Cognitive ARtifacts for Machine Learning << CarML >>

github.com/rai-project/carml

September 19th 2017

Abdul Dakkak, Cheng Li

ILLINOIS

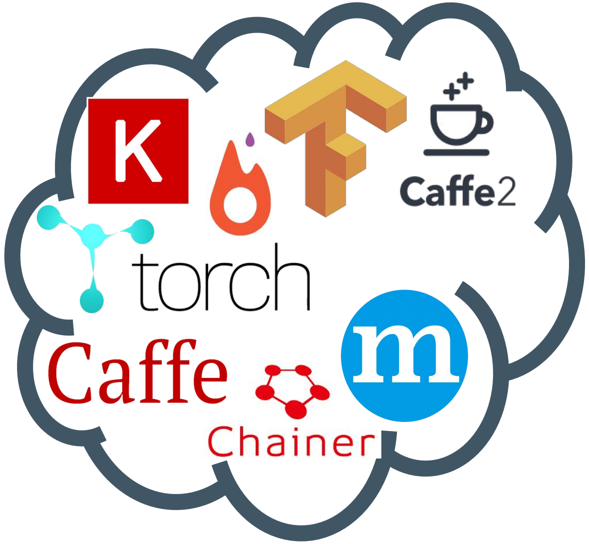
C3SR center for cognitive computing systems research

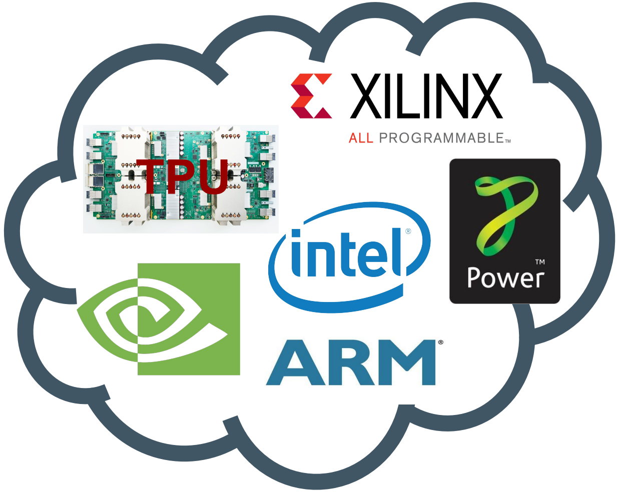
Motivation

Motivation

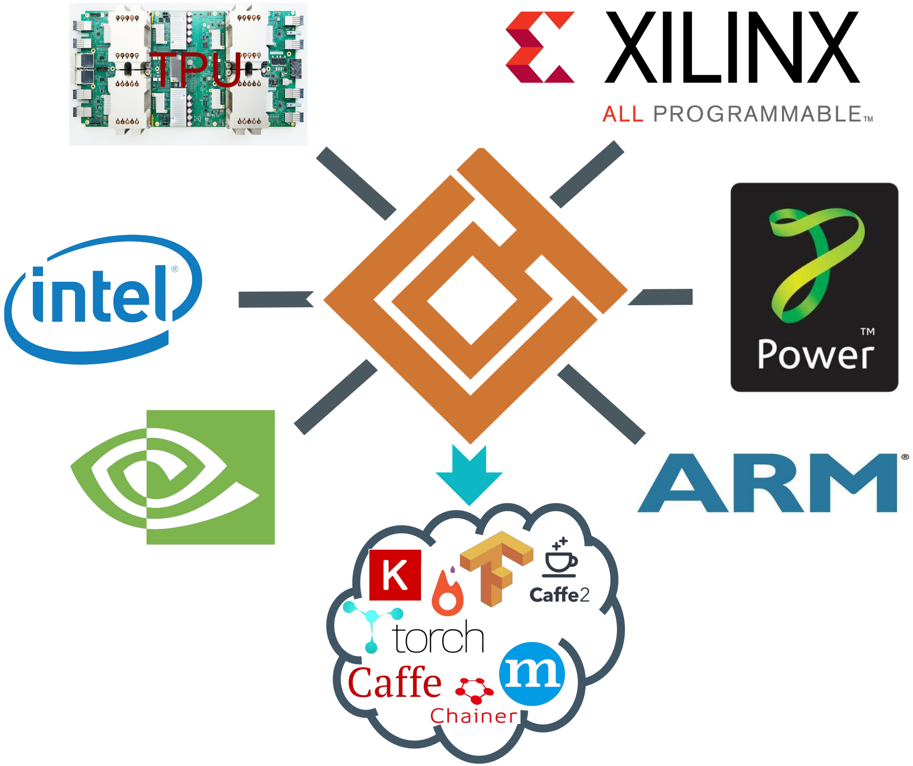
Diverse models, frameworks, and hardware infrastructures complicate deployment and usage

- Framework/model compatibility
- Software compatibility
- Hardware compatibility
- Hardware and system configuration





What we would like to have

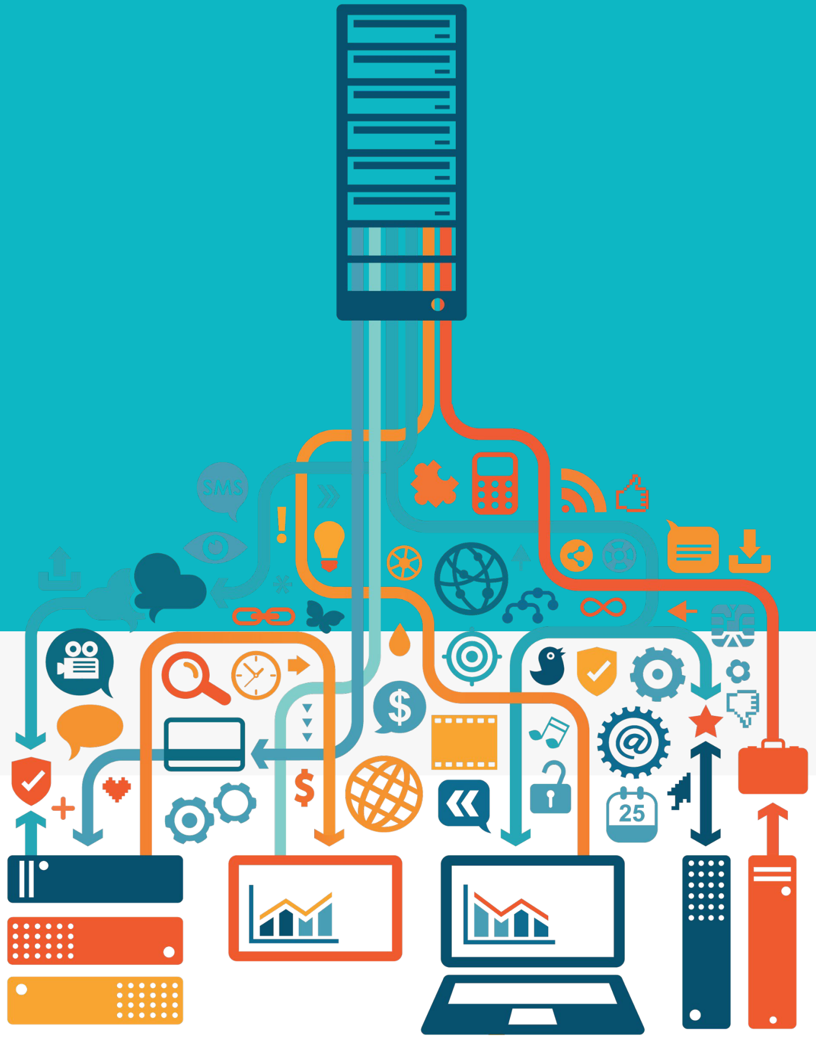


CarML - Cognitive ARtifacts for Machine Learning

An open source distributed platform to easily deploy and benchmark machine learning frameworks and models across hardware infrastructures, through a common interface.

- An experimentation platform for ML users
- A deployment platform for ML developers
- A benchmarking platform for systems architects

Impact



CarML for ML Users

CarML is a platform allowing users to evaluate and consume ML models and algorithms

- Try ML models with a click
- Optimize model, algorithm, and hardware selection based on:
  - Own dataset's accuracy
  - Cost, power, latency constraints
- Validate model's accuracy

CarML for ML Developers

CarML is a deployment platform allowing the public to try developer's models and get feedback

- Supports different input modalities
- Publishing ML model does not require writing code
  - Model defined by a manifest file
- Adding an ML framework requires writing a CarML predictor wrapper

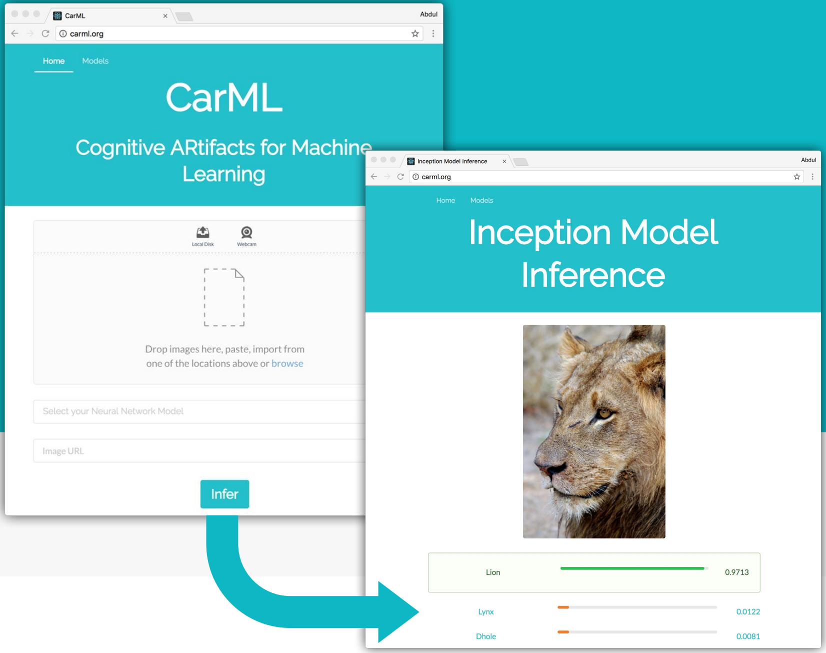
CarML for System Architects

CarML is a benchmarking platform to profile and understand system bottlenecks

- Distributed tracing and health monitoring
- Run real world end-to-end workloads on different hardware
- Informs research in:
  - Memory persistent objects for model loading
  - Near Memory Acceleration for preprocessing
  - Customized inference processors

Demo

www.carml.org



CarML

Cognitive ARtifacts for Machine Learning

Drop images here, paste, import from one of the locations above or browse

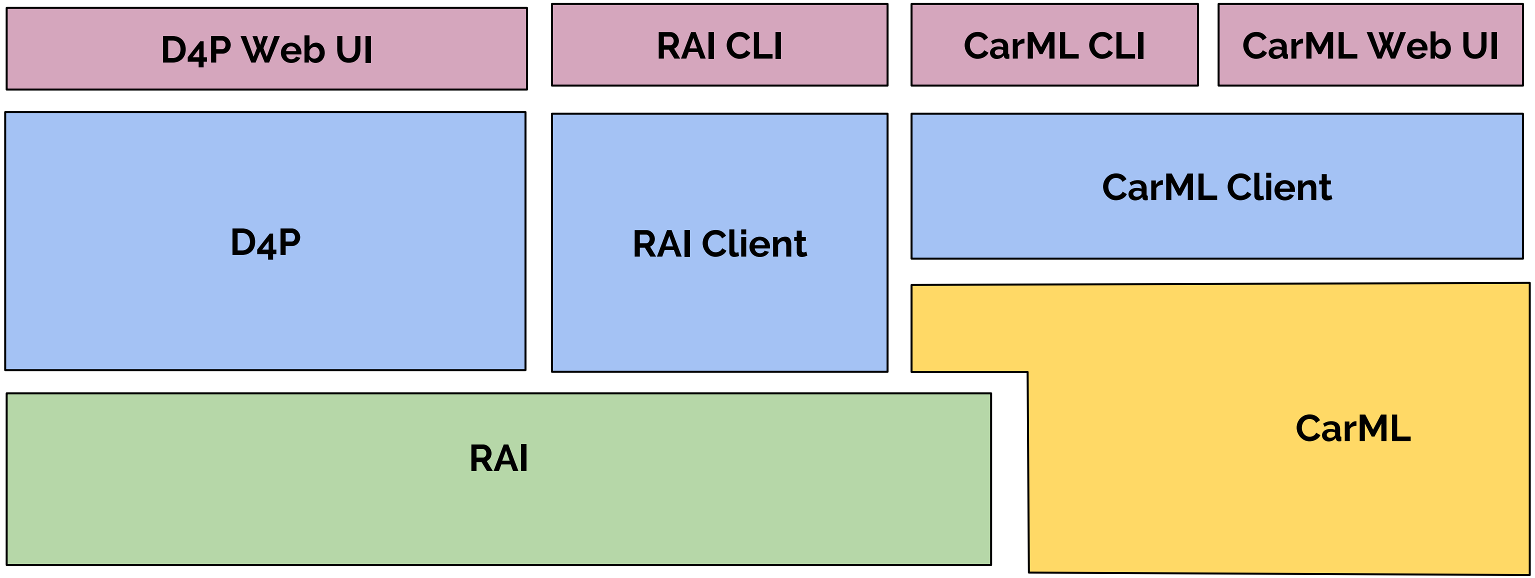
Select your Neural Network Model

Image URL

Inception Model Inference

Lion	0.9713
Lynx	0.0122
Dhole	0.0081

Architecture



D4P Web UI

RAI CLI

CarML CLI

CarML Web UI

D4P

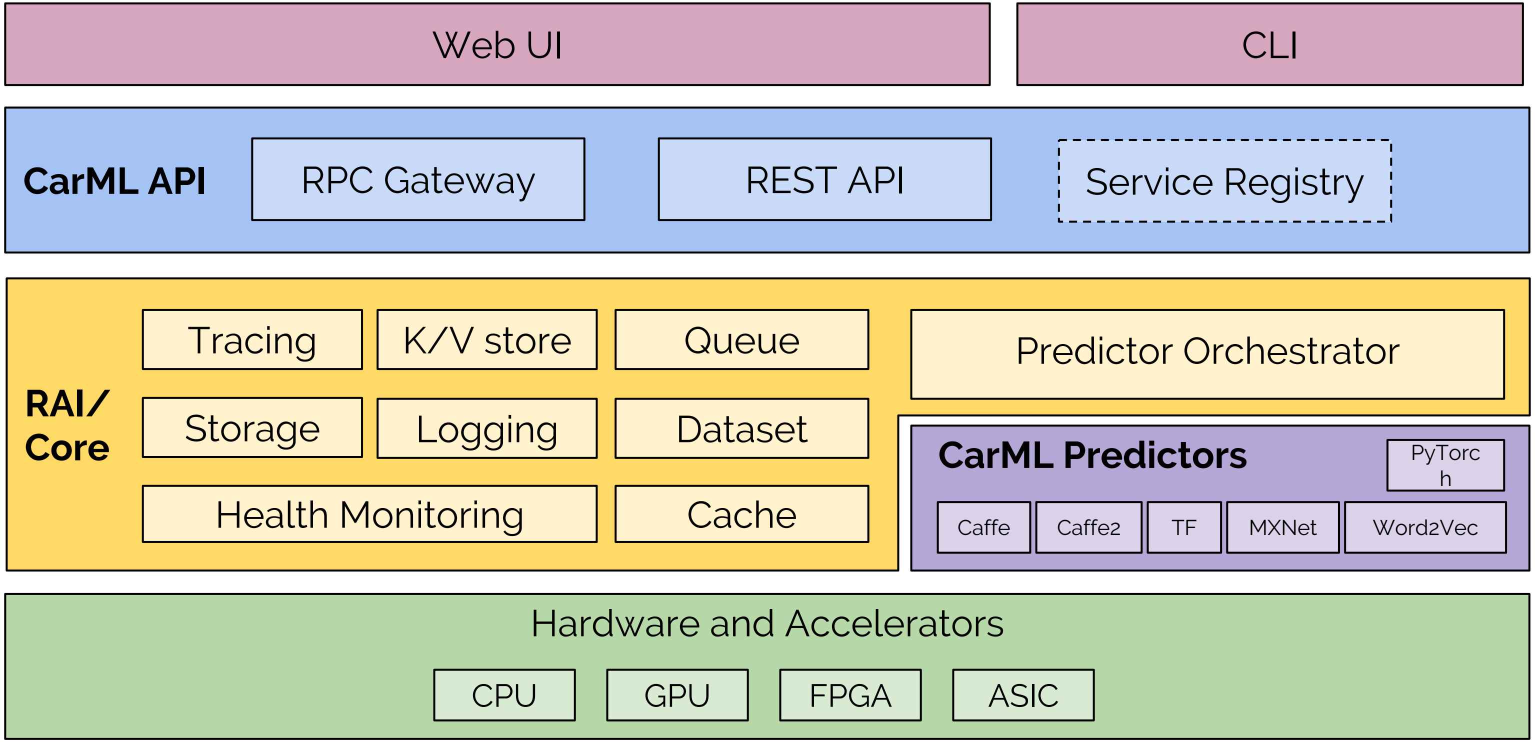
RAI Client

CarML Client

RAI

CarML

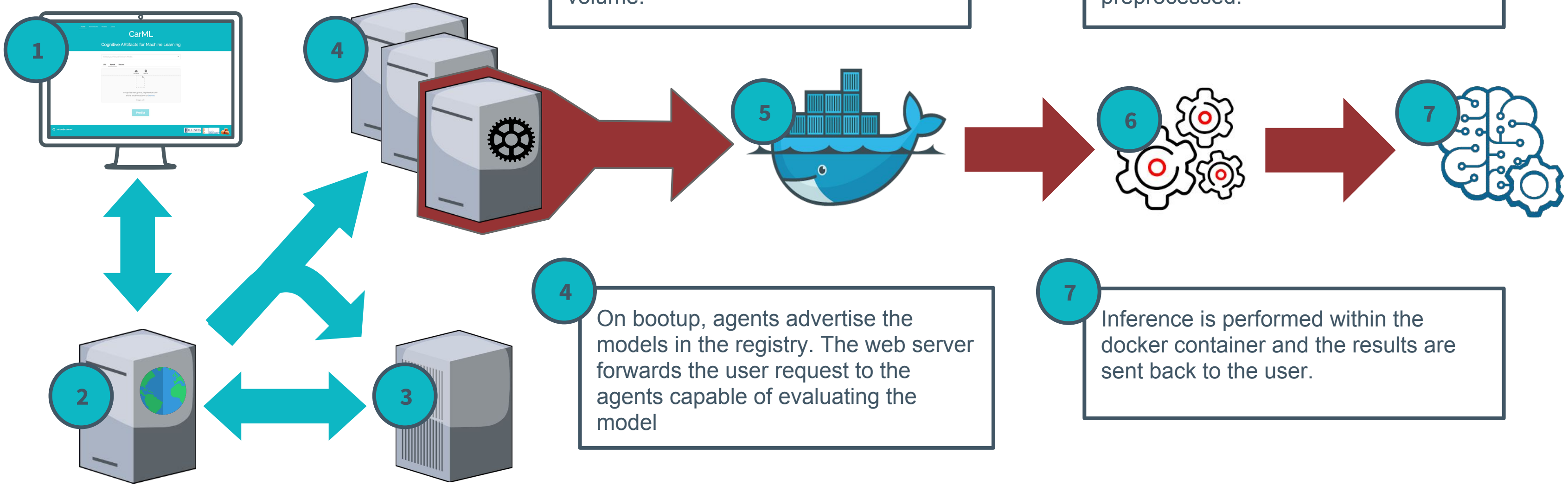
CarML Architecture



1 A use selects models and inputs through web UI or API.

5 The agent starts a docker container for the request. Data that is shared across executions are mounted as a shared volume.

6 Within the docker container, the model is downloaded, loaded into memory, and the user's inputs are preprocessed.



4 On bootup, agents advertise the models in the registry. The web server forwards the user request to the agents capable of evaluating the model

7 Inference is performed within the docker container and the results are sent back to the user.

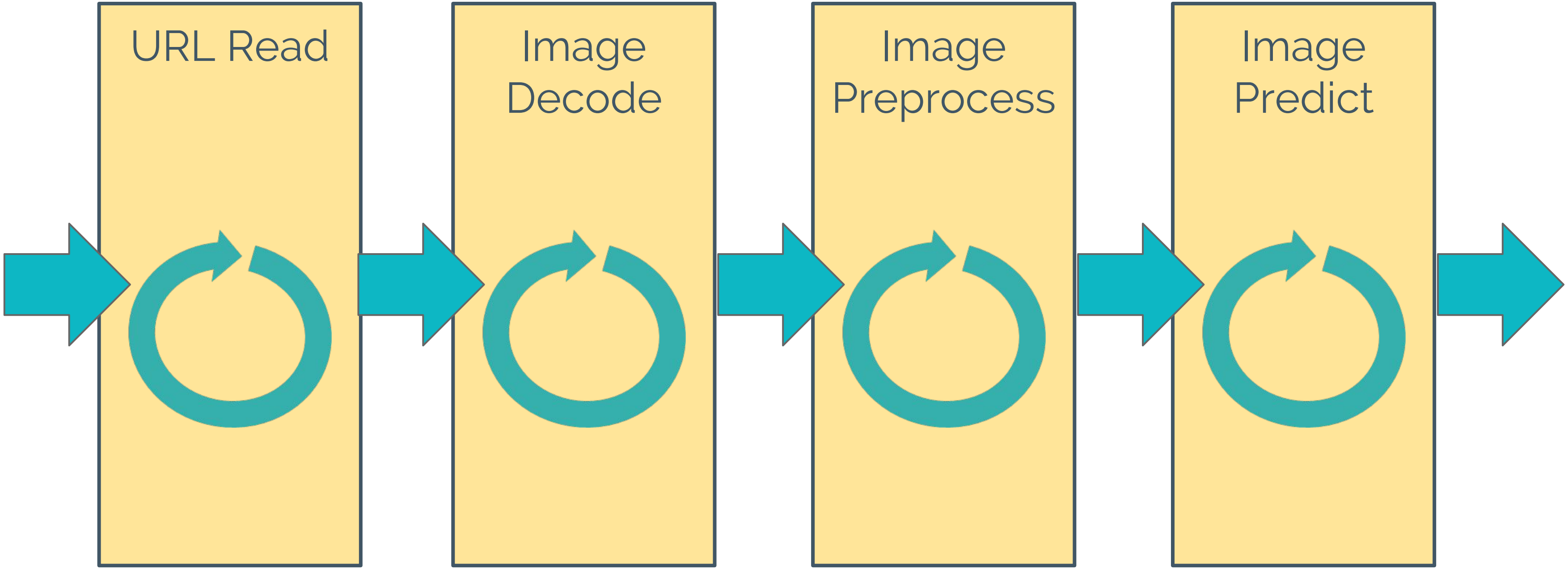
2 The web server accepts the user inputs and interacts with registry, tracing, and agents

3 The web server queries the agent registry to retrieve the address. The tracing process is started, if profiling is enabled.

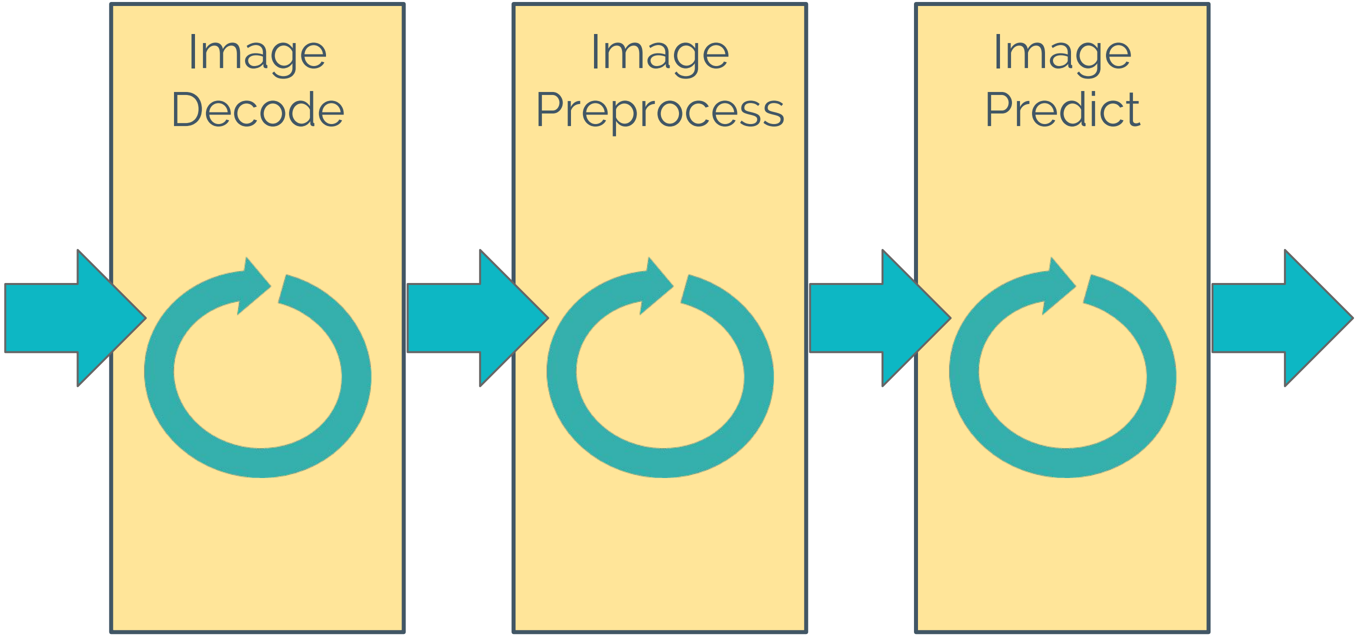
CarML Scalability

- A distributed and resilient system where the web server, registry, tracer, and agents can span nodes
- Horizontal and vertical scaling

CarML Image URL Predict Pipelining

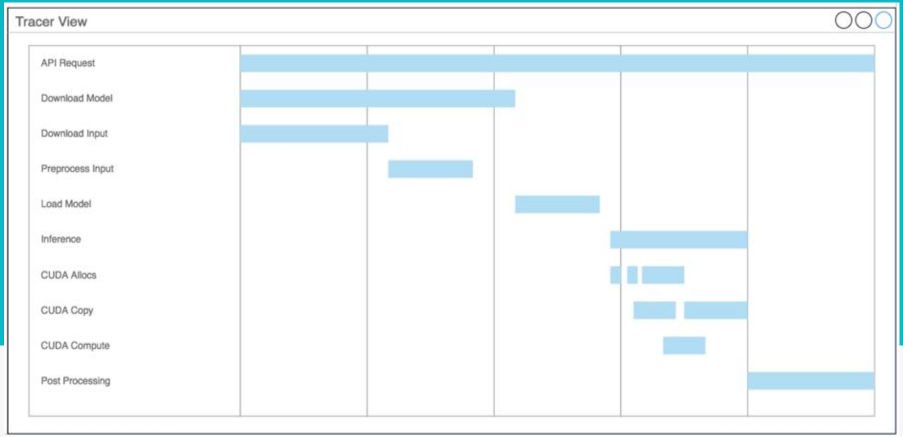


CarML Image Data Predict Pipelining



Tracing Demo

52.44.160.49:9411



Cognitive ARchitecture for Machine Learning

<< CarML >>

Cognitive AbstRaction for Machine Learning

<< CarML >>

Future Work

CarML Pipeline Builder

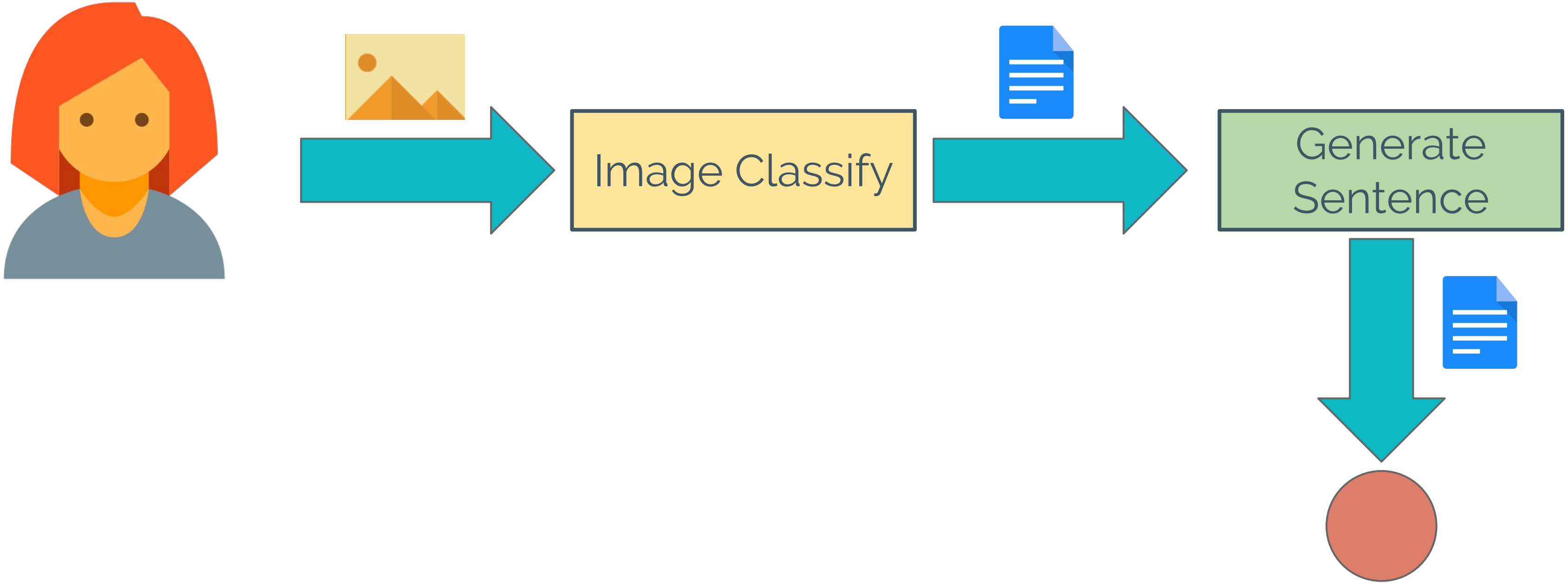


Image Classify

Generate Sentence

CarML Pipeline Builder

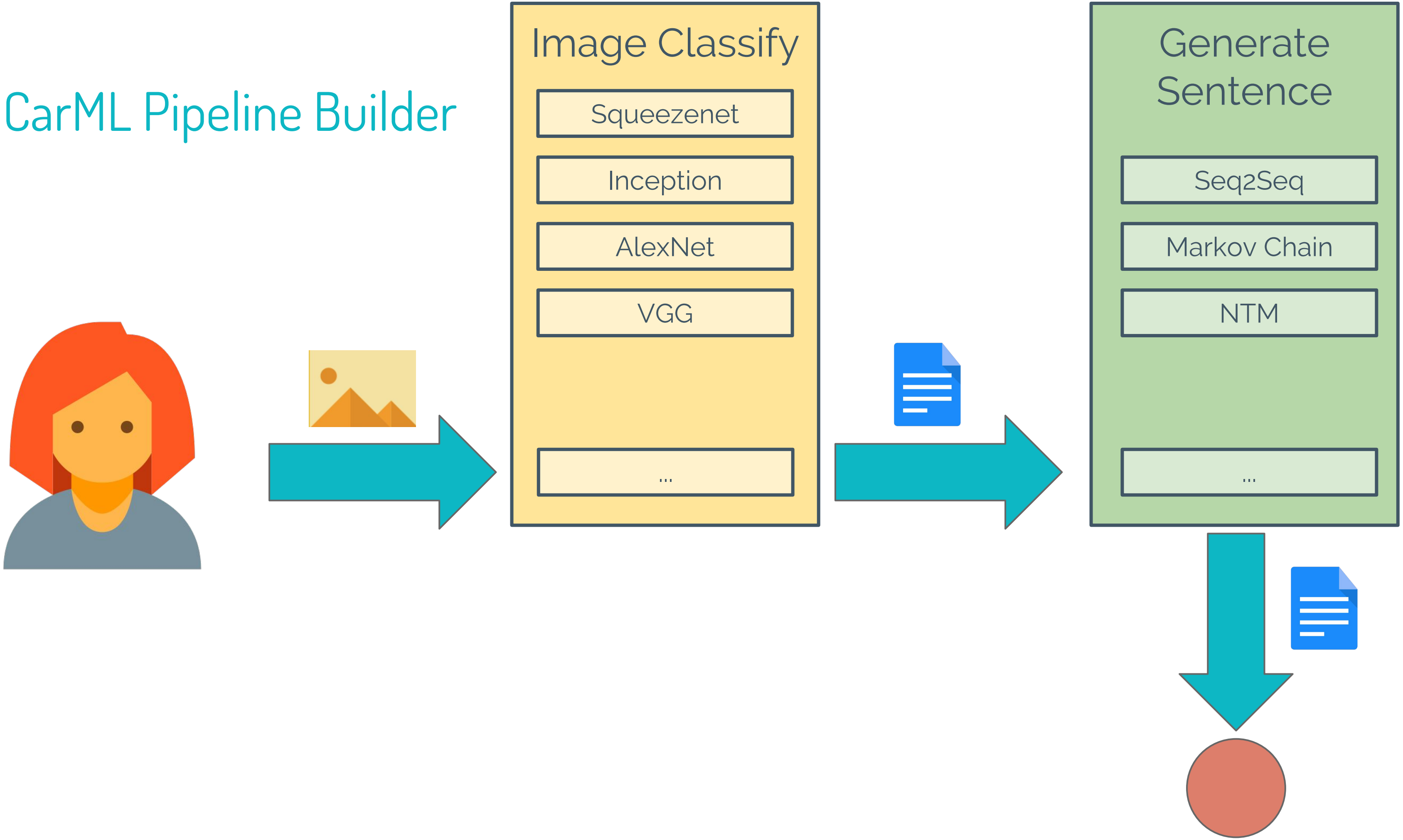


Image Classify

Squeezenet

Inception

AlexNet

VGG

...

Generate Sentence

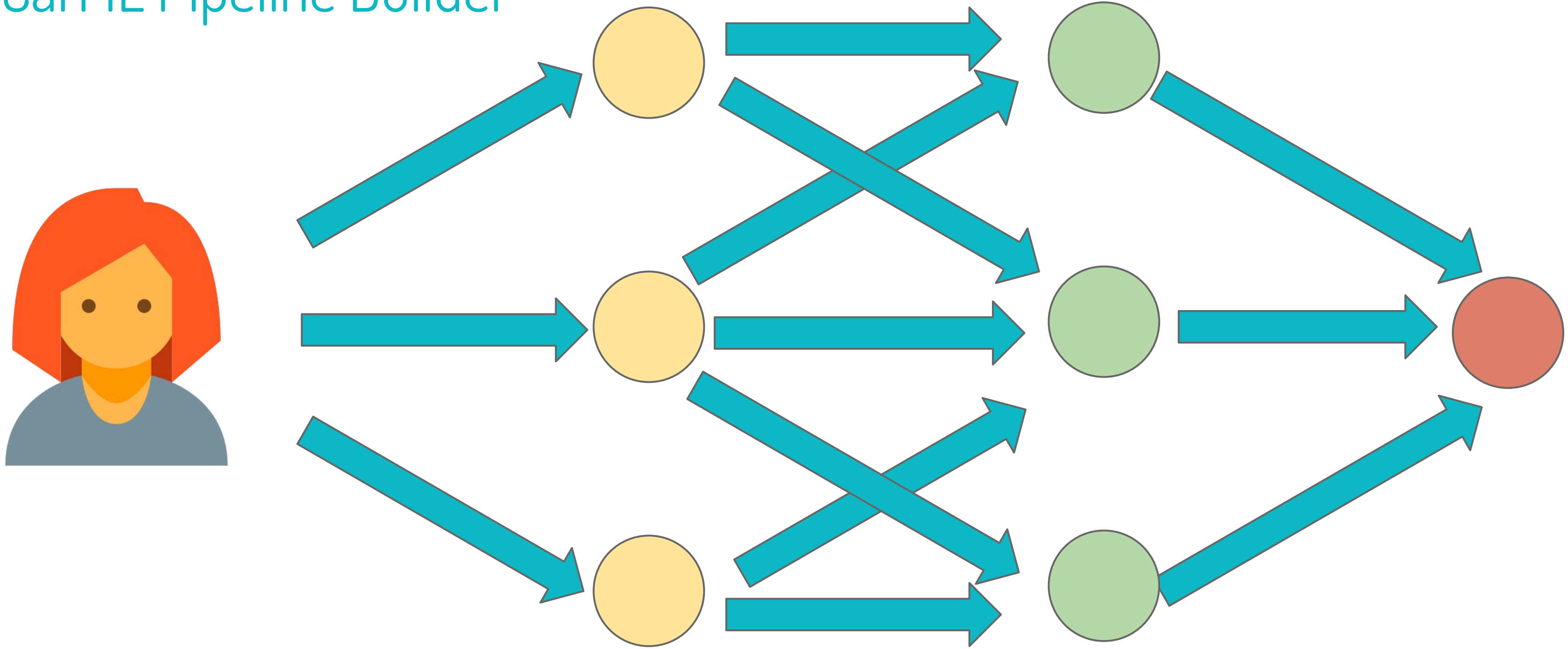
Seq2Seq

Markov Chain

NTM

...

CarML Pipeline Builder



Near Future

Next Steps

- Add more builtin models, frameworks, and datasets
- Perform profiling and characterize workloads across frameworks and machines
- Scheduling and resources management

Conclusion

- CarML simplifies ML deployment and usage
- CarML informs system designs based on real world end-to-end usage of ML models
- The objective is for CarML to be the **hub** to develop, evaluate, and experiment with ML/DL models

Questions/Comments

Thank you

CarML.org