BRATISLAVA

OpenCamp

GAP
DATA
INSTITUTE

Open Data Science: Python, R & Julia

(OpenCamp 2019, FIIT STU; #OpenCampBA #OpenCamp)

13. 4. 2019

Radovan Kavický, GapData Institute

Talk Structure

□ **1. Introduction & Why Open Data Science?/Motivation** (5-7 min.)

- What is Data Science, History & Data Science Workflow

□ **2. Main (tools), Where to start & How to do it?** (17-20 min.)

- R & Data Science (IDE-s and tools)

- Python & Data Science (IDE-s and tools)

- Julia & Data Science (IDE-s and tools)

- Why should people care? (PyData worldwide)

□ **3. PyData Bratislava (where we are & our activities), Partners, Meetups & Conferences** (5-7 min.)

□ **4. Conslusion (the end of talk +vision), Potential & Results** (7-10 min.)

- **How to "unlock" the insights hidden in data and how to use it to transform not only public administration or business, but whole society and economy towards the insight & knowledge based?**

- Open Data, Open Government Partnership, Open Public Administration & Open Data Science

- 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, skczTUG

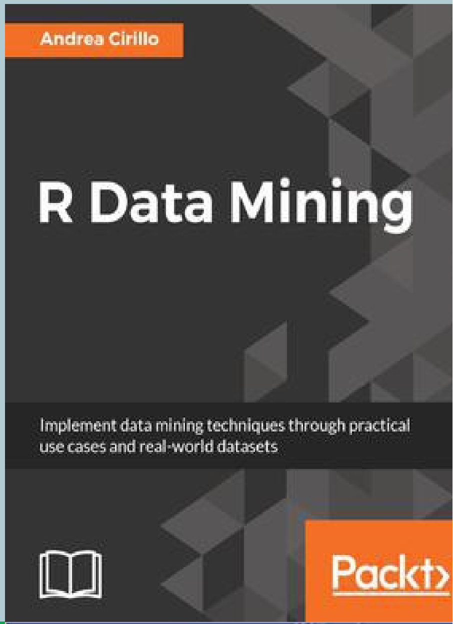
DataCamp

LEARN 2CODE

BASECAMP

GOPAS

GAP DATA INSTITUTE

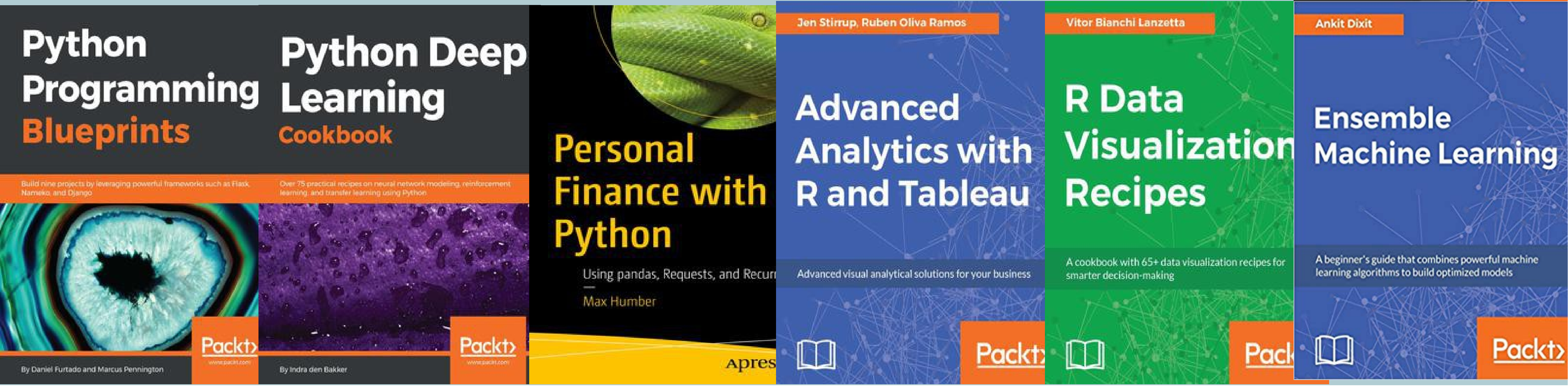


Andrea Cirillo

R Data Mining

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

Packt>



Python Programming Blueprints

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

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

By Indra den Bakker

Personal Finance with Python

Using pandas, Requests, and Recurr

Max Humber

Apres

Jen Stirrup, Ruben Oliva Ramos

Advanced Analytics with R and Tableau

Advanced visual analytical solutions for your business

Vitor Bianchi Lanzetta

R Data Visualization Recipes

A cookbook with 65+ data visualization recipes for smarter decision-making

Ankit Dixit

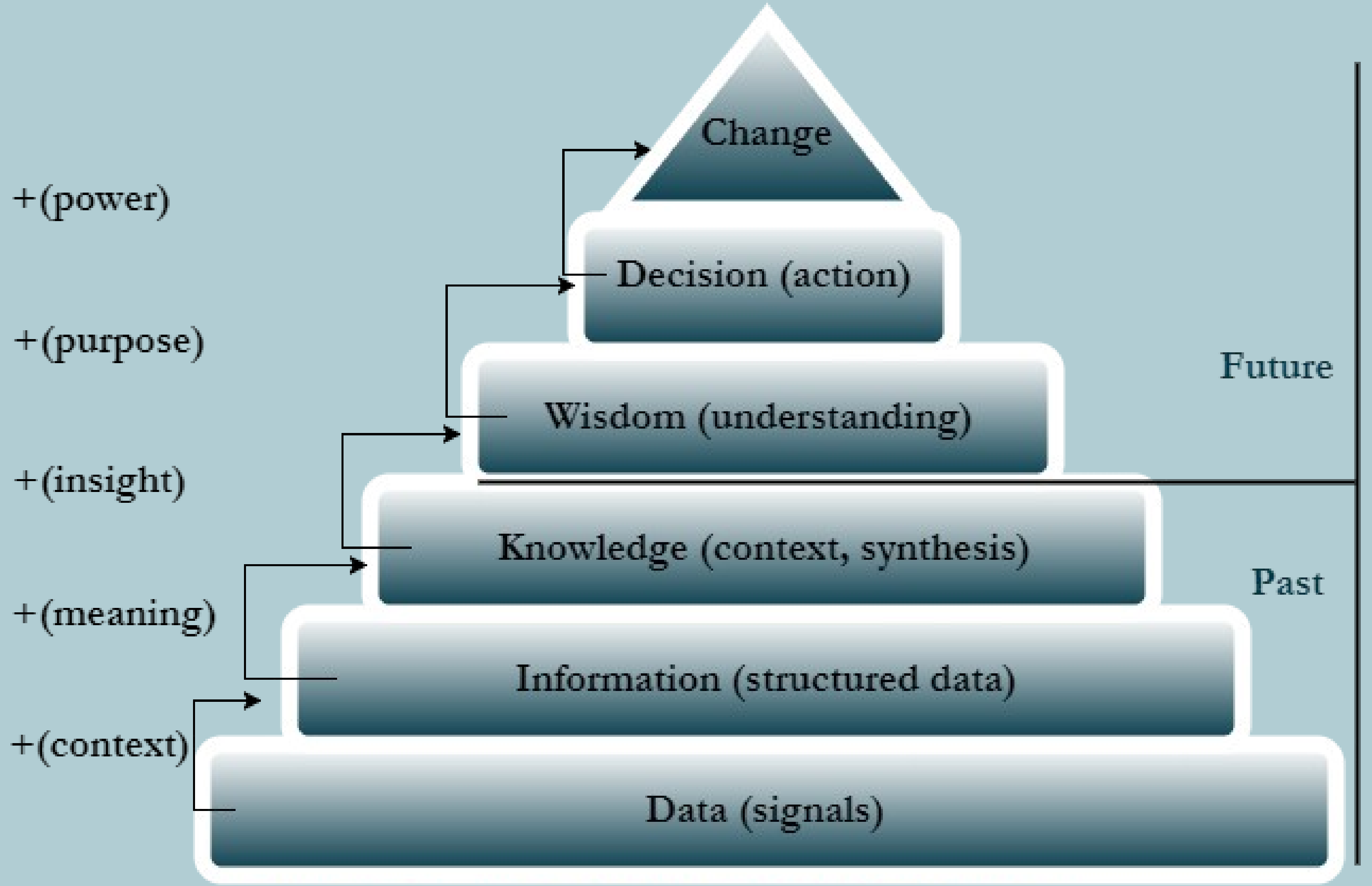
Ensemble Machine Learning

A beginner's guide that combines powerful machine learning algorithms to build optimized models

What is (Open) Data Science?

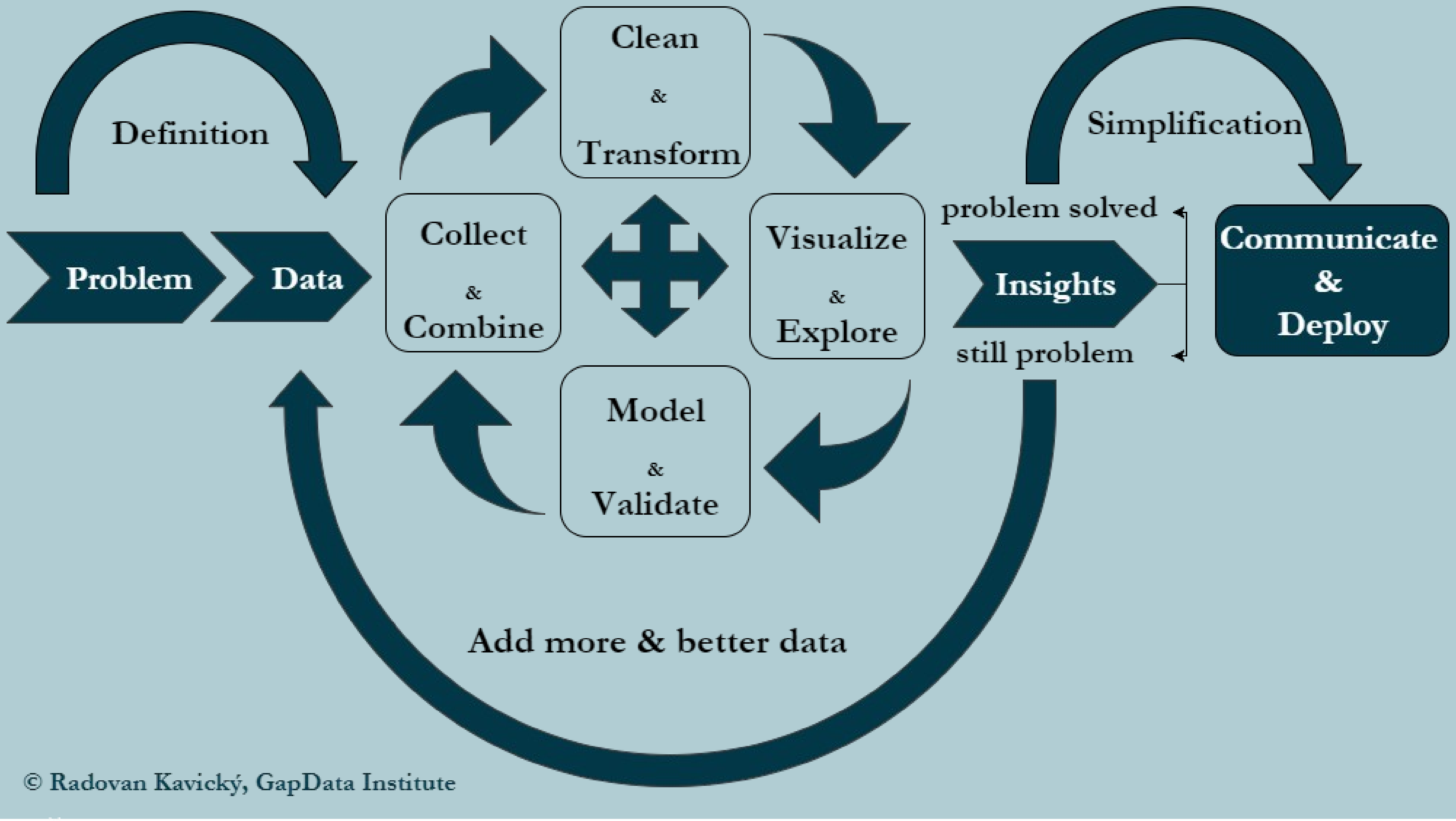
- John Tukey, Bell Labs (+John von Neumann, 50's), DJ Patil (1st US Chief Data Scientist + Jeff Hammerbacher “coined term Data Science”, 2012)
- collection of scientific results and methods for transformation of data from raw form to meaningful information, knowledge and wisdom, which should support better decisions
- Data Science ≠ Big Data
- Data Science: Statistics + programming; data analysis + Computer Science, modelling + Econometrics, Big Data, ML/DL
- Anaconda (open-source distribution of Python&R)

Data Science Pyramid



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

Data Science Process & Workflow



Definition

Problem

Data

Collect & Combine

Clean & Transform

Visualize & Explore

Model & Validate

Insights

problem solved

still problem

Simplification

Communicate & Deploy

Add more & better data

© Radovan Kavický, GapData Institute

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

R & Data Science (tools)

R - Studio (IDE for R)

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

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

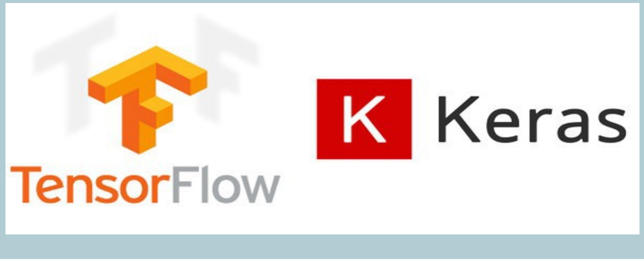
dplyr, tidyr (subsetting, joining, data manipulation)

ggplot2, rBokeh, ggvis (data visualization)

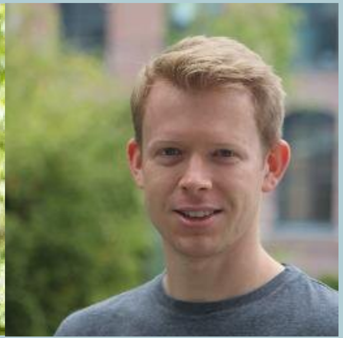
caret – modelling (regression, classification)

H2O.ai, Tensorflow & Keras – ML & DL

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



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), PyStan (Bayes/C++), PyMC3 (Bayes/statistical modeling), Keras (TensorFlow/DL)

pandas

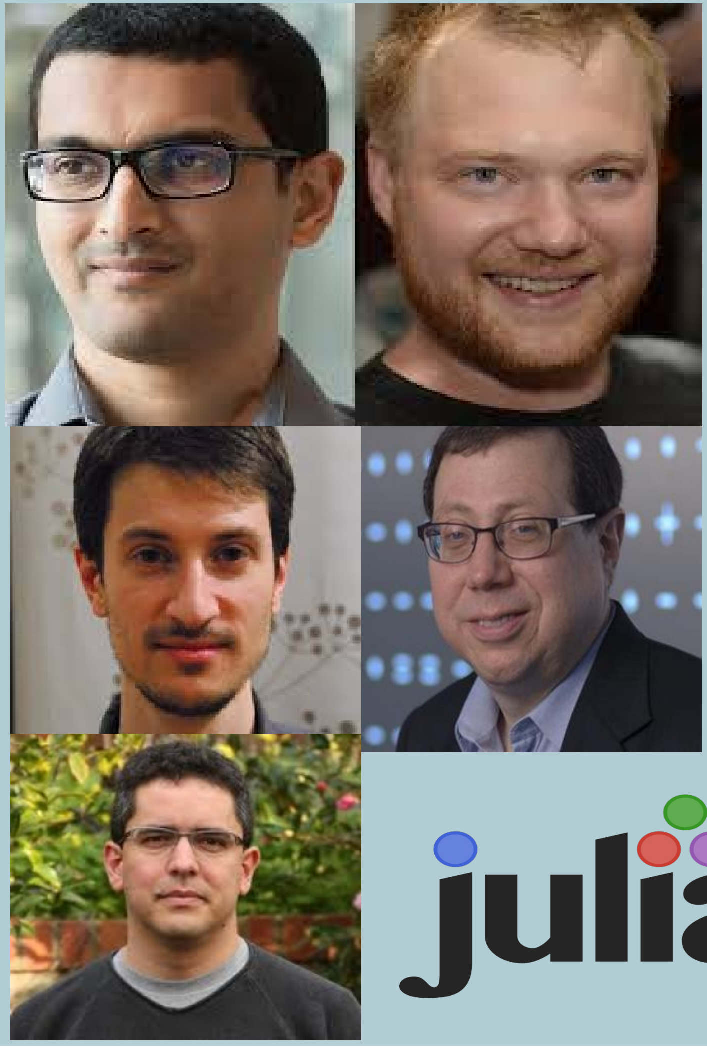
$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$

SM StatsModels

Statistics in Python

PyMC3

Julia & Data Science



- 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

Julia & Data Science (tools)

IDE's for Julia – Jupyter Notebooks (IJulia kernel), Juno (enhanced Atom), JuliaBox (online Jupyter)

Data Collection/Import – Feather (read/write binary file format/non-csv/Apache Arrow/Wes McKinney), PyCall (call any Python function/library, also for imports), Rcall.jl (call any R function/library)

Data Visualization – Gadfly.jl (static, gg), Bokeh.jl (interactive/d3.js like), Plotly.jl (declarative dataviz), PyPlot.jl, Plots.jl

Data Cleaning & Transform – Pandas.jl, datacleaner/Missings.jl (automate cleaning and handling missing values)

Data Modeling – StatsModels (DataFrames.jl) +StatsBase.jl, MLBase.jl, ScikitLearn.jl (ML), Keras (TensorFlow/DL), Mocha.jl (DL)

PyData Bratislava in 2019

WeAreDevelopers

Human Level AI

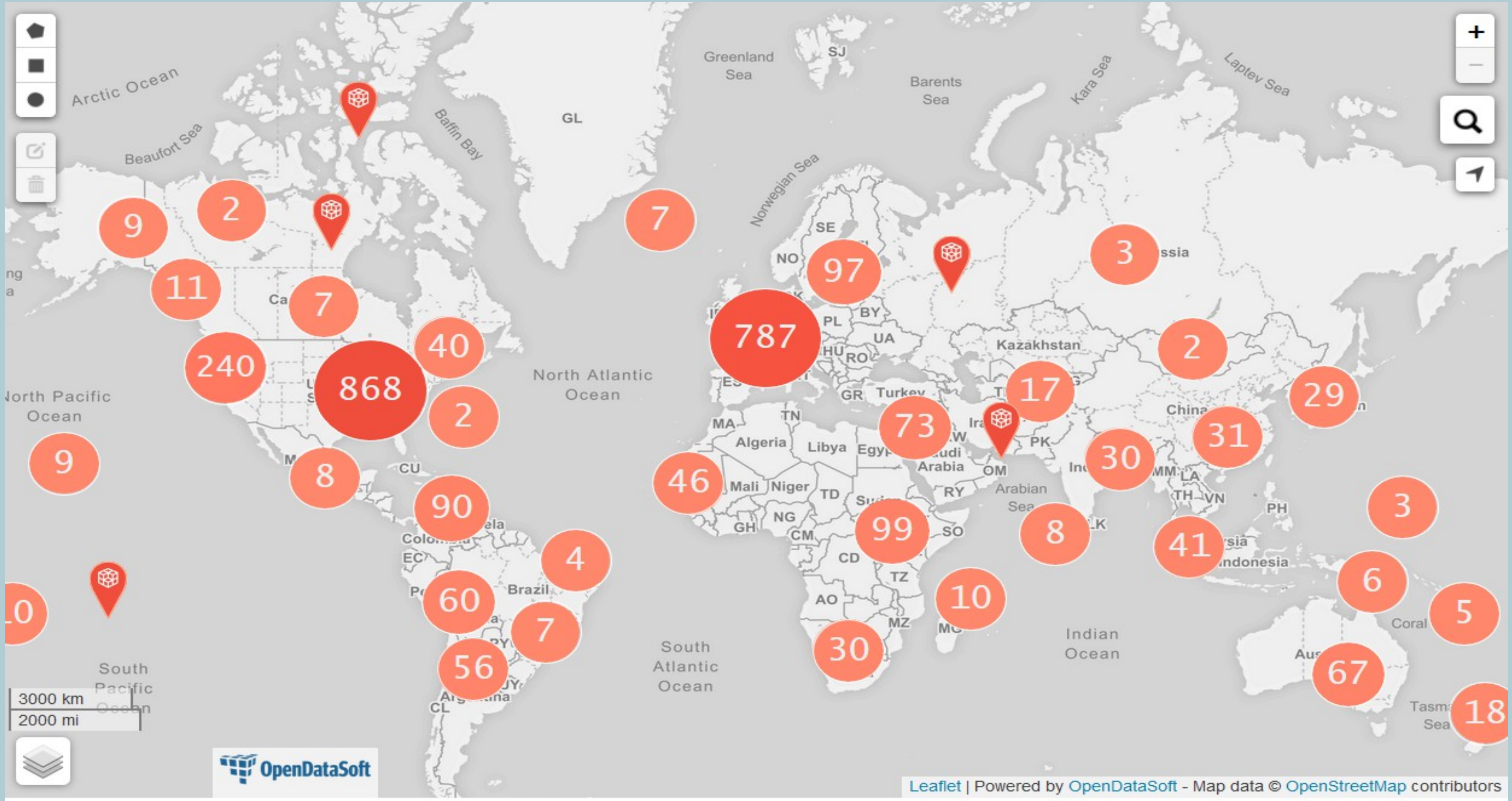
AI Startup Showcase

Machine Learning Prague

WEBEXPO

- 200+ members
- 500+ on mailing list
- 800+ followers
- Partners: robime.it, Nervosa
- Sponsors: H2O.ai, Kiwi Security, SwissRe, ERNI
- Conferences (satRday, PyData Bratislava)
- And more to come. Stay tuned.

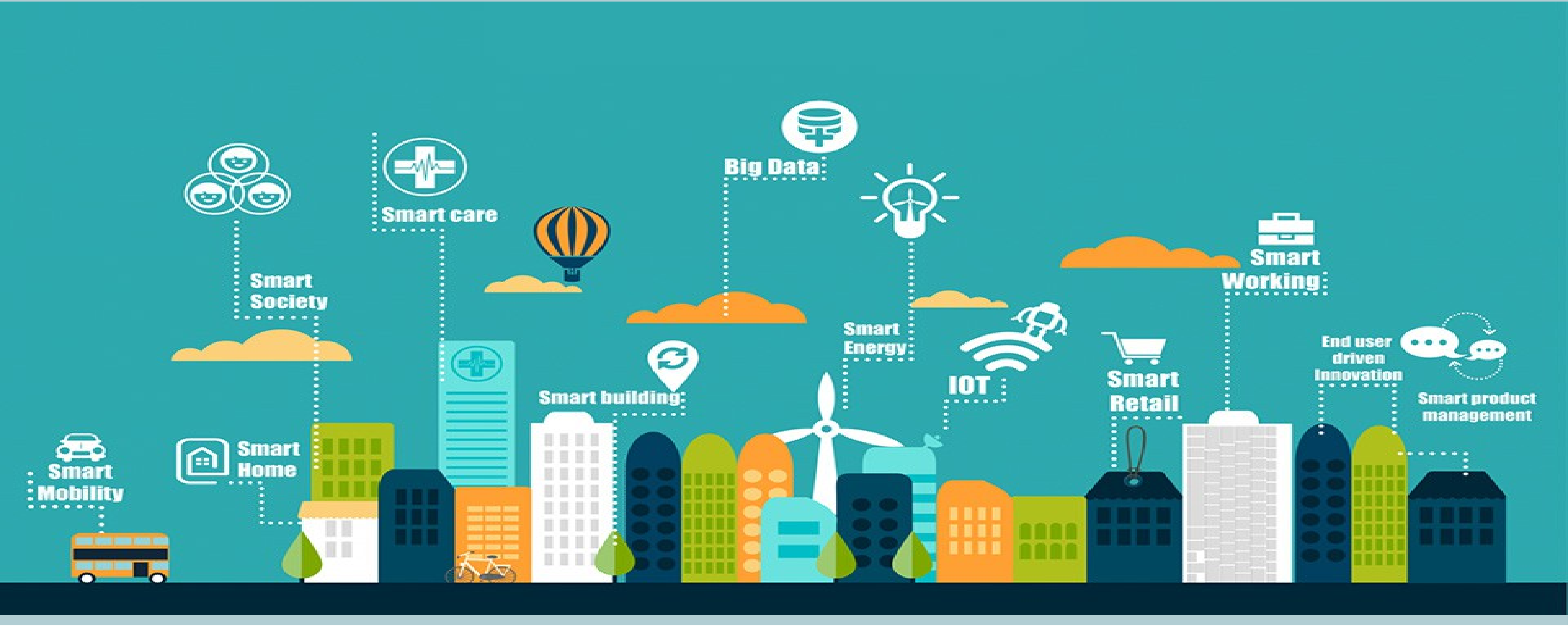
How many “Open Data” portals are there?



- Europe (787 portals)
- Slovakia (6):
- Datanest AFP
- Map Client Zbgis
- Slovak Hydrometeorological Institute
- Slovak Republic National Data Portal
- State Geological Survey of Slovakia
- Statistical Office SK

Source: https://www.opendatasoft.com/a-comprehensive-list-of-all-open-data-portals-around-the-world/

Why smart cities… it’s necessary.



#AllForJan Hackathon (cooperation with FIIT STU)

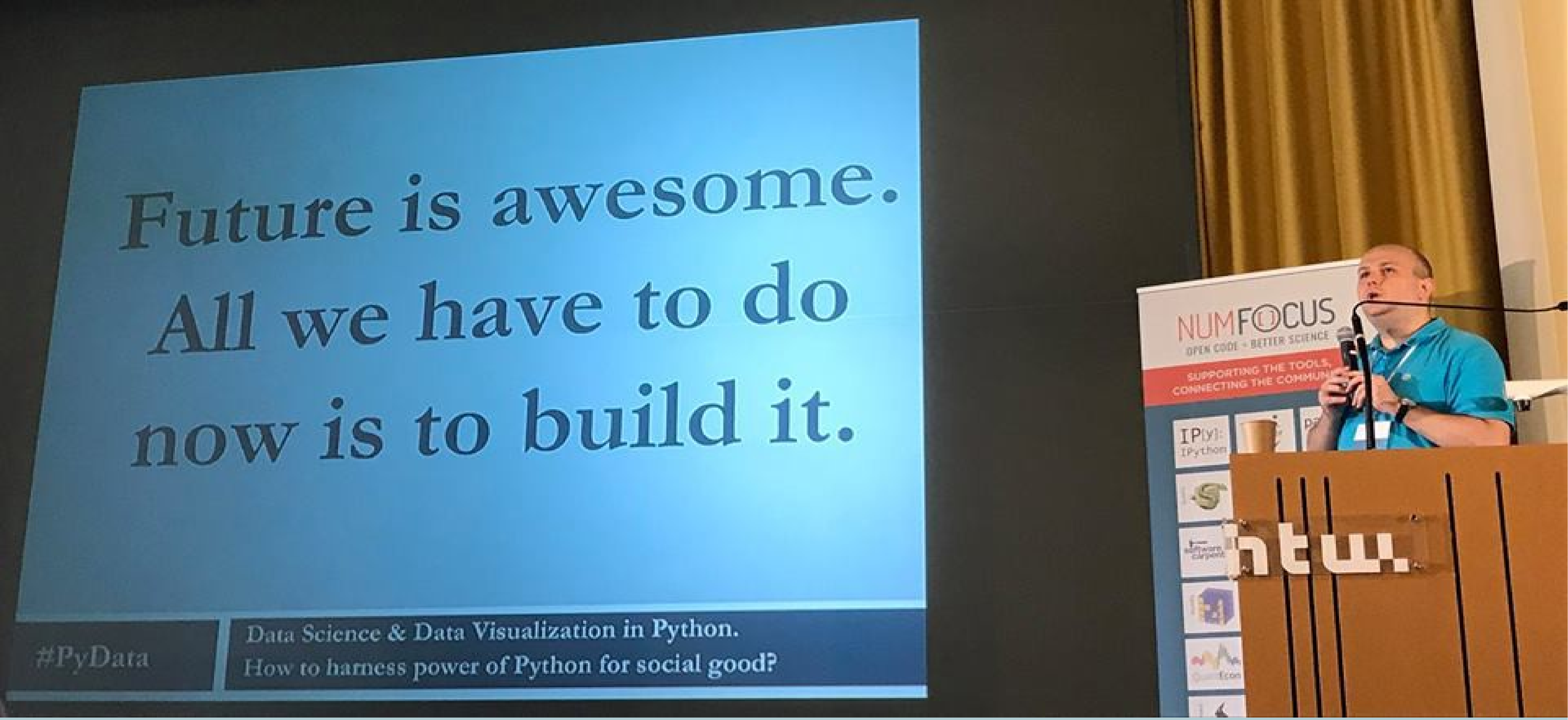
Source: https://github.com/AllForJan

- 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.

- 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/opencamp2019bratislava
- https://github.com/radovankavicky/OpenCampBratislava2019

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

Thank you for your attention

PyData
Bratislava

Contact:

Radovan Kavicky

radovan.kavicky@gapdata.org

radovan.kavicky@gmail.com

+420 777 595 262 (CZ)

+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

GAP
DATA
INSTITUTE

#OpenCampBA

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

Oh, and one more thing… :)

Slovakia julia users group

- Meetup.com

https://www.meetup.com/Julia-Users-Group-Slovakia/

- Facebook

https://www.facebook.com/groups/379292635993253/

Hashtag: #JUGSlovakia

More to come... soon. Stay tuned.