By Joaquin Delgado, PhD.
and Diana Hu

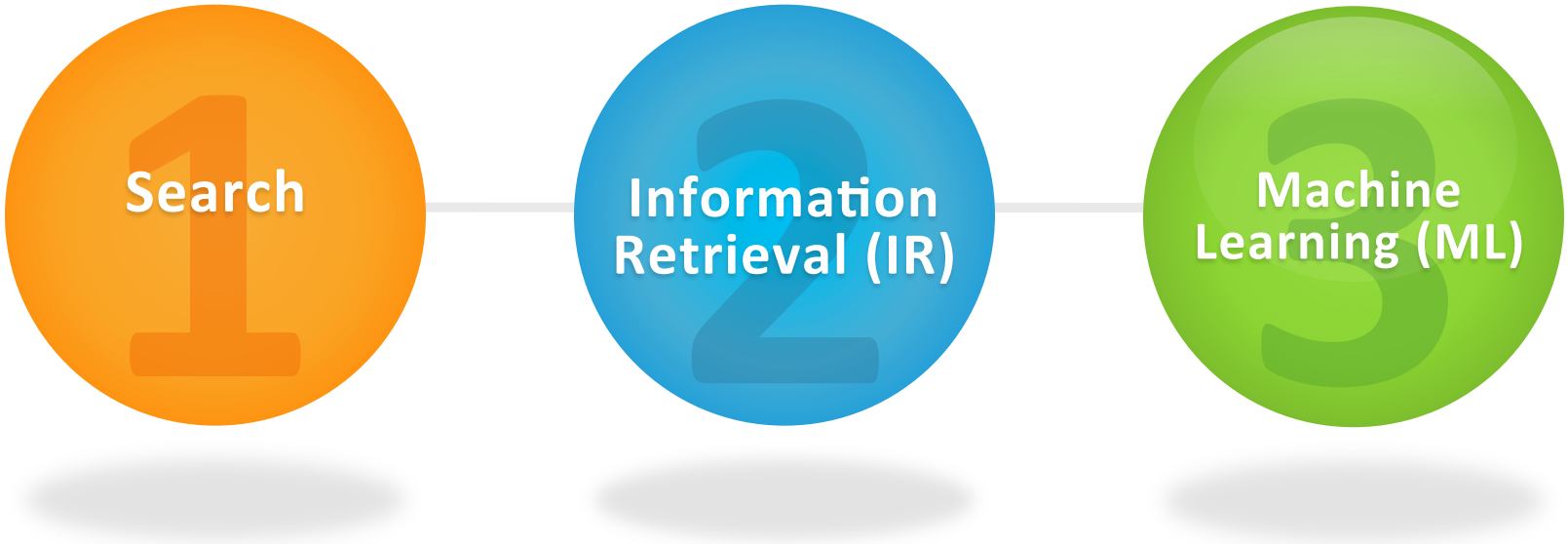
ML-Scoring

Where Search Meets Machine Learning

Disclaimer

The content of this presentation are of the authors' personal opinion and does not officially represent their employer's view in anyway. Included content is especially not intended to convey the views of OnCue or Verizon.

Finding Commonalities

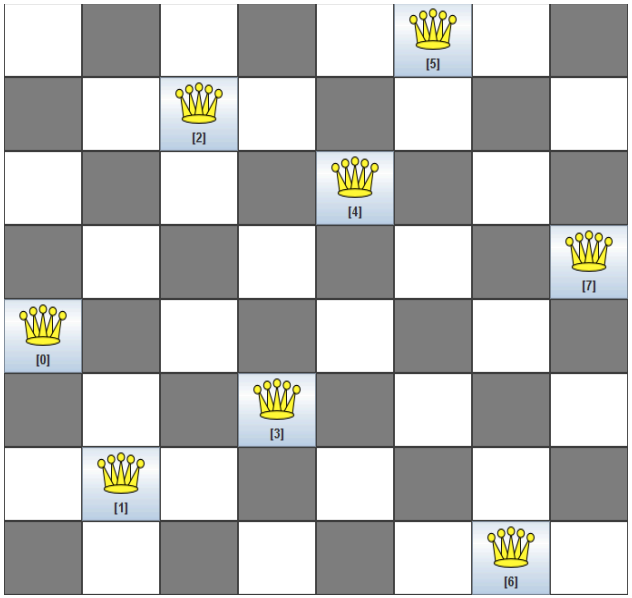


Can you map these to other things?

Search is a method for solving CSPs. The objective is to find all possible solutions (objects) that satisfy a set of constraints (query) on variables (fields). Optionally, scoring is used to rank results

Predicate Logic and Declarative Languages Rock!

Constraint Satisfaction Problem



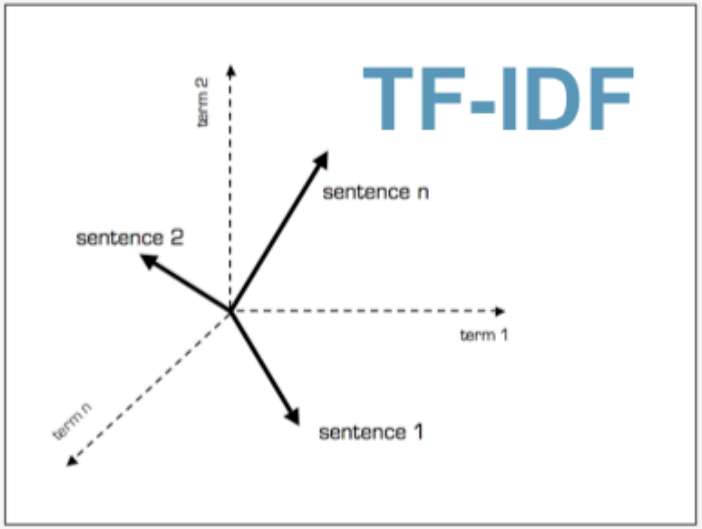
# Queens	# possible solutions	# feasible solutions	# optimal solutions	# optimal out of # possible
4	256 (4^4)	2	2	1 *out of* 128
8	8^8	64	64	1 *out of* 262144
16	16^16	14772512	14772512	1 out of 1248720872503
32	32^32	?	?	?
n	n^n	?	# fesasible	?

N-Queens Problem: Where to place N-Queens on a chessboard that don't interfere with (cannot capture) each other?

Information Retrieval is search + *relevance ranking* (scoring) applied to text documents and fields; often referred to as text or keyword search.

Have you heard of Bag-of-Words? Vector Space Representation? What about TF-IDF?

Ranking in the Vector Space Model



Language Model

$P(\text{optimization} \mid \text{search, engine}) >> P(\text{walking} \mid \text{search, engine})$

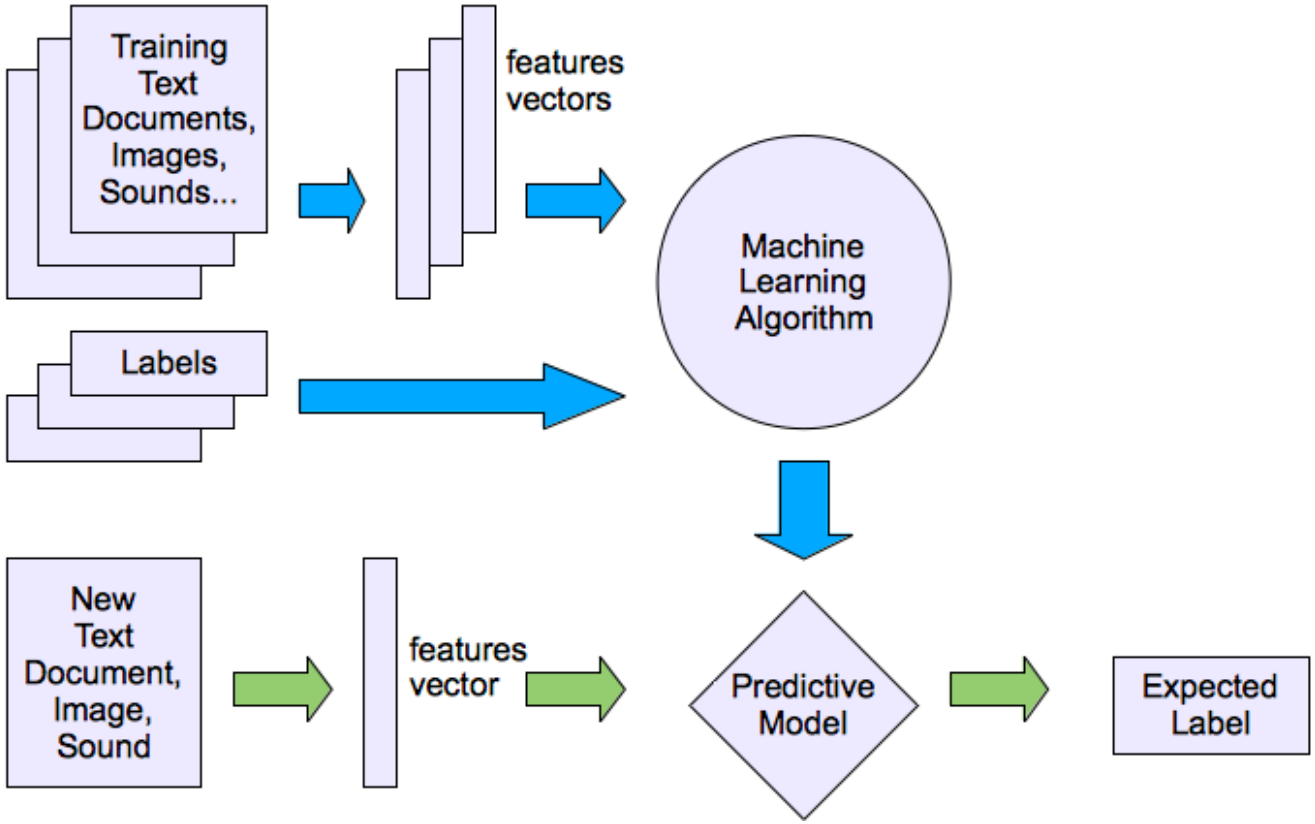
Probability Ranking Principle

$P(R = 1 \mid d, q)$ or $P(R = 0 \mid d, q)$

Machine Learning and in particular *supervised learning* refer to techniques used to learn how to classify or score previously unseen objects based on training data

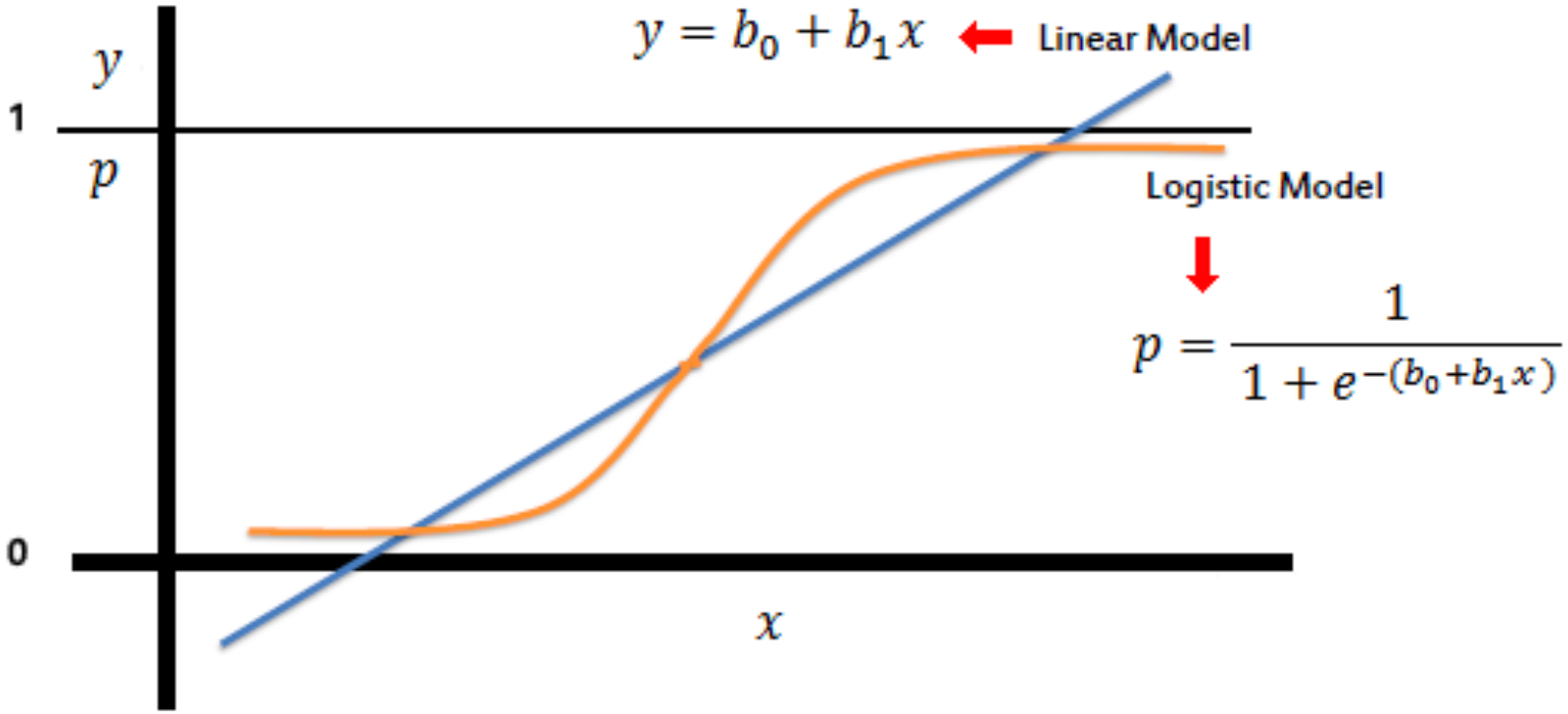
Inference and Generalization are the Key!

Generating a Predictive Model from Supervised Learning



Expected labels are generated for a classification task; instead this same process can be used to learn and generate a score (also called a *regression task*)

Example of ML Algorithm: Logistic Regression



Examples of Other Search-based Systems

- Database Systems
- AI Systems
- E-Commerce Systems
- Recommender Systems
- Advertising Systems

Search Engines : The Big Hammer!

- Search engines are largely used to solve non-IR search problems, and here is why:
  - Widely Available
  - Fast and Scalable
  - Integrates well with existing data stores (SQL and No-SQL)

But Are We Using the Right Tool?

- **Search Engines** were originally designed for IR.
- More complex non-IR search tasks sometimes require a **two phase approach**

Information Retrieval based Scoring

Ranking Formula

Called Lucene Similarity

Score of a document for a given query

Normalized doc length, shorter docs are more likely to be relevant than longer docs

$$score_{q,d} = norm(q) \times \sum_{t\ in\ q} \sqrt{tf_{t,d}} \times idf_t^2 \times norm(d,field) \times boost(t)$$

Can be ignored (was an attempt to make query scores comparable across indices, it's there for backward compatibility)

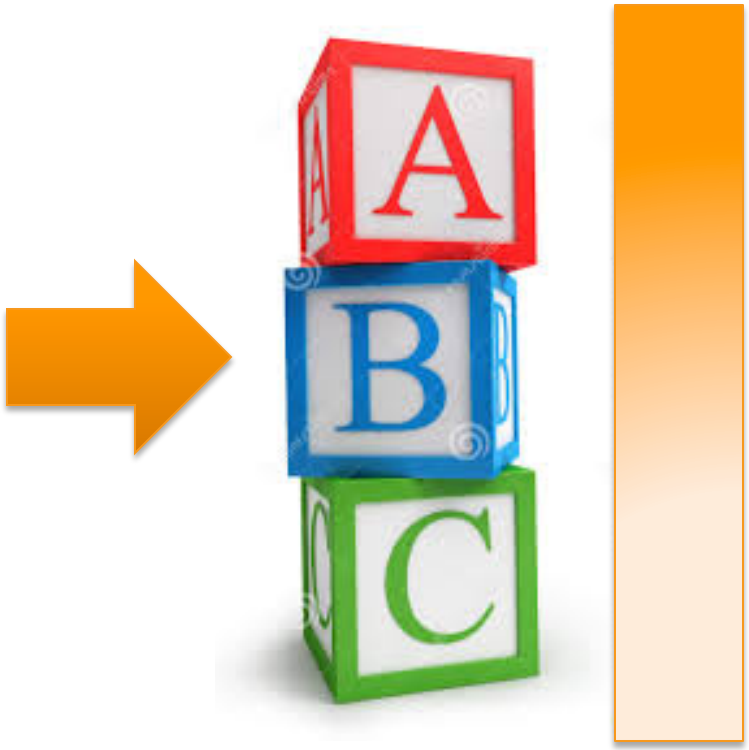
Core TF/IDF weight

Boost of query term t

Complex Scoring: Two Phase Approach

Filter

Rank

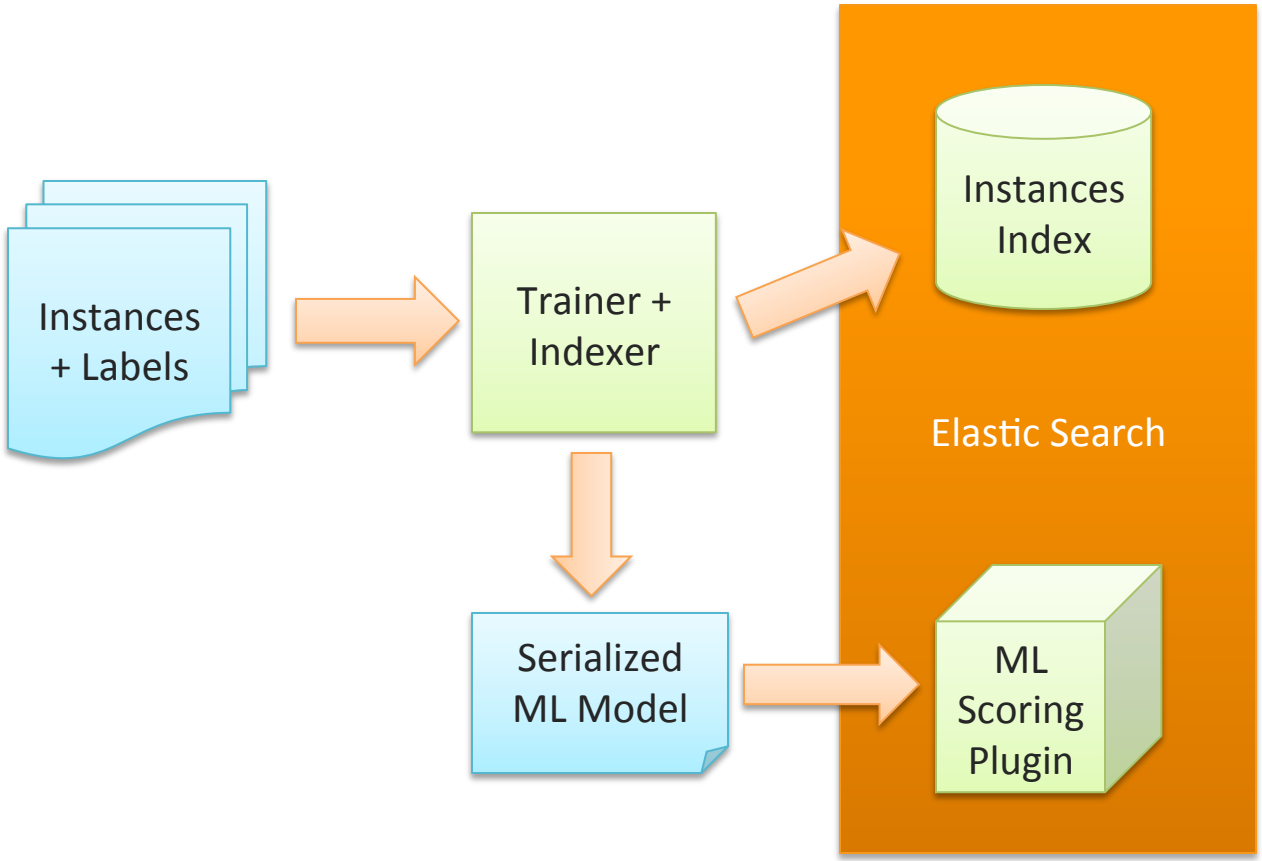


What is ML-Scoring?

- Create an Elastic Search (ES) document index of instances
- Train a supervised learning ML model from a subset of instances + labels
- Generate a ES plugin that uses the ML model to srore documents

An Open Source POC!

Introducing ML-Scoring



Search Predictor (ML-Scoring) Query

```
{
  "query": {
    "function_score": {
      "query": {
        "match_all": {}
      },
      "functions": [
        {
          "script_score": {
            "script": "search-predictor",
            "lang": "native",
            "params": {}
          }
        }
      ],
       "boost_mode": "replace"
    }
  }
}
```

Support for Various ML Libraries

Only Linear Models for now…

https://github.com/sdhu/elasticsearch-prediction

Potential Issues

- Performance
  - It may be a problem if the search space is very large and/or the computation too intensive
- Operations
  - Code running on a key infrastructure
  - Versioning and binary compatibility

Questions?