

Introduction to the summer school on Clouds and Big Data 2015

Seif Haridi
KTH/SICS



Big Data a collection of massive amount of information whether structured or unstructured

Big Compute large scale processing required to extract value from Big Data

Machine Learning algorithms to discover patterns and hidden concepts implicit in data

Predictive analytics using machine learning to predict future outcome or unknown data points

Objectives (Scientific)

- Understand the state-of-art of technology in the area of Big Data analytics
- Systems for massive data
- Algorithms for massive data

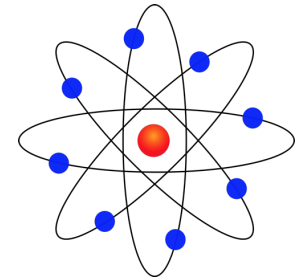
Objectives (Industrial)

- Get to know how the emerging technology creates new business opportunities
- Use cases
- Work out your own business ideas

What makes Big Data and Data Science relevant today ?

- Rapid growth of massive datasets
- Online activity, Science, Sensors, Networks

Google™



facebook



What makes Big Data and Data Science relevant today ?

- Distributed Clusters are Pervasive (clouds)
- Maturity of Big data storages and processing technologies
- Mature Methods for Common Problems e.g., classification, regression, collaborative filtering, clustering



What makes Big Data and Data Science relevant today ?

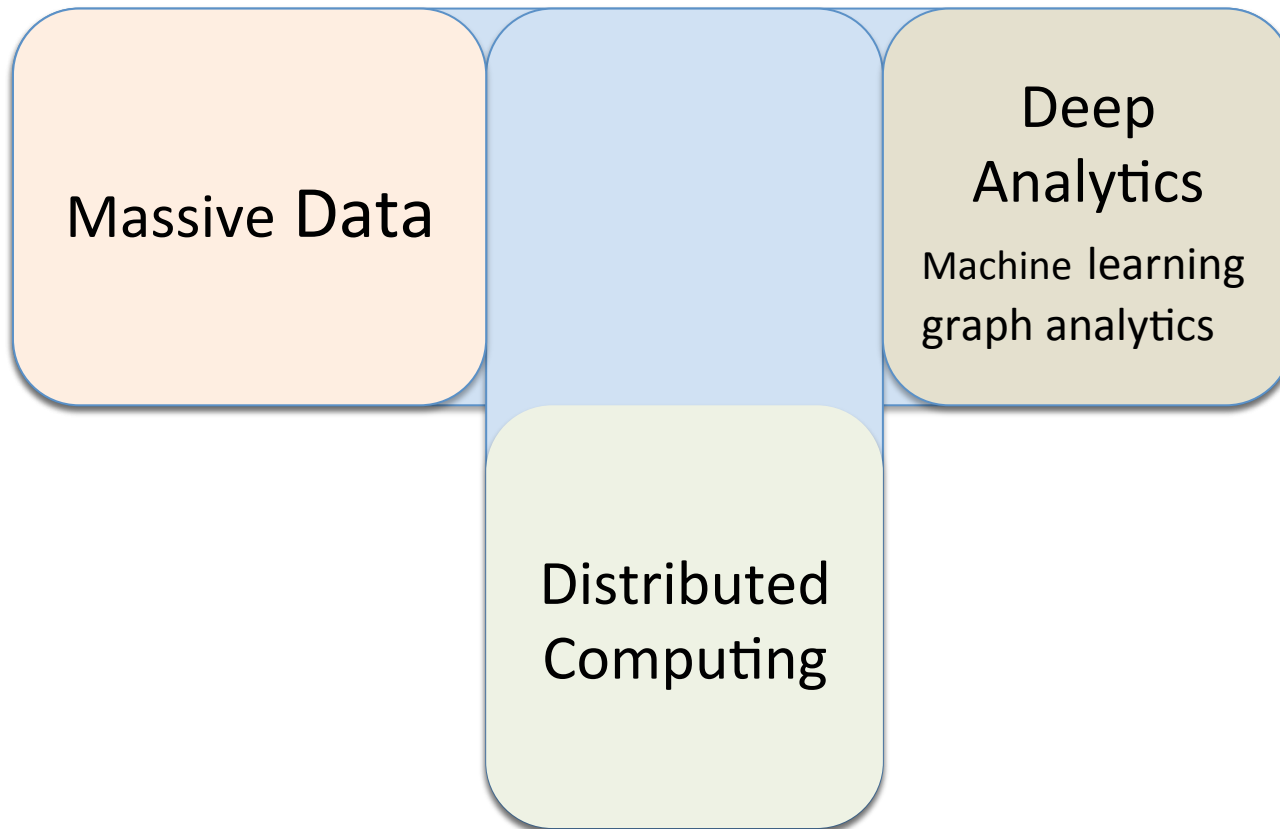
- Distributed Clusters are Pervasive (clouds)
- Maturity of Big data storages and processing technologies
- Mature Methods for deep analytics
- classification, regression, collaborative filtering, clustering, graph analytics



Google Cloud Platform



So What is BIG data science?



Summer School on "Cloud and Big Data"

Stockholm, August 3-14

	August 3, 2015	August 4, 2015	August 5, 2015	August 6, 2015	August 7, 2015	August 8, 2015	August 9, 2015
8:00		I&E Breakfast Project selection and review, BMC and lean methods	I&E Breakfast Importance of roles in teams and hiring	I&E Breakfast Overview of business models in the cloud space	I&E Breakfast The perfect pitch	I&E Breakfast Summary of first week	Social Events (Kayaking, Hiking)
9:00	Welcome Speech Seif Haridi (SICS/KTH) Introduction to I&E Björn Hovstadius (SICS/EIT Digital BDA)	Hadoop Ecosystem Sameh El-Ansary (Nile University, Egypt)	FLINK Volker Markl (TU Berlin, Germany)	SPARK Paco Nathan (Databricks, USA)	EIT ICT Labs Tua Huomo	Streaming Paris Carbone and Gyula Fora (SICS, Sweden)	
10:00	Coffee Break						
10:30	Introduction to I&E Björn Hovstadius (SICS/EIT Digital BDA)	Hadoop Ecosystem Sameh El-Ansary (Nile University, Egypt)	FLINK Ufuk Celebi and Max Michels (Data Artisans, Germany)	SPARK Paco Nathan (Databricks, USA)	Big Data and Deep learning: Scalability and Fault-tolerance of Parallel and Distributed Infrastructures Divyakant Agrawal (UCSB, USA)	Streaming Paris Carbone and Gyula Fora (SICS, Sweden)	
11:30	Company presentations Rickard Coster (Ericsson)	Hadoop Ecosystem Sameh El-Ansary (Nile University, Egypt)	FLINK Ufuk Celebi and Max Michels (Data Artisans, Germany)	SPARK Paco Nathan (Databricks, USA)	Mining Democracy Matthias Grossglauser (EPFL, Switzerland)	Stateful distributed dataflow graphs Peter Pietzuch (Imperial College London, UK)	
12:30	Lunch						
14:00	Company presentations Jean-Michel Gaullier (Novartis)	Locality-sensitive hashing (1) Jeffrey D. Ullman (Stanford, USA)	FLINK Ufuk Celebi and Max Michels (Data Artisans, Germany)	Clustering Jeffrey D. Ullman (Stanford, USA)	Using Social Media for Health Studies Ingmar Weber (QCRI, Qatar)	Machine Learning Anders Holst and John Ardelius (SICS, Sweden)	
15:00	Additional case and company presentations	Locality-sensitive hashing (2) Jeffrey D. Ullman (Stanford, USA)	Counting distinct elements in a Stream Jeffrey D. Ullman (Stanford, USA)	Graph algorithms Jeffrey D. Ullman (Stanford, USA)	Google Cloud Dataflow Cosmin Arad (Google, USA)	Machine Learning Anders Holst and John Ardelius (SICS, Sweden)	
16:00	Coffee Break						
16:30	Team formation and Case selection	Sampling and filtering data streams Jeffrey D. Ullman (Stanford, USA)	PageRank Jeffrey D. Ullman (Stanford, USA)	Designing good MapReduce algorithms Jeffrey D. Ullman (Stanford, USA)	Google Cloud Dataflow Cosmin Arad (Google, USA)	SPARK Machine Learning Paco Nathan (Databricks, USA)	
17:30	Homeworks						

	August 10, 2015	August 11, 2015	August 12, 2015	August 13, 2015	August 14, 2015
9:00 - 12:30	Team building revisited; market research; competitor analysis, IPR	Value chains, Marketing, go-to-market strategy, branding	Financing your idea, Venture financing, risk analysis	Presenting, stage training, pitch rehearsal	Final preparation for presentations
12:30	Lunch				
14:00 - 16:30	Free time in Stockholm	Project work in teams			Final presentations to the Jury
17:30		Pitch training and feedback			Project reports due on Aug 31

Resources

- <http://ictlabs-summer-school.sics.se/>
- <https://canvas.instructure.com/courses/940184>