PyData [/]

Schedule

The PyData Berlin 2017 call for proposals received a good response and we're very happy with the program. Below is the full list of talks and tutorials accepted to PyData Berlin 2017. We will publish a full schedule later in May.

Keynote Speakers
Panel Discussion
Talks
Tutorials

Keynote Speakers

Panel Discussion Talks Tutorials

Barbara Plank

Veronika Valeros

Toby Walsh *Many Fears About AI Are Wrong*

Should you be worried about progress in Artificial Intelligence? Will Artificial Intelligence destroy jobs? Should we fear killer robots? Does

Artificial Intelligence threaten our very existence? Artificial Intelligence (AI) is in the zeitgeist. Billions of dollars are being poured into the field, and spectacular advances are being announced regularly. Not surprisingly, many people are starting to worry where this will all end. The Chief Economist of the Bank of England predicted that Artificial Intelligence will destroy 50% of existing jobs. Thousands of Artificial Intelligence researchers signed an Open Letter predicting that Artificial Intelligence could transform warfare and lead to an arms race of "killer robots". And Stephen Hawking, Elon Musk and others have predicted that Artificial Intelligence could end humanity itself. What should you make of all these predictions? Should you worry? Are you worrying about the right things? And what should we do now to ensure a safe and prosperous future for all?

Panel Discussion

Keynote Speakers Talks Tutorials

We will have a panel discussion on *Ethics in Machine Learning*. Stay tuned for more details.

Talks

Keynote Speakers Panel Discussion Tutorials

Abhishek Thakur, *Is That a Duplicate Quora Question?*
Quora released its first ever dataset publicly on 24th Jan, 2017. This dataset consists of question pairs which are either duplicate or not.

Duplicate questions mean the same thing. In this talk, we discuss methods which can be used to detect duplicate questions using Quora dataset. Of course, these methods can be used for other similar datasets.

Aisha Bello, *Spying on my Network for a Day: Data Analysis for Networks.*
In this talk I would show how I used open source tools like Moloch, Wireshark, Bokeh and Jupyter to analyse my home network data for the day. Not just the volume of data generated daily, but how interesting it is to leverage data tools to discover when your network is experiencing downtime which could be as a result of packet loss, poorly placed Access points or just proximity away from your rout

Alexander Weiss, *Size Matters! A/B Testing When Not Knowing Your Number of Trials*
If an A/B test's statistical evaluation is based on the number of trials and successes for each variation, what happens if you don't know the number of trials for your experiment? Although this question sounds rather theoretical, we faced it in practice. This talk is about our search for the right question to ask.

Alexey Grigorev, *Large Scale Vandalism Detection in Knowledge Bases*
Wikidata is a Knowledge Base where anybody can add new information. Unfortunately, it is targeted by vandals, who put inaccurate or offensive information there. To fight them, Wikidata employs moderators, who manually inspect each suggested edit. In this talk we will look into how we can use Machine Learning to automatically detect vandalic revisions and help the moderators.

Amit Steinberg, *Data Analytics and the new European Privacy Legislation*
They upcoming privacy legislation of the EU will radically change the way we do data analytics, restricting the processing of personally identifiable

data. We will go through common data processing scenarios and learn how the new legislation will affect them, offering practical solutions.

Andreas Dewes, *Fairness and transparency in machine learning: Tools and techniques*
This talk will try to answer a simple question: When building machine learning systems, how can we make sure that they treat people fairly and can be held accountable? While seemingly trivial, this question is not easy to answer, especially when using complex methods like deep learning. I will discuss tools and techniques that we can use to make sure our algorithms behave as they should.

Benedikt Koehler, *Crunching Blockchain data with Python and Jupyter*
Bitcoin and the Blockchain are one of the most hyped technology trends right now. In this talk, we'll take a look at the data side of the Blockchain. With many practical example this talk will demonstrate how to download the Blockchain, analyze it with Python to examine the network structure and also to trace some of the famous events in Bitcoin history - involving pizza and the FBI.

Carlotta Schatten, *Towards Pythonic Innovation in Recommender Systems*
Recommender Systems are nowadays ubiquitous in our lives. Although many implementations of basic algorithms are well known, recent advances in the field are often less documented. This talks aims at reviewing available Recommender Systems libraries in Python, including cutting edge Time- and Context-aware state of the art models.

Daniele Rapati, *Engage the Hyper-Python - a rattle-through many of he ways you can make a Python program faster*
A fast paced high-level overview of speed optimisation in Python. What makes a program "slow"? How to tell what is making your program slow. Common speed-up paradigms: parallelization, alternatives to the regular

Python interpreter and asynchronous processing.

David Soares Batista, *Semi-Supervised Bootstrapping of Relationship Extractors with Distributional Semantics*
Semi-supervised bootstrapping techniques for relationship extraction from text iteratively expand a set of initial seed relationships while limiting the semantic drift. This talk presents an approach to bootstrap relationship instances using word embeddings to find similar relationships. Results show that relying on word embeddings achieves a better performance than using TF-IDF weighted vectors.

Emily Gorcenski, *Polynomial Chaos: A technique for modeling uncertainty*
Parametric uncertainty is broadly difficult to quantify. In particular, when those parameters don't fit nice distributions it can be hard to generate reasonable simulations. Polynomial chaos is a somewhat obscure technique that leverages a natural connection between probability distributions and orthogonal polynomial families. This talk will demonstrate the technique and its applications.

Florian Wilhelm, *“Which car fits my life?” - mobile.de’s approach to recommendations*
As Germany’s largest online vehicle marketplace mobile.de uses recommendations at scale to help users find the perfect car. We elaborate on collaborative & content-based filtering as well as a hybrid approach addressing the problem of a fast-changing inventory. We then dive into the technical implementation of the recommendation engine, outlining the various challenges faced and experiences made.

Françoise Provencher, *Biases are bugs: algorithm fairness and machine learning ethics*
Biases are bugs. They need to be found, fixed, and learnt from. A mix of good ethics and good engineering practices can get us a long way

towards that goal. In this talk you'll learn what biases are, what software tools can help, and how to adopt engineering practices that can make your algorithms fairer.

Hendrik Heuer, *Data Science for Digital Humanities: Extracting meaning from Images and Text*
Analyzing millions of images and enormous text sources using machine learning and deep learning techniques is simple and straightforward in the Python ecosystem. Powerful machine learning algorithms and interactive visualization frameworks make it easy to conduct and communicate large scale experiments. Exploring this data can yield new insights for researchers, journalists, and businesses.

Héctor Andrade Loarca, *Fast Multidimensional Signal Processing using Julia with Shearlab.jl*
Shearlab is a Julia Library with toolbox for two- and threedimensional data processing using the Shearlet system as basis functions which generates a sparse representation of cartoon-like functions with applications on Signal Processing, Compressed Sensing, 3D Imaging, MRI Imaging and a lot more, with visible improvements with respect of the Wavelet Transform in representing multidimensional data.

Irina Vidal Migallon, *Deep Learning for detection on a phone: how to stay sane and build a pipeline you can trust*
Deploying a deep model on a mobile device to be used for real-time detection is not quite trivial yet. Defining your Deep Learning architecture, gathering the right data, designing your training process, evaluating your models and turning this into a pipeline that keeps everyone on the team (somewhat) sane - these all have their pitfalls.

Jonathan Ronen, *Social Networks and Protest Participation: Evidence from 130 Million Twitter Users*
Data mining social networks for evidence of political participation. A

demonstration of python being used to data mine the twitter conversations around the #JeSuisCharlie hashtag, and analyzing it to learn about real world protest behavior.

Karan Saxena, *Ranking News Summary Bots by Semantic Document Relatedness*
Semantic relatedness, or similarity between documents plays an important role in many textual applications. Text understanding starts with the challenge of finding machine-understandable representation that captures the semantics of texts. We explore the issue of document similarity with various existing language models, examining them in the task by ranking different News Summary Bots.

Karolina Alexiou, *Patterns for Collaboration between Data Scientists And Software Engineers*
The talk is going to present, with examples, how a software engineer team can work together with data scientists (both in-house and external collaborators) in order to leverage their unique domain knowledge and skills in analyzing data, while supporting them to work independently and making sure that their work can be constantly tested/evaluated and easily integrated into the larger product.

Lev Konstantinovskiy, *Find the text similiarity you need with the next generation of word embeddings in Gensim*
What is the closest word to "king"? Is it "Canute" or is it "crowned"? There are many ways to define "similar words" and "similar texts". Depending on your definition you should choose a word embedding to use. There is a new generation of word embeddings added to Gensim open source NLP package using morphological information and learning-to-rank: Facebook's FastText, VarEmbed and WordRank.

Matti Lyra, *Evaluating unsupervised models for text*
Unsupervised models in natural language processing (NLP) have become

very popular recently. Word2vec, GloVe and LDA provide powerful computational tools to deal with natural language and make exploring large document collections feasible. We would like to be able to say if a model is objectively good or bad, and compare different models to each other, this is often tricky to do in practice.

Max Humber, *Patsy: The Lingua Franca to and from R*
How to build R-like statistical models in Python with Patsy and scikit-learn.

Miguel Vaz, *Hedging portfolios with Reinforcement Learning*
Math finance models, such as the famous Black Scholes model for option pricing, produce useful (and beautiful) results, sometimes relying on strong assumptions. Machine learning allows for more flexible models, but are they as good as the traditional ones? With a few toy examples, we'll see how an exploring agent compares with traditional approaches.

Miroslav Batchkarov, *Gold standard data: lessons from the trenches*
The first stage in a data science project is often to collect training data. However, getting a good data set is surprisingly tricky and takes longer than one expects. This talk describes our experiences in labelling gold-standard data and the lessons we learnt the hard way. We will present three case studies from natural language processing and discuss the challenges we encountered.

Nick Radcliffe, *Developments in Test-Driven Data Analysis*
Test-driven data analysis fuses and builds upon the ideas of test-driven development and reproducible research to support higher quality data analysis. This talk will extend the foundation parts of TDDA with extensions including tight constraints on string fields with automatically discovered regular expressions and automatically discovered relationships between datasets.

Oliver Eberle, *Where are we looking? Prediciting human gaze using deep networks.*
Which features in an image draw our focus to a specific area while neglecting others entirely? This fascinating question has been motivating researchers for decades but also sparked interest in design and marketing. Thus, saliency models aim at identifying locations that stand out from their visual neighbourhood. Using tensorflow and matplotlib this talk will shed some light on these features..

Radovan Kavicky, *Data Science & Data Visualization in Python. How to harness power of Python for social good?*
Python as an Open Data Science tool offers many libraries for data visualization and I will show you how to use and combine the best. I strongly believe that power of data is not only in the information & insight that data can provide us, Data is and can be really beautiful and can not only transform our perception but also the world that we all live in.

Rafael Schultze-Kraft, *Building smart IoT applications with Python and Spark*
In this talk I will present how we use Python, PySpark and AWS as our preferred data science stack for the Internet of Things, which allows us to efficiently develop and deploy smart data applications on top of IoT sensor data. We use these technologies to analyse and model IoT timeseries data, as well as to build automated and scalable data pipelines for smart IoT data applications in the cloud.

Raphael Pierzina, *Kickstarting projects with Cookiecutter*
Cookiecutter is a command-line utility that generates projects from templates. You can use Cookiecutter to create new Python projects and to generate the initial code for different types of applications. In this talk, I will give an introduction to Cookiecutter, how to install it from PyPI, how to use it in the CLI, and finally how to author your own

template.

Rebecca Bilbro, *Yellowbrick: Steering Machine Learning with Visual Transformers*
Yellowbrick is a new library that extends Scikit-Learn's API to incorporate visualizations into machine learning. While the machine learning workflow is increasingly being automated with gridsearch, APIs, and GUIs, in practice, human intuition outperforms exhaustive search. By visualizing model selection, we can not only steer towards robust models, but also avoid common pitfalls and traps.

Robert Meyer, *Analysing user comments on news articels with Doc2Vec and Machine Learning classification*
I used the Doc2Vec framework to analyze user comments on German online news articles and uncovered some interesting relations among the data. Furthermore, I fed the resulting Doc2Vec document embeddings as inputs to a supervised machine learning classifier. Can we determine for a particular user comment from which news site it originated?

Ross Kippenbrock, *Finding Lane Lines for Self Driving Cars*
Self-driving cars might not be in our everyday lives yet, but they are coming! Analyzing images and figuring out where the lane lines are on a given roadway is one of the core competencies of any respectable self-driving car. Humans do this with ease and this talk will show you how to find these lines using Python and OpenCV.

Sirin Odrowski, *Introduction to Search*
Search engines make archives, inventories, websites, large publishing platforms, and the internet navigable. Using the search engine we built at WordPress.com, a platform where more than 90 million blog posts and web pages are published per month, as an example, I will explain how search engines work and how their performance can be evaluated.

Stefan Otte, *On Bandits, Bayes, and swipes: gamification of search*
The talk will show how to use active learning to work with Small Data. Active learning is an underappreciated subfield of ML where the algorithm actively gathers labeled data, e.g. it can query the user for the most informative data. I will discuss the basics of active learning theory, and look at a case study showing how to use active learning and tailor it to a practical problem.

Tal Perry, *A word is worth a thousand pictures: Convolutional methods for text*
Those folks in computer vision keep publishing amazing ideas about you to apply convolutions to images. What about those of us who work with text? Can't we enjoy convolutions as well? In this talk I'll review some convolutional architectures that worked great for images and were adapted to text and confront the hardest parts of getting them to work in Tensorflow .

Thomas Kober, *What does it all mean? - Compositional distributional semantics for modelling natural language*
Distributional semantic word representations have become an integral part in numerous natural language processing pipelines in academia and industry. An open question is how these elementary representations can be composed to capture the meaning of longer units of text. In this talk, I will give an overview of compositional distributional models, their applications and current research directions.

Trent McConaghy, *Blockchains for Artificial Intelligence*
This talk describes the various ways in which emerging blockchain technologies can be helpful for machine learning / artificial intelligence work, from audit trails on data to decentralized model exchanges.

Ulrike Thalheim, *When the grassroots grow stronger - 2017 through the eyes of German open data activists*

The talks will give an overview about recent legislative changes that will lead to much more Open Data in Germany. In a second part, the talk will show the implications to the existing Open Data community. Three very different examples of current Open Data projects from the Code for Germany community will be presented.

Vaibhav Singh, *Machine Learning to moderate ads in real world classified's business*
In todays world of online business, it is difficult to moderate all the content coming to your site. In this talk we share our experiences on how we built machine learning models to moderate 100+ million classified ads every month. Audience will get a chance to experience a real world of content moderation and a race to beat online fraudsters and scammers.

Vincent D. Warmerdam, *TNaaS - Tech Names as a Service*
In this talk I will explain how I built a service that generates Pokemon names. You'd be surprised how hard it is to do this properly and how easy it is to do it practically.

Tutorials

Keynote Speakers Panel Discussion Talks

Adrin Jalali, *The path between developing and serving machine learning models.*
As a data scientist, one of the challenges after you develop and train your model, is to deploy it in production where other systems would use the output of the model in real time. In this tutorial we use PipelineIO, to deploy a cluster on the cloud, which gives us a JupyterHub to develop our method, and uses PMML to persist and deploy and serve the model.

Alexander Hendorf, *Introduction to Data-Analysis with Pandas*
Pandas is the Swiss-Multipurpose Knife for Data Analysis in Python. With Pandas dealing with data-analysis is easy and simple but there are some things you need to get your head around first as Data-Frames and Data-Series. The tutorial provides a compact introduction to Pandas for beginners for I/O, data visualisation, statistical data analysis and aggregation within Jupiter notebooks.

Alexandru Agachi, *Introductory tutorial on data exploration and statistical models*
This tutorial will focus on analyzing a dataset and building statistical models from it. We will describe and visualize the data. We will then build and analyze statistical models, including linear and logistic regression, as well as chi-square tests of independence. We will then apply 4 machine learning techniques to the dataset: decision trees, random forests, lasso regression, and clustering.

Bhargav Srinivasa Desikan, *Topic Modelling (and more) with NLP framework Gensim*
Topic Modelling is an information retrieval technique to identify topics in a large corpus of text documents. This tutorial will guide you through the process of analysing your textual data through topic modelling - from finding and cleaning your data, pre-processing using spaCy, applying topic modelling algorithms using gensim - before moving on to more advanced textual analysis techniques.

David Higgins, *Introduction to Julia for Scientific Computing and Data Science*
Developed at MIT, with a focus on fast numerical computing, Julia has a syntactical complexity similar to that of Python or Matlab but a performance orders of magnitude faster. We will present an introduction to the language, followed by a sampling of some of our favourite packages. The focus is on which aspects of Julia are currently

ready for use by numerical computing and data scientists.

Dr. Kristian Rother, *Best Practices for Debugging*
This tutorial introduces concepts and techniques for systematic debugging. Participants will debug example programs with different kinds of bugs and with increasing difficulty.

Gerrit Gruben, *Leveling up your Jupyter notebook skills*
Most of us regularly work with Jupyter notebooks, but fail to see obvious productivity gains involving its usage. Did you know that the web interface works like a modal editor such as VIM? Do you know that you can actually profile AND debug code in notebooks? How about setting formulas or use pre-made style settings for visualizations? Let us go through the tricks of the trade together!

Jo-fai Chow, *Introduction to Machine Learning with H2O and Python*
H2O.ai is focused on bringing AI to businesses through software. Its flagship product is H2O, the leading open source platform that makes it easy for financial services, insurance and healthcare companies to deploy machine learning and predictive analytics to solve complex problems. This tutorial aims to demonstrate the basic usage of H2O with worked examples in Python.

Stephen Simmons, *Pandas from the Inside / "Big Pandas"*
Pandas is great for data analysis in Python. It promises intuitive DataFrames from R; speed like numpy; groupby like SQL. But there are plenty of pitfalls. This tutorial looks inside pandas to see how DataFrames actually work when building, indexing and grouping tables. You will learn how to write fast, efficient code, and how to scale up to bigger problems with libraries like Dask.

Subscribe to Receive PyData Updates

email address

Submit

Tickets

Get Now

IMPRESSUM

PyData is a group for users and developers of data analysis tools to share ideas and learn from each other. We gather to discuss how best to apply Python tools, as well as those using R and Julia, to meet the evolving challenges in data management, processing, analytics, and visualization. PyData groups, events, and conferences aim to provide a venue for users across all the various domains of data analysis to share their experiences and their techniques. PyData is organized by NumFOCUS.org, a 501(c)3 non-profit in the United States.

NumFOCUS
P.O. Box 90596
Austin, TX 78709
info@numfocus.org
NumFOCUS Executive Director Leah Silen (+1 512-222-5449; leah@numfocus.org)

You can find more detailed information about NumFOCUS on its website (https://www.numfocus.org/about/legal/).

You can also reach the local Berlin organisers at
info@pydata.berlin
+49 (0)30 568 22 153

CONTACT

ACCOUNT

Login (/berlin2017/account/login/)
Sign up (/berlin2017/account/signup/)

THE PYDATA CONFERENCE SERIES IS BROUGHT TO YOU BY

NUMFOCUS (http://numfocus.org)
OPEN CODE = BETTER SCIENCE