SKILup℠ DAYS

by: DevOps Institute
ADVANCING THE HUMANS OF DEVOPS

Antonio Linari
Chief Technology Officer, NA
expert.ai
Email: alinari@expert.ai
Twitter: @advancedlogic

Cloud Native on the Edge

October 15th, 2020

expert.ai

About me

- Software Developer since I was 11
- 1st bug solved at 11 (a syntax error ;))
- 20 years in technology
- 15 years Natural Language Processing practitioner

Github: advancedlogic

- Go-freeling 830*
- GoOSe 363*

Agenda

- Edge Cluster (RPI, Nvidia)
- Natural Language Processing on ARM64
- Docker/Docker Swarm (KUBE in the future)
- Full stack: Golang and Svelte.js
- OpenFaaS (Function as a Service)
- DEMO

Green AI

Roy Schwartz* ◇ Jesse Dodge*◇♣ Noah A. Smith◇♡ Oren Etzioni◇

◇Allen Institute for AI, Seattle, Washington, USA
♣ Carnegie Mellon University, Pittsburgh, Pennsylvania, USA
♡ University of Washington, Seattle, Washington, USA

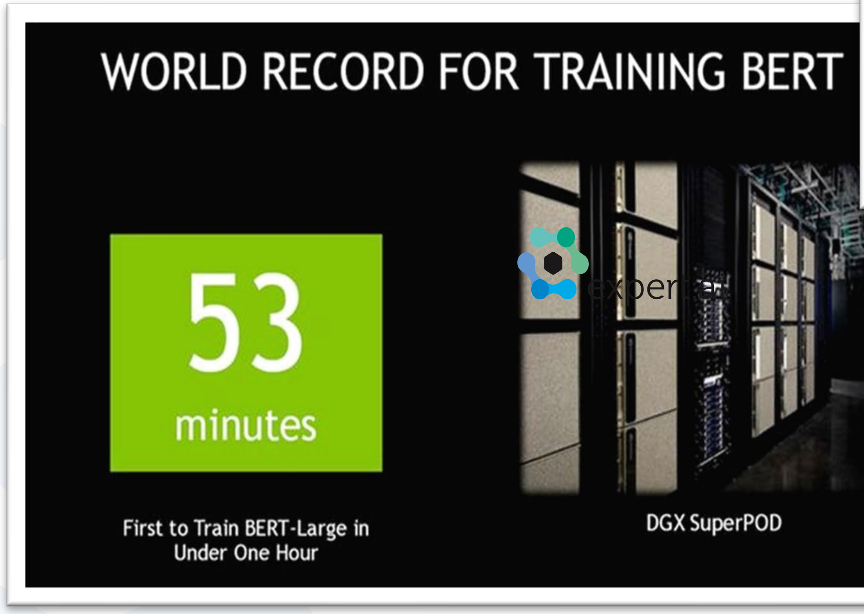
July 2019

Abstract

The computations required for deep learning research have been doubling every few months, resulting in an estimated 300,000x increase from 2012 to 2018 [2]. These computations have a surprisingly large carbon footprint [40]. Ironically, deep learning was inspired by the human brain, which is remarkably energy efficient. Moreover, the financial cost of the computations can make it difficult for academics, students, and researchers, in particular those from emerging economies, to engage in deep learning research.

This position paper advocates a practical solution by making **efficiency** an evaluation criterion for research alongside accuracy and related measures. In addition, we propose reporting the financial cost or "price tag" of developing, training, and running models to provide baselines for the investigation of increasingly efficient methods. Our goal is to make AI both greener and more inclusive—enabling any inspired undergraduate with a laptop to write high-quality research papers. Green AI is an emerging focus at the Allen Institute for AI.

The $40M infrastructure

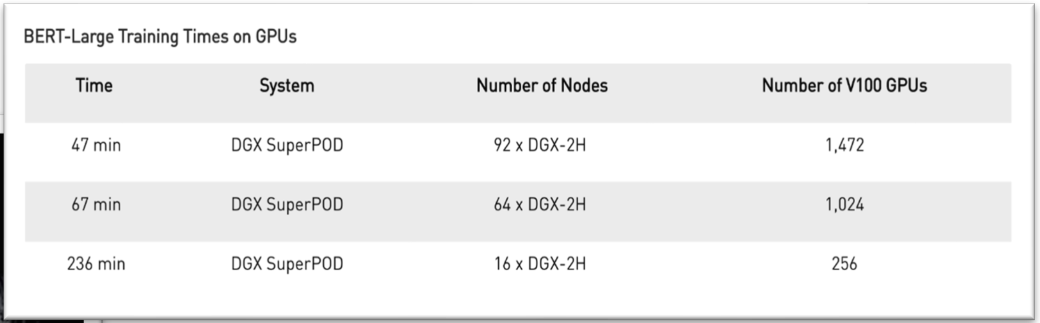


WORLD RECORD FOR TRAINING BERT

53 minutes

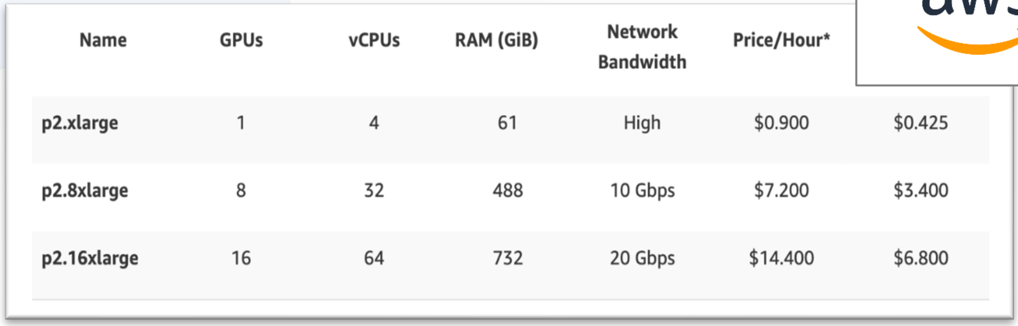
First to Train BERT-Large in Under One Hour

DGX SuperPOD



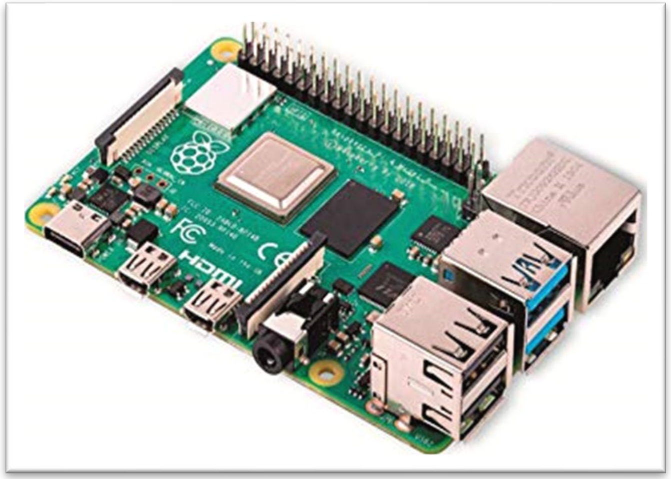
BERT-Large Training Times on GPUs

Time	System	Number of Nodes	Number of V100 GPUs
47 min	DGX SuperPOD	92 x DGX-2H	1,472
67 min	DGX SuperPOD	64 x DGX-2H	1,024
236 min	DGX SuperPOD	16 x DGX-2H	256

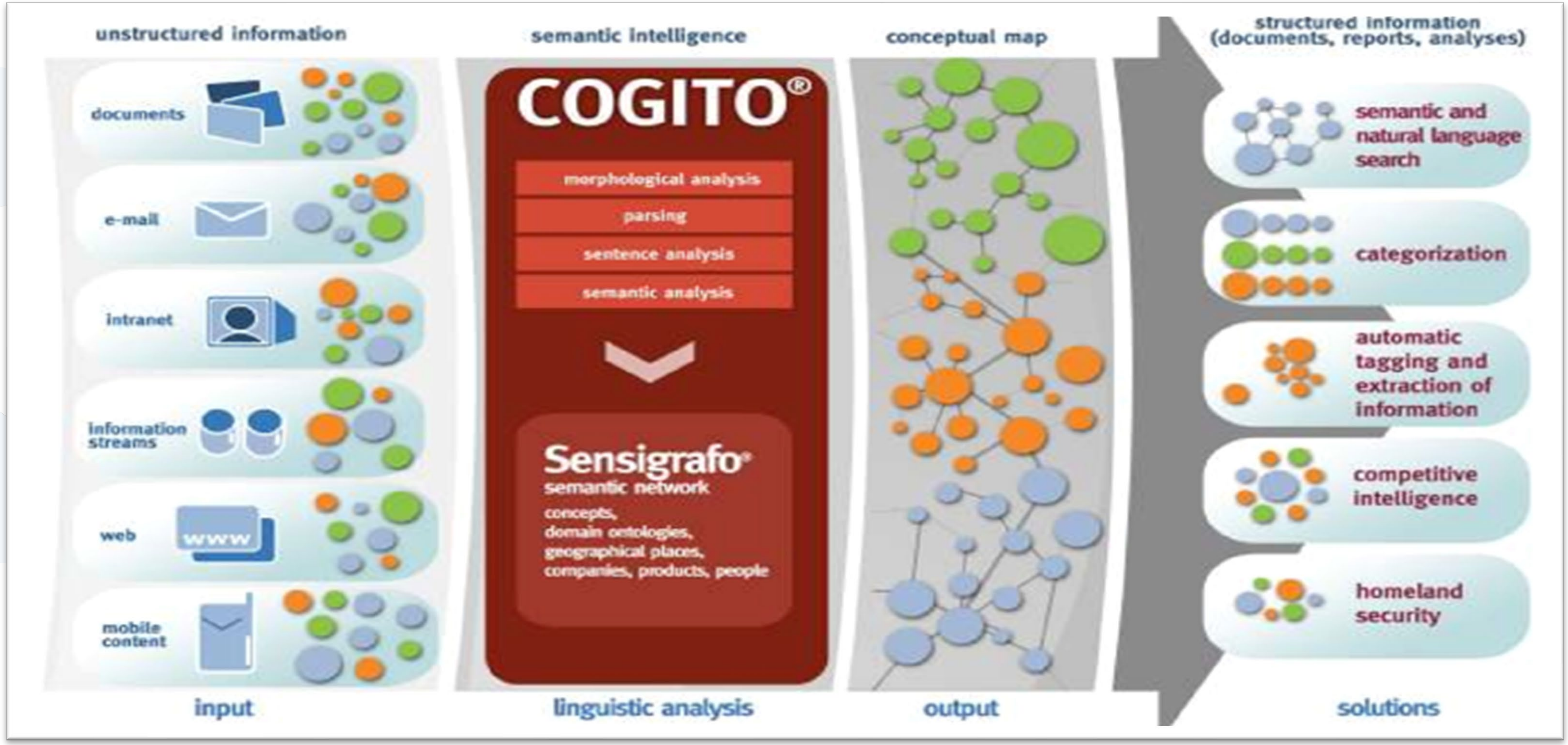


Name	GPUs	vCPUs	RAM (GiB)	Network Bandwidth	Price/Hour*	
p2.xlarge	1	4	61	High	$0.900	$0.425
p2.8xlarge	8	32	488	10 Gbps	$7.200	$3.400
p2.16xlarge	16	64	732	20 Gbps	$14.400	$6.800

The $40sh infrastructure

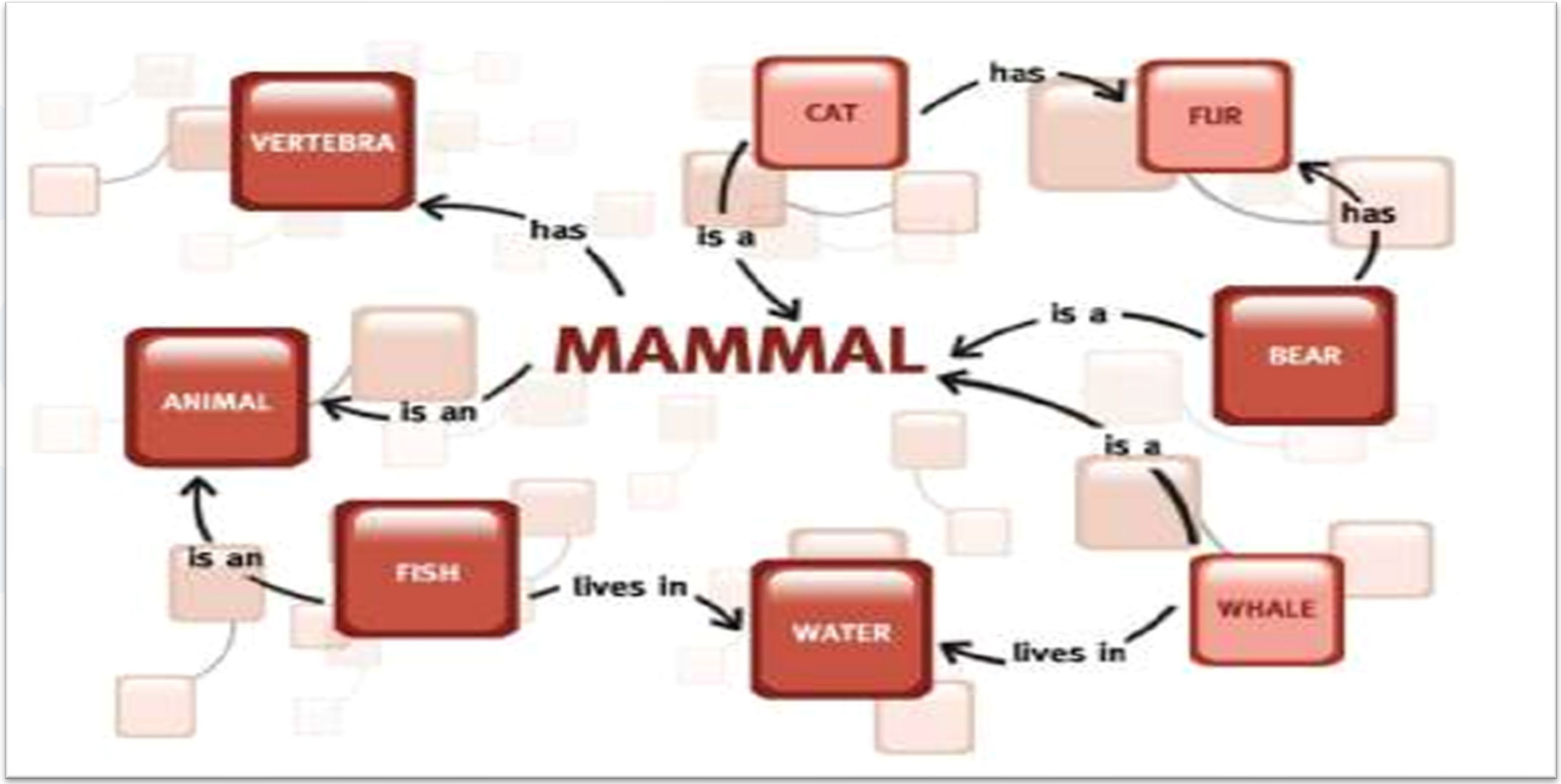


Word Sense Disambiguation

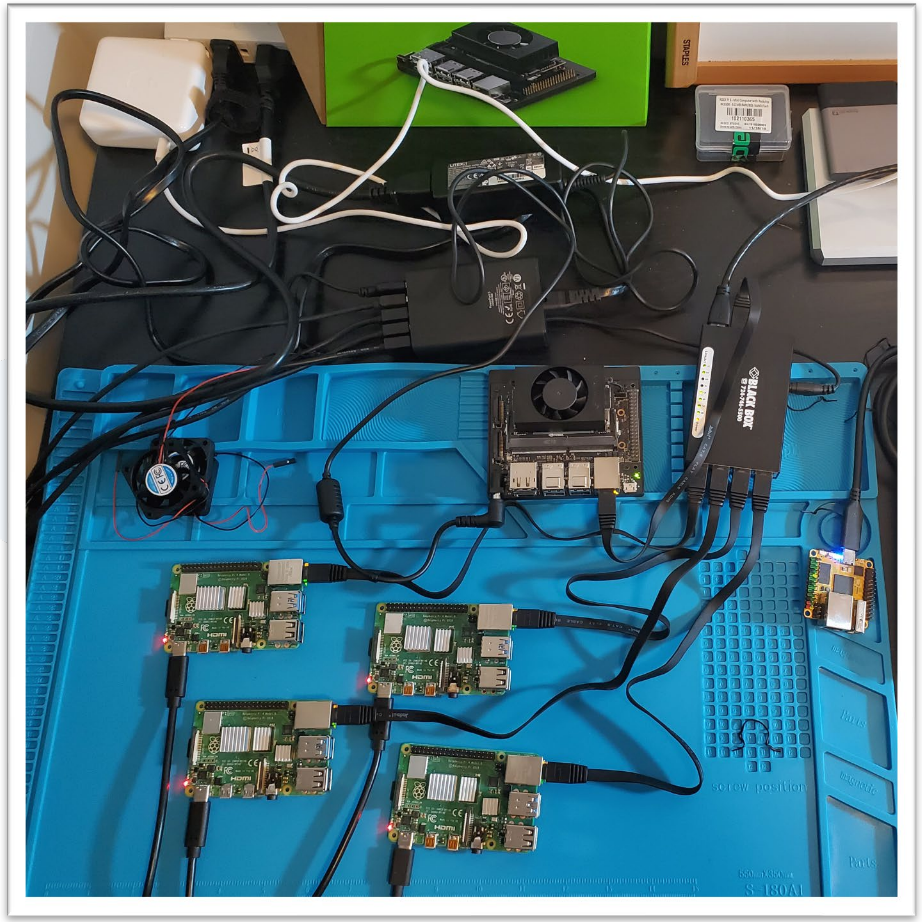


expert.ai

Semantic Network (Sensigrafo)



expert.ai



Docker Swarm Cluster:

1 Nvidia Jetson Xavier (Master)

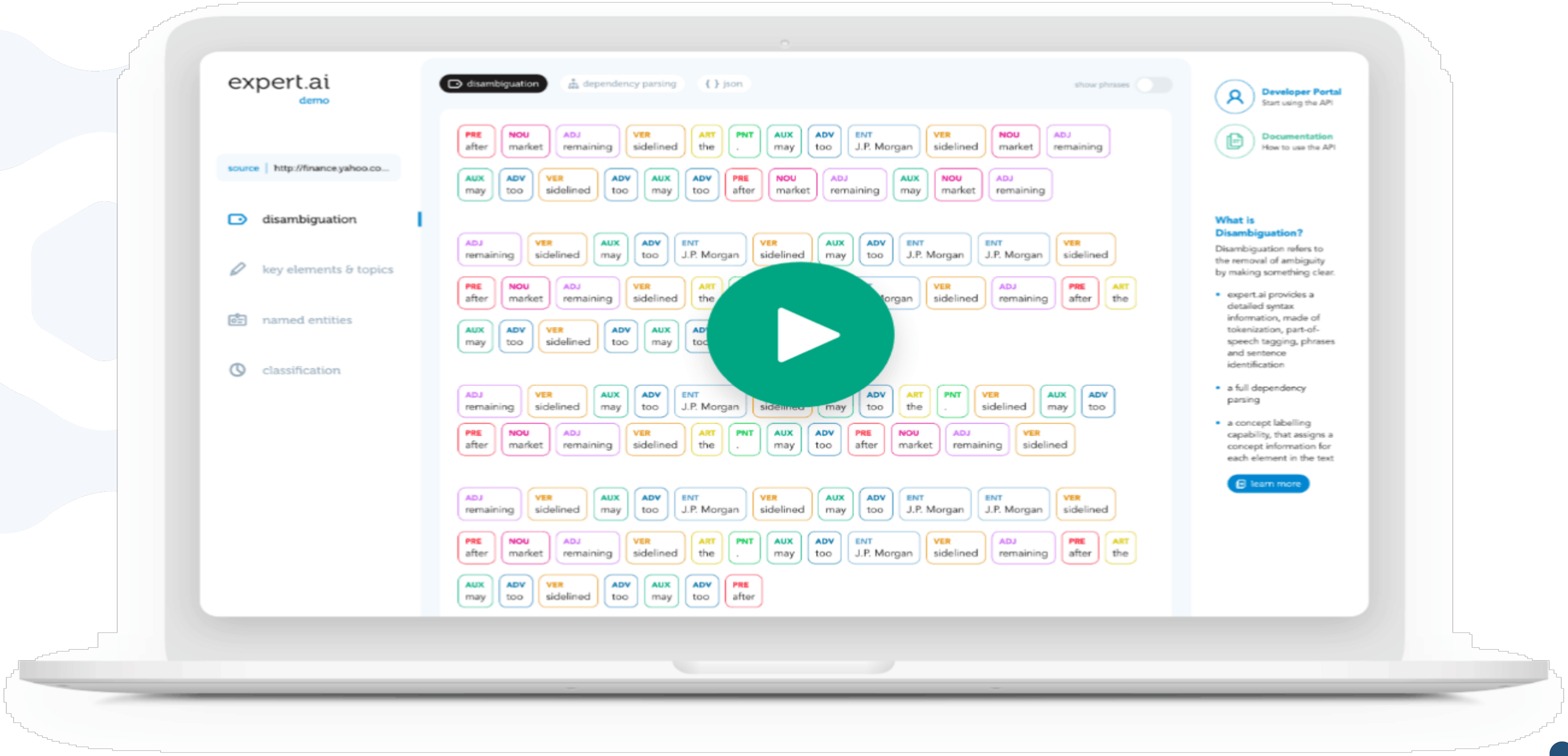
1 Nvidia Jetson Nano (Register)

4 raspberry pi (Nodes)

+1 rockpi s

SVELTE

Demo


expert.ai
demo
source | http://finance.yahoo.co...
disambiguation
key elements & topics
named entities
classification
disambiguation
dependency parsing
{ } json
show phrases
PRE after NOU market ADJ remaining VER sidelined ART the PNT . AUX may ADV too ENT J.P. Morgan VER sidelined NOU market ADJ remaining
AUX may ADV too VER sidelined ADV too AUX may ADV too PRE after NOU market ADJ remaining AUX may NOU market ADJ remaining
ADJ remaining VER sidelined AUX may ADV too ENT J.P. Morgan VER sidelined AUX may ADV too ENT J.P. Morgan ENT J.P. Morgan VER sidelined
PRE after NOU market ADJ remaining VER sidelined ART the Morgan VER sidelined ADJ remaining PRE after ART the
AUX may ADV too VER sidelined ADV too AUX may AD too
ADJ remaining VER sidelined AUX may ADV too ENT J.P. Morgan sidelined may ADV too ART the PNT . VER sidelined AUX may ADV too
PRE after NOU market ADJ remaining VER sidelined ART the PNT . AUX may ADV too PRE after NOU market ADJ remaining VER sidelined
ADJ remaining VER sidelined AUX may ADV too ENT J.P. Morgan VER sidelined AUX may ADV too ENT J.P. Morgan ENT J.P. Morgan VER sidelined
PRE after NOU market ADJ remaining VER sidelined ART the PNT . AUX may ADV too ENT J.P. Morgan VER sidelined ADJ remaining PRE after ART the
AUX may ADV too VER sidelined ADV too AUX may ADV too PRE after
Developer Portal
Start using the API
Documentation
How to use the API
learn more

What is Disambiguation?

Disambiguation refers to the removal of ambiguity by making something clear.

- expert.ai provides a detailed syntax information, made of tokenization, part-of-speech tagging, phrases and sentence identification
- a full dependency parsing
- a concept labelling capability, that assigns a concept information for each element in the text

expert.ai

THANK YOU!

Meet me in the Network Chat Lounge for questions

SKILup℠ DAYS by: DevOps Institute ADVANCING THE HUMANS OF DEVOPS