PyData
Berlin 2017

GAP
DATA
INSTITUTE

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

(PyData Berlin 2017, HTW Berlin; #PyData)

1. 7. 2017

Radovan Kavický, GapData Institute

Talk Structure

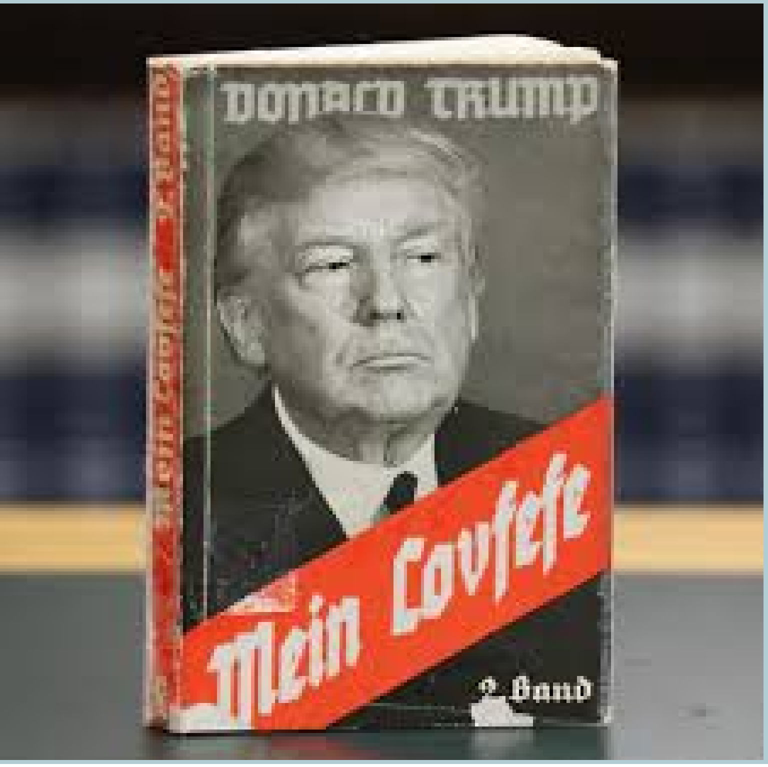
- **1. Introduction (non-Python part) & Why to do it?/Motivation** (7-10 min.)
  - **Why Data Science and why more people should be really interested?**
  - **Open Data, Open Government Partnership, Open Public Administration & all the advantages of Open Data Science & Python.**
- **2. Main (Python part) & How to do it?** (17-20 min.)
  - **Python & Data Science (Data Science Workflow, IDE-s)**
  - **Data Science & Visualization Tools (best libraries will be shown for the specific purpose; focus particularly on Bokeh, Seaborn, Plotly and Jupyter Notebook)**
  - **How to "unlock" the insights hidden in data and how to use it to transform not only public administration or business, whole society and economy towards the insight & knowledge based?**
- **3. Conclusion (the end of talk +vision), Potential & Results** (7-10 min.)
  - **Data-Driven Approach. Everywhere. Now.**

About me

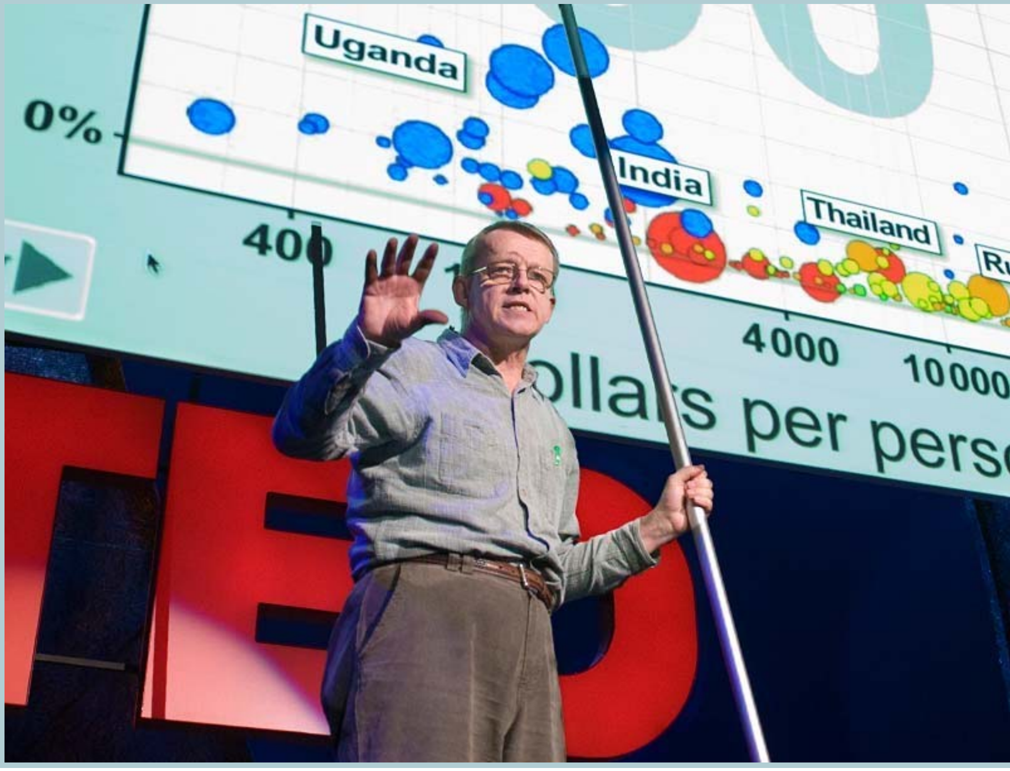
- Economist (Macro, Finance, Statistics)
- R, Python, Tableau > Matlab, SAS, Stata
- Data Science & Open Data & Public Policy
- PyData Bratislava, R <- Slovakia, skczTUG (Tableau User Group)

Picture the media shows us & the reality

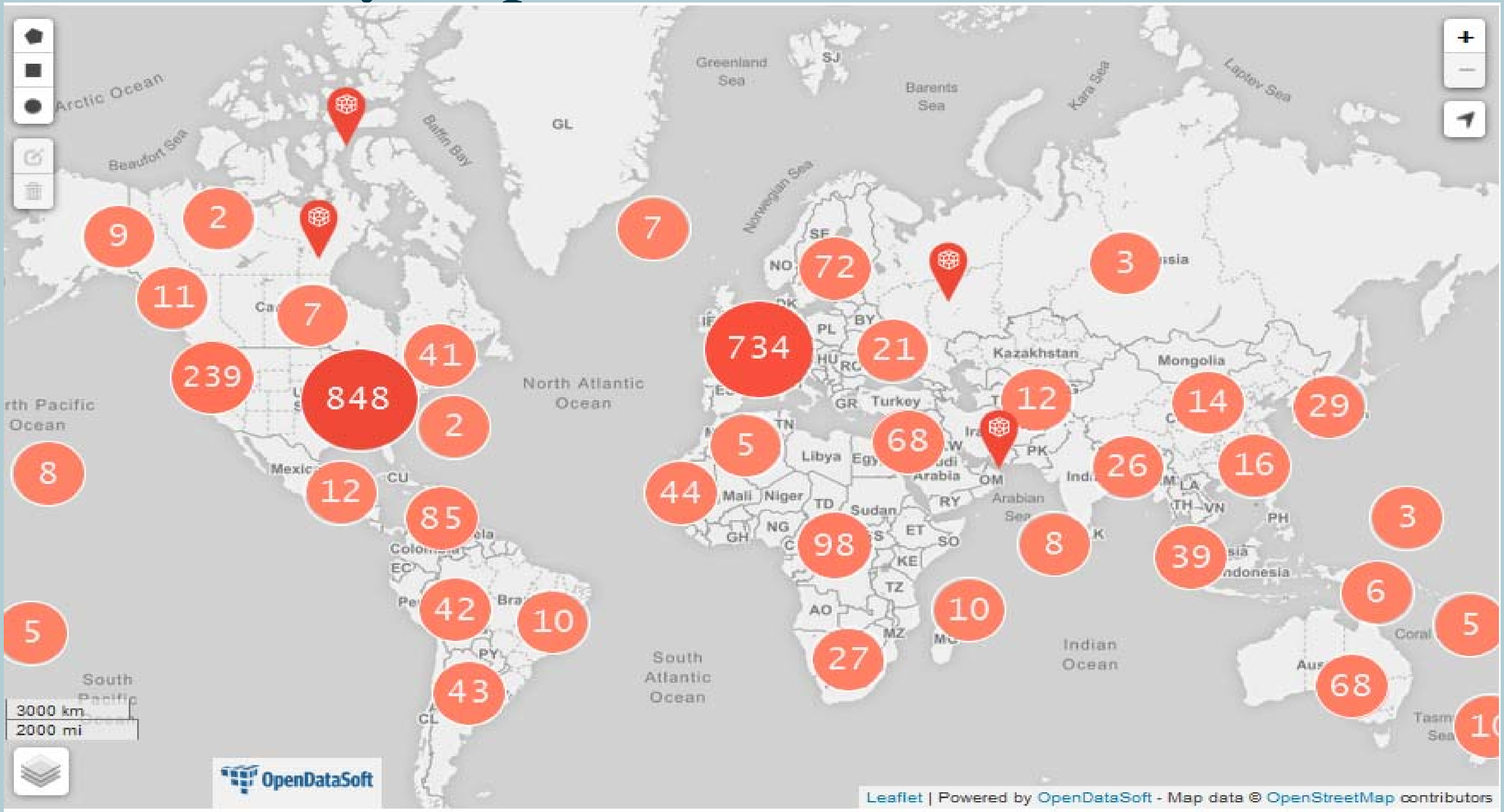
- Post-Factual World & Fake News
- Post-Truth Politics & Public Policy done via Twitter (21st century!)
- Some Funny Guy with little hands tweeting “covfefe” at 1:00am



vs.

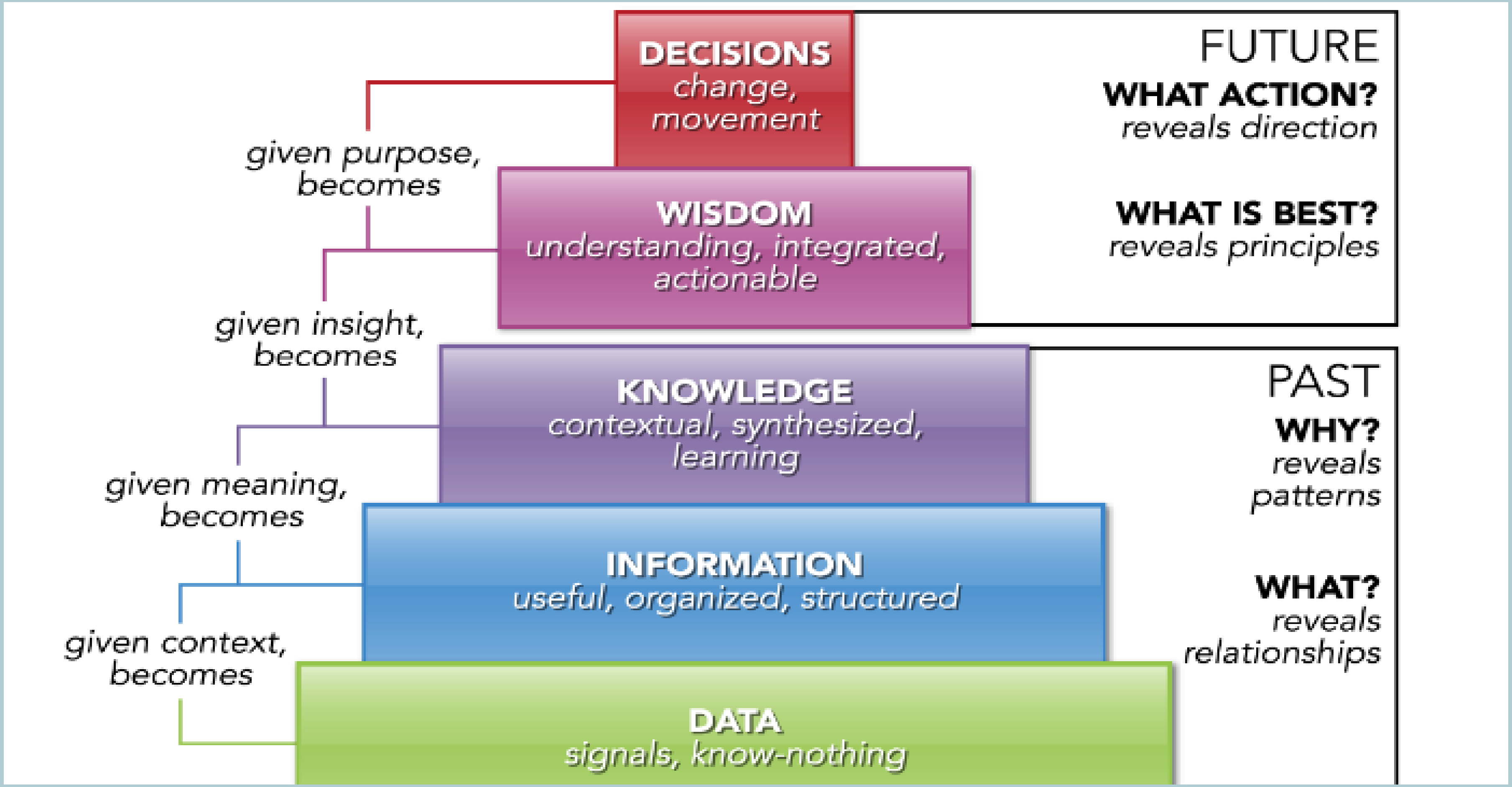


How many “Open Data” are there?



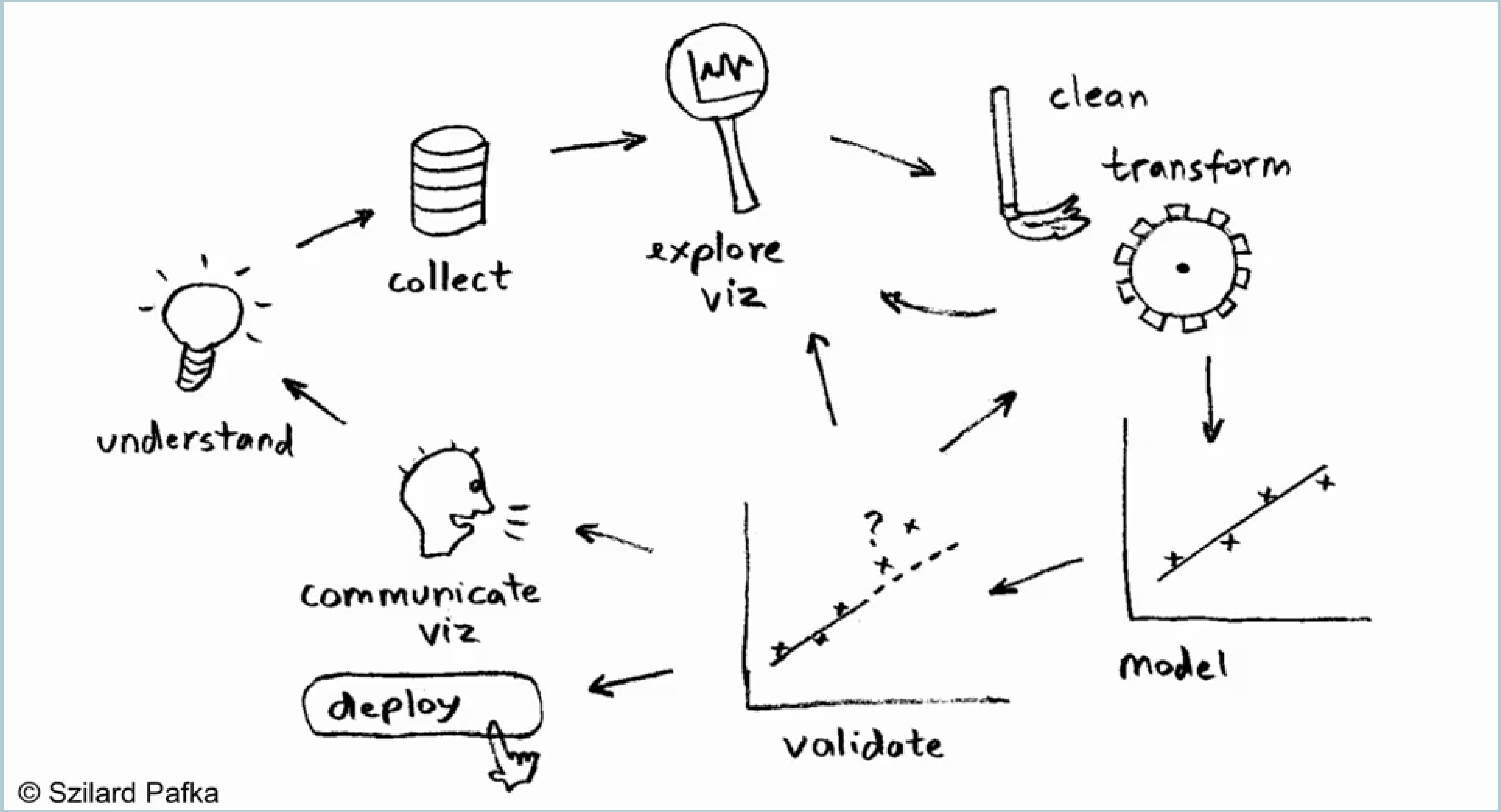
Source: https://www.opendatasoft.com/a-comprehensive-list-of-all-open-data-portals-around-the-world/

Data, Information, Knowledge, Wisdom, Decision Pyramid



Source: https://www.linkedin.com/pulse/handy-concept-better-dikwd-pyramid-peter-j-korsten

How to do it? (Data Science Process/Workflow)



Source: https://blog.dominodatalab.com/video-huge-debate-r-vs-python-data-science/

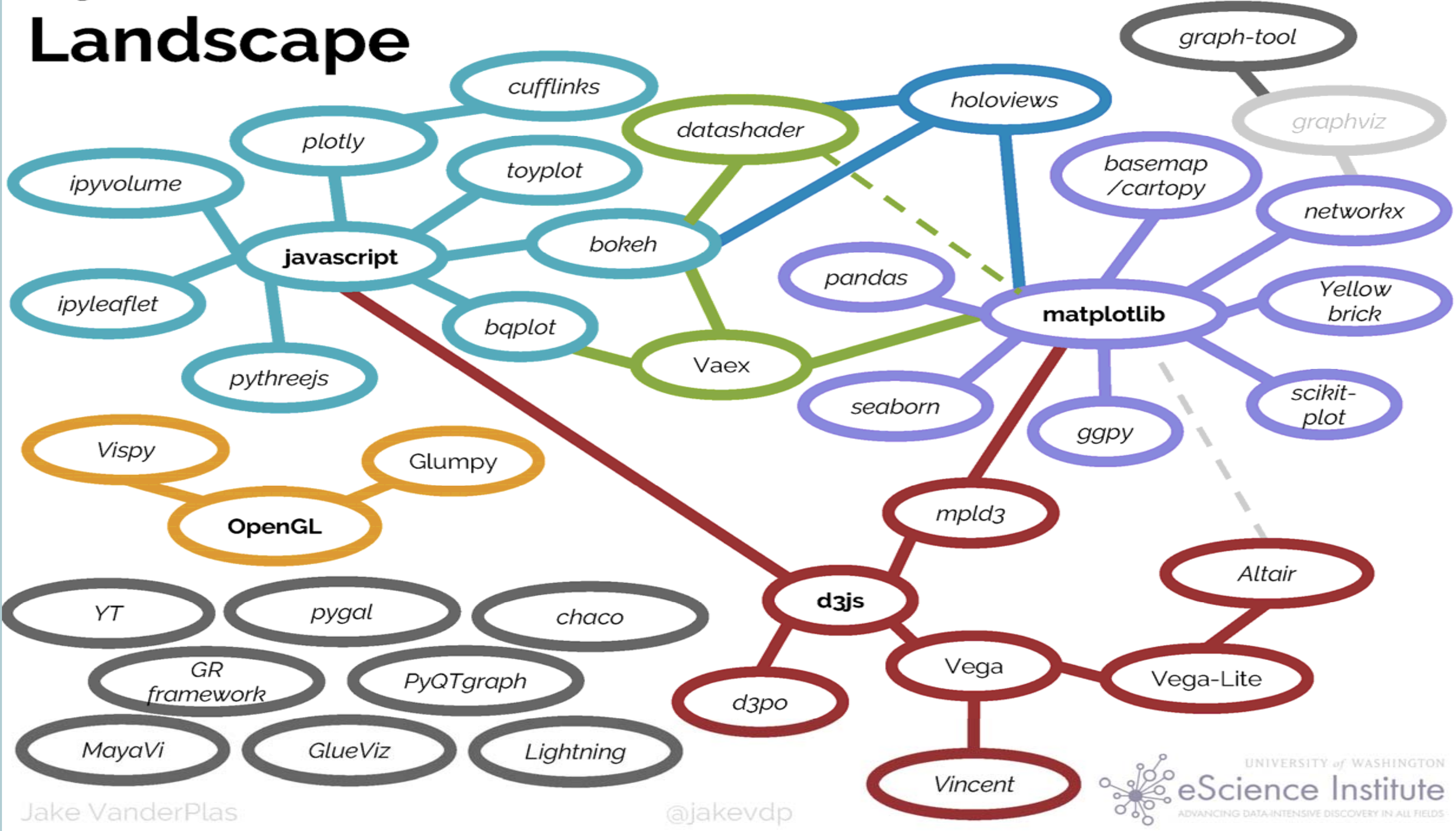
IDE-s for Python, R & Tableau as Data Science Platform (Data Science Toolbox)

- □ Anaconda/Spyder
- □ Rodeo (IDE for Python & R)
- □ Jupyter Notebook
- □ PyCharm (IDE for Python)
- □ R-Studio (IDE for R +rpy2)
- □ Wing + VIM/NeoVim (IDE for Python)
- □ Tableau Public (Python & R code implementation +TabPy)

Python (best tools @ Data Science)

- Data Collection – Feather (binary file format/non-csv/Apache Arrow/Wes McKinney), Ibis (remote/Hadoop +SQL & bridge it with pandas), ParaText (parallel reading of csv/C++)
- Data Visualization – Seaborn (matplotlib based/static), Bokeh (interactive/d3.js like), Plotly (declarative dataviz), Altair (static/js), Geoplotlib (maps)
- Data Cleaning & Transform - datacleaner (automate cleanining your data in Pandas), Blaze (NumPy/pandas-like), xarray (pandas-like), Dask (parallel computing)
- Data Modeling – StatsModels +Patsy (describe statistical models), PyStan (Bayes/C++), PyMC3 (Bayes/statistical modeling), Keras (TensorFlow/DL)

Python's Visualization Landscape



Source: https://speakerdeck.com/jakevdp/pythons-visualization-landscape-pycon-2017

Jupyter Notebook +Seaborn

- IPython/Project Jupyter (Julia, Python + R)
  - Jupyter notebook (.ipynb rendering via web browser) +nbviewer
- Jupyter as Data Science Front-End f.e. in Bloomberg, Microsoft, IBM Watson, and Netflix
- Very powerful tool for collaborative data science and making analysis actionable.
- JupyterHub (Jupyter notebook server/multi-user data science teams)
- Seaborn (dataviz/matplotlib based/static)

Bokeh +D3.js

- Interactive visualization (JavaScript-like) library and platform Python
- Gallery: http://bokeh.pydata.org/en/latest/docs/gallery.html
- Bindings with R (rBokeh) +Scala (bokeh-scala)
- D3 (DataViz, JavaScript)
- Goal is to provide elegant, concise construction of novel graphics in the style of Protovis/D3, while delivering high-performance interactivity over large data to thin clients
- standalone HTML documents, or server-backed apps
- No Java-Script; HoloViews & GeoViews (annotate data +visualize/render with Bokeh)

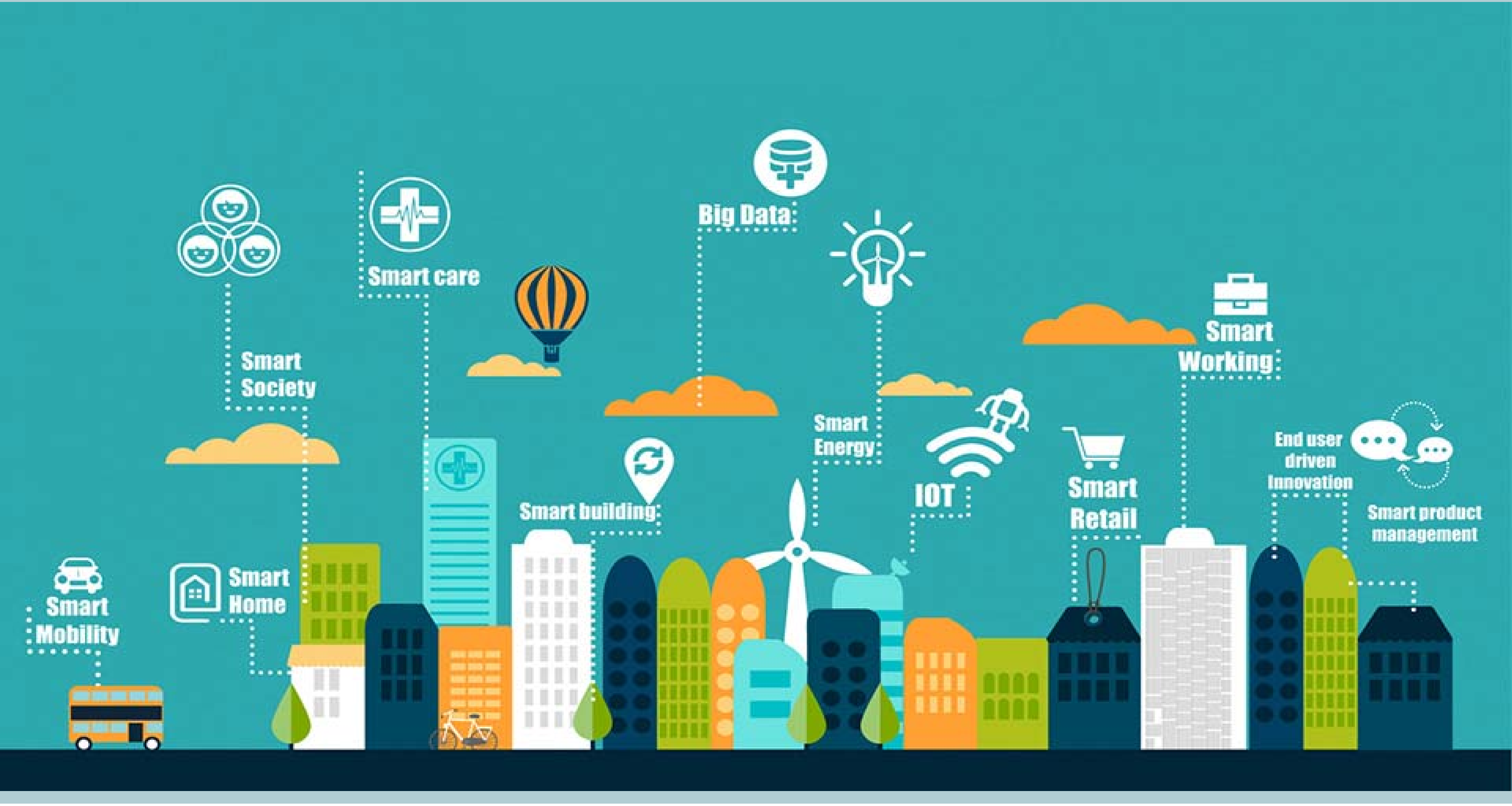
Plotly +Dash (R’s Shiny for Python)

- Plotly (descriptive data visualization .json converter)
- DataViz @ Matlab, R, Python, Julia, Perl, Arduino, REST
- Python +Django framework; only front end is JS
- Not just plots/data visualizations, but also presentations and dashboards/web-apps
- Dash (Interactive, reactive web apps in pure Python/js-free), analytical web application

Potential of Data Science

- Citizen Data Scientists
- Open Data (only the 1st necessary step)
- Smart Cities
- Economic Reforms (any area)
- Data-Driven Public Policy Making
- Data Visualization (Interactive DataViz Tools)
- Data-Driven mobility/self-driving vehicles
- Data Science in the center of Digital transformation
- Age of Data with Data as the new currency
- Nearly any current problem can & will be solved by data

Why smart cities… why not? It's necessary.



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

GapData Institute (GDI)

- 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/pydata2017berlin
- https://github.com/radovankavicky/PyDataBerlin2017

Thank you for your attention

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/messages/py-data/

https://github.com/radovankavicky

https://github.com/GapData/PyDataBratislava

@radovankavicky, @PyDataBA, @GapDataInst

PyData Bratislava

GAP DATA INSTITUTE

#PyData

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