Chapter BD+LT:I

I. Introduction

- □ Language Technologies
- Big Data Processing Architectures
- □ ML/AI/Data Tools Landscape

Goals of Language Technology

1. Aid humans in writing.

Correcting mistakes, formulating and paraphrasing text, transcription.

2. Identify texts related to spoken or written requests. Text information retrieval, semantic text similarity, question answering.

- 3. Make sense of texts without reading the originals. Categorization, information extraction, summarization, translation.
- 4. Instruct, and be advised by a computer.

Audio interfaces (e.g., dialog systems, robotics), learning and assessment.

- Ļ
- 5. Converse with computers as if they were human.

Turing test, conversational AI and chatbots, computational humor. What is the nature of language and its relation to (artificial) intelligence?

A Brief History

1940s Machine Translation (Shannon & Weaver, 1949)

1950s Turing Test (Turing, 1950)

Generative Grammars (Chomsky, 1957 & 1965)

1960s Chatbots: ELIZA (Weizenbaum, 1966), SHRDLU (Winograd, 1968)

- 1970s-80s Ontologies & Symbolic AI
 - 1990s Statistical NLP
 - 2000s Neural Language Models (Bengio et al., 2003) IBM Watson (2006–2011)

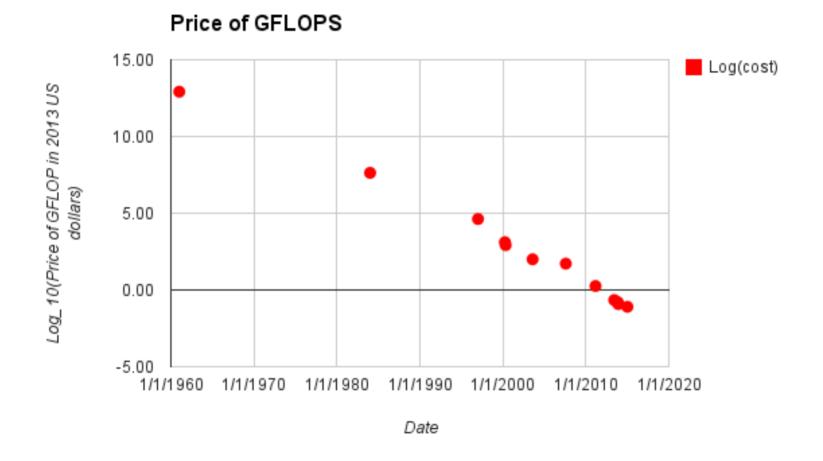
A Brief History (continued)

2010s Word Embeddings (Mikolov et al., 2013; Pennington et al., 2014)
 Deep Recurrent & Convolutional Nets in NLP

 (e.g. LSTM; Hochreiter & Schmidhuber, 1997)
 Sequence-to-sequence models (Sutskever et al., 2014)
 Attention (Bahdanau et al., 2015)

- 2017 Self-attention; Transformer (Vaswani et al., 2017)
- 2018 Pre-training: ELMo (Peters et al., 2018), BERT (Devlin et al., 2018), GPT (Radford et al., 2018)
- 2020 Large language models, zero-shot transfer: GPT-3 (Brown et al., 2020), T-Zero (Sanh et al., 2021), GPT-NeoX-20B (Black et al., 2022)

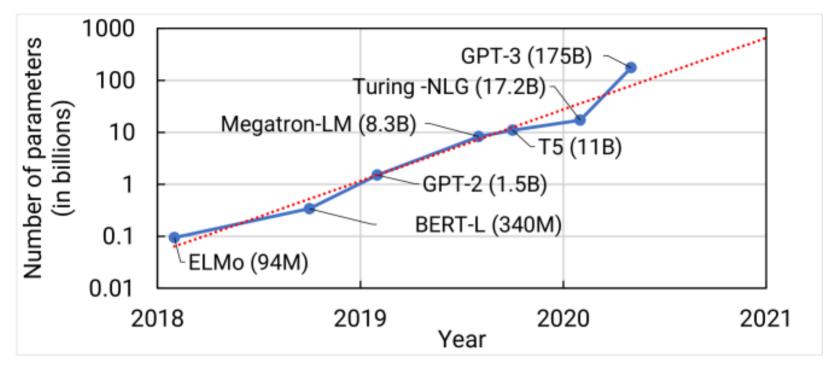
Cost of Compute



Source: [aiimpacts.org], pricing data from Wikipedia

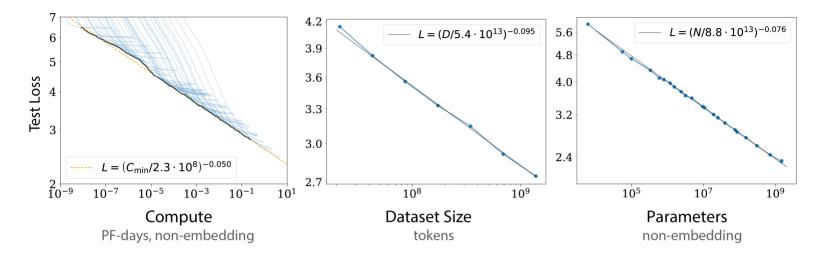
BD+LT:I-5 Introduction

Language Model Size



(Narayanan, 2021)

Language Technologies Scaling Laws



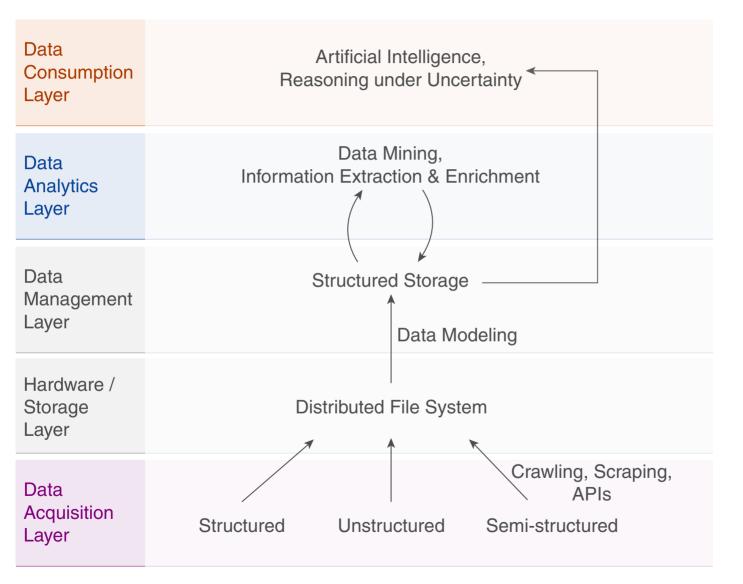
- □ More parameters as compute budget increases (Kaplan et al., 2020)
- □ Amount of training data needs to grow accordingly (Hoffmann et al., 2022)

Compute & Data Requirements

 GPT-NeoX-20B: Open-source autoregressive language model, trained for 2.5 months on 96 Nvidia A-100 GPUs (Black et al., 2022)

□ The Pile: 825GiB open-source curated text dataset (Gao et al., 2020)

Big Data Architecture Stack



Big Data Tools ML, AI, and Data ("MAD") Landscape

INFOACTOUCTUDE

INFRASTRUCTURE	ANALYTICS	MACHINE LEARNING & ARTIFICIAL INTELLIGENCE APPLICATIONS – ENTERPRISE			
STORAGE	BI PLATFORMS VISUALIZATION	DATA SCIENCE ML PLATFORMS	SALES MARKETING - MARKETING - B2C CUSTOMER EXPERIENCE / SERVICE HUMAN CAPITAL		
touders and the second	loöker 🚐 marana 🔽 🚌 🔅 + oble ov 🖬 PowerBl	NOTEBOOKS PLATFORMS Chinadar colab			
Standard Contract Con	T Thought Spail.		Card growense Arrise Signment mitpionel @Amplitude @HEMALLA Zendeck @Amplitude Chicker growense gitaldidek		
Pasescolar () wassel metarine Pwotel () formio () interview () () () () () () () () () () () () ()		O terretter 🚯 Status 💽 mentanter	A size peoples Street streetly Separate Control Streetly Separate		
A store parameter	ATSCALE QIKQ 2201 MODEN DODY		Careful tactual week Careful Ampenty & Ottommach Astarp Quids athrit mechanicares Careful BWode&Wendy		
Arrelation powerry	G Moratise & birst France O Observable	Meetparnel Court @ COCALC ALL ALL ALL ALL ALL ALL ALL ALL ALL	colective)) has more a house 201 @ Blaecore invoca ^O brage @ cause taila@ knows @yterCC DRESTA @ Onewy childle		
Contro VAST Division Reproper Booto Free Junits	12 million 2 mil	Oneverse Sellet ■ seesable ▲KNIPE ▲Naistratio E dotDoto datmo S one. Soboje obviewsky al	engens wholese creations (1995400) Convert Street Comments with Spirit Typelowed meengage		
		anno entro entro son colge comoya	LEGAL REGTECH & FINANCE AUTOMATION & RPA SECURITY		
ROBMS NoSQL DATABASES REALTIME GRAPH DBs DATABASES REALTIME GRAPH DBs	DATA ANALYST PLATFORMS AUGMENTED ANALYTICS	DATA GENERATION MODEL BUILDING FEATURE STORE DEPLOY MODEL & LABELLING MONITORING			
	Microsoft @pentoho alteryx TroughtSpot.	anaron rectanicaturi di HIVE II Weights & Blases TECTON @ recreases DUCTION & OBSERVA BILITY	TLADSSHT Concernence Sector Automatic Automatic Sector Automatic		
ZT AS STATE OFACLE ImageDB ACC Dimensioned Dimensioned IBM	Mode Datameer - Bitter	scole Upwork soppen Tig: Cons earon 🙆 Google Cloud	Sherviny Strategy Control Cont		
Filter Cluster Contents Transmisse [ROCKSET] ORACLE "Orientation		Gonty spennon islation ⊕ isiter ⊕ intervent SRASOC ● martin Victo ● martin			
Back Couldone grating Billion volton Sinte II- CickHouse Altimeters	ENDOR I Fower til ATTIV/O narrative @ science	ALIEVELE & Store Ain Store Contaction Scotter	TREAD > Miles a control + and a control + an		
And RECORDERED OF The Cash	incorta. guivus Acces efacet mechane>	USecretial Sharewark Register Bactory Classics Annual Annual State	💠 EN EO S U M 🔯 RECLEM al 🕐 Payment, 32 minute ALEVICE 👌 for Asser 19 Security National Architector 🗖 CODE42 👙 High		
ArangoDB - CRATE.10 ProgCA @ represent	switchboard obvieusly.ai 🏞 Airtoble cliff.ol	A commence Weakly Broom Collinson Conversion WHYLASS			
MPP DBs T FIL/ELT/ BEVERSE FIL DATA INTEGRATION DATA COVERNANCE		COMPUTER VISION SPEECH NLP SYNTHETIC MEDIA	APPLICATIONS - INDUSTRY		
TERADATA	AND DISCOVERY STORE Splunk>		INTELLIGENCE STITCH FIX A Instancert		
X dbt talend alteryx	M metaphor		Oxondr MediaMath Source Optimize Optimize SHEIN FAIRE SHEIN FAIRE SHEIN		
VERTICA Z meter S Rich & MATULION highbouch Maddet TEALIUM imploge IBM Profile	atian Trace Contrages Sumologic	🗣 hay 🖏 benade 🔯 transfer Validation 🖸 descript	Criteol. LAStern. Markety Opendore Buddening Phanese Thousand a second standard		
	Supergrain solarwinds kabana	Contanta de Presidentes de PRIMER	STATUTE STATUTE Child		
Contion at the come of Papers of restorement of Attuning ZALENI import.ie OKERA Manage	Stervina 🔅 🛞 logz.io 👔 logdna	Vanivers trax 🖉	Cectara da HNANCE - INVESTING		
Regnitio Strenden were Stere @Any.	(+) SELECT STAR • transform • Creakeds = rLogRhythm	Second grante and supertone an	CIDert gungum Selfin Sancauces		
Exasol Metters Stores a News	Barreda @Castor ZQL OBSERVE	Converse ALL CONTRACT Projection Contraction Contraction			
PRIVACY & DATA OBSERVABILITY MGMT / MONITORING SERVER CLUSTER SVCS	QUERY SEARCH	HORIZONTALAI GPU DBS & CLOUD AI HARDWARE	HEALTHCARE UFE SCIENCES TRANSPORTATION AGRICULTURE _ INDUSTRIAL OTHER		
SECURITY Crymer record storager Crymer record storag	ENGINE elasticsearch Endes Exacts	See Watson & OpenAI Cocepting KinEtica Google TPU arm @			
	(7 dremio	Paperspace Canada GRAPHCORE			
First Copernary 20 Instant Comments Comments Solarwinds # seventeers octific A	Starburst	Voyagerus Doublessee Vicorious.	amentania Atylescent Okazar Casser and Athlescent Athle		
Same (a) harmond DATA QUALITY >> (min) Character (Characteria	🚴 ahana 😢 Lucidworks 👩 swiftype 🛛 ATTIV/O	A Numenta 🖸 Million norologics Petruum			
skyllow 🖳 talend SODA Olione a 着 wetter		-CRIMENT OSADO I) Affectivo	Same Same Same Same Boxes Instro		
Control Digoroutry season Processory Pr		PG Storn	ezna General La A o E A States Conservi		
Statch Annua prober Viene Lines Operating Contracts	VARADA MAANA & QUICKWIT CHAOSSEARCH	Prostern Pros II NAME PROVIDE CONTRACT STORE	Change Strate (Learning Director Crisp. Change Over Crisp. Unexterned University Crisp. Constructions		

MACHINE LEARNING, ARTIFICIAL INTELLIGENCE, AND DATA (MAD) LANDSCAPE 2021

MACHINE LEADNING & ADTIFICIAL INTELLICENCE

						01	EN SOURCE						
FRAMEWORKS	FORMAT _ QUERY / DATA FLOW -	DATA	DATABASES	ORCHESTRATION INFF	DATA OPS	STREAMING &	MLOPS & INFRA	AL/ MACHINE LEARNING / DEEP LEARNING		SEARCH	LOGGING & MONITORING -	VISUALIZATION C	COLLABORATION 1 SECURITY
Anter South Anter and	spoil store 🕺 🐧	L Uber	🜉 🔊 kaugett 📦 vils Franklan Arveit	Revenues Presenter STR.		MESSAGING LANGUAGES I	nthan announ	All Inducting Learning Direct Learning Transformer & Operation State Strengthere & Operation State Strengthere & Operation State Arrestore & The Strengthere & Space (State Arrestore & The Strengthere & Space (State)	👷 📓 Kern BERT 🕅 Kaner Caffe	eetcert Sol?	rindicused Kitzana Antonio	3 CO Separat malalitätik	Grates Apache Ranger IK
deflink YARN 782% O katemete	See And A Company of C	Daraback Suprit Amundaen	🗇 influed) presto 🕅 🏐 dividi 💇 🔬 Generati	Source V Flyte	atanan Doer Masses	Anna dettek	· · · · · · · · · · · ·	Connection 👯 @ OpenAI © Pythensuppose theamo	annen annen onsten Enterte villes	Tucana. * Sphirx	Logitah Orientus Young	Mendone Parts Investment	🚓 🥖 🌆 saam
Service CORP	······································	Ay magda.	28608 🧟 🌺 💁 🕬 🛤 rick	Concercor 🍐 terra 🔞 p	Risos 🌰 www.s	see Sille	DVC SECTOR AND	2000 neon" 🖄 💴 🖉 🕬 🦛 🖼			Mart Obstan Witten	Burning Instants (20)	
And Had Hart A	Verte res	w dan	Harris Marrie Corner Opinot me a Fabiat	6 ¹ (²)	ngo 🕹 Paquat Hanara	Aparla Embrado () () (172)	The State Name	Arrester The You'r Control SpaCy grader	Alace 1.2 😰 remot	faarte	A	DUKET C	ANALONDA .

DATA SOURCES & APIs						DATA RESOURCES			
DATA MARKETPLACES	INANCIAL & ECONOMIC DATA	AIR/SPACE/SEA	PEOPLE / ENTITIES	LOCATION INTELLIGENCE	OTHER	DATA SERVICES	INCUBATORS & SCHOOLS RESEARCH		
& DISCOVERY BI	Nomberg () HOMON HOMON D DOW JONES Quant. D British	O DERAMON AROBOTICS & Spire 🕐 SSA	Z zoominto acxiem tijoperion.	FOURSQUARE Omopbox Omopbox	MAGENET LOOPING	C DUANTUMBLACK 00 Boox Alien Hamilton	Susauceser O SENERAL ASSENSEY O DataCamp & OpenAI Google Research focebook research		
A reason of a second	Sarra Cheven and Street and Stree	WINDWARD Presenter O TIS.	arender 🕲 kryside View 👙 a andward	Piecelle 🍘 esri 🎫 🚺 🔒 Radar 🖄 Mopilary			A Dataine galvanize 🖉 who 'nstart,		
	kignite Etwann skirrigst predata MX tink ¹ ⊟ marner netwar \$******* \$*KAIKO converters	€ ExcLabs 0,1701 ≤ Symposium	Quantcast ⊑tstis ⊗secceani Demyst melissα ≷zignat…		VERITORE 📕 comscore	DataKind moreces			

Version 3.0 - November 2021

© Matt Turck (@mattturck), John Wu (@john_d_wu) & FirstMark (@firstmarkcap)

ANALYTICS

mattturck.com/data2021

FIRSTMARK 🥏

[mattturck.com/data2021]

Big Data Tools ML, AI, and Data ("MAD") Landscape

INFRASTRUCTURE	ANALYTICS	MACHINE LEARNING & ARTIFICIAL INTELLIGENCE	APPLICATIONS - ENTERPRISE			
	INFORMATION	MACHINE LOADING A RIFFLUAL MILLIA MILLIAR CALLEDONCE				
Other Transform	Monool Operator alleryx ThoughtSpot. Monoo Datameer Children Monoo Datameer Children Monoo Datameer Monool Datameer M	A LABLING A LABLING	Description Description <thdescription< th=""> <thdescription< th=""></thdescription<></thdescription<>			
	ADMC CANAGE MITRIC COAMANTICS ASD DRCOVING Stort Stort Bit metaphor Coamantics Splunk> Charles Coamantics Splunk> Secondaria Trace Stortgram Bit metaphor Coamantics Splunk> Secondaria Stortgram Stortgram Distance taxa Vanstorn Stortgram Distance taxa Vanstorn Stortgram Distance taxa Vanstorn Stortgram Stortgram Ital Stortgram	COMULTINGNA STREET OF THE STRE	ADVERSING DOLLARS CONTACTORS DOL			
PRVX 4 S COURT DAA 0500 MARLIN MAAL MARKEN MAAL MARKEN <td>OUEP SAEC* WINNE Reserved: Offernini Reserved: Statutoriti Reserved: Statutoriti Reserved: Autoriti Linithworks Statutoriti Statutoriti Autoriti Linithworks Statutoriti Statutoriti Autoriti Linithworks VRRADA MANDIA</td> <td>NORDOVILLAI GTUDESS ACQUO ANALWESS Image: Source of the s</td> <td></td>	OUEP SAEC* WINNE Reserved: Offernini Reserved: Statutoriti Reserved: Statutoriti Reserved: Autoriti Linithworks Statutoriti Statutoriti Autoriti Linithworks Statutoriti Statutoriti Autoriti Linithworks VRRADA MANDIA	NORDOVILLAI GTUDESS ACQUO ANALWESS Image: Source of the s				

MACHINE LEARNING, ARTIFICIAL INTELLIGENCE, AND DATA (MAD) LANDSCAPE 2021

	OPEN SOURCE	
FRAMEWORKS FORMAT QUERY / DATA FLOW DATA DATABASES ORCHESTRATION	INFRA DATA OPS	EARCH LOGGING & MONITORING VISUALIZATION COLLABORATION SECURITY
And the set of the set		Solo States and States
Caretoria Careto		ancana "Sphirx Lington Onester Your Statement See 5 - 10 the another
🔛 😵 HERES 👉 2000 F CADE 🖉 🛄 VILL 👘 🖓 MORES. 280 CB 🥸 🌺 🙆 Marken 🚍 crick 🔇 MOREN 🖉 🕬		
A Real REAL Contract Contract Opinion Real C	👔 orgo 🖕 neutrine Augustania () 5170.2 😨 📠 💮 💭 Interna Station Anneche 📼 Yarr () 1070. Booling Residue Abort D 😰 mare	ដ 🚵 🔬 🕺 🕺 👘 👘

DATA SOURCES & APIs						DATA RESOURCES			
DATA MARKETPLACES	FINANCIAL & ECONOMIC DATA	AIR/SPACE/SEA	PEOPLE / ENTITIES	LOCATION INTELLIGENCE	OTHER	DATA SERVICES	INCUBATORS & SCHOOLS	RESEARCH	
& DISCOVERY	Biomberg () howsenantin () (ow senis Quant, Quant,	🔿 выязыны Анарабтись 🛆 spire 🕐 солч	Z zoominto acxiem Seperan	FOURSQUARE Omopbox Omopbox			STATALESENT O GENERAL ASSENDLY S DataCamp	OpenAI Google Research focebook research	
S DADEM C descert	State Change Street State Z State Arts	WINDWARD Characters	erencer GinsideView 🧔 Bardward	macatel weest had a kadar schophary		kapple ElectrifAi fractales (Ress.	A Datalite galvanize 🖉 Herri "INSIGHT,	MIRI V MIRI 🕅 🔤	
	xignite exercise carriest predata MX tink? Tournees	€Exalabe 0,1708 % Syspectre	Quantost ⊑tasis ⊜secones Demyst melissa ≷zianai ~		VERTONE 🗮 comiscore	DataKind moneccus		AIZ ANTHROPIC Salk III	

Version 3.0 - November 2021

© Matt Turck (@mattturck), John Wu (@john_d_wu) & FirstMark (@firstmarkcap)

mattturck.com/data2021

FIRSTMARK 📂

[mattturck.com/data2021]

ML, AI, and Data ("MAD") Landscape Open Source Tools



[mattturck.com/data2021]