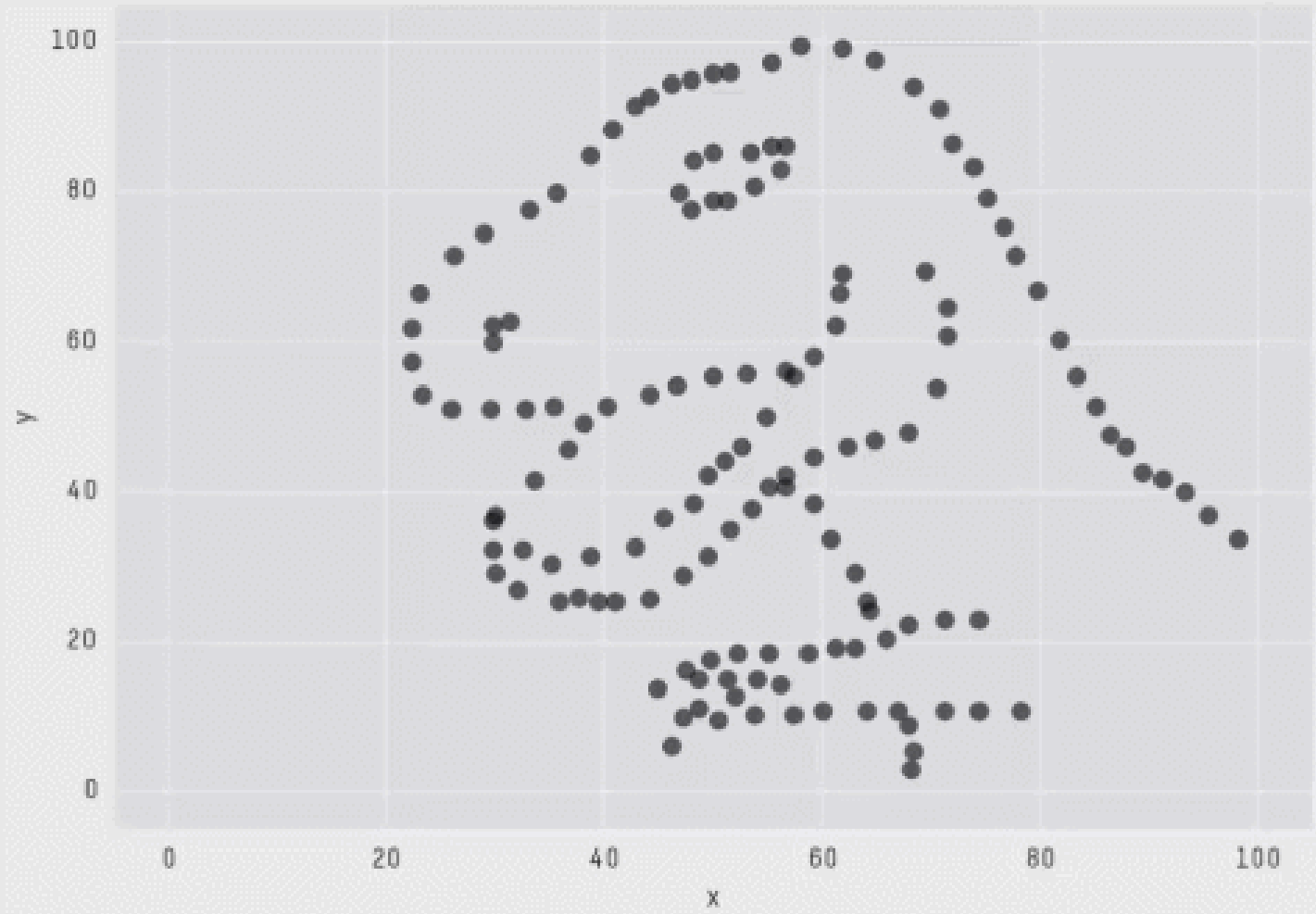


X Mean: 54.2659224
Y Mean: 47.8313999
X SD : 16.7649829
Y SD : 26.9342120
Corr. : -0.0642526

Thinking Data

Yas Suttakulpiboon: Chulalongkorn Business School

Isariya (Yas) Suttakulpiboon

PhD Risk Management and Insurance
- Georgia State University

MSc Mathematical Risk Management
- Georgia State University

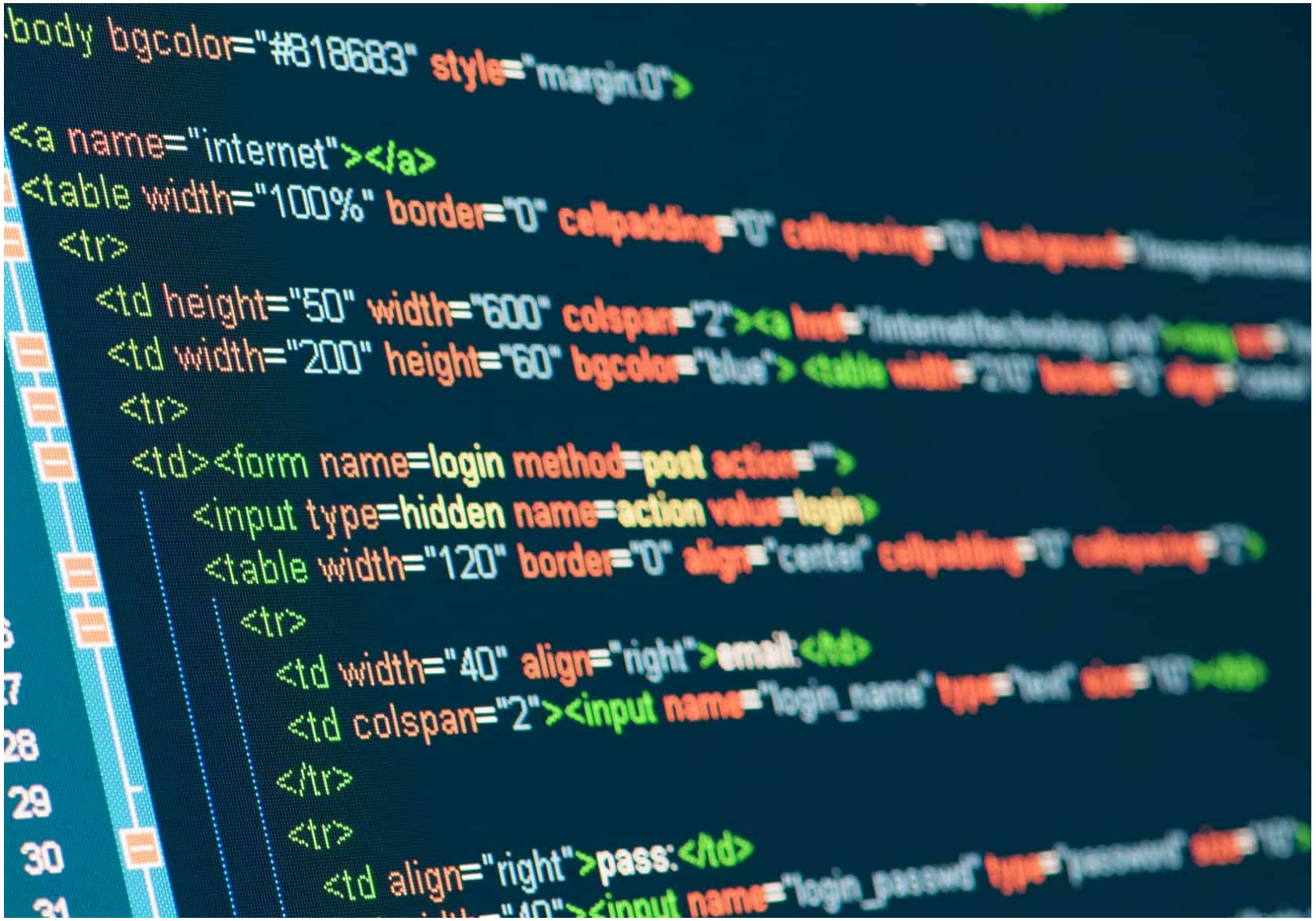
BA Economics (Summa Cum Laude)
- Thammasat University

Working Experience:
- Tax & Trade Policy Consulting
- Lecturer at Georgia State University
- Guest lecturers at Thammasat U, NIDA, Siam U, PSU, Naresuan U
- Risk workshops for PTT, PTG, CPF, TIP, SCB, KTB, GSB and others

Research Interest:
- Insurance Economics
- Enterprise Risk Management
- Risk Modelling

tableau®


Power BI


<body bgcolor="#818683" style="margin:0">

<table width="100%" border="0" cellpadding="0" cellspacing="0"
<tr>
<td height="50" width="600" colspan="2"><a
<td width="200" height="60" bgcolor="blue">
<tr>
<td><form name=login method=post action="">
<input type=hidden name=action value=login>
<table width="120" border="0" align="center"
<tr>
<td width="40" align="right">email:</td>
<td colspan="2"><input name="login_name"
</tr>
<tr>
<td align="right">pass:</td>
<input name="login_passwd"

Thinking Data

AM

- Asking Game
- The “Right” Questions
- Data Value Canvas

PM

- Exploring Deeper Insights in your Data

“The Right Questions help your audience make better decision for their lives/organizations/communities”

Asking Game

แบ่งทีมกันครับ


WEEEEEEEEEEEE


TMB


krungsri
กรุงศรี


SCB


KBank


Krungthai
กรุงไทย


ธนาคารออมสิน
รัฐบาลเป็นประกัน
ธนาคาร
ออมสิน
Government Savings Bank

ข้อมูลที่ให้มา

- Row ID: หมายเลขแถวข้อมูล
- Domain Name: ชื่อ Website ทำการใช้บัตร
- Referring Domain Name: ชื่อ Website อ้างอิงก่อนการใช้บัตร
- View Pages: จำนวนหน้าเข้าชม ณ Domain Name (รวมการขั้นตอนการใช้บัตร)
- View Duration: จำนวนนาทีเข้าชม ณ Domain Name (รวมการขั้นตอนการใช้บัตร)
- Trans Date: ปี-เดือน-วันที่ การใช้บัตร
- Trans Hours: ชั่วโมงนาฬิกา (หน่วย 24 ชั่วโมง) การใช้บัตร

ข้อมูลที่ให้มา

- Product Name: รายละเอียดของสินค้า / บริการ
- Product Category: หมวดของสินค้า / บริการ
- Product Sub-Category: หมวดย่อยของสินค้า / บริการ
- Total Spending: ค่าใช้จ่ายของสินค้า / บริการ ผ่านบัตร Credit
- Household Size: จำนวนคนในครอบครัวของผู้ถือบัตร
- Income: กลุ่มรายได้ของผู้ถือบัตร
- Income Text: คำอธิบายกลุ่มรายได้ผู้ถือบัตร
- Have Children: ผู้ถือบัตรมีบุตรหรือไม่
- Zip Code: ที่อยู่ของผู้ถือบัตร

กรณีศึกษา ข้อมูลการใช้บัตร Credit Card

- ข้อมูลการใช้บัตร Credit บนร้านค้า Online บางส่วนของธนาคาร A ในเดือน พฤศจิกายน 2561

- ถ้าเราอยากทำข่าวจากข้อมูลนี้...

“เราจะเล่าเรื่องอะไรดีเพื่อให้เป็นประโยชน์แก่ผู้อ่าน?”

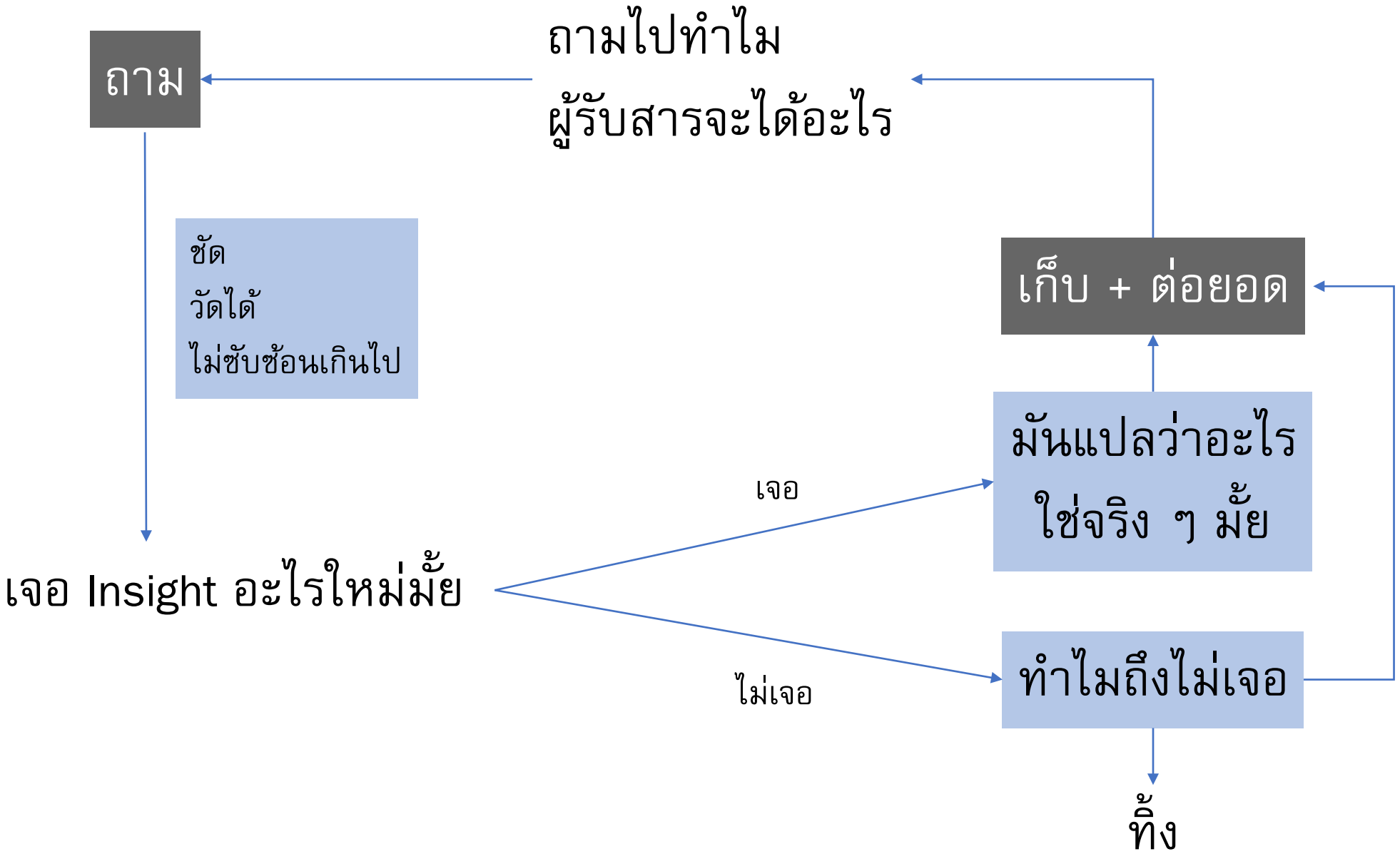
“เราจะเล่าเรื่องอะไรดีเพื่อให้เป็นประโยชน์แก่ผู้อ่าน?”

Asking Game

- ให้แต่ละทีมช่วยกันคิดว่าถ้าต้องการ สร้างประโยชน์จากข้อมูล แต่ละทีมจะต้อง ถามคำถามอะไรบ้าง? โดยจะต้องเป็นคำถามที่สามารถตอบได้โดยข้อมูลที่ให้มาเท่านั้น

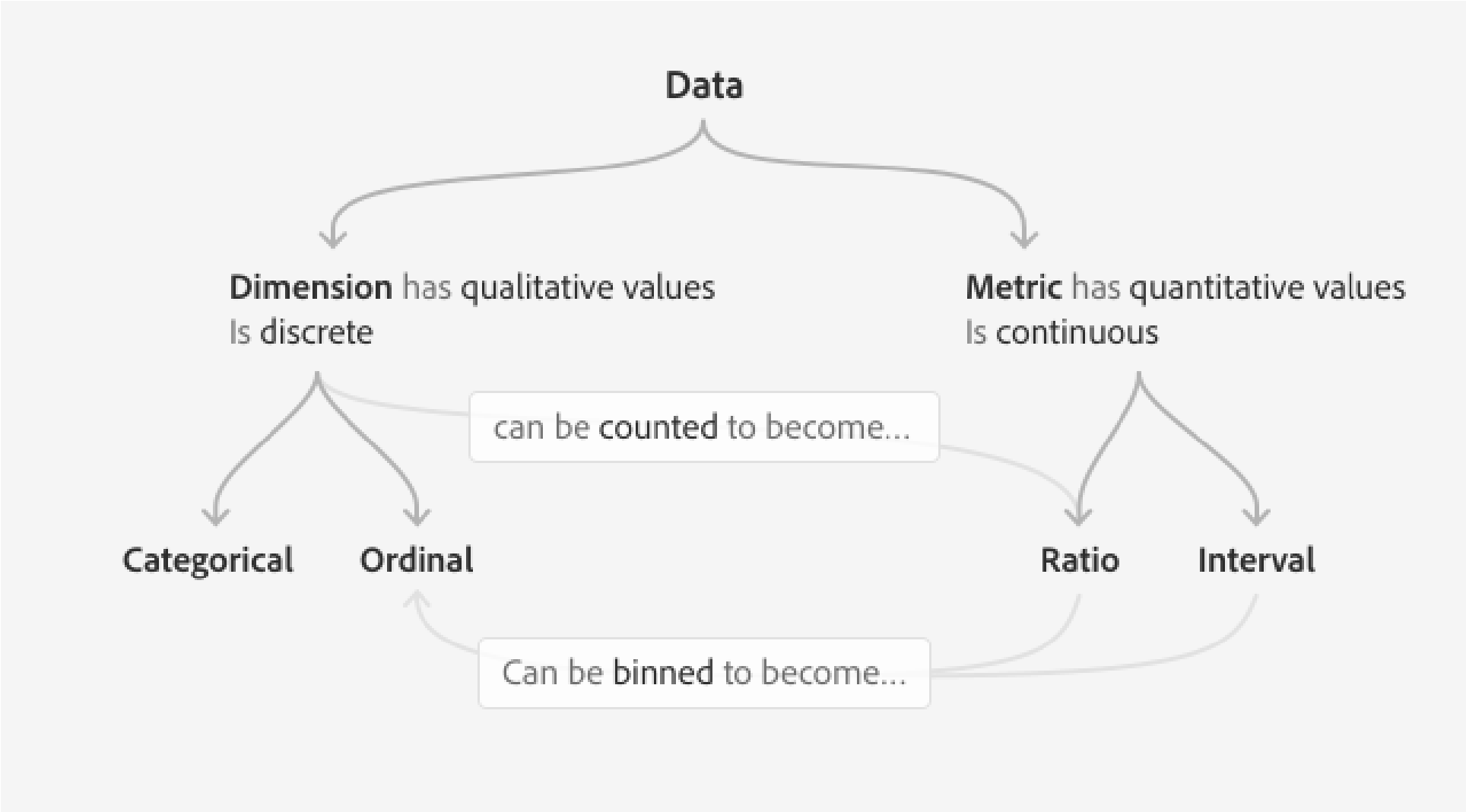
- สุ่มแต่ละทีมถามทีมละ 1 คำถามไม่ซ้ำกันเพื่อให้ผมตอบคำถามผ่าน Excel

- แล้วเราจะมาค้นพบกันว่า คำถามที่ดีและนำไปสู่เรื่องราวที่สร้างประโยชน์ได้จริง ควรเป็นคำถามแบบใด


ถาม
ถามไปทำไม
ผู้รับสารจะได้อะไร
ชัด
วัดได้
ไม่ซับซ้อนเกินไป
เก็บ + ต่อยอด
มันแปลว่าอะไร
ใช่จริง ๆ มั้ย
เจอ
เจอ Insight อะไรใหม่มั้ย
ไม่เจอ
ทำไมถึงไม่เจอ
ทิ้ง

Discussion Time

Digging Insights with Data



Source: https://spectrum.adobe.com/page/data-visualization-fundamentals/

SCALE	CATEGORICAL	ORDINAL	INTERVAL	RATIO
Example	**Country** (US, Japan, Mexico)	**Status** (Extinct, Endangered, Threatened)	**Temperature** (32°, 54°, 68°)	**Height** (1.65 m, 3.1 m, 2.01 m.)
The **ranking** of the values is known		x	x	x
Has a **mode** (most frequent value)	x	x	x	x
Has a **median** (middle value) Has a **percentile**		x	x	x
Has a **mean** (average value)			x	x
Can quantify the **difference** between values			x	x
Has a **spread** (standard deviation, variance)				x
Can multiply and divide values				x
Has a “true” zero				x

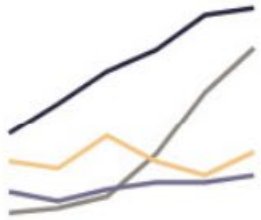
Source: https://spectrum.adobe.com/page/data-visualization-fundamentals/

กรอง	เจาะ	เทียบ	ชน

กรอง	เจาะ	เทียบ	ชน
เอาข้อมูลที่ “ไม่สำคัญ” หรือ “สกปรก” ออกไป	ดูตัวแปรเดียวอย่างละเอียด เช่น ความถี่ ผลรวม ค่าเฉลี่ย การกระจายข้อมูล ตำแหน่งข้อมูล	เปรียบเทียบระหว่างตัวแปรเดียวกันในหลายประเภท/ชนิด หรือดูสัดส่วน	ดูความสัมพันธ์ระหว่างตัวแปรสองตัวขึ้นไป

LINE CHART

Line charts are commonly used for time-series relationships with continuous data. They show trends, acceleration, deceleration, and volatility.



LINE

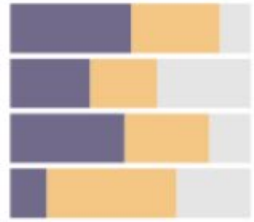
BAR CHART

Bar charts are best used for data with long category labels. Bar charts are usually used to compare different categories or parts of a whole.

BAR

STACKED

GROUPED

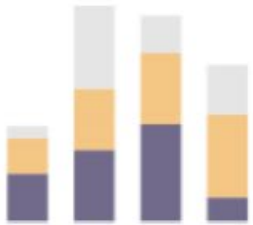


100% STACKED

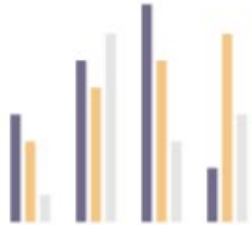
COLUMN CHART

Column charts are best used to show change over time (percentage variation), compare different categories, or compare parts of a whole.

COLUMN



STACKED



GROUPED

100% STACKED

Source: https://github.com/glosophy/CatoDataVizGuidelines/blob/master/PocketStyleBook.pdf

SCATTER PLOT

Scatter plots show the relationship between groups based on two dimensions. They are best used to show correlations between two sets of data.

SCATTER

GROUPED SCATTER

DOT PLOT

PIE CHART

Pie charts are best used for making part-to-whole comparisons with discrete or continuous data. They only do well when working with a small dataset.

PIE

DOUGHNUT

SEMICIRCLE

AREA CHART

Area charts show time-series relationships, but they are different than line charts in that they can also represent volume.

AREA

STACKED

100% STACKED

STREAM AREA

Source: https://github.com/glosophy/CatoDataVizGuidelines/blob/master/PocketStyleBook.pdf

MAPS

Maps can display both categorical or continuous data using intensity of color to represent values of geographic areas.

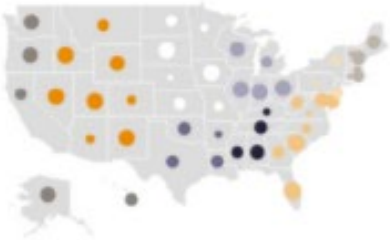
HEAT MAP

GROUPED

BUBBLE CHARTS

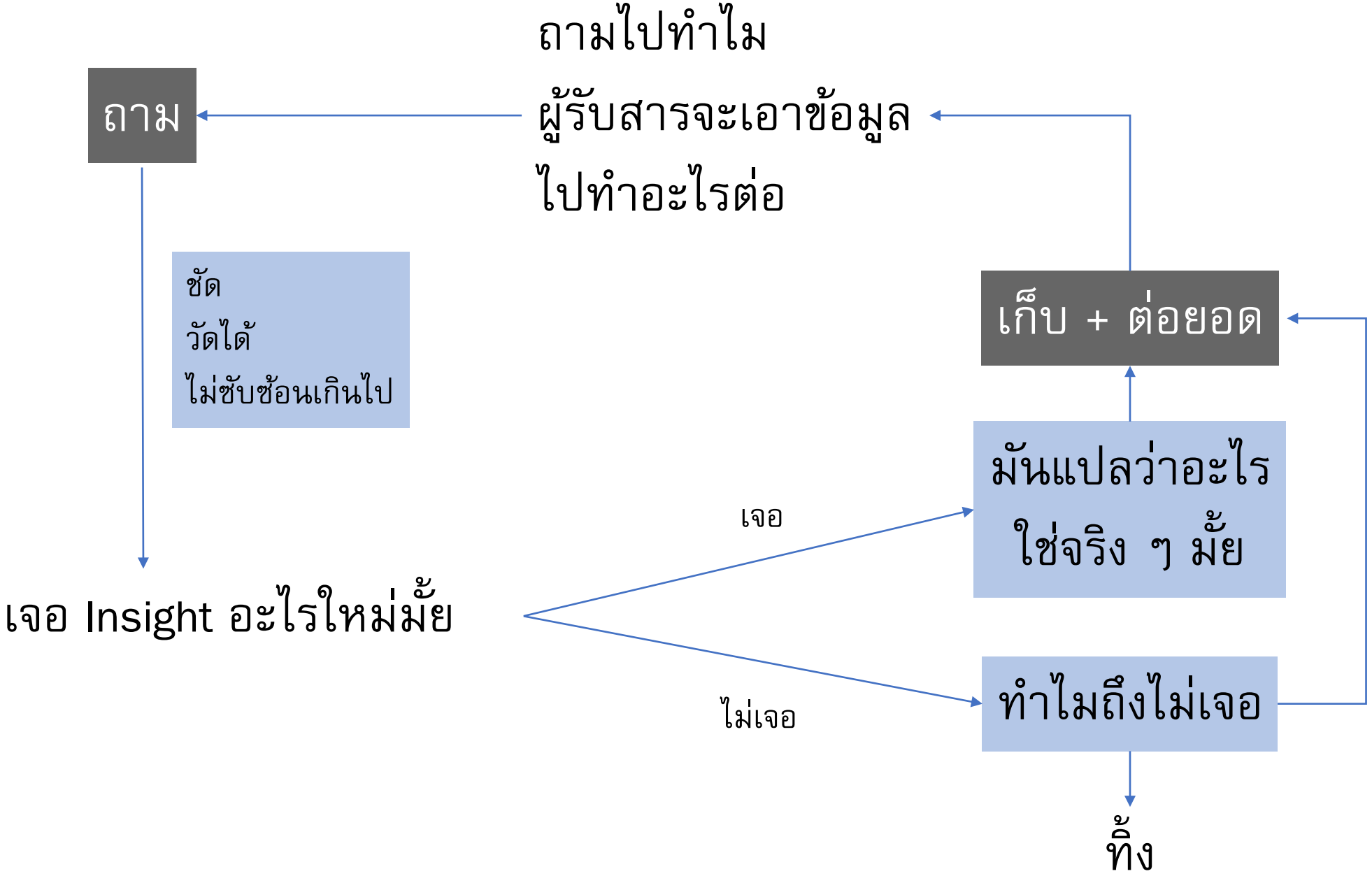
Bubble charts are good for showing nominal comparisons or ranking relationships.

BUBBLE PLOT



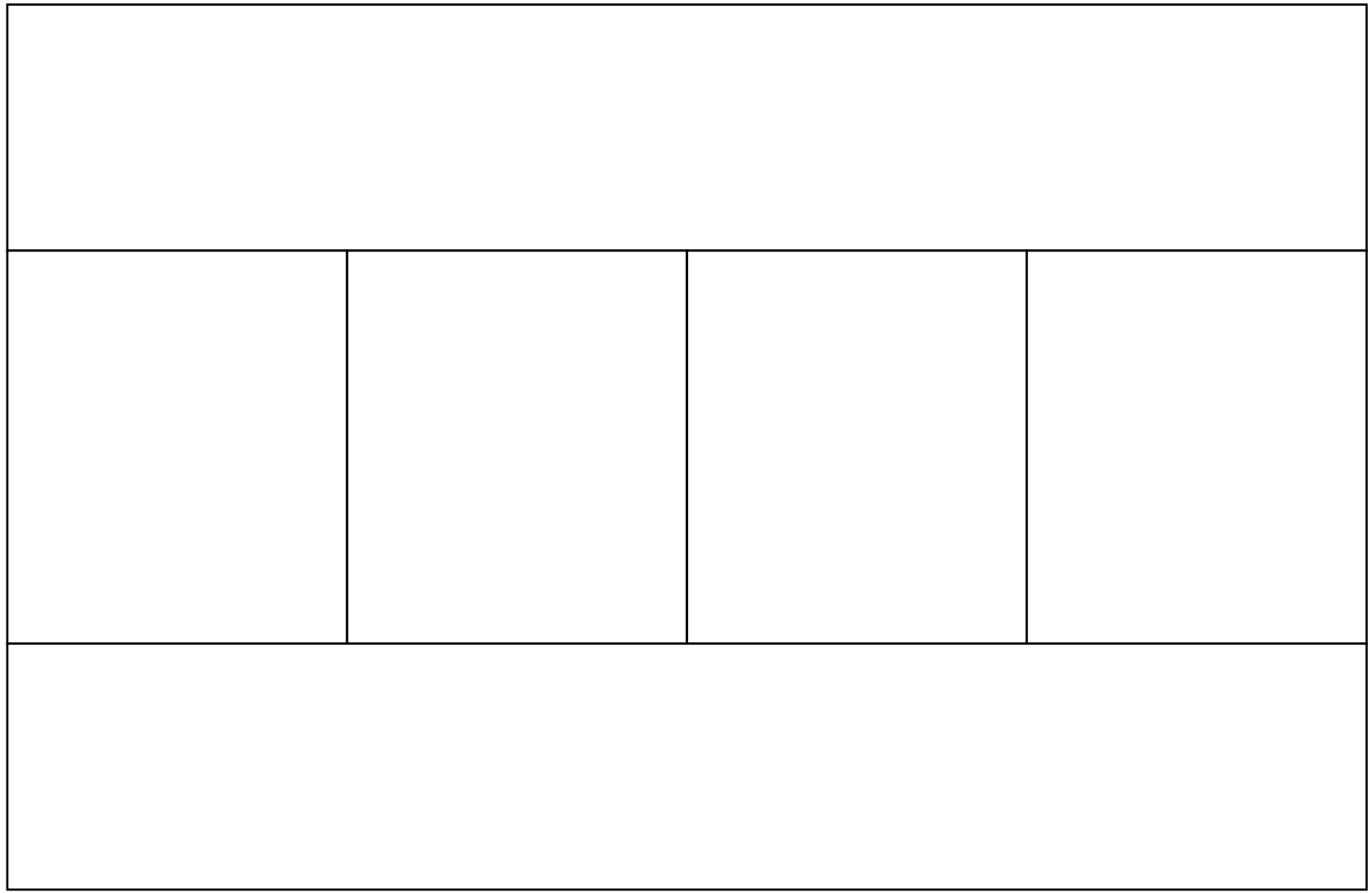
BUBBLE MAP

Source: https://github.com/glosophy/CatoDataVizGuidelines/blob/master/PocketStyleBook.pdf


ถามไปทำไม
ผู้รับสารจะเอาข้อมูล
ไปทำอะไรต่อ
ถาม
ชัด
วัดได้
ไม่ซับซ้อนเกินไป
เจอ Insight อะไรใหม่มั้ย
เจอ
ไม่เจอ
มันแปลว่าอะไร
ใช่จริง ๆ มั้ย
ทำไมถึงไม่เจอ
เก็บ + ต่อยอด
ทิ้ง

Data Value Canvas


Find freedom on this canvas.



Data Creator

Audience Value Generation

What's your audience's goal?

What insights they need to know to inform / make decisions?

How they can generate values using insights?

Data Acquisition

What

Where

When

Who

How

How Many

Data Integration

How "dirty"

is your data?

Analysis

List out exact

questions

and variables/models

needed

Potential bias and error

traps?

Delivery

Key visuals

Key statistics

Key takeaways

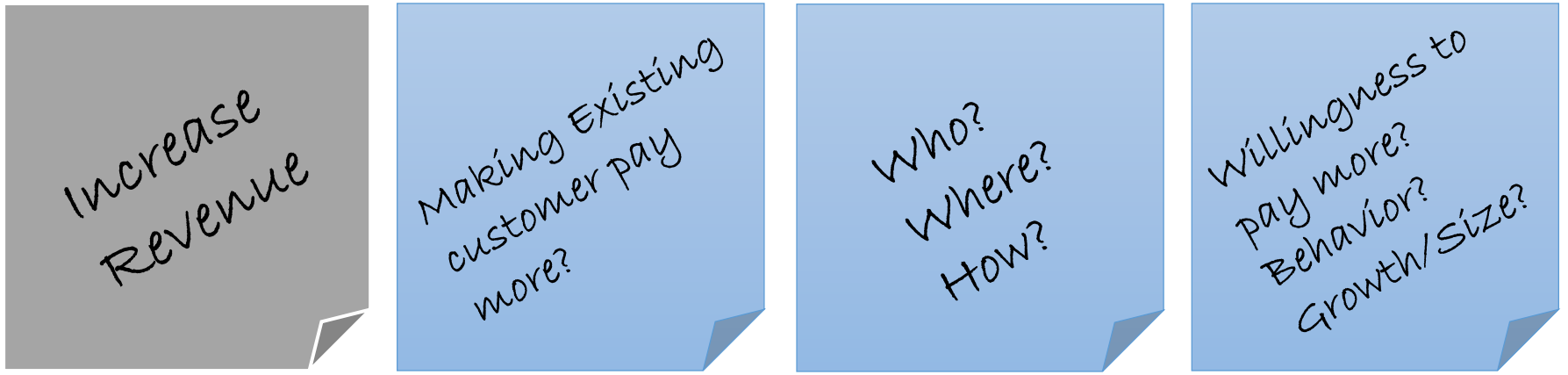
Data Governance

Does your organization promote behaviors for good data practice?

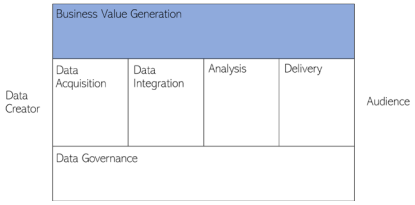
Data Principles / Quality / Privacy / Life Cycle / Ethical Use

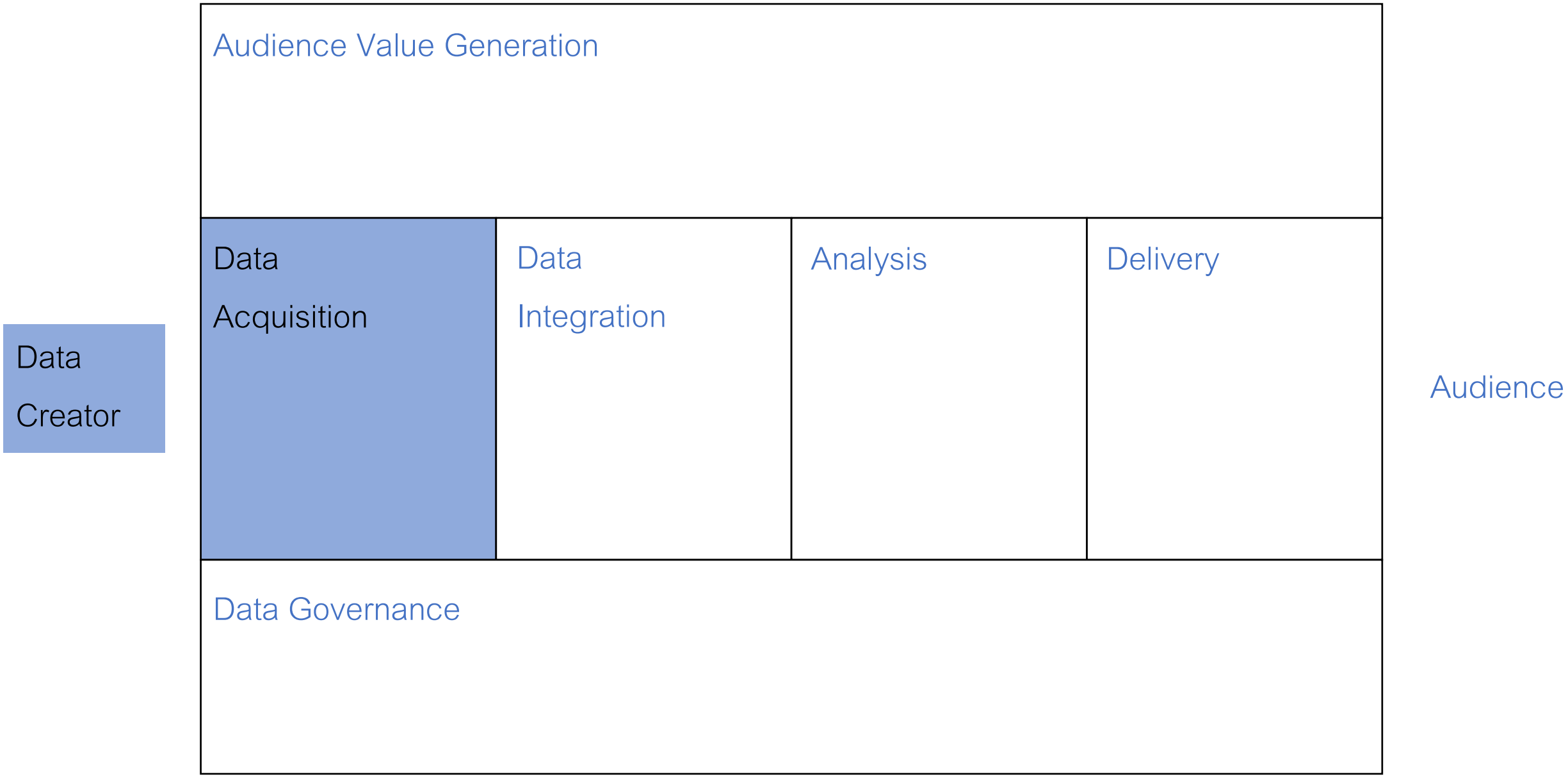
Audience

Audience Value Generation



Are these impactful / insightful / lead to audience use?




Audience Value Generation
Data
Acquisition
Data
Integration
Analysis
Delivery
Data Governance
Data
Creator
Audience

Data Creator

Internal / External?

Free / Costly to Acquire?

Reliable?

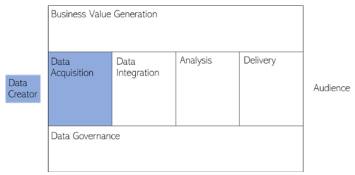
Survey

Third Party Database

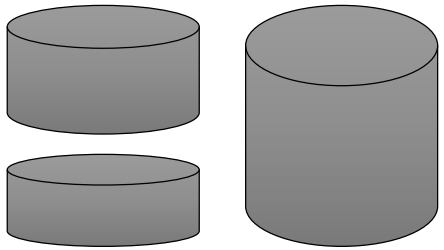
Web Scraping / Social Hearing

Others

Data Acquisition



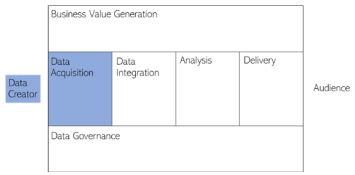
Data Acquisition

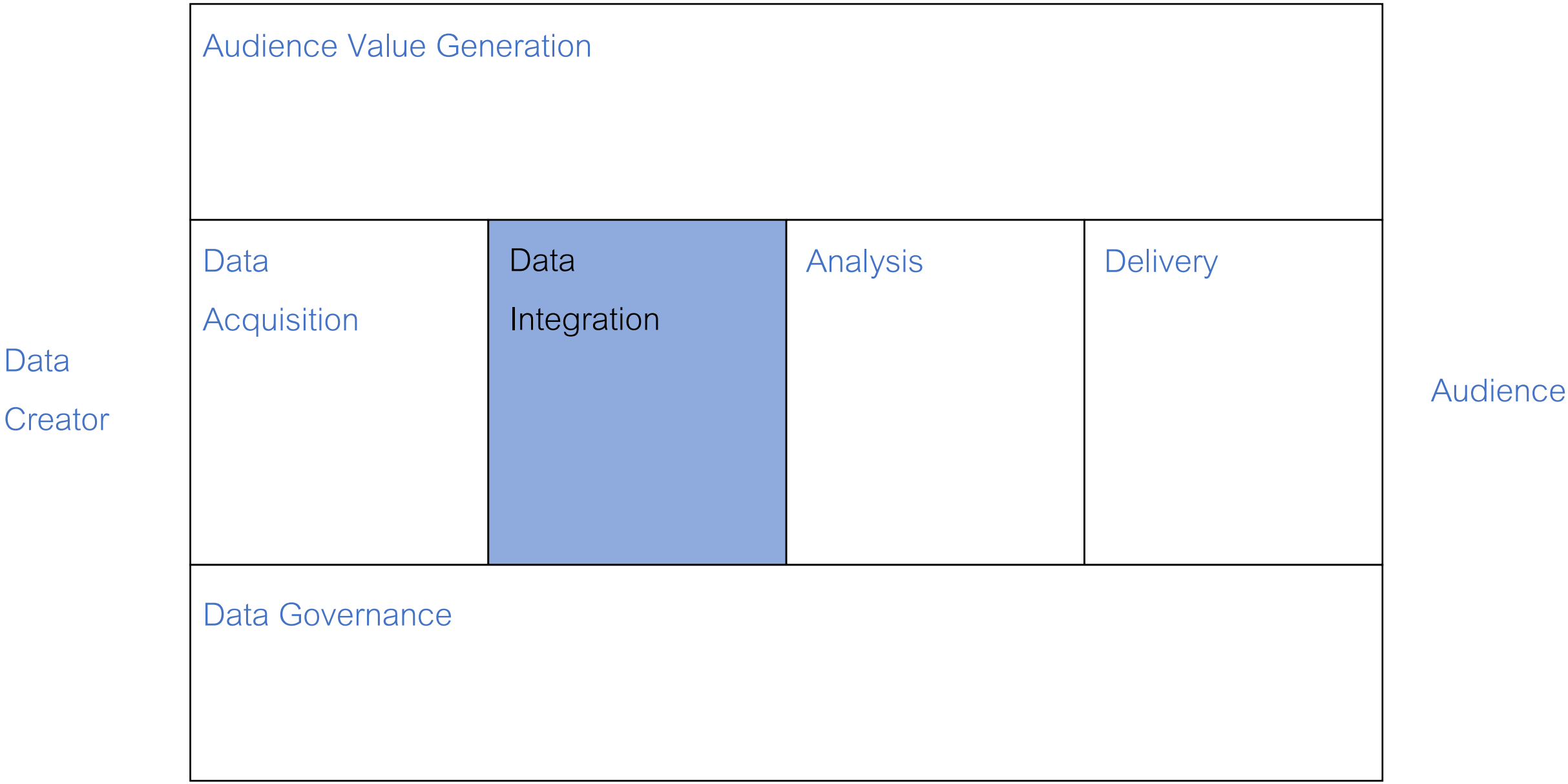


What Where When

How many

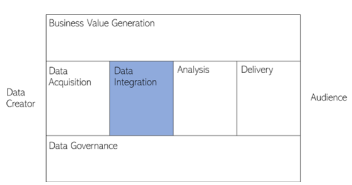
Ensure wide/deep/logical/relevant data acquisition process




Audience Value Generation
Data
Acquisition
Data
Integration
Analysis
Delivery
Data Governance
Data
Creator
Audience

Dirty Data

Clean Data



Typos?

Outliers?

Outdated?

Duplicates?

Dirty Data

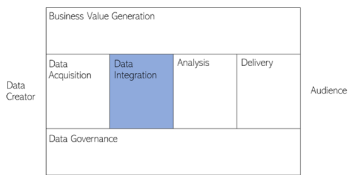
Clean Data

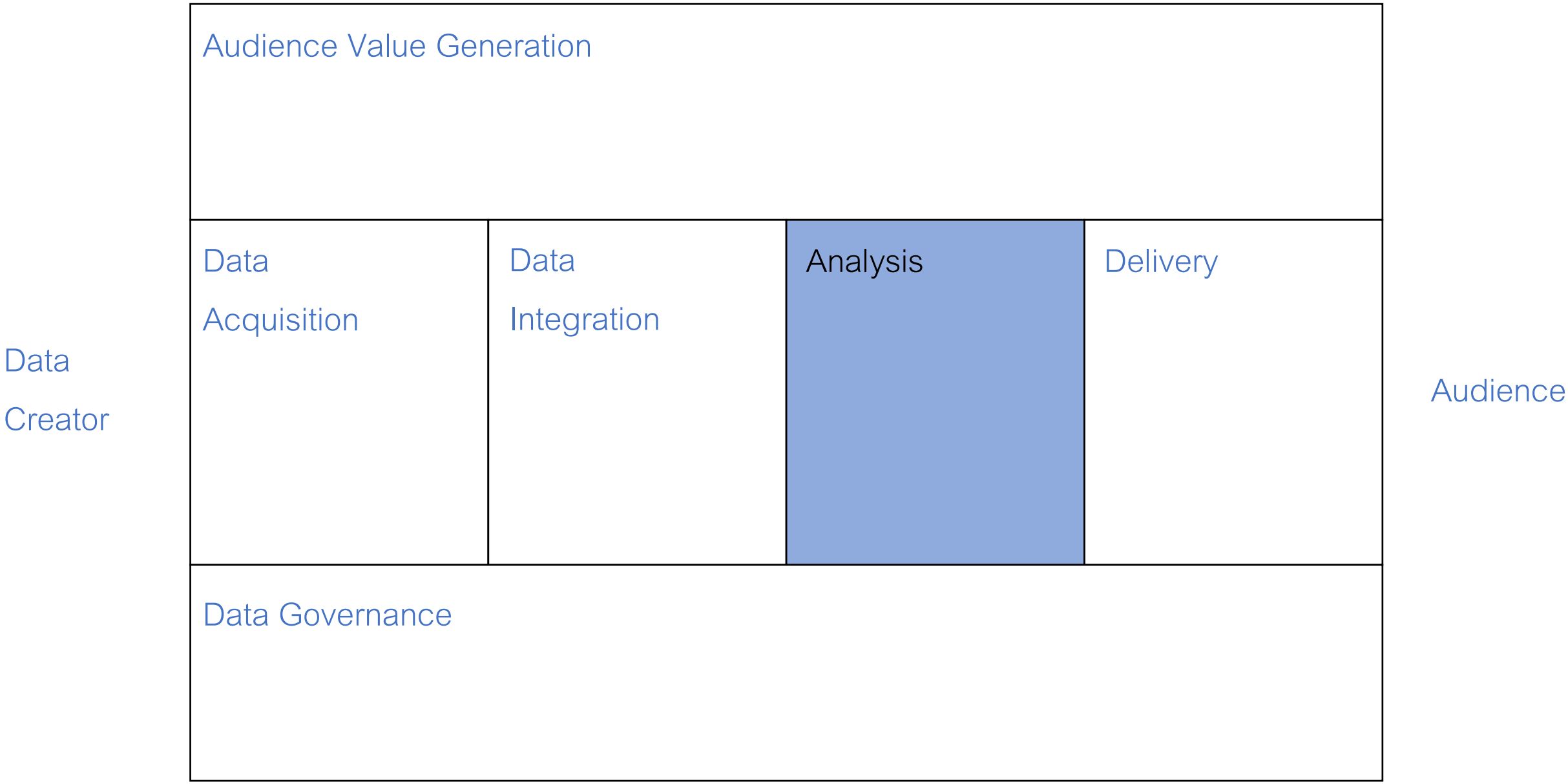
Incorrect Field?

Missing Values?

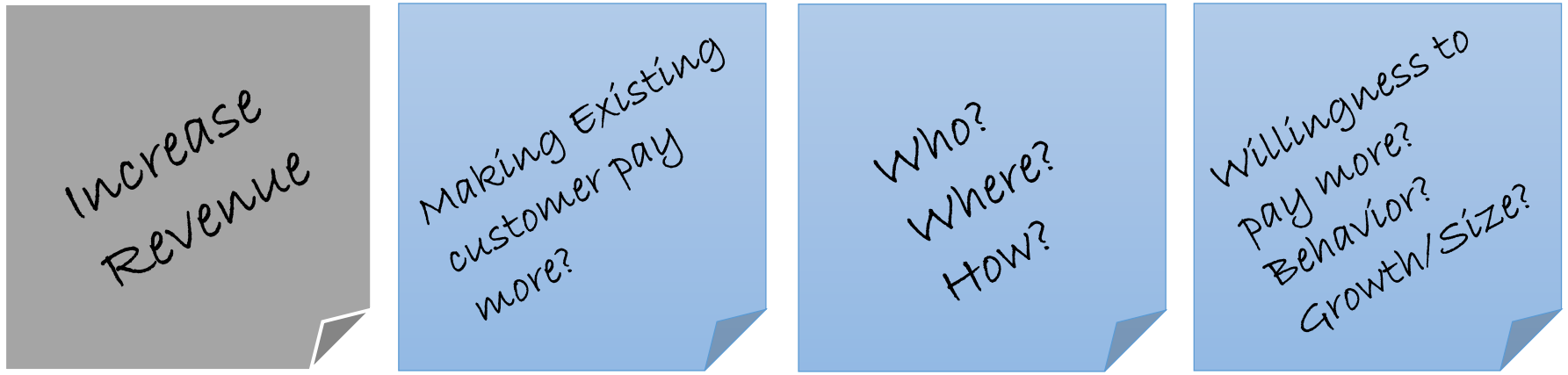
Integrated Properly?

Represent Your Target?

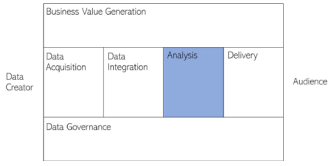



Audience Value Generation
Data
Acquisition
Data
Integration
Analysis
Delivery
Data Governance
Data
Creator
Audience

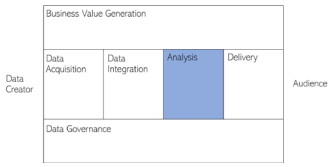
Analysis



List out the exact questions and variables needed to answer audience's questions



Human Bias Trap




$300


$4,000
$300

STEP 1: PICK YOUR PLAN

MAX

$199
a month

10,000 Minutes a month

- 3 Toll Free and/or Local Numbers
- UNLIMITED Existing Number Transfers
- UNLIMITED Voice Studio Services
- UNLIMITED Extensions
- ALL FEATURES INCLUDED

FREE Activation

MOST POPULAR

GROW

$49
a month

2,000 Minutes a month

- 2 Toll Free and/or Local Numbers
- 2 Existing Number Transfers
- UNLIMITED Extensions
- ALL FEATURES INCLUDED

$25 activation Fee

START

$9.95
a month

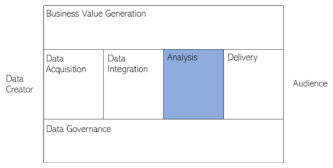
100 Minutes a month

- 1 Toll Free and/or Local Number
- UNLIMITED Extensions
- ALL FEATURES INCLUDED

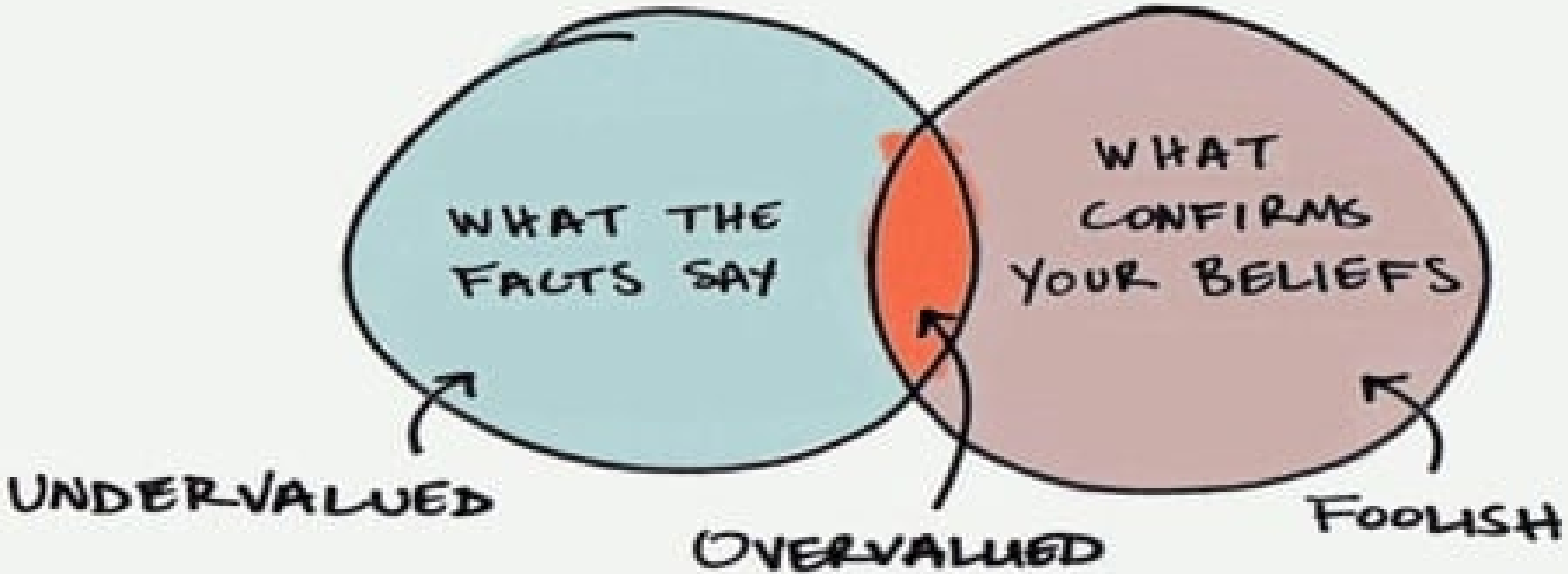
$25 activation Fee

Human Bias Trap

Anchoring Effect

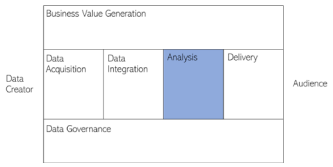


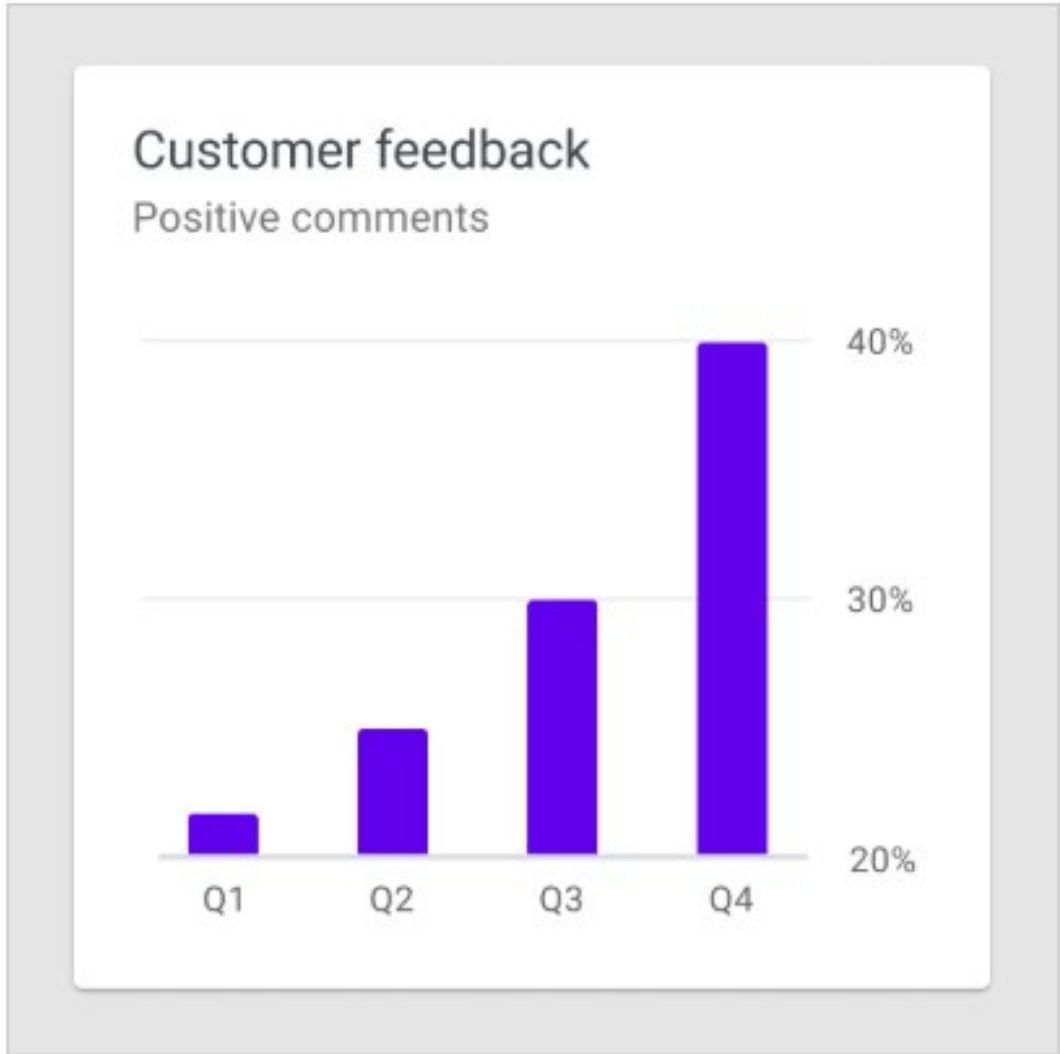
CONFIRMATION BIAS

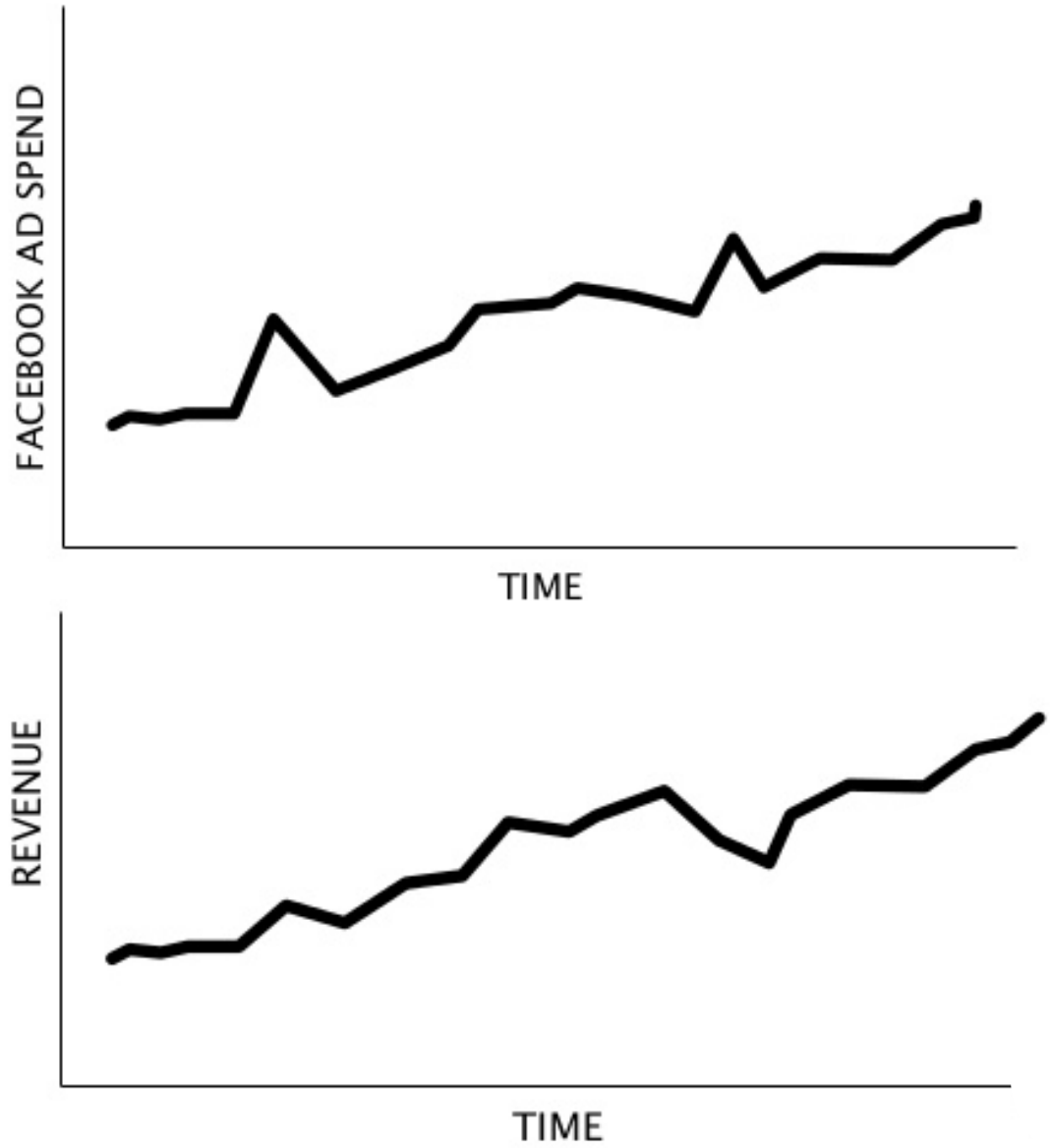


Human Bias Trap

Confirmation Bias

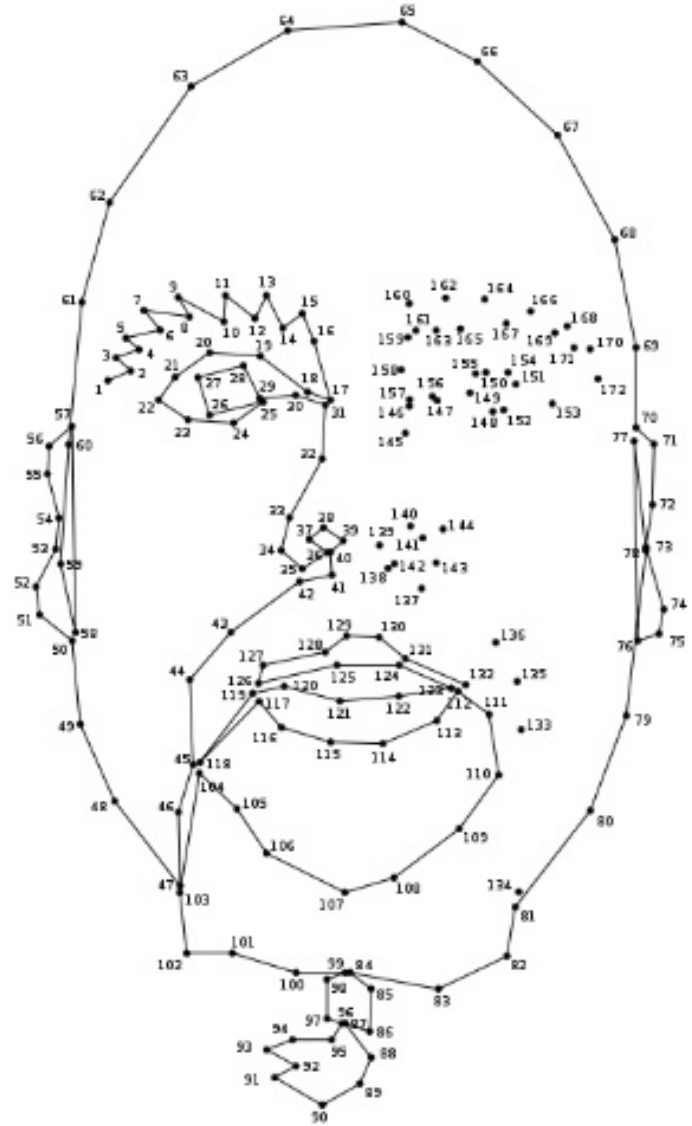



Customer feedback
Positive comments
40%
30%
20%
Q1
Q2
Q3
Q4


FACEBOOK AD SPEND
TIME
REVENUE
TIME

We see patterns where none exist

Clustering Illusion

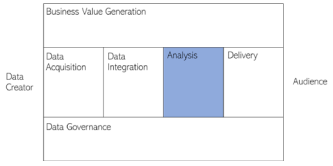


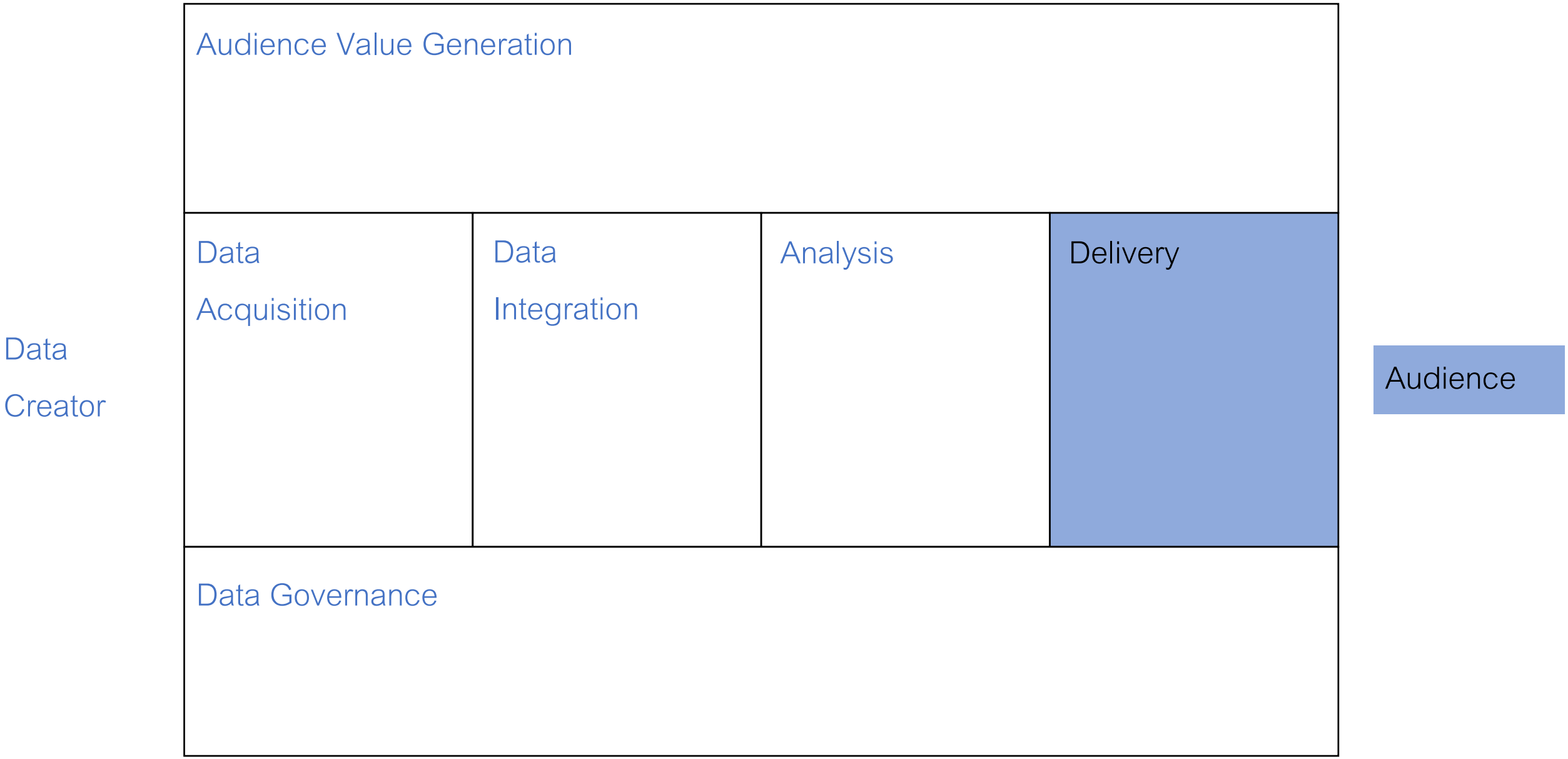
Human Bias Trap

Anchoring Effect

Confirmation Bias

Clustering Illusion




Audience Value Generation
Data
Acquisition
Data
Integration
Analysis
Delivery
Data Governance
Data
Creator
Audience

Delivery

Audience

“Know Your Data Audience”

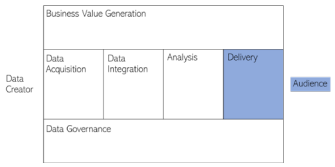
Millennials

Baby Boomers

Thai SMEs

Policymakers

Others



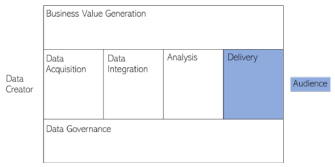
Delivery

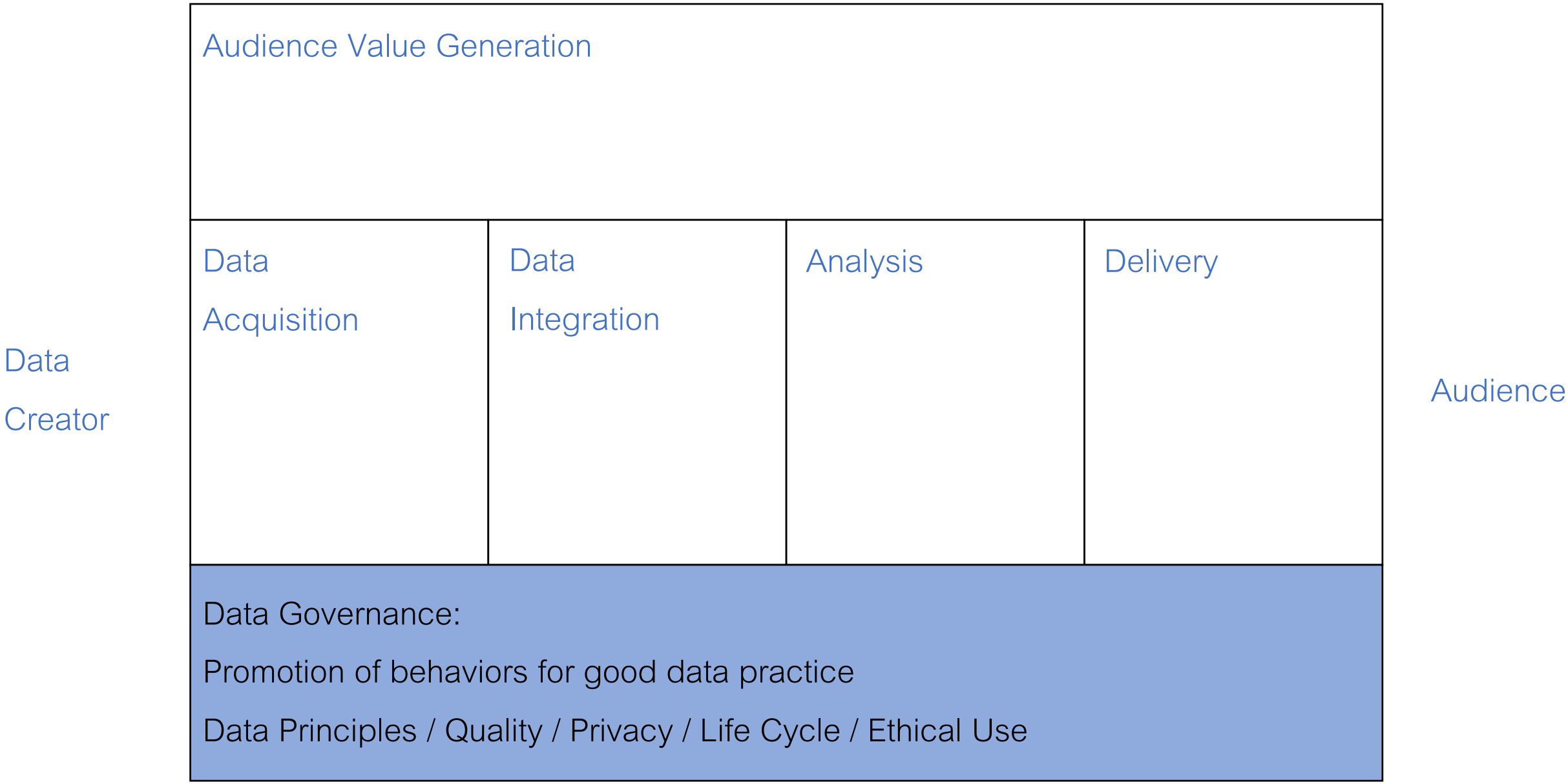
Audience

Key visuals

Key statistics

Key takeaways




Audience Value Generation
Data
Acquisition
Data
Integration
Analysis
Delivery
Data
Creator
Audience
Data Governance:
Promotion of behaviors for good data practice
Data Principles / Quality / Privacy / Life Cycle / Ethical Use

Question?

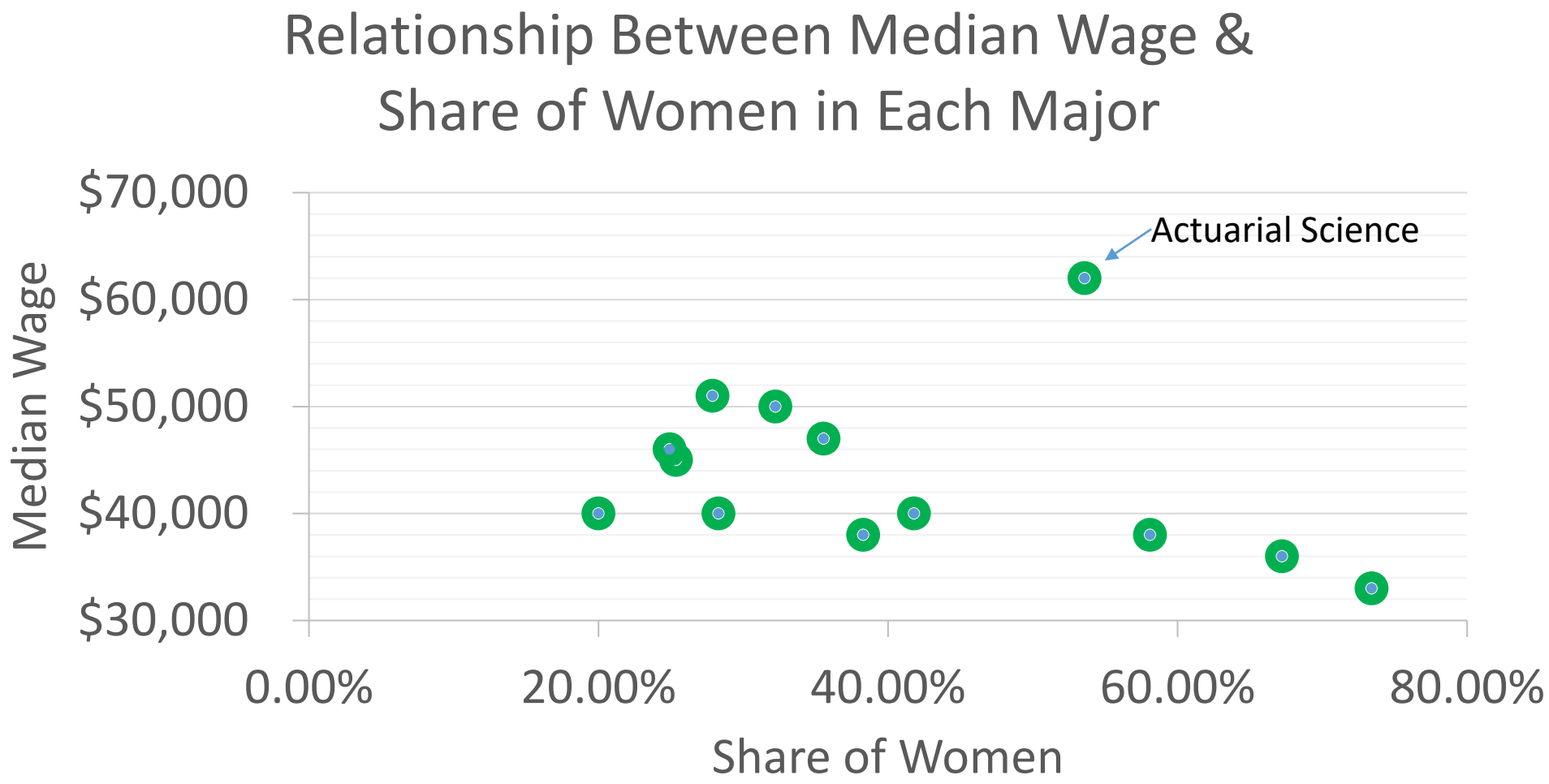
Dig Deeper Insights in your Data

Kids Explain Why Women Are Paid Less Than Men

Statistics of 2016 American Recent Graduates' Median Wage and Share of Women
By Their Major of Studies (Source: Pew Research Center)

Major	%Women	MedianWage
ACCOUNTING	25.36%	$ 45,000
ACTUARIAL SCIENCE	53.57%	$ 62,000
BUSINESS ECONOMICS	24.92%	$ 46,000
BUSINESS MANAGEMENT AND ADMINISTRATION	58.09%	$ 38,000
FINANCE	35.55%	$ 47,000
GENERAL BUSINESS	41.79%	$ 40,000
HOSPITALITY MANAGEMENT	73.40%	$ 33,000
HUMAN RESOURCES AND PERSONNEL MANAGEMENT	67.22%	$ 36,000
INTERNATIONAL BUSINESS	28.29%	$ 40,000
MANAGEMENT INFORMATION SYSTEMS AND STATISTICS	27.88%	$ 51,000
MARKETING AND MARKETING RESEARCH	38.29%	$ 38,000
MISCELLANEOUS BUSINESS & MEDICAL ADMINISTRATION	20.00%	$ 40,000
OPERATIONS LOGISTICS AND E-COMMERCE	32.22%	$ 50,000

Can you spot a relationship between %Women & Median Wage

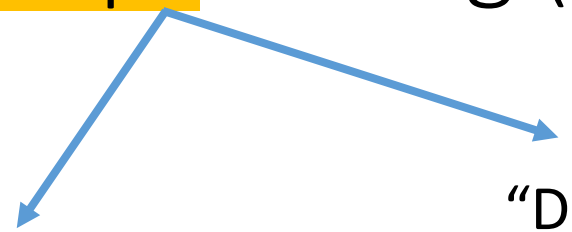


Can you spot a relationship between %Women & Median Wage

Case Study:
Do wage gap exist among female millennials?

Better Tools to Explore Data Insights?

Regression = A statistical process for estimating the relationships among (random) variables



"Magnitude"
- Large
- Small
- Zero

"Direction"
- Positive relationship
- Negative relationship
- No relationship at all?

Why regression?

- Regression is a standard way to show a relationship among variables
- Regression is flexible to model the relationship
- Regression is extremely useful in business & policymaking

Regression in 5 Steps

1. Frame the question & Form the hypotheses
2. Pick a regression model that could answer the question
3. Collect, clean, summarize & visualize the data
4. Use the model in (2) and data in (3) to run a regression
5. Evaluate & interpret the results

Iterate some or all of steps 1-5 if necessary

Case Study: Recent Grad's Median Wage

- Step 1: Frame the question & Form the hypotheses
  - Question: Does share of women in college major explains recent graduate's median wage?
  - Hypothesis: Share of women in college major is negatively related to recent graduate's median wage
    - Why?
    - Theory? Have other people looked into this? Your own observation?

Case Study: Recent Grad's Median Wage

- Step 2: Pick a regression model that could answer the question

X: Explanatory (Independent) Variable

  - Question: Does share of women in college major explains recent graduate's median wage?

Y: Dependent Variable

  - Linear Regression Model:

$$\mathbf{Y} = \boldsymbol{\alpha} + \boldsymbol{\beta}\boldsymbol{X} + \boldsymbol{\varepsilon}$$

Median Wage $= \boldsymbol{\alpha} +$ $\boldsymbol{\beta}$**(Share of women) +** **Other unknown variables that potentially explain Y but we can just ignore it for now**

Case Study: Recent Grad's Median Wage

- Step 2: Pick a regression model that could answer the question

- Hypothesis: Share of women in college major is negatively related to recent graduate's median wage

$$\mathbf{Y} = \boldsymbol{\alpha} + \boldsymbol{\beta X} + \boldsymbol{\varepsilon}$$

If $\boldsymbol{\beta}$ is ***negative & significant*** then we can conclude that our hypothesis is "believable"

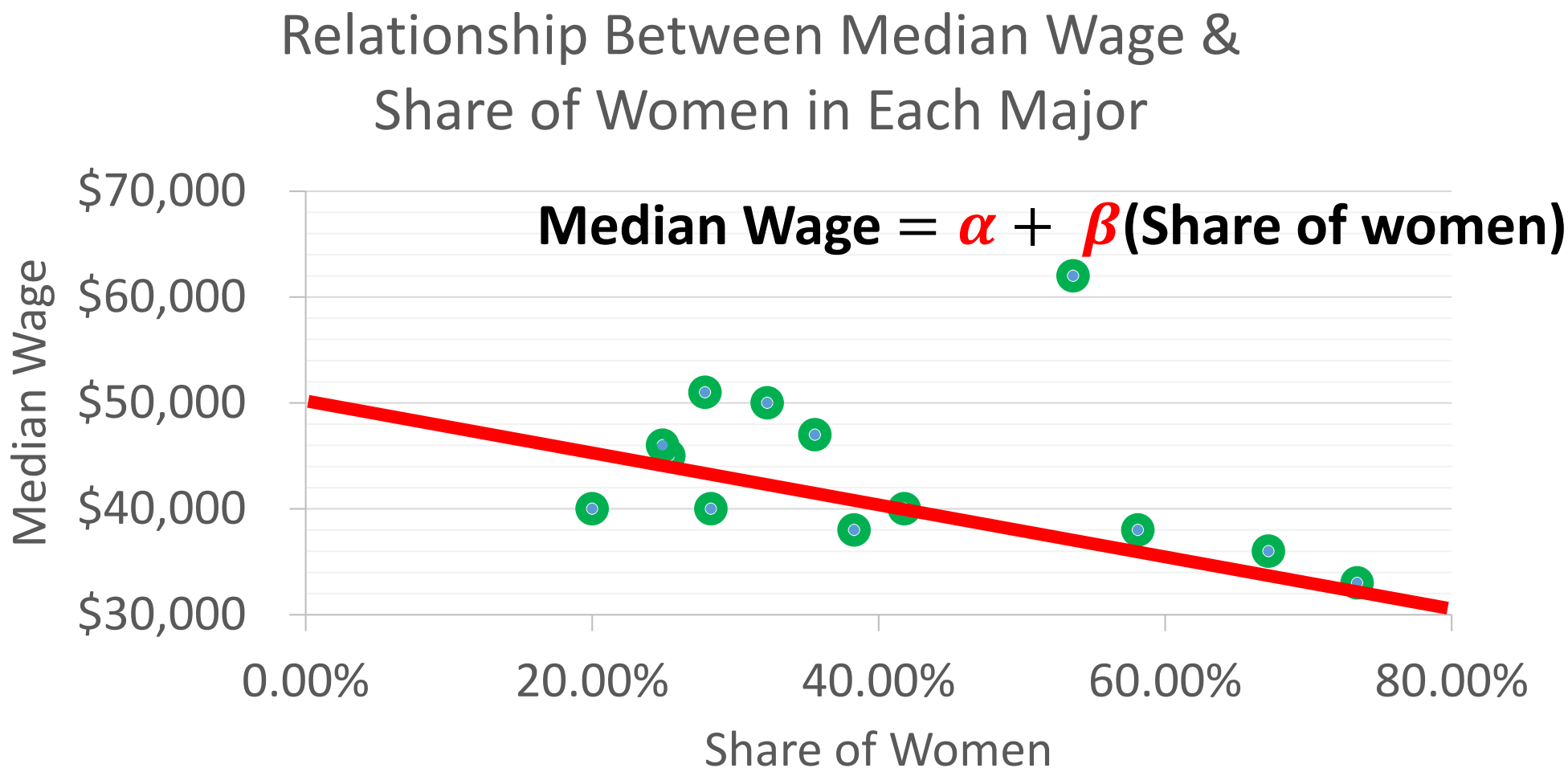
Step 3: Collect, clean, summarize & visualize the data

Statistics of 2016 American Recent Graduates' Median Wage and Share of Women
By Their Major of Studies

Major	%Women	MedianWage
ACCOUNTING	25.36%	$ 45,000
ACTUARIAL SCIENCE	53.57%	$ 62,000
BUSINESS ECONOMICS	24.92%	$ 46,000
BUSINESS MANAGEMENT AND ADMINISTRATION	58.09%	$ 38,000
FINANCE	35.55%	$ 47,000
GENERAL BUSINESS	41.79%	$ 40,000
HOSPITALITY MANAGEMENT	73.40%	$ 33,000
HUMAN RESOURCES AND PERSONNEL MANAGEMENT	67.22%	$ 36,000
INTERNATIONAL BUSINESS	28.29%	$ 40,000
MANAGEMENT INFORMATION SYSTEMS AND STATISTICS	27.88%	$ 51,000
MARKETING AND MARKETING RESEARCH	38.29%	$ 38,000
MISCELLANEOUS BUSINESS & MEDICAL ADMINISTRATION	20.00%	$ 40,000
OPERATIONS LOGISTICS AND E-COMMERCE	32.22%	$ 50,000

Source: Pew Research Center

Step 3: Collect, clean, summarize & visualize the data



Case Study: Recent Grad's Median Wage

- Step 4: Use the model in (2) and data in (3) to run a regression
  - Use Data Analysis >> Regression
  - Watch the videos again for detail

Case Study: Recent Grad's Median Wage

- Step 5: Evaluate & interpret the results

- You will consider **3** numbers

1. Coefficients
   - These are your parameters in the regression model (α, β)
2. P-Values
   - These will tell you the significance of the relationships (coefficients)
   - Normally, ***lower than 0.1 or 0.05*** is considered significant and useful.
3. Adjusted R Square
   - This will tell you an overall usefulness of the model. The higher Adjusted R Square, the better the model is.

Case Study: Recent Grad's Median Wage

- Regression Result

SUMMARY OUTPUT	
Regression Statistics	
Multiple R	0.3142
R Square	0.0988
Adjusted R Square	0.0168
Standard Error	7,708.39
Observations	13
ANOVA	

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	71618685.3	71618685.3	1.205310547	0.295713206
Residual	11	653612083.9	59419280.36		
Total	12	725230769.2			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	49270.2959	5641.6626	8.7333	0.0000	36853.0802	61687.5116	36853.0802	61687.5116
%Women	-14150.4697	12889.0639	-1.0979	0.2957	-42519.1081	14218.1687	-42519.1081	14218.1687

Regression Model:
Median Wage = **49270.2959 – 14150.4697** x (Share of Women)

The coefficient of share of women is ***negative but not significant.*** It appears that the share of women is ***negatively related*** with the median wage but the relationship is ***very weak*** and we ***should not believe*** that there is a negative relationship between the two variables.

The overall relationship is weak & insignificant coefficient suggesting ***iteration.***

Case Study: Recent Grad's Median Wage

- Regression Result without An Actuarial Science Major Data

SUMMARY OUTPUT	
Regression Statistics	
Multiple R	0.6984
R Square	0.4877
Adjusted R Square	0.4365
Standard Error	4,270.40
Observations	12

Regression Model:

Median Wage = **50918.5804 – 22625.9071** x (Share of Women)

Without the actuarial science major, the coefficient of share of women is ***negative and significant***. It is highly probable that the higher the share of women within a college major (except the actuarial science major), the lower the median wage. The overall significance is good i.e. ***Share of women can explain median wage 43.65%***

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	173636759.2	173636759.2	9.521478036	0.011528342
Residual	10	182363240.8	18236324.08		
Total	11	356000000			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	50918.5804	3142.2208	16.2046	0.0000	43917.2761	57919.8847	43917.2761	57919.8847
%Women	-22625.9071	7332.5294	-3.0857	0.0115	-38963.8009	-6288.0134	-38963.8009	-6288.0134

Case Study Summary

5-Step Regression	
Step 1: Question & Hypothesis	
Step 2: Data	
Step 3: Regression Model	
Step 4: Run the regression using Excel	
Step 5: Interpret & Evaluate the Results	

Review: Regression in 5 Steps

1. Frame the question & Form the hypotheses
2. Pick a regression model that could answer the question
3. Collect, clean, summarize & visualize the data
4. Use the model in (2) and data in (3) to run a regression
5. Evaluate & interpret the results

Iterate some or all of steps 1-5 if necessary

Regression Study Guide: No Textbook Needed!

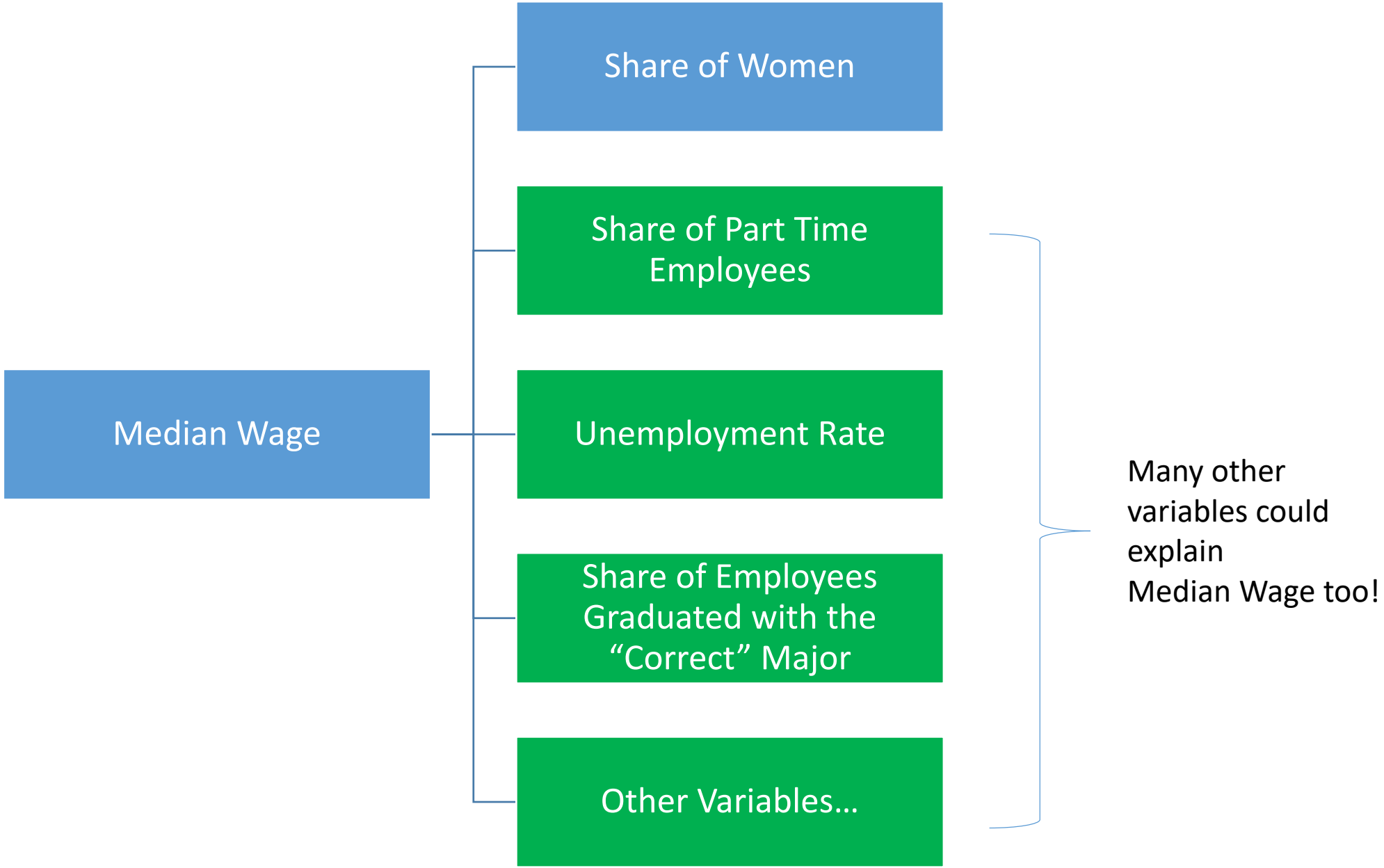
- Good background in regression analysis:
  - *การวิเคราะห์ถดถอย* ***Simple Regression***
    - https://www.youtube.com/watch?v=qE_6D4gScs8
- Must-See Videos:
  - *การวิเคราะห์ถดถอยพหุ # แนวคิดเบื้องต้น*
    - https://www.youtube.com/watch?v=jEp-0m8_89k
- Optional Lecture Note:
  - Penn State's ***Simple Linear Regression Lesson 1***
    - https://onlinecourses.science.psu.edu/stat501/node/250

Basic Regression: Extended

Extension to the Simple Linear Regression Model

Simple Linear Regression

$$\mathbf{Y} = \boldsymbol{\alpha} + \boldsymbol{\beta}\boldsymbol{X} + \boldsymbol{\varepsilon}$$


Median Wage
Share of Women
Share of Part Time Employees
Unemployment Rate
Share of Employees Graduated with the "Correct" Major
Other Variables…
Many other variables could explain Median Wage too!

We want to simultaneously explain Median Wage with

(1) %Women

(2) %PartTime

(3) %Umemployed

(4) %CollegeJob

Use explanatory (independent) variables… **to explain dependent variables**

Major	Major_category	%Women	%PartTime	%Unemployed	%CollegeJob	MedianWage
MISCELLANEOUS FINE ARTS	Arts	0.4102	0.3195	0.0894	0.2192	50000
COMMERCIAL ART AND GRAPHIC DESIGN	Arts	0.3744	0.2357	0.0968	0.4138	35000
FILM VIDEO AND PHOTOGRAPHIC ARTS	Arts	0.687	0.3307	0.1058	0.217	32000
MUSIC	Arts	0.4446	0.4114	0.076	0.2654	31000
FINE ARTS	Arts	0.667	0.3178	0.0842	0.3179	30500
VISUAL AND PERFORMING ARTS	Arts	0.6974	0.3848	0.1022	0.2687	30000
STUDIO ARTS	Arts	0.5848	0.3342	0.0896	0.2431	29000
DRAMA AND THEATER ARTS	Arts	0.6295	0.3698	0.0775	0.1612	27000
ACTUARIAL SCIENCE	Business	0.5357	0.0784	0.0957	0.7552	62000
MANAGEMENT INFORMATION SYSTEMS AND STATISTICS	Business	0.2788	0.1293	0.0582	0.4958	51000
OPERATIONS LOGISTICS AND E-COMMERCE	Business	0.3222	0.1008	0.0479	0.2725	50000
FINANCE	Business	0.3555	0.123	0.0607	0.2935	47000
BUSINESS ECONOMICS	Business	0.2492	0.1456	0.0964	0.2111	46000
ACCOUNTING	Business	0.2536	0.1394	0.0697	0.1853	45000
GENERAL BUSINESS	Business	0.4179	0.1545	0.0729	0.1863	40000
MISCELLANEOUS BUSINESS & MEDICAL ADMINISTRATION	Business	0.2	0.1876	0.072	0.1729	40000
INTERNATIONAL BUSINESS	Business	0.2829	0.1888	0.0962	0.2126	40000
BUSINESS MANAGEMENT AND ADMINISTRATION	Business	0.5809	0.1526	0.0722	0.1688	38000
MARKETING AND MARKETING RESEARCH	Business	0.3829	0.1746	0.0612	0.172	38000
HUMAN RESOURCES AND PERSONNEL MANAGEMENT	Business	0.6722	0.1538	0.0596	0.1726	36000
HOSPITALITY MANAGEMENT	Business	0.734	0.1717	0.0612	0.0669	33000

Remember: Regression in 5 Steps

1. Frame the question & Form the hypotheses
2. Pick a regression model that could answer the question
3. Collect, clean, summarize & visualize the data
4. Use the model in (2) and data in (3) to run a regression
5. Evaluate & interpret the results

Iterate some or all of steps 1-5 if necessary

Regression in 5 Steps

1. Frame the question & Form the hypotheses

   - Question
     -
   - Hypotheses
     -
     -
     -
     -

Regression in 5 Steps

2. Pick a regression model that could answer the question

- Your expectation:
  -
  -
  -
  -

Regression in 5 Steps

3. Collect, clean, summarize & visualize the data
   - Use Data Analysis &…
     - Descriptive Statistics
     - Scatter Plot
     - Correlation ***
     - Etc.
   - Watch the videos for detail

Correlation (r) = Show how variables are related

Correlation Formula
(It is just another summary statistics like mean, variance etc.)

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

How to compute correlation in Excel:
Data >> Data Analysis >> Correlation

	%Women	*%PartTime*	*%Unemployed*	*%CollegeJob*	*MedianWage*
%Women	100.00%				
%PartTime	45.72%	100.00%			
%Unemployed	20.35%	48.66%	100.00%		
%CollegeJob	-10.29%	-22.19%	23.17%	100.00%	
MedianWage	-54.89%	-71.98%	-18.02%	57.54%	100.00%

r is between -1 and 1

r > 0: Positive Relationship

r< 0: Negative Relationship

IMPORTANT FOOTNOTE

Correlation ≠ Causation

Y & X move together
(same or different direction)

You know whether
X causes Y or
Y causes X
**IT'S A BOLD CLAIM!!

IMPORTANT FOOTNOTE

IMPORTANT FOOTNOTE

Question: Is regression better than correlation?

- ***Answer: There are pros and cons***

Regression	Correlation
A little harder to estimate	Simpler to estimate
Show both direction & magnitude	Show direction only
Can show many-to-one relationships	One-to-one relationship only
Can disentangle other indirect or joint relationships	Cannot disentangle other indirect or joint relationships

IMPORTANT FOOTNOTE

Question: Can a regression model capture 'causation'?

- ***Answer: Not necessary***
- The usual regression model doesn't capture causation.
- You need a careful regression modeling design to capture the causation effect
  - Difference-in-difference design
  - Regression discontinuity design
  - Instrument variable (IV design)
  - Random experiment design… just like in the science lab experiment!
  - Structural modeling design

My doctoral research!

- But that doesn't mean regression or correlation isn't useful

Result with One Explanatory Variable

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.55					
R Square	0.30					
Adjusted R Square	0.26					
Standard Error	7751.81					
Observations	21					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	1	492231698	492231698	8.191497623	0.00997496	
Residual	19	1141720683	60090562.26			
Total	20	1633952381				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	52843.66	4943.98	10.69	0.00	42495.79	63191.53
%Women	-28606.24	9994.91	-2.86	0.01	-49525.84	-7686.64

Regression Model:
Median Wage = 52843.66 – 28606.24 x (%Women)

Looking at the data from Arts & Business major recent grads, the "share of women" coefficient is highly significant & negative as before. The Adjusted R Square = 0.26 meaning these variables can explain median wage 26%.

Result with All 4 Explanatory Variables

Regression Model:

Median Wage = 48045.74 – 14409.18 x (%Women)
– 44565.57 x (%Part Time)
+ 13749.36 x (%Unemployed)
+ 26451.96 x (%College Job)

The regression coefficients are highly significant and follow the hypotheses ***except %Umemployed***, however, its coefficient is not significant (p-value = 0.87; higher than 0.1). Adjusted R Square = 0.70 meaning these variables can explain median wage 70%.

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.87					
R Square	0.76					
Adjusted R Square	0.70					
Standard Error	4940.93					
Observations	21					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	4	1243347729	310836932.2	12.73254399	7.55929E-05	
Residual	16	390604652	24412790.75			
Total	20	1633952381				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	48045.74	5768.01	8.33	0.00	35818.11	60273.37
%Women	-14409.18	7165.69	-2.01	0.06	-29599.77	781.41
%PartTime	-44565.57	14438.09	-3.09	0.01	-75172.94	-13958.20
%Unemployed	13749.36	81970.21	0.17	0.87	-160019.72	187518.44
%CollegeJob	26451.96	8456.37	3.13	0.01	8525.25	44378.67

Want to Improve the Adjusted R Square? Try dropping the insignificant variable(s)

Regression Model:

Median Wage = 48707.90 – 14440.45 x (%Women)
– 43278.11 x (%Part Time)
+ 27017.69 x (%College Job)
+ 26451.96 x (%College Job)

Once removed the %Umemployed, the Adjusted R Square improves (before 0.70, now = 0.7183) meaning these variables can explain median wage 71.83%... A little better than the previous model

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.8721
R Square	0.7605
Adjusted R Square	0.7183
Standard Error	4797.6190
Observations	21

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	1242660866	414220288.6	17.99616049	1.61034E-05
Residual	17	391291515.2	23017147.95		
Total	20	1633952381			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	48707.90	4083.50	11.93	0.00	40092.48	57323.33
%Women	-14440.45	6955.50	-2.08	0.05	-29115.27	234.37
%PartTime	-43278.11	11874.15	-3.64	0.00	-68330.38	-18225.84
%CollegeJob	27017.69	7529.75	3.59	0.00	11131.30	42904.07

Extension to the Simple Linear Regression Model

Simple Linear Regression

$$\mathbf{Y} = \boldsymbol{\alpha} + \boldsymbol{\beta}\boldsymbol{X} + \boldsymbol{\varepsilon}$$

When you have more than one X

$$\mathbf{Y} = \boldsymbol{\alpha} + \boldsymbol{\beta}_1\boldsymbol{X}_1 + \boldsymbol{\beta}_2\boldsymbol{X}_2 + \boldsymbol{\varepsilon}$$

We want to explain Median Wage with

(1) %Women,

(2) %PartTime,

(3) %Umemployed,

(4) %CollegeJob

(5) Major Category… not a number!!!

Use explanatory (independent) variables… **to explain dependent variables**

Major	Major_category	%Women	%PartTime	%Unemployed	%CollegeJob	MedianWage
MISCELLANEOUS FINE ARTS	Arts	0.4102	0.3195	0.0894	0.2192	50000
COMMERCIAL ART AND GRAPHIC DESIGN	Arts	0.3744	0.2357	0.0968	0.4138	35000
FILM VIDEO AND PHOTOGRAPHIC ARTS	Arts	0.687	0.3307	0.1058	0.217	32000
MUSIC	Arts	0.4446	0.4114	0.076	0.2654	31000
FINE ARTS	Arts	0.667	0.3178	0.0842	0.3179	30500
VISUAL AND PERFORMING ARTS	Arts	0.6974	0.3848	0.1022	0.2687	30000
STUDIO ARTS	Arts	0.5848	0.3342	0.0896	0.2431	29000
DRAMA AND THEATER ARTS	Arts	0.6295	0.3698	0.0775	0.1612	27000
ACTUARIAL SCIENCE	Business	0.5357	0.0784	0.0957	0.7552	62000
MANAGEMENT INFORMATION SYSTEMS AND STATISTICS	Business	0.2788	0.1293	0.0582	0.4958	51000
OPERATIONS LOGISTICS AND E-COMMERCE	Business	0.3222	0.1008	0.0479	0.2725	50000
FINANCE	Business	0.3555	0.123	0.0607	0.2935	47000
BUSINESS ECONOMICS	Business	0.2492	0.1456	0.0964	0.2111	46000
ACCOUNTING	Business	0.2536	0.1394	0.0697	0.1853	45000
GENERAL BUSINESS	Business	0.4179	0.1545	0.0729	0.1863	40000
MISCELLANEOUS BUSINESS & MEDICAL ADMINISTRATION	Business	0.2	0.1876	0.072	0.1729	40000
INTERNATIONAL BUSINESS	Business	0.2829	0.1888	0.0962	0.2126	40000
BUSINESS MANAGEMENT AND ADMINISTRATION	Business	0.5809	0.1526	0.0722	0.1688	38000
MARKETING AND MARKETING RESEARCH	Business	0.3829	0.1746	0.0612	0.172	38000
HUMAN RESOURCES AND PERSONNEL MANAGEMENT	Business	0.6722	0.1538	0.0596	0.172	36000
HOSPITALITY MANAGEMENT	Business	0.734	0.1717	0.0612	0.0669	33000

How?
Convert Text to Number using Dummy Variable

N types = N dummy variables

Major_category	Business_Dummy	Arts_Dummy	%Women	%PartTime	%Unemployed	%CollegeJob	MedianWage
Arts	0	1	0.4102	0.3195	0.0894	0.2192	50000
Arts	0	1	0.3744	0.2357	0.0968	0.4138	35000
Arts	0	1	0.687	0.3307	0.1058	0.217	32000
Arts	0	1	0.4446	0.4114	0.076	0.2654	31000
Arts	0	1	0.667	0.3178	0.0842	0.3179	30500
Arts	0	1	0.6974	0.3848	0.1022	0.2687	30000
Arts	0	1	0.5848	0.3342	0.0896	0.2431	29000
Arts	0	1	0.6295	0.3698	0.0775	0.1612	27000
Business	1	0	0.5357	0.0784	0.0957	0.7552	62000
Business	1	0	0.2788	0.1293	0.0582	0.4958	51000
Business	1	0	0.3222	0.1008	0.0479	0.2725	50000
Business	1	0	0.3555	0.123	0.0607	0.2935	47000
Business	1	0	0.2492	0.1456	0.0964	0.2111	46000
Business	1	0	0.2536	0.1394	0.0697	0.1853	45000
Business	1	0	0.4179	0.1545	0.0729	0.1863	40000
Business	1	0	0.2	0.1876	0.072	0.1729	40000
Business	1	0	0.2829	0.1888	0.0962	0.2126	40000
Business	1	0	0.5809	0.1526	0.0722	0.1688	38000
Business	1	0	0.3829	0.1746	0.0612	0.172	38000
Business	1	0	0.6722	0.1538	0.0596	0.1726	36000
Business	1	0	0.734	0.1717	0.0612	0.0669	33000

*** BUT when you run regression, only include N-1 dummy variables in your regression model

Major_category	Business_Dummy	Arts_Dummy	%Women	%PartTime	%Unemployed	%CollegeJob	MedianWage
Arts	0	1	0.4102	0.3195	0.0894	0.2192	50000
Arts	0	1	0.3744	0.2357	0.0968	0.4138	35000
Arts	0	1	0.687	0.3307	0.1058	0.217	32000
Arts	0	1	0.4446	0.4114	0.076	0.2654	31000
Arts	0	1	0.667	0.3178	0.0842	0.3179	30500
Arts	0	1	0.6974	0.3848	0.1022	0.2687	30000
Arts	0	1	0.5848	0.3342	0.0896	0.2431	29000
Arts	0	1	0.6295	0.3698	0.0775	0.1612	27000
Business	1	0	0.5357	0.0784	0.0957	0.7552	62000
Business	1	0	0.2788	0.1293	0.0582	0.4958	51000
Business	1	0	0.3222	0.1008	0.0479	0.2725	50000
Business	1	0	0.3555	0.123	0.0607	0.2935	47000
Business	1	0	0.2492	0.1456	0.0964	0.2111	46000
Business	1	0	0.2536	0.1394	0.0697	0.1853	45000
Business	1	0	0.4179	0.1545	0.0729	0.1863	40000
Business	1	0	0.2	0.1876	0.072	0.1729	40000
Business	1	0	0.2829	0.1888	0.0962	0.2126	40000
Business	1	0	0.5809	0.1526	0.0722	0.1688	38000
Business	1	0	0.3829	0.1746	0.0612	0.172	38000
Business	1	0	0.6722	0.1538	0.0596	0.1726	36000
Business	1	0	0.734	0.1717	0.0612	0.0669	33000

Regression Model & Hypotheses

Pick a regression model that could answer the question

$$MedianWage = \alpha + \beta_1 \%Women + \beta_2 \%Parttime + \beta_3 \%Unemployed + \beta_4 \%CollegeJob + \beta_5 BusinessDummy + \varepsilon$$

- Your expectation:
  - $\beta_1 < 0$
  - $\beta_2 < 0$
  - $\beta_3 < 0$
  - $\beta_4 > 0$
  - $\beta_5 > 0$ (business major grads earn more than the art major grads)

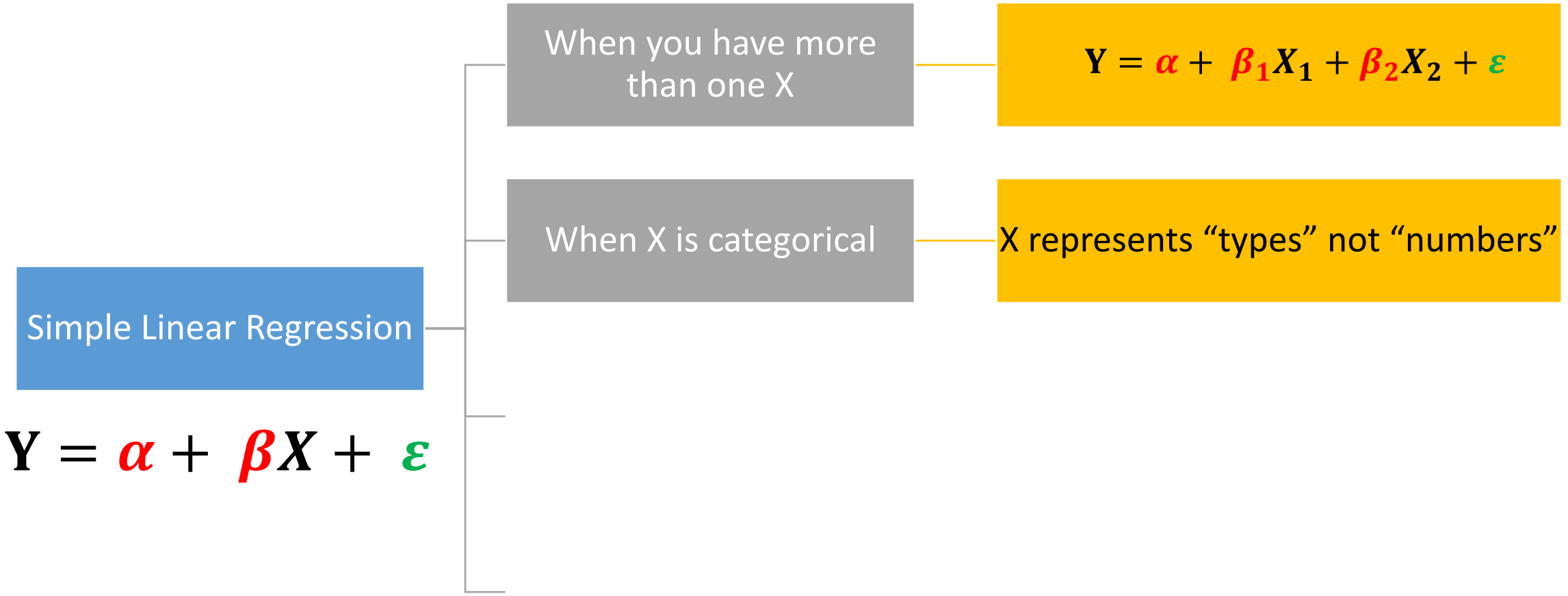
Regression Model:

Median Wage = 48045.74 + 483.15 x (Business Dummy)
– 14356.20 x (%Women)
– 42430.92 x (%Part Time)
+ 14626.26 x (%Unemployed)
+ 26795.66 x (%College Job)

The regression coefficient for business dummy is ***positive but not significant*** (p-value = 0.9484… much higher than 0.1). It means that there's no significant different between business major recent grad's & arts major recent grad's median salary.

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.87					
R Square	0.76					
Adjusted R Square	0.68					
Standard Error	5102.23					
Observations	21					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	5	1243460389	248692077.7	9.553028584	0.000296902	
Residual	15	390491992.4	26032799.5			
Total	20	1633952381				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	47095.7636	15620.8990	3.0149	0.0087	13800.6056	80390.9217
Business_Dummy	483.1505	7344.4424	0.0658	0.9484	-15171.1580	16137.4589
%Women	-14356.1973	7443.3306	-1.9287	0.0729	-30221.2809	1508.8863
%PartTime	-42430.9219	35710.4486	-1.1882	0.2532	-118545.9413	33684.0975
%Unemployed	14626.2558	85689.4020	0.1707	0.8667	-168016.3811	197268.8927
%CollegeJob	26795.6551	10176.0411	2.6332	0.0188	5105.9369	48485.3733

Extension to the Simple Linear Regression Model



Interaction Variable

X_1 or X_2 alone sometimes are not enough to explain Y

But X_1 times X_2 can significantly explain Y

$$\mathbf{Y} = \boldsymbol{\alpha} + \boldsymbol{\beta_1 X_1} + \boldsymbol{\beta_2 X_2} + \boldsymbol{\beta_3 X_1 X_2} + \boldsymbol{\varepsilon}$$

Interaction Variable

Create a new variable called

College_PartTime = %CollegeJob x %PartTime

Major_category	Business_Dummy	%Women	%Unemployed	%PartTime	%CollegeJob	College_PartTime	MedianWage
Arts	0	0.4102	0.0894	0.3195	0.2192	0.0700344	50000
Arts	0	0.3744	0.0968	0.2357	0.4138	0.09753266	35000
Arts	0	0.687	0.1058	0.3307	0.217	0.0717619	32000
Arts	0	0.4446	0.076	0.4114	0.2654	0.10918556	31000
Arts	0	0.667	0.0842	0.3178	0.3179	0.10102862	30500
Arts	0	0.6974	0.1022	0.3848	0.2687	0.10339576	30000
Arts	0	0.5848	0.0896	0.3342	0.2431	0.08124402	29000
Arts	0	0.6295	0.0775	0.3698	0.1612	0.05961176	27000
Business	1	0.5357	0.0957	0.0784	0.7552	0.05920768	62000
Business	1	0.2788	0.0582	0.1293	0.4958	0.06410694	51000
Business	1	0.3222	0.0479	0.1008	0.2725	0.027468	50000
Business	1	0.3555	0.0607	0.123	0.2935	0.0361005	47000
Business	1	0.2492	0.0964	0.1456	0.2111	0.03073616	46000
Business	1	0.2536	0.0697	0.1394	0.1853	0.02583082	45000
Business	1	0.4179	0.0729	0.1545	0.1863	0.02878335	40000
Business	1	0.2	0.072	0.1876	0.1729	0.03243604	40000
Business	1	0.2829	0.0962	0.1888	0.2126	0.04013888	40000
Business	1	0.5809	0.0722	0.1526	0.1688	0.02575888	38000
Business	1	0.3829	0.0612	0.1746	0.172	0.0300312	38000
Business	1	0.6722	0.0596	0.1538	0.1726	0.02654588	36000
Business	1	0.734	0.0612	0.1717	0.0669	0.01148673	33000

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.906					
R Square	0.821					
Adjusted R Square	0.744					
Standard Error	4575.215					
Observations	21					
ANOVA						
	df	*SS*				
Regression	6	1340896142	223			
Residual	14	293056238.6	20932588.47			
Total	20	1633952381				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	50366.46	14089.18	3.57	0.00	20148.17	80584.74
Business_Dummy	-5315.95	7113.22	-0.75	0.47	-20572.28	9940.38
%Women	-19368.97	7067.33	-2.74	0.02	-34526.90	-4211.05
%PartTime	8263.19	39717.77	0.21	0.84	-76922.94	93449.33
%Unemployed	-4039.27	77323.89	-0.05	0.96	-169882.52	161803.98
%CollegeJob	56605.76	16558.25	3.42	0.00	21091.84	92119.68
College_PartTime	-273895.57	126951.34	-2.16	0.05	-546179.11	-1612.02

Regression Model:

Median Wage = 50366.46 - 5315.95 x (Business Dummy)
– 19368.97 x (%Women)
+ 8263.19 x (%Part Time)
- 4039.27 x (%Unemployed)
+ 56605.76 x (%College Job)
- 273895.57 x (%College x %PartTime)

The regression coefficient for the interaction variable is ***negative and significant*** (p-value = 0.05… lower than 0.1). It shows that part-time job paid less than full-time job among the field with high % of college grads work in their own major.

Extension to the Simple Linear Regression Model

Simple Linear Regression

$$\mathbf{Y} = \boldsymbol{\alpha} + \boldsymbol{\beta}\boldsymbol{X} + \boldsymbol{\varepsilon}$$

- When you have more than one X
  - $\mathbf{Y} = \boldsymbol{\alpha} + \boldsymbol{\beta_1}\boldsymbol{X_1} + \boldsymbol{\beta_2}\boldsymbol{X_2} + \boldsymbol{\varepsilon}$
- When X is categorical
  - X represents “types” not “numbers”
- When X can be multiplied with another X
  - $\mathbf{Y} = \boldsymbol{\alpha} + \boldsymbol{\beta_1}\boldsymbol{X_1} + \boldsymbol{\beta_2}\boldsymbol{X_2} + \boldsymbol{\beta_3}\boldsymbol{X_1}\boldsymbol{X_2} + \boldsymbol{\varepsilon}$

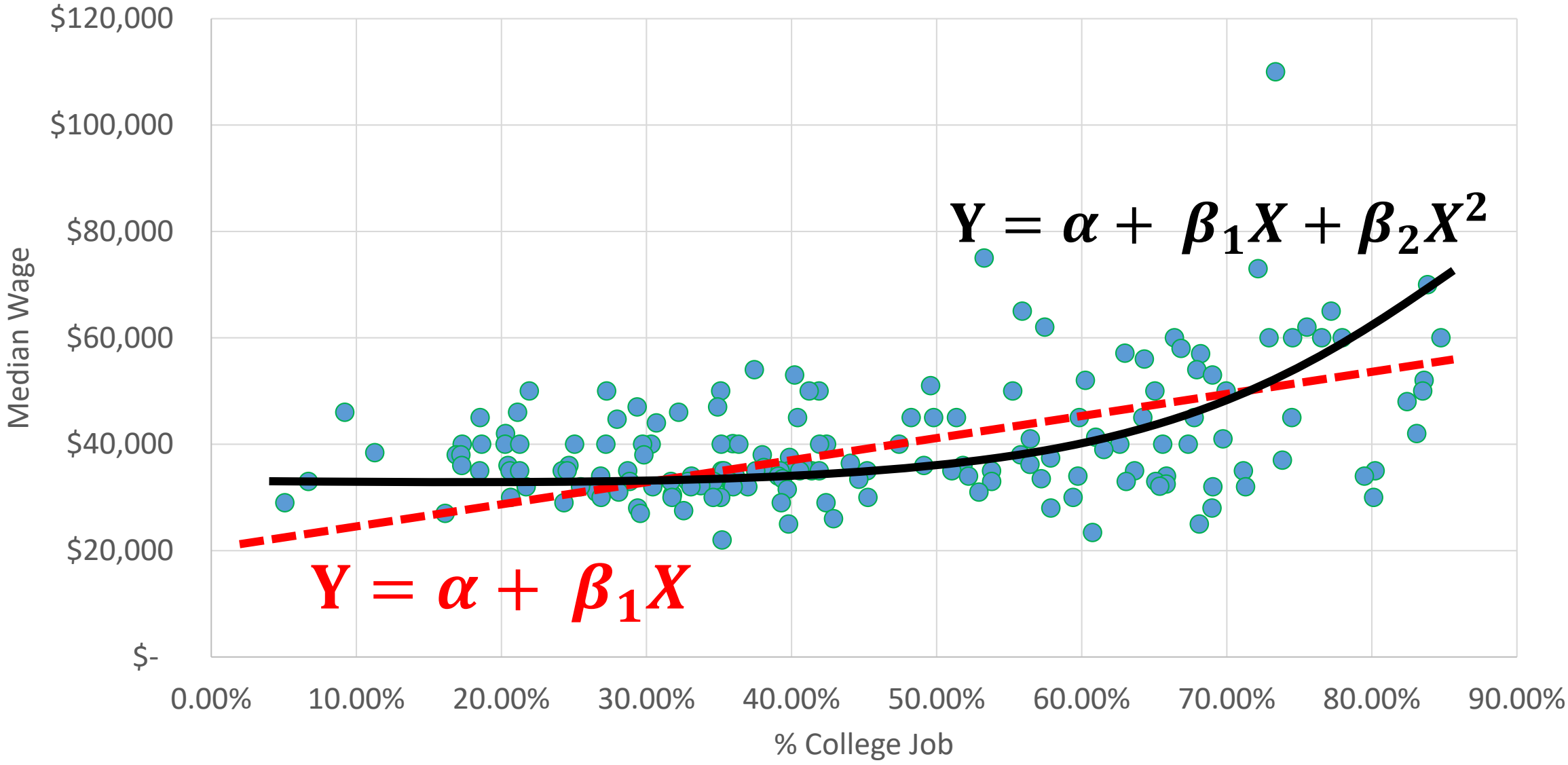
Non-linear Model?

$$\mathbf{Y} = \boldsymbol{\alpha} + \boldsymbol{\beta_1 X} + \boldsymbol{\beta_2 X^2} + \boldsymbol{\beta_3 X^3} + \boldsymbol{\varepsilon}$$

$$\boldsymbol{log(Y)} = \boldsymbol{\alpha} + \boldsymbol{\beta_1 log(X)} + \boldsymbol{\varepsilon}$$

And many others…

Red Line or Black Line?



SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.404					
R Square	0.163					
Adjusted R Square	0.158	←	The basic linear regression model gives you adjusted r square = 15.8%. It seems fine so far…			
Standard Error	10553.529					
Observations	172					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	1	3695081697	3695081697	33.17634889	3.85867E-08	
Residual	170	18934087372	111376984.5			
Total	171	22629169070				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	29288.117	2050.665	14.282	0.000	25240.070	33336.164
%CollegeJob	23820.373	4135.559	5.760	0.000	15656.709	31984.036

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.445					
R Square	0.198					
Adjusted R Square	0.189					
Standard Error	10362.256					
Observations	172					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	2	4482567672	2241283836	20.87316297	7.91964E-09	
Residual	169	18146601398	107376339.6			
Total	171	22629169070				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	40061.309	4458.651	8.985	0.000	31259.484	48863.135
%CollegeJob	-30485.419	20459.979	-1.490	0.138	-70875.473	9904.634
%CollegeJob_square	56918.633	21017.805	2.708	0.007	15427.376	98409.891

However, it's actually better to include the square term in the model because of the higher adjusted r square & significant squared variable

Key Takeaway

Data Creator

Audience Value Generation

What's your audience's goal?

What insights they need to know to inform / make decisions?

How they can generate values using insights?

Data Acquisition

What

Where

When

Who

How

How Many

Data Integration

How "dirty"

is your data?

Analysis

List out exact

questions

and variables/models

needed

Potential bias and error

traps?

Delivery

Key visuals

Key statistics

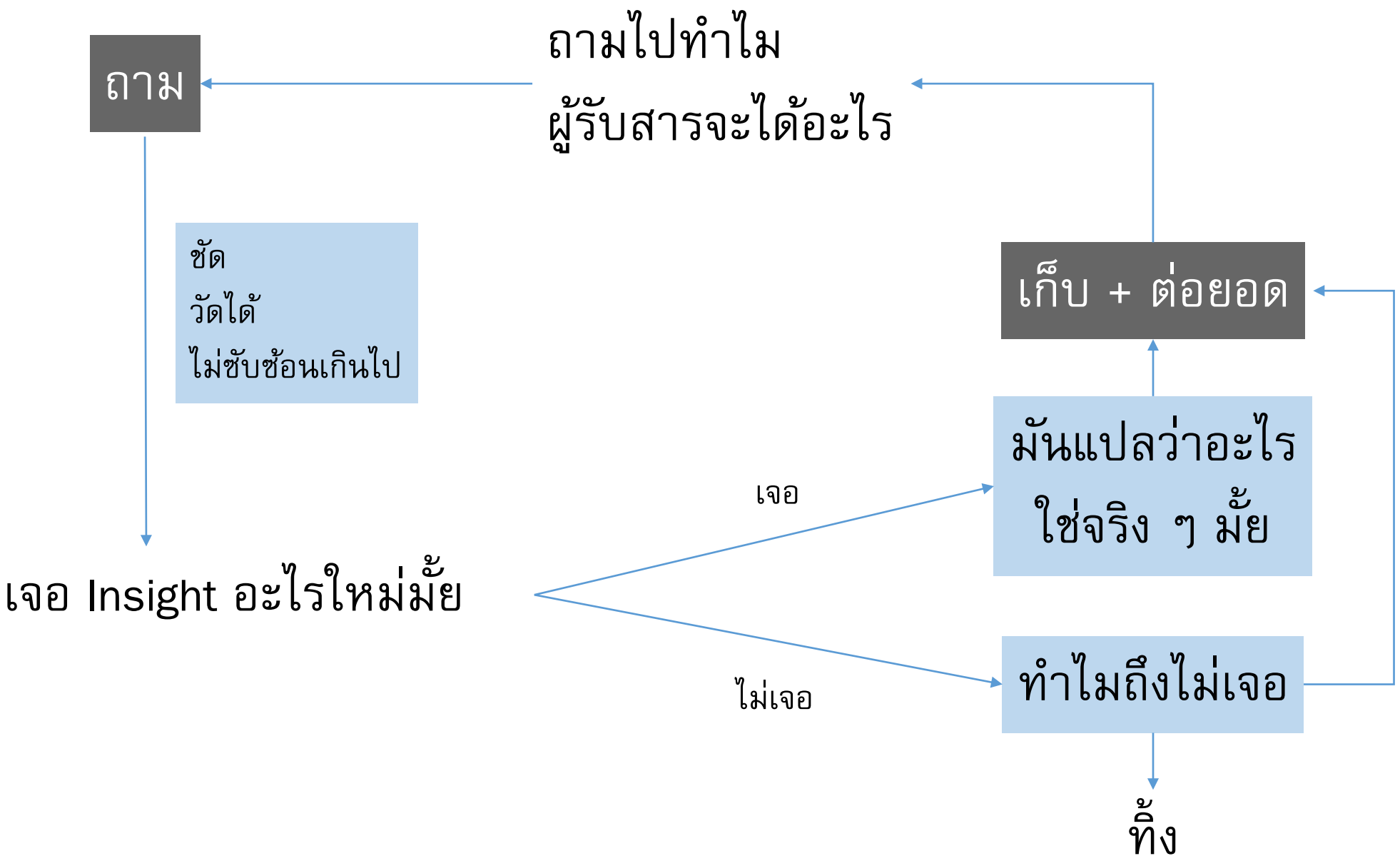
Key takeaways

Data Governance

Does your organization promote behaviors for good data practice?

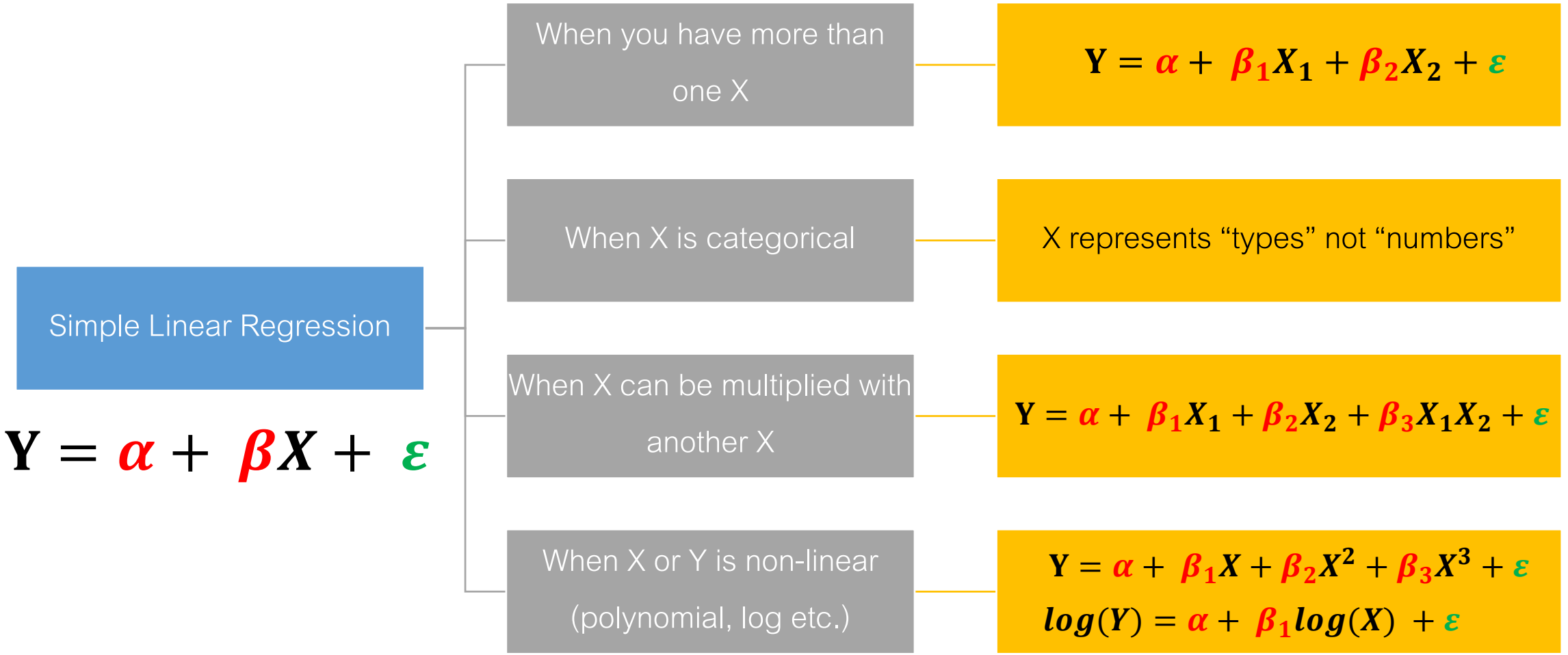
Data Principles / Quality / Privacy / Life Cycle / Ethical Use

Audience


ถาม
ถามไปทำไม
ผู้รับสารจะได้อะไร
ชัด
วัดได้
ไม่ซับซ้อนเกินไป
เก็บ + ต่อยอด
มันแปลว่าอะไร
ใช่จริง ๆ มั้ย
เจอ
เจอ Insight อะไรใหม่มั้ย
ไม่เจอ
ทำไมถึงไม่เจอ
ทิ้ง

กรอง	เจาะ	เทียบ	ชน
เอาข้อมูลที่ “ไม่สำคัญ” หรือ “สกปรก” ออกไป	ดูตัวแปรเดียวอย่างละเอียด เช่น ความถี่ ผลรวม ค่าเฉลี่ย การกระจายข้อมูล ตำแหน่งข้อมูล	เปรียบเทียบระหว่างตัวแปรเดียวกันในหลายประเภท/ชนิด หรือดูสัดส่วน	ดูความสัมพันธ์ระหว่างตัวแปรสองตัวขึ้นไป

Gain Deeper Insights with Regression Analysis



Question?