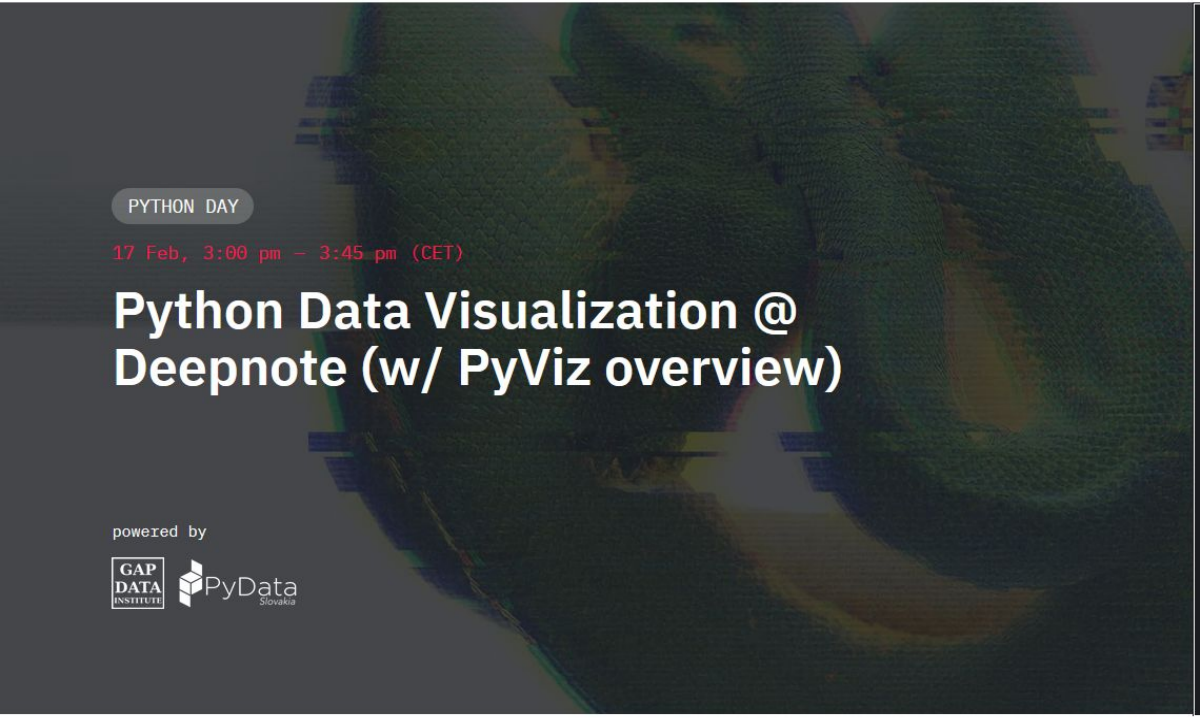


Python Data Visualization @ Deepnote (w/ PyViz overview)

#WeAreDevs #WeAreDevsLIVE

17. 2. 2021

Radovan Kavický, GapData Institute & PyData Slovakia #PythonDay

Talk Structure

- **1. Introduction, History of Python, DataViz & Data Science + Why Python?/Motivation** (8-10 min.)
- **Humble Beginnings of Data Visualization & Science**
- **What is Data Science & Data Science Workflow**
- **Data Visualization Landscape**
- **2. Data Visualization in Python (tools), Where to start & How to do it?** (10-12 min.)
- **Data Visualization in Python (PyData & PyViz overview)**
- **Why should people care? (PyData Worldwide & PyData movement/educational programme)**
- **3. Common DataViz Problems and pitfalls/mistakes (and how to resolve them)** (8-10 min.)
- **4. Conclusion (the end of talk +vision), Potential (our activities, partners, meetups) & Results** (2-5 min.)
- **Future of Data Visualization & Science (AI/ML/NLP in/outside Python)** (1-2 min.)
- **5. Discussion/Q&A** (5-6 min.)

Slides (this talk) > tiny.cc/WeAreDevsPython2021, Deepnote (Live-Code) > http://bit.ly/PythonDayDeepnote

Slides & code (later today) https://github.com/radovankavicky/WeAreDevs2021

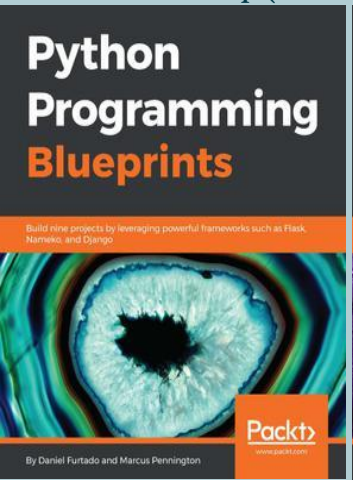
About me

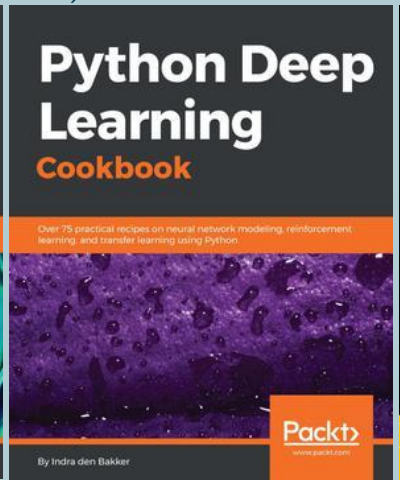
- Economist (Macro, Finance); Slovak Economic Association (SEA)
- Principal Data Scientist (GapData), Consulting (public, private)
- Member of PSF, Slovak.AI, CLAIRE, EU AI Alliance, TAILOR & UDSC
- Data Science Instructor @ DataCamp, BaseCamp.ai, Learn2Code
- Founder of PyData Slovakia/Bratislava (#PyDataBA), R <- Slovakia (#RSlovakia), Julia Users Group Slovakia (#JUGSlovakia) & SK/CZ Tableau User Group (#skczTUG)

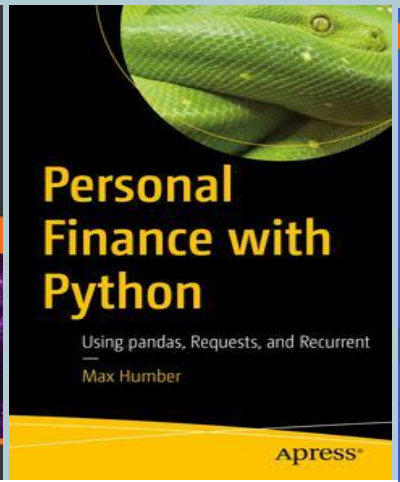
DataCamp

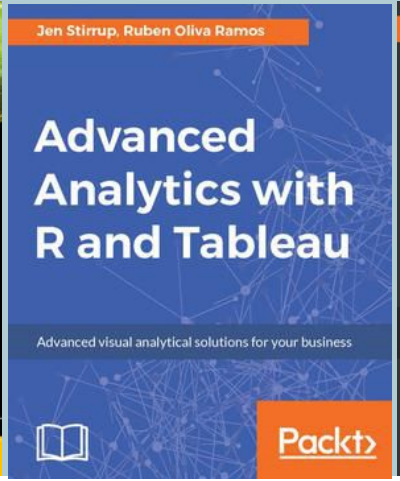
LEARN 2CODE

BASECAMP









Python Data Visualization @ Deepnote (w/ PyViz overview)

slido

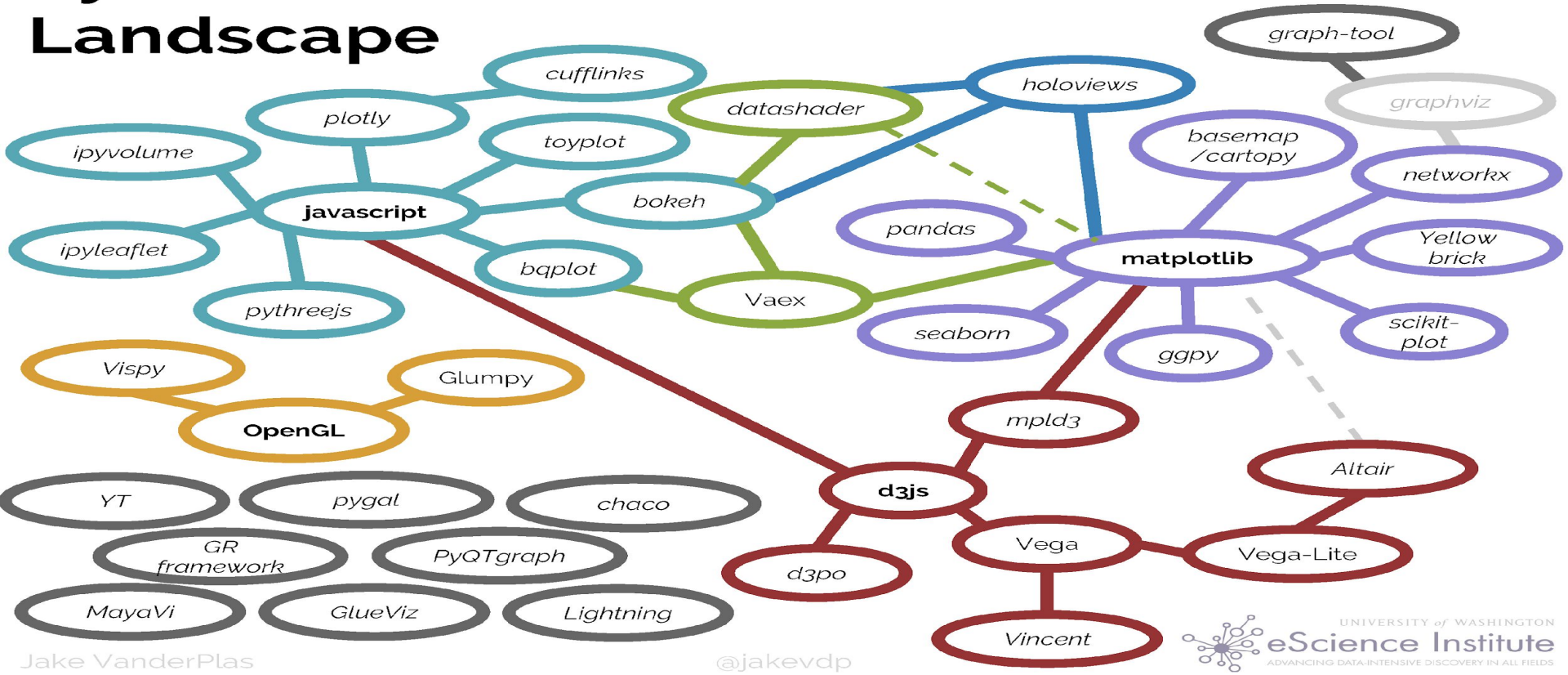
(Are you/Do you see yourself as) Python Developer or Data Scientist?

ⓘ Start presenting to display the poll results on this slide.

Where & When did Python start?

- 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

Python's Visualization Landscape

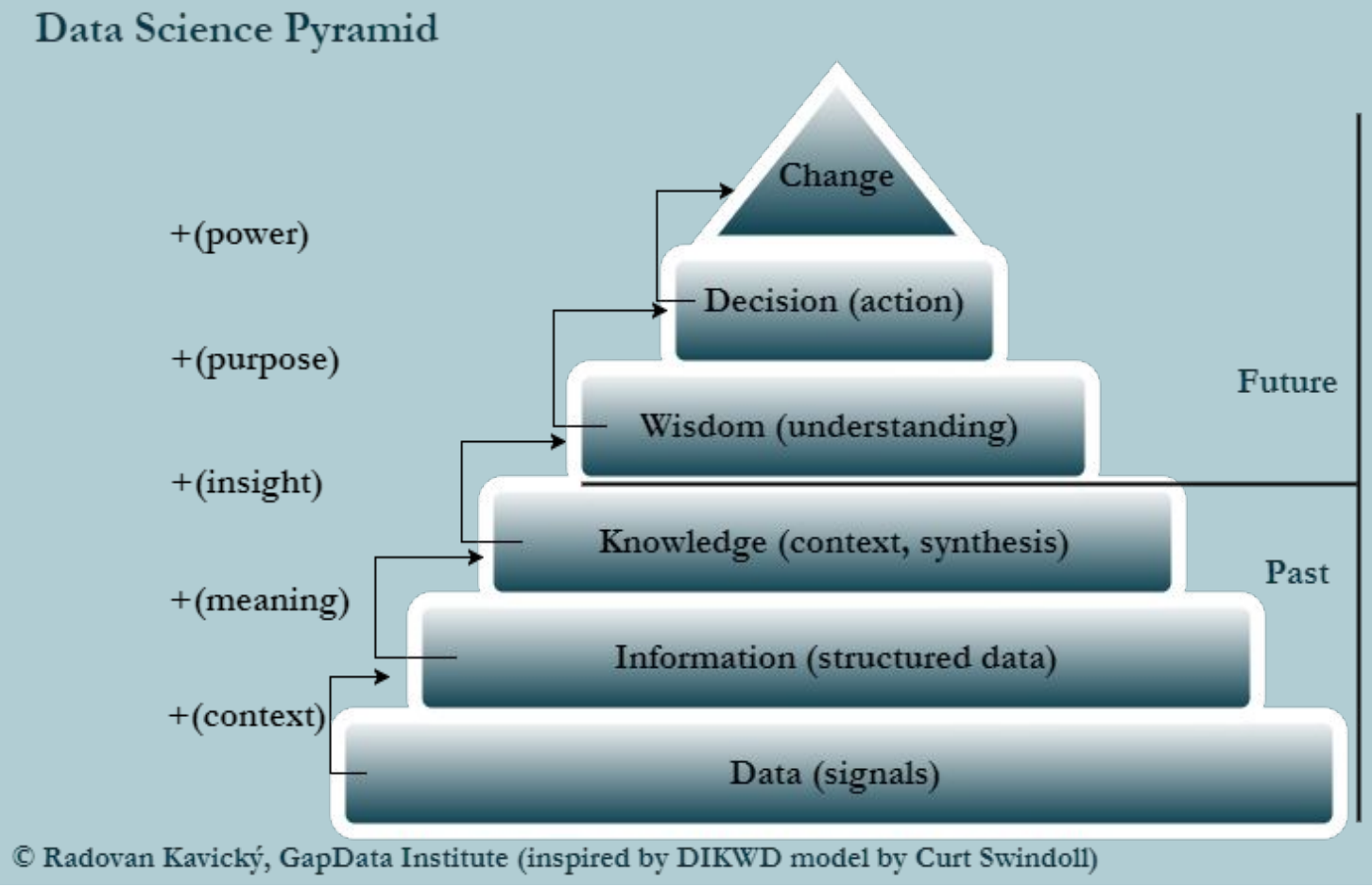


Source: https://speakerdeck.com/jakevdp/pythons-visualization-landscape-pycon-2017, OR https://pyviz.org/overviews/index.html

slido

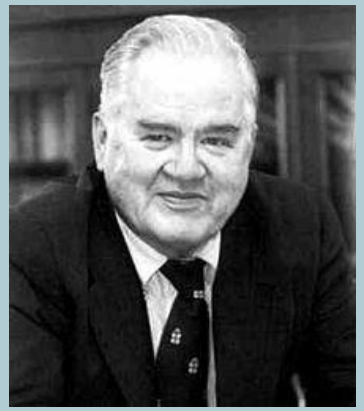
In which area you use Python?

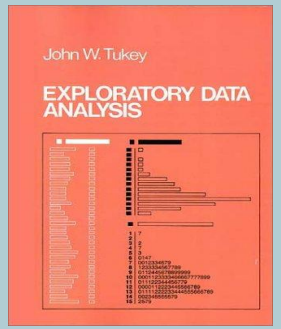
ⓘ Start presenting to display the poll results on this slide.



Python Data Visualization @ Deepnote (w/ PyViz overview)

History of Data Science & Data Visualization





- 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

John W. Tukey (1977), Exploratory Data Analysis

"The greatest value of a picture is when it forces us to notice what we never expected to see."

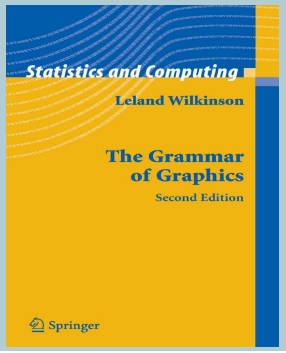
Leland Wilkinson (1999, 2005), The Grammar of Graphics (gg, ggplot2)

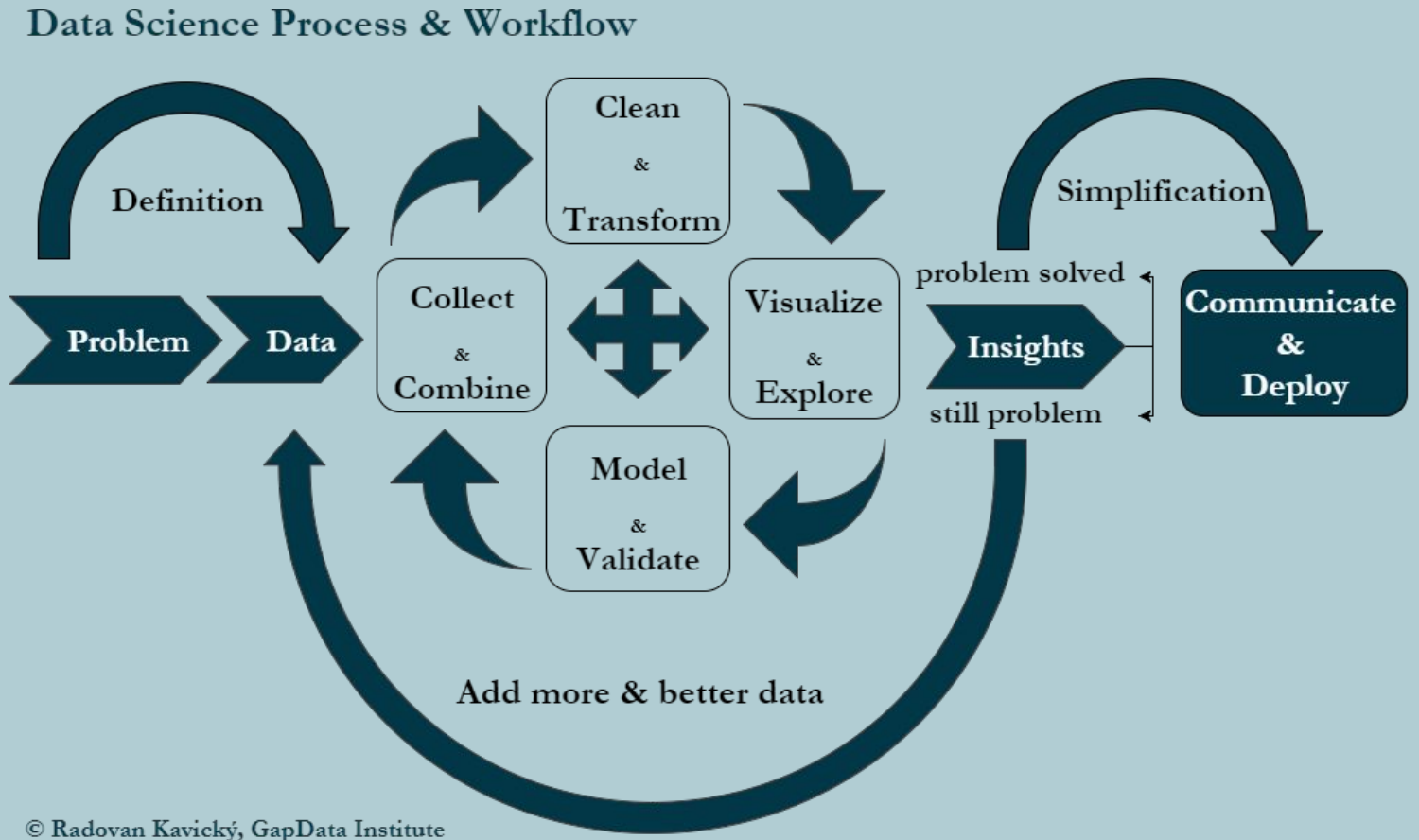
"This book is not about good taste, practice, or graphic design, however. This book focuses on rules for constructing graphs mathematically and then representing them as graphics aesthetically."

- More on History of Data Science: Donoho, David (2015, Stanford), https://courses.csail.mit.edu/18.337/2015/docs/50YearsDataScience.pdf

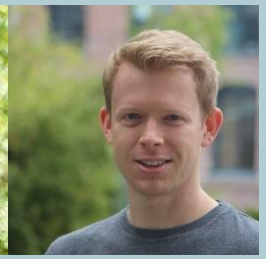
Radovan Kavický - Data Science with Python: Past, present and future

https://www.youtube.com/watch?v=8mBI3iioT8A




Data Science Process & Workflow
Definition
Problem
Data
Collect & Combine
Clean & Transform
Model & Validate
Visualize & Explore
Simplification
problem solved
Insights
still problem
Communicate & Deploy
Add more & better data
© Radovan Kavický, GapData Institute

Python Data Visualization (PyData & PyViz overview)



- Travis Oliphant (NumPy + SciPy/Matlab alternative)

1995 (Numeric), 2006 (Numpy)

- John Hunter (matplotlib) 2003
- Wes McKinney (pandas)

January 2008, Python, C, Cython

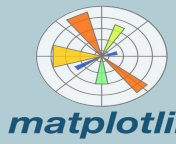
- Anaconda as one of the Python distributions (1000 selected libraries for Data Science)
- NumFocus
- PyData conferences and local meetups (PyData Bratislava)

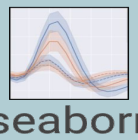
PyData Bratislava

ANACONDA

PyData Slovakia

Python Data Visualization (PyData & PyViz overview)





- Matplotlib (dataviz/static)
  - works well with pandas/included
- seaborn (dataviz/static)
  - more options than Matplotlib
- Altair (dataviz/Vega-based, JSON format)
- Bokeh (Python & JavaScript-based, D3-like)
- Plotly (JavaScript-based, universal)
- Datashader (big data/billions of points/DARPA)
- Dash (dashboards in Python)

slido

Is Python your only programming language? (pick also another that you use, Julia, R, C, C++, Java, JavaScript, other)

ⓘ Start presenting to display the poll results on this slide.

Python & Data Science (tools)

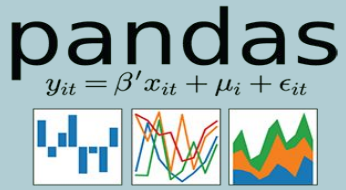
IDE's – Jupyter Notebooks (IPython kernel), PyCharm, JupyterLab

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

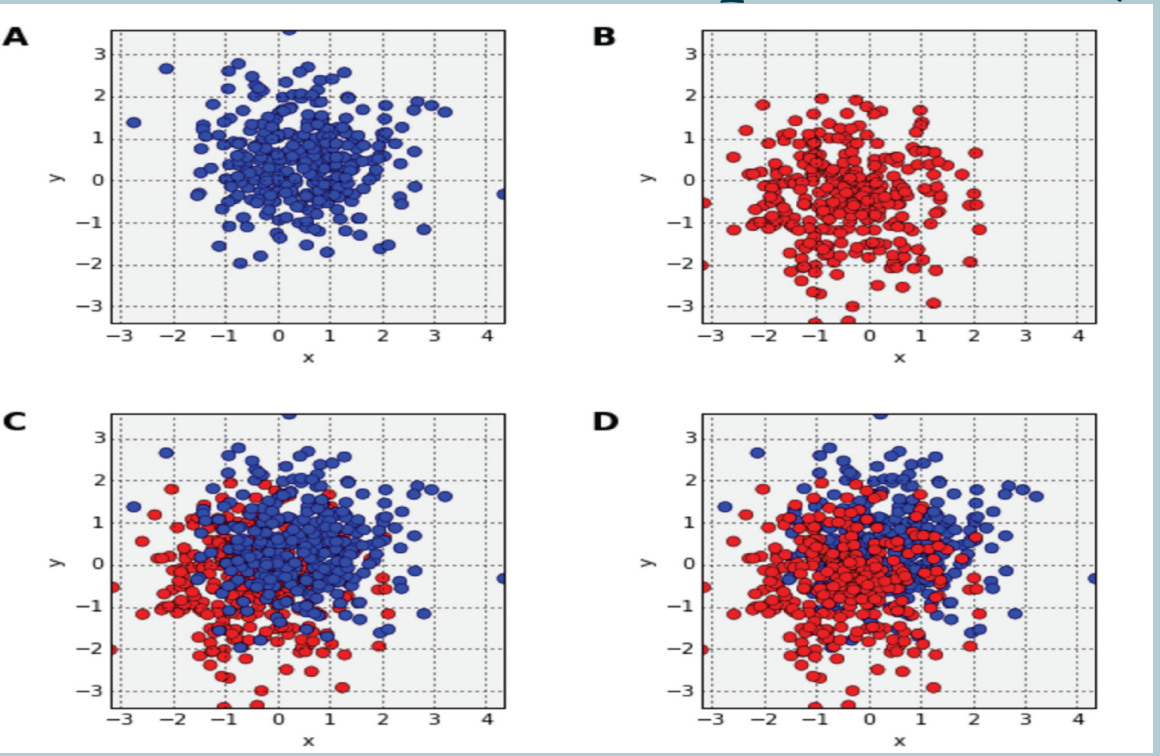
Data Visualization – Seaborn (matplotlib based/static), Bokeh (interactive/d3.js like), Plotly (declarative dataviz), Altair (static/js)

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)

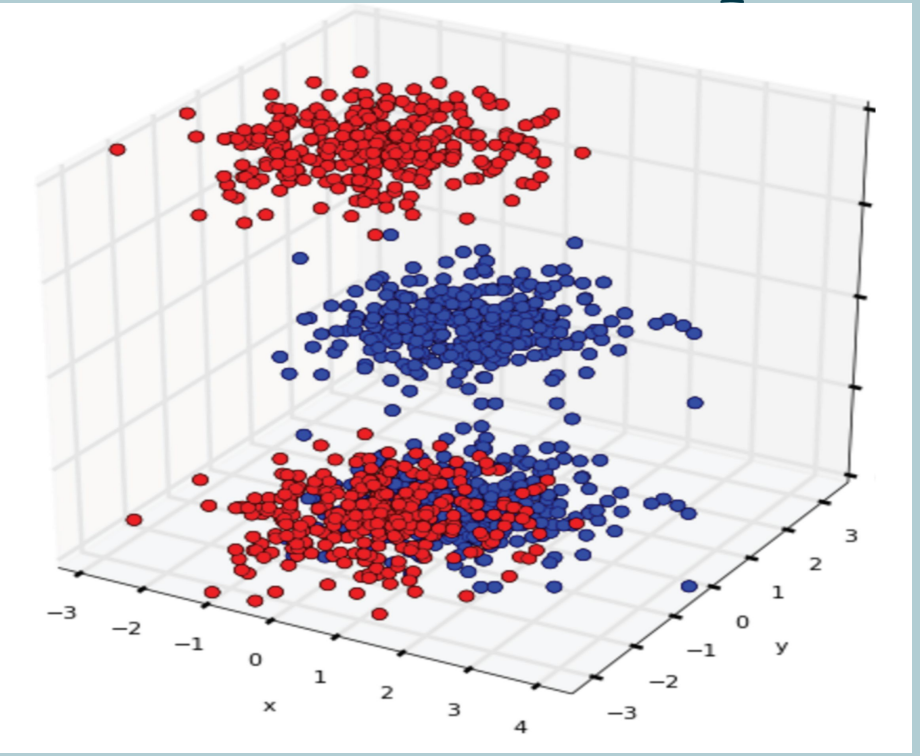


Common data-vis problems (overplotting)



- order matters
- last data plotted over-plots the previous
- Solution:
- set alpha/alpha transparency
- jittering (add a little random noise to the date)
- change shape (data point) or reduce size
- remove fill color (for data points)
- reduce amount (scale)

Common data-vis problems (overplotting)



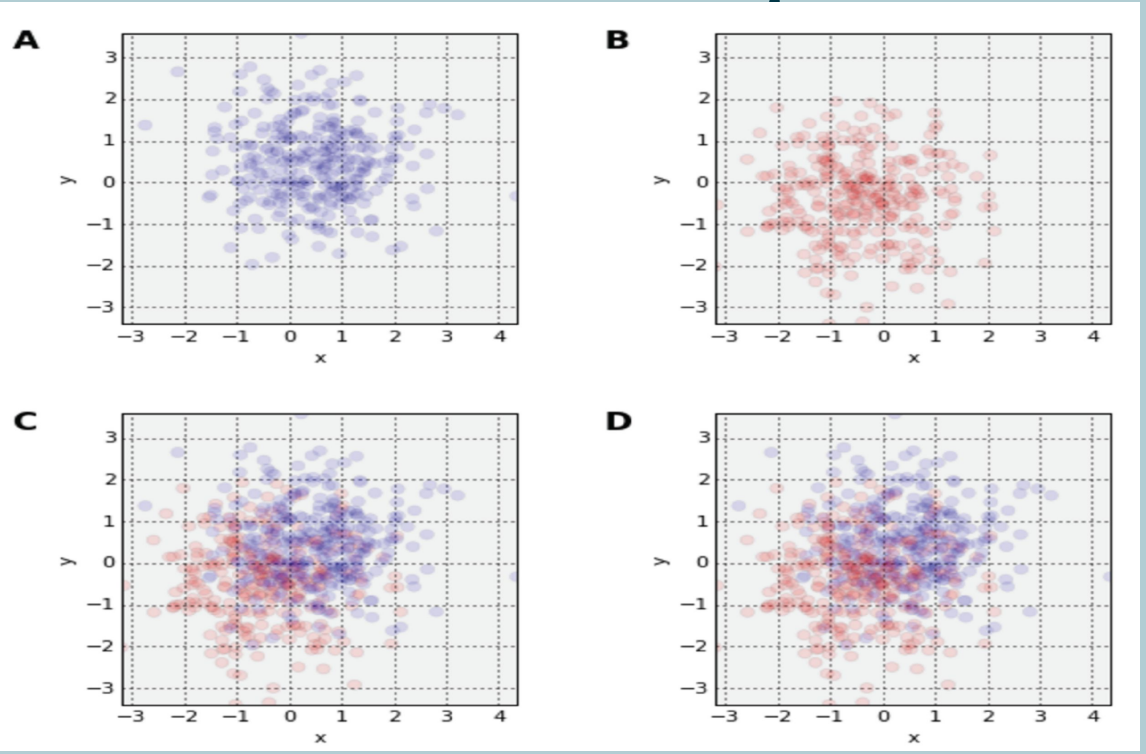
- occlusion (visually absorbed information)
- typical for 3D datavis (data within the physical space/ behind the object the user is viewing)
- Solution:
- set alpha/alpha transparency or semi/partial-transparency
- rotation (re-plotting)
- jittering (mostly useful for 2D)
- scale/remove the data plotted

slido

Which IDE for Python do you use? (multiple choice/you can select more than one that you use)

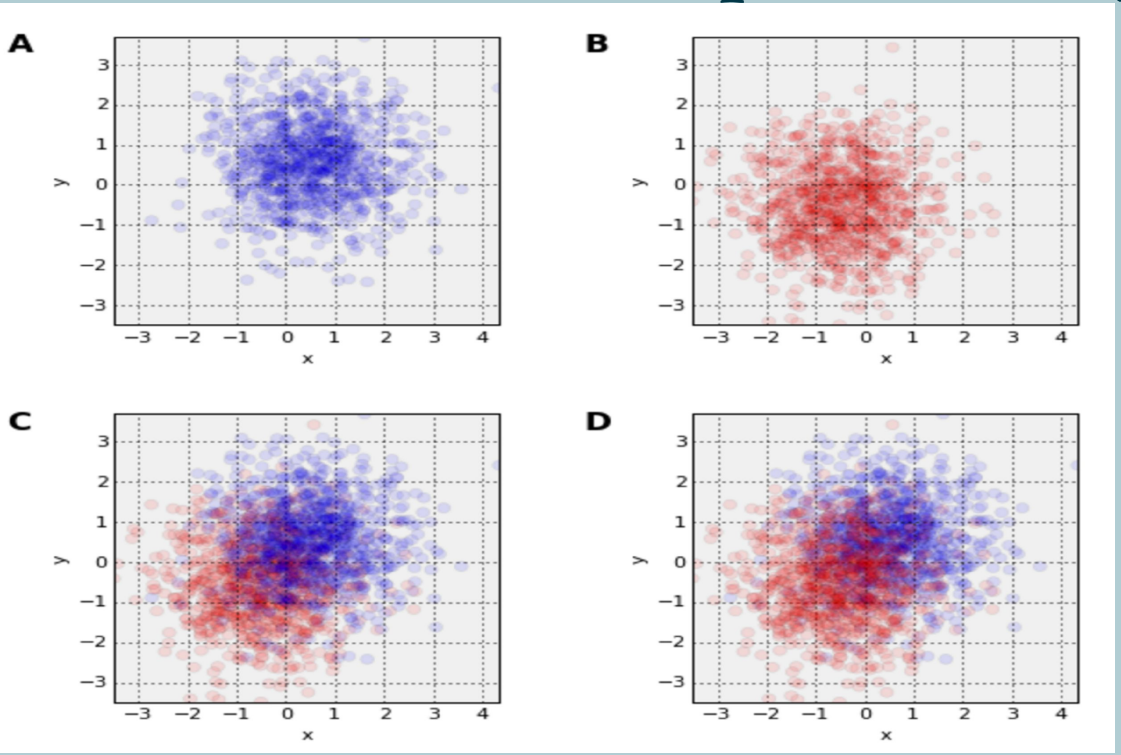
ⓘ Start presenting to display the poll results on this slide.

Common data-vis problems (saturation)



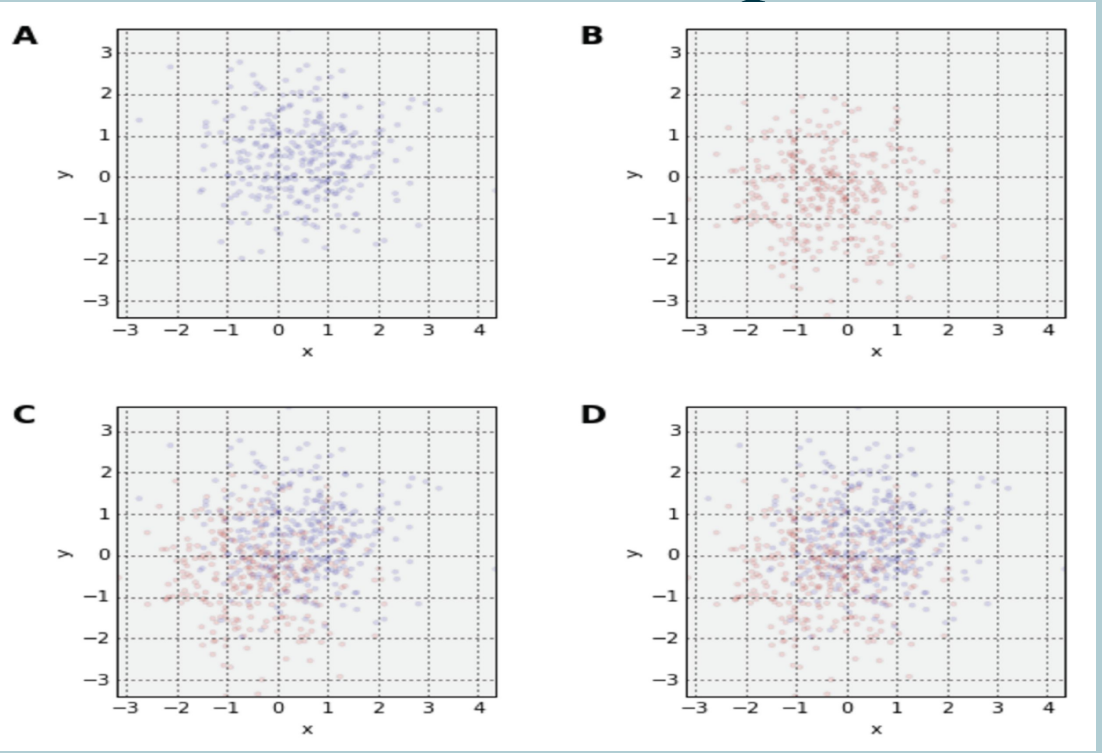
- intensity of colour
- f. e. if you set the alpha to 0.1, you have only about 10 points that can overlap before saturation (+first plotted data loss)
- > 10 overlapping points identical
- Solution:
- smaller points
- change shape no fill of data points

Common data-vis problems (saturation)



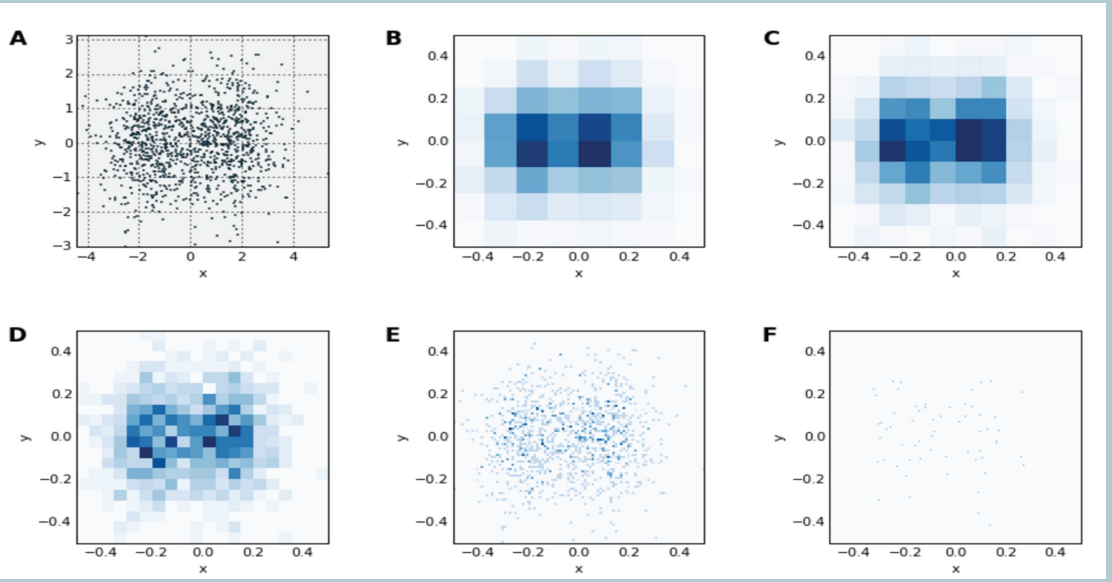
- with same alpha value, but more points (same problem)
- hard to know when the data are misrepresented
- Solution:
- again smaller points can help (+higher resolution)
- although shape and size of the point is tricky to pick

Common data-vis problems (saturation)

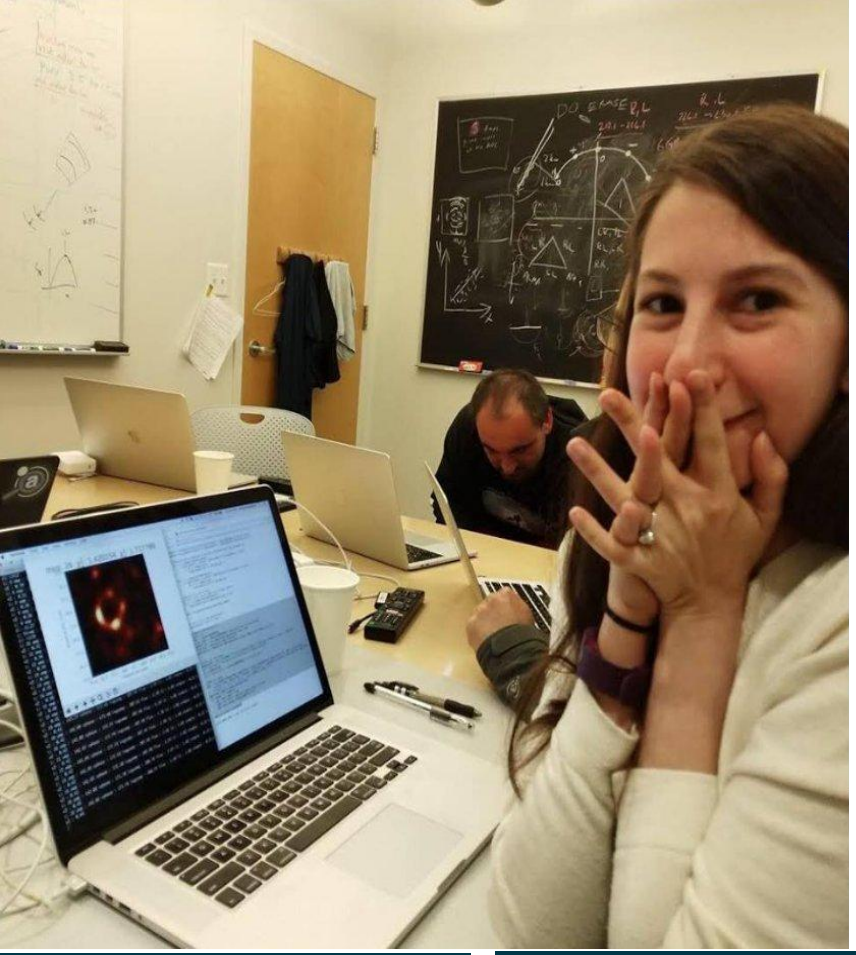


- smaller points you cannot see clearly
- one pixel per data point
- Solution:
- higher resolution or scaling +different type of visualization, f.e. heatmap
- auto-range on pixels (data-shades)
- add zooming/interactivity for user

Common data-vis problems (binning)



- how many bins or pixels should represent a data point
- tricky to set the right binning
- two merged normal distributions look different under different binning
- Solution:
- heatmaps + scale/data-shading and zooming in/out



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)

Future of DataViz (in Python too) ML/AI (NLP and voice)

alphaa.ai

DataRobot

- AI/ML augmented analytics (scenarios for data-driven decision making), BI & AI worlds will collide
- build and deploy highly accurate machine learning models without writing a single line of code and transparent/viz too
- experiment, simulate and compare different scenarios using various models/scenarios to identify the best strategy or test ideas before committing resources, voice & text interaction
- focus on the right data to analyze, get predictive insights with explanations in your dashboards, H2O.ai AutoViz (Leland Wilkinson/ex-Tableau)
- run simulations to get actionable prescriptive guidance on what to do next and get instant visual response
- Alphaa.ai (world's 1st AI voice analyst/not regular "chatbot")

In past we could use:

- TabPy (connect to Tableau Server from Python)
- Rserve (TCP/IP binary R server)

H2O.ai

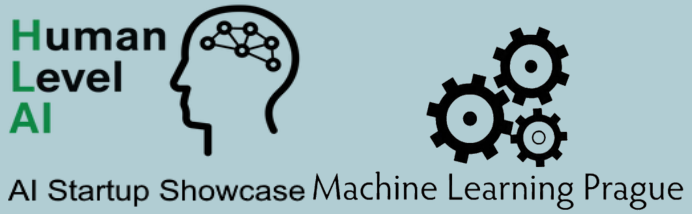
tableau

slido

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

ⓘ Start presenting to display the poll results on this slide.

PyData Slovakia/Bratislava & activities within CEE + V4



- 2000+ members
- 5000+ on mailing list
- 900+ followers
- Partners: O'Reilly, NexTech, robime.it, Learn2Code, sli.do
- Sponsors: H2O.ai, SwissRe, Kiwi.com
- WeAreDevelopers World Congress (“GAPDATA-25” & “PYDATA-25”), Berlin/Vienna
- OSCON 2021 (“PCGAPDATA” for 25% discount)
- Our Conferences (satRday, PyData Slovakia)

#PythonDay @ sli.do

Python Data Visualization @ Deepnote (w/ PyViz overview)

GapData Institute (GDI) and how to support us.

GAP DATA INSTITUTE

NUMFOCUS
OPEN CODE = BETTER SCIENCE

PyData Bratislava

- 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/WeAreDevsPython2021
- https://github.com/radovankavicky/WeAreDevs2021
- Deepnote (Live-Code) > http://bit.ly/PythonDayDeepnote

PAY by square

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

slido

How did you like this talk? (It was: 1 star/low, 2 stars/ok, 3 stars/meh/average, 4 stars/high, 5 stars/awesome)

ⓘ Start presenting to display the poll results on this slide.

Thank you for your attention

Slides (this talk): tiny.cc/WeAreDevsPython2021

https://github.com/radovankavicky/WeAreDevs2021

Deepnote (Live-Code) > http://bit.ly/PythonDayDeepnote

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

FB: https://www.facebook.com/groups/356635138031671

@radovankavicky, @PyDataBA, @GapDataInst

GAP DATA INSTITUTE

#wearedevslive @ sli.do

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


That's all Folks!