Conclusions and Future Directions

Overview and Introduction

Knowledge Extraction

Knowledge Cleaning

Q&A

Break

Ontology Mining

Applications

Conclusion and Future Directions **10 min**

Q&A

What’s So Special about PKGs?

Primarily text heavy	• Textual product profiles. Other modalities complement text. • Explicit natural language handling is critical.
Other modalities	• Product images can provide important additional signals.

What's So Special about PKGs?

Primarily text heavy	• Textual product profiles. Other modalities complement text. • Explicit natural language handling is critical.
Other modalities	• Product images can provide important additional signals.
Noise	• Product profiles and corresponding facts can be noisy. • Need explicit noise handling, and data cleaning steps.

What's So Special about PKGs?

Primarily text heavy	• Textual product profiles. Other modalities complement text. • Explicit natural language handling is critical.
Other modalities	• Product images can provide important additional signals.
Noise	• Product profiles and corresponding facts can be noisy. • Need explicit noise handling, and data cleaning steps.
Hierarchical structure	• The product taxonomy is mainly hierarchical in nature.

What's So Special about PKGs?

Primarily text heavy	• Textual product profiles. Other modalities complement text. • Explicit natural language handling is critical.
Other modalities	• Product images can provide important additional signals.
Noise	• Product profiles and corresponding facts can be noisy. • Need explicit noise handling, and data cleaning steps.
Hierarchical structure	• The product taxonomy is mainly hierarchical in nature.
Dynamic taxonomy	• Constantly emerging product categories. • Automatic taxonomy enrichment.

What's So Special about PKGs?

Primarily text heavy	• Textual product profiles. Other modalities complement text. • Explicit natural language handling is critical.
Other modalities	• Product images can provide important additional signals.
Noise	• Product profiles and corresponding facts can be noisy. • Need explicit noise handling, and data cleaning steps.
Hierarchical structure	• The product taxonomy is mainly hierarchical in nature.
Dynamic taxonomy	• Constantly emerging product categories. • Automatic taxonomy enrichment.
User activity	• User search logs, product complement, co-view and substitute purchases can be very useful.

Knowledge Extraction Takeaways

- **Definition**: Find values for a given product and a set of attributes.
- **Recipe**: Sequence tagging.
- **Key to Success**: Scale up in different dimensions (#attributes, #categories).
- **Applicability to other domains**: Domains like finance, biomedical etc, where the “subject” is known.

Knowledge Cleaning Takeaways

- **Definition**: Finding wrong attribute values.
- **Recipe**: Identify data inconsistency column-wise, row-wise, source-wise and across sources.
- **Key to Success for Products**:
  - Leverage rich textual information of unstructured data as context
  - Solution with aware of taxonomy.
- **Applicability to Other Domains:** Domains like: medical, legal, etc.
  - Domains with heavy text data.
  - Rich taxonomy information.

Ontology Enrichment Takeaways

- **Definition**: discover relations between product categories and attributes.
  - Attribute Applicability: “Is an attribute applicable to one product category?”
  - Attribute Importance: “Is an attribute important when people are making their purchase decisions?”
- **Recipe**: Text Mining and Graph Mining.
- **Key to Success for Products**: Leverage both seller/customer inputs.
- **Applicability to other domains**:
  - An increasing variety of relationships or predicate diversity.
  - Quantify the relation strength.

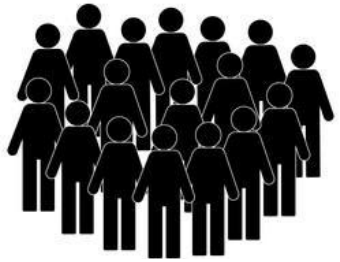
Applications Takeaways

- Applications of product knowledge graphs can make use of:
  - The structured factual information for each product.
  - The product connections in the overall graph structure.

- The graph structure also allows the utilization of graph level constructs, like knowledge graph embeddings, which is useful for many applications.

- General applications of knowledge graphs include recommendation systems, search, among others.

An End-to-End Pipeline



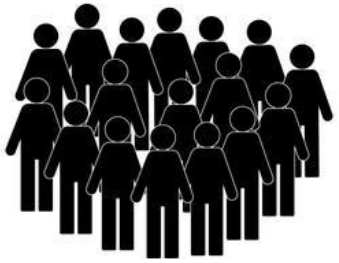
Understand domain and attributes, and generate LOTS OF training data

Identify product taxonomy and attributes

Train and fine-tune models

Extraction Models

Postprocess extraction results to further improve data quality



Pre-publish evaluation as gatekeeper to guarantee high quality data

An End-to-End Pipeline

Automatic Training Data Generation

Distant supervision, Data programming

Automatic Taxonomy Enrichment

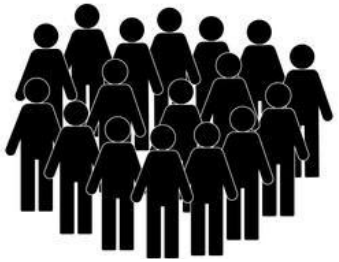
Category classification, attribute identification

Train and fine-tune models

Extraction Models

Benchmarking

Postprocess extraction results to further improve data quality



Pre-publish evaluation as gatekeeper to guarantee high quality data

An End-to-End Pipeline

Automatic Training Data Generation

Distant supervision, Data programming

Automatic Taxonomy Enrichment

Category classification, attribute identification

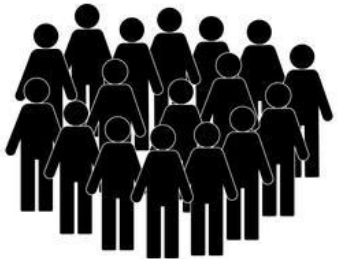
Train and fine-tune models

Extraction Models

Benchmarking

Deep Learning Data Cleaning

Postprocess extraction results to further improve data quality



Pre-publish evaluation as gatekeeper to guarantee high quality data

An End-to-End Pipeline

Automatic Training Data Generation

Distant supervision, Data programming

Automatic Taxonomy Enrichment

Category classification, attribute identification

Train and fine-tune models

Extraction Models

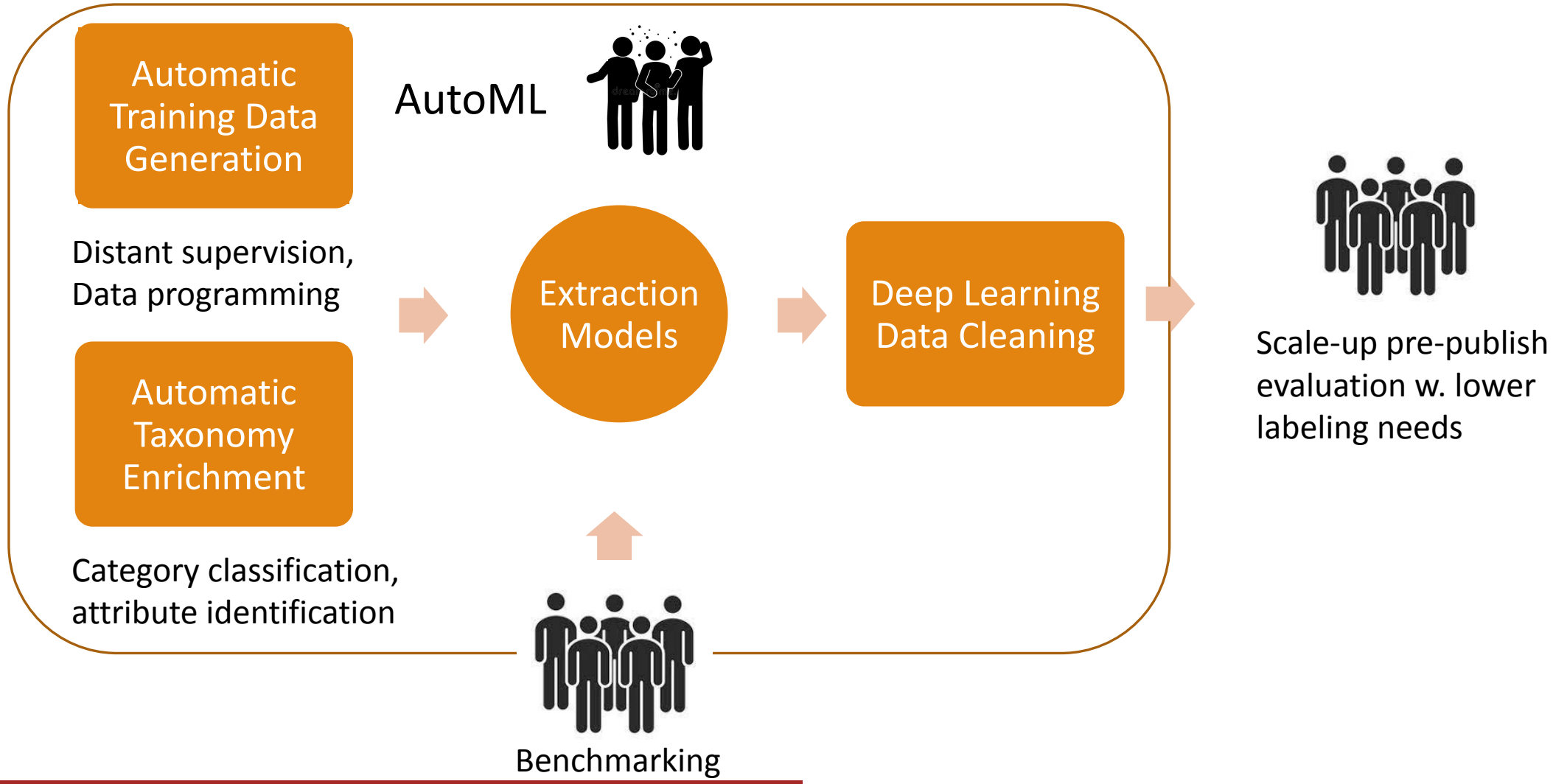
Benchmarking

Deep Learning Data Cleaning

Postprocess extraction results to further improve data quality

Scale-up pre-publish evaluation w. lower labeling needs

An End-to-End Pipeline



Practical Tips

- **Training data**
  - Mainly distant and weak supervision approaches.
  - Some manual rules to enhance quality is a good investment!
  - Check values distribution, and any outliers.

- **Evaluation:**
  - Two-step evaluation process:
    1. Annotate benchmarks to iterate while model training.
    2. Evaluate a predictions sample when model is ready.
  - Update model based on benchmarks.
  - Post-processing rules when manual intervention is unavoidable.

Practical Tips

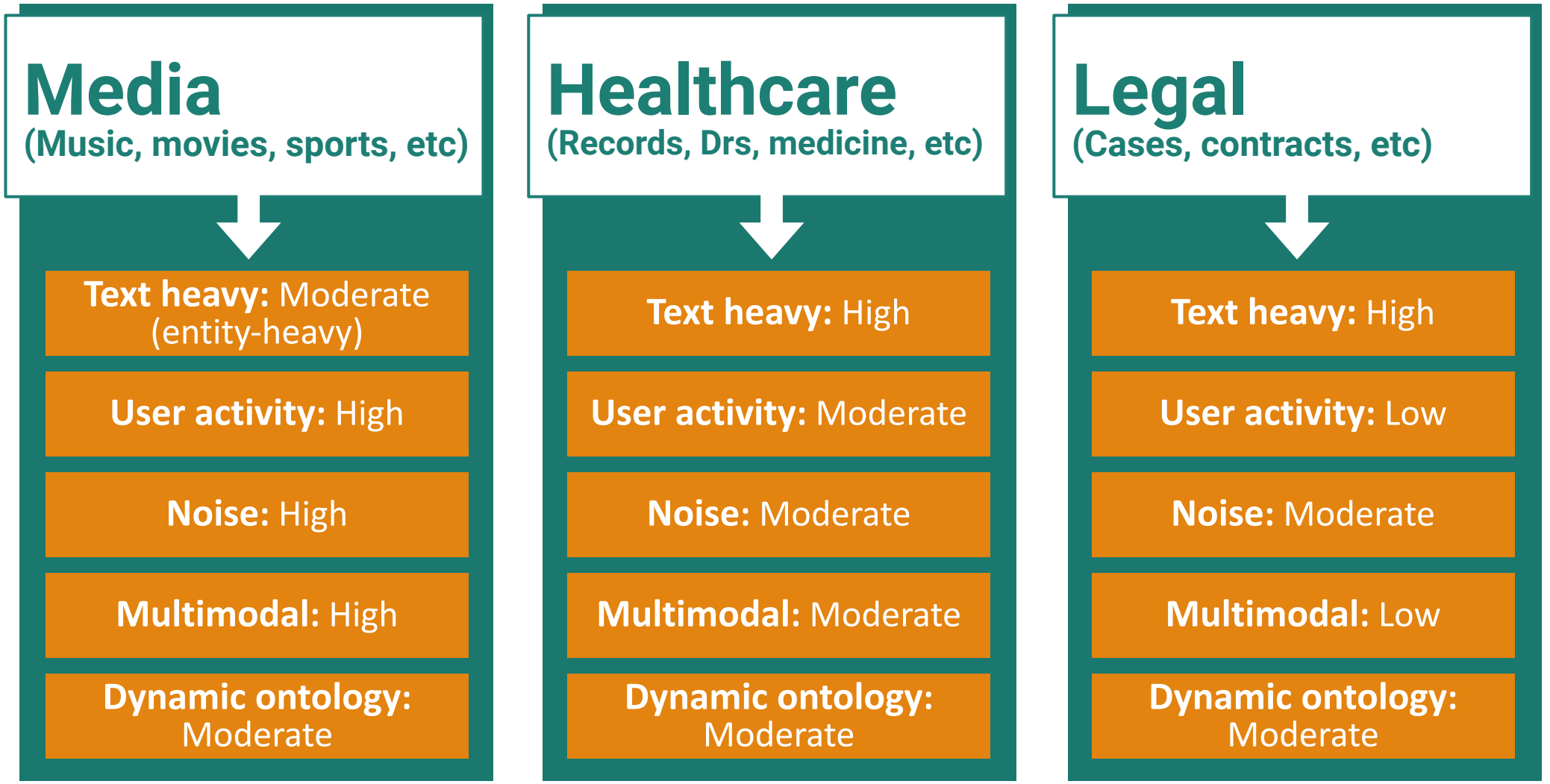
- **Modeling scope**
  - Categorical classification: When target space is closed and small, and when handling implicit values.
  - Textual extraction: In open-world cases, and when target values tend to be mentioned explicitly.

- **Prediction confidence**
  - We set thresholds based on prediction confidence to filter out predictions, and balance precision and recall

Practical Tips

- **Human in the loop**
  - We strive for scale and automation, while maintaining accuracy.
  - Achievable, through balancing automation and human input, at the right place.
  - Empower humans with the right tools and analytics tools.

Applicability to other Domains



Future Directions

We identified the following themes for future directions:

- **Training data:**
  - Make better use of unlabeled and seed datasets.
  - Enhance data quality through better data programming methods.

- **Ensembling and multitask methods**:
  - Ensemble data cleaning methods, syntactic, semantic, graph, etc.
  - Ensembling tagging and classification methods.
  - Taxonomy Enrichment and Relation Discovery **in one shot.**

Future Directions

- **Multi-modal/multi-source signals**:
  - Better handling of multi-modal extraction.
  - Better utilization of user logs, like search, co-purchase, etc.

- **Personalization**
  - Better embedding users, venders, brands, etc.
  - Better connection with customer behavior.

- **Connect private to public data**
  - Incorporate common sense knowledge like ConceptNet to clean the data.

Questions

Overview and Introduction

Knowledge Extraction

Knowledge Cleaning

Q&A

Break

Ontology Mining

Applications

Conclusion and Future Directions

Q&A **10 min**