



Bringing NLP to Production

(an end to end story about some multi-language NLP services)

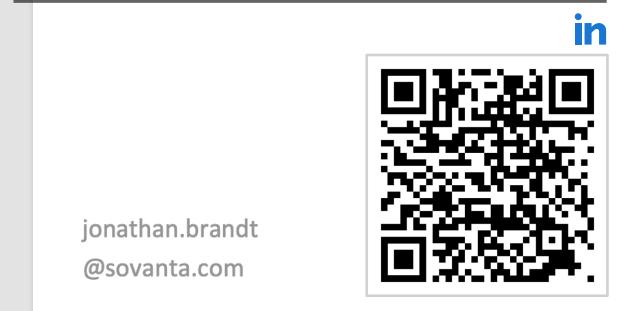
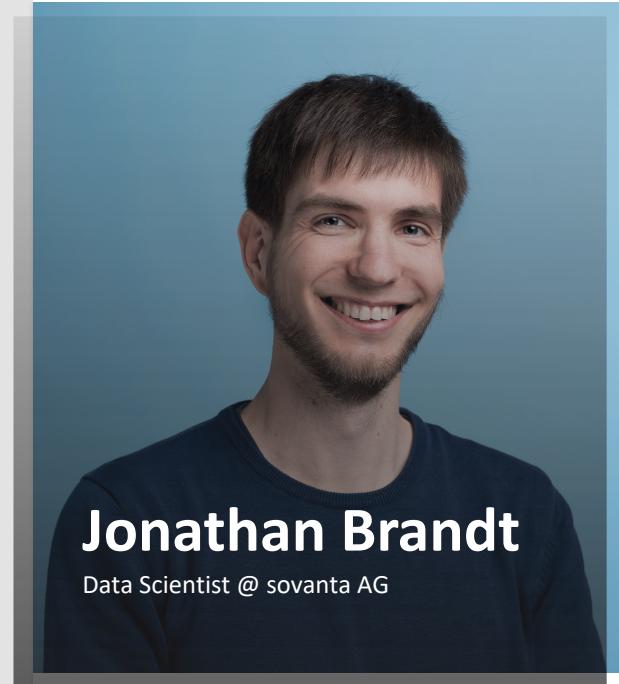
sovanta AG, 18.04.23

Larissa Haas, Jonathan Brandt

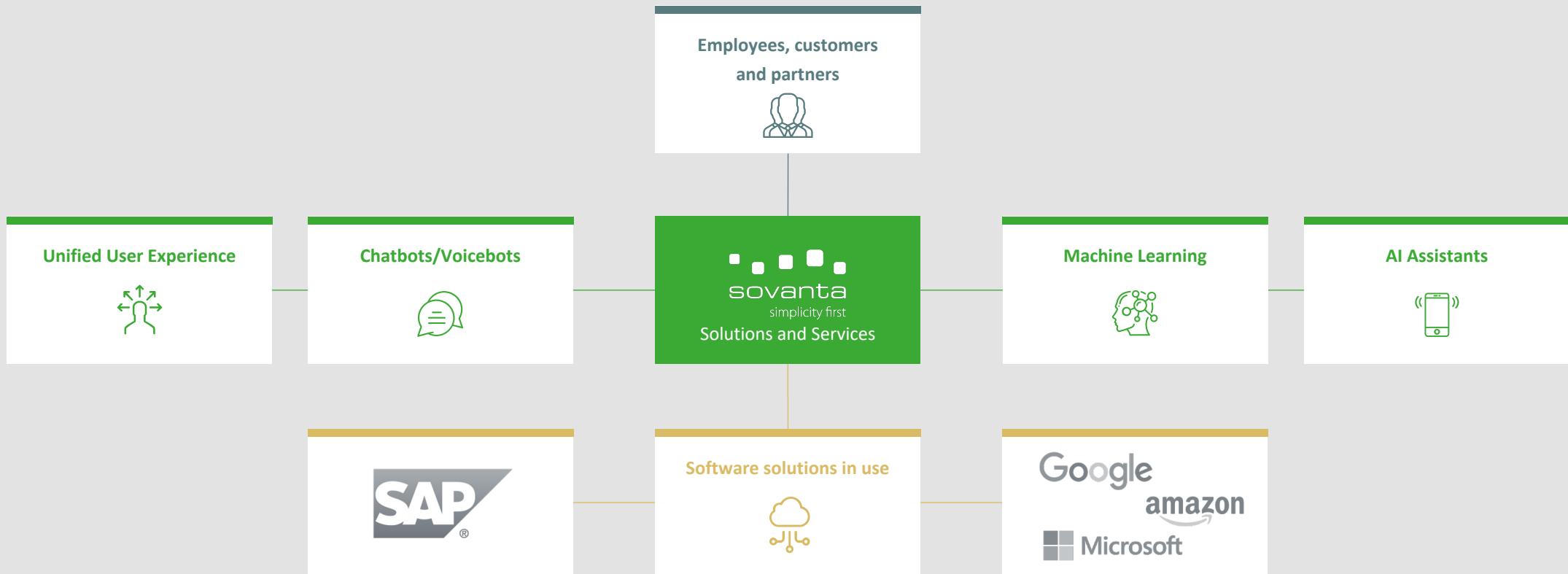


Agenda

1. Journey to NLPOps
2. Use Case 1 - Productsearch
3. Use Case 2- Text classification
4. Checklist to take away



User Experience closes the Business Gap

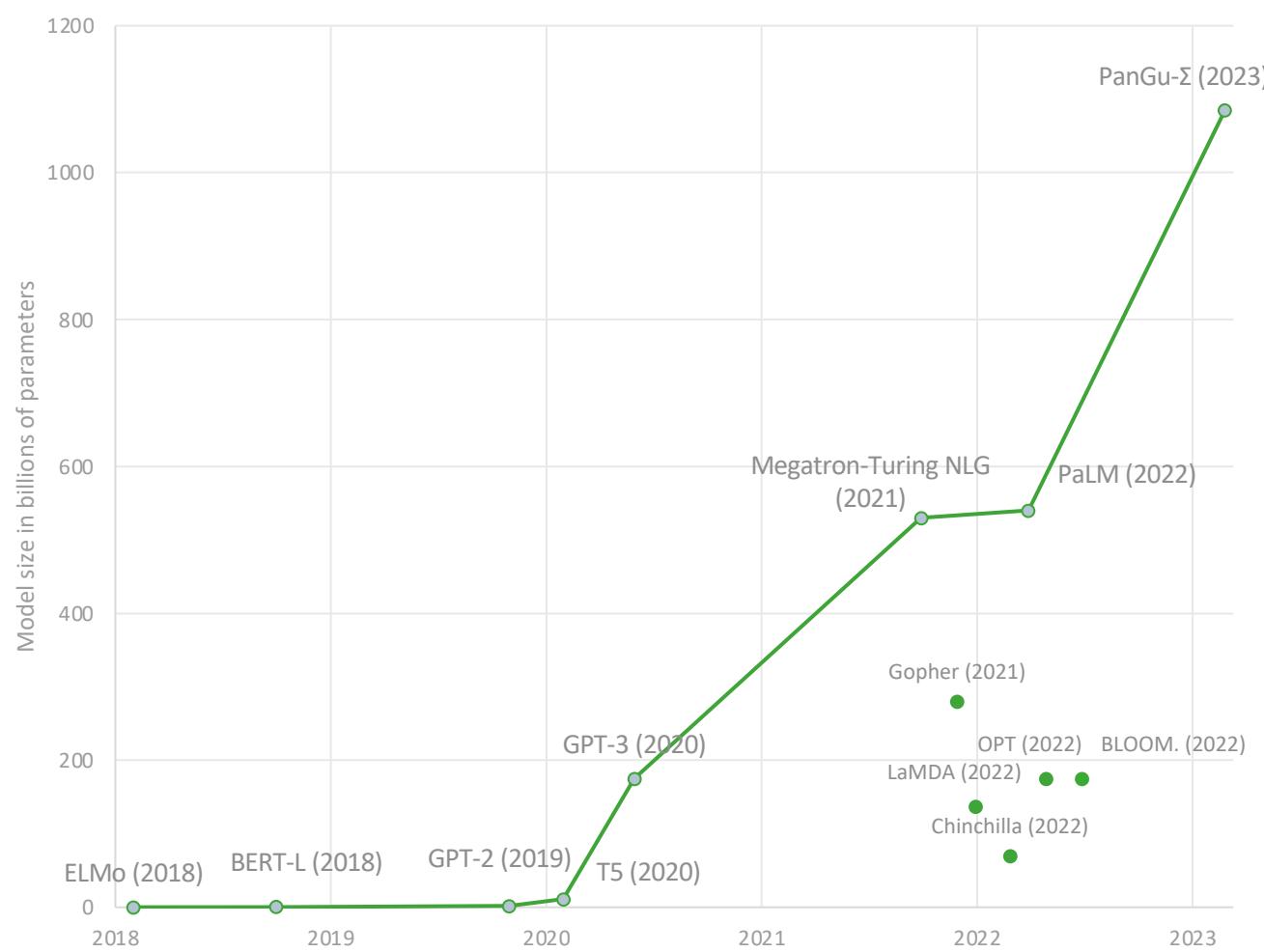




Our locations worldwide



Germany Heidelberg	150+ Customers
Germany Hamburg	500+ Projects
Germany Cologne	2 Mio. User
Spain Valencia	200 Employees
Argentina Buenos Aires	





Journey to NLPOps

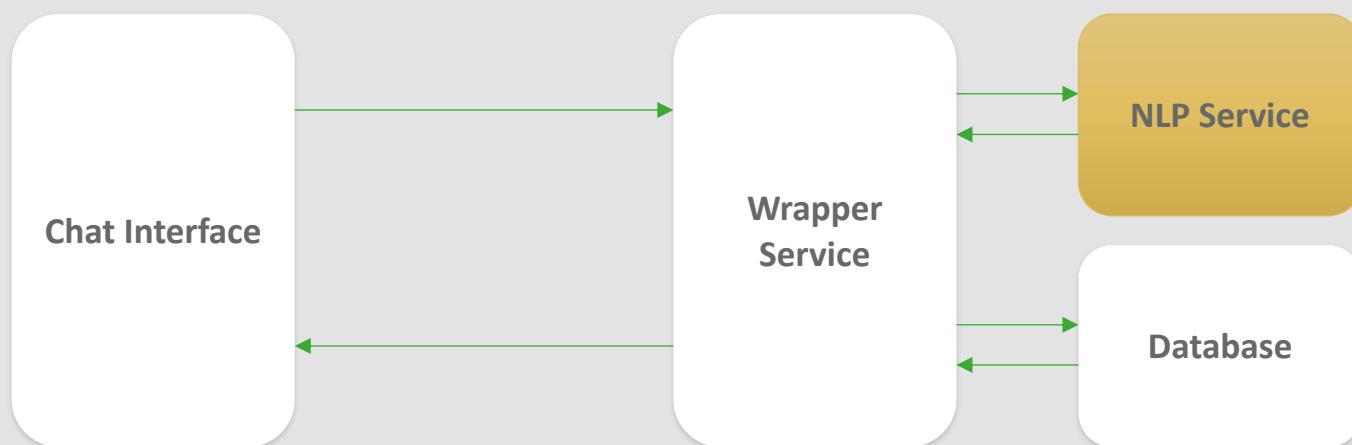




Deployment

NLPOps Checklist	
<input checked="" type="checkbox"/>	Model Type
<input checked="" type="checkbox"/>	Languages
<input checked="" type="checkbox"/>	Size of Dependencies
<input checked="" type="checkbox"/>	Retraining Frequency and Location
<input checked="" type="checkbox"/>	Request Frequency and Availability
<input checked="" type="checkbox"/>	Response Times
<input checked="" type="checkbox"/>	Running Perspective
<input checked="" type="checkbox"/>	Monitoring / Security
<input checked="" type="checkbox"/>	Running Context / Existing Infrastructure

Use Case 1 – Productsearch



I want to buy
vegan pizza

[
 {"text": "vegan", "ner": "property"},
 {"text": "pizza", "ner": "product"}
]

- Pizza XYZ
- Pizza ABC

Here are your
vegan pizza
options: ...

Use Case 1 – Productsearch

Feature	Characteristic
Model Type	SpaCy Model
Languages	One language
Size of Dependencies	ca. 300 MB
Retraining Frequency and Location	none
Request Frequency and Availability	
Response Times	
Running Perspective	PoC

Cloud Foundry (self-hosted)

Advantages:

- No need to care about infrastructure
- Complete service is pushed
- Quick (re)starts

Drawbacks:

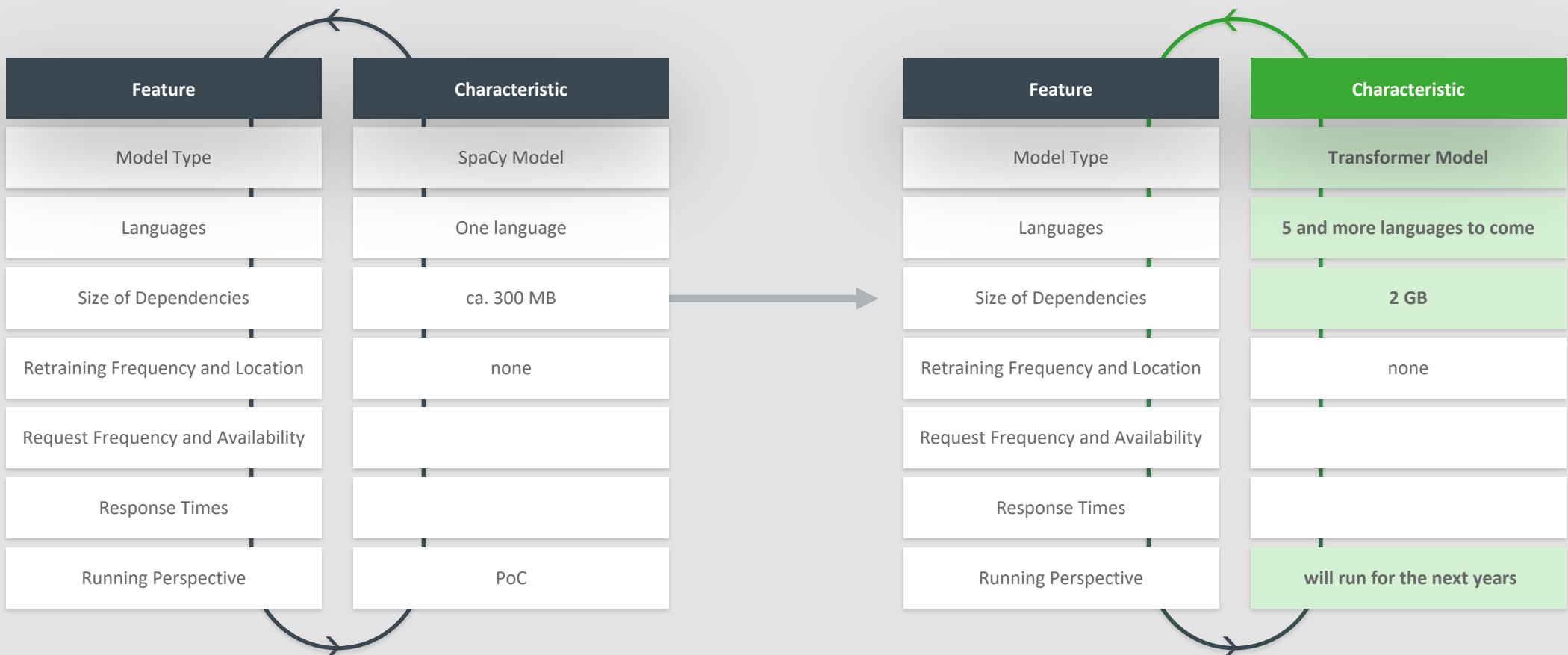
- Dependent on buildpack: only certain python versions available
- 1 GB upload limitation (with all dependencies)
- 4 GB disk space & limited memory allocation

```
! manifest-qx.dev.yaml
1  applications:
2    - name:           -connector-dev
3      type: python
4      path:          -connector
5      parameters:
6        memory: 250
7        disk_quota: 250
8        instance: 1
9        health-check-type: process
10       buildpacks:
11         - https://github.com/cloudfoundry/python-buildpack
12       provides:
13         - name:           -connector-binding
14         properties:
15           url: ${default-url}
16
```



Use Case 1 – Productsearch

Updating Requirements



Use Case 1 – Productsearch

Feature	Characteristic
Model Type	Transformer Model
Languages	10 languages
Size of Dependencies	ca. 2 GB
Retraining Frequency and Location	none
Request Frequency and Availability	
Response Times	
Running Perspective	will run for the next years

Azure Stack (Kubernetes / Container Instances)

Advantages:

- Quick, easy starting of containers without configuring clusters
- Known environment with Docker and deployment yaml

Drawbacks = Requirements:

- Docker or CI/CD pipeline
- Registry



Use Case 2 – Text classification



no error was
detected with the
device

das Sensorkabel ist
kaputt

... "text": "no error was detected with the
device", "lang": "EN" ...

... "text": "das Sensorkabel ist kaputt", "lang":
"DE" ...

"pred": 0 – not relevant

"pred": 1 – relevant

Use Case 2 – Text classification

Feature	Characteristic
Model Type	"plain" NLP
Languages	English + German
Size of Dependencies	ca. 300 MB
Retraining Frequency and Location	none
Request Frequency and Availability	scheduled scans + rare manual requests
Response Times	
Running Perspective	PoC

Cloud Foundry (SAP Business Technology Platform)

Advantages:

- Hosting existing application
- Easy communication between app and model
- BTP offers additional services

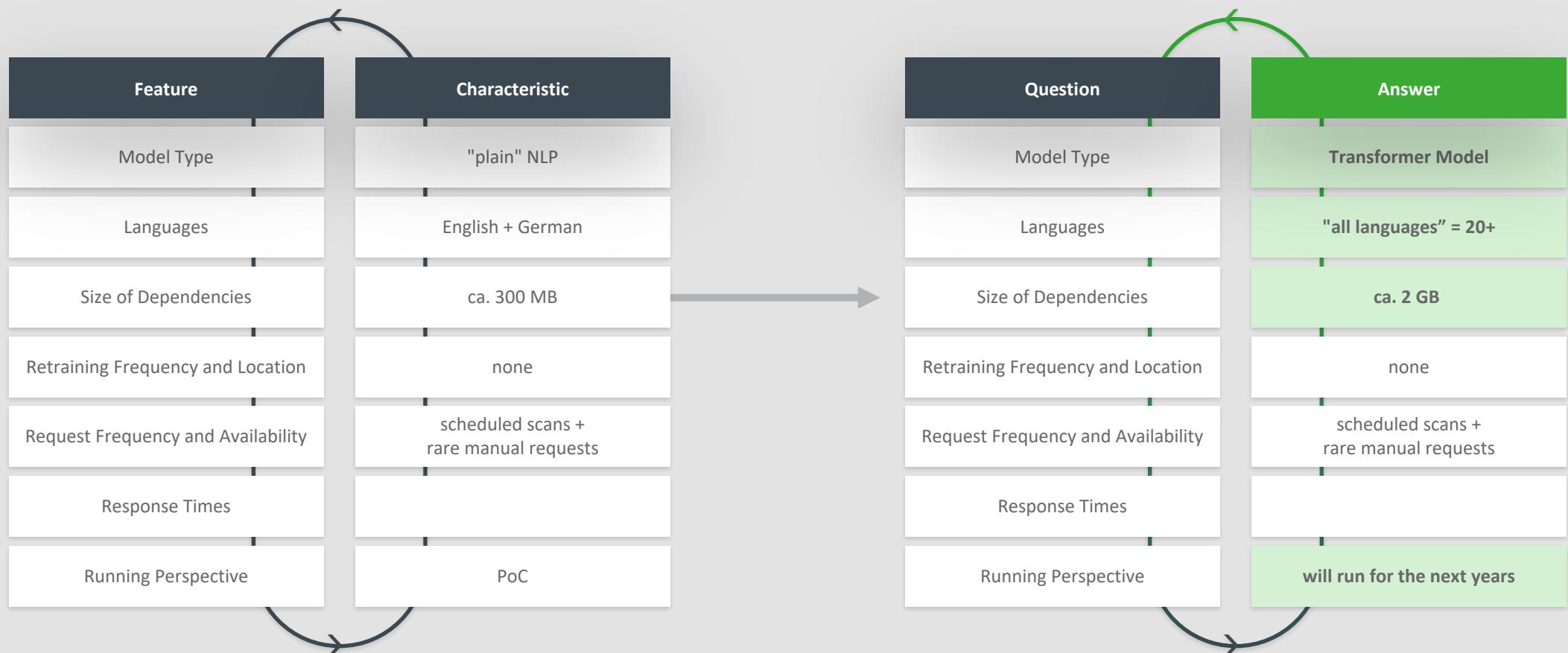
Drawbacks:

- 1.5 GB upload limitation (with all dependencies)
- 4 GB disk space & 8 GB memory allocation
- CPU only

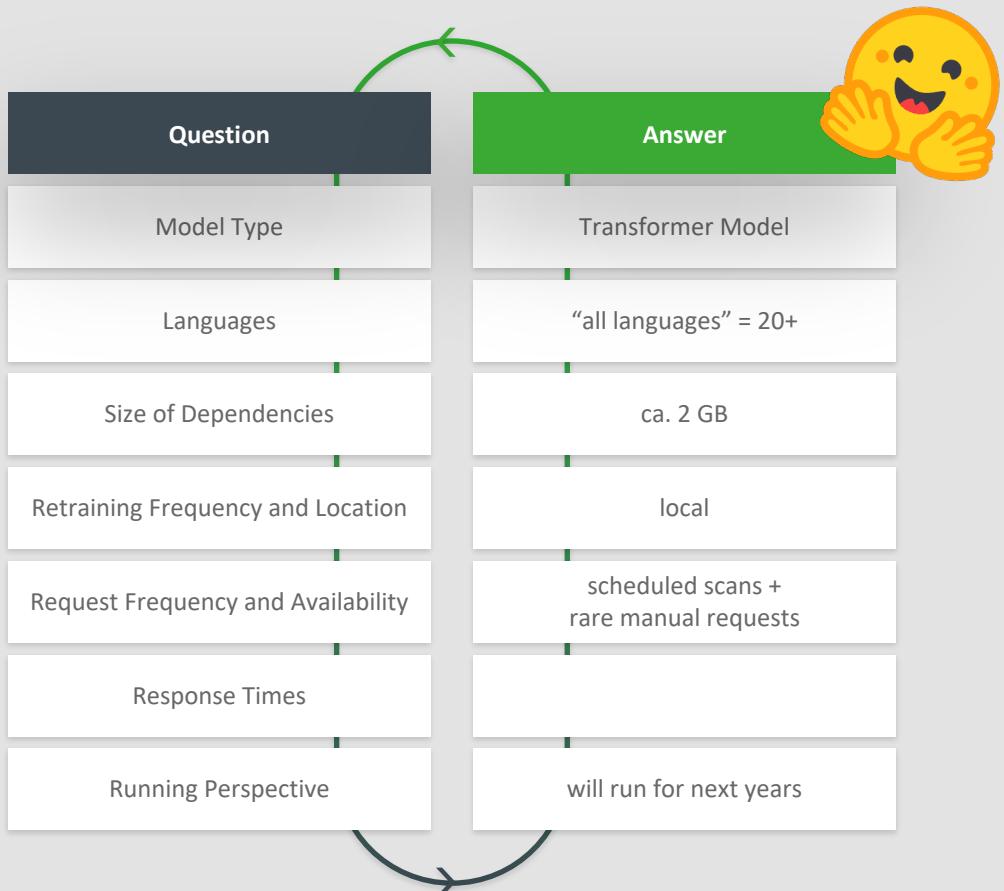
CLOUD FOUNDRY

Use Case 2 – Text classification

Updating Requirements



Use Case 2 – Text classification

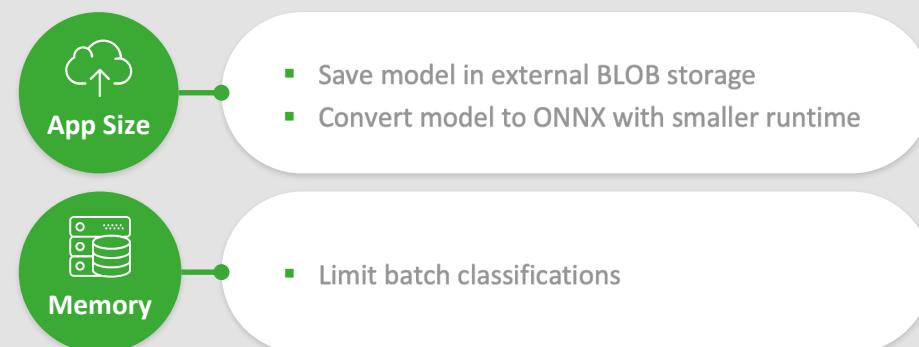


Cloud Foundry (SAP Business Technology Platform)

Preferred:

- Dedicated AI runtime not available
SAP AI Core (Managed Kubernetes Cluster)

Workarounds





NLPOps Checklist



Model Type



Languages



Size of Dependencies



Retraining Frequency and Location



Request Frequency and Availability



Response Times



Running Perspective



Monitoring / Security



Running Context /
Existing Infrastructure

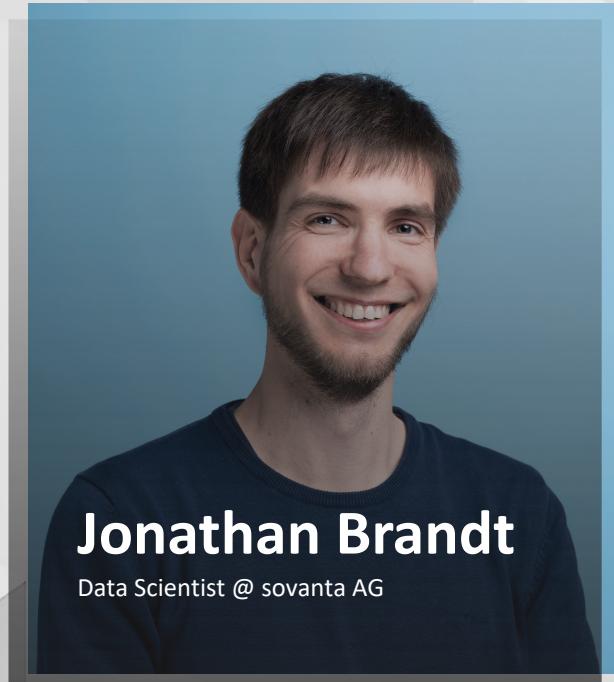
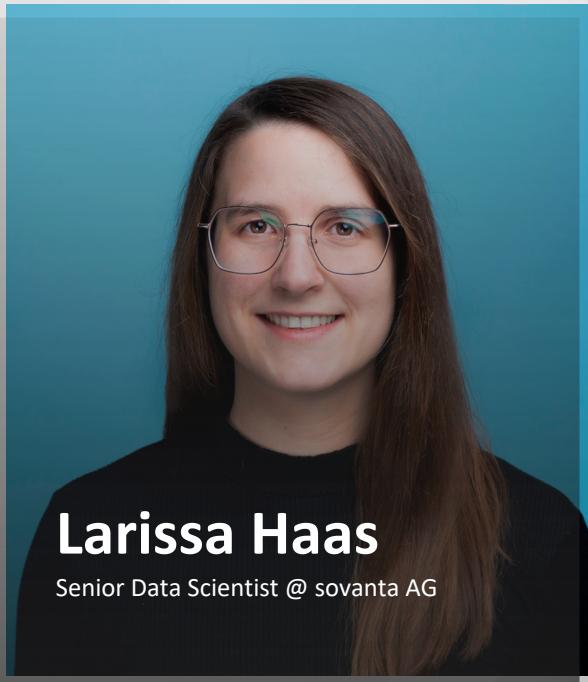
Key Takeaways and Further Reading

- Think about your goal with the model beforehand
- Clarify all options/restrictions you have
- Make yourself familiar with condensing mechanisms that work for your model
- Take this as guideline, there is no black and white

[Read more here](#)

Questions?

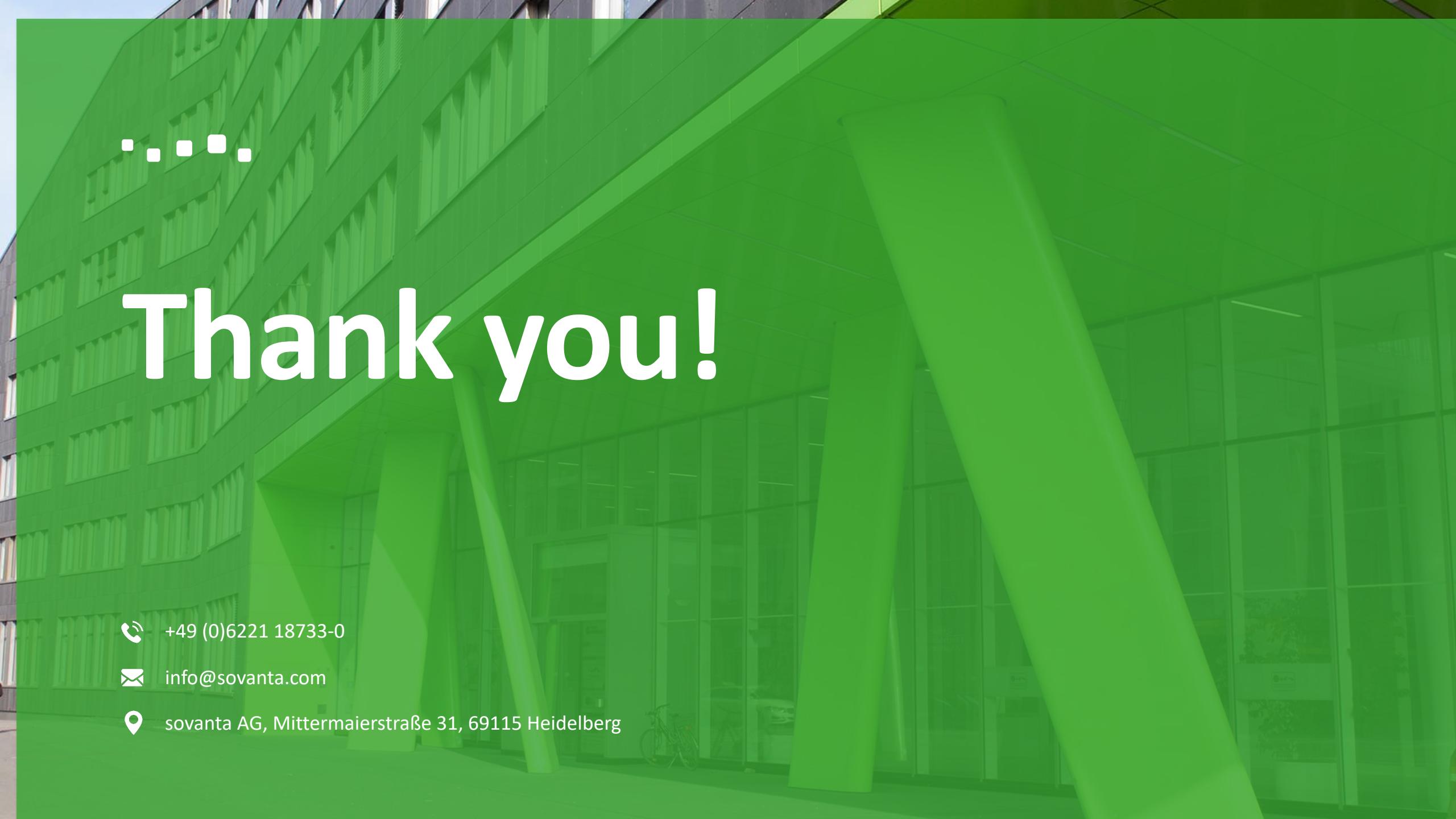
Find the talk slides here:



sovanta

/ We are hiring!





...
.

Thank you!

 +49 (0)6221 18733-0

 info@sovanta.com

 [sovanta AG, Mittermaierstraße 31, 69115 Heidelberg](#)