	49.112	0.81177	20.5528	79.552

- Wide, spreadsheet-oriented table
- Variables encoded in column names (e.g., L1_BMD, L1_Area)
- Multiple variables represented across columns
- Adding new measurements requires new columns





Tidy Data Principles

- Each variable in a column
- Each observation in a row
- Each cell contains a single value
- Summary statistics should not be mixed with raw observations

country	year	cases	population
Afghanistan	1999	745	15467071
Afghanistan	2000	2666	20995360
Brazil	1999	31737	17206362
Brazil	2000	80488	17404898
China	1999	211258	127215272
China	2000	210766	128042583

variables

country	year	cases	population
Afghanistan	1999	745	15467071
Afghanistan	2000	2666	20995360
Brazil	1999	31737	17206362
Brazil	2000	80488	17404898
China	1999	211258	127215272
China	2000	210766	128042583

observations

country	year	cases	population
Afghanistan	1999	745	15467071
Afghanistan	2000	2666	20995360
Brazil	1999	31737	17206362
Brazil	2000	80488	17404898
China	1999	211258	127215272
China	2000	210766	128042583

values





The Same Data in Tidy Format

Subject	Sex	Age	Test_Phase	Region	Measure	Value	Units
S0001	M	40	PRE_TEST	L1	BMD	0.911662	g/cm^2
S0001	M	40	PRE_TEST	L1	Area	23.78829	cm^2
S0001	M	40	PRE_TEST	L1	BMC	48.72902	g
S0001	M	40	PRE_TEST	L2	BMD	0.918842	g/cm^2
:	:	:	:	:	:	:	:
S0003	M	35	PRE_TEST	L2	BMC	49.11204	g
S0003	M	35	PRE_TEST	L3	BMD	0.811774	g/cm^2
S0003	M	35	PRE_TEST	L3	Area	20.55281	cm^2
S0003	M	35	PRE_TEST	L3	BMC	79.55212	g

- Each column represents a single, explicit variable
- The same operations now apply uniformly across the dataset





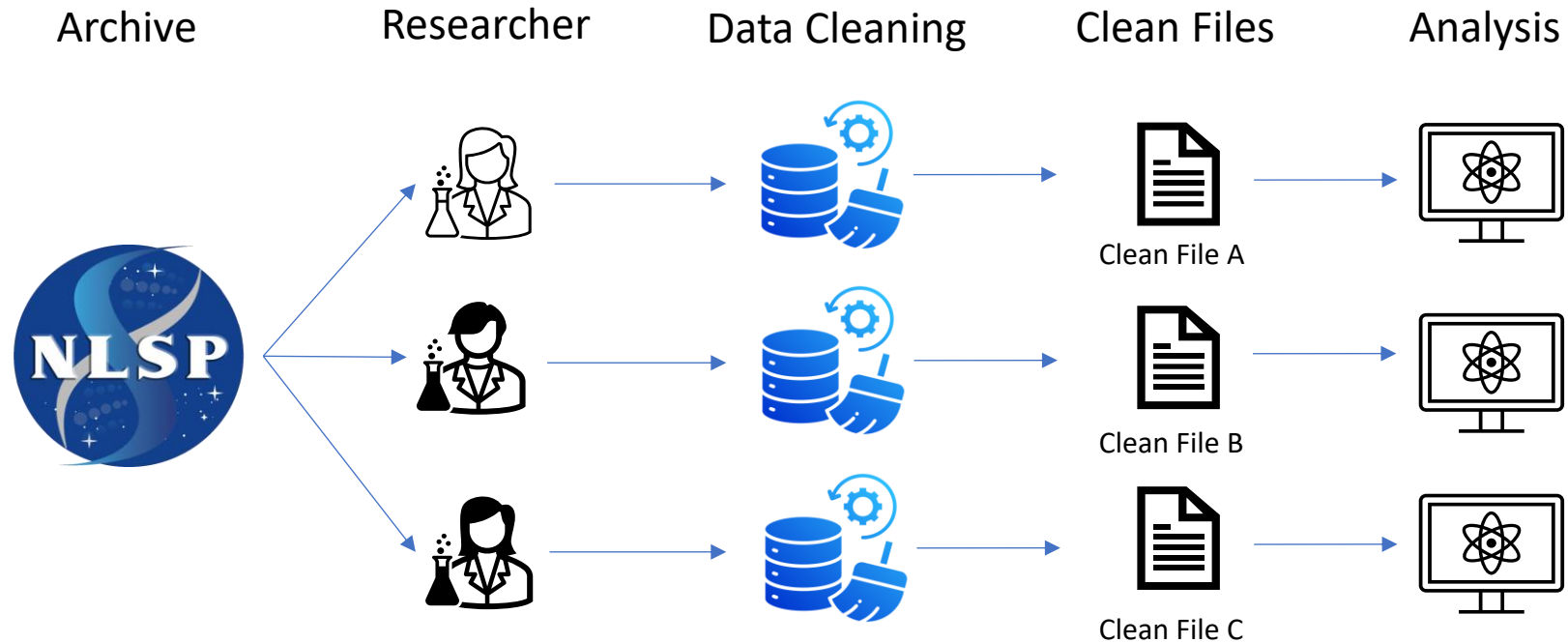
Data Preparation for Modern Analysis

- Finding significant trends in data requires a large enough amount to not be purely anecdotal
- Analysts must reshape tables, isolate variables, and separate observational and summary data for every sheet used in analysis
- Preparation effort has a direct relationship with the amount of data required for insight
- Steps must be repeated for each new user



From Raw to Reusable

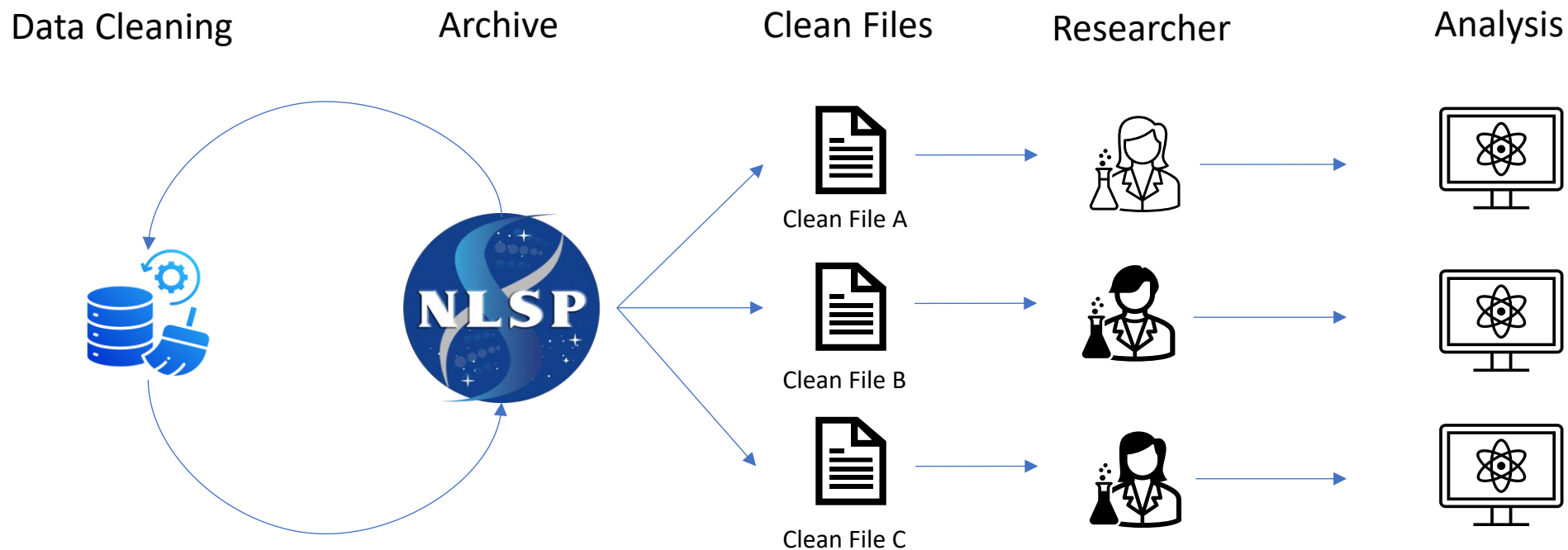
“Data preparation accounts for about 80% of the work of data scientists”





From Raw to Reusable

“Data preparation accounts for about 80% of the work of data scientists”





Legacy Data, Modern Use

- High-value human research data is now machine-readable and analysis-ready
- Legacy analog datasets (HDT, countermeasures, exercise) have been standardized into tidy format, the standard format for data submissions since 2021
- Eliminates manual data preparation and restructuring
- Enables direct use in Python, R, and BI tools
- Supports cross-study comparison and integration





Public Data, Open Science

- Released as publicly available datasets
- Reduces request timelines and removes access barriers
- Enables use by NASA researchers, external investigators, and citizen scientists
- Standardized structure supports reproducibility and broader collaboration
- Aligns with FAIR principles: Findable, Accessible, Interoperable, Reusable
- Preserves and extends the value of historical human research data while enabling open science

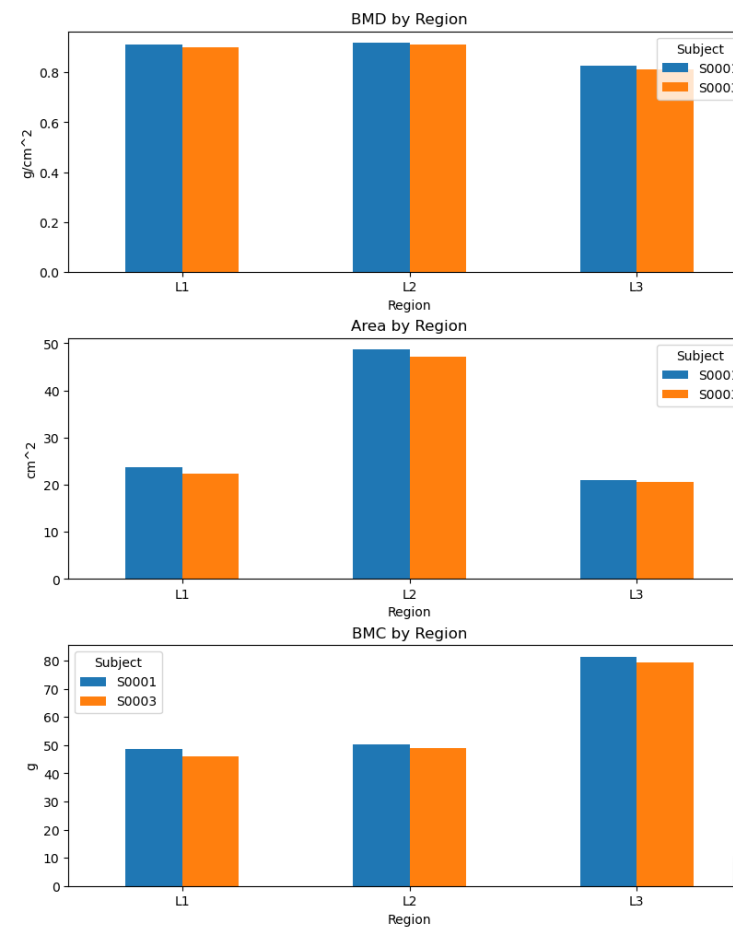




Enabling Reuse for Modern Use Cases

Data Analytics

- Rapid exploration of physiological changes across campaigns
- Comparison of outcomes across subjects, studies, and timepoints
- Identification of relationships and correlations between isolated variables (EDA)
- Move from data access to insight without preprocessing delays

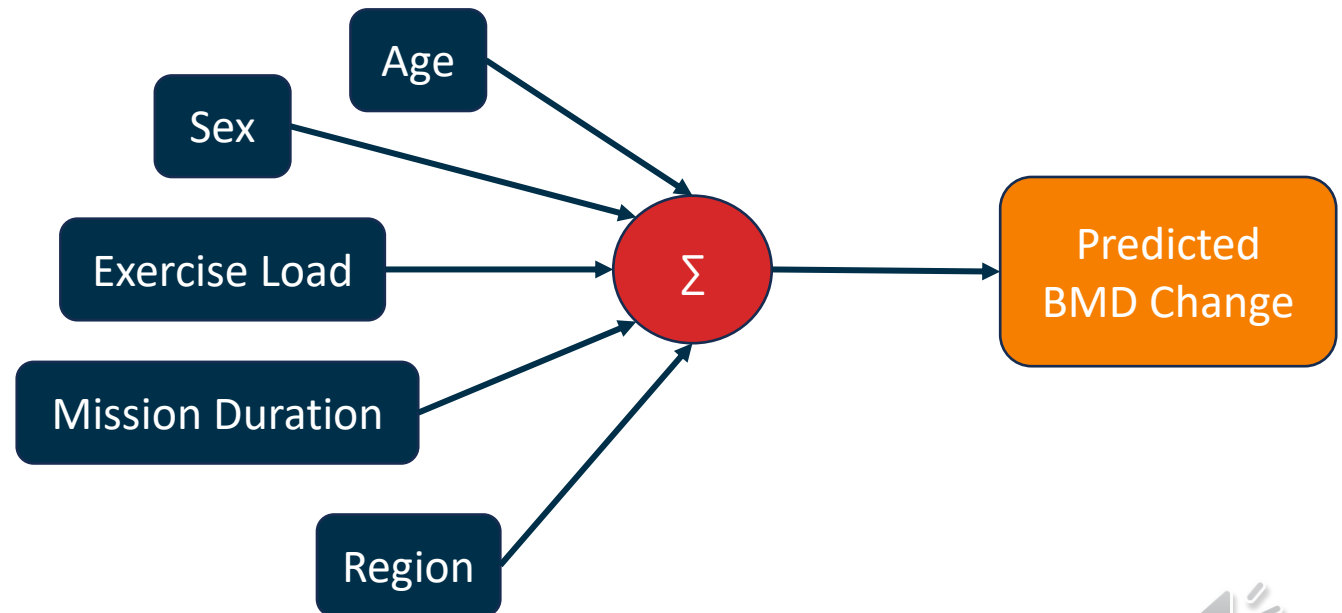




Enabling Reuse for Modern Use Cases

Machine Learning

- Variables map directly to model-ready features
- Combined datasets enable richer modeling of risk factors
- Consistent features across studies and timepoints
- New data integrates without pipeline rework
- Faster path from data to prediction

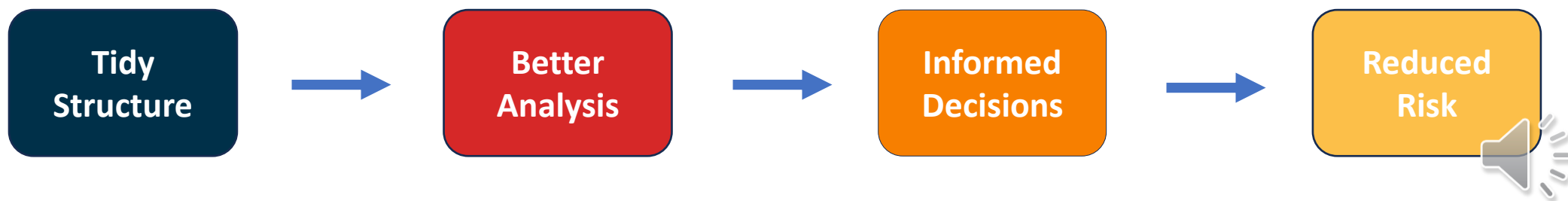




Enabling Reuse for Modern Use Cases

Artificial Intelligence Pipelines

- Tidy structure aligns with AI input preference, no decoding required
- Variables are isolated and atomic, enabling AI to learn relationships without parsing combined or ambiguous fields
- Missing data is explicit, cutting down on AI misreads as zeros or errors
- Adding new observations to a tidy dataset doesn't change its structure, no need for retraining when new data is added





Publicly Available Tidy Datasets

- UTMB Campaign 1 & 3
 - Multi-study bed rest campaigns simulating microgravity's physiological effects across bone, muscle, and cardiovascular systems in healthy subjects
- CFT70
 - 70 days of 6 degree head-down tilt with integrated Countermeasure and Functional Testing
- MR035G
 - Longitudinal bone density tracking via DXA, measuring loss and recovery across the bed rest timeline
- MR079G
 - Isokinetic muscle strength and endurance testing of the knee, ankle, and back before and after bed rest





Future Work

- Tidy data format has been required for all submissions since 2021
- Identify predecessor datasets to modern Standard Measures
- Modernize these legacy datasets into tidy structures
- Align historical data with current data collection standards
- Expand the amount of reusable, analysis-ready LSDA data



Thank You

