Correlated Simulated Data for Decennial System of Systems (SoS) Development and Test Privacy by Design

Todd Johnsson, ExactData , Beverly Harris, U.S. Census Bureau

- Leverage of patented data generation platform, Dynamic Data Generator™
- Developed Census SoS specific data models for DDG
- Model of Models (MoM architecture) analogous to SoS architecture
- DDG engine and models, with automated data generation within AWS, stored in S3 bucket, pulled to Census Simulated Data Repository through Trusted Internet Connection (TIC)
- Outside-in data AND System-to-System data, in ingestible formats
- Data types: Survey responses, Paradata, Administrative records, Human Resource records, Field Operations records, Address product (MAFX), Geographic products (GRFC, GRFN)
- Characteristics: Correlated, Longitudinal, Artifacts, Ground Truth, Realistic
- Title 13 Compliance: INHERENTLY COMPLIANT, Privacy by Design
- Can generate dataset to most any US Geography within a day
- Able to generate a full country size dataset within a week (140M households, 350M people, 10+TB)
- 1000’s of files, consistent and keeping pace with ICD and MDR updates/versions – just a matter of regenerating data (from same model, with updated formatting)

Correlated Simulated Data for Decennial System of Systems Development and Test Privacy by Design

Todd Johnsson, ExactData

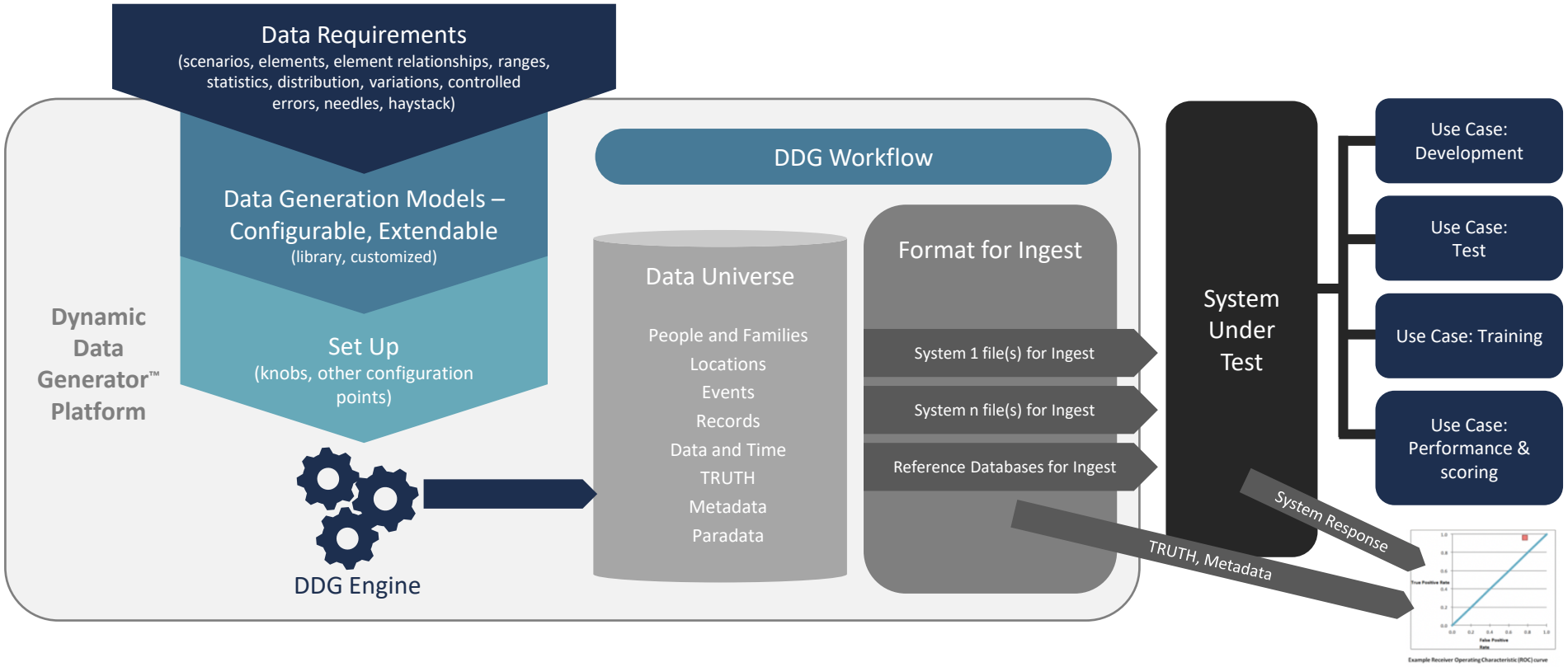
Beverly Harris, U.S. Census Bureau

The Simulated Data Requirements

- For Development and Test of 2020 Decennial Census SoS (System of Systems)…
- Needed data to simulate inputs/outputs of these systems, and data that could be introduced at SoS (System of Systems) entry points, and be processed throughout entire SoS. 1000's of output files in native file formats for direct ingestion.
- Needed data to be consistent across the systems (ex. Respondent data consistent with Admin Record data).
- Needed to not violate Title 13 (privacy of data provided to Census Bureau)
- Needed scale to simulate entire country – 140M addresses, datasets up to 10TB in size.
- Needed interwoven "fraud" scenarios
- Needed longitudinally consistent paradata– data/timestamps of Census website events.

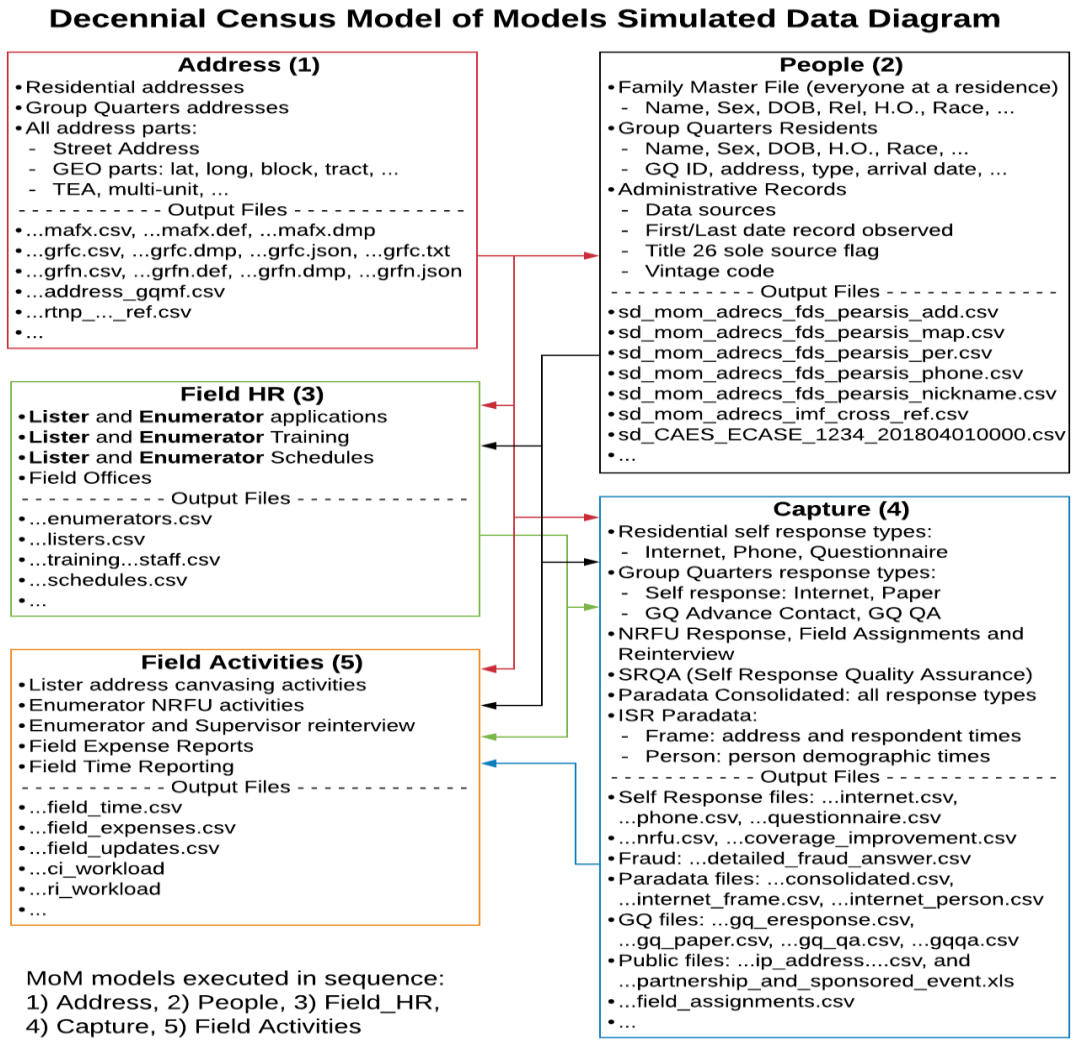
How to Generate Correlated Simulated Data

- Making data is easy… Making really good data is really hard

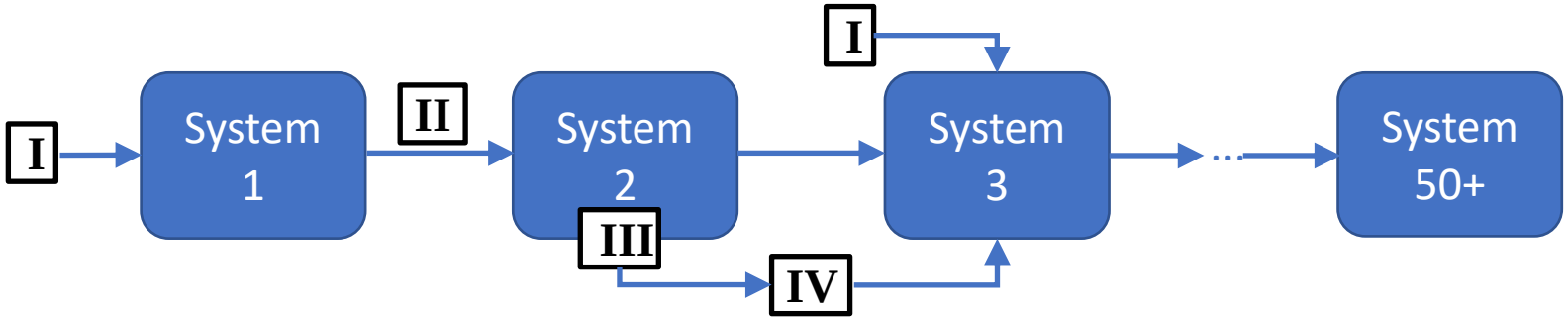


2020 Decennial Census Simulated Data

MoM Architecture



2020 Decennial Census
Simulated Data
Simulated Data Integration to System of Systems



I. Input Data to System of Systems
II. Interface Data between Systems (producer -> consumer)
III. Internal State of System
IV. External Data informed by System processing

2020 Decennial Census
Simulated Data
Data Types

- Survey responses
- Paradata
- Administrative records
- Human Resource records
- Field Operations records
- Address product (MAFX)
- Geographic products (GRFC, GRFN)

2020 Decennial Census
Simulated Data
Data Characteristics

- The simulated data itself was correlated across inputs and systems
- Was longitudinally consistent
- Had happy and non-happy path scenarios
- Had intended patterns
- Had event-level artifacts
- Had intended aggregate-level statistics
- Had ground truth
- Had the interrelated complexity and realism that sophisticated data processing application development and test require

2020 Decennial Census
Simulated Data
Title 13 Compliance

- The data sets are NOT a derivation of production data
- Fundamentally NOT de-identified data
- PII and Individuals cannot be re-identified, inherently
- If there IS a data breach in the Dev and Test environments, NO PII at risk
- Data is inherently safe
- Privacy by Design

2020 Decennial Census Simulated Data

Capabilities

- Datasets could be generated within a day, with millions of correlated records, to most any defined US geography
- Generated to scale (full country, 140 million families, 350 million people, 10+ TB) – these took ~ 1 week to generate
- Each dataset contained 1000's of files in various formats , types, schemas, MetaData Registry (MDR) versions conforming to Interface Control Documents for each Census Operational Delivery
- Census data models are….
  - **Configurable** – set "knobbed" configuration points, amongst 1000's of configuration points
  - **Extendable** – build into data model new scenarios requested from development teams and learned from test teams
  - **Maintainable** – can evolve gracefully with System of Systems developments progression
- Datasets maintained in AWS S3 bucket, pulled to Census through Trusted Internet Connection (TIC), to Census Simulated Data Repository, for real-time availability

11 Counties
(2018 Census Test)
800K households
100 output files

203 Counties
(representative sampling)
9M households
1,000 output files

Eastern Time Zone
70M households
10,000 output files

United States
140M households
30,000 output files

- Output File Formats
  - CSV
  - Oracle Dump
  - JSON
  - XML
  - Excel
  - Pipe Delimited Text
  - Image
  - Paper

United States Census Bureau

Additional Information

United States Census Bureau

2020 Decennial Census
Simulated Data
Quality: How do we know the data is good?

- Data model is made to generate to a specification, including a statistical definition
  - Like “real” is but one statistical definition
  - Many times the need is for a different specific statistical makeup
  - Data Model is made and the generated data is designed for the development and test needs of the System Under Test
- Utilization of “test driven development” techniques to ensure data model code passes “tests” as being developed
- With data generated from a model
  - “Fix once, fixed forever”
  - Quality Assurance (model is reliably repeatable) vs. Quality Controlled
- Automated analytics to validate requirements

2020 Decennial Census
Simulated Data
Expert Analysis Enabled
Fraud Detection

- Patterns of Fraudulent behavior emerge in dataset
- Relevant events, artifacts, records are “tagged” with precise understanding of their time-stamp and location (ground truth)
- Test becomes… Can your Fraud Detection System identify Fraud with high precision and low escape rates?
- ROC (Receiver Operating Characteristic) and Confusion Matrix analyses enabled
  - **True Negative** – System under Test (SUT) classified the event as “did NOT happen”, correctly.
  - **False Negative** – SUT classified event as “did NOT happen”, incorrectly
  - **True Positive** – SUT classified event as “Did happen”, correctly
  - **False Positive** – SUT classified event as “Did happen”, incorrectly

2020 Decennial Census
Simulated Data
Expert Analysis Enabled
Identity Resolution for Survey Responses and Administrative Records Integration

- Two streams of simulated data. Respondent data and Administrative Record data.
- All correlated, including longitudinally
- Can make very challenging, just like real life
- Examples
  - Names expressed differently
  - Moving in interim between tax record date and Census day
  - Getting married in interim between tax record date and Census day
  - Can make respondent data incomplete
- Leveraging ground truth, again enables ROC and Confusion Matrix analyses and drive to optimization
- Opportunity to more fully leverage in future surveys and integration of Administrative Records

2020 Decennial Census
Requirements: How did we develop?

- Input
  - Interface Control Documents (ICD)
  - MetaData Registry (MDR)
  - Subject Matter Expert descriptions of business logic and data correlations
  - Obfuscated/hand-crafted sample data
- Process: Gap Analysis
  - Evaluation of every data element “field” in ICD
- A benefit of Gap Analysis and/or generated simulated data was to reveal gaps between ICD and software, for reconciliation