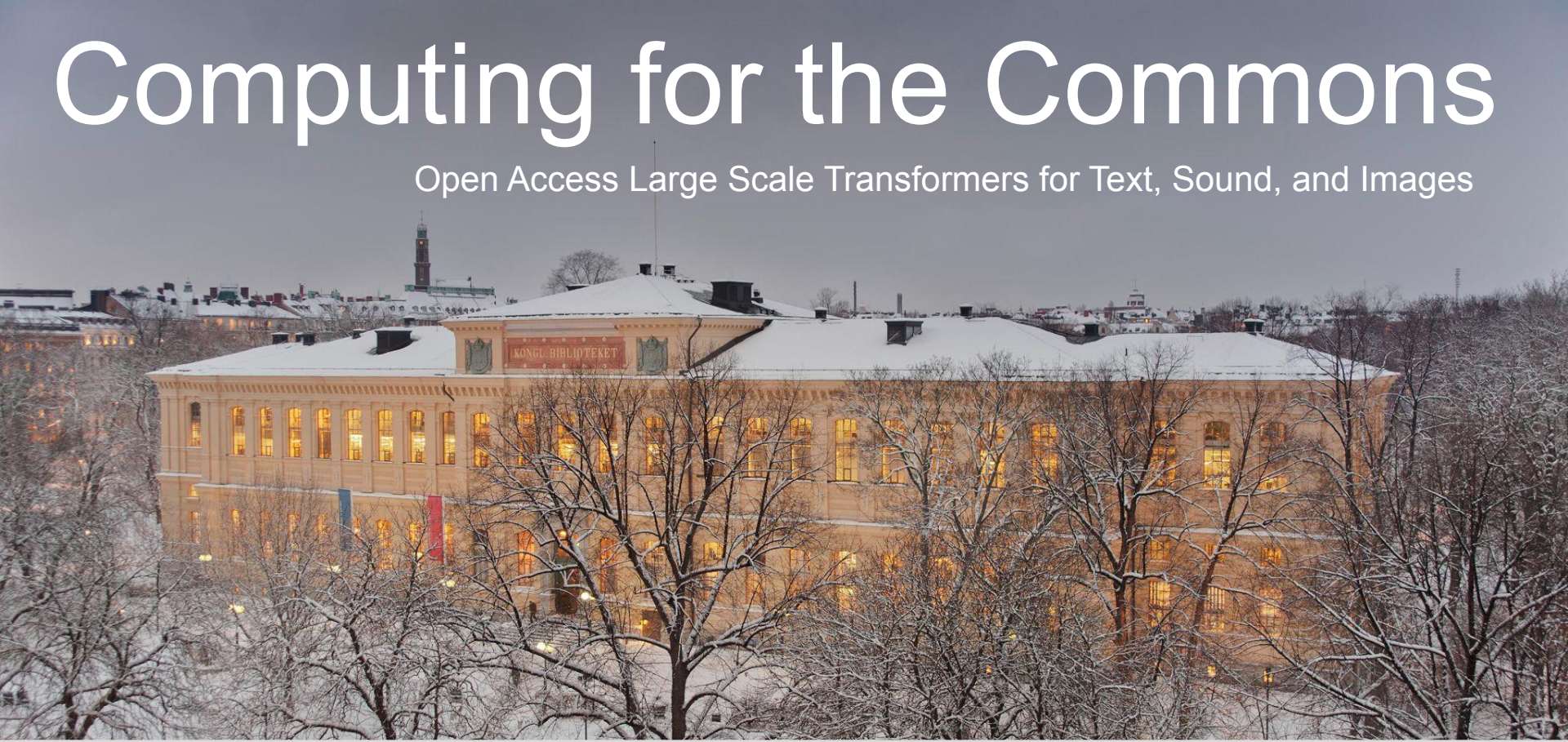
Computing for the Commons

Open Access Large Scale Transformers for Text, Sound, and Images



Love Börjeson, PhD
Director of KBLab, National Library of Sweden (KB)
love.borjeson@kb.se

KBLab

KUNGL BIBLIOTEKET

The National Library, KB - A Governmental Hoarder

Sweden has legal deposit laws installed in 1661 for everything printed.

During the the twentieth century, the law was gradually extended to include all modalities (text, sound, images and videos) and all formats (physical and digital).

As a result, the KB has vast and ever growing collections, closing in on 26 Petabyte of data.

KB has the the largest, broadest and deepest collection of humanistic data for the Swedish language.

The collections includes objects such as..:

- Books
- Commercial leaflets
- Computer games
- School photos
- Hand-written manuscripts
- TV and radio broadcasts
- Digital newspapers

...and etc.

The not so Common Commons

The KB collections are part of the commons - a national resource that should to the greatest possible extent be made available for ***anyone*** to use

Caveat: data in the collections typically fall under copyright and GDPR restrictions

Access directly to the collections can be granted on premises at the library

However: data such as KB's has an even larger potential as raw material in the current explosion of artificial intelligence

How can this potential be unlocked?

The Transformer Transformation

- In 2018 Google released a text based language understanding model, Bidirectional Encoder Representations from Transformers, commonly known as ***BERT***
- BERT was the first language model that could perform on par with, or even better than, humans on several linguistic tasks
- BERT is characterized by its transformer architecture of a very large network of artificial neurons.
- To train this network, you typically need a LOT of data and a LOT of computational resources

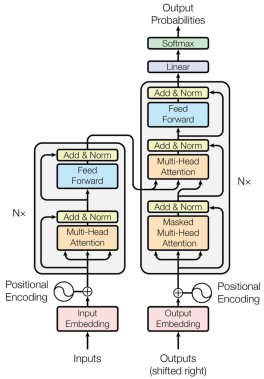
-> Not very accessible for the commons...

- However, BERT, and subsequent transformer-based models, have a very attractive feature - their general capability can be transferred to downstream specific applications

Data

Model with general ability

Training



Transferred capability

Domain and/or function-specific model

Hand annotated data

contentSkip to site indexPoliticsSubscribeLog InSubscribeLog InToday's PaperAdvertisementSupported ORG byF.B.I. Agent Peter Strzok PERSON , Who Criticized Trump PERSON in Texts, Is FiredImagePeter Strzok, a top F.B.I. GPE counterintelligence agent who was taken off the special counsel investigation after his disparaging texts about President Trump PERSON were uncovered, was fired. CreditT.J. Kirkpatrick PERSON for The New York TimesBy Adam Goldman ORG and Michael S. SchmidtAug PERSON . 13 CARDINAL , 2018WASHINGTON CARDINAL — Peter Strzok PERSON , the F.B.I. GPE senior counterintelligence agent who disparaged President Trump PERSON in inflammatory text messages and helped oversee the Hillary Clinton PERSON email and Russia GPE investigations, has been fired for violating bureau policies, Mr. Strzok PERSON 's lawyer said Monday DATE . Mr. Trump and his allies seized on the texts — exchanged during the 2016 DATE campaign with a former F.B.I. GPE lawyer, Lisa Page — in PERSON assailing the Russia GPE investigation as an illegitimate "witch hunt." Mr. Strzok PERSON , who rose over 20 years DATE at the F.B.I. GPE to become one of its most experienced counterintelligence agents, was a key figure in the early months DATE of the inquiry.Along with writing the texts, Mr. Strzok PERSON was accused of sending a highly sensitive search warrant to his personal email account.The F.B.I. GPE had been under immense political pressure by Mr. Trump PERSON to dismiss Mr. Strzok PERSON , who was removed last summer DATE from the staff of the special counsel, Robert S. Mueller III PERSON . The president has repeatedly denounced Mr. Strzok PERSON in posts on

Fine-tuning

Matchen spelas på Djurgårdens idrottsplats mellan Dj

Compute

Computation time on cpu: 0.039 s

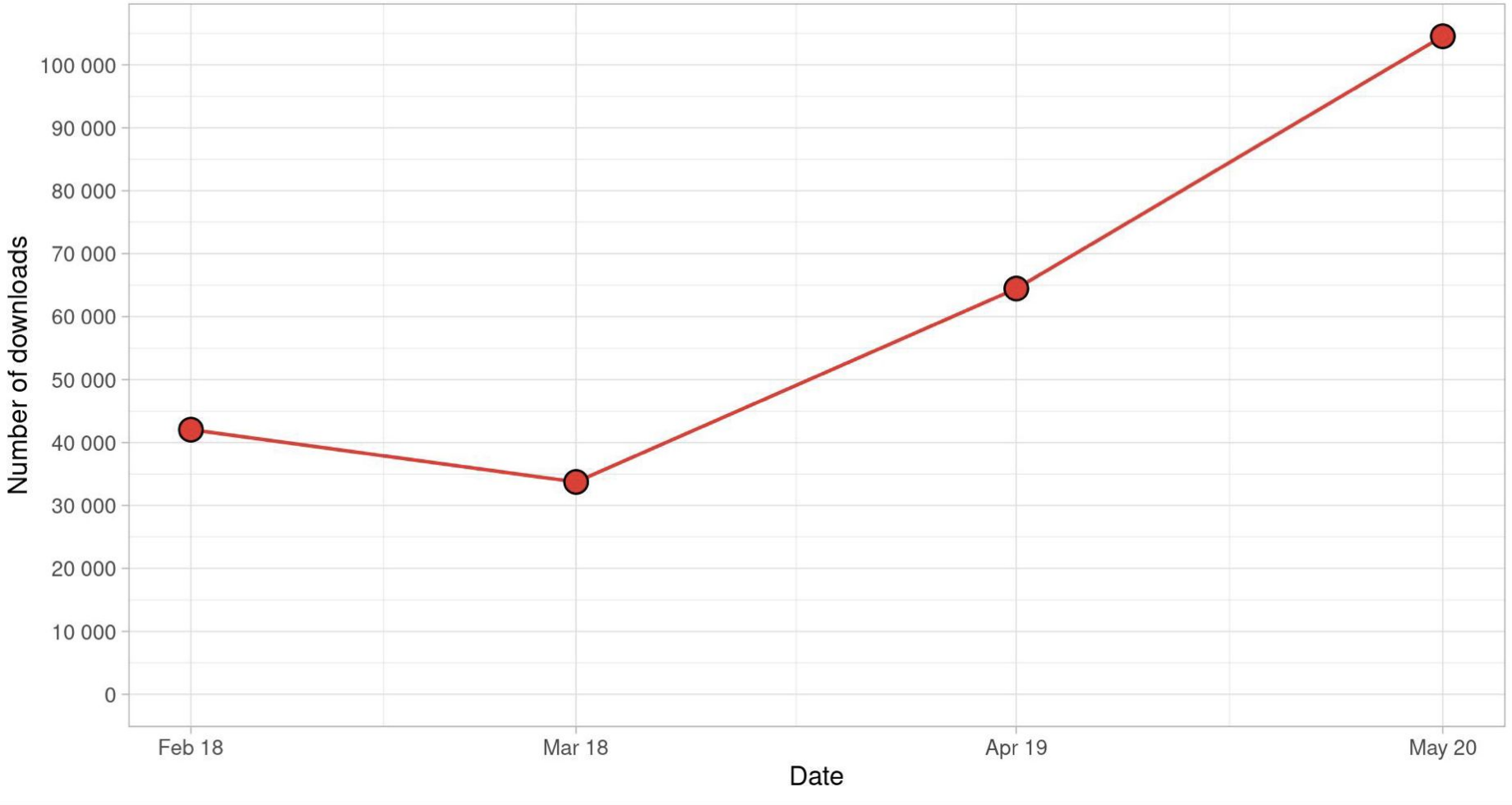
Matchen spelas på Djurgårdens idrottsplats LOC mellan Djurgårdens IF ORG och AIK ORG klockan 17 TME

English, Chinese and Multilingual. And Swedish.

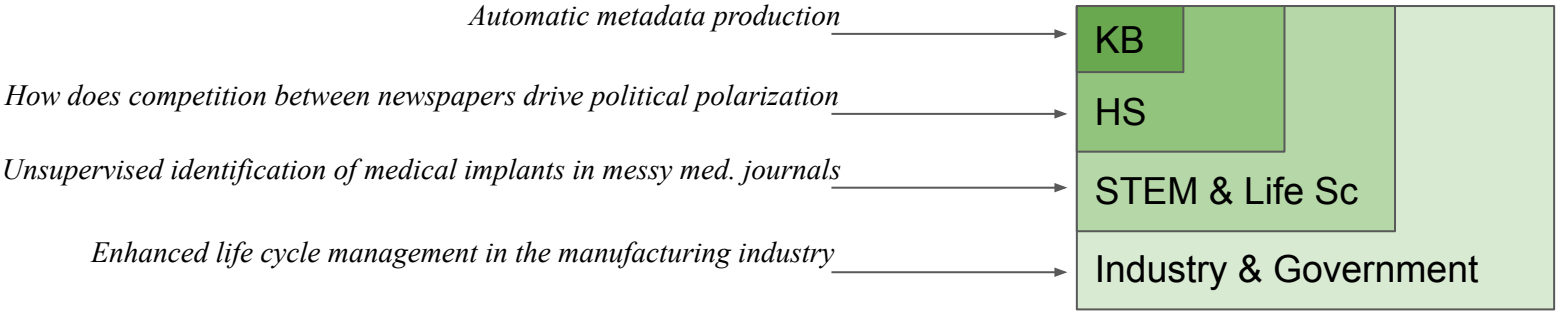
- Google released their models in an English, Chinese and Multilingual.
- KBLab, like a parasite, used the BERT architecture to train a Swedish BERT - **KB-BERT** -that outperforms Google's multilingual model that includes Swedish
- All models are published openly, free for anyone to use
- Estimated economical societal value of KB-BERT alone is several times bigger than the whole budget for the National Library
- KBLab train models across modalities, and KBLab models for text and sound are typically SOTA for the Swedish language
    - We cheat! And use superior KB data…
- Currently 28 transformer models under CC0 from KBLab available at Hugginface

KBLab's models is a way to ***transfer the full potential of KB's collection to the commons*** and to contribute to the digital transformation of society, ultimately supporting high quality research and democratic development.

Do the models really get used?


Number of downloads
100 000
90 000
80 000
70 000
60 000
50 000
40 000
30 000
20 000
10 000
0
Feb 18
Mar 18
Apr 19
May 20
Date

Examples of downstream applications for KB-BERT



None of these applications where foreseen ->

...the biggest potential is inherently unknown

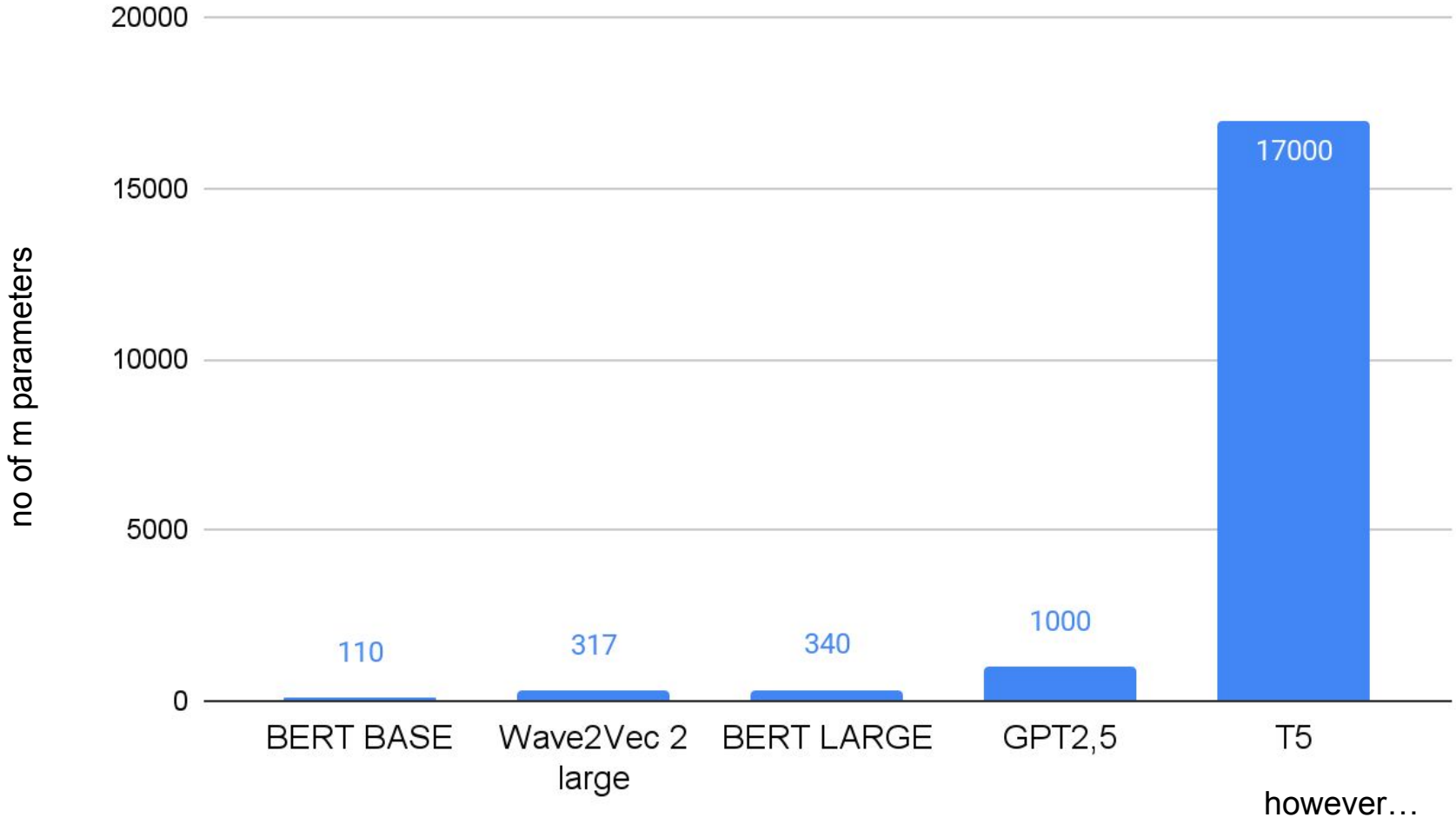
However, training takes computational resources.

… and with the constant increase of model sizes, the need for computational resources is ever growing.

Solution? High Performance Computing

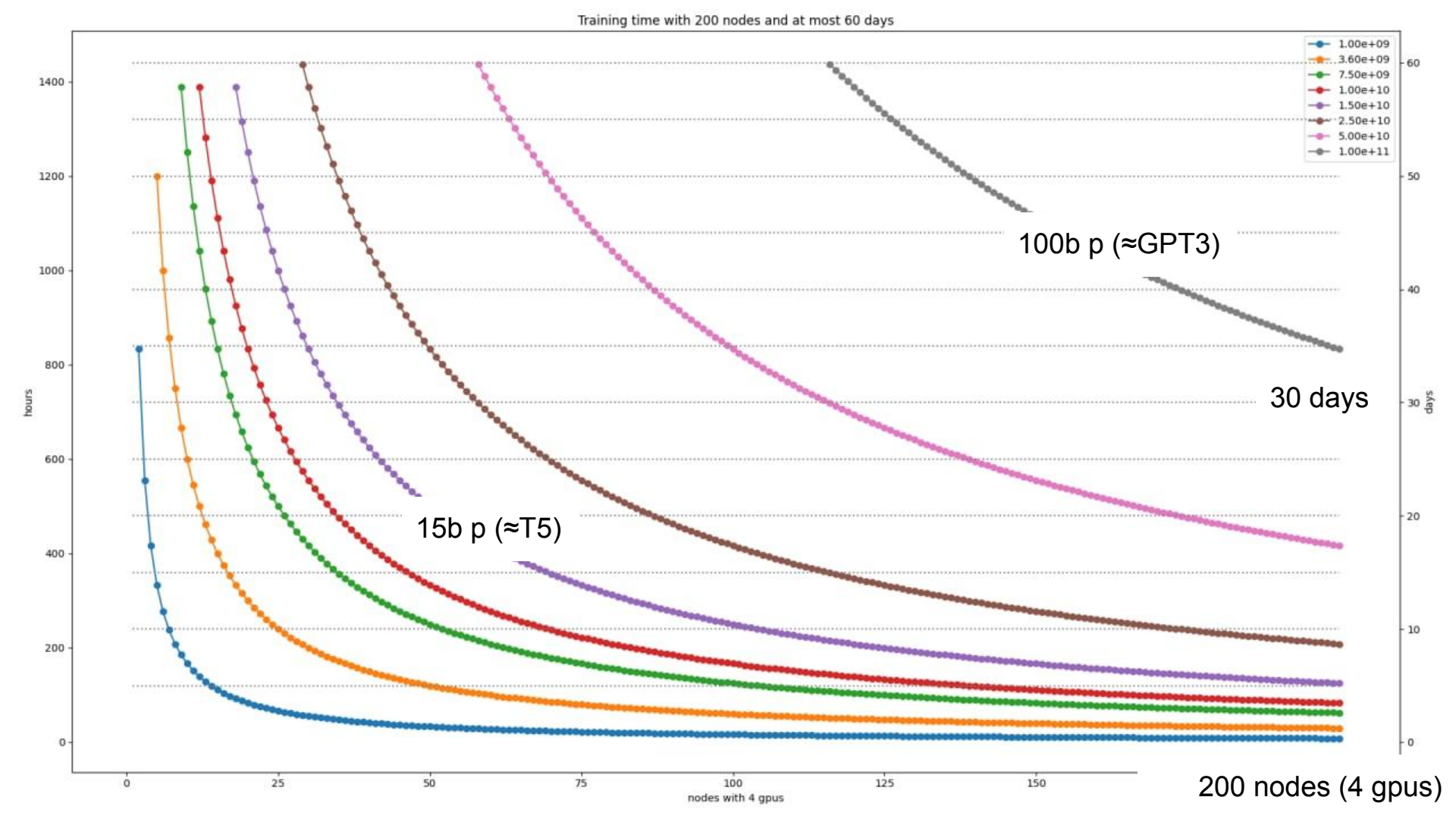
EuroHPC JU system

KBLab is the first public administration in Europe that successfully applied, received access and completed AI-models within the EuroHPC JU systems, thus (for a brief moment in time) putting Sweden on the front line among government agencies in Europe.


20000
15000
10000
5000
0
no of m parameters
110
317
340
1000
17000
BERT BASE
Wave2Vec 2 large
BERT LARGE
GPT2,5
T5

however…

…and GPT3 is certainly not the end of it


Training time with 200 nodes and at most 60 days
1.00e+09
3.60e+09
7.50e+09
1.00e+10
1.50e+10
2.50e+10
5.00e+10
1.00e+11
100b p (≈GPT3)
30 days
15b p (≈T5)
200 nodes (4 gpus)
hours
days
nodes with 4 gpus
0
200
400
600
800
1000
1200
1400
0
10
20
30
40
50
60
0
25
50
75
100
125
150

Were to now?

KBLab is currently using HPC Vega (Slovenia) and are inte process of porting code/modeling frameworks to LUMI (Finland). We are also applying for using MeluXina (Luxenburg).

Modeling outlook:

- Really large text model (>10b param)
- Multimodality
- Cross-lingual (revisited)

EVERY GPU-hour we can lay our hands on will be used to the benefit of society, i.e. will be used to transfer the potential of the library's collection to the Commons.

Thanks!

Love Börjeson, PhD
Director of KBLab, National Library of Sweden
love.borjeson@kb.se

KBLab

KUNGL. BIBLIOTEKET