Automatic Generation of Python Programs for Creating SDTM Datasets
The views and opinions expressed in this presentation are those of the author(s) and do not necessarily reflect the official policy or position of the organization where the author belongs.
The main topic of this presentation is automatic generation of Python programs for creating CDISC SDTM datasets.

Before diving into the main topic, this presentation will explain:
- Why Python is a good option for data science in drug development
- How programming practice is changing with Generative AI
- How Digital Data Flow is impacting approach to data processing
Background and Trends
A study conducted in 2020 for 1,000 data scientists in the United States showed that **Python** was the most popular language among SAS, R and Python, in any industry.

It also shows that Python is more popular in younger generation (with less years of experience in data science).

Reference: Burtch Works LLC "2020 SAS, R, or Python Survey Results: Which Tool do Data Scientists & Analytics Pros Prefer?"
Generative AI

- Generative AI (more precisely LLM: Large Language Model) is significantly changing the way how programs are developed.
  - Code Generation
  - Code Completion
  - Test Code Generation and more

- ChatGPT, the most famous LLM-based generative AI, is built with Python and best at generating Python codes among other languages.

- Another famous LLM-based generative AI, Copilot, is trained on all languages that appear in public repositories (see the top 50 pull requests in Q3 of 2023 on the left).

<table>
<thead>
<tr>
<th># Ranking</th>
<th>Programming Language</th>
<th>Percentage (YoY Change)</th>
<th>YoY Trend</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Python</td>
<td>17.715% (+0.55%)</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Java</td>
<td>11.086% (-0.172%)</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Go</td>
<td>10.817% (+0.950%)</td>
<td>✓</td>
</tr>
<tr>
<td>4</td>
<td>JavaScript</td>
<td>10.065% (+0.892%)</td>
<td>✓</td>
</tr>
<tr>
<td>5</td>
<td>C++</td>
<td>9.712% (-0.348%)</td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>R</td>
<td>0.056%</td>
<td></td>
</tr>
</tbody>
</table>

Can ChatGPT create programs automatically? Please explain the language you are best at along with its reason.

Yes, ChatGPT can automatically create programs. Python is the language that I'm most proficient in, and there are several reasons for this:

- Another famous LLM-based generative AI, Copilot, is trained on all languages that appear in public repositories (see the top 50 pull requests in Q3 of 2023 on the left).
Advanced data analysis is another powerful feature of ChatGPT that allows uploading files, write and execute Python codes, and create graphs with instructions in natural language.
Digital Data Flow (DDF) Overview

- Digital Data Flow (DDF) is an initiative by TransCelerate that aims to create digitized study protocol and automate creation of study assets.

- CDISC is collaborating with TransCelerate to develop a standard model for Study Definitions Repository as part of their journey to achieve end-to-end automation.

Note: The content in this slide has been created by summarizing contents on the TransCelerate web site at the author's own discretion.
DDF Achieves More SDTM Automation

(1) Activities/BCs are populated via Study Builder or read from Study Protocol.

(2) Activities with BCs are mapped to forms when a draft eCRF Spec is created from SDR.

(3) Activities with BCs are mapped to SDTM domains when a draft SDTM Spec is created from SDR.

(4) SDR knows which BCs are referenced from which forms/domains.

Digital Data Flow

Metadata Repository

© 2023 Fujitsu Limited.
Do they conflict or complement to each other?

Digital Data Flow

- Protocol
- CRF Spec
- SDTM Spec
- SDTM Datasets
- Statistical Analysis Plan
- Mock Shells
- ADaM Spec
- ADaM Datasets
- TLFs
- Study Report
A well-known example of R-based SDTM automation

Up to 22 Reusable Algorithms

<table>
<thead>
<tr>
<th>Only as Algorithms</th>
<th>Only as Sub-Algorithms</th>
<th>Algorithm &amp; Sub-Algorithms</th>
</tr>
</thead>
<tbody>
<tr>
<td>03_AE_AEREL</td>
<td>11_MERGE</td>
<td>01_ASSIGN_NO_CT</td>
</tr>
<tr>
<td>07_DATASET_LEVEL</td>
<td>18_REMOVE_DUP</td>
<td>02_ASSIGN_CT</td>
</tr>
<tr>
<td>09_IF_THEN_ELSE</td>
<td>19_GROUP_BY</td>
<td>05_HARDCODE_CT</td>
</tr>
<tr>
<td>17_WHODRUG_FA</td>
<td>20_NEED_USER_INPUT</td>
<td>06_HARDCODE_NO_CT</td>
</tr>
<tr>
<td>13_RELREC</td>
<td>08_NOTSUBMITTED</td>
<td></td>
</tr>
<tr>
<td>14_RELREC_CONDITION</td>
<td>15_MULTIPLE_RESPONSES</td>
<td></td>
</tr>
<tr>
<td>21_NONCRF_LAB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>22_NONCRF_PKC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>23_PAIRED_VARS</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Reference: F. Hoffmann-La Roche AG "OAK Garden - SDTM Automation The flourishing Data Transformation Engine"
Similar components and processes to OAK Garden, but more variety of methods.

- Metadata Repository
- Transformation Engine
- Trial Design
- SDTM Spec
- SDTM Spec (Executable)
- Define-XML
- SDTMDatasets
- Annotated CRF

Standards and Terminologies
CRF and eDT Specs
Biomedical Concepts

Topic Methods:
- assignTopic
- copy
- copyDesign
- copyUnique
- extrt
- joinUnique
- lookupConcept
- lookupConceptNormalized
- lookupDomainREGEX
- n1mapTopic
- se
- sv

Non-Topic Methods:
- bfi
- calc
- coeval
- comment
- concat
- concatDate
- convertToStd
- copy
- copyDesign
- copyIf
- copyToSUPP
- …and more

55 Methods and 286 Use Cases*

*: Pre-configured method calls
The 55 methods are being implemented in Python.
Generative AI is multipurpose, and thus your ideas matter to make it valuable.

Python sits in a unique position in the AI era.

Digital Data Flow is built on CDISC standards and has good familiarity with data processing.

After all, there are many chances of automation in your job with the rise of LLM-based generative AI and Digital Data Flow.
Which part of the process below can be further automated?
My Answer

- Which part of the process below can be further automated?
Thank you

Please enjoy chatting in our booth to find more information.