Data Science Challenges in Computational Chemistry

M. Sc. Julia Jasper

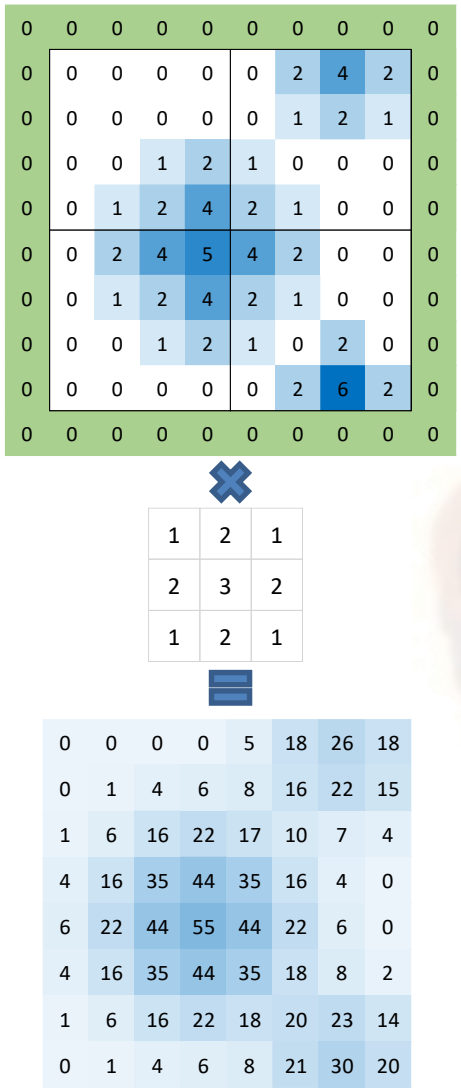
M. Sc. Nicolas Tielker

M. Sc. Yannic Alber

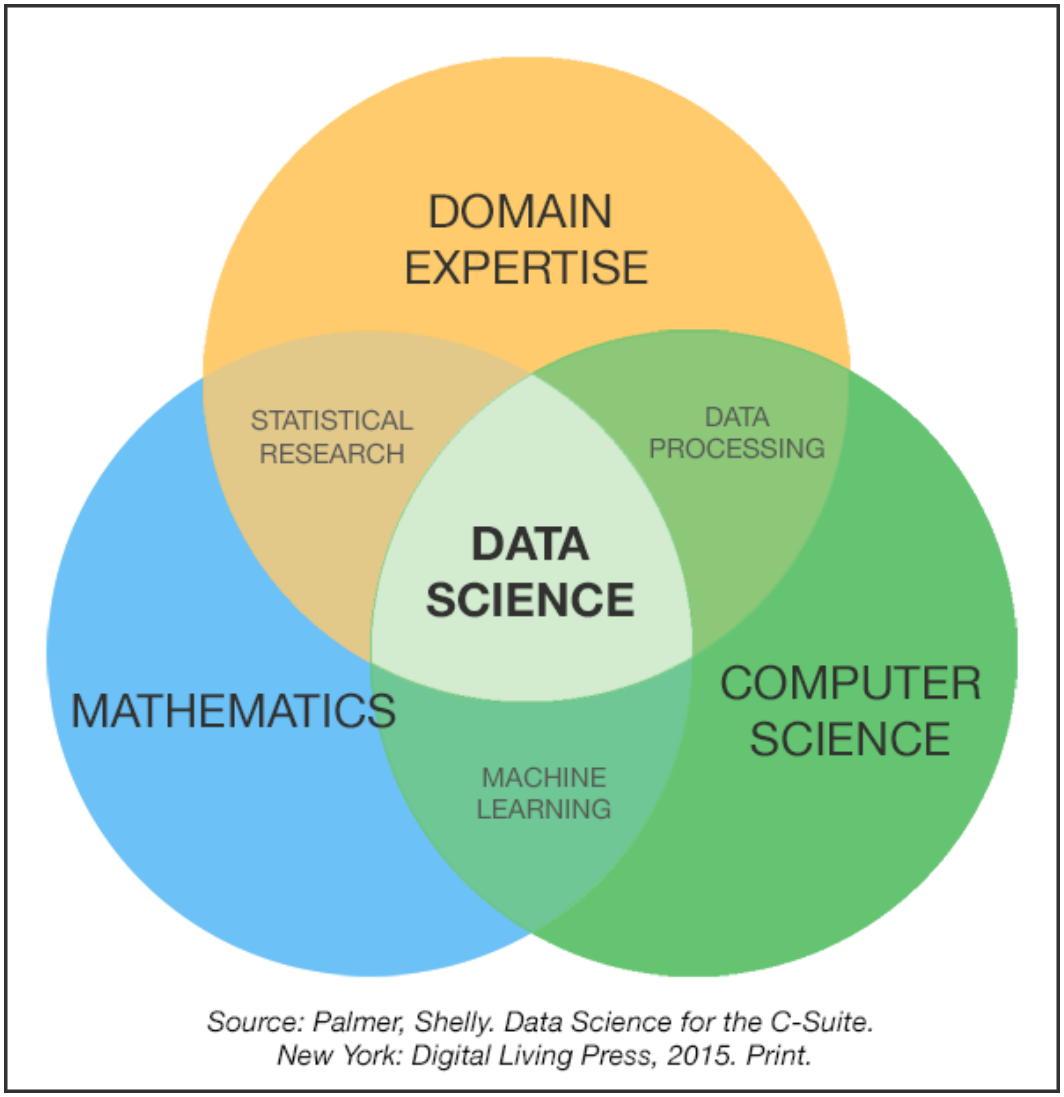
Prof. Dr. Stefan M. Kast

Fakultät für Chemie und Chemische Biologie

Physikalische Chemie III

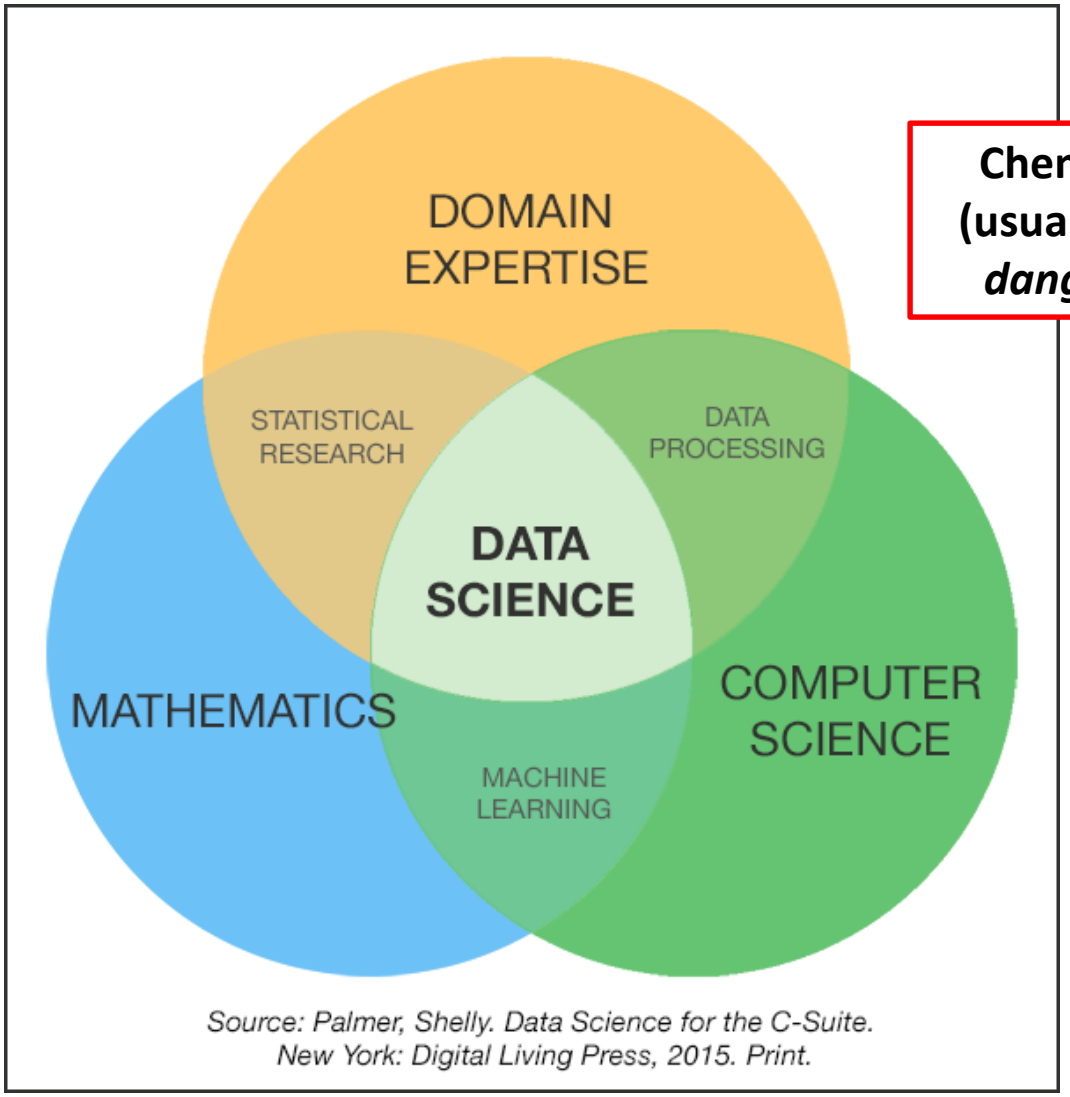


Data & Science in Chemistry



Source: Palmer, Shelly. Data Science for the C-Suite. New York: Digital Living Press, 2015. Print.

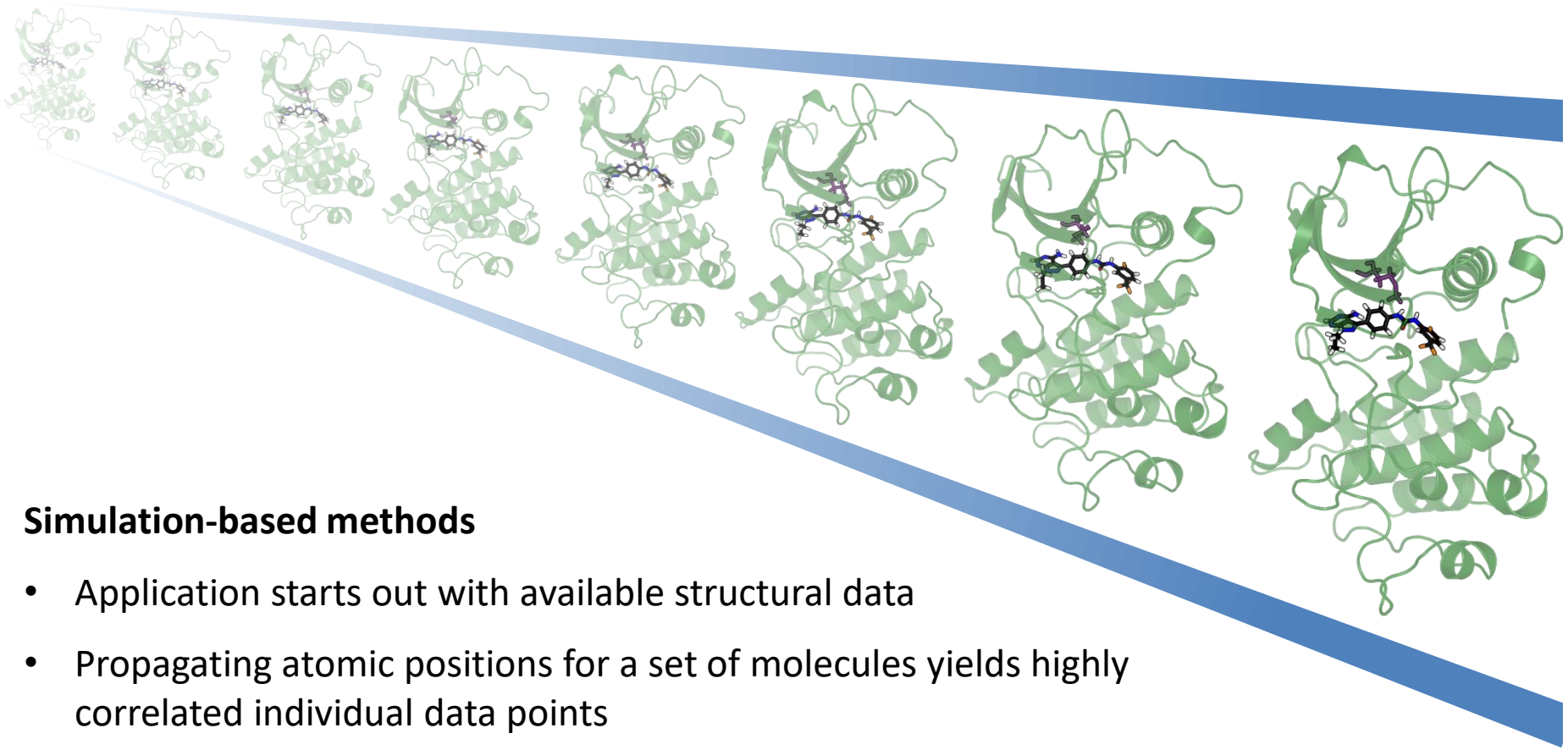
Data & Science in Chemistry



Source: Palmer, Shelly. Data Science for the C-Suite. New York: Digital Living Press, 2015. Print.

Chemists are (usually) in the *danger zone*

Molecular Simulations



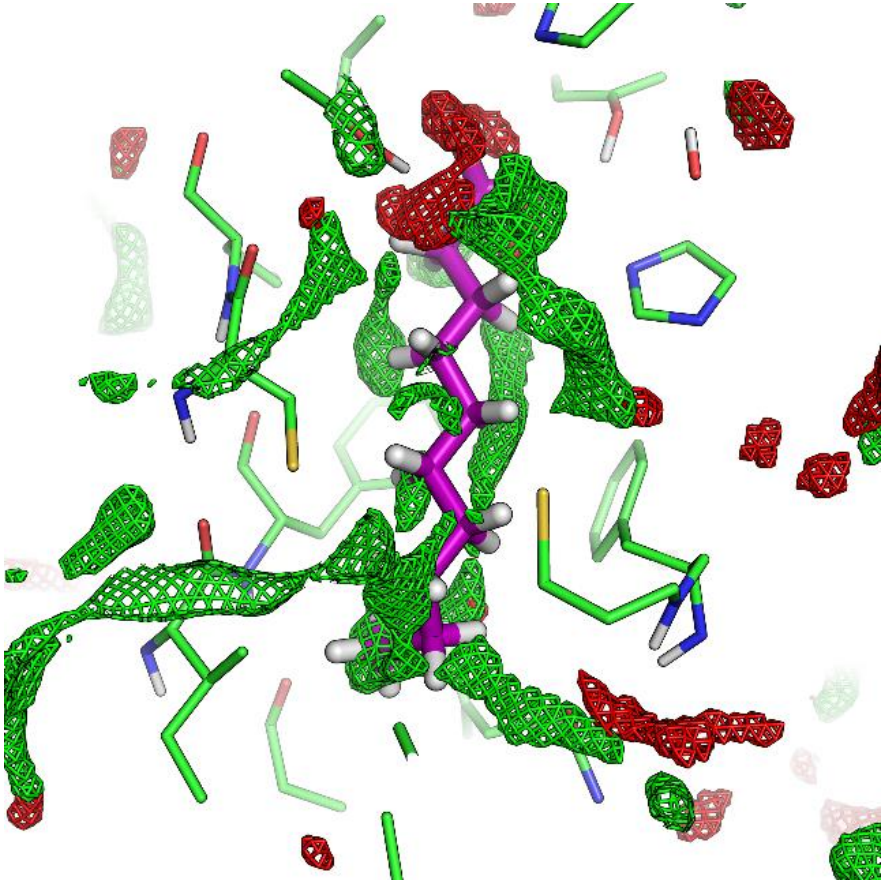
Simulation-based methods

- Application starts out with available structural data
- Propagating atomic positions for a set of molecules yields highly correlated individual data points
- Millions of snapshots (results have to be extracted from trajectory)
- Tools for analysis often originate from statistics (Markov chain models, correlation analysis)

Molecular Simulations – Volumetric Data

Data structures

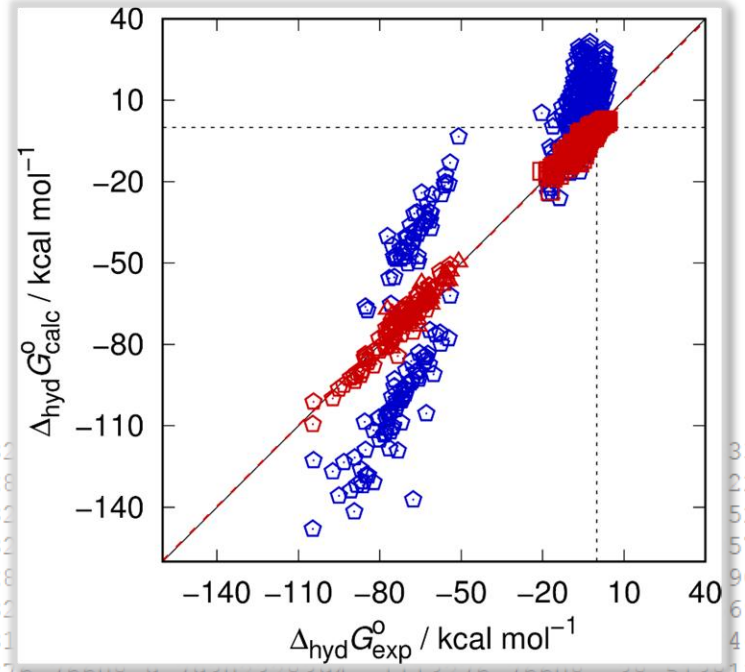
- Distribution of solvent around a solute
- Volumetric data on 3D grids (density or energy value at each point)
- Large files → storage problem
- Volumetric data collapsed to scalars
- Both data types can be used for machine learning



Quantum-Mechanical Data

Compensating for physical approximations

- Clean data but errors caused by physical approximations
- Complexity unaccounted for by high level data fitted through regression models
- Expensive and slow (larger systems can be unfeasible due to exponential scaling)

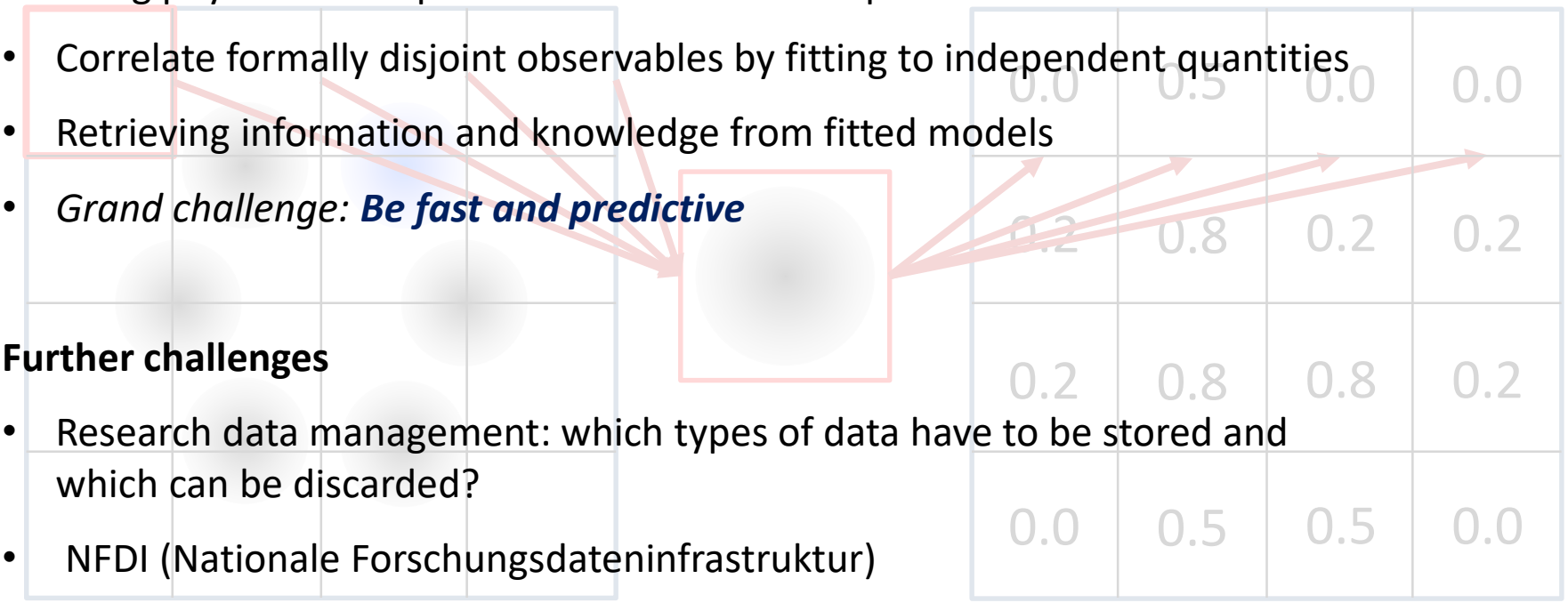



```
neutrals,SM18,micro003,S01,-1113242.20834,-1109904.41787,True,-1113
neutrals,SM18,micro007,S01,-1113271.22061,-1109943.6758,True,-111328
neutrals,SM18,micro009,S01,-1113286.03263,-1109986.46144,True,-11132
neutrals,SM18,micro009,S02,-1113285.21502,-1109985.75923,True,-11132
neutrals,SM18,micro013,S01,-1113288.44152,-1109988.1452,True,-111328
neutrals,SM18,micro013,S02,-1113292.41729,-1109991.79991,True,-11132
neutrals,SM18,micro016,S01,-1113326.8789,-1110017.70494,True,-111331
neutrals,SM18,micro017,S01,-1113286.5591,-1109980.09421,True,-111327
neutrals,SM18,micro017,S02,-1113299.38421,-1109986.88735,True,-1113277.34918,22.0350310684,-1113277.34918,-22.7456950335
neutrals,SM18,micro021,S01,-1113291.49703,-1109975.65747,True,-1113300.21216,-8.71513052342,-1113300.21216,-55.7448275096
neutrals,SM18,micro021,S02,-1113296.38389,-1109986.76957,True,-1113296.10686,0.277026900096,-1113296.10686,-43.6335236424
neutrals,SM18,micro025,S01,-1113299.92996,-1109994.60138,True,-1113302.0032,-2.073242163,-1113302.0032,-44.7435672945
neutrals,SM18,micro028,S01,-1113320.71764,-1110012.35759,True,-1113307.42336,13.2942850167,-1113307.42336,-32.2799362811
neutrals,SM18,micro030,S01,-1113261.02623,-1109922.22863,True,-1113285.14878,-24.1225548924,-1113285.14878,-72.5185245315
neutrals,SM18,micro038,S01,-1113294.75044,-1109993.19018,True,-1113285.19193,9.55851844168,-1113285.19193,-33.1260027151
neutrals,SM18,micro038,S02,-1113295.27963,-1109993.53174,True,-1113285.03792,10.2417119694,-1113285.03792,-32.3006789651
neutrals,SM18,micro042,S01,-1113318.84815,-1110009.24597,True,-1113310.82114,8.02701119503,-1113310.82114,-35.6188153991
neutrals,SM18,micro047,S01,-1113328.58087,-1110019.82569,True,-1113314.55627,14.0245954398,-1113314.55627,-27.37161076
```

Regression Models on Synthetic Data

Practical applications of Data Science

- Correcting errors within the data generated using physics-based methods by training low level models
- Using physics to compensate for insufficient experimental data sources
- Correlate formally disjoint observables by fitting to independent quantities
- Retrieving information and knowledge from fitted models
- *Grand challenge:* ***Be fast and predictive***



Further challenges

- Research data management: which types of data have to be stored and which can be discarded?
- NFDI (Nationale Forschungsdateninfrastruktur)
- Integration of Data Science in curricula