This is an idea for the workshop that I'd like to receive feedback for. There are many different components in a data science architecture. If massive amounts of data need to be handled, the final solution can be a complex one.

Data science processes include data exploration, data transformation, data modeling, deploying machine learning models, and data visualization. The idea to be presented is to have radical automation and radical coupling between these processes and and within this processes.

Infrastructure is key for these processes and can be automated. Infrastructure as code. The idea is to have infrastructure, software and storage deployed automatically depending on the needs of the project.

It is recognized three instances when human interaction is required. Entering the credentials to connect to cloud providers or onprem servers. Second, to classify metadata after automated data exploration. The metadata classification will help to create and study derived variables from the data. Metadata can tell what's the target variable, the ids in the data, relationships among data sources, etc. This information can be used for a second round of automated data exploration and also for feature engineering.

Maximizing automation would allow organizations without he knowledge to analyse data, or with the right infrastructure to use systems that can help them to understand the data they have. The necessary human intervention is because the credentials are needed, the expert knowledge domain is added through the metadata tagging after an initial data exploration.

A system is proposed where cohesion between infrastructure and software is radically used. Automated data exploration, automated feature engineering, and automated machine learning is deployed in infrastructure that is automatically deployed.

The processes can be coordinated in a distributed manner. The status is keeped locally and can be check at low cost, so any new machine joining the processes can verify the status by checking the metadata status and the log of actions carried by the system.

A centralized metadata database keeps the status of the completed actions and current status of the system.

This type of system would help many companies to achieve data governance, and data understanding of their data sources.

Radical automation is suggested on infrastructure and software for data science.