ROCKET:
RObust Concept and Knowledge Extraction from Text

Shi Yu, Glenn Fung
August 2016

Outlook

- Brief Intro to AmFam and SDA
- Background / Motivation
- Potential Business Value
- Relevant Data: Text in AmFam
- Overview of ROCKET
  - Active learning
  - Predictive model development
  - Modeling techniques
- Use-case: Internet search in the claim process

About American Family Insurance (AmFam)

Madison, Wisconsin – based American Family Insurance is the nation's third-largest mutual property/casualty insurance company and ranks 332th on the Fortune 500 list.

The company sells American Family-brand products, including auto, homeowners, life, business and farm/ranch insurance, through its exclusive agents in 19 states.

American Family affiliates (The General, Homesite and AssureStart) also provide options for consumers who want to manage their insurance matters directly over the Internet or by phone.

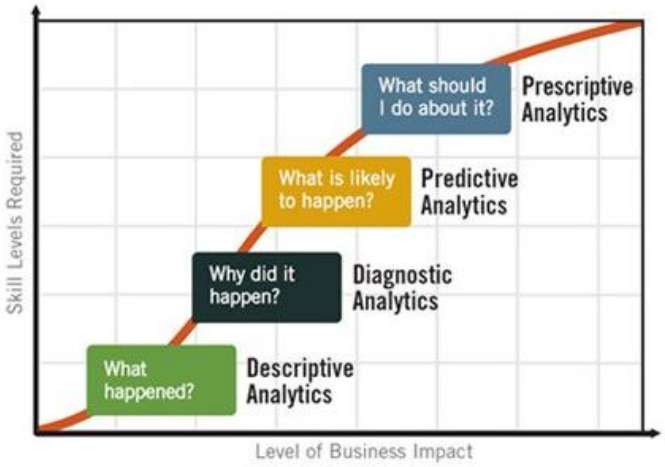
SDA – Providing Cutting Edge Analytics

In order to accelerate advancement of big data analytics capabilities and business transformation, AmFAM created a center of excellence for predictive analytics

Strategic Data & Analytics

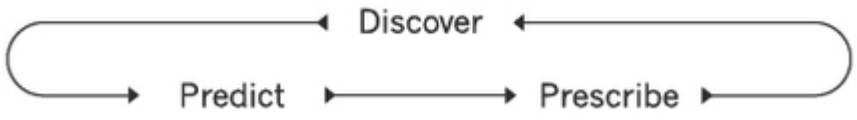
Vision

To be the advanced analytics organizational champion, supporting the organization in becoming a superior analytic competitor.



Mission

To create competitive advantage and economic value by bringing together data innovation, advanced analytics and business acumen to optimize or transform our business models.



Machine Learning inside SDA

Machine Learning (ML)

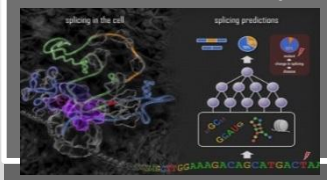
"Machine learning is the modern science of finding patterns and making predictions from data based on work in multivariate statistics, data mining, pattern recognition, and advanced/predictive analytics."

ML techniques are the core technologies behind :

Self-driving cars

Smart web search

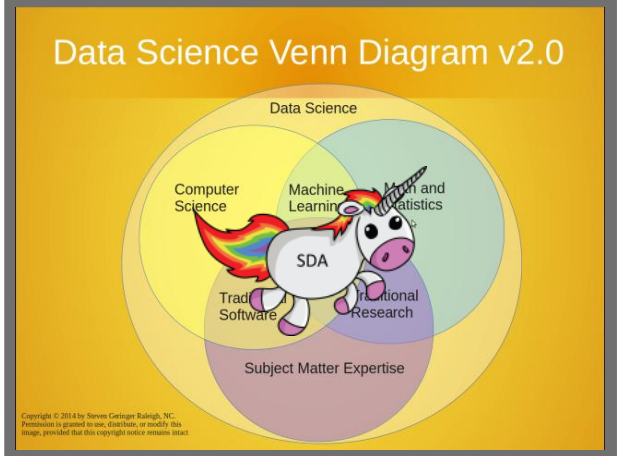
Genome discovery



Speech recognition

We SDA have a world-class team with the following (ML) expertise:

- Natural language processing
- Advance modeling: Random forests, SVMs, etc.
- Probabilistic graphical models and Bayesian techniques
- Deep learning
- Big data – parallel processing



Customer satisfaction, Subrogation, Total loss:

- Advanced modeling
- Text analytics
- Temporal mining

Recommender system for insurance :

- Collaborative filtering
- Bayesian networks
- Low-rank matrix factorizations

Text score for claim notes:

- Sentiment analysis
- Big data: 145MM notes
- Advance modelling (Multiple instance learning)

Customer Speech Sentiment analysis:

- Signal processing
- Speech-to-text conversion

Fraud detection, Roof classification:

- Image processing
- Deep learning

Analytics Applications in Insurance

With the complex nature of insurance risk exposures, and deep customer service interactions, insurance offers rich opportunities for advanced analytics

Customer & Marketing	Risk	Operations	Financial
• Customer experience & behavior • Propensity • Product Affinity • Retention & conversion • Lifetime value	• Pricing • Underwriting • Catastrophe modeling • Claims severity & Fraud • Reserving • Product Design	• Sales, UW & Claims operational performance • HR – Employee Engagement attraction & retention	• Portfolio optimization • Financial modeling/ forecasting • Economic modeling

Knowledge Extraction from Text (I)

"A fundamental task in text analysis is to ascertain or infer that a piece of text (a sentence, passage, or document) refers to a particular, given topic or concept."*

*Automated Identification of Medical Concepts and Assertions in Medical Text
Rómer Rosales, Faisal Farooq, Balaji Krishnapuram, Shipeng Yu, and Glenn Fung

Knowledge Extraction from Text (II)

- The topic can vary widely; for example we can assert if there is evidence that a party was interested in a rental car after reporting the loss:
  - *ERIC CALLED ME AND WANTS RENTAL…* can be accepted as a statement of the fact that ERIN (A party in the claim) is interested in a rental
- A similar task is to determine the polarity of the text; that is, whether the piece of text represents positive or negative evidence about the topic. For example:
  - *The insured reported a neck injury* clearly provides positive evidence about a neck injury, while *The insured does not seems to have any significant injuries* provides negative evidence about any injury type (neck injury in this case).

Examples of use-cases for AmFam

Use-cases for the insurance domain

- Relevance of a concept or question:
  - Is there evidence in this claim note that this person had a neck injury?
  - Is this claim related to wind or hail?
- Entity extraction from unstructured text
  - Places, people names, roles (insured, lawyer, claimant, etc.), dollar amounts
- To detect predefined relations among entities
  - Does this Phone number belong to this lawyer?
  - Is Pedro Jones a Doctor?

Business Value of the Technology (some examples)

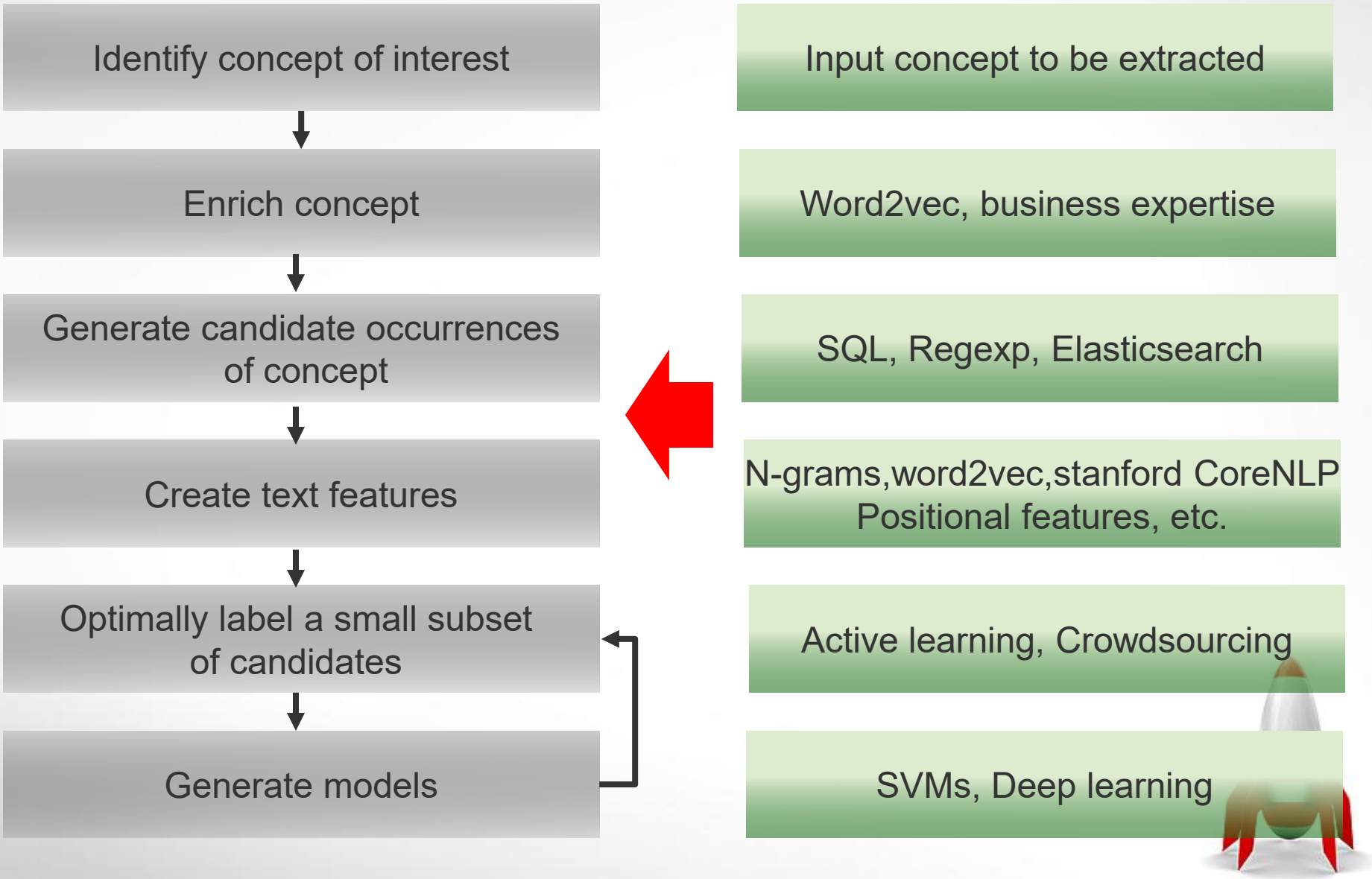
- Unstructured data sources processing and model-ready structural data preparation → Reduce intensive manual labor
- Complex sentiment and emotion analysis → Brand reputation monitoring
- To detect specific concerns expressed by customer in social media → Improve customer satisfaction
- Provide recommendations by detecting actionable events from unstructured data: calls transcripts, emails, notes → Better customer experience and improved information for Upsell / cross-sell

Description of relevant data: Text in AmFam

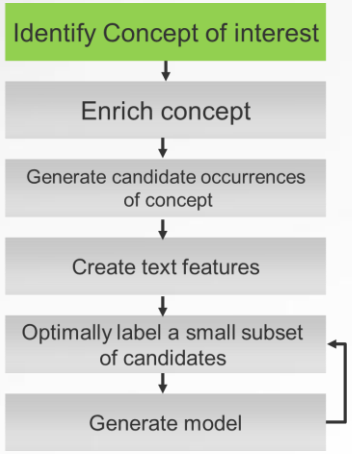
Unstructured data is ubiquitous inside AmFam:

- Claims notes (160MM+) notes available
- Touchpoint (claim survey) comments
- FNOL (first notice of loss) descriptions
- Social media feeds (Sysomos, Networked Insights)
- Call transcriptions (voci)
- Code, project charters, documentations, emails, etc.

Overview of the ROCKET framework



Define / Identify Concept

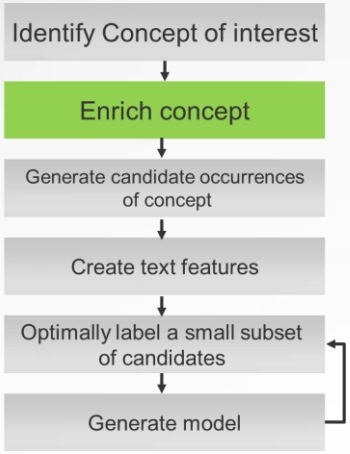


- This is the initial input to the framework
- Consists in defining the concept to be extracted
- We will use the following example for the rest of

The presentation: We want to find notes that are relevant to the concept of neck injury

Concept terms set = {neck,injury}

Using Word2vec to expand Concepts



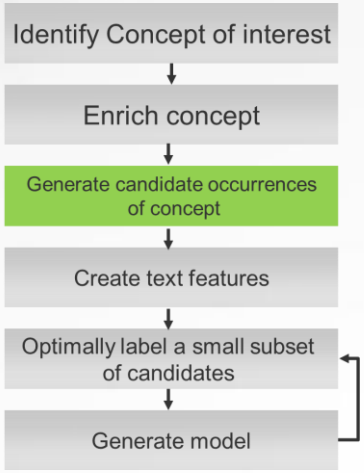
Concept terms set={neck, injury}

Word2vec, business expertise

Expanded concept terms set

{neck, injury, pain, lesion, cervix, whiplash, stiff, spine}

Candidate Generation



160MM claim notes

SQL,Regexp,Elasticsearch

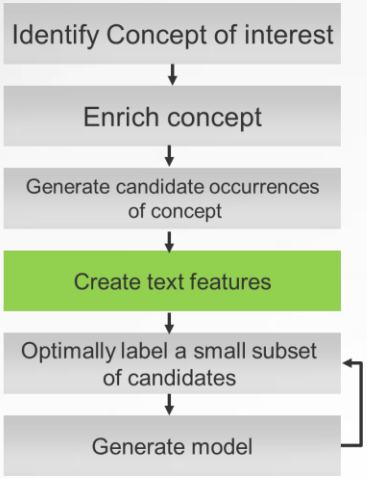
10K notes that contain the concept terms

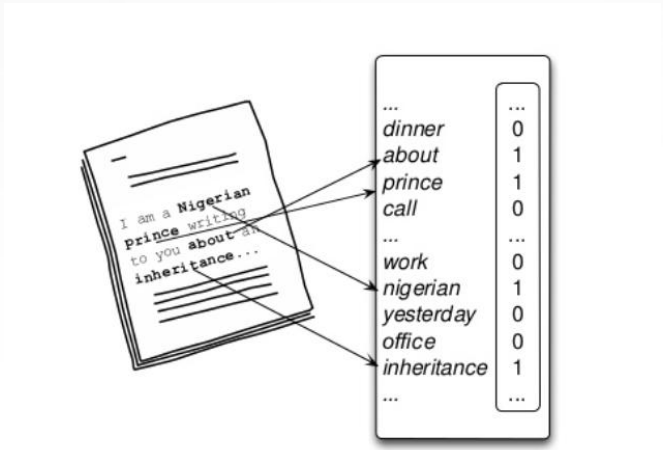
{neck, injury, pain, lesion, cervix, whiplash, stiff, spine}

Note that not all notes are relevant to Neck injury!
Example:
"The insured say that the Last time he saw the collar was in his wife neck"

Here we want maximum Recall!

Using NLP for Feature Generation



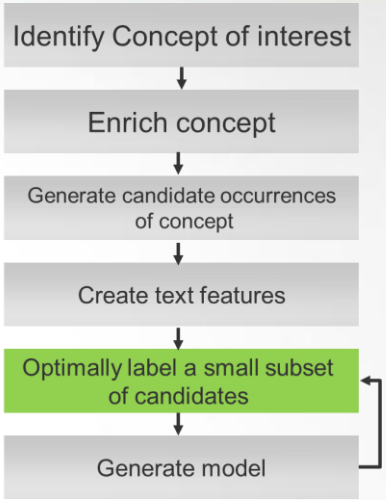


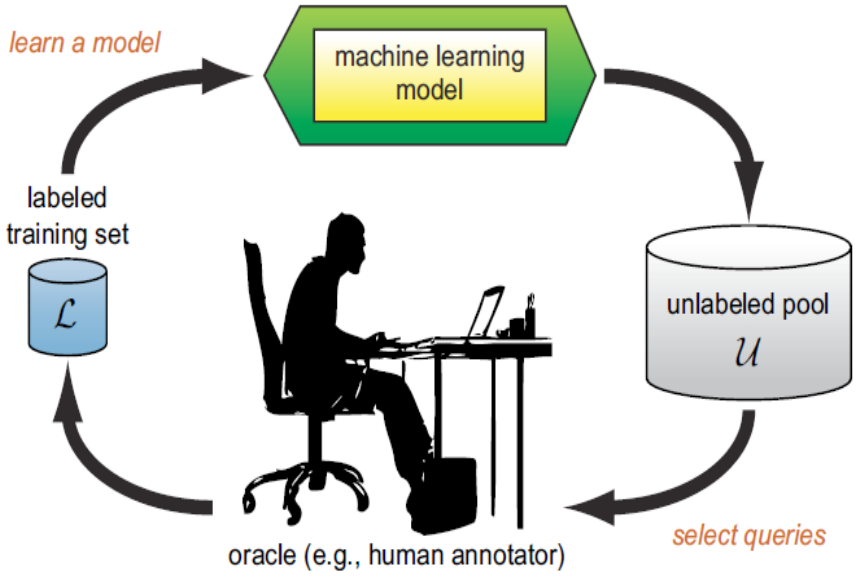
In this step we use NLP different techniques to represent each one of the 10k candidates as a vector

- N-grams: the basic bag of word representation
- Word2vec: we can use this to map similar terms to the same vector position to improve generalization e.g.: policyholder → driver / enterprise
- NER to recognize people names, places, dates, etc.
- Spatial features to measure the relative position of every word to the concept terms.

Using Active Learning to optimally label data

- Difficult and time-consuming to label all the data manually
- Adaptive learning paradigm for optimal labelling
- Manually label first k text examples
- Learn model, score it on all the unlabeled candidates
- Pick the best next *k* candidates to label
- Repeat the two previous steps until performance converges





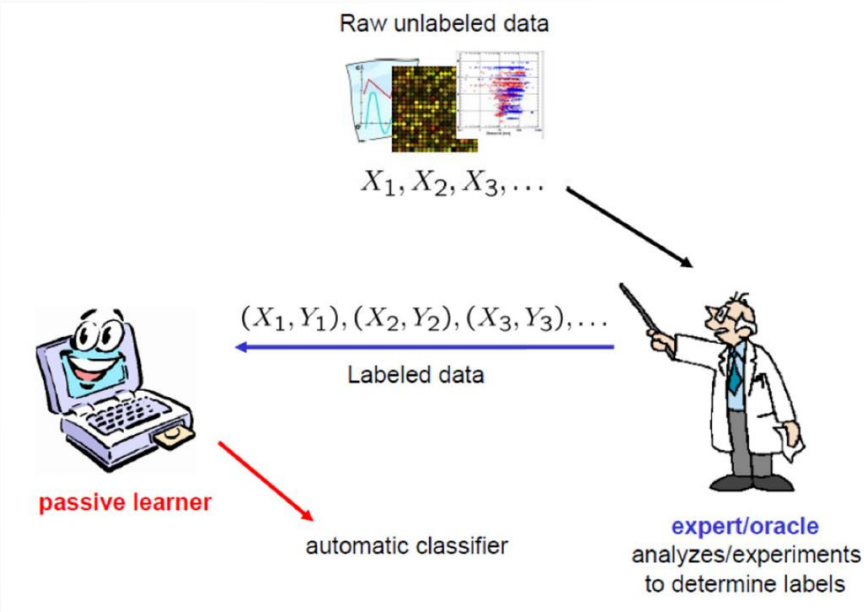
We will leverage the NEXT Crowdsourcing / Active Learning platform from the UW

Active Learning

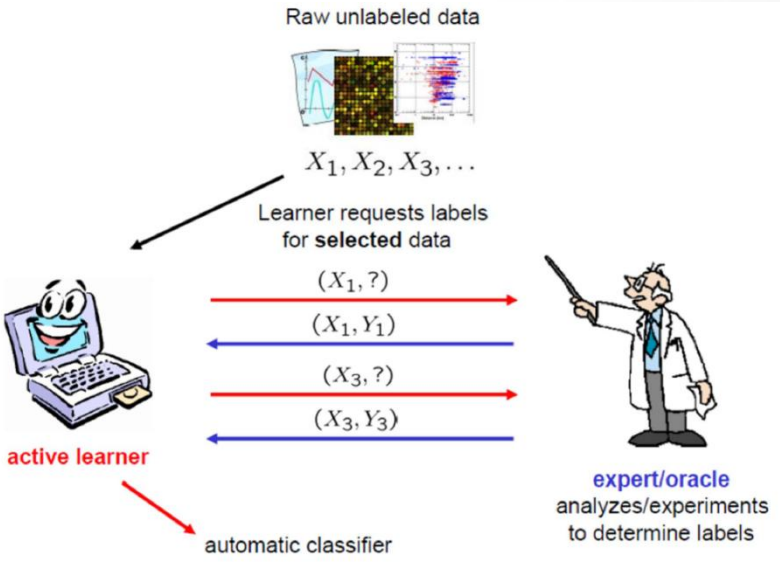
Instead of having annotators label all the training data, we would like to intelligently choose instances to be labeled -- called ***active learning.***

Passive Learning vs. Active Learning

Standard supervised learning



Active learning



Taken from Prof. Yi Zhan lectures

What is NEXT?

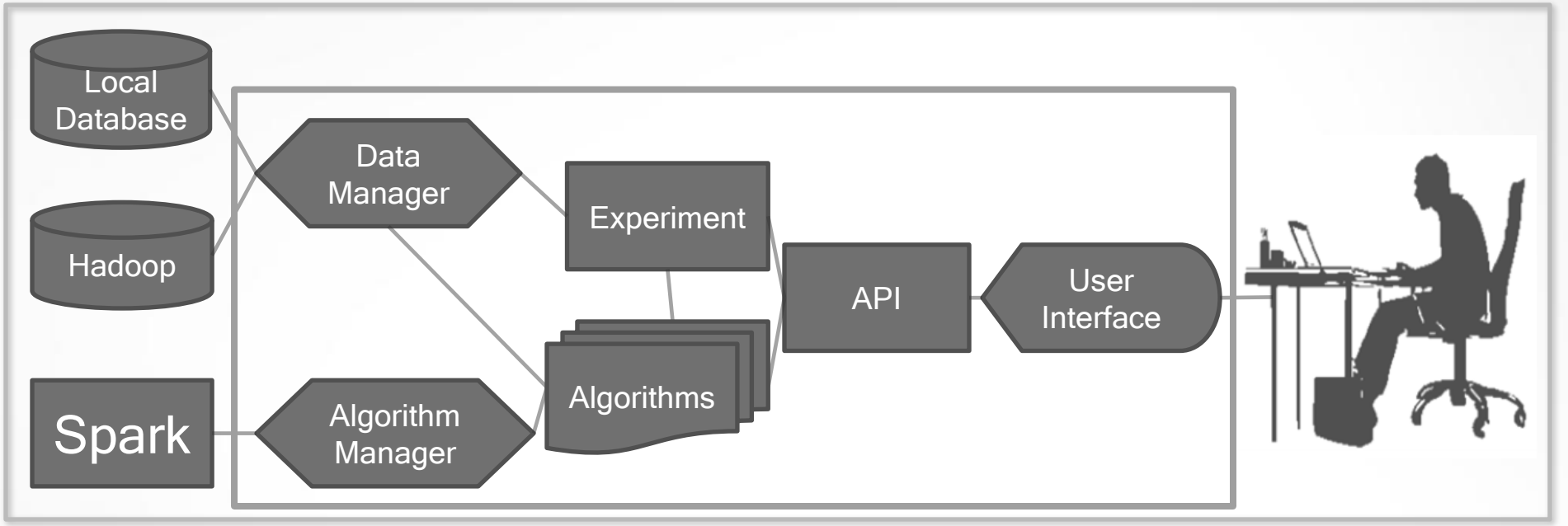
"NEXT is a machine learning system that runs in the cloud and makes it easy to develop, evaluate, and apply active learning in the real-world."

http://nextml.org/

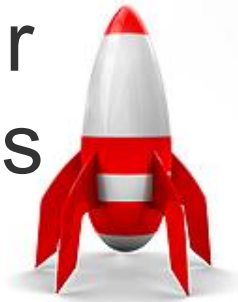
Source: https://github.com/nextml/next/

NEXT provides the architecture to enrich **machine-learning** with human feedback

NEXT

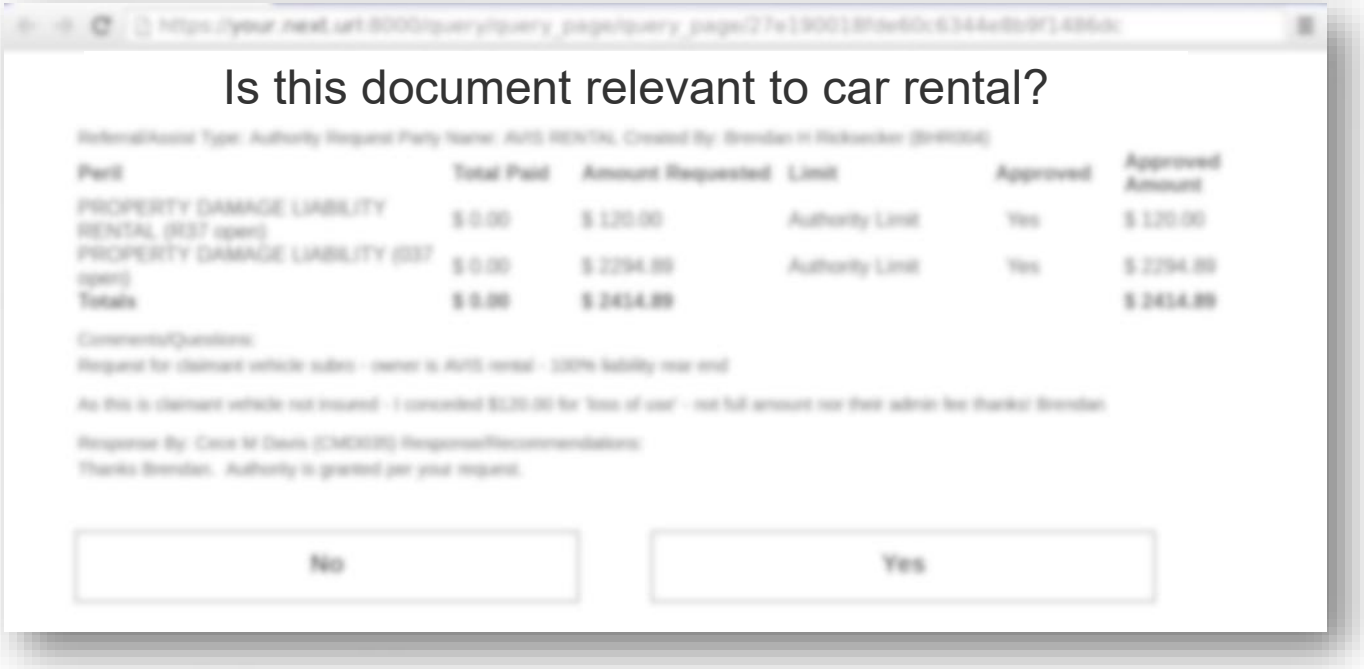


NEXT is open-source, integrates with our existing code, and can scale to our needs



NEXT

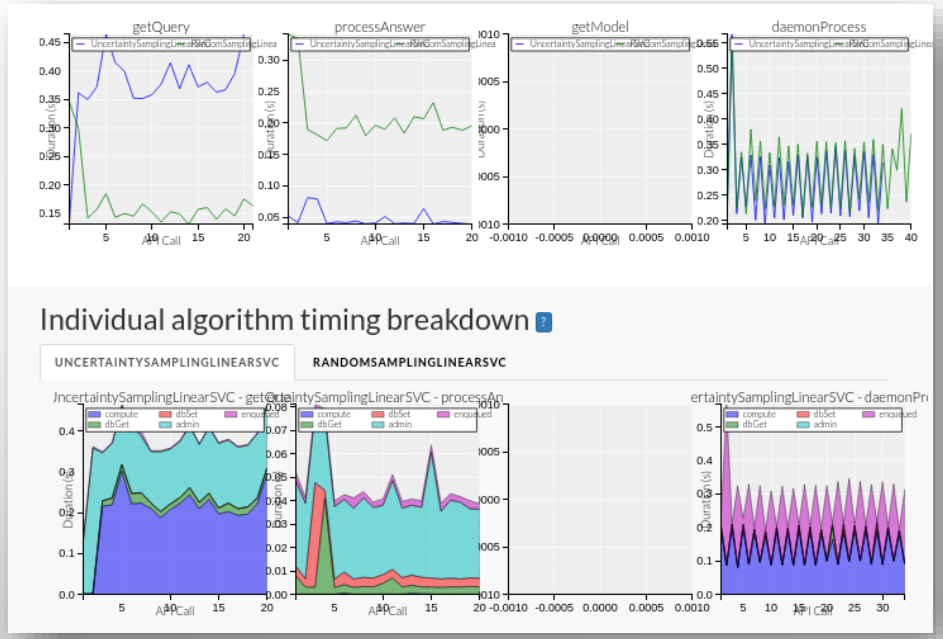
NEXT starts a server and provides a URL to send to users to **crowdsource** data:



NEXT automatically processes the answers and updates the algorithms

NEXT

NEXT provides real-time diagnostics and performance data via customizable dashboards

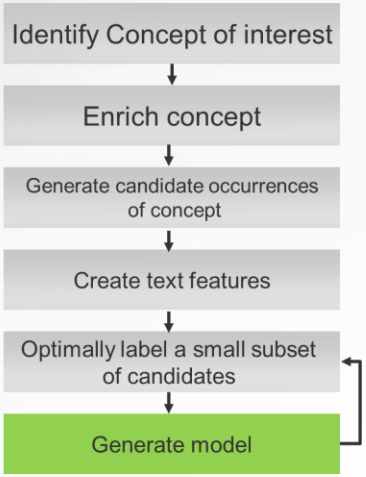


NEXT

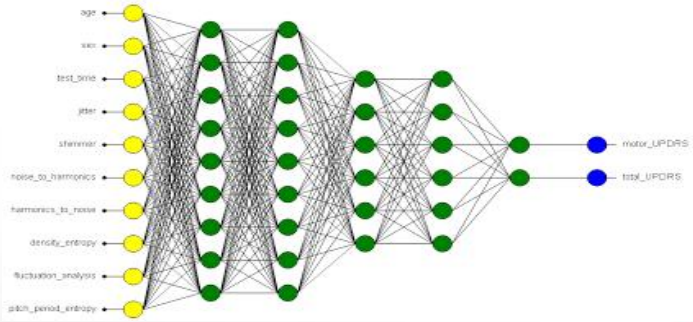
NEXT can be used anywhere where human feedback is needed in machine learning:

- Semi-supervised and supervised machine learning
- Text mining, image processing, speech recognition
- Acquiring labels for unlabeled data
- Running multiple algorithms in parallel to compare performance
- Evaluating and validating existing models

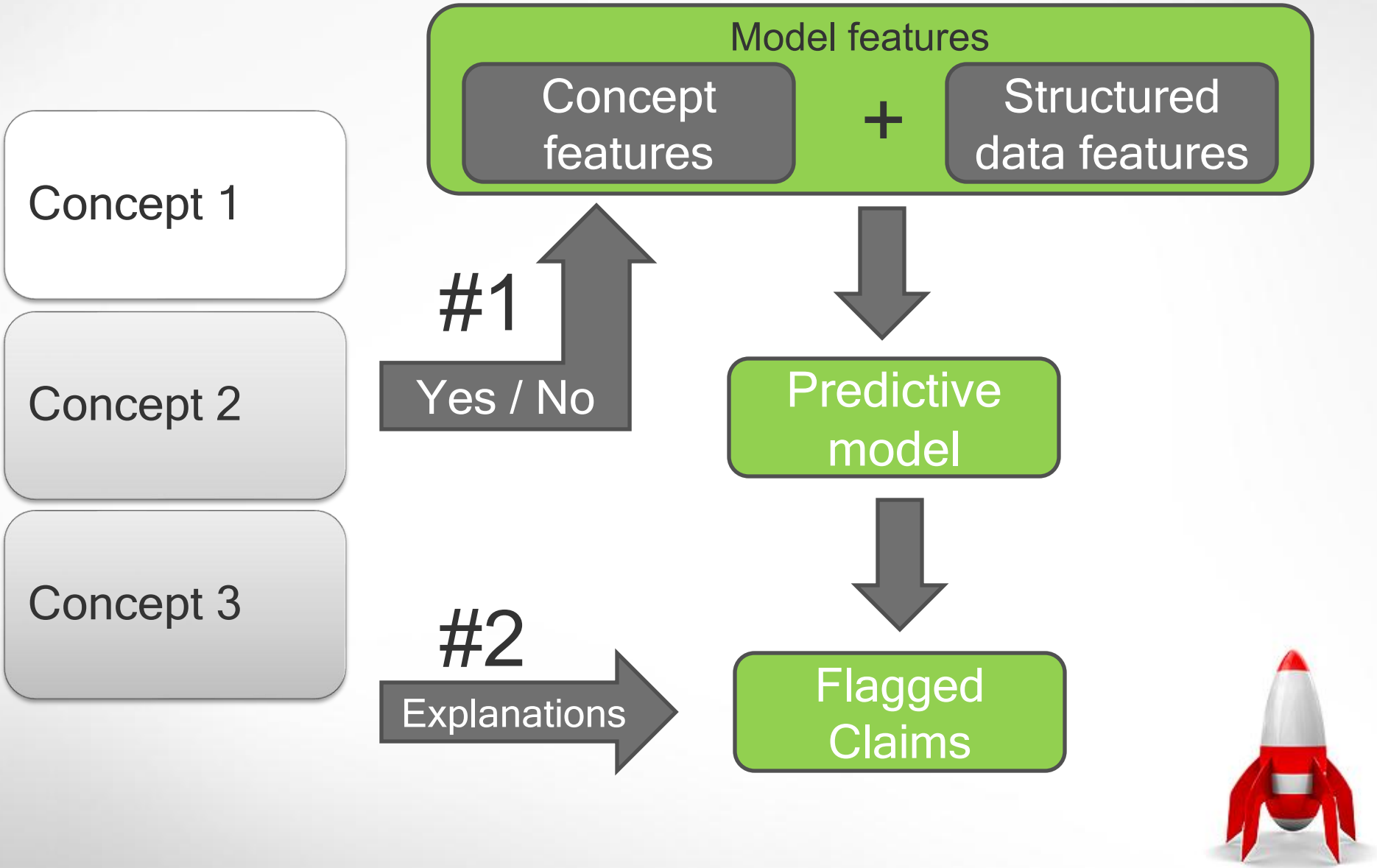
Classifiers



- For the active learning process it is preferable to consider models that:
  - Can be trained in a fast manner to reduce waiting time for human feedback (labeling)
  - Can be retrained in an incremental fashion since the labeled data set grows incrementally (updating)
- We consider several models (Logistic Regression, SVMs, LS-SVMS, Random forest, Bayesian approaches) for the AL loop.
- Once the AL loop converges a final model we will consider more complex models such as deep learning, semi-supervised approaches.



Use-case: Using claims-related concepts for insurance predictive models



Examples from the “Car Rental” concept

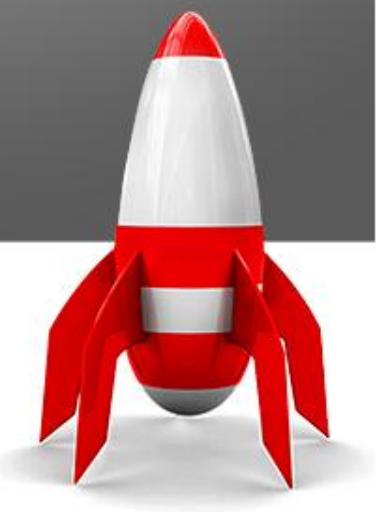
True negatives

Payment Instructions: 00/022025/$500 ded/driver side front and rear door/rental:30/750/no lien
CONTACT: Insured Luisa B reported fnol from (xxx) xxx-xxxx . Claim acknowledged Advised contact 1-2 business days FACTS OF LOSS: Three car accident insured was middle vehicle and was pushed into claimant vehicle in front. Insured explained she was stopped in traffic when claim behind hit her. REPAIR OPTIONS: Vehicle is located at Tow Yard sending PD adj APO insured plans to have claimant carrier handle repairs and rental RENTAL: N/A CLAIM PROCESS: Explained coverages with AMFAM. OTHER PERTINENT INFORMATION: You have the legal right to choose a repair shop to fix your vehicle. Your policy will cover the reasonable costs of repairing your vehicle to its pre-accident condition no matter where you have the repairs made
Insured called in claim from cell: XXX-XXX-XXXX (only number for insured). Insured pulled out in front of claimant in intersection. Date of Loss: 12/18/2010 @ 10:37am. Location: 15th Street And XXX Street - Hannibal MO. Hannibal Police report #XXXX-XXXX. No claimant info known. Insured driver: XXXX XXXXX (00) 00 driving the 2004 Buick Century Custom. Damage: passenger fender and front bumper headlight smashed in hood pushed over passenger front door does not open properly. Color: light brown. Mileage is approx. 78k. MO plate #XXXX. VEA: 10 points. Insured has estimates from Hannibal Paint & Body for $2898.32 and from Gibbon's B/S for $2603.80. Insured would like to use Gibbon's. Advised due to amount of damage PD inspection is needed. Explained process. Offered rental advised it is available during the repairs. Insured declined stating she is now nervous to drive and would rather not have rental. Advised if she changes her mind it will be available for her. Claims process explained.
Referral/Assist Type: General Party Name: XXXX XXX Created By: XXX A XXX (KXXXX) Comments/Questions: ERIC CALLED ME AND WANTS RENTAL FROM BAUM - WAS SICK AND SHOP TOLD HER NOT TO BE DRIVING HER VEHICLE '

True positives

Conclusions

- We have proposed a semi-automatic framework for robust concept extraction from text – ROCKET
- ROCKET can be extended with few effort to perform specialized knowledge extraction in lots of analytical applications
- Human interaction requires few or non training (depending on the concept to be learned)



BACK UP