Data Science and AI

Ross J. Schuchard
ross.schuchard@nps.edu
CS 4000: Harnessing AI
2 August 2021

Outline

- What is Data Science?
- Data Science Workflow
- Data Science in DoD
- Trust in Data – It’s all about the data
- A Cautionary Warning
- Questions

Data Science: Reality [or] Hype

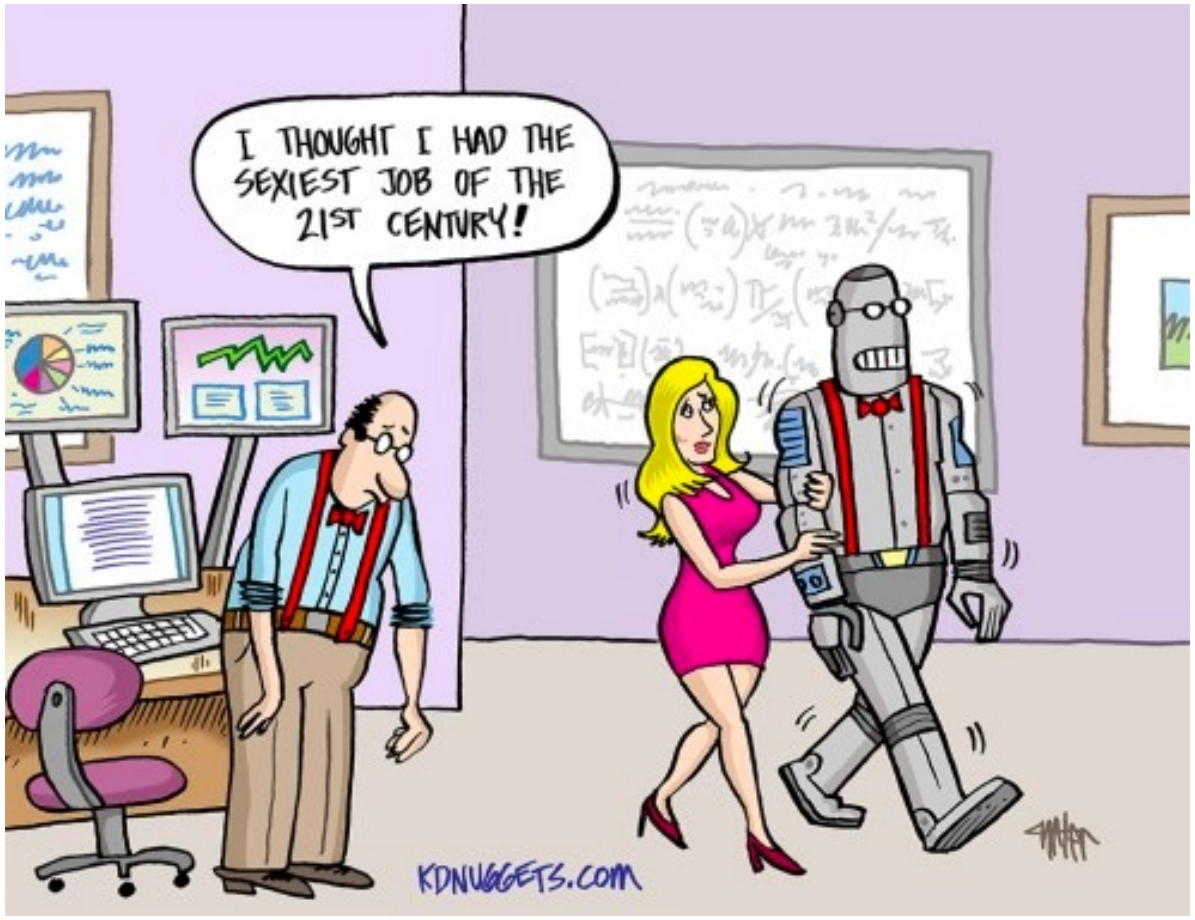
Data Scientist: The Sexiest Job of the 21st Century

Harvard Business Review

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

HBR (2012)



KD Nuggets (2018)

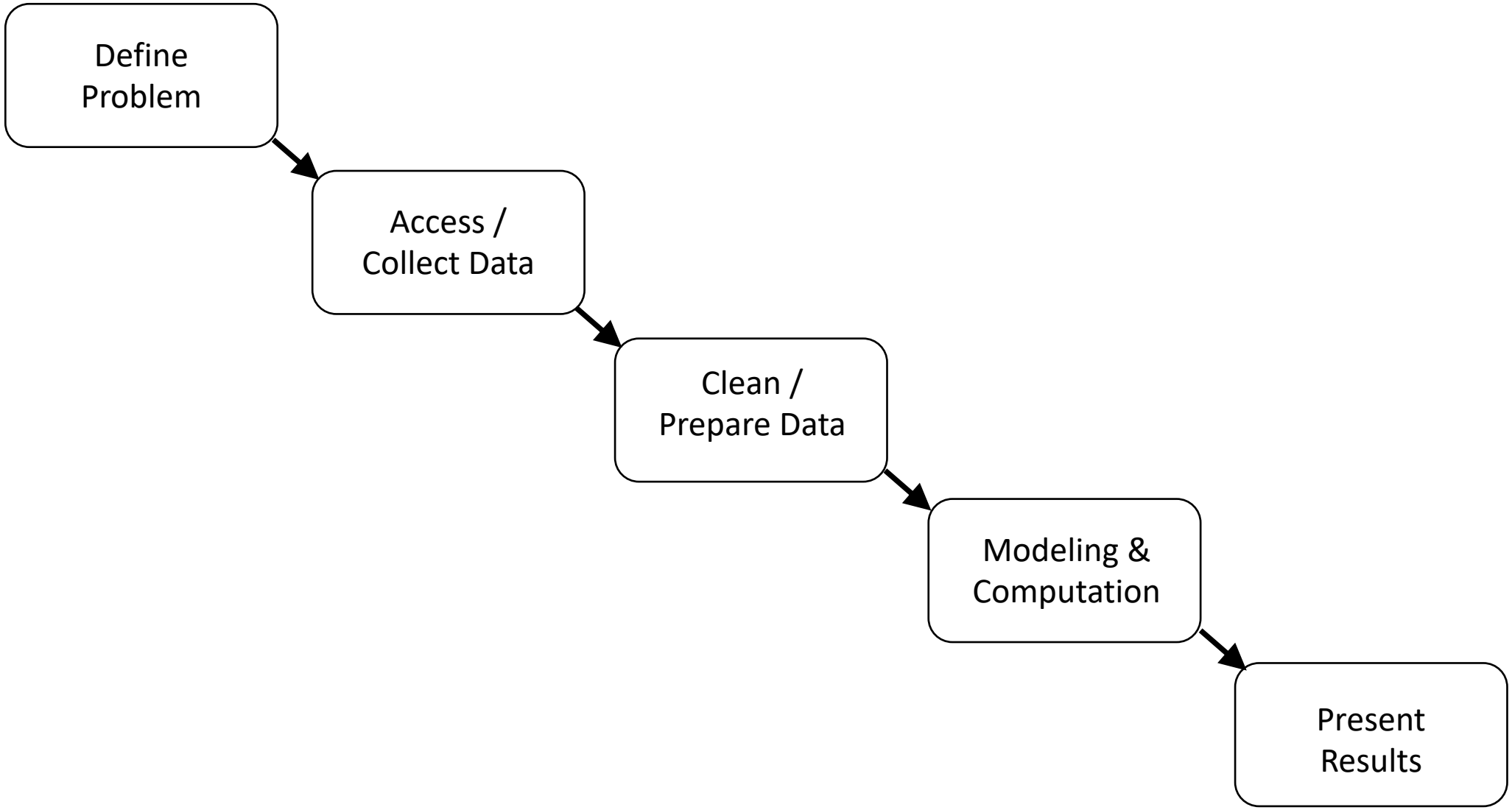
What is Data Science?

Data science studies the analysis of data, focuses on building models and validating them against data. The models can be used not only for prediction, but also for explaining a phenomenon and for performing simulations with the phenomenon.

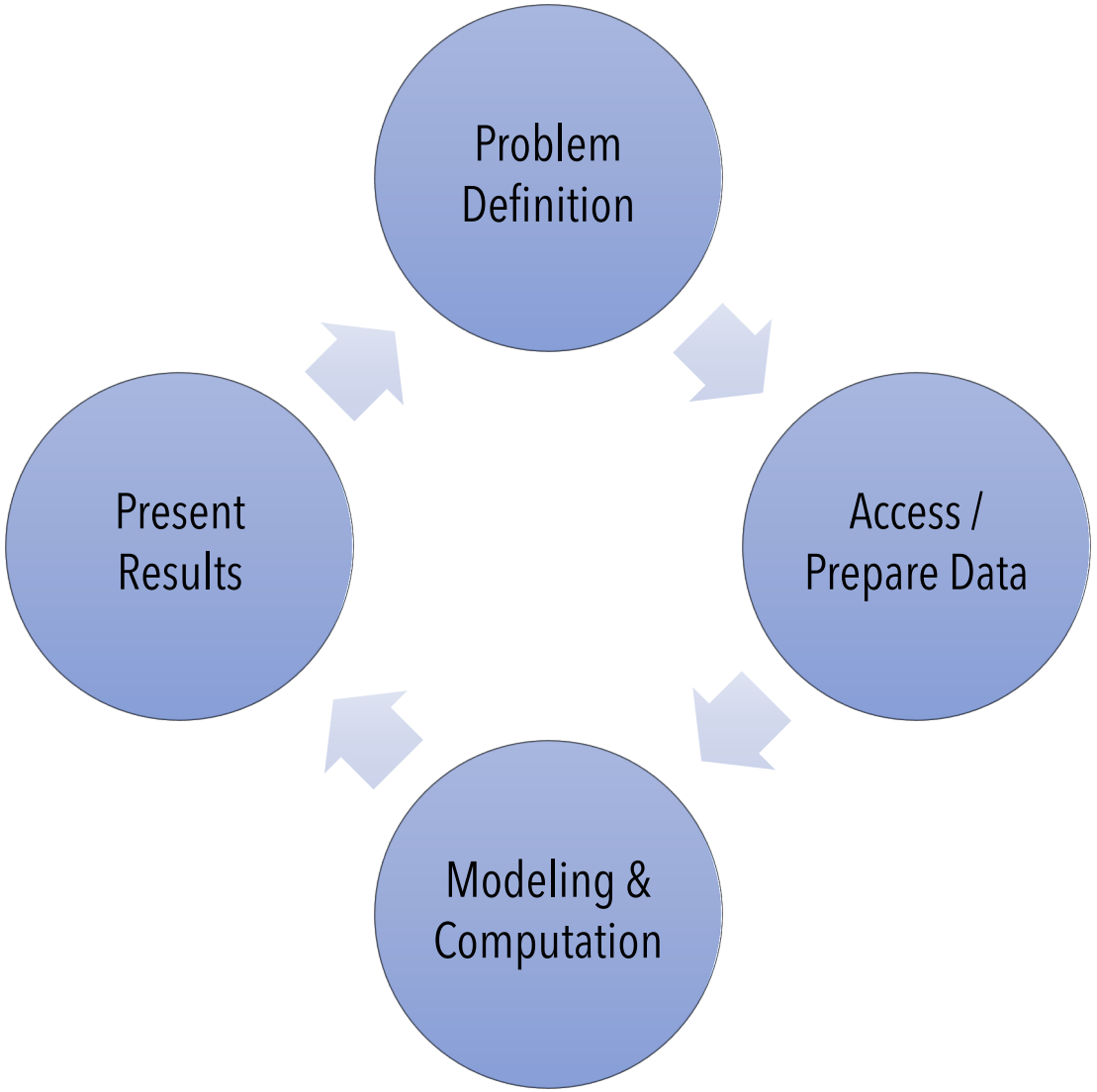
Positive accomplishments (so far):

- Image recognition
- Email spam filtering
- Recommender systems
- Predictive Maintenance
- Military Medicine: Rapid Analysis of Threat Exposure (RATE)

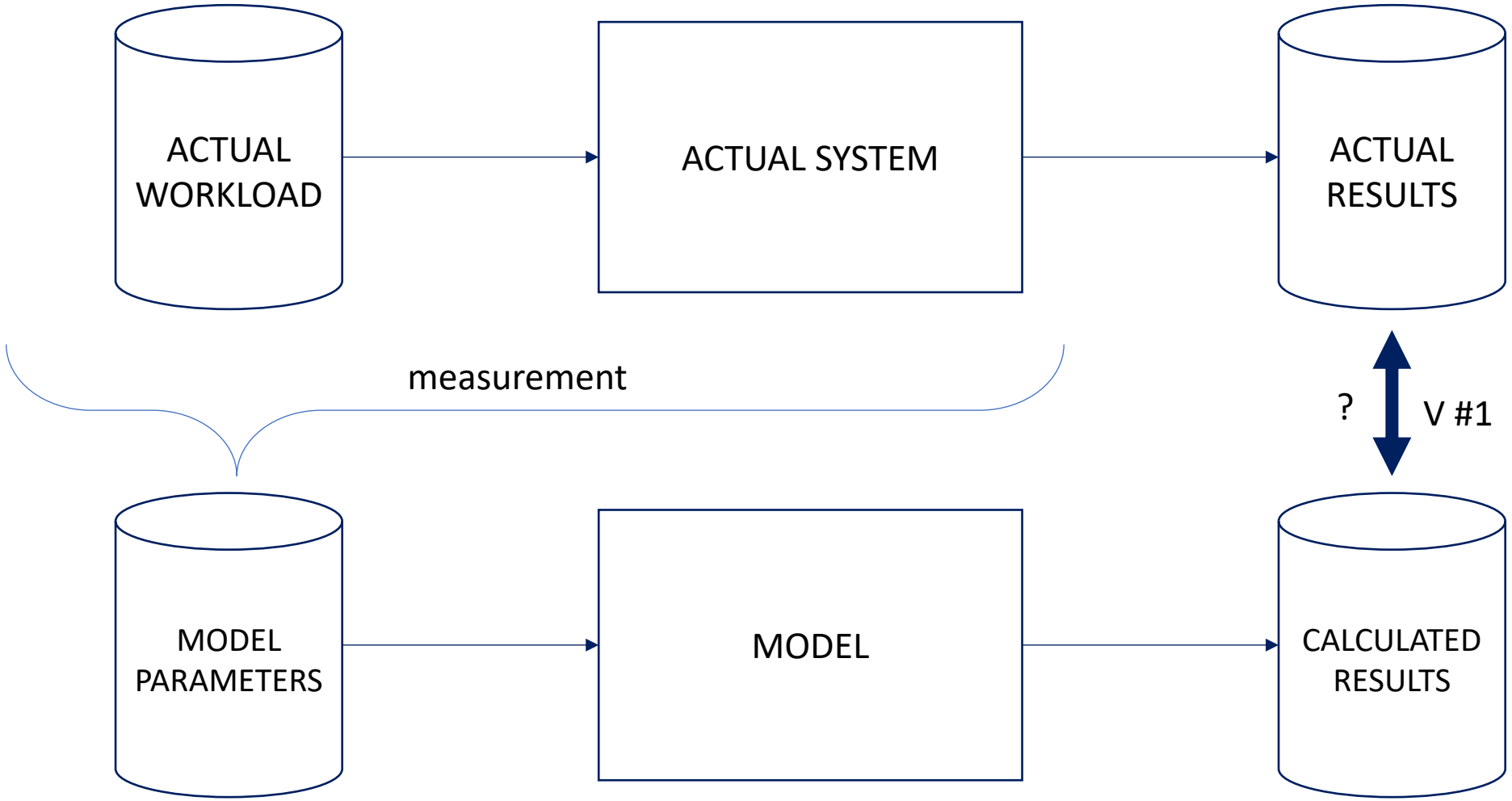
Data Science Workflow

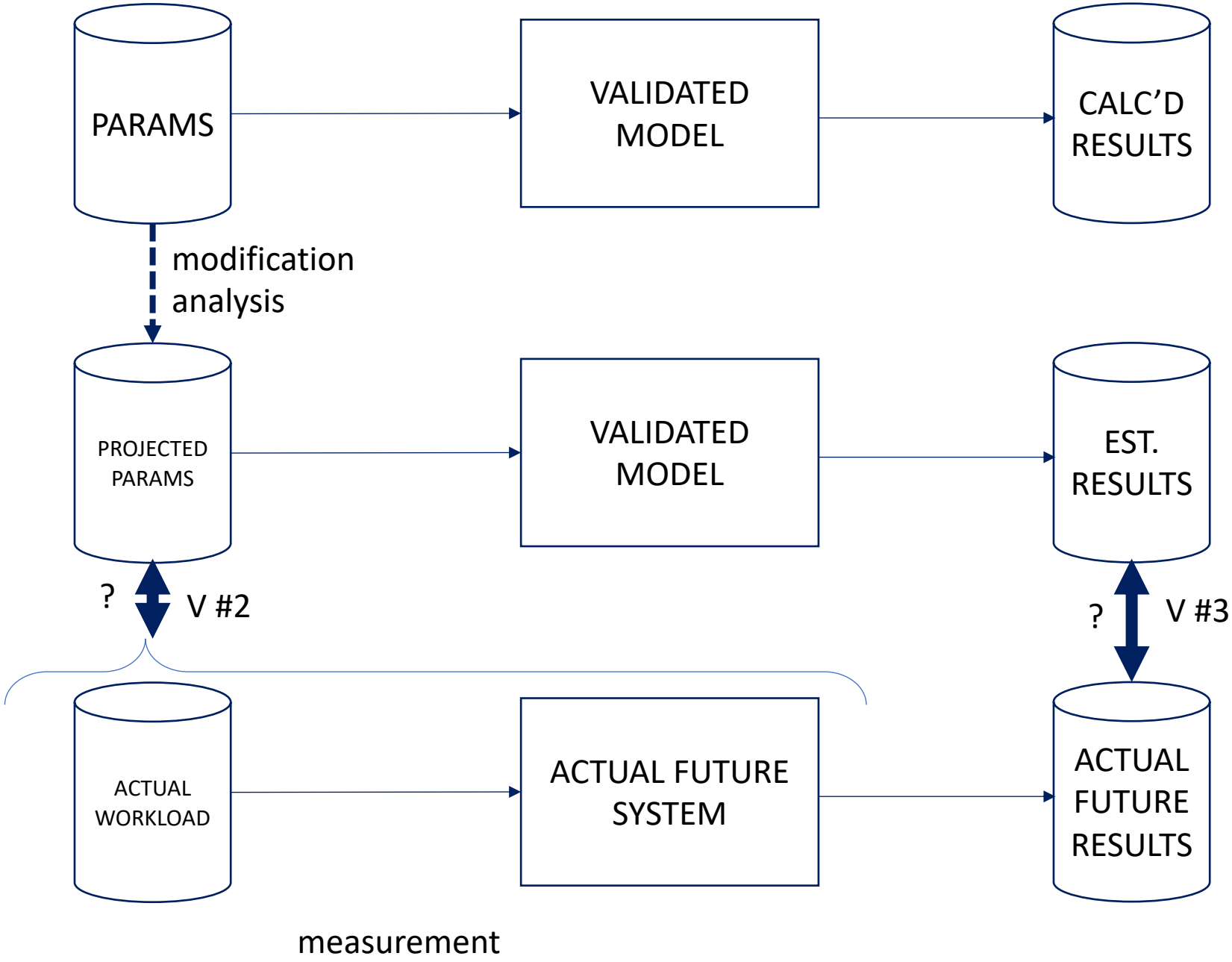


Data Science Workflow



Getting to a validated model …




PARAMS
VALIDATED MODEL
CALC'D RESULTS
modification analysis
PROJECTED PARAMS
VALIDATED MODEL
EST. RESULTS
?
V #2
?
V #3
ACTUAL WORKLOAD
ACTUAL FUTURE SYSTEM
ACTUAL FUTURE RESULTS
measurement

DoD Efforts – Applying Data Science

Joint Artificial Intelligence Center
(est. 2018)

Army AI Integration Center
(est. 2019)

SOCOM Data Engineering Lab
(est. 2019)

Data Science and Analytics Group

NPS DSAG
(est. 2018)

*NOTE: Not a comprehensive survey, but focused on ties to NPS

DoD Efforts – Build a Data Science Workforce

- Graduate School Course/Certification Offerings
  - NPS (resident & non-resident offerings)
  - AFIT
  - Focused efforts to send officers to new civilian graduate institutions
- Short Course Instruction
  - NPS Data Science and Analytics Group (DSAG)
  - Army FA49 Continuing Education Offerings
- Formal Data Science Skill Identification
  - Army approved personnel development skill identifier (PDSI) `R1J` (2019)

Jaywalking Billionaire?

SOURCE: Medium

Establishing Trust

- Selection Bias
- Labeling Bias
- Explainability

Selection Bias

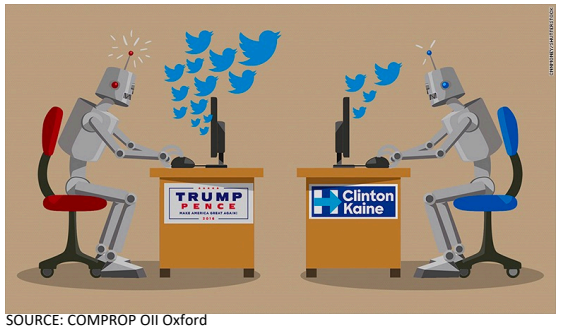
- Basic issue of not having the right data OR not knowing you have the wrong data.
- Is the data representative of the specified problem?
  - Network analysis of host-based intrusion data
  - Specific problem scope success vs. general adaptation failure
- Macro versus micro data selection
  - Simpson’s Paradox

Labeling Bias

- Human labeling
  - Non-native English speakers manually judging sentiment of text
  - Non-medical professionals assessing coughing audio and polyp images
- Machine Labeling
  - Automation considerably more efficient
  - Significant risk of data type misinterpretation
- Hybrid approaches: ‘human-in-the-loop’ OR ‘human-on-the-loop’
- Tradeoff analysis evaluating efficiency/correctness (when to accept risk)

Limitations of Automated Labeling: Social Bots

SOURCE: Krebs



SOURCE: COMPROP OII Oxford

Bot Detection

- Wide variety of **supervised** and **unsupervised** algorithmic approaches
- Focus extensively on detecting ever-increasing sophistication in bots
- Typically develop and train around 'ground truth' use-case

"The main takeaway from the DARPA challenge is that a bot-detection system needs to be semi-supervised. "

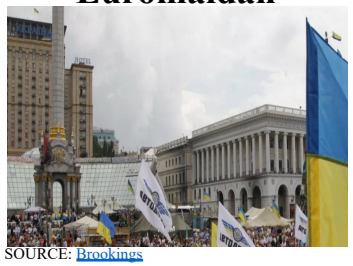
– Subrahmanian et al. (2016)

Bot Analysis

- Mostly qualitative analysis with increasing usage of stats/NLP
- Few works capitalize on detection efforts; many use ex post facto 'lists'
- Isolated views of single or small-scale use-cases (***e.g. Pew 2018***)

Why is this important?

Euromaidan

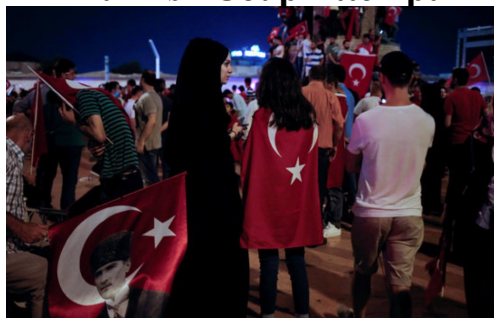


SOURCE: Brookings

Tahrir Square

SOURCE: BBC

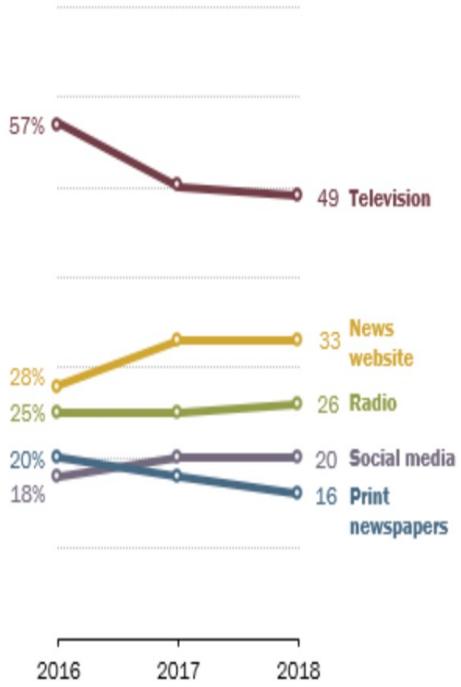
Turkish Coup Attempt



SOURCE: NYTimes

More Americans get news often from social media than print newspapers

% of U.S. adults who get news often on each platform



SOURCE: PEW RESEARCH

Masks, cash and apps: How Hong Kong's protesters find ways to outwit the surveillance state

SOURCE: WASHPOST

How the U.S. Is Fighting Russian Election Interference

SOURCE: NYT

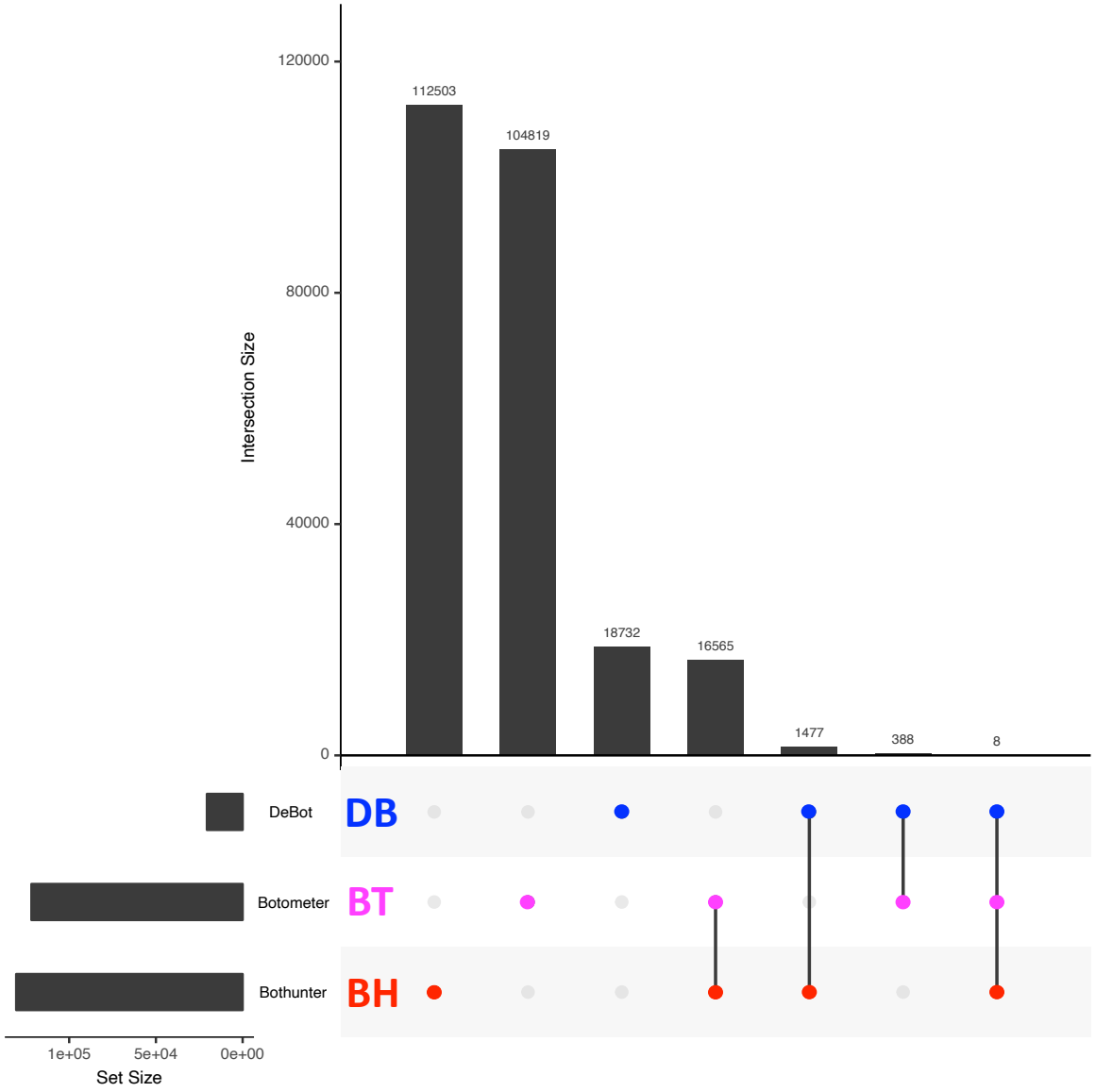
Brexit: Vote Leave broke electoral law, says Electoral Commission

17 July 2018 Share

Leave campaigns spending investigations

SOURCE: BBC

Classification Results



SOURCE: Schuchard & Crooks, Insights into Elections (2021)

DETECTION OVERLAP VOLUME

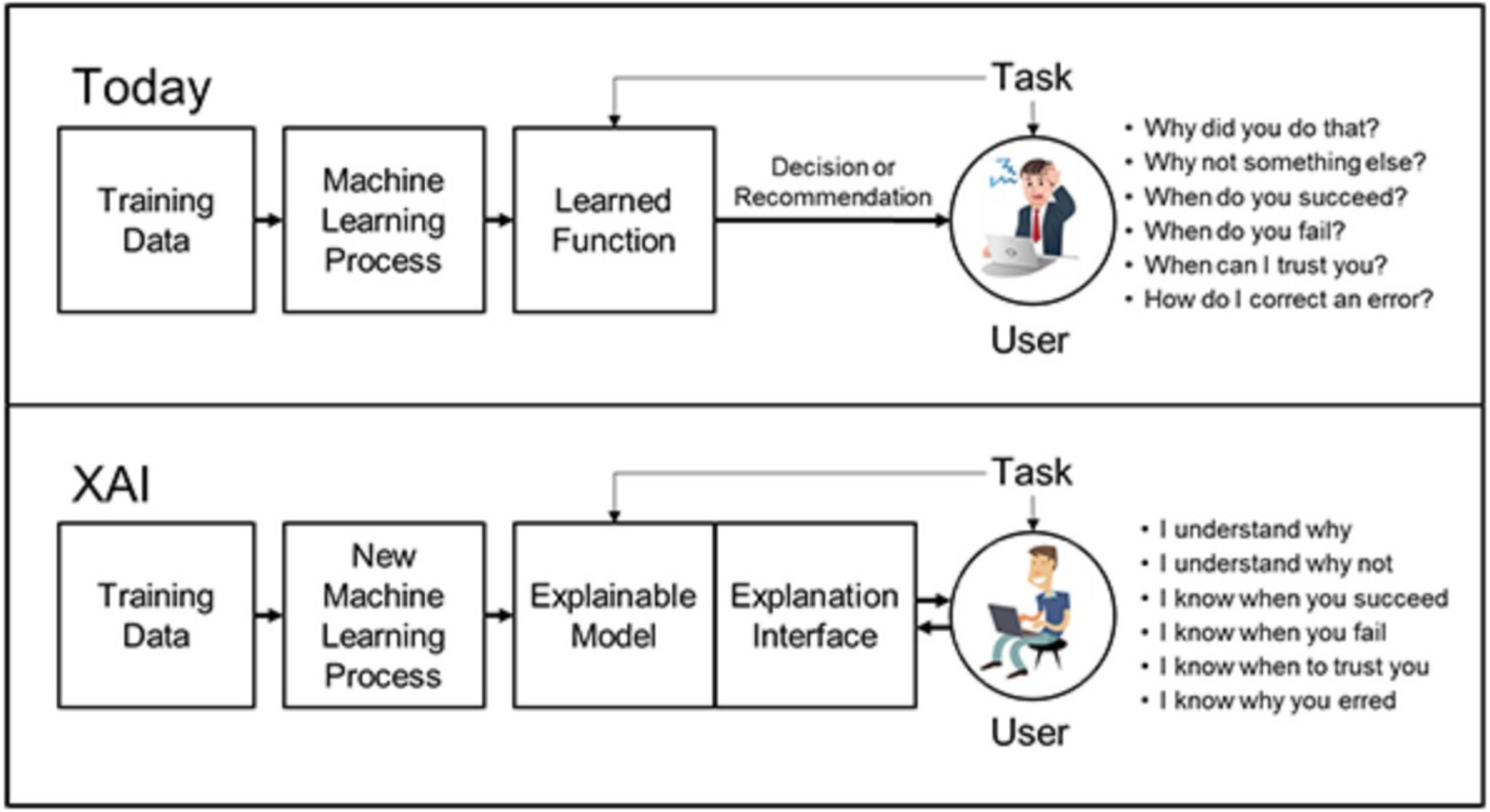
Botometer (BT) ∩ Bot-hunter (BH)
16,585 accounts

DeBot (DB) ∩ Bot-hunter (BH)
1,477 accounts

DeBot (DB) ∩ Botometer (BT)
388 accounts

DeBot (DB) ∩ Botometer (BT) ∩ Bot-hunter (BH)
8 accounts

Explainability



SOURCE: DARPA Explainable AI (XAI) Project

Explainability Example

- In 2008, Google researchers launched Google Flu Trends (GFT) to use search trends to predict the onset of the flu
- Research published in *Nature* (2009) stated GFT achieved a 97% accuracy rate when compared to ground-truth CDC data
- In 2013, GFT overestimated flu rates by over 140% during the peak of flu season.

A cautionary warning …

- DoD (and beyond) continues to invest heavily in Data Science to develop/expand expertise and enable better industry engagement.
- Most investments are in their earliest stages and rapidly evolving as we (DoD) adjust to ‘seeing ourselves’ through data for the first time.
- Models are developed based on past data. To use them for prediction we must forecast the future model parameter values.
- Forecasting future model parameters is a highly uncertain process because of the complexities, contingencies, and surprises of the real world.
- Thus, a well validated model may still be a poor predictor.

Questions

Ross J. Schuchard
ross.schuchard@nps.edu