Essential Tools for Getting Started with Data Science

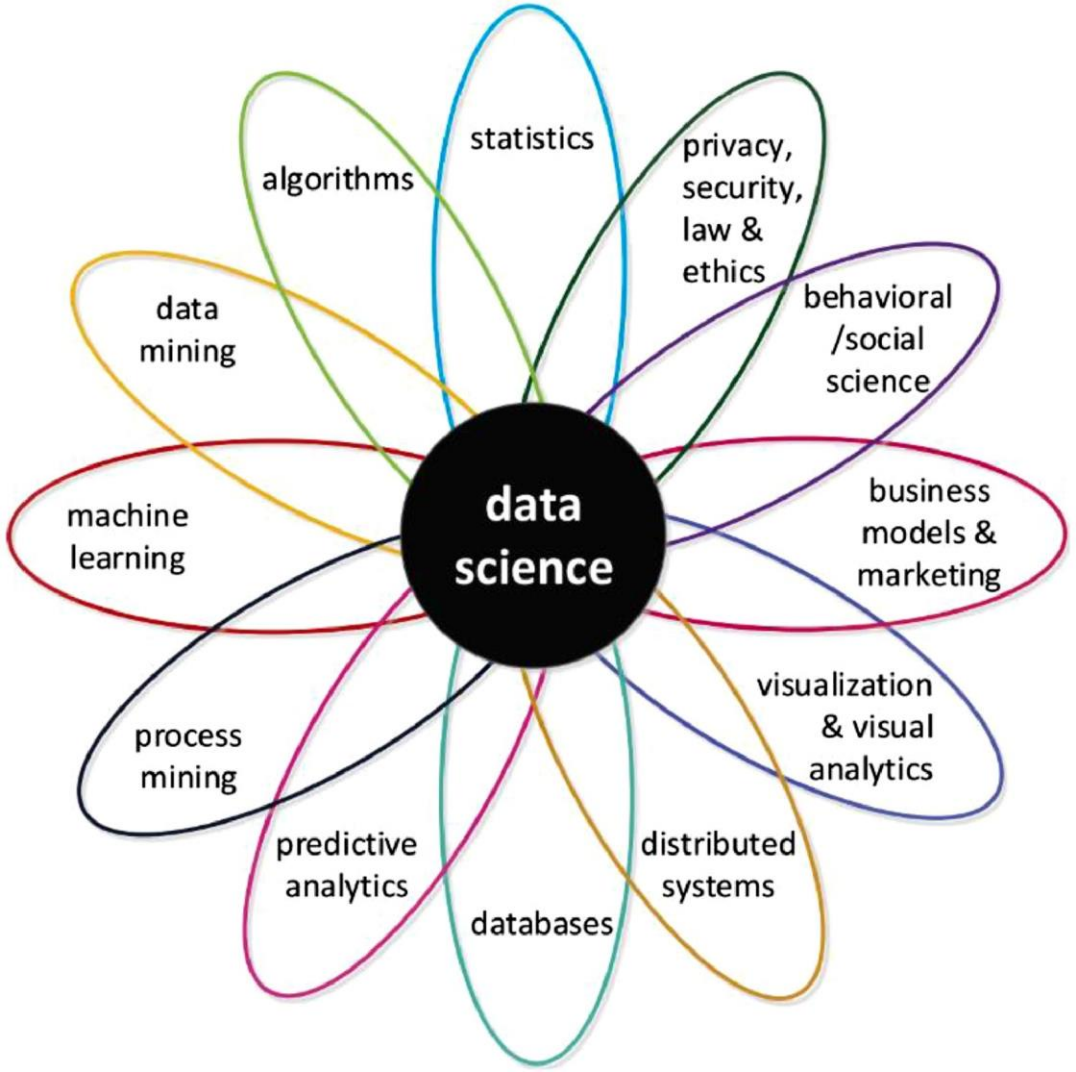
Andres Patrignani
Associate Professor in Soil Water Processes
Department of Agronomy
Kansas State University

ASA-CSSA-SSSA Webinar – 20 February 2025

What is Data Science?

Data science is the process of **extracting insights** from data by **collecting, cleaning, analyzing, and visualizing** results through **reproducible workflows**.

"Reproducible research is the idea that code and data from scientific publications are available so that others can analyze the data, reproduce key findings from the original paper, and use the code and data in developing future research" (Stodden at al., 2014; Skaggs et al., 2015)



The ingredients of data science.
Figure from Van Der Aalst, W., & van der Aalst, W. (2016). *Data science in action* (pp. 3-23). Springer Berlin Heidelberg.

No code options

Microsoft Excel / Google Sheets

- **Basic data cleaning, analysis, and visualization**.
- Can handle functions, pivot tables, and simple machine learning via add-ons.

Tableau / Power BI

- Drag-and-drop tools for **interactive dashboards**.
- Great for visual storytelling.
- Integration with tools like OneDrive.

KNIME / Orange

- **Node-based data analysis.** Each task has a specific node. Each node has "ports" for data input and outputs. Nodes that accomplish some specific pipeline can be grouped into a component.
- Community-based workflows.
- No-code statistics, machine learning, data processing, and visualization.

QGIS / ArcGIS

- Excellent for spatial data analysis.
- Graphical user interface with option to add code for implementing advanced custom features, automation, and reproducibility.

Code-based options

Multipurpose high-level programming languages
Steeper learning curve, but offer the greatest flexibility

- **MATLAB**: A programming environment designed for numerical computing, data analysis, and visualization, widely used in engineering and applied sciences. *https://www.mathworks.com*

- **Python**: A versatile, general-purpose programming language for data science, machine learning, and scientific computing. Strong community support. *https://www.anaconda.com/download*

- **R**: A statistical computing language specialized for data analysis, visualization, and machine learning, with a rich ecosystem of packages. *https://cran.rstudio.com*

- **Julia**: A high-performance language designed for numerical and scientific computing with native support for parallelism and seamless integration with Python and C libraries. *https://julialang.org*

Essential Skills and Common Python Modules in the Data Science Ecosystem

Interactive Plots

- Bokeh
- Plotly
- Altair
- Ipywidgets

Dashboards

- Streamlit
- Dash
- Panel
- Gradio

Apps

- Tkinter
- PySimpleGUI
- Kivy
- PyQt

Coding Environments

- Jupyter Notebooks
- Spyder
- PyCharm
- VSCode

Databases

- TinyDB
- MySQL
- MongoDB
- Cassandra
- Redis
- PostgreSQL

Essential coding skills

- Data types
- Data structures
- Understand objects
- Operations with dates
- Control flow
- Functions
- Modules

Core libraries

- Numpy
- Pandas
- Matplotlib
- Seaborn
- SciPy
- SQLite
- JSON

Geospatial Analysis

- Geopandas
- Shapely
- Folium
- Rasterio
- Xarray
- PySAL
- Cartopy
- GDAL
- Google Earth Engine

Web Development

- Bottle
- Flask
- Fast API
- Web2Py
- Django
- PyScript

Stats and Machine Learning

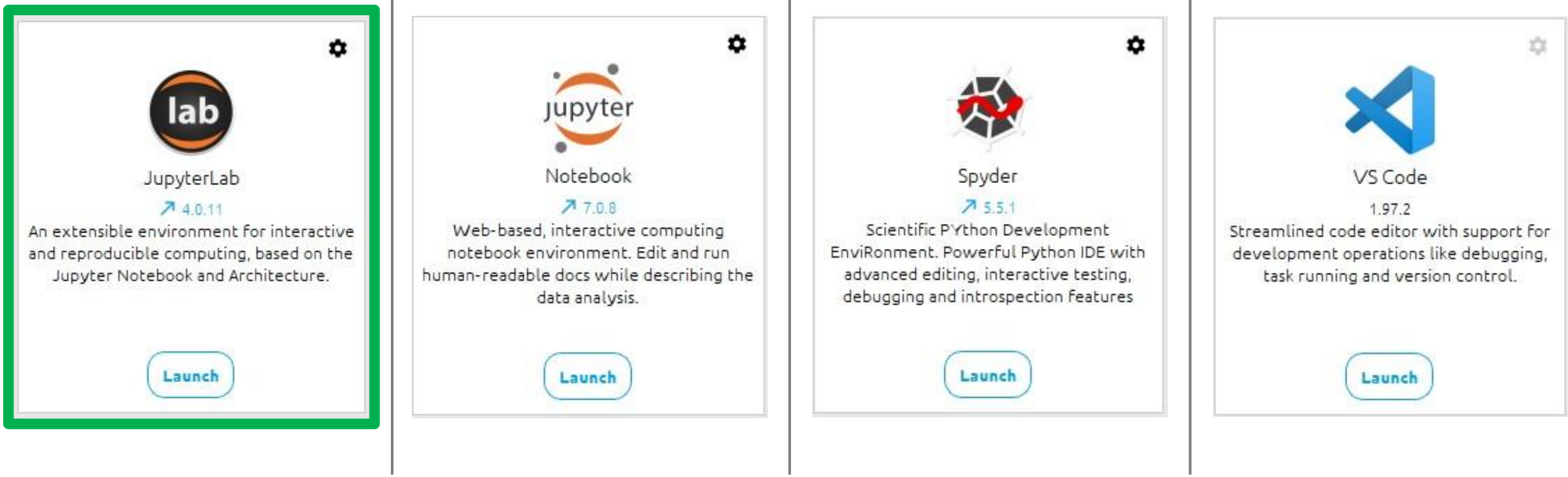
- Statsmodels
- Scikit-learn
- Pytorch
- Keras/Tensorflow

Image Analysis

- Pillow
- Scikit-image
- OpenCV

Created by Andres Patrignani, 2025

Coding environment



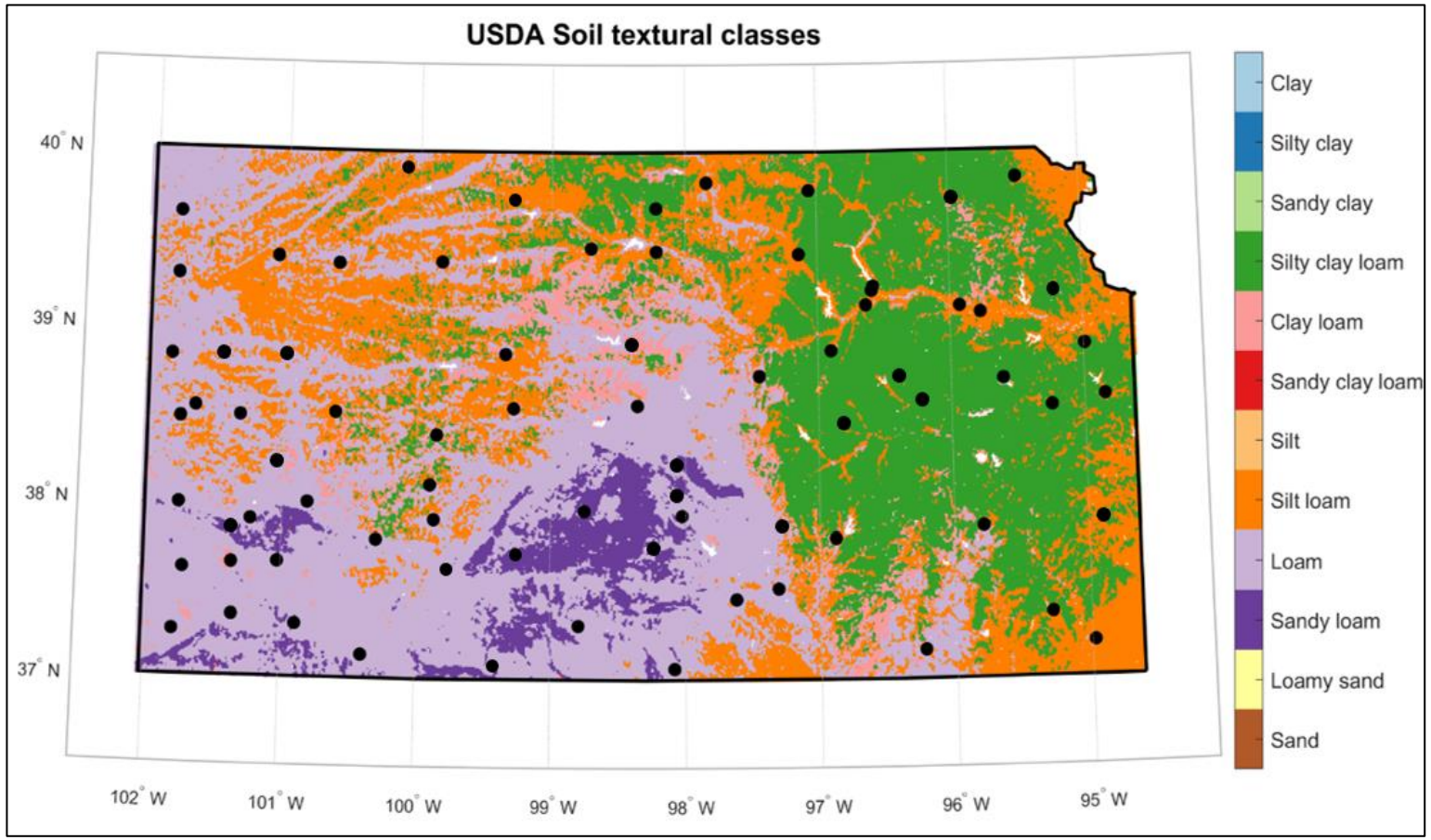
Today you will learn how to:

- open and use a Jupyter Notebook to write Python code
- read tabular weather data from a local spreadsheet
- implement simple computations
- create publication quality figures

You can find more examples related to agronomy and soil science at:

- PyNotes in Agriscience: https://soilwater.github.io/pynotes-agriscience
- PyNotes in Google Earth Engine: https://soilwater.github.io/pynotes-gee
- Precision Ag Workshop 2024: https://github.com/soilwater/precisionag-workshop-2024

Overview of the Kansas Mesonet



Established in 1986 by Kansas State Research and Extension. About 90 in situ stations monitoring essential climate variables, including precipitation, air temperature, humidity, solar radiation, wind speed, soil moisture and soil temperature.

Reference: Patrignani, A., Knapp, M., Redmond, C., & Santos, E. (2020). Technical overview of the Kansas Mesonet. Journal of Atmospheric and Oceanic Technology, 37(12), 2167-2183. https://doi.org/10.1175/JTECH-D-19-0214.1

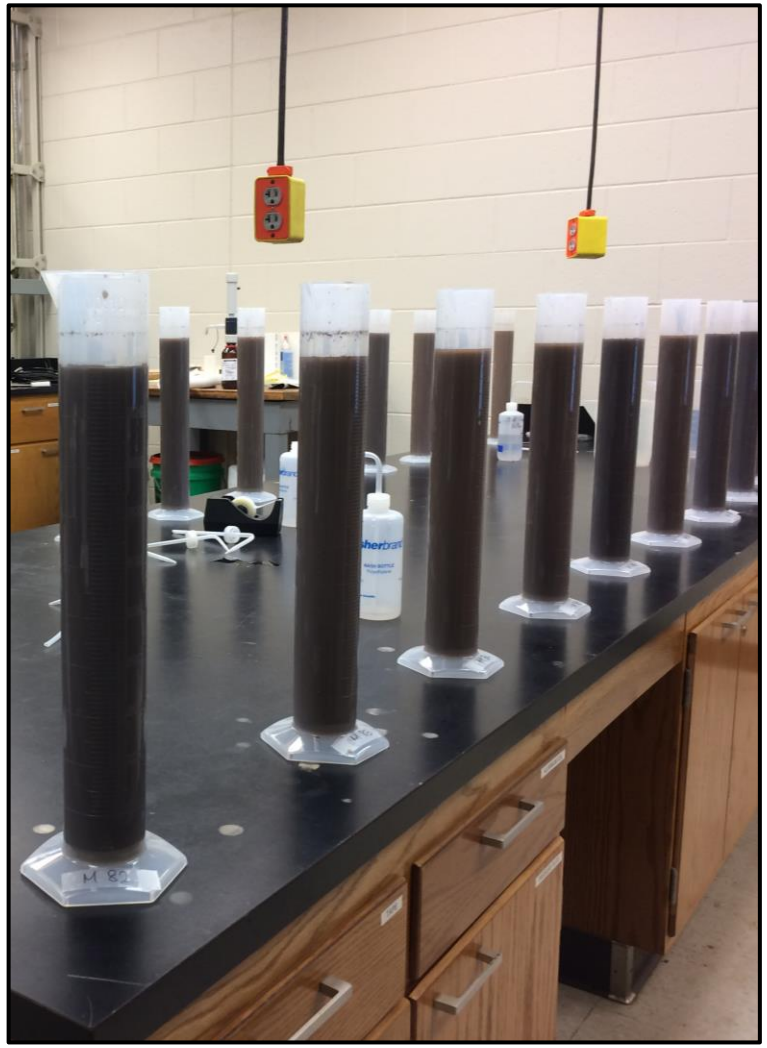
Side view of the Ashland Bottoms station.

Soil Physical Properties Dataset

Sampling near Clay Center, KS

Soil water retention curves

Soil texture



Parker, N., Kluitenberg, G. J., Redmond, C., & Patrignani, A. (2022). A database of soil physical properties for the Kansas Mesonet. *Soil Science Society of America Journal*, *86*(6), 1495-1508. https://doi.org/10.1002/saj2.20465