INTRODUCTION TO NATURAL LANGUAGE PROCESSING

THEORY AND APPLICATION FOR ENGINEERING

Thurston Sexton

Knowledge Extraction and Application Project

Systems Integration Division

Engineering Laboratory

National Institute of Standards and Technology

U.S. Department of Commerce

DISCLAIMER

The use of any products described in any presentation does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that products are necessarily the best available for the purpose.

BACKGROUND: PROJECT/PROGRAM OVERVIEW

Knowledge Extraction *and* ***Application***

- Much of manufacturing know-how is computationally inaccessible, within informally-written documents
- Create human-centric data pipelines to extract value from existing unstructured data at minimal labor cost
- Develop guidelines for using semi-structured data in KPI creation, functional taxonomy prediction, and customized worker training paths

BACKGROUND: MAINTENANCE WORK-ORDER DATA

"HP coolant pressure at 75 psi"

"Hyd leak at saw atachment"

"Major hydraulic leak at Sp#6 horseshoe"

"Clamping spool guard broken"

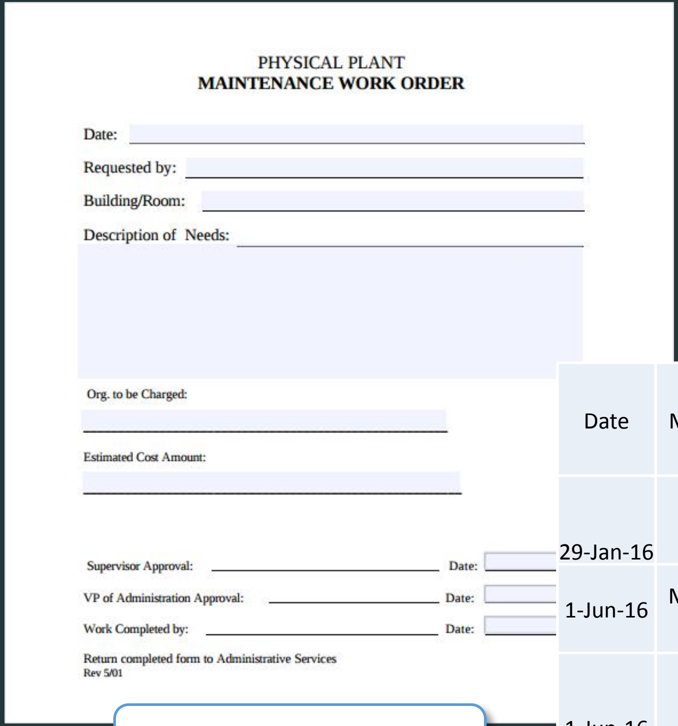
"Replaced seal in saw attachment but still leaking – Reapirs pending with ML"

"Bad Gauge / Low pressure lines cleaned ou"

"Replaced – Operator could have done this!"

"Repaired horseshoe seals"

BACKGROUND: CURRENT MWO DATA ENTRY



PHYSICAL PLANT
MAINTENANCE WORK ORDER

Date:

Requested by:

Building/Room:

Description of Needs:

Org. to be Charged:

Estimated Cost Amount:

Supervisor Approval: Date:

VP of Administration Approval: Date:

Work Completed by: Date:

Return completed form to Administrative Services
Rev 5/01

WORK ORDER FORMS

SPREADSHEETS

Date	Mach	Description	Issued By	Date Up	Maint Tech Assigned	Resolution
29-Jan-16	H15	St#14 tool detect INOP	JS	29-Nov-16	SA	Slug detector at station 14 not working. Would not recognize “Start” signal.
1-Jun-16	Mitsu FT	Brakes worn -Not stopping when in gear	AB	28-Jun-16	Steve A	Repaired
1-Jun-16	H8	St#7 rotator collet broken -wait for Bob B to show him how to remove	JS	8-Jun-16	John Smith	Machine went offline on 6/8 -Mark removed and instructed Bob B on removal/install process

Do “AI” to it! (…?)

Natural Language Processing (et al.) as Engineering Tools

TODAY'S TALK: TAKE-HOME

- NLP "Theory" Basics
  a. Data **models** and engineering **assumptions**
  b. NLP "**Tasks**" and **approaches**
  c. **Metrics** and **Evaluation**
- Contextualize NLP techniques, paradigms
  a. How NLP concepts interface with "Engineering Practice"
  b. Continuous interaction between experts (domain ←→ NLP)

TODAY'S TALK: STRUCTURE

Engineering Practice

- Goal & Approach
- Assumptions
- Measure & Evaluate
- Validate

"State the methods followed and why."

"State your assumptions."

"Apply adequate factors of safety."

"Always get a second opinion."

Hutcheson, M. L. (2003). *Software testing fundamentals: Methods and metrics*. John Wiley & Sons.

TODAY’S TALK: STRUCTURE

Engineering Practice

- Goal & Approach — *“State the methods followed and why.”*
- Assumptions — *“State your assumptions.”* — Start Here
- Measure & Evaluate — *“Apply adequate factors of safety.”*
- Validate — *“Always get a second opinion.”*

Hutcheson, M. L. (2003). *Software testing fundamentals: Methods and metrics*. John Wiley & Sons.

ASSUMPTIONS

That turn “Natural Language” into something to “Process”

ASSUMPTIONS: RULE-BASED VS. NUMERICAL

Some very successful ways to “process” natural language involve **rules**.

Assume a language model based on known “logic”:

- Pattern Matching (e.g. regex), “coding”, etc.
- Clear definitions and transparent assumptions (iterate!)
- Can be **powerful** and **efficient**
- Can be **brittle** and **labor**-intensive

Newer techniques assume the text and its **statistical** properties **alone**

ASSUMPTIONS: THE CONTEXT SPECTRUM

- How do we turn text into "numbers"?
- Traditional techniques come in two "flavors"
  a. Bag-of-Words *(Global Frequency and Context)*
  b. Markov Model *(Local Sequence Probability)*
- Opposite answers to the question:

*"How much does **global** vs. **local** matter to you and/or this text?*

ASSUMPTION: GLOBAL FREQUENCY & CONTEXT

Basic Bag-of-Words

Words in similar ***contexts*** *<u>are</u>* ***similar****.*

- *Hydraulic leak at saw attachment*
- *Worn seal caused leak, replaced seal.*
- *Replaced saw, operator could have done this...*

	Hyd.	leak	saw	seal	rep.	...
Doc 1	1	1	1	0	0	...
Doc 2	0	1	0	2	1	...
Doc 3	0	0	1	0	0	...

- Remarkably Powerful
- Similarity is “vector directional”
  - Documents or Terms
  - → Cosine Similarity

ASSUMPTION: GLOBAL FREQUENCY & CONTEXT

Basic Bag-of-Words

Words in similar ***contexts*** *<u>are</u>* ***similar****.*

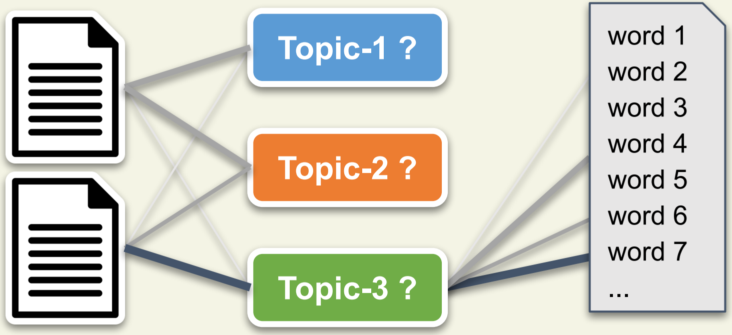
- *Hydraulic leak at saw attachment*
- *Worn seal caused leak, replaced seal.*
- *Replaced saw, operator could have done this...*

	Hyd.	leak	saw	seal	rep.	...
Doc 1	1	1	1	0	0	...
Doc 2	0	1	0	2	1	...
Doc 3	0	0	1	0	0	...

- Remarkably Powerful
- Similarity is "vector directional"
  - Documents or Terms
  - → Cosine Similarity

Modifications

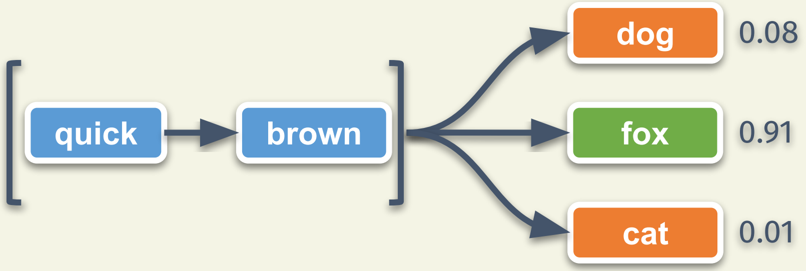
- Re-weighting schemes
  - Normalization, TF-IDF
  - Ties to informational entropy
- Dimension Reduction & **Topics**
  - Some "latent" set of topics:
    "Stuff we talk about" has less variety than "words we have"
  - Acronym soup
    PCA,SVD,LSA,NMF,LDA,TSNE,UMAP



ASSUMPTION: LOCAL SEQUENCE PROBABILITY

Markov Model

Next “states” (read: token/character) is conditionally dependent on the past:

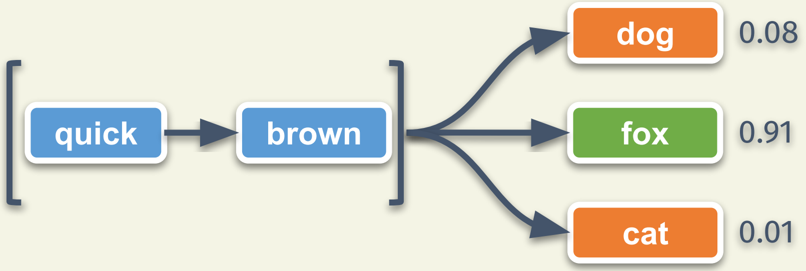


- Useful to generate text and estimate cond. probabilities
- High preference for observed sequences (precision)

ASSUMPTION: LOCAL SEQUENCE PROBABILITY

Markov Model

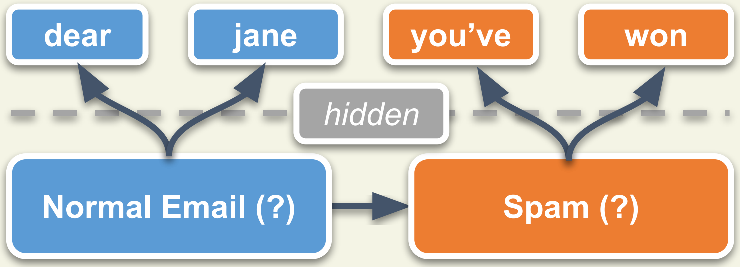
Next "states" (read: token/character) is conditionally dependent on the past:



- Useful to generate text and estimate cond. probabilities
- High preference for observed sequences (precision)

Hidden Markov Model

What we "observe" are emissions from a sequence of states we cannot observe.



- Used for last-gen. language models, bio-informatics, etc.
- Modular! See: GMMs, Bayes-nets…

ASSUMPTIONS: MODERN EMBEDDINGS

But... neural-nets?!

- We like the global context, but also want local sensitivity...
- Neural Nets can be "trained" to find a **vector space** model that **balances** both
  a. **Trained** is the operative term
  b. Packages/tools that let us "embed" text have **already trained** on a textual corpus
- You are assuming your text is "like" *that* text

*Otherwise these are an **approach**—and require proper design!*

ASSUMPTIONS: MORE ON "MODERN EMBEDDINGS"

- *Word2Vec (2013)* trains on a *word*-level
  - Continuous Bag-of-Words (**CBOW**): target word from local context
  - **Skip-Gram**: local context from target word
  - Maintains semantic linearity ("word algebra") — also see GloVe (2014)

lunch + night - day → dinner

better - good + bad → worse

wine + barley - grapes → beer

coffee - drink + snack = pastry

ASSUMPTIONS: MORE ON "MODERN EMBEDDINGS"

- *Word2Vec (2013)* trains on a *word*-level
  - Continuous Bag-of-Words (**CBOW**): target word from local context
  - **Skip-Gram**: local context from target word
  - Maintains semantic linearity ("word algebra") — also see GloVe (2014)

lunch + night - day → dinner

better - good + bad → worse

wine + barley - grapes → beer

coffee - drink + snack → pastry

- *BERT (2018)* is a *sub-word* model...**context** (sentence) dependent!
  - Can capture separate semantic meaning (homophones) and out-of-vocab.
  - State-of-the-art in 2019; used for your Google searches.

GOALS & APPROACH

NLP Tasks and “The Pipeline”

GOALS & APPROACHES: OVERVIEW

- Typical NLP Tasks
  (and their image-processing relatives)
  a. Document Grouping, Classification
  b. Keyword Extraction, Multi-Label Classification
  c. Named Entity Recognition and Parts-of-Speech
- The NLP "Pipeline"
  a. Preprocessing
  b. Analyses

GOAL: DOCUMENT TYPING

- Clustering (Unsupervised)
  - Detect “natural groupings” for analysts to parse
  - Also: interpreting topic models
  - May or may not be relevant, but a useful tool

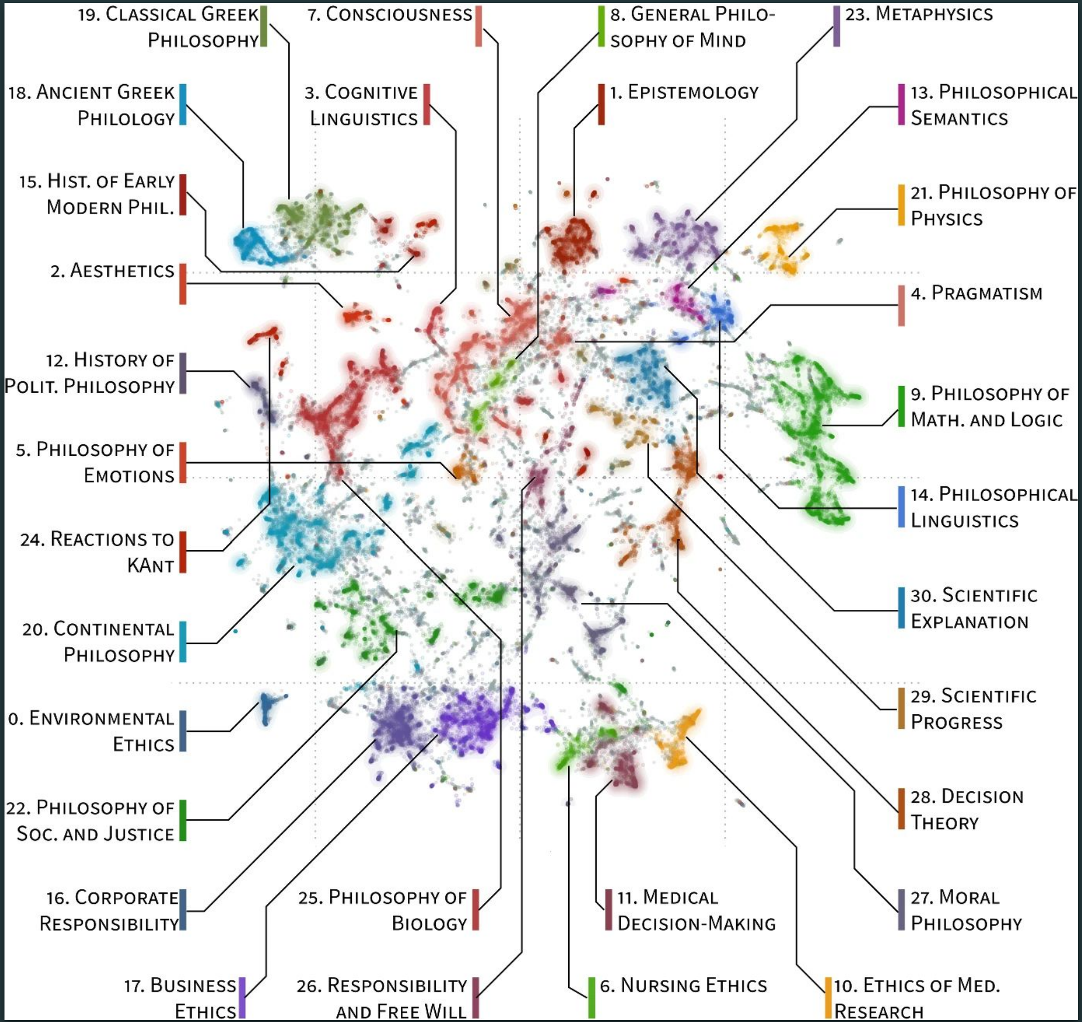
The Structure of Recent Philosophy

Noichl, M. Modeling the structure of recent philosophy. *Synthese* **198**, 5089–5100 (2021). https://doi.org/10.1007/s11229-019-02390-8

Each "dot" is a paper.

- Embed to 2-dimensions (UMAP)
- Clustering (HDBScan)
- Interpret, synthesize (hard)

Fully interactive online:
https://homepage.univie.ac.at/maximilian.noichl/full/zoom_final/index.html



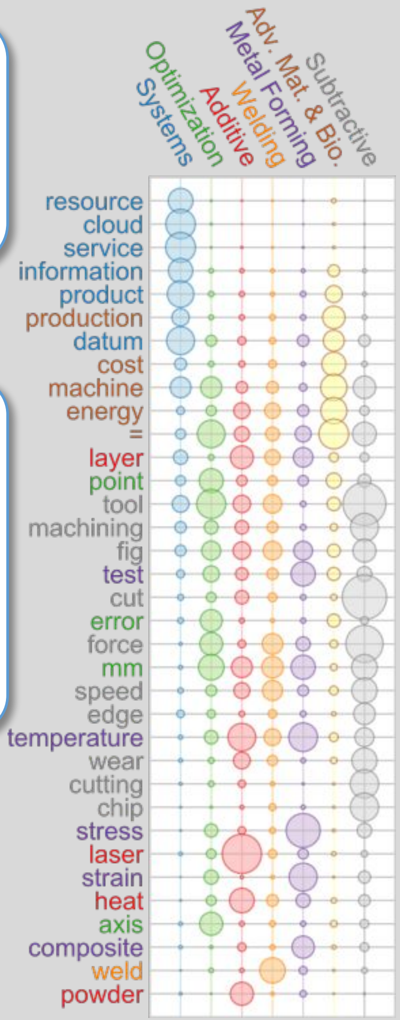
MSEC: A Quantitative Retrospective

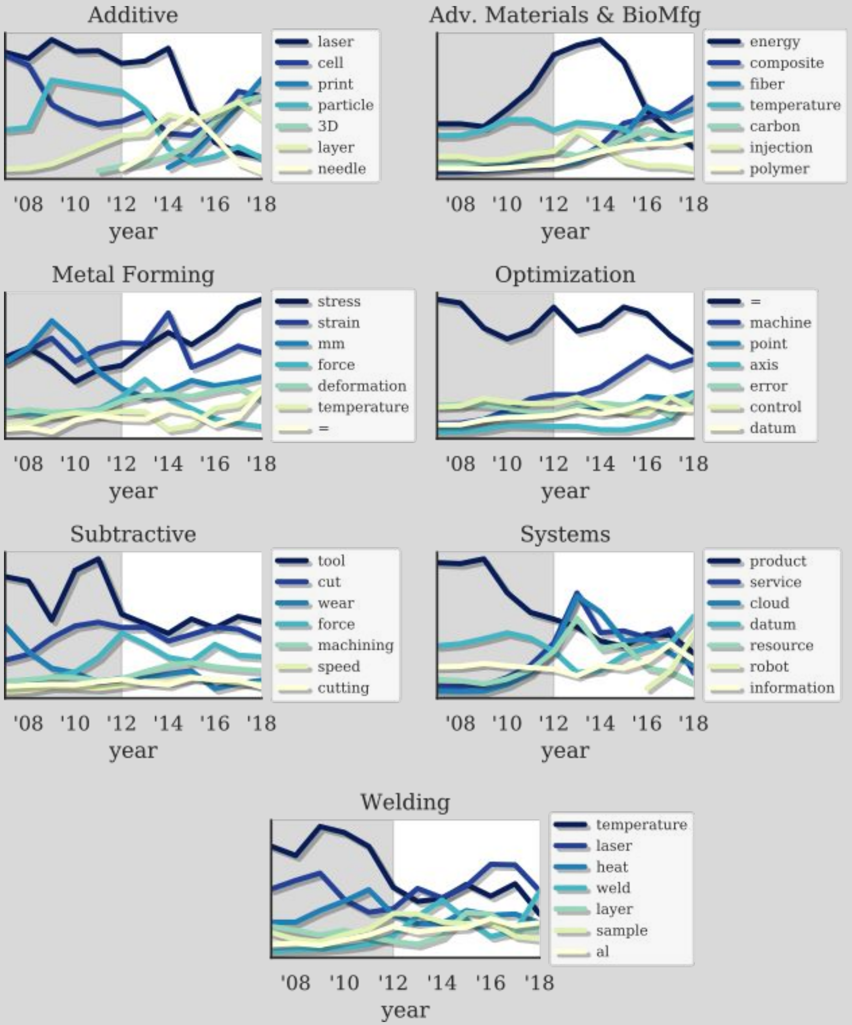
Sexton, T, Brundage, MP, Dima, A, & Sharp, M. "MSEC: A Quantitative Retrospective." September 2020
https://doi.org/10.1115/MSEC2020-8440

Topic Models as an approach to typing:

- Useful understanding
- LDA for static
- Dynamic LDA over time

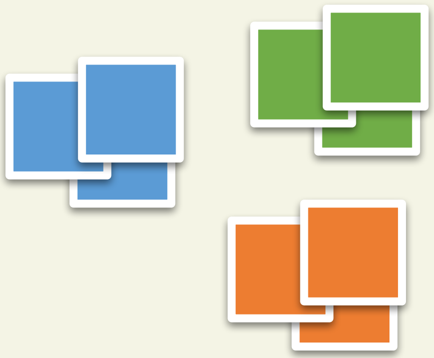
We had to name the topics.



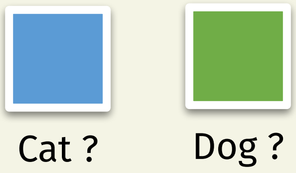


GOAL: DOCUMENT TYPING

- Clustering (Unsupervised)
  - Detect “natural groupings” for analysts to parse
  - Also: interpreting topic models
  - May or may not be relevant, but a useful tool



- Classification (Supervised)
  - Labels required: 1 per category (mutually exclusive)
  - Can be useful for recommendations: “relevant vs. not”
  - Images: “*is this a stoplight?*” or “*which animal?*”, etc.



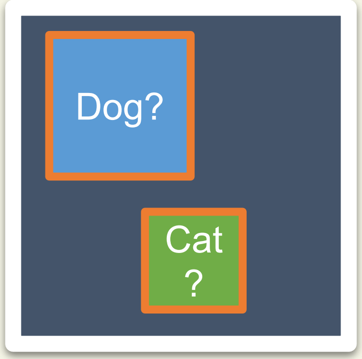
GOAL: DOCUMENT KEYWORDS

- Keyword Extraction (Unsupervised)
  - Use statistical properties to find “important terms”
  - Also see: text summarization
  - TF-IDF (sum), TextRank (graph-based), YAKE, +more

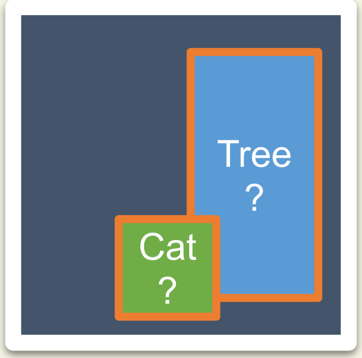
- Multi-Label Classification (Supervised)
  - Labels required: **multiple**-per-document (multiset)
  - Several ways to train, can use domain-knowledge
  - **Harder** problem, but maybe easier to **make** training data...
  - Images: “*What animals are present?*”

GOAL: ENTITY RECOGNITION

- Named Entity Recognition
  - Find **text spans** that contain **keywords**, and **annotate** them
  - Predetermined vocabulary/taxonomy (usually 2-levels)
  - E.g. “I went to New York [LOC]” or “They owe me $25 [CURR]”
  - Images: *“highlight and label the animals…”*



- Parts-of-Speech
  - Automatic determination of **grammar** information
  - SVO triples, dependency parsing, etc.
  - Can be used to “mine” **knowledge graphs**
  - Domain/language-dependent… hard with technical text!



GOALS: OTHERS WORTH MENTIONING

Wide variety of other tasks:

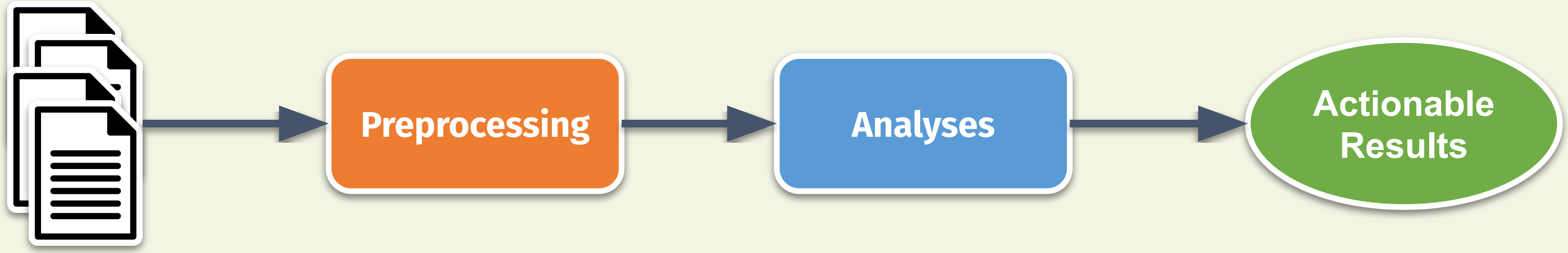
- Sentiment Analysis
- Seq2Seq & Machine Translation
- Reading complexity and writing quality, inclusivity
- Question Answering
- Text Synthesis

What does it take to get to this point?

PROCESS: "THE PIPELINE"

In theory, the NLP Pipeline is a

- Sequential progression, that
- Provides usable insight



Impossible to outline the number of variations on this "theme"... Here's:

- A common sequence — a *day-in-the-life* of your analyst.
- Benefits and drawbacks of each step

PROCESS: TEXT PREPROCESSING

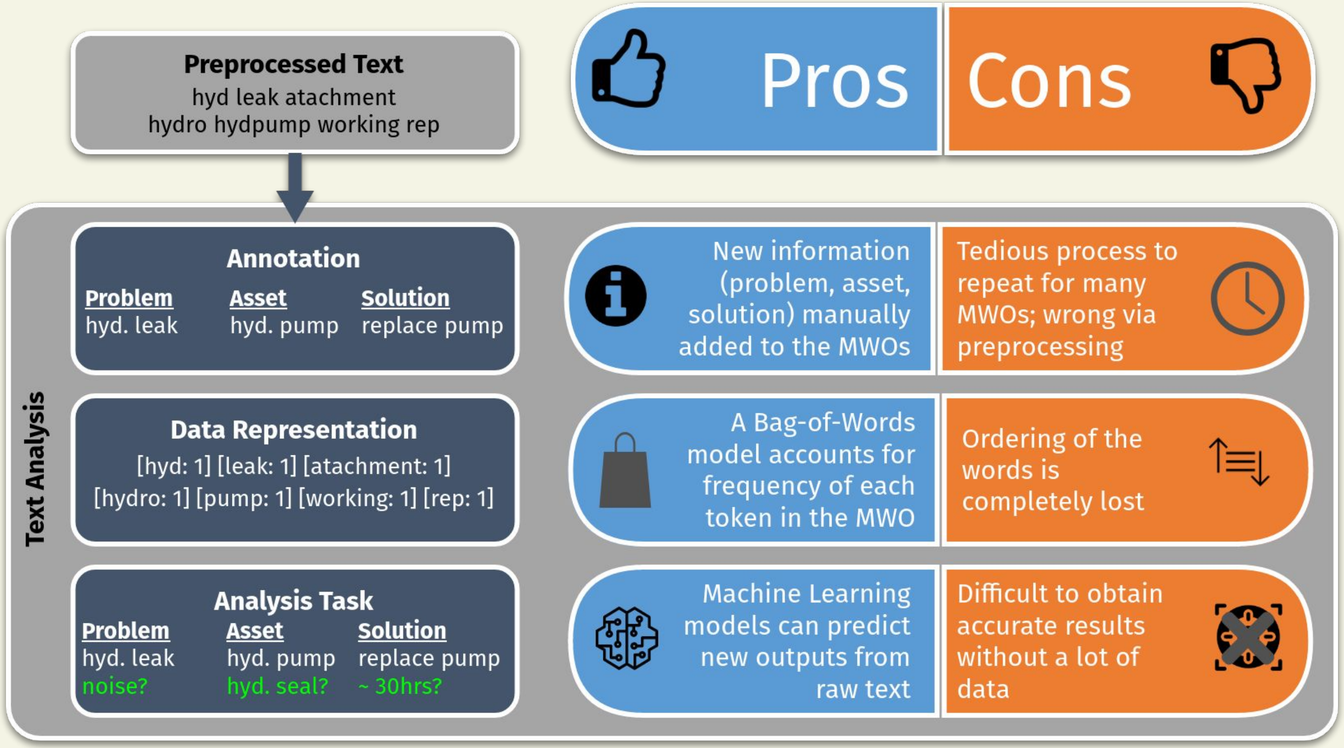
Raw Text
Hyd leak at saw atachment Hydromat Saw 012, hydpump not working, rep with new HS012

Preprocessing

	Pros	Cons
Tokenizer {hyd}{leak}{at}{saw}{atachment}{hydromat}{012}{,}{hydpump}{not}{working}{rep}{with}{new}{hs012}	Each word successfully broken into an individual token	Hydpump not tokenized correctly
Stop Word Remover {hyd}{leak}{atachment}{hydromat}{012}{,}{hydpump}{working}{rep}{hs012}	All stop words removed	NOT, SAW, NEW have important meaning and were removed
Cleaner {hyd}{leak}{atachment}{hydromat}{hydpump}{working}{rep}	Punctuation removed	HS012 is an asset number; still removed
Stemmer hyd, hydraulic, hydromat, hyrdaulic → hyd, hydraul, hydro, hyda	Hydromat correctly linked to hydro	Hyd, Hydraulic, Hyrdaulic not linked

Technical language processing: Unlocking maintenance knowledge. Brundage, M. P., Sexton, T., Hodkiewicz, M., Dima, A., & Lukens, S. (2021). *Manufacturing Letters*, *27*, 42-46.
Image adapted from original.

PROCESS: TEXT ANALYSES



Technical language processing: Unlocking maintenance knowledge. Brundage, M. P., Sexton, T., Hodkiewicz, M., Dima, A., & Lukens, S. (2021). *Manufacturing Letters*, *27*, 42-46.
Image adapted from original.

MEASURE & EVALUATE

Importance of metrics and knowing what gets evaluated

MEASURE & EVALUATE: OVERVIEW

Key skill of the analyst or engineer is knowing how to **translate:**

Qualitative *needs and constraints* → ***Quantitative*** *metrics and evaluations*

- What do I want to measure?
  - Do **my assumptions** conflict with the measurement?
  - Do the **metric's assumptions** conflict with my goal/process?
  - Will **multiple metrics** provide a broader insight? (yes)
- What constitutes progress toward, or success in, my goal?
  - Have I encoded my (stakeholder) expectations (preferences) sufficiently?
  - Do I have parameters to tune (continuously and/or iteratively)?

Most important: have I **transparently documented** my decisions for **iteration**?

MEASURE

What do I need to measure? Have I “done my homework”?

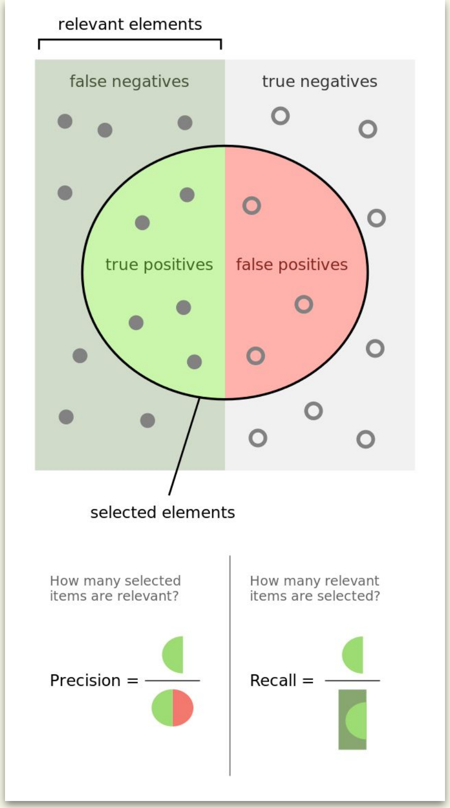
- Similarity or Distance
  - Discrete options, spellings: *Levenstein, Hamming, SymSpell, Jaccard*
  - Vector/Geometry: *Euclidean, Mahalanobis, Minkowski*
  - Distributions: *Kullback-Leibler, Earth-mover/Wasserstein, Cross-Entropy*
- Quality
  - Annotation coverage, label/class imbalance (rare-event?)
  - “Usefulness”: *topic perplexity, (B/A) Information Criterion*
  - Inter-rater agreement: *Fleiss’* κ, *Kendall’s* τ, *graph-based?*
- Importance
  - Information content: *Shannon Entropy, log-odds, lift, sum-TFIDF*
  - Centrality: *degree, betweenness, spectral (e.g. TextRank),*

EVALUATE: PRECISION & RECALL

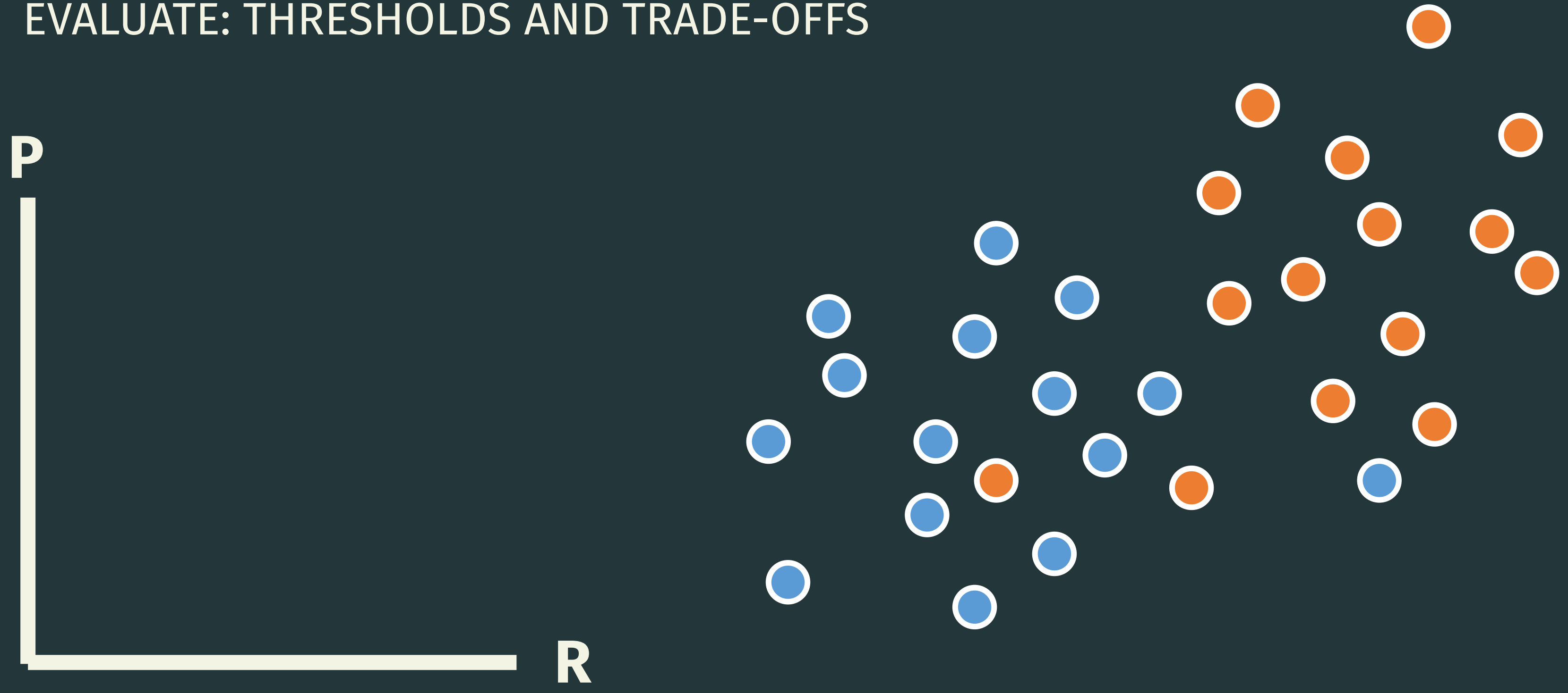
NLP often involves *multilabel* or *imbalanced* classification.

→ Accuracy is **unfair** or **overly optimistic**

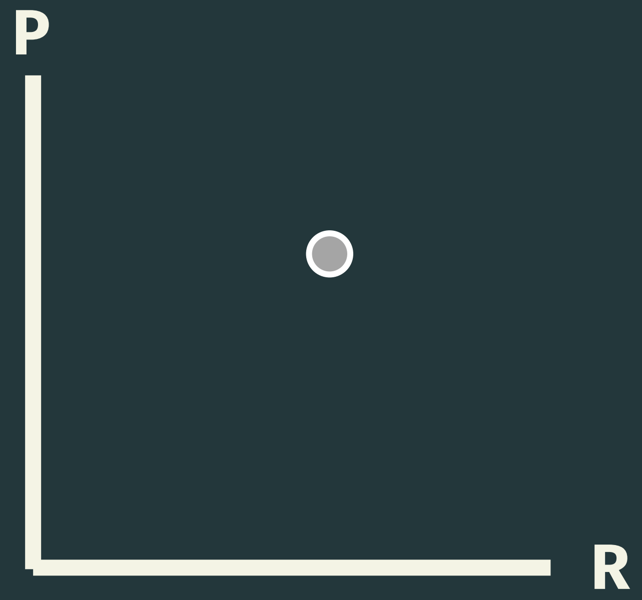
- Precision
  - Also *Positive Predictive Value (PPV):* [TP/(TP+FP)]
  - “Of things **predicted** X, how many **are** X?”
- Recall
  - Also *True Positive Rate* or *Sensitivity:* [TP/(TP+FN)]
  - “Of the things that **are** X, how many were **predicted** X?”
- F-Score
  - Harmonic mean of Precision & Recall:
  - Explicitly combines our preferences for the two
  - Parameter **β** (usually 1) : assign **β**-times more importance to Recall than precision.

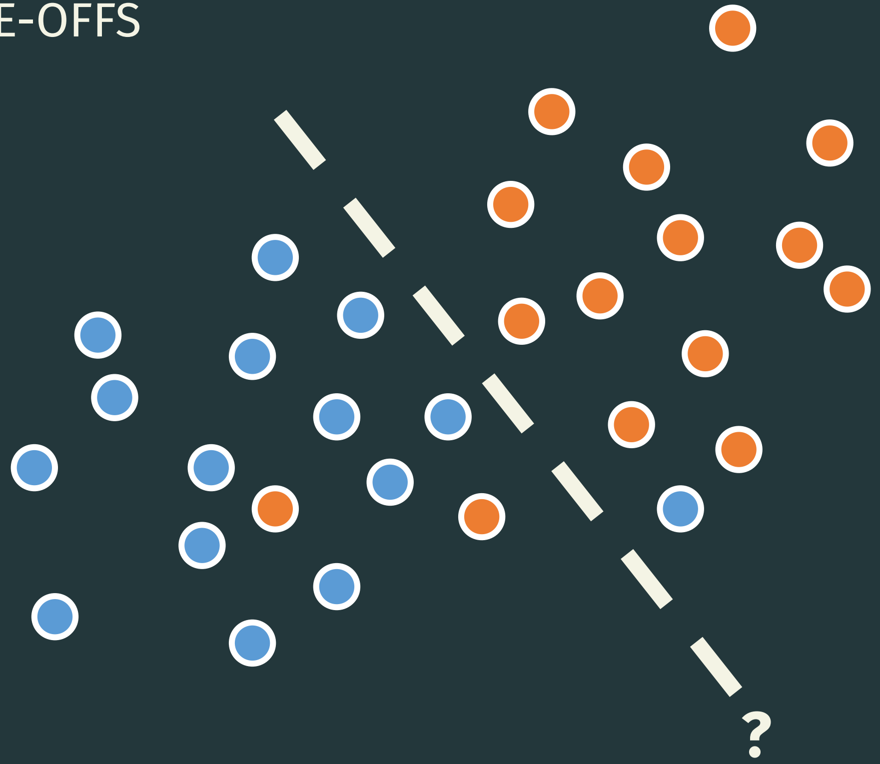


EVALUATE: THRESHOLDS AND TRADE-OFFS

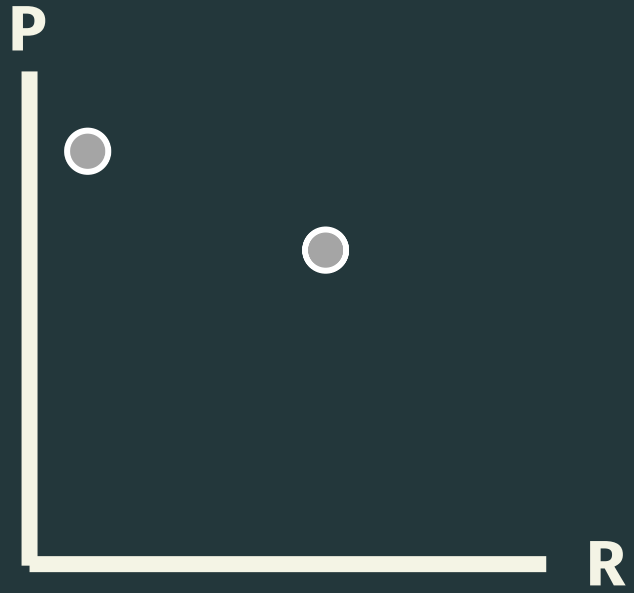


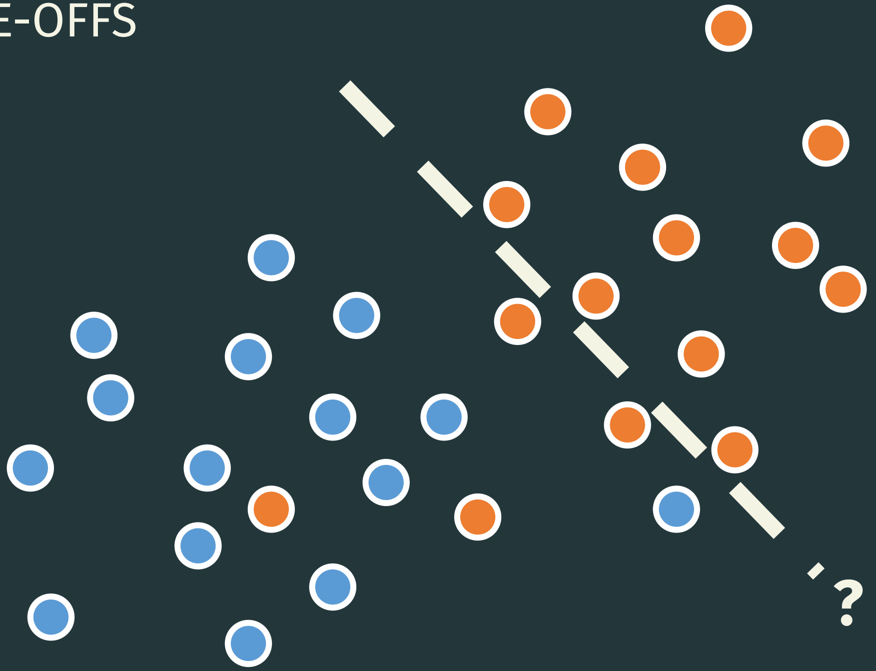
EVALUATE: THRESHOLDS AND TRADE-OFFS



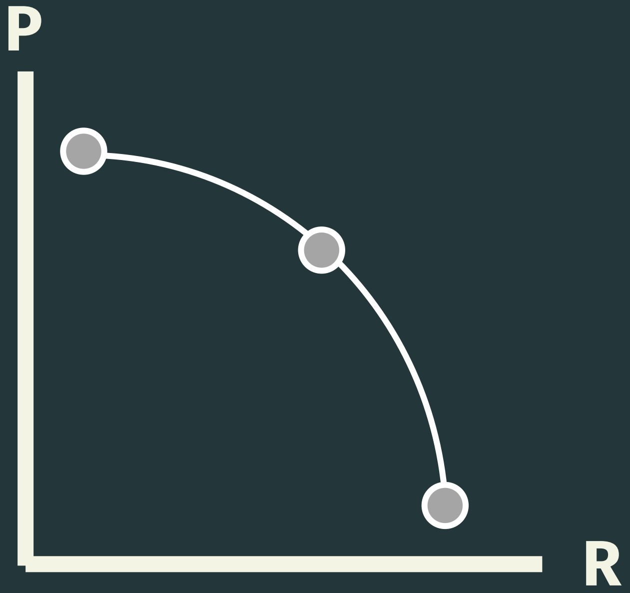


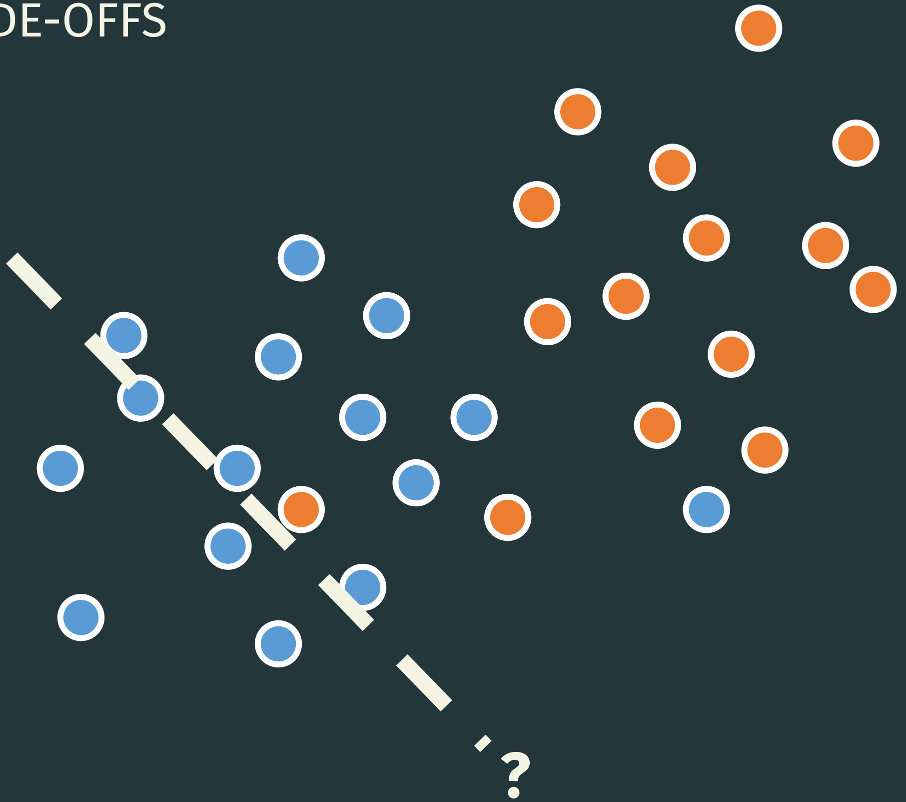
EVALUATE: THRESHOLDS AND TRADE-OFFS



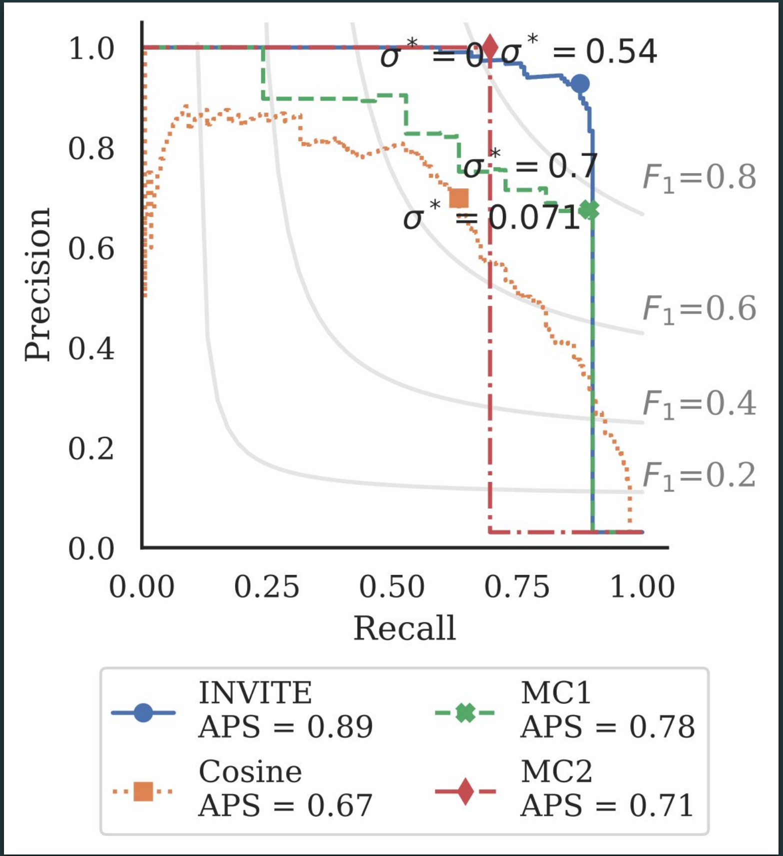


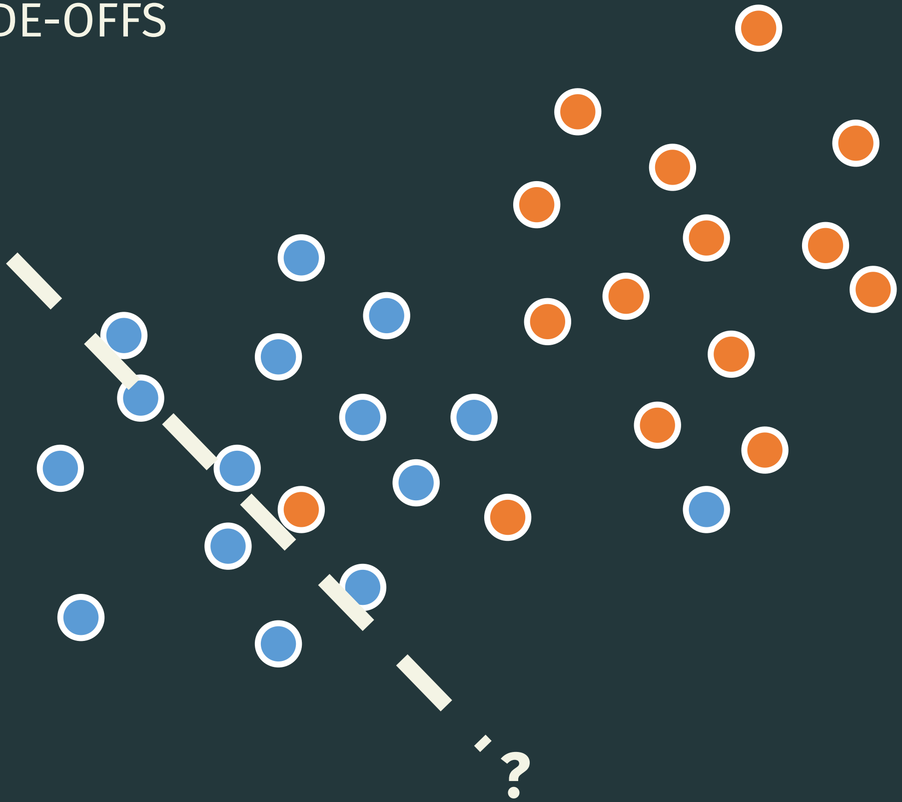
EVALUATE: THRESHOLDS AND TRADE-OFFS





EVALUATE: THRESHOLDS AND TRADE-OFFS





Sexton, T., and Fuge, M. (January 13, 2020). "Organizing Tagged Knowledge: Similarity Measures and Semantic Fluency in Structure Mining." ASME. *J. Mech. Des.* March 2020; 142(3): 031111. https://doi.org/10.1115/1.4045686

EVALUATE: SUMMARY

Do your **homework**

If there's something you want to measure, a metric may exist.

Metrics **evaluate**

Use fundamentals to design metrics that assess what matters.

Metrics **communicate**

Confusion is never the answer; strive for mutual understanding.

Remember that NLP is working on data *for humans, by humans.*

*Be **transparent** and **reproducible**.*

VALIDATION

The “open problem” of human-in-the-loop, domain-specific NLP

VALIDATION: PROBLEMS

So far we have glossed over some very common problems:

- Interpreting topic models can be fraught [1]
- Out-of-the-box tools are pre-trained on *very* different text
- There is not enough data to train custom models
- Too hard to hand-annotate the data we have
- No existing standard annotation to apply, no ontology we agree on
- Events of interest are far too rare (unclear if over-sampling applies)
- ...

In most Engineering Design and Reliability tasks, we *validate:*

Sanity checks, second opinions, processes for oversight and collaboration

[1] *Chang, Jonathan, et al.*
"Reading tea leaves: How humans interpret topic models."
Neural information processing systems. Vol. 22. 2009.

VALIDATION: RE-ASSESSING “THE PIPELINE”

Reality is never as clean as “The Pipeline”.

> *“In practice, the line between input and output are not well defined. An analyst might use intermediary tasks and representations to enrich annotations and cascade into further tasks. A holistic approach to improving one component will inevitably improve the others; a stolid adherence to a given pipeline can prevent progress all-around.*
>
> *[...]*
>
> *By lowering barriers to entry for text analysis through the development of efficiency-boosting tools and a more human-centered annotation approach, engineers have a unique opportunity to simultaneously learn from other domains and improve on their processes. A new approach is needed to adapt NLP methods to industry use cases in a scalable and reproducible way.*[1]

→ View NLP as a socio-technical system rather than as an algorithmic pipeline.

[1]Brundage, Michael P., et al.
"Technical language processing: Unlocking maintenance knowledge."
Manufacturing Letters 27 (2020): 42-46.

VALIDATION: TECHNICAL LANGUAGE PROCESSING

Enter ***Technical Language Processing***

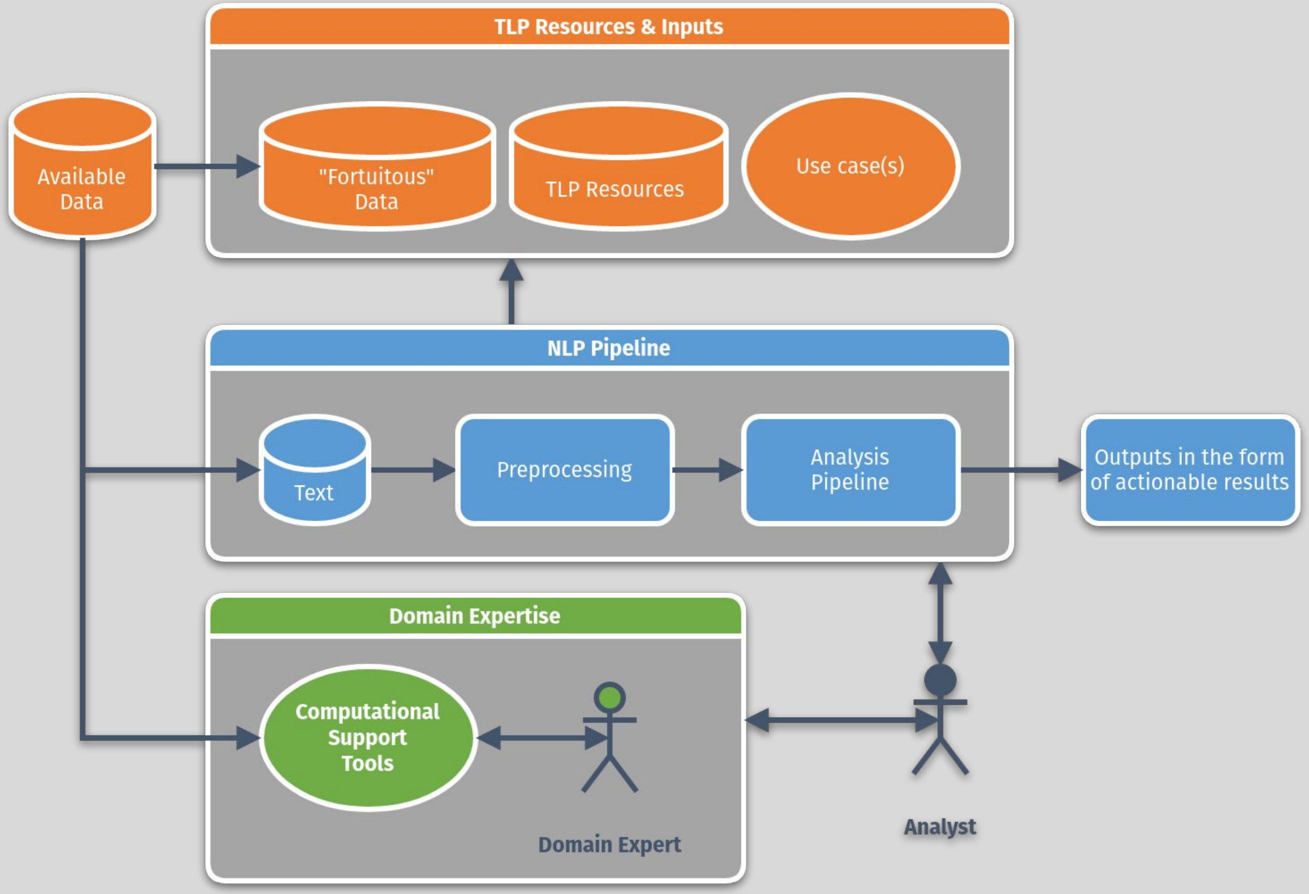
- NLP Techniques do not *always* adapt well to engineering text

- Current NLP solutions need to be adapted ***correctly*** for use in technical domains

- TLP is a methodology to tailor NLP solutions to engineering text and industry use cases in a scalable and reproducible way

Adapting Natural Language Processing for Technical Text
Dima, Alden, et al.
Applied AI Letters: e33.
Image adapted from original

- How the TLP approach to meaning and generalization differs from NLP
- How data quantity and quality can be addressed
- Potential risks of *not* adapting NLP



VALIDATION: GET INVOLVED

Plan for Distributed Collaboration in the TLP CoI

I. GitHub Organization (just started): **TLP-CoI**
   A. Documentation - best practices for TLP, theory, etc
   B. Networking - curated list for state-of-the-practice: awesome-tlp
   C. Collaboration - base or forks for open tool repositories

II. Events:
   A. Past Workshop (slides):
   B. TLP-COI Slack Workspace - QR code →
   C. Other options? Webinars? Let us know!

THANK YOU

Thurston Sexton

thurston.sexton@nist.gov

National Institute of
Standards and Technology
U.S. Department of Commerce