Econometrics at Scale:
Spark Up Big Data in Economics

Benjamin Bluhm
(benjaminbluhm@gmail.com)

Jannic Alexander Cutura
(cutura@finance.uni-frankfurt.de)

Motivation

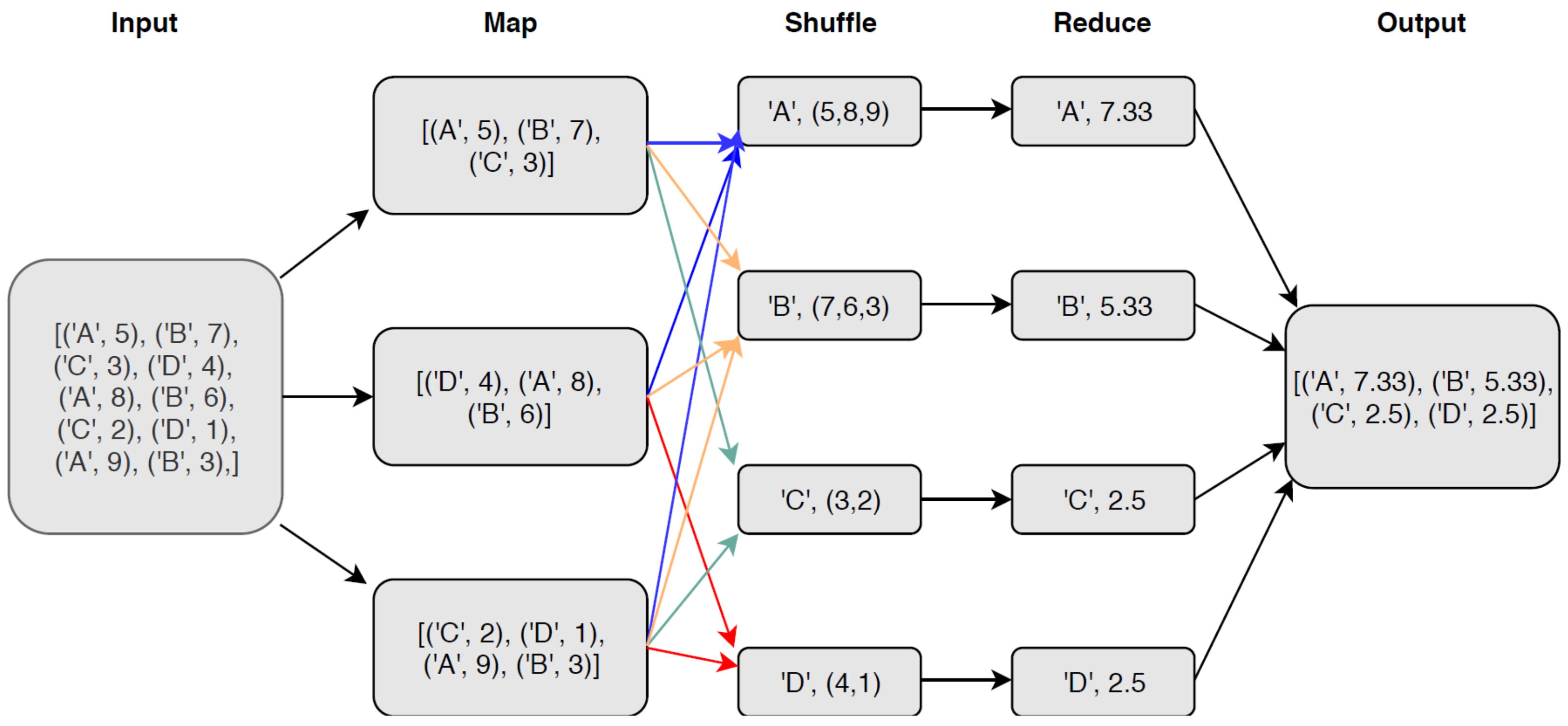
- The amount of **data available** for research is ever **increasing**
- **New econometrics techniques** borrowered from the **machine learning** literature gained considerable attention over the past years
- "**Diabolic loop**" between research questions and perceivable data handling
- Ability to handle and analyse datasets that are too large to fit in memory crucial to leverage 21 century opportunities
- Little guidance for economists on how to handle big data
- Contribution:
  - **Lower the threshold** for usage of cloud computing for economic research
  - Provide an **accessible overview** of the concepts as well as ready-to-use minimal examples in **"econspeak"**

What is Big Data?

- **Volume** refers to the vast amounts of data generated every second
- **Velocity** refers to the speed at which new data is generated and the speed at which data moves around while it is being generated, without ever putting it into databases.
- **Variety** refers to the different types of data we can now conversations, photos, sensor data, video or voice recordings and bring them together with more traditional, structured data.
- **Veracity** refers to the messiness or trustworthiness of the data.. The volumes often make up for the lack of quality or accuracy.
- **Value**: Then there is another V to take into account when looking at Big Data: Value
- **Working Definitions**: Data is Big Data if you run into **memory limits** in your retail grade computer

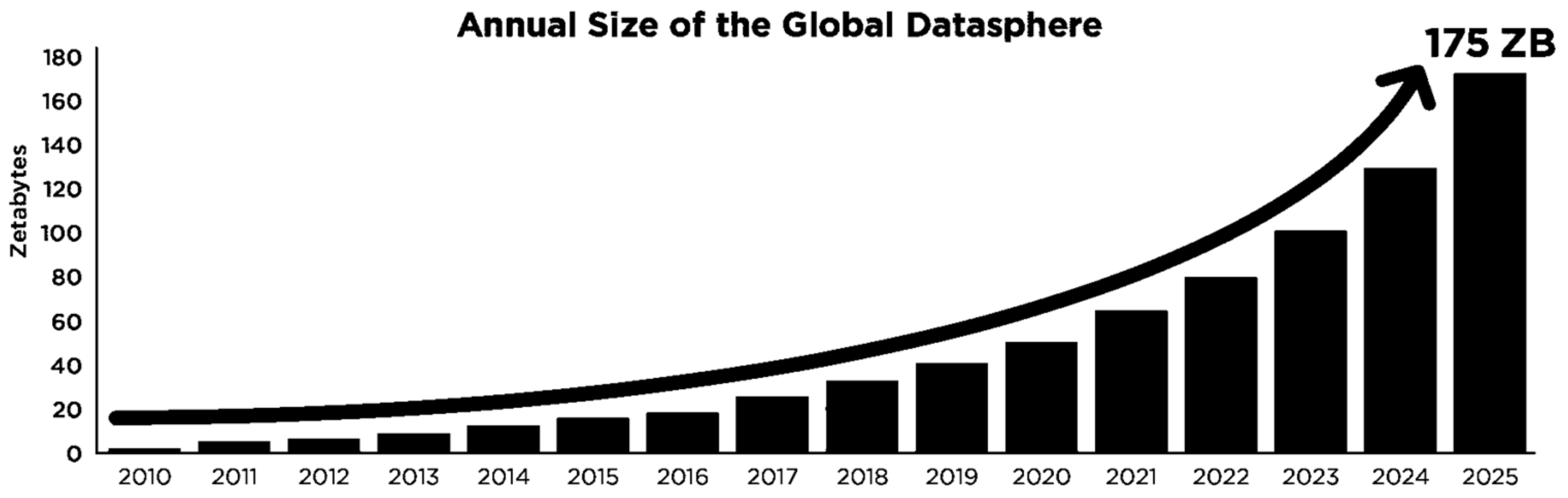
Spark Architecture

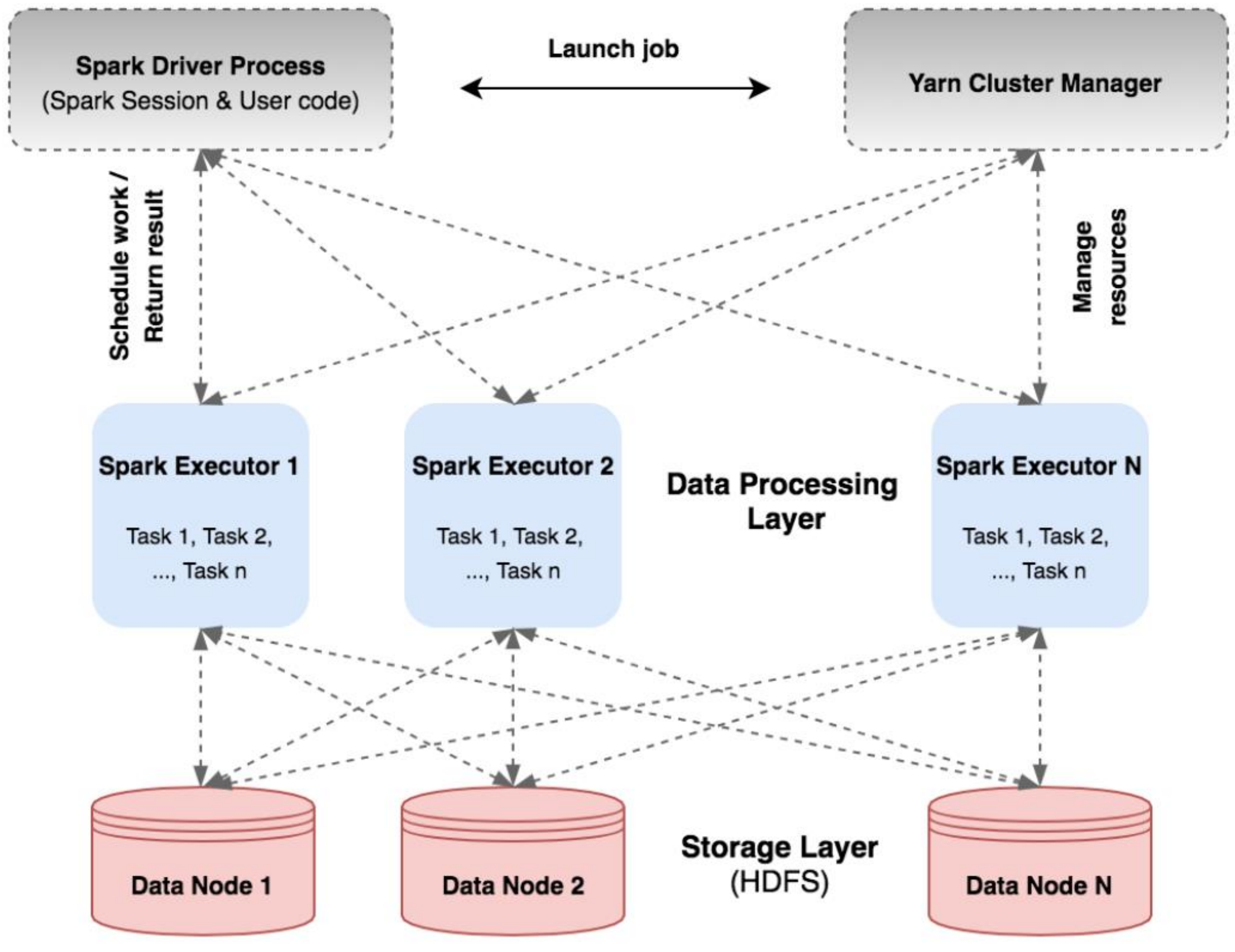
- Spark solves memory problems by distributing the computation across clusters
  - Your computer: **Take the data to the computation**
  - Spark: **Bring the computation to the data**
- Is based on efficient implementations of the map-reduce framework
- How do you compute the mean of large dataset?


Input
Map
Shuffle
Reduce
Output
[('A', 5), ('B', 7), ('C', 3), ('D', 4), ('A', 8), ('B', 6), ('C', 2), ('D', 1), ('A', 9), ('B', 3),]
[(A', 5), ('B', 7), ('C', 3)]
[('D', 4), ('A', 8), ('B', 6)]
[('C', 2), ('D', 1), ('A', 9), ('B', 3)]
'A', (5,8,9)
'B', (7,6,3)
'C', (3,2)
'D', (4,1)
'A', 7.33
'B', 5.33
'C', 2.5
'D', 2.5
[('A', 7.33), ('B', 5.33), ('C', 2.5), ('D', 2.5)]

Cloud Environments

aws
salesforce
ORACLE CLOUD
phoenixNAP GLOBAL IT SERVICES
SAP Cloud Platform
Azure
KAMATERA PERFORMANCE CLOUD
cloudera Ask Bigger Questions
Google Cloud





Econometrics on Spark

- Spark allows to use your **existing data handling and analysis pipelines** in a distributed fashion
- We demonstrate ease of use for many standard empirical setups
  - Understanding data
  - Microeconometrics
  - Paneleconometrics
  - Time Series Econometrics

	OLS			Probit			Logit		
	Spark (local)	base R (local)	Spark (AWS)	Spark (local)	base R (local)	Spark (AWS)	Spark (local)	base R (local)	Spark (AWS)
Loan granted	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Income (000$)	0.0001***	0.0001***	0.0000061944***	0.0003***	0.0003***	0.0002***	0.0006***	0.0006***	0.0002***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Male	0.0250***	0.0250***	0.0298***	0.0639***	0.0639***	0.070***	0.0982***	0.0982***	0.07006***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Race									
White	0.01383***	0.01383***	0.00829***	0.0354***	0.0354***	0.0264***	0.0613***	0.0613***	0.02645***
	(0.0113)	(0.0113)	(0.0000)	(0.0109)	(0.0109)	(0.0000)	(0.0061)	(0.0061)	(0.0000)
Black	-0.1486***	-0.1486***	-0.1450***	-0.3823***	-0.3823***	-0.3629***	-0.6060***	-0.6060***	-0.3629***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
# Observations	147,329	147,329	137,819,151	147,329	147,329	137,819,151	147,329	147,329	137,819,151
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Loan Purpose FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Runtime (min)	**0.71**	**0.007**	**9.90**	**0.146**	**1.556**	**33.69**	**0.026**	**0.995**	**19.833**

Conclusion

- Cloud Computing is a useful tool for research in economics and social sciences more broadly
- Cloud technologies enables usage of existing data handling pipelines to be executed on (i) arbitrary large datasets in (ii) a fraction of computing time
- Intuitive explanations of Spark using econpseak alongside fully reproducible minimal examples

HOUSE OF FINANCE
Goethe-Universität Frankfurt

GOETHE UNIVERSITÄT FRANKFURT AM MAIN