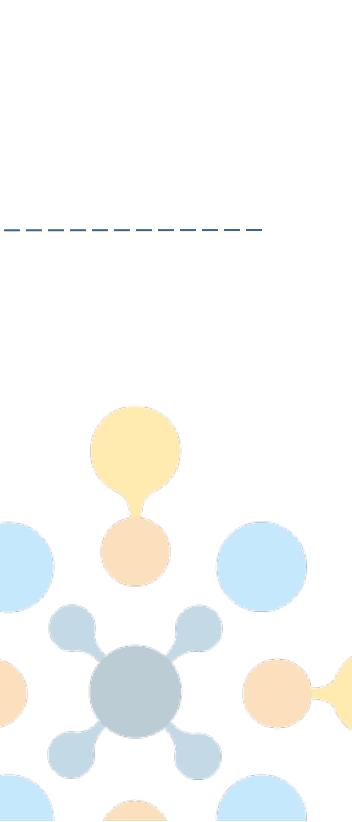


Preliminary Guidelines For **Combining Data Integration and Visual Data Analysis**

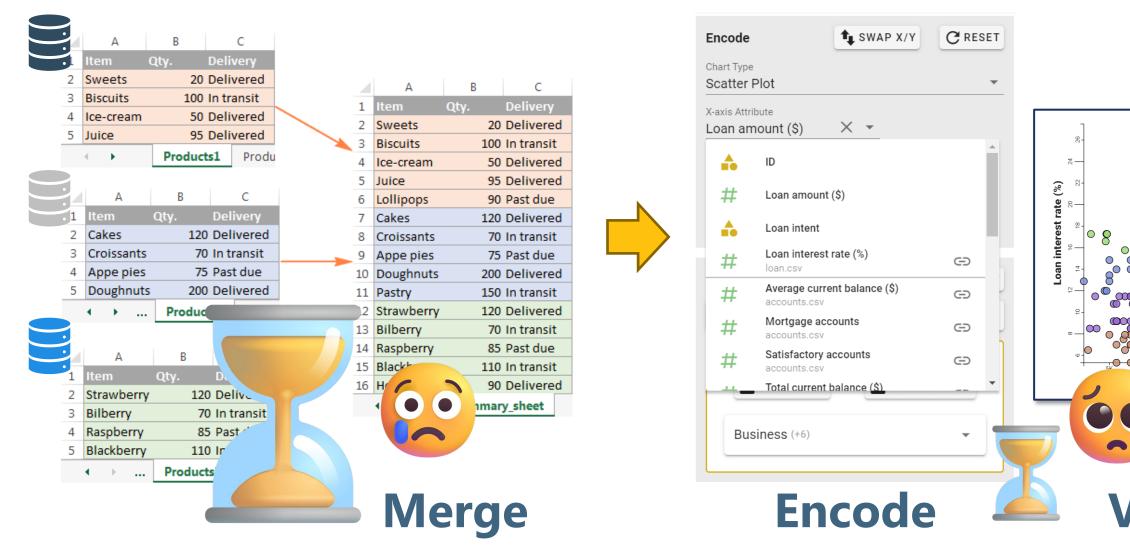
Adam Coscia* Ashley Suh** Remco Chang** Alex Endert*



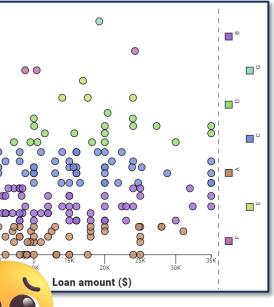




A data integration + visual analytics scenario







Visualize

A data integration + visual analytics scenario

4 Ice-cream Products1 4 Appe pies ... Produce

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

	✓ → … Products	Merge	Encode	
5	Blackberry 110 Ir			
4	Raspberry 85 Past		Business (+6)	
3	Bilberry 70 In transit	imary_sheet		
2	Strawberry 120 Deliv			



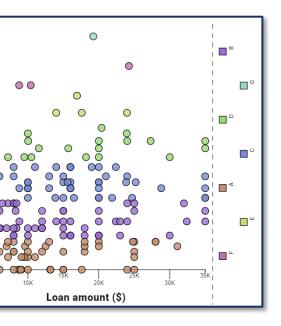


Visualize

A data integration + visual analytics scenario

							\nearrow							
<u> </u>	Α	вс							Encode		SWAP X/Y	C RESE	т	
	Item	Qty. Delivery							Oh e et Ture e				_	
2	Sweets	20 Delivered							Chart Type Scatter				-	
3	Biscuits	100 In transit		1.140	A	B C				1 101				
4	Ice-cream	50 Delivered			em (veets	Qty. Delivery 20 Delivered			X-axis Attri		~			
5	Juice	95 Delivered					_		Loan an	nount (\$)	× •			-38
		Products1 Produ			scuits	100 In transit 50 Delivered	_			ID			*	
					e-cream lice	95 Delivered	_			ID				
·	Α	B C			ollipops	90 Past due	_		#	Loan amount	(\$)			(%)
1	Item	Qty. Delivery			akes	120 Delivered	_				×-/			∞_
2		120 Delivered			roissants	70 In transit	-			Loan intent				Loan interest rate
3		70 In transit			ppe pies	75 Past due	-			Loan interest	rate (%)			°≏ _
4	Appe pies	75 Past due			oughnuts	200 Delivered	-		#	loan.csv		Ð		an i
5	Doughnuts			11 Pa	-	150 In transit	-		#	Average curre	nt balance (\$)	Ð		• Ľ
	۰ ۰ ۰				rawberry	120 Delivered			++	accounts.csv		0		2-
				13 Bil		70 In transit	_		* #	Mortgage acc	ounts	Ð	1	∞
					aspberry	85 Past due	_			accounts.csv			1	
	A	B C	1		ackberry	110 In transit	_		#	Satisfactory a accounts.csv	ccounts	Ð		- e +
1	Item	Qty. Delivery			oneyberry	90 Delivered	_			Total current	palance (\$)		-	
2	Strawberry			4	·	Summary_sheet			_		_		-	
3	Bilberry	70 In transit	·			,								
4	Raspberry	85 Past due							Bus	siness (+6)		•		
5	Blackberry	110 In transit						~						
	·	Products3 +												
							-							
									/ e 	rde	+ En	CO		2
										3-				-





Visualize

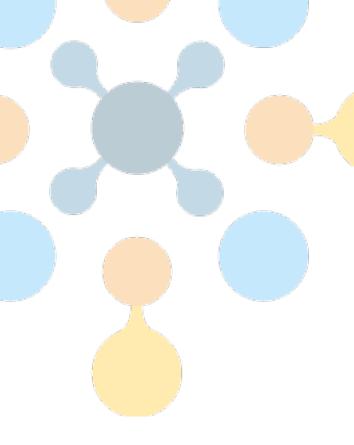
Two open research questions

- 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 Automatic **"in-situ**" integration with Excel VS data integration built-in

v						=			
Λ	А	В	C		^ =	A	В	С	D
		Control	Highest degre				Admission rate (%)	Admission yield (%)	Admissions to
2	C0001	Public	Doctoral degr		2	C1279			
3	C0002	Public	Doctoral degr	_	3	C1306	52	36	5
4	C0003	Private	Doctoral degr		4	C0369	73	23	4
5	C0004	Public	Doctoral degr		5	C1460	81	44	4:
6	C0005	Public	Doctoral degr		6	C0040	61	43	
7	C0006	Public	Doctoral degr		7	C0583	68	21	18
8	C0007	Public	Bachelor's de		8	C0290	9	53	20
9	C0008	Public	Doctoral degr		0	C1071	- 70	26	A.

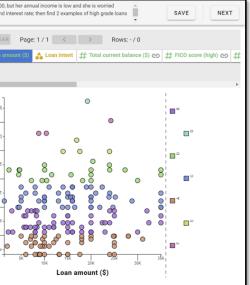
Leave the interface to get data



Integrate directly in the interface, without leaving the tool

ttributes	primary.csv	Encode	t∎ swap x/	C RESET	Elaborate	\times
tD.	+	Chart Type Scatter Plo		-	🔒 ID 🕂	# 4
🛱 Loan amount (\$)		X-axis Attribute				
Loan intent						
dded			oan amount (\$)			
# Total current balance	(\$)		oan intent			-
FICO score (high)	Ð		oan interest rate (%) an.csv	Θ	5	Loan interest rate (%)
FICO score (low)	Ð		verage current balance (\$) ccounts.csv	Ð	r I	st rai
# Loan interest rate (%)		# 。	fortgage accounts ccounts.csv	Ð	1	Itere
# Annual income (\$)	G	# 。	atisfactory accounts ccounts.csv	Ð		an ir
Job title	G		otal current balance (<u>\$).</u>			Ĕ
# Tax liens	Ð	Busine	SS (+6)	-		

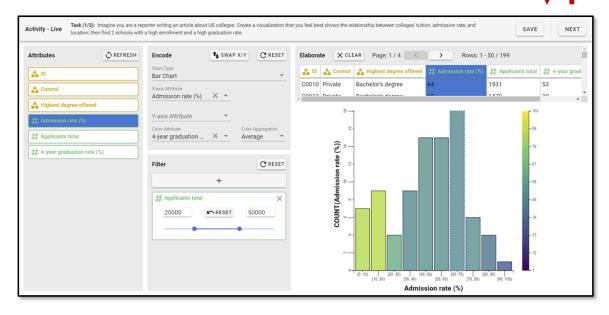




Manual **"ex-situ**" data integration with Excel VS data integration built-in

D Control Highest degree Admission rate (%) Admission yield (%) Admissions to שון C0001 Public Doctoral deg 2 C1279 C0002 Public Doctoral deg 3 C1306 52 36 C0003 Private Doctoral deg 4 C0369 73 23 C0004 Public Doctoral deg 81 5 C1460 44 6 C0005 Public Doctoral deg 6 C0040 61 43 C0006 Public Doctoral deg 7 C0583 68 21 8 C0007 Public Bachelor's d 8 C0290 9 53

Leave the interface to get data



Automatic "in-situ"

Integrate directly in the interface, without leaving the tool

Attributes		Encode		🛊 SWAP X/Y	C RES		Elaborate	
	+					÷	🔥 ID 🕂	
		X-axis Attr						
		Loan ar		× *		-		
Added		#	Loan amount (\$					
		•	Loan intent					24
		#	Loan interest ra Ioan.csv	te (%)				%)
		#	Average current accounts.csv	t balance (\$)				st rat
		#	Mortgage accor accounts.csv	unts		1		Loan interest rate
		#	Satisfactory act accounts.csv					oan ir
			Total current ba	lance (<u>\$).</u>				¹² Lc
					v			

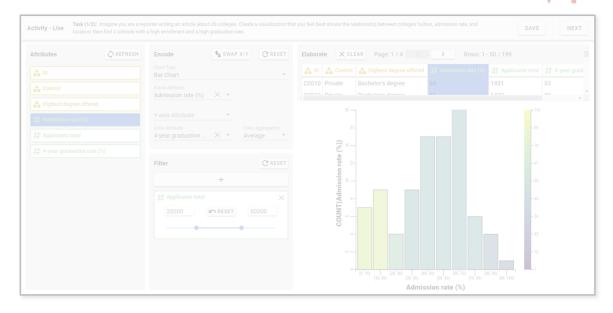




Manual "ex-situ" data Automatic **"in-situ**" integration with Excel data integration built-in

I	v									
	Λ	Α					A		С	D
			Control	Highest degre			U	Admission rate (%)	Admission yield (%)	Admissions to
	2	C0001	Public	Doctoral degr		2	C1279			
	3	C0002	Public	Doctoral degr	_	3	C1306	52	36	8
	4	C0003	Private	Doctoral degr		4	C0369	73	23	4
	5	C0004	Public	Doctoral degr		5	C1460	81	44	41
	6	C0005	Public	Doctoral degr		6	C0040	61	43	2
	7	C0006	Public	Doctoral degr		7	C0583	68	21	18
	8	C0007	Public	Bachelor's de		8	C0290	9	53	26
	C			Doctoral degr		0	C1071	70	7 <i>C</i>	ла

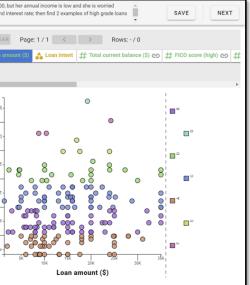
Leave the interface to get data





Attributes	primary.csv	Encode		SWAP X/	C RESET	Elaborate	\times
📤 ID	+	Chart Type Scatter			Ŧ	📤 ID +	#1
# Loan amount (\$)		X-axis Attr		(-			
Loan intent		Loan ar	nount (\$) ×		^	<	
dded		#	Loan amount (\$)		- 1		
# Total current balance (\$)	Ð	•	Loan intent		- 1		_
# FICO score (high)	Ð	#	Loan interest rate loan.csv	(%)	Ð		e (%)
# FICO score (low)	Ð	#	Average current b accounts.csv	alance (\$)	Ð		st rat
# Loan interest rate (%)	Ð	#	Mortgage account accounts.csv		Ð		Loan interest rate (%)
# Annual income (\$)	6	#	Satisfactory acco accounts.csv		Ð		an ir
Job title	9		Total current bala	nce (<u>\$)</u>			Ĕ
# Tax liens	e	Bu	siness (+6)		×		
TT							





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)

• **Procedure** (counter-balanced interface/task):

- Practice \rightarrow Task 1/2 \rightarrow Task 2/1 • #1 \rightarrow
- #2 Practice \rightarrow Task 1/2 \rightarrow Task 2/1 \rightarrow Practice \rightarrow Task 1/2 \rightarrow Task 2/1
- Practice \rightarrow Task 1/2 \rightarrow Task 2/1



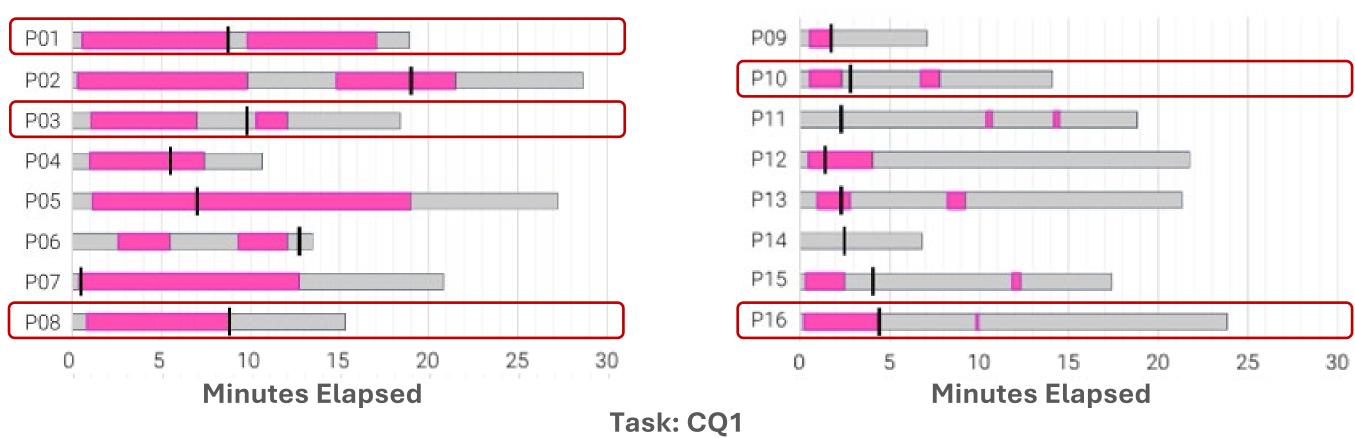
Separated Interface Combined Interface

Time Spent Integrating by Interface and Task

Separated Interface

Combined Interface

Integrating





Ē

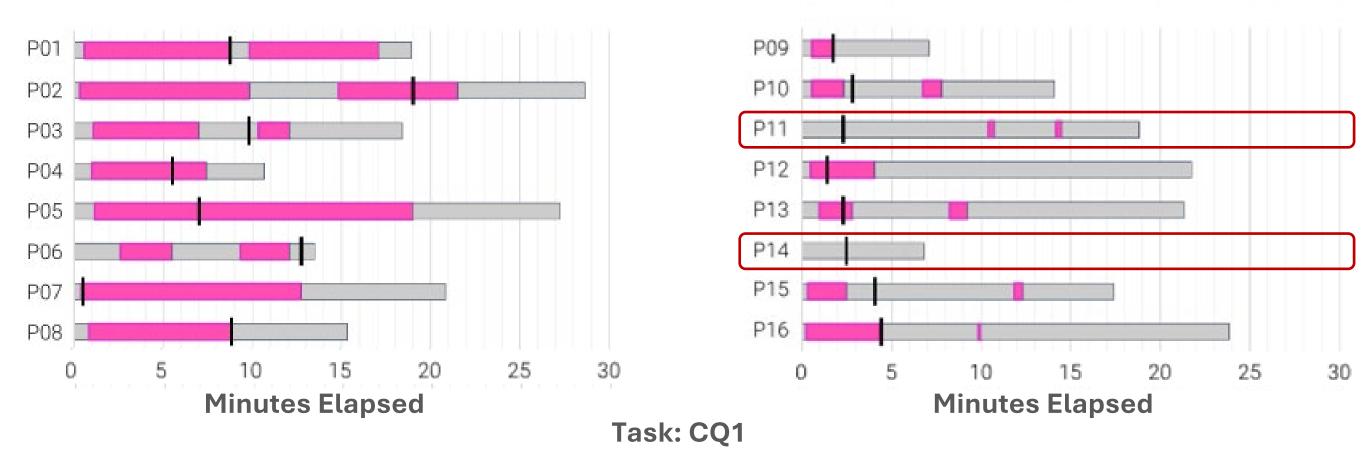
Not Integrating Analysis Started

Time Spent Integrating by Interface and Task



Integrating

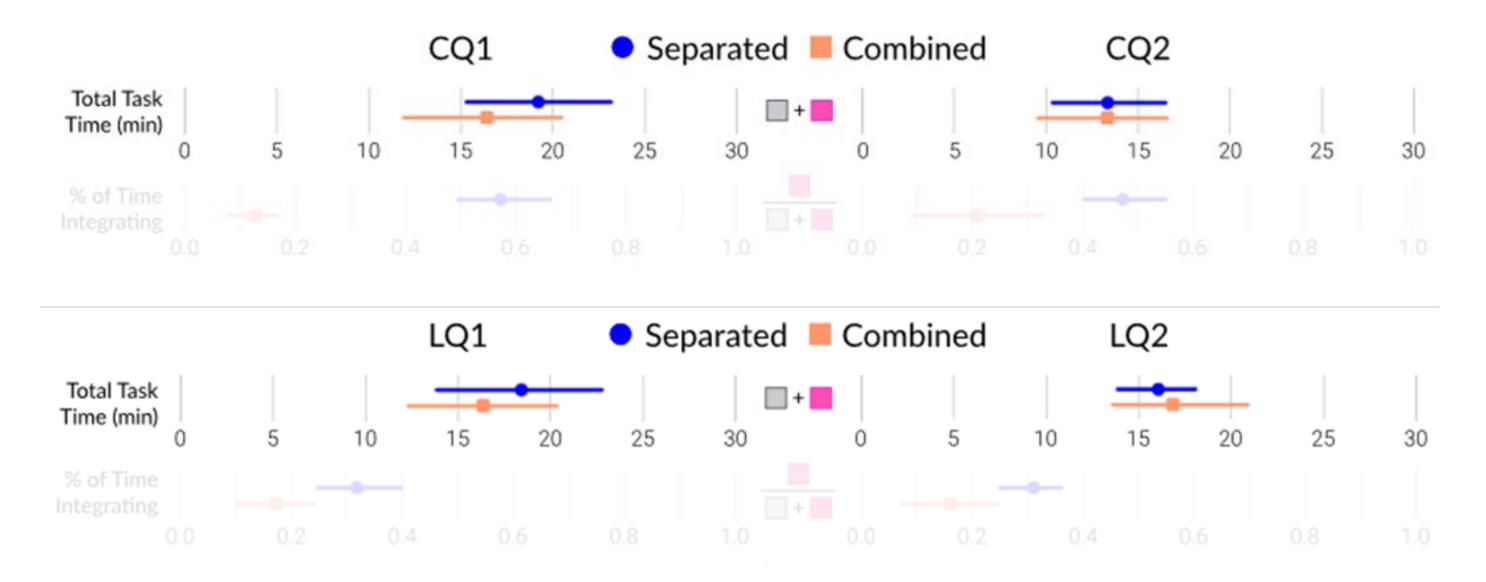
Separated Interface



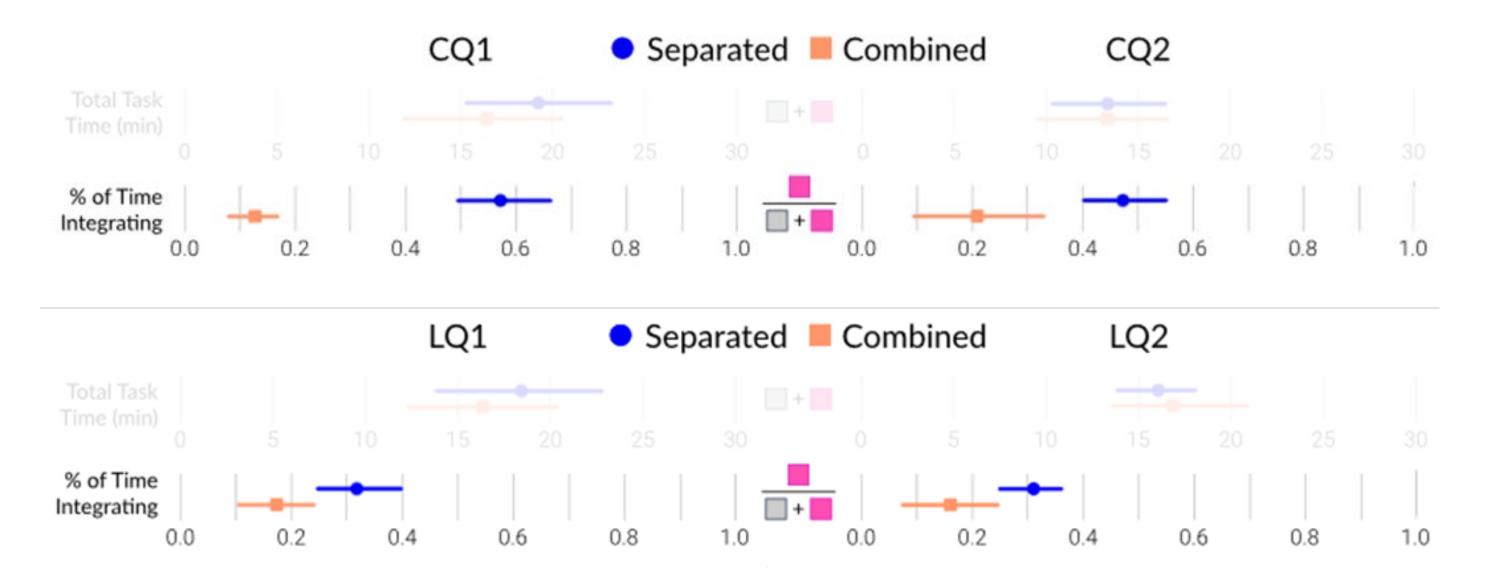


F

Not Integrating Analysis Started

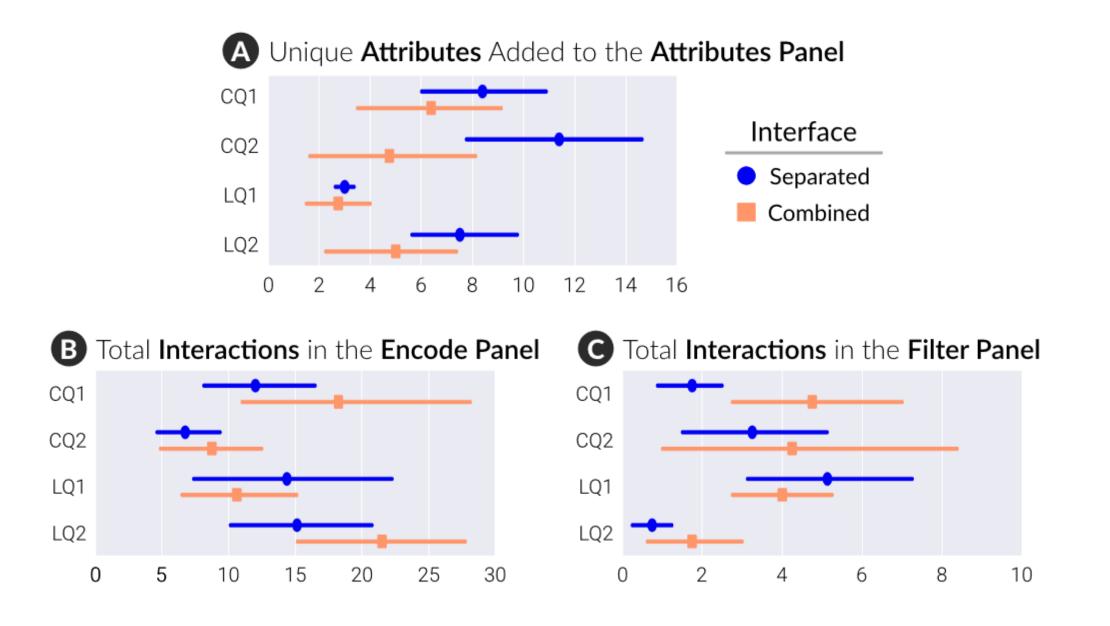








Study results | Attribute interactions





Study results | Participant behaviors

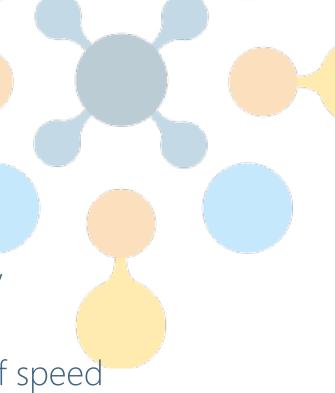
Satisficing

- 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** preprocessing 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





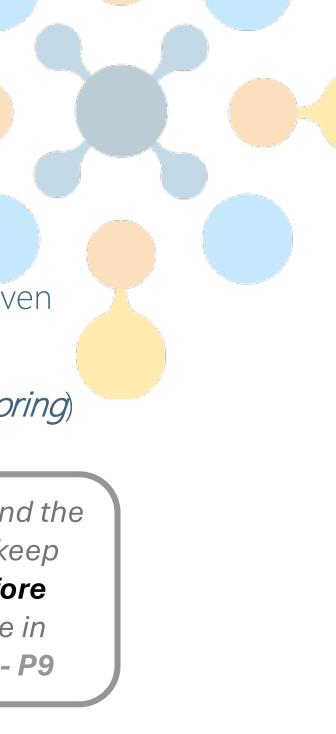
Study results | Participant behaviors

• 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)

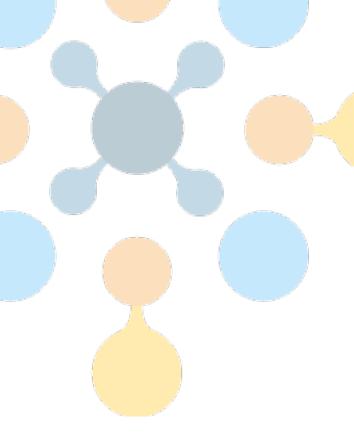
"When I work on visualizations, I think of it as a **two-step process**: I find the attributes first, then make the visualizations. Otherwise, it's a lot to keep track of and think about... I'm just in the habit of making my list **before** visualizing... I think of the tasks as **separate**... I think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





- 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**





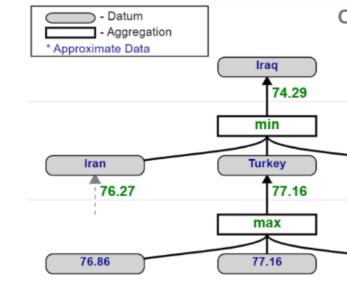
1. Show where and how data are being integrated

Use in-situ integration for exploring the space of attributes
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

shares border with

Jordan

74.29

life expectancy

76.53

1. Show where and how data are being integrated

2. Use in-situ integration for exploring the space of attribute **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)



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



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

X Challenge: Manual preferred

Solution: Provide manual data prep \checkmark for important joins, like in Tableau

"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

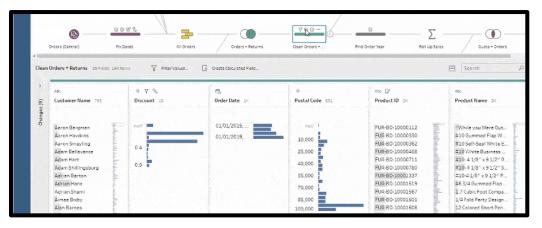


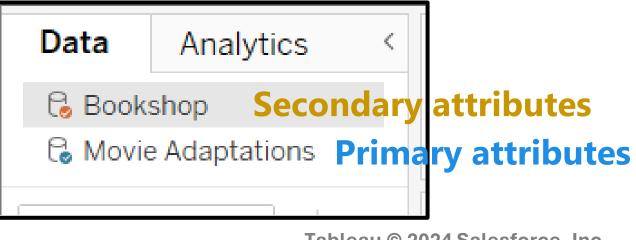
Tableau © 2024 Salesforce, Inc.

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

Challenge: Automated preferred

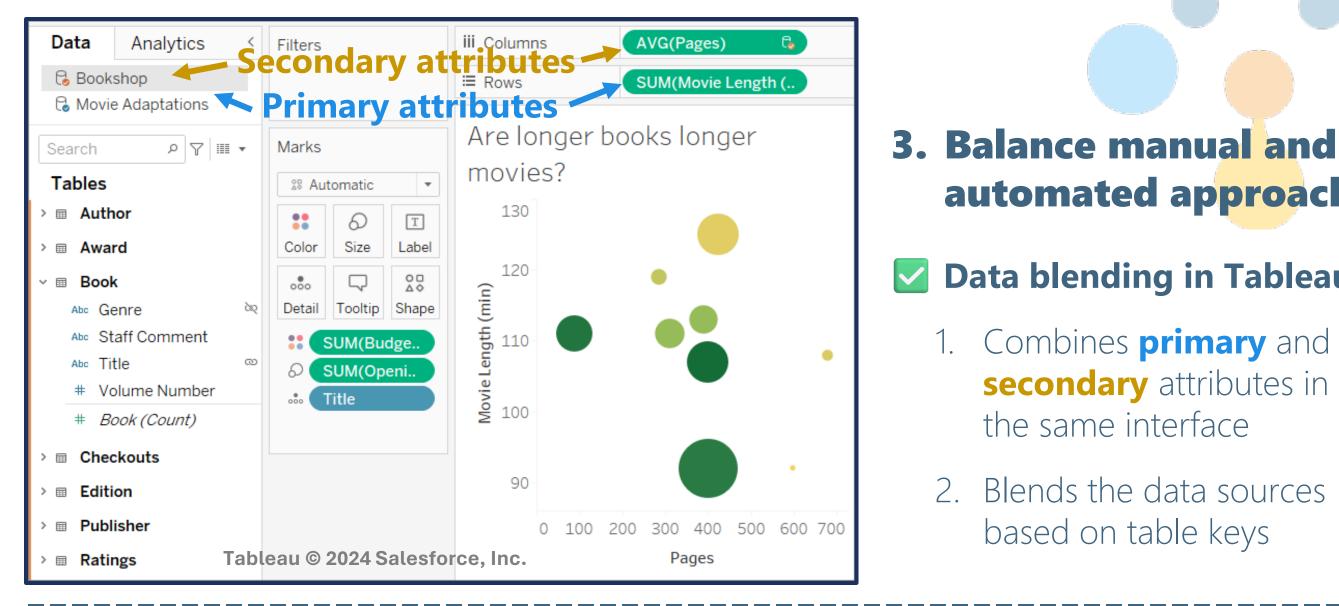
"It takes a long time to do **manual** integration. When I open a file, I have thoughts about what it may contain. It's not the same operation to find and use the attribute, unlike in the [Combined] interface." - P2

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





Secondary attributes Tableau © 2024 Salesforce, Inc.





automated approaches

Data blending in Tableau

secondary attributes in

Blends the data sources

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: \checkmark

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



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



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



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



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





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

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







