

A large, stylized teal graphic on the left side of the slide, consisting of several thick, wavy lines that curve and flow downwards, resembling a stylized 'S' or a series of connected loops.

Can AI/ML Drive Data Analysis and Data Mapping in Clinical Trials?

**Philip Ross, Ph.D.
Strategic Clinical Analytics**

(Opinions expressed are those of Philip Ross)

revvity
signals



The Software as a Service Platform for Pharma & Biotech

Viral adoption across our industry

Research

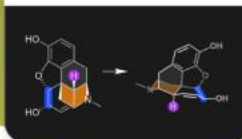
Large Molecule
Research

Small Molecule
Research

Formulations
Research

Testing

Assay Development



Development

Large Molecule
Development

Small Molecule
Development

Formulations
Development

Safety

Animal testing

Clinical & Manufacturing

Translational

Biomarker Discovery + Patient Stratification

Clinical Research
Research

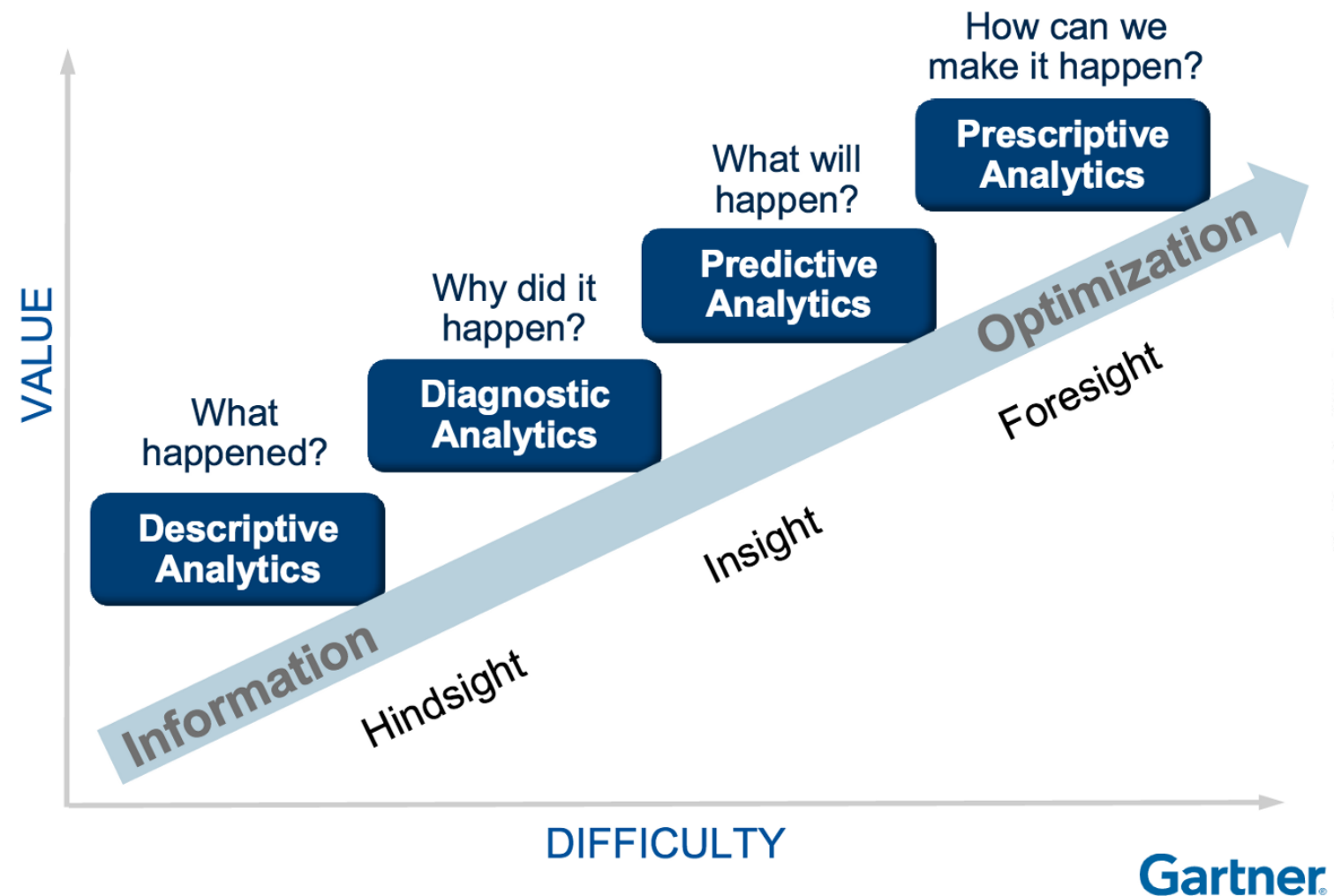
Clinical Operations
Risk Based Monitoring

CMC
QA / QC

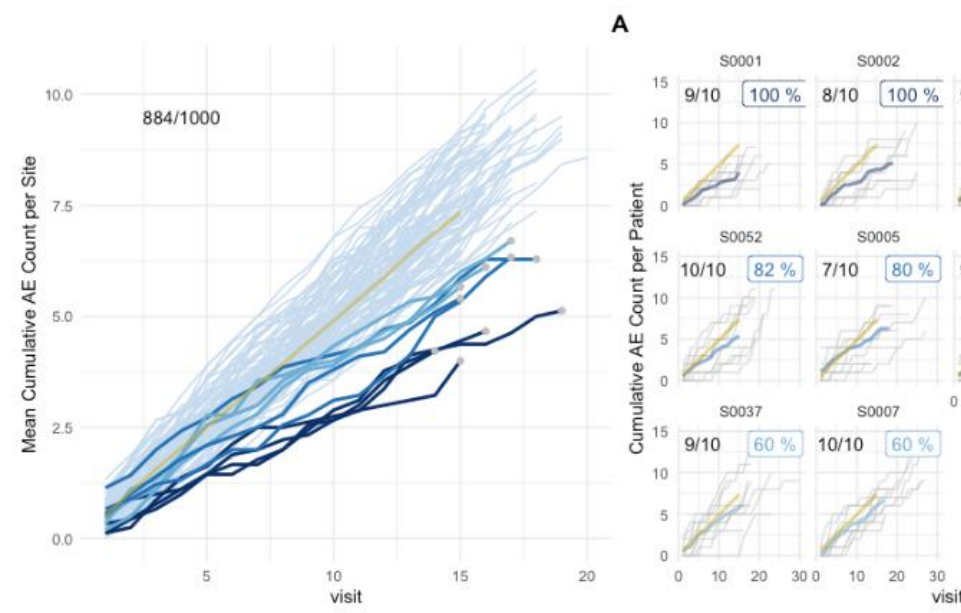


Data science as a spectrum... where does AI/ML fit?

Gartner Analytic Ascendancy Model

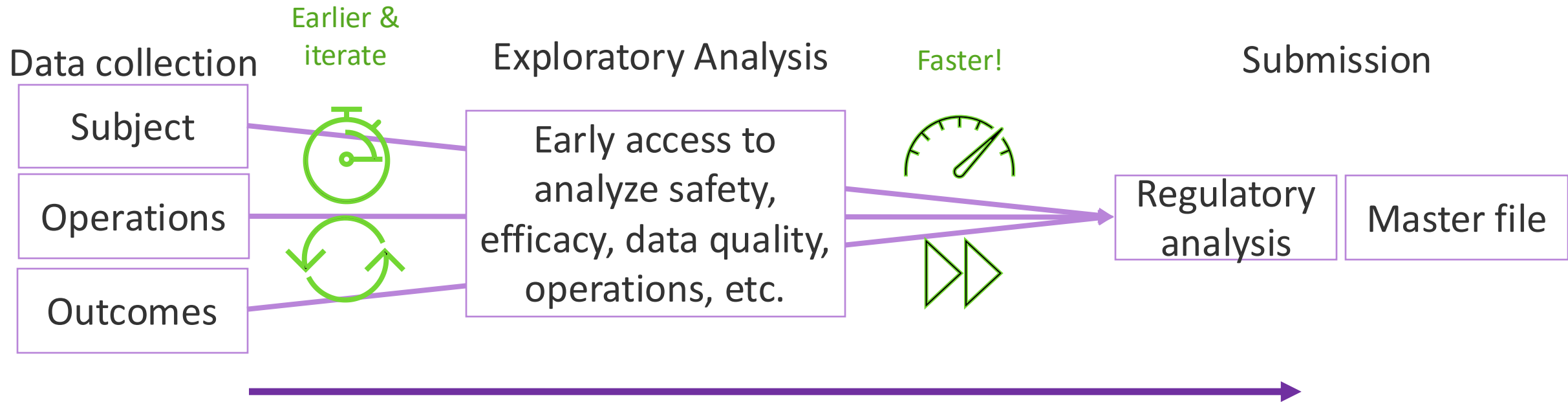


“What *should* be happening in my data?”



IMPALA Industry group

Clinical trials suffer from lack of access to analysis ready data & reproducible decision-making



- AI/ML is powerful and growing
- Examples: AI/ML to accelerate data mapping and data analysis

DATA

PEOPLE

DECISIONS

Real-time data access

Connect to all data

Reusable data maps

Transform any data

Robust self-service analytics

Interactive Visualizations

 Spotfire integration

Cross-study analysis

Traceable decisions

Track data changes

Query Workflow

Issue Management

1. How can AI automatically transform clinical datasets to a standard format such as CDASH/SDTM?

Data Automapping

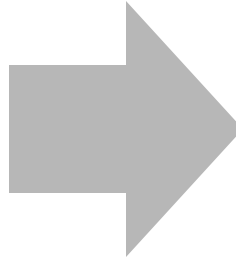
Data from Clinical Trials

- Variety of formats and schemas
- Many activities require a standard format
 - Visualizing with Spotfire
 - Storage in a database
 - Submission to the FDA
- Standardizing requires many manual hours

Automapping example

*Original AEACN
column*

Action Taken
Not Changed
Interrupted
Reduced
Discontinued



We want CDASH format

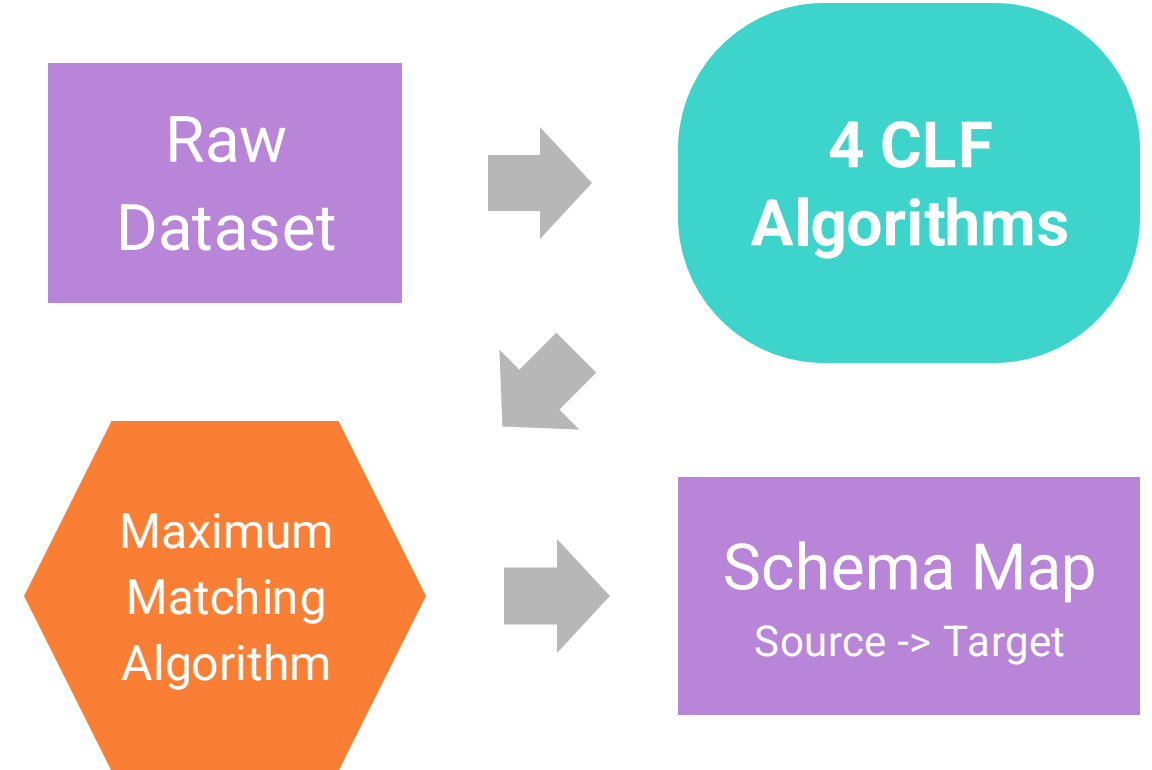
AEACN
DOSE NOT CHANGED
DRUG INTERRUPTED
DOSE REDUCED
DRUG WITHDRAWN

Goals of AI Automapper

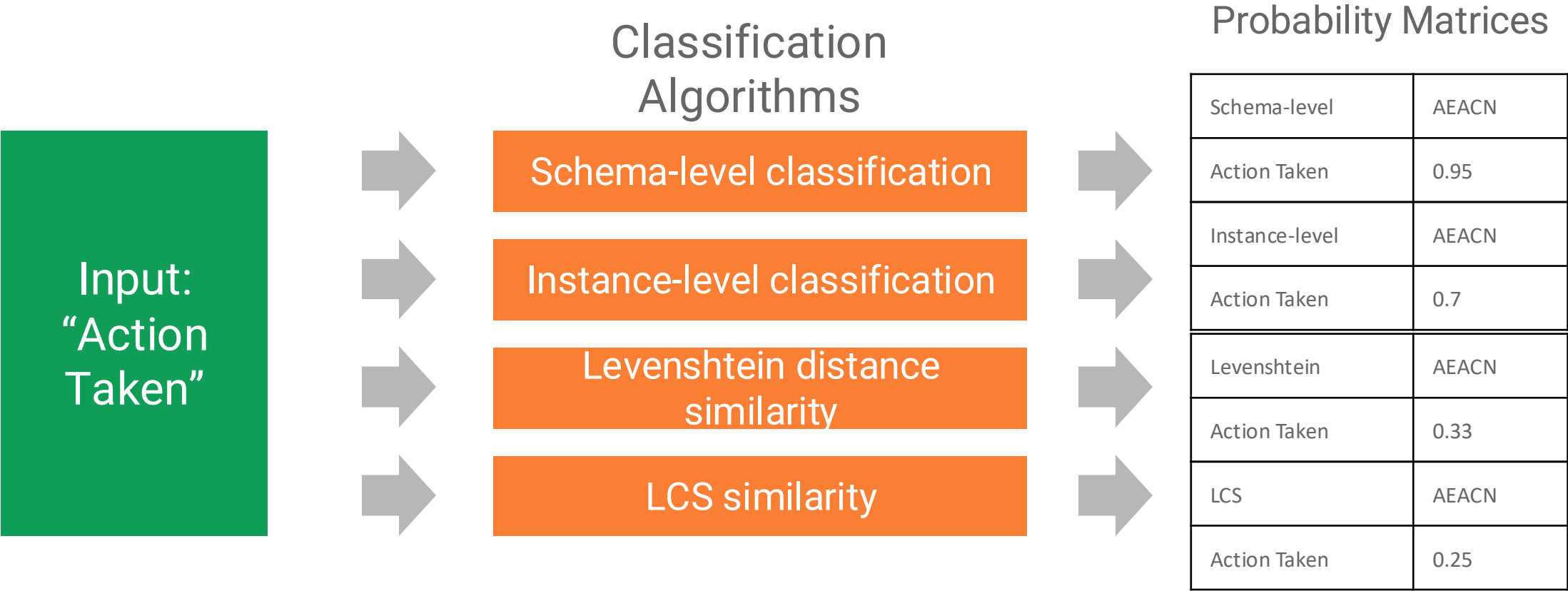
- Leverage the power of machine learning to predict schema mappings and data transformations
- Automatically transform around 80% of the dataset accurately
- Facilitate manual transformation of the remaining 20% and provide suggested mappings
- Learn from user input

Step 1: Schema Mapping

- Used 4 classification algorithms
- Trained on mostly simulated data
- “Bag-of-words” approach to features
 - Frequency of a-z characters
 - Frequency of 2 letter sequences
- Predict optimal 1-1 mapping of source and target columns



Schema Mapping Process



Schema Mapping Process (cont.)

Probability Matrices

Schema-level	AEACN
Action Taken	0.95
Instance-level	AEACN
Action Taken	0.7
Levenshtein	AEACN
Action Taken	0.33
LCS	AEACN
Action Taken	0.25



Weighted
combination
of the four
probability
matrices



Maximum
matching
algorithm



90%
sure?

No



"NA"

Yes



"AEACN"

	AEACN	STUDY
Action Taken	0.93	0.6
Study ACN	0.94	0.98

Step 2: Data Transformation

Automatically transform the mapped CDASH or SDTM columns

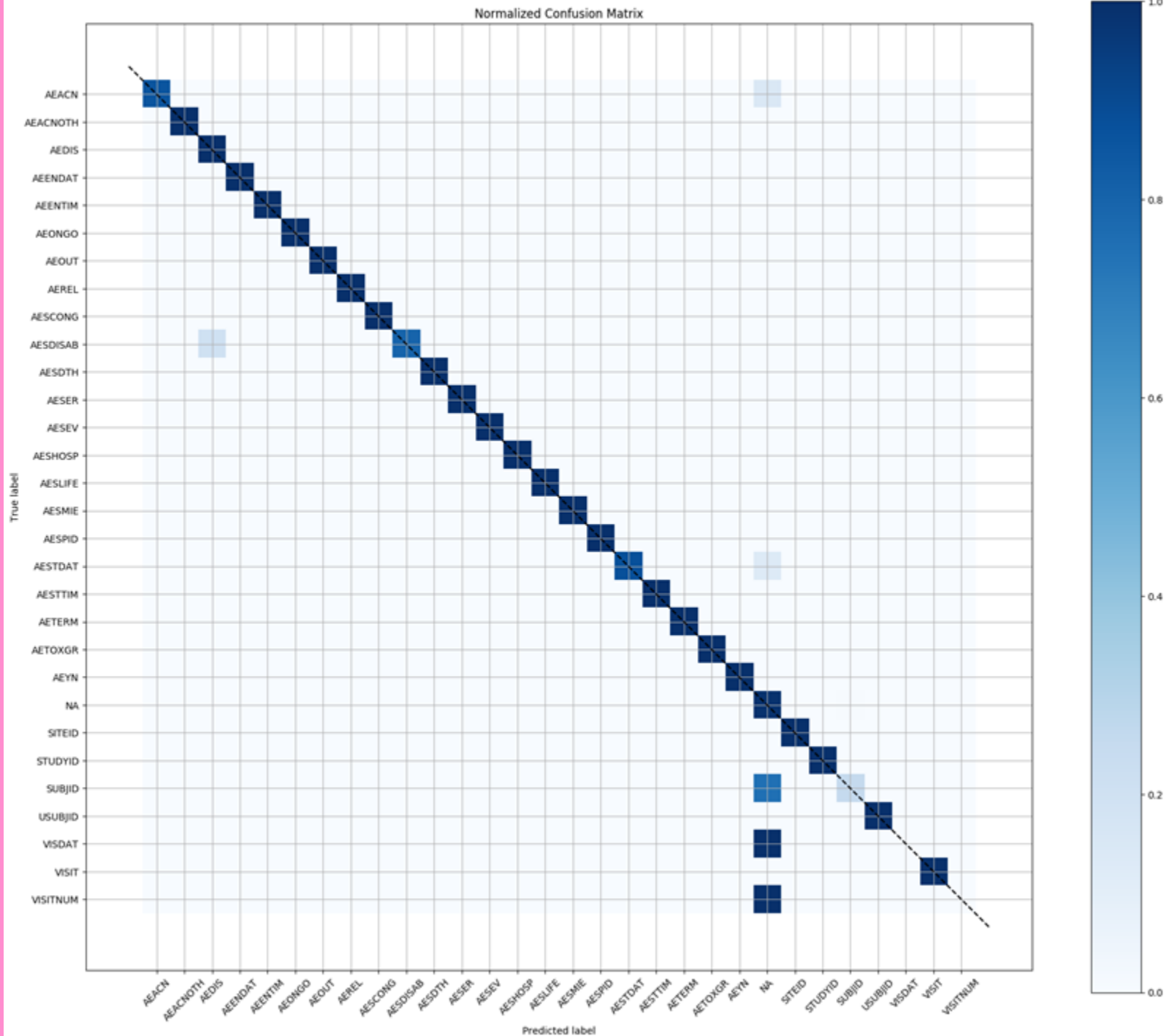
Two primary transformations:

Dates

Categories

Schema Mapping Results

- Accuracy = 97.3%
- Tested across 8 AE tables
- Classifies 35 standard CDASH columns
- Columns with no optimal CDASH mapping are labelled 'NA'
- Prefers classifying a column as 'NA' rather than incorrectly labelling it as a CDASH column



Automapping conclusions

- Significant improvements in data mapping is achievable
 - Patent granted in 2022
 - Time savings in programming
 - Quality improvements
- Challenges remain in:
 - Infrastructure
 - Building maps vs. loading annotated CRF



US 20190286620A1

(19) **United States**
(12) **Patent Application Publication** (10) **Pub. No.: US 2019/0286620 A1**
Al-Haimi et al. (43) **Pub. Date: Sep. 19, 2019**

(54) **METHODS AND SYSTEMS FOR
AUTOMATING CLINICAL DATA MAPPING
AND TRANSFORMATION**

(71) Applicant: **PerkinElmer Informatics, Inc.**,
Waltham, MA (US)

(72) Inventors: **Abdul Al-Haimi**, Waterloo (CA); **Chad
Millen**, Guelph (CA)

(21) Appl. No.: **16/356,187**

(22) Filed: **Mar. 18, 2019**

Related U.S. Application Data

(60) Provisional application No. 62/645,133, filed on Mar.
19, 2018.

Publication Classification

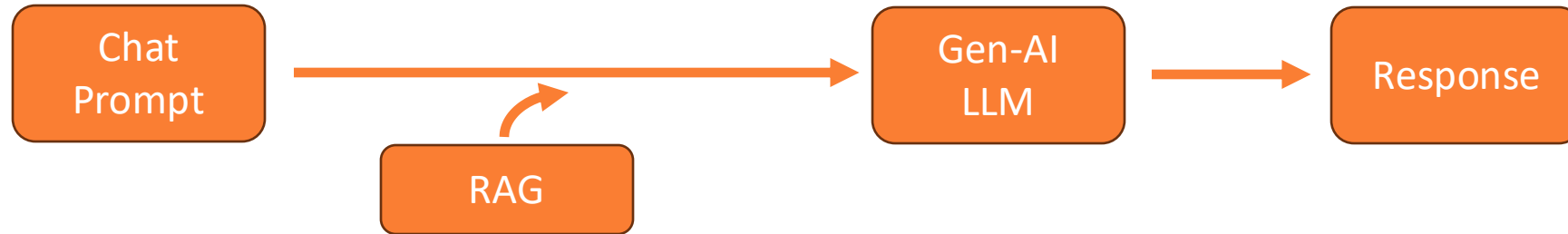
(51) **Int. Cl.**
G06F 16/21 (2006.01)
G06N 3/08 (2006.01)
G06N 3/04 (2006.01)
G16H 10/20 (2006.01)

(52) **U.S. Cl.**
CPC **G06F 16/211** (2019.01); **G16H 10/20**
(2018.01); **G06N 3/0454** (2013.01); **G06N**
3/08 (2013.01)

(57) **ABSTRACT**

Aspects discussed herein relate to employing deep learning to automate mapping and transformation of a source data set to a target data schema. A system may utilize deep learning algorithms to determine a mapping from the source schema to the target schema through identifying the source schema and creating a correspondence between source fields and target fields, and a corresponding data transformation. Artificial neural networks, configured as schema-level and instance-level classifiers, may generate a set of predictions based on the fields of the source data set and fields of the target data schema. These predictions may be combined with other predictions based on other criteria (such as similarity between the fields) to generate a complete prediction of a schema mapping. Similarly, deep learning techniques may be employed to determine an appropriate data transformation to transform source data content to an appropriate format for corresponding fields of the target schema.

Generative AI – What is it?



- Summarize past information and write documents
 - Leveraging trained information
 - Relies on Token Similarity and similar algorithms
 - Varied outputs based on Prompt
- Can AI leverage new information? Yes, but...
 - Simplest is to include information in the prompt
 - Emerging techniques seek to make more information available to AI
 - Retraining is a more expensive option, dependent on available training data

Simple prompt with current data

- Describe this data

Unique.Subject.Identifier Dictionary.Derived.Term Severity/Intensity 01-706-1049 SYNCOPES SEVERE 01-706-1049 SYNCOPES MODERATE 01-709-1424 SYNCOPES MODERATE 01-710-1166 SYNCOPES MODERATE 01-710-1166 SYNCOPES MODERATE 01-713-1043 NAUSEA MODERATE 01-713-1043 NAUSEA MODERATE 01-716-1071 SYNCOPES MODERATE 01-717-1344 NAUSEA MODERATE 01-718-1066 SYNCOPES SEVERE 01-718-1066 SYNCOPES SEVERE 01-718-1170 SYNCOPES SEVERE 01-718-1427 NAUSEA MODERATE 01-718-1427 NAUSEA MODERATE 01-718-1427 NAUSEA MODERATE 01-718-1427 NAUSEA MODERATE 01-718-1427 NAUSEA MODERATE 01-718-1427 NAUSEA SEVERE

Response to simple prompt with current data

- This data appears to be a medical or clinical dataset with three columns:
 1. Unique Subject Identifier: This column contains unique codes for each subject or patient, such as "01-706-1049" or "01-718-1427".
 2. Dictionary Derived Term: This column lists medical conditions or symptoms experienced by the subjects. In this dataset, we see two terms: "SYNCOPE" (fainting or passing out) and "NAUSEA".
 3. Severity/Intensity: This column indicates the severity level of the condition, with three possible values: "MODERATE" or "SEVERE".
 4. Some subjects appear multiple times in the dataset, suggesting multiple occurrences or reports of symptoms.
 5. Syncope is reported as both moderate and severe, while nausea is mostly reported as moderate with one severe case.
- This type of data is commonly used in medical research to track and analyze the frequency and severity of symptoms or side effects in a study population.

Mostly correct, item 3 counts 3 values
but displays the correct 2 values

Prompts using RAG with AI

- Using the knowledge sources, list the dictionary terms that might be of concern
 - The dictionary terms that might be of concern include "SYNCOPE" and "NAUSEA" with varying severity levels such as SEVERE and MODERATE
- Using the ai training, which of the two dictionary terms are likely to be of greater health concern and what are the health concerns?
 - Based on the information provided, SYNCOPE appears to be of greater health concern as it is listed with both severe and moderate severity/intensity, whereas NAUSEA is mostly listed with moderate severity and only one instance of severe severity.
 - Specific health concerns for each term:
 - SYNCOPE: This term refers to fainting or a temporary loss of consciousness, which can be severe and may indicate underlying health issues such as cardiovascular problems.
 - NAUSEA: This term refers to the feeling of sickness with an inclination to vomit, which is generally less severe but can still be uncomfortable and may indicate gastrointestinal issues or other health conditions.

Prompt 3 using RAG with AI – counting and ranking

- List the subjects with the most events and the total number of events, corresponding dictionary terms and severity for each subject
 - Subject 01-718-1427:
 - Total number of events: 6 **correct**
 - Dictionary term: NAUSEA
 - Severities: MODERATE and SEVERE
 - Subject 01-718-1066:
 - Total number of events: 3 **Incorrect** - only 2 events, 3 patients with 2 AEs
 - Dictionary term: SYNCOPÉ
 - Severity: SEVERE

Prompt 4 using RAG with AI

- What data is needed for an eDISH plot?
 - An eDISH plot typically requires data on laboratory measurements, particularly those related to liver function. This includes parameters such as lymphocytes and mean corpuscular hemoglobin (MCH) levels **Incorrect**
- Do eDISH plots typically plot peak alt and bili?
 - Yes, eDISH plots typically plot peak ALT (Alanine Aminotransferase) and bilirubin levels to assess liver function and detect potential drug-safety issues. **correct**
- Are there peak ALT and bilirubin in the Knowledge file?
 - Yes, there are peak ALT and bilirubin values in the Knowledge file. For ALT, the peak value is 40 U/L . For bilirubin, the peak value is 0.8 mg/dl . **Incorrect** - 129 and 7.3, respectively

Prompt 5 using RAG with AI – listing works for 1195 rows of data

- Using the knowledge sources, list the dictionary terms that might be of concern?
 - The dictionary terms of concern in clinical products from Revvity include: **extrapolation** – to all Revvity products
 - Application site erythema **Correct**- simply listed unique dictionary terms
 - Diarrhoea
 - Application site pruritus
 - Erythema
 - Atrioventricular block second degree
 - Fatigue
 - ...

AI summary of FDA Validation Rules

- Good morning, AI. You are a clinical data manager. Are you familiar with FDA validation rules?
 - Good morning! As an AI language model, I don't have personal experiences or specific professional roles. However, I can provide information about FDA validation rules based on my training data.
 - FDA validation rules are an important aspect of clinical data management and are used to ensure the quality, integrity, and reliability of data collected during clinical trials. These rules are designed to meet regulatory requirements and Good Clinical Practice (GCP) standards. **correct**
 - While I can discuss general concepts related to FDA validation rules, it's important to note that specific rules and their implementation may vary depending on the study, therapeutic area, and regulatory requirements. If you have any specific questions about FDA validation rules, I'd be happy to try to answer them or provide general information on the topic.

AI – summary of FDA AI guidance

- Here's a summary of the key points from the FDA's draft guidance on the use of Artificial Intelligence (AI) to support regulatory decision-making for drug and biological products:
 1. **Scope:** The guidance covers the use of AI models in the drug product life cycle to produce information or data supporting regulatory decisions on safety, effectiveness, or quality of drugs.
 2. **Risk-Based Credibility Assessment Framework:** The guidance proposes a 7-step process to establish and assess the credibility of an AI model output for a specific context of use (COU) based on model risk.
 3. **Model Risk:** Defined as a combination of model influence and decision consequence.
 4. **Credibility Assessment Plan:** Should include detailed descriptions of the model, data used for development, training process, and evaluation methods.
 5. **Life Cycle Maintenance:** Emphasizes the importance of ongoing monitoring and management of AI models, especially in pharmaceutical manufacturing.
 6. **Early Engagement:** Strongly encourages sponsors to engage early with FDA to discuss credibility assessment activities and potential challenges.
 7. **Engagement Options:** Provides various options for sponsors to engage with FDA depending on the intended use of the AI model.
- The guidance aims to provide a framework for establishing the credibility of AI model outputs in regulatory decision-making while emphasizing a risk-based approach and the importance of early engagement with the FDA.

correct

Prompt Engineering

- Zero-, One-, and Few-Shot Prompting
 - Simple question, Question with one example, Question with a few examples
- Role, Style, Context and Emotion Prompting
 - You are a data manager
- Rephrase and Respond, Reread, System Self-Ask Prompt
 - AI takes actions to improve responses
- Chain-of-Thought Prompt (Let's Think Step-by-Step)
 - Break this problem into smaller steps. At each step, explain the reason for the suggestion
- Step-Back Prompt (Broad prompt followed by specific prompt)
 - Explain the key factors for drug approval. Based on this, what are the key factors for approval of an Alzheimer's drug?

Expected reasons for erroneous AI responses

- Lack of prompt clarity
 - Too short or too long
- Noun-phrase routes drive AI token similarity matching
 - Avoid pronouns or ambiguity; noun specificity is key
 - Expect generalization when specific matches are not in training
- Lack of current information
 - Historical training data
 - Some data can be in prompt; too much data can skew tokenization giving poor counting
- Use additional prompts to clarify or correct incomplete responses
- Check factual references, counts, details
 - Literature citations often have incorrect details due to semantic matching (pulling information from a bin of tokens)

Summary

- AI/ML can be useful to recommend data mapping and transformation
- Gen-AI LLMs – can we make use of the massive development investments
 - LLM best knows historical information (training)
 - Describe FDA Validation Rules or AI guidance
 - Describe how to program a visualization or overall data features
 - Challenges when asking for exact data values (smallest, outliers)
 - Prompt engineering
 - Exploring Retrieval-, SQL- and Table-Augmented Generation
 - Use as an Assistant (or Intern) – recommend, subject to expert review

Thank you!

For more information:

<https://revvitysignals.com/products/clinical-development/signals-clinical>

