

Shutay Consulting

Research & Data Sciences



DATA SCIENCE DEMOCRATIZATION

Strategy & Execution

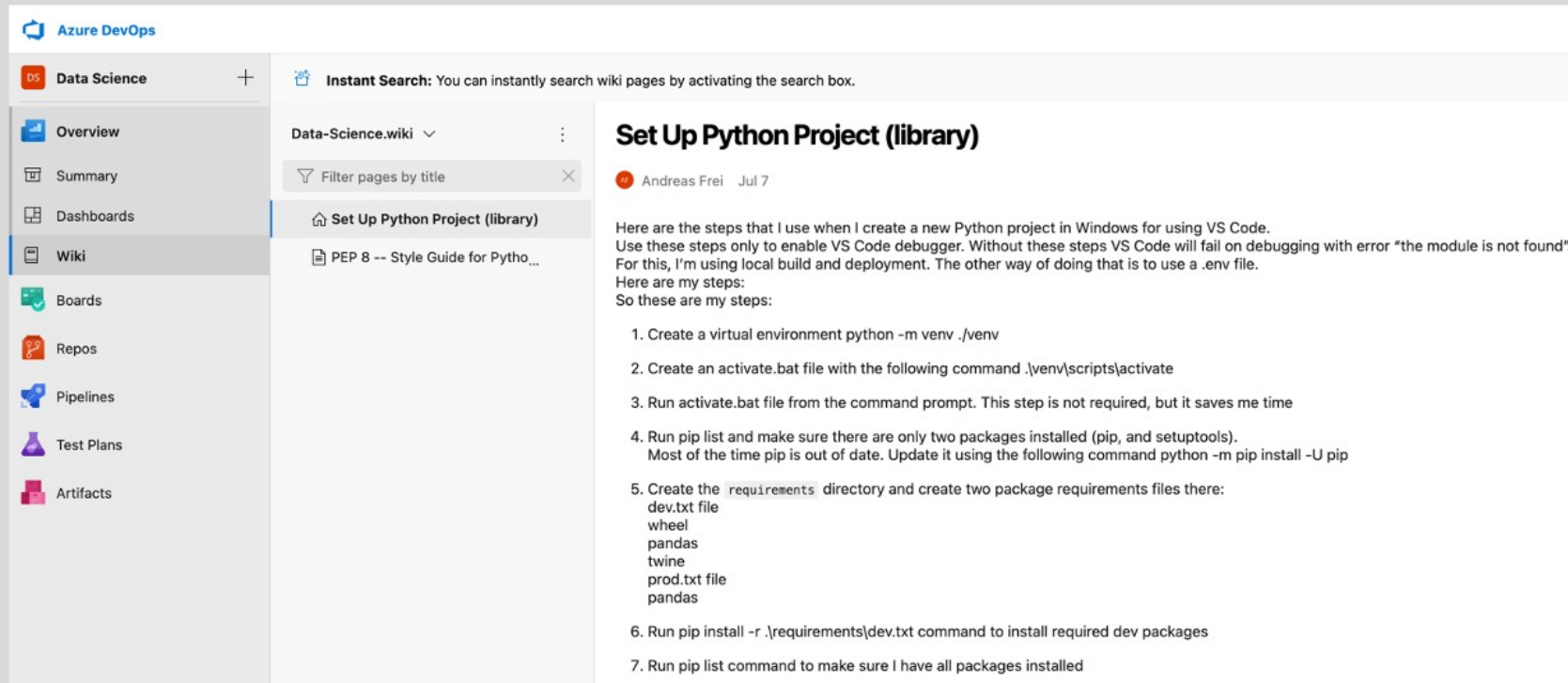
What is data science democratization?

Empowering expert and citizen data scientists and data engineers

- Develop a community of practice
 - Provide resources & training for the data science community
 - Create an organizational knowledge center
 - Provide guidance and oversight for citizen and junior data scientists
- Facilitate data science within the organization
 - Collaborate with Tech to design and build a modern data platform
 - Provide POV on data & tool selection (e.g., ML pipelines and user-friendly platforms)
 - Upskill to create citizen data scientists
- Data science governance
 - External evaluations and adoption of best practices
 - Version control, data drift, model drift
 - Data integrity and security

Community of practice examples

- Shareable repository consisting of searchable research findings, proprietary modules, training materials, sharable Jupyter notebooks, code repositories, templates & protocols outlining best practices, methods, & standards, etc.
- Data science upskilling with for all levels (e.g., Microsoft modules)
- Standardized and regularly updated content for onboarding new data scientists
- Standardized assessments & interview protocols for vetting candidates and assessing current employee capabilities and interest (e.g., Harvey ball assessment against data science taxonomy)
 - Compare to strategic roadmap to identify gaps and build plan around hiring and upskilling



The screenshot shows the Azure DevOps interface for the Data Science Wiki. The left sidebar contains navigation links: Overview, Summary, Dashboards, Wiki (selected), Boards, Repos, Pipelines, Test Plans, and Artifacts. The main content area displays the 'Set Up Python Project (library)' page by Andreas Frei, dated Jul 7. The page includes an 'Instant Search' bar and a list of pages, with 'Set Up Python Project (library)' selected. The page content provides instructions for setting up a Python project in Windows using VS Code, including steps for creating a virtual environment, activating it, and installing dependencies.

Azure DevOps

Data Science +

Instant Search: You can instantly search wiki pages by activating the search box.

Data-Science.wiki ▾

Filter pages by title ✕

Set Up Python Project (library)

PEP 8 -- Style Guide for Python...

Set Up Python Project (library)

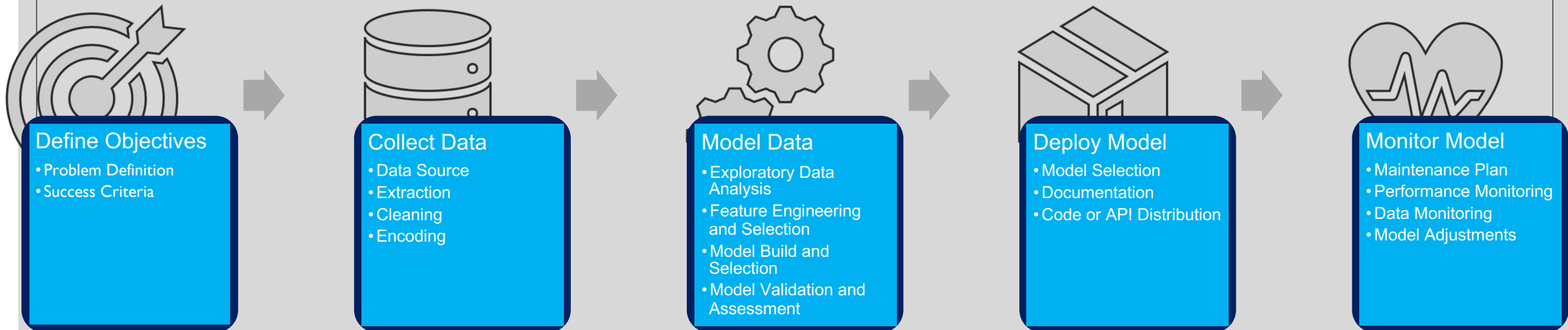
Andreas Frei Jul 7

Here are the steps that I use when I create a new Python project in Windows for using VS Code. Use these steps only to enable VS Code debugger. Without these steps VS Code will fail on debugging with error "the module is not found". For this, I'm using local build and deployment. The other way of doing that is to use a .env file. Here are my steps:
So these are my steps:

1. Create a virtual environment `python -m venv .\venv`
2. Create an activate.bat file with the following command `.\venv\scripts\activate`
3. Run activate.bat file from the command prompt. This step is not required, but it saves me time
4. Run pip list and make sure there are only two packages installed (pip, and setuptools).
Most of the time pip is out of date. Update it using the following command `python -m pip install -U pip`
5. Create the requirements directory and create two package requirements files there:
dev.txt file
wheel
pandas
twine
prod.txt file
pandas
6. Run `pip install -r .\requirements\dev.txt` command to install required dev packages
7. Run pip list command to make sure I have all packages installed

AutoML & automated data pipelines

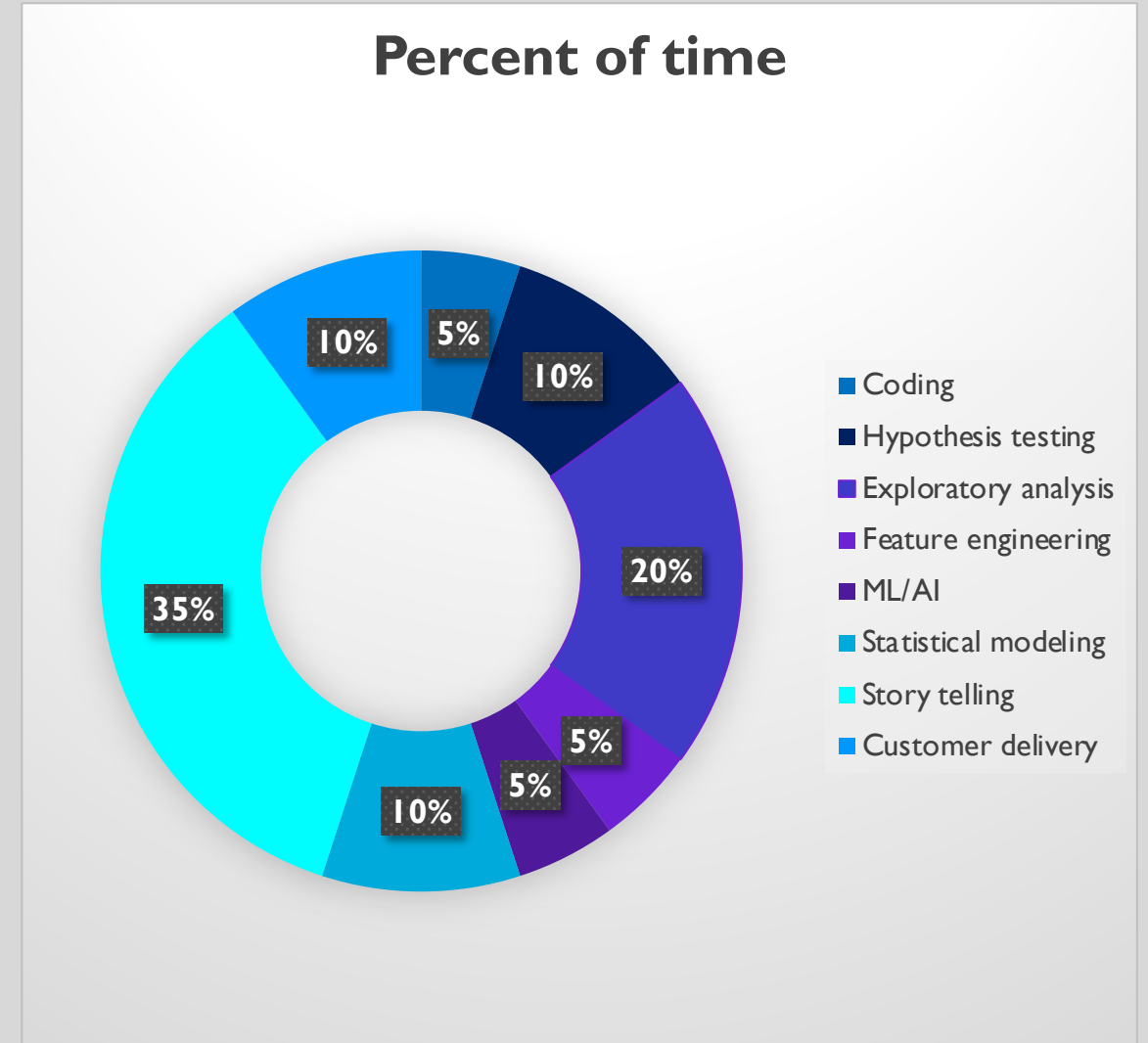
AutoML provides a framework and architecture to support model building and deployment with significantly less effort for those who are strong business analysts, but who lack deep technical expertise (e.g., coding, ML/AI).



The citizen data scientist

Citizen data scientist persona

- Business SME / domain expertise
- Can specify business requirements
- Interest in more advanced techniques
- Lack formal data science training
 - Tend to have low or no coding skills
 - May have some statistical knowledge, but lack ML & AI expertise
- Have strong capability to “story tell”
- Tend to have good “soft skills”



Data science facilitation benefits

- The Citizen Data Scientist is an “upskilled” version of the Business Analyst:
 - Stronger ML modeling and validation skills
 - Stronger coding skills (e.g., SQL & Python)
- Benefits of “upskilling” include:
 - Less reliance on expensive and hard to get data scientists
 - Efficiency gains and quicker turn-arounds
 - More effective story telling leading to improved decision-making
 - Employees feel empowered
 - Potential to improve retention
 - Data scientist relieved from support role
 - Citizen data scientist has more autonomy than business analyst

Data science governance

- Version control for models, data, and code
- Reliability and validity methods and standards for hypothesis testing
- Best practices in data exploration and cleaning
 - E.g., outliers, anomalies, missing values, inconsistent formats, etc.
- Best practices in feature engineering
- Best practices in model training, validation, and monitoring
- Best practices in data visualization
- Coding standards
- Documentation standards

Democratization example plan

- Identify 2-3 business SMEs and begin upskilling using resources such as Microsoft training for Auto ML
- Set up Azure ML workspace and train data science team on using the Azure Auto ML SDK via Microsoft Data science team to work with the business SMEs to leverage the no code/ low code Azure ML Studio
 - Provide guidance, supervision, and consultation
- Begin creating Wikis in our Azure DevOps with data science content and resources
 - Publish data science projects for transparency and so we don't reinvent the wheel
 - Publish methods & standards for conducting POCs, etc. to ensure best practices are employed
 - Publish tutorials and other related content
- Work closely with the business to identify and prioritize our data science roadmap
- Provide opportunities for analysts to leverage the platform for quick hypothesis tests
- Include data engineering in the data science meetings, especially when conducting POC design sessions
- Consider conducting 1-2 hackathons per year
- Provide mentorship and consultation to employees regarding the science and application of data science
- Consider having an external evaluation of our maturity 6-12 months out