

Big Data Analytics

Gregor Kasieczka
(gregor.kasieczka@uni-hamburg.de)

IDT-UM Meeting
September 21, 2020

CLUSTER OF EXCELLENCE
QUANTUM UNIVERSE



Background

- *ErUM-Data Pilot Project* (IDT-UM) successfully started in 2017
- Running until 30.09.2021
- How to continue?
- Idea: Two follow-up projects
- *Federated infrastructures and cloud computing*
 - ~Areas A+B of IDT-UM
 - Kick-Off:
<https://indico.desy.de/indico/event/27012/>
 - Coordinated by Alexander Schmidt (next presentation)
- *Big Data Analytics*
 - ~Areas C+D of IDT-UM
 - Kick-Off:
<https://indico.desy.de/indico/event/27015/>

Innovative Digitale Technologien für die Erforschung von Universum und Materie

Gemeinsamer Antrag von Gruppen aus den Bereichen
Elementarteilchenphysik, Hadronen- und Kernphysik und Astroteilchenphysik

- Rheinisch-Westfälische Technische Hochschule Aachen, Prof. Dr. Martin Erdmann
- Rheinische Friedrich-Wilhelms-Universität Bonn, PD Dr. Philip Bechtle
- Friedrich-Alexander-Universität Erlangen-Nürnberg, Prof. Dr. Gisela Anton
- Goethe Universität Frankfurt am Main, Prof. Dr. Volker Lindenstruth
- Albert-Ludwigs-Universität Freiburg, Prof. Dr. Markus Schumacher
- Georg-August-Universität Göttingen, Prof. Dr. Arnulf Quadt
- Universität Hamburg, Jun.-Prof. Dr. Gregor Kasieczka
- Karlsruher Institut für Technologie, Prof. Dr. Günter Quast
- Johannes Gutenberg-Universität Mainz, Prof. Dr. Volker Büscher
- Ludwig-Maximilians-Universität München, Prof. Dr. Thomas Kuhr
- Bergische Universität Wuppertal, Prof. Dr. Christian Zeitnitz

Assoziierte Partner sind

- CERN, Dr. Markus Elsing
- DESY, Dr. Volker Gülzow
- GridKa, Dr. Andreas Heiss
- GSI Helmholtzzentrum für Schwerionenforschung, Darmstadt, Dr. Kilian Schwarz
- Forschungszentrum Jülich, Dr. Elisabetta Prencipe
- Westfälische Wilhelms-Universität Münster, PD Dr. Christian Klein-Bösing

For reference: Area C/D of IDT-UM (2017)

Themenbereich C: Deep Learning, Erkenntnisgewinn durch fundierte datengetriebene Methoden

Themenbereich D: Ereignisrekonstruktion: Kosten- und Energieeffiziente Nutzung von Computing-Ressourcen

C1) Sensornahe Verarbeitung von Daten <ul style="list-style-type: none">• Signalfilter, Rauschunterdrückung• Verarbeitung von zeitabhängigen Signalen	C2) Objektrekonstruktion <ul style="list-style-type: none">• Spur- und Clusterrekonstruktion, Jetbildung, Ereignisrekonstruktion• Fragestellungen für Anordnung, Reihenfolge, Zuordnungen von Daten• Optimierungen zur Extraktion kleiner Signale bei großem Untergrund
C3) Netzwerkbeschleunigte Simulationen <ul style="list-style-type: none">• Generative adversarial networks, Anpassung von Simulationen an Datenverteilungen• Evaluationsverfahren für die Qualität der Netzwerksimulationen	C4) Qualität von Netzwerkvorhersagen <ul style="list-style-type: none">• Reduzierung experimenteller systematischer Unsicherheiten• Spezielle Lernstrategien• Vorhersagenrelevante Information• Unsicherheiten von Vorhersagen

D1) Spurfindung <ul style="list-style-type: none">• alternative Algorithmen, z.B. zellulärer Automat• alternative Architekturen, z.B. GPUs	D2) Parameterbestimmung <ul style="list-style-type: none">• Verknüpfung GenFit2-ACTS
D3) Neutrinoexperimente <ul style="list-style-type: none">• dünnbesetzte Detektorinstrumentierung• variable Signalzeit als kritische Information	

Environment

- Best fit in anticipated Aktionsplan ErUM-Data
 - Not released yet
- -> Consider submission as application in ErUM-Pro
 - Motivate large and urgent need towards BMBF
 - Expect (and want) to move to ErUM-Data funding line
 - Expected deadline 1.11.2020
- Of course adjust if Aktionsplan ErUM-Data is released in the meantime
- Difference to other formats:
 - PUNCH/NFDI: Focus on infrastructure, projects that reach beyond our communities
 - ErUM-Pro CMS/ATLAS/LHCb/... applications: Experiment and analysis specific work and developments
 - Federated infrastructures and cloud computing: As the name says - computing infrastructure work across experiments
 - This: Development of new analysis methods across experiments/theory

Content

- **Area I:** *Encoding symmetries and understanding network decision*
- **Area II:** *Resource efficient generation and simulation (accelerating Monte Carlo simulation, generative ML models, reduction of the simulation gap,...)*
- **Area III:** *Improved reconstruction and pattern recognition (tracking, ACATS, realistic and time-dependent conditions,...)*
- **Area IV:** *Real-time decision making (data processing on hardware, machine learning on FPGAs, triggering, fast event reconstruction,...)*

Preliminary expressions of interest

	I (Symmetry)	II (Generation)	III (Reconstruction)	IV (Real-time)
<i>KIT - Astroteilchenphysik (Haungs)</i>	X	X	X	X
<i>LMU (Kuhr)</i>		X		
<i>Fast event-generator racing team (Heidelberg, Plehn)</i>		X		
<i>Fast event-generator racing team (Göttingen, Schumann)</i>		X		
<i>Fast event-generator racing team (KIT, Gieseke)</i>		X		
<i>Fast event-generator racing team (Dortmund, Kröniger)</i>		X		
<i>Aachen (Erdmann)</i>	X			
<i>Goettingen (Quadt)</i>	X	X	X	
<i>Hamburg (Haller, Kasiieczka, Schleper)</i>		X		X
<i>Mainz (Schmitt)</i>	X			X
<i>LHCb (Hansmann-Menzemer)</i>			X	
<i>ACTS (Bernlochner)</i>			X	
<i>KHuK (Schwarz)</i>	X		X	X
<i>FIAS (Steinheimer/Zhou)</i>	X	X		X
<i>Erlangen (Funk)</i>	X	X	X	
<i>Information Field Theory (Ensslin)</i>	X			
<i>Dresden (Straessner, Siegert)</i>		X		X
<i>Siegen (Cristinziani)</i>			X	

Organisation and Timeline

- Please sign up for mailing list
<https://lists.desy.de/sympa/info/big-data-analytics>
(big-data-analytics@desy.de)
for simple communication
 - **Also alert potentially interested groups not involved yet**
- Now:
 - Fix participants and define work-packages and concrete projects
- Until Mid-October: Finalise general part of application and requested FTE
- Be prepared for submission on 1.11., still can move back depending on ErUM-Data

Thank you!