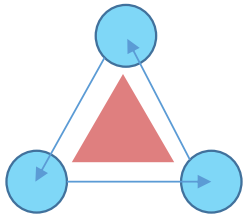


Indicator Frameworks

Joshua Tan (University of Oxford), Christine Kendrick (City of Portland),
Abhishek Dubey (Vanderbilt University), and Sokwoo Rhee (NIST)

Applied Category Theory @ NIST

March 15, 2018

Three ideas

- Use the **data** you have.
- **Science** is a (natural) language.
- **Correlation** “correlates” with **causation**.

Start with the data you have,
not the data you want.

Why indicator frameworks?

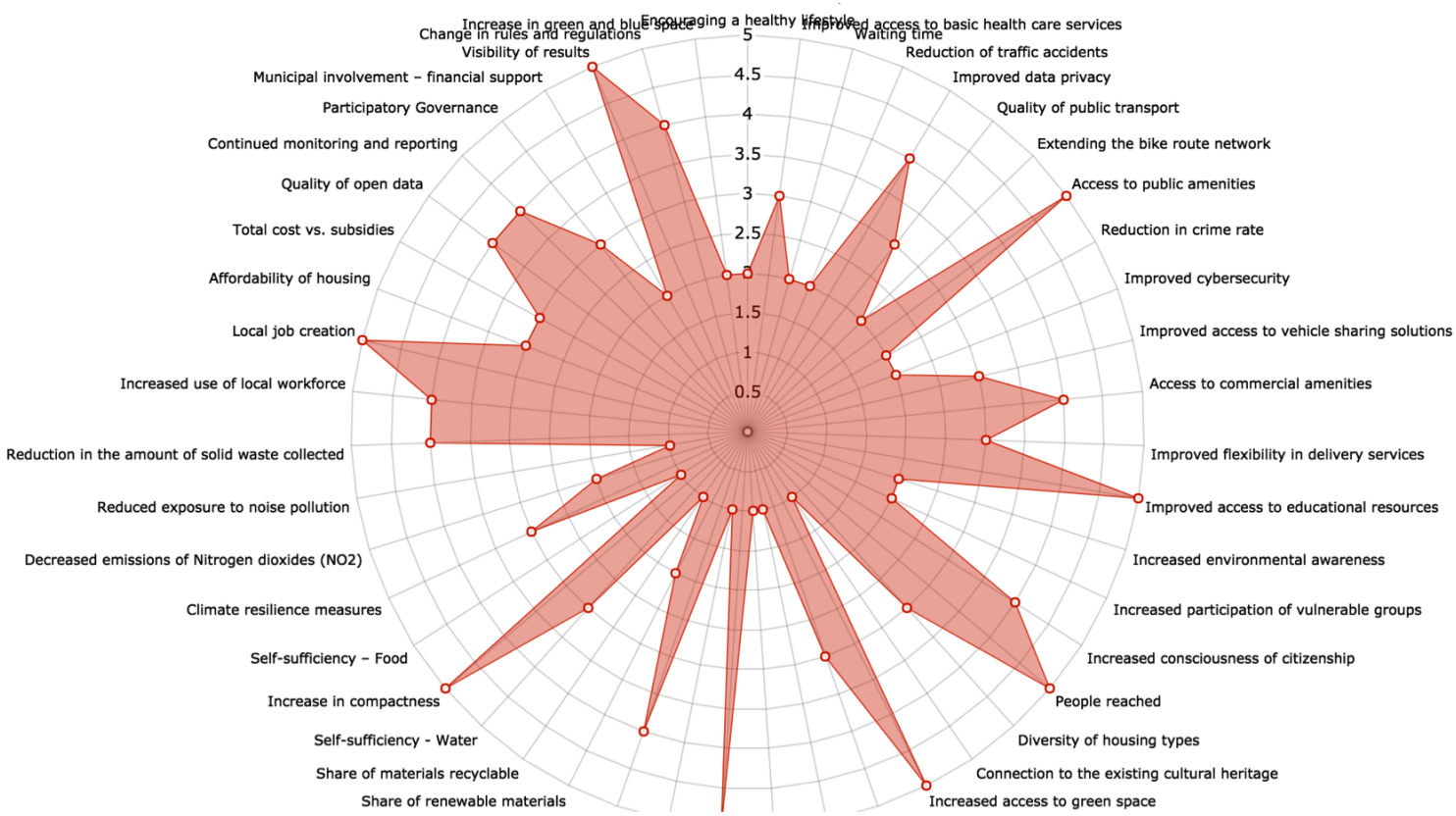
- Sets the strategic **priorities** of your project, department, or city
- Communicate **progress** to stakeholders
- Enable **policy reactions** to data, especially in the optimization of processes
- **Simplify** your situation

* but mainly: they're everywhere, and they're simple

What are indicators?

- Basically: a column of data, usually time-series.

Indicator: "Access to public amenities"
Description: It is presumed that nearby availability of amenities leads to a lively neighbourhood and less car use. Amenities in the urban environment make an area more enjoyable and contribute to its desirability. It is assumed that these factors contribute to the success of smart city projects.
Definition: The extent to which public amenities are available within 500m
Calculation: Likert scale (1-5) 1. No amenities: no public amenities whatsoever are available (e.g. no basic nor additional). 2. Relatively few amenities: only few basic public amenities are available (e.g. a small park). 3. A reasonable number of amenities: basic public amenities are available including a few important amenities such as a park and a community center. 4. A sufficient number of amenities: basic public amenities are widely available (e.g. open green spaces, public recreation) as well as many important public amenities (theatres). 5. Relatively many amenities.
Expected data source: Google maps; project documentation and/or interviews with project leader, planning documents
Expected availability: High (everyone can access google maps); other relevant information should be available at the city planning office
Collection interval: After the project, but can also be used ex-ante to evaluate plans
Expected reliability: Because of the subjectivity that cannot be excluded, this indicator is not 100% reliable.
Expected accessibility: As a component of a successful project and selling point in a marketing sense, it is expected that this information will be accessible. No sensitivities expected.


Encouraging a healthy lifestyle
Improved access to basic health care services
Waiting time
Reduction of traffic accidents
Improved data privacy
Quality of public transport
Extending the bike route network
Access to public amenities
Reduction in crime rate
Improved cybersecurity
Improved access to vehicle sharing solutions
Access to commercial amenities
Improved flexibility in delivery services
Improved access to educational resources
Increased environmental awareness
Increased participation of vulnerable groups
Increased consciousness of citizenship
People reached
Diversity of housing types
Connection to the existing cultural heritage
Increased access to green space
Share of renewable materials
Share of materials recyclable
Self-sufficiency - Water
Increase in compactness
Self-sufficiency – Food
Climate resilience measures
Decreased emissions of Nitrogen dioxides (NO2)
Reduced exposure to noise pollution
Reduction in the amount of solid waste collected
Increased use of local workforce
Local job creation
Affordability of housing
Total cost vs. subsidies
Quality of open data
Continued monitoring and reporting
Participatory Governance
Municipal involvement – financial support
Visibility of results
Change in rules and regulations
Increase in green and blue space
5
4.5
4
3.5
3
2.5
2
1.5
1
0.5

Problems with the one-by-one approach

- It's expensive.
- It's subjective.
- It's ad hoc.
- ~~Is it worth it?~~
- There's just not enough data.*

* data that you **want** vs data that you **have**

Towards a science of measuring systems

- Cities are **cyber-physical systems**
- Cities are **systems of systems**
- These models are *mathematical descriptions*. They do not measure anything, per se.
- Other mathematical descriptions:
  - Network approaches
  - Economic models
  - Game theory

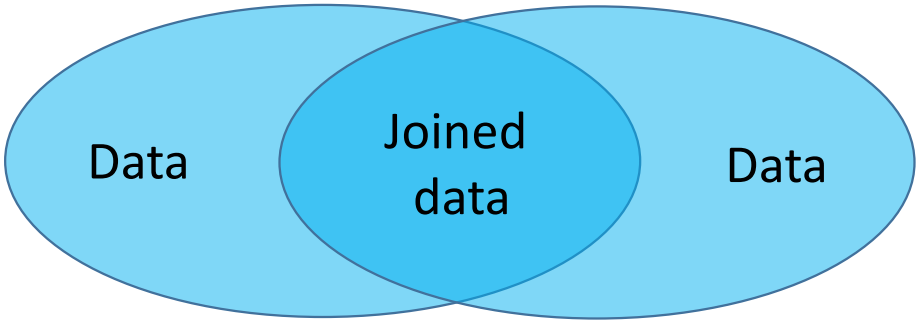
Problems with the systems approach

- People end up building toy models
  - See 90% of academic studies
- Or they build highly specific, technical models… thus not “systems”
- Or they build giant, unwieldy models
  - 175 indicators in CITYKeys, 212 in the Boston Indicators Project, etc.
  - More than 43 (!) indicator frameworks built for “smart and sustainable cities”
- The point is: models are useful locally, but they’re hard to sustain on bigger systems

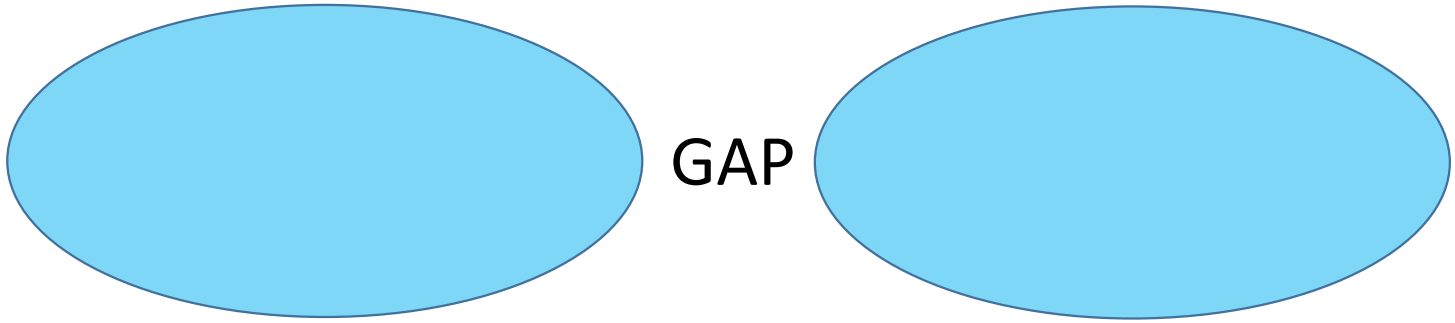
Towards a science of measuring cities

- Problem: the **models** fail to describe the world perfectly. (Duh)
- Problem: the **data** doesn't either.*

* cities are **complex**: there isn't enough raw data to describe all the interactions



Joined data



How do we join this data?

The goal

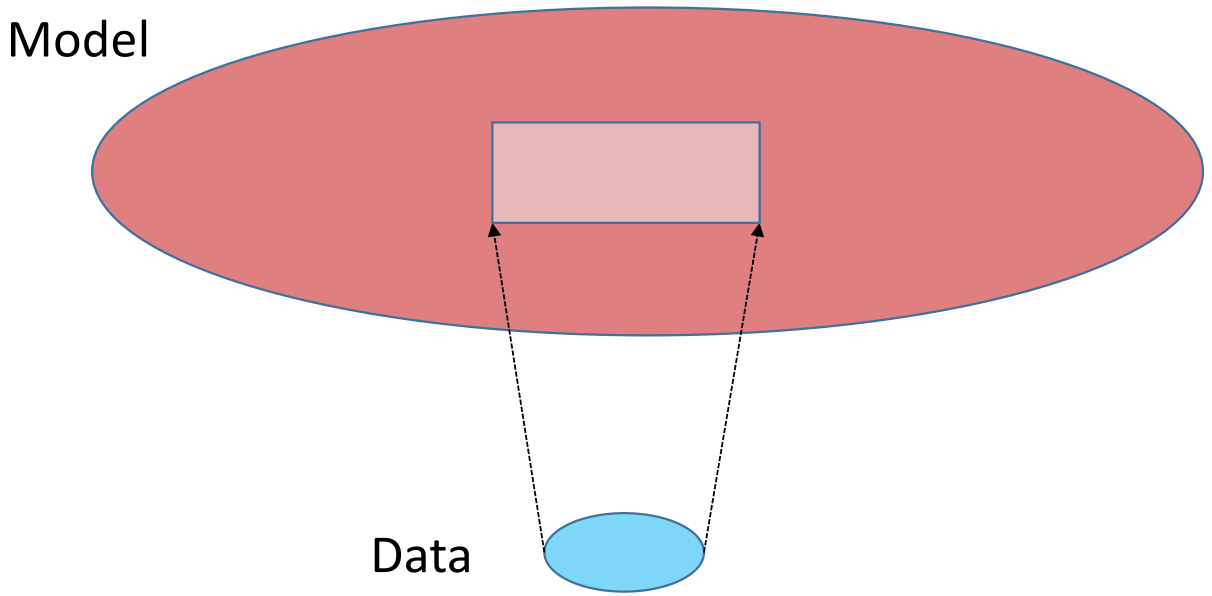
Compute the correlation between two indicators even *without* data.

- Step 1: give a **mathematical semantics** for indicator sets.
- Step 2: test whether indicator sets can be upgraded to synthesize "**models** over **data**".

Abstract indicator frameworks

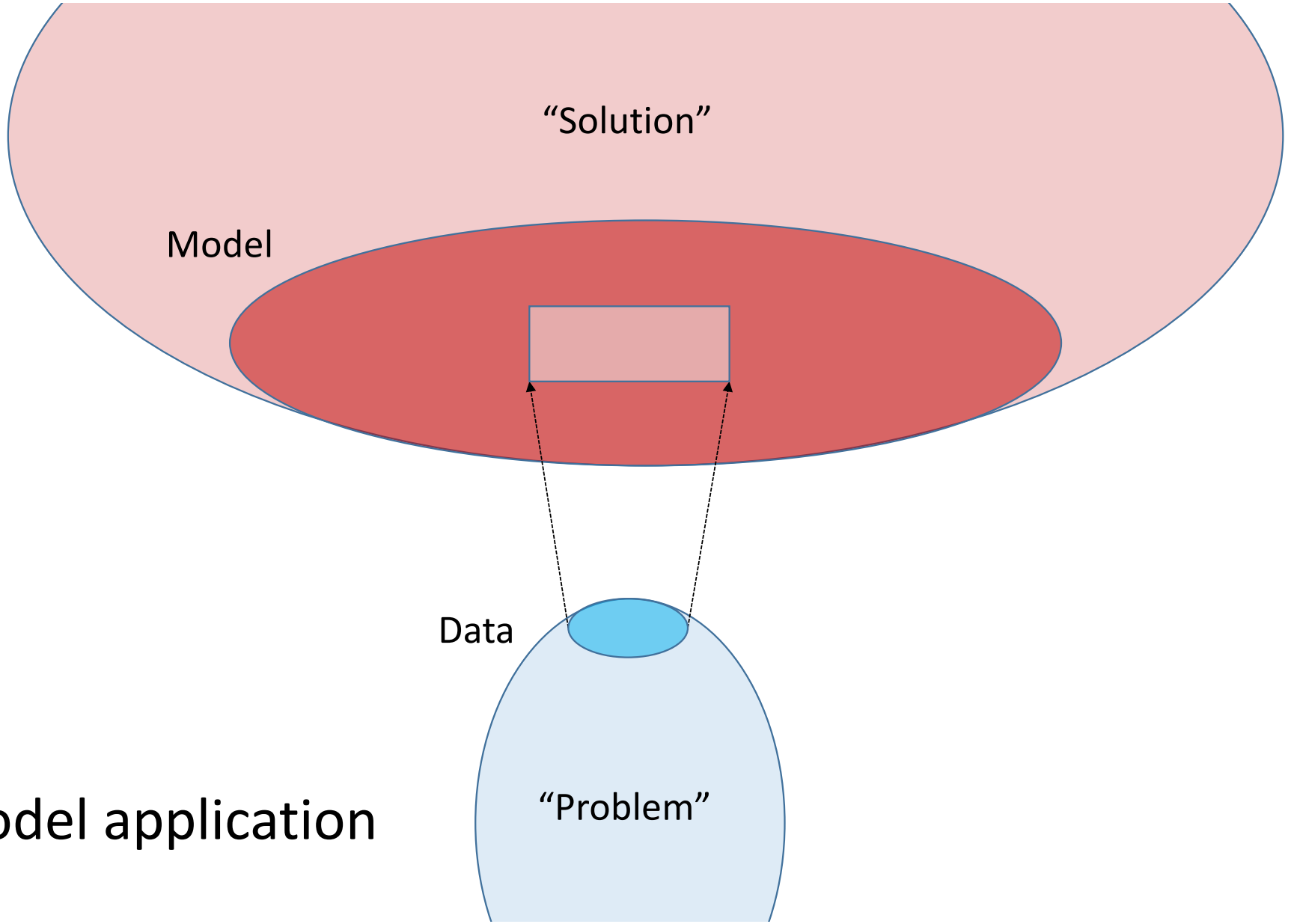
- Whiteboard!

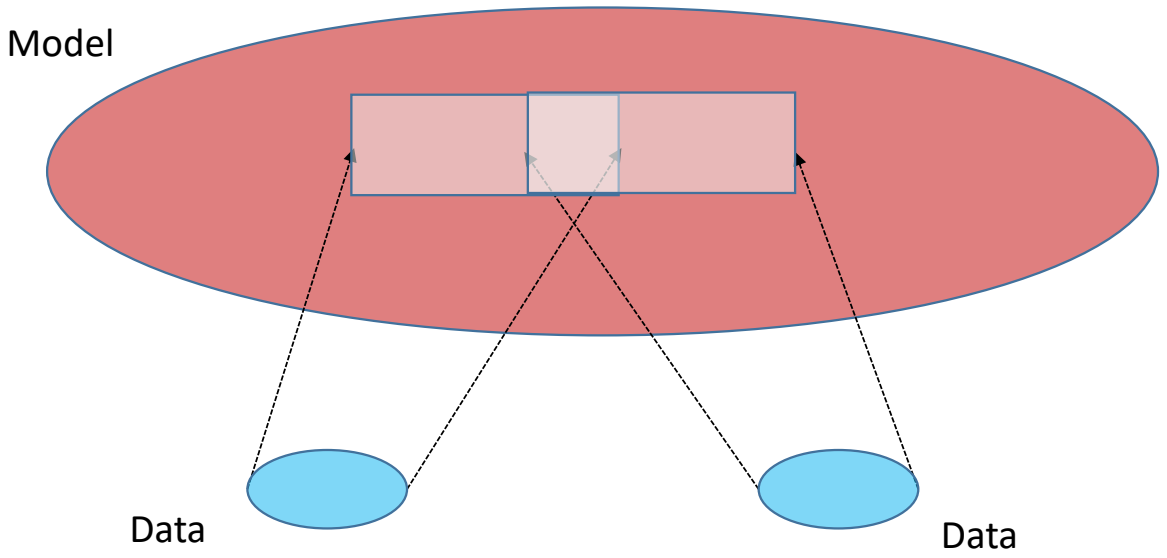
Data informs models,
while models constrain data.



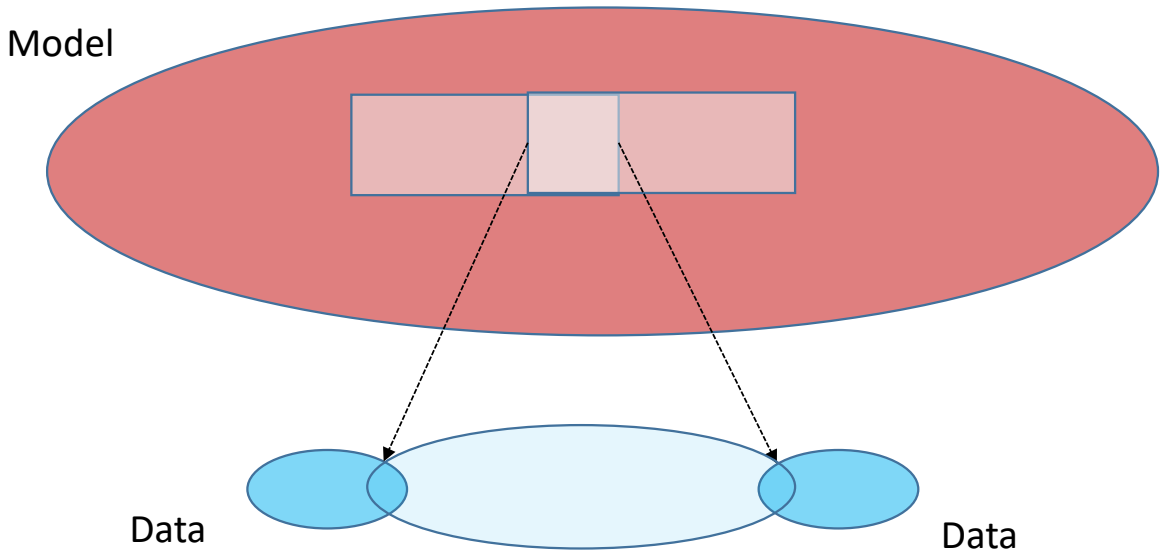
Model construction

Model application

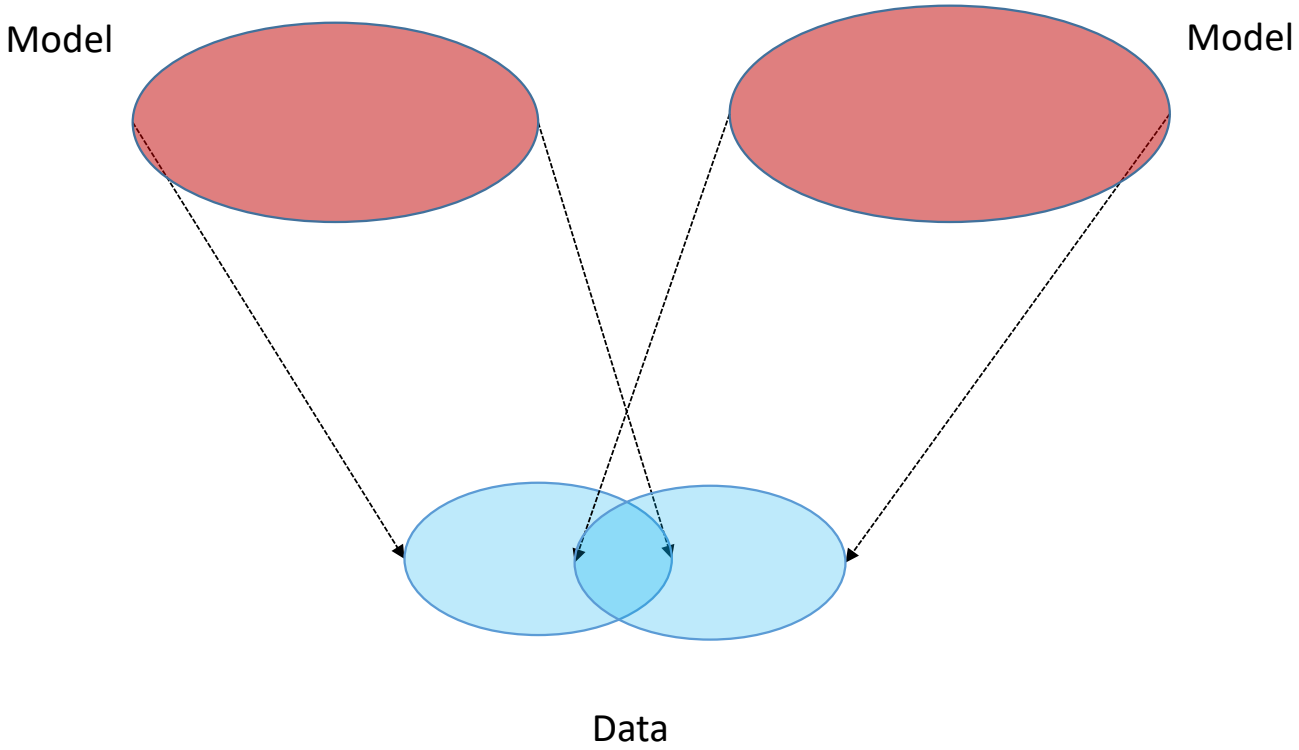




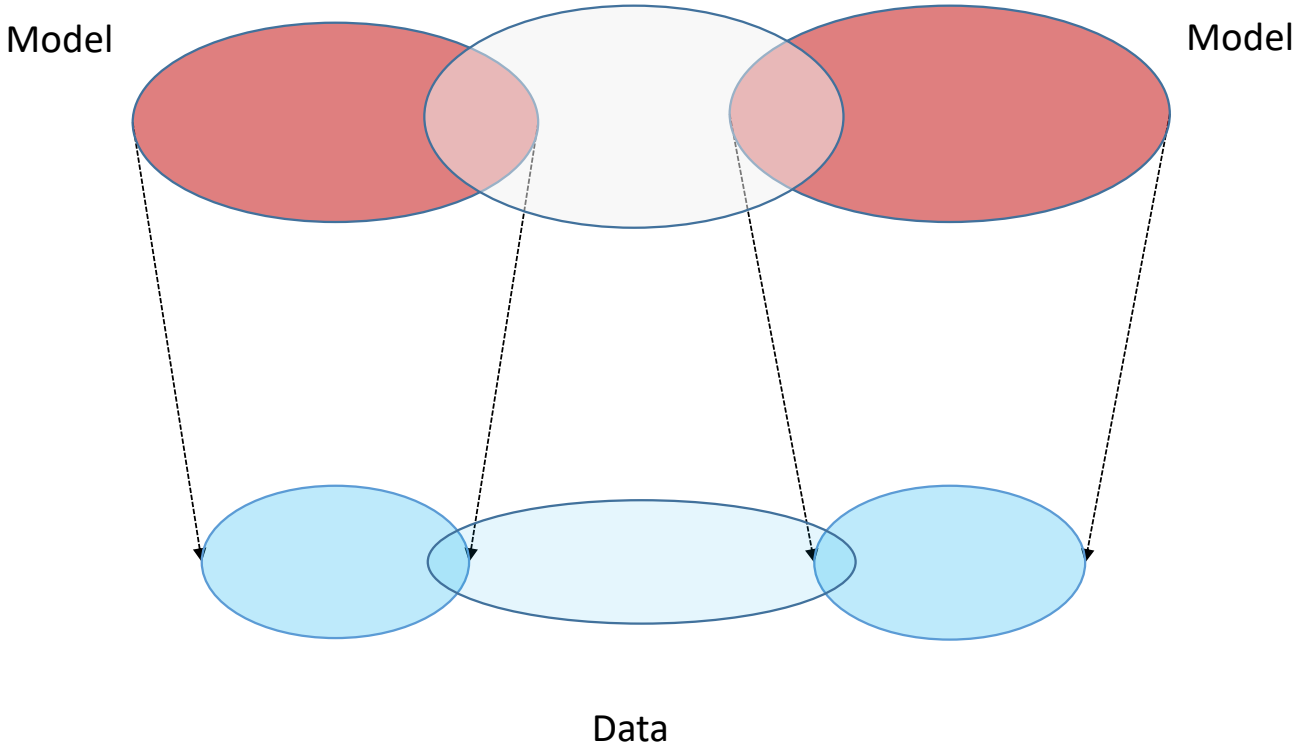
Model validation



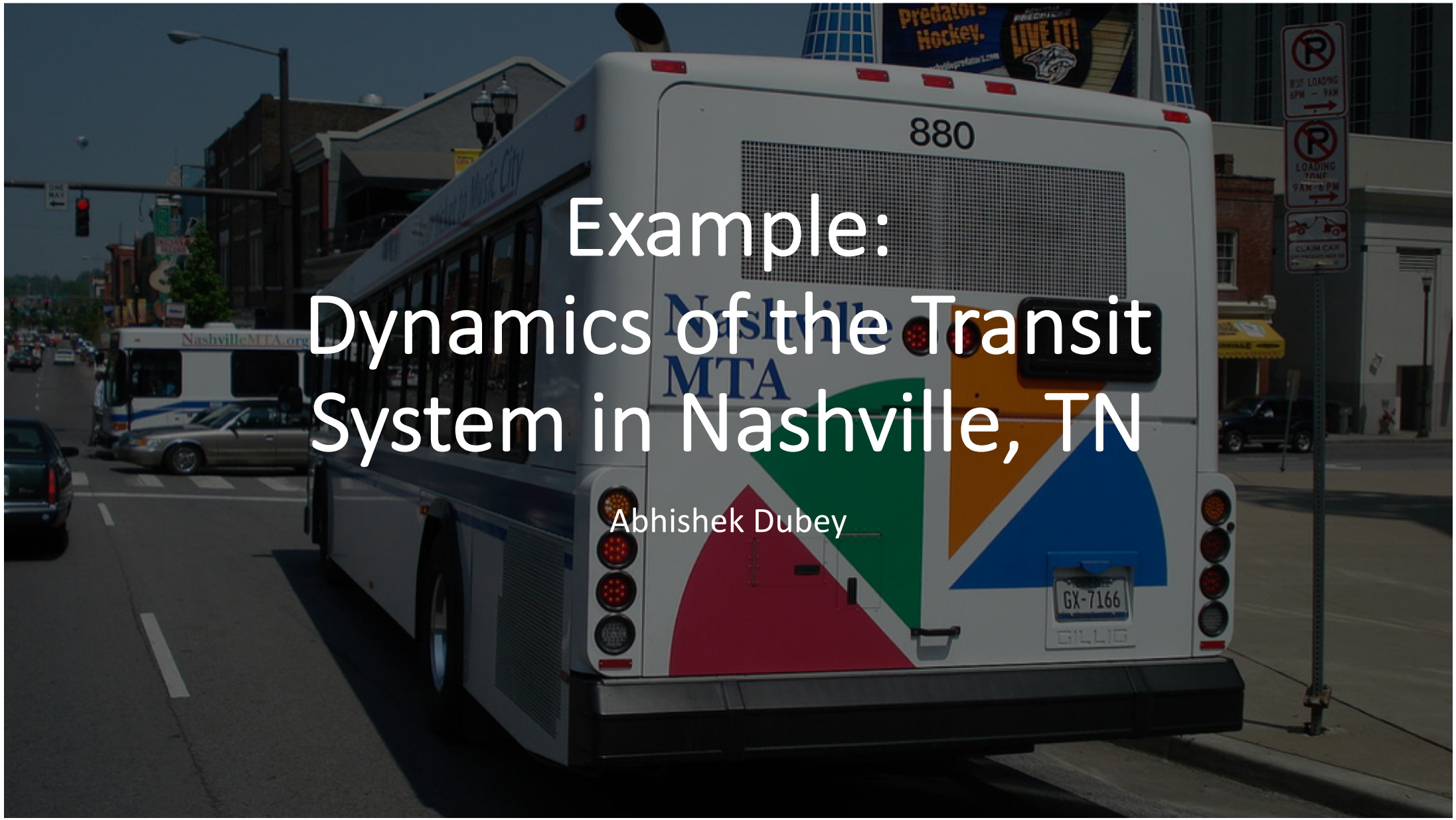
“Integrating under a model”



Data simulation & integration



The general case?



Example: Dynamics of the Transit System in Nashville, TN

Abhishek Dubey

Date	Weekday (Mon=0)	Segment	Direction	Route	Actual Travel (Seconds)	Scheduled Travel (Seconds)	Actual Arrivel Time (24hr)	Trip_ID	Driver_ID	Jam Factor (0~10)	Actual Traffic Speed (mph)	Free Flow Speed Limit (mph)
20161011	1	MCC5_11 - HFOGG	FROM DOWNTOWN	1	826.99998	360	15:35:39	126346	1683	3.487592973	13.24569763	19.82003285
20161011	1	HFOGG - MTWD	FROM DOWNTOWN	1	580.99998	600	15:48:05	126346	1683	3.937953674	11.96260066	18.78503114
20161011	1	MTWD - 100O	FROM DOWNTOWN	1	805.99998	660	16:01:31	126346	1683	2.793803409	17.05275554	23.36003872
20161011	1	MCC5_11 - HFOGG	FROM DOWNTOWN	1	802.99998	420	16:38:47	126347	1683	5.28474093	10.79350626	19.82003285
20161011	1	HFOGG - MTWD	FROM DOWNTOWN	1	648	600	16:49:35	126347	1683	4.822438899	10.89503857	18.78503114
20161011	1	MTWD - 100O	FROM DOWNTOWN	1	507.99996	600	16:58:03	126347	1683	0.024484444	26.24559907	23.36003872
20161011	1	MCC5_11 - HFOGG	FROM DOWNTOWN	1	900.99996	420	17:37:25	126348	1683	4.263709706	12.1785496	19.82003285
20161011	1	HFOGG - MTWD	FROM DOWNTOWN	1	609.99996	540	17:47:35	126348	1683	4.862932428	10.63845745	18.78503114

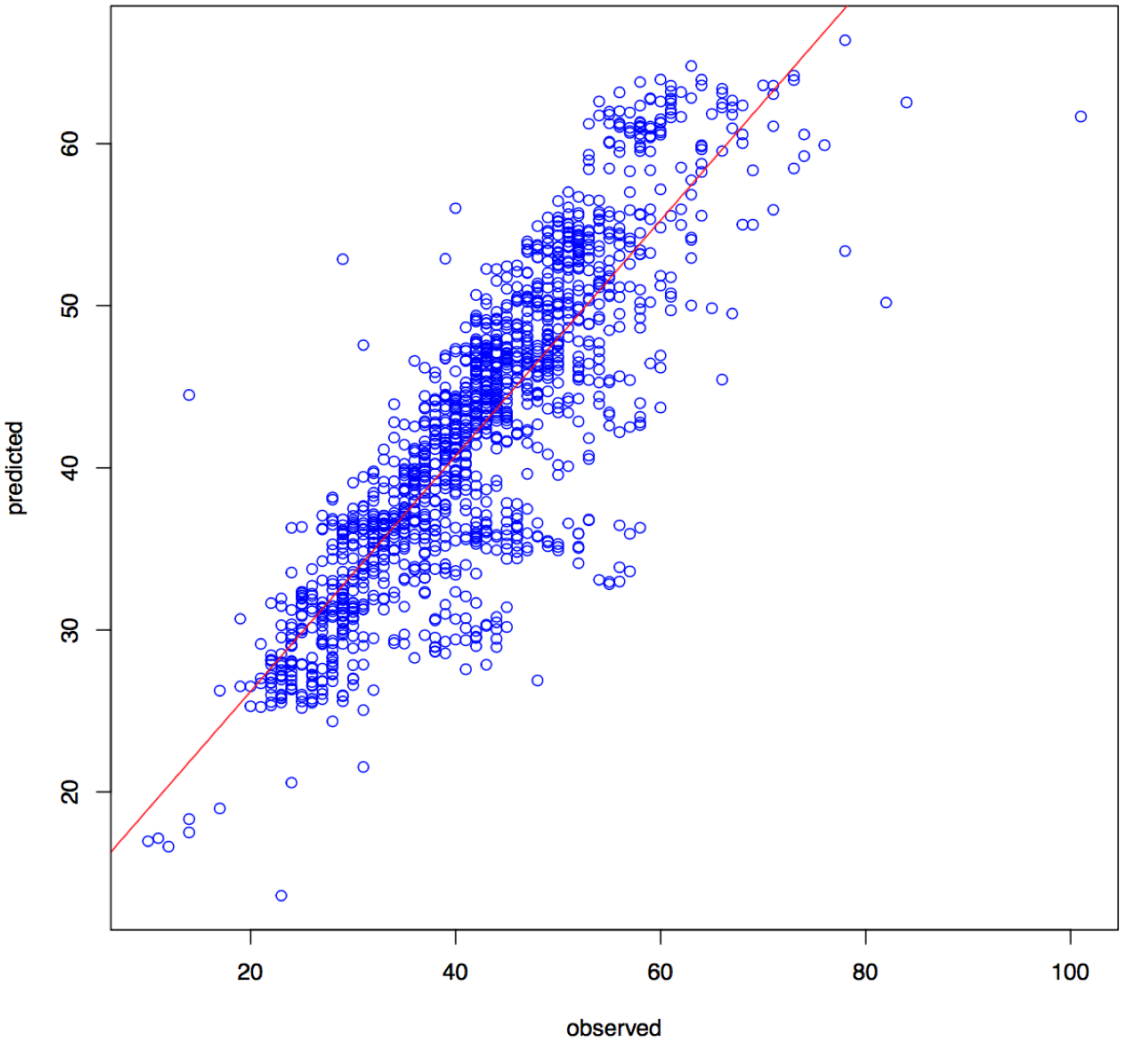


Figure 2. Observed Versus Predicted Travel Time.

Scheduled Travel Time

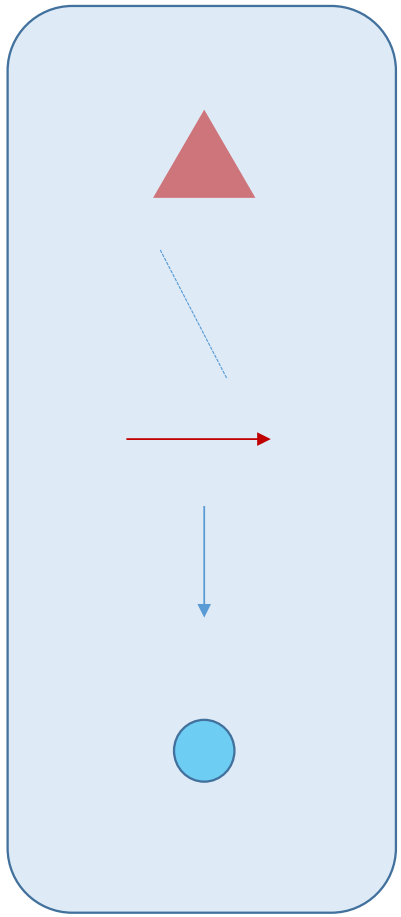
Predicted Travel Time

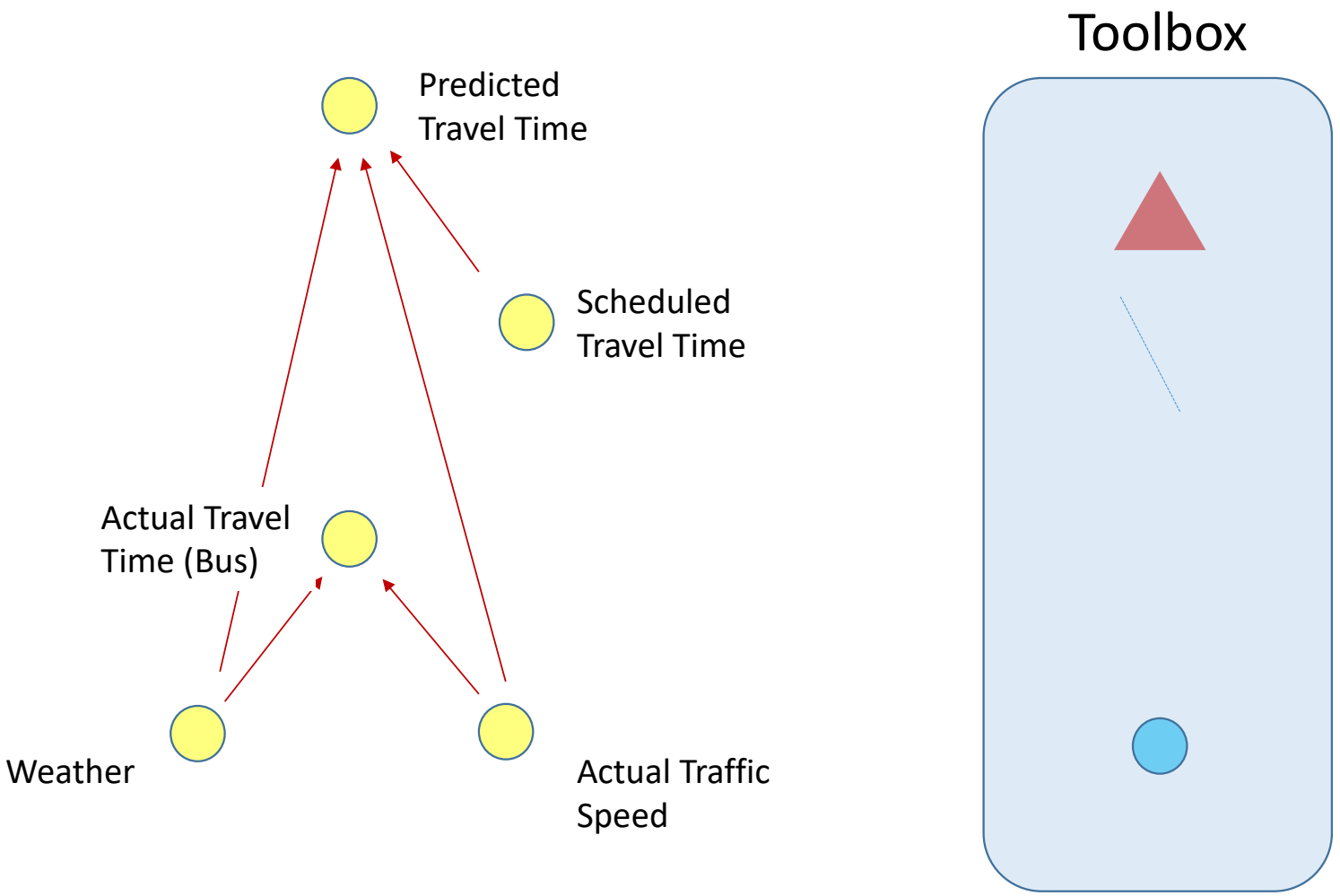
Actual Travel Time (Bus)

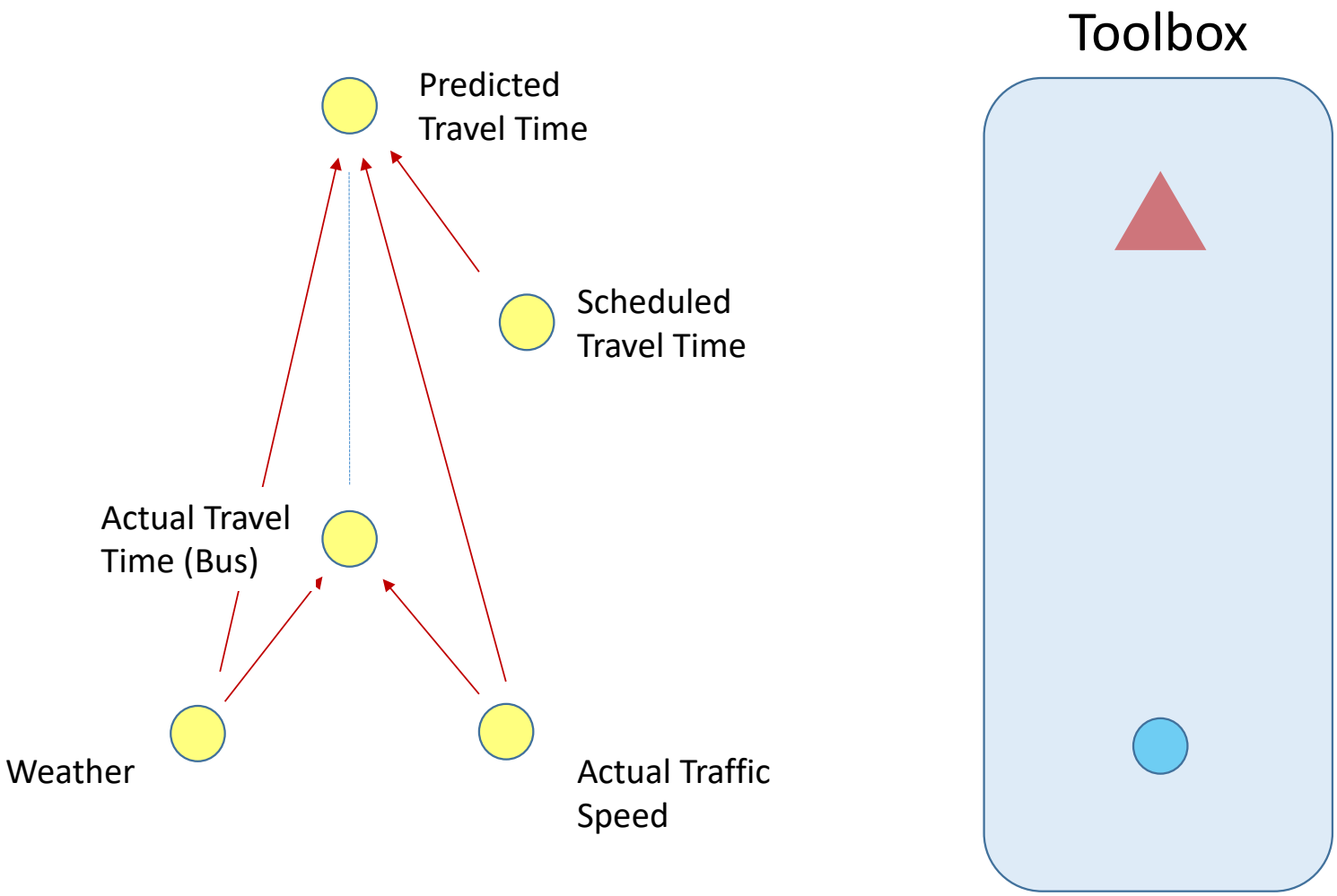
Weather

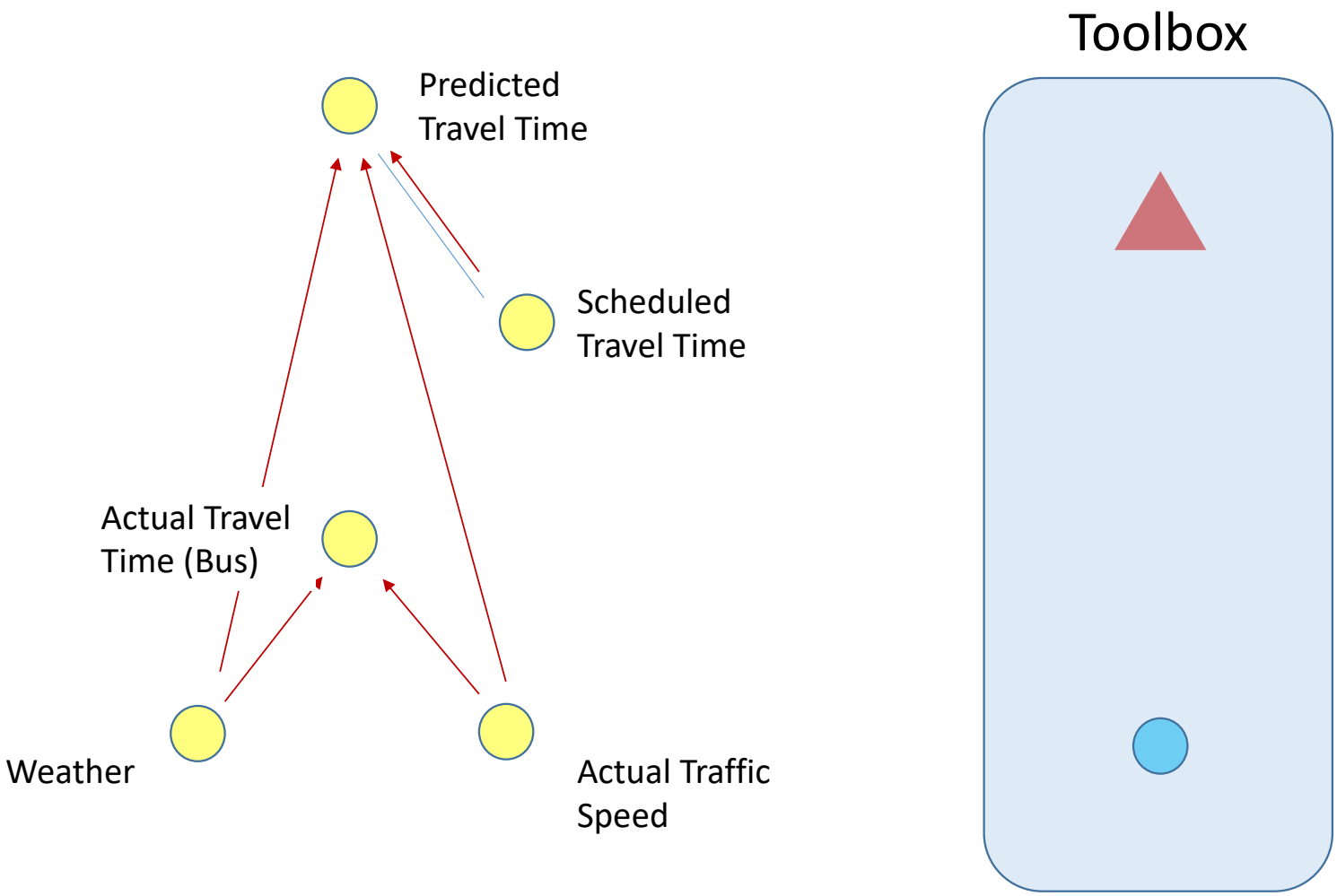
Actual Traffic Speed

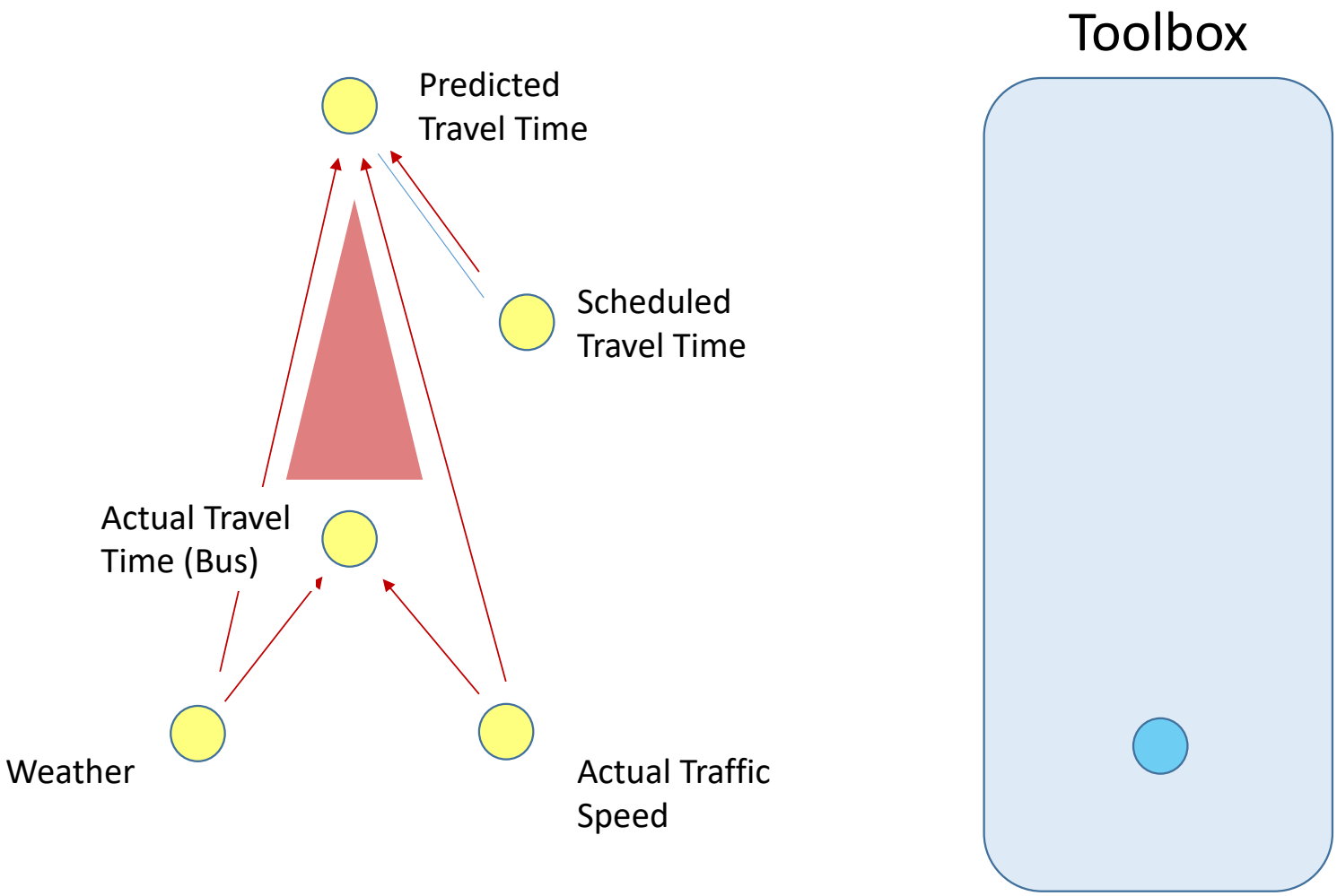
Toolbox

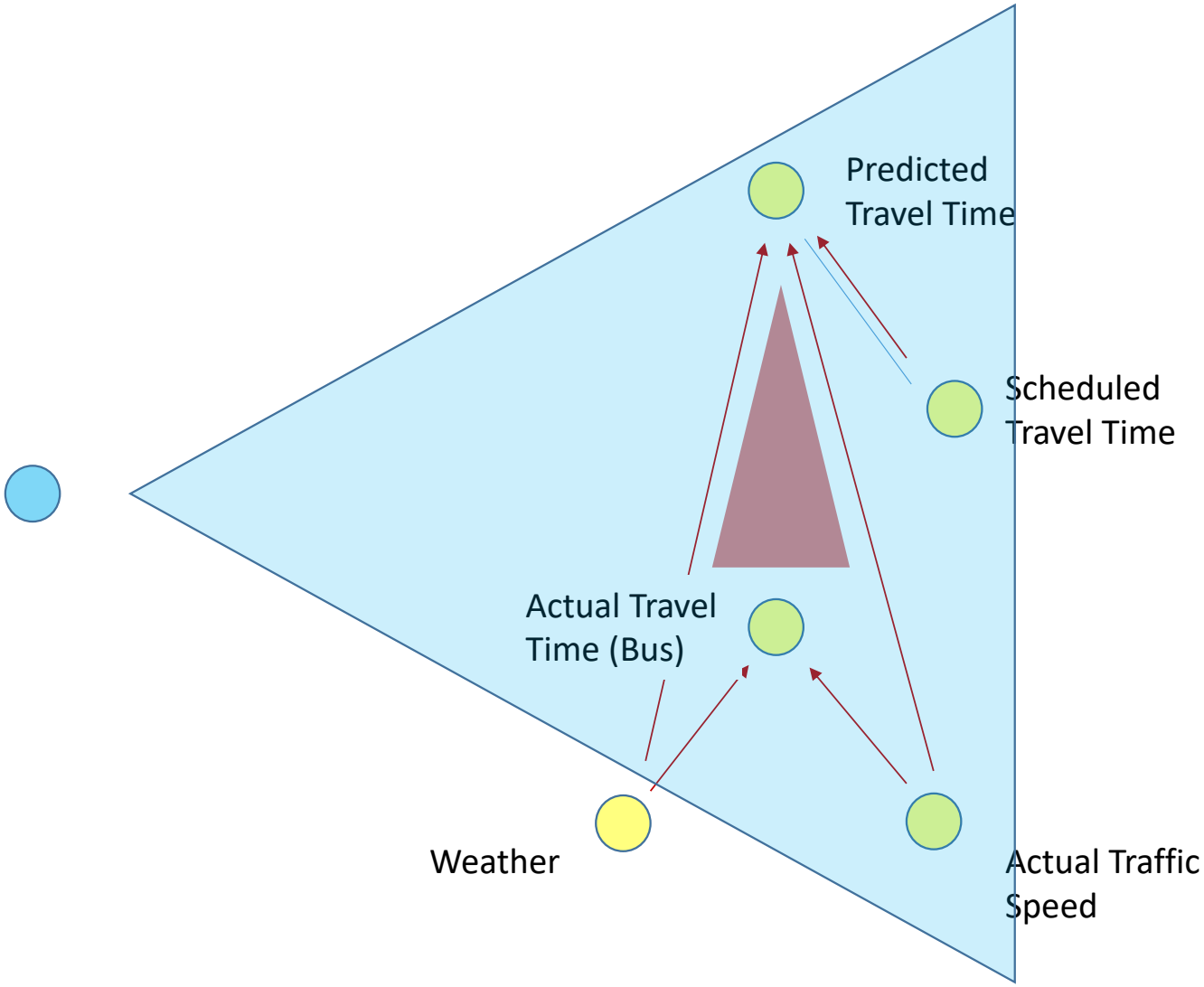



Predicted
Travel Time
Scheduled
Travel Time
Actual Travel
Time (Bus)
Weather
Actual Traffic
Speed
Toolbox

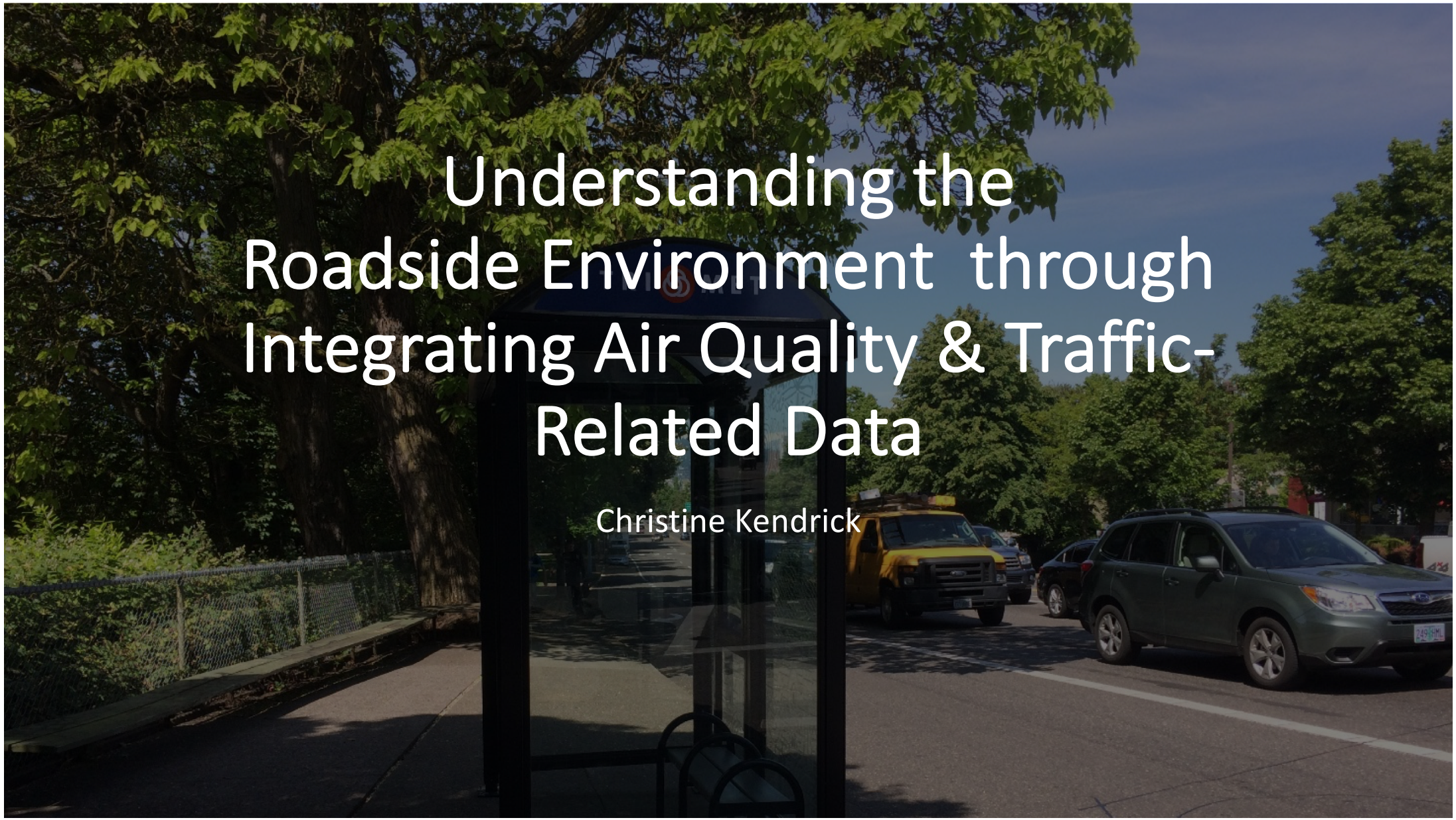

Predicted
Travel Time
Scheduled
Travel Time
Actual Travel
Time (Bus)
Weather
Actual Traffic
Speed
Toolbox


Predicted
Travel Time
Scheduled
Travel Time
Actual Travel
Time (Bus)
Weather
Actual Traffic
Speed
Toolbox


Predicted
Travel Time
Scheduled
Travel Time
Actual Travel
Time (Bus)
Weather
Actual Traffic
Speed
Toolbox


Predicted
Travel Time
Scheduled
Travel Time
Actual Travel
Time (Bus)
Weather
Actual Traffic
Speed

Toolbox



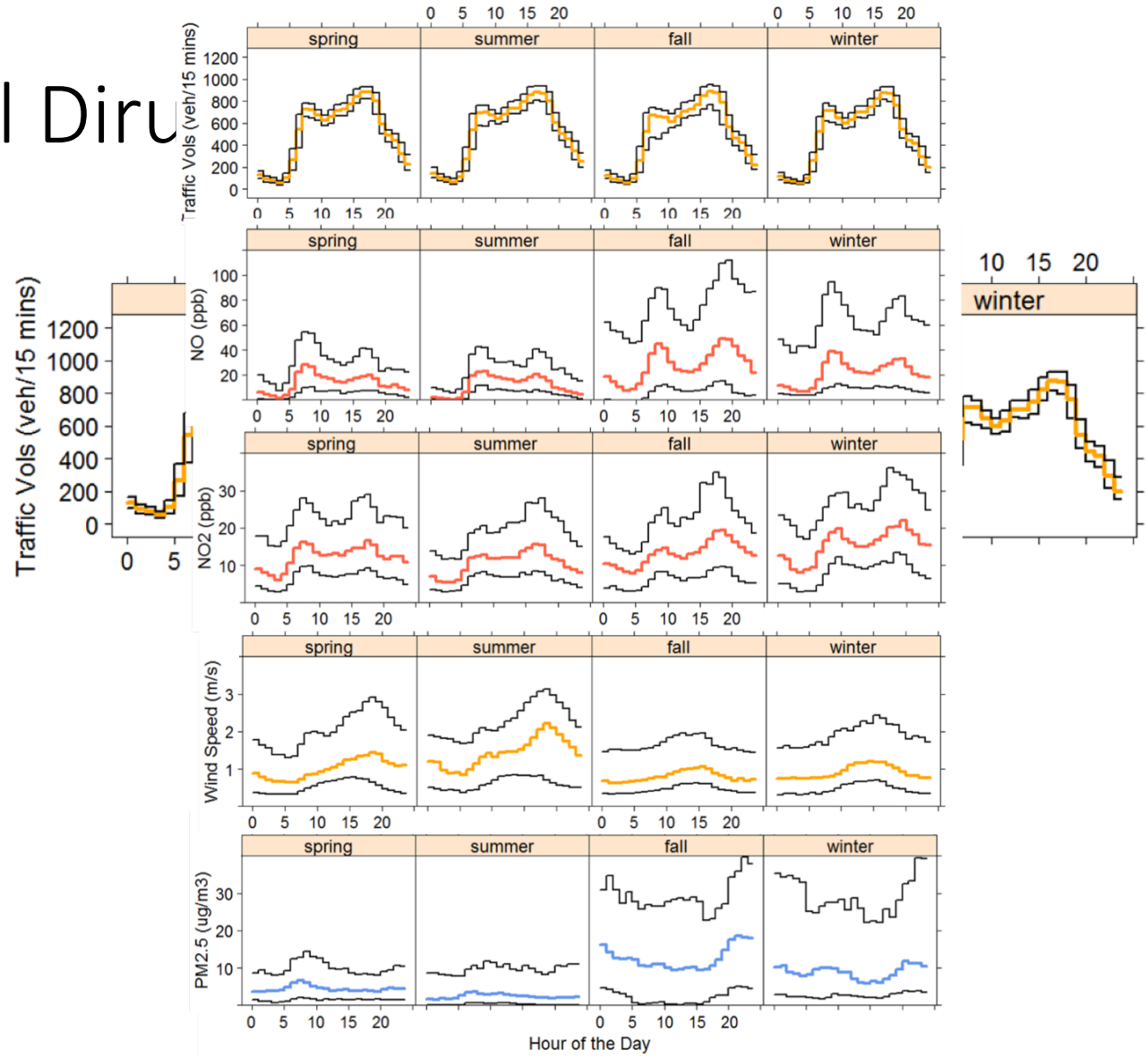
Understanding the Roadside Environment through Integrating Air Quality & Traffic-Related Data

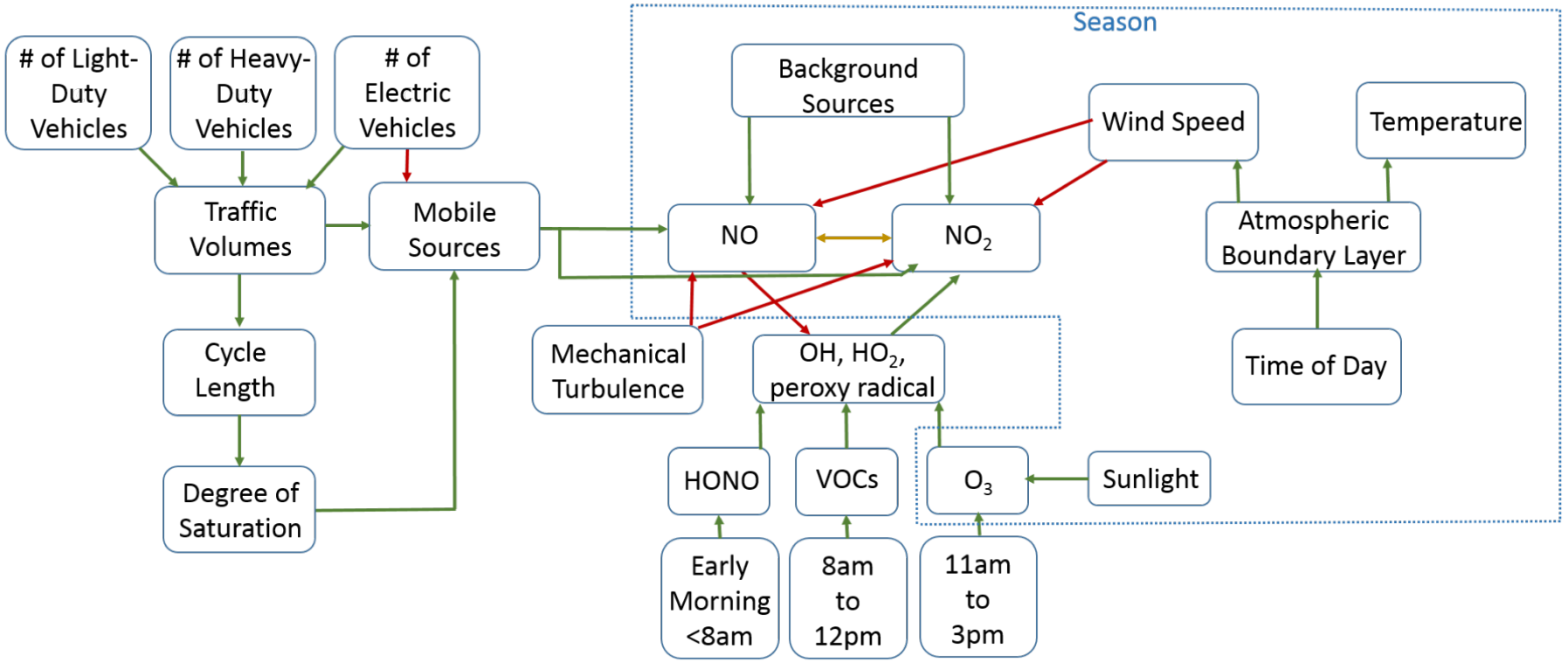
Christine Kendrick

Findings

Season	Time Period	Coefficient NO per 100 vehicles per 15 mins	Standard Error of Coefficient NO per 100 vehicles per 15 mins	Adjusted r^2	Coefficient NO_2 per 100 vehicles per 15 mins	Standard Error of Coefficient NO_2 per 100 vehicle per 15 mins	Adjusted r^2
Fall	Morning	6.3 ** (7.9)**	1.4 (0.4)	0.1 (0.16)	1.2** (1.6)**	2.3 (0.1)	0.14 (0.24)
Winter	Morning	9.4** (11.2)**	1.4 (0.7)	0.14 (0.24)	1.5** (1.9)**	0.4 (0.2)	0.17 (0.26)
Spring	Morning	6.3** (6.7)**	0.8 (0.3)	0.41 (0.43)	2.5** (2.3)**	0.3 (0.2)	0.27 (0.28)
Summer	Morning	4.6** (4.4)**	(0.4) (0.2)	(0.45) (0.37)	1.3** (1.6)**	0.3 (0.1)	0.25 (0.23)
Fall	Evening	-1.3 (-1.6)	2.5 (1.1)	<0.001 (0.005)	1.2* (0.9)	0.6 (0.2)	0.05 (0.03)
Winter	Evening	-0.04 (0.09)	2.6 (1.3)	0.002 (0.001)	-0.4 (-0.08)	0.8 (0.4)	0.05 (0.03)
Spring	Evening	2.1 (2.2)**	1.1 (0.5)	0.03 (0.04)	0.9 (0.9)**	(0.8) (0.3)	0.007 (0.02)
Summer	Evening	2.1* (2.9)**	0.8 (0.3)	0.02 (0.07)	1.6* (1.9)**	0.7 (0.2)	0.03 (0.05)

Seasonal Diru




Season
# of Light-Duty Vehicles
# of Heavy-Duty Vehicles
# of Electric Vehicles
Traffic Volumes
Mobile Sources
Cycle Length
Degree of Saturation
Background Sources
Wind Speed
Temperature
NO
NO2
Atmospheric Boundary Layer
Mechanical Turbulence
OH, HO2, peroxy radical
Time of Day
HONO
VOCs
O3
Sunlight
Early Morning <8am
8am to 12pm
11am to 3pm

Thank you.

https://indicator-frameworks.github.io

or **joshua.tan@cs.ox.ac.uk**

Operational indicator frameworks

- An indicator is a column of numeric data values. They are typically to used to measure inputs, immediate outcomes, and long-term impacts of city projects. We assume that all indicators are time-varying.

id,id_wasp,id_secret,frame_type,frame_number,sensor,**value**,timestamp,raw,parser_type
44637,city1,408414489,128,132,noise,**50**,"2016-08-10 06:00:29",noraw,0
44679,city1,408414489,128,138,noise,**52**,"2016-08-10 06:02:24",noraw,0
44742,city1,408414489,128,143,noise,**51**,"2016-08-10 06:04:00",noraw,0
44777,city1,408414489,128,149,noise,**55**,"2016-08-10 06:05:55",noraw,0
44819,city1,408414489,128,152,noise,**60**,"2016-08-10 06:06:53",noraw,0
44875,city1,408414489,128,160,noise,**62**,"2016-08-10 06:09:27",noraw,0

- An operational indicator framework is just a list of indicators, sometimes organized hierarchically.

Abstract indicator frameworks, v1

- An abstract indicator framework is composed of:
  1. A **R**-valued matrix whose columns represent indicators and rows represent **data**
  2. An inner product operation between indicators, understood as their sample **correlation**
- The set of all abstract indicator frameworks forms something called a **category**

Mathematical background: category theory

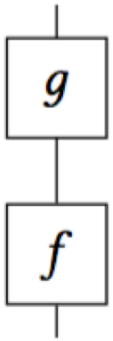
- Category theory was originally invented in the 1960s to integrate different aspects of mathematics, especially topology and algebra. It is now being tested by a variety of different agencies, like NIST and DARPA, as a language for modeling and integrating large, heterogeneous **systems**, e.g. Nextgen or CASCADE.
- A *category* **C** is a collection of objects, called the objects of C, along with a collection of maps, called the morphisms of C, satisfying certain properties.
- A *functor* F : **C** -> **D** between two categories is a map taking objects of C to objects of D, and morphisms of C to morphisms of D in a compatible way.

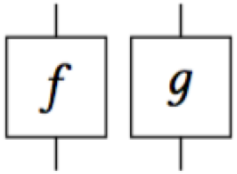
Definition 3.4. The category of $\mathbb{R}$-valued data tables, Data, is defined by the following data:

(1) objects $\mathcal{X} = (\mathcal{X}, \Omega_{\mathcal{X}}, \mathbb{I}_{\mathcal{X}})$ of Data are $m \times n$ tables of $\mathbb{R}$-valued data vectors whose rows are assigned an index key given by $\mathbb{I}_{\mathcal{X}} : \Omega_{\mathcal{X}} \to \mathbb{R}$ and whose columns, $B_{\mathcal{X}} = \{X_1, ..., X_n\}$, represent indicators
(2) morphisms $f : \mathcal{X} \to \mathcal{Y}$ are linear transformations of the column values of $\mathcal{X}$ by vector addition (of other columns in $\mathcal{X}$) and scalar multiplication
(3) the composition is just the matrix product
(4) the tensor product of $\mathcal{X} \otimes \mathcal{Y}$ is the integrated table of their data values over a table of linkages, $S \subset \Omega_{\mathcal{X}} \times \Omega_{\mathcal{Y}}$

Mathematical background: monoidal categories

- A category with a tensor operation $\otimes$: **C** × **C** -> **C**, satisfying certain properties, is called a *monoidal category*.
- Essentially the tensor allows you to compare morphisms in parallel, while composition allows you to compare morphisms in series.





Abstract indicator frameworks, v2

- We want to abstract from the data management aspect. It's the choice of the indicators that is important, not the individual rows of data underneath.
- Many of operations on indicators are purely statistical, e.g. correlation, so we would like to define them in a general context.
- We especially want to emphasize correlation, because it emphasizes relations *between* indicators rather than the indicators themselves.
- We want to set the stage for linking **models** to **data**.
- This motivates the following definition:

Definition 3.2. The category of random variables, Rand, is defined by the following data:

(1) objects are finite-dimensional Hilbert spaces

$$\mathcal{X} = L^2(\Omega_{\mathcal{X}}, \Sigma_{\mathcal{X}}, \mathbb{P}_{\mathcal{X}})$$

of square-integrable random variables (under the equivalence relation $X_1 \sim X_2$ if $\mathbb{P}_{\mathcal{X}}(X_1 = X_2) = 1$) with inner product $\langle X, Y \rangle = E(XY)$, defined over probability spaces $(\Omega_{\mathcal{X}}, \Sigma_{\mathcal{X}}, \mathbb{P}_{\mathcal{X}})$, with an associated basis $\mathcal{B}_{\mathcal{X}} = \{X_1, X_2, \ldots, X_n\} \cup \mathbf{1}$, where $\mathbf{1}$ is the random variable with constant value 1.

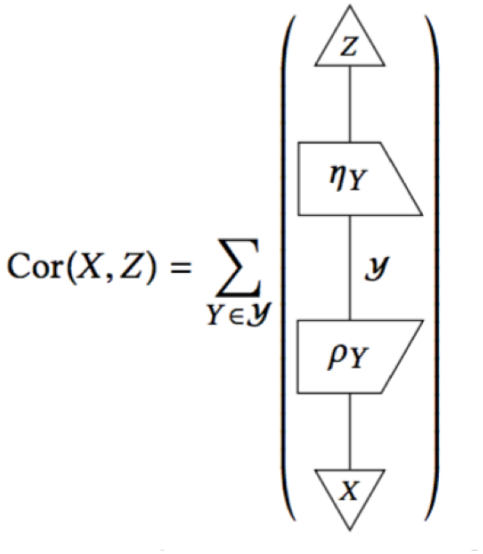
(2) morphisms $F : \mathcal{X} \to \mathcal{Y}$ are bounded linear operators

(3) the composition is the usual composition of bounded linear operators

(4) the tensor product of $\mathcal{X}$ and $\mathcal{Y}$ is the pushout over their joint support in $\Omega_{\mathcal{X}} \times \Omega_{\mathcal{Y}}$

Defining new indicator frameworks in **Rand**

- Given two indicator frameworks **X** and **Z** in **Rand**, one can write "formulas" in **Rand** that describe new indicator frameworks, exactly analogous to how one defines mediating (or confounding) variables in statistics. In the diagrammatic calculus, these look something like this:



Abstract indicator frameworks, v3

- The causal diagram serves as a primitive **model** of a given context.
- The category **Rand** of random variables serves as a (still primitive) **semantics of how we use data**.
- The idea: define constraints in **Rand**, and thus indicator frameworks, by mapping the causal theory into **Rand**; this creates a **model** of the causal theory in **Rand**., i.e. an indicator framework based on the causal theory.

Definition 3.7. The category Ind of abstract indicator frameworks is defined by the following data:

(1) an object I of Ind is a strong symmetric monoidal functor $C \to$ Rand from a causal theory C to the category of random variables.
(2) a morphism η between abstract indicator frameworks is a natural transformation of strong symmetric monoidal functors

Future Work

- More examples!
- Constructions besides mediating frameworks
- Improved semantics on **Rand**
- True “hybrid indicator frameworks” for CPS models

Example: Shot Spot in South Bend

- Courtesy of Santiago Garces, CIO of South Bend, Indiana
- Target indicator: reduce crime
- Target indicator: reduce gun crime and group-related activity
- Target indicator: target interventions at specific group members
- Means: Incorporate Shot Spot information with 911 dispatch calls. “Whenever, a shot incident is detected and a resident also calls 911 to inform of the incident, the dispatch is classified differently than when a signal is detected but not accompanied with a resident's call.”

Example: Shot Spot in South Bend

Indicators Used

1. **The total number of shootings involving a group member, compared to a 3 year rolling average;**
2. the ratio of group member involved shootings, compared to the total number of criminal assault shootings;
3. **Number of shooting incidents recorded both by Shot Spotter and residents**
4. **Number of direct interventions with group members, or close social relatives**
5. Number of call-ins (large meetings where notorious group members are presented with the opportunity to get involved with social services, or communicated the enforcement action alternative)
6. Number of enforcement actions, interventions directed at executing warrants and investigations against the most violent group
7. Ratio of shots where a resident called as a shot is detected by Shot Spotter (proxy for community trust and collaboration with the Police Department)
8. Percentage/ s-value of number of complaints relative to calls for service

Example: Shot Spot in South Bend

- Things we need to model:
  - Frameworks that draw on indicators from a number of different sources: ISO 37120 for overview, FBI uniform crime reporting program, 911 CAD system, Shot Spotter, internal databases for group activity and “interventions”
  - Comparisons of indicators as indicators, e.g. the total # of shootings compared to a 3 year rolling average
  - Logical operations on indicators, e.g. # of shootings recorded by Shot Spotter AND local residents, # of direct interventions with group members OR close relatives
  - Relationship between indicator and `sub-indicators’
  - Properties of indicators, like how difficult it is to supply that indicator

Non-temporal indicators

- E.g., location-varying indicators like energy use per building:

Property Name	Reported	Property Type	Address	ZIP	Gross Area (sq ft)	Site EUI (kBTU/sf)
MEEI -Longwood	Yes	Ambulatory Surgical Center	800 Huntington Ave	02115	76,300	173.1
Prime Motor Group	Yes	Automobile Dealership	1525-1607 VFW Parkway	02132	150,000	28.7
New England Center for Homeless Veterans	Yes	Barracks	17 Court St.	02108	130,000	49.8