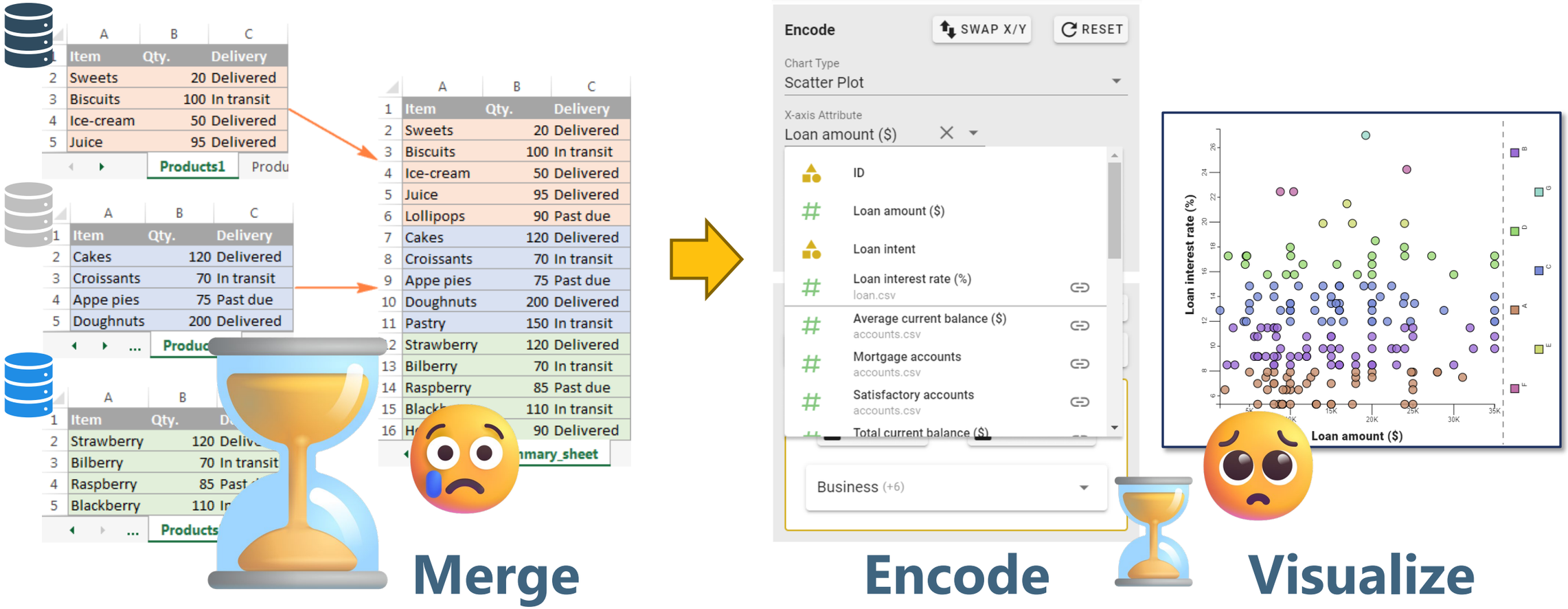
VIS2024

Preliminary **Guidelines** For Combining Data Integration and Visual Data Analysis

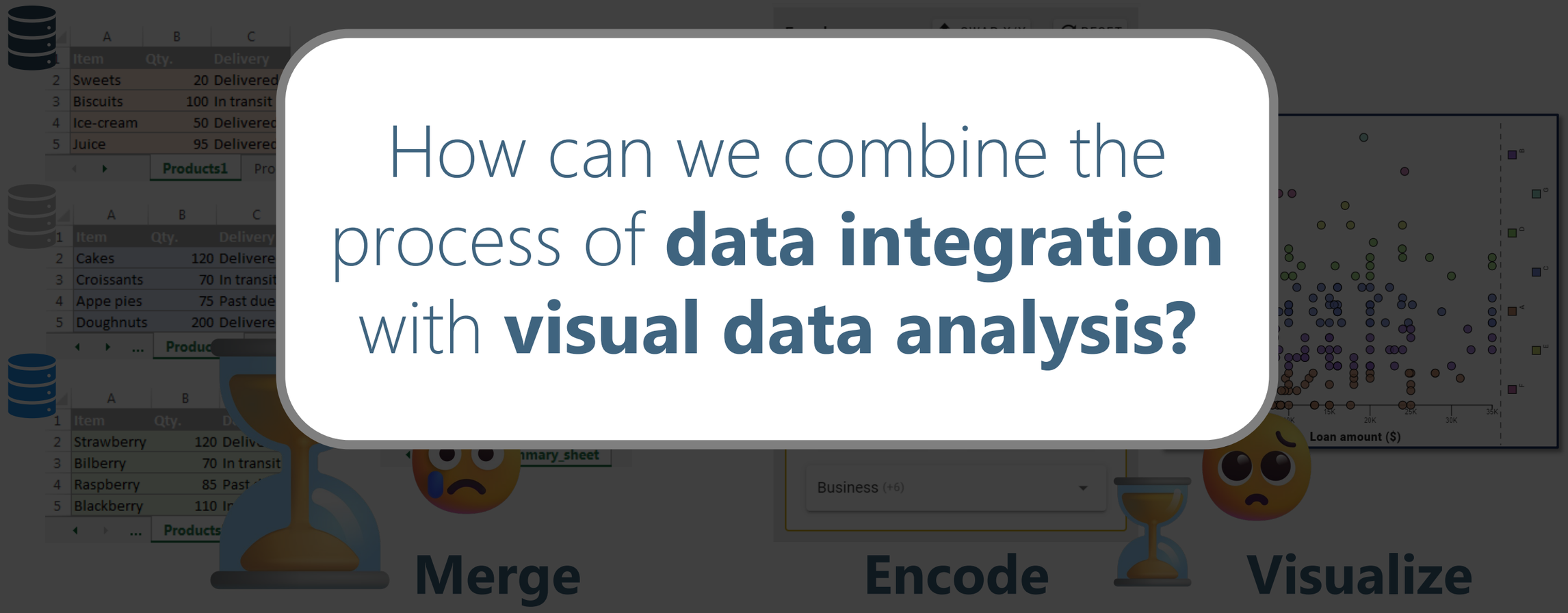
Adam **Coscia*** Ashley **Suh**** Remco **Chang**** Alex **Endert***

*Georgia Tech **Tufts University

A **data integration** + **visual analytics** scenario

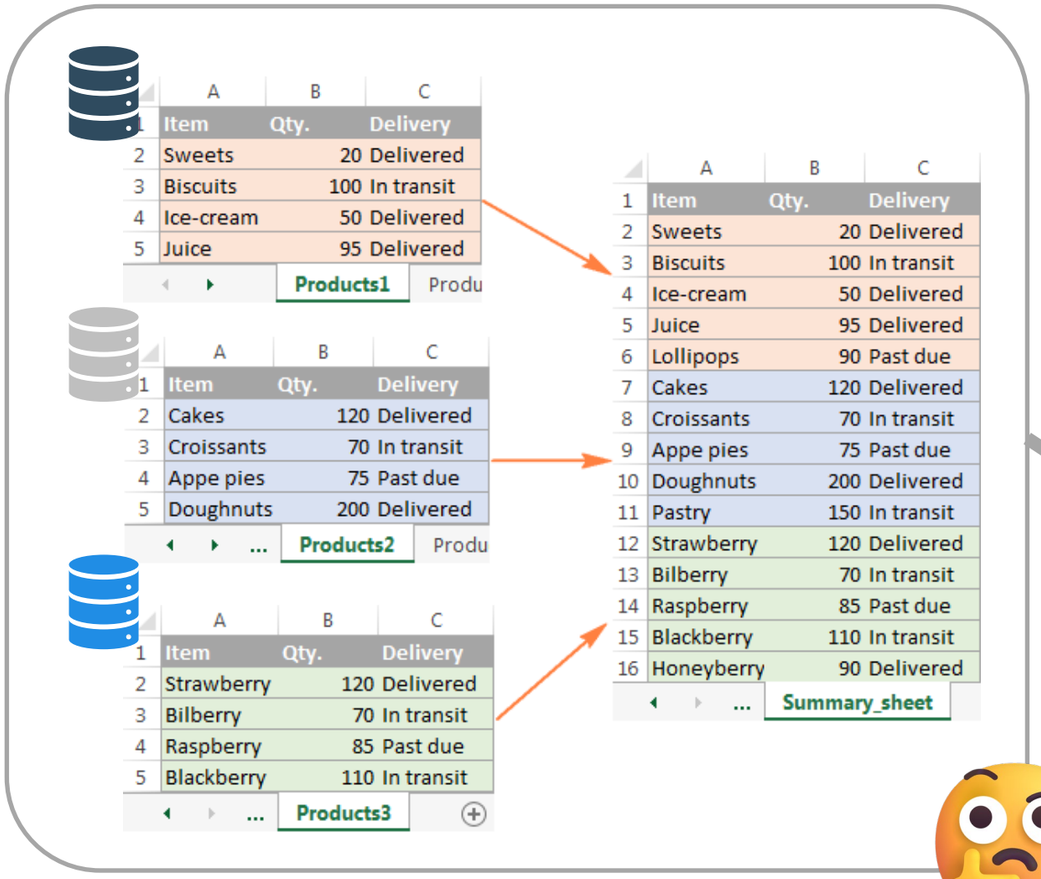


A **data integration** + **visual analytics** scenario



How can we combine the process of **data integration** with **visual data analysis?**

A **data integration** + **visual analytics** scenario

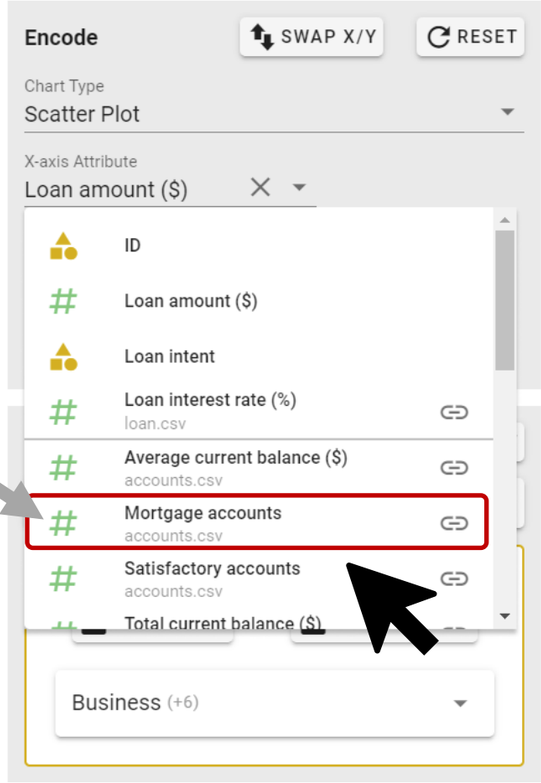


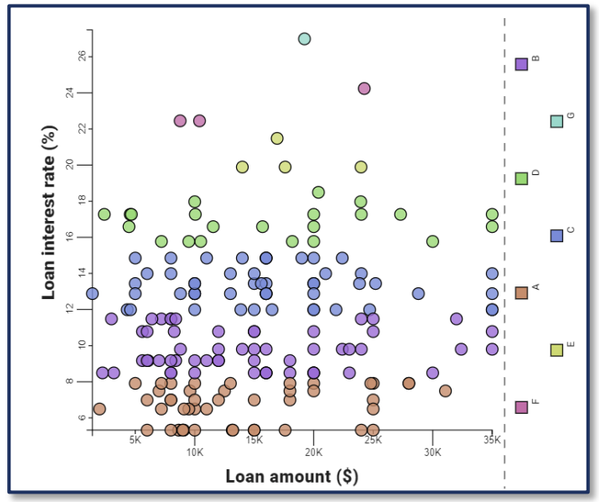
Item	Qty.	Delivery
Sweets	20	Delivered
Biscuits	100	In transit
Ice-cream	50	Delivered
Juice	95	Delivered

Item	Qty.	Delivery
Cakes	120	Delivered
Croissants	70	In transit
Appe pies	75	Past due
Doughnuts	200	Delivered

Item	Qty.	Delivery
Strawberry	120	Delivered
Bilberry	70	In transit
Raspberry	85	Past due
Blackberry	110	In transit

Item	Qty.	Delivery
Sweets	20	Delivered
Biscuits	100	In transit
Ice-cream	50	Delivered
Juice	95	Delivered
Lollipops	90	Past due
Cakes	120	Delivered
Croissants	70	In transit
Appe pies	75	Past due
Doughnuts	200	Delivered
Pastry	150	In transit
Strawberry	120	Delivered
Bilberry	70	In transit
Raspberry	85	Past due
Blackberry	110	In transit
Honeyberry	90	Delivered





Merge + Encode **Visualize**

Two open research questions

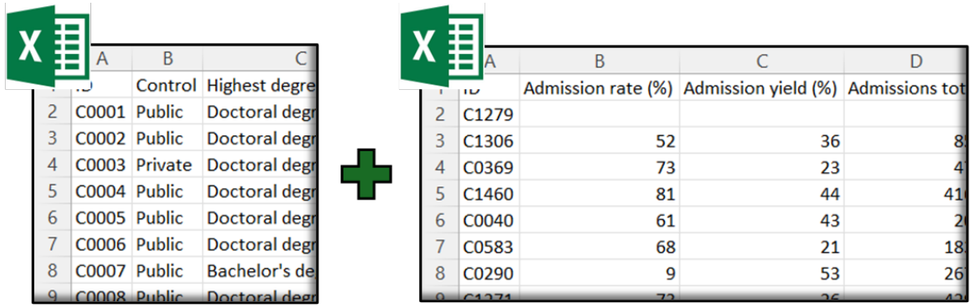
1. Where and how should data integration operations be supported in tandem with visual analytics operations?
2. How will incorporating data integration into an on-going visual analytics process affect user behaviors?

Goal: Contribute **preliminary guidelines** for incorporating **data integration** into an active **visual analytics** process

Manual **"ex-situ"** data integration with Excel

VS

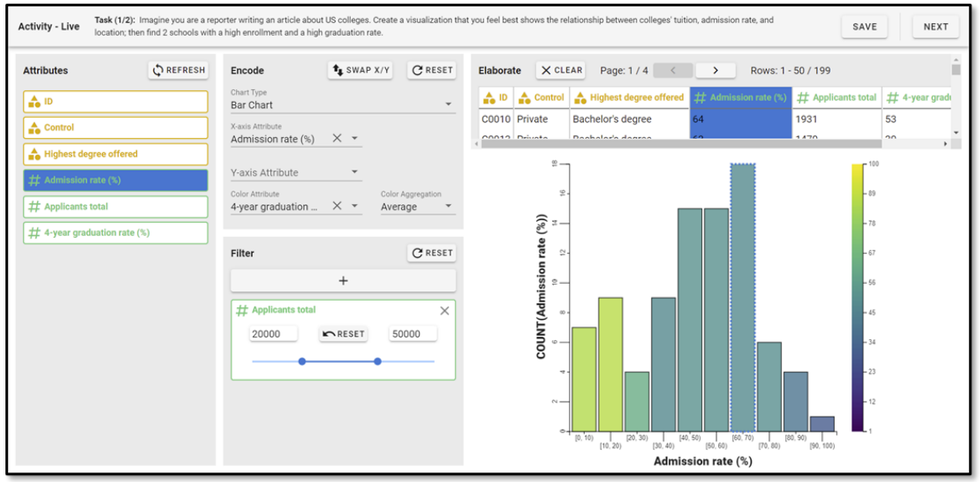
Automatic **"in-situ"** data integration built-in



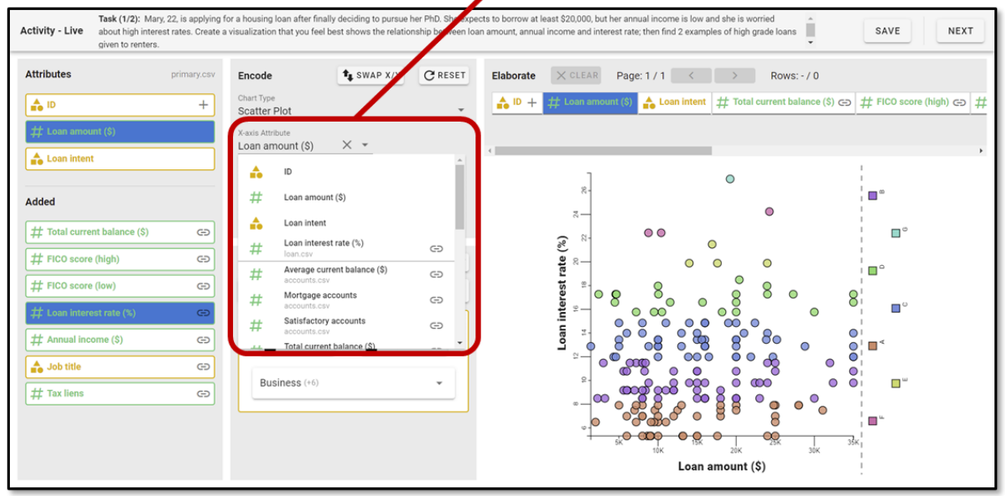
	A	B	C
		Control	Highest degre
2	C0001	Public	Doctoral degr
3	C0002	Public	Doctoral degr
4	C0003	Private	Doctoral degr
5	C0004	Public	Doctoral degr
6	C0005	Public	Doctoral degr
7	C0006	Public	Doctoral degr
8	C0007	Public	Bachelor's de
9	C0008	Public	Doctoral degr

	A	B	C	D
		Admission rate (%)	Admission yield (%)	Admissions tot
2	C1279			
3	C1306	52	36	8
4	C0369	73	23	4
5	C1460	81	44	41
6	C0040	61	43	2
7	C0583	68	21	18
8	C0290	9	53	26

Leave the interface to get data



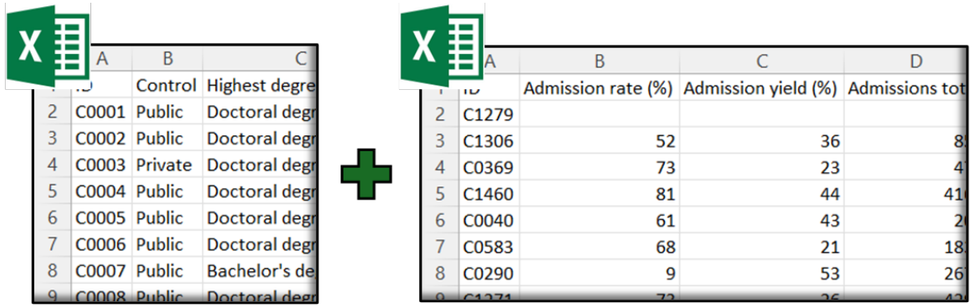
Integrate directly in the interface, without leaving the tool



Manual **"ex-situ"** data integration with Excel

VS

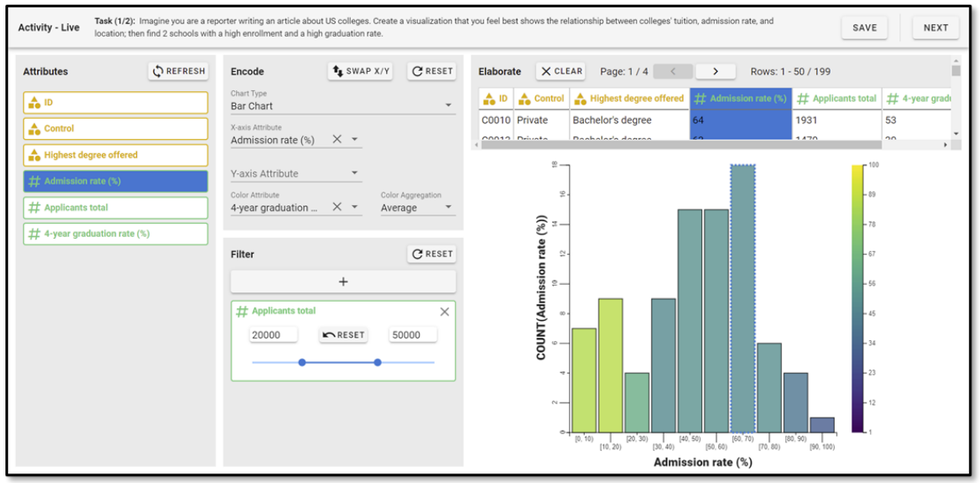
Automatic **"in-situ"** data integration built-in



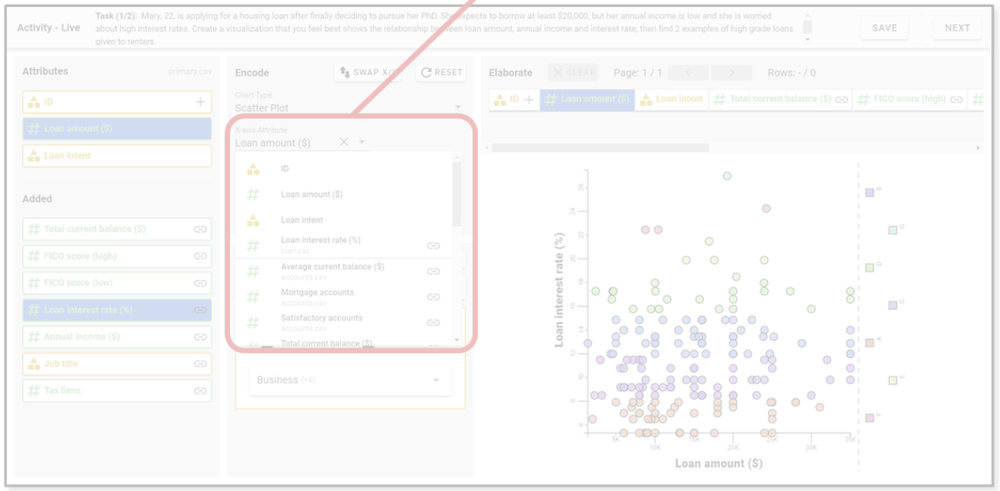
	A	B	C
		Control	Highest degre
2	C0001	Public	Doctoral degr
3	C0002	Public	Doctoral degr
4	C0003	Private	Doctoral degr
5	C0004	Public	Doctoral degr
6	C0005	Public	Doctoral degr
7	C0006	Public	Doctoral degr
8	C0007	Public	Bachelor's de
9	C0008	Public	Doctoral degr

	A	B	C	D
		Admission rate (%)	Admission yield (%)	Admissions tot
2	C1279			
3	C1306	52	36	8
4	C0369	73	23	4
5	C1460	81	44	41
6	C0040	61	43	2
7	C0583	68	21	18
8	C0290	9	53	26

Leave the interface to get data



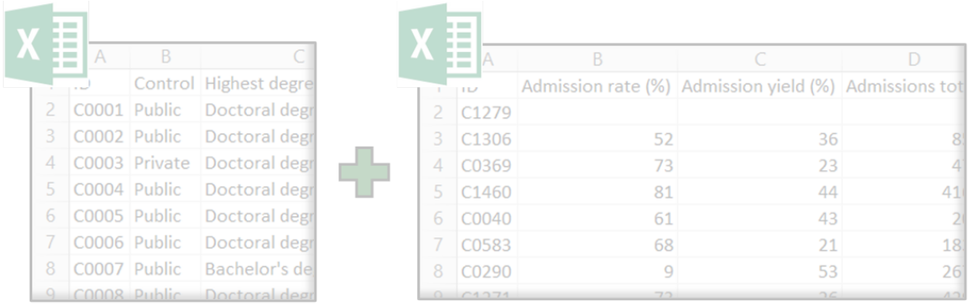
Integrate directly in the interface, without leaving the tool



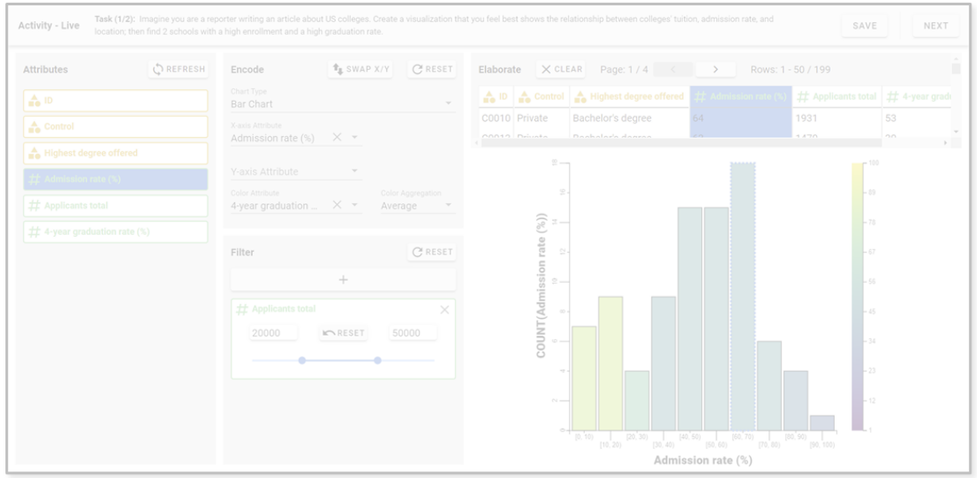
Manual **"ex-situ"** data integration with Excel

VS

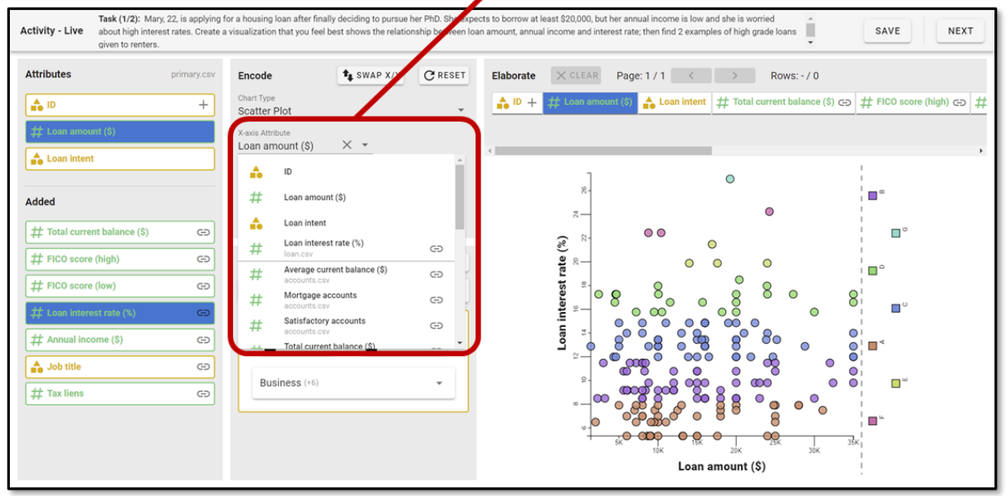
Automatic **"in-situ"** data integration built-in



Leave the interface to get data



Integrate directly in the interface, without leaving the tool



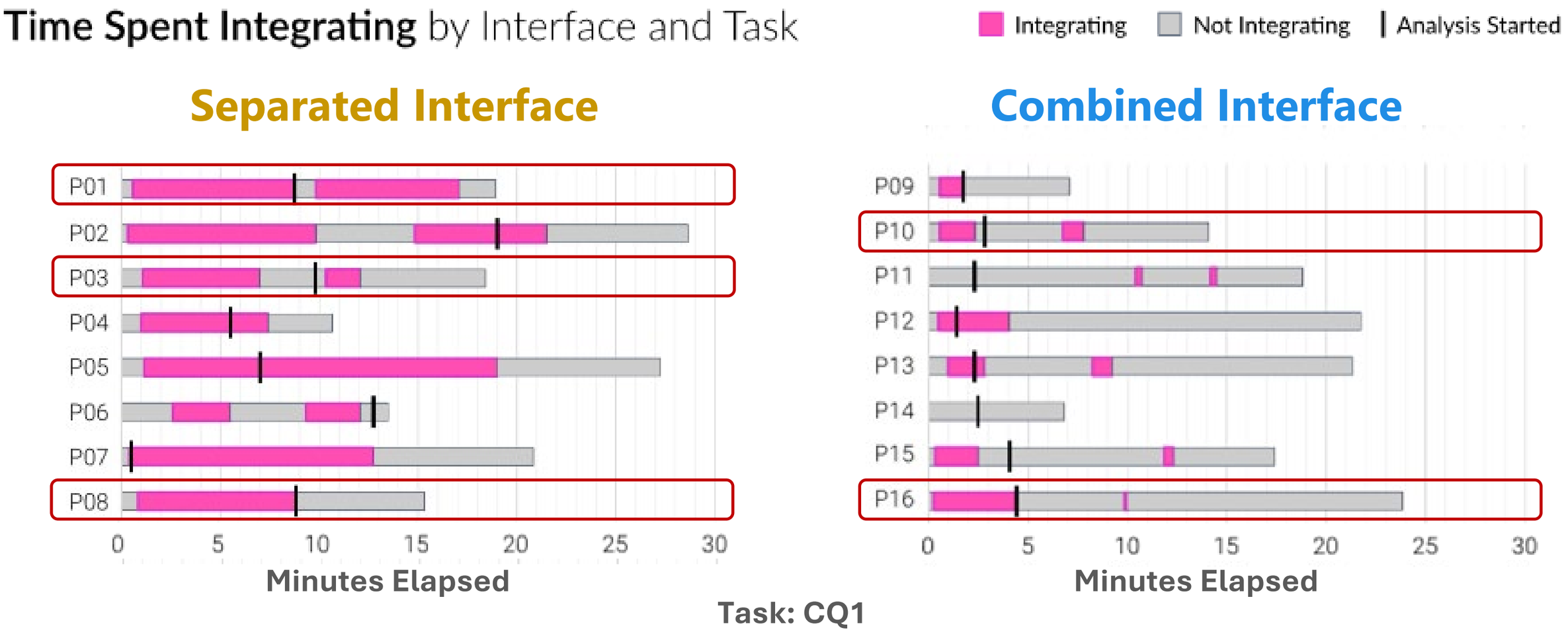
Study design

- **Participants**: 16 university students (P1-16)
  - **Fields:** Computer Science (8), Analytics (4), Human-Computer Interaction (2), Human-Centered Computing (1), and Industrial Design (1)
  - **Experience:** Tableau (15), Python/Matplotlib (11), R/ggplot2 (6), Microsoft Power BI (4), D3.js (2), SAS (2), and AWS Quicksight (1)

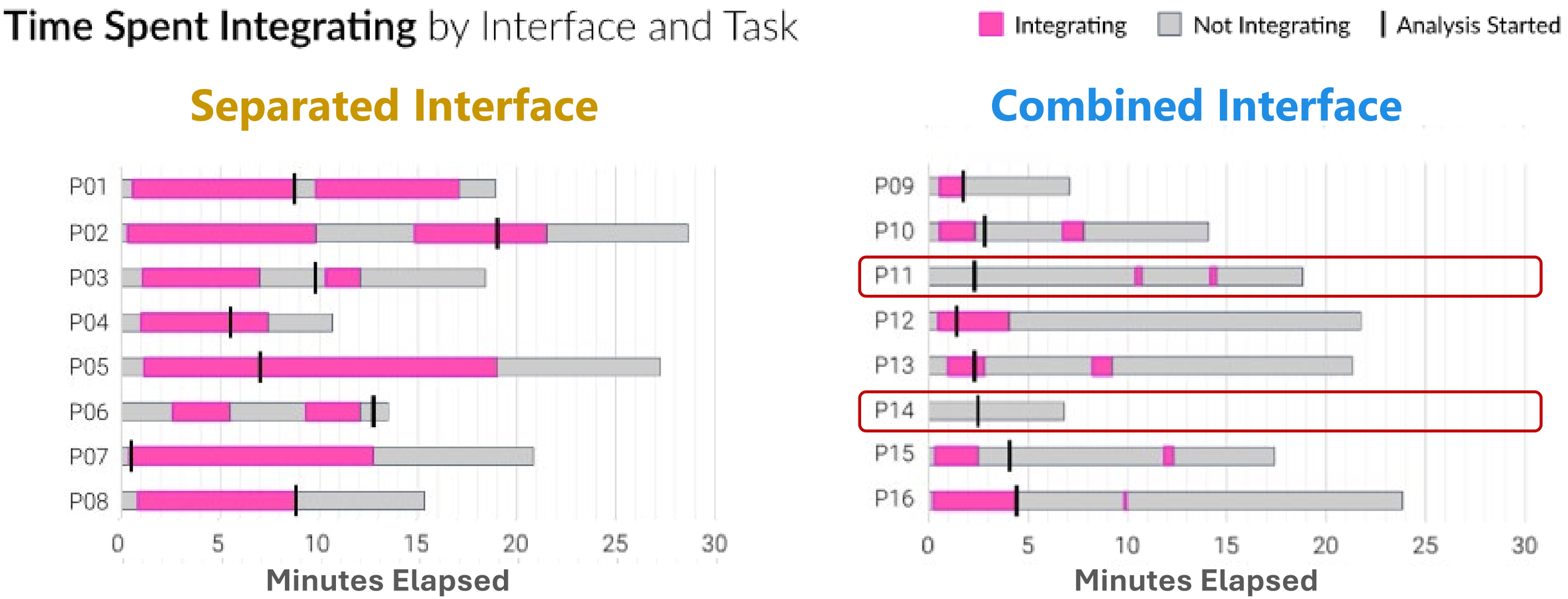
Separated Interface
Combined Interface

- **Procedure** (counter-balanced interface/task):
  - #1 Practice → Task 1/2 → Task 2/1 → Practice → Task 1/2 → Task 2/1
  - #2 Practice → Task 1/2 → Task 2/1 → Practice → Task 1/2 → Task 2/1

Study results | Time spent integrating

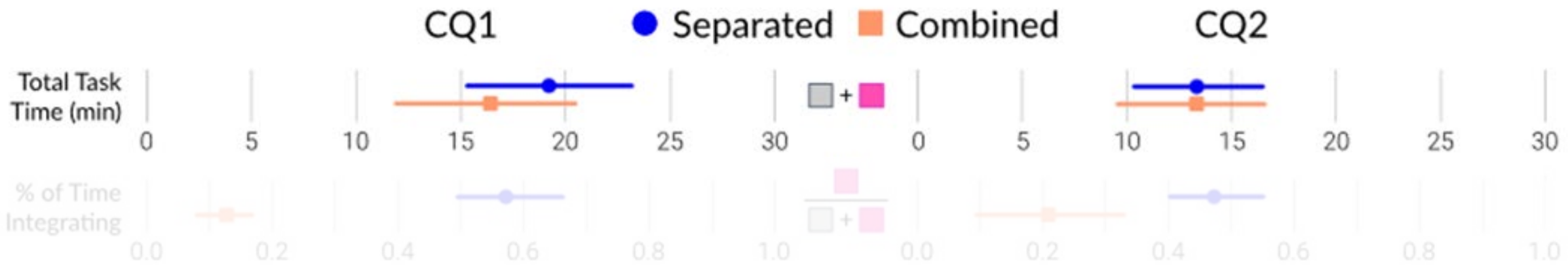


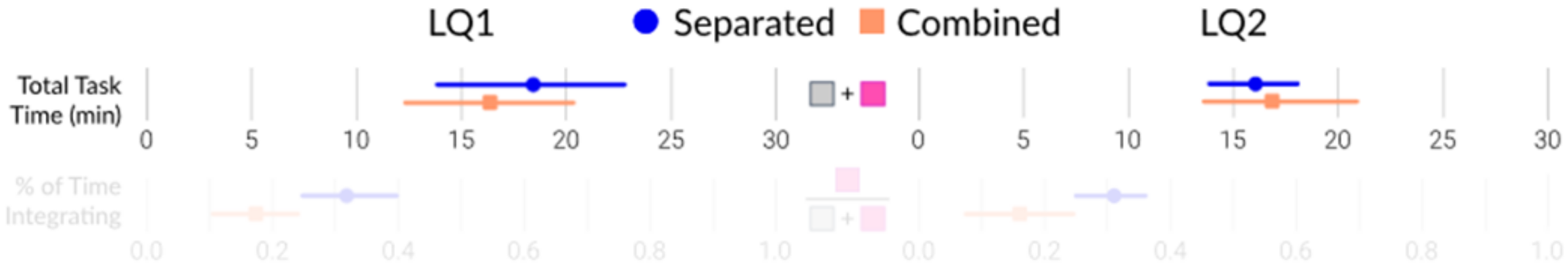
Study results | Time spent integrating



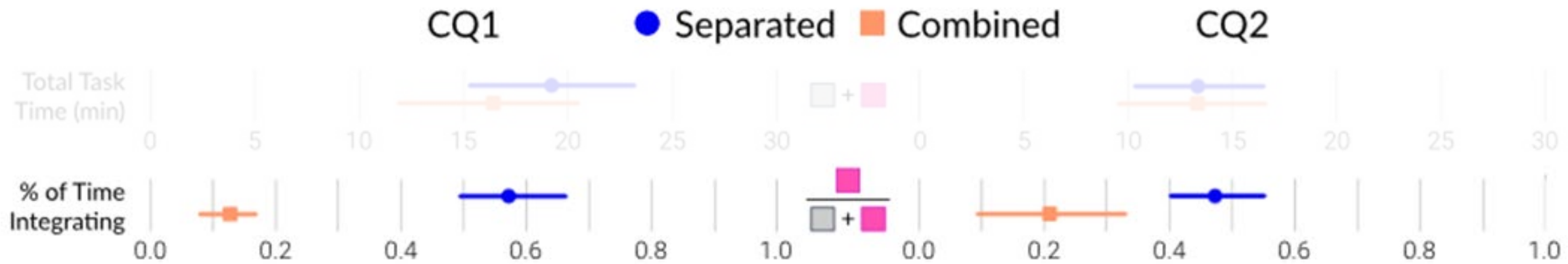
Task: CQ1

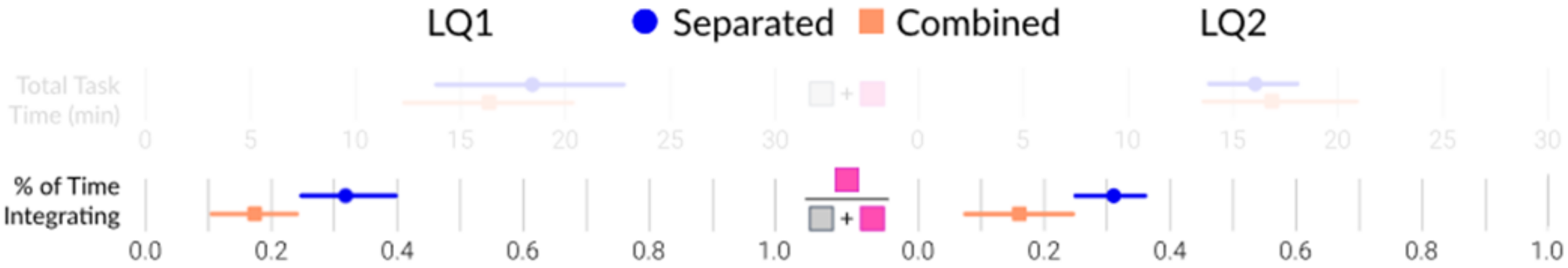
Study results | Time spent integrating



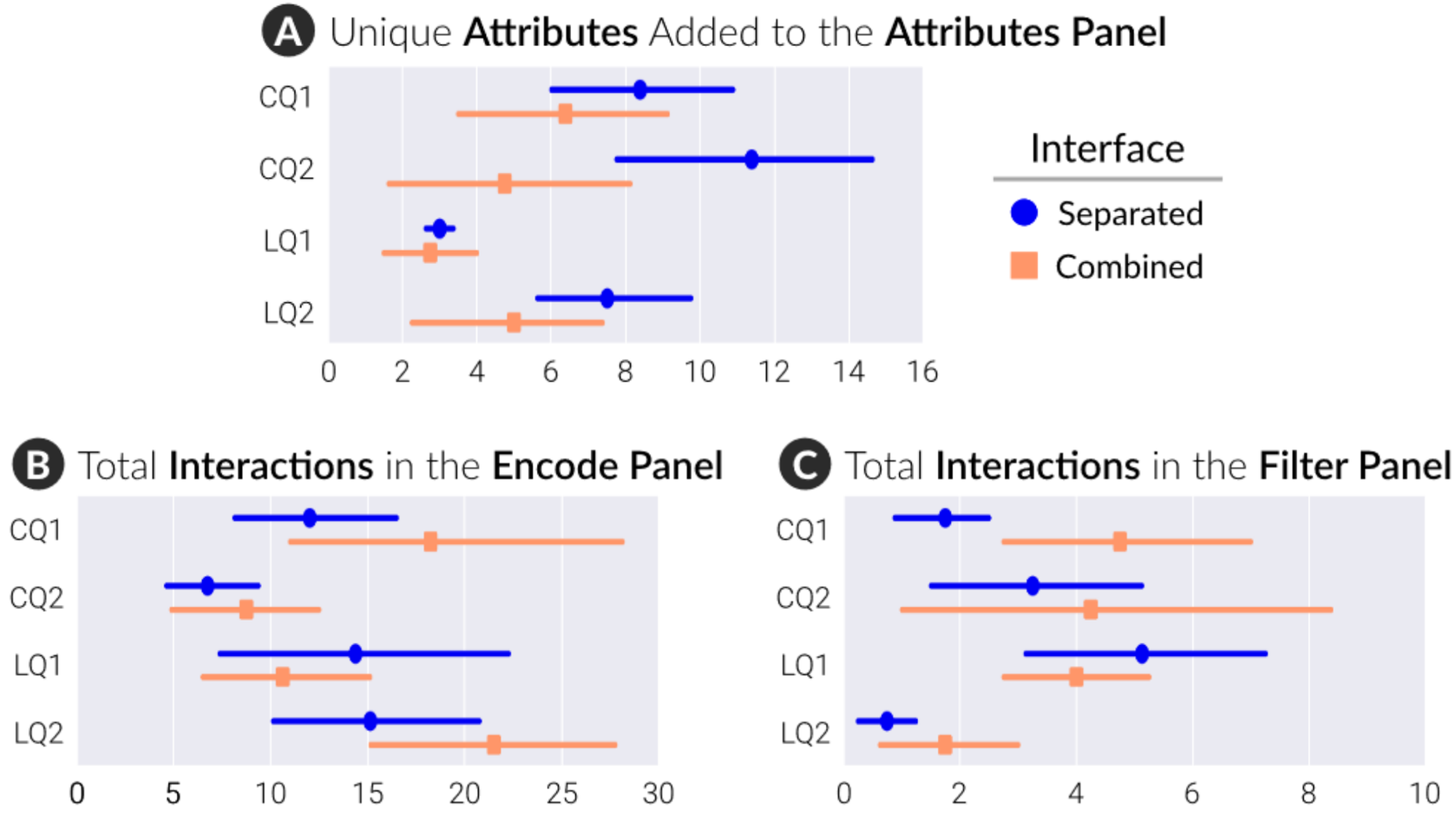


Study results | Time spent integrating





Study results | Attribute interactions



Study results | Participant behaviors

- **Satisficing**
  - Some participants prioritized insight generation over data processing, potentially missing important attributes
  - While others used integration to gain additional insights at the cost of speed

*"I had **less time** to decide which attributes to use and spent **more time** pre-processing data. I prefer the **[Combined]** interface more. In visual data analysis, it's more important to gain insights." - **P5***

*"In terms of **accuracy** and **insights**, the **[Separated]** interface was better. For workflow, the **simplicity** of the **[Combined]** interface was better... I think it all comes down to how much you **trust** the data." - **P9***

Study results | Participant behaviors

- **Exhibiting bias**
  - Some participants visualized the same subset of "familiar" attributes even when integrating new ones was a single click (*confirmation bias*)
  - Others explicitly stuck to their initial integrated set of attributes (*anchoring*)

*"When I work on visualizations, I think of it as a **two-step process**: I find the attributes first, then make the visualizations. Otherwise, it's a lot to keep track of and think about… I'm just in the habit of making my list **before** visualizing… I think of the tasks as **separate**… I think my experience in **Tableau** makes me expect to have to connect data in sheets first." - **P9***

Discussion | Design guidelines

1. **Show where and how data are being integrated**
2. **Use in-situ integration for exploring the space of attributes**
3. **Balance manual and automated approaches**

Discussion | Design guidelines

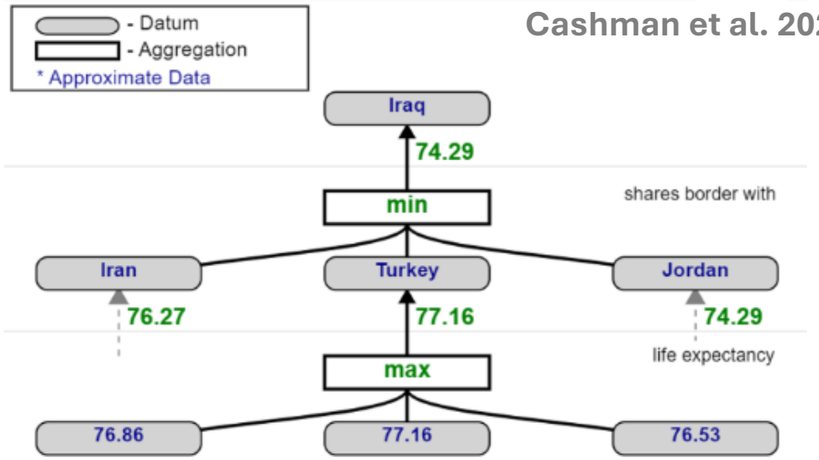
1. **Show where and how data are being integrated**
2. **Use in-situ integration for exploring the space of attributes**
3. **Balance manual and automated approaches**

❌ **Challenge:** "Anonymous" integration

*"In the **[Separated]** interface, I had to manage column names and [avoid] manual **errors**... I feel like the **[Combined]** interface would do a better job of overcoming [copy-and-paste errors]." - **P1***

✅ **Solution:** Integration "pop-up" windows

Cashman et al. 2020 *CAVA*



Discussion | Design guidelines

1. **Show where and how data are being integrated**
2. **Use in-situ integration for exploring the space of attributes**
3. **Balance manual and automated approaches**

❌ **Challenge:** Too many attributes

> *"I would often look for just the attributes I felt like were **relevant** to the task. I **ignored** the rest because I had to go through the tables to find them [in the **Combined** interface]." – **P10***

✅ **Solution:** Automatically determine subset of relevant attributes to show

1. Limit the number of in-situ attributes shown at once (**attributes on demand**)
2. Use **semantic relevance** to suggest related attributes (e.g., with a knowledge graph)

Discussion | Design guidelines

1. Show where and how data are being integrated
2. **Use in-situ integration for exploring the space of attributes**
3. Balance manual and automated approaches

❌ **Challenge:** High cost of integration

*"I copied the values into the **wrong file** because so many windows were open [in the **Separated** interface]. That **wasted my time**." - P2*

✅ **Solution:** Visual data "scents"

*"I didn't know what all attributes were [in the **Combined** interface], but I checked the **names of the files** for the attributes in order to **choose** which attributes to use" - P5*

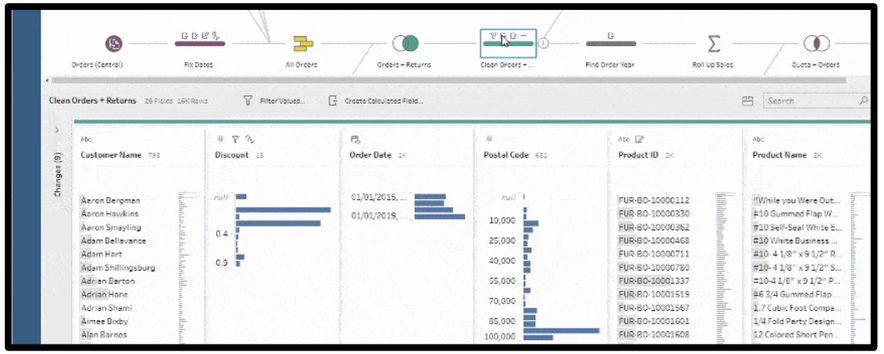
Discussion | Design guidelines

1. Show where and how data are being integrated
2. Use in-situ integration for exploring the space of attributes
3. **Balance manual and automated approaches**

❌ **Challenge:** Manual preferred

*"Since I wasn't the one doing the joins [in the **Combined** interface], it was **harder to remember** the attributes that were available to me. I would have remembered them if I had to **manually join** the attributes." - **P2***

✅ **Solution:** Provide manual data prep for important joins, like in Tableau



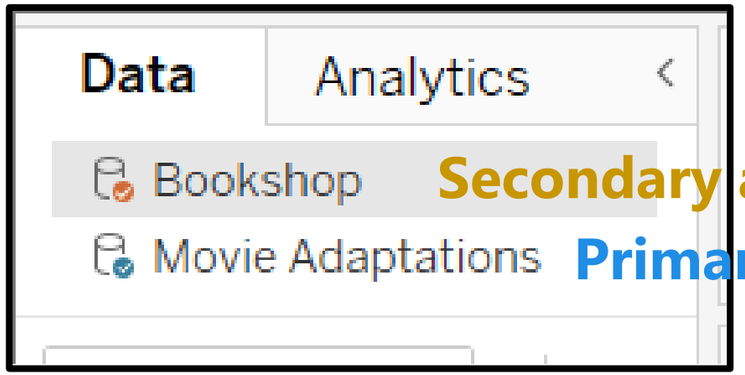
Discussion | Design guidelines

1. Show where and how data are being integrated
2. Use in-situ integration for exploring the space of attributes
3. **Balance manual and automated approaches**

❌ **Challenge:** Automated preferred

*"It takes a long time to do **manual** integration. When I open a file, I have thoughts about what it may contain. It's not the same operation to **find** and **use** the attribute, unlike in the **[Combined]** interface." - P2*

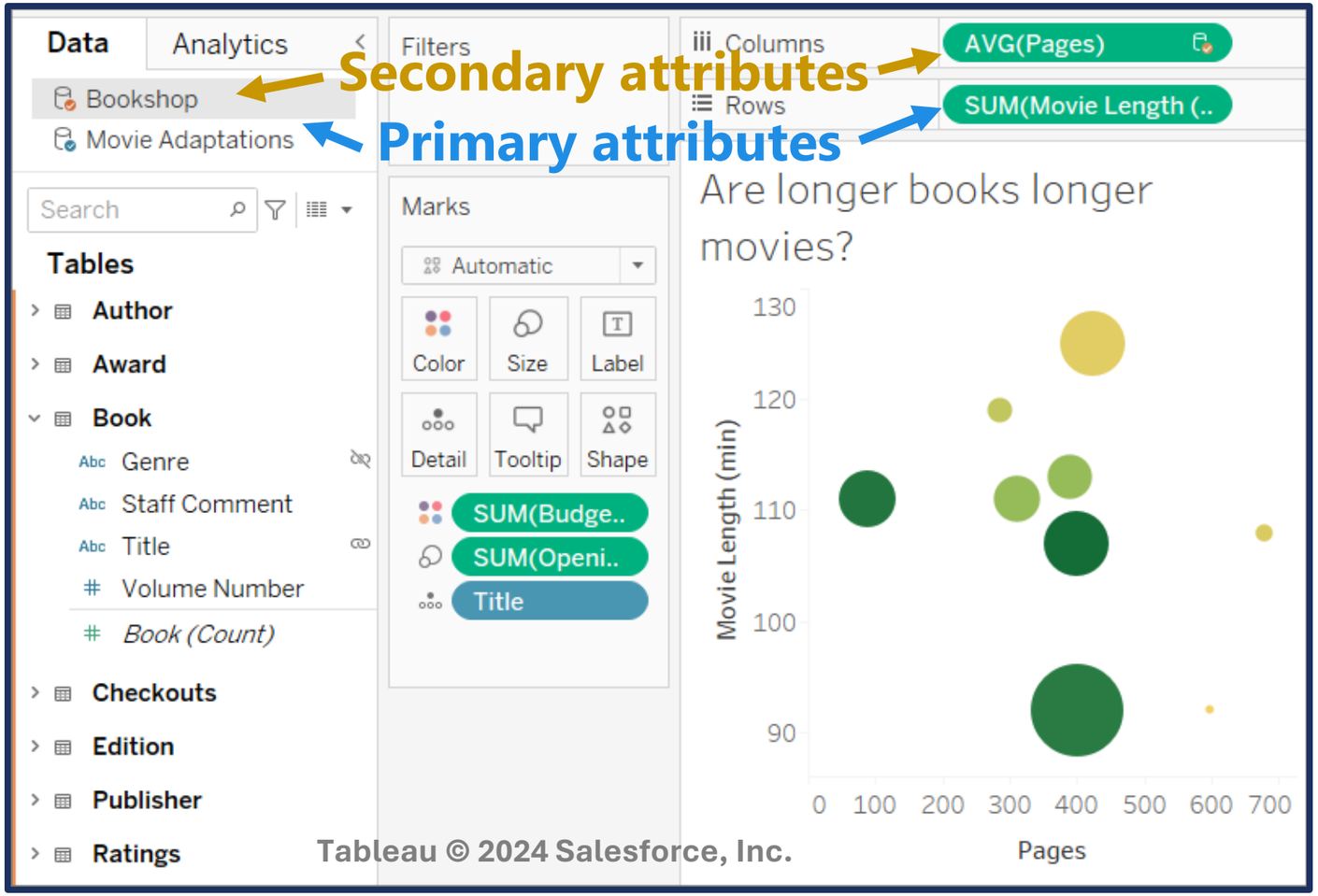
✅ **Solution:** Allow data *"blending"* for trivial integration steps, like in Tableau



Secondary attributes

Primary attributes

Discussion | Design guidelines



3. Balance manual and automated approaches

✅ **Data blending in Tableau**

1. Combines **primary** and **secondary** attributes in the same interface
2. Blends the data sources based on table keys

Discussion | Design guidelines

1. **Show where and how data are being integrated**
   - ✅ Use integration "**pop-up**" windows to avoid "**anonymous**" integration
   - ✅ Show only relevant **subset** of attributes to avoid **satisficing**
2. **Use in-situ integration for exploring the space of attributes**
   - ✅ Use visual "**scents**" to support **sensemaking** during in-situ integration
3. **Balance manual and automated approaches**
   - ✅ Provide **manual** integration for **important** joins that need verification
   - ✅ Allow **automated** integration for **trivial** steps (e.g., blending in Tableau)

Discussion | Revisiting our questions

- *Where and how should data integration operations be supported in tandem with visual analytics operations?*
  - Several integration **strategies**: before analysis, on the fly, & switching between
  - **Time spent** on tasks + interactions **not** significantly different b/w interfaces

✅ **In-situ integration could enable analysts to explore attributes faster than analogous ex-situ strategies, leaving more time for analysis tasks**

Discussion | Revisiting our questions

- *How will incorporating data integration into an on-going visual analytics process affect user behaviors?*
  - Participants used integration to generate and track **hypotheses** and **insights**
  - Yet we observed **satisficing** and **biases** in participants' analytical behaviors

✅ **Supporting integration in visual analytics tools will require:**

- **transparency up front about what and how data are integrated**
- **balancing both automated and manual approaches**

Discussion | Limitations & future work

Types of integration

- Deduplication, entity resolution, operation latency, data quality (e.g., missingness)

Task requirements

- Task performance (e.g., correctness), dataset size, performing "real" integration

Users' experience

- Different analysis backgrounds/experience, effects of familiarity with domain

Discussion | Limitations & future work

Types of integration

- Deduplication, entity resolution, operation latency, data quality (e.g., missingness)

Task requirements

- Task performance (e.g., correctness), dataset size, performing "real" integration

Users' experience

- Different analysis backgrounds/experience, effects of familiarity with domain

Discussion | Limitations & future work

Types of integration

- Deduplication, entity resolution, operation latency, data quality (e.g., missingness)

Task requirements

- Task performance (e.g., correctness), dataset size, performing "real" integration

Users' experience

- Different analysis backgrounds/experience, effects of familiarity with domain

Discussion | Limitations & future work

Types of integration

- Deduplication, entity resolution, operation latency, data quality (e.g., missingness)

Task requirements

- Task performance (e.g., correctness), dataset size, performing "real" integration

Users' experience

- Different analysis backgrounds/experience, effects of familiarity with domain

VIS2024

Preliminary **Guidelines** For Combining Data Integration and Visual Data Analysis

Adam **Coscia*** Ashley **Suh**** Remco **Chang**** Alex **Endert***

*Georgia Tech **Tufts University

Study results | Time spent integrating