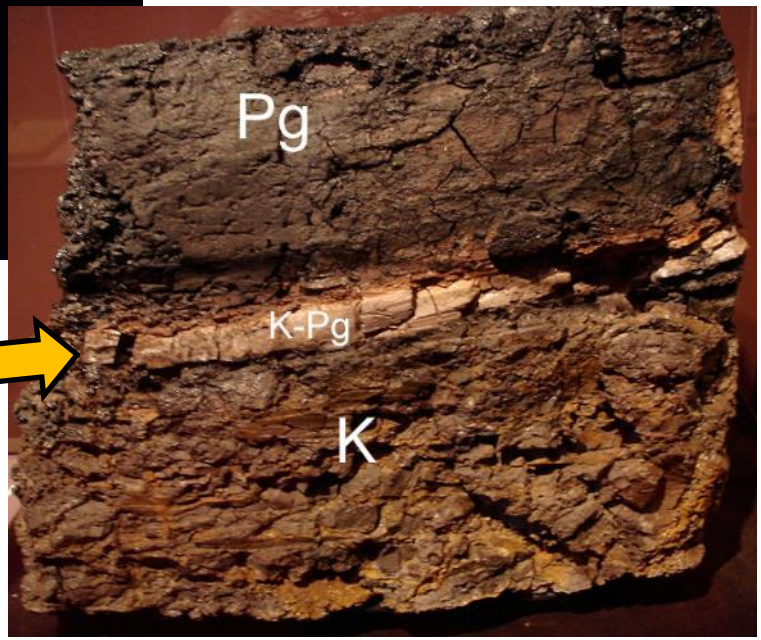
The Relevance of "Data Science" for Survey Research: Finding Solutions in an Ever-Changing Data World

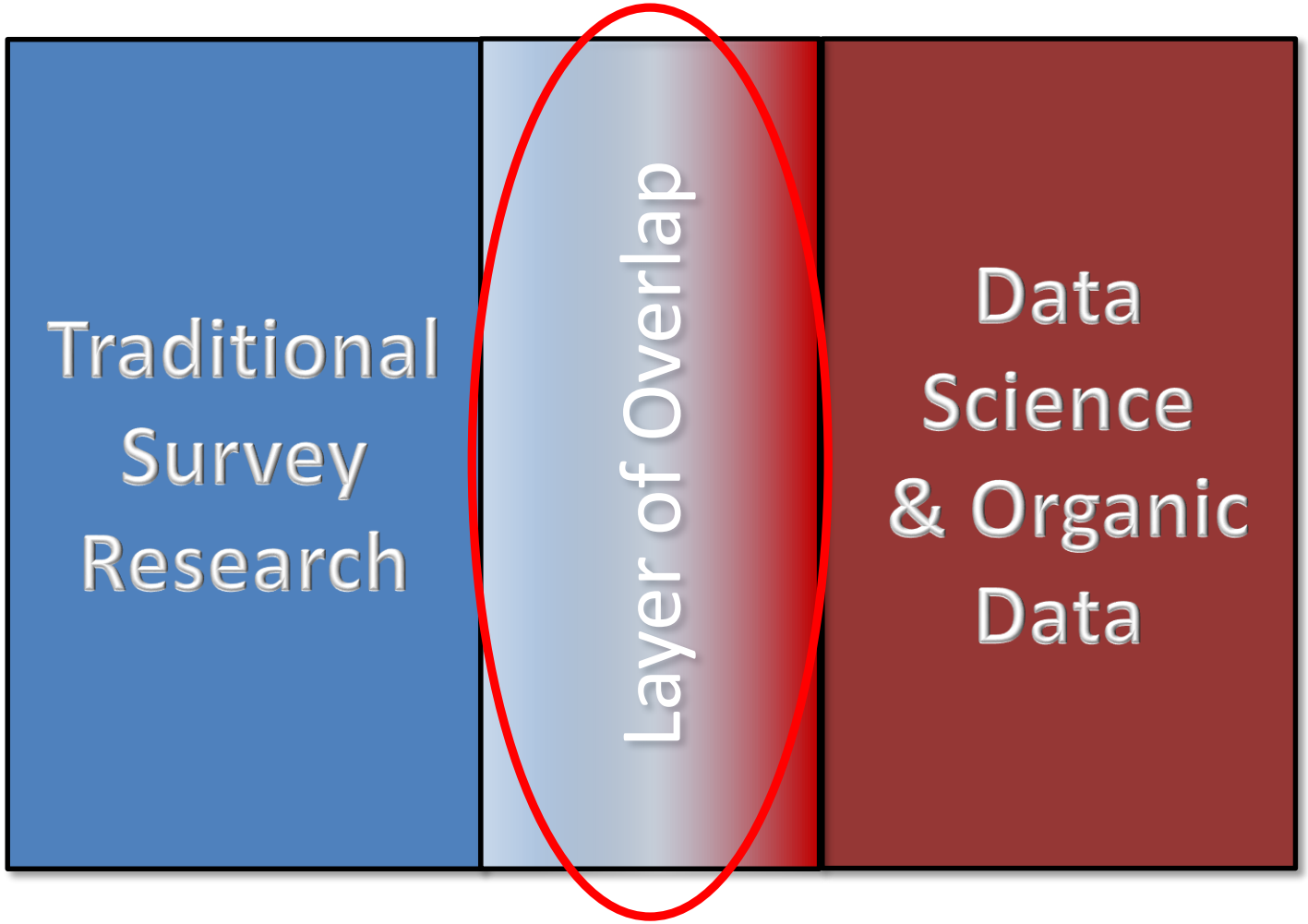
Michael Link, Ph.D.
Division VP,
Data Science, Surveys, & Enabling Technologies
Abt Associates

2019 International Conference and Workshop on Survey Research Methodology
Taipei, Taiwan
8/8/19

66 Million Years Ago …



What Do We Find at the Survey - Data Science Boundary?



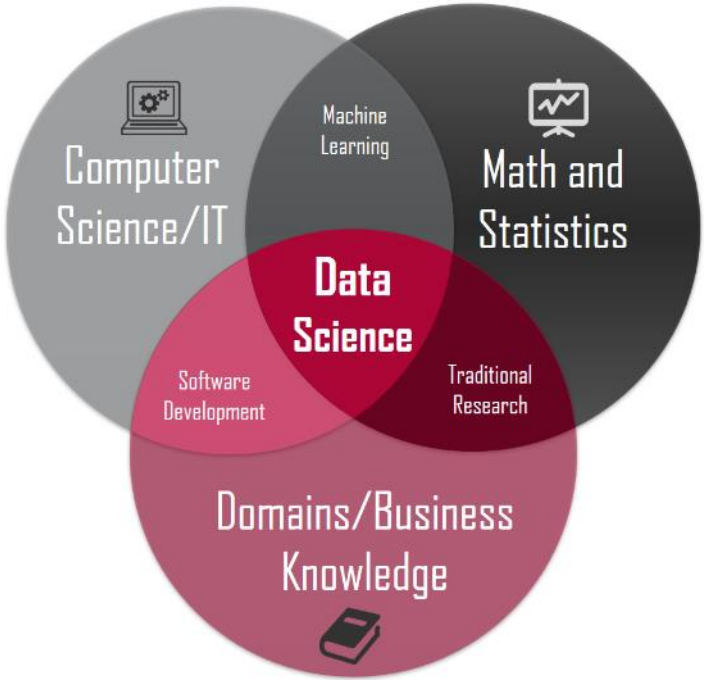
Questions of Interest for Today …

- What is “Data Science”?
- How are data science techniques and new forms of data influencing survey research methodology?
- What are some of the cautions in this new era?

I. WHAT IS DATA SCIENCE?

What is “Data Science”?

“Data Science: a multi-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data.”*



* Wikipedia (Why not? Best definition I’ve found!)

Data Science brings changes in …

- Data we can leverage (“organic data”)
- Tools we use (e.g., Machine Learning, Natural Language Processing, Image Recognition)
- Opportunities & cautions for the field of survey research methods

New Sources of Data

- Administrative data
- Transaction records (banking, purchase, etc.)
- Medical Records
- Social Media
- Bluetooth enabled devices
- Mobile devices
- Location info: GPS
- Geo-info: Satellites, planes, drones
- Visual: pictures & video
- Wearable devices
- Sensors / Internet of Things (IOT)

Design vs Organic Data

Design Data

- Traditional data (e.g. surveys)
- From a census or survey
- Collected from specific populations
- For specific purposes
- Often collected by those who will use them
- Respondents asked to answer questions
- Researchers control the data

Organic Data

- Arise out of the information ecosystem
- Often massive
- Close to “real time” measures
- Not designed for research purpose – input or output of “machine”/platform
- Collection unobtrusive to those being measured
- Researchers do not control data

Adopted from Robert Groves (2011). “Census Directors Blog: Designed Data and Organic Data”. Accessed at: https://www.census.gov/newsroom/blogs/director/2011/05/designed-data-and-organic-data.html

Types of Organic Data

	Structured Data - Administrative Records	Other Structured Data	Semi-Structured Data	Unstructured Data
Definition	Data with a fixed format easily exportable to a data set for analysis with minimal scrubbing required	Highly organized data easily placed in a data set but require additional scrubbing or transformation before analysis	Data that may have some structure but not complete and cannot be placed in a relational database; requires substantial cleaning	Data which have no standard analytic structure and must have data extracted and transformed before use
Examples	• Govt programs • Commercial transactions • Credit card / bank records • Medical records • University / school records	• E-commerce transactions • Mobile phone GPS • Roadside / Weather / pollution sensors	• Computer logs • Text messages • Email • Fitbit / wearable data • Internet of Things	• Social media data • Pictures / videos • Traffic webcams • Drone data • Satellite / radar images

Adopted from: National Academies of Sciences, Engineering, & Medicine. (2017). *Innovations in Federal Statistics: Combining Data Sources While Protecting Privacy*. Washington, DC: The National Academies Press.

New Data Tools

- Machine Learning
- Natural Language Processing
- Image Recognition

New Tools Facilitate:

- Extraction
- Formatting
- Cleaning
- Parsing
- Analysis

… of complex structured & unstructured data sources.

Tool 1: Machine Learning

- “Machine Learning” – Form of artificial intelligence (AI) that allows a computer application to become more accurate in predicting outcomes without being explicitly programmed, often relying on patterns and inference instead.

- Primary outcomes: predicting classification/groups or clusters/values

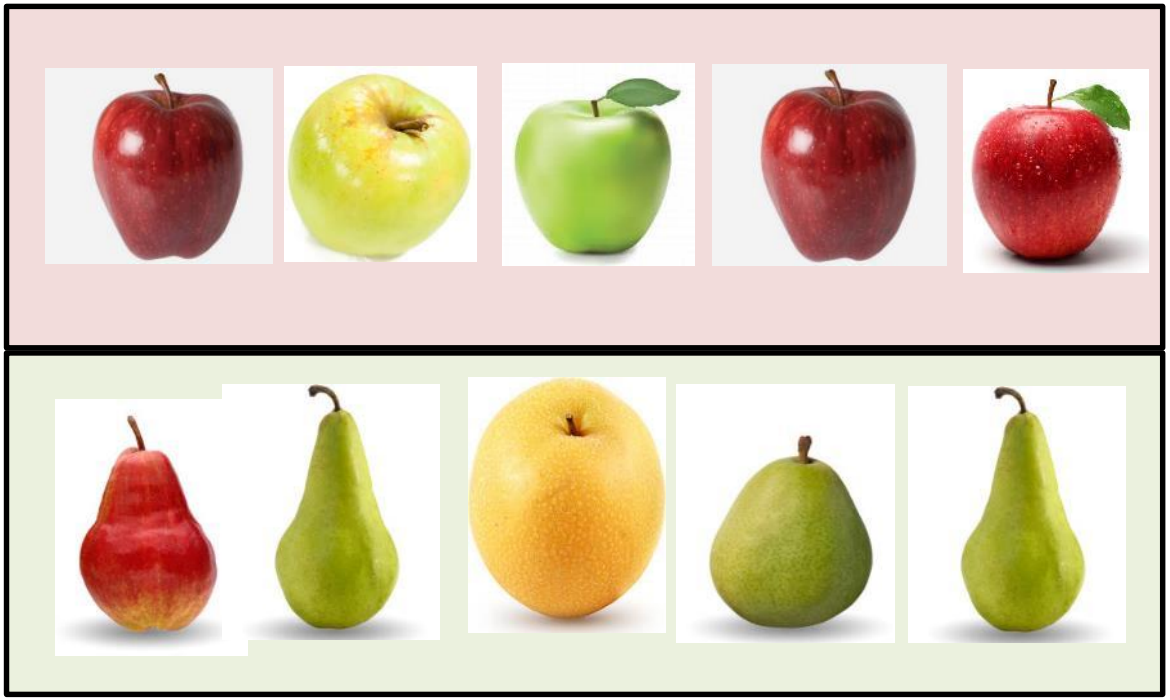
- Use cases:
  - Recommender engines (like Amazon, Netflix)
  - “Fake News” bot detection on social media sites
  - Exploring large volumes of visual data for exo-planet discovery

- In the survey world:
  - Potential interviewer falsification alerts
  - Improving area sampling via satellite imagery (especially in developing world)
  - Exploring new sources of data (social media, medical records, published reports, etc.) for new insights into attitudes and behaviors

Apples Or Pears?

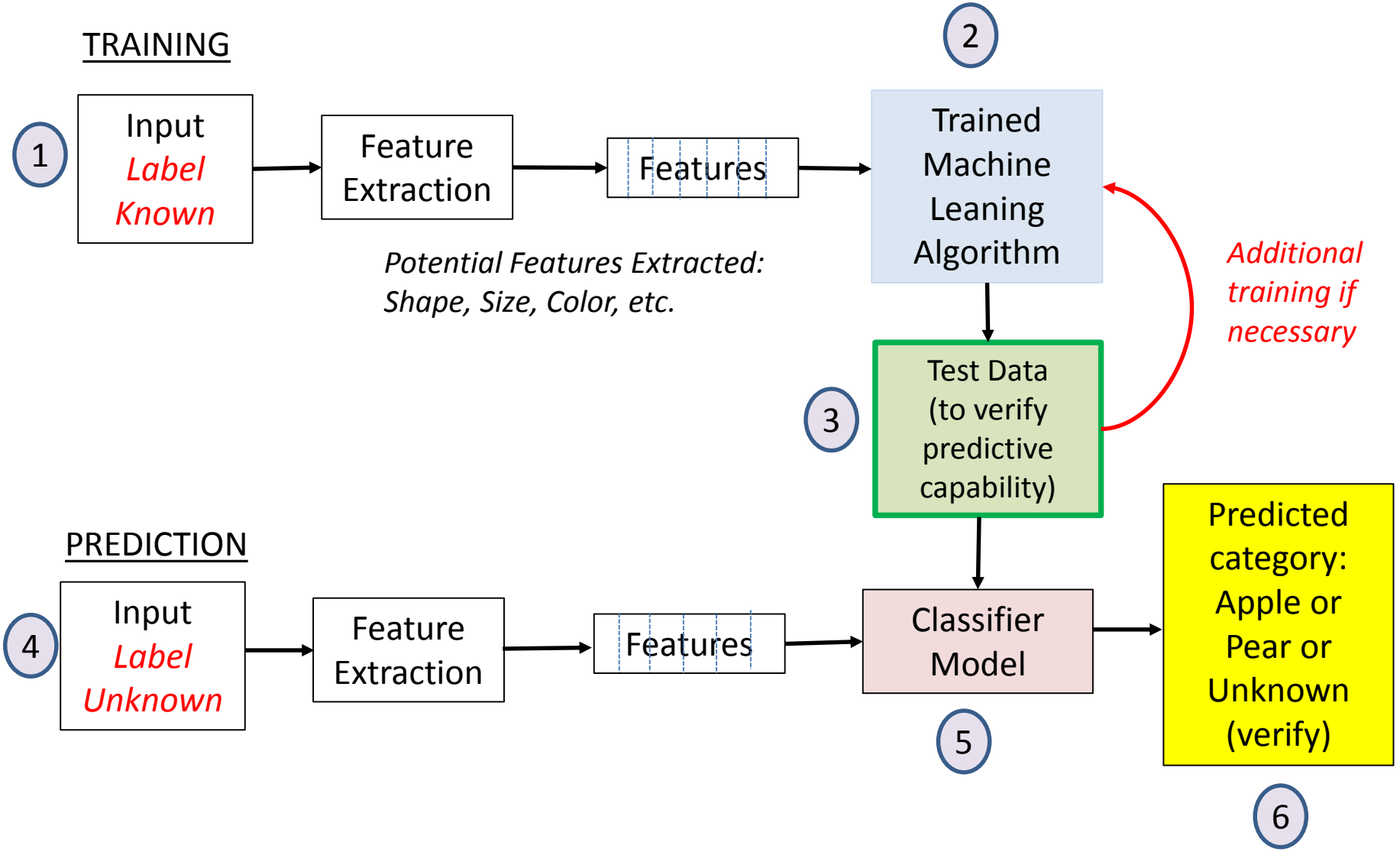
How can we classify these unknown images as apples or pears??

Develop “Training Set” with known values to “training” the ML algorithm

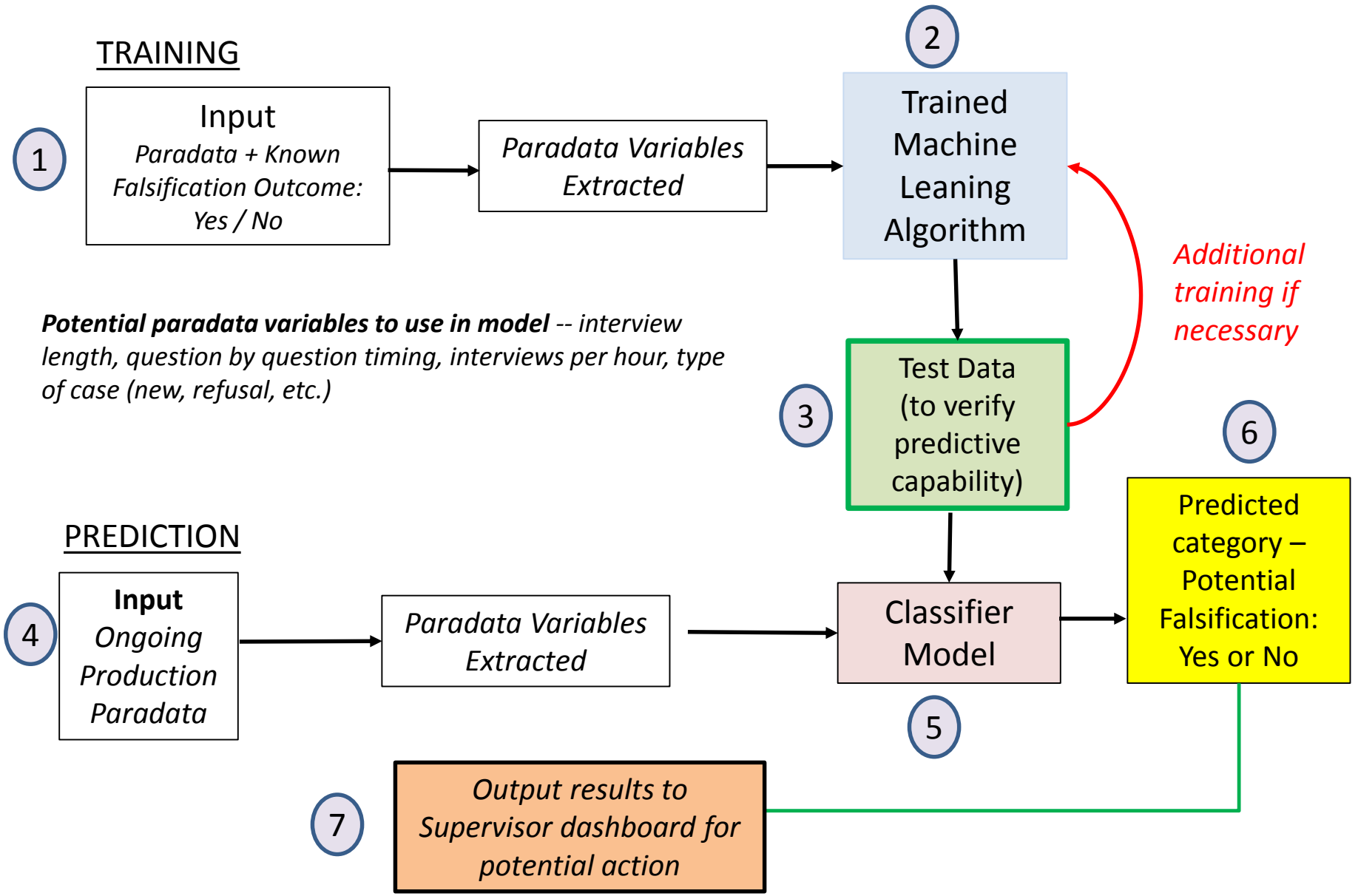


Features we can use to identify and categorize and apple or pear: shape, size, color, etc.

Machine Learning Workflow



ML Applied to Survey Example: Detecting Potential Telephone Interviewer Falsification



Tool 2: Natural Language Processing

- "Natural Language Processing" – Class of AI techniques for extracting, processing, and analyzing various forms of human communication, particularly speech and text
  - "Text Analytics": Set of methods for extracting and analyzing information from written sources

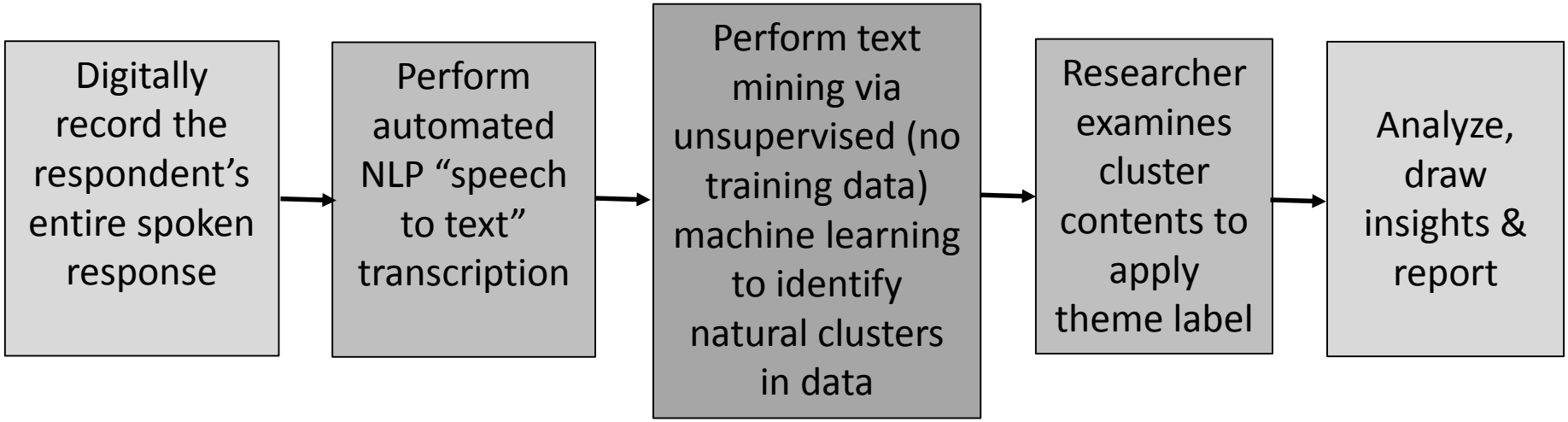
- Primary outcomes: Conversion of speech or text from recordings, documents, social media, websites, PDFs, letters to gov't, etc. into analyzable formats

- Use cases:
  - Google search
  - Amazon's Alexa, Apple's Siri
  - Analyzing arrays of medical records to understand disease origins and spread

- In the survey world:
  - Automatic coding of open-ended survey responses
  - Exploring new sources of data (social media, medical records, published reports, etc.) for new insights into attitudes and behaviors
  - Development of interactive, human-computer training modules for interviewers

NLP Applied to Survey Example: "Theme Modeling" Open-Ended Responses

"In your own words, how would you describe the most important problem facing the country?"



Deeper insights into respondent's thinking leveraging both digital recording technology, large cloud data storage, and data science analytics

Tool 3: Image Recognition

- Set of AI techniques for extracting data & insights from visual media:
  - Image Recognition (IR) – Class of AI techniques for extracting information from still, video or streaming images
  - Object Recognition – IR methods form making sense of patterns in image pixels to identify “objects” of interest (ex., households, cars, etc.)

- Primary outcomes: object of interest present/not present; count of objects; location within image of objects (called “object detection”)

- Use cases:
  - Facial recognition (ex., log into new iPhone)
  - Autotagging of people or products in vast number of images
  - Key feature of self-driving cars

- In the survey world:
  - Leveraging aerial images (satellite/drone) to develop unique population sample frames
  - Replace survey self-reports with image capture of purchases (consumer studies) / portion sizes (health studies)
  - Add greater contextual data to surveys of community safety or environmental exposures,

Applying Object Recognition & Machine Learning

Goal: Counting cars automatically using video

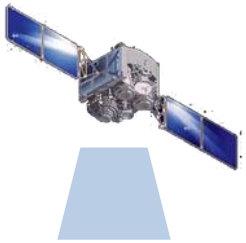
EXAMPLE TRAINING DATASET

Images with Known Vehicles

Images without Vehicles

EXAMPLE OUTCOMES

Survey Example – Developing Unique Frames from Aerial Images



Goal: Survey households with children who live within 1 KM of a playground

Object Recognition: extract key features identifying playgrounds

Trained Machine Learning Algorithm: categorize images

Validation by Researcher: Confirm all positive, & subset of negatives

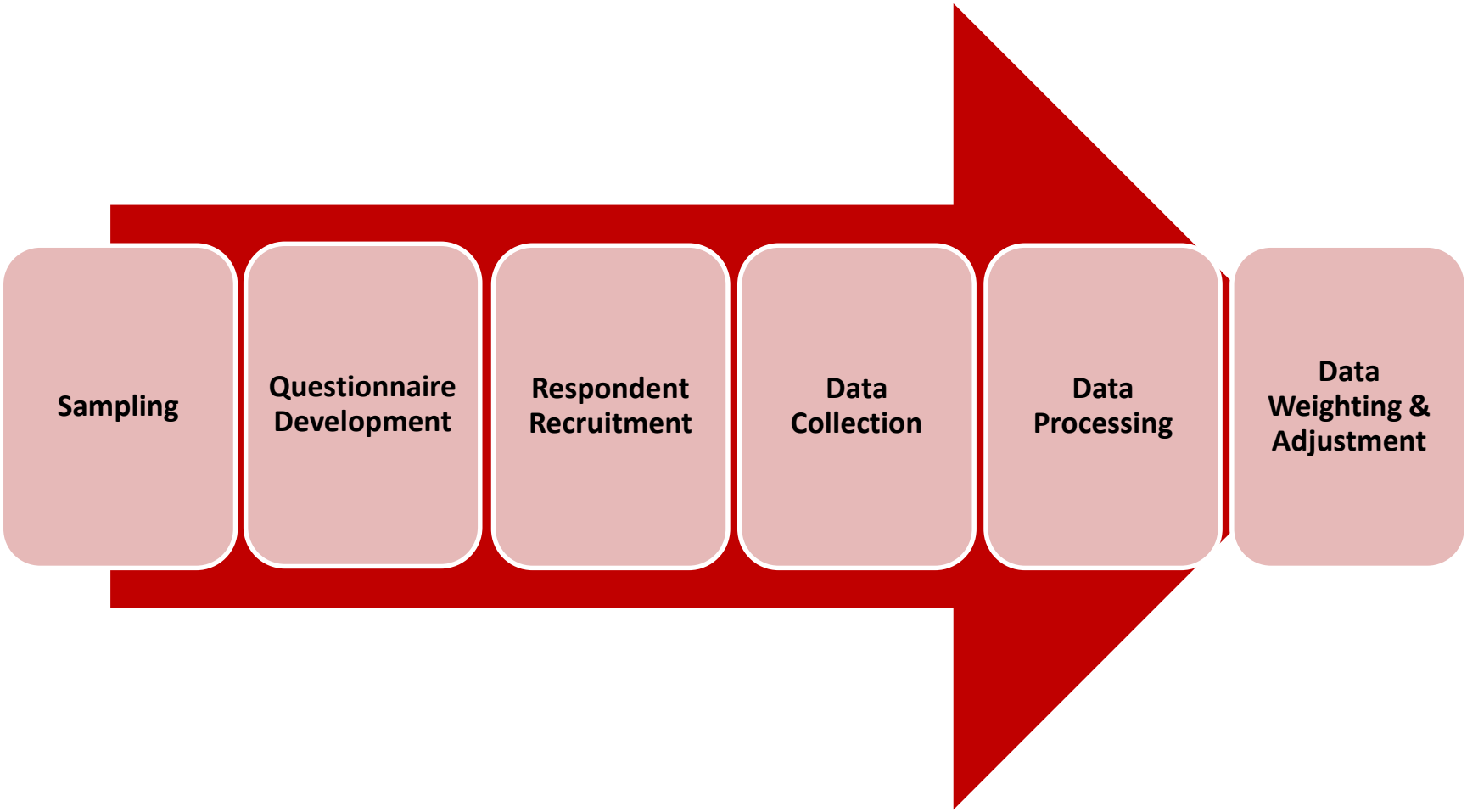
Geocode positives

Use registry of addresses to select addresses within 1 KM of geocoded playground

Refine sample as needed and conduct survey!

II. HOW ARE DATA SCIENCE AND NEW FORMS OF DATA INFLUENCING SURVEY RESEARCH?

Basic Steps in Survey Process



Survey Sampling

- **Enhance more traditional frames:**
  - Supervised ML leveraging external data to improve quality of standard frame (such as address-based sample frames)
  - Unsupervised machine learning to partition known populations groups into optimal groupings for stratification or other designs
  - Supervised ML to create / improve sample frames

- **Build New Frames Leveraging Aerial Data:**
  - Enhance counting & listing efforts via drones (Krotki, 2018)
  - Satellite images to develop frames in developing countries (Chew et al., 2018)
  - Build unique sample frames from satellite and other aerial data on windmills, playgrounds (Eck et al, 2018)

Questionnaire Development

- Automated question-check systems based on ML algorithms
  - Survey Quality Predictor http://sqp.upf.edu
    - Tests reliability & validity of questions
  - Questionnaire Understanding Aide (QUAID)
    - Assesses wording, syntax, semantics of questions

- ML algorithms to identify real-time respondent behaviors and help reduce them (ex. Item-level response, straight-lining, speeding for self-administered surveys

Respondent Recruitment

Help drive Responsive / Adaptive Designs

- Response propensity
- Response propensity by mode
- Characteristic propensity
- Panel attrition
- Allocation of incentives by identifying those most likely to respond with and without incentive
- Mining field interviewer notes & observations

Additional Insights:

- ✓ Approaches more successful the richer the paradata or external linked data;
- ✓ Best suited for ongoing panels or longitudinal studies w multiple waves
- ✓ Problematic when “training data” don’t mirror unknown population data

Data Collection

- Automating Interviewer Training
  - AvaTalk-TI: Speech recognition, NLP & ML to allow interviewers to practice refusal avoidance with “automated respondent”
- Potential interviewer falsification
  - ML for outlier patterns in timings
  - NLP of recordings to identify “turns of conversation”
- Monitoring field interviewers
  - GPS data to develop metrics – routes to homes, timing, efficiencies, also potential falsification

Data Processing

- NLP to auto-code open-ended text responses or recorded responses
  - Key word counts (from a priori list)
  - Sentiment analysis (pos/neg)
  - “Topic modelling” with unsupervised ML – find natural patterns & clusters
- Coding complex concepts such as Industry & Occupation codes
- Imputation models:
  - ML models can often reduce time & costs
- Linking external data to survey data
  - Validation of measures
  - Alternative measures (e.g., Fitbit for physical activity – steps, sleep, etc.)
  - Extend variables for analysis (e.g., Statistics Netherlands webscraping data for online prices, housing data, job information – to inform survey data)

Data Weighting & Adjustment

- Non-response Weighting:
  - Unsupervised ML to create weight classes based on broad set of potential predictors (not just demographics)

III. CAUTIONS IN THIS NEW ERA

Recognize and Address the Challenges

- Lot of hype, little published
  - Conferences currently best source
- Beware the “file cabinet” phenomena – most focus is on what works, not the hundreds of approaches that did not work
- Key areas of concern:
  - Data issues
  - How we apply data science techniques
  - Transparency of methods

Data Issues

- Do we understand the origins of the data we use? The people & concepts it does (or does not) represent?
- Access to data – some easier/some harder; access over time?
- Changes in the "platform" can cause changes in the data (e.g., the "Perpetual, Dynamic Algorithm Dilemma")
- "Fake data" / Bots

How We Apply Data Science Techniques

- Despite the focus on “learning,” algorithms are programmed by people – how we train and interpret algorithms matters

- Different sources of training data, different outcomes
  - ✓ Biases in training data will lead to biases in outcomes

- How we assess outputs - algorithms often evaluated on speed or predictive capacity (% assigned to a category rather than “unknown”) --- often not validated with external source
  - – Few agreed upon metrics to measure potential bias

- Human communication is complex – making sense of it in a valid, reliable manner is difficult

Transparency of Methods

- Need documentation of data used, variables, methods, outcomes, interpretations, potential sources of error (as we do for most surveys)
  - "Black box" approach is not acceptable

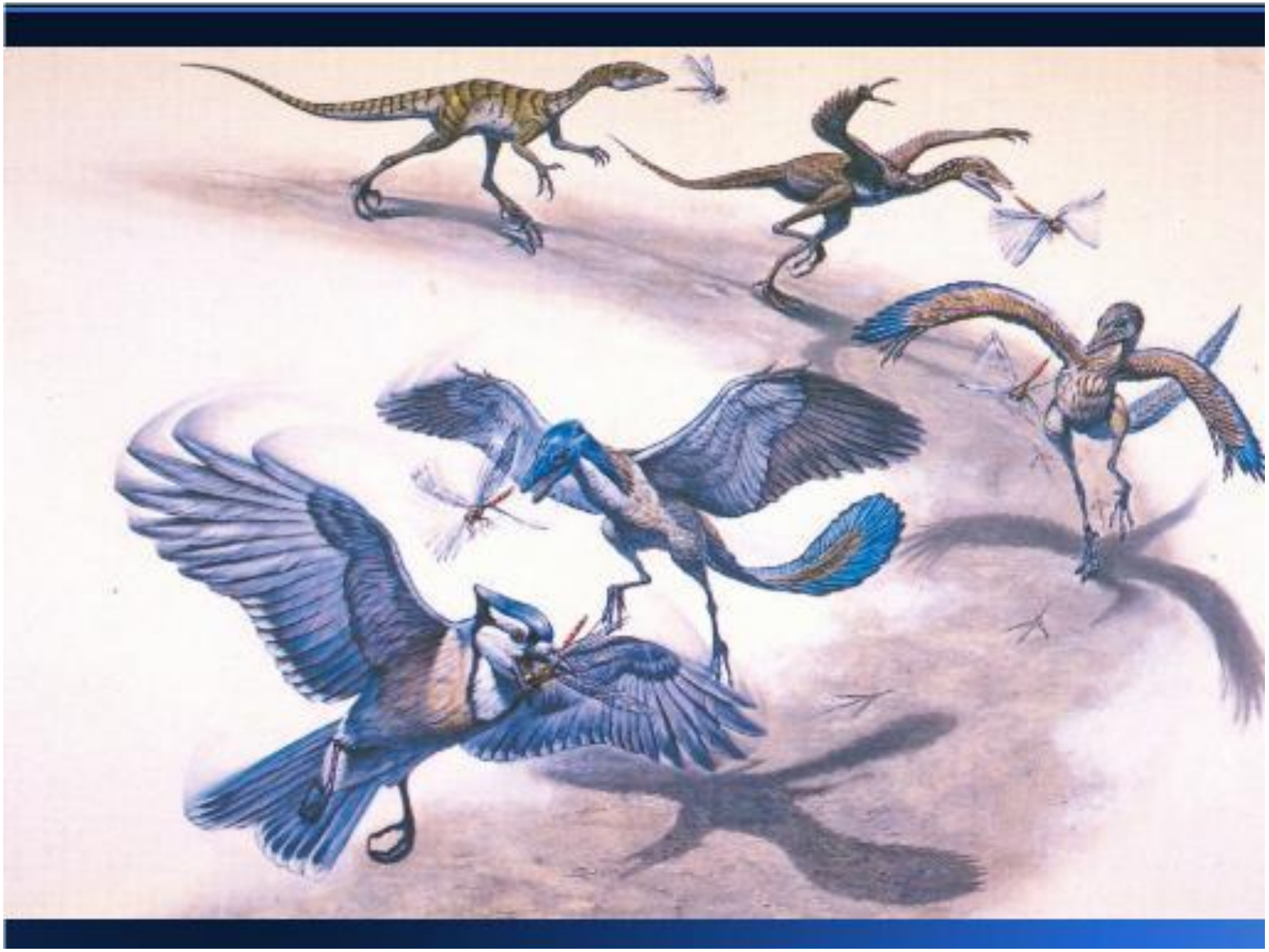
- Need to be able to replicate results and ensure approaches can be reproduced & scaled (not just "one-off")

New Era Requires More Robust Assessment Evaluation Framework

- Discipline needs to broaden our “Total Survey Error” framework to a “Total Error” framework
  - Sampling & non-sampling “traditional” concerns
  - Plus: measure of error in data sources, extraction, processing, filtering, linkage, etc.

- Recognize the many new sources of error & bias and develop standardized, agreed-upon ways to measure & document

Don’t Bury Surveys Yet: Evolution Can Happen!



Michael Link, Ph.D.

Division VP,
Data Science, Surveys, & Enabling Technologies
Abt Associates

Contact Info:
Michael_Link@Abtassoc.com
Twitter: @MLink01