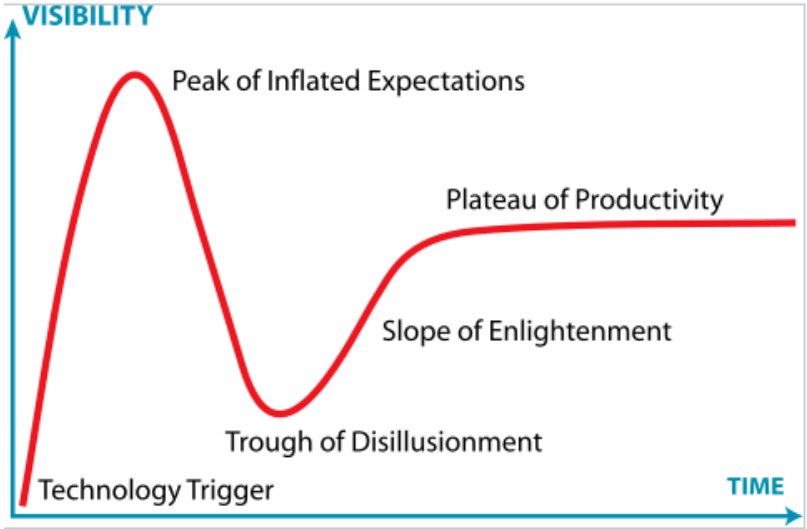
UMD DATA605: Big Data Systems

Lesson 1.3: Is Data Science Just Hype?

Instructor: Dr. GP Saggese, gsaggese@umd.edu

Is Data Science Just Hype?

- **Big data (or data science) is everywhere**
  - *"Any process where interesting information is inferred from data"*
- Data scientist called the "sexiest job" of the 21st century
  - The term has becoming very muddled at this point
- **Is it all hype?**

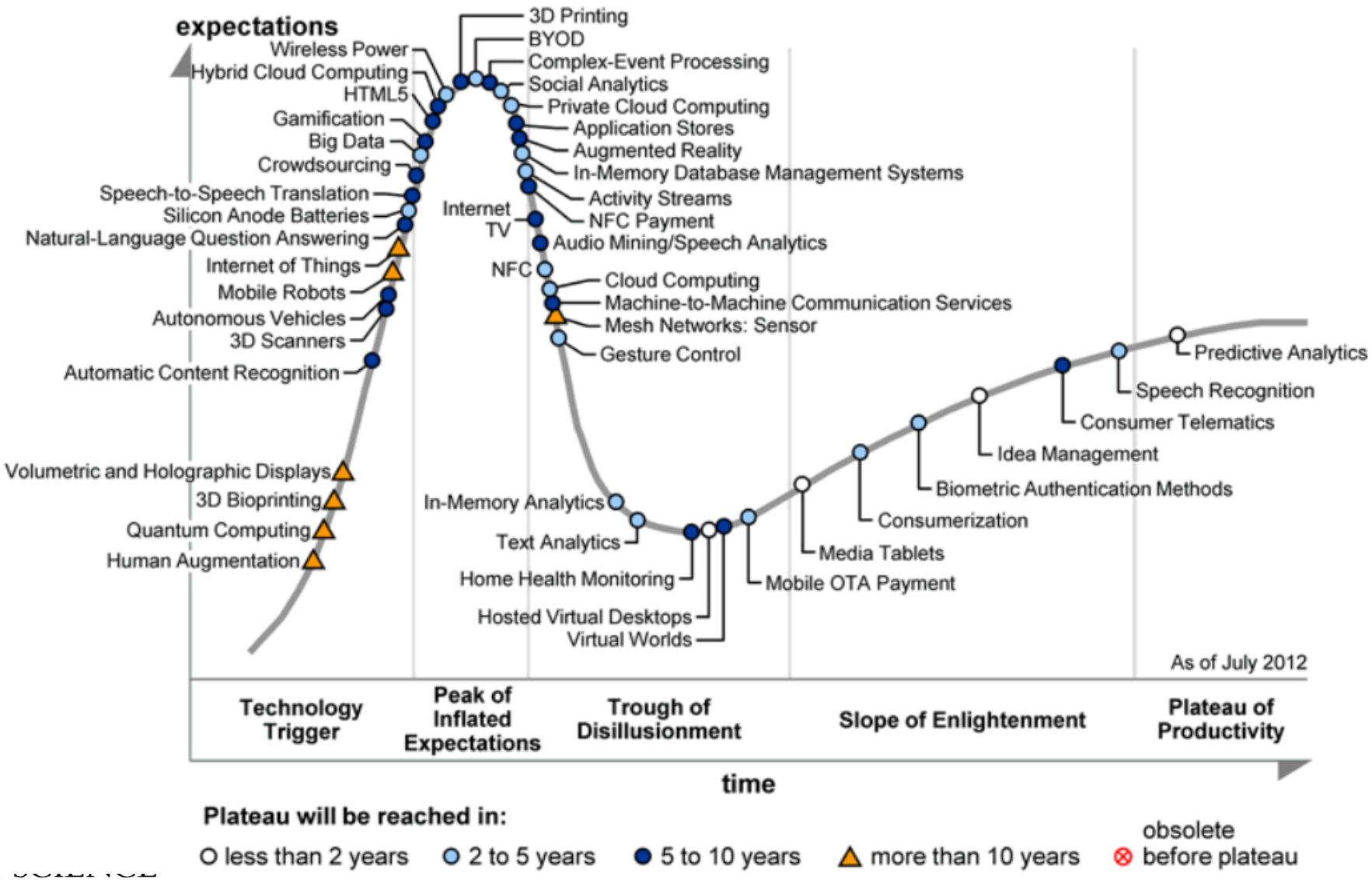


Is Data Science Just Hype?

- **No**
  - Extract insights and knowledge from data
  - Big data techniques revolutionize many domains
    - E.g., education, food supply, disease epidemics
- **But**
  - Similar to what statisticians have done for years
- **What is different?**
  - More data is digitally available
  - Easy-to-use programming frameworks (e.g., Hadoop) simplify analysis
  - Cloud computing (e.g., AWS) reduces costs
  - Large-scale data + simple algorithms often outperform small data + complex algorithms

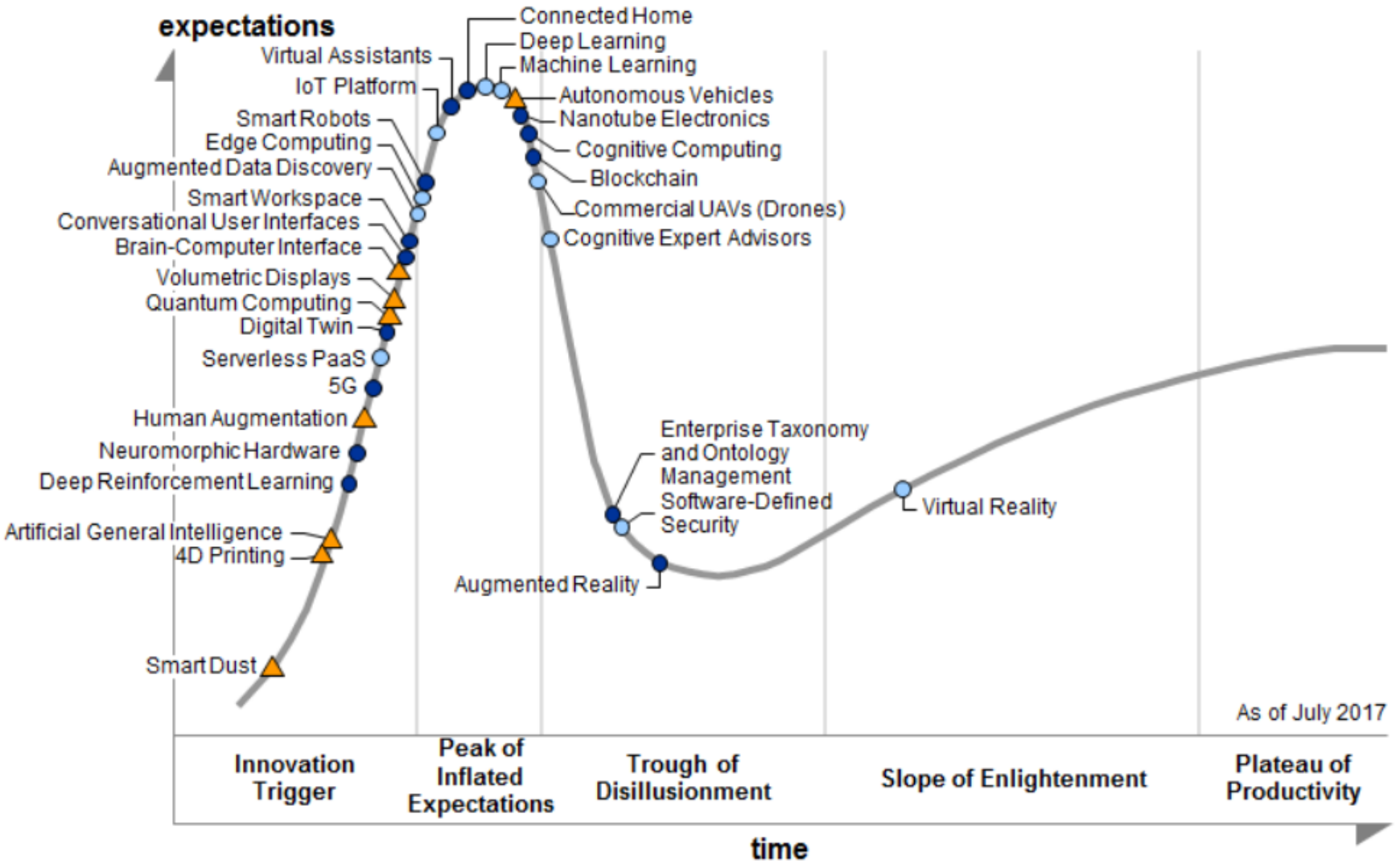
What Was Cool in 2012?

- Big data, Predictive analytics



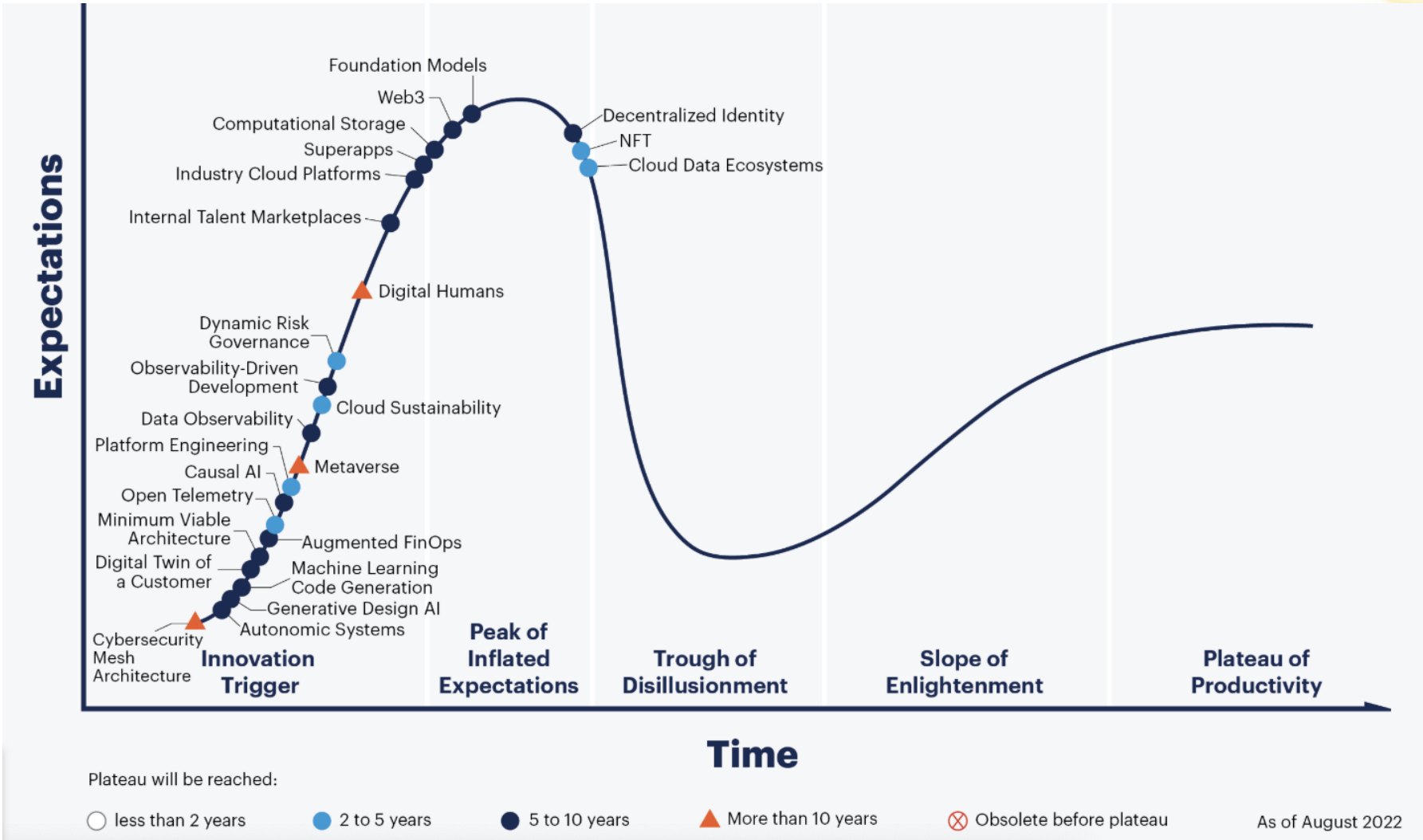
What Was Cool in 2017?

- Deep learning, Machine learning



What Was Cool in 2022?

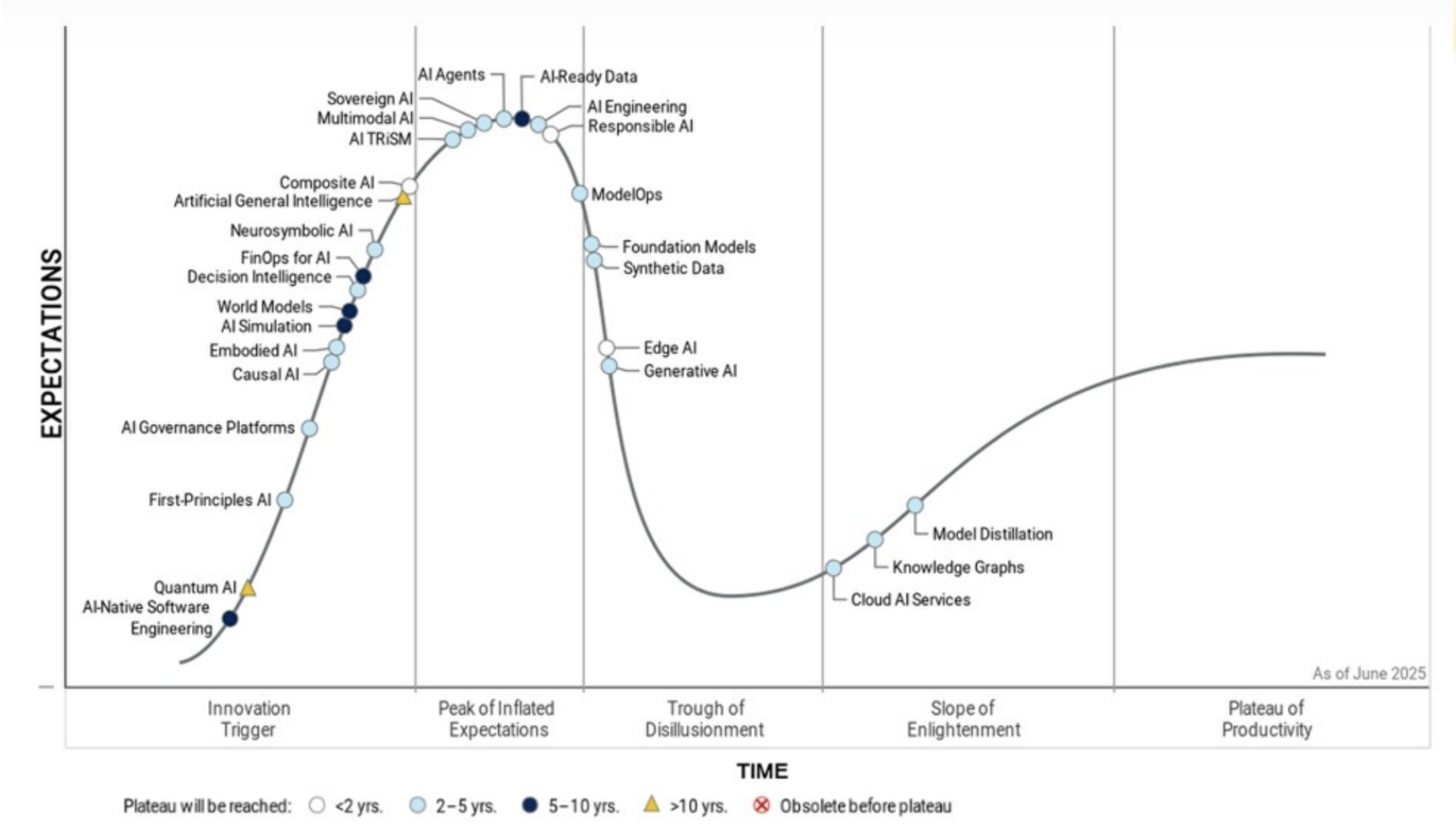
- Causal AI



SCIENCE ACADEMY

What Was Cool in 2025?

- Causal AI, Decision intelligence



SCIENCE ACADEMY

Key Shifts Before/After Big-Data

- **Datasets: small, curated, clean → large, uncurated, messy**
  - Before:
    - Statistics based on small, carefully collected random samples
    - Costly and careful planning for experiments
    - Hard to do fine-grained analysis
  - Today:
    - Easily collect huge data volumes
    - Feed into algorithms
    - Strong signal overcomes noise
- **Causation → Correlation**
  - Goal: determine cause and effect
  - Causation hard to determine → focus on correlation
    - Correlation is sometimes sufficient
    - E.g., diapers and beer bought together
- **"Data-fication"**
  - = converting abstract concepts into data
  - E.g., "sitting posture" data-fied by sensors in your seat
  - Preferences data-fied into likes

Examples: Election Prediction

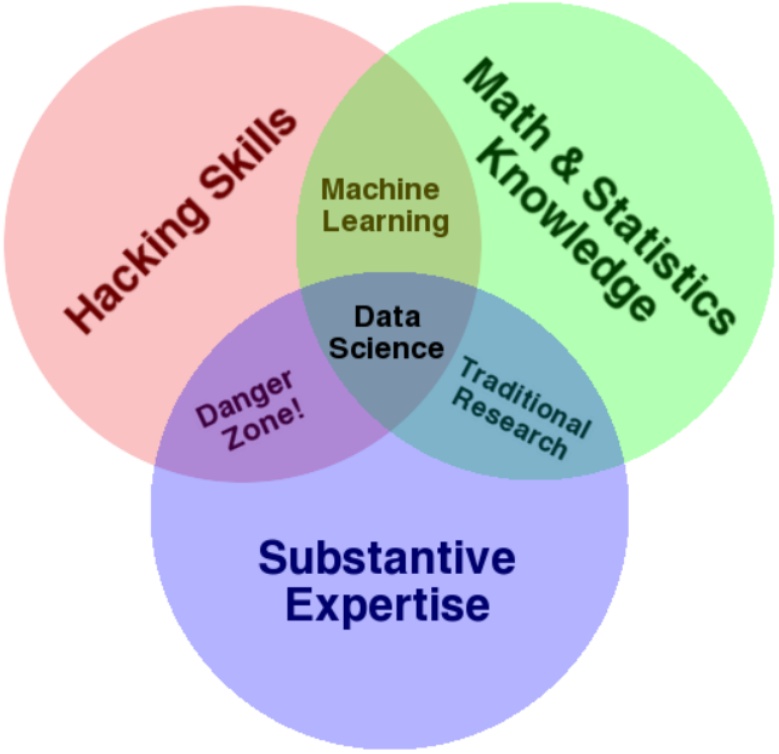
- Nate Silver and the 2012 Elections
  - Predicted 49/50 states in 2008 US elections
  - Predicted 50/50 states in 2012 US elections
- **Reasons for accuracy**
  - Multiple data sources
  - Historical accuracy incorporation
  - Statistical models
  - Understanding correlations
  - Monte-Carlo simulations for electoral probabilities
  - Focus on probabilities
  - Effective communication

Examples: Google Flu Trends

- 5% to 20% of the US population gets the flu annually; **40k deaths**
  - Early warnings help in prevention and control
- **Google Flu Trends**
  - Provided early flu outbreak alerts via search query analysis
    - Analyzed 45 search terms
    - Used IP to determine location
  - Predicted regional flu outbreaks 1-2 weeks before CDC
  - Operated from 2008 to 2015
- **Caveat: accuracy issues**
  - Claimed 97% accuracy
  - Lower out-of-sample accuracy (overshot CDC data by 30%)
  - People search about flu without confirmed diagnosis
    - E.g., searching "fever" and "cough"

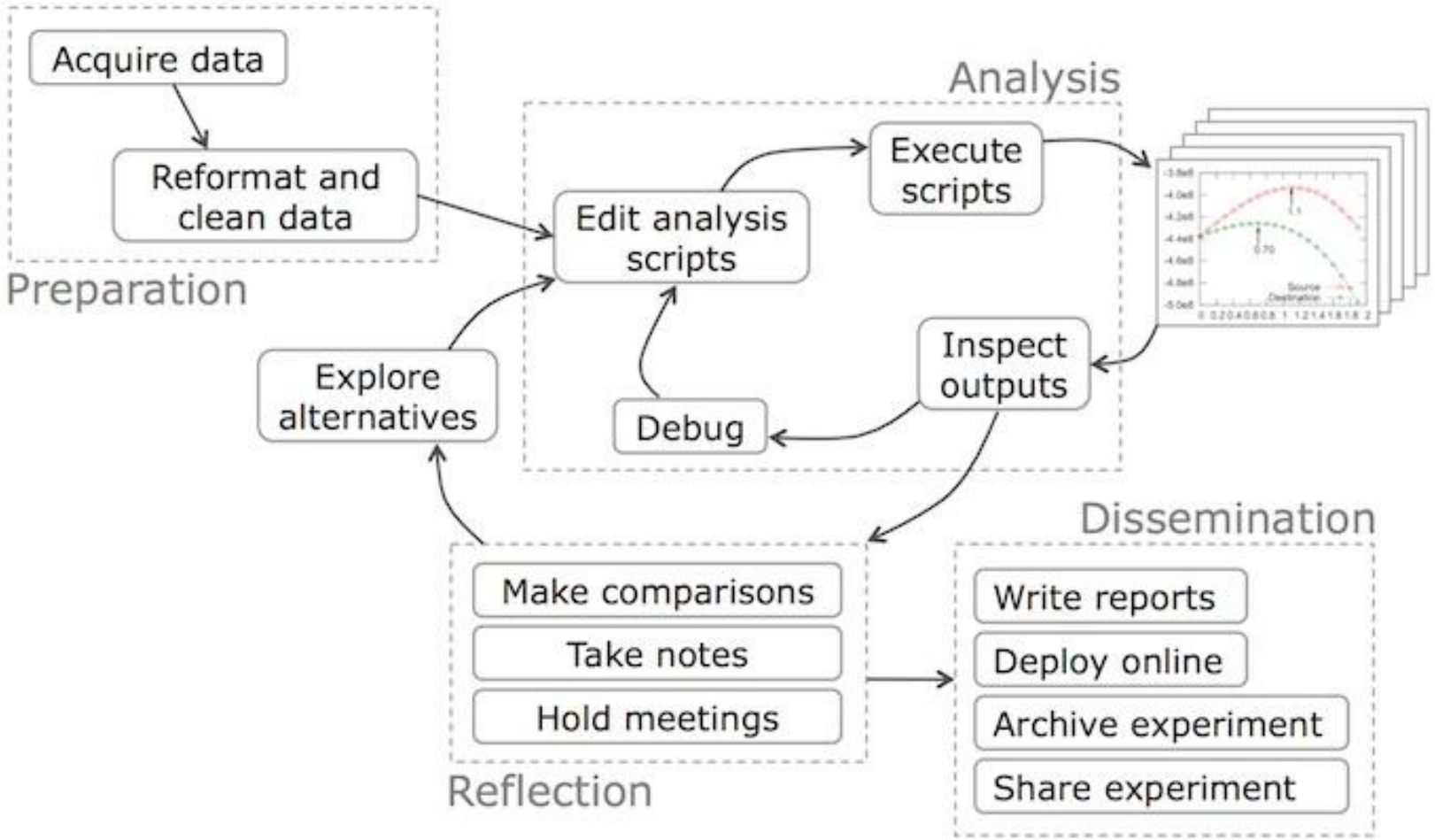
Data Scientist

- Ambiguous, ill-defined term
- From Drew Conway's Venn Diagram



Typical Data Scientist Workflow

- From Data Science Workflow



Where Data Scientist Spends Most Time

- 80-90% of the work is data cleaning and wrangling
- "Janitor Work" in Data Science

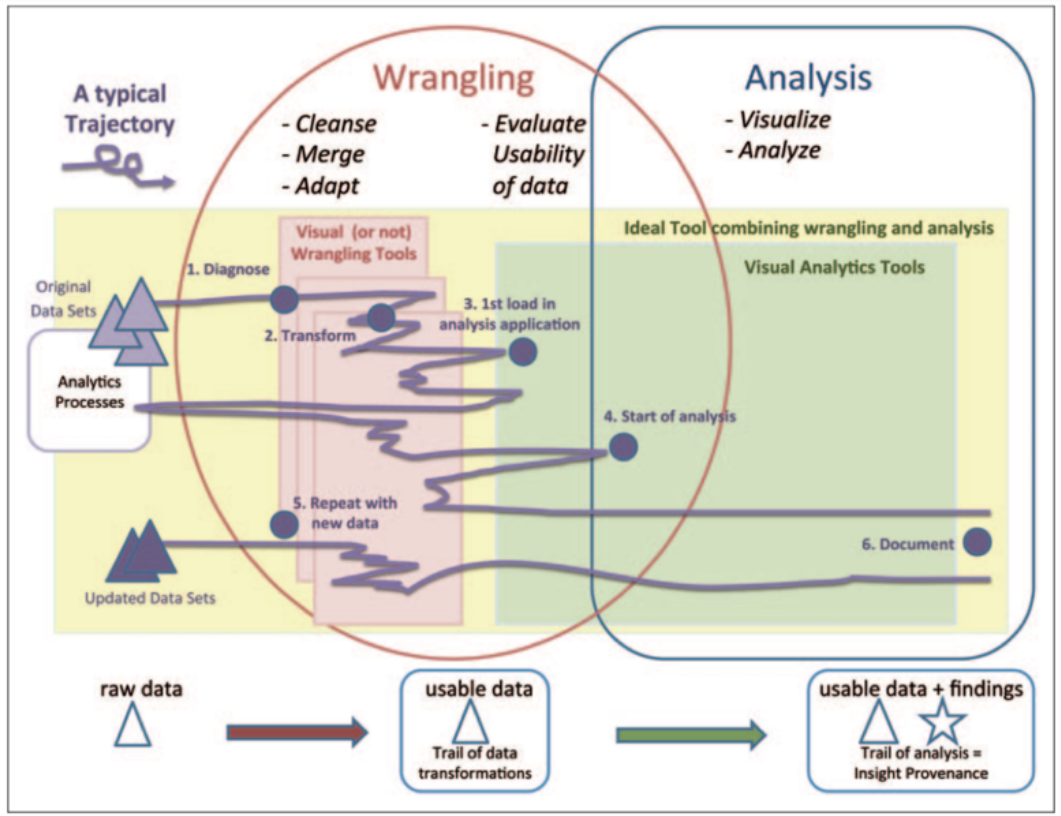


Figure 1. The iterative process of wrangling and analysis. One or more initial data sets may be used and new versions may come later. The wrangling and analysis phases overlap. While wrangling tools tend to be separated from the visual analysis tools, the ideal system would provide integrated tools (light yellow). The purple line illustrates a typical iterative process with multiple back and forth steps. Much wrangling may need to take place before the data can be loaded within visualization and analysis tools, which typically immediately reveals new problems with the data. Wrangling might take place at all the stages of analysis as users sort out interesting insights from dirty data, or new data become available or needed. At the bottom we illustrate how the data evolves from raw data to usable data that leads to new insights.

What a Data Scientist Should Know

- **Data grappling skills** ← *DATA*605
  - Move and manipulate data with programming
  - Scripting languages (e.g., Python)
  - Data storage tools: relational databases, key-value stores
  - Programming frameworks: SQL, Hadoop, Spark
- **Data visualization experience**
  - Draw informative data visuals
  - Tools: `D3.js`, plotting libraries
  - Know what to draw
- **Knowledge of statistics**
  - Error-bars, confidence intervals
  - Python libraries, Matlab, R
- **Experience with forecasting and prediction**
  - Basic machine learning techniques
- **Communication skills** ← *DATA*605
  - Tell the story, communicate findings