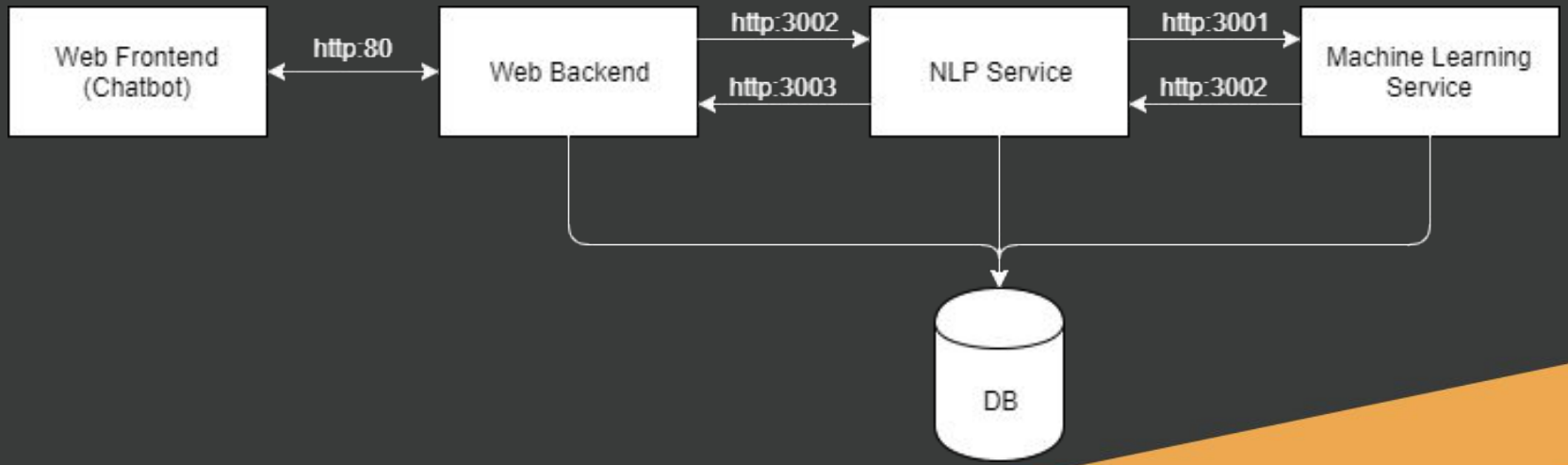
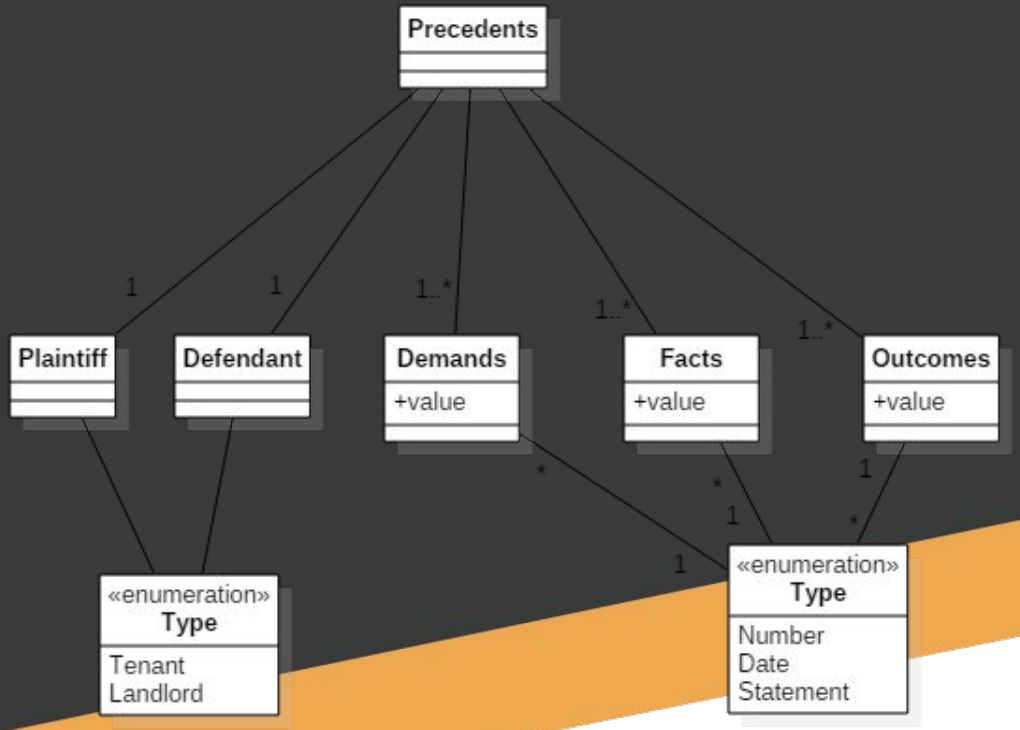
PROCEUS

AI Legal Chatbot

Overall Architecture



Dataset Used (Domain Model)



Dataset Used (Raw Input)

XXXXXXX
Locateur - Partie demanderesse
c.
XXXXXXX
Locataires - Partie défenderesse

D É C I S I O N

[1] Le locateur réclame 1 060 $ de loyer, plus 1 798,96 $ en dommages-intérêts.
[2] Il s'agit d'un bail du 1er juillet 2011 au 30 juin 2012 au loyer mensuel de 530 $.
[3] Le locateur déclare que les locataires ont quitté le logement en août 2012. Le logement fut reloué pour le 15 octobre 2012. Les locataires doivent 1 060 $, soit le loyer des mois de juillet, août 2012 plus 795 $ représentant une indemnité de loyer pour septembre et la moitié du mois d'octobre 2012 ainsi qu'un mois de loyer (530 $) gratuit pour faciliter le locateur pour un sous total de 2 385 $.
[4] Le locateur réclame également 40 $ pour la publicité, 69,13 $ pour l'électricité, 100 $ pour repeindre un mur orangé.
[5] Le tribunal conclut que le loyer est impayé et les dommages sont dus.

POUR CES MOTIFS, LE TRIBUNAL :
[6] CONDAMNE les locataires à payer au locateur la somme de 2 594,13 $ plus les intérêts au taux légal et l'indemnité additionnelle prévue à l'article 1619 du Code civil du Québec à compter du 8 novembre 2012, plus les frais judiciaires de 78 $.

Date de l'audience :
25 août 1888

Dataset Format

- Majority given in .doc format
- Easier to manipulate in XML
  - Convertible using openoffice conversion utility tool
- Simple parser can extract the fact/precedent/outcome sets based on [#] Prefix
- In our case, 45 000 precedents were converted to TXT format, which was less consistent

Initial Attempt

- Use both unsupervised and supervised learning techniques
- Clustering to perform extract features from precedents
- Classification & regression to perform predictions

Data Pipeline (Facts/Demands/Outcomes)

Remove entities	Remove Stopwords, stem + lemmatization	Vectorize words	Vectorize sentences	Cluster Sentence vectors
Replace Variable entities with constants: I.e. - 500$ -> MONEY - June 12 -> DATE - Julie -> NAME - etc		Obtain vector representation of words. Word2Vec was used (French)	Obtain vector representation of sentences. Mean of all word vectors was used.	HDBSCAN (DBSCAN variant) was used - Clusters to be as density-based - Unknown # of clusters - Scales well with dataset

Data Pipeline (Precedents)

Create precedent vector	Train Regressors & Classifiers to predict outcomes	Ask user Questions to populate input	Return Prediction to user
Each precedent was turned into a vector using cluster membership	Input vector composed of: - Demands - Facts Output vector: - Outcomes	Web chatbot used to populate inputs	Return a UI dashboard of the results. - Predictions - Additional Resources

Problems with this approach

- Relied heavily on correct clustering
  - DBSCAN clusters ended up with too many data points classified as noise
  - KNN, GMM, Meanshift ended up with “noisy” clusters
- No visible progress while we improved fact/outcome/demand clustering
  - Puts project success in peril

Actual Implementation

- Use regular expressions to extract facts/demands/outcomes
- Validate correctness of regular expressions using clustered sentences.

Feature Extraction

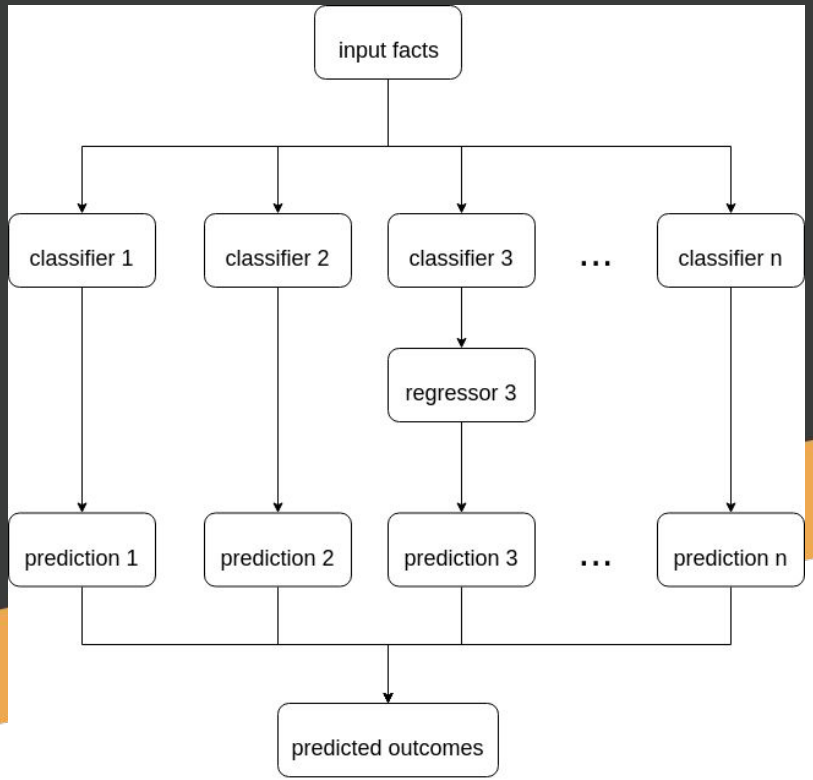
Inspect Sample documents	Write regex to match statements	Evaluate regex correctness	Use entity extraction to extract values	Create Precedent Vectors
• Understand writing style of legal documents. • Find common patterns among various documents	Verify validity of regex using: • Fact Coverage • Manual inspection of regex matches •	Metrics used: • % coverage of cluster sentences • % of statement coverage per precedent	Extract the following entities: • Money • Date • Number	Obtain vector representation of precedents by using regex match.

Predicting Outcomes + Precedent Similarity

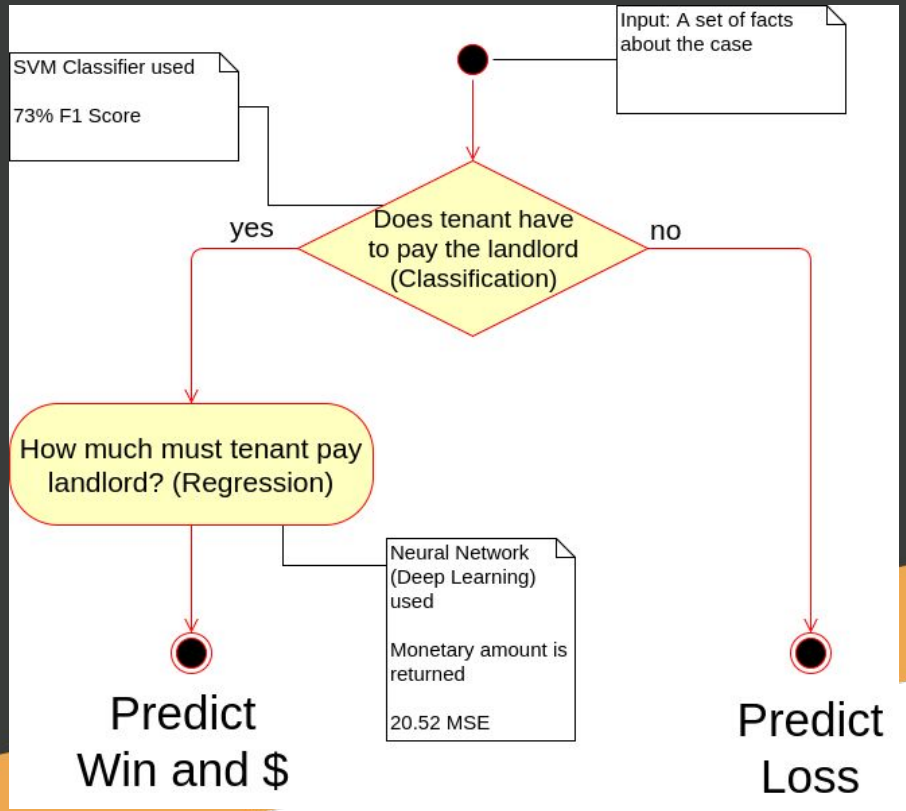
- Current Statistical ML Predictions
  - Classification
    - Tenant Expulsion
    - Lease Termination
  - Regression
    - Monetary amount that the tenant must pay the landlord (Classification & Regression)
    - Indemnity fee that must be paid (Classification & Regression)
  - Similarity
    - Finding similar precedents to your case

Making Predictions

- Series of classifiers make one prediction per outcome
- Certain outcomes require regression as well as classification



Prediction: Amount tenant owes landlord



Classification Predictions

Prediction	f1
additional_indemnity_date	[0.87275506 0.87743835]
additional_indemnity_money	[0.87868732 0.74466434]
declares_housing_inhabitable	[0.999875 0.]
declares_resiliation_is_correct	[0.95337609 0.69552941]
orders_expulsion	[0.97162192 0.9619883]
orders_immediate_execution	[0.91473297 0.81991443]
orders_resiliation	[0.97318436 0.9659671]
orders_tenant_pay_first_of_month	[0.99673981 0.]
rejects_landlord_demand	[0.98689956 0.5255814]
rejects_tenant_demand	[0.98617865 0.42245989]
tenant_ordered_to_pay_landlord	[0.72910864 0.68399675]
tenant_ordered_to_pay_landlord_legal_fees	[0.80152357 0.69338422]

Precedent Similarity

- Determine which precedents are similar to your case
- Experimenting with a few similarity
  Metrics:
  - Cosine Similarity
  - Nearest Neighbour with
    mahalanobis distance
- Limitations:
  - No access to evaluation metrics (Expensive to obtain manually)

Thank you!