EZ-ETL -> OKDOCTOR

Ontology-Driven Data Analysis
Supporting Evidence Based Medicine

Vijay Gaikwad and Tom Vandermolen
Sponsors: Mala Sarat Chandra, Mike Doane

Problem: Cleaning "Dirty" Data

Big Data = Big Problems

- Unstructured, "dirty" data is everywhere
- No formalized method to turn unstructured data into useful structured data
- Process is repeated for every problem
- Very difficult to automate the process of finding relationships and structure in textual data

80% of a data scientist's time is spent preparing the data *before analysis*

The Extract, Transform, Load (ETL) process is usually still carried out MANUALLY

Solution: EZ-ETL

An ontology-driven Extract, Transform, Load (ETL) process, EZ-ETL combines Semantic Web and Machine Learning techniques to extract concepts from any text data source and translate them into concepts specific to the customer knowledge domain.

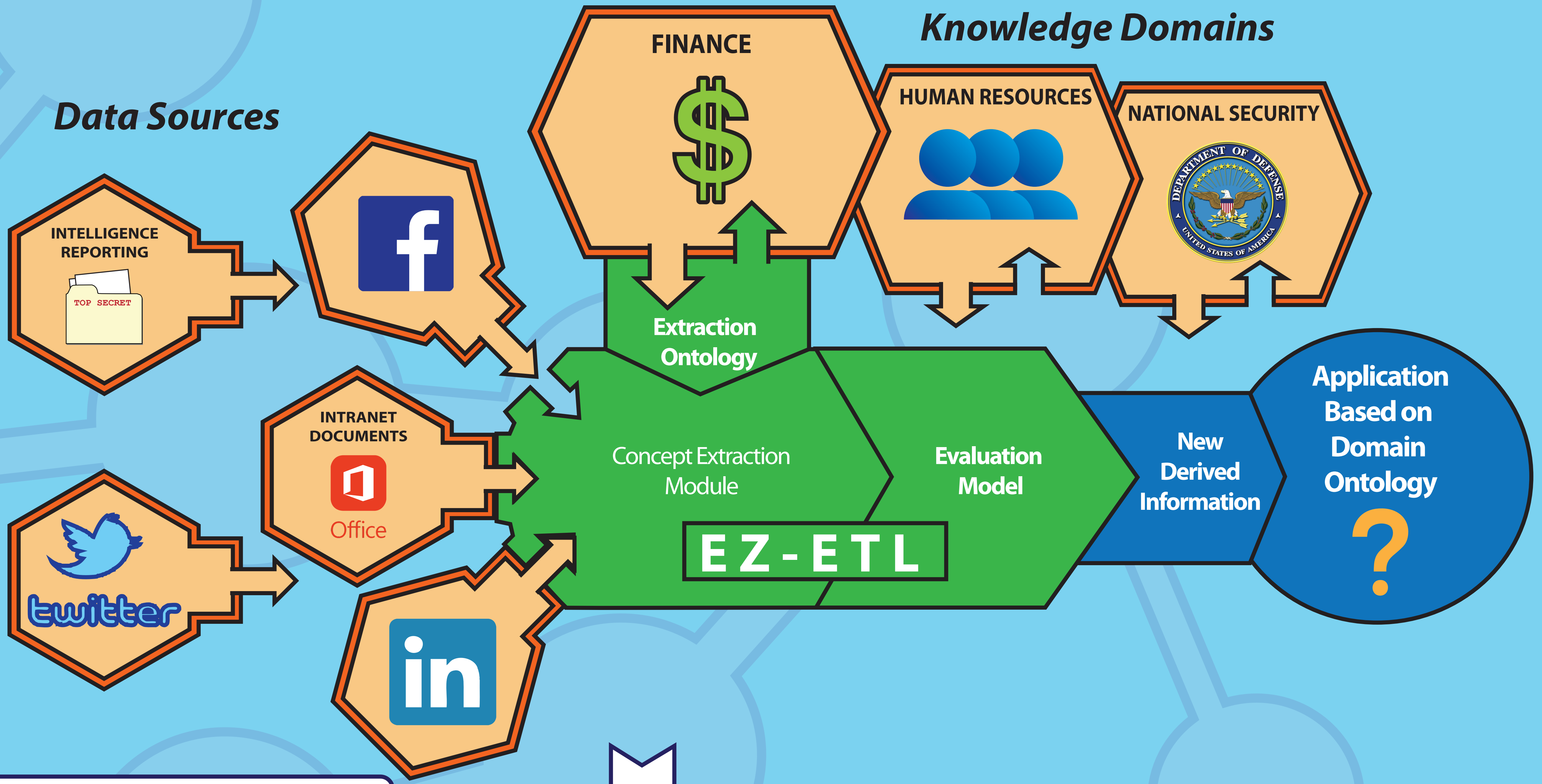
- Uses a customer-provided Domain Ontology as a "schema" to structure unclean data sources
- The same customer-based domain ontology drives:
  - -- Back-end data processing
  - -- Front-end user application
- Allows re-use of existing rich ontologies

A PLATFORM SOLUTION

This approach can be used for any general data problem, with any type of text-based data--the only variable is the Domain Ontology.

Future Work:

- Incorporate more complex Machine Learning techniques
- Incorporate more advanced Natural Language Processing
- Exploit the network of relationships in the Domain Ontology



Knowledge Domains

FINANCE

HUMAN RESOURCES

NATIONAL SECURITY

Data Sources

INTELLIGENCE REPORTING

TOP SECRET

INTRANET DOCUMENTS

Office

twitter

in

Extraction Ontology

Concept Extraction Module

Evaluation Model

EZ-ETL

New Derived Information

Application Based on Domain Ontology

?

EZ-ETL Applied to the Domain of Evidence Based Medicine

Problem: Evaluating Evidence For Evidence Based Medicine

Evidence Based Medicine (EBM) requires practitioners to assess the quality of published medical research for relevance and validity to their specific field. However, there is an overwhelming volume of research being published, and evaluating research requires careful reading of each report.

3,000 Research articles published on PubMed *every day*

How can practitioners keep up? How can they identify and use quality research?

Solution: OKDoctor

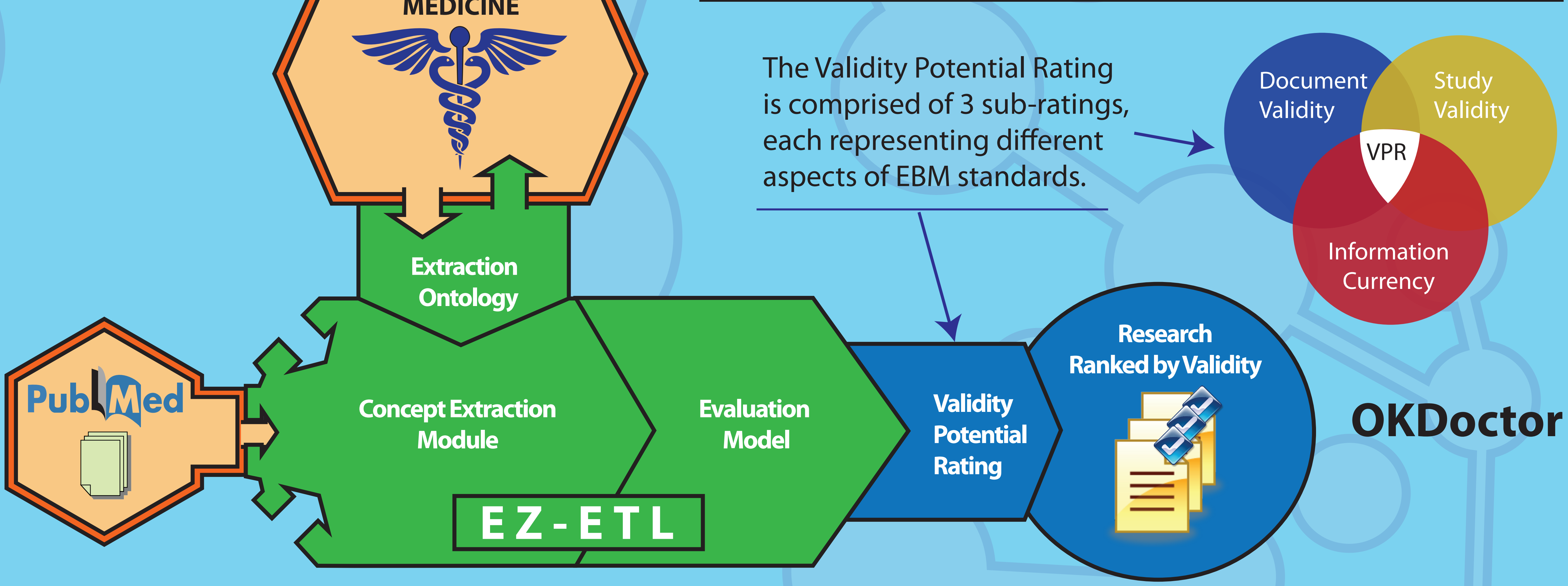
Existing solutions provide either too much or not enough information. OKDoctor complements both approaches by providing an evaluation metric--the Validity Potential Ranking--to help EBM practitioners quickly find the evidence with the highest validity.

The Validity Potential Ranking:

- Rates whether the article has the earmarks of a valid study
- Helps practitioner find most valid documents

Future Work:

- More advanced Natural Language Processing
- Enhance relevance and improve extraction accuracy
- Expert feedback on VPR equation and factors



EVIDENCE BASED MEDICINE

The Validity Potential Rating is comprised of 3 sub-ratings, each representing different aspects of EBM standards.

Document Validity

Study Validity

VPR

Information Currency

PubMed

Extraction Ontology

Concept Extraction Module

Evaluation Model

EZ-ETL

Validity Potential Rating

Research Ranked by Validity

OKDoctor

Technology and Assistance Provided by:

poolparty

SoMe Lab
Social Media Lab @ UW