WHATSNEXT: Guidance-enriched Exploratory Data Analysis with Interactive, Low-Code Notebooks

Chen Chen, Jane Hoffswell, Shunan Guo, Ryan Rossi,

Yeuk-Yin Chan, Fan Du, Eunyee Koh, Zhicheng Liu

DEPARTMENT OF COMPUTER SCIENCE

Adobe

Computation Notebooks are Popular for EDA

Data Science by using pandas liabrary

In [2]:
```
import pandas as pd
import numpy as np
import matplotlib as mpl
```

In [2]:
```
df = pd.DataFrame([[38.0, 2.0, 18.0, 22.0, 21, np.nan],[19, 439, 6, 452, 226,232]],
                  index=pd.Index(['Tumour (Positive)', 'Non-Tumour (Negative)'], name='Actual Label:'),
                  columns=pd.MultiIndex.from_product([['Decision Tree', 'Regression', 'Random'],['Tumour', 'Non-Tumour']], names=['Model:', 'Predicted:']))
df.style
```

Out[2]:

Model:	Decision Tree		Regression		Random	
Predicted:	Tumour	Non-Tumour	Tumour	Non-Tumour	Tumour	Non-Tumour
Actual Label:						
Tumour (Positive)	38.000000	2.000000	18.000000	22.000000	21	nan
Non-Tumour (Negative)	19.000000	439.000000	6.000000	452.000000	226	232.000000

In [3]:
```
weather_df = pd.DataFrame(np.random.rand(10,2)*5,
                          index=pd.date_range(start="2021-01-01", periods=10),
                          columns=["Tokyo", "Beijing"])

def rain_condition(v):
    if v < 1.75:
        return "Dry"
    elif v < 2.75:
        return "Rain"
    return "Heavy Rain"

def make_pretty(styler):
    styler.set_caption("Weather Conditions")
    styler.format(rain_condition)
    styler.format_index(lambda v: v.strftime("%A"))
    styler.background_gradient(axis=None, vmin=1, vmax=5, cmap="YlGnBu")
    return styler

weather_df
```

Out[3]:

	Tokyo	Beijing
2021-01-01	3.621969	4.508378
2021-01-02	0.109078	3.062590
2021-01-03	0.493561	1.459806
2021-01-04	3.307517	0.808751
2021-01-05	3.709100	2.608226
2021-01-06	4.165616	0.519589
2021-01-07	4.695969	1.588812
2021-01-08	4.540735	0.477025
2021-01-09	3.149028	3.671971
2021-01-10	3.041754	0.944977

In [4]:
```
weather_df.loc["2021-01-04":"2021-01-08"].style.pipe(make_pretty)
```

Out[4]:

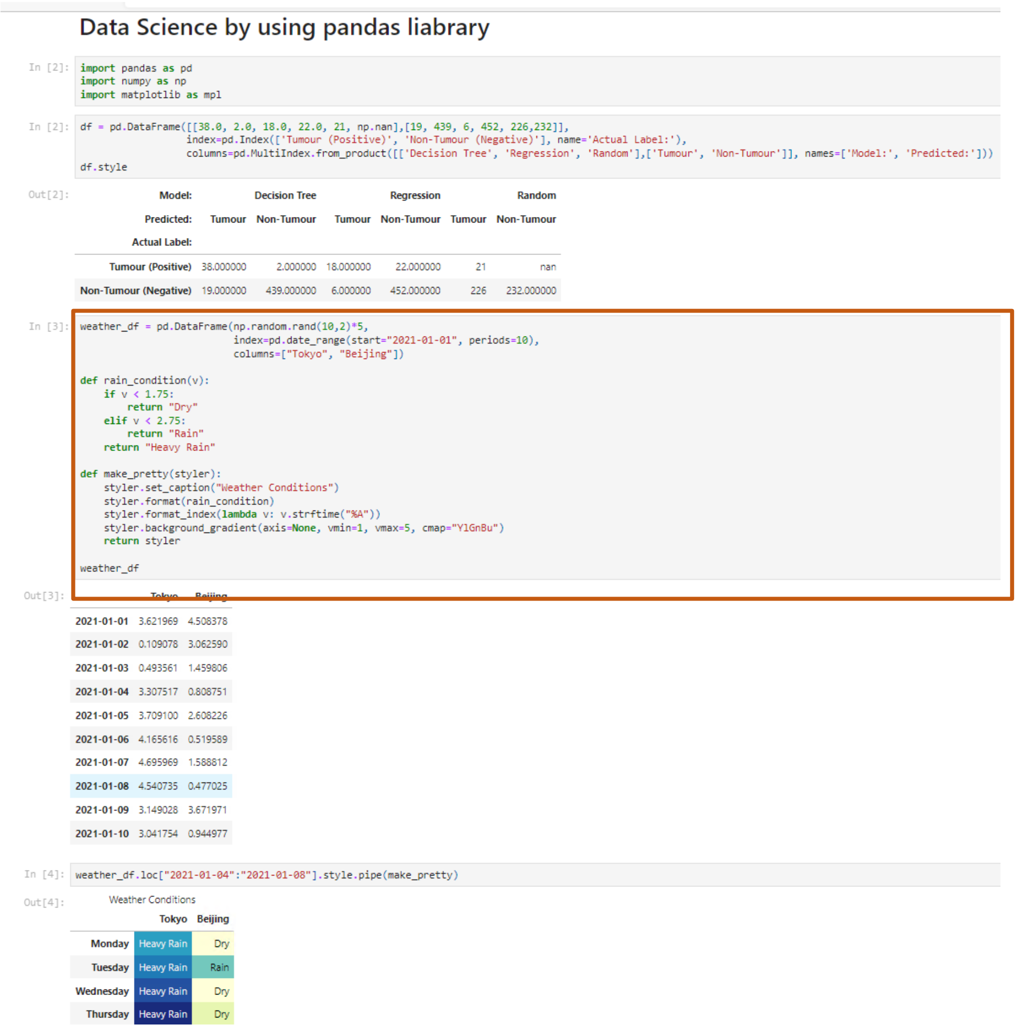
Weather Conditions

	Tokyo	Beijing
Monday	Heavy Rain	Dry
Tuesday	Heavy Rain	Rain
Wednesday	Heavy Rain	Dry
Thursday	Heavy Rain	Dry

https://openexchange.intersystems.com/package/iris-python-apps

- Multimodality
  - text, code, visualizations, and tables
- Flexibility
  - changes reflected in real-time

Computation Notebooks Limitations



Data Science by using pandas liabrary

In [2]:
```python
import pandas as pd
import numpy as np
import matplotlib as mpl
```

In [2]:
```python
df = pd.DataFrame([[38.0, 2.0, 18.0, 22.0, 21, np.nan],[19, 439, 6, 452, 226,232]],
                  index=pd.Index(['Tumour (Positive)', 'Non-Tumour (Negative)'], name='Actual Label:'),
                  columns=pd.MultiIndex.from_product([['Decision Tree', 'Regression', 'Random'],['Tumour', 'Non-Tumour']], names=['Model:', 'Predicted:']))
df.style
```

Out[2]:

Model:	Decision Tree		Regression		Random	
Predicted:	Tumour	Non-Tumour	Tumour	Non-Tumour	Tumour	Non-Tumour
Actual Label:						
Tumour (Positive)	38.000000	2.000000	18.000000	22.000000	21	nan
Non-Tumour (Negative)	19.000000	439.000000	6.000000	452.000000	226	232.000000

In [3]:
```python
weather_df = pd.DataFrame(np.random.rand(10,2)*5,
                          index=pd.date_range(start="2021-01-01", periods=10),
                          columns=["Tokyo", "Beijing"])

def rain_condition(v):
    if v < 1.75:
        return "Dry"
    elif v < 2.75:
        return "Rain"
    return "Heavy Rain"

def make_pretty(styler):
    styler.set_caption("Weather Conditions")
    styler.format(rain_condition)
    styler.format_index(lambda v: v.strftime("%A"))
    styler.background_gradient(axis=None, vmin=1, vmax=5, cmap="YlGnBu")
    return styler

weather_df
```

Out[3]:

	Tokyo	Beijing
2021-01-01	3.621969	4.508378
2021-01-02	0.109078	3.062590
2021-01-03	0.493561	1.459806
2021-01-04	3.307517	0.808751
2021-01-05	3.709100	2.608226
2021-01-06	4.165616	0.519589
2021-01-07	4.695969	1.588812
2021-01-08	4.540735	0.477025
2021-01-09	3.149028	3.671971
2021-01-10	3.041754	0.944977

In [4]:
```python
weather_df.loc["2021-01-04":"2021-01-08"].style.pipe(make_pretty)
```

Out[4]:

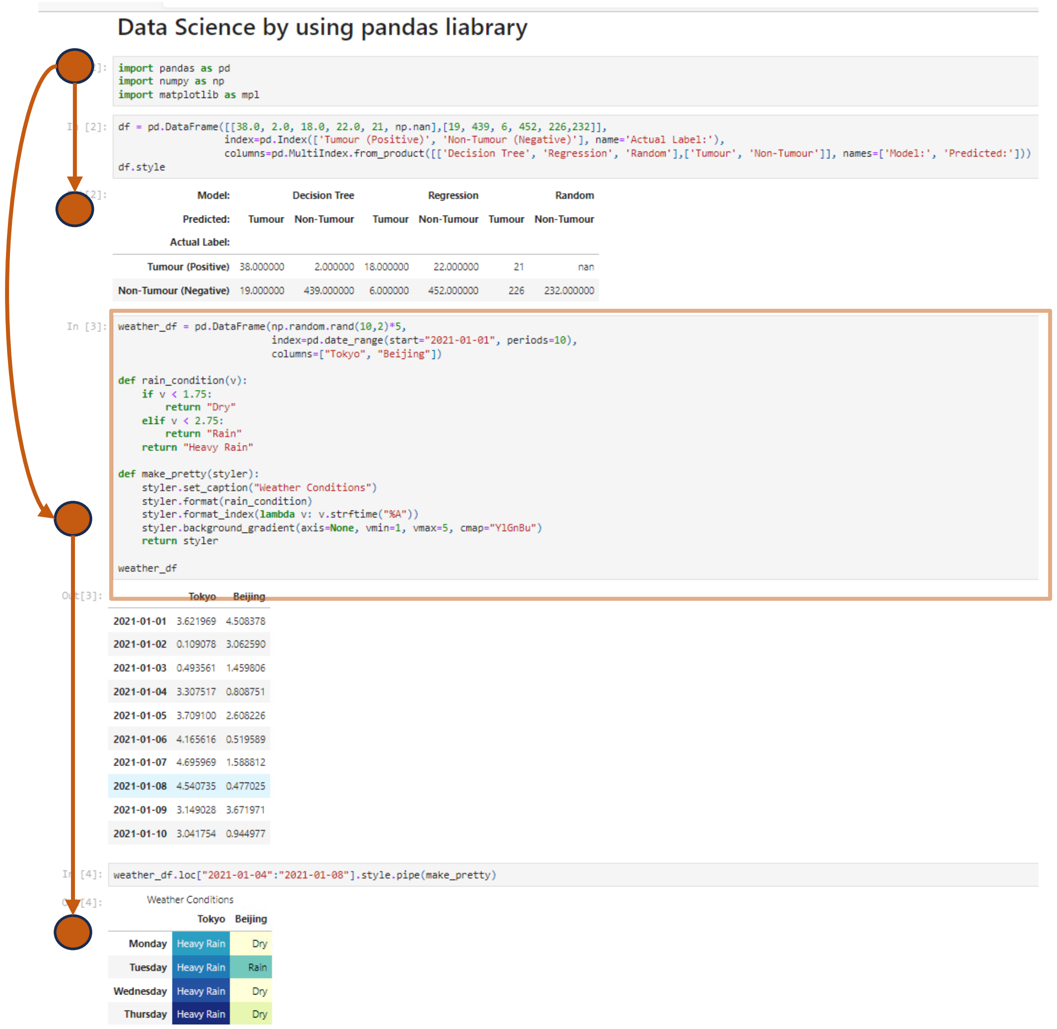
Weather Conditions

	Tokyo	Beijing
Monday	Heavy Rain	Dry
Tuesday	Heavy Rain	Rain
Wednesday	Heavy Rain	Dry
Thursday	Heavy Rain	Dry

https://openexchange.intersystems.com/package/iris-python-apps

- Notebooks' **code-reliance** limits its usage by inexperienced programmers

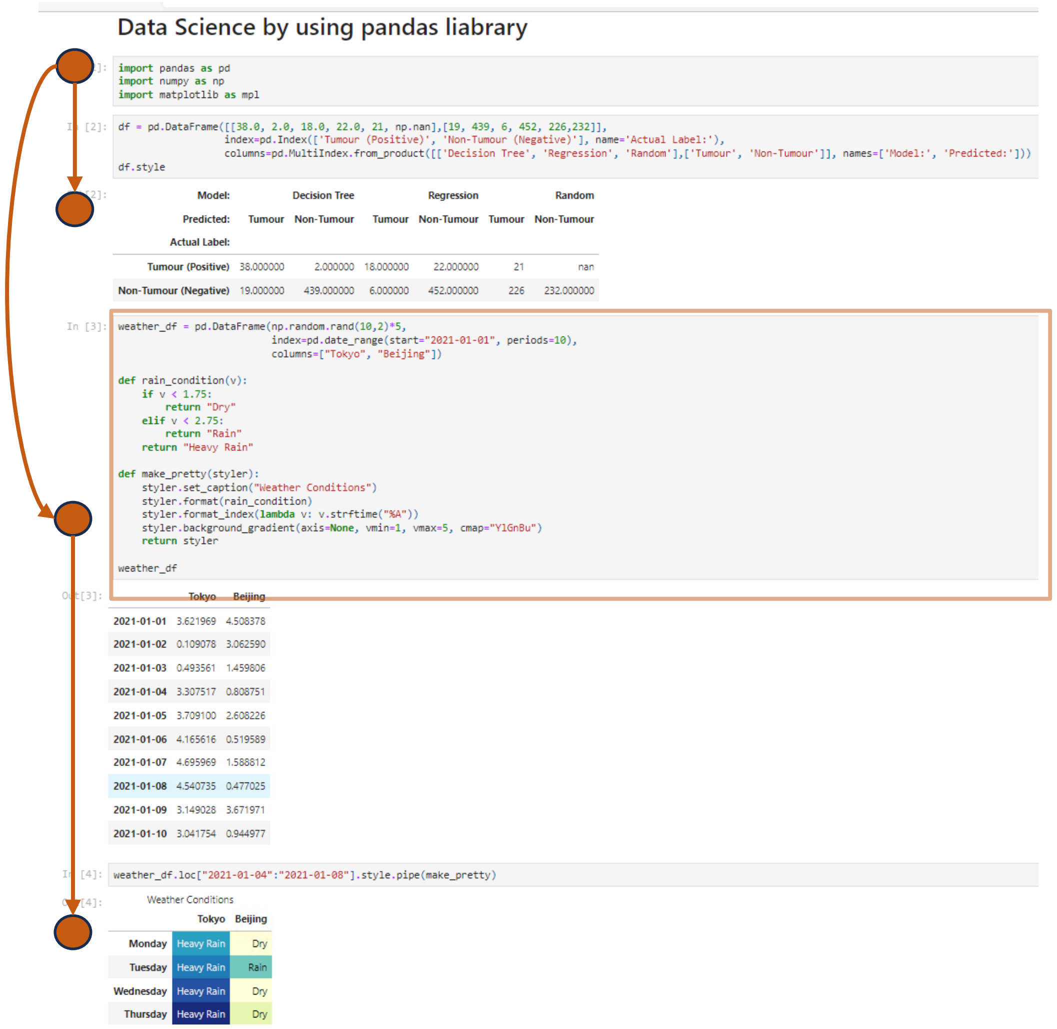
Computation Notebooks Limitations



https://openexchange.intersystems.com/package/iris-python-apps

- Notebooks' **code-reliance** limits its usage by inexperienced programmers
- Notebooks present **a single, interleaved thread**, which may not capture the user's analysis flow

Computation Notebooks Limitations

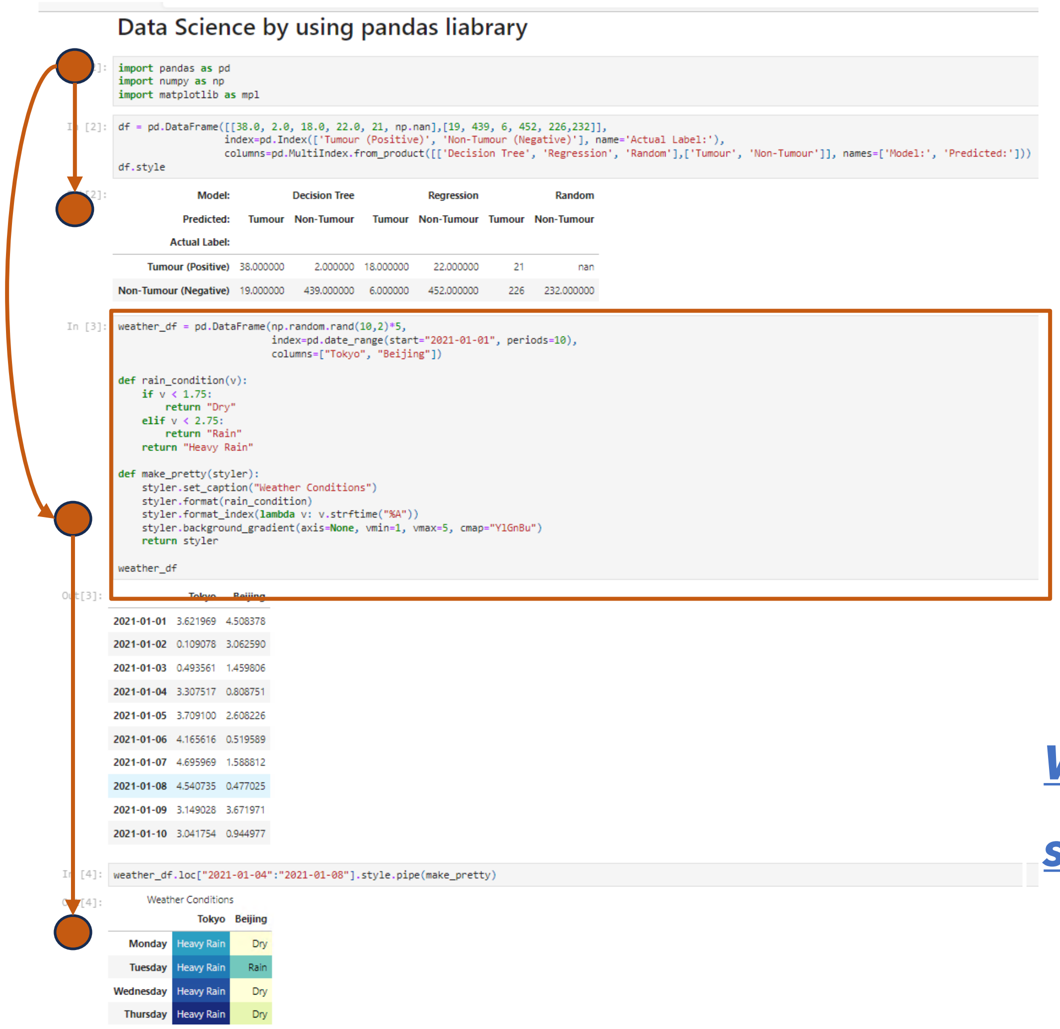


https://openexchange.intersystems.com/package/iris-python-apps

- Notebooks' **code-reliance** limits its usage by inexperienced programmers
- Notebooks present **a single, interleaved thread**, which may not capture the user's analysis flow

much worse when you have many cells…

Computation Notebooks Limitations



https://openexchange.intersystems.com/package/iris-python-apps

- Notebooks' **code-reliance** limits its usage by inexperienced programmers
- Notebooks present **a single, interleaved thread**, which may not capture the user's analysis flow

We aim at an interactive notebook framework to support efficient low-code exploratory data analysis

Design Goals

- Notebooks’ **code-reliance** limits its usage by inexperienced programmers
  - **DG1: Low-code:** support users with varying levels of programming expertise.
  - **DG2: Insight-driven:** help quickly synthesize compound data insights.

Design Goals

- Notebooks' **code-reliance** limits its usage by inexperienced programmers
  - **DG1: Low-code:** support users with varying levels of programming expertise.
  - **DG2: Insight-driven:** help quickly synthesize compound data insights.
- Notebooks present **a single, interleaved thread**, which may not capture the user's analysis flow
  - **DG3: History:** help recall and navigate efficiently with visual cues and interactions.
  - **DG4: Structure:** reveal the analytic dependencies between cells.

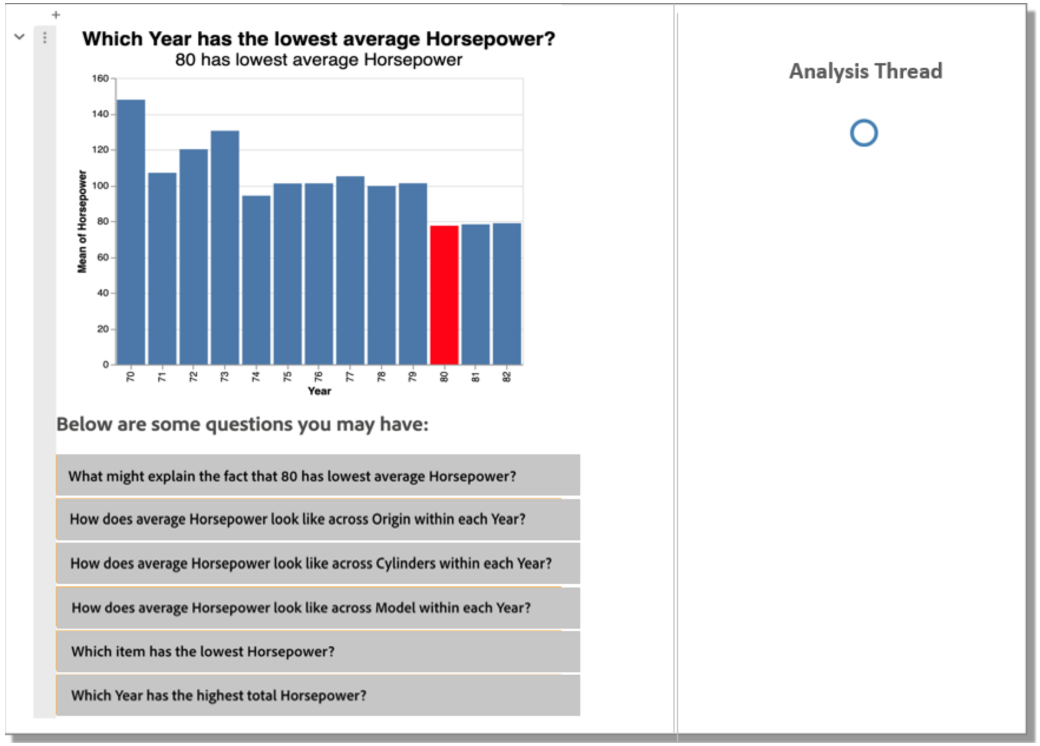
WhatsNext:
Guidance-enriched EDA with Interactive, Low-Code Notebooks

We use the following car sales dataset to present a usage scenario of WhatsNext.

Model	MPG	Cylinders	Displacement	Horsepower	Weight	Acceleration	Year	Origin
volkswagen 1131	26	4	97	46	1835	20.5	70	Europe
volkswagen supe	26	4	97	46	1950	21	73	Europe
volkswagen rabb	43.1	4	90	48	1985	21.5	78	Europe
vw rabbit c (dies	44.3	4	90	48	2085	21.7	80	Europe
vw dasher (diese	43.4	4	90	48	2335	23.7	80	Europe
fiat 128	29	4	68	49	1867	19.5	73	Europe
toyota corona	31	4	76	52	1649	16.5	74	Japan
chevrolet chevet	29	4	85	52	2035	22.2	76	US
mazda glc delux	32.8	4	78	52	1985	19.4	78	Japan

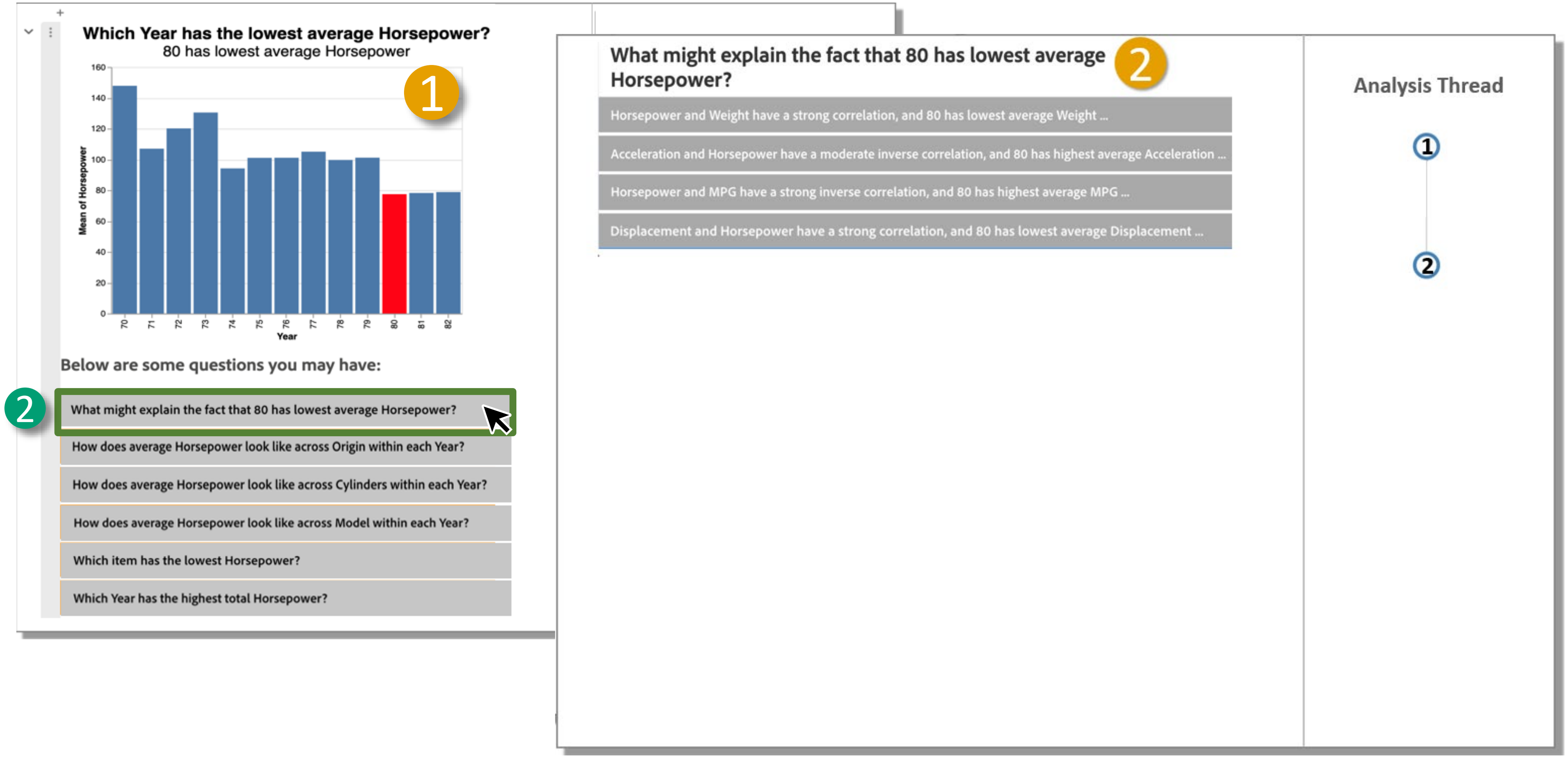
https://goo.gl/9G1egz

WhatsNext: A Use Case



DG1: Low-code: support users with varying levels of programming expertise.

WhatsNext: A Use Case

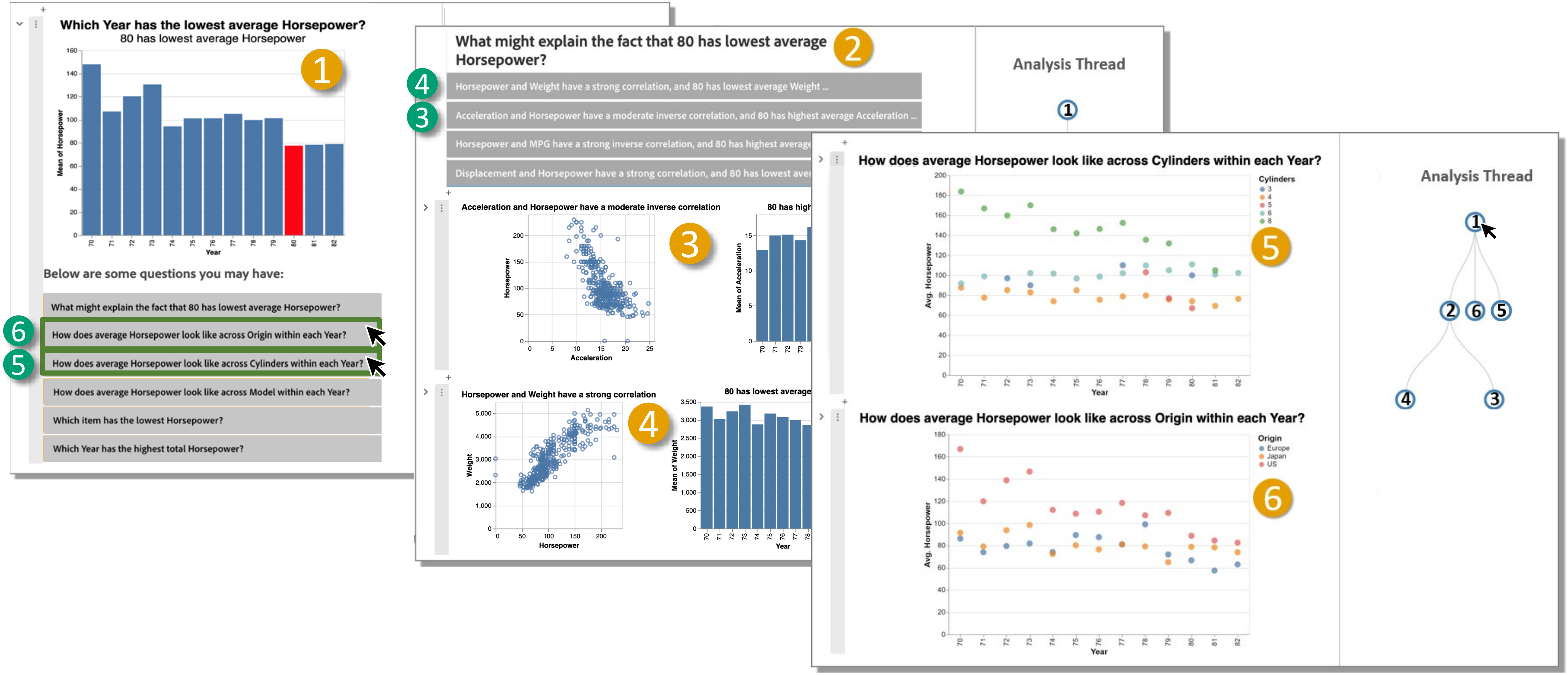


DG2: Insight-driven: help quickly synthesize compound data insights.

WhatsNext: A Use Case

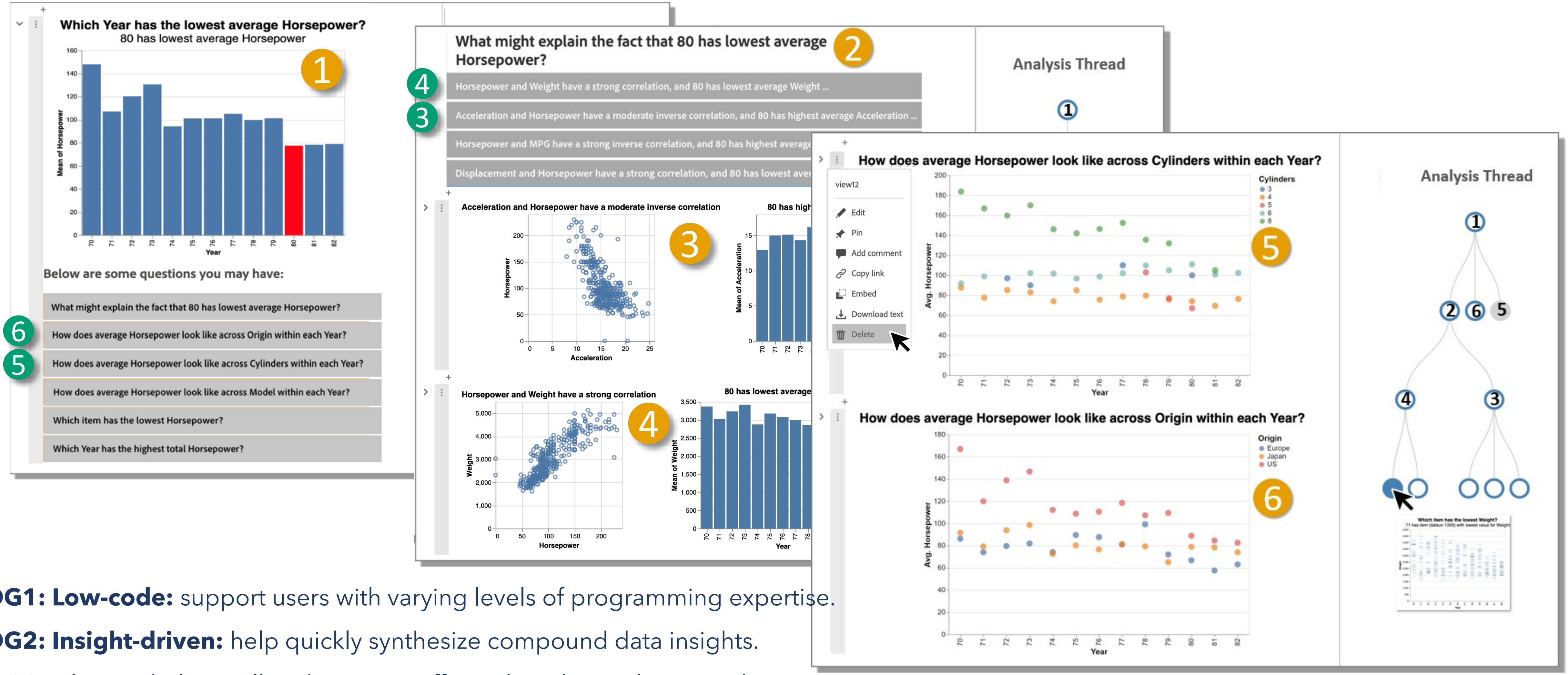
DG4: Structure: reveal the analytic dependencies between cells.

WhatsNext: A Use Case



DG3: History: help recall and navigate efficiently with visual cues and interactions.

WhatsNext: A Use Case



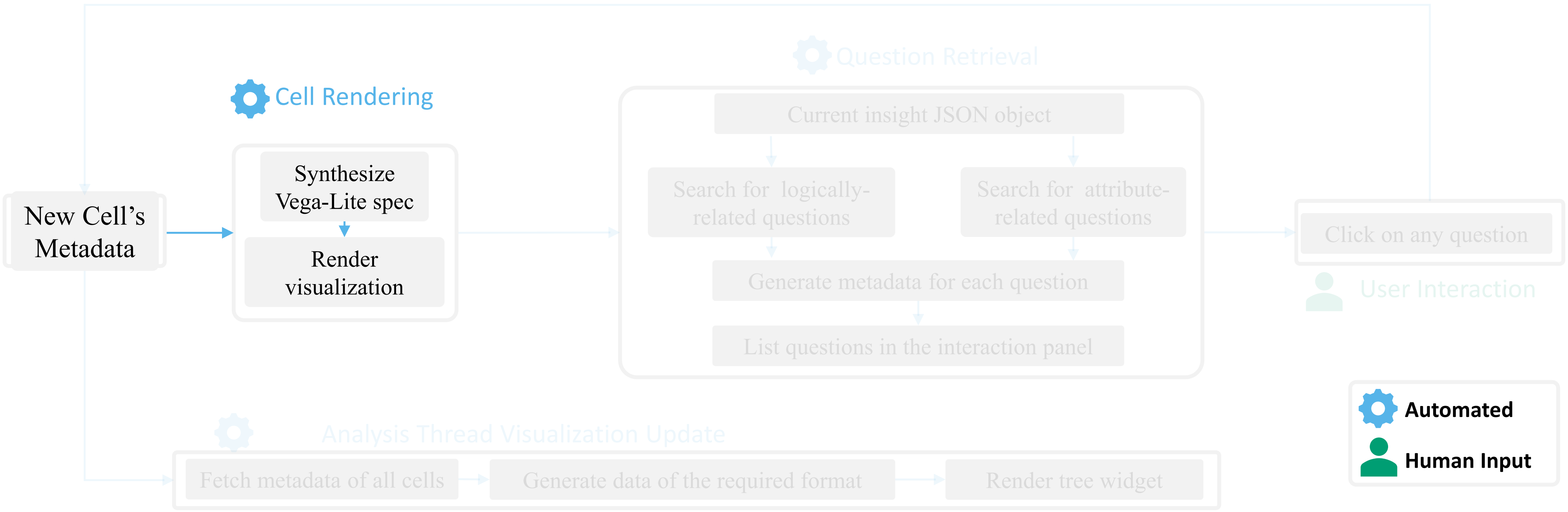
DG1: Low-code: support users with varying levels of programming expertise.

DG2: Insight-driven: help quickly synthesize compound data insights.

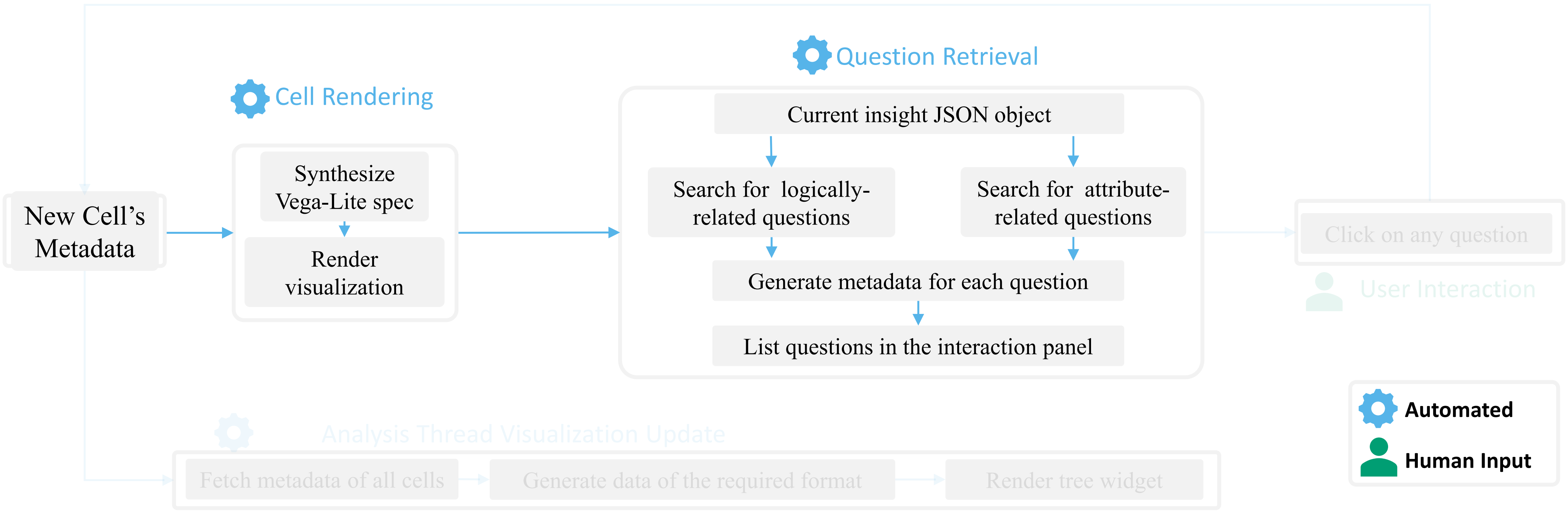
DG3: History: help recall and navigate efficiently with visual cues and interactions.

DG4: Structure: reveal the analytic dependencies between cells.

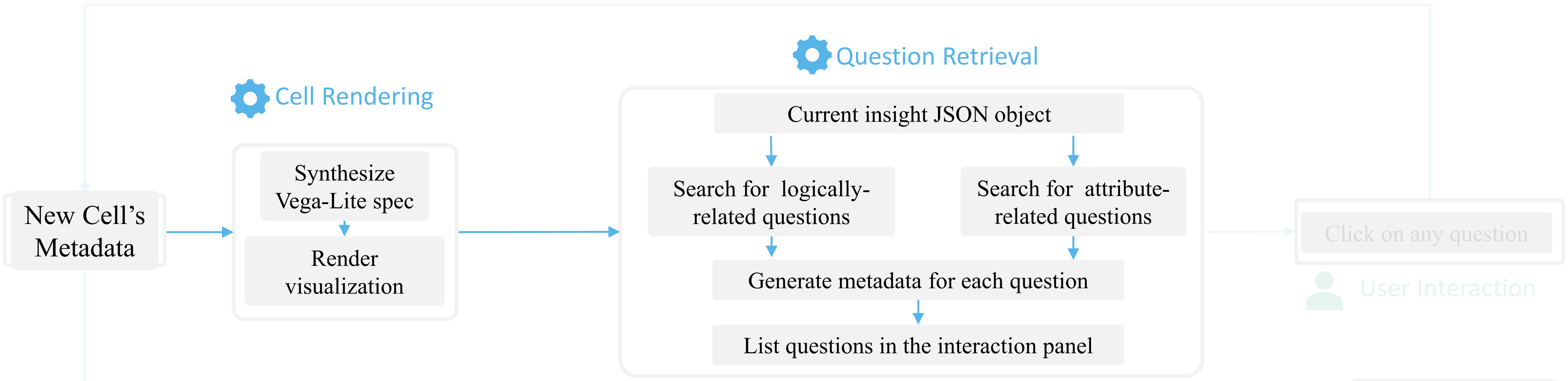
WhatsNext: Pipeline



WhatsNext: Pipeline

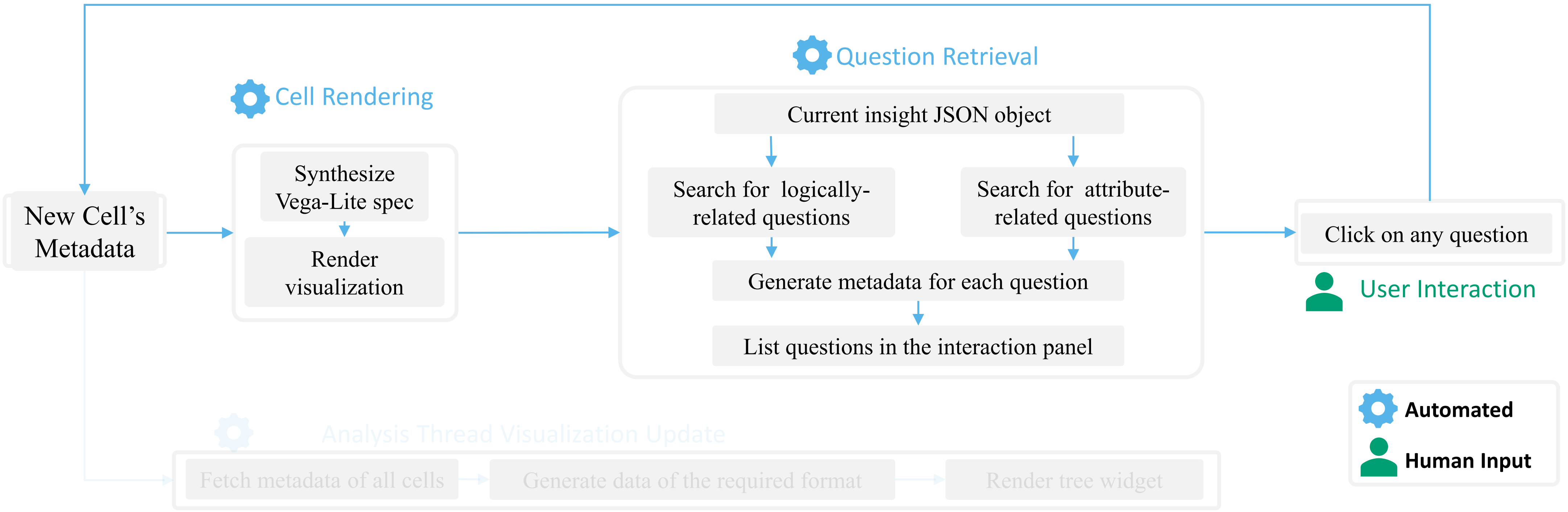


WhatsNext: Pipeline

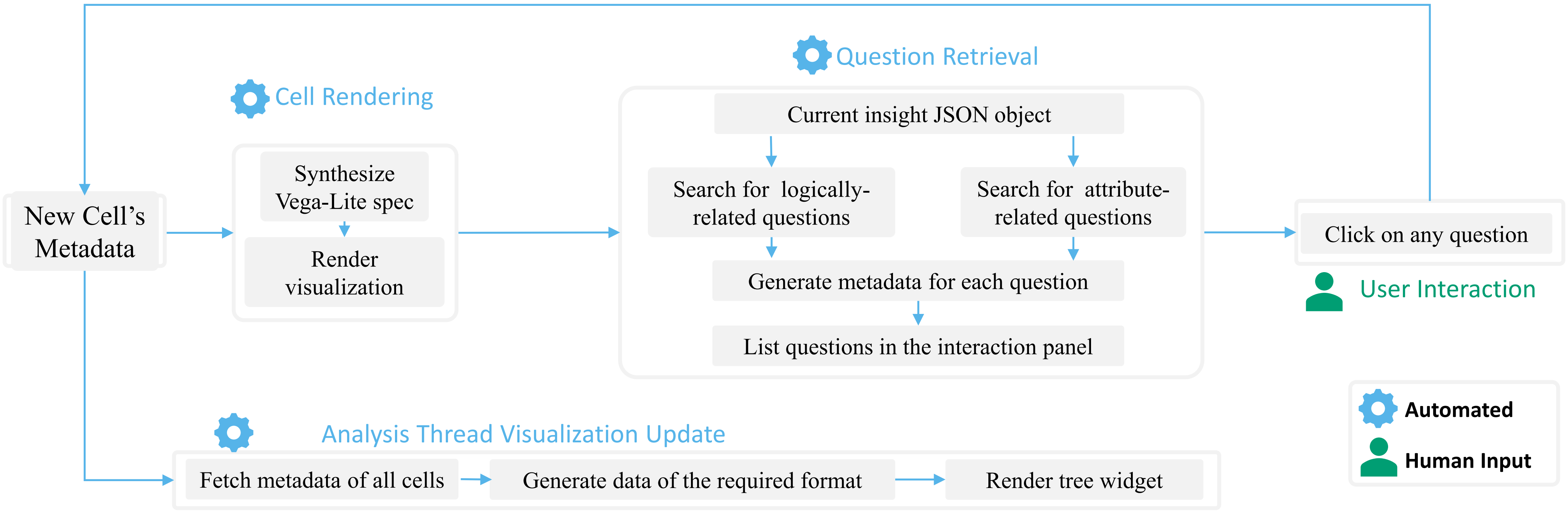


Given Insight Example	Logically-related Insight Example	Converted Question
[Ext] *"Cars from the year 1980 have the* ***lowest*** *average Weight"*	[Ext] *"Cars from the year 1980 have the* ***lowest*** *average Horsepower"* + [Cor] *"Horsepower and Weight have a strong* ***correlation****"*	*"Why do cars from the year 1980 have the* ***lowest*** *average Weight?"*
	[Ano] *"There are three* ***anomalies*** *regarding Weight in the year 1980"*	
	[Ext] *"Cars from Japan in the year 1980 have the* ***lowest*** *average weight"*	
[Cor] *"Horsepower and Weight have a strong* ***correlation****"*	[Cor] *"Weight and Displacement have a strong* ***correlation****"* + [Cor] *"Horsepower and Displacement have a strong* ***correlation****"*	*"Why do Horsepower and Weight have a strong* ***correlation****?"*
[Ano] *"The car 'renault 18i' appears to be an* ***outlier*** *regarding Horsepower"*	[Dis] *"Most values for Horsepower are in the* ***range*** *[75.0, 125.0]"*	*"What is the major value* ***range*** *of Horsepower?"*
[Dis] *"Most values for Horsepower are in the* ***range*** *[75.0, 125.0]"*	[Ano] *"The car 'renault 18i' seems an* ***outlier*** *regarding Horsepower"*	*"What are potential* ***outliers*** *regarding Horsepower?"*
	[Dis] *"Most values for Horsepower in 1980 are in the* ***range*** *[70.0, 120.0]"*	*"What is the* ***distribution*** *of Horsepower in the year 1980?"*

WhatsNext: Pipeline



WhatsNext: Pipeline



Hover: to display a screenshot tooltip
Click: go to the corresponding cell

Summarization

We introduce WHATSNEXT, an interactive notebook environment for low-code data exploration that

- augments a standard notebook cell with a **no-code interaction panel** showing recommended follow-up analysis questions or actions;
- utilize a set of **insight-driven heuristics** to synthesize follow-up questions to help reveal / explore / integrate data insights;
- visualizes the **analysis hierarchy** to help users trace the history of diverging analysis threads.

Discussion & Future Work

- Improve connections between the notebook and analysis structure.
  - More **annotations** to show hidden features, such as insight type / recommendation type / instructions…

Discussion & Future Work

- Improve connections between the notebook and analysis structure.
  - More **annotations** to show hidden features, such as insight type / recommendation type / instructions…
- Support user-customized EDA trajectories.
  - Support **user-specified analysis** intents
  - Maintain ecological synchronization

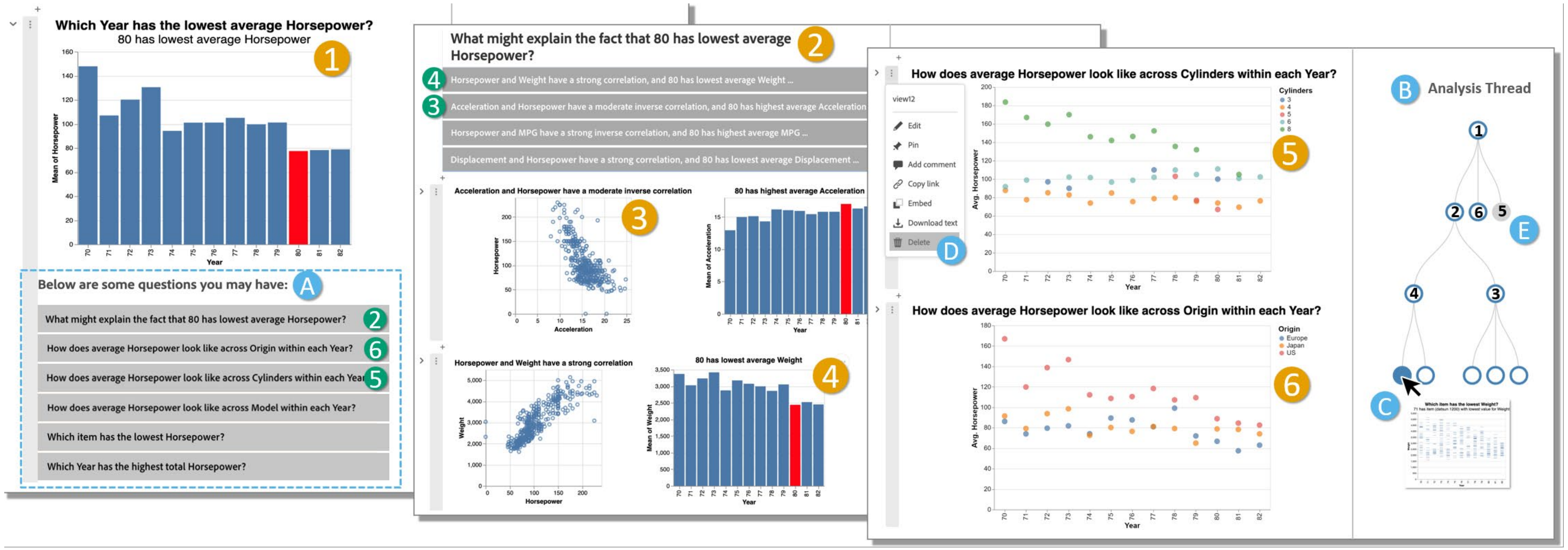
Discussion & Future Work

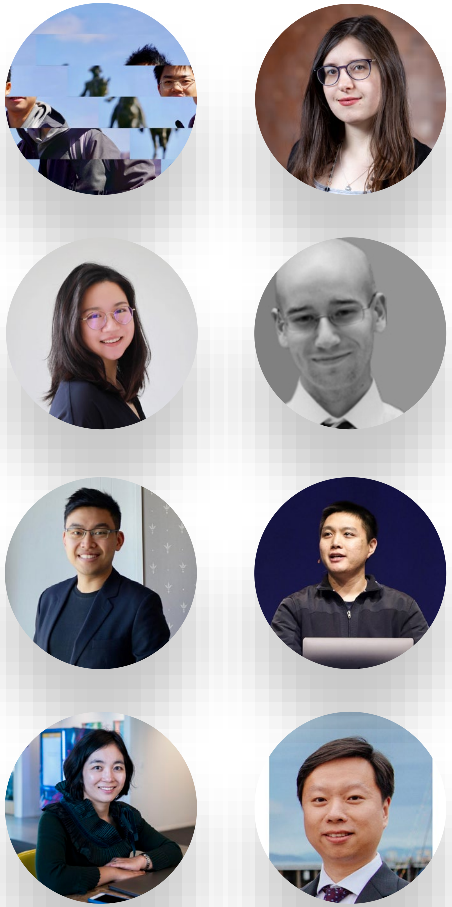
- Improve connections between the notebook and analysis structure.
  - More **annotations** to show hidden features, such as insight type / recommendation type / instructions…
- Support user-customized EDA trajectories.
  - Support **user-specified analysis** intents
  - Maintain ecological synchronization
- Support thread **decomposition, switching, and saving**.

Discussion & Future Work

- Improve connections between the notebook and analysis structure.
  - More **annotations** to show hidden features, such as insight type / recommendation type / instructions…
- Support user-customized EDA trajectories.
  - Support **user-specified analysis** intents
  - Maintain ecological synchronization
- Support thread **decomposition, switching, and saving**.
- Evaluate **usability** through a comparative study

WHATSNEXT: Guidance-enriched Exploratory Data Analysis with Interactive, Low-Code Notebooks





Contact: cchen24@umd.edu