GAP
DATA
INSTITUTE

Data Science in Python: Past, Present & Future

(PyData Track @ PyConLT 2019, SMK; #PyConLT19 #PyConLT)

25. 5. 2019

Radovan Kavický, GapData Institute & PyData Bratislava #PyDataBA

Talk Structure

▯ **1. Introduction, History of Data Science & Why Python for Data Science?/Motivation** (5-8 min.)

- **Humble Beginnings of Data Science**
- **What is Data Science & Data Science Workflow**
- **History of Python & Future of Data Science**

▯ **2. Main (tools), Where to start & How to do it?** (10-12 min.)

- **Python & Data Science (IDE-s and tools)**
- **Why should people care? (PyData Worldwide & PyData movement/educational programme)**

▯ **3. PyData Bratislava (where we were/are & our activities), Partners, Meetups & Conferences** (2-5 min.)

▯ **4. Conclusion (the end of talk +vision), Potential & Results** (2-5 min.)

- **Future of Data Science (in/outside Python)**
- **Data-Driven Approach. Everywhere. Now. (Citizen Data Science)**

About me

- Economist (Macro, Finance, Statistics)
- Principal Data Scientist, Consulting (public, private)
- R, Python, Julia, Tableau > Matlab, SAS, Stata
- Data Science & Open Data & Public Policy
- PyData Bratislava, R <- Slovakia, SlovakiaJUG, skczTUG football/soccer (FC Economist), Effective Altruism Slovakia

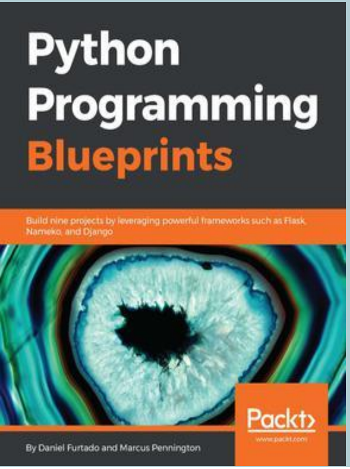
DataCamp

LEARN 2CODE

GAP DATA INSTITUTE

BASECAMP

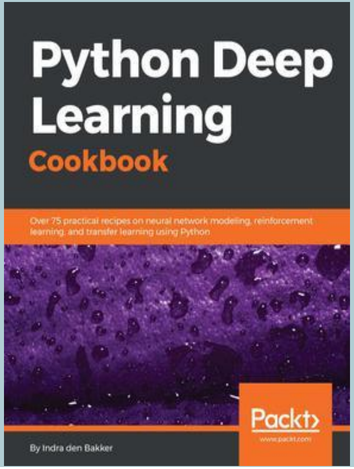
GOPAS


Python Programming Blueprints

Build nine projects by leveraging powerful frameworks such as Flask, Nameko, and Django

Packt>

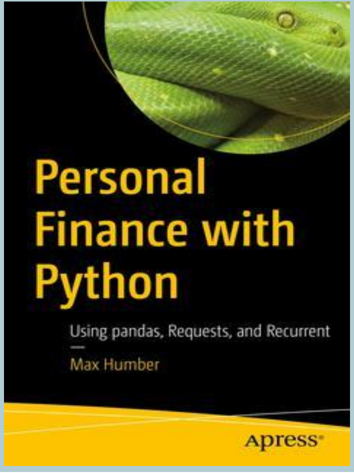
By Daniel Furtado and Marcus Pennington


Python Deep Learning Cookbook

Over 75 practical recipes on neural network modeling, reinforcement learning, and transfer learning using Python

Packt>

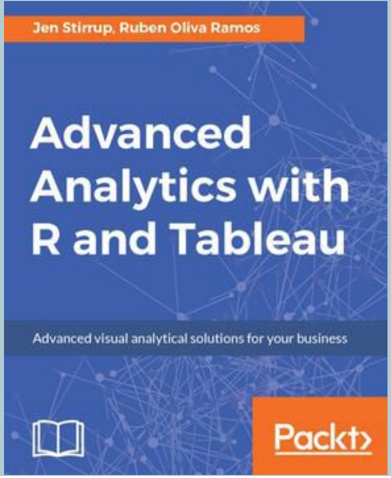
By Indra den Bakker


Personal Finance with Python

Using pandas, Requests, and Recurrent

Max Humber

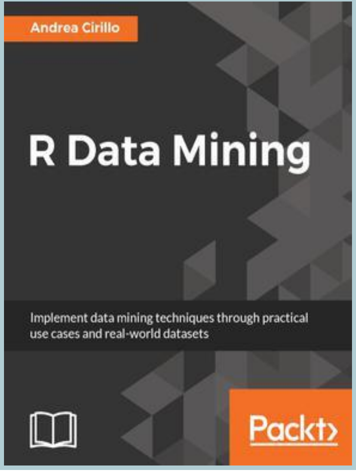
Apress


Jen Stirrup, Ruben Oliva Ramos

Advanced Analytics with R and Tableau

Advanced visual analytical solutions for your business

Packt>


Andrea Cirillo

R Data Mining

Implement data mining techniques through practical use cases and real-world datasets

Packt>

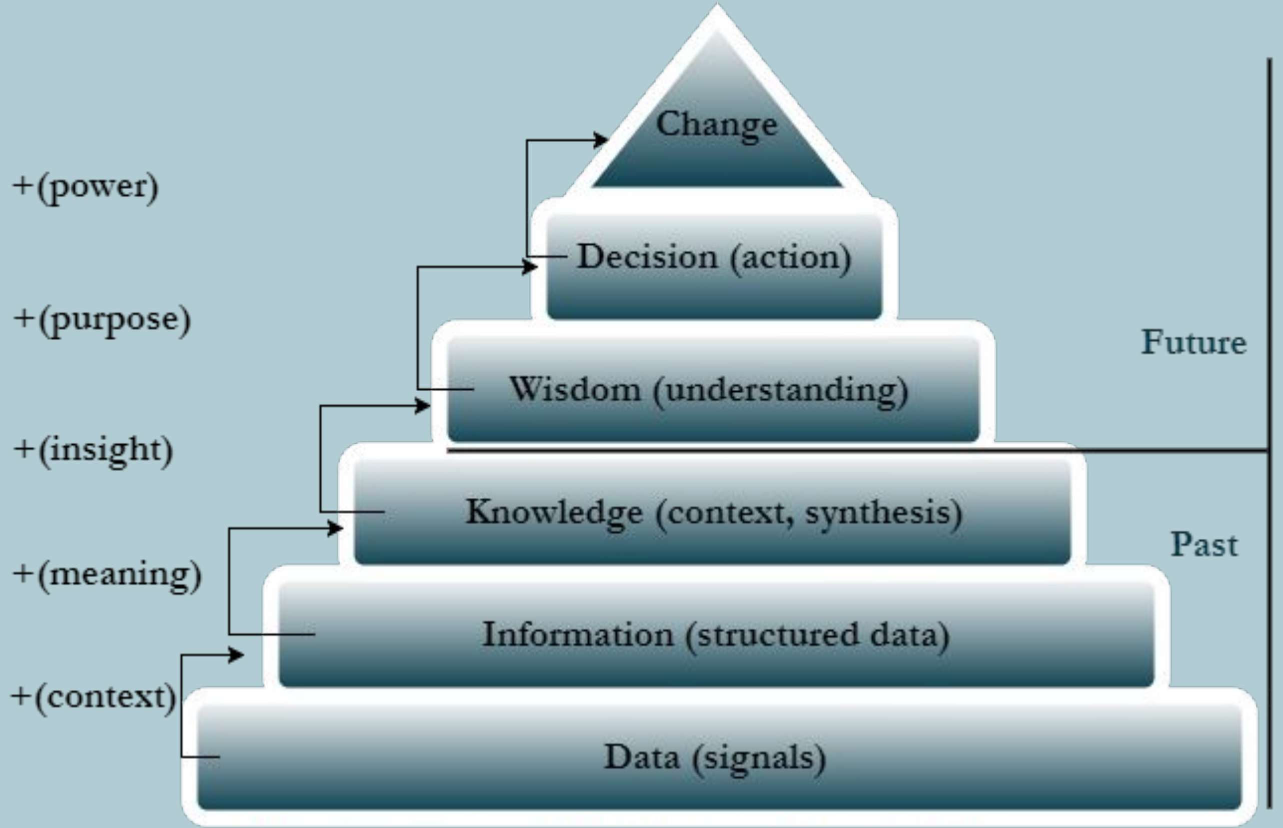
Join at
slido.com
#PyConLT

Python Developer or Python Data Scientist?

How old is Data Science & who “coined” it?

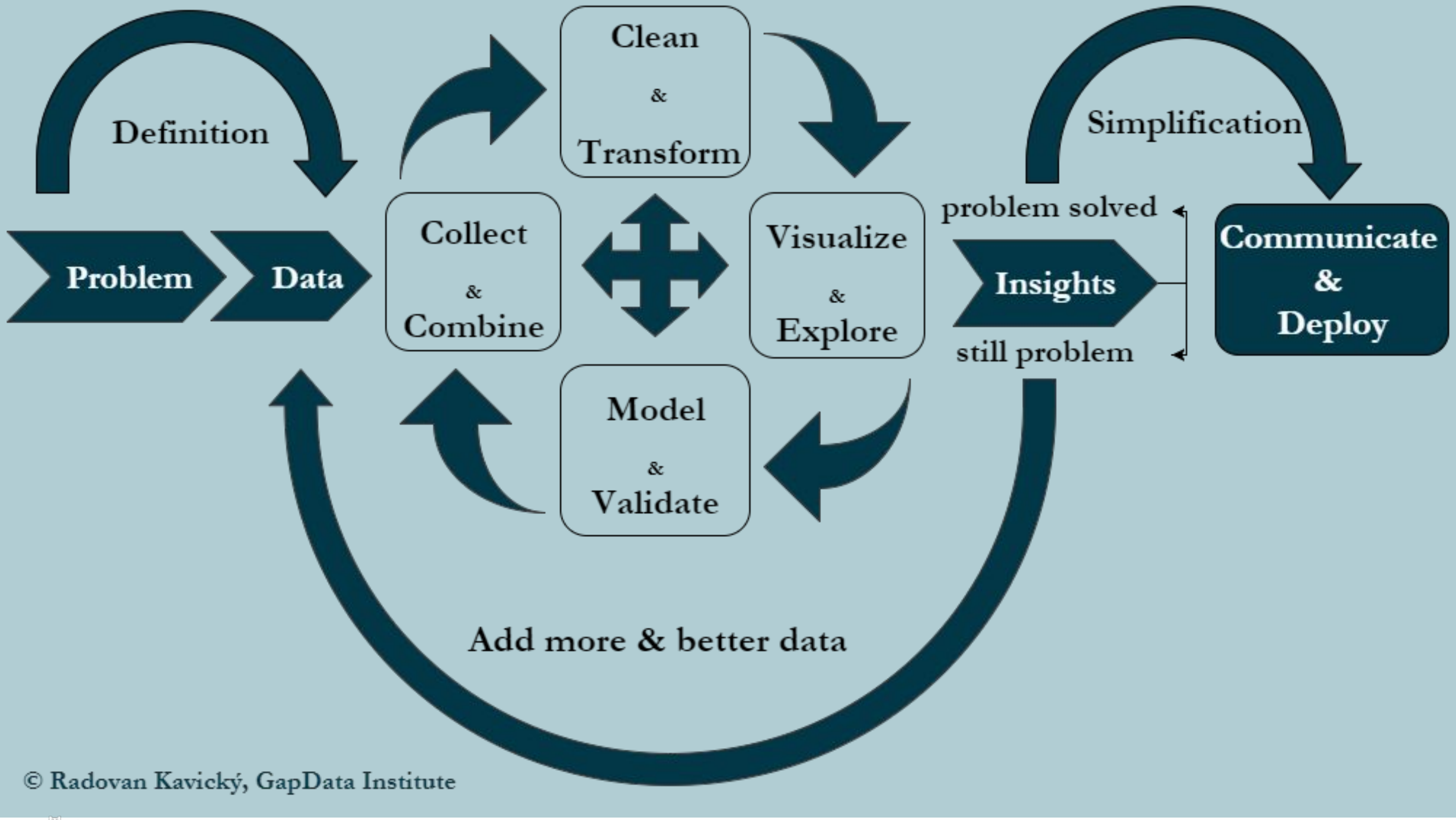
- John Tukey, Bell Labs (+John von Neumann, 50's), DJ Patil (1st US Chief Data Scientist + Jeff Hammerbacher “coined term Data Science”, 2012)
- collection of scientific results and methods for transformation of data from raw form to meaningful information, knowledge and wisdom, which should support better decisions
- Data Science ≠ Big Data
- Data Science: Statistics + programming; data analysis + Computer Science, modelling + Econometrics, Big Data, ML/DL
- More on History of Data Science: Donoho, David (2015, Stanford), https://courses.csail.mit.edu/18.337/2015/docs/50YearsDataScience.pdf

Data Science Pyramid



© Radovan Kavický, GapData Institute (inspired by DIKWD model by Curt Swindoll)

Data Science Process & Workflow



© Radovan Kavický, GapData Institute

Where & When did Python

- 80’s (idea), December 1989 (beginning), Guido van Rossum (Netherlands/Python is Dutch, CWI Research Institute)
- 1994 (1.0), 2000 (2.0), 2008 (3.0/3000/Py3K), 2020 (4.0 +end 2.0)
- hobby project during holidays
- Monty Python Flying Circus
- ABC language (cancelled +SETL), Amoeba (OS) needed script language
- First PC (ideal), universal a multiplatform+ easy to use, hard to master
- Not for data analysis and Data Science (s-l-o-w)
- IPython/Jupyter Notebook (Fernando Perez) –JUlia, PYthon, teR

R & Data Science

S-PLUS® 6

- Programming language S - John Chambers, Bell Labs (USA)

+ Rick Becker, Allan Wilks

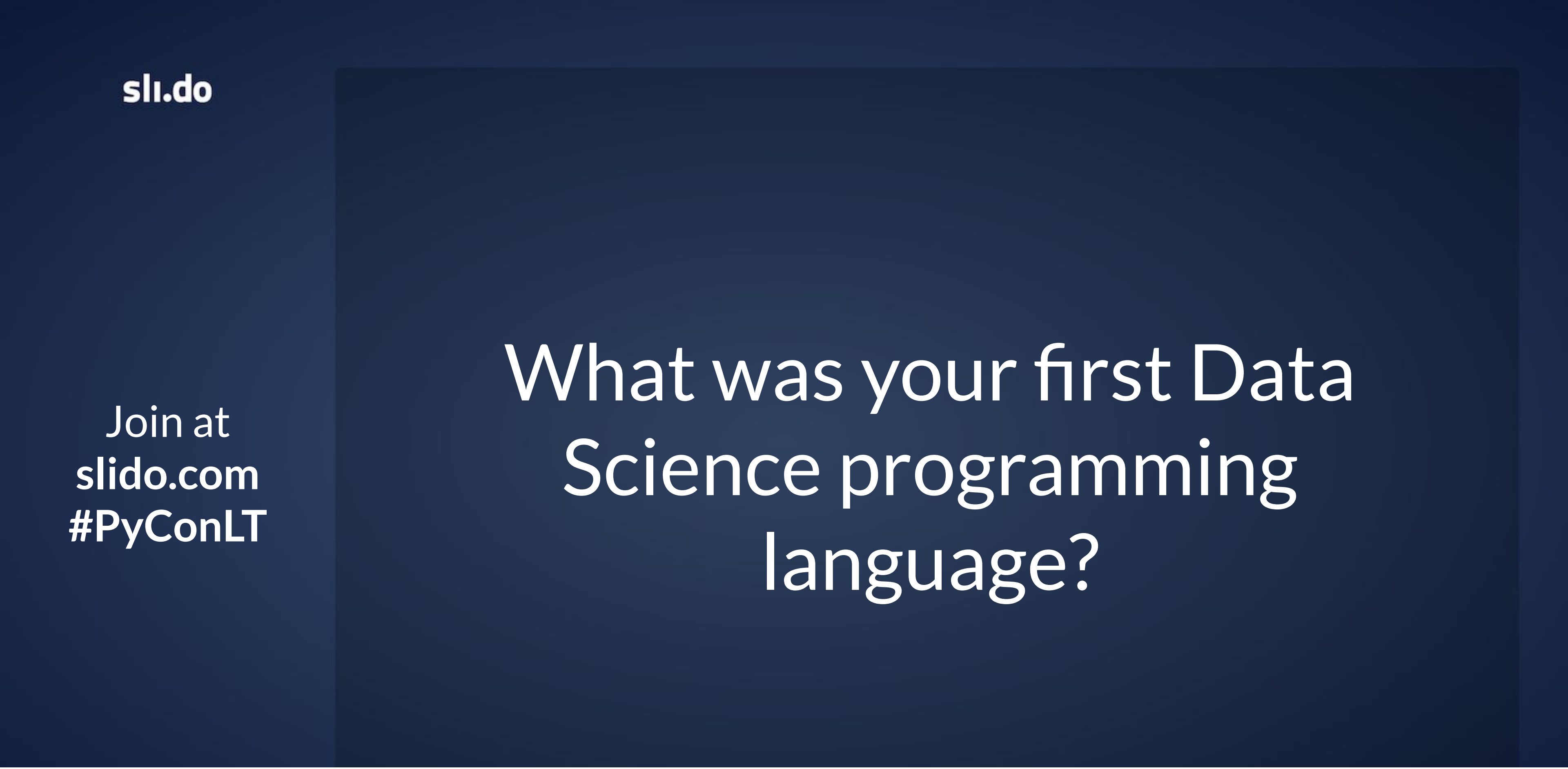
1976, Fortran subroutines, later S-Plus (commercial version)

- R - Robert Gentleman a Ross Ihaka (New Zealand, University of Auckland)

August 1993, C + Fortran

- R as implementation of S (didn't want to pay license)

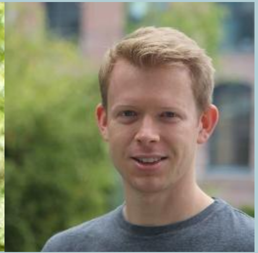
THE UNIVERSITY OF AUCKLAND NEW ZEALAND



Join at
slido.com
#PyConLT

What was your first Data Science programming language?

Python & Data Science



- Travis Oliphant (NumPy + SciPy/Matlab alternative)

1995 (Numeric), 2006 (Numpy)

- John Hunter (matplotlib)

2003

- Wes McKinney (pandas)

January 2008, Python, C, Cython

- Anaconda as one of the Python distributions (1000 selected libraries for Data Science)
- NumFocus
- PyData conferences and local meetups (PyData Bratislava)

Python & Data Science (tools)

IDE's – Jupyter Notebooks (IPython kernel), PyCharm, JupyterLab

Data Collection – Feather (binary file format/non-csv/Apache Arrow/Wes McKinney), pandas datareader

Data Visualization – Seaborn (matplotlib based/static), Bokeh (interactive/d3.js like), Plotly (declarative dataviz), Altair (static/js)

Data Cleaning & Transform - datacleaner (automate cleaning your data in Pandas), Blaze (NumPy/pandas-like), Dask (parallel computing)

Data Modeling – StatsModels +Patsy (describe statistical models), PyMC3 (Bayes/statistical modeling), Keras (TensorFlow & PyTorch/DL)

R vs/& Python “war” in Data Science

R - Studio (IDE for R & Python)

Jupyter Notebook (Irkernel) + r-essentials (conda)

Anaconda (open-source distribution of Python&R)

Tidyverse (tidyverse.tidyverse.org, Magrittr, %>%)

dplyr, tidyr (subsetting, joining, data manipulation)

ggplot2, rBokeh, ggvis (data visualization)

caret – modelling (regression, classification)

H2O.ai, Tensorflow & Keras – ML & DL

Shiny, blogdown, R Markdown (publication, RMd, blog)

R Studio®

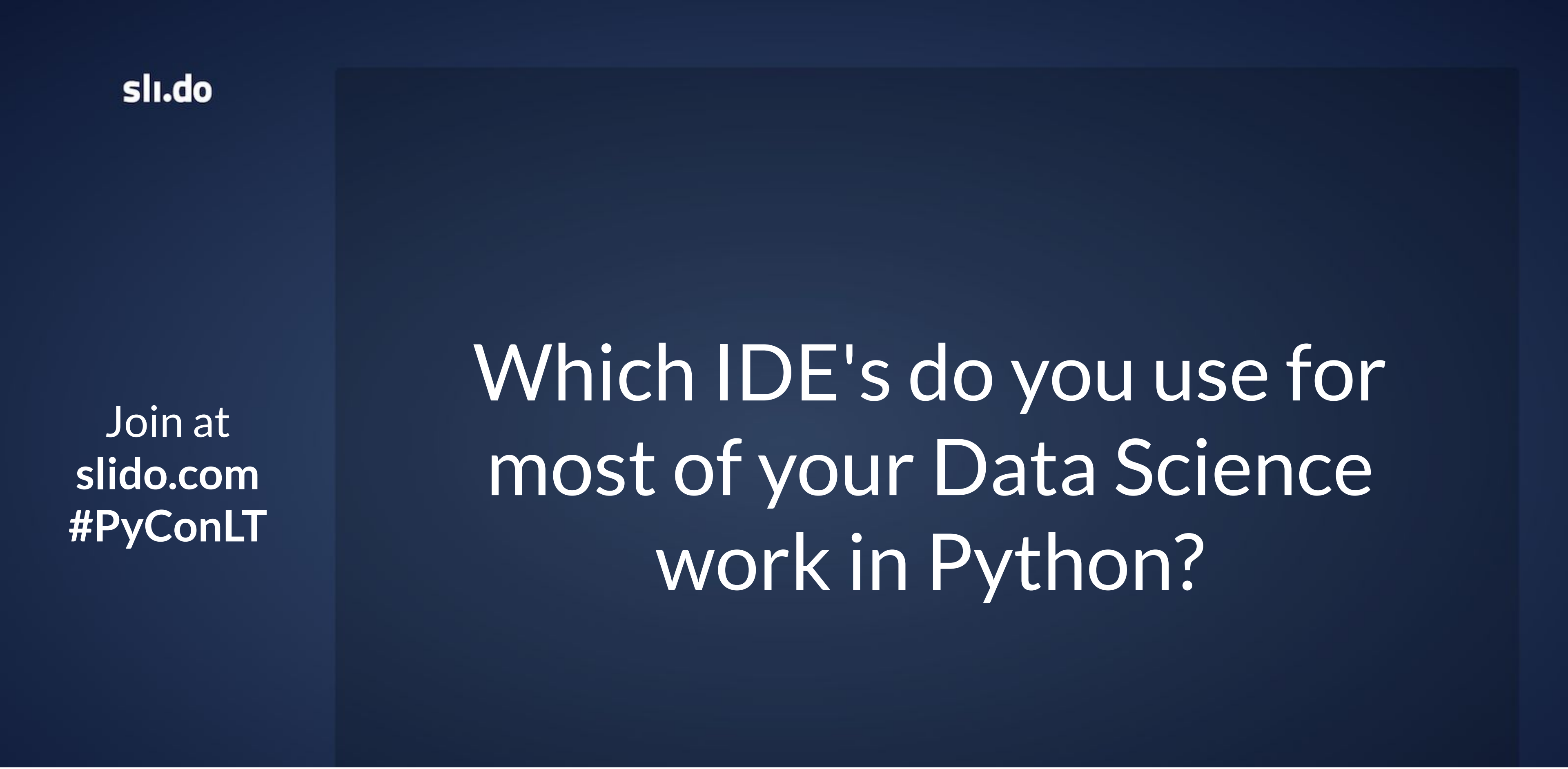
tidyverse

TensorFlow

K Keras

dplyr

data.table



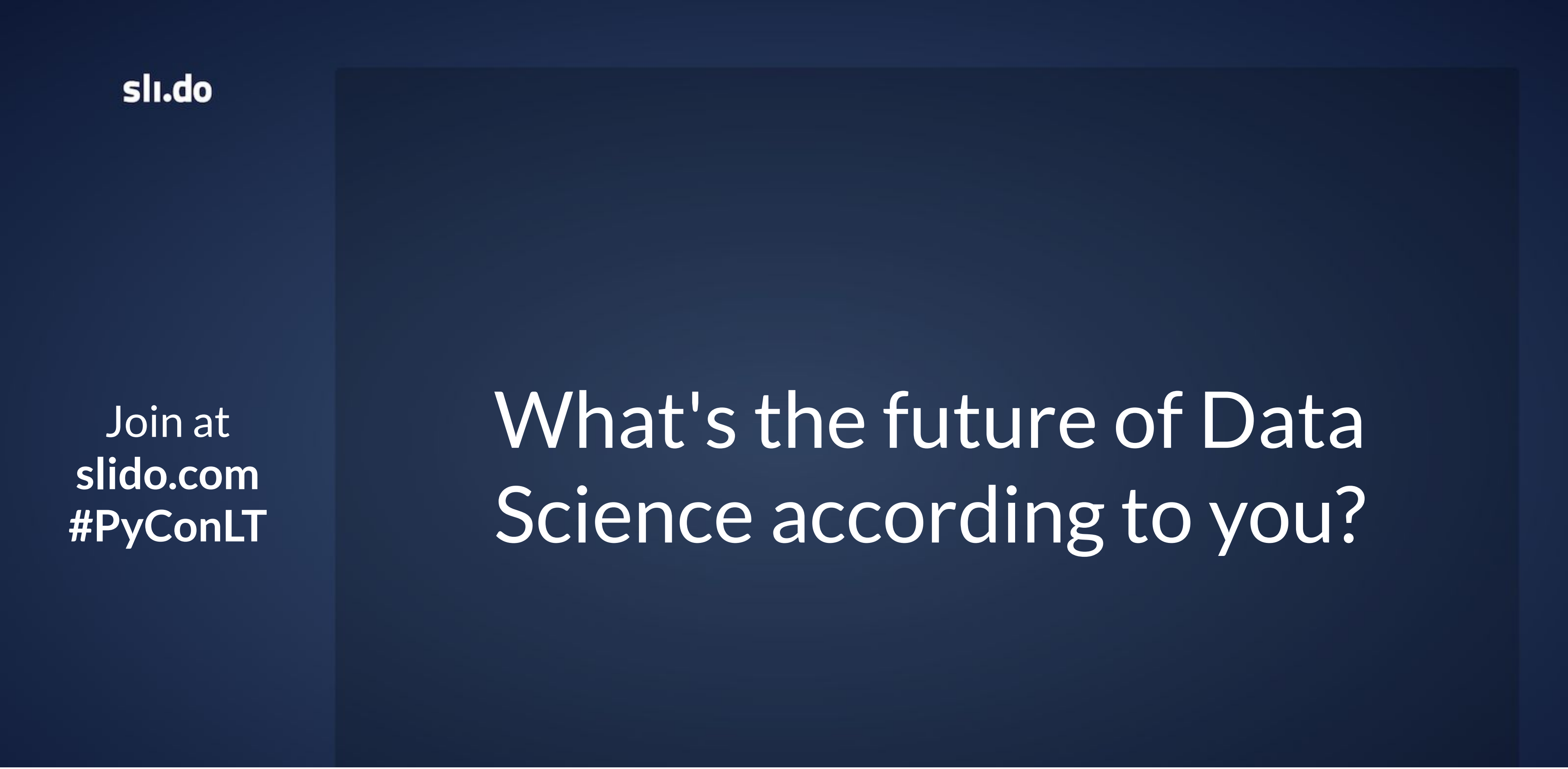
sli.do

Join at
slido.com
#PyConLT

Which IDE's do you use for most of your Data Science work in Python?

Future of Data Science (Pythonic or Future is Female/Julia?)

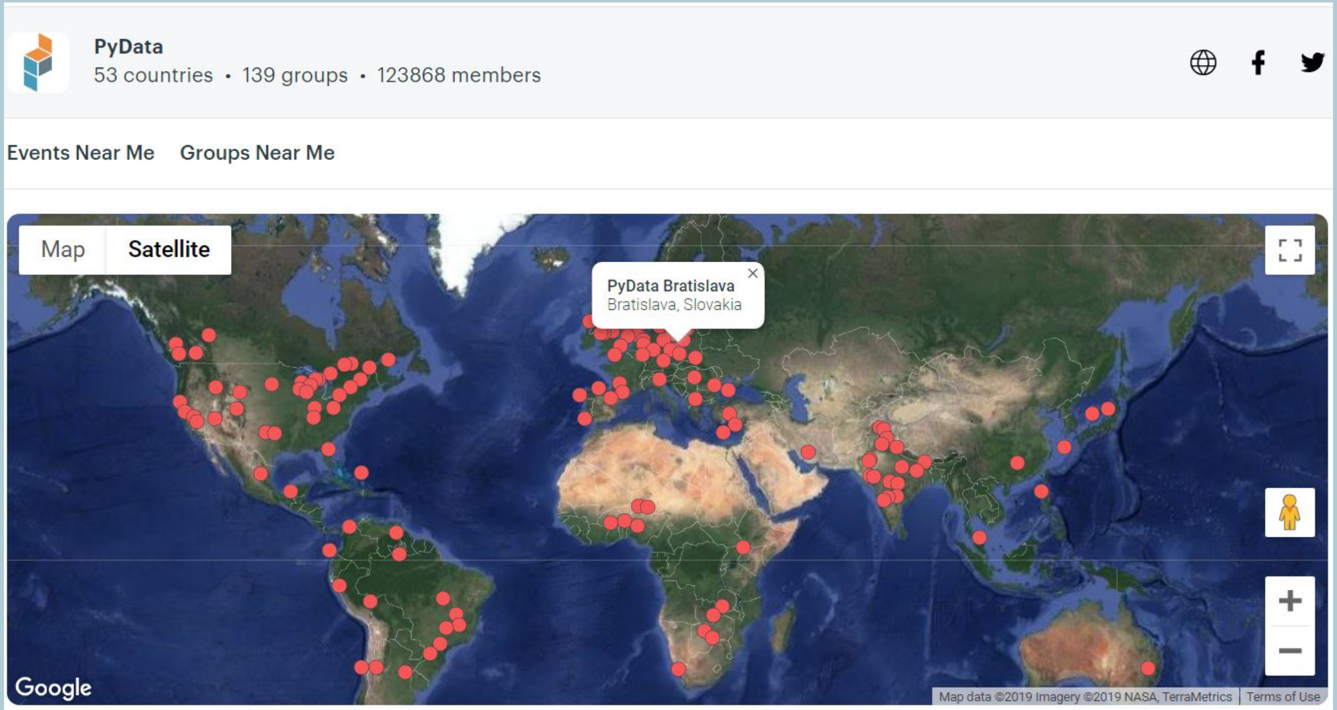
- Fast as C, syntax/easy to read as Python, dynamic as Ruby, mathematic notation as Matlab, statistics as in R, string processing as Perl, true macros like Lisp
- Creators (2009, 2012, MIT): Viral B. Shah, Stefan Karpinski, Jeff Bezanson, prof. Alan Edelman
- Julia Computing (developers 1.0 – summer 2018)
- Future of Data Science & designed with Data Science in mind
- Python not for analysis and Data Science (s-l-o-w)
- Julia is (f-a-s-t/ LLVM compiler)
- Solves "two-language problem" - prototype (slow, dynamic), delivery (quick, static)
- IPython/Jupyter (Fernando Perez) –JUlia, PYthon, teR



Join at
slido.com
#PyConLT

What's the future of Data Science according to you?

PyData as global movement

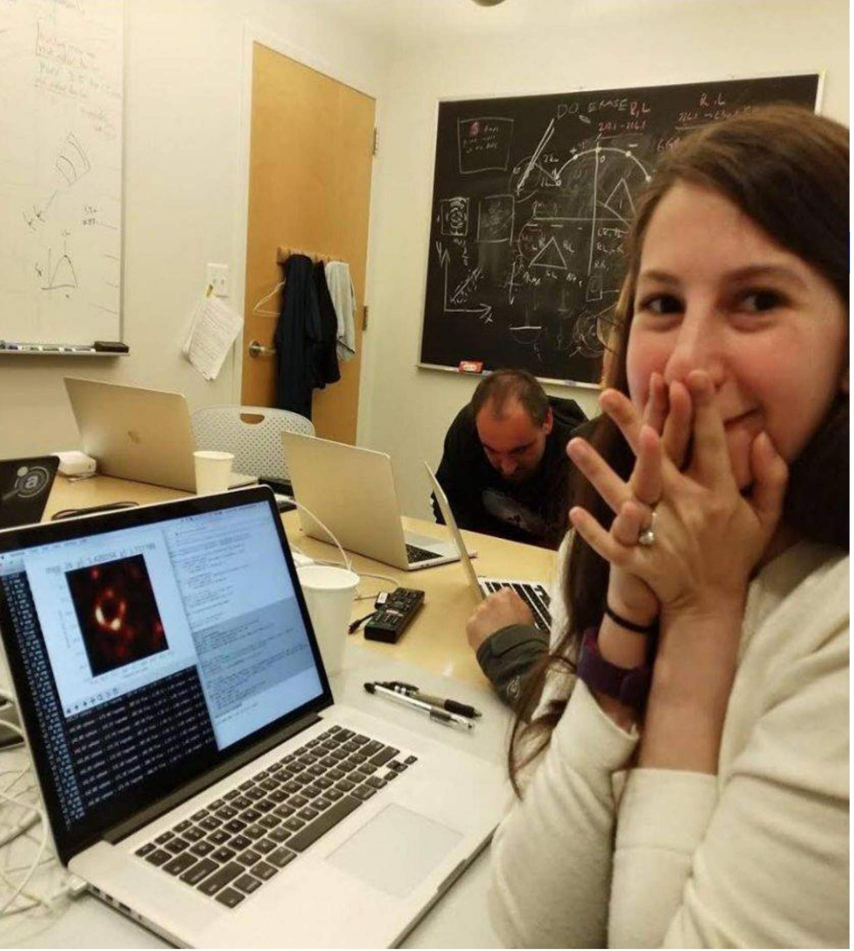


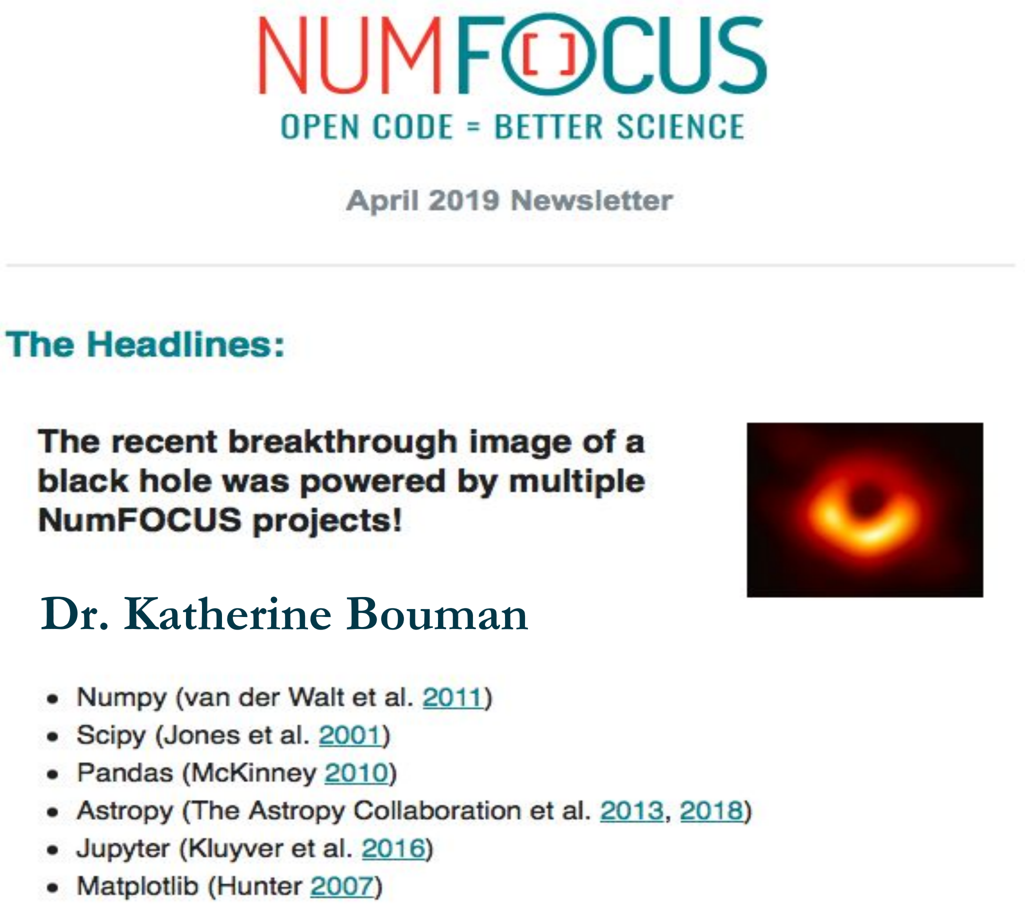
PyData
53 countries • 139 groups • 123868 members

Events Near Me Groups Near Me

Source: https://www.meetup.com/pro/pydata/

- More Meetups outside US than within
- Slovakia (40th globally):
- Few days after us LA (as 41st & 11th in US)
- Before us: Poland (Warsaw, Wroclaw, Krakov), Austria (Vienna, few months)
- After us: Czechia (Prague, 2 years), Hungary (Budapest, 1 year)





NUMFOCUS

OPEN CODE = BETTER SCIENCE

April 2019 Newsletter

The Headlines:

The recent breakthrough image of a black hole was powered by multiple NumFOCUS projects!

Dr. Katherine Bouman

- Numpy (van der Walt et al. 2011)
- Scipy (Jones et al. 2001)
- Pandas (McKinney 2010)
- Astropy (The Astropy Collaboration et al. 2013, 2018)
- Jupyter (Kluyver et al. 2016)
- Matplotlib (Hunter 2007)

PyData Bratislava & activities within CEE + V4

WeAreDevelopers

O'REILLY OSCON

Human Level AI

AI Startup Showcase

Machine Learning Prague

techsummit.sk bratislava

WEBEXPO

- 300+ members
- 600+ on mailing list
- 800+ followers
- Partners: O'Reilly, robime.it, Learn2Code, sli.do
- Sponsors: H2O.ai, SwissRe, ERNI, Kiwi Security
- WeAreDevelopers World Congress (“GAPDATA-25” & “PYDATA-25”), Berlin, 6.-7. June
- OSCON 2019 (“PCGAPDATA” for 25% discount)
- TechSummit, Bratislava, 29.-30. May (“pydatabratislava20” for 20% discount)
- Our Conferences (satRday, PyData Bratislava)

#AllForJan Hackathon (cooperation with FIIT STU)

Source: https://github.com/AllForJan

- Data journalism
- Public Policy, Activism & Data
- Useful tools for journalist
- Agricultural data (dotations)
- Vice-dean FIIT Pištek & GapData Institute
- LOD (linked open data), transparency check
- All outputs public (GitHub)
- Over 50 developers & data scientists

GapData Institute (GDI) and how to support us.

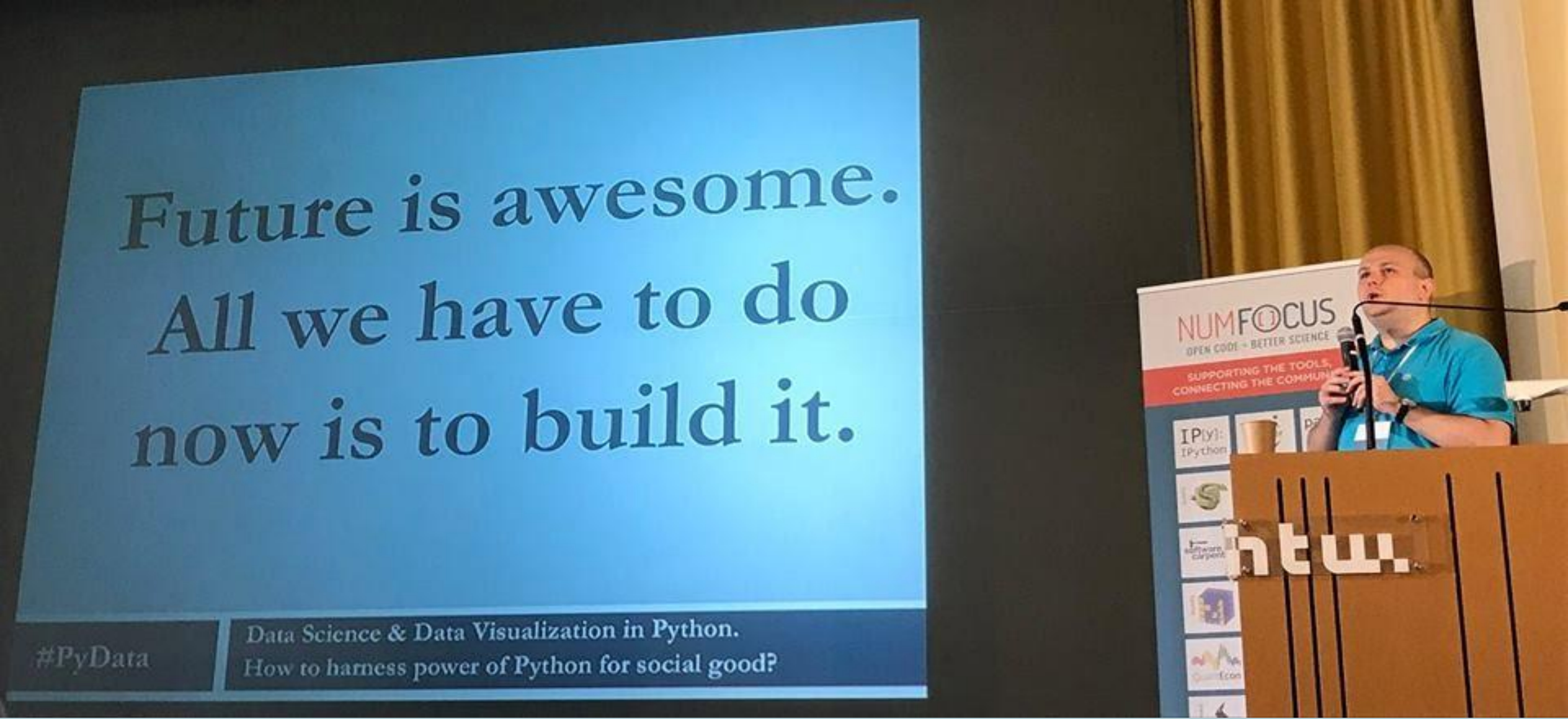
GAP DATA INSTITUTE

- Economic Research & Public Policy & Data Science think-tank (data-tank)
- Data. Think. Change.
- GapData Institute (GDI) is a non-profit nonpartisan research institution harnessing power of data & wisdom of economics for public good.
- Transparent account (from day #1; SK7383300000002200933920 https://www.fio.sk/ib2/transparent?a=2200933920)
- Partnership (openness, transparency)
- Slides (this talk): tiny.cc/pyconLT2019vilnius
- https://github.com/radovankavicky/PyConLT2019

NUMFOCUS
OPEN CODE = BETTER SCIENCE

PyData
Bratislava

PAY by square



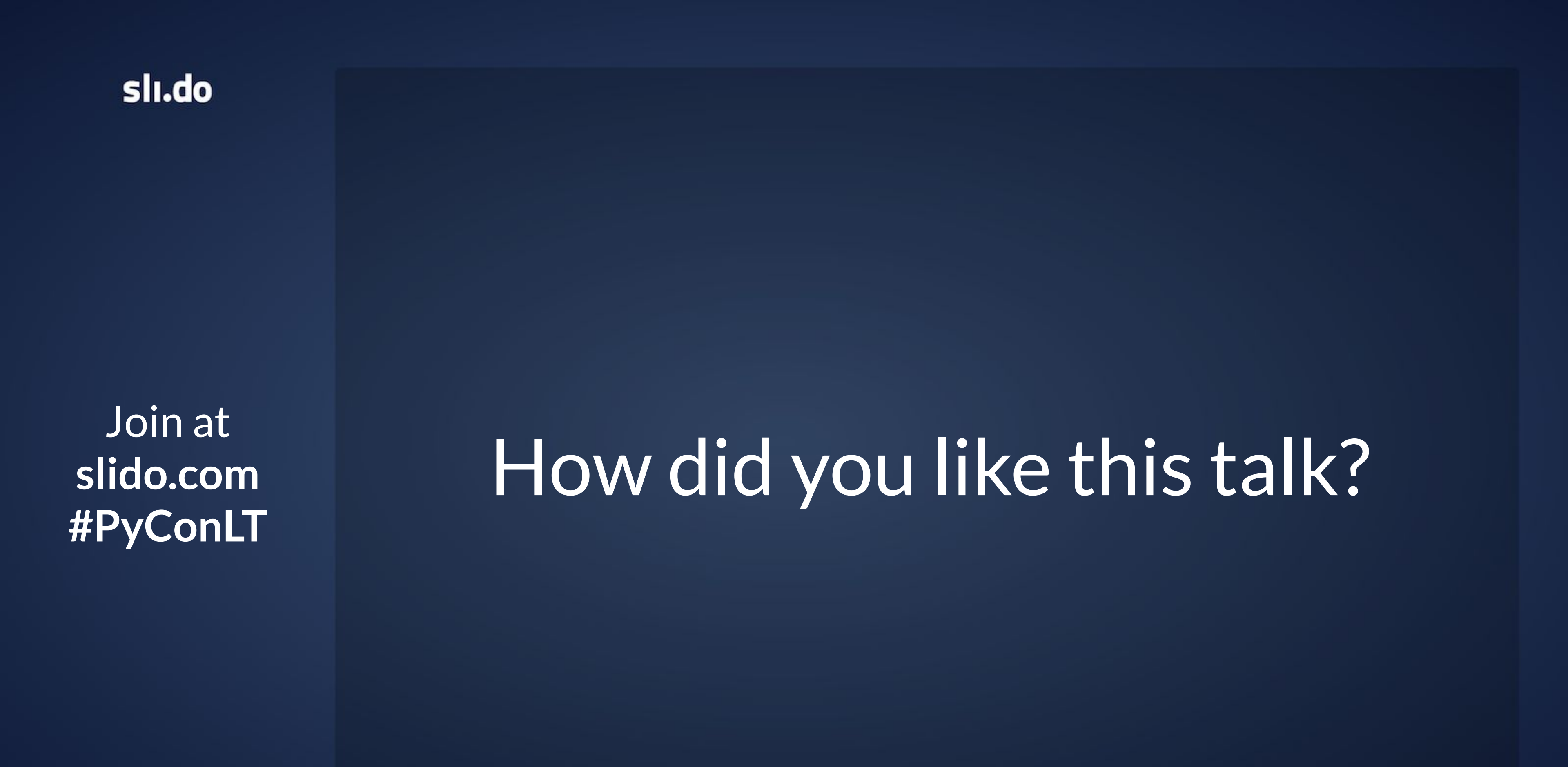
Future is awesome. All we have to do now is to build it.

#PyData

Data Science & Data Visualization in Python.
How to harness power of Python for social good?

NUMFOCUS
OPEN CODE = BETTER SCIENCE
SUPPORTING THE TOOLS, CONNECTING THE COMMUN
IP[y]: IPython
htw.

Source: PyData Berlin 2017, Talk on YouTube:
https://www.youtube.com/watch?v=I5578BhU4sE



sli.do

Join at
slido.com
#PyConLT

How did you like this talk?

Thank you for your attention

Contact:

Radovan Kavicky

radovan.kavicky@gapdata.org

radovan.kavicky@gmail.com

+421 949 716 214 (SK)

http://www.linkedin.com/in/radovankavicky

https://gapdata.slack.com/

Invitations (e-mail): https://gapdata.herokuapp.com/

https://github.com/radovankavicky

https://github.com/GapData/PyDataBratislava

@radovankavicky, @PyDataBA, @GapDataInst

PyData Bratislava

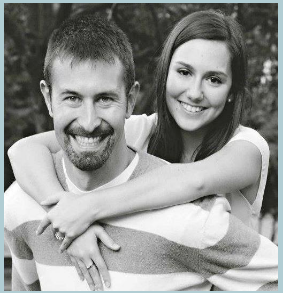
GAP DATA INSTITUTE

#PyConLT

In case you have any question, feel free to ask.

SatRday Bratislava 2019 in September & PyData Bratislava 2019 in December?

- Saturday, 28th of September 2019, Bratislava Slovakia
- Chester Ismay (DataCamp, Curriculum Lead/Head of Content Development, Oregon/US)
- Szilard Pafka (EP0CH, Chief Data Scientist, Founder: Data Science.LA)
- R Consortium, NumFocus
- PyData conferences and local meetups (PyData Bratislava)
- PyData Bratislava conference/track (December 2019 maybe?)



#PyDataBA & #RSlovakia

Data Science in Python: Past, Present & Future