Understanding the World Through Systems & Data

Jennifer Schellinck[1], Patrick Boily[2]

[1]Adjunct Professor, Institute of Cognitive Science, Carleton University
Principal, Sysabee

[2]Manager, Centre for Quantitative Analysis and Decision Support, Carleton University
Adjunct Professor, Department of Mathematics and Statistics, University of Ottawa
President, Idlewyld Analytics and Consulting Services

“What if the only valid model of the Universe
is the Universe itself?”

(unknown)

Contents

Systems and Data

- Thinking in Systems Terms
- Identifying Gaps in Knowledge
- Conceptual Models
- Relating the Data to the System

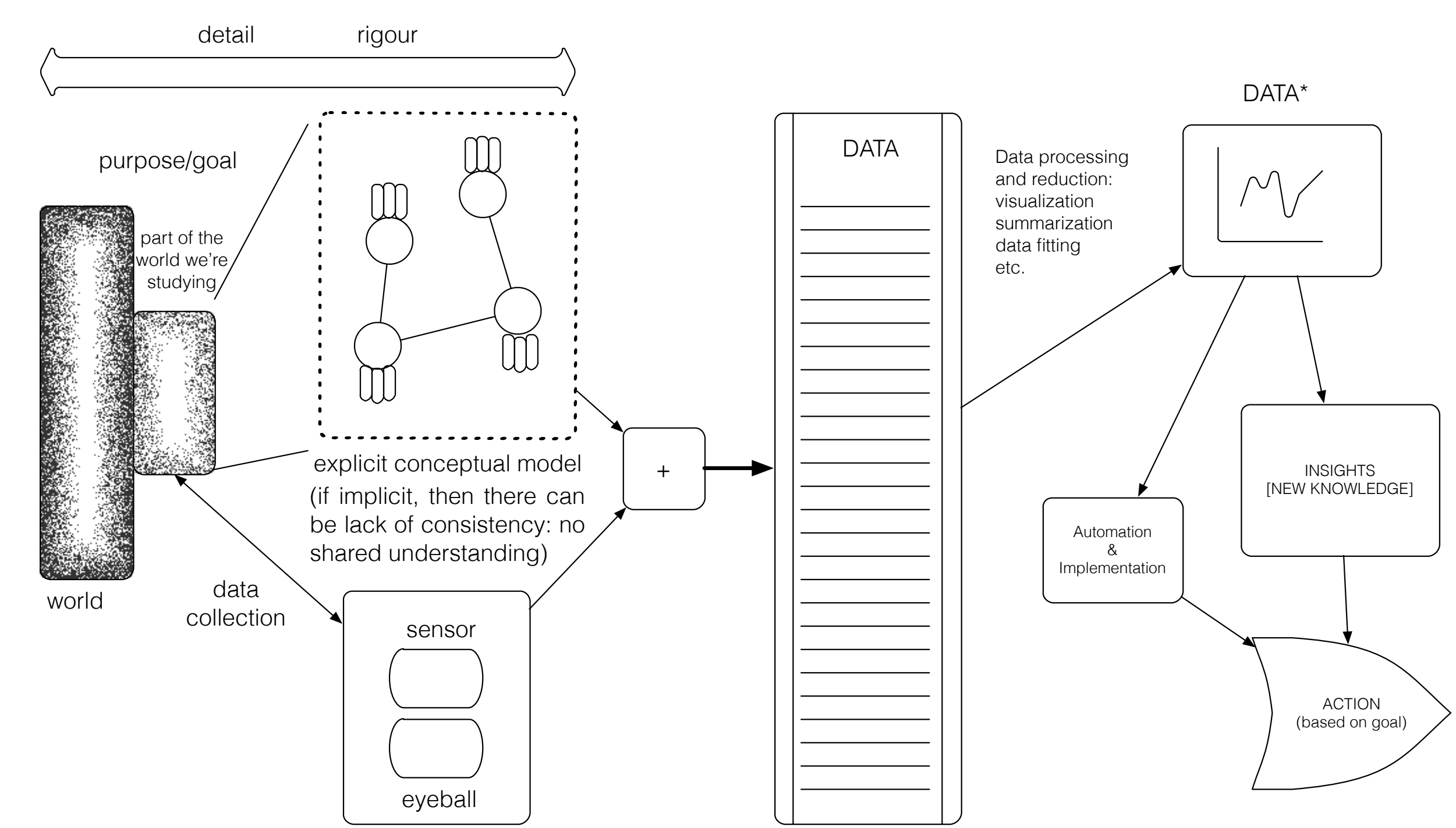
Case Study – BASA

- The BASA World
- Goals and Purposes
- System Model

Case Study – BASA (cont.)

- BASA Dataset
- Data Collection
- Data Exploration
- Data Consolidation
- System Model (reprise)
- Data Consolidation (reprise)
- Insights
- Informed Implementation and Action

Systems and Data


detail
rigour
purpose/goal
part of the world we're studying
world
explicit conceptual model (if implicit, then there can be lack of consistency: no shared understanding)
data collection
sensor
eyeball
+
DATA
DATA*
Data processing and reduction: visualization summarization data fitting etc.
Automation & Implementation
INSIGHTS [NEW KNOWLEDGE]
ACTION (based on goal)

Thinking in Systems Terms

In order to understand how various aspects of the World interact with one another, we need to **carve out chunks** corresponding to the aspects and define their **boundaries**.

Working with other intelligences requires **shared understanding** of what is being studied.

A **system** is made up of **objects** with **properties** that potentially change over time. Within the system we perceive **actions** and **evolving** properties leading us to think in terms of **processes**.

Thinking in Systems Terms

Objects themselves have various properties. Natural processes generate (or destroy) objects, and may change the properties of these objects over time.

We **observe**, **quantify**, and **record** particular values of these properties at particular points in time.

This generates data points, capturing the **underlying reality** to some degree of **accuracy** and **error** (biased or unbiased).

Take-Away: certain aspects of the Universe can be approximated with the help of systems.

Take-Away: system models provide the basis under which data is identified and collected.

Take-Away: but data itself is approximate and selective.

Identifying Gaps in Knowledge

A **gap in knowledge** is identified when we realize that what we thought we knew about a system proves incomplete (or false).

This might happen repeatedly, at any moment in the process:

- data cleaning
- data consolidation
- data analysis

The solution is to be flexible. When faced with such a gap, **go back**, **ask questions**, and **modify the system representation**.

Take-Away: it’s going to happen, no matter what. Be prepared and ready to re-visit your set-up regularly.

Conceptual Models

Exercise:

- Assume that an acquaintance has just set foot in your living space for the first time.
- You are on the phone with them but not currently at home.
- Explain to them how to go about preparing a cup of sugar.

Conceptual models are built using methodical investigation tools

- diagrams
- structured interviews
- structured descriptions
- etc.

Take-Away: we often only rely on implicit conceptual modeling … but that way lies danger.

Relating the Data to the System

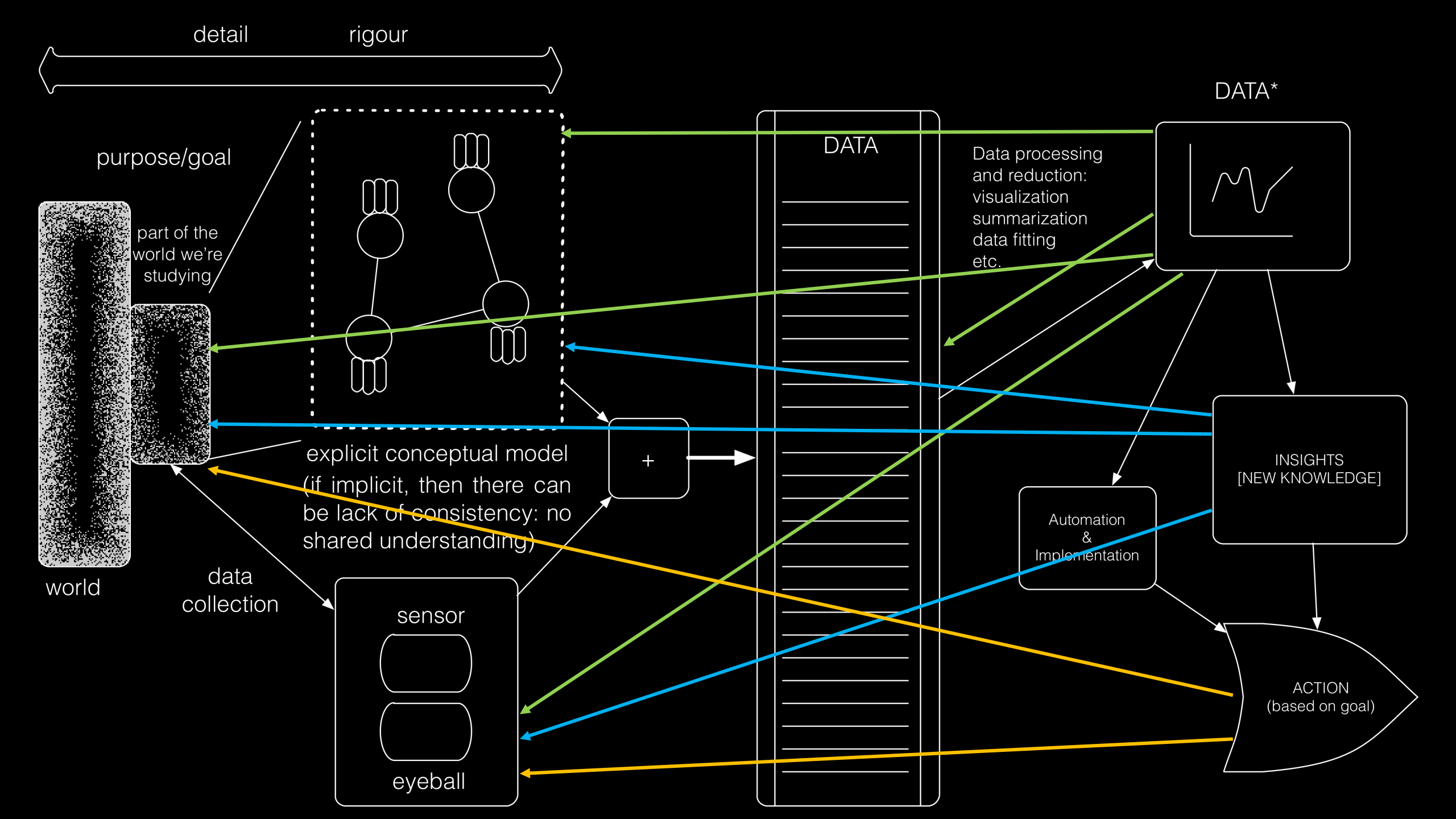
Is the data which has been collected and analyzed going to be of any use when it comes to understanding the system?

This question can only be answered if we understand:

- how the data is **collected**
- the **approximate nature** of both data and system
- what the data **represents** (observations and features)

Is the combination of system and data sufficient to understand the aspects of the world under consideration?

Take-Away: if the data, the system, and the world are out of alignment, insights might prove useless.


detail
rigour
purpose/goal
part of the world we're studying
world
data collection
explicit conceptual model (if implicit, then there can be lack of consistency: no shared understanding)
+
sensor
eyeball
DATA
DATA*
Data processing and reduction: visualization summarization data fitting etc.
INSIGHTS [NEW KNOWLEDGE]
Automation & Implementation
ACTION (based on goal)

Case Study – BASA

The BASA World

The *Borealian Aeronautic Security Agency* (BASA) runs **pre-board screening** of passengers and crew for all flights departing the nation's airfields.

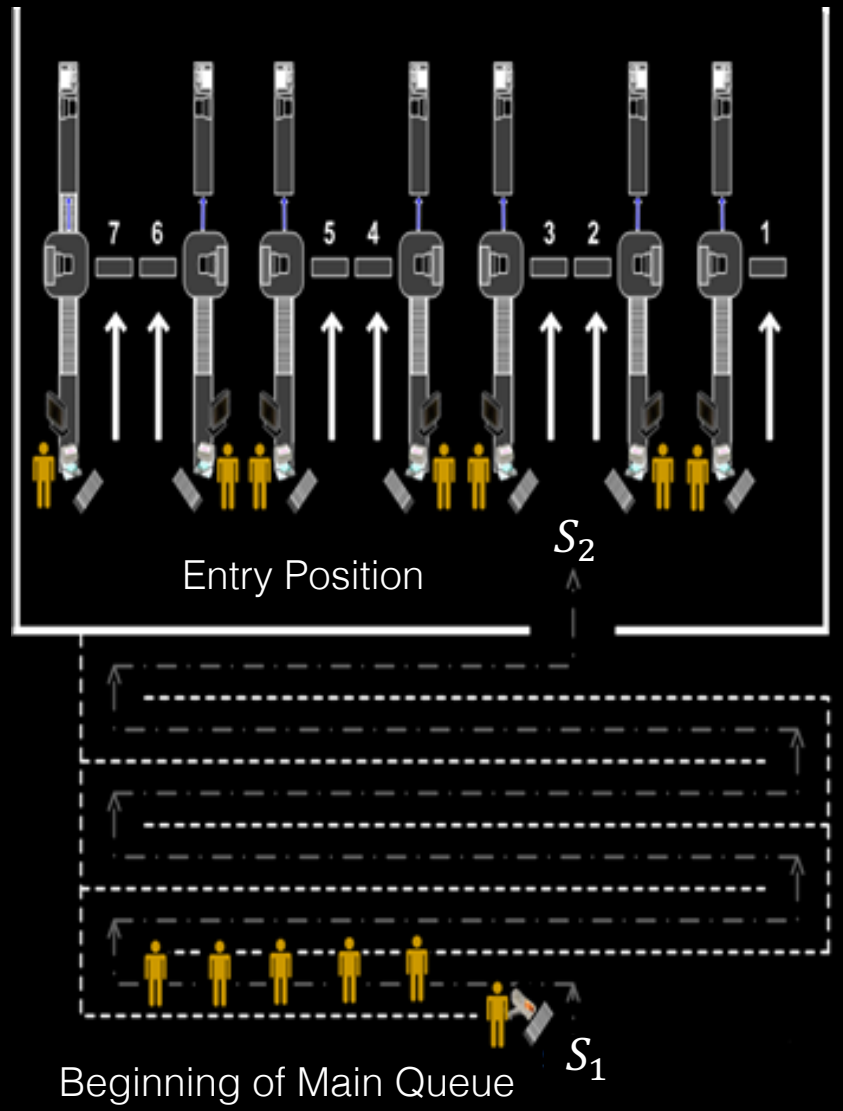
4 Major **Airfields**

- Auckland
- Chebucto
- Saint-François
- Queenston

The BASA World

The screening process is structurally similar at each airfield:

1. Passengers arrive at the beginning of the main queue
2. Boarding passes may or may not be scanned at S_1
3. Passengers enter the main queue
4. Boarding passes are scanned at S_2
5. Passengers are directed to a server entry position
6. Passengers and carry-on luggage are screened by a server



Goals and Purposes

Some factors influence the PBS wait time, including:

- **schedule intensity** of departing flights
- **passenger volume** on these flights
- number of **servers** and **processing rates** at a given airfield, etc.

There might also be

- **yearly**, **seasonal**, **time-of-day**, **day-of-week interaction effects** (among others) depending on the airfield, the flight destination, etc.
- trend **level shifts** in the number of passengers, flights, destinations, etc.

Goals and Purposes

Ultimately, BASA is seeking an **in-depth understanding** of their data to help make Borealian airfields as **efficient** and **secure** as possible.

At the most basic level, BASA is **seeking answers** to **questions:**

- Are there any insights which could be gleaned by visualizing data?
- What do anomalous observations look like at the passenger, flight, and active server levels?
- In what circumstances are passengers not scanned at S_1?
- When do passengers typically arrive to be scanned at S_1?

Goals and Purposes

Questions (continued):

- On average, how long do passengers wait in the main queue? What factors affect the waiting time?
- Does server performance change according to traffic patterns?
- Is it possible to forecast passenger arrival patterns based on flight schedule?
- Is it possible to predict main queue waiting times given specific arrival patterns, flight schedule, and server vacation policy?
- Is it possible to set a server vacation policy to control waiting times based on predicted arrival patterns?
- Do passengers ever miss flights because of the waiting time? Can we find factors that are linked with missed flights?

Goals and Purposes

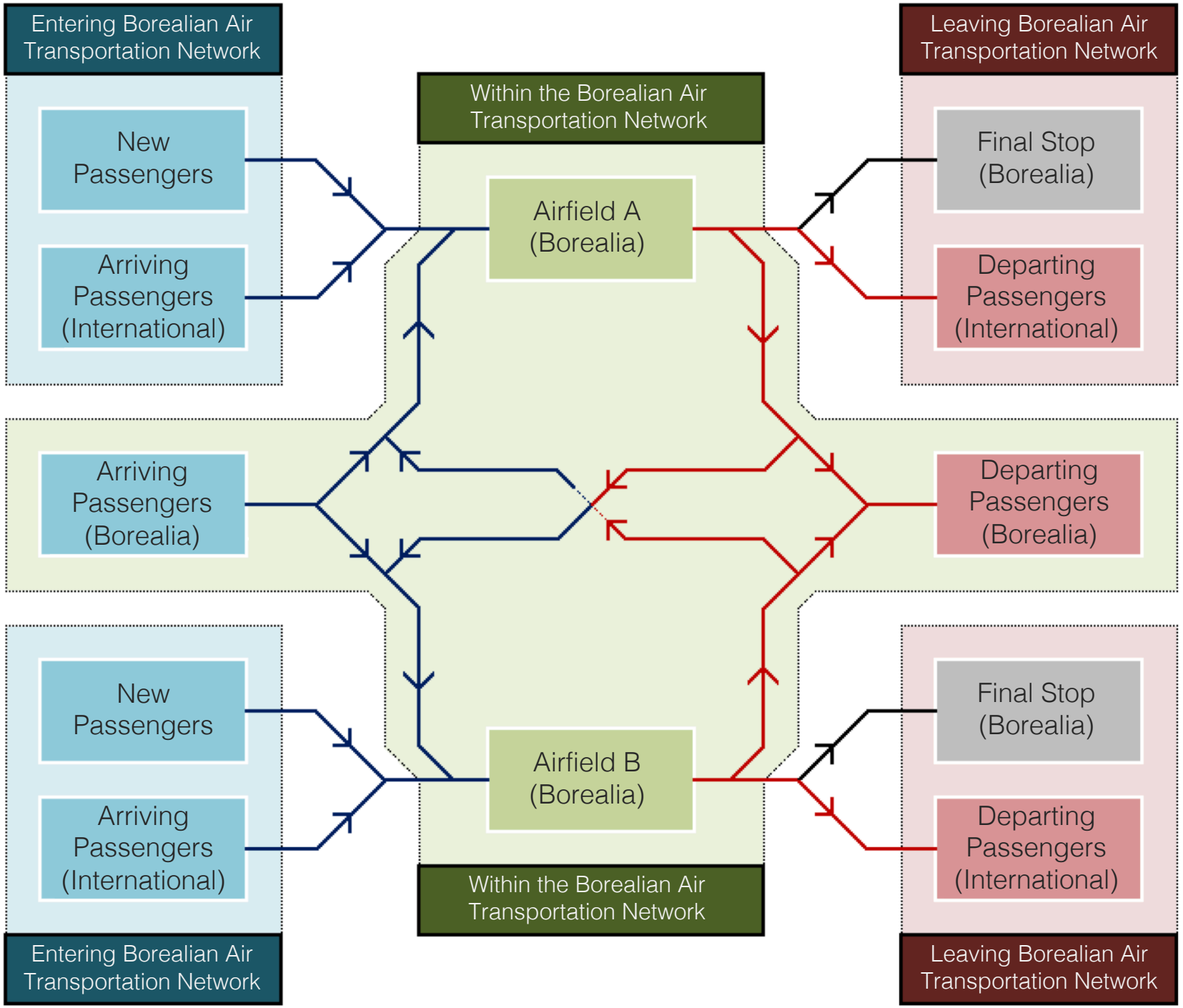
Questions (continued):

- Based on the size, schedule, and final destination of the passengers, what flights are most similar? Most dissimilar?
- Do the number of flights and number of passengers exhibit seasonal patterns or trend level shifts?
- Is there any way to detect if servers or airfields are not reporting their data correctly, either through fraud, or incompetence?
- Can we predict the effects that temporarily shutting down an airfield or modifying the number of flights between airfields could have on the Borealian network?
- Can anything else insightful be said about the data?

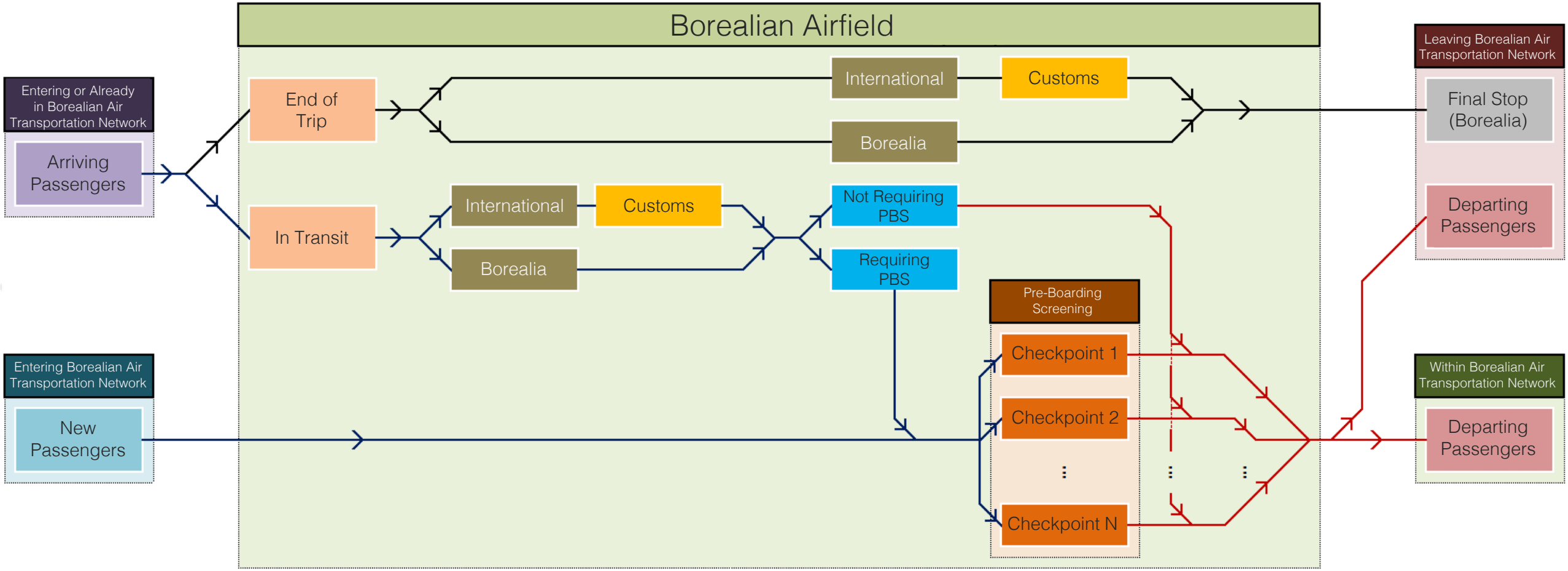
System Model

Most of us have flown on a blimp, so we have an **implicit model of the system**.

Can we formalize our understanding of the system (without reference to SMEs)?



System Model



BASA Dataset

Available data, from 20X6 to 20X8 (about ~10M observations):

- Airfield
- Passenger ID
- Scan at S_2
- Wait Time
- C_{start}: # servers at S_1
- C_0: # servers at S_2
- Scheduled Departure
- Actual Departure
- Destination City and Country

Data Collection

The data has already been collected. **Scanner readings** are meant to be made both at S_1 and at S_2, but a fair chunk ($\sim 45\%$) of passengers were not scanned at S_1.

Flight information (scheduled and actual departure time) and **passenger information** (passenger ID and final destination) are collected separately at the airline level, and merged by BASA.

Piece of cake, right? We have data, we understand the system: let's get to it and answer the questions already!

Take-Away: what could possibly go wrong?

Data Exploration

Some initial questions:

- How much data is there?
- What is wrong with the data?
- What is ‘going on’ with the data?
- What can we 'see' in the data?

Field Name	Description	DataType	Format	NA indicator in Raw Data	Field Precision
Airfield	Four airfields of interest: Auckland [AUC], Chebucto [CWL], Queenston [QUE], Saint-François [SAF]	String (factor)	AAA	Blank/NA	-
S2	The date and time at which passengers exited the main queue	Datetime	YYYY-MM-DD HH:MM:SS	Blank/NA	minute
Wait_Time	The interval of time spent in the main queue	Integer	#...	Blank/NA	minute
C_Start	The reported number of active servers when a passenger entered the main queue	Integer	#...	Blank/NA	-
C0	The reported number of active servers when a passenger exited the main queue	Integer	#...	Blank/NA	-
C_avg	The average reported number of active servers during the period spent in the main queue	Float	#.###...	Blank/NA	-
Sch_Departure	The scheduled departure time of each passenger's flight	Datetime	YYYY-MM-DD HH:MM:SS	Blank/NA	minute
Act_Departure	The actual departure time of each passenger's flight	Datetime	YYYY-MM-DD HH:MM:SS	Blank/NA	minute
BFO_Dest_City	The city code for passenger's final destination airfield	String (factor)	AAA### OR AAA	Blank/"."	-
BFO_Destination_Country_Code	The country code for passenger's final destination airfield	String (factor)	AAA	Blank/"."	-
Order	Order of S2	Integer	#...	Blank/NA	-
Pass_ID	A unique passenger ID	Integer	#...	Blank/NA	-
Departure_Date	Date of actual departure (derived field)	Date	YYYY-MM-DD	Blank/NA	-
Departure_Time	Time of actual departure (derived field)	Integer	#...	Blank/NA	-
Time_of_Day	Time slot of actual departure (derived field)	String (factor)	# - A*	0 - NO DATA	-
Period_of_Week	of actual departure (derived field)	String (factor)	# - A*	0 - NO DATA	-
Day_of_Week	Day of actual departure (derived field)	String (factor)	# - A*	0 - NOD	-
Month	Month of actual departure (derived field)	String (factor)	## - A*	00 - NOD	-
Season	Season of actual departure (derived field)	String (factor)	# - A*	0 - NO DATA	-
Year	Year of actual departure (derived field)	integer	####	Blank/NA	-

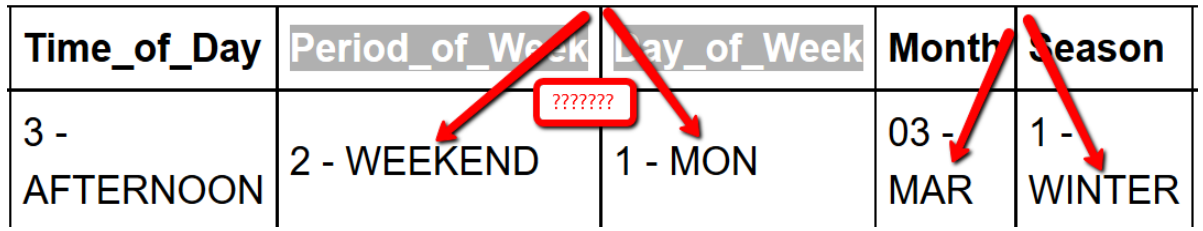
Data Exploration

Some initial questions:

- How much data is there?
- What is wrong with the data?
- What is ‘going on’ with the data?
- What can we 'see' in the data?

Field	Number of non-missing entries	Number of missing entries	%Missing
Airfield	9,984,687	-	0%
S2	9,984,687	-	0%
Wait_Time	5,450,590	4,534,097	45%
C_Start	4,588,266	5,396,421	54%
C0	8,792,155	1,192,532	12%
C_avg	4,588,266	5,396,421	54%
Sch_Departure	9,792,456	192,231	2%
Act_Departure	9,792,456	192,231	2%
BFO_Dest_City	9,858,613	126,074	1%
BFO_Destination_Country_Code	9,858,613	126,074	1%
order	9,984,687	-	0%
Pass_ID	9,984,210	477	0%
Departure_Date	9,792,456	192,231	2%
Departure_Time	9,792,456	192,231	2%
Time_of_Day	9,792,456	192,231	2%
Period_of_Week	9,792,456	192,231	2%
Day_of_Week	9,792,456	192,231	2%
Month	9,792,456	192,231	2%
Season	9,792,456	192,231	2%
Year	9,792,456	192,231	2%

Time_of_Day	Period_of_Week	Day_of_Week	Month	Season
3 - AFTERNOON	2 - WEEKEND	1 - MON	03 - MAR	1 - WINTER



Data Exploration

Tables and visualizations are fairly “rough and ready.”

Goal: start connecting the data (and what it is telling you) to your knowledge of the system.

Does it match what you **already know** about the system?

Is anything **unexpected** and/or **interesting**?

	Month					
Year	01 - JAN	02 - FEB	03 - MAR	04 - APR	05 - MAY	06 - JUN
1899	0	0	0	0	0	0
1900	154	0	0	0	0	0
2026	0	0	0	0	0	0
2027	153526	158477	275123	277215	269409	267474
2028	252028	289019	320808	302880	288543	280781
2029	273617	287340	334000	310066	289620	282898
2030	3459	0	0	0	0	0

	Month					
Year	07 - JUL	08 - AUG	09 - SEP	10 - OCT	11 - NOV	12 - DEC
1899	0	0	0	0	0	176
1900	0	0	0	0	0	0
2026	0	0	0	0	0	7
2027	288016	295477	263719	264050	241385	268915
2028	309786	317122	278978	284296	252702	271046
2029	316108	326890	278122	292092	251657	267706
2030	0	0	0	0	0	0

Scanned Passengers, by Year

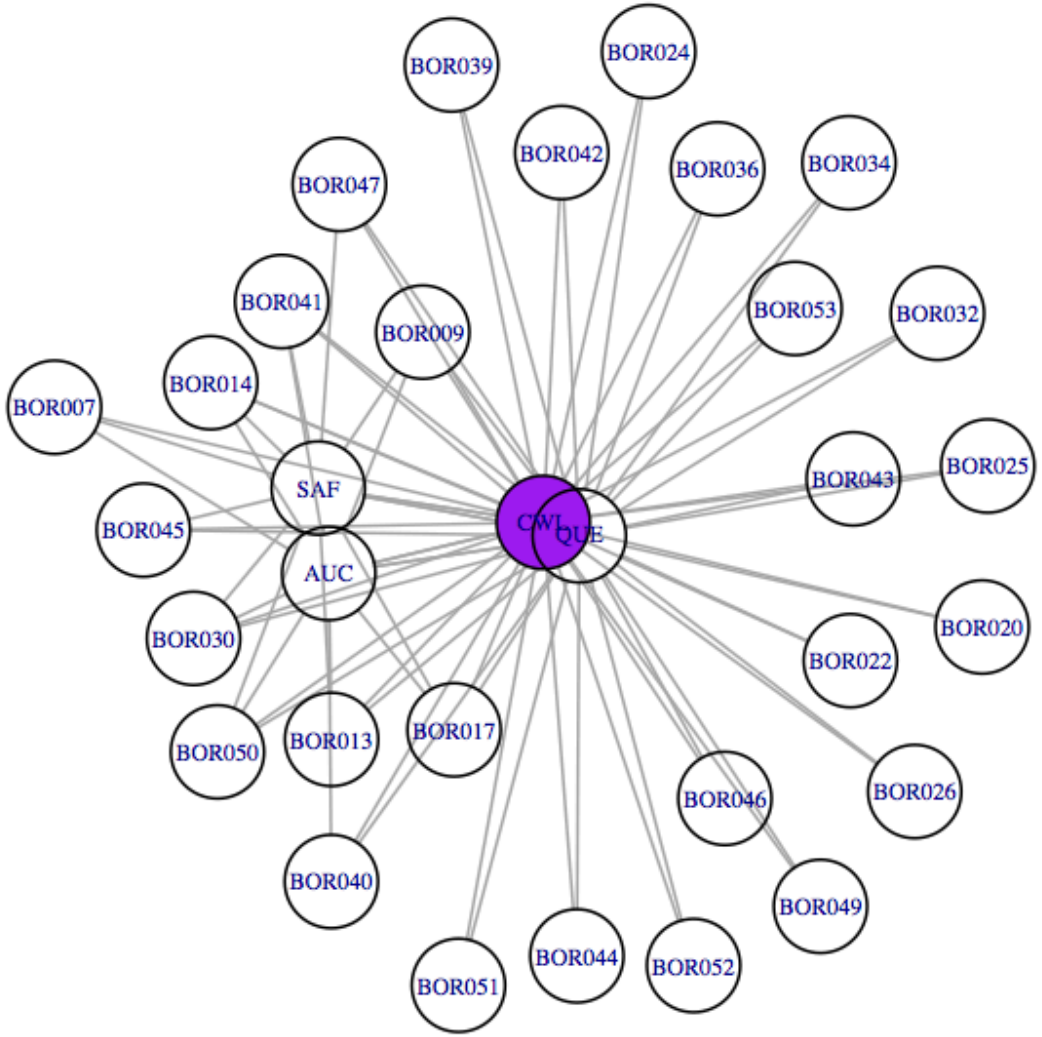
Data Exploration

Tables and visualizations are fairly “rough and ready.”

Goal: start connecting the data (and what it is telling you) to your knowledge of the system.

Does it match what you **already know** about the system?

Is anything **unexpected** and/or **interesting**?



Final Destination Pairs, by Airfield

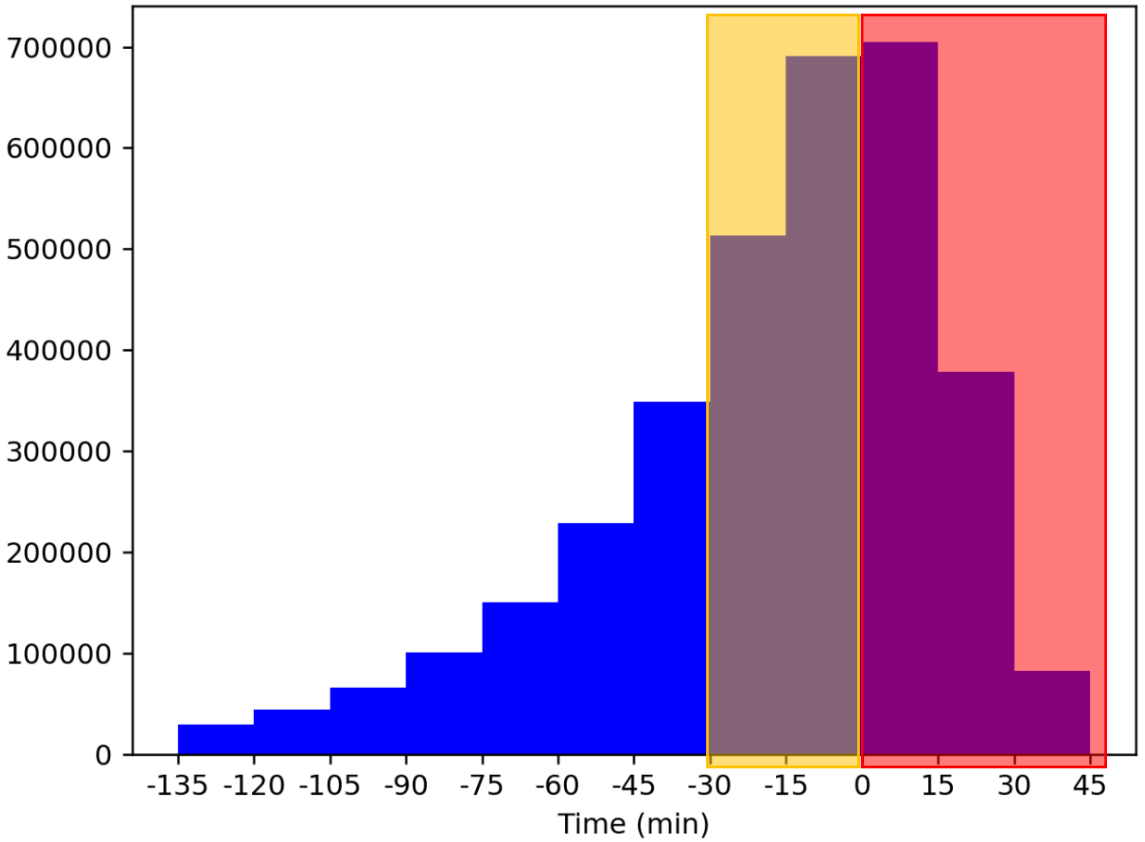
Data Exploration

Tables and visualizations are fairly “rough and ready.”

Goal: start connecting the data (and what it is telling you) to your knowledge of the system.

Does it match what you **already know** about the system?

Is anything **unexpected** and/or **interesting**?



S_2 – Sch_Departure (positive means late arrival)

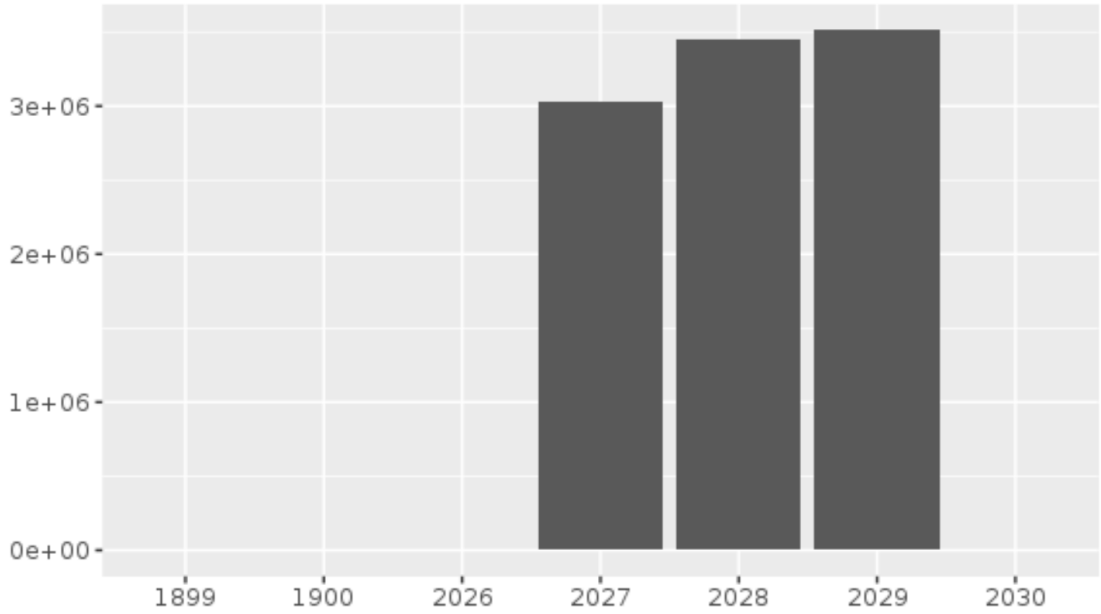
Data Exploration

Tables and visualizations are fairly “rough and ready.”

Goal: start connecting the data (and what it is telling you) to your knowledge of the system.

Does it match what you **already know** about the system?

Is anything **unexpected** and/or **interesting**?

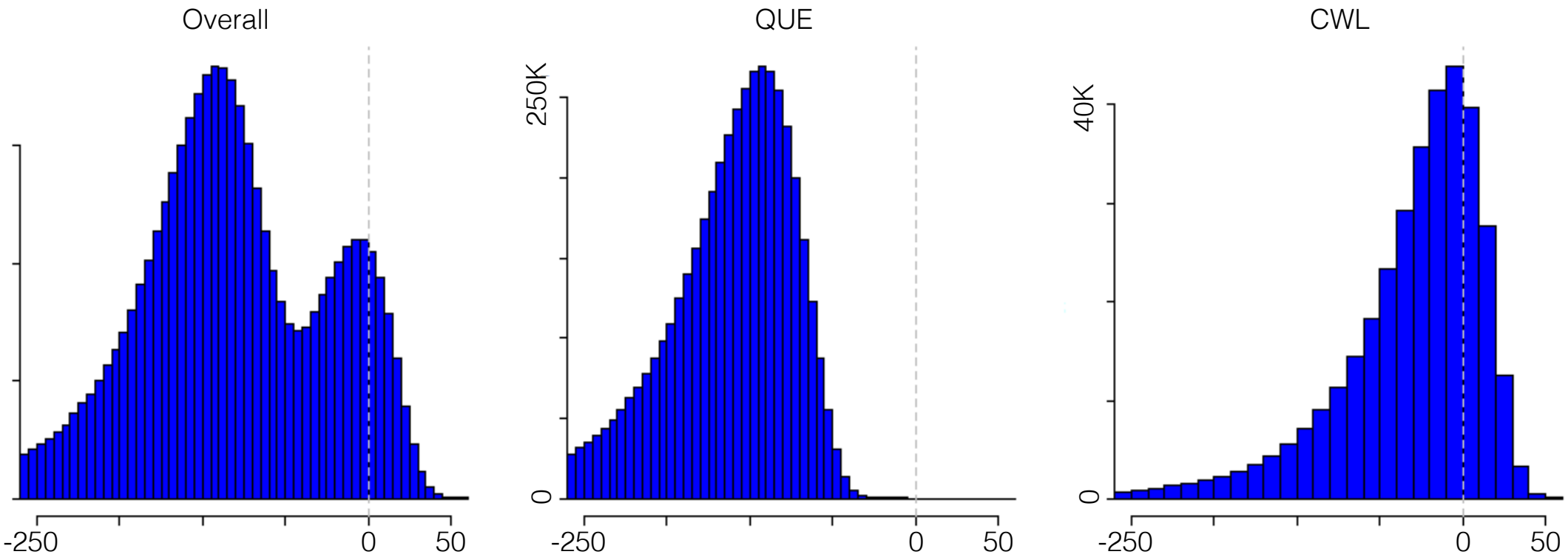


Frequency of S_2 Scans by Year, for All Airfields

The most exciting phrase to hear in science, the one that heralds new discoveries, is not “Eureka!” but “That’s funny …”

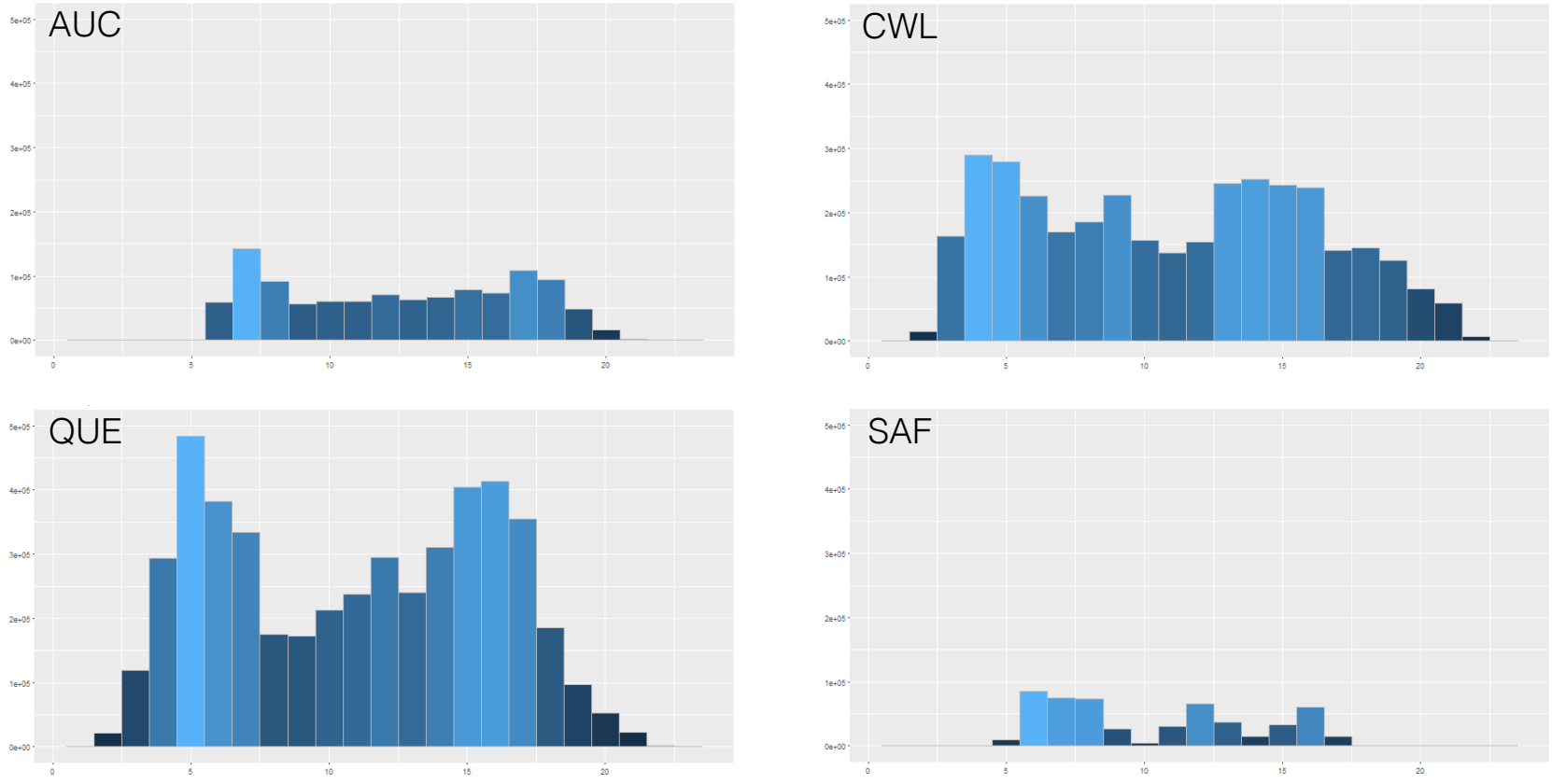
(attributed to Isaac Asimov)

Data Consolidation



Distribution of S_2 Relative to Actual Departure Time
(in min)

Data Consolidation



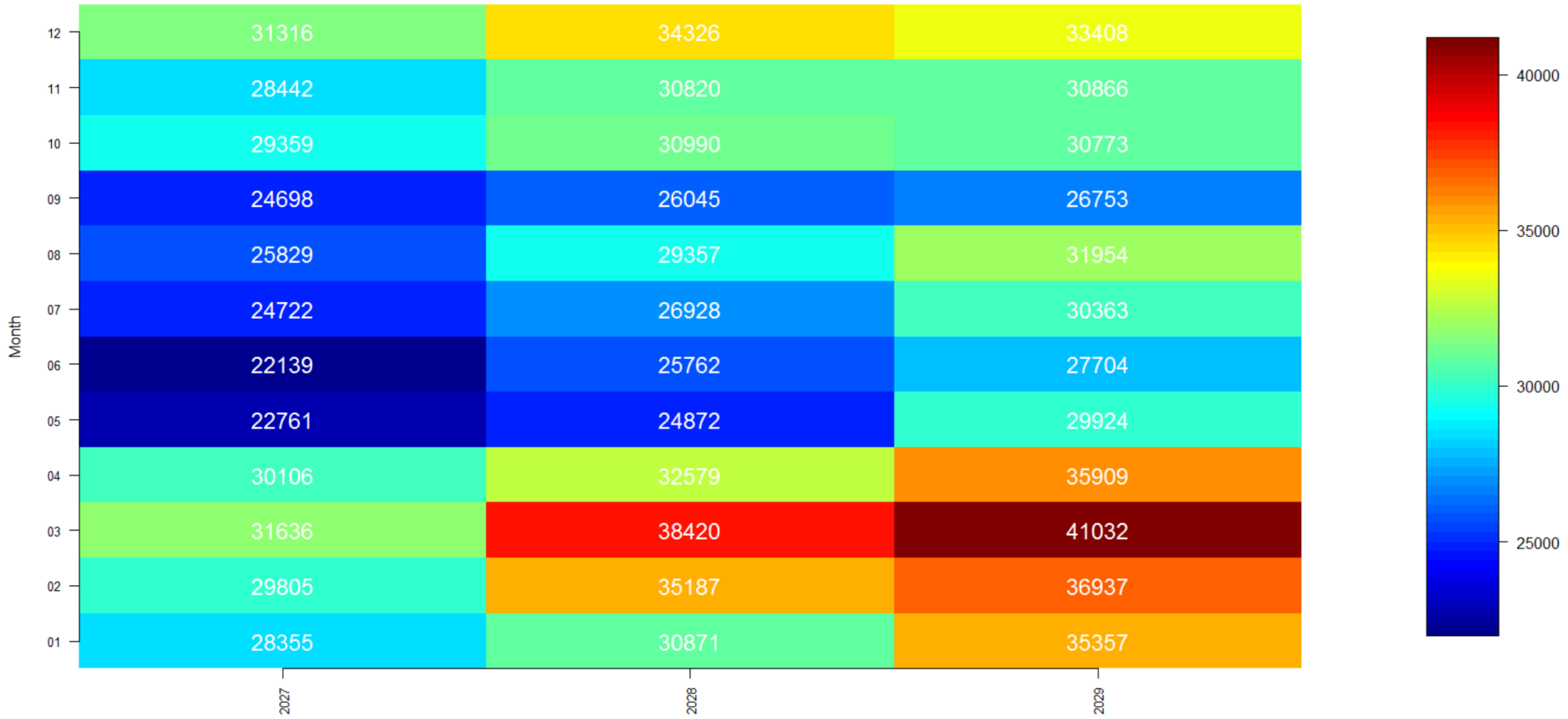
Distribution of Total # of Passenger Going Through PBS Throughout the Day, by Airfield

Data Consolidation

		Destination		
		Domestic	International	Total
Origin	AUC	7.3%	3.4%	10.6%
	CWL	25.2%	10.4%	35.6%
	QUE	34.1%	14.3%	48.5%
	SAF	2.1%	3.2%	5.3%
	Total	68.7%	31.3%	100.0%

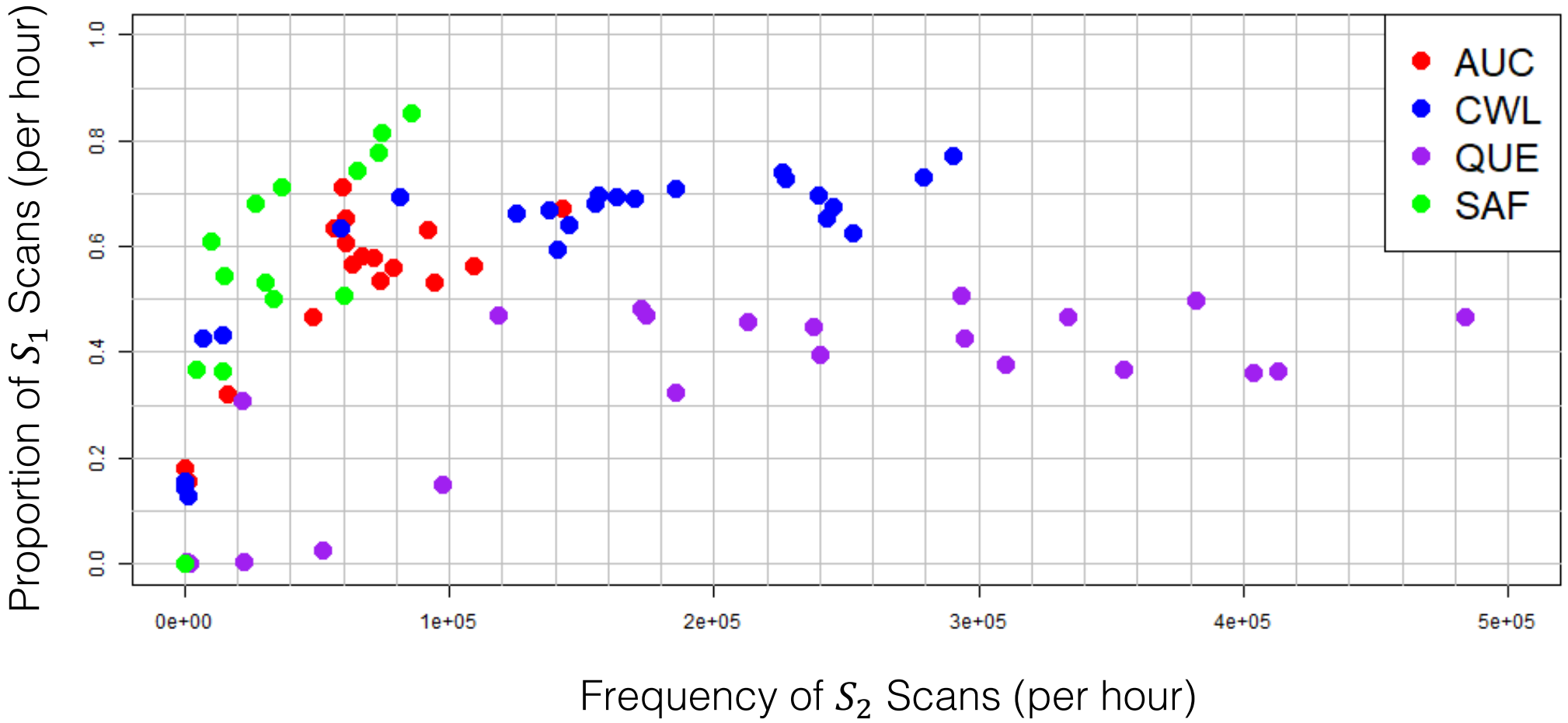
Percentage Domestic and International Travellers, by Airfield

Data Consolidation



Traffic Density, per Year, per Month (AUC)

Data Consolidation



Take-Away: a first pass at data consolidation may prove illuminating – but it will also likely reveal serious gaps in the knowledge of the system.

System Model: Q&A with Clients

STAGE 1: CHECK - IN

When passengers check in for traveling, they will be assigned a passenger ID. All passengers (including infants and young children) will be assigned a passenger ID.

CORRECT? ☐

For a given airport, Passenger IDs are unique - no two people should ever have the same passenger ID, and the same person taking different flights at different times will always have a different passenger ID.

CORRECT? ☐

The passenger ID is connected to flight information (e.g. flight number, scheduled departure time) for the flight the passenger is booked on (e.g. via a foreign key pointing to a flight information database). Passengers will always be connected to a specific flight.

CORRECT? ☐

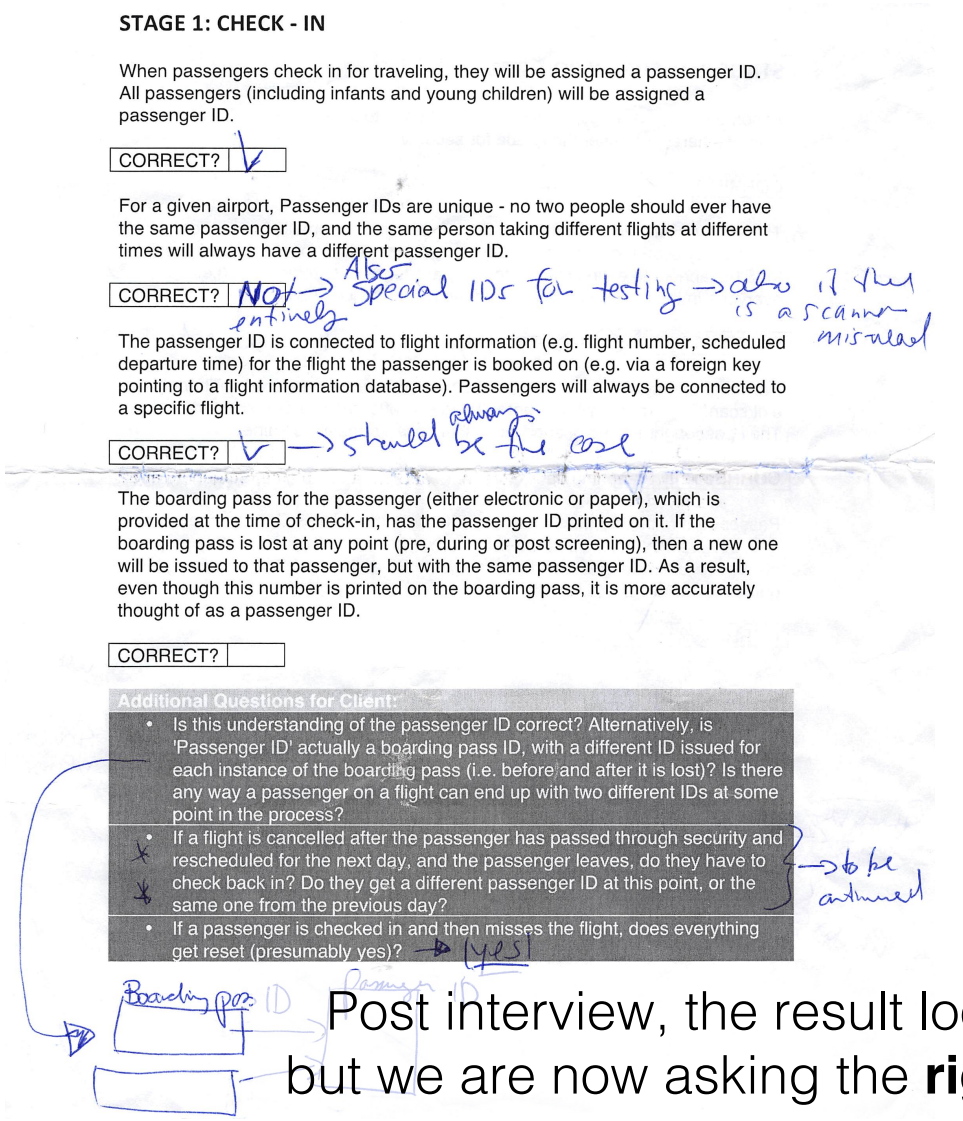
The boarding pass for the passenger (either electronic or paper), which is provided at the time of check-in, has the passenger ID printed on it. If the boarding pass is lost at any point (pre, during or post screening), then a new one will be issued to that passenger, but with the same passenger ID. As a result, even though this number is printed on the boarding pass, it is more accurately thought of as a passenger ID.

CORRECT? ☐

Additional Questions for Client:

- Is this understanding of the passenger ID correct? Alternatively, is 'Passenger ID' actually a boarding pass ID, with a different ID issued for each instance of the boarding pass (i.e. before and after it is lost)? Is there any way a passenger on a flight can end up with two different IDs at some point in the process?
- If a flight is cancelled after the passenger has passed through security and rescheduled for the next day, and the passenger leaves, do they have to check back in? Do they get a different passenger ID at this point, or the same one from the previous day?
- If a passenger is checked in and then misses the flight, does everything get reset (presumably yes)?

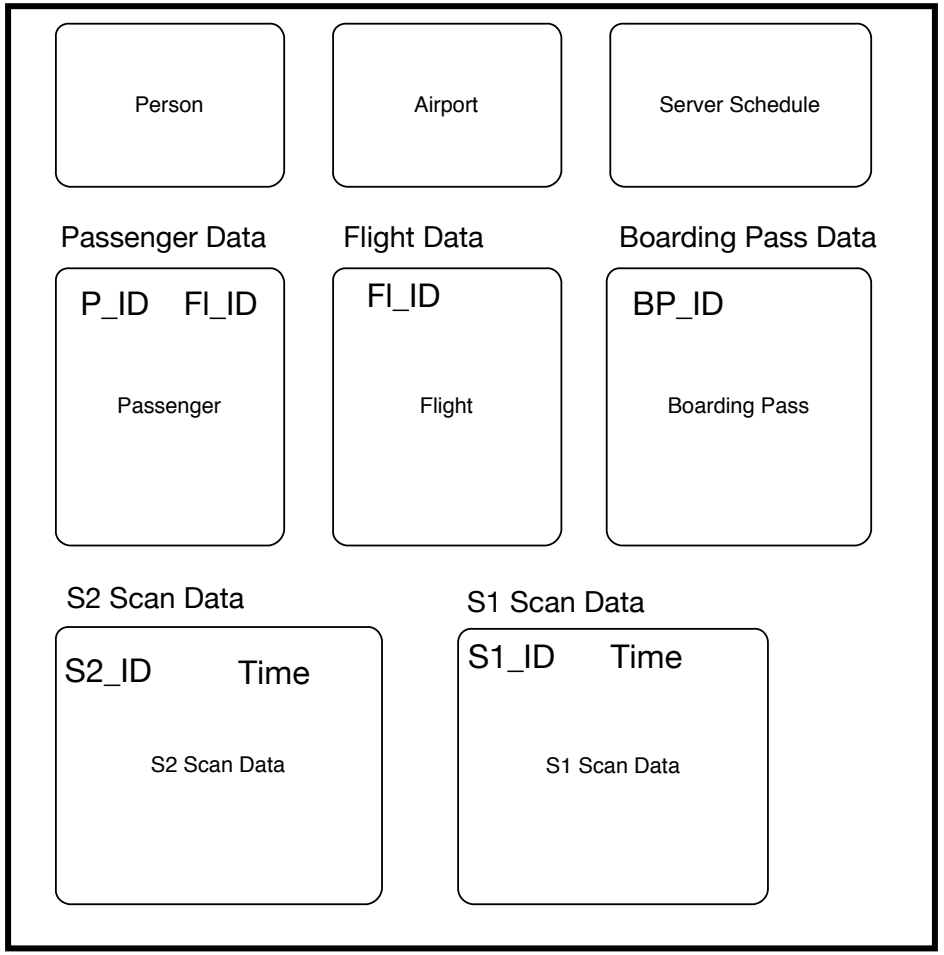
Prior to the in-depth interview: our understanding is '**neat**' but… **wrong**?



Post interview, the result looks **messy** – but we are now asking the **right** questions!

System Model: Expand, Make Explicit

```
=>Basic Concepts and Required Base Tables

Flight: A flight is travel by an airship from one location to another.
The airship takes off from the first location and lands at the second.
Flights can exist without passengers.

Passenger: A passenger is someone who has a ticket that allows them to
board an airship in one (specific) location and disembark from an
airship to enter another (specific) location

Base Tables:

We could, in principle, hold all of the necessary data to derive
information about about flights, trips and routes (see below) in a
small number of base tables:

Flight Table: information about airships that depart from one airfield
and arrive at another (e.g. time of departure, time of arrival)
Passenger Table: Information about individuals who boards an airship
at an airfield, travel in the airship and then disembark at an
airfield
Flight-Passenger Table: Information about which passengers have been
on which flights
Boarding Pass Table: Table with boarding pass information, linking
boarding passes to flights and passengers
Scanning Tables: Tables linking scans to boarding passes
```



Explicitly define concepts and create a formal model that can be **shared**.

Data Consolidation: Derived Datasets

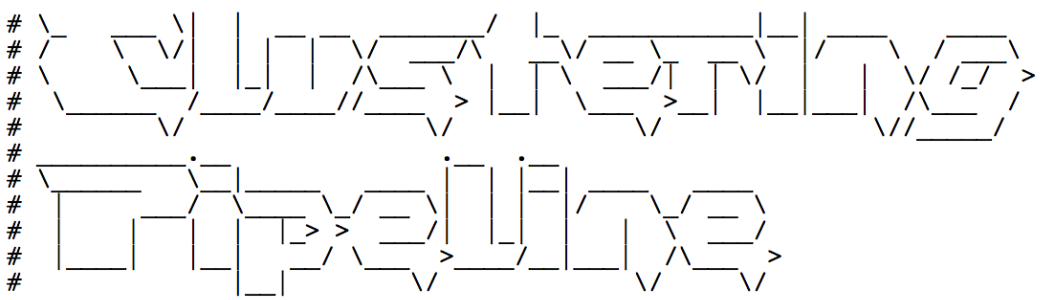
Having defined the concept of ‘a flight’ (within the system), we can generate a dataset of **flights** from the original dataset of passengers.

This also holds for **routes**, **trips**, etc.

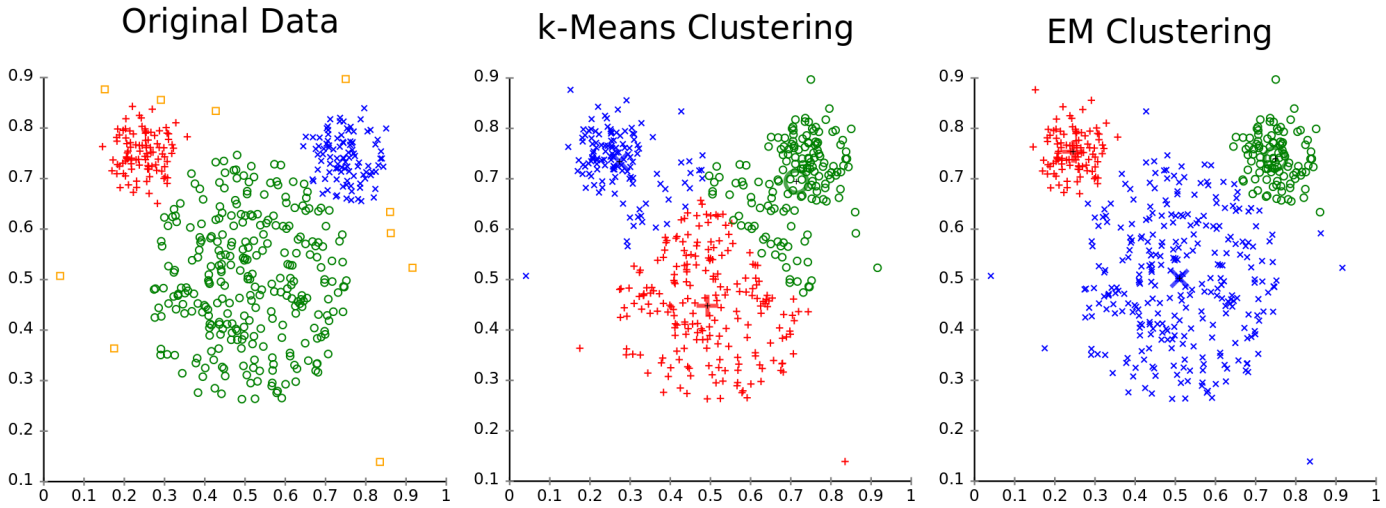
```
#-----------------------------------------------------#
#------------------ Description ----------------------#
#-----------------------------------------------------#

# Airfield: Originating airfield
# Flight_ID: Unique identifier for flights
# Sch_Departure: Scheduled departure time
# Act_Departure: Actual departure time
# Time_of_Day: Night, Morning, Afternoon, Evening, for Act_Departure
# Period_of_Week: Weekday and Weekend, for Act_Departure
# Day_of_Week: Monday, ... , Sunday, for Act_Departure
# Month: January, ... , December, for Act_Departure
# Season: Winter, Spring, Summer, Autumn, for Act_Departure
# Year: Year (2027 - 2029) for Act_Departure
# tot_pass: Total number of passengers on a flight
# N: Number of passengers with Wait_Time information
# min: min(Wait_Time) for a flight
# mean: average(Wait_Time) for a flight
# median: median(Wait_Time) for a flight
# max: max(Wait_Time) for a flight
# mean_WTL: proportion of passengers with Wait_Time > 4-hours
# mean_City_Flag:  proportion of passengers with BFO_Dest_City
information
# mode_BFO_Dest_City: mode(BFO_Dest_City) for a flight
# sum_city_mode: number of passengers with final destination mode city
# N_of_Dest_City: number of distinct BFO_Dest_City for a flight
# mode_BFO_Dest_Country_Code: mode(BFO_Dest_Country_Code) for a flight
# sum_city_mode: number of passengers with final destination mode
country
# N_of_Dest_Country: number of distinct BFO_Dest_Country_Code for a
flight
# Delay_in_Seconds: (Act_Departure - Sch_Departure) in seconds

# Note 1: N_of_Dest_City and N_of_Dest_Country includes the
mode_BFO_Dest_City/Country
```

Data Consolidation: Clustering

We can generate models that describe and structure the data. For example, we might write computer code that will **detect** and **define** cluster structures within the data.



```
#steps:

#transform dataset
#pick fields you want to use for clustering
#fix missing values
#gen. similarity/dissimilarity measures
#5a pick clustering parameters
#5b cluster
#validate/evaluate results
#visualize results
```



Clustering: finding meaningful subsets (clusters) of the dataset

Data Consolidation: Clustering

Analysis provides a condensed structuring of the data set, that can be further understood using visualization techniques. Ideally, this deepens understanding of the system as well.

Affinity Propagation Clustering Results

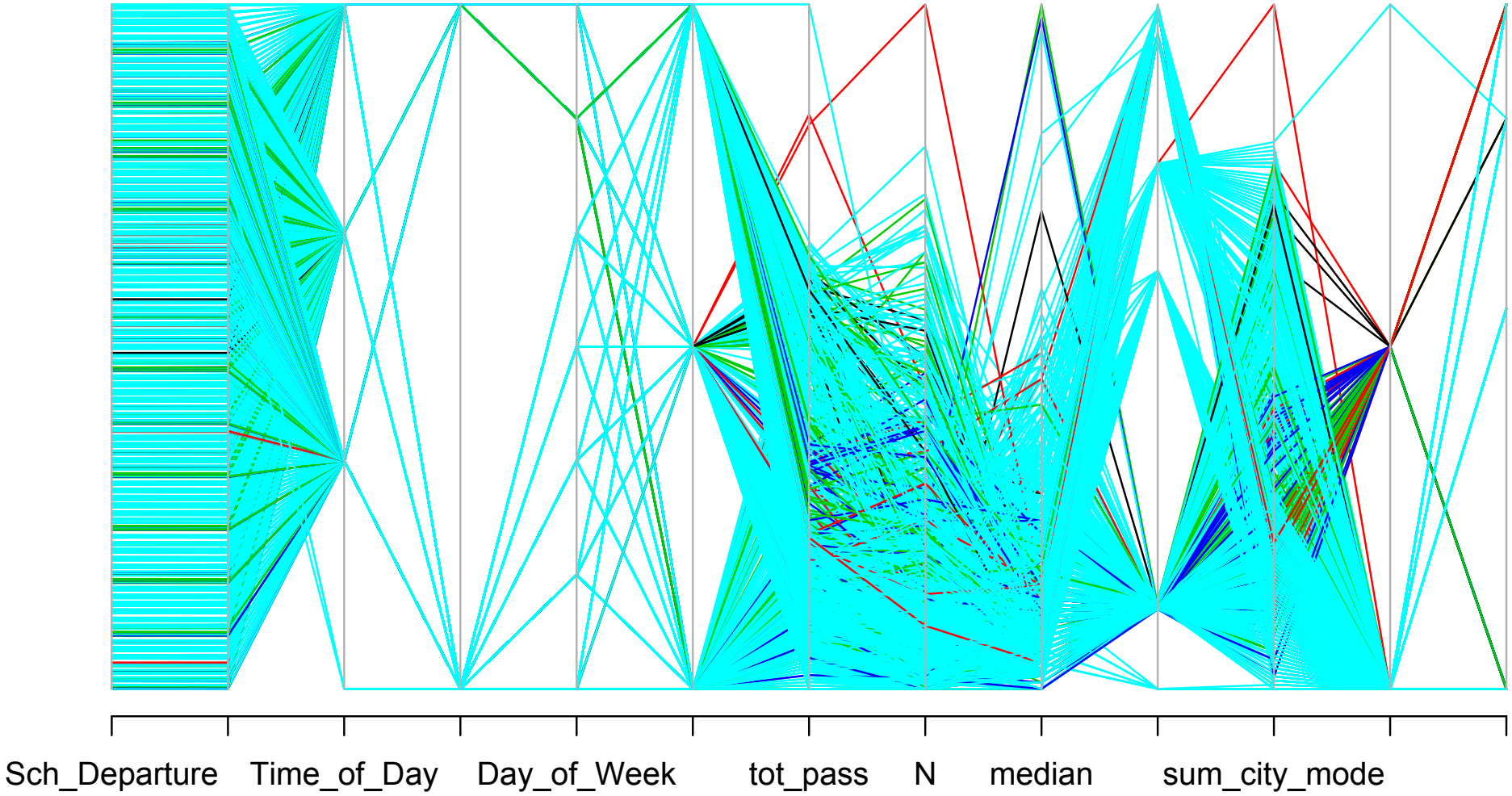
Cluster	1	2	3	4	5
Size	14	29	192	83	2057

DBSCAN Clustering Results

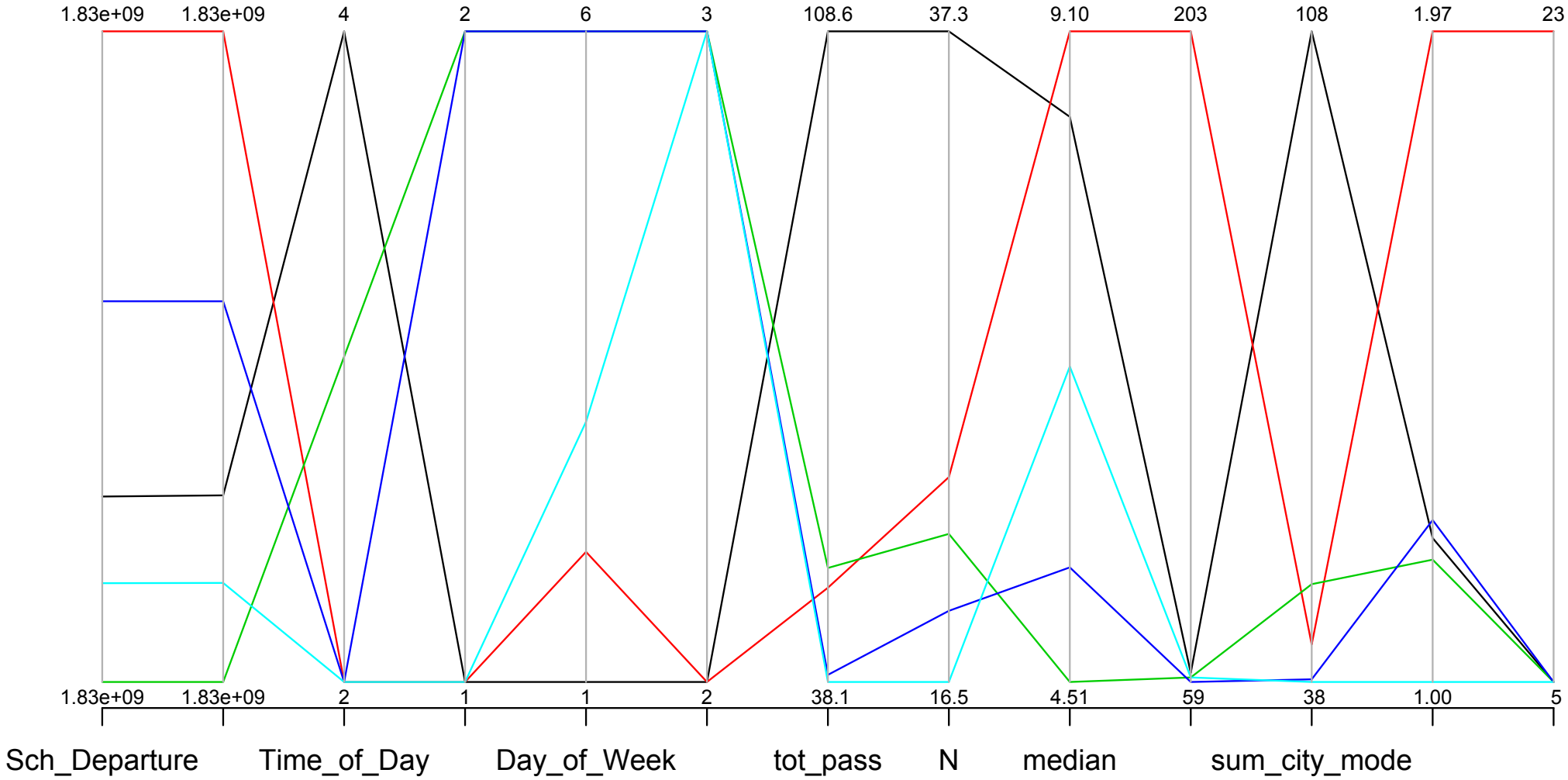
Cluster	Anomalies	1
Size	206	2169

Cluster	Size	Cluster Description
1	14	Evening weekday flights in February that go to SAF and mostly one destination city
2	29	Weekday flights with 2 final destination cities
3	192	Weekend flights that fly within Borealia and go to mostly one city
4	83	Weekend flights that go to CWL
5	2057	Flights that only go to one city

Data Consolidation: Clustering



Data Consolidation: Clustering



Data Consolidation: Queueing Systems

Modeling can also use data in more traditional pathways

Given

- arrival data
- processing data
- checkpoint utilization reports

can we predict the **number of servers** c that need to be opened in order to meet a **pre-defined performance level** (i.e. 90% of passengers waiting less than 15 min at PBS)?

Are there seasonal, daily, airport, or checkpoint components?

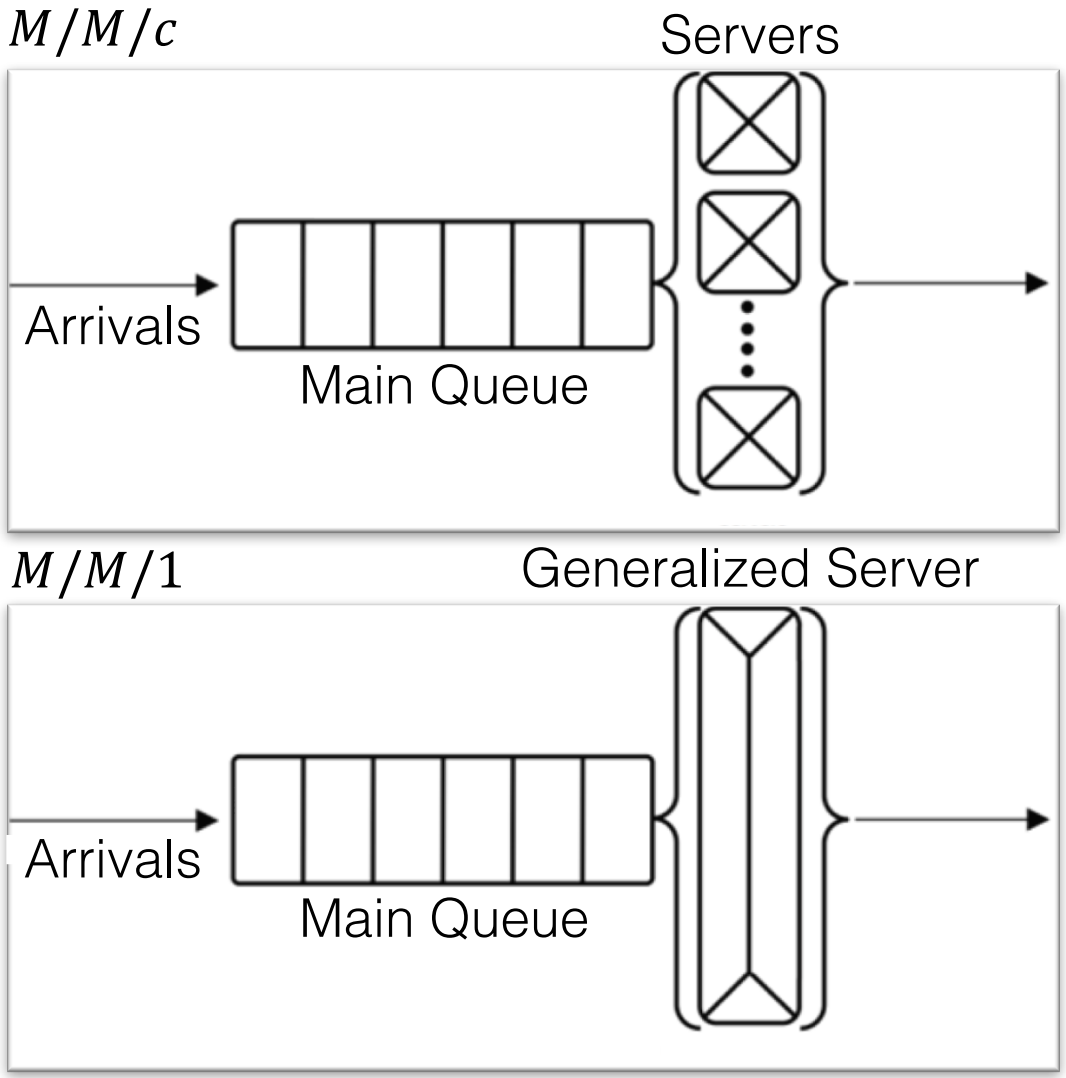
Data Consolidation: Queueing Systems

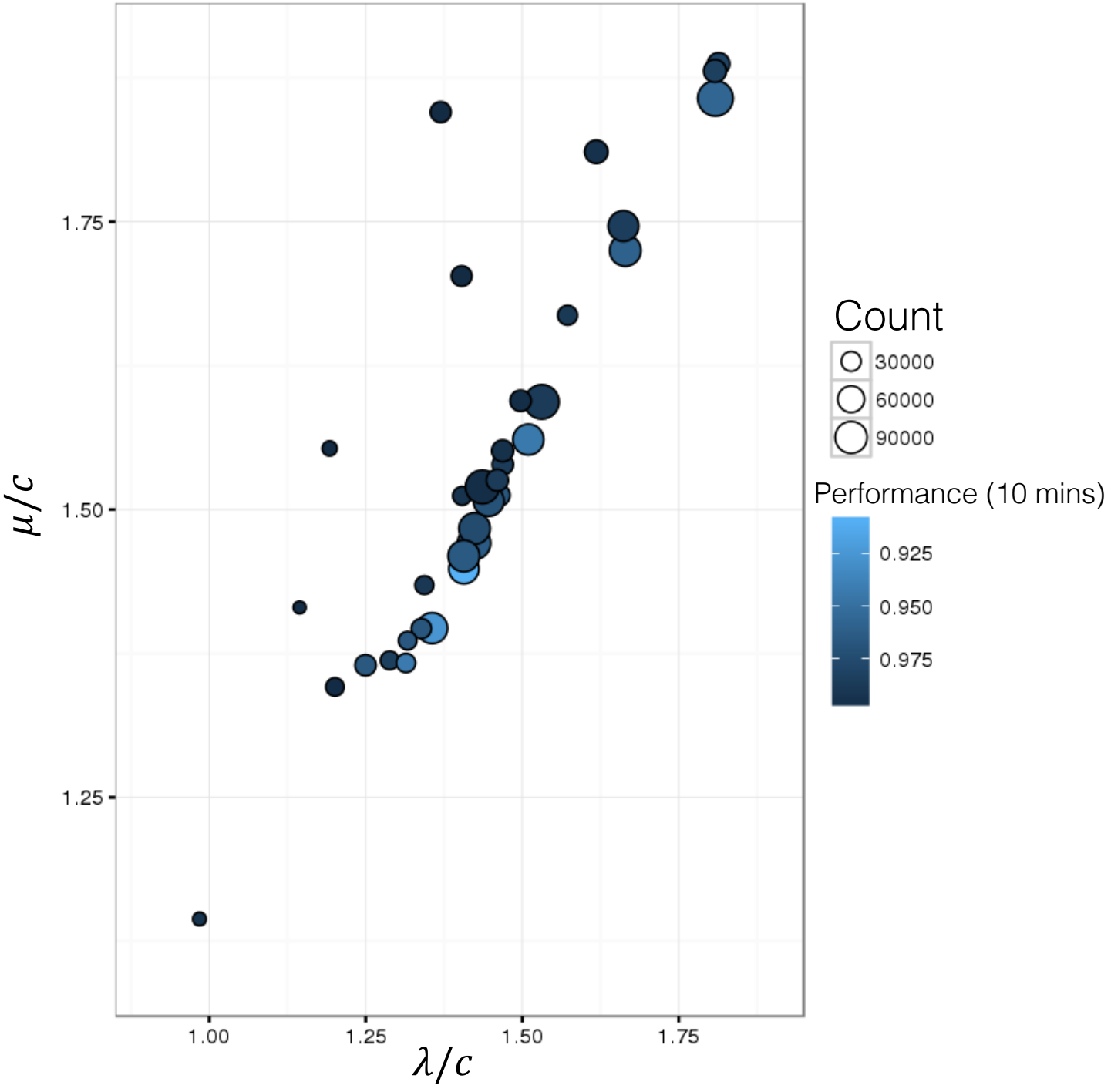
Methods

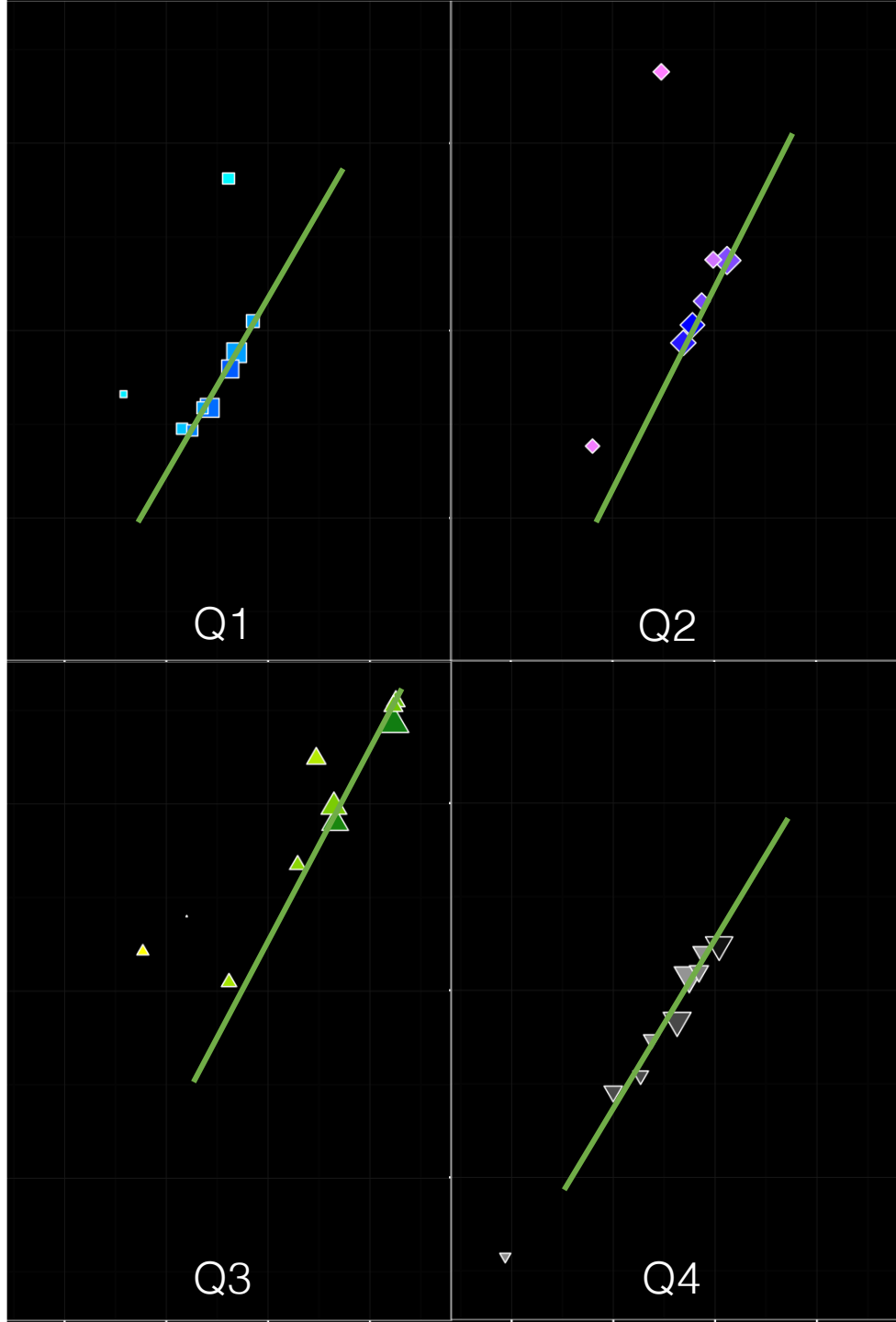
- queueing models
- statistical analysis of large datasets
- code optimization

Challenges

- Varying dataset quality
- difficulty in predicting arrivals
- unknown server vacation policies
- volume of observations to process




1.75
1.50
1.25
μ/c
1.00
1.25
1.50
1.75
λ/c
Count
30000
60000
90000
Performance (10 mins)
0.925
0.950
0.975

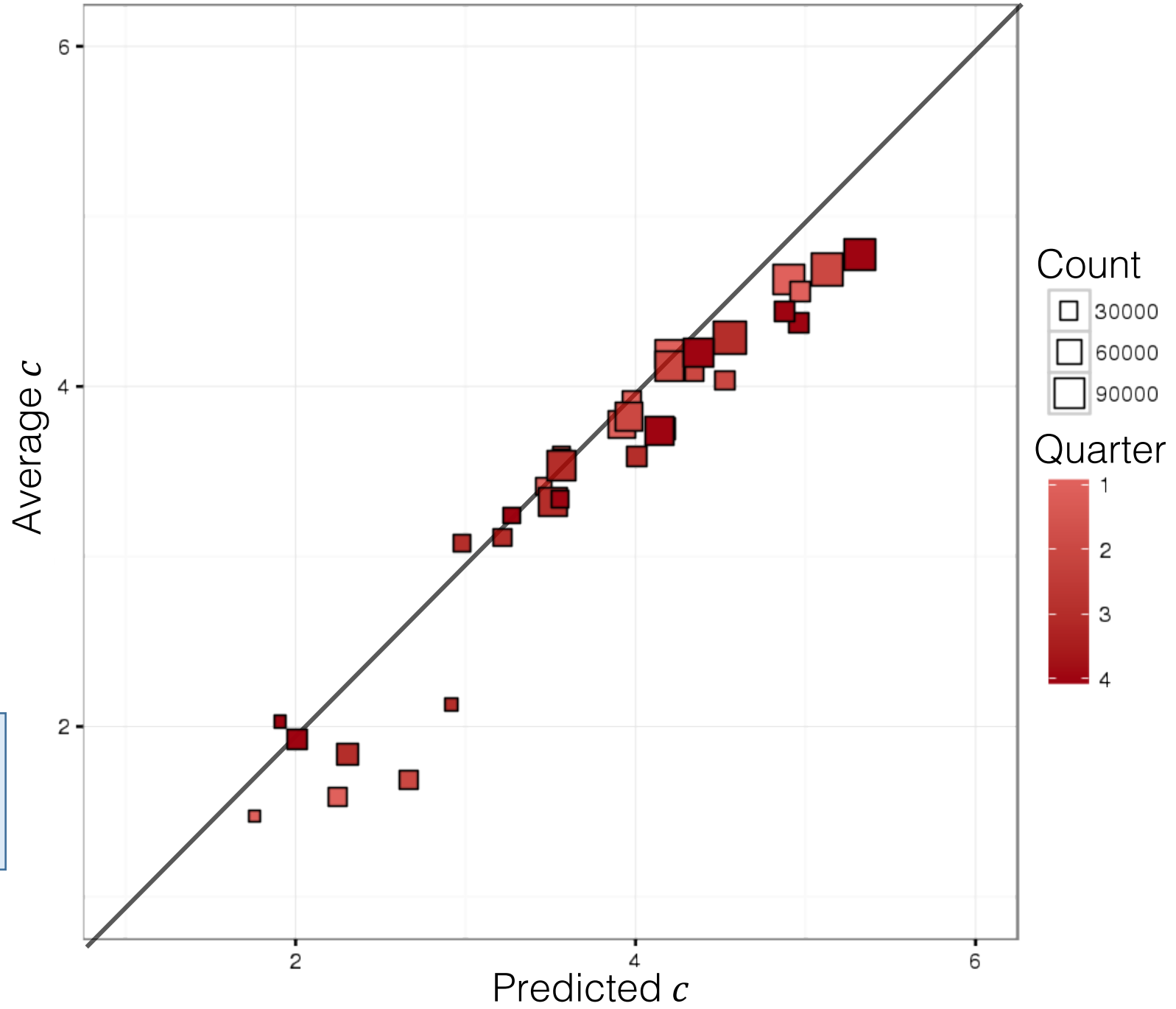

Q1
Q2
Q3
Q4

In a nutshell, we can find the processing rate μ in 2 ways:

- $\mu = ac + b\lambda$ from the data (time-intensive)
- $\mu = \frac{1}{x} W_0\left(\frac{\lambda x e^{\lambda x}}{1-p}\right)$ from queueing theory, for a prescribed (x, p)

Prediction:

$$\hat{c} = \frac{1}{ax}\left[W_0\left(\frac{\lambda x e^{\lambda x}}{1-p}\right) - b\lambda x\right]$$



Insights

Would we have predicted, on **theoretical grounds**, that:

- flights are more similar than dissimilar (in some clustering schema)?
- arrival rates per server and processing rates per server had a roughly linear relationship?
- the number of servers could be predicted with surprising accuracy by a theoretical formula?

Are these insights obvious, **after analysis**?

Are they counter-intuitive?

Take-Away: insights don’t always shatter pre-conceptions.

Data confirmation of domain expertise does not invalidate the analytical process, just as analytical refutation of some aspect of common knowledge is not a death-knoll for domain expertise.

Informed Implementation and Action

Insight for the sake of insight might be sufficient for specific purposes, but data analysis is costly – if insights are not also **actionable**, is the process worth it for your organization?

Can any of the insights be used by BASA to

- improve the data collection process (PBS scans)?
- set server vacation policy?
- modify the flight schedule?
- allocate PBS resources?
- identify candidate flights for re-routing?
- etc.

Take-Away: having a strategy to use any eventual insight provided by data analysis is going to go a long way towards helping you understand and control your “world”.

It will also a play a crucial role in your evaluation of the process, in your ability to answer the question: “was it worth it?”