Statistical Thinking and Big Data Challenges
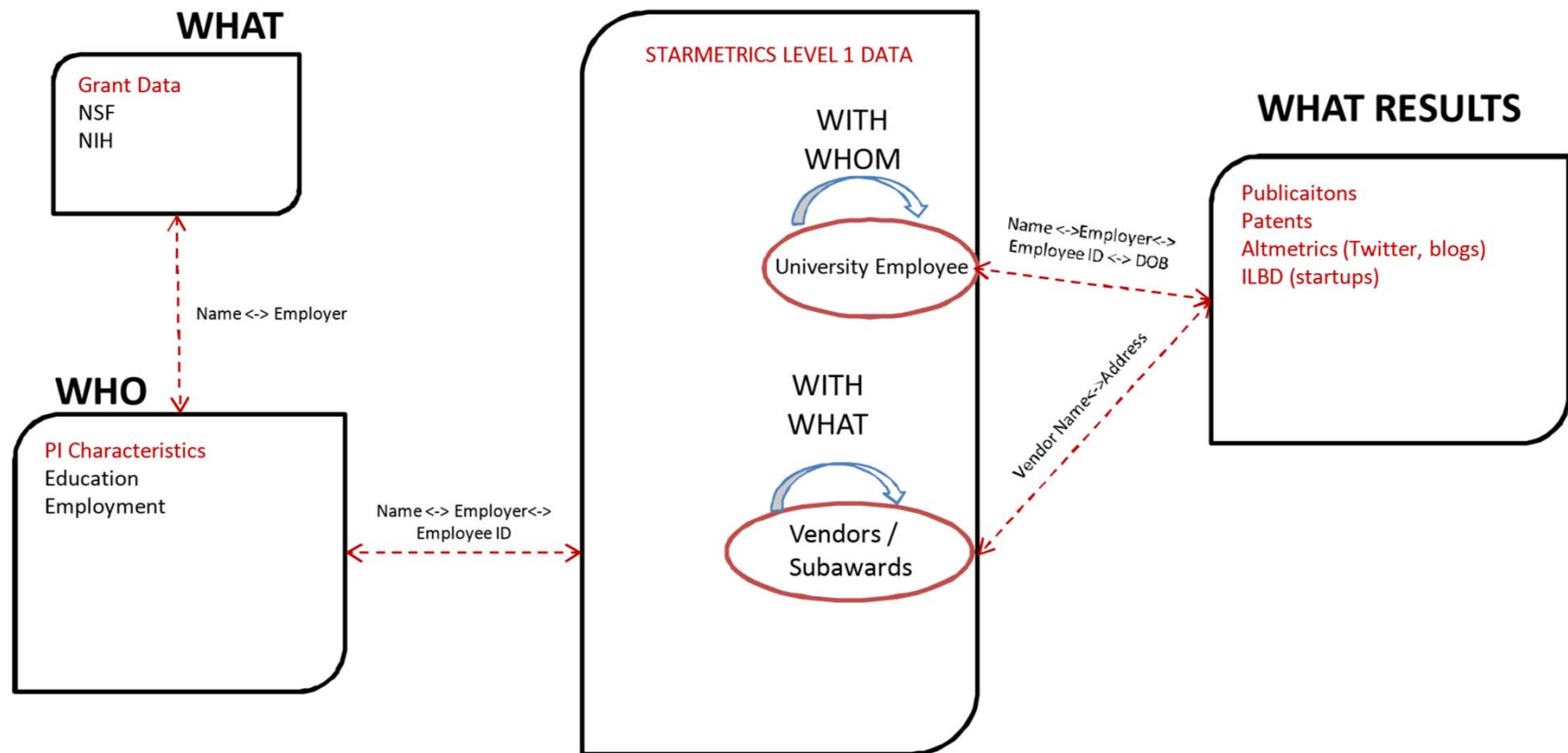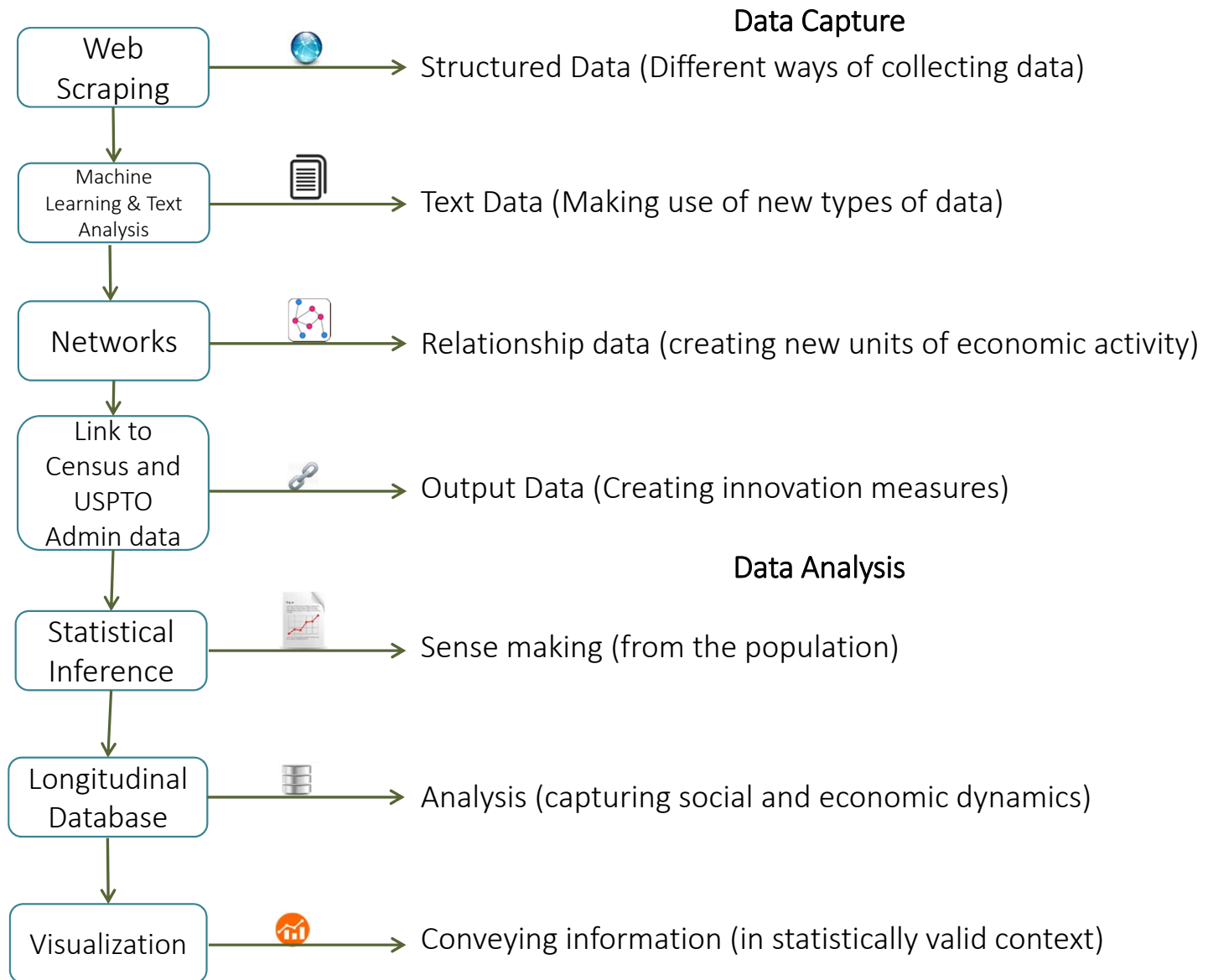
Frauke Kreuter (JPSM – University of Maryland)

Big Data for Federal Agencies Julia Lane + Ron Jarmin
What is the link between federal R&D and innovation?

Data Capture

Web Scraping → Structured Data (Different ways of collecting data)

Machine Learning & Text Analysis → Text Data (Making use of new types of data)

Networks → Relationship data (creating new units of economic activity)

Link to Census and USPTO Admin data → Output Data (Creating innovation measures)

Data Analysis

Statistical Inference → Sense making (from the population)

Longitudinal Database → Analysis (capturing social and economic dynamics)

Visualization → Conveying information (in statistically valid context)
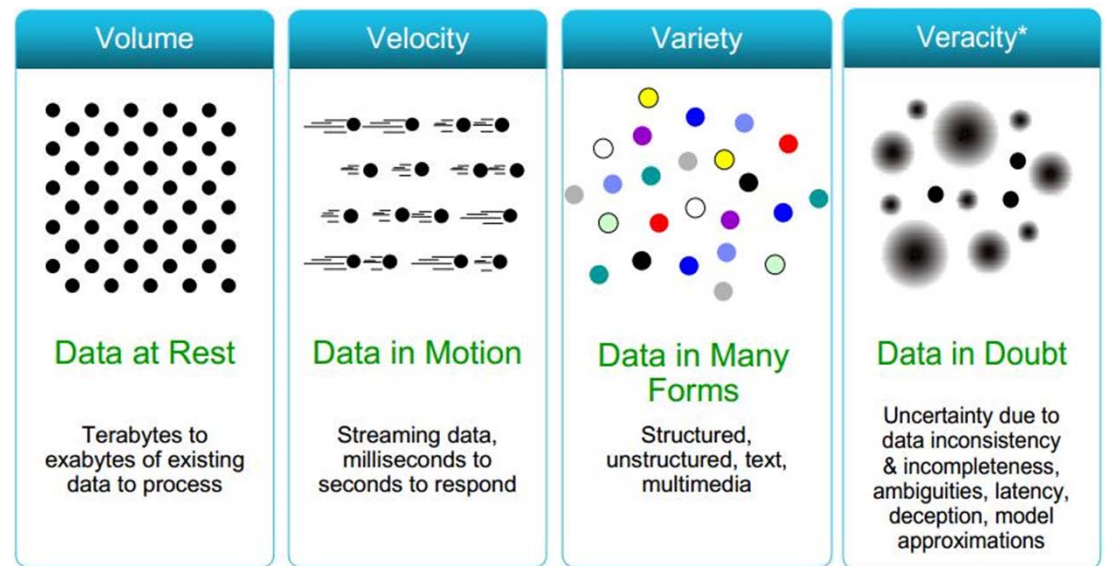
Data quality continues to be an issue

New/different types of data generated as by-product (e.g., smartphones, social media, satellites)
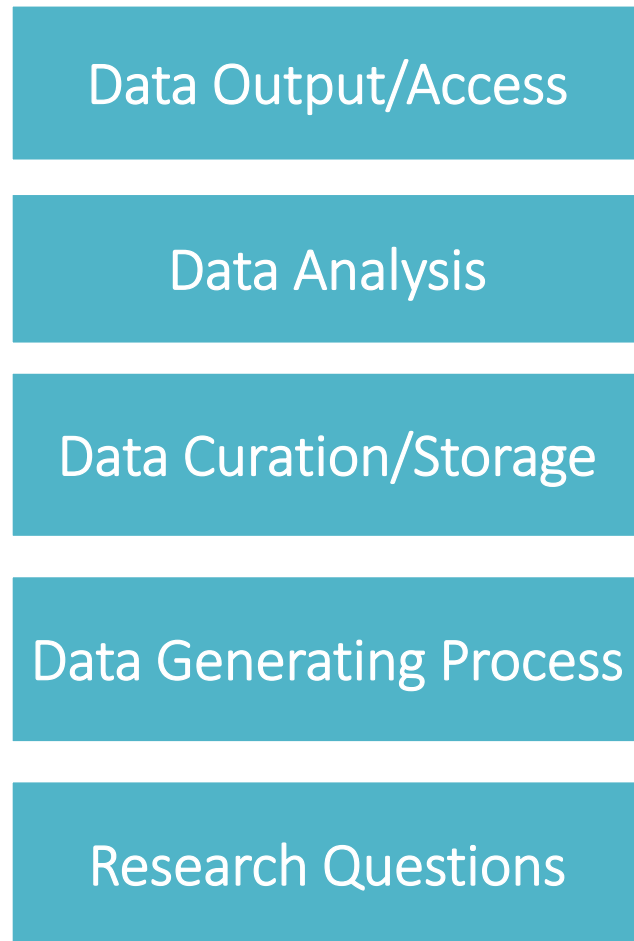
Fundamental changes in collection, availability, integration and dissemination of data

Paradigm shift for those who in the past relied primarily on survey research

Lack of people with skills to collect data, build modern statistical products and handle data veracity

Volume	Velocity	Variety	Veracity*
Data at Rest	**Data in Motion**	**Data in Many Forms**	**Data in Doubt**
Terabytes to exabytes of existing data to process	Streaming data, milliseconds to seconds to respond	Structured, unstructured, text, multimedia	Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

http://www.rosebt.com/blog/data-veracity

Data Output/Access	Visualization, disclosure control, ethics, privacy
Data Analysis	Statistical methods, advanced analytics including Bayesian, hierarchical, small area estimation
Data Curation/Storage	Practical training in data base management, editing, coding, imputation, etc.
Data Generating Process	Data collection from surveys, social media etc. data integration, record linkage, statistical matching
Research Questions	Market research, public policy, economics, journalism, public health, sociology, etc.

... teamwork needed

http://jointprogram.umd.edu/
http://survey-data-science.net/

fkreuter@umd.edu