Automating travel booking requests

David S. Batista - November 2021

Team

Sebastian Mika
VP Data

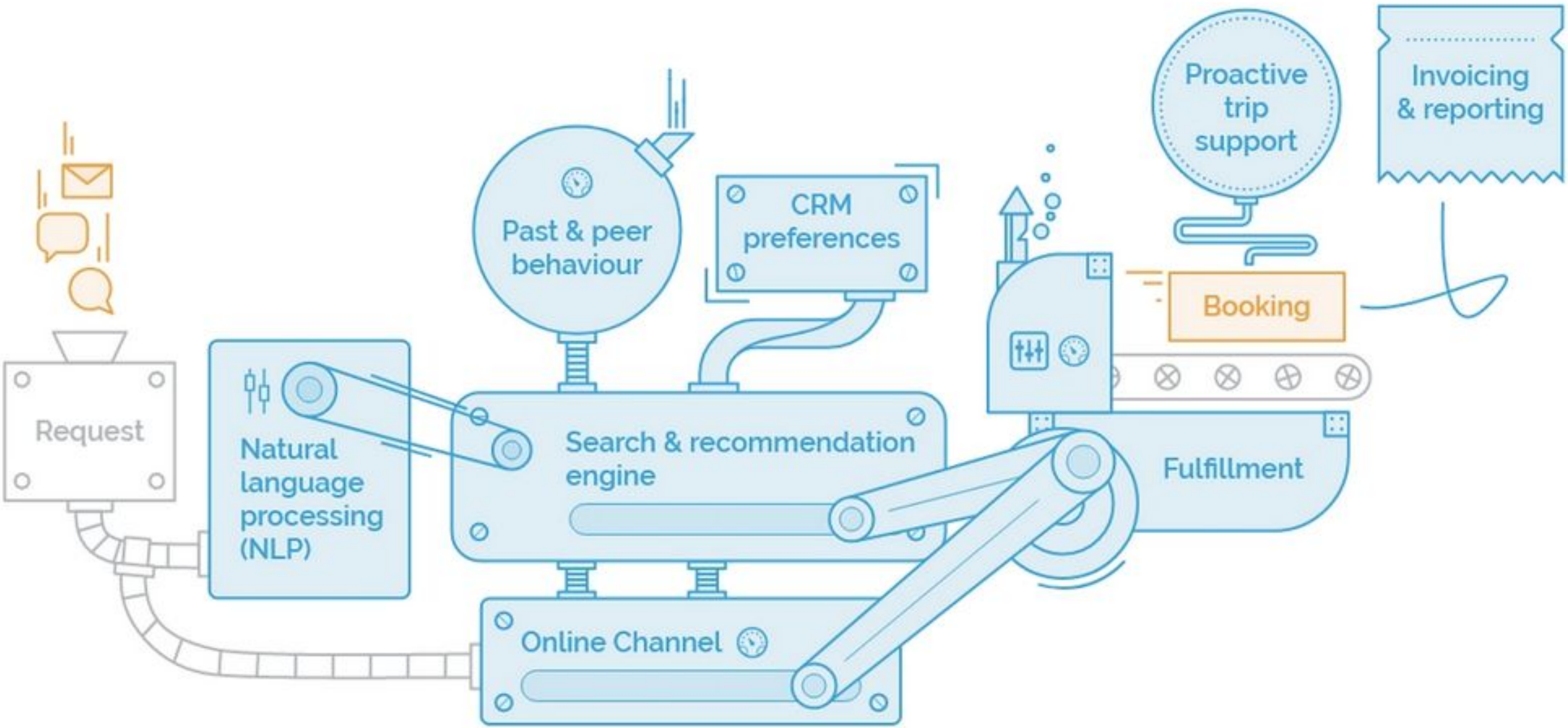
David Batista
Lead NLP Engineer

Sandip Mukherjee
ML Engineer

Scott Martens
Senior ML Engineer

+ team of 4 remote worker annotators

What is the Automation?



Hello Comtravo,

I need a train from Berlin to Munich, next Thursday around 9:00
and back to Berlin on Friday at 18:00.

Best Regard,
Mr Muster from Muster Inc.

Hello Comtravo,

I need a train from Berlin to Munich, next Thursday around 9:00 and back to Berlin on Friday at 18:00.

Best Regard,
Mr Muster from Muster Inc.

Booking: 94.3% **Not a Booking: 5,7%**

~~Hello Comtravo,~~

I need a train from Berlin to Munich, next Thursday around 9:00 and back to Berlin on Friday at 18:00.

~~Best Regard,~~
~~Mr Muster from Muster Inc.~~

Booking: 94.3% **Not a Booking: 5,7%**

~~Hello Comtravo,~~

I need a train from **Berlin** to **Munich**, next **Thursday around 9:00** and back to **Berlin** on **Friday at 18:00**.

~~Best Regard,~~
~~Mr Muster from Muster Inc.~~

Booking: 94.3% **Not a Booking: 5,7%**

Trip 1

origin: Berlin
destination: Munich
dpt_time: Thursday around 9:00

Trip 2

origin: Munich
destination: Berlin
dpt_time: Friday at 18:00

Trip 1

origin: 08011155
destination: 08000261
dpt_time: Thursday around 9:00

Trip 2

origin: 08000261
destination: 08011155
dpt_time: Friday at 18:00

Trip 1

origin: 08011155
destination: 08000261
dpt_time: 2021-12-02T09:00

Trip 2

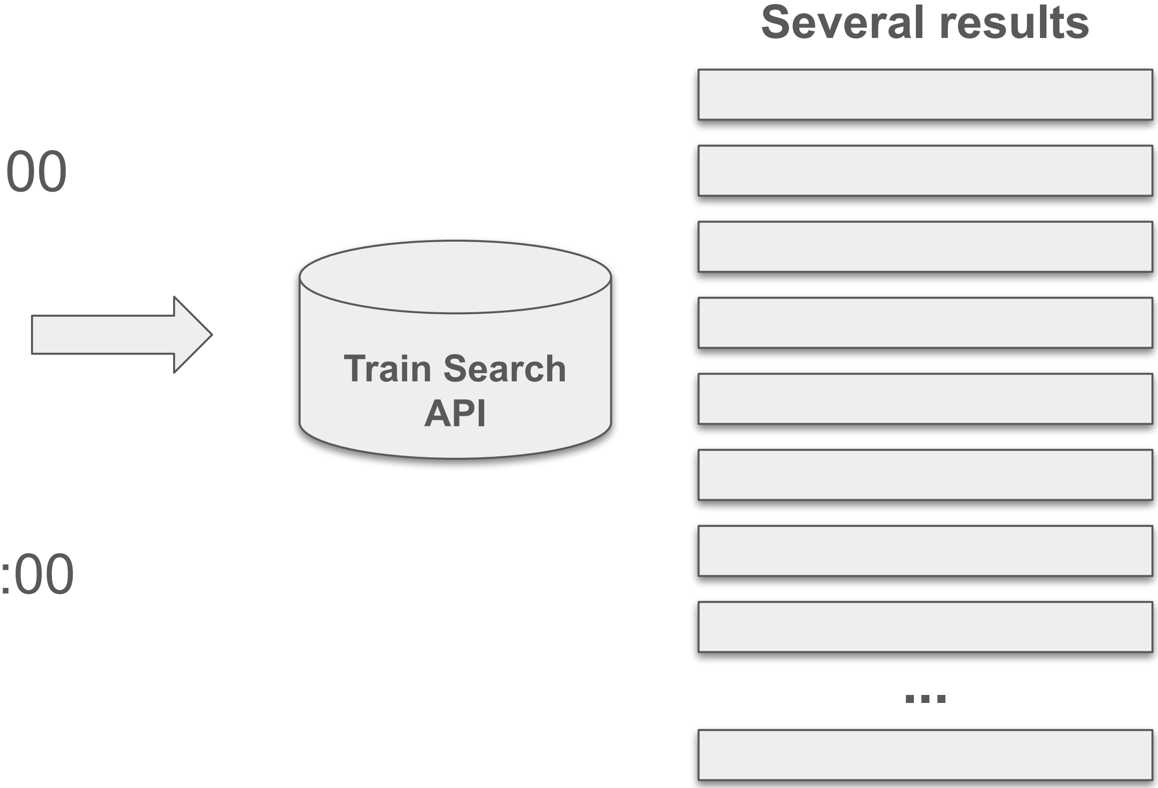
origin: 08000261
destination: 08011155
dpt_time: 2021-12-03T18:00

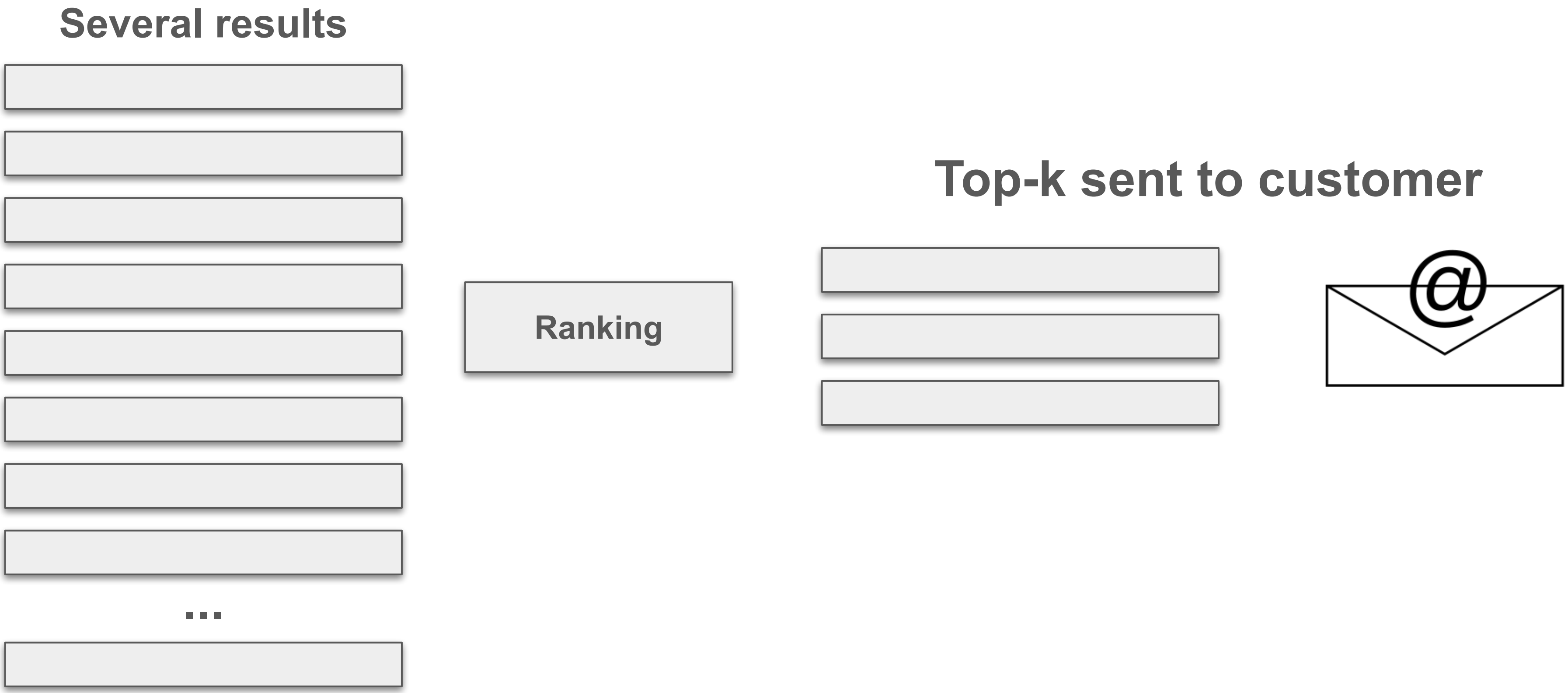
Trip 1

origin: 08011155
destination: 08000261
dpt_time: 2021-12-02T09:00

Trip 2

origin: 08000261
destination: 08011155
dpt_time: 2021-12-03T18:00





Several results

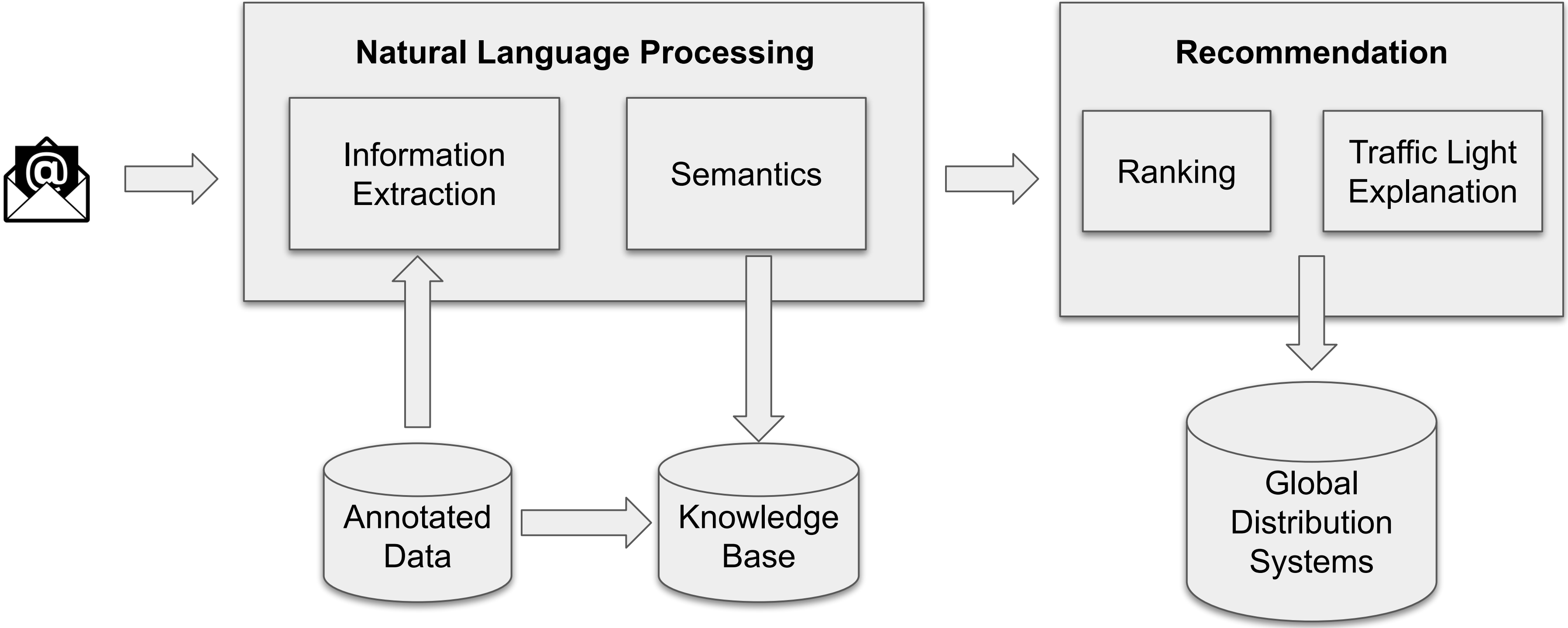
...

Ranking

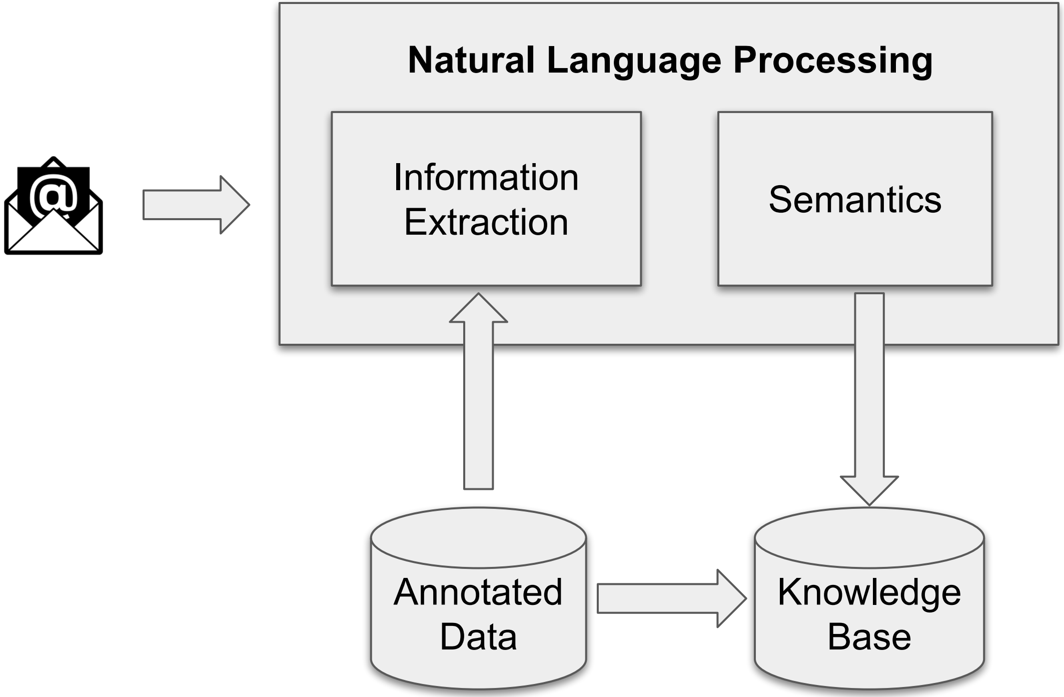
Top-k sent to customer

How does this all comes together?

Overview


Natural Language Processing
Information Extraction
Semantics
Recommendation
Ranking
Traffic Light Explanation
Annotated Data
Knowledge Base
Global Distribution Systems

Overview


Natural Language Processing
Information Extraction
Semantics
Annotated Data
Knowledge Base

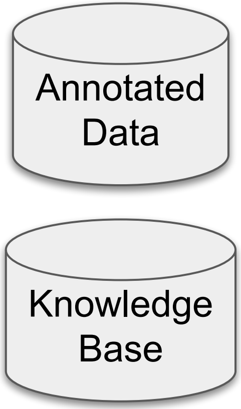
Annotated Data: surface strings

Annotated Data

- Named-Entities
  - Localities: airports, train stations, cities, full addresses
  - Temporal Expression: check-in/out, departure, time intervals
  - Airlines: *TAP, Lufthansa, EW*
  - Flight Numbers: “*LH123”, TP31”*
  - Train Types and Numbers:: “*ICE123”*
  - Hotel Names: *“Ibis”, “Motel One”, etc.*
  - Hotel Category/Stars: “*only 3 or 4 stars hotel please”*
  - Prices: “*between 80 and 120 EUR per night”*

- 40+ Named-Entity types

Annotated Data: semantics



Add semantics to the named-entities:

- time expressions → date format YY/MM/DD
- *‘Berlin’* - BER (airport), 8011160 (train station code), lat/lon
- *‘Adlon Hotel’* - #1234 (identifiers in hotel providers)

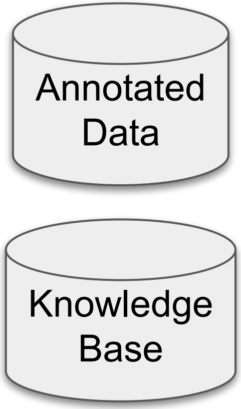
Knowledge Bases:

- Open source data : WikiData, GeoNames, DB Open Data
- Acquired data sets and payed APIs access

Annotators use the KB to “semantify” the named-entities:

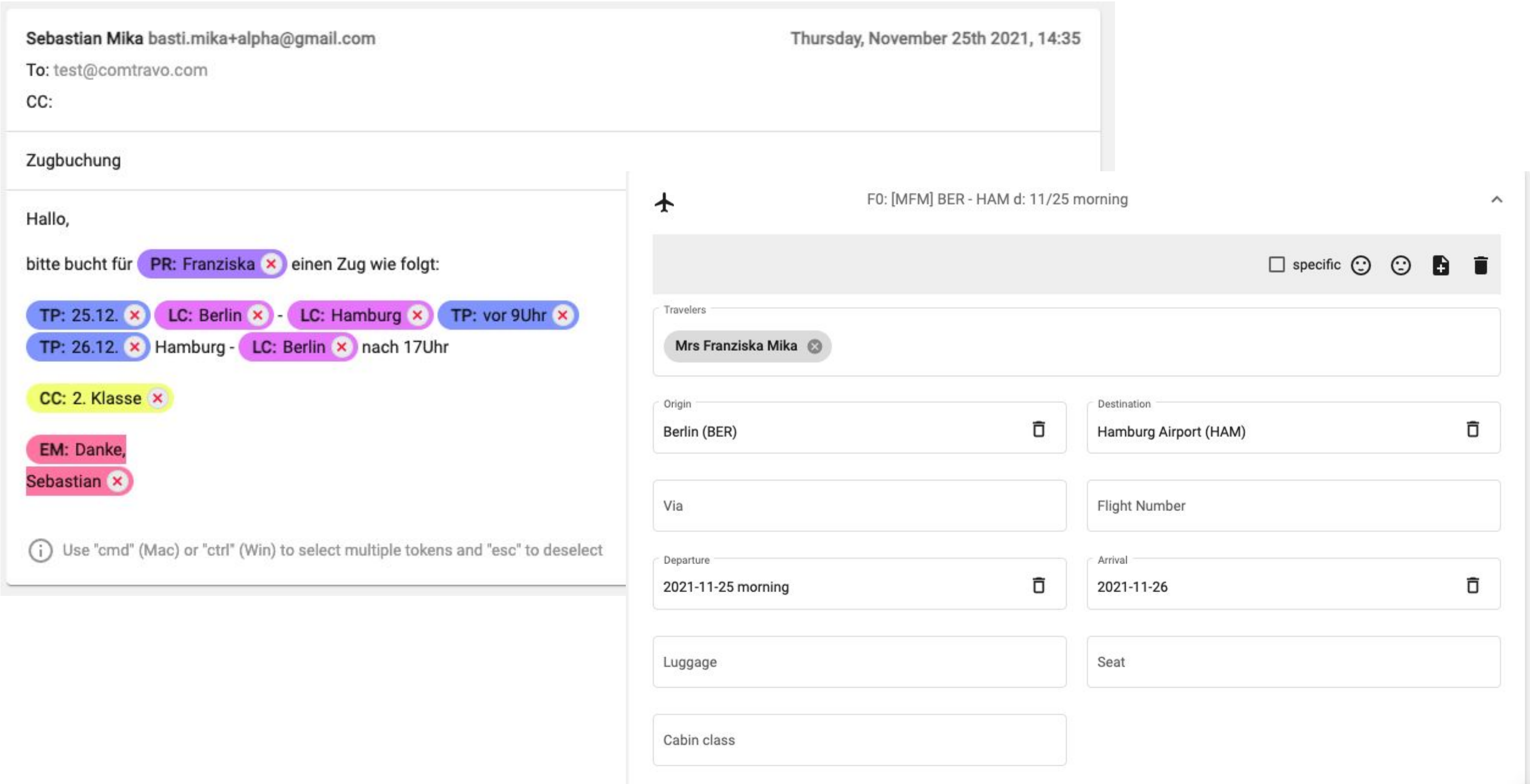
- real-world concepts, identifiers in the Knowledge Base
- geographic coordinates, etc.

Annotated Data: our own tool



Annotated Data

Knowledge Base



Sebastian Mika basti.mika+alpha@gmail.com

Thursday, November 25th 2021, 14:35

To: test@comtravo.com

CC:

Zugbuchung

Hallo,

bitte bucht für PR: Franziska ✕ einen Zug wie folgt:

TP: 25.12. ✕ LC: Berlin ✕ - LC: Hamburg ✕ TP: vor 9Uhr ✕
TP: 26.12. ✕ Hamburg - LC: Berlin ✕ nach 17Uhr

CC: 2. Klasse ✕

EM: Danke,
Sebastian ✕

ⓘ Use "cmd" (Mac) or "ctrl" (Win) to select multiple tokens and "esc" to deselect

F0: [MFM] BER - HAM d: 11/25 morning

□ specific

Travelers: Mrs Franziska Mika

Origin: Berlin (BER)

Destination: Hamburg Airport (HAM)

Via

Flight Number

Departure: 2021-11-25 morning

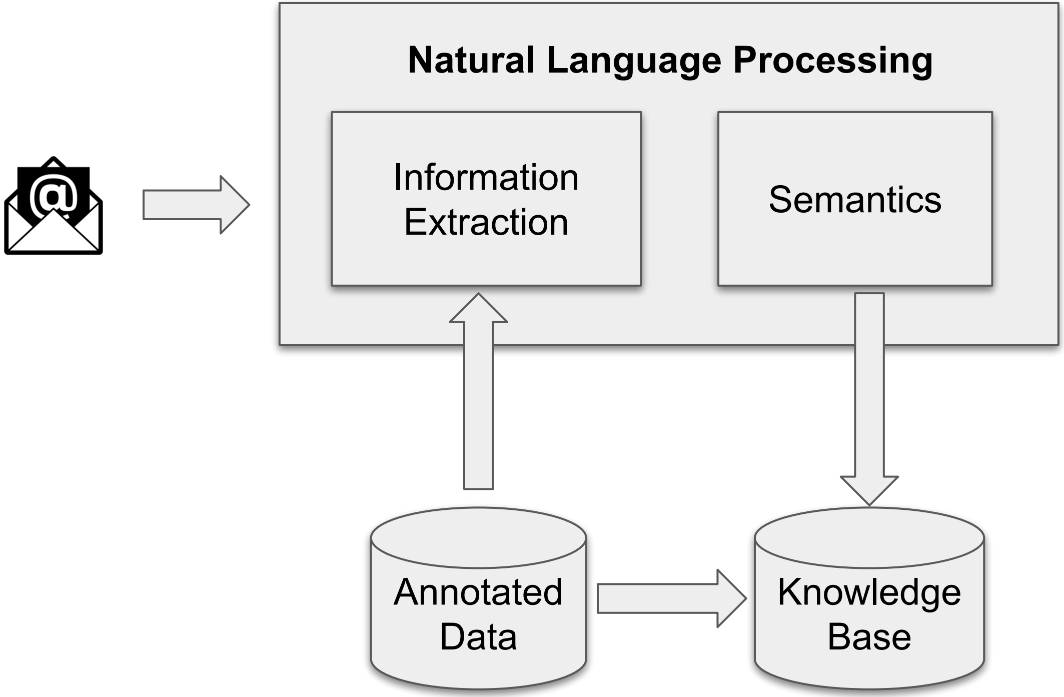
Arrival: 2021-11-26

Luggage

Seat

Cabin class

Overview



Natural Language Processing

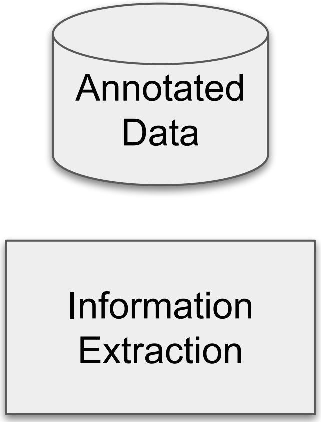
Information Extraction

Semantics

Annotated Data

Knowledge Base

Natural Language Processing - Information Extraction



Annotated Data

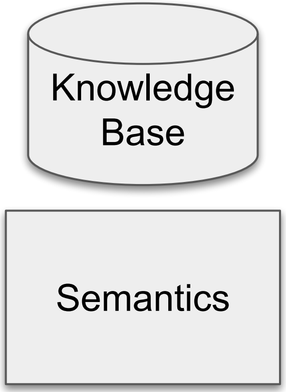
Information Extraction

- **Document level classifier**: is this email a booking ?
    - `DistilBert`

- **Named-Entity recognition**: supervised model + pattern-based rules
    - `BERT: Transformer`
        - Language Model pre-trained model on Wikipedia
        - Fine-tuned on our annotated data

    - Regex based
        - triggered after the model predictions
        - specific entities - low annotated samples
        - important clues in the message we need to get

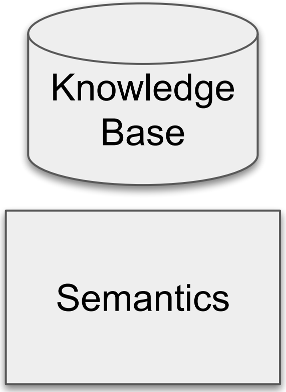
Natural Language Processing - Semantics



- ***‘Berlin’***
  - *BER* - IATA code for the Berlin airport
  - 8011160 - IBNR code for Berlin Central Train station
- ***‘Park Inn’***: associate an identifier from booking.com, HRS, etc.
- **Options**: room type, luggage, rebookable, cancelable

- Cleaning/Adjusting the output of the tagger
- Analyzing relationships between tagged entities
- Querying Knowledge Bases + Reasoning
- Using pre-computed priors from the KB + our internal data

Natural Language Processing - Semantics


Knowledge Base

Semantics

Need a hotel for two nights in ***Frankfurt***

Frankfurt am Main	Hessen	0.93
Frankfurt (Oder)	Brandenburg	0.2
..	..	..

Need a flight to ***New York****,* ***JFK*** *departing from either* ***Düsseldorf*** *or* ***Cologne***

Time Expressions: ct-parse

- https://github.com/comtravo/ctparse
- rules, regular expression, supervised modeling ~ **PCFG**
- Normalization of English and German time expressions

Natural Language Processing - Outcome

Please book me two nights from the 2nd of December to the 4th in the Quality Hotel am Tierpark

check-in	02.12.2021	🟢
check-out	04.12.2021	🟢
location	Tierpark	🟡
hotel	Quality Hotel	🟡

Natural Language Processing - Outcome

I need a train on Wednesday 1st of December from Frankfurt to Hamburg, please first train of the day.

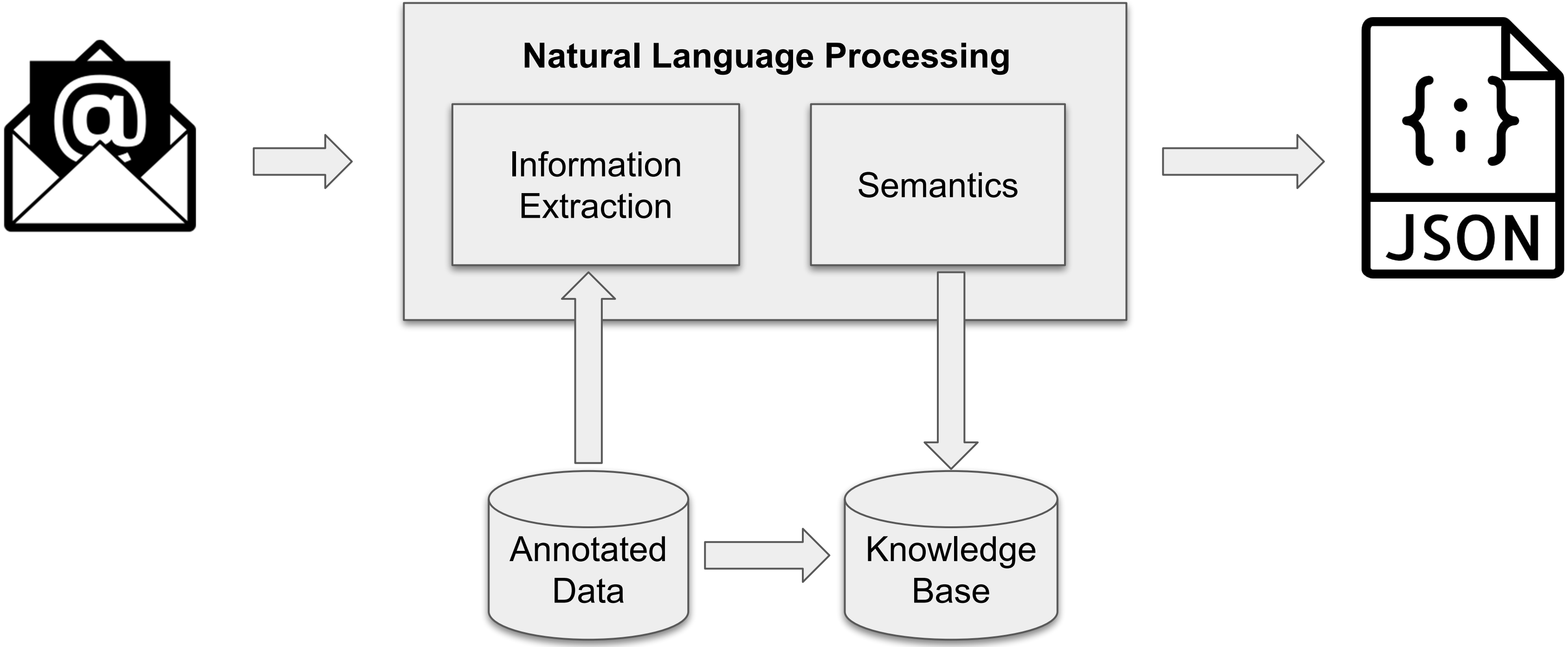
dpt_time	01.12.2021	🟡
origin	800321	🟢
destination	800123	🟢

Natural Language Processing - Outcome

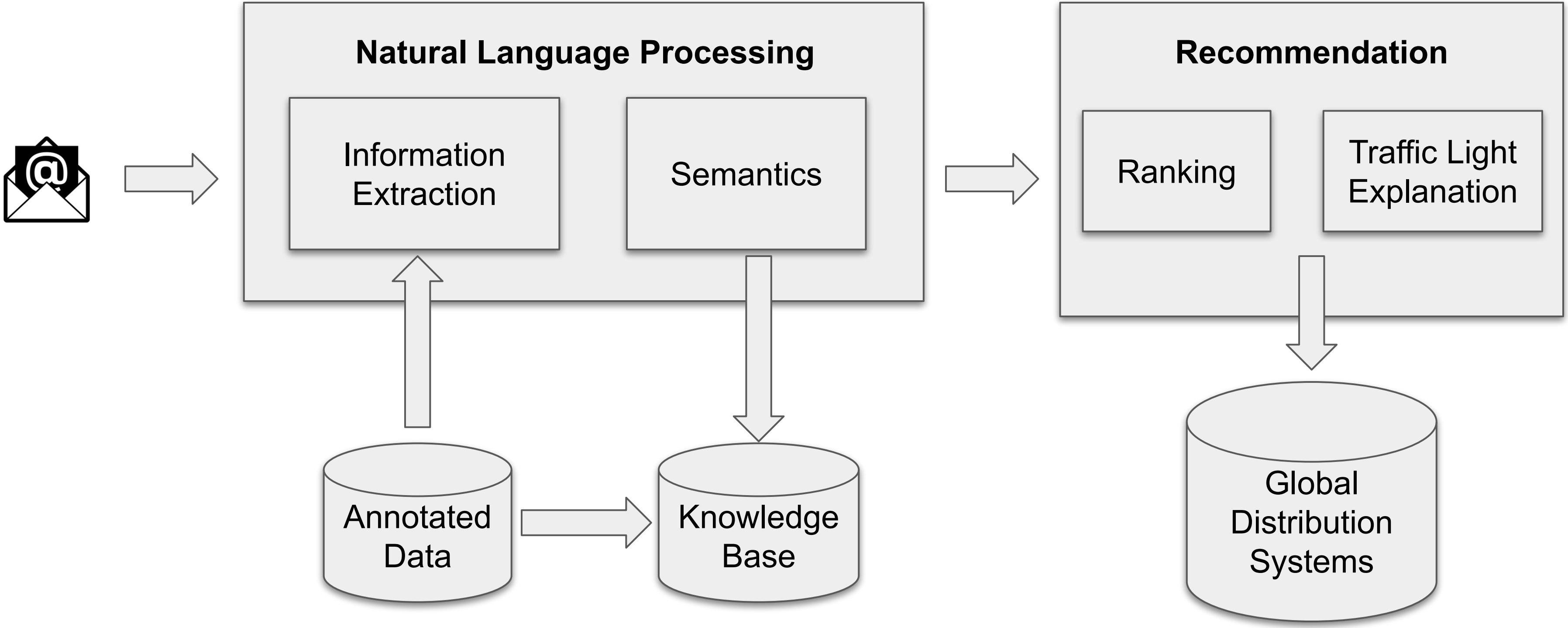
I need a flight on the 11th of December from Berlin to Lisbon.

dpt_time	11.12.2021	🟢
origin	BER	🟢
destination	LIS	🟢

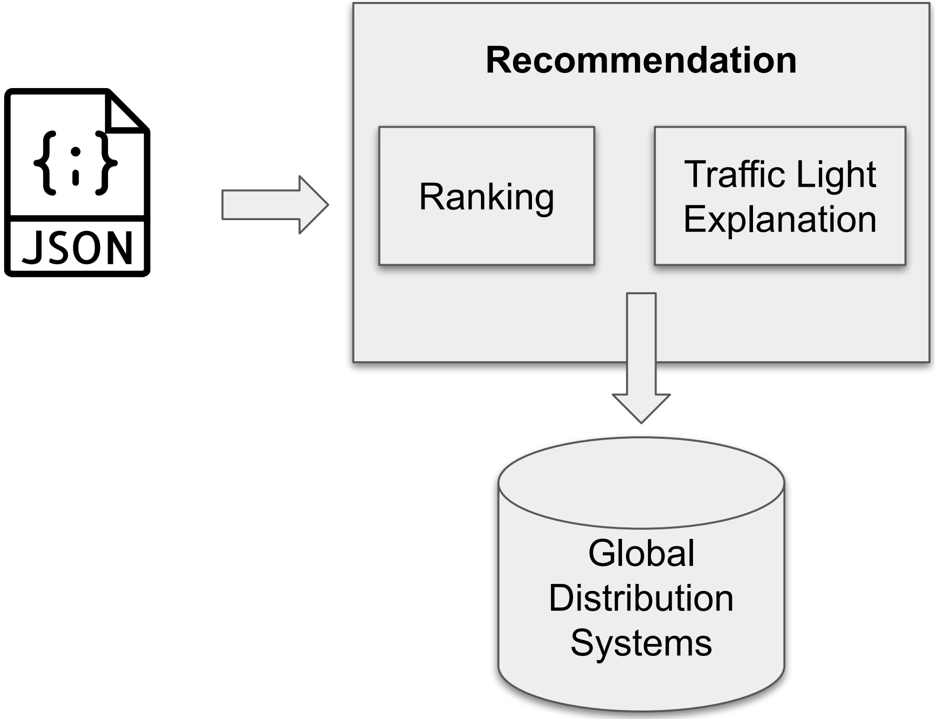
Natural Language Processing



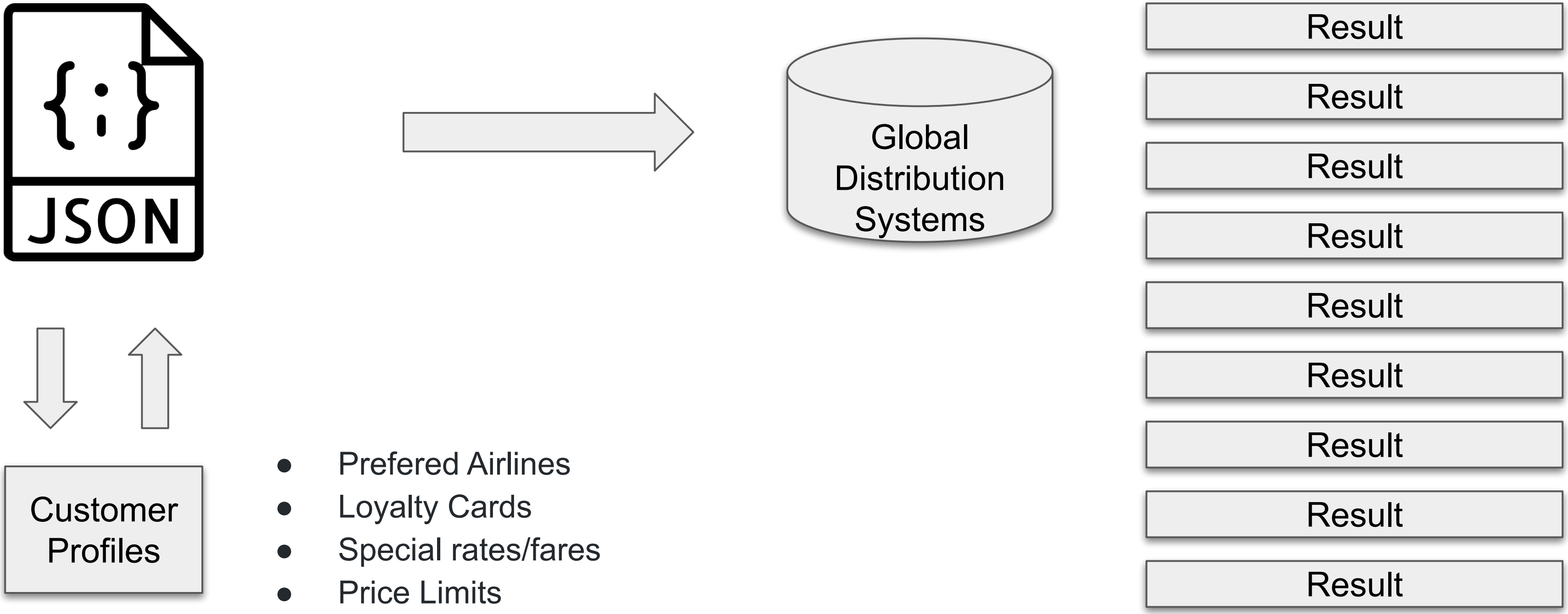
Overview


Natural Language Processing
Information Extraction
Semantics
Recommendation
Ranking
Traffic Light Explanation
Annotated Data
Knowledge Base
Global Distribution Systems

Overview


JSON
Recommendation
Ranking
Traffic Light Explanation
Global Distribution Systems

Recommendation - Search



- Prefered Airlines
- Loyalty Cards
- Special rates/fares
- Price Limits

Recommendation - Ranking

From hundreds of results, how to select the best ones ?

Multiple Objectives:

- the cheapests train tickets
- but also for the shortest trip time
- the lowest possible number of changeovers

NOTE: assume that each objective is a quantity we want to minimize

Pareto Efficiency/Optimality: finds solutions in a multi-objective setting

Recommendation - Pareto Efficiency

- No single objective can be further improved without hurting others objectives
- A and B are pareto efficient while C is not
- Pareto-efficient solutions are not unique: pareto frontier

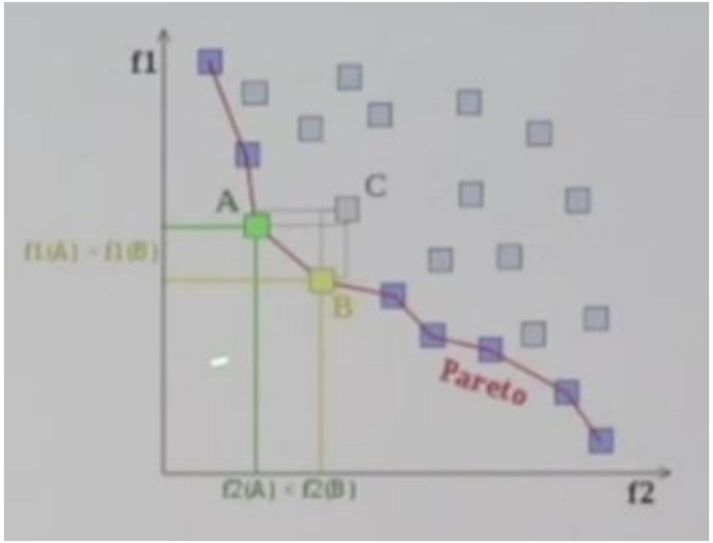
Formally:

Two results, r_i, r_j with *K* objectives $f_1, \cdots, f_k$

$$r_i = (f_1^i, \cdots, f_k^i), r_j = (f_1^j, \cdots, f_k^j)$$

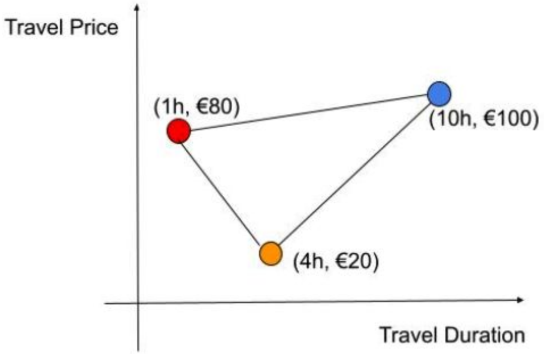
r_i dominates r_j iff $f_1^i \leq f_1^j, \cdots, f_k^i \leq f_k^j$

r_i is Pareto-efficient iff there exists no r_j which dominates r_i



Recommendation - Pareto Efficiency

	Name	Price	Duration
🔴	Train 1	80 euros	1 hour
🟠	Train 2	20 euros	4 hours
🔵	Train 3	100 euros	10 hours



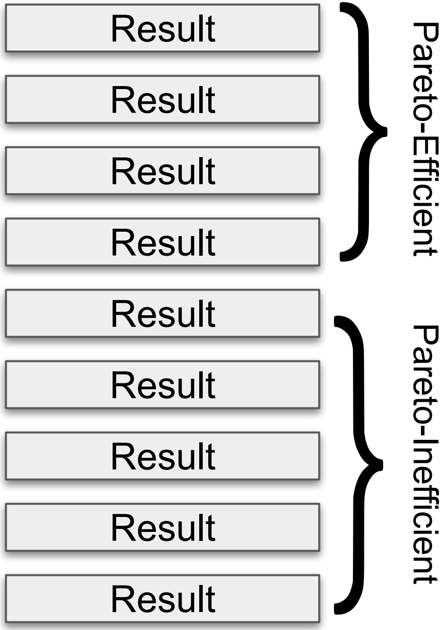
Recommendation - Pareto Ranking - Algorithm

Simple Cull Algorithm

- for every result item X
  a. compare X with every other result item
  b. if there is at least one result item which is better than X
     i. X is marked as **Pareto-Inefficient**
     ii. otherwise X is marked as **Pareto-Efficient**

- Each item will either be **Pareto-Efficient** or **Pareto-Inefficient**

Recommendation - Outcome



- Rank both sets - using a predefined precedence between objectives

- Select top-*k* and inspect them, how good are the selected results?

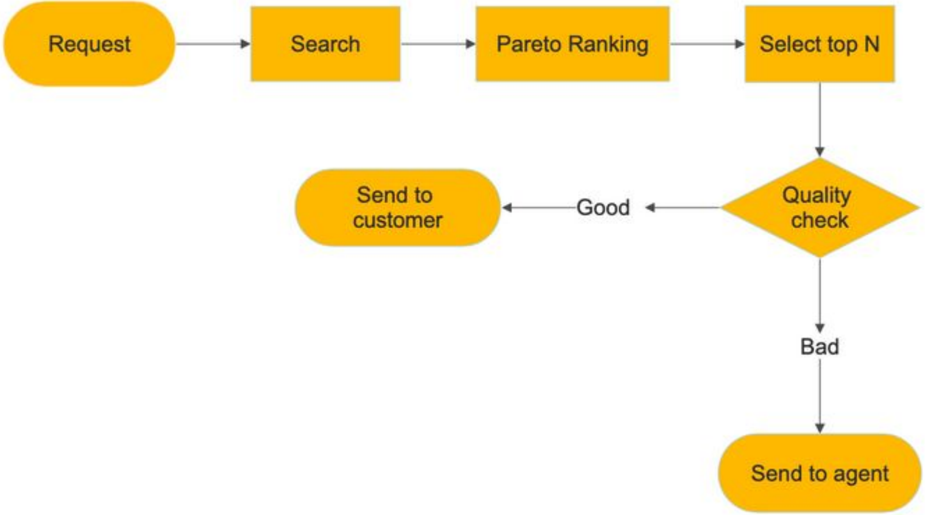
- We relax a bit the constraints to avoid not having any results, e.g:
  - a specific Hotel
  - a flight before 12:00

- Compare each of the top-*k* against the customer requests:
  - 🟢 if all preferences/objectives fulfilled
  - 🟡 otherwise

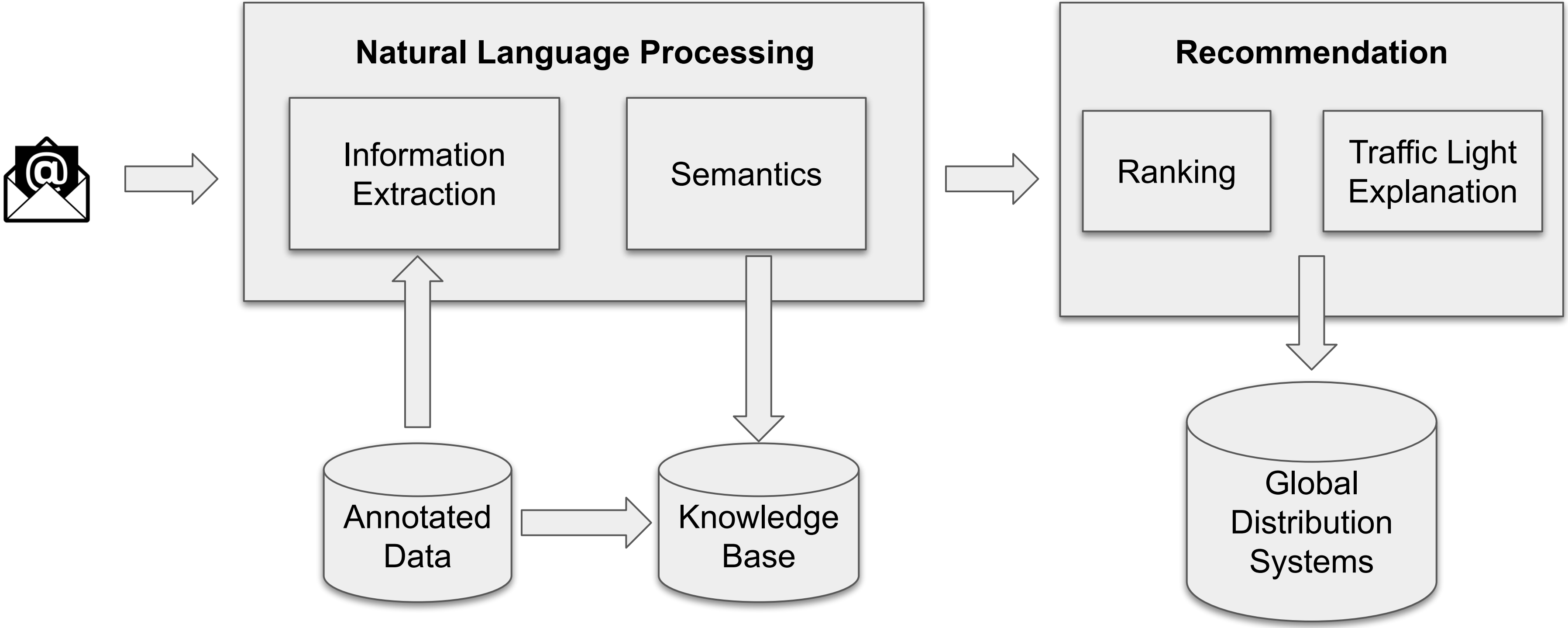
Recommendation - Outcome

- Combine Outcomes from
  - Natural Language Processing
  - Recommendation

- Depending on this outcomes we either:
  - Email the customer with top-*k* options
  - Shift handling to a travel agent:
    - detailing the reason(s)



Overview



Natural Language Processing

Information Extraction

Semantics

Recommendation

Ranking

Traffic Light Explanation

Annotated Data

Knowledge Base

Global Distribution Systems

Lessons learned: best practices

- Annotated data:
  - Training and evaluation of models
  - New annotations every week
  - Periodically run checks, statistical analysis

- Logging:
  - logs each relevant step in the pipeline
  - Events to S3 + ETL runs periodically indexing everything on Elasticsearch
  - input to search APIs and returns results algorithm did

- Error analysis:
  - Have an established process
  - Can be a framework to inspect all the logs referring every booking request