Open discussion on the potential standard dataset

Albert Cabellos

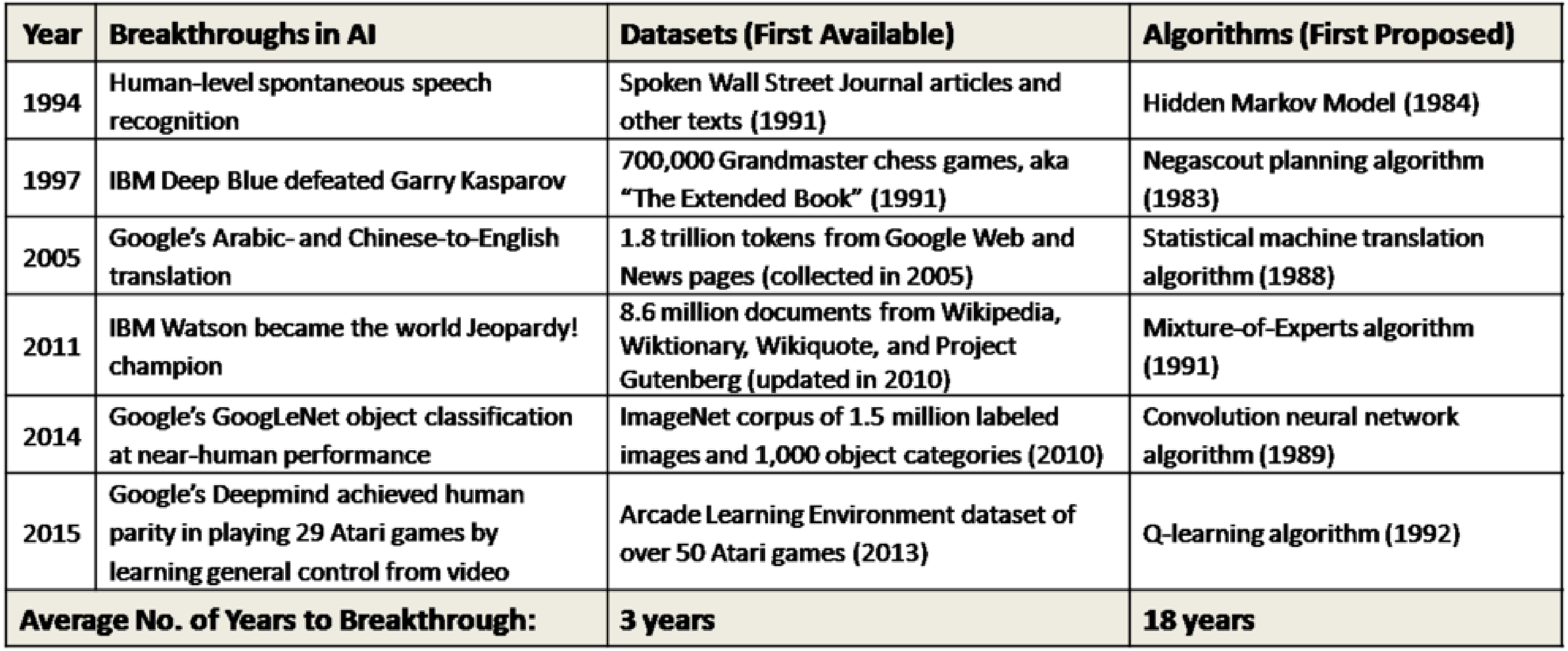
IETF Berlin, (proposed) NML WG

July 2016

Scope & Context

- In NML#3 (Athens) we discussed about the standardization of datasets for Network ML
- In this presentation:
  1. Bring attention and discuss the importance of datasets in NML
  2. Discuss the importance of datasets in other ML fields
  3. Trigger discussion about standardizaton of datasets for NML

How important are datasets?



Year	Breakthroughs in AI	Datasets (First Available)	Algorithms (First Proposed)
1994	Human-level spontaneous speech recognition	Spoken Wall Street Journal articles and other texts (1991)	Hidden Markov Model (1984)
1997	IBM Deep Blue defeated Garry Kasparov	700,000 Grandmaster chess games, aka "The Extended Book" (1991)	Negascout planning algorithm (1983)
2005	Google's Arabic- and Chinese-to-English translation	1.8 trillion tokens from Google Web and News pages (collected in 2005)	Statistical machine translation algorithm (1988)
2011	IBM Watson became the world Jeopardy! champion	8.6 million documents from Wikipedia, Wiktionary, Wikiquote, and Project Gutenberg (updated in 2010)	Mixture-of-Experts algorithm (1991)
2014	Google's GoogLeNet object classification at near-human performance	ImageNet corpus of 1.5 million labeled images and 1,000 object categories (2010)	Convolution neural network algorithm (1989)
2015	Google's Deepmind achieved human parity in playing 29 Atari games by learning general control from video	Arcade Learning Environment dataset of over 50 Atari games (2013)	Q-learning algorithm (1992)
Average No. of Years to Breakthrough:		**3 years**	**18 years**

Table from: Datasets over Algorithms (SpaceMachine, March 2015)
http://www.spacemachine.net/views/2016/3/datasets-overalgorithms
Original content from: Wissner-Gross, Alexander (2016). Datasets Over Algorithms. Edge. Retrieved from: https://www.edge.org/response- detail/26587

How important are datasets?

Datasets might be the key limiting factor to the development of new AI techniques

Table from: Datasets over Algorithms (SpaceMachine, March 2015)

http://www.spacemachine.net/views/2016/3/datasets-overalgorithms

Original content from: Wissner-Gross, Alexander (2016). Datasets Over Algorithms. Edge. Retrieved from: https://www.edge.org/response- detail/26587

Benefits of public datasets for Network ML

- ML-based algorithms do not provide guarantees (as opposed to traditional networking algorithms)
  - How can we make sure that our newly trained AI algorithm will work in different (untrained) scenarios?
- Provides a benchmark
  - Is the new algorithm better than the old one?
- Encourages research
- Allows for reproductible research

Datasets in AI and Networking

- Several AI fields have well-known public datasets, examples:
  - IMAGENET hosts 14M images for computer Vision. IMAGENET Challenge
  - Yahoo News Feed including 20M anonymized user-data
- The networking field has also a long tradition of public datasets, examples:
  - The CAIDA Anonymized Internet Traces 2012 Dataset
  - RIPE Atlas
  - CRAWDAD: A Community Resource for Archiving Wireless Data At Dartmouth

Open discussion

- Should this WG promote public datasets for NML?
- How NML datasets are different from already existing network datasets?
- What are the privacy implications of such datasets?
  - Are there other associated risks?
- Can we help by developing a standard?
- If so, what are the relevant aspects of such standard?
  - Traffic features
  - Benchmark
  - Anonymization techniques