A Perspective on Autonomous Experimentation and Discovery

Marcus Michael Noack
What is Autonomous Data Acquisition

CAMERA Workshop on AE and Main Takeaways

Basic Gaussian-Process-Driven Autonomous Data Acquisition

Mathematical Optimization for AE and ML

Bringing Autonomous Discovery to the Community
What is Autonomous Data Acquisition
The Human-Driven Experiment Loop

Detector image

Sample
Depends on synthesis, processing and environmental parameters

Instrument
Depends on instrument parameters

Select sample parameters

Trigger measurement

A function over the parameter space

Automated data analysis

Human Scientist
<table>
<thead>
<tr>
<th></th>
<th>Random</th>
<th>Scanning</th>
<th>Intuition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inefficient</td>
<td>Prev. data not used/Non-optimal</td>
<td>Prev. data not used/Non-optimal measurements</td>
<td>Not autonomous/human attention required</td>
</tr>
<tr>
<td>Uninformative</td>
<td>No metric for quality</td>
<td>No metric for quality</td>
<td>No metric for quality</td>
</tr>
<tr>
<td>Biased</td>
<td></td>
<td></td>
<td>Relies on past experience Lack of reproducibility</td>
</tr>
</tbody>
</table>
The Autonomous Experiment Loop

Automated Sample Preparation
- In Situ/Ex Situ
- 3D Printing

Automated Data Acquisition
- Robotics
- Remote Access

Communication
- By File
- ZMQ
- S3

Automated Data Analysis/Dim. Reduction
- PCA
- NMF
- NNs
- Integration
- Peak Finding

Autonomous Decision Making
- Stochastic Processes
- Reinforcement Learning
- Optimization

Active Learning
ML (AI)

Supervised Learning
- Labeled data

Unsupervised Learning
- Unlabeled data

Active Learning

Mathematics:
- Function Appr.
- Optimization
- Topology
- Lin. Algebra
- Abstract Algebra
- Calculus
- Probability Theory
- Statistics

A Venn Diagram for ML and Active Learning
What is Autonomous Data Acquisition

CAMERA Workshop on AE and Main Takeaways

Basic Gaussian-Process-Driven Autonomous Data Acquisition

Mathematical Optimization for AE and ML

Bringing Autonomous Discovery to the Community
CAMERA Workshop on AE and Main Takeaways
Autonomous Discovery for Science and Engineering

A three-day workshop for sharing recent developments in autonomous methods, sponsored by CAMERA — The Center for Advanced Mathematics for Energy Research Applications

Dates: April 20th – 22nd, 2021

Methods and Algorithms: Gaussian Processes, Neural Networks, Reinforcement Learning, New Math, Optimization, Data Analytics and Infrastructure, UQ

Tutorials: gpCAM, Summit, CamLink, Escalate, Bluesky, Atinary SDLabs, DataFed, Cameo, AtomAI, ART


https://autonomous-discovery.lbl.gov/
DOI: https://doi.org/10.2172/1818491
The Organizing Committee

- Marcus Noack, LBNL
- Petrus Zwart, LBNL
- Apurva Mehta, SLAC
- Daniela Ushizima, LBNL
- James Sethian, UCB
- Kevin Yager, BNL
- Simon Billinge, Columbia U.
- Héctor García Martín, LBNL
- Nicholas Schwarz, ANL
- B. Reeja Jayan, CMU
- Eva M. Herzig, Bayreuth U.
- Alex Hexemer, LBNL
- Bobby Sumpter, ORNL
- Martin Boehm, ILL
- Sergei Kalinin, ORNL
- Aaron Gilad Kusne, NIST
What is your background?
512 responses

- Physics 11.4%
- Mathematics 20.3%
- Chemistry 9.7%
- Biology 8.2%
- Materials Sciences 21.9%
- Data Science
- Environmental Science
- Geoscience

- Computer Science
- Chemical Engineering
- Chemical Engineering
- Mechanical Engineering
- Engineering
- Computer science
- engineering
- Electrical Engineering

- Computational Biology
- Radiation Therapy
- Lead a Data Science team at PNNL b.
- Computer Science, Data Science
- X-ray science
- Bioengineering
- computer science
- Chemical Engineer

- Computer Science + Math
- PhD student in Materials Science /
- Chemistry, Biology, Environmental s
- Mechanical Engineering + Materials
- Energy Sciences
- Physics/geoscience
- mechanical engineering
- Robotics

- Chemical Physics
- Computer science and control
- IT
- computational biophysics
- chemical engineering
- Operations Research
- Math, Stats, biology, and environment...
- Quantum Computing

- computational mechanics
- Application to IT Infrastructure
- Engineering in support of Joint Genome Institute.
- Computer Engineering
- Optimization, Chemical engineering, machine learning
- social sciences, ethics
Take-Home Message 1: Gaussian processes and reinforcement learning are the most popular techniques for control, in different data regimes.

Take-Home Message 4: Efficient mathematical optimization under uncertainty and subject to constraints has come far but remains a challenge.

Take-Home Message 2: The data-analysis step is more and more often done automatically data-science tools (PCA, NN, Clustering,...).

Take-Home Message 3: Instrument systems are increasingly built with automation in mind.

Take-Home Message 5: Data-management systems are emerging using lower-level tool (control, optimization, ...) to allow for standardization.
Takeaways and Achievements: Application

X-Ray Scattering

Autonomous Analytics and Control in X-ray Scattering

Kevin G. Yager
Center for Functional Nanomaterials, Brookhaven National Laboratory

Materials Science

Autonomous Research Systems for Materials Development

Benji Marynauskas
1 Air Force Research Laboratory
benji.marynauskas@us.af.mil

Neutron Scattering

Improving the Measurement Strategy in Neutron Spectroscopy with Machine Learning

Martin Binder1,2, Tobias Weber1, Vassiliki Leog1, Marcus Noack1, Paolo Malnani1
1 Leibniz Institute for Solid State and Materials Research (IFW), Dresden, Germany
2 Humboldt University, Berlin, Germany

Interpretability and uncertainty quantification for machine learning analysis of neutron medium scattering

Keith T. Butler1,2,3, Maosheng Diao1,2,3, Jaydeep Thayyilagum1,3, Toby G. Porritt1,2
1 SciML, Scientific Computing Department, STFC Rutherford Appleton Laboratory, Oxford, UK
2 International Centre for Simulation of Materials, University of Oxford, Oxford, UK
3 Department of Chemistry, University of Oxford, Oxford, UK

Spectroscopy

SEM Image Deconvolution with Unsupervised Machine Learning for Better Defect Inspection and Metrology

Baspajyoti Dey1,3,4,5, Rosalia Holthus1,2, Karen Keffel1,2, Glenn Leinner1,2, Jean sewer1,2, Philippe Leoni3,4, Magdi A. Baydoun5
1 imec, Kapeldreef 15, 3001 Leuven, Belgium
2 The Center for Advanced Computer Studies, University of Louisiana at Lafayette, Louisiana, USA
3 Department of Electrical and Computer Engineering, University of Louisiana at Lafayette, Louisiana, USA
4 bapajyoti.dey@imec.be

Biology

Developing Fabricated Ecosystems to Harnessed Plant-Microbe Interactions

Peter F. Auseler1, Trent Northrup1
1 Environmental Geotechnics and Sustainable Biology, Lawrence Berkeley National Laboratory, Berkeley, CA, 94720
2 Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA, 94720
3 TNorthup@lbl.gov

Vision for the integration of fabricated ecosystems with field
Challenges

The Role of Co-Design: Much of the work is performed through co-design teams, bringing together needed expertise. The work has aspects of theory, modeling, algorithm design, data analysis, workflow, and software engineering.

Integrating Across Required Expertise: Teams (or in some cases, individuals) working in autonomous design often take on all the required roles, which requires a large breadth of expertise: it is challenging for a team to excel in all the necessary aspects.

Sharing Developments: There are significant opportunities to share advances across autonomous efforts. However, there is often inconsistent nomenclature and problem formulation.

Workflow and Infrastructure: For the most part, individual efforts center around homegrown workflows and infrastructures. There are opportunities to build work tools and infrastructures that can be shared.

Software: Understandably, many of the efforts described within are aimed at solving a particular set of scientific problems, and the emphasis is not on generalizable software. Openly-available software that is well-documented and properly maintained would be a step forward.

Shared Testbeds and Reproducible Research: It is challenging to cross-test different algorithms and methodologies with common accessible (FAIR) datasets with maintained standards.

Data: Data ≠ data. The structure of data has to be discussed before data taking and ML applications.

Compute Resources: The field requires a new kind of compute-resource allocation, which keeps resources available throughout the experiment.
What is Autonomous Data Acquisition

CAMERA Workshop on AE and Main Takeaways

Basic Gaussian-Process-Driven Autonomous Data Acquisition

Mathematical Optimization for AE and ML

Bringing Autonomous Discovery to the Community
Basic Gaussian-Process-Driven Autonomous Data Acquisition
Gaussian Process Regression in a Nutshell

\[ p(f) = \frac{1}{2\pi^{d/2}} |\Sigma|^{-1/2} \exp\left[-\frac{1}{2}(f - \mu)^T \Sigma^{-1} (f - \mu)\right] \]

\( \mathcal{H} = \{ f(x) : f(x) = \sum_{i}^{N} \alpha_i k(x_i, x), \forall \alpha \in \mathbb{R}^N, x \in \mathbb{R}^n \} \)

\[
m(x_0) = \mu + k^T (K + I_e)^{-1} (y - \mu) \]

\[
\sigma^2(x_0) = k(x_0, x_0) - k^T (K + I_e)^{-1} k
\]
acquisition function
acquisition function
acquisition function
acquisition function
Our Very First Experiment: A Nanoparticle Stain Mapping Experiment
Facility: NSLS2, CFN @ BNL | Technique: SAXS | Achievement: Commissioning experiment

A Kriging-Based Approach to Autonomous Experimentation with Applications to X-Ray Scattering
Marcus M. Noack, Kevin G. Yager, Masafumi Fukuto, Gregory S. Doerk, Ruipeng Li & James A. Sethian
Autonomous SAXS Exploration of Nanoscale Ordering in a Blade-Coated Polymer-Grafted Nanorod Film

Facility: AFRL and NSLS II | Technique: SAXS | Achievement: 15% of data required, higher resolution in areas of interest

Grain Size

J. Streit, R. Vaia (AFRL), M. Fukuto, R. Li (BNL/NSLS-II), K. Yager (BNL/CFN), M. Noack (LBNL/CAMERA)
Autonomous Steering of ARPES Data Acquisition
Facilities: ALS @ LBNL | Technique: ARPES | Achievement: 12% of data required

K-Means-Driven Gaussian Process Data Collection for Angle-Resolved Photoemission Spectroscopy
Autonomous Control of Synchrotron Infrared Spectroscopy
Facility: ALS @ LBNL | Technique: IR Spec. Micr. | Achievement: ~5% of data required, collected in ~10% of the time, materials targeted
Autonomous Scanning Tunneling Spectroscopy
Facility: Molecular Foundry @ LBNL | Technique: STS Microscopy | Achievement: ~4% of data required, ~35 hrs vs ~1 mo. acq. time

Scanning Probe Microscopy

STM / STS
- structure and electronic properties

Investigate Next Frontiers in 2D Quantum Materials

SiC(0001)

Graphene

WS₂

STS Experiment

Classification

Thomas et al., arXiv:2110.03351 (2021)
The Power of the RKHS: Domain-Informed Symmetry Constraints — Six-Fold Symmetry

\[ k(x_i, x_j) = \frac{1}{36} \sum_{\phi} \sum_{\theta} \tilde{k}(R_\phi x_i, R_\theta x_j) \]
Physics Knowledge in the Form of Periodicity for X-Ray Scattering

Facility: NIST and NSLS II | Technique: SAXS/GISAXS | Achievement: Use of non-stationary kernels to learn and exploit local characteristics

Kevin Yager, Masafumi Fukuto, Jonathan Seppala @ CFN, BNL, NIST
Physics-Based Model

Data-Driven GPR

Physics-Aware GPR

Physics-Aware Prediction of Lattice Thermal Conductivity of Alloys

Facility: SLAC @ Stanford | Technique: Diffusivity, Heat Capacity and Density Measurements | Achievement: Physics-informed GP-driven steering
Targeted Autonomous Neutron Scattering
Facility: ILL, France | Technique: Inelastic Neutron Scattering | Achievement: More efficient exploration, experiment time decreased from several days to one night
Mathematical Optimization for AE and ML
The Traditional Training/Optimization Workflow needs a Large Number of Function Evaluations and Blocks the Main Thread
Minimizing Number of Function Evaluations: Asynchronous Distributed Training

Kill/Restart/Ingest New Data
Optimization of the Log–Likelihood and Acquisition Functions with HGDL

Using DASK, pytorch and GPUs for High Performance Asynchronous Distributed Training

\[ \log(L(D; \phi, \mu)) = \]
\[ -\frac{1}{2} (y - \mu(x))^T (K(\phi) + I_e)^{-1} (y - \mu(x)) - \frac{1}{2} \log(|K(\phi) + I_e|) \]

HGDL leads to:
1. a set of different interpretations of the data
2. a set of optimal measurements
3. HPC readiness of training and prediction
4. Asynchronous training
Mathematical Optimization for AE and ML

Basic Gaussian-Process-Driven Autonomous Data Acquisition

What is Autonomous Data Acquisition

CAMERA Workshop on AE and Main Takeaways

Bringing Autonomous Discovery to the Community
Bringing Autonomous Discovery to the Community
fvGP: A flexible multi-task Gaussian process tool

HGDL: Asynch. Distributed Optimizer
The products are three separate APIs that are built on top of each other:

- **HGDL**: Asynchronous Distributed Optimizer
- **fvGP**: A flexible multi-task Gaussian process tool
- **gmC**: Multi-Task Global Optimization
pip install gpcam

```python
from gpcam.autonomous_experimenter import AutonomousExperimenterGP
from instrument import instrument
import numpy as np

parameters = np.array([[3.0, 45.8], [4.0, 47.0]])
init_hyperparameters = np.array([1, 1, 1])

hyperparameter_bounds = np.array([[0.01, 100], [0.01, 100.0], [0.01, 100]])

my_ae = AutonomousExperimenterGP(parameters, instrument, init_hyperparameters,
                    hyperparameter_bounds, init_dataset_size=10)

my_ae.train()

my_ae.go()

def instrument(data):
    for entry in data:
        entry["value"] = np.sin(np.linalg.norm(entry["position"]))
    return data
```

More information: [gpcam.lbl.gov](https://gpcam.lbl.gov)
A Perspective on Autonomous Experimentation and Discovery

Marcus Michael Noack
MarcusNoack@lbl.gov
gpcam.lbl.gov