On the intricacies of generating individuals' cellular network datasets

STEM workshop day

20 Dec 2022

Anne Josiane Kouam
4th Year PhD

Supervisors: Aline Carneiro, Alain Tchana

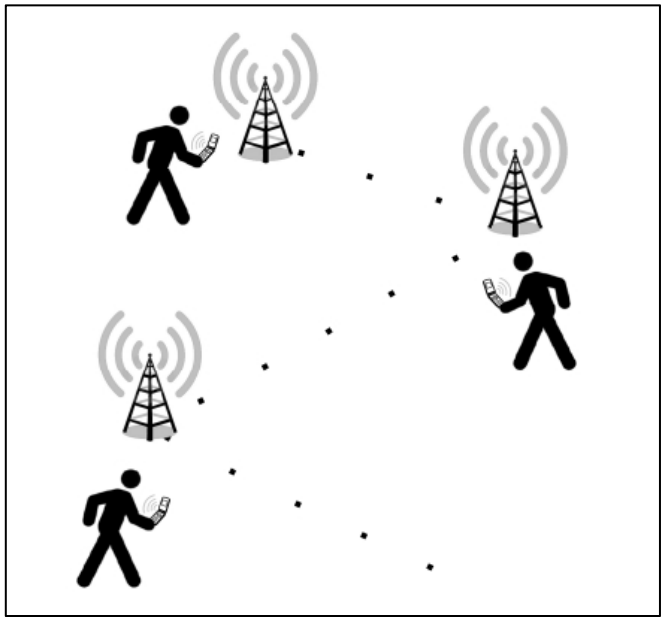
1- Cellular network datasets

1- Cellular network datasets

Description

Network events metadata collected per operator

- Call Detail Records (CDRs)
- eXtended Data Records (XDRs)
- Charding Data Records (CDRs)



Billing purposes but not only...

CDRs description

Time, event, caller ID, cell_ID, called ID, call duration, data volume

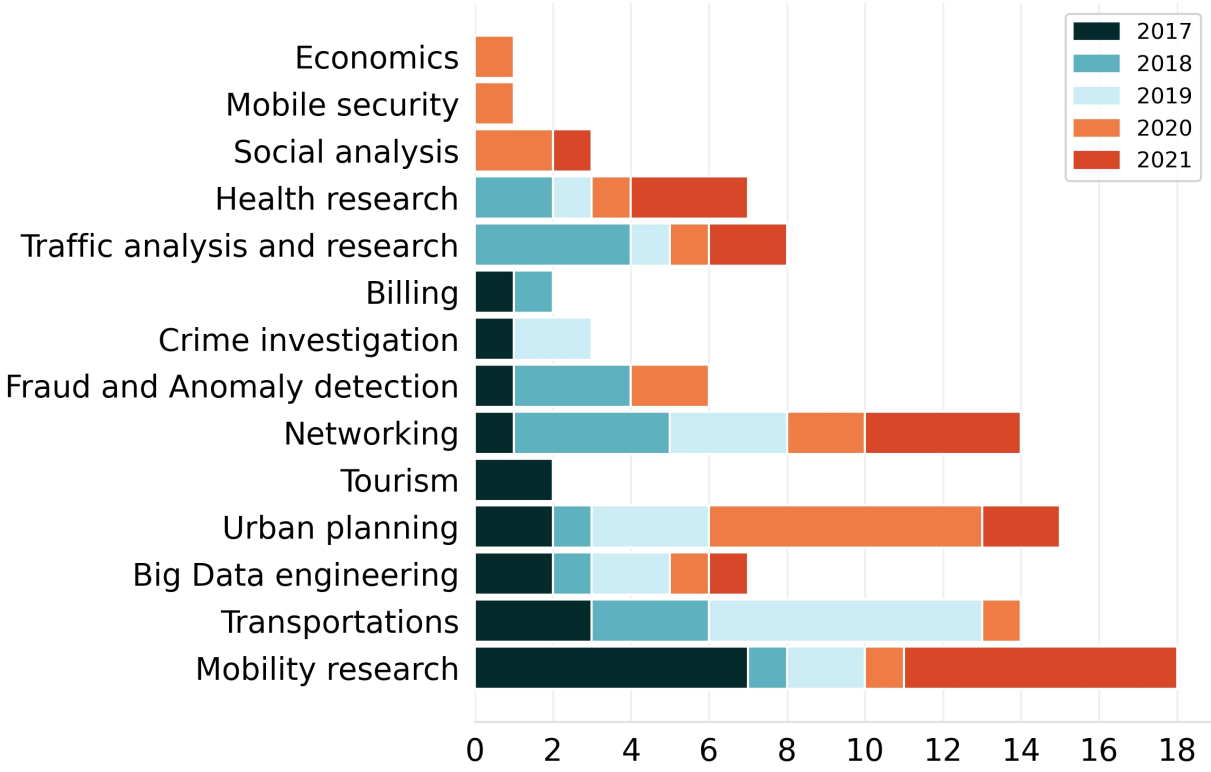
12:00:05, CALL, 0987980517, 15, 0999798777, 300, 0

12:15:12, DATA, 0987980515, 15, / , 0, 20

12:30:00, SMS, 0987980515, 20, 0987980515, 0, 0

1- Cellular network datasets

Utility



- 1022 papers - Google Scholar - **last 5** years
- **14 domains** from a sample of 100

1- Cellular network datasets

Exploitation limitations

Accessibility

need for strict agreements with mobile operators

Usability

data aggregation limits related analyses' preciseness

Privacy

even anonymized, CDRs hold users' habits sensitive information

Flexibility

Limitation in terms of population size, CDRs duration, or geographical areas

Realistic CDRs generation as a goal

2- CDRs generation

2- CDRs generation

Problem formalization

Generation

- Synthetizing timestamped datasets from X_0 to X_T only from a given context c
- Model the joint conditional probability $P(X_1, X_2, \ldots, X_T | c)$
- Equivalent to $P(X_1|c)P(X_2|X_1, c)\ldots P(X_T|X_1, X_2, \ldots, X_{T-1}, c)$

harder than

>

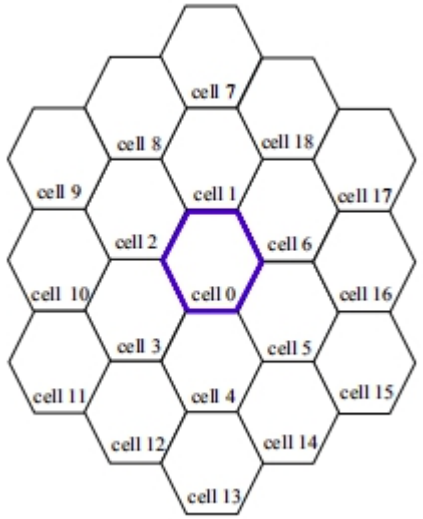
Prediction

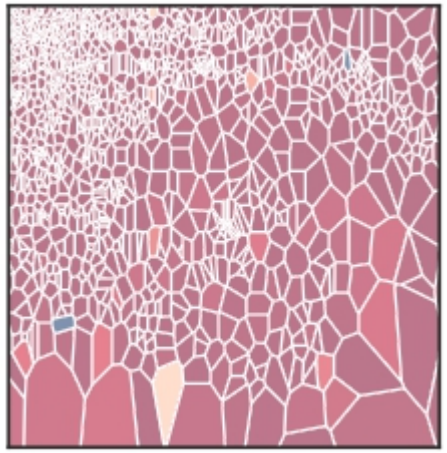
- Estimating the next X_{T+1} based on the records from 0 to T
- Model the conditional probability $P(X_t | X_{t-1}, c)$

2- CDRs generation

Additional requirements

1. Controllability and modeling arbitrary network topologies



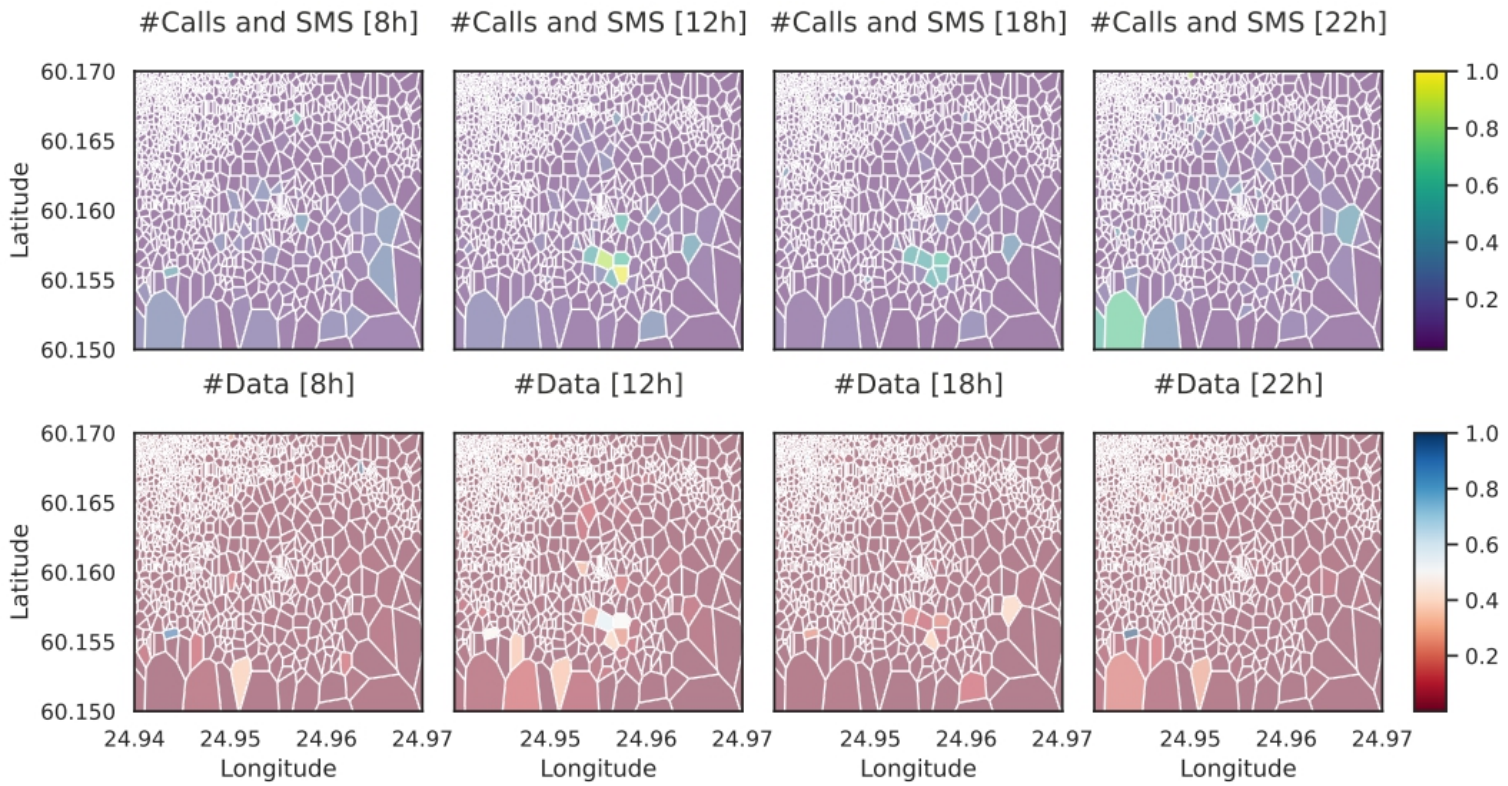


- Used for a variety of use cases
- Possibility to modify output CDRs according to parameters, e.g., duration, population size, mobility area
- Mobility area => urbanization level, layout, infrastructure
- Strong **dependency** between **CDRs** and **network topology**

2- CDRs generation

Additional requirements

2. Modeling spatiotemporal correlations



- **Human activity in space fluctuations** as a function of the **time of the day**
- Aggregated counts of events at 8h, 12h, 18h, and 22h in Helsinki, EU
  - Working zones
  - Residential zones
  - Leisure zones

2- CDRs generation

Additional requirements

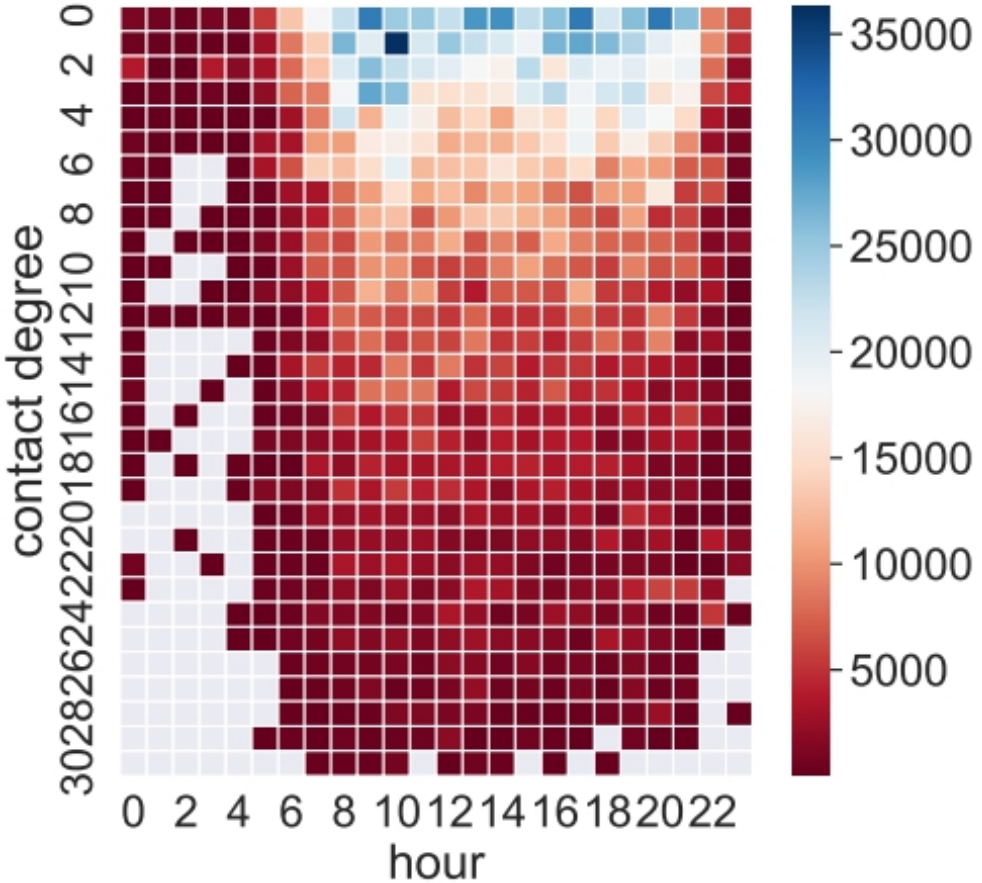
3. Modeling social interactions

- Calls and SMS => human interactions
- Are to be captured
  - How many contacts
  - Who are these contacts
  - How are the interactions with those

2- CDRs generation

Additional requirements

4. Modeling inter-features correlation

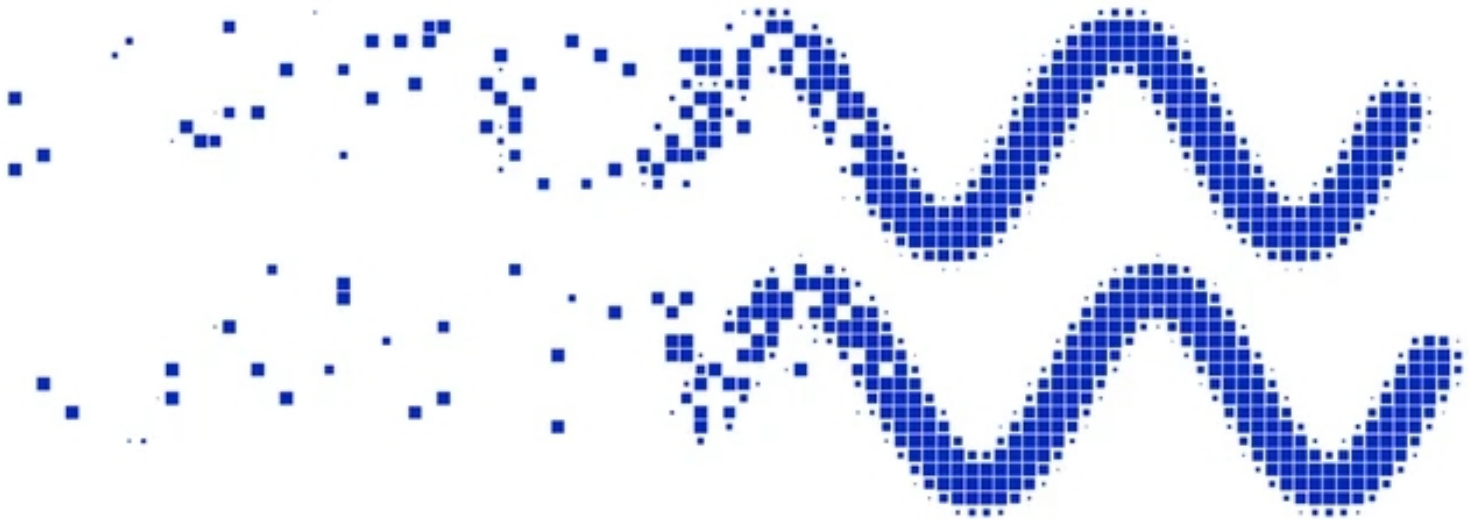


- CDRs features are part of a whole
  - Traffic (what, when, how)
  - Mobility (where)
  - Social (whom)
- Inherent correlation
- Total week call duration (how) varying with the contact degree (whom) and the time (when)

2- CDRs generation

Additional requirements

5. Modeling temporal dynamics

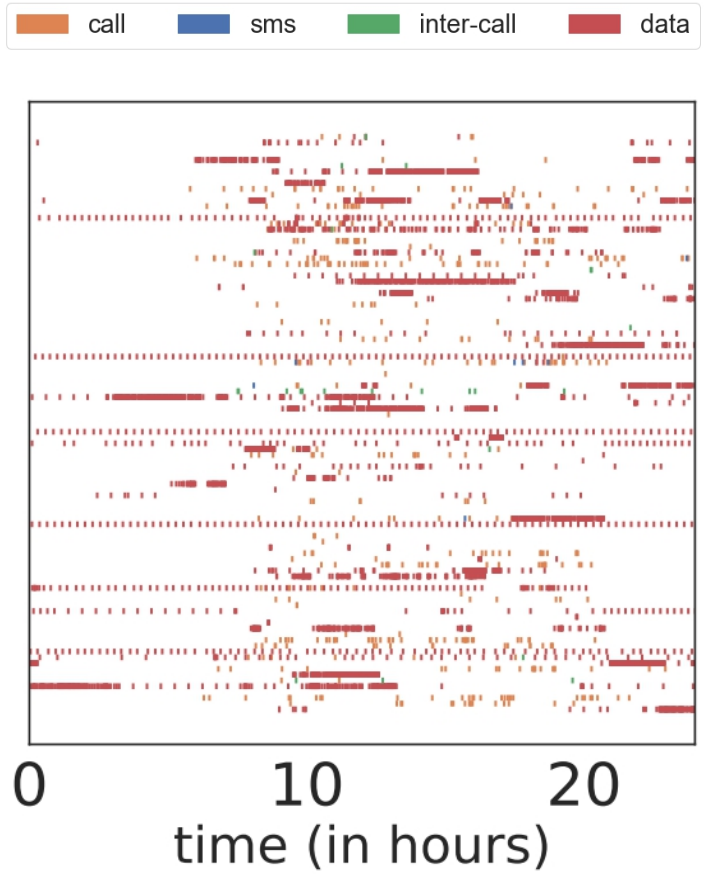


- Long-term dynamics correlated to regular human activities
  - peak and off-peak daily hours
  - week days and week ends
  - vacations and working periods
- Circardian rythm

2- CDRs generation

Additional requirements

6. Modeling individuality and heterogeneity



- Capturing heterogeneity is a goal
- 100 randomly selected users in real traces
  - type of events made
  - inter-event times
- Dependency between each component (i.e., user)

3- Approaches

3- Approaches

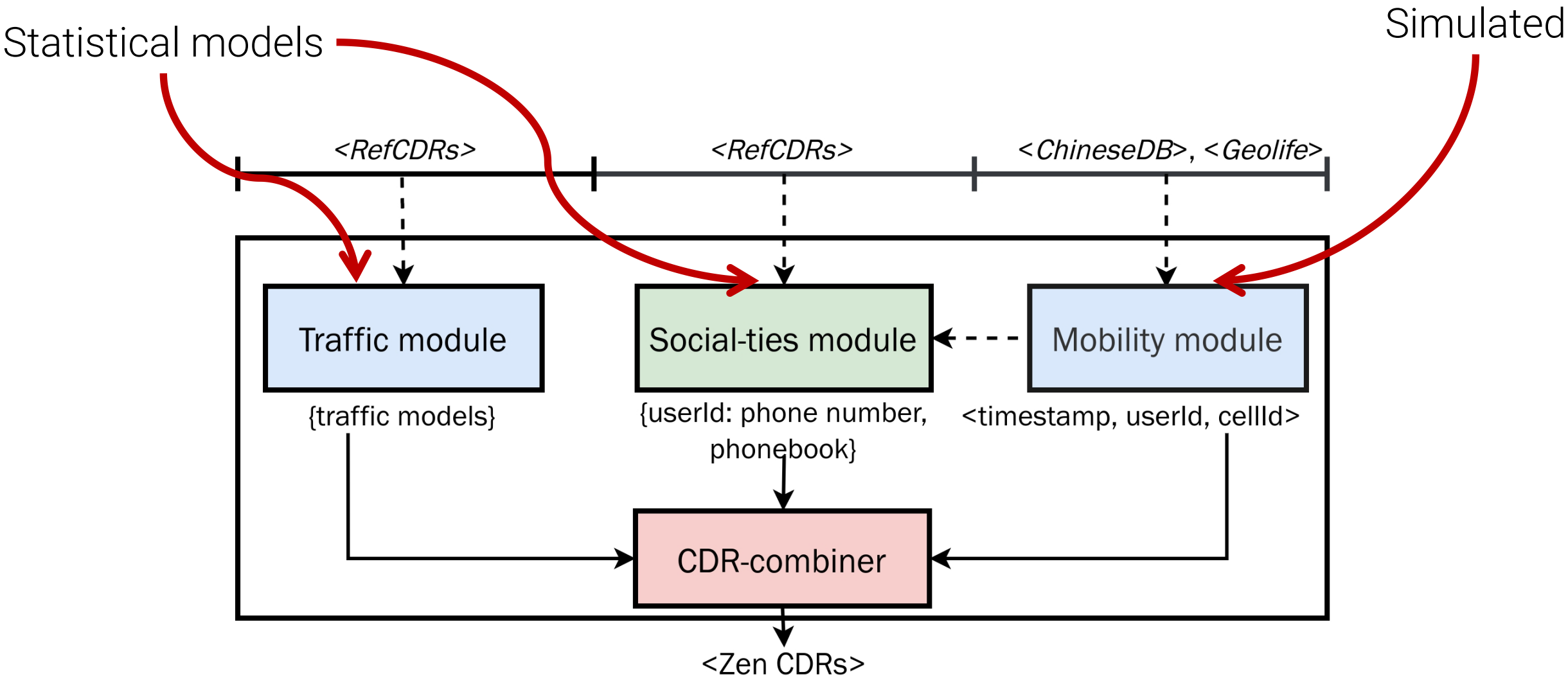
Overview

Controllability and modeling arbitrary network topologies	
Modeling spatiotemporal correlations	
Modeling social interactions	
Modeling inter-features correlation	
Modeling temporal dynamics	
Modeling individuality and heterogeneity	

- A1- Statistical + Simulation
- A2- LSTM-based + Simulation
- A3- GAN-inspired

3- Approach 1

Statistical + Simulation

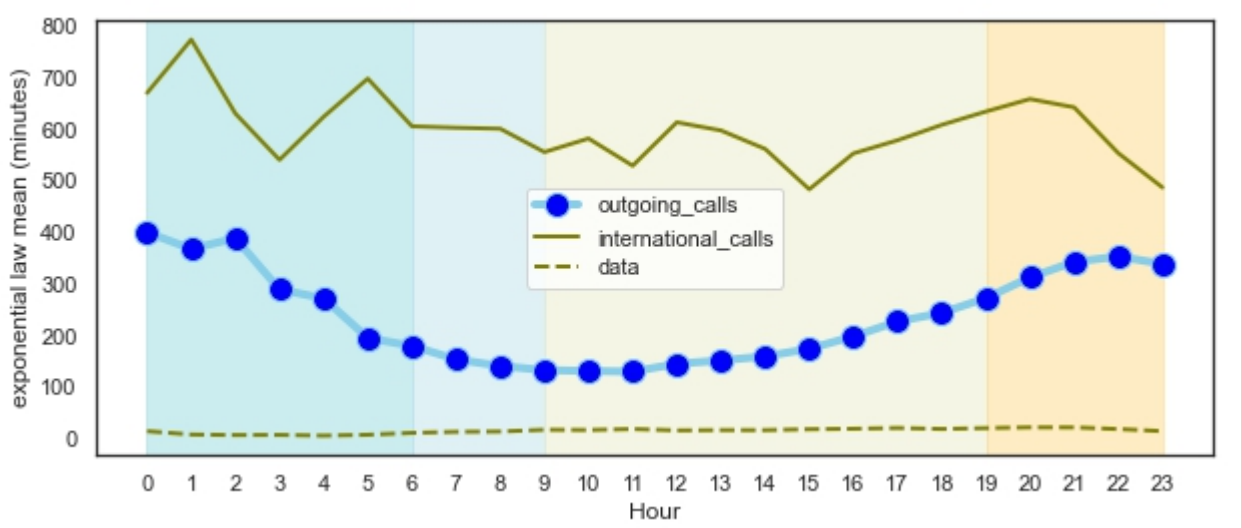


Traffic modeling
Event (type & time) generation

- Three event-types
  - local outgoing calls
  - international outgoing calls
  - data

- Model 1
  - Non parametric regression of IET exponential distribution

user_id timestamp event_type *corresponding_user duration/size*



Traffic modeling
Event (type & time) generation

- Three event-types
  - local outgoing calls
  - international outgoing calls
  - data

- Model 1
  - Non parametric regression of IAT exponential distribution

- Model 2
  - Conditional probability table
  - $P(X_{i+1}$ in $Interval_b / X_i$ in $interval_a)$

user_id timestamp event_type corresponding_user duration/size

$Interval_a$	0-6	6-9	9-19	19-23	None
0-6	0.022	0.457	0.333	0.042	0.145
6-9	0.015	0.33	0.48	0.036	0.135
9-19	0.015	0.282	0.276	0.184	0.242
19-23	0.03	0.41	0.34	0.04	0.18

Traffic modeling
Social interactions

- Contacts are sorted from the closest to the least

user_id timestamp event_type ***corresponding_user*** *duration/size*

	number	call_count	contact1	contact2	contact3	contact4	contact5	contact6	contact7	contact8	...	contact151
0	00013334a2466a58e8a583a98e74bf53ffe29881	3	2	1	0	0	0	0	0	0	...	0
1	00038d28bfb6f2baa9dc5e8e8281fc582594044f	52	6	6	4	3	3	2	2	2	...	0
2	0003c590893b15ec0036c1b309fe3d12d6f6d872	89	29	10	5	3	3	3	3	3	...	0
3	000af2bd14eba092727bb348b69b0c607fe2a540	4	2	2	0	0	0	0	0	0	...	0
4	0015d51dfc07a4cd4e771e05dbae0f0971d2e4ac	5	2	1	1	1	0	0	0	0	...	0
...	...	...	...	...	...	...	...	...	...	...	...	...
13359	ffe9acd9b4f843abfb2a04cca194f598d84896cb	5	4	1	0	0	0	0	0	0	...	0
13360	ffeed1a032cf1456ddf281de6c07b8ff21473f17	7	3	1	1	1	1	0	0	0	...	0
13361	fffdaabc43d4d32bd98f5ed9af2600c5e0b8face	77	10	9	4	4	4	3	3	3	...	0
13362	fffdd2b9ea5c42b83b0079fd6adfeb931b928fd2	536	35	23	22	22	22	17	16	16	...	0
13363	fffe90fefa573c3f93d4e015ffca38ee47603274	72	11	8	8	7	6	4	3	3	...	0

13364 rows × 162 columns

0.473	0.199	0.126	0.090	0.070	0.057	0.048	...	0.002

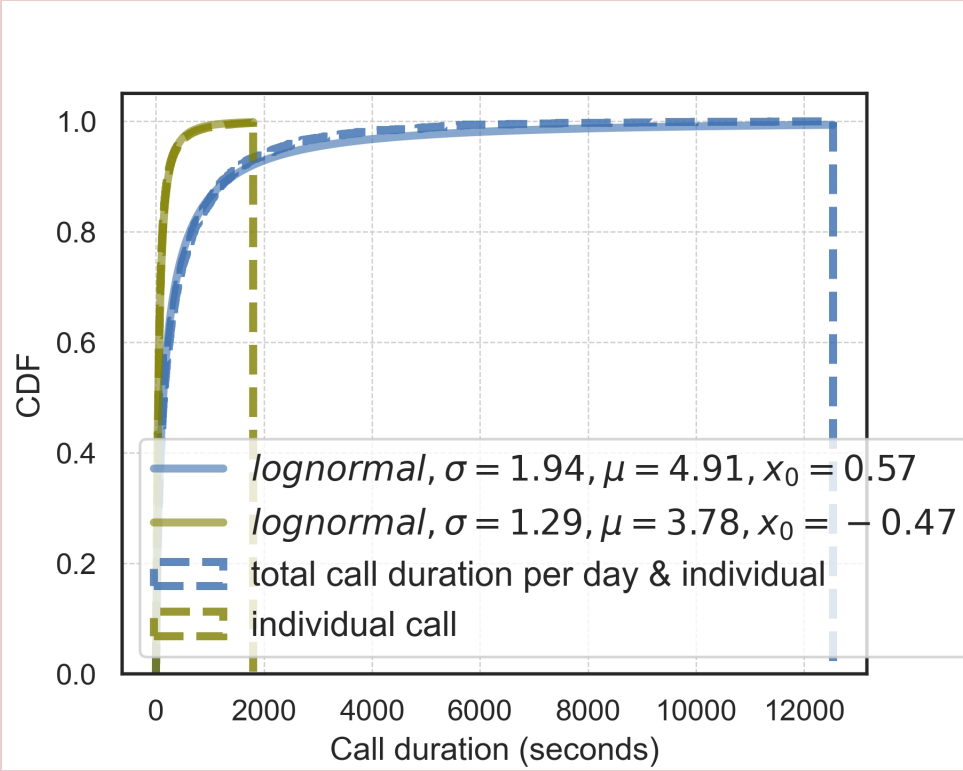
- Alternative modeling approaches
  - chinese table or stick breaking process

Traffic modeling
Metrics

- Statistical test (call duration)

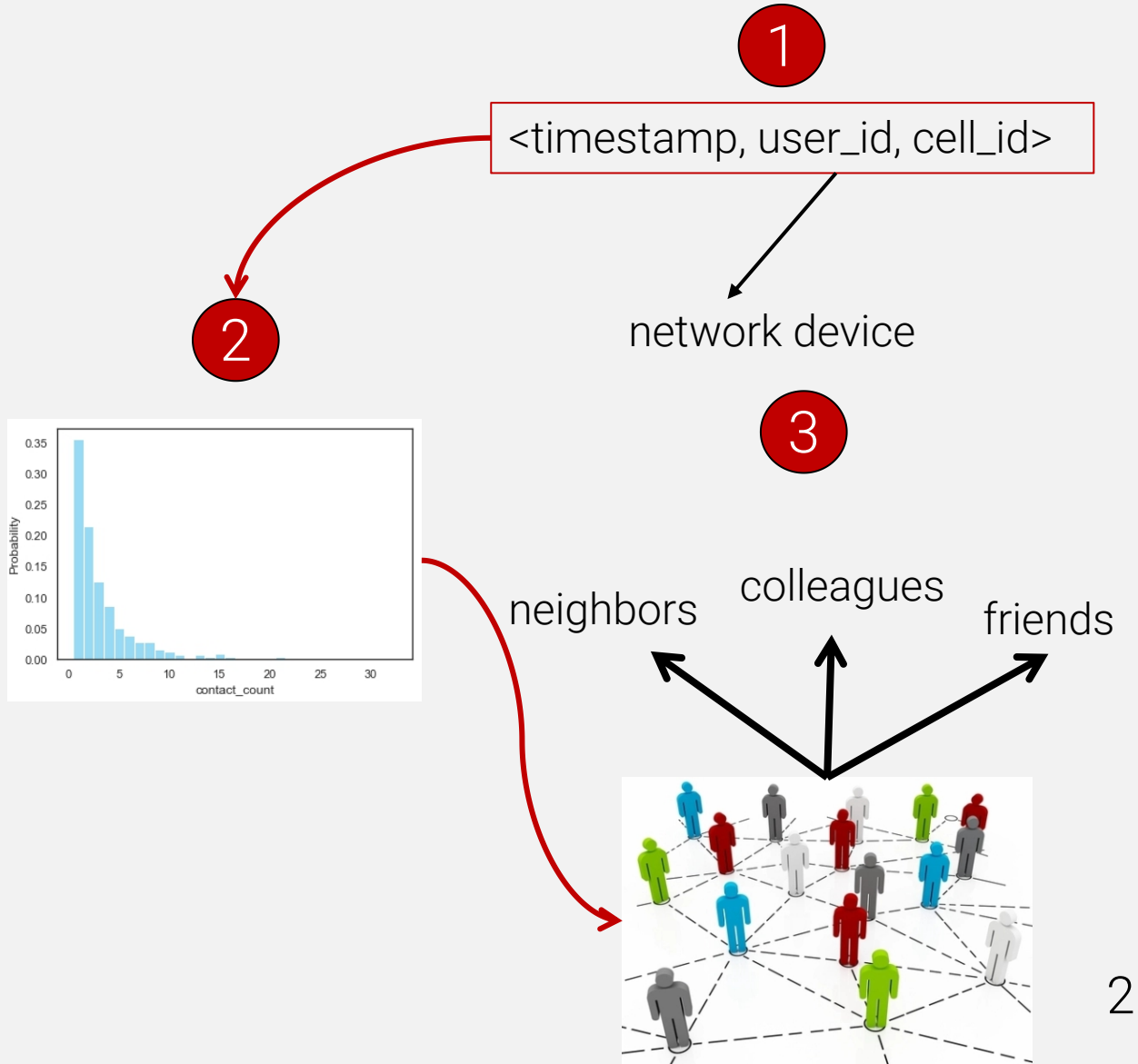
$D \sim LogNormal$

user_id *timestamp event_type corresponding_user* ***duration/size***



Modeling social ties

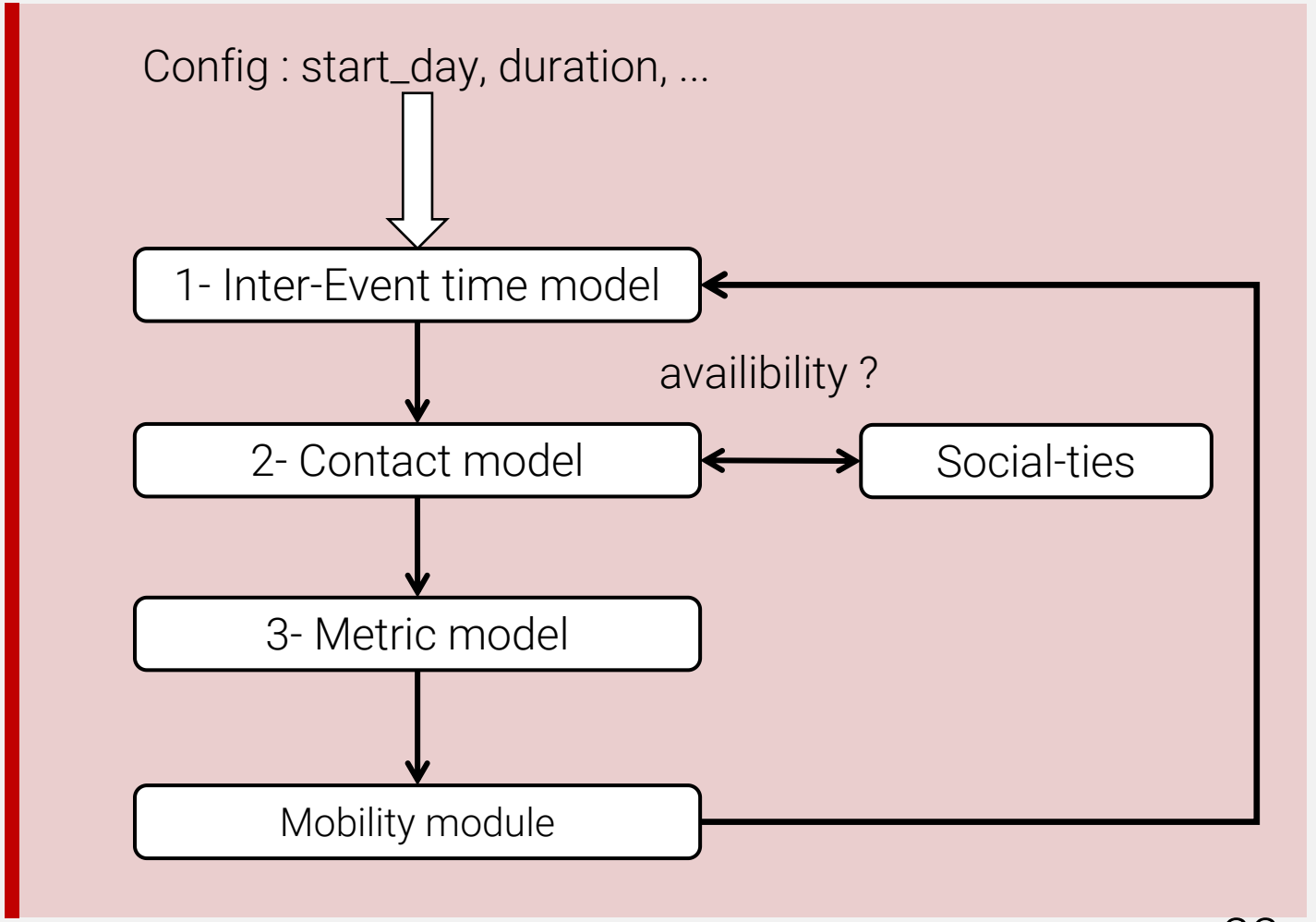
- Mobility users to cellular nodes
- List of correspondents for each node



CDR-combiner

Combination between traffic and mobility

- No strategy,i.e., all users implement the same traffic models



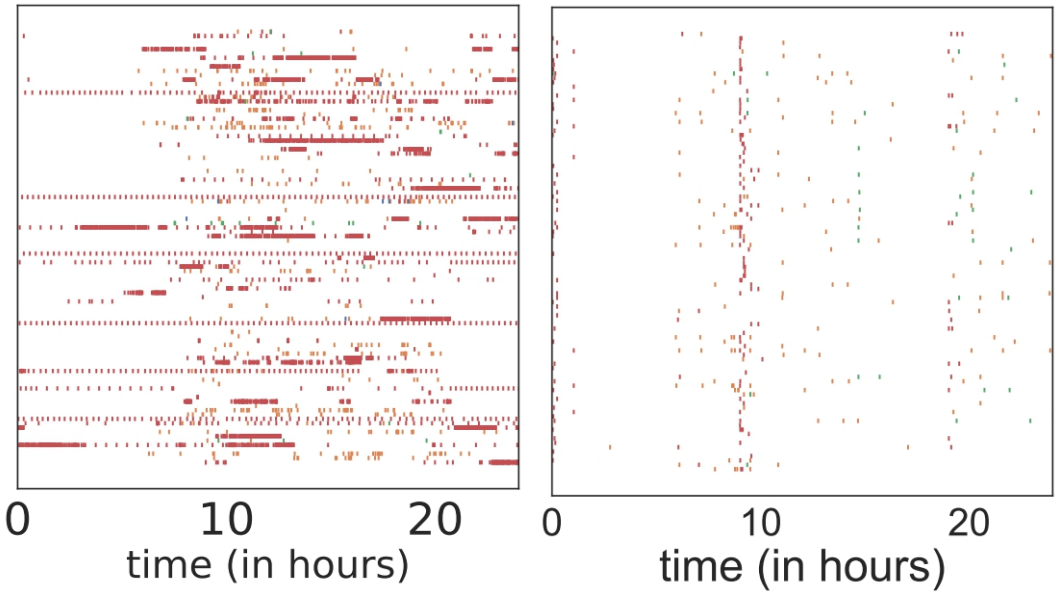
3- Approach 1

Validation

Controllability and modeling arbitrary network topologies	~
3- Modeling spatiotemporal correlations	X
Modeling social interactions	/
1- Modeling inter-features correlation	X
Modeling temporal dynamics	/
2- Modeling individuality and heterogeneity	X

Reproduced distributions related to users' traffic

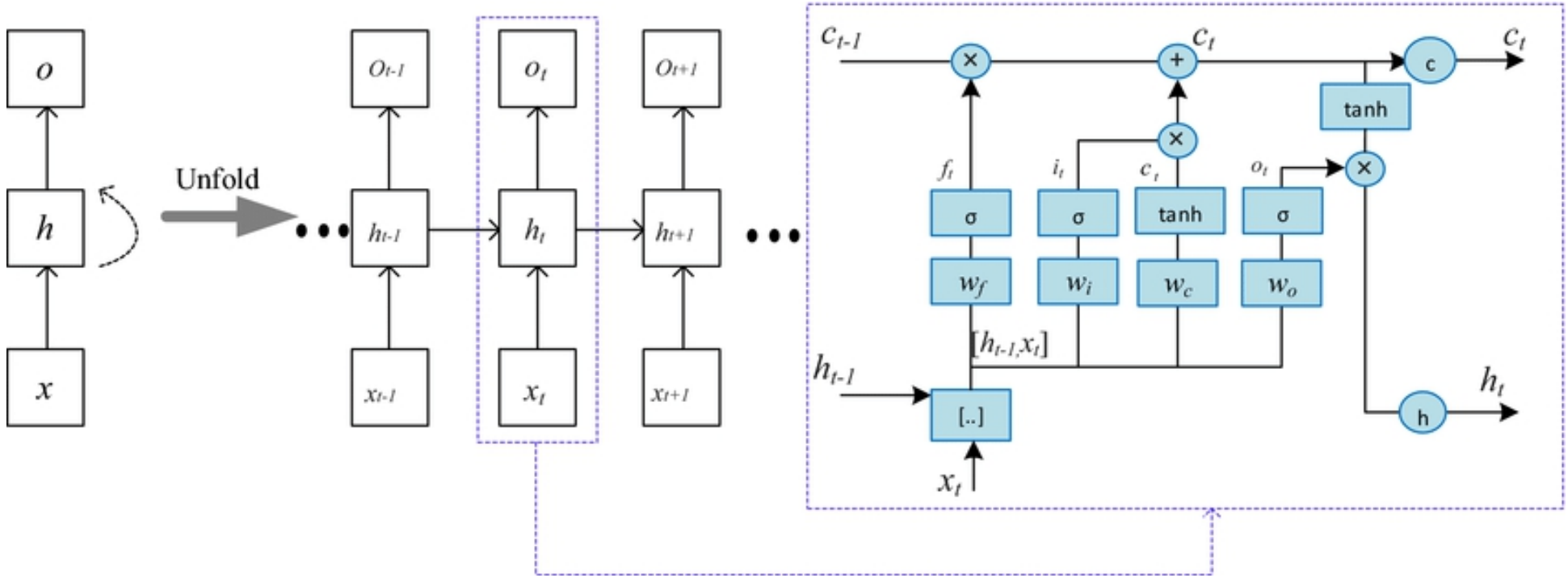
- Number of calls per user
- Call duration



3- Approach 2

LSTM-based

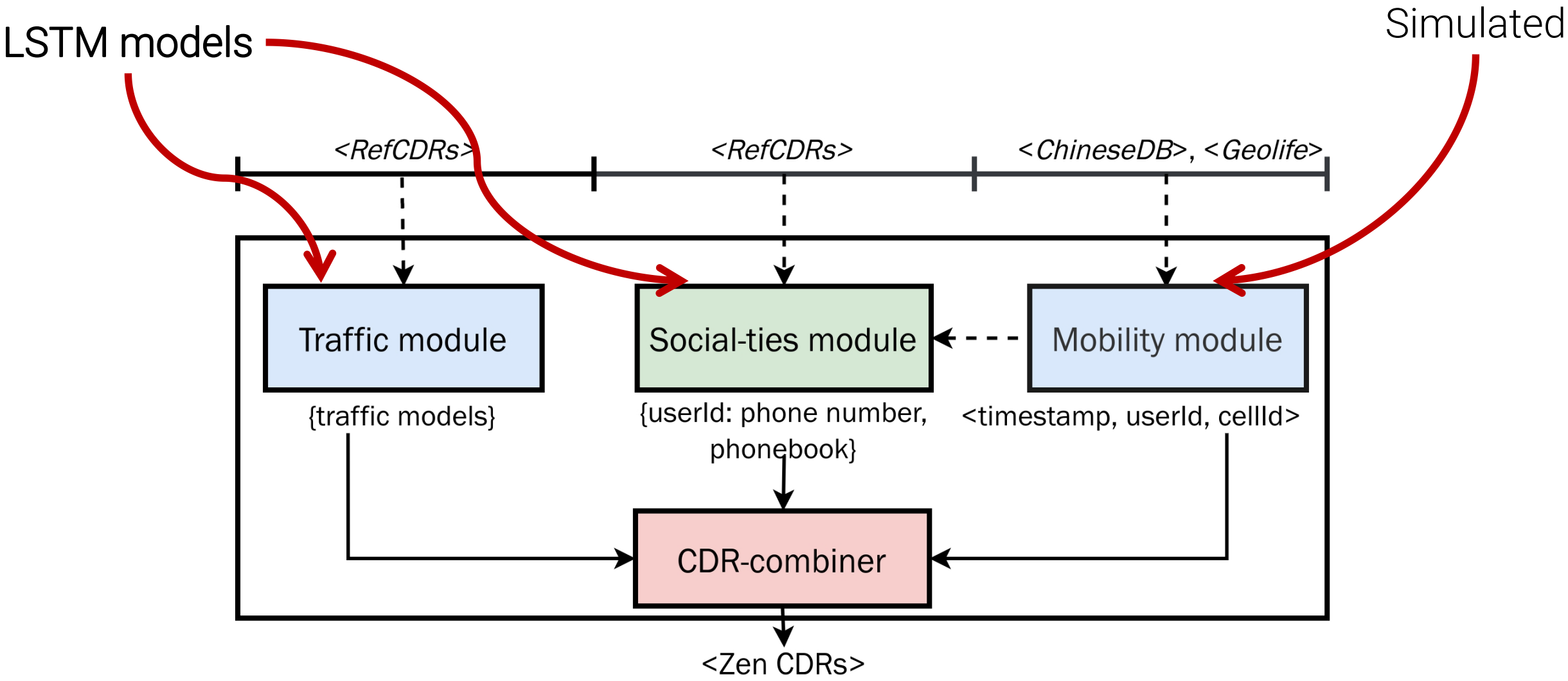
Long-Short Term Memory NN are a form of RNN



- The global distribution is learnt but each sequence is considered => heterogeneity
- Deep learning has enough expressive power to capture inter-features correlation

3- Approach 2

LSTM-based



Zen: traffic module

- Each user corresponds to a sequence of timestamped events

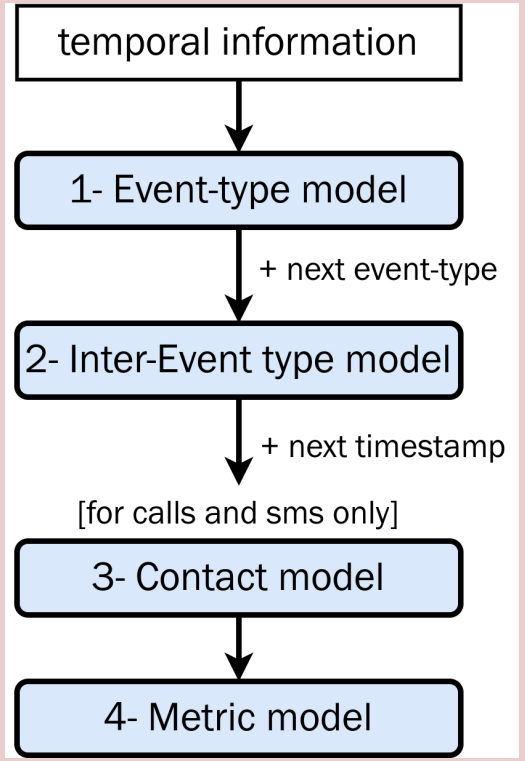
- Modeling traffic individual traffic behavior
  - *what : event-type*
  - *when: time*
  - *whom : social interactions*
  - *how : metric*

- *Long Short Term Memory*-based modeling

- Model training

Traffic CDR Trace Format

timestamp user_id event_type corresponding_user duration/size



Zen: traffic module

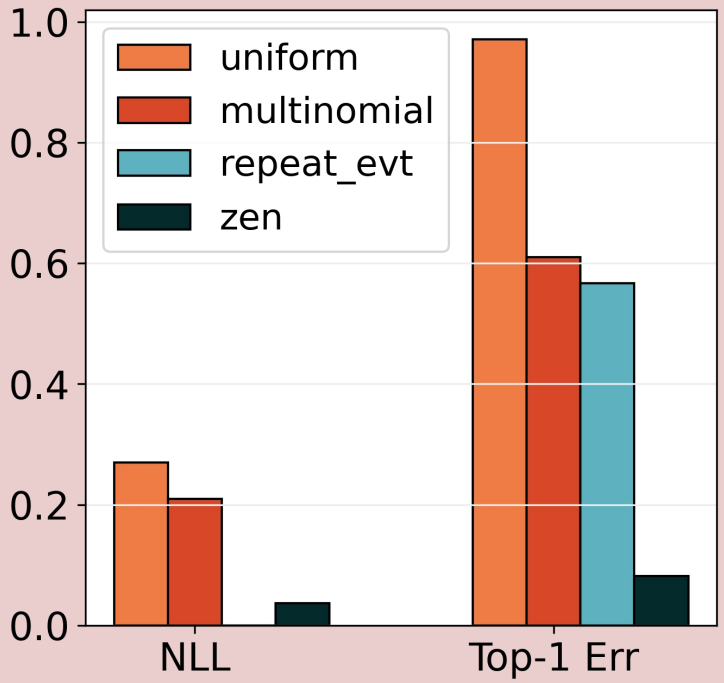
Model 1:
(event_type, timestamp) => next_event_type

Features

- event_type : one-hot encoded
  - SMS, DATA, CALL, INTER_CALL
- timestamp
  - dow : 0-6 one-hot encoded
  - hod : 0-23 one-hot encoded
  - sod : (sin, cos) encoded

Traffic CDR Trace Format

timestamp user_id event_type corresponding_user duration/size



Zen: traffic module

Model 2:
(next_event_type, timestamp) => IET interval

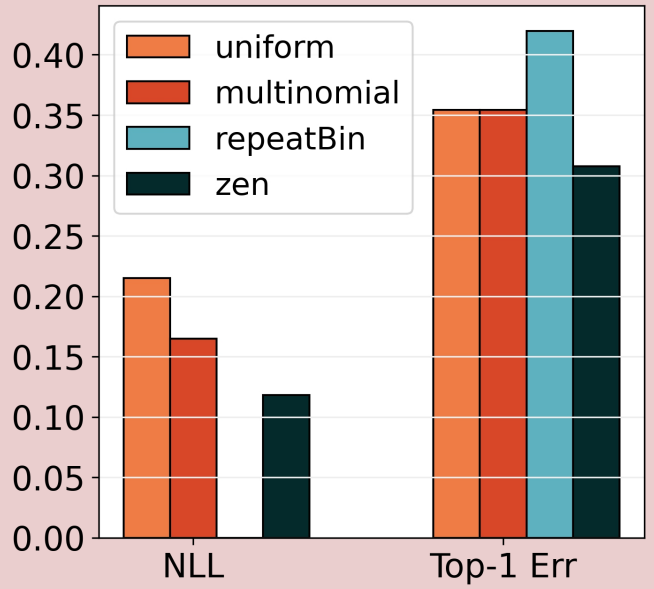
Discretized time bins
[0-30min]]30min-24h], and > 24h

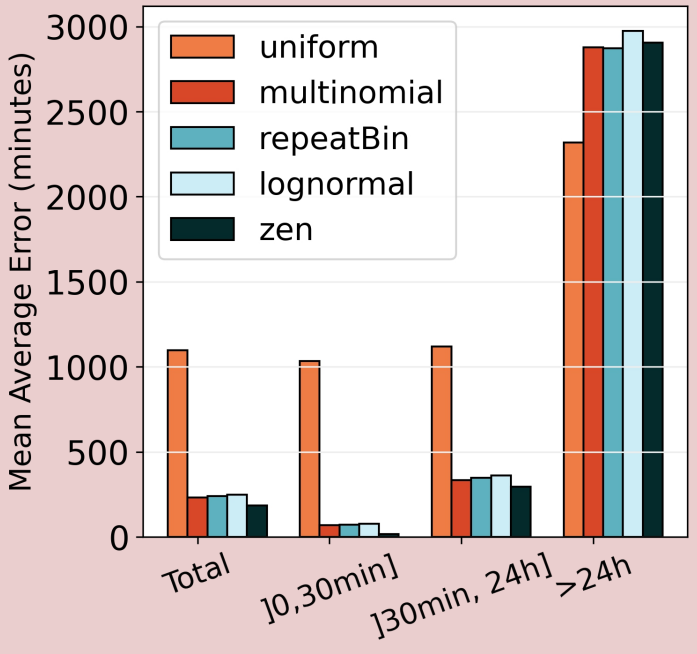
Features
- event_type : one-hot encoded
- timestamp
  - dow : 0-6 one-hot encoded
  - hod : 0-23 one-hot encoded
  - sod : (sin, cos) encoded

Continuous estimation
Distributions of IET in each interval

Traffic CDR Trace Format

timestamp user_id event_type corresponding_user duration/size





Zen: traffic module

Model 3:
(event_type, timestamp) => friendship_degree (fd)

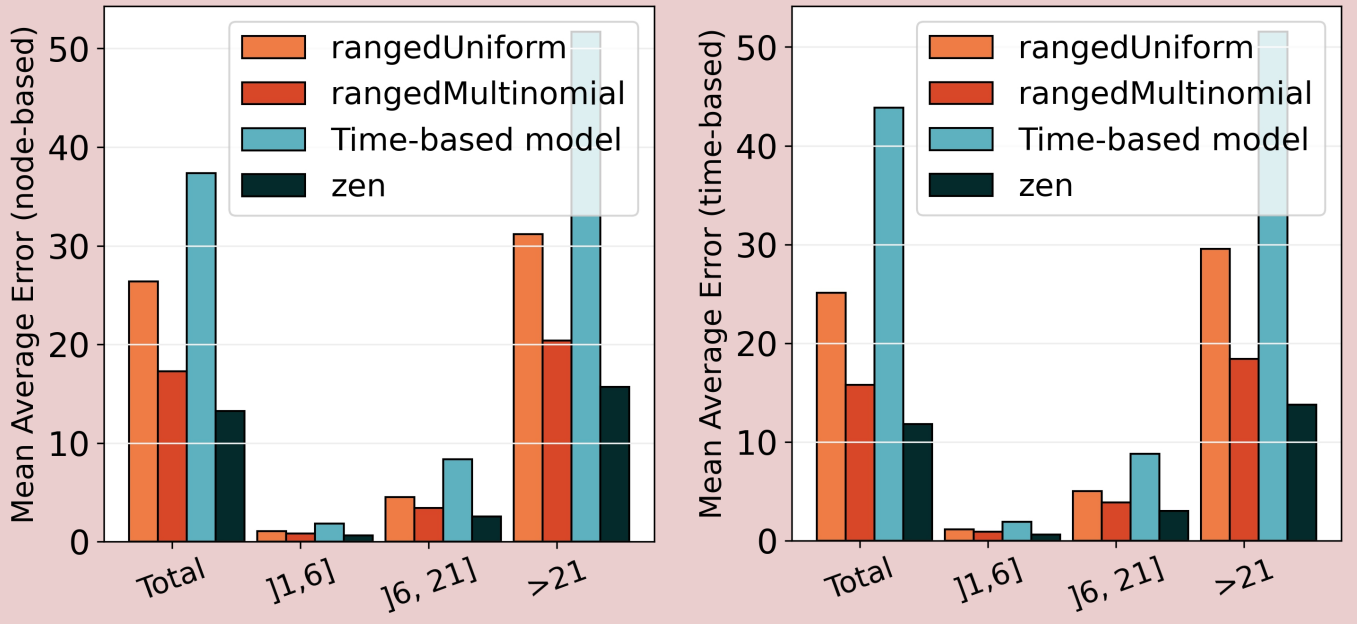
fd: strength of the relationship with each contact

Features

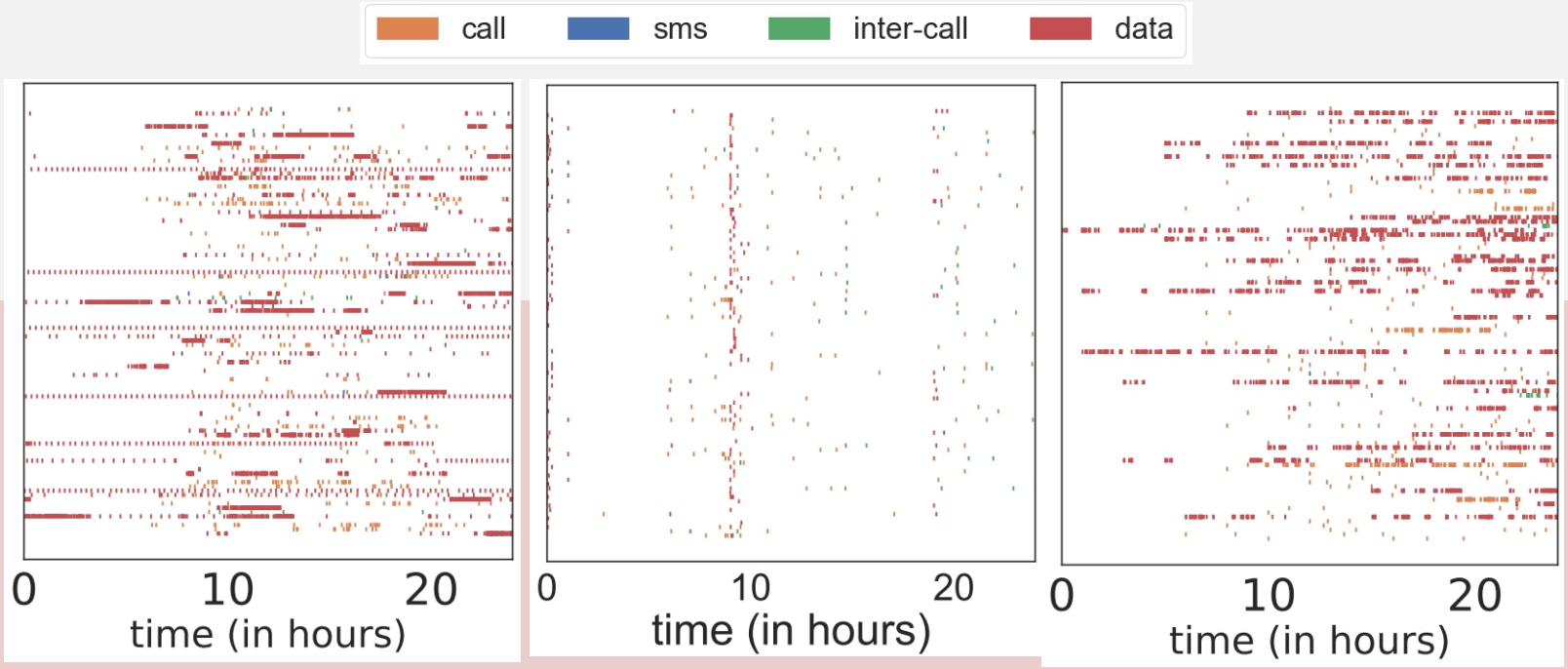
- event_type : one-hot encoded
- timestamp
  - dow : 0-6 one-hot encoded
  - hod : 0-23 one-hot encoded
- #contacts

Traffic CDR Trace Format

timestamp user_id event_type corresponding_user duration/size



Zen: traffic module



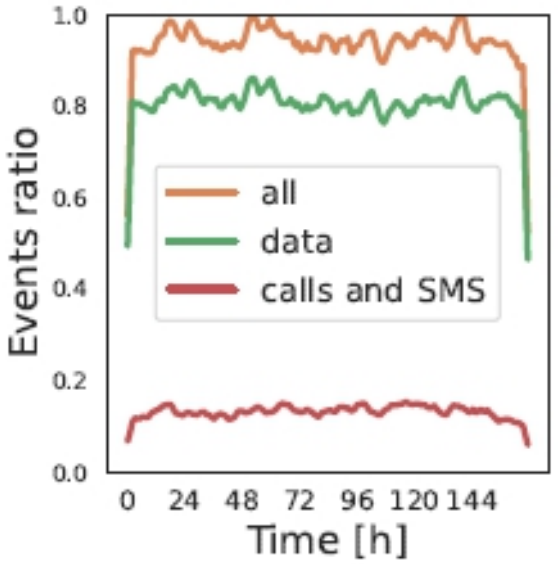
Real traces | Statistically generated | Zen-generated

100 users randomly selected
Users active

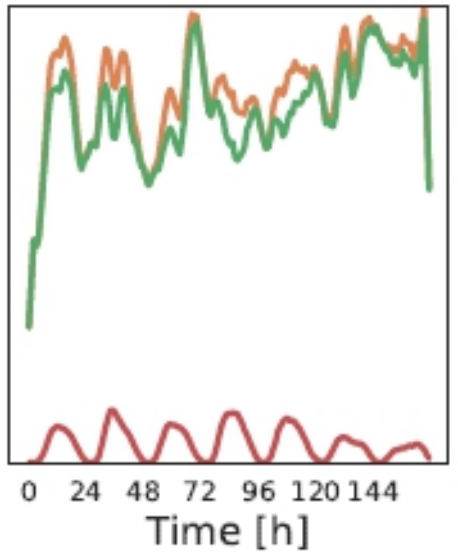
3- Approach 2

Validation

Controllability and modeling arbitrary network topologies	~
3- Modeling spatiotemporal correlations	X
Modeling social interactions	/
1- Modeling inter-features correlation	√
4- Modeling temporal dynamics	X
2- Modeling individuality and heterogeneity	~



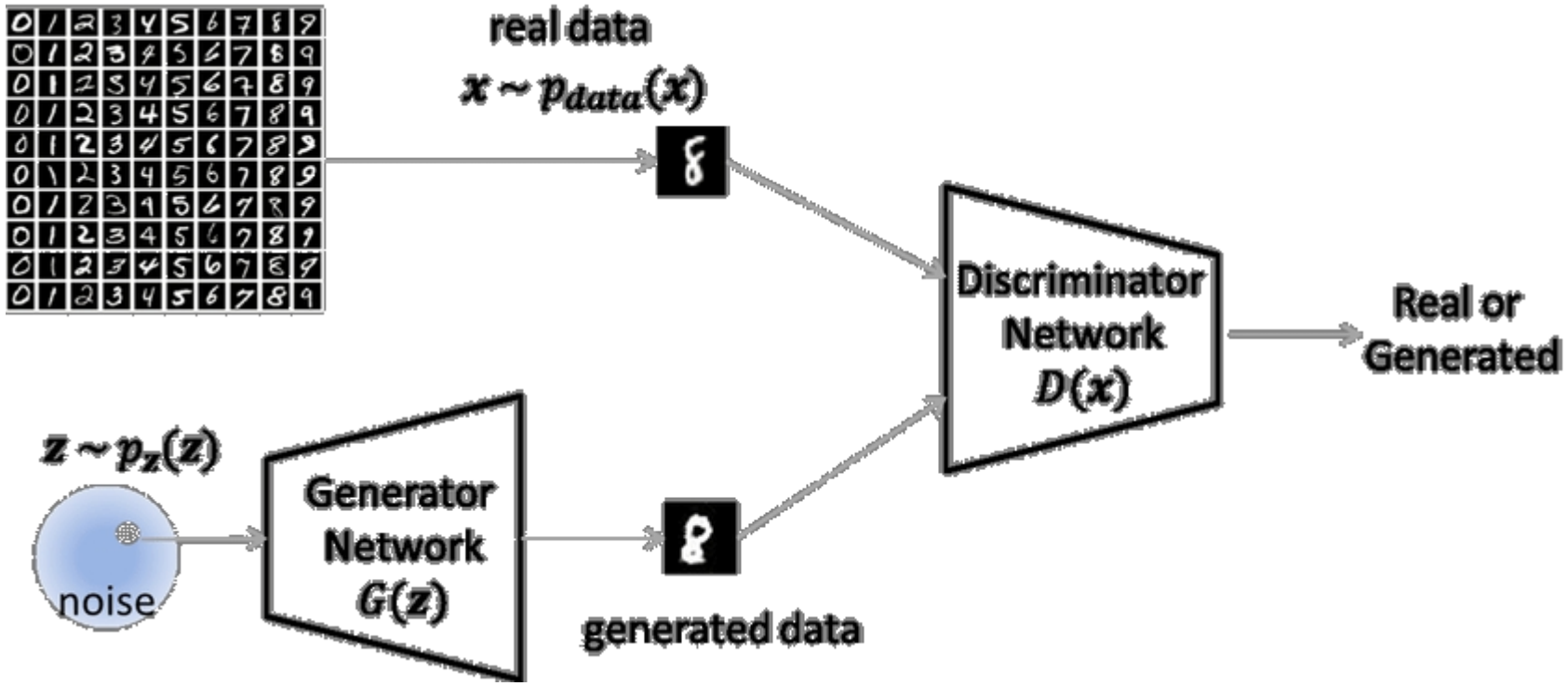
zen



real

3- Approach 3

GAN inspired



- Expected access to complete traces with per individual traffic and mobility
- Goal: capture any existing correlation between traffic and mobility behaviors

3- Approach 3

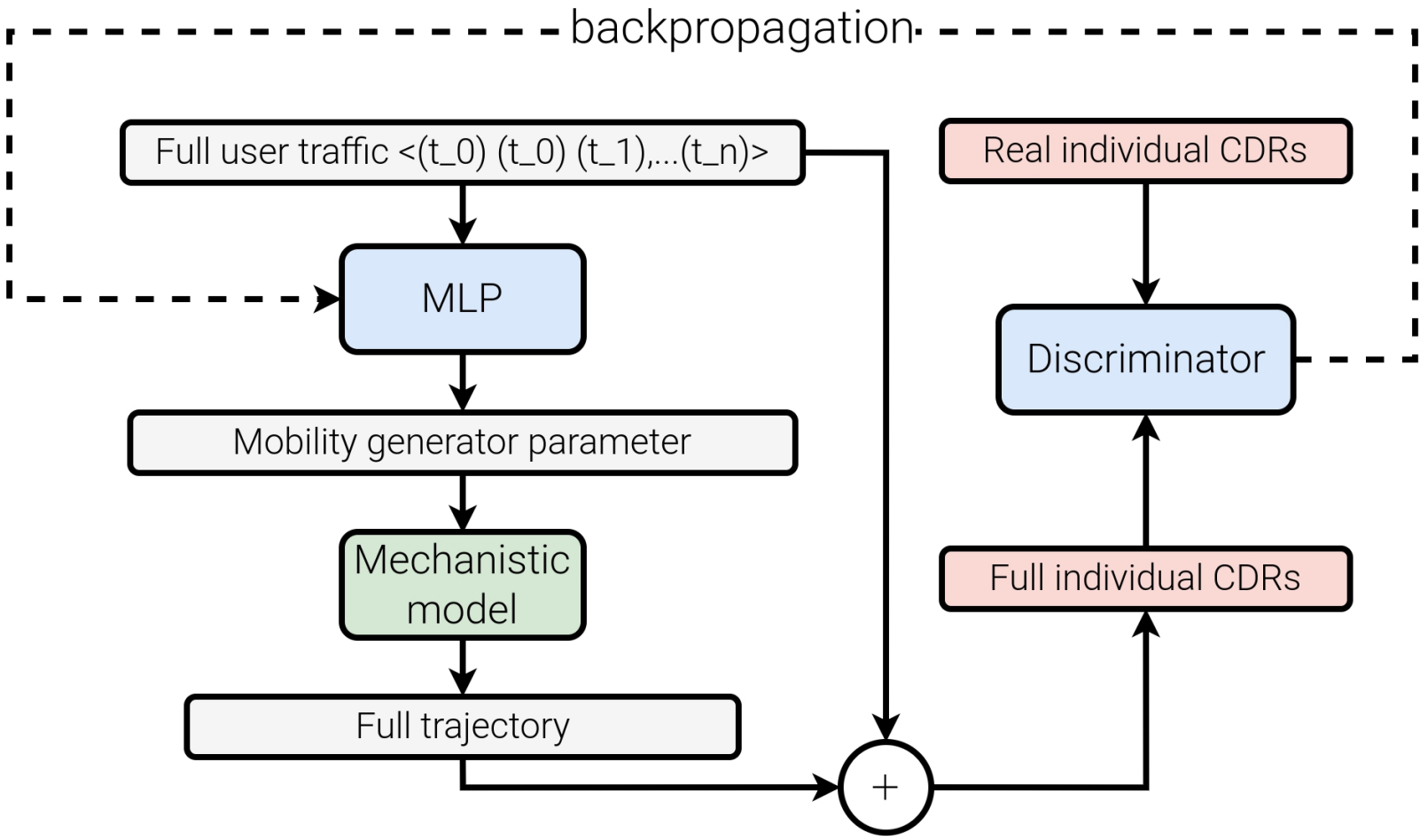
GAN inspired

Generating separately mobility and traffic from the same dataset and merging realistically

- Traffic modeling: Zen
  - What (call/data/sms) => (data)
  - When: time discretization in 10min/5min intervals and prediction of the sequence of number of events per slot
    - Challenges: lot of zeros
    - High dimensionality e.g., 1008 time slots in a week for 10min long
- Mobility modeling: mechanistic model (e.g., TimeGeo or DITRAS/d-EPR)
- Merging
  - train a model mapping traffic sequence to mechanistic model parameters
    - e.g. TimeGeo (n_w,β_1, β_2) or d-EPR markov model parameters
    - Possibility to use a GAN that'll validate or not output CDRs
      - Challenge: encode an individual's CDR sequence

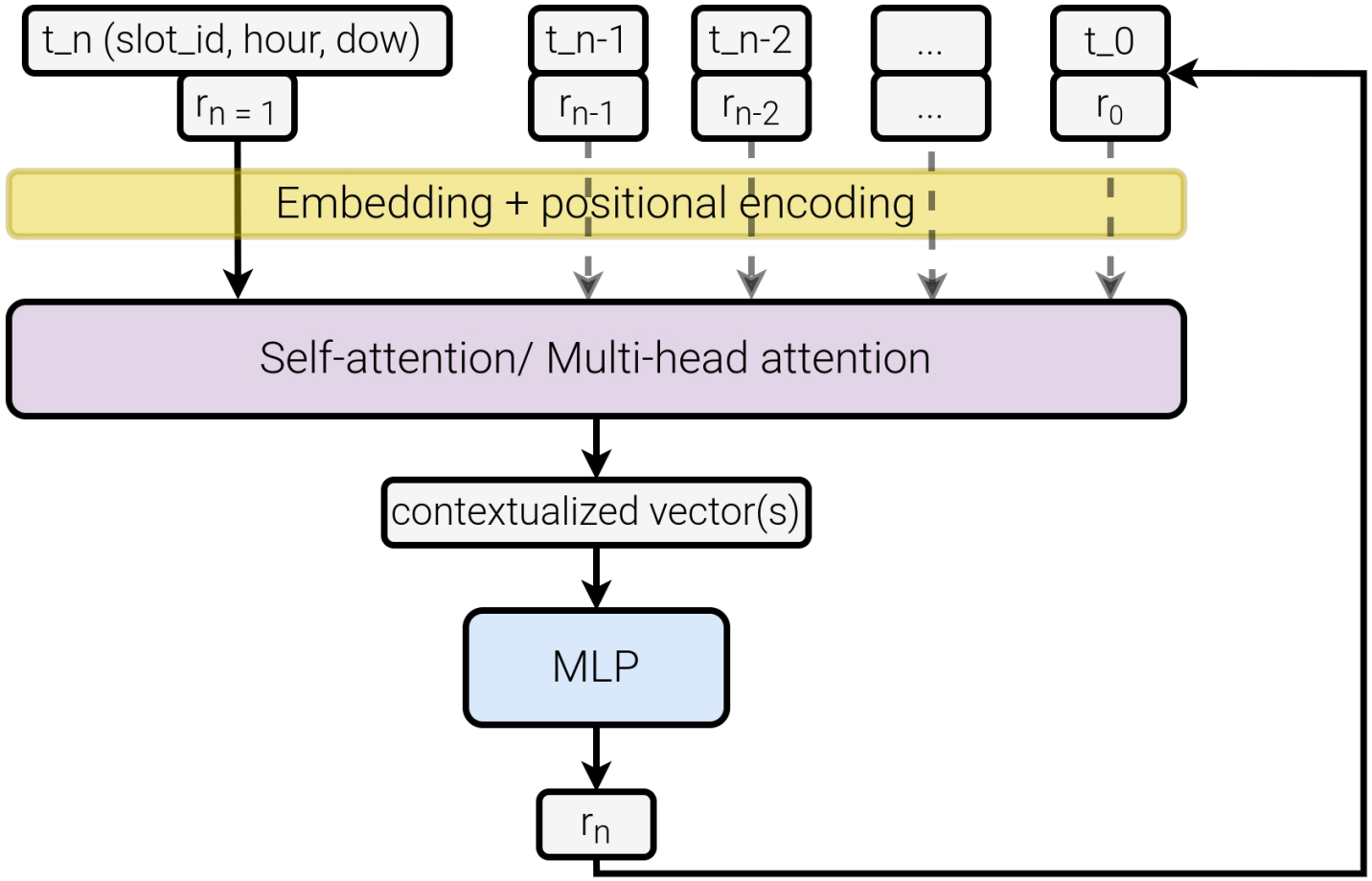
3- Approach 3

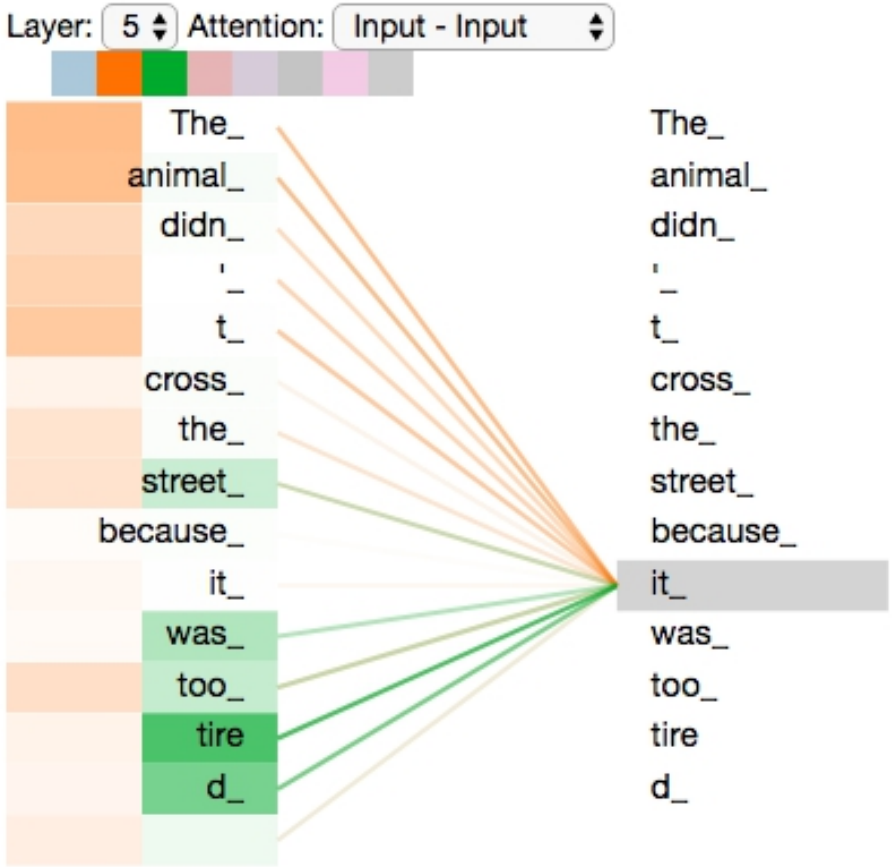
GAN inspired



3- Approach 3

GAN inspired





To conclude

- Research is all but straightforward
- The road to “Eurêka” may be long
- Identifying the requirements and challenges is important
- Models can never be perfect
  - There is always room for improvement
- Strong mathetical background may help

Thanks !

Are there any questions ?