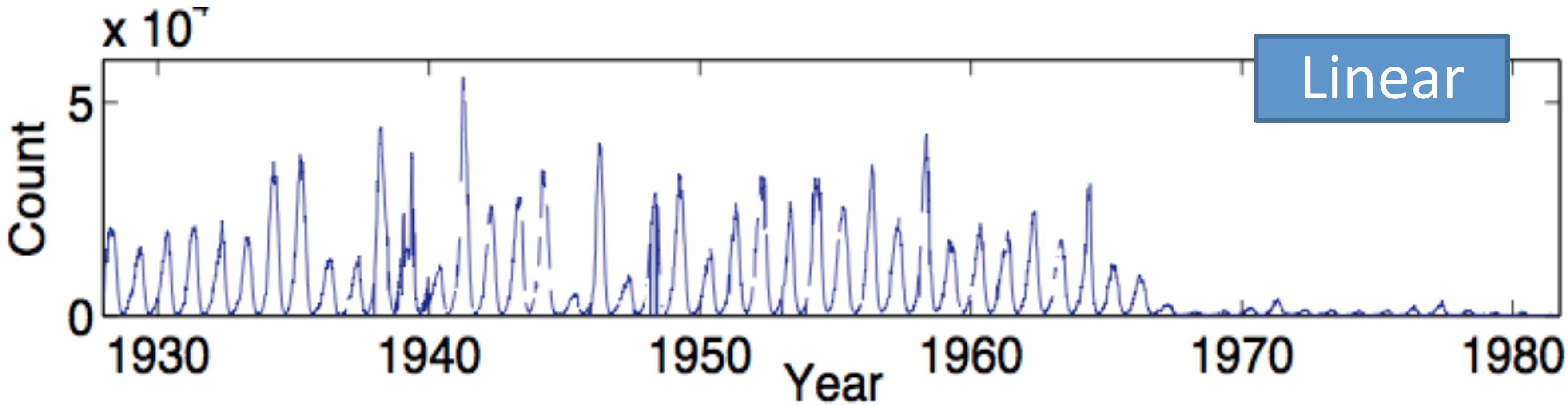
FUNNEL: Automatic Mining of Spatially Coevolving Epidemics

Yasuko Matsubara, Yasushi Sakurai (Kumamoto University)
Willem G. van Panhuis (University of Pittsburgh)
Christos Faloutsos (CMU)

Motivation

Given: Large set of epidemiological data

e.g., Measles cases in the U.S.

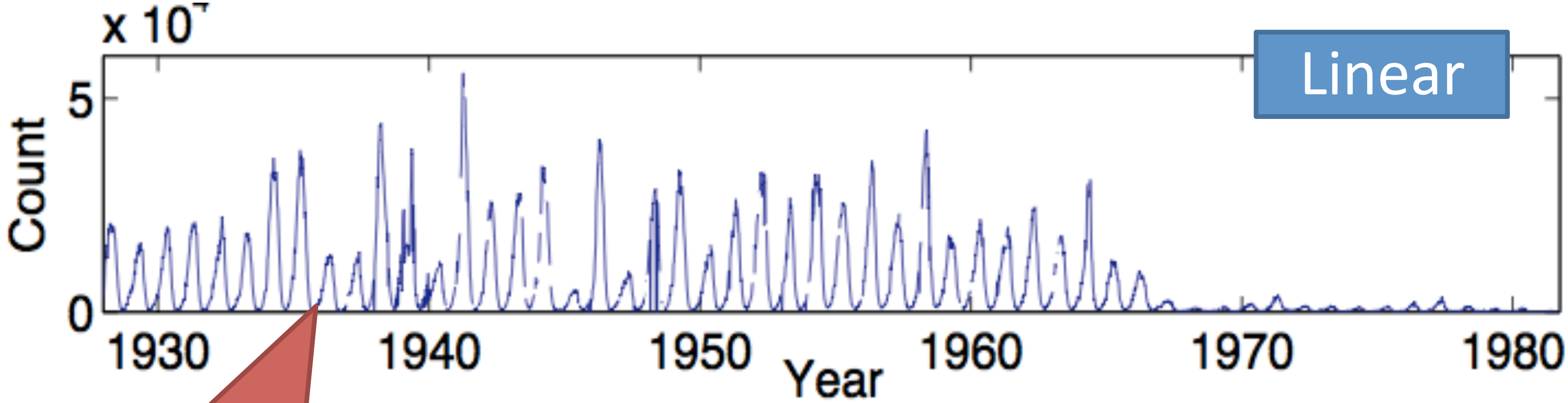


(Weekly)

Motivation

Given: Large set of epidemiological data

e.g., Measles cases in the U.S.



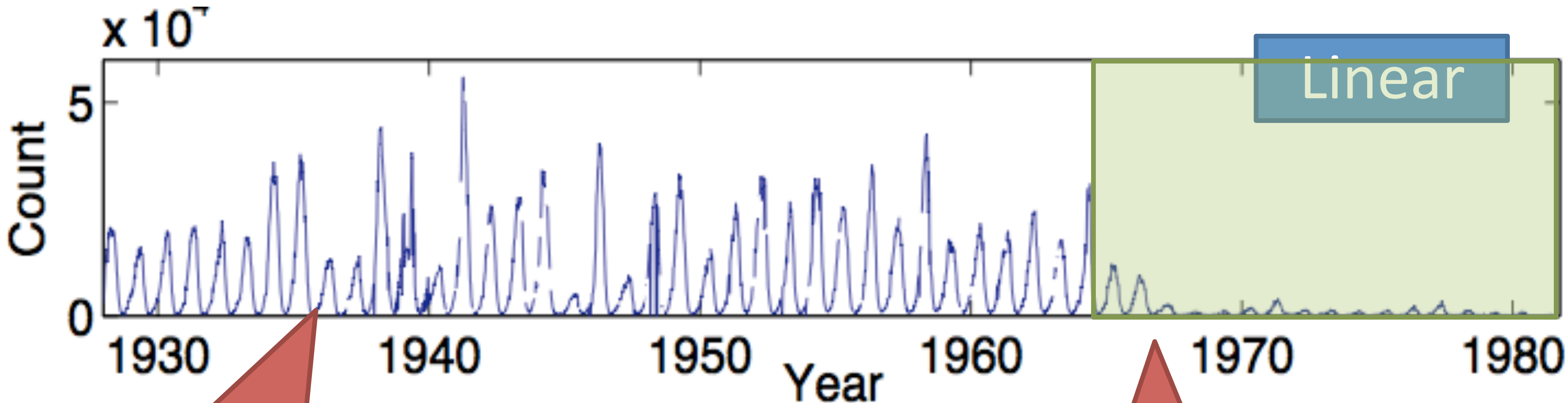
Yearly periodicity

(Weekly)

Motivation

Given: Large set of epidemiological data

e.g., Measles cases in the U.S.



Yearly periodicity

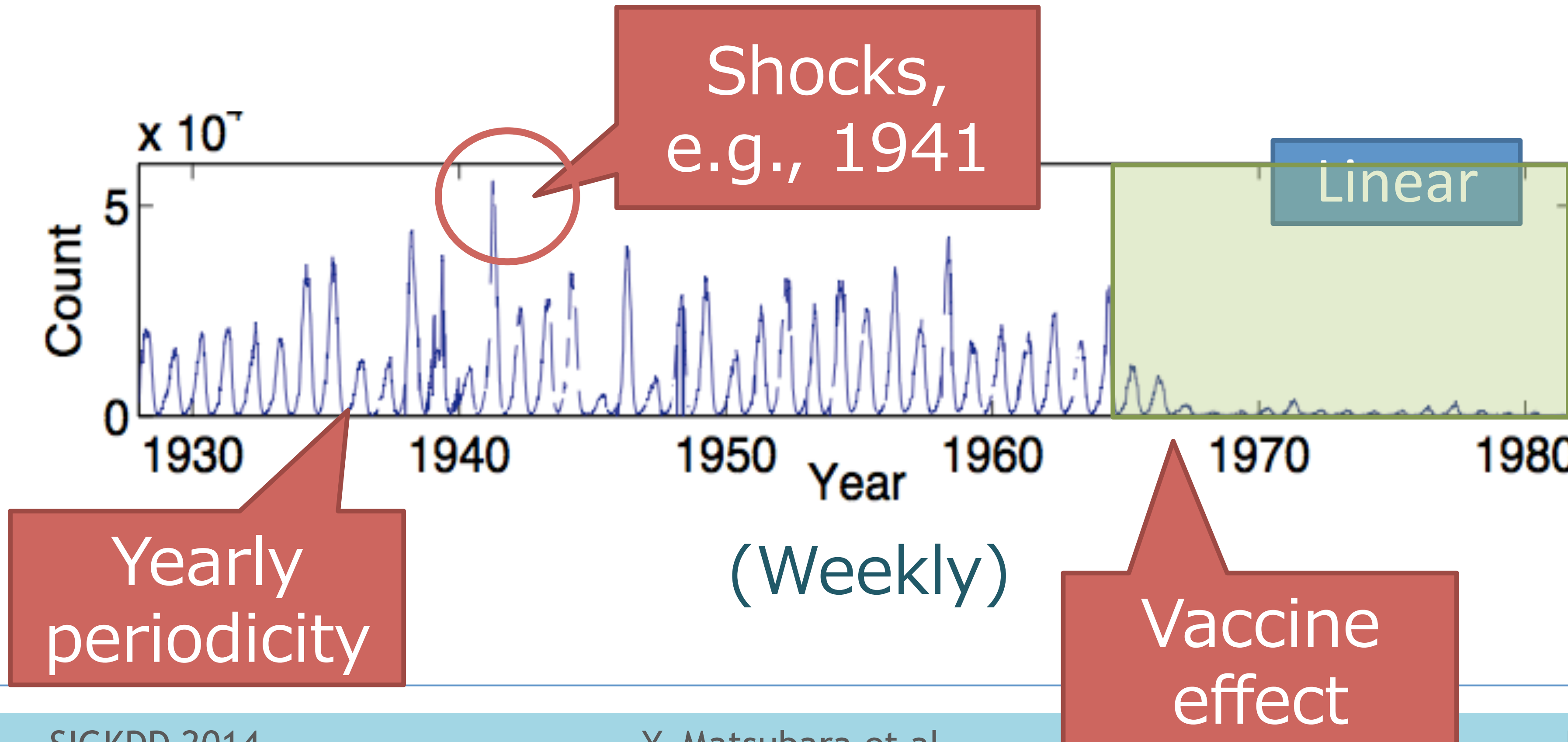
(Weekly)

Vaccine effect

Motivation

Given: Large set of epidemiological data

e.g., Measles cases in the U.S.

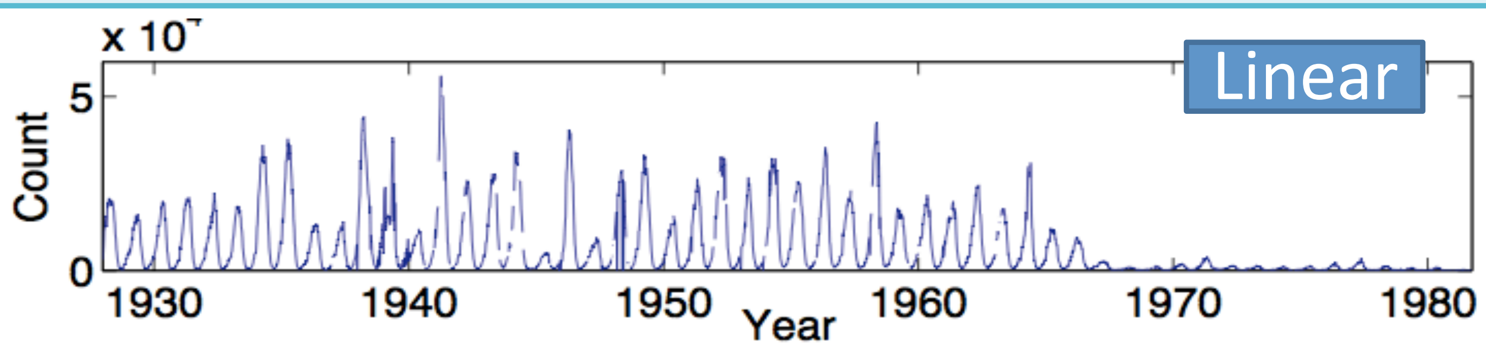


Motivation

Given: Large set of epidemiological data

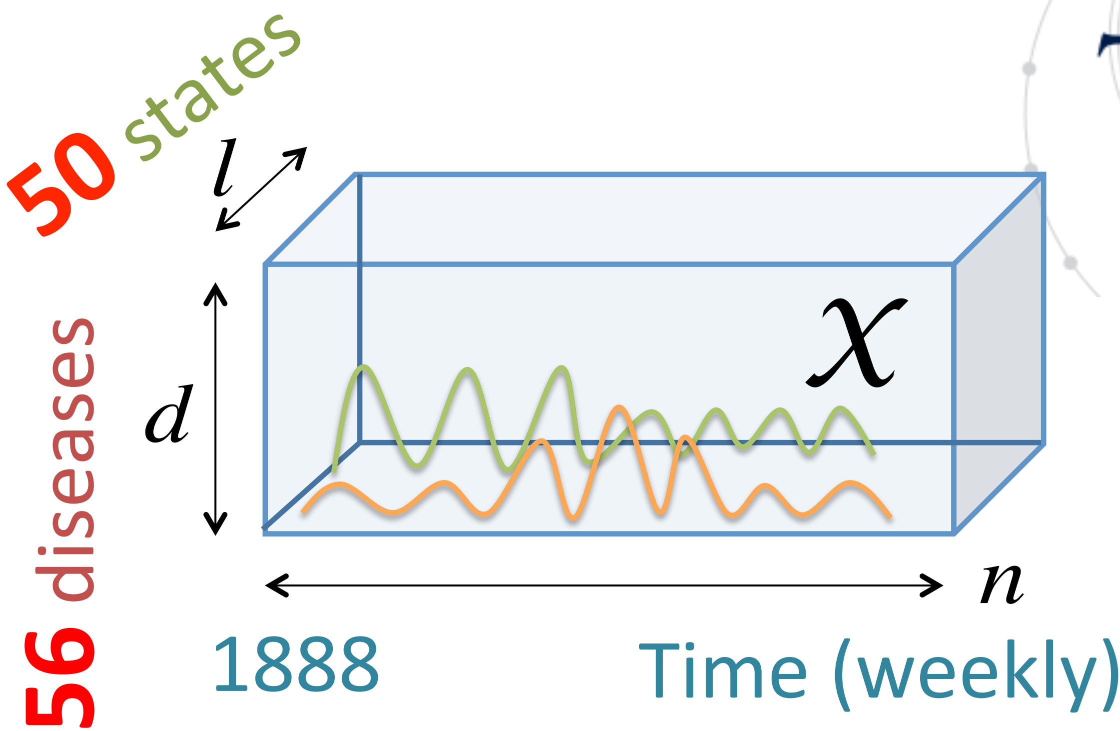
e.g., Measles cases in the U.S.

Goal: summarize all the epidemic time-series, **"fully-automatically"**



Data description

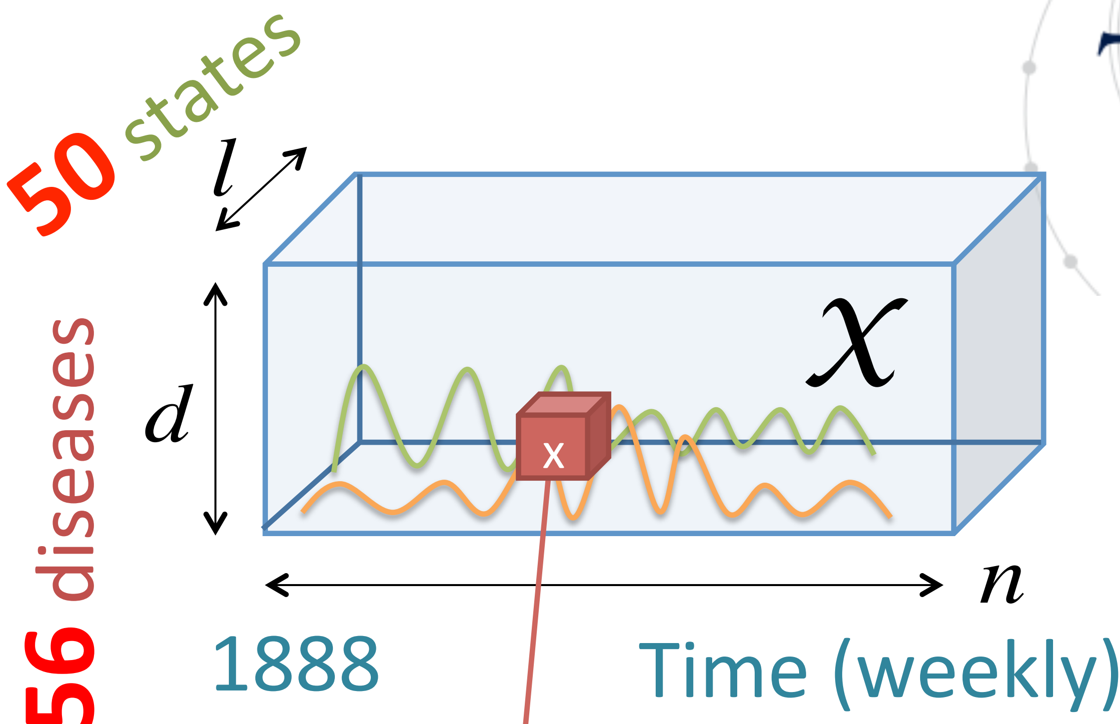
Project Tycho: infectious diseases in the U.S.



(> **125** years)

Data description

Project Tycho: infectious diseases in the U.S.



(> **125** years)

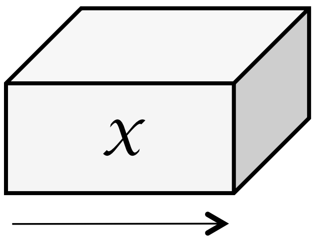
Element x : # of cases

e.g., ‘measles’, ‘NY’, ‘April 1-7, 1931’, ‘4000’

Problem definition

Given:

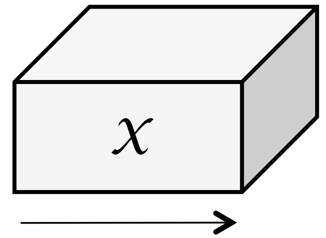
Tensor $\mathcal{X}$ (disease x state x time)



Problem definition

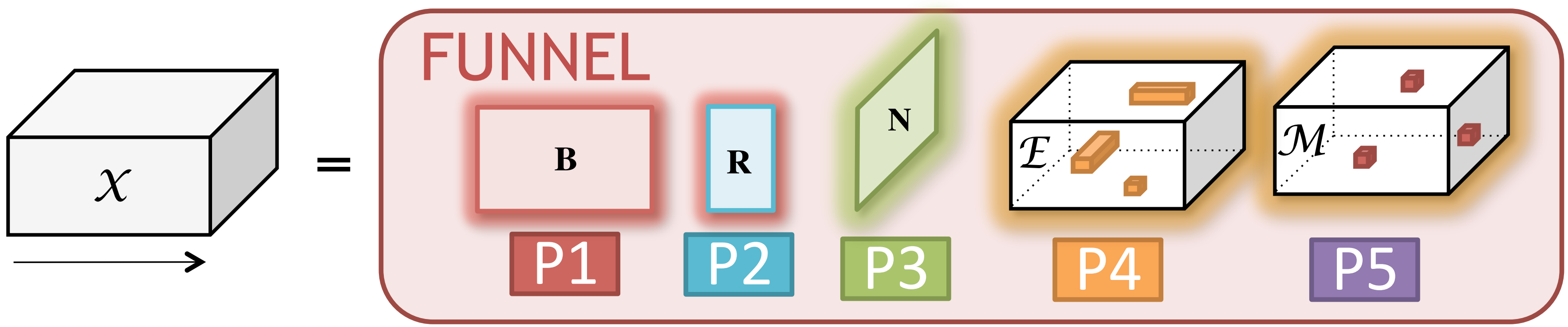
Given:

Tensor $\mathcal{X}$ (disease x state x time)

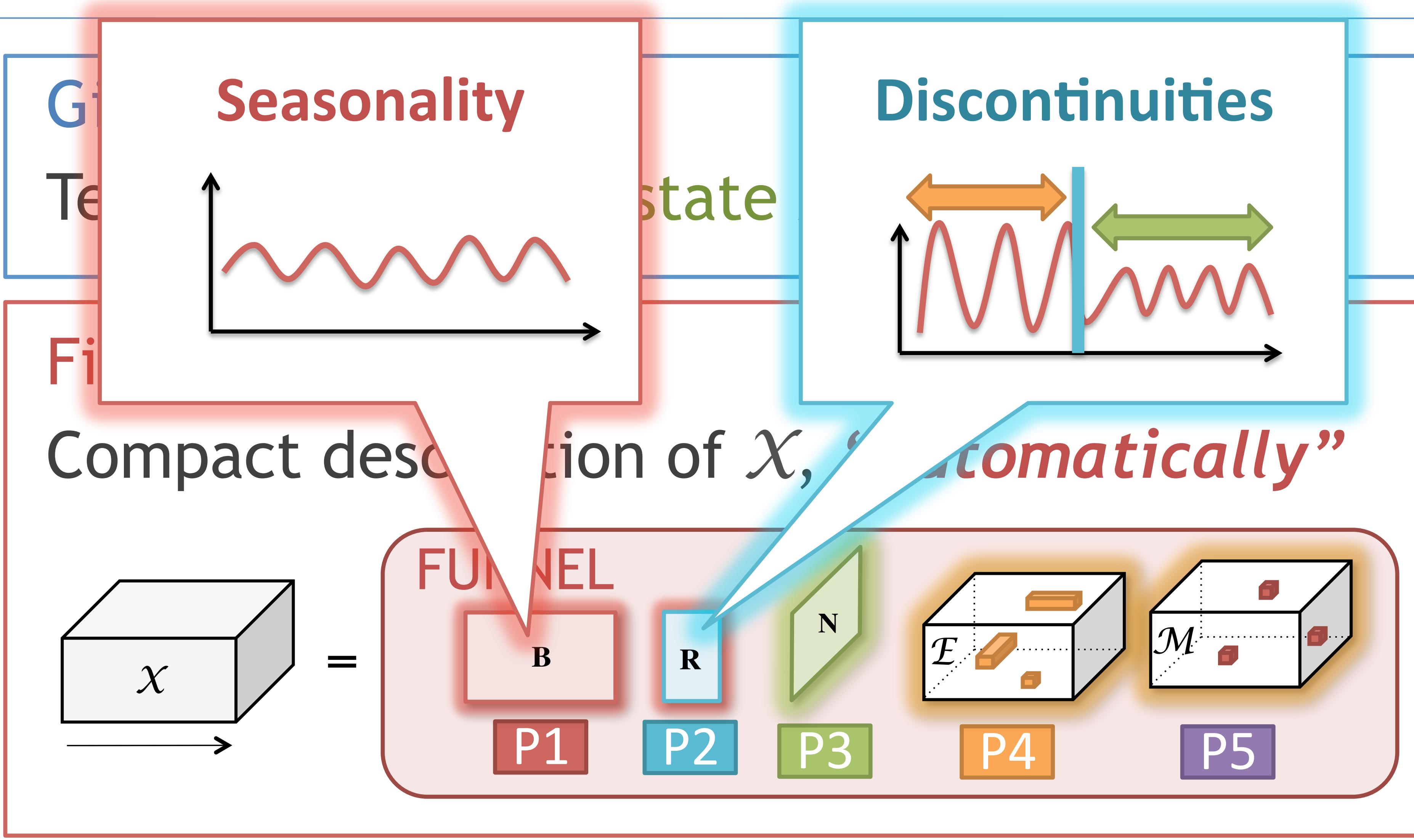


Find:

Compact description of $\mathcal{X}$, ***"automatically"***



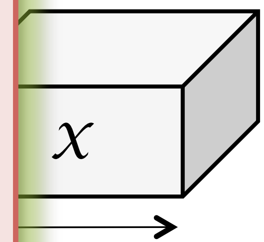
Problem definition



Problem definition

Given:

Tensor $\mathcal{X}$



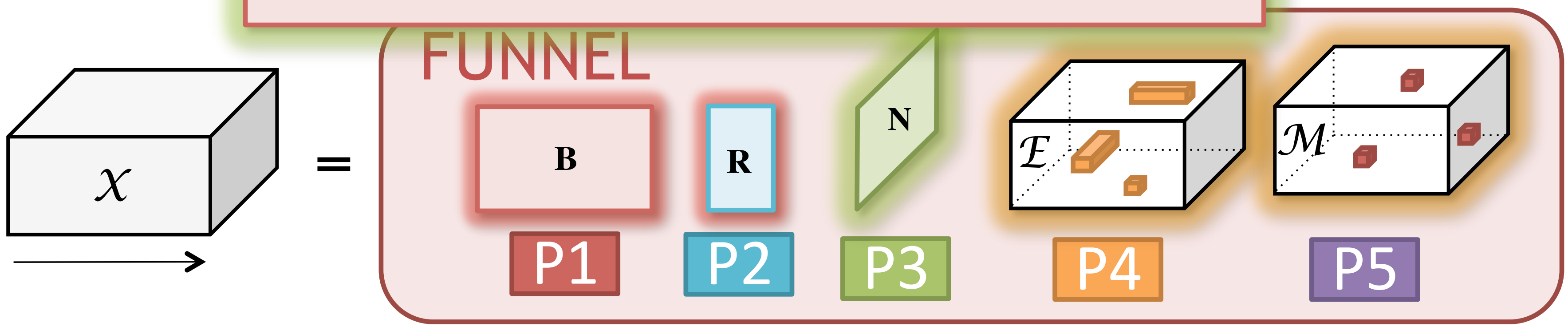
Find:

Compac

ically”

NO magic numbers !

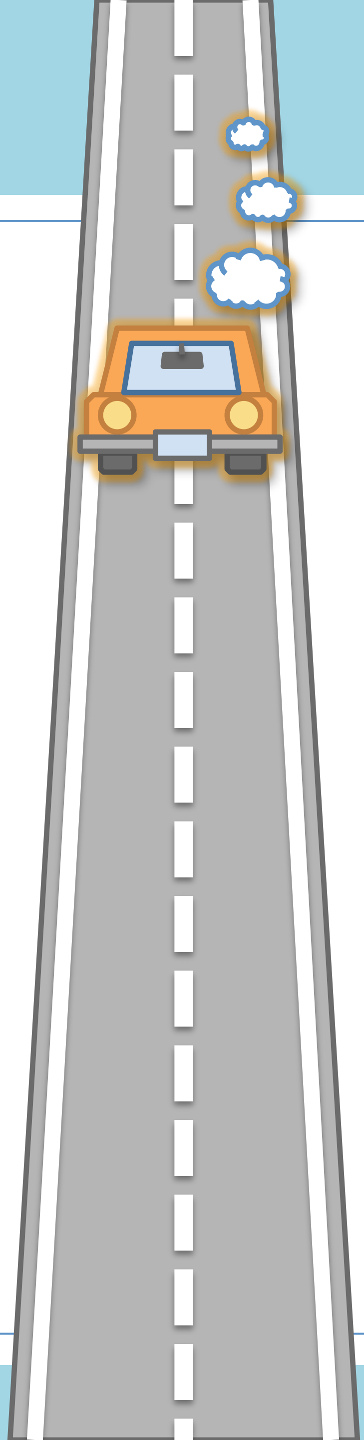
Parameter-free!



Roadmap

✔ Motivation

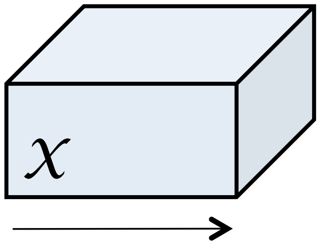
- Modeling power of FUNNEL
- Overview – main ideas
- Proposed model – idea #1
- Algorithm – idea #2
- Experiments
- Discussion
- Conclusions



Modeling power of FUNNEL

Questions about epidemics

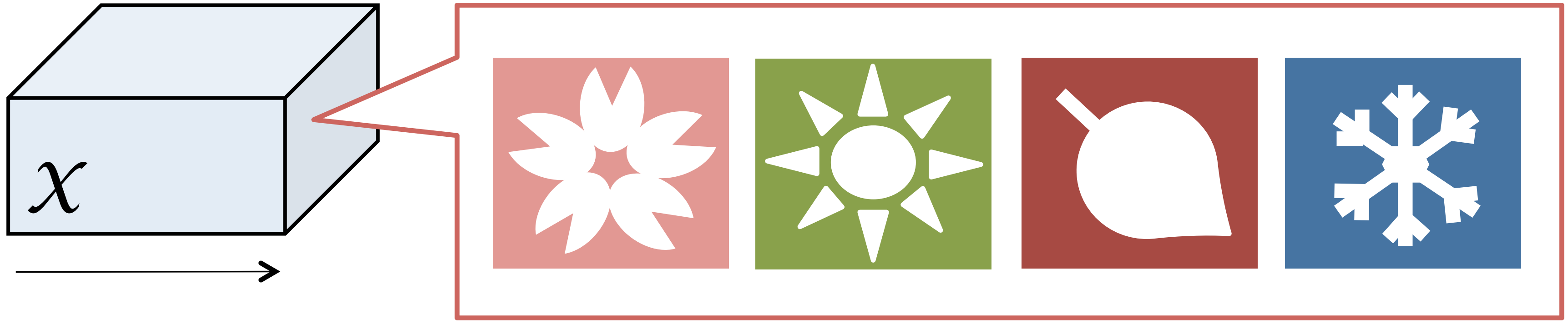
Q1 Q2 Q3 Q4 Q5



Q1

Are there any periodicities?

If yes, when is the peak season?



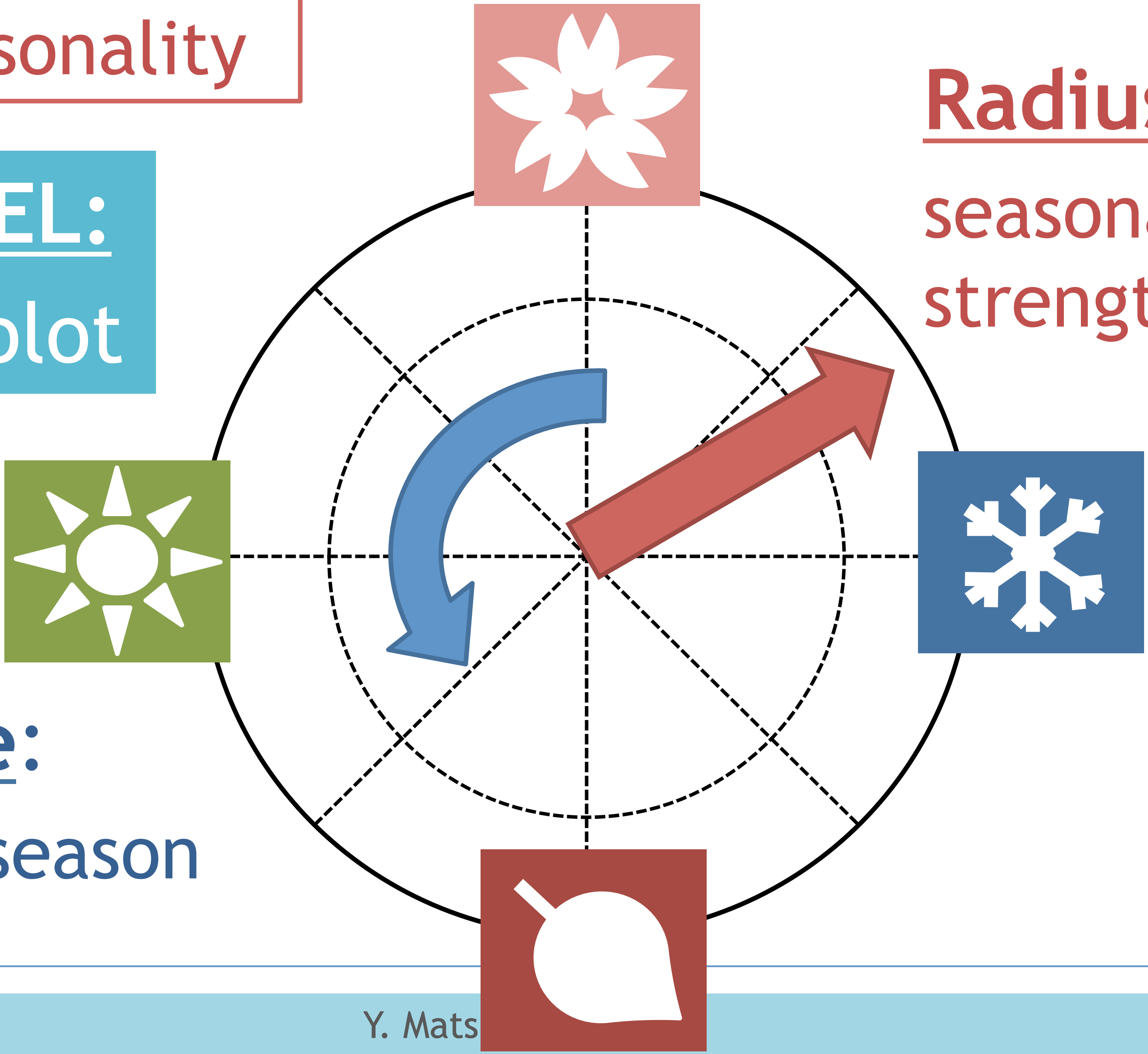
Answers

Q1 Q2 Q3 Q4 Q5

P1 Seasonality

FUNNEL:

Polar plot



Radius:

seasonality strength

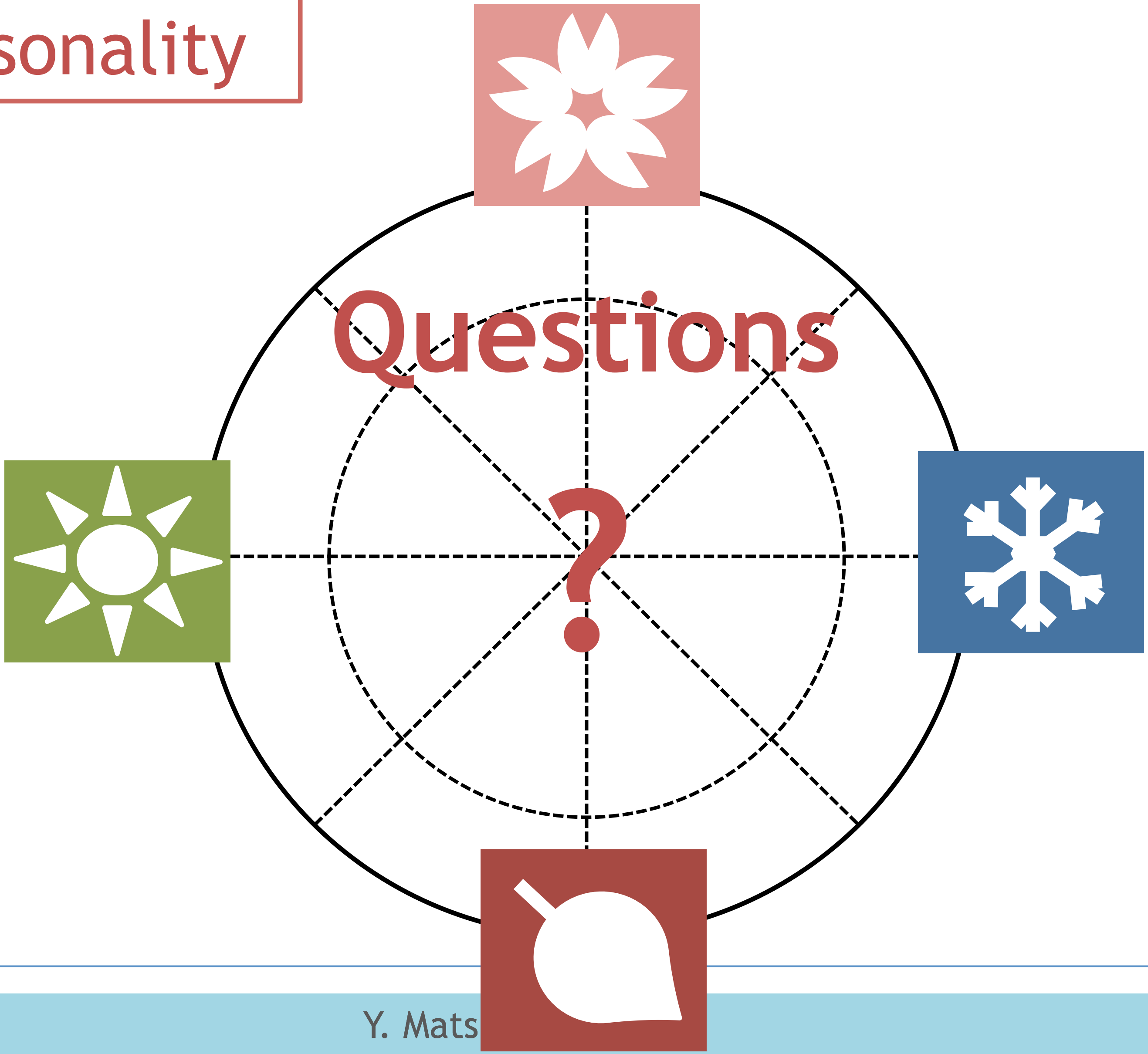
Angle:

peak season

Answers

Q1 Q2 Q3 Q4 Q5

P1 Seasonality



Answers

Q1 Q2 Q3 Q4 Q5

P1 Seasonality

Q: Does Influenza have seasonality? If yes, when?

Answers

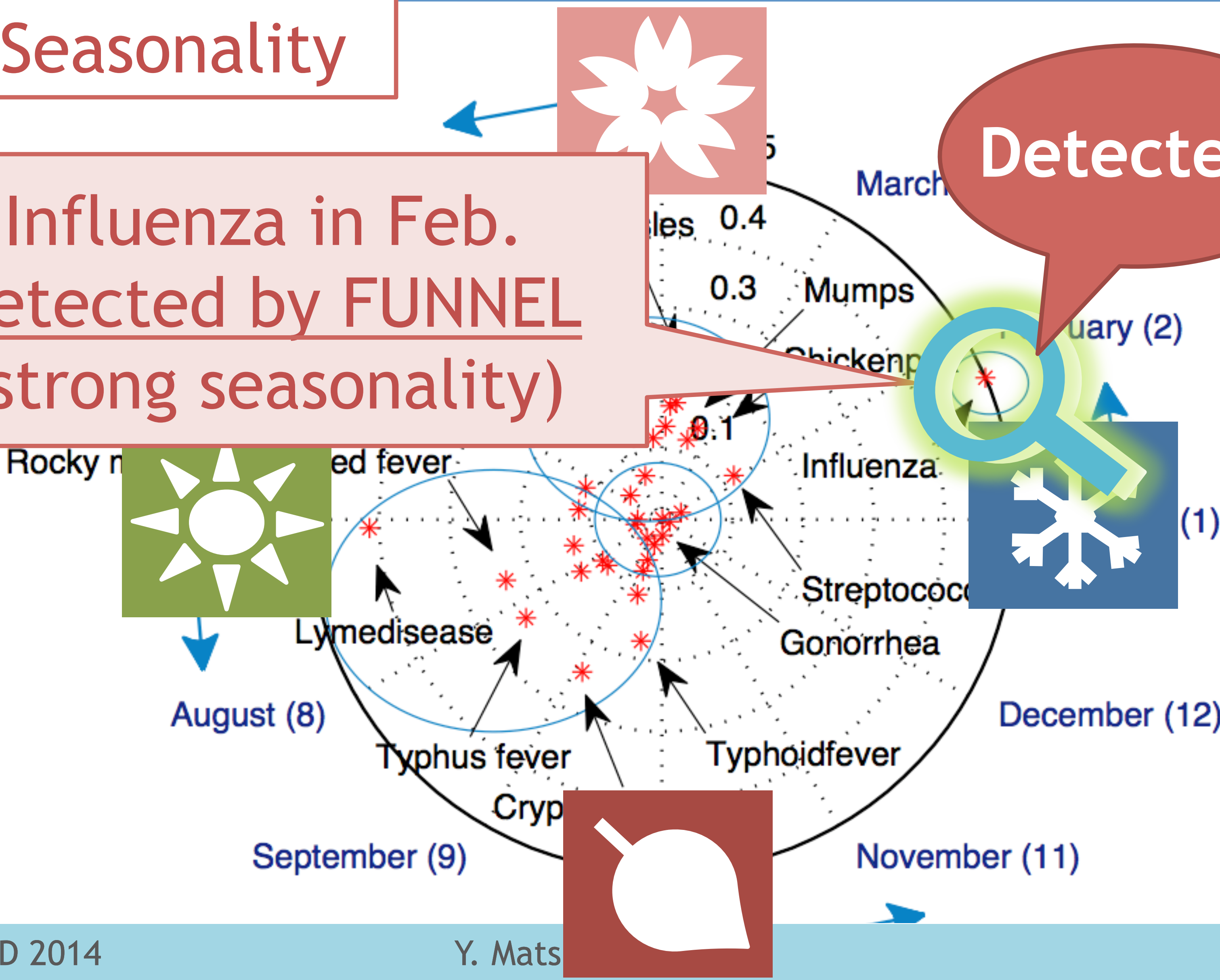
P1 Seasonality

Answers

Q1 Q2 Q3 Q4 Q5

P1 Seasonality

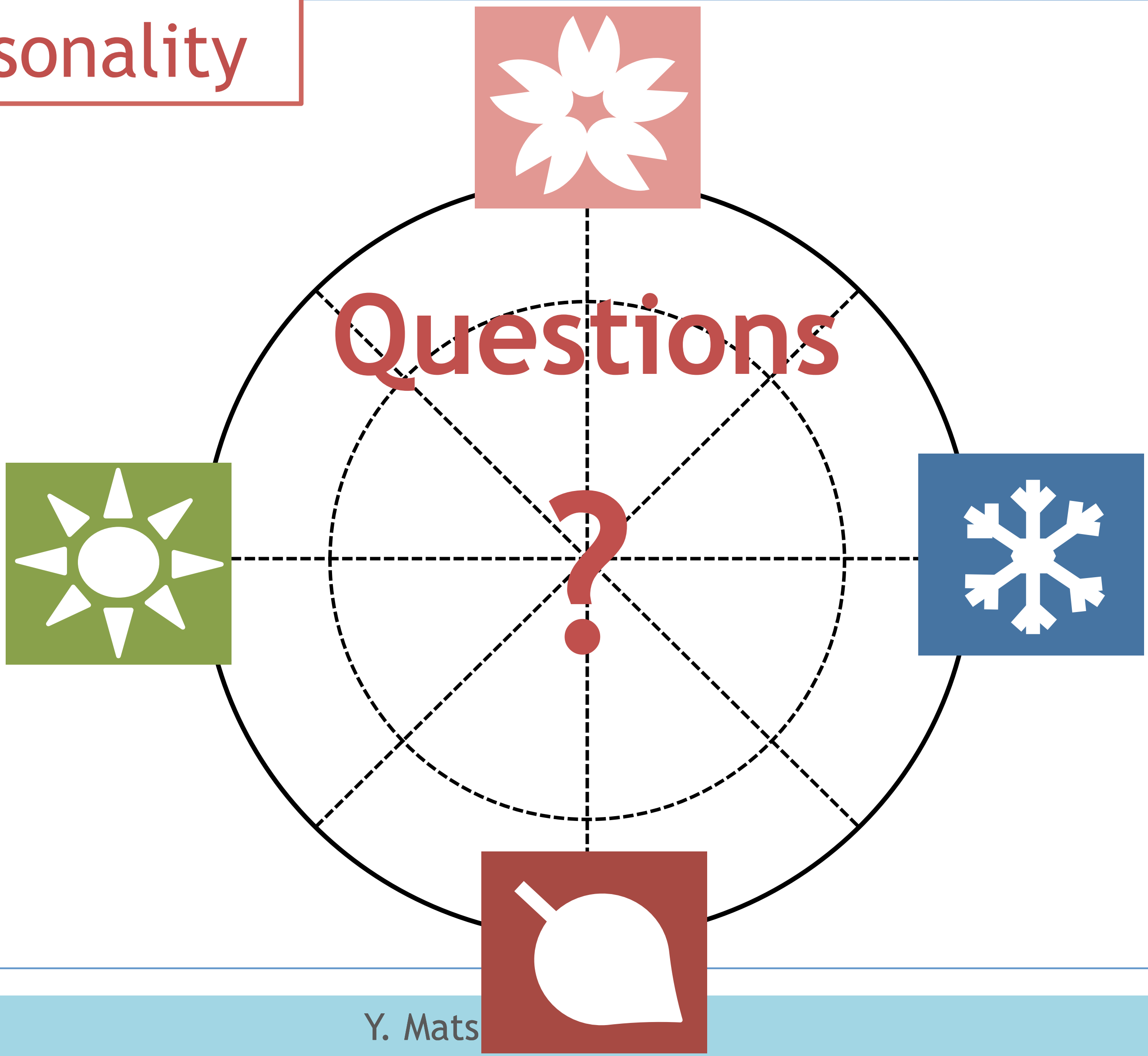
Influenza in Feb.
<u>Detected by FUNNEL</u>
(strong seasonality)



Answers

Q1 Q2 Q3 Q4 Q5

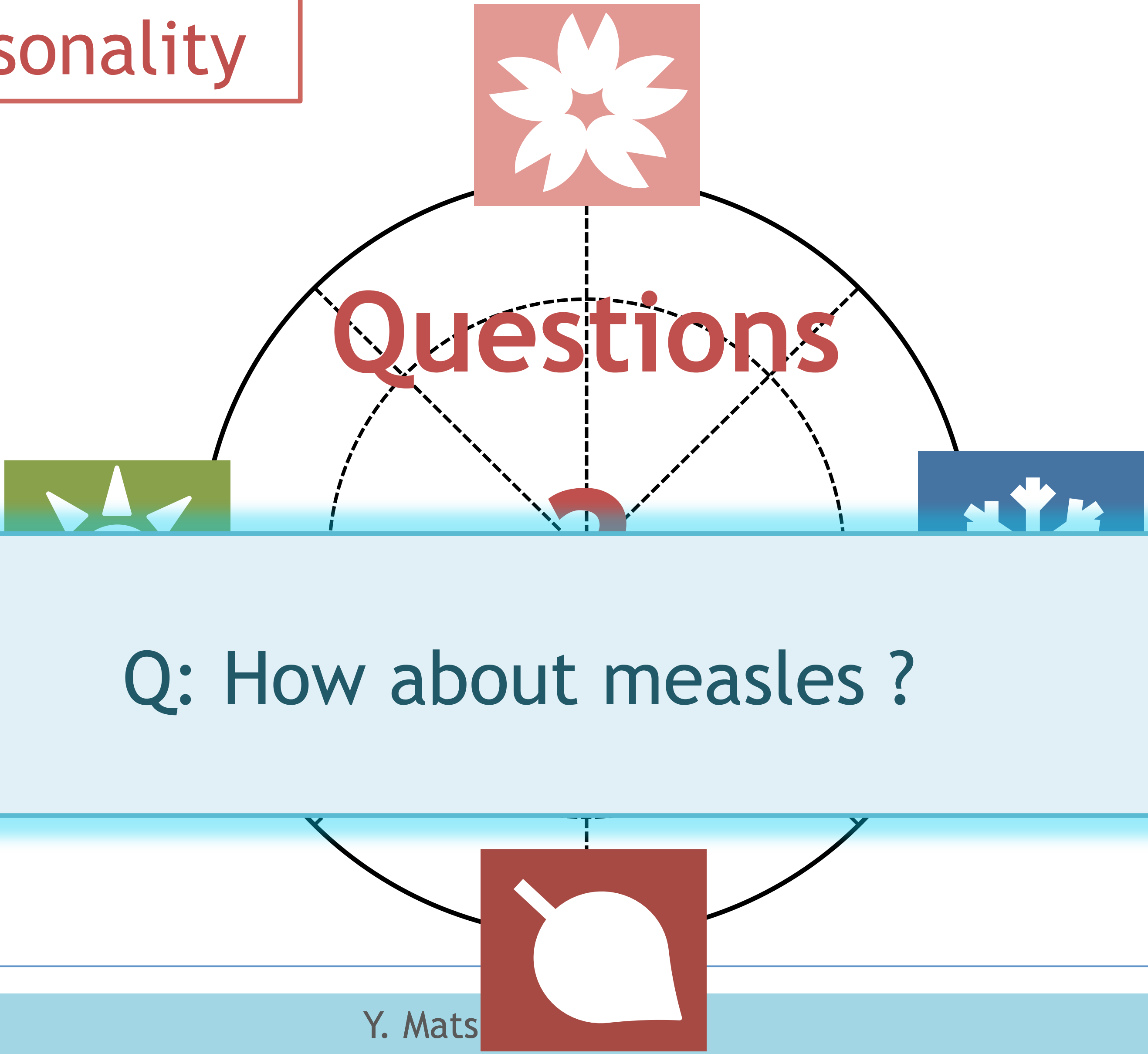
P1 Seasonality



Answers

Q1 Q2 Q3 Q4 Q5

P1 Seasonality

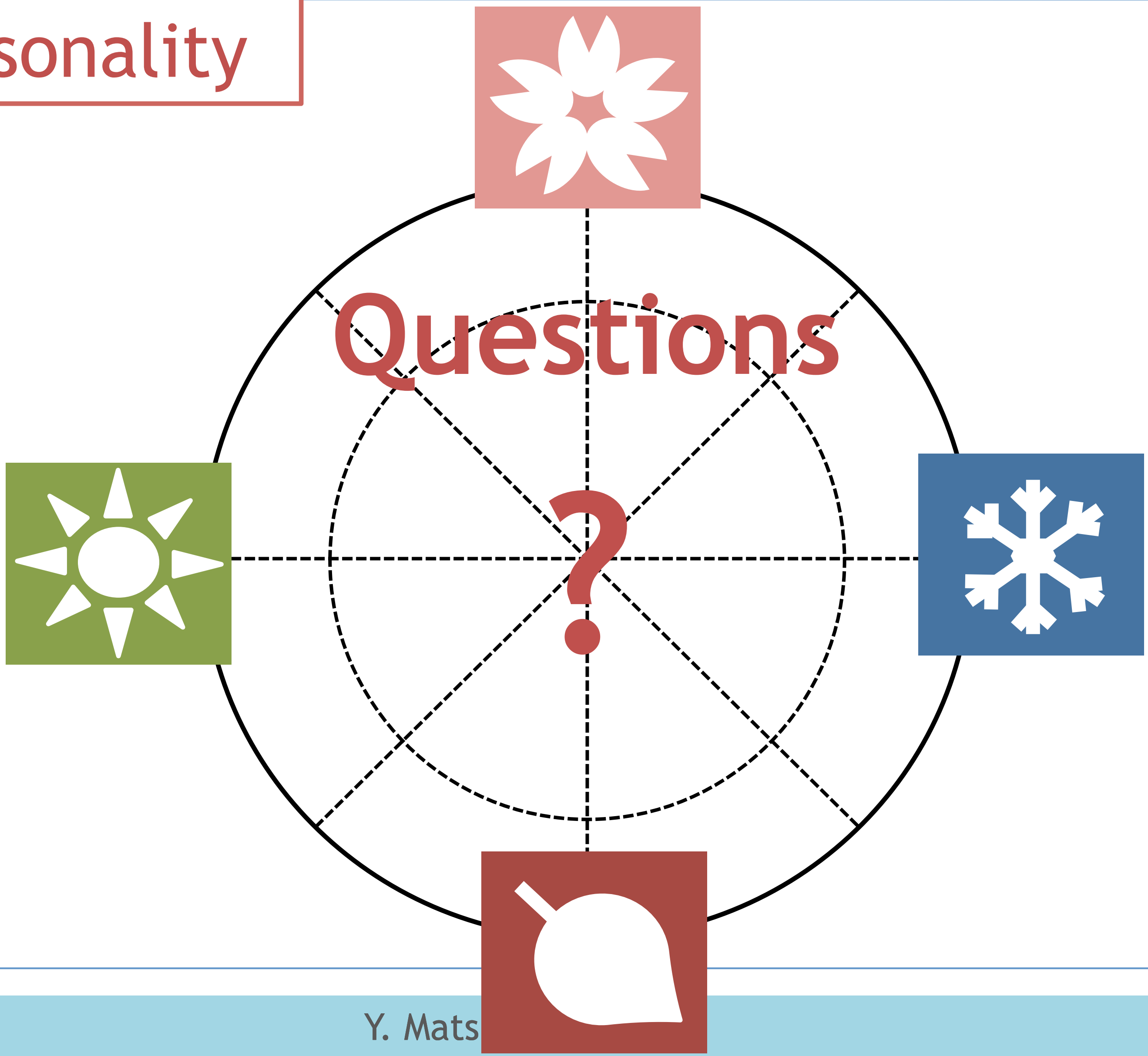


Q: How about measles ?

Answers

Q1 Q2 Q3 Q4 Q5

P1 Seasonality

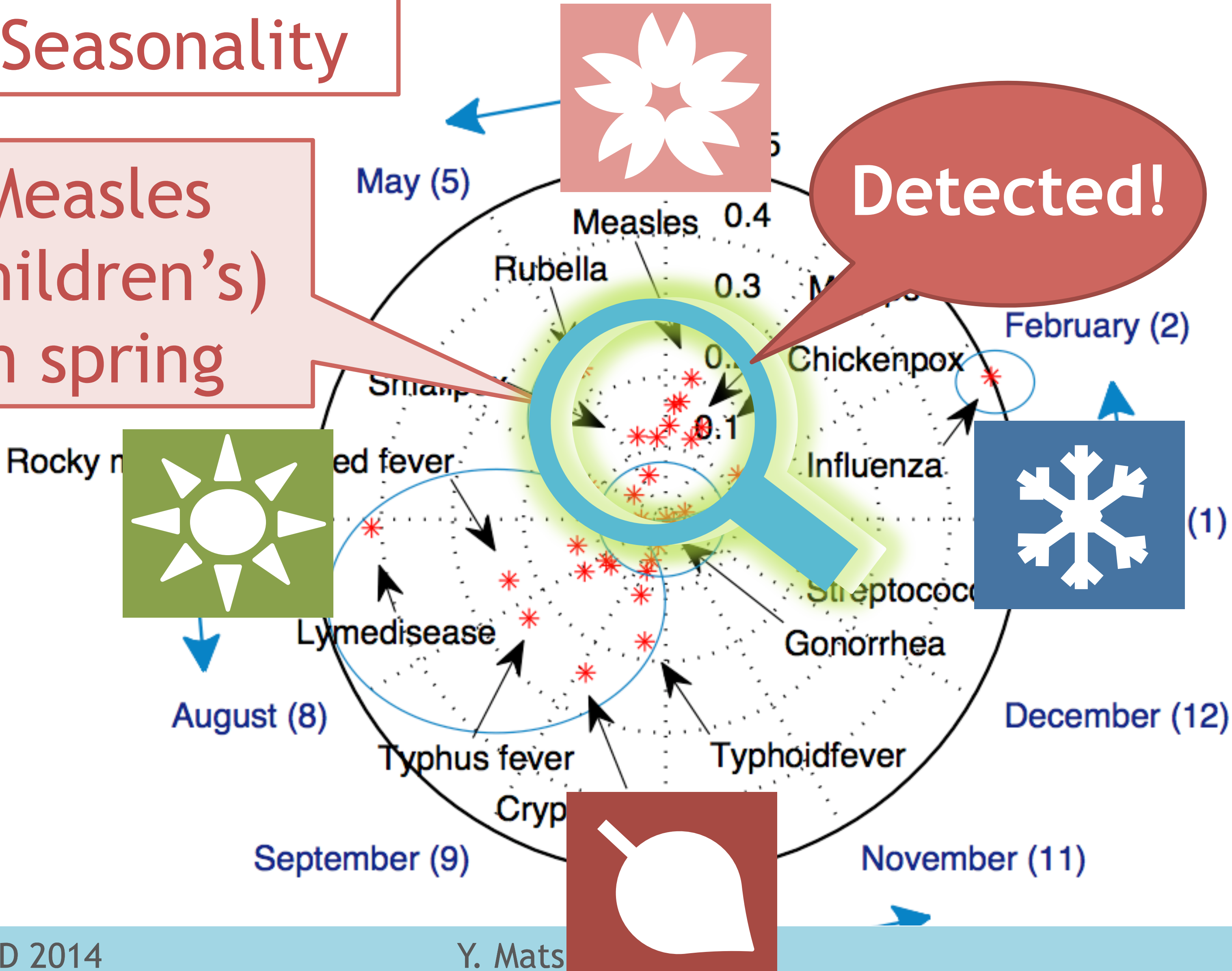


Answers

Q1 Q2 Q3 Q4 Q5

P1 Seasonality

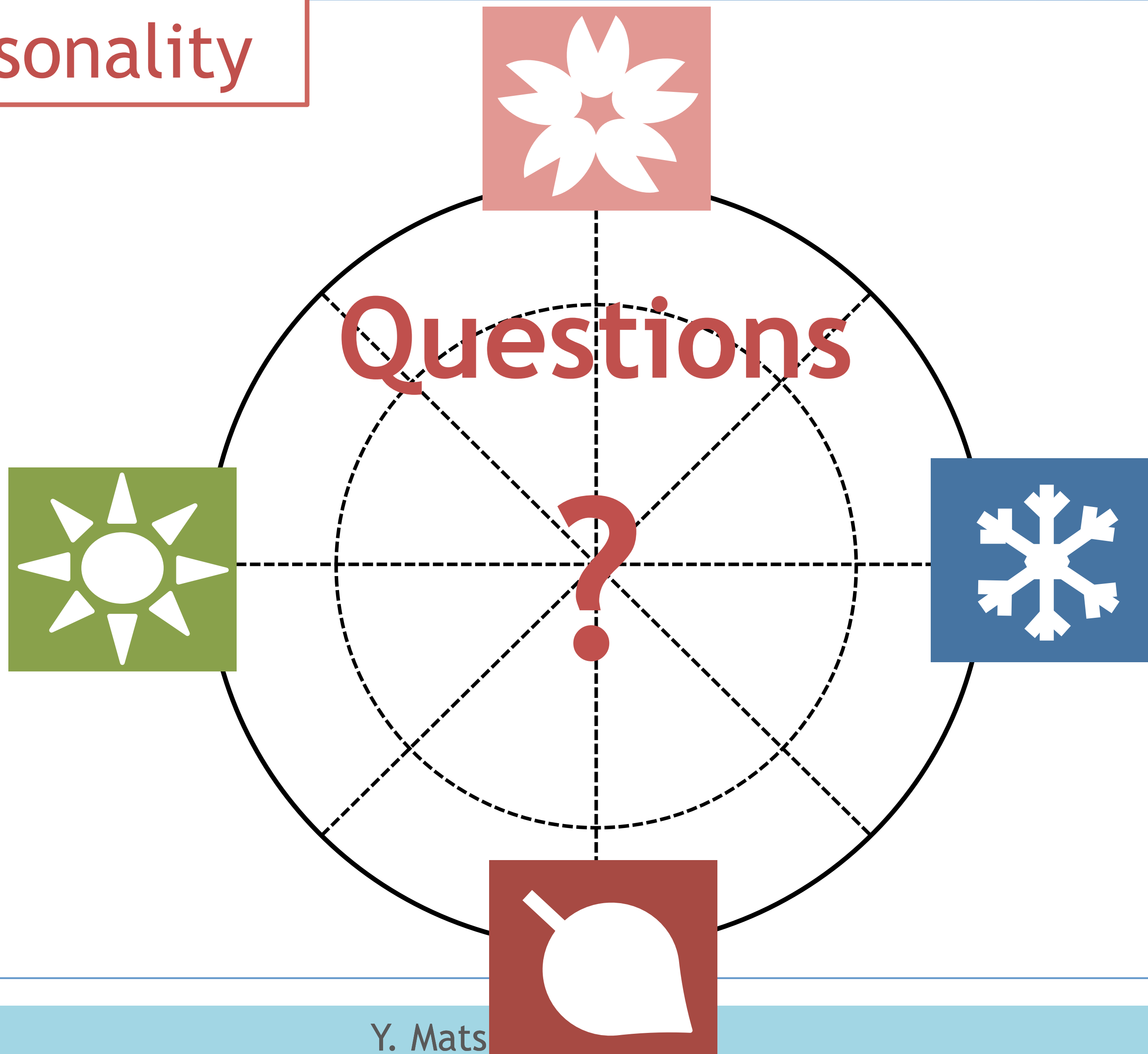
Measles (children's) in spring



Answers

Q1 Q2 Q3 Q4 Q5

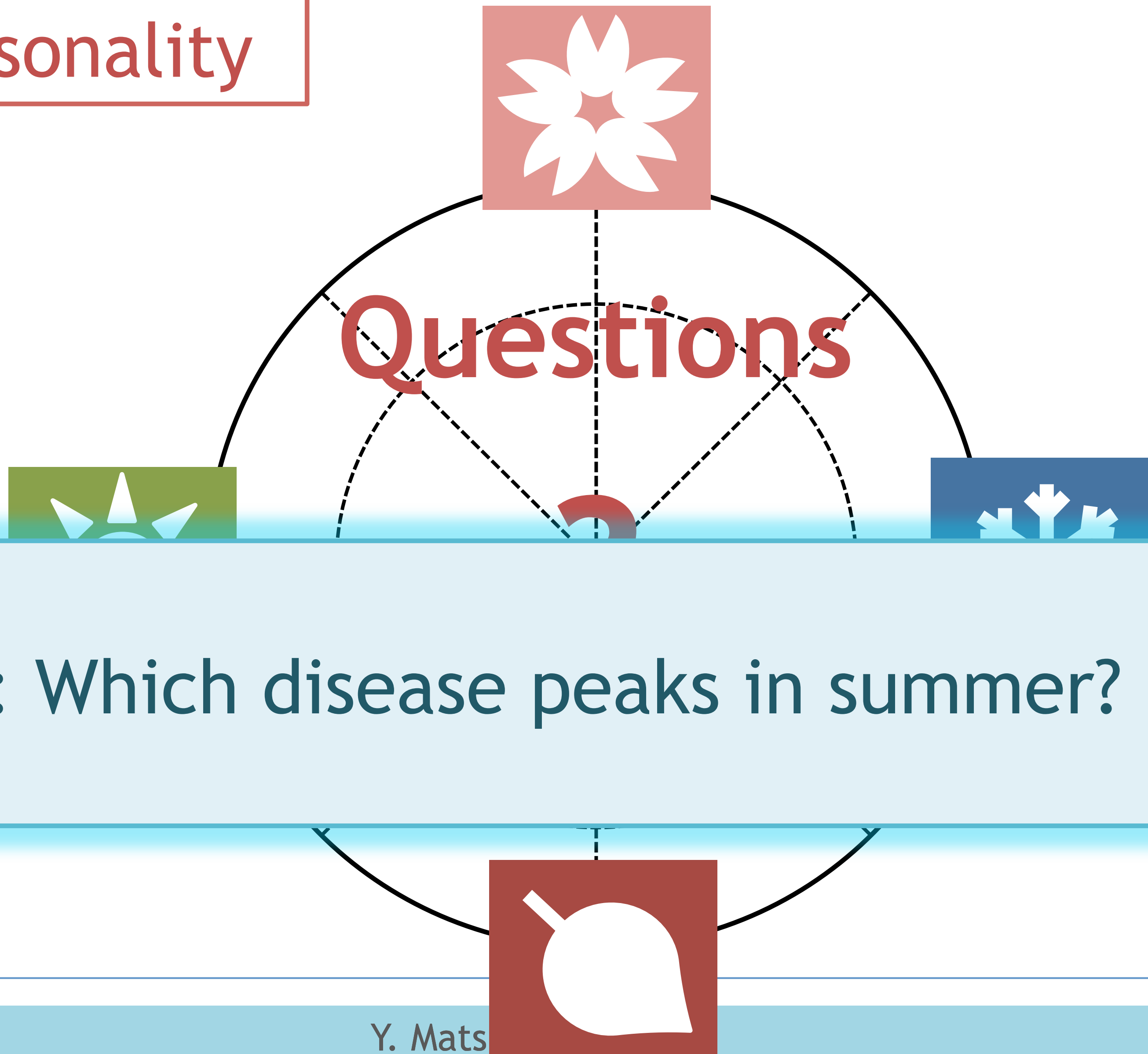
P1 Seasonality



Answers

Q1 Q2 Q3 Q4 Q5

P1 Seasonality



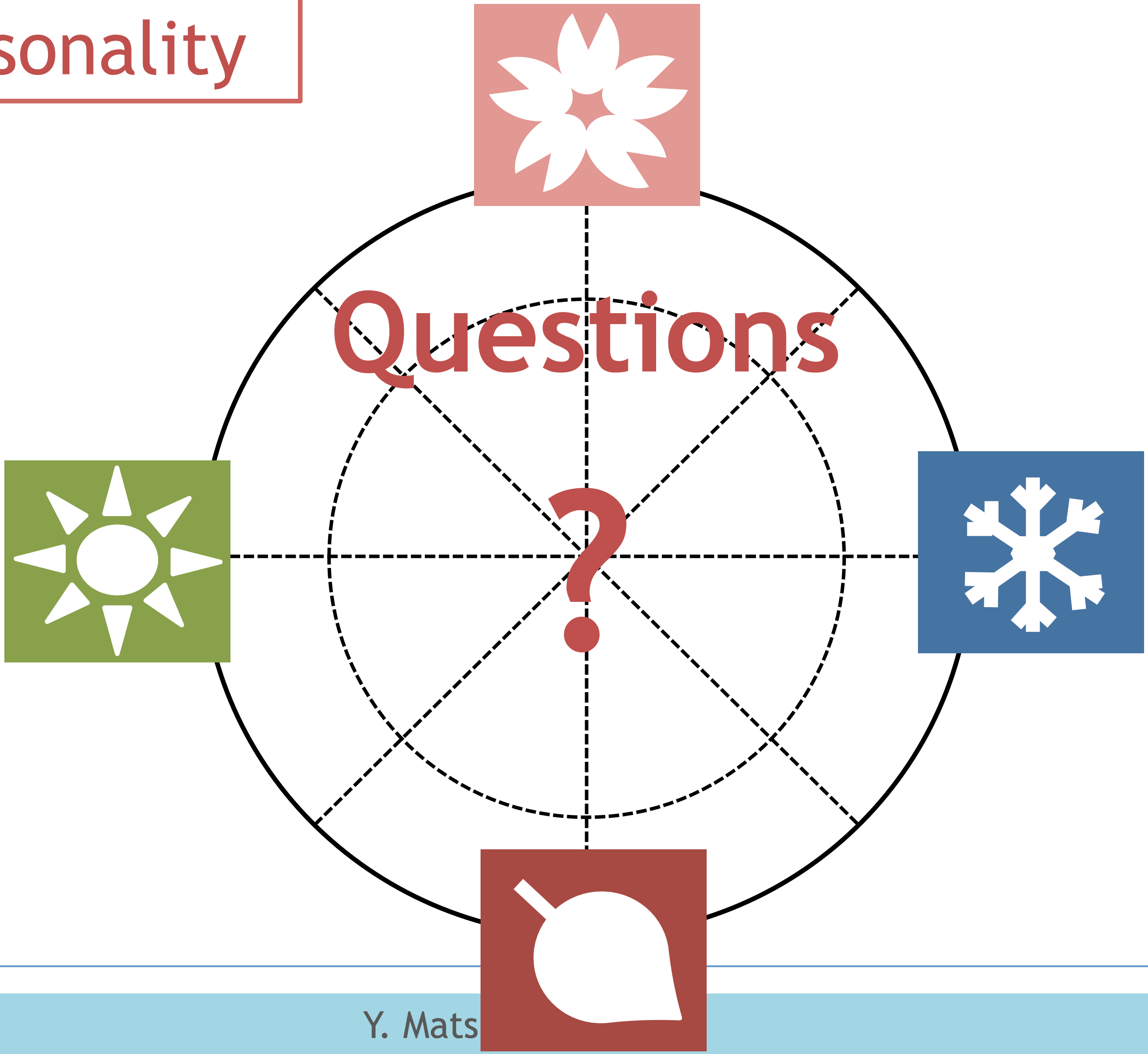
Questions

Q: Which disease peaks in summer?

Answers

Q1 Q2 Q3 Q4 Q5

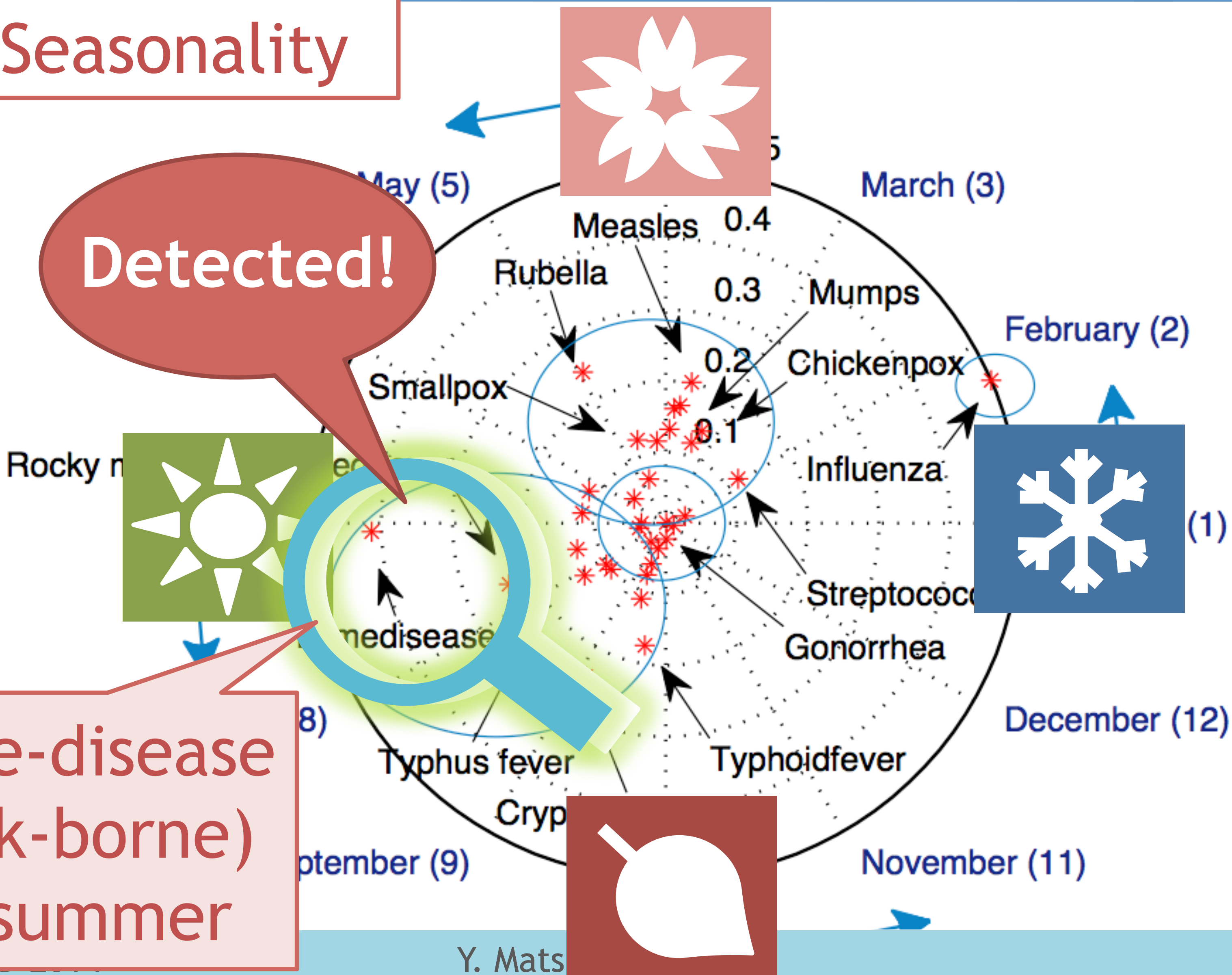
P1 Seasonality



Answers

Q1 Q2 Q3 Q4 Q5

P1 Seasonality

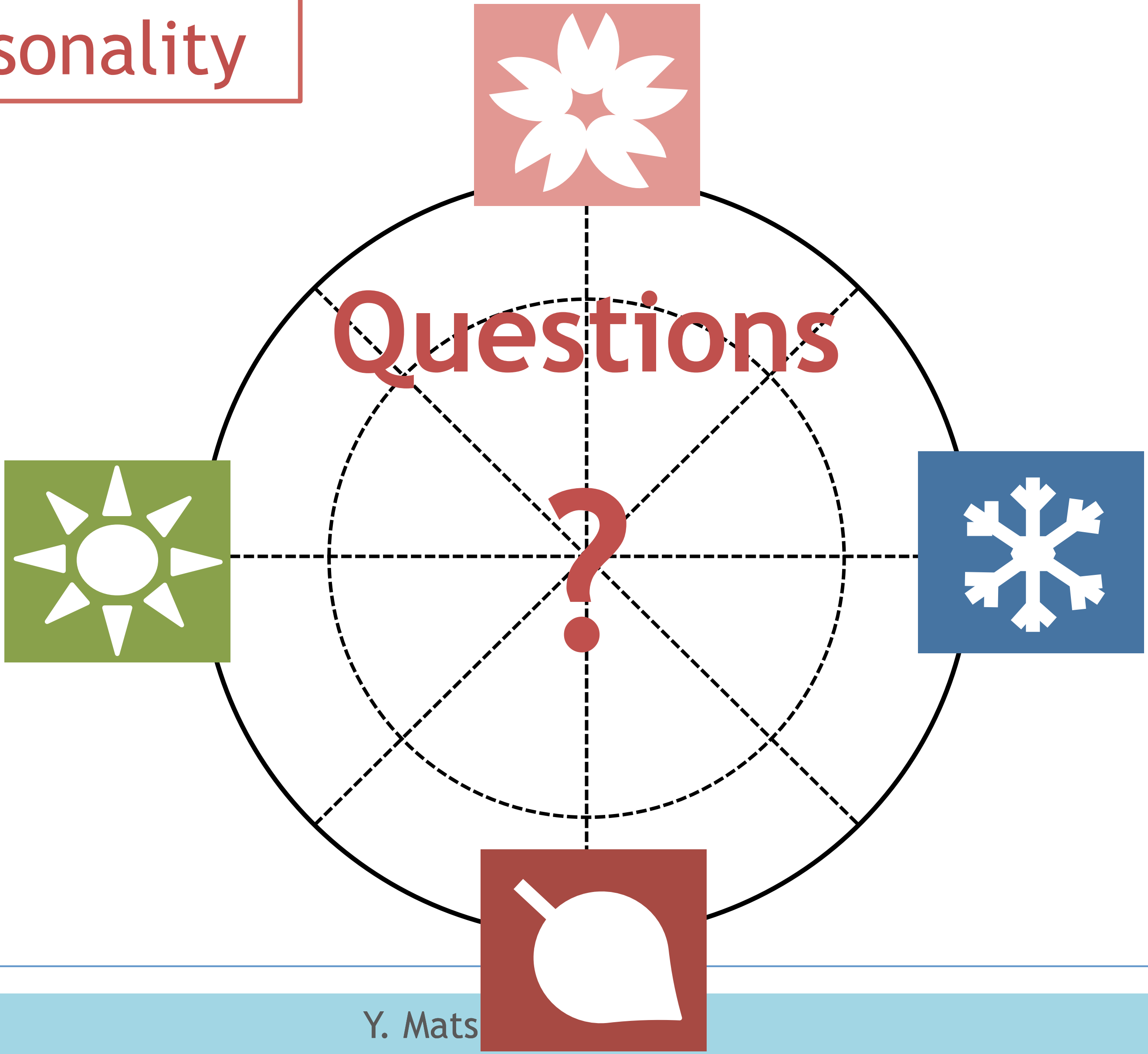


Lyme-disease (tick-borne) in summer

Answers

Q1 Q2 Q3 Q4 Q5

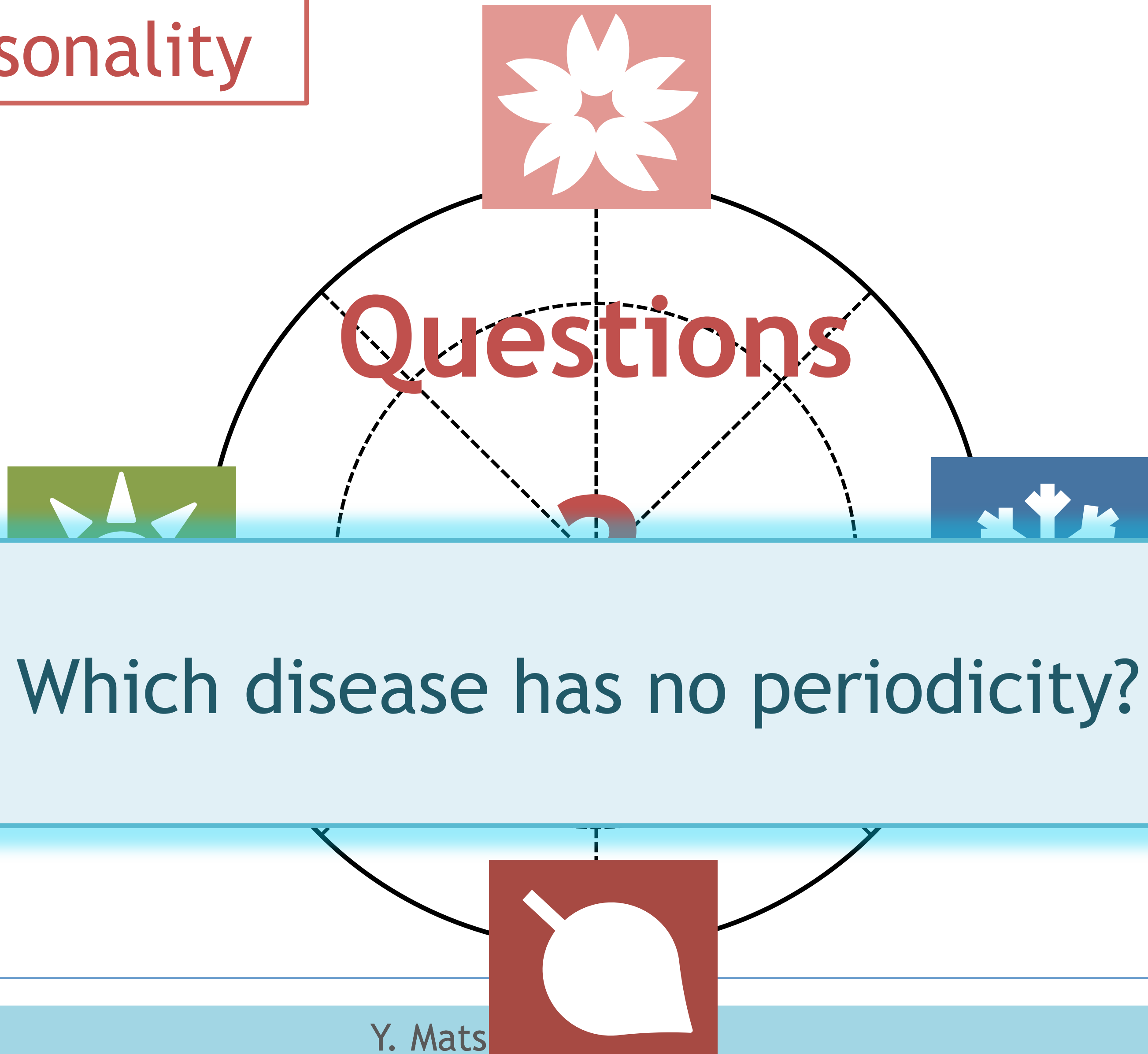
P1 Seasonality



Answers

Q1 Q2 Q3 Q4 Q5

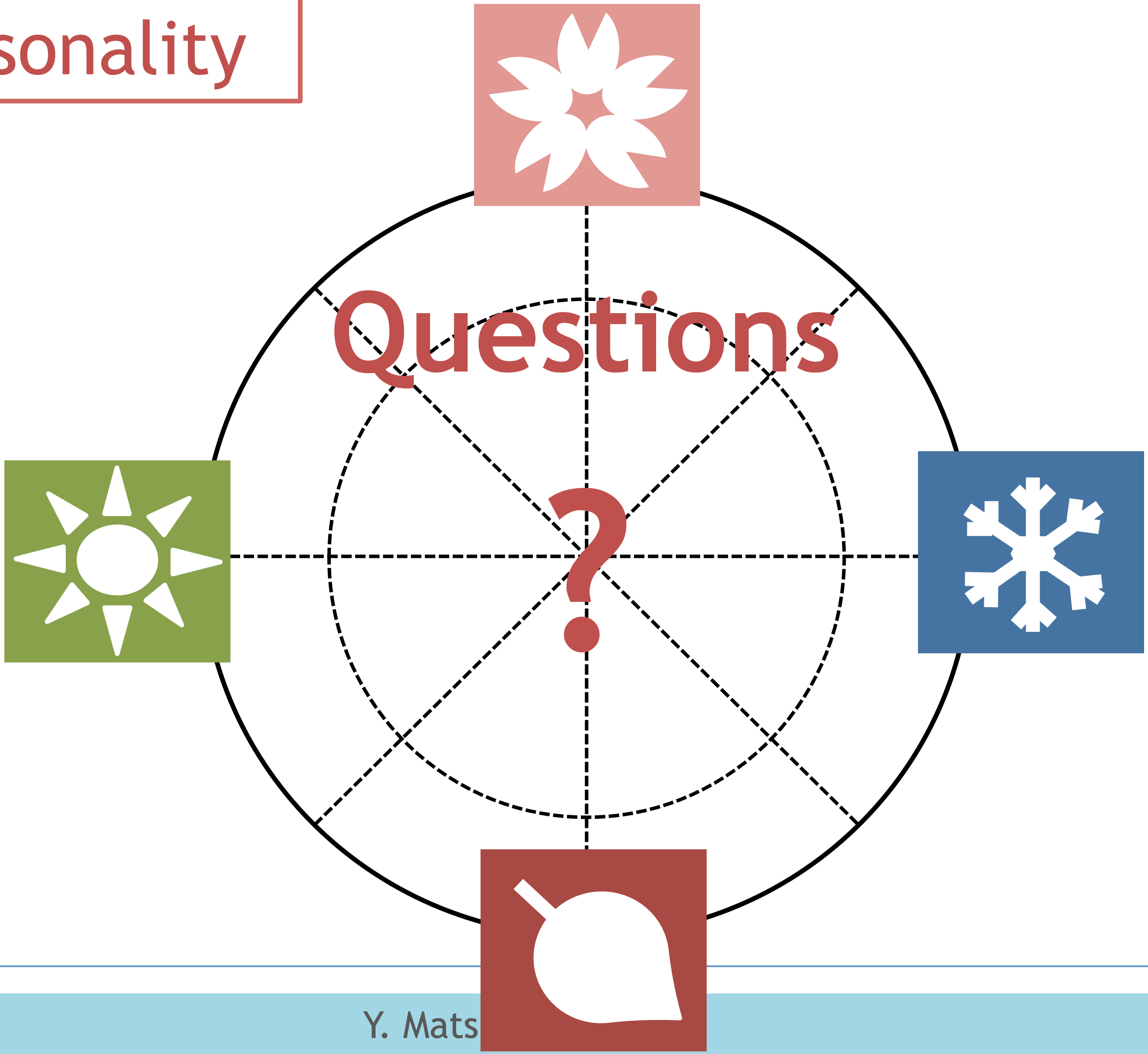
P1 Seasonality



Q: Which disease has no periodicity?

Answers

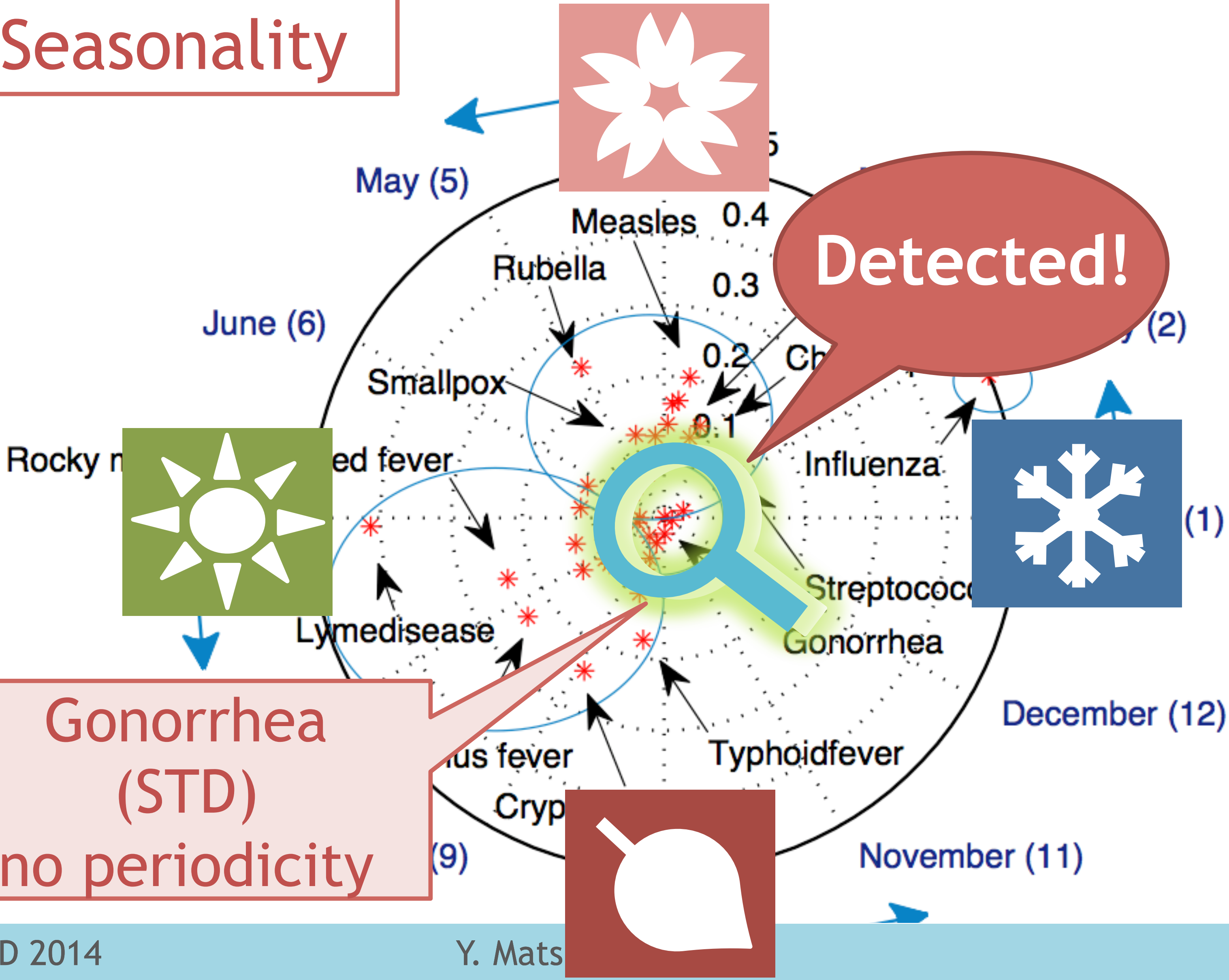
P1 Seasonality



Answers

Q1 Q2 Q3 Q4 Q5

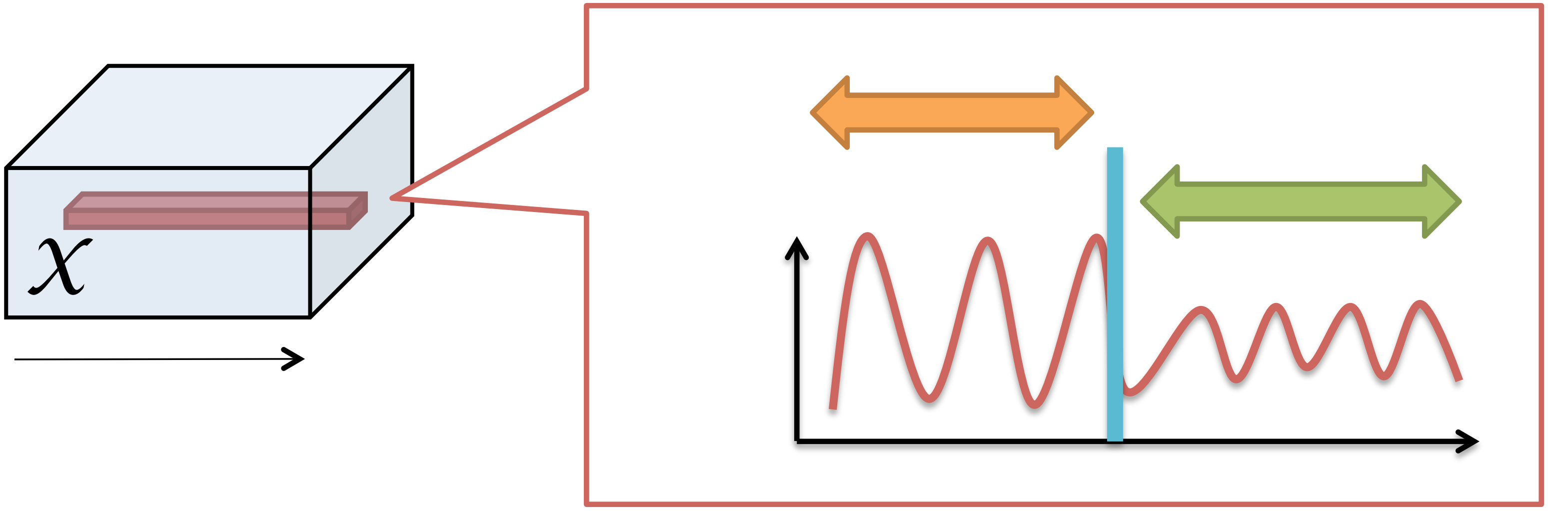
P1 Seasonality



Gonorrhea (STD)
no periodicity

Q2

Can we see any <u>discontinuities</u>?

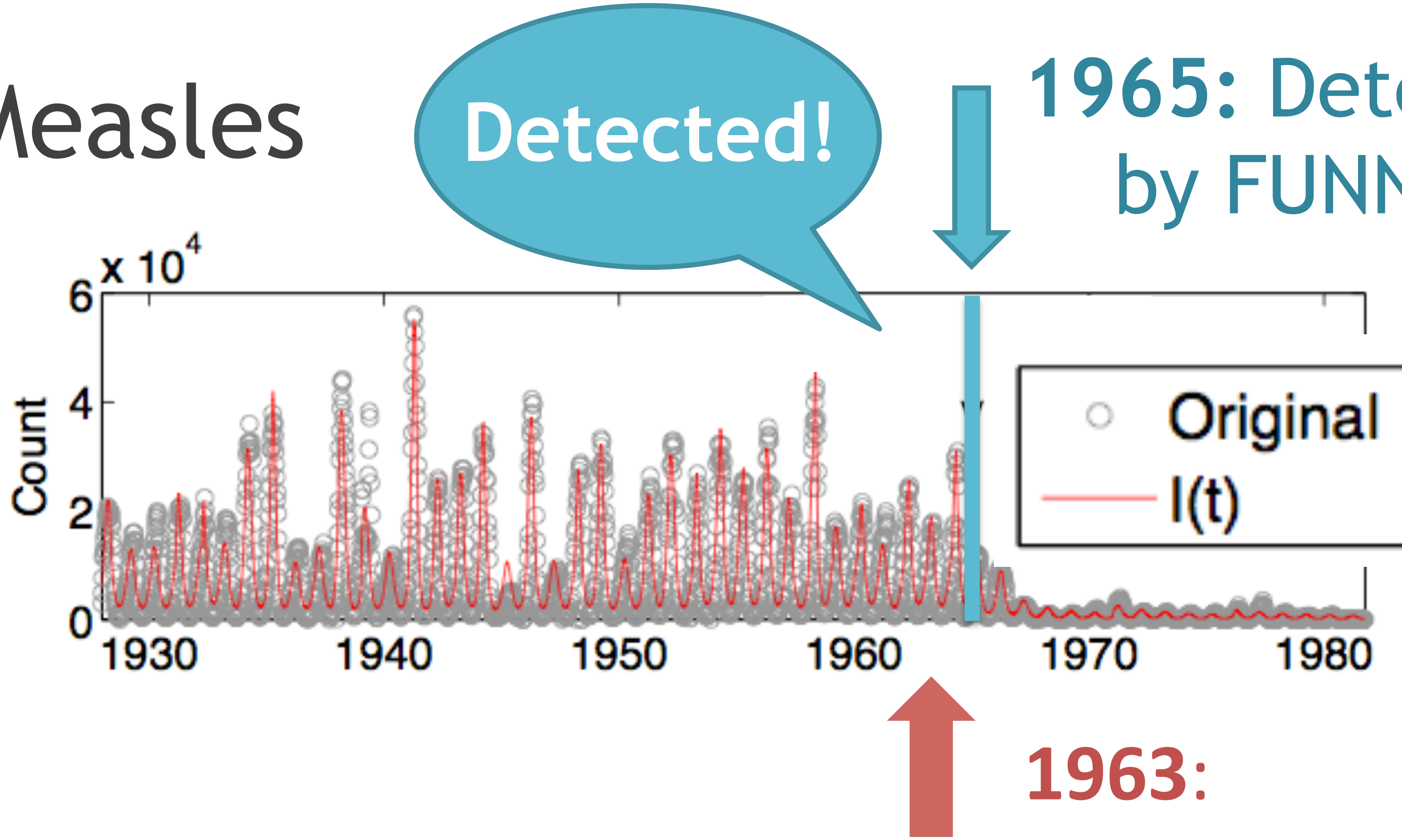


Answers

Q1 Q2 Q3 Q4 Q5

P2 Disease reduction effect

Measles



1965: Detected by FUNNEL

1963: Vaccine licensure

Questions

Q1 Q2 Q3 Q4 Q5

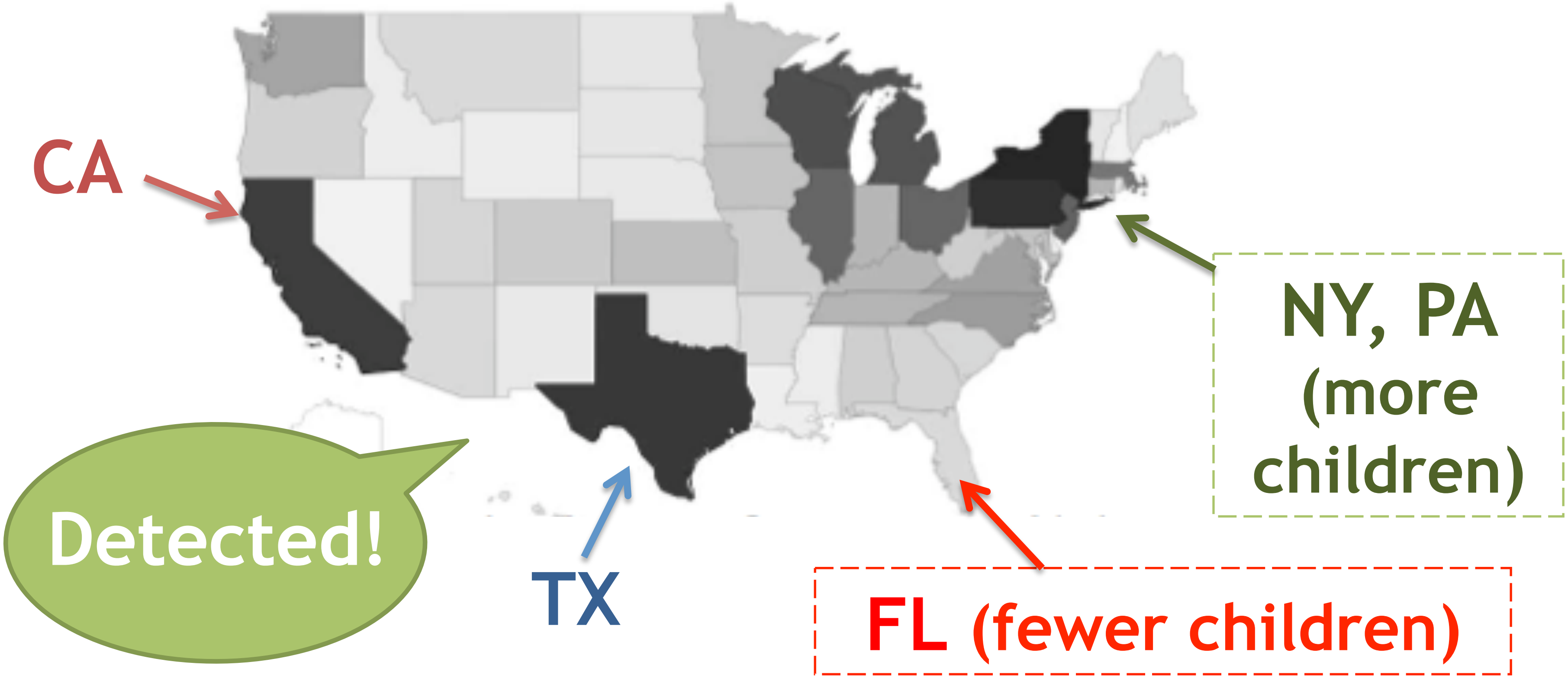
Q3

What's the <u>difference</u> between measles in <u>**NY**</u> and in <u>**FL**</u>?

Answers

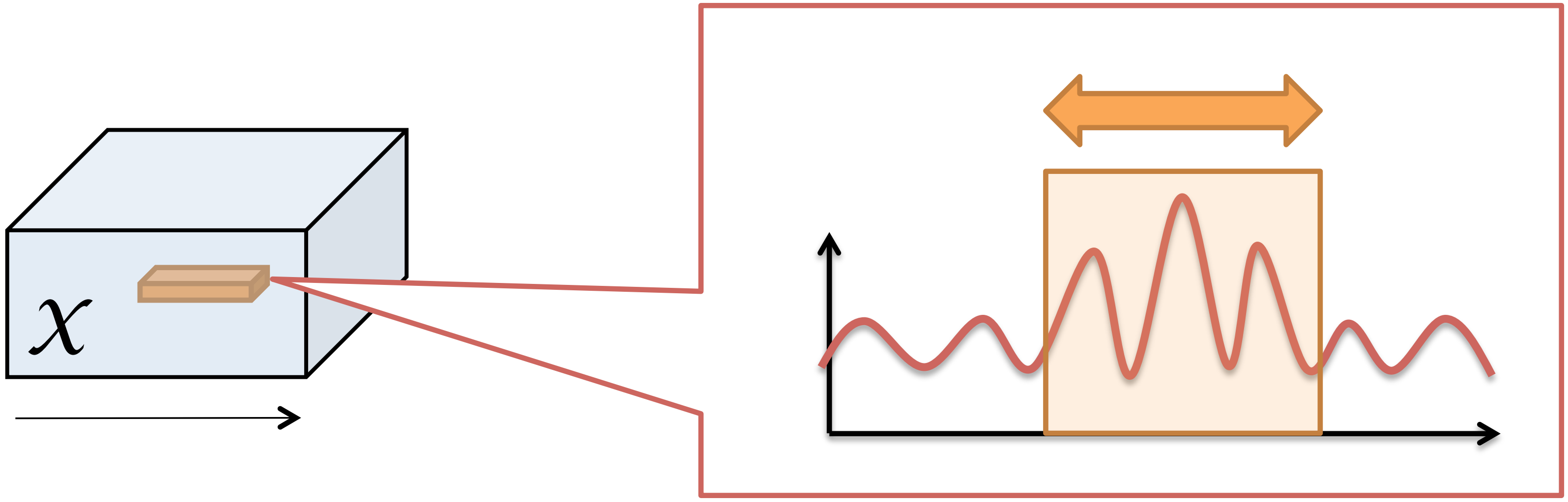
P3 area sensitivity

FUNNEL's guess of susceptibles (measles)



Q4

Are there any external shock events, like wars?

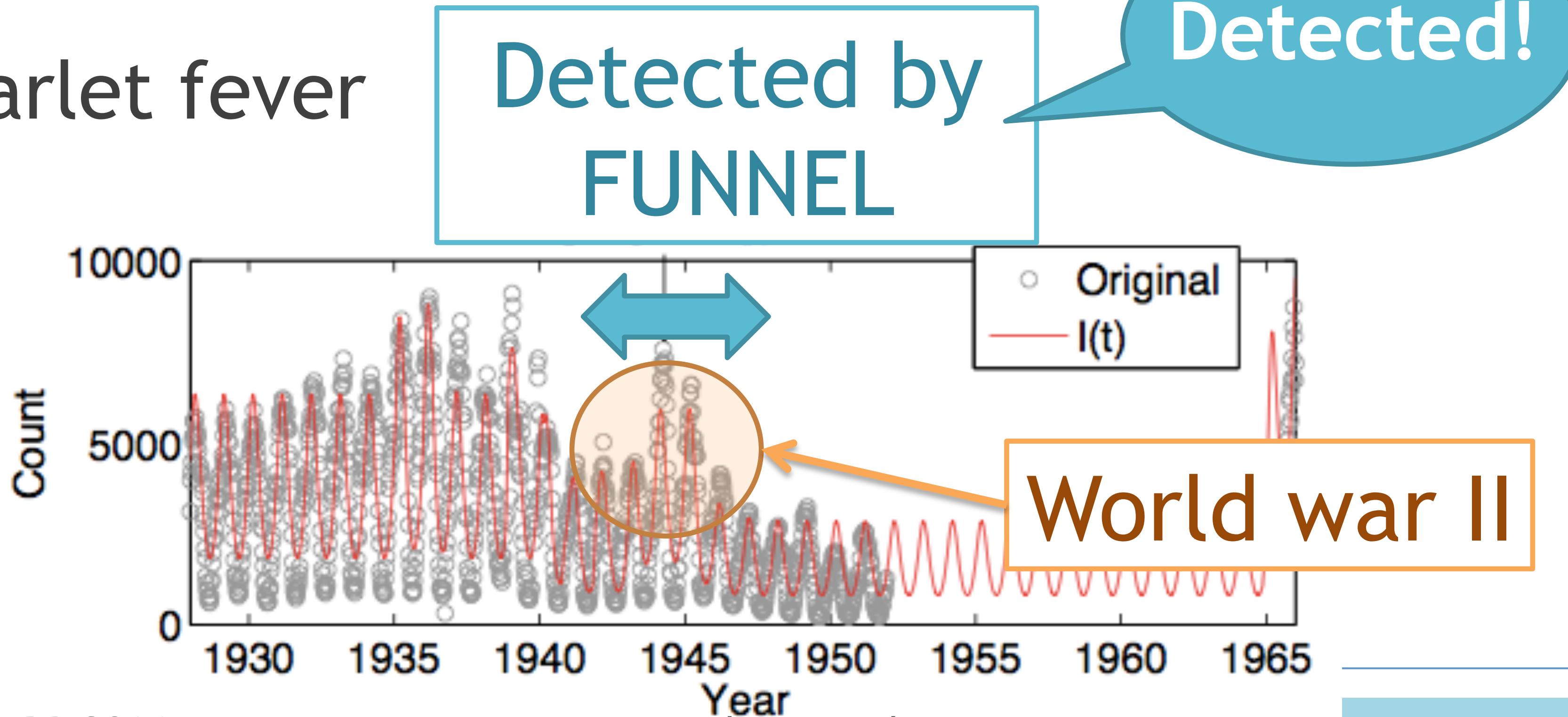


Answers

P4 external shock events

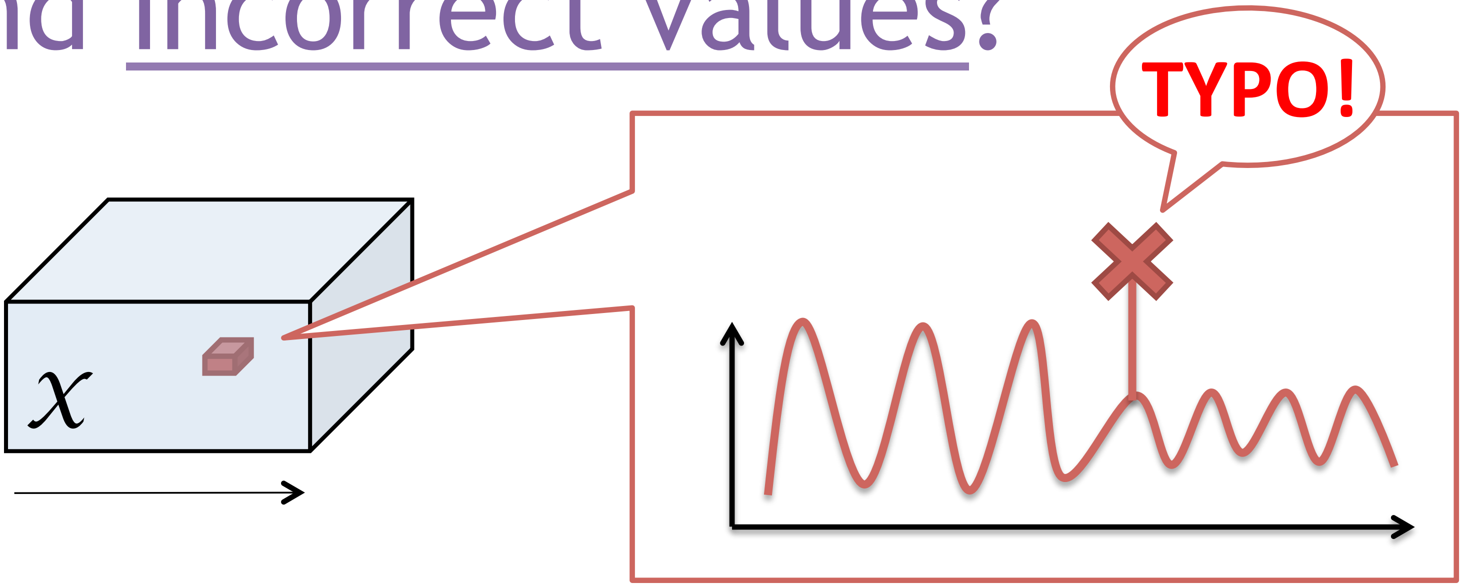
Funnel can detect external shocks "**fully-automatically**" !

Scarlet fever



Q5

How can we remove mistakes and incorrect values?



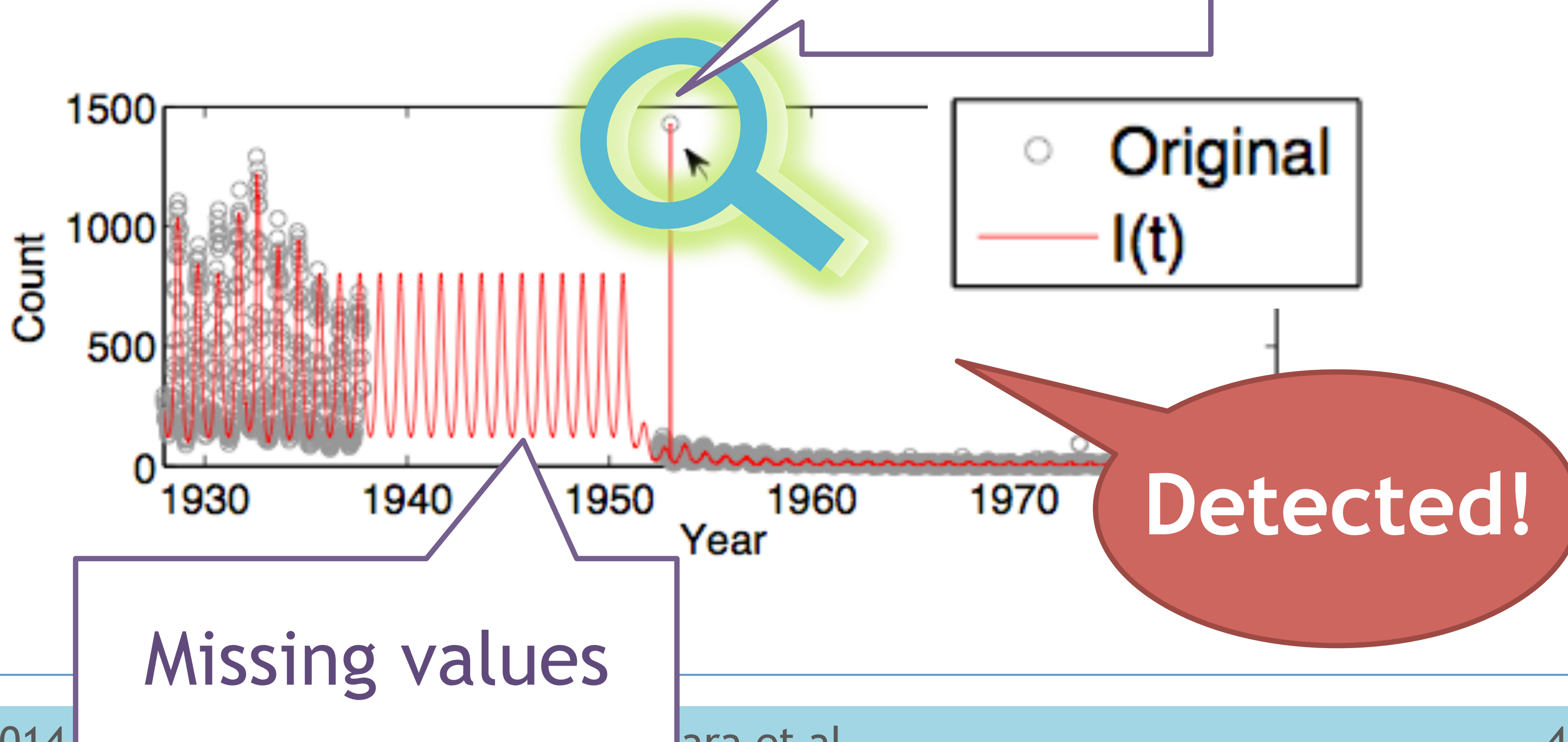
Answers

Q1 Q2 Q3 Q4 Q5

P5 mistakes

It can also detect typos, “**automatically**” !!

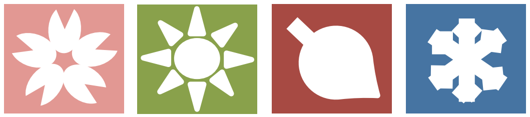
Typhoid fever cases



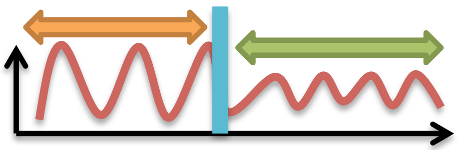
Modeling power of FUNNEL

Our model can capture 5 properties

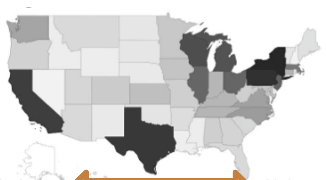
P1 Seasonality



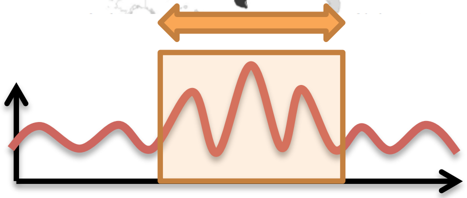
P2 Disease reductions



P3 Area sensitivity



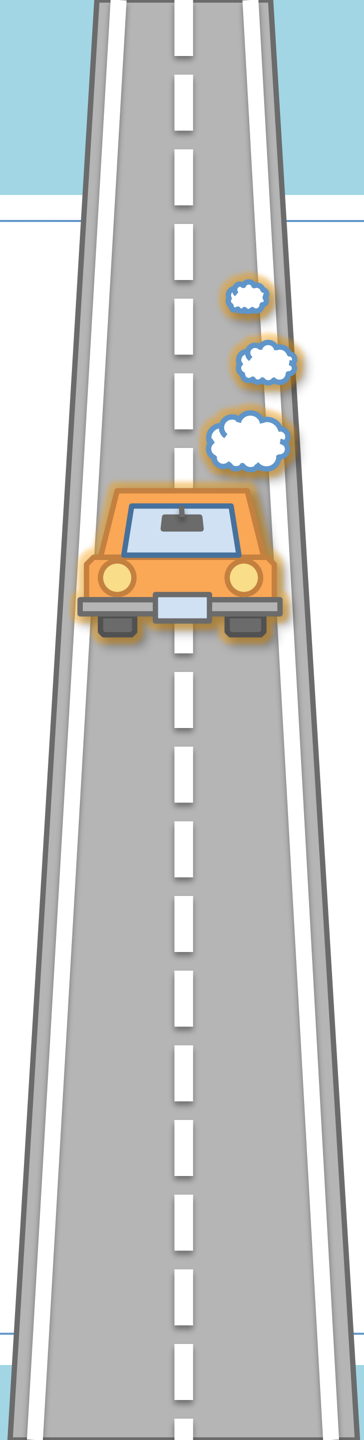
P4 External events



P5 Mistakes

Roadmap

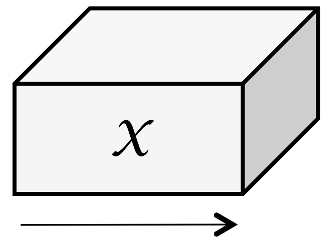
- ✔ Motivation
- ✔ Modeling power of FUNNEL
- Overview - main ideas
- Proposed model - idea #1
- Algorithm - idea #2
- Experiments
- Discussion
- Conclusions



Problem definition

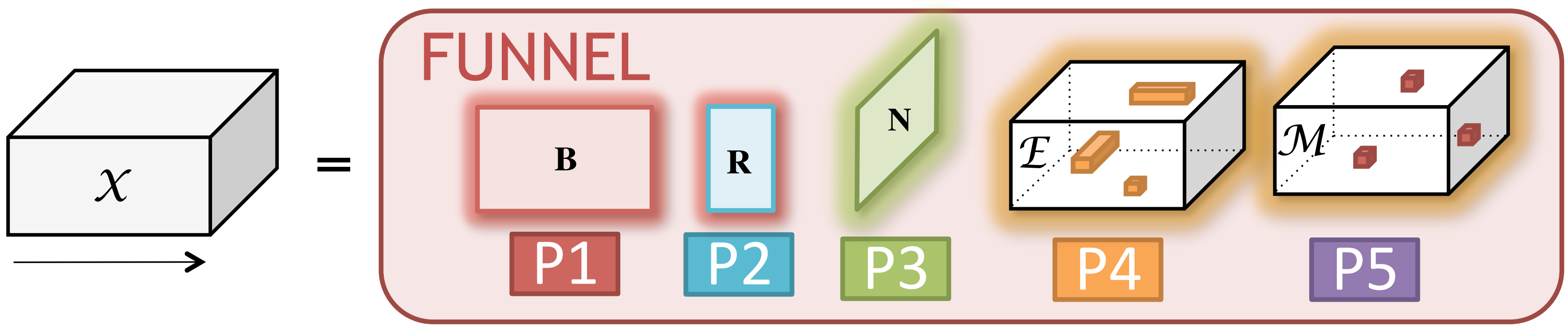
Given:

Tensor $\mathcal{X}$ (disease x state x time)



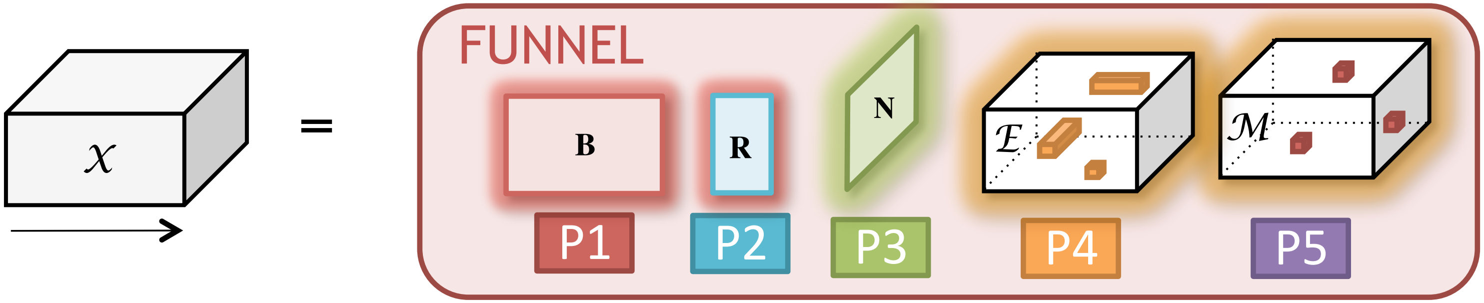
Find:

Compact description of $\mathcal{X}$, ***"automatically"***



Two main ideas

Idea #1: Grey-box model

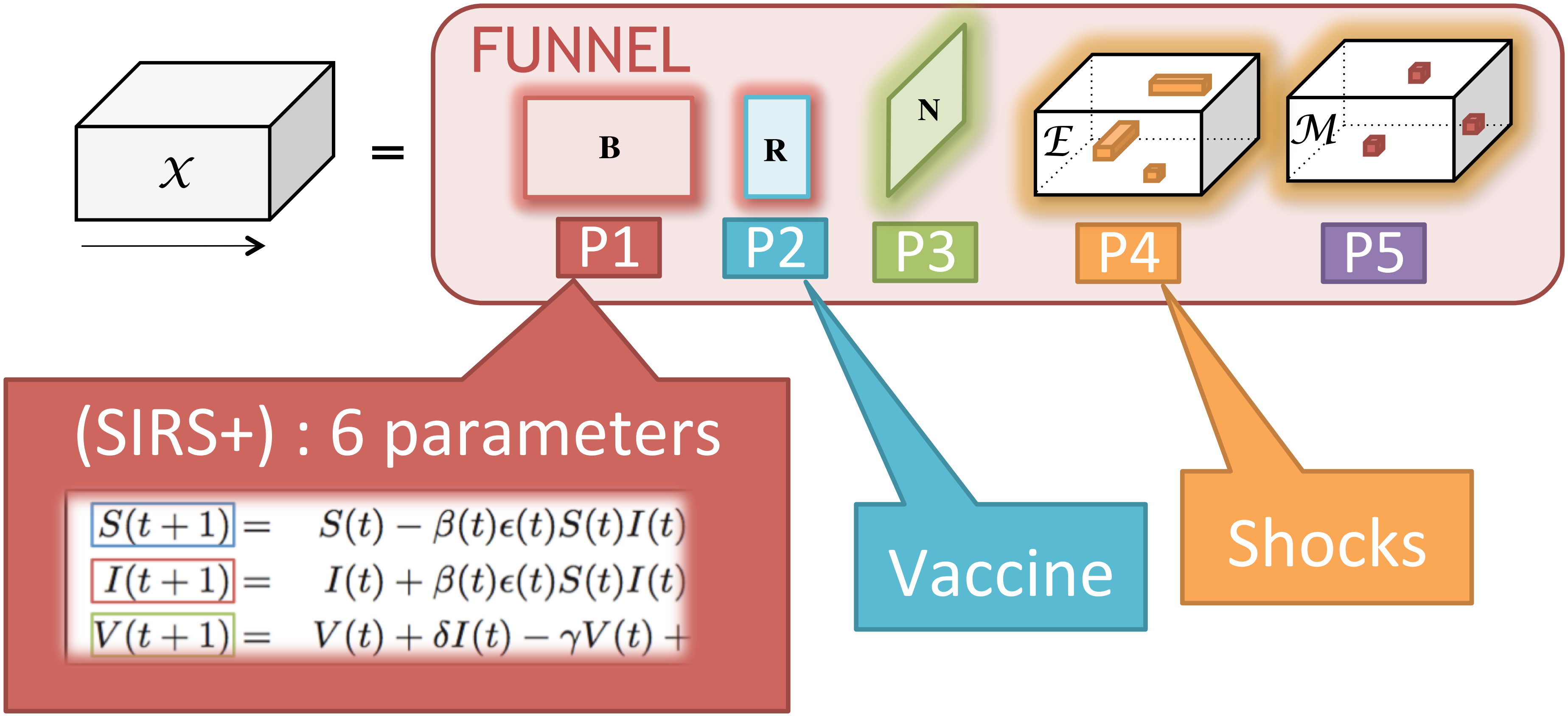


Idea #2: MDL for fitting

NO magic numbers !
(parameter-free)

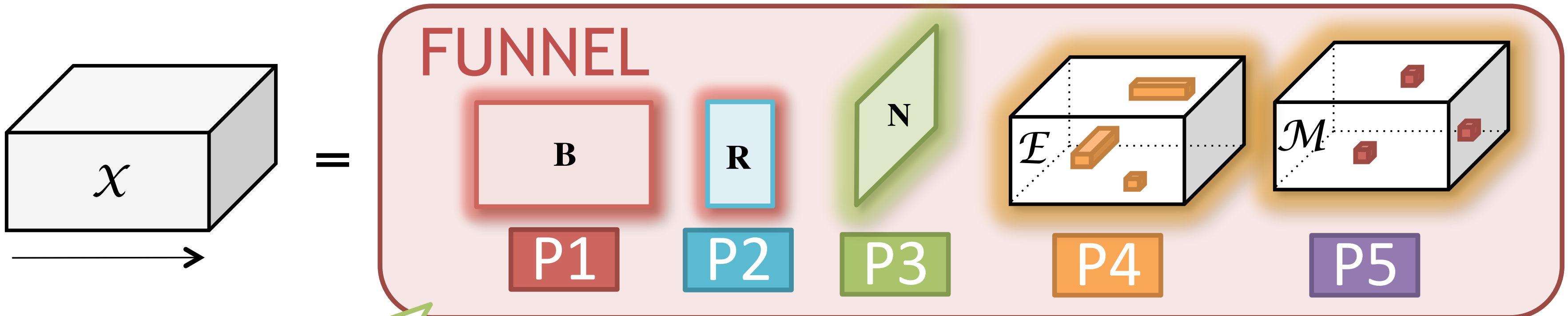
Two main ideas

Idea #1: Grey-box model - domain knowledge



Two main ideas

Idea #2: Fitting with MDL -> parameter free!



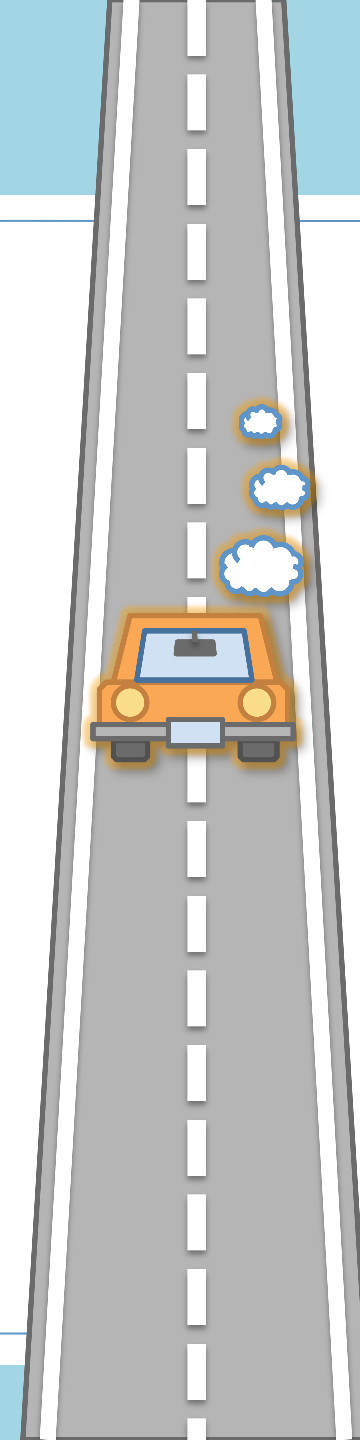
$$Cost_T(\mathcal{X};\mathcal{F}) = \log^*(d) + \log^*(l) + \log^*(n) + Cost_M(\mathbf{B}) + Cost_M(\mathbf{R}) + Cost_M(\mathbf{N}) + Cost_M(\mathcal{E}) + Cost_M(\mathcal{M}) + Cost_C(\mathcal{X}|\mathcal{F})$$

NO magic numbers

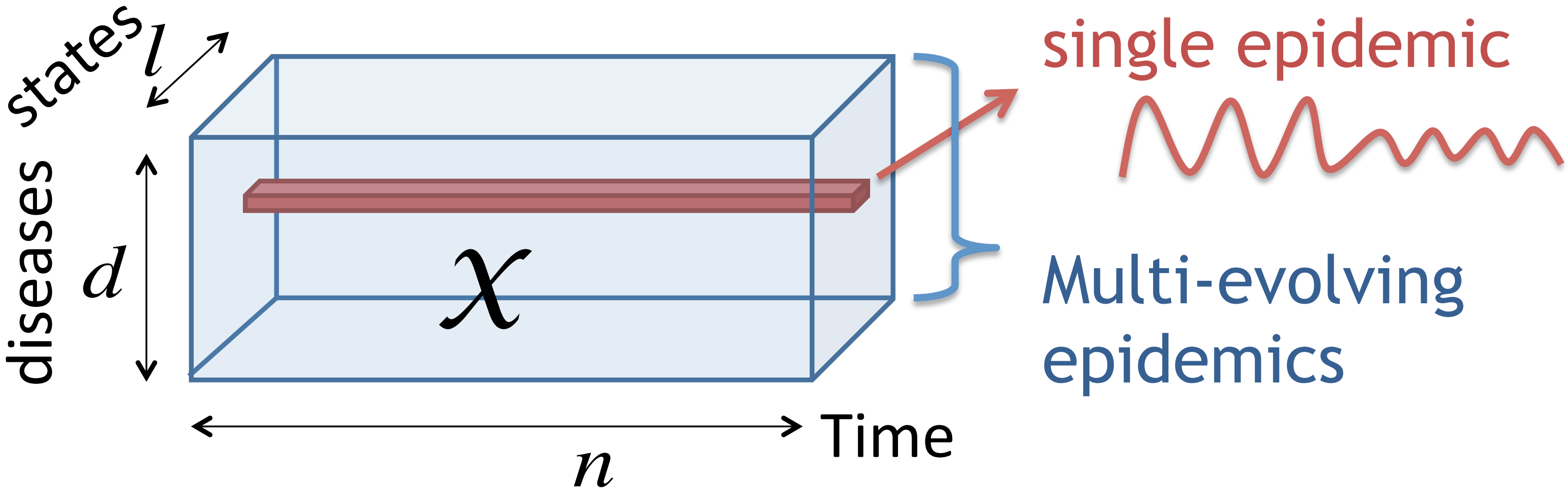
Parameter-free!

Roadmap

- ✔ Motivation
- ✔ Modeling power of FUNNEL
- ✔ Overview - main ideas
- Proposed model - idea #1
- Algorithm - idea #2
- Experiments
- Discussion
- Conclusions

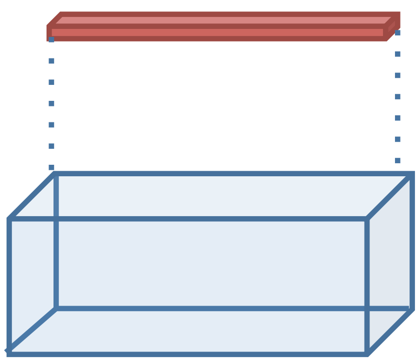


Proposed model: FUNNEL

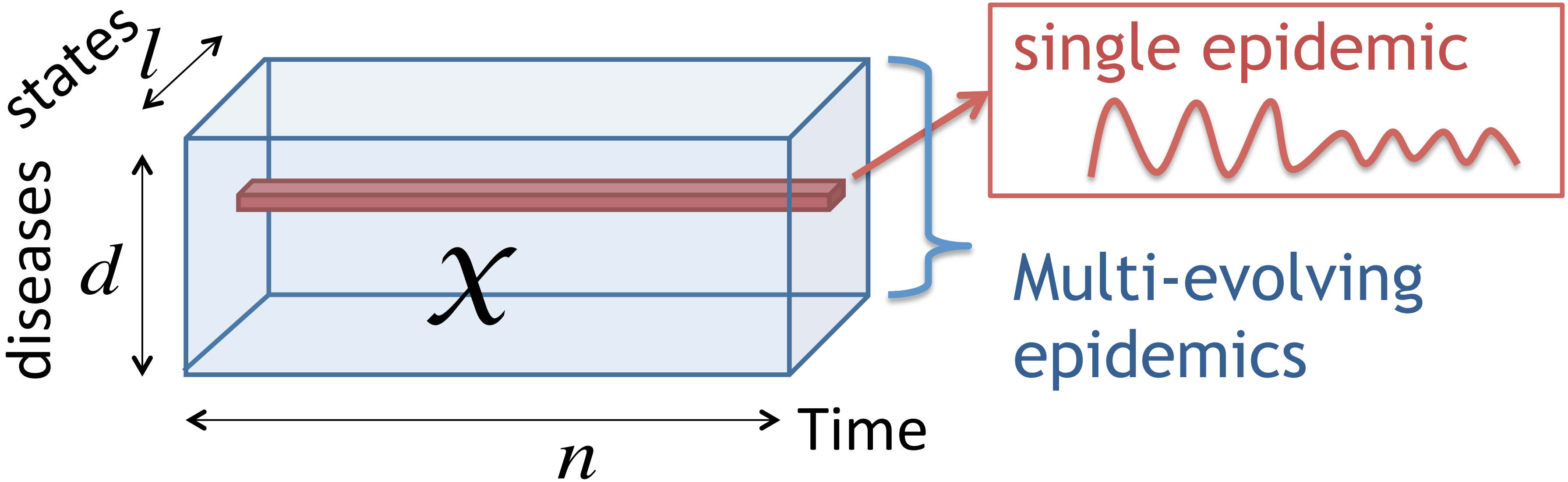


(a) FUNNEL-single

(b) FUNNEL-full

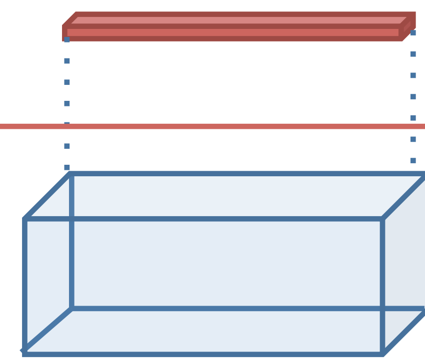


Proposed model: FUNNEL



(a) FUNNEL-single

(b) FUNNEL-full

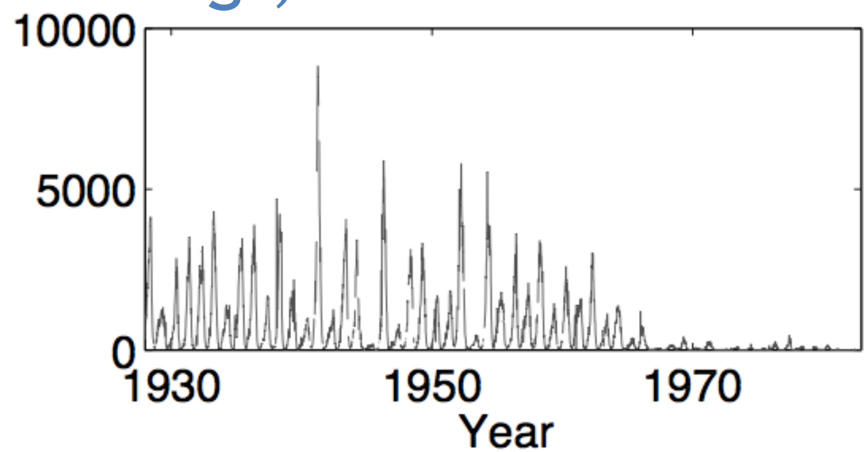


FUNNEL - with a single epidemic

Given:

"single" epidemic sequence

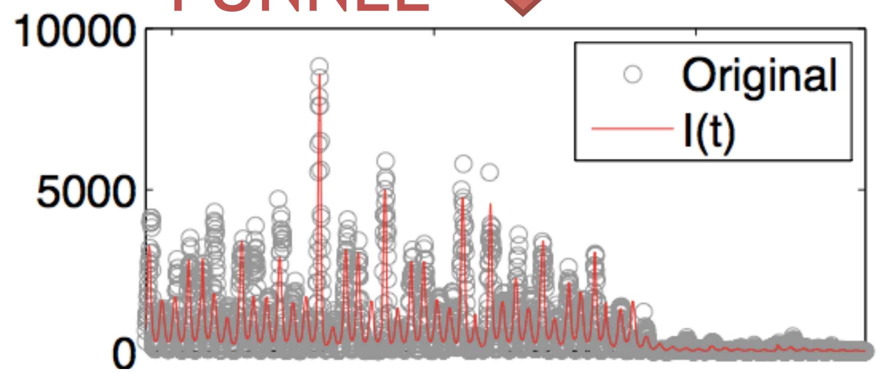
e.g., measles in NY



Find:

nonlinear equation, model parameters

FUNNEL



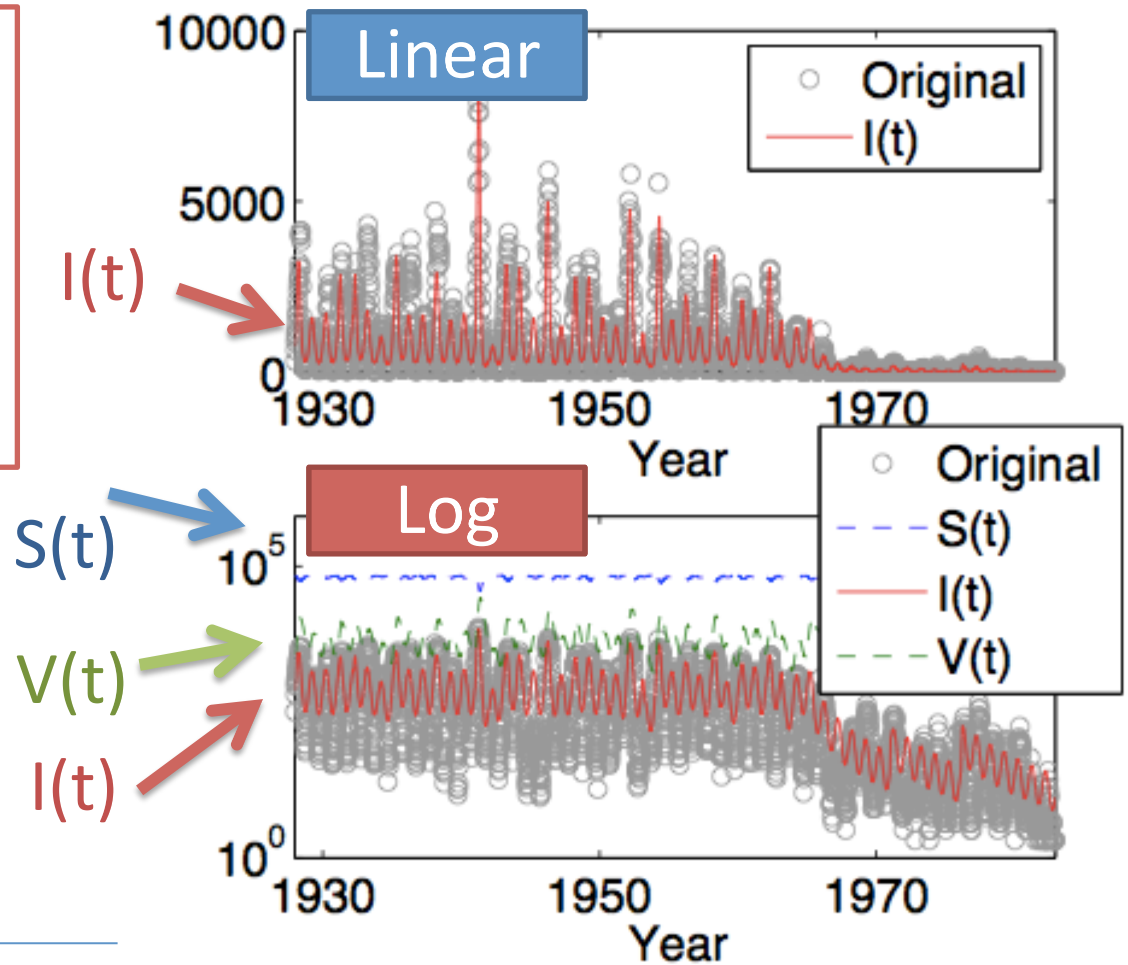
FUNNEL - with a single epidemic

Details

With a single epidemic: Funnel-RE

People of 3 classes

- **S** : Susceptible
- **I** : Infected
- **V** : Vigilant/ vaccinated



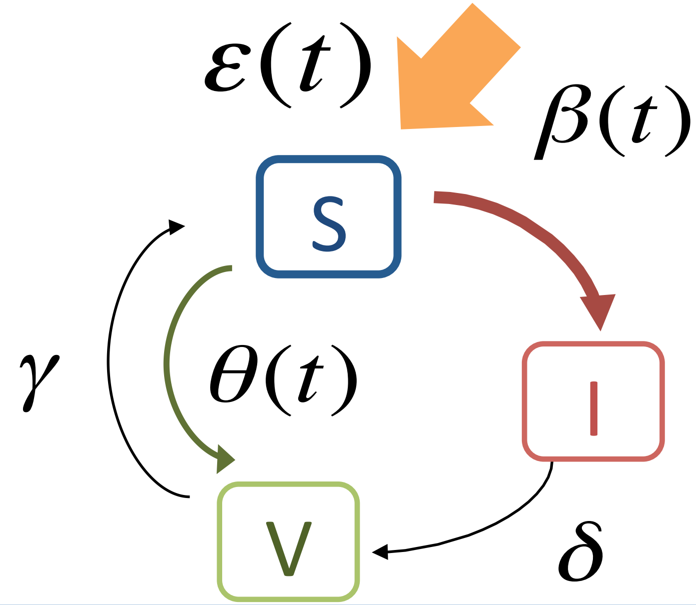
FUNNEL - with a single epidemic

Details

With a single epidemic: Funnel-RE

$$
\begin{aligned}
S(t+1) &= S(t) - \beta(t)\epsilon(t)S(t)I(t) + \gamma V(t) - \theta(t)S(t) \\
I(t+1) &= I(t) + \beta(t)\epsilon(t)S(t)I(t) - \delta I(t) \\
V(t+1) &= V(t) + \delta I(t) - \gamma V(t) + \theta(t)S(t) \qquad (3)
\end{aligned}
$$

S(t) : susceptible
I (t) : Infected
V(t) : Vigilant /Vaccinated



FUNNEL - with a single epidemic

Details

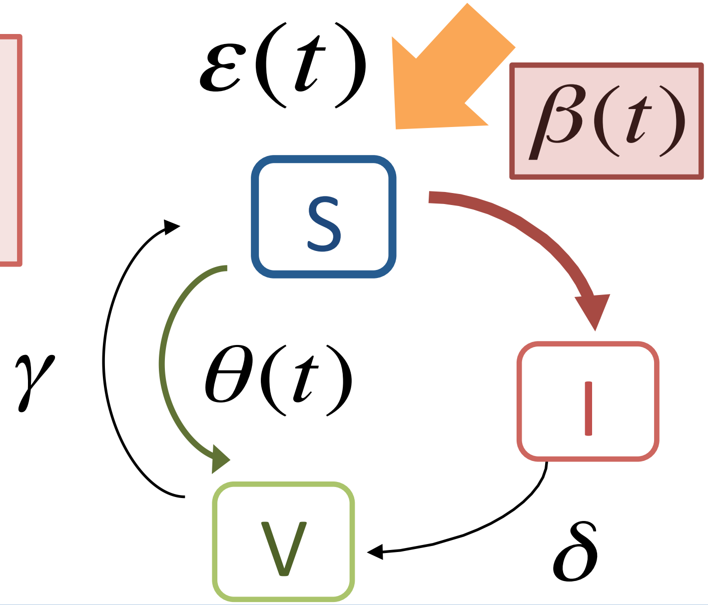
With a single epidemic: Funnel-RE

$$S(t+1) = S(t) - \beta(t)\epsilon(t)S(t)I(t) + \gamma V(t) - \theta(t)S(t)$$
$$I(t+1) = I(t) + \beta(t)\epsilon(t)S(t)I(t) - \delta I(t)$$
$$V(t+1) = V(t) + \delta I(t) - \gamma V(t) + \theta(t)S(t) \quad (3)$$

$\beta(t)$: strength of infection
(yearly periodic func)

$$\beta(t) = \beta_0 \cdot \left(1 + P_a \cdot cos\left(\frac{2\pi}{P_p}(t + P_s)\right)\right)$$

$$P_p = 52$$



FUNNEL - with a single epidemic

Details

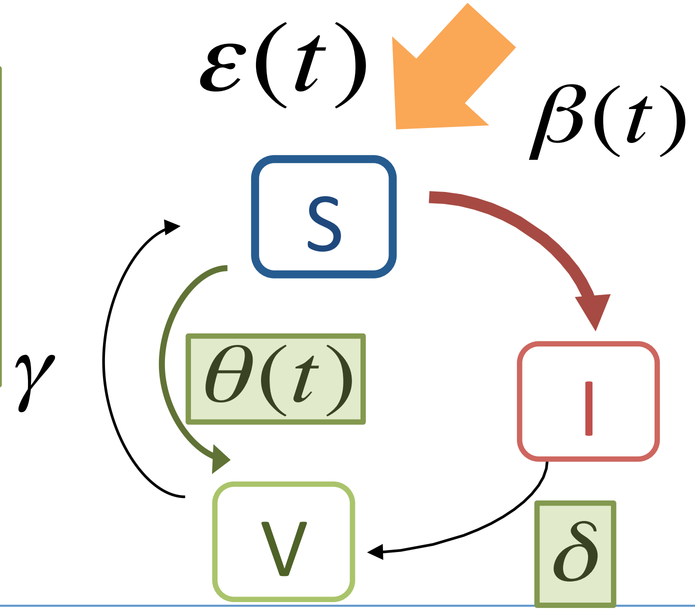
With a single epidemic: Funnel-RE

$$\begin{aligned} S(t+1) &= S(t) - \beta(t)\epsilon(t)S(t)I(t) + \gamma V(t) - \theta(t)S(t) \\ I(t+1) &= I(t) + \beta(t)\epsilon(t)S(t)I(t) - \delta I(t) \\ V(t+1) &= V(t) + \delta I(t) - \gamma V(t) + \theta(t)S(t) \qquad (3) \end{aligned}$$

δ : healing rate

$\theta(t)$: disease reduction effect

$$\theta(t) = \begin{cases} 0 & (t < t_\theta) \\ \theta_0 & (t \geq t_\theta) \end{cases}$$



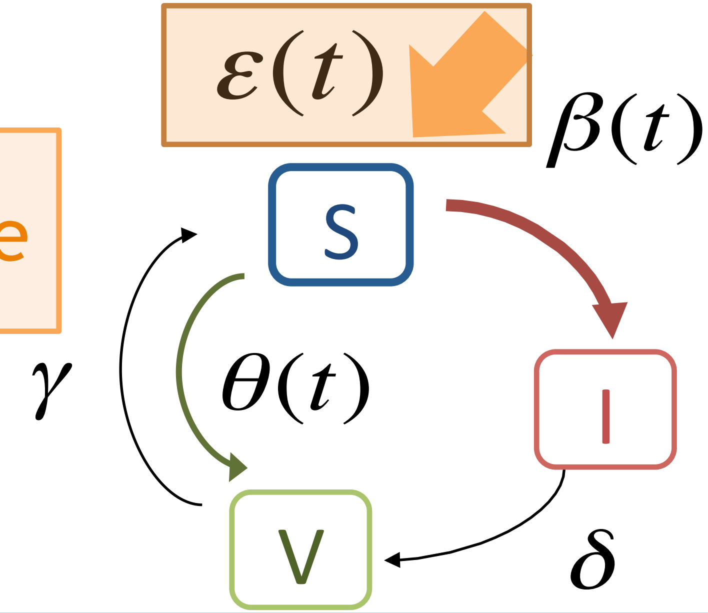
FUNNEL - with a single epidemic

Details

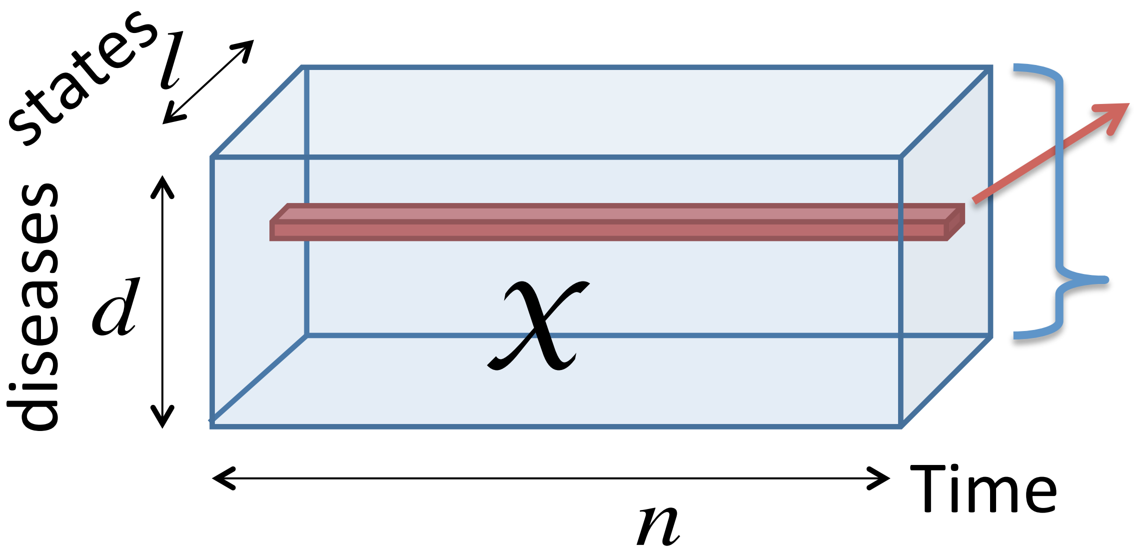
With a single epidemic: Funnel-RE

$$
\begin{aligned}
S(t+1) &= S(t) - \beta(t)\epsilon(t)S(t)I(t) + \gamma V(t) - \theta(t)S(t) \\
I(t+1) &= I(t) + \beta(t)\epsilon(t)S(t)I(t) - \delta I(t) \\
V(t+1) &= V(t) + \delta I(t) - \gamma V(t) + \theta(t)S(t) \qquad (3)
\end{aligned}
$$

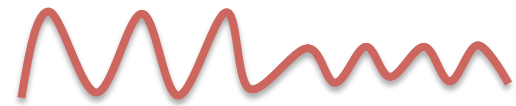
$\varepsilon(t)$: temporal susceptible rate



Proposed model: FUNNEL



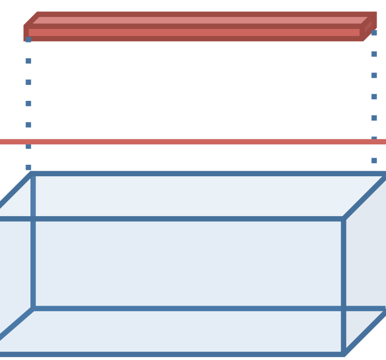
single epidemic



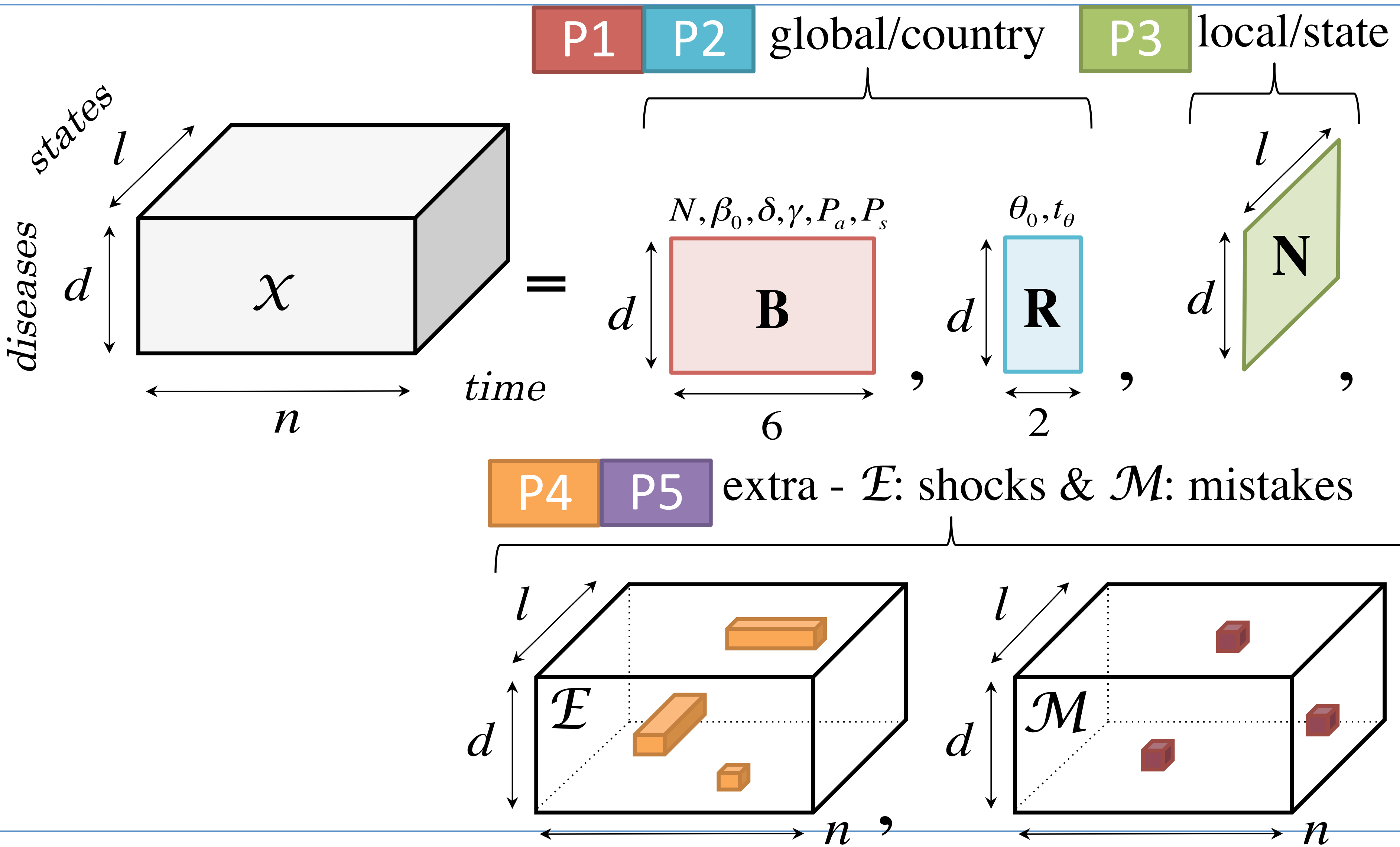
Multi-evolving epidemics

(a) FUNNEL-single

(b) FUNNEL-full

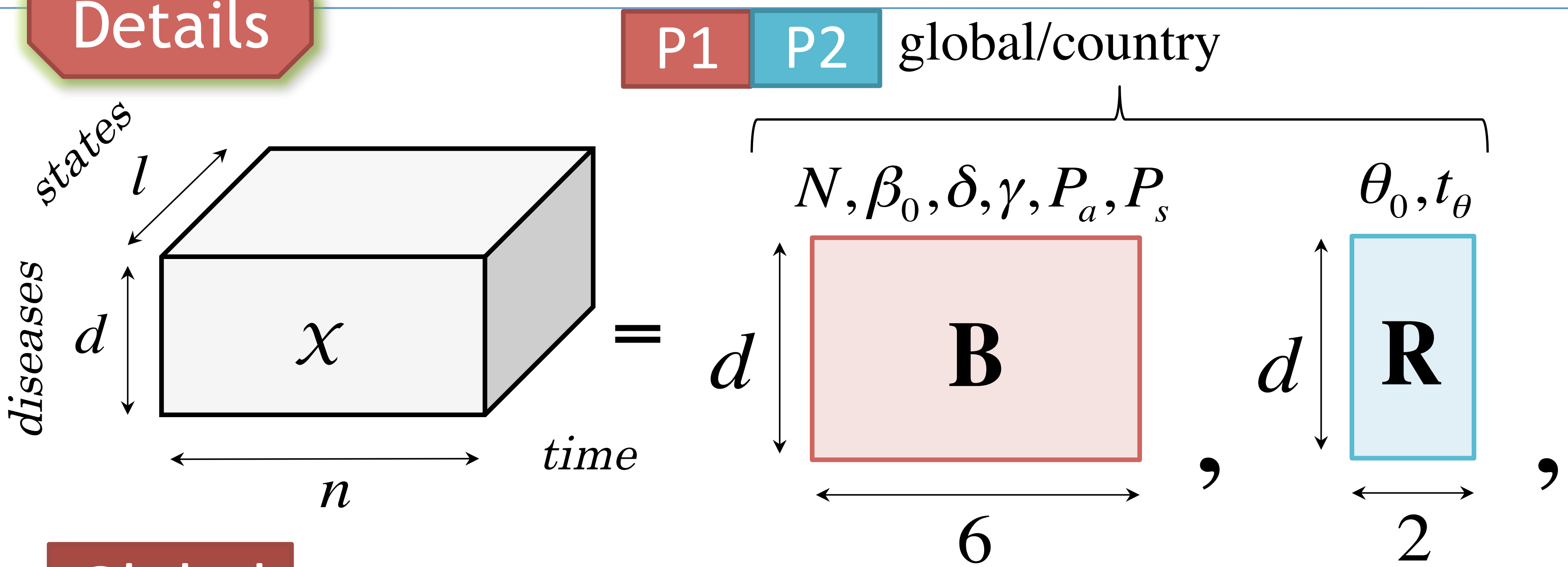


Proposed model: FUNNEL-full



Proposed model: FUNNEL-full

Details

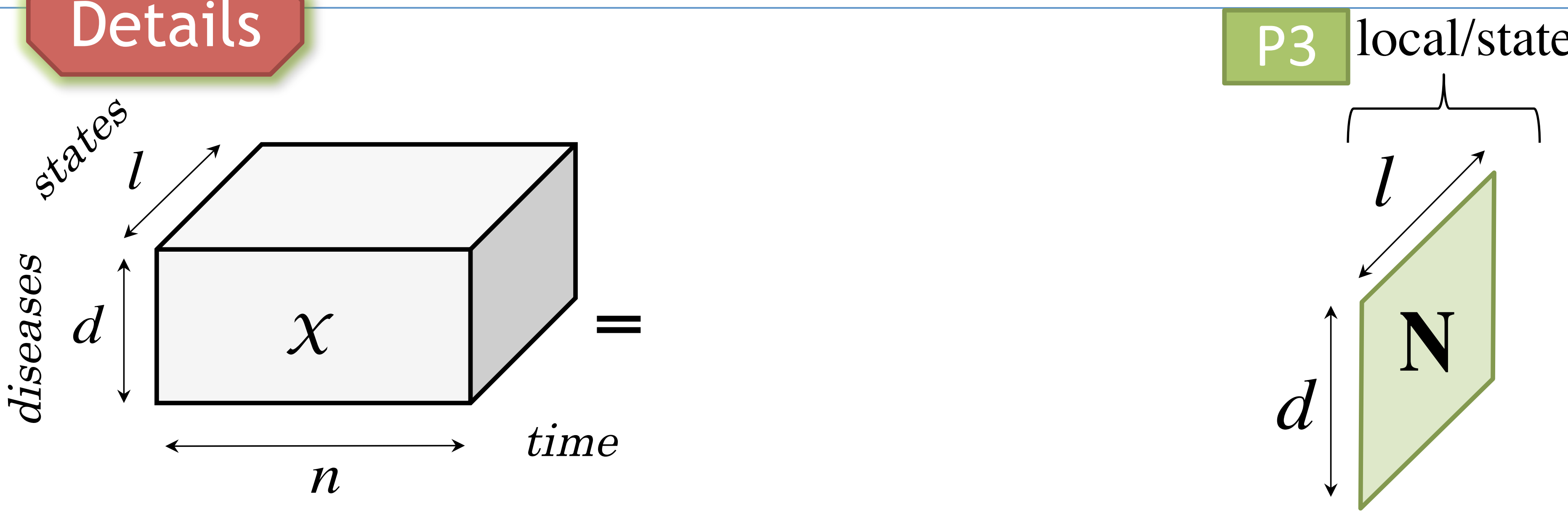


Global

- P1 Base matrix **B** (d x 6)
- P2 Disease reduction matrix **R** (d x 2)

Proposed model: FUNNEL-full

Details



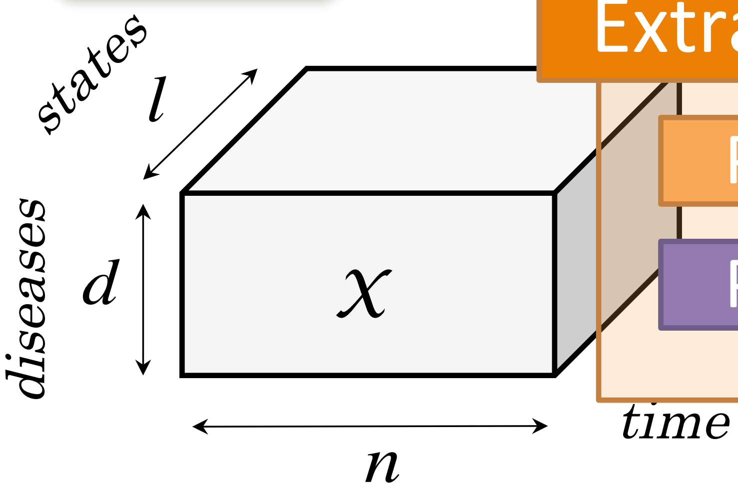
Local

P3 Geo-disease matrix $\mathbf{N}$ (d x l)

$\mathbf{N} = \{N_{ij}\}_{i,j=1}^{d,l}$: potential population of disease i in state j

Proposed model: FUNNEL-full

Details

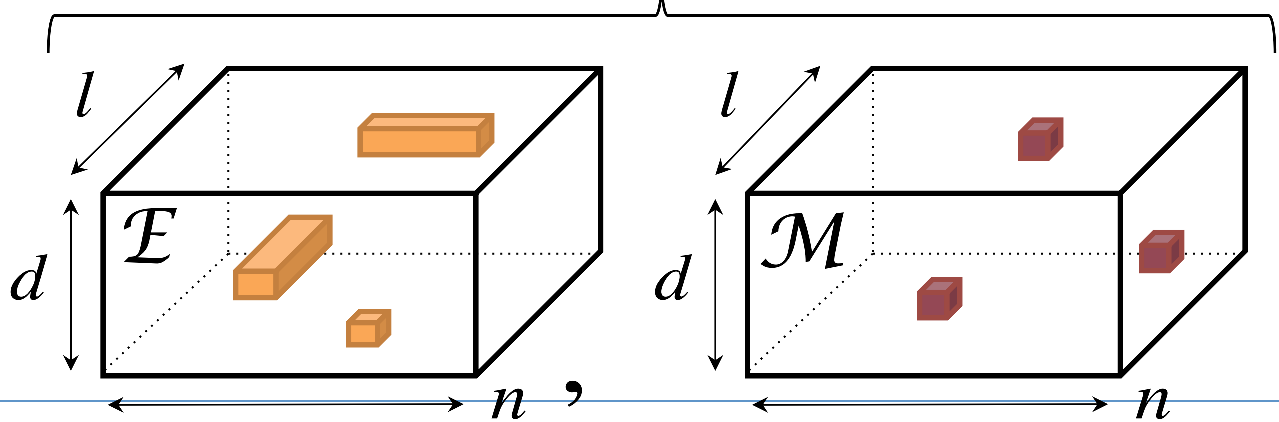


Extra

P4 External shock tensor $\mathcal{E}$

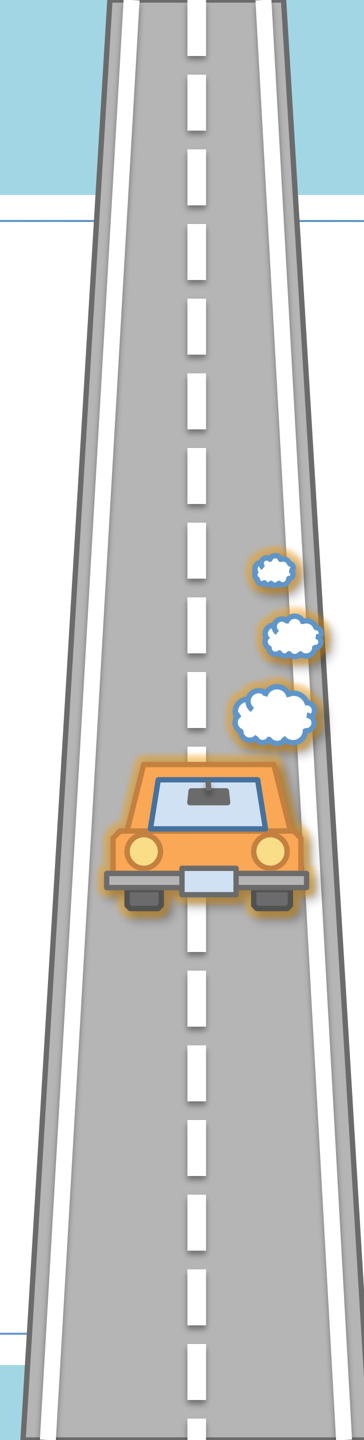
P5 Mistake tensor $\mathcal{M}$

P4 P5 extra - $\mathcal{E}$: shocks & $\mathcal{M}$: mistakes



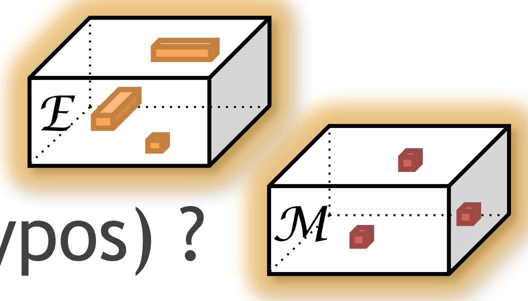
Roadmap

- ✔ Motivation
- ✔ Modeling power of FUNNEL
- ✔ Overview – main ideas
- ✔ Proposed model – idea #1
- **Algorithm – idea #2**
- Experiments
- Discussion
- Conclusions

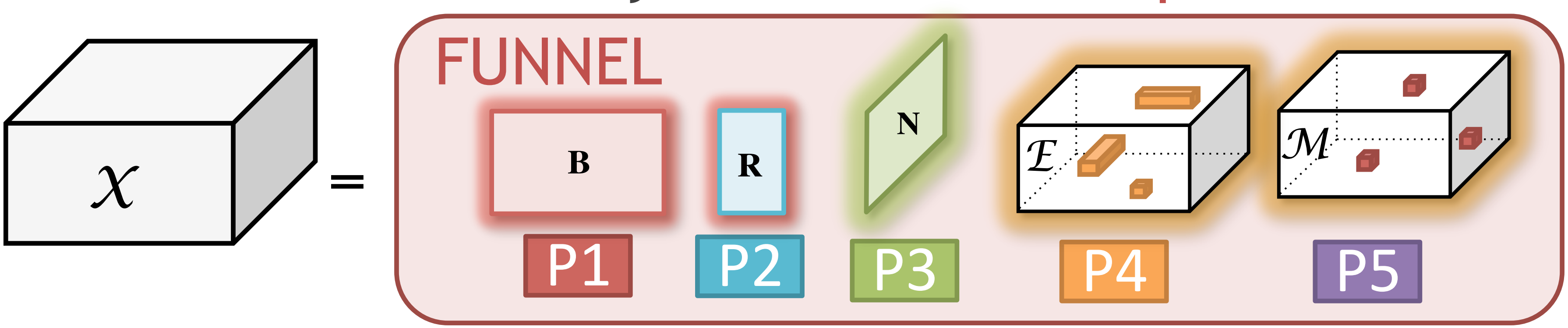


Challenges

Q1. How to automatically
- find "external shocks" ?
- ignore "mistakes" (i.e., typos) ?

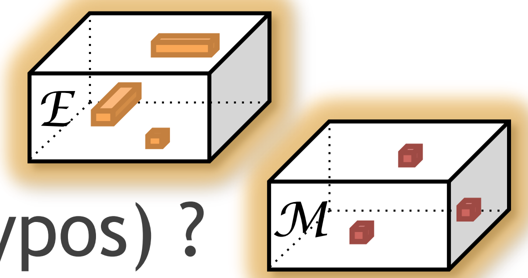


Q2. How to efficiently estimate model parameters ?



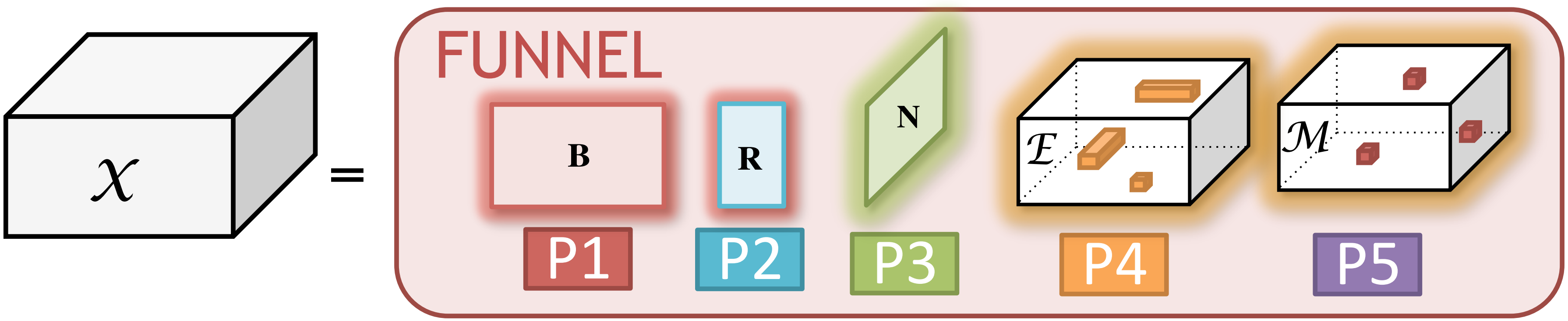
Challenges

Q1. How to automatically
- find "external shocks" ?
- ignore "mistakes" (i.e., typos) ?



Idea (1) : Model description cost

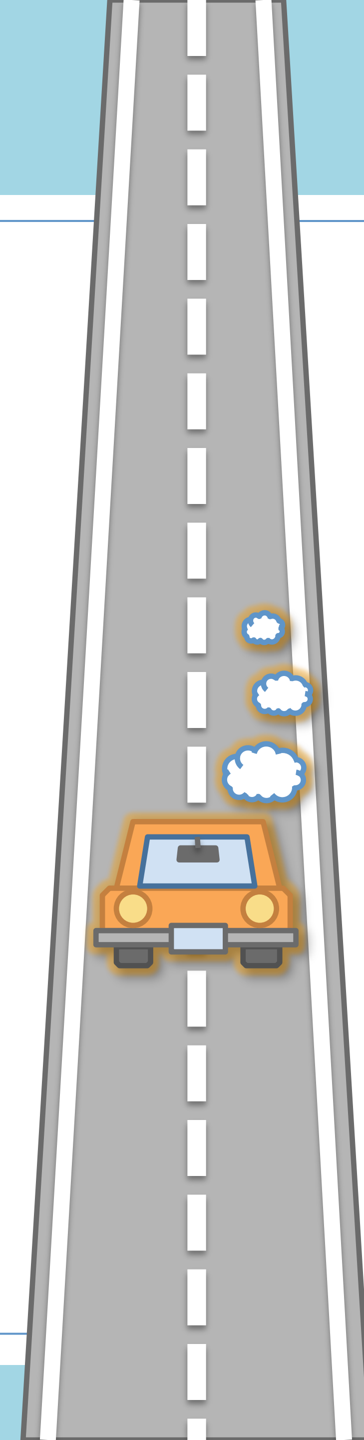
Q2. How to efficiently estimate model parameters ?



Idea (2): Multi-layer optimization (linear)

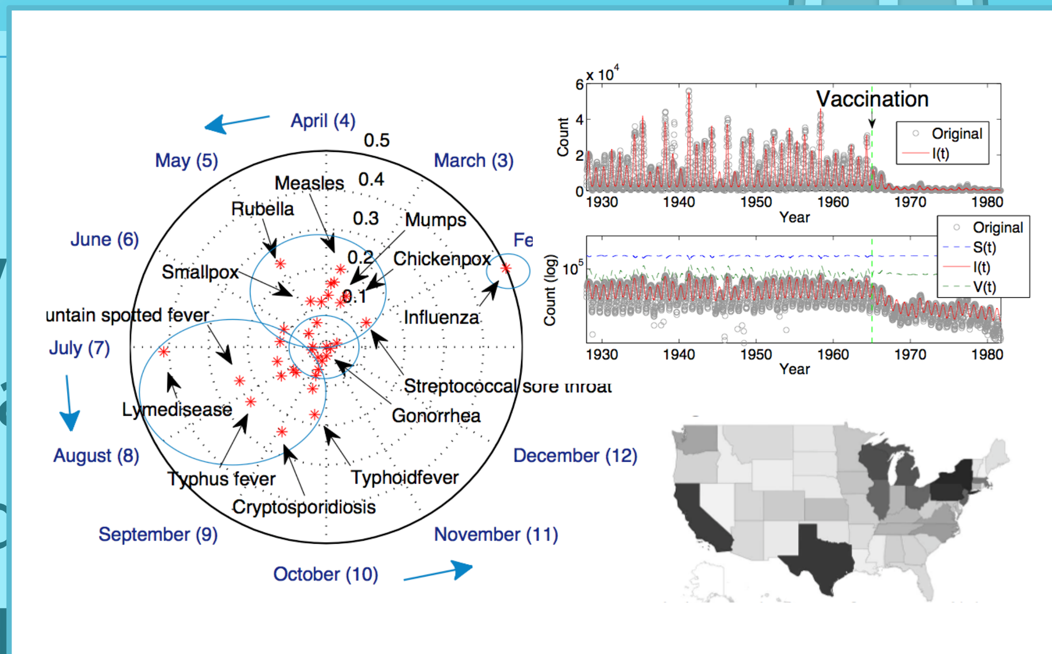
Roadmap

- ✔ Motivation
- ✔ Modeling power of FUNNEL
- ✔ Overview – main ideas
- ✔ Proposed model – idea #1
- ✔ Algorithm – idea #2
- Experiments
- Discussion
- Conclusions



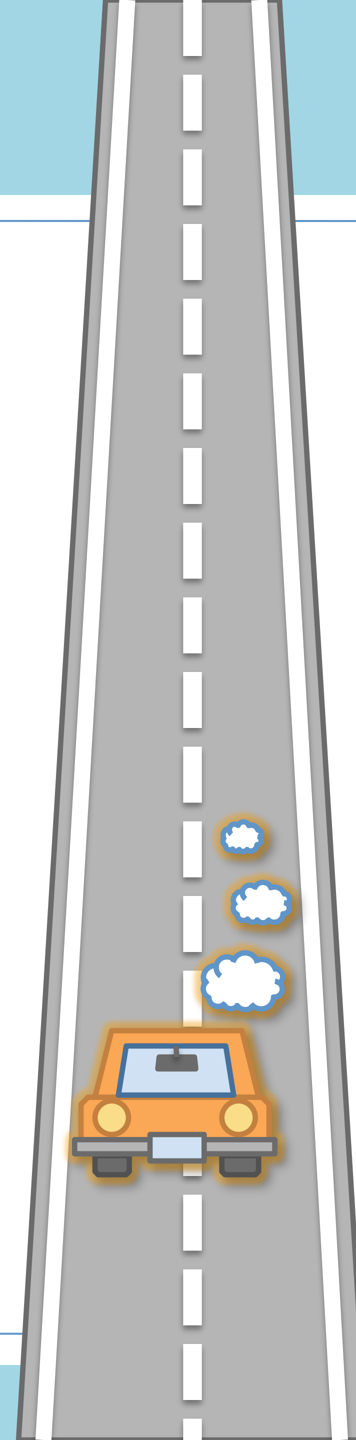
Roadmap

- ✔ Motivation
- ✔ Modeling pow
- ✔ Overview - ma
- ✔ Proposed moc
- ✔ Algorithm - id
- ✔ Experiments
- – Discussion
- – Conclusions



Roadmap

- ✔ Motivation
- ✔ Modeling power of FUNNEL
- ✔ Overview – main ideas
- ✔ Proposed model – idea #1
- ✔ Algorithm – idea #2
- ✔ Experiments
- **Discussion**
- Conclusions

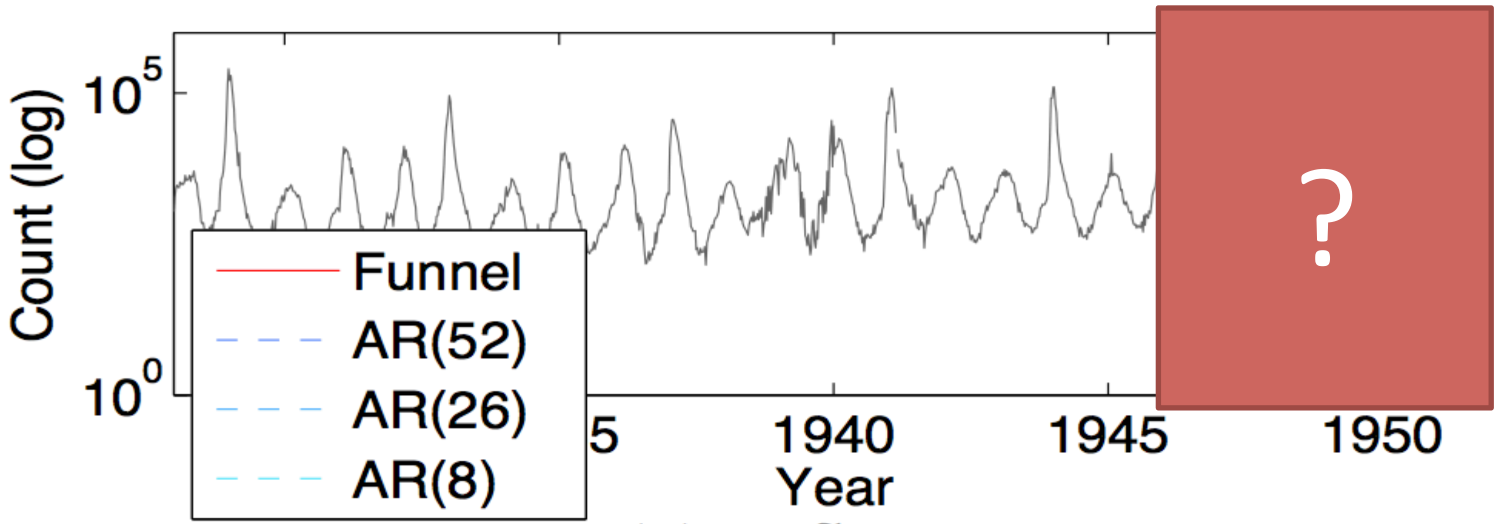


(a) FUNNEL at work - forecasting

Forecasting future epidemics

Train:
2/3 sequences

Forecast:
1/3 following years



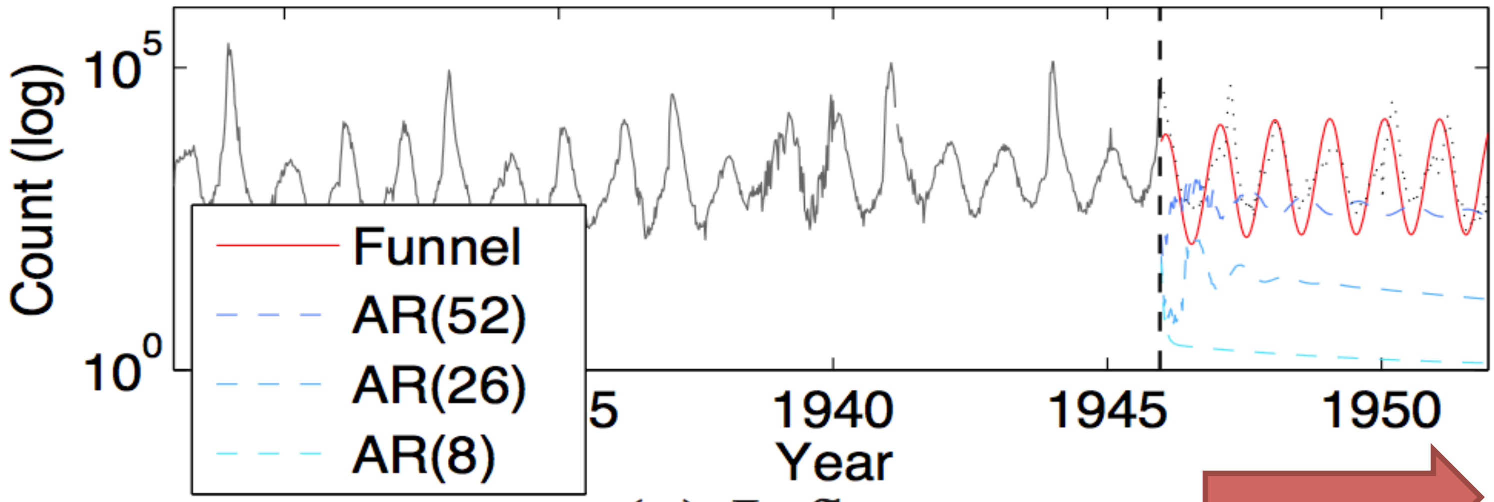
(a) Influenza

(a) FUNNEL at work - forecasting

Forecasting future epidemics

Train:
2/3 sequences

Forecast:
1/3 following years



(a) Influenza

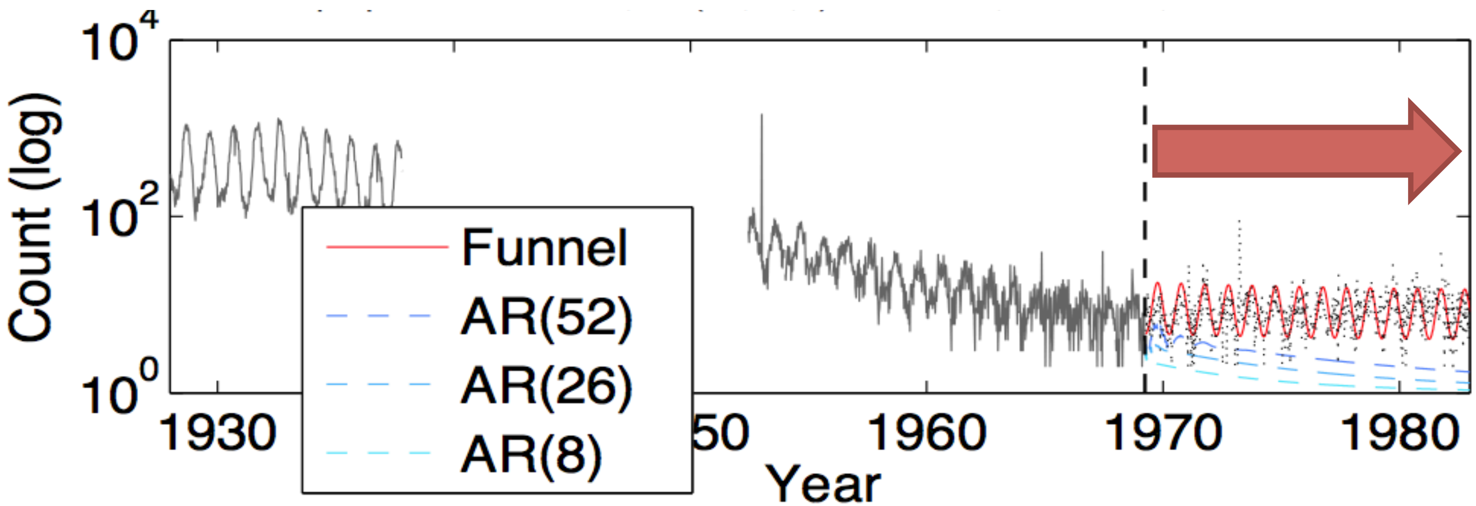
Funnel can capture future epidemics (AR: fail)

(a) FUNNEL at work - forecasting

Forecasting future epidemics

Train:
2/3 sequences

Forecast:
1/3 following years

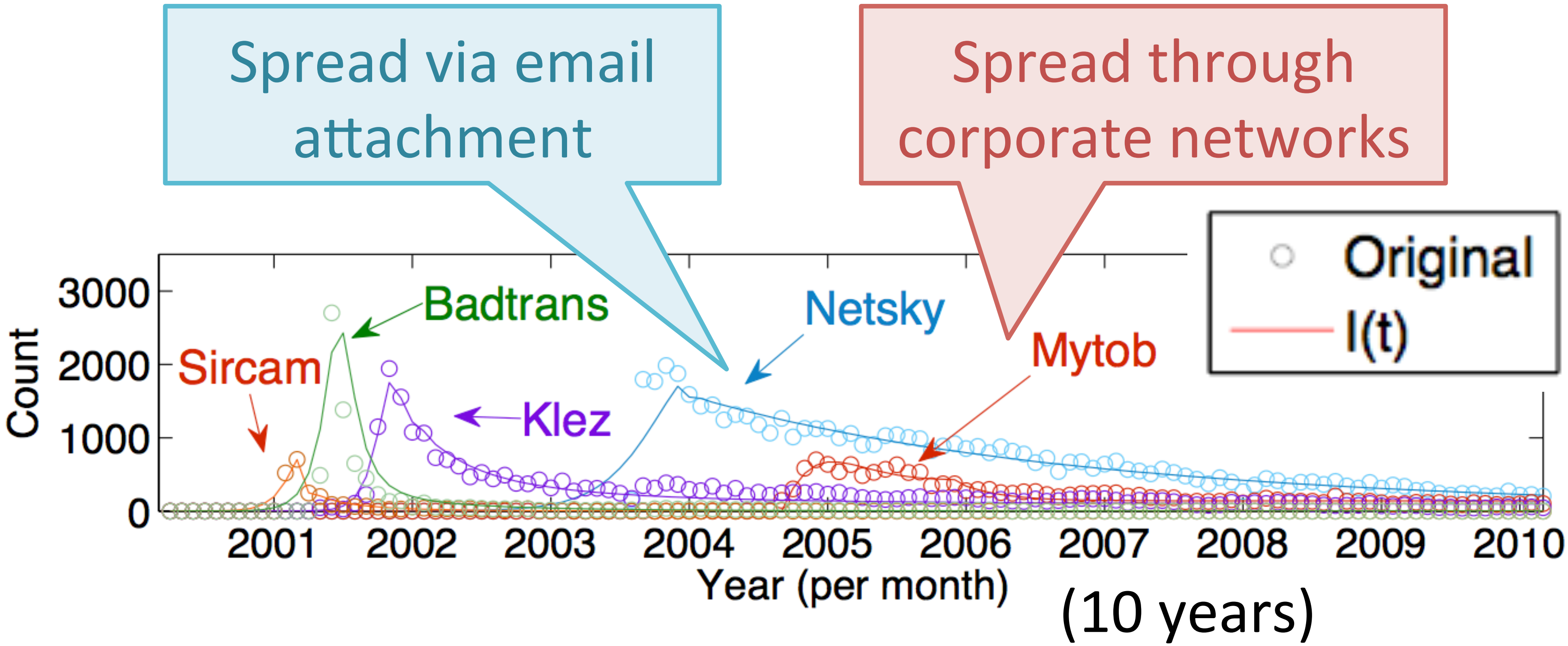


(c) Typhoid fever

Funnel can capture future epidemics (AR: fail)

(b) Generality of FUNNEL

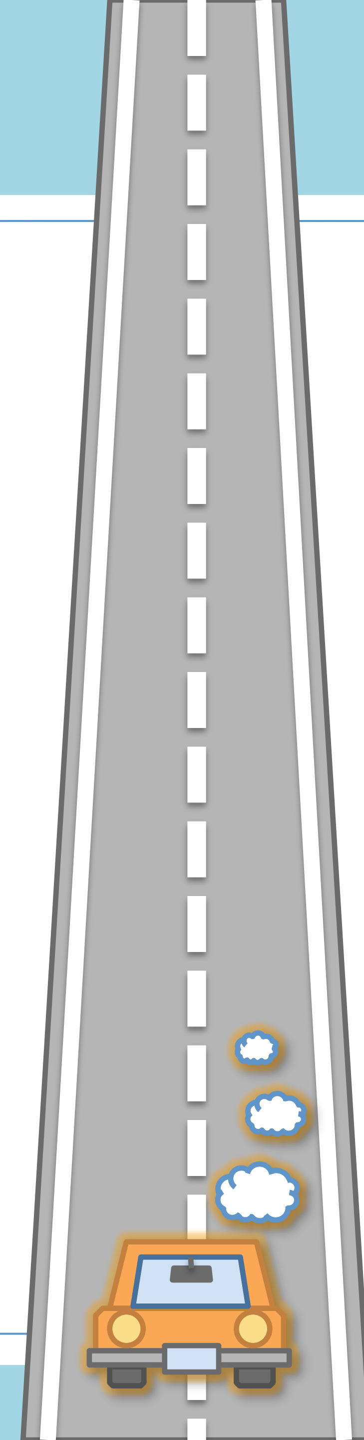
Epidemics on computer networks



Funnel is general: it fits computer virus very well!

Roadmap

- ✔ Motivation
- ✔ Modeling power of FUNNEL
- ✔ Overview – main ideas
- ✔ Proposed model – idea #1
- ✔ Algorithm – idea #2
- ✔ Experiments
- ✔ Discussion
- **Conclusions**



Conclusions

FUNNEL has the following advantages

✔ Sense-making

Captures all essential aspects: P1 P2 P3 P4 P5

✔ Fully-automatic

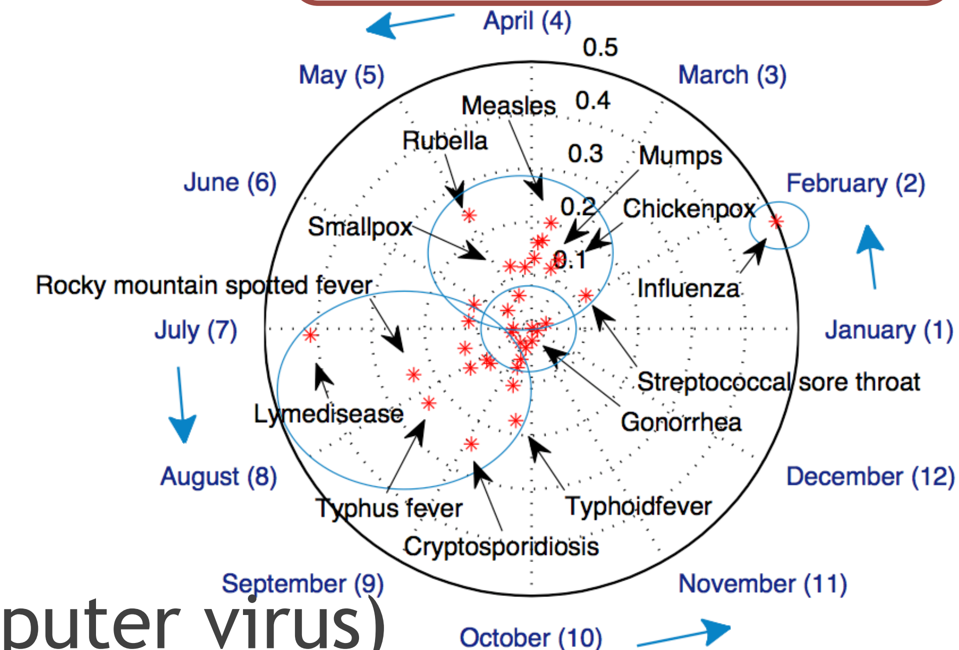
No training set

✔ Scalable

It scales linearly

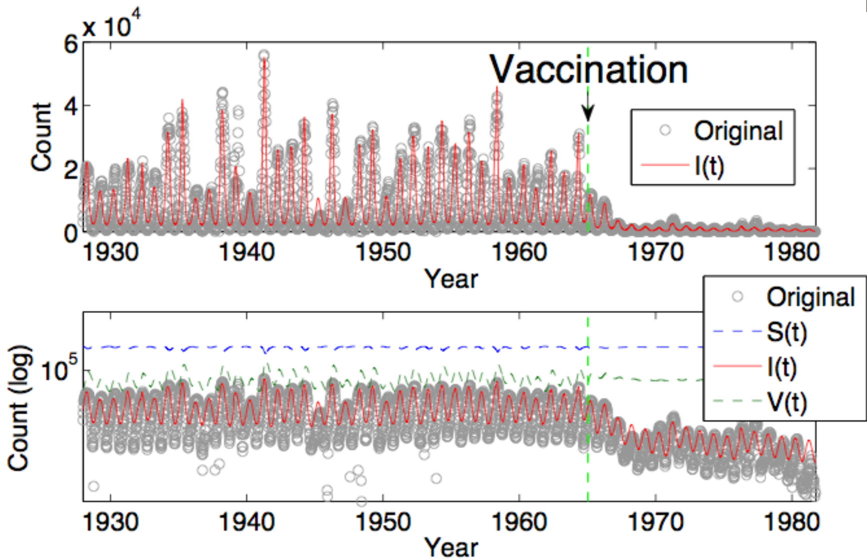
✔ General

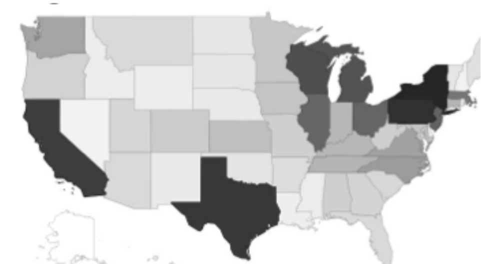
Real epidemics (+ computer virus)

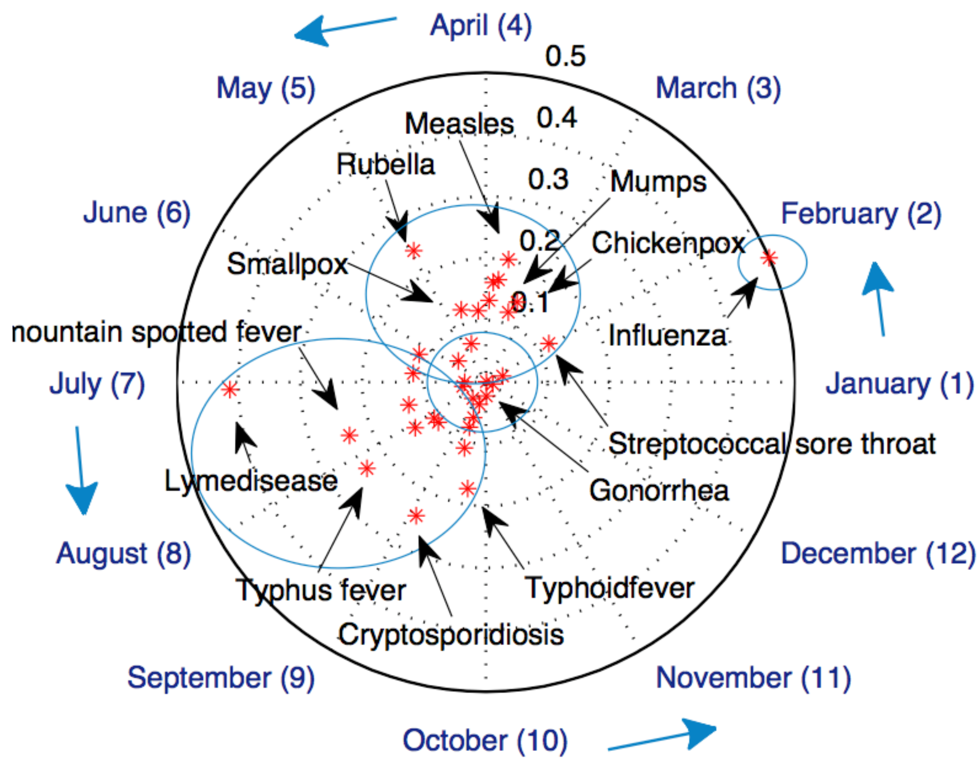


Thank you!







Data: http://www.tycho.pitt.edu/

Code: http://www.cs.kumamoto-u.ac.jp/~yasuko/software.html

FUNNEL: Automatic Mining of Spatially Coevolving Epidemics

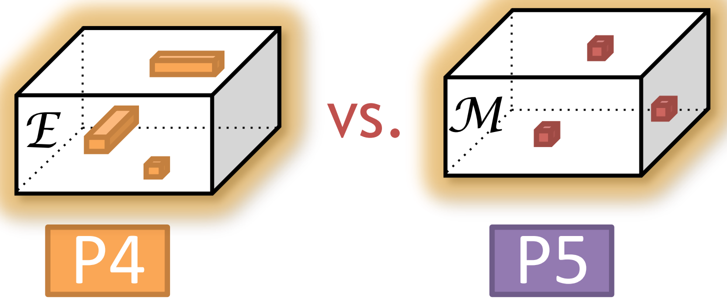
Yasuko Matsubara, Yasushi Sakurai (Kumamoto University)
Willem G. van Panhuis (University of Pittsburgh)
Christos Faloutsos (CMU)

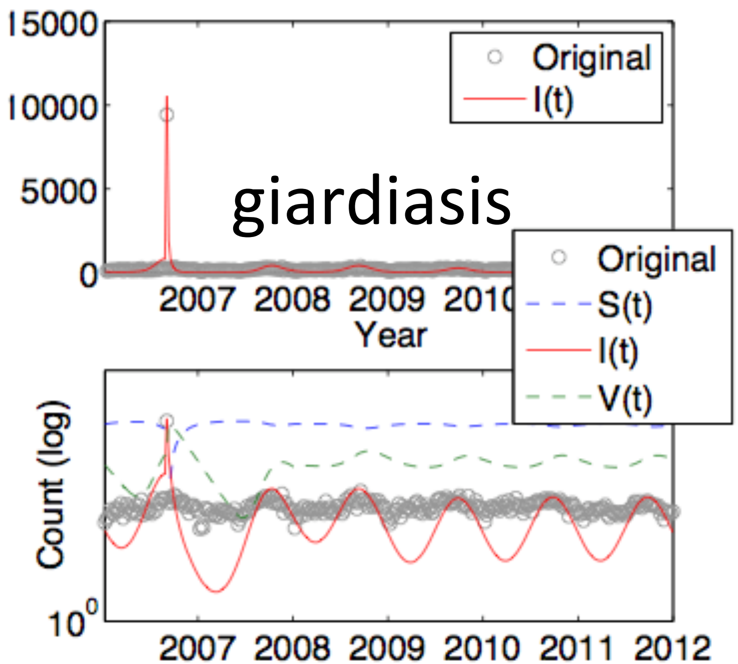
Proposed model: FUNNEL-full

Details

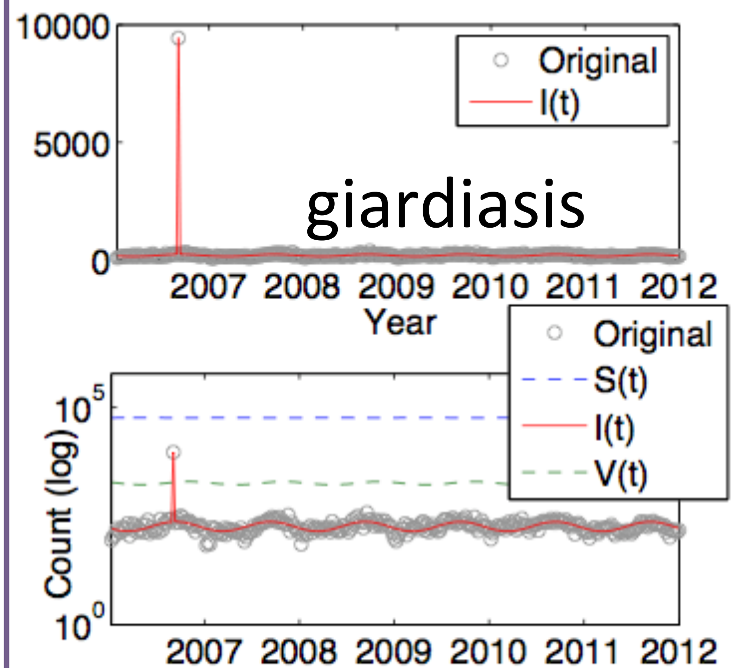
What's the difference??

External shock vs. Mistake





P4 External shock fitting



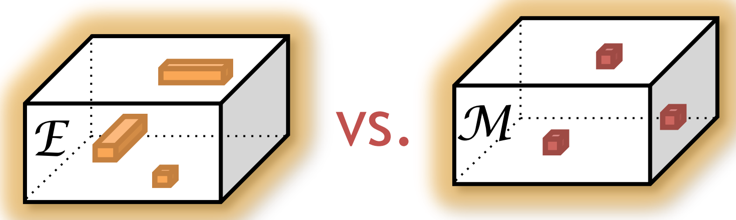
P5 Mistake fitting

Proposed model: FUNNEL-full

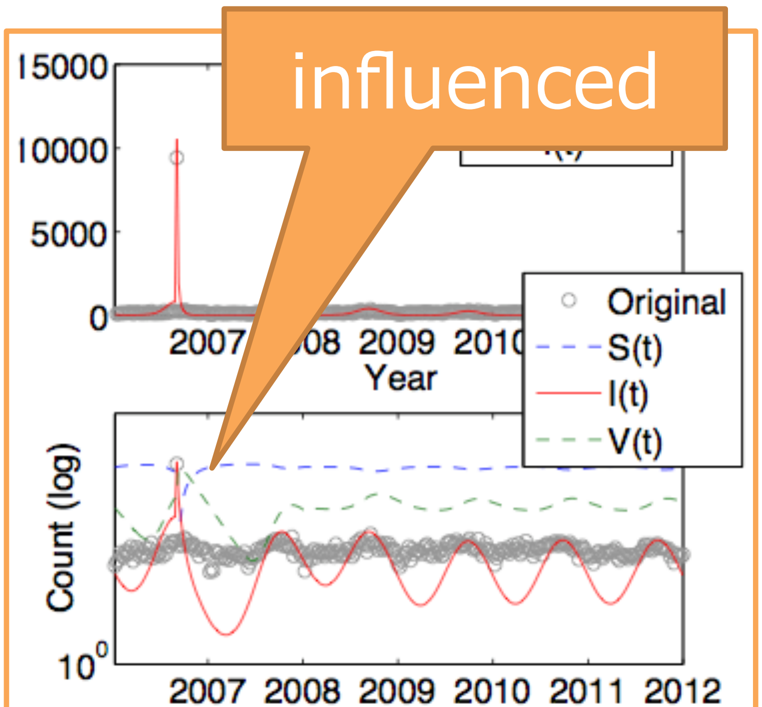
Details

What's the difference??

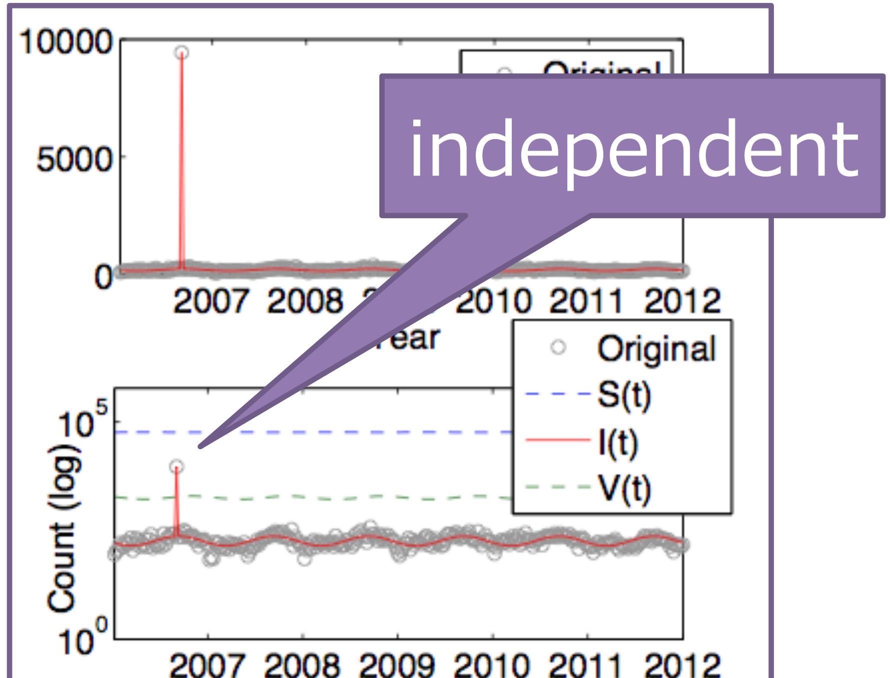
External shock vs. Mistake



P4 P5



P4 External shock fitting

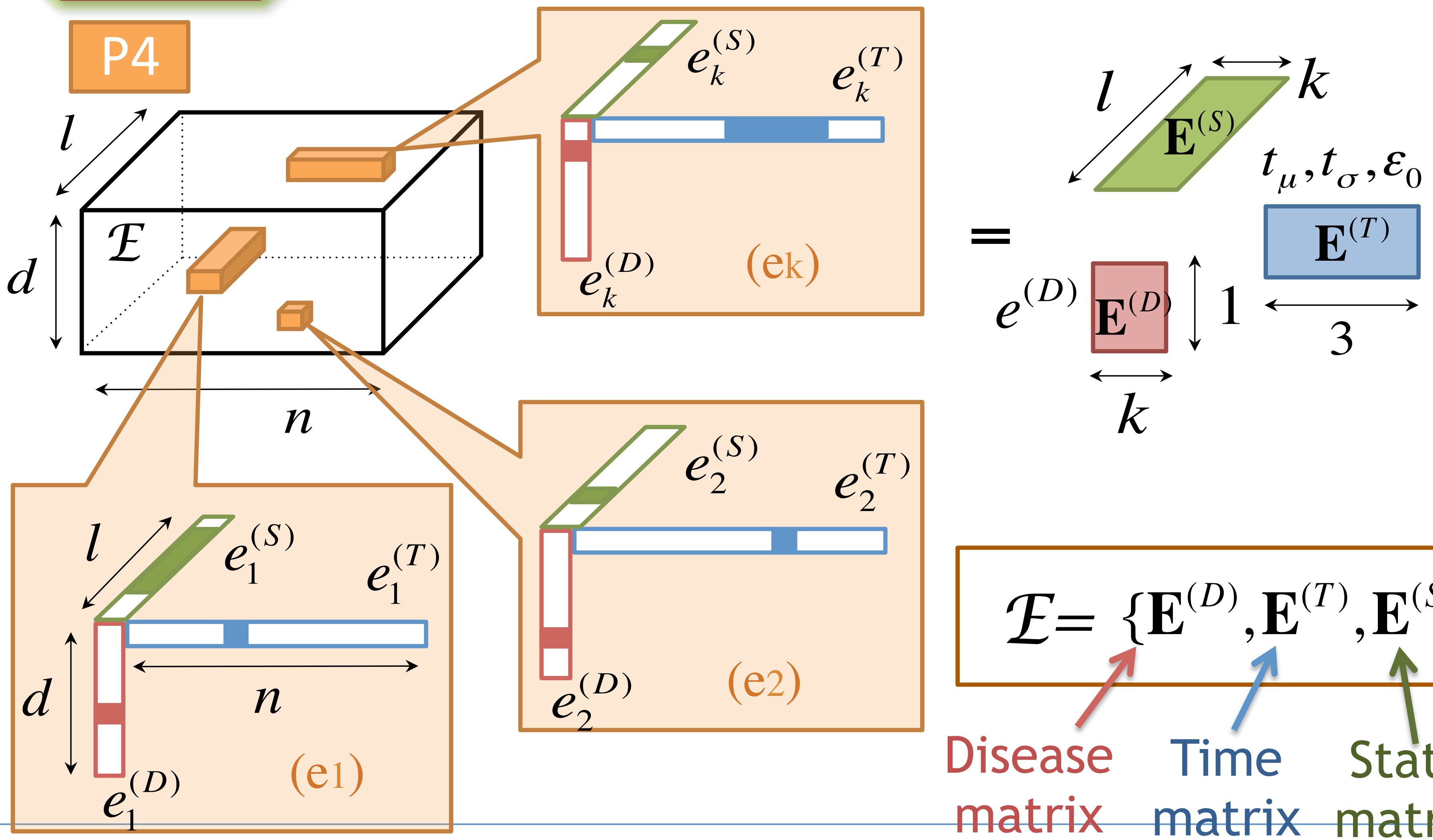


P5 Mistake fitting

Proposed model: FUNNEL-full

Details

P4



$$\mathcal{E} = \{\mathbf{E}^{(D)}, \mathbf{E}^{(T)}, \mathbf{E}^{(S)}\}$$

Disease matrix

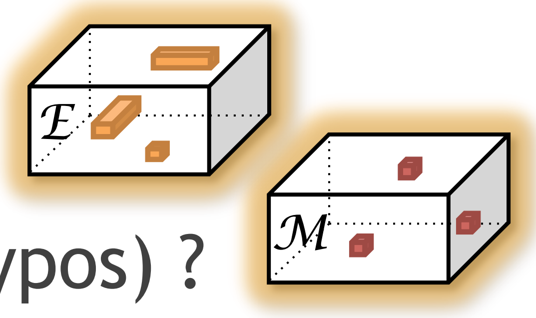
Time matrix

State matrix

Idea (1): Model description cost

Q1. How should we

- find "external shocks" ?
- ignore "mistakes" (i.e., typos) ?



Idea (1) : Model description cost

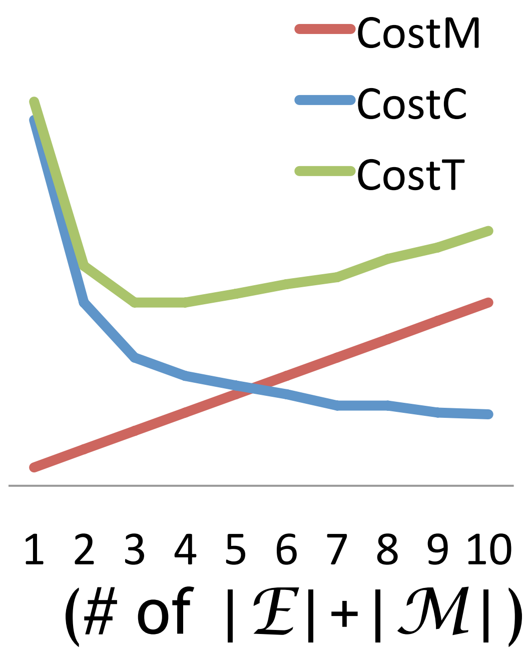
- Minimize coding cost
- find "optimal" # of externals/mistakes
- "automatically"

Idea (1): Model description cost

Idea: Minimize encoding cost!

$$\min \left(\mathrm{Cost}_M(\mathcal{F}) + \mathrm{Cost}_c(\mathcal{X} \mid \mathcal{F}) \right)$$

Model cost   Coding cost



Good compression ⟲ Good description

Idea (1): Model description cost

Details

Total cost of tensor $\mathcal{X}$, given $\mathcal{F}$

$$\mathcal{F} = \{\mathbf{B}, \mathbf{R}, \mathbf{N}, \mathcal{E}, \mathcal{M}\}.$$

$$Cost_T(\mathcal{X}; \mathcal{F}) = \log^*(d) + \log^*(l) + \log^*(n) + Cost_M(\mathbf{B}) + Cost_M(\mathbf{R}) + Cost_M(\mathbf{N}) + Cost_M(\mathcal{E}) + Cost_M(\mathcal{M}) + Cost_C(\mathcal{X}|\mathcal{F})$$

Idea (1): Model description cost

Details

Total cost of tensor $\mathcal{X}$, given $\mathcal{F}$

$$\mathcal{F} = \{\mathbf{B}, \mathbf{R}, \mathbf{N}, \mathcal{E}, \mathcal{M}\}.$$

Dimensions of X

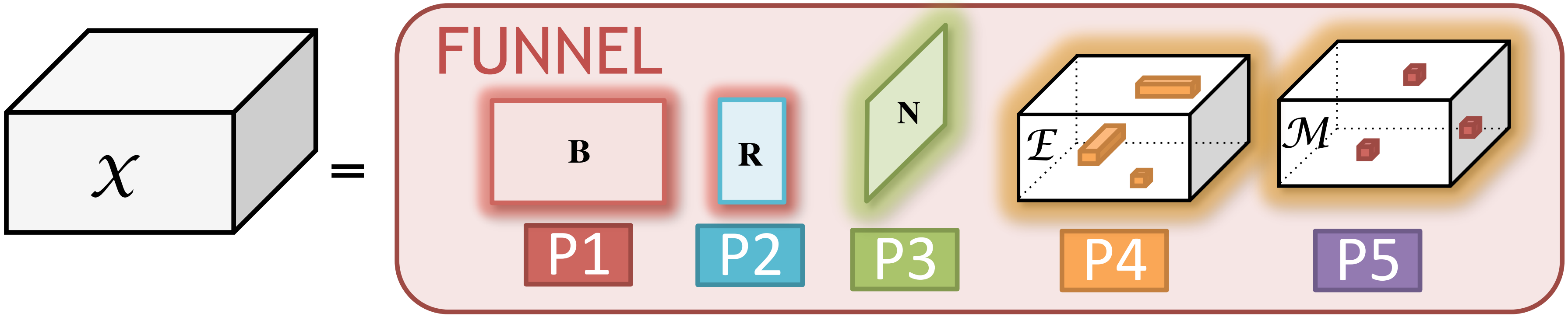
$$Cost_T(\mathcal{X}; \mathcal{F}) = \log^*(d) + \log^*(l) + \log^*(n) + Cost_M(\mathbf{B}) + Cost_M(\mathbf{R}) + Cost_M(\mathbf{N}) + Cost_M(\mathcal{E}) + Cost_M(\mathcal{M}) + Cost_C(\mathcal{X}|\mathcal{F})$$

Model description cost of F

Coding cost of X given F

Idea (2): Multi-layer optimization

Q2. How to efficiently estimate model parameters ?

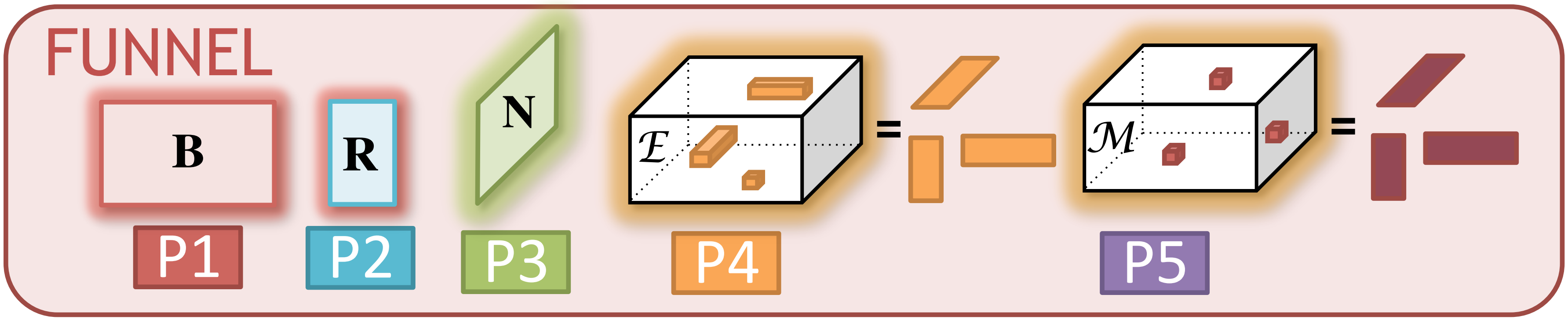


Idea (2): Multi-layer optimization

- Find "optimal" solution w.r.t.
  - Global level parameters
  - Local level parameters

Idea (2): Multi-layer optimization

Find "optimal" solution w.r.t. Global Local

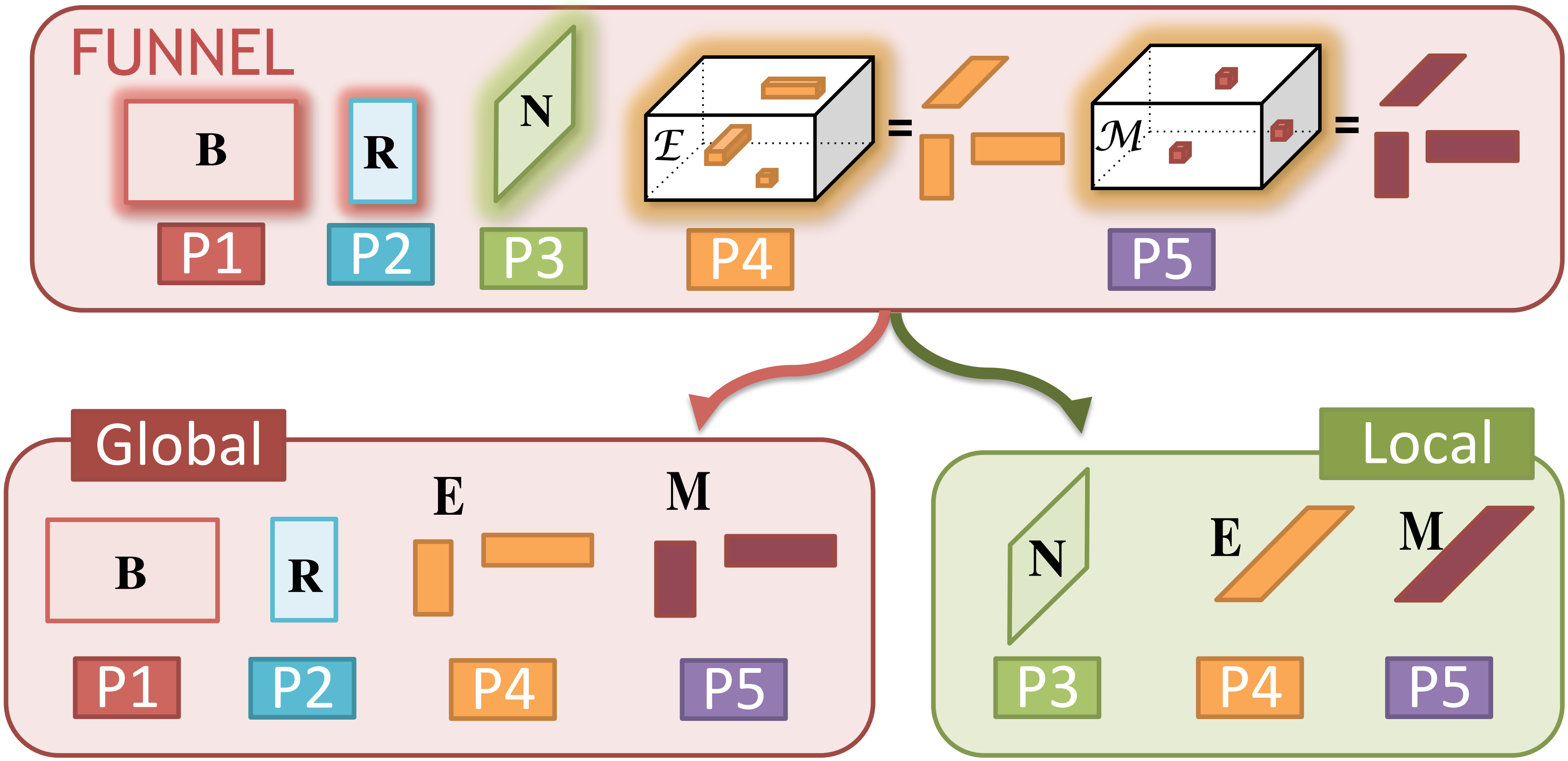


Global

Local

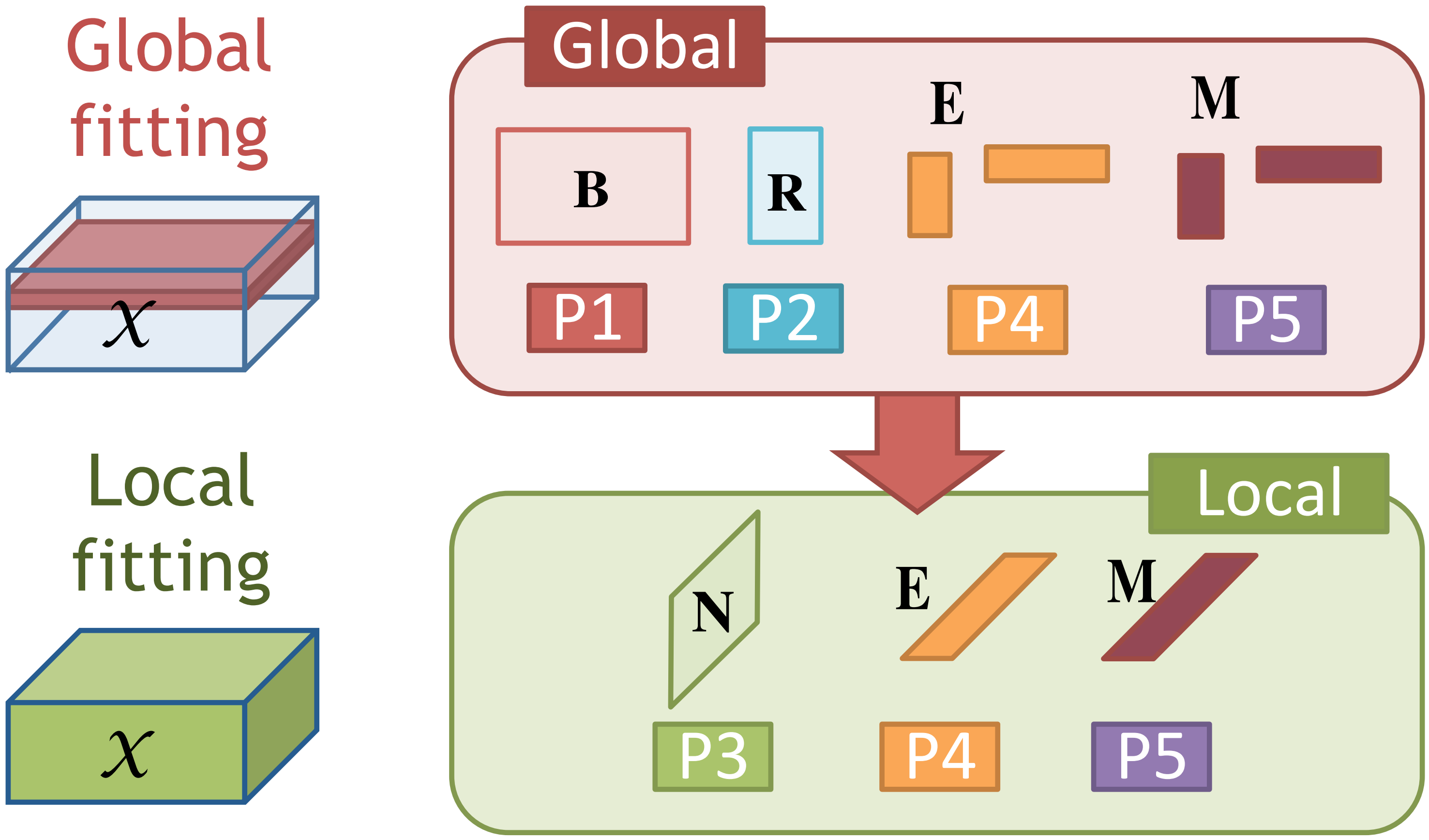
Idea (2): Multi-layer optimization

Find "optimal" solution w.r.t. Global Local



Idea (2): Multi-layer optimization

Multi-layer fitting algorithm



Experiments

We answer the following questions…

Q1. Sense-making

Can it help us understand the given epidemics?

Q2. Accuracy

How well does it match the data?

Q3. Scalability

How does it scale in terms of computational time?

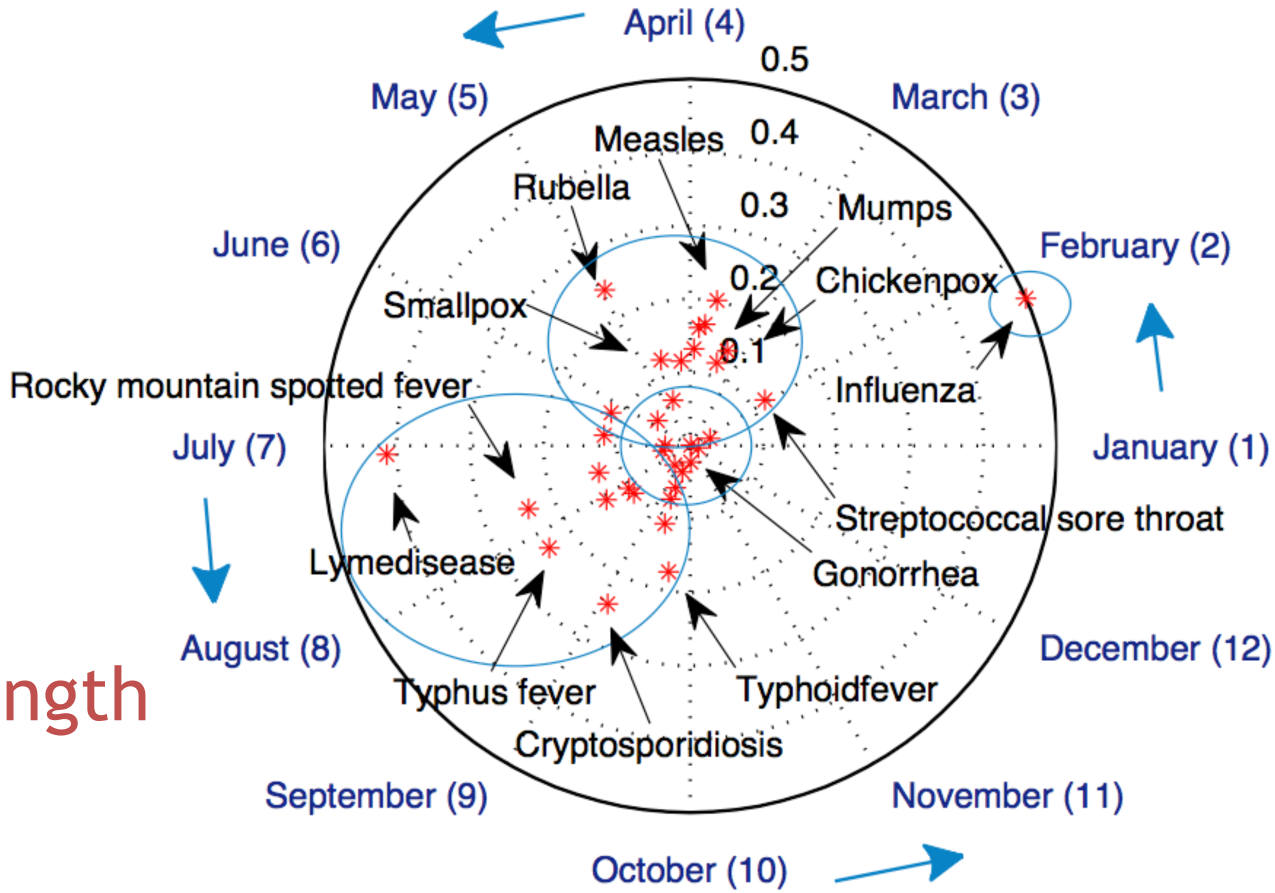
Q1. Sense-making

Our preliminary observations:

- P1 yearly periodicity
- P2 disease reduction effects
- P3 area specificity and sensitivity
- P4 external shock events
- P5 mistakes, incorrect values

Q1. Sense-making

P1 Disease seasonality

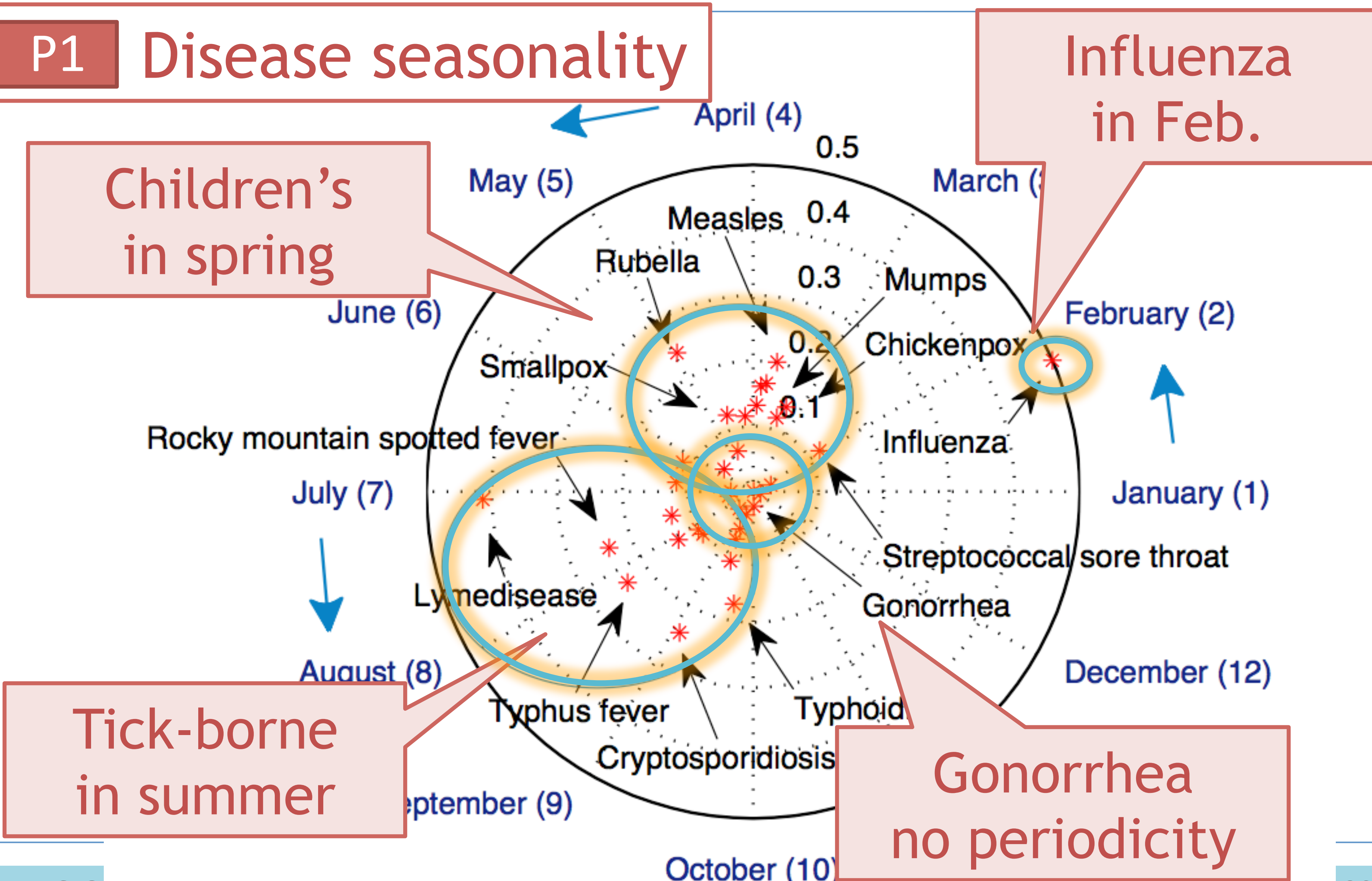


Radius:
seasonality strength

Angle:
peak season

Q1. Sense-making

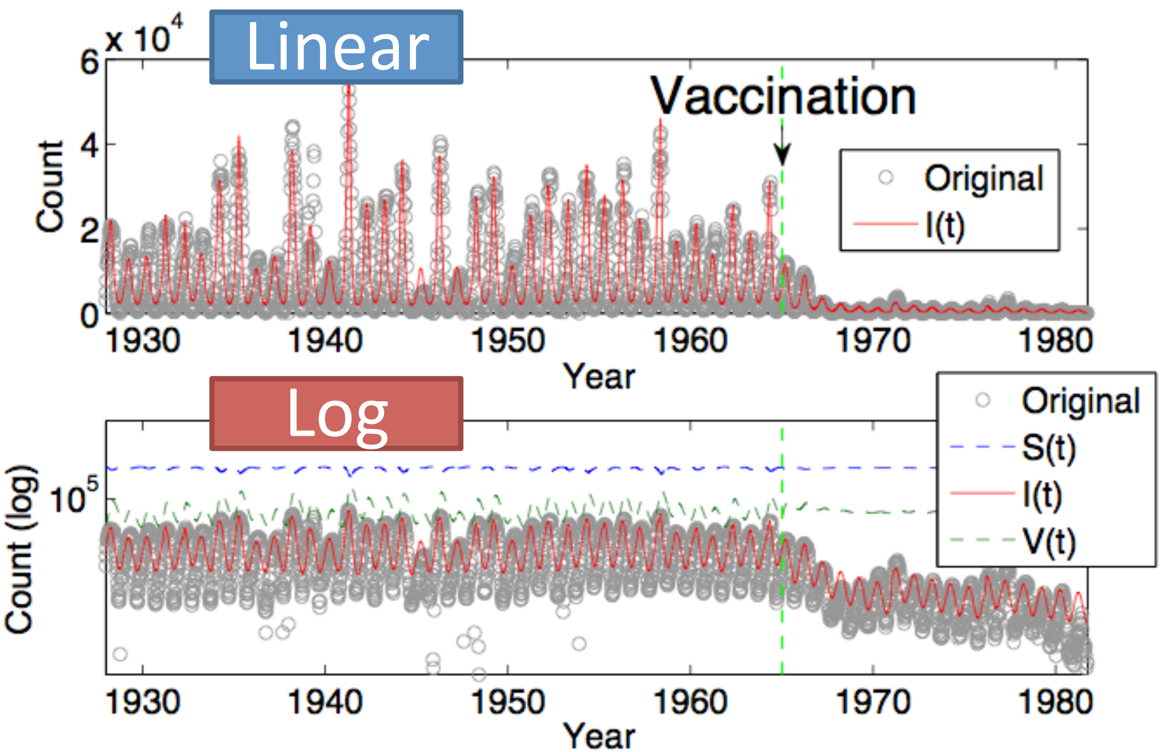
P1 Disease seasonality

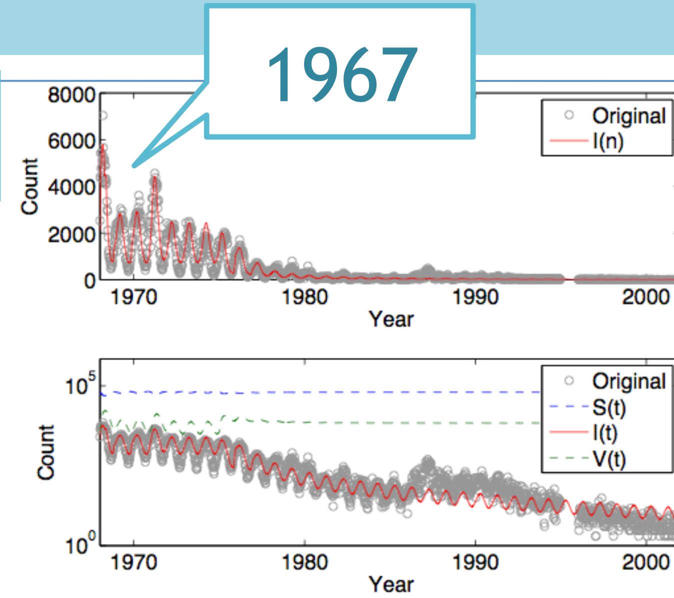


Q1. Sense-making

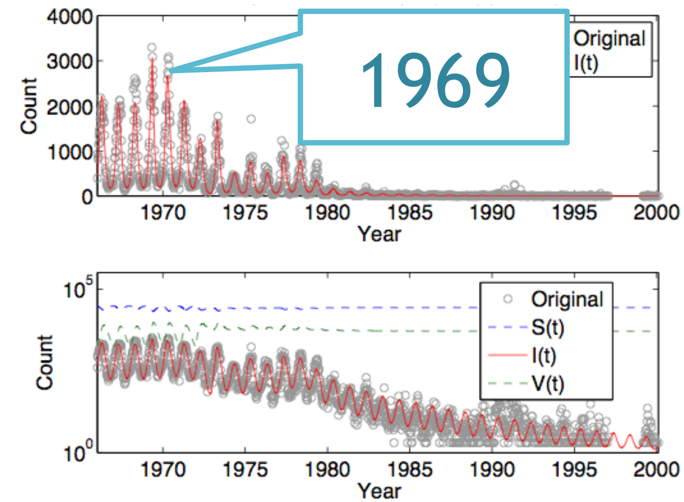
P2 Disease reduction effect

Measles (vaccine licensure: 1963)





(c) Mumps (**P1**), (**P2**), (**P4**)

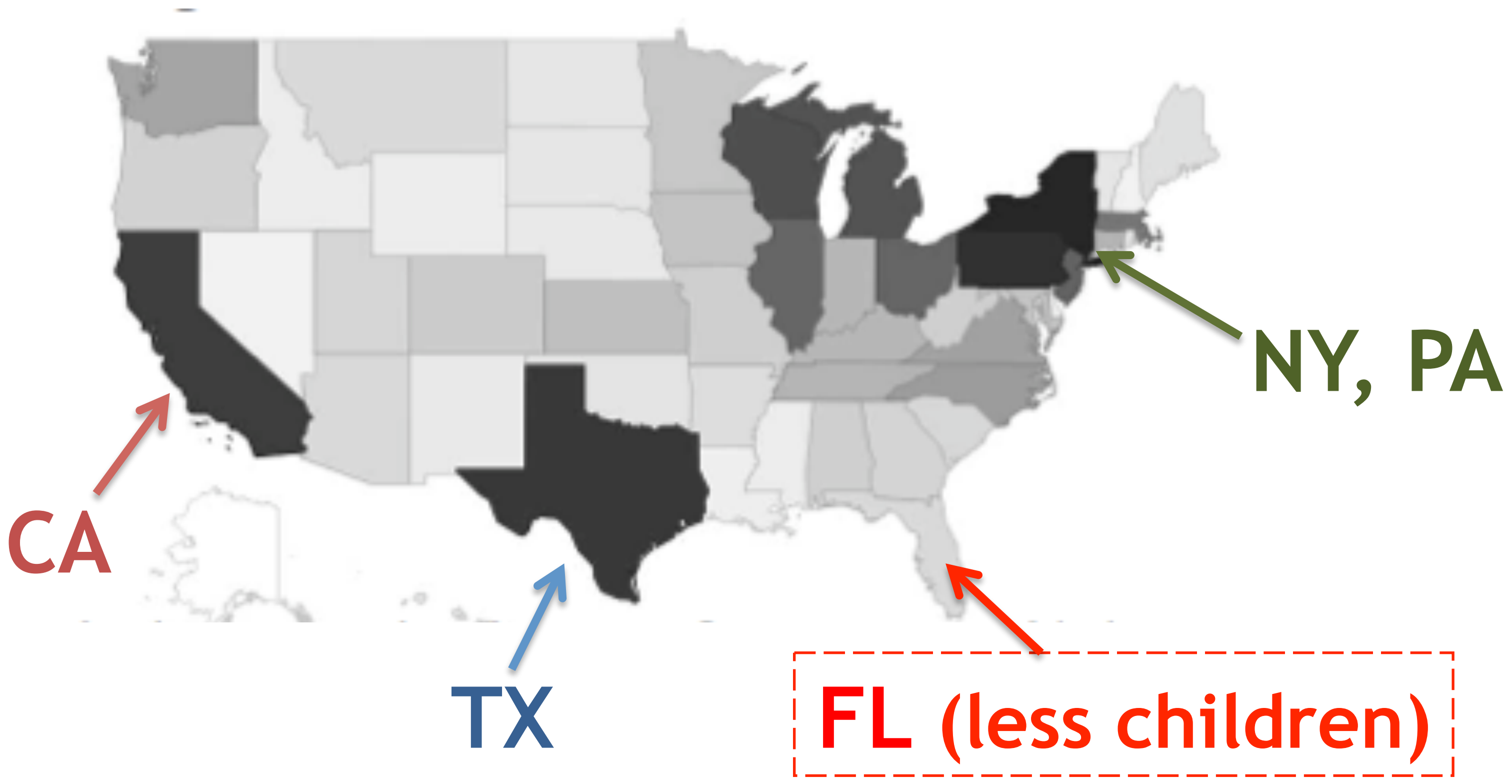


(d) Rubella (**P1**), (**P2**), (**P4**)

Q1. Sense-making

P3 area specificity and sensitivity

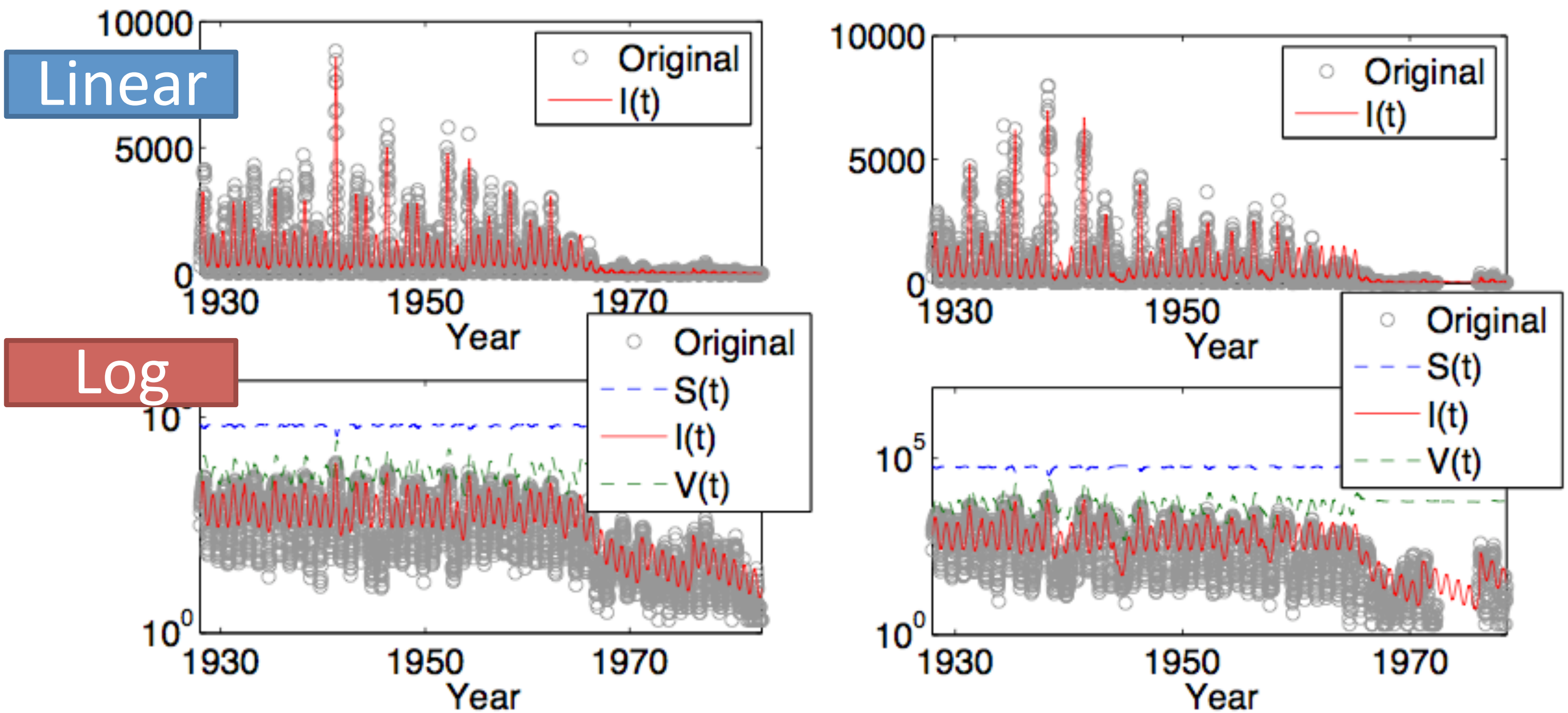
Potential population of susceptibles (measles)



Q1. Sense-making

P3 area specificity and sensitivity

Measles in NY and PA

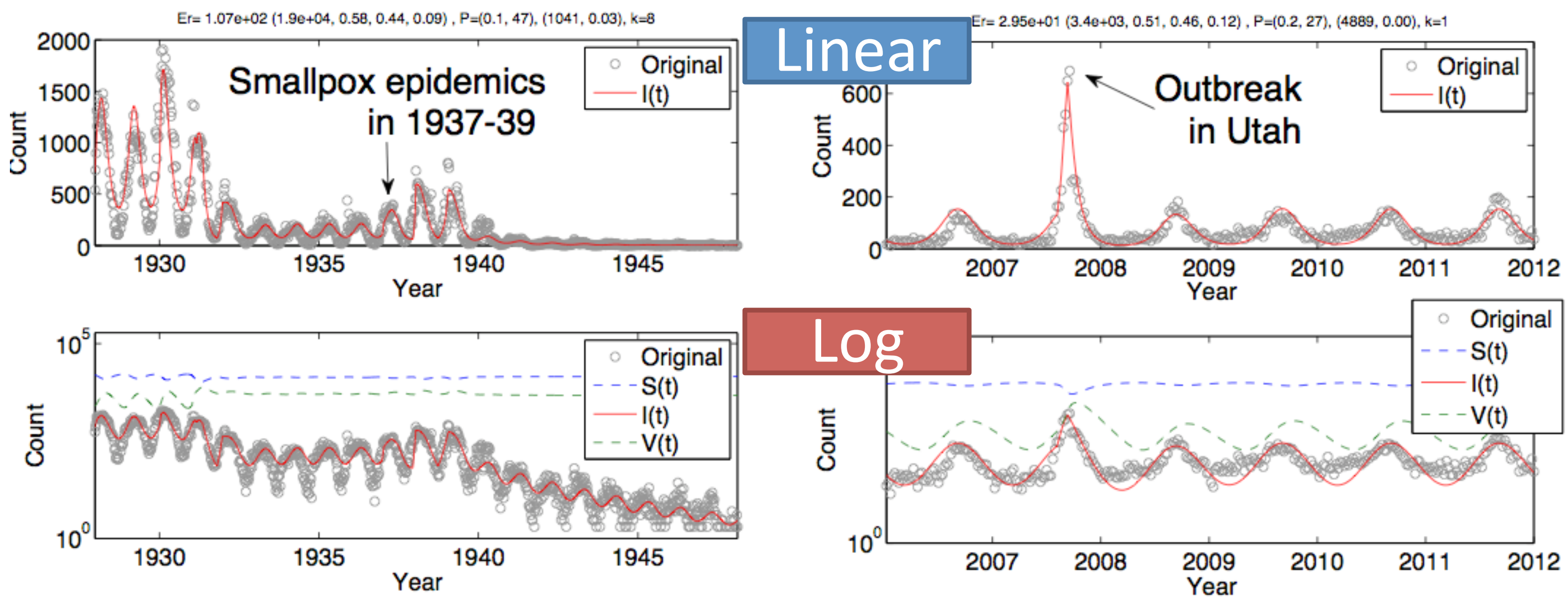


(a) New York State (NY) (b) Pennsylvania (PA)

Q1. Sense-making

P4 external shock events

We can detect external shocks "**automatically**" !!



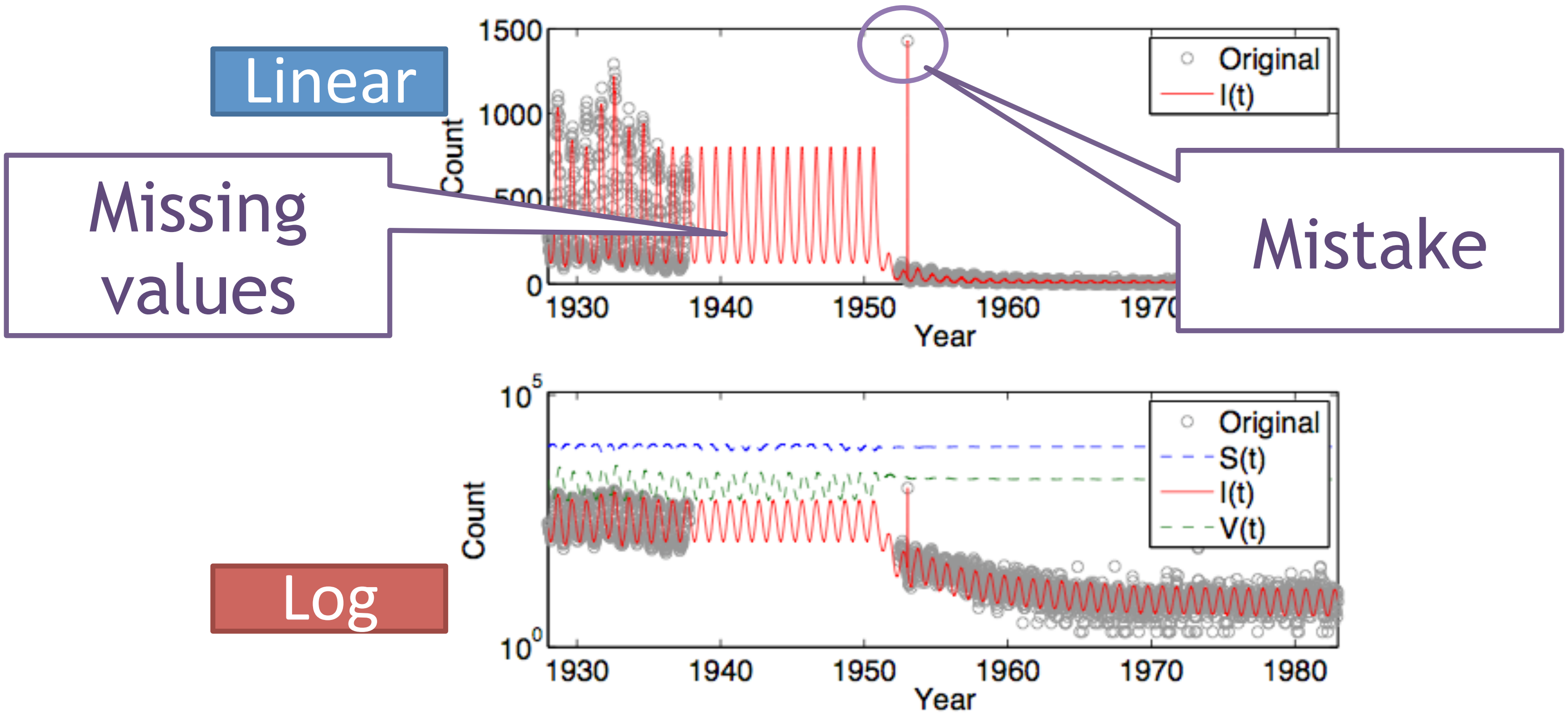
(h) Smallpox (**P1**), (**P2**), (**P4**)

(j) Cryptosporidiosis (**P1**), (**P3**), (**P4**)

Q1. Sense-making

P5 mistakes, incorrect values

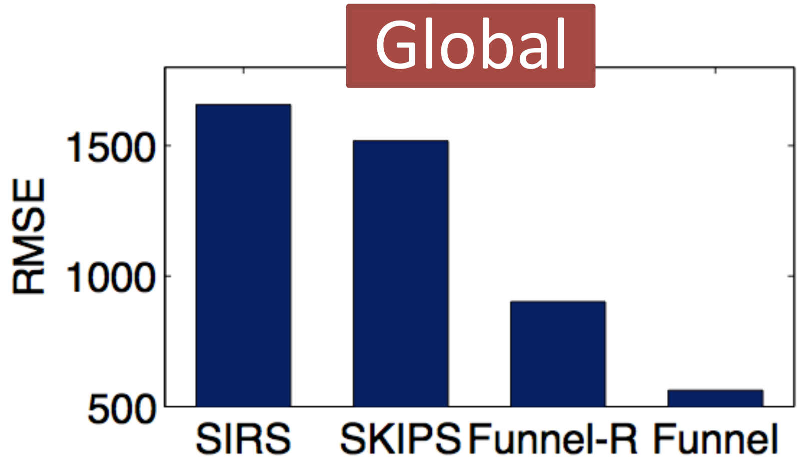
We can also detect typos "**automatically**" !!



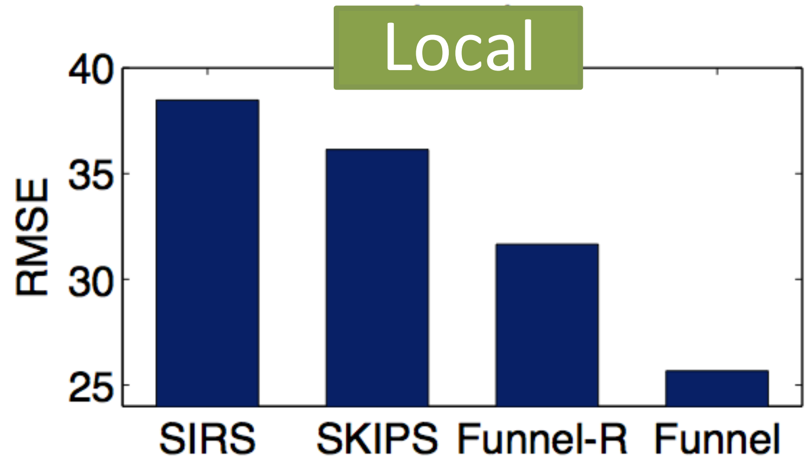
Q2. Accuracy

Fitting accuracy for Global Local sequences

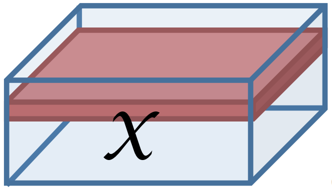
(lower is better)



(a) Global fitting

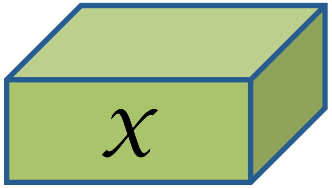


(b) Local fitting



$$\{\bar{\boldsymbol{x}}_i(t)\}_{i,t}^{d,n}$$

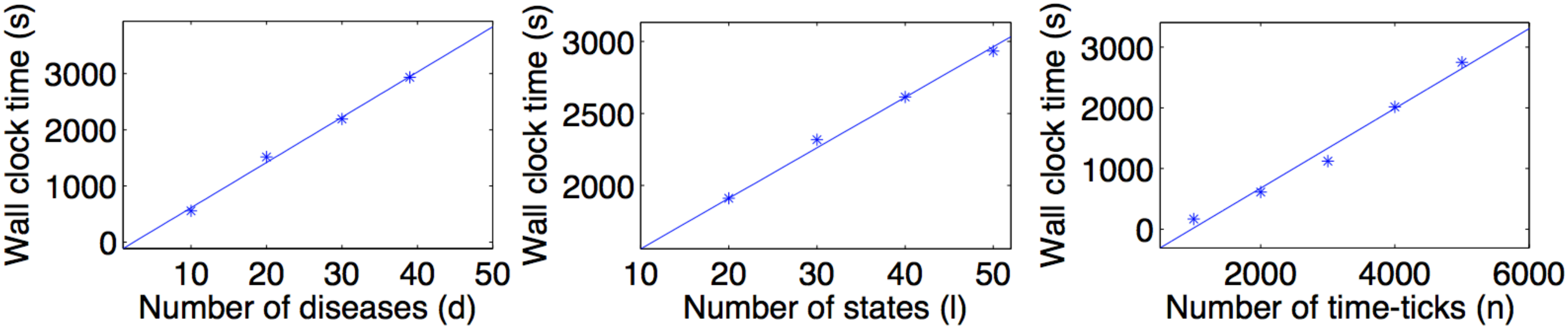
$$\bar{x}_i(t) = \sum_{j=1}^{l} x_{ij}(t).$$



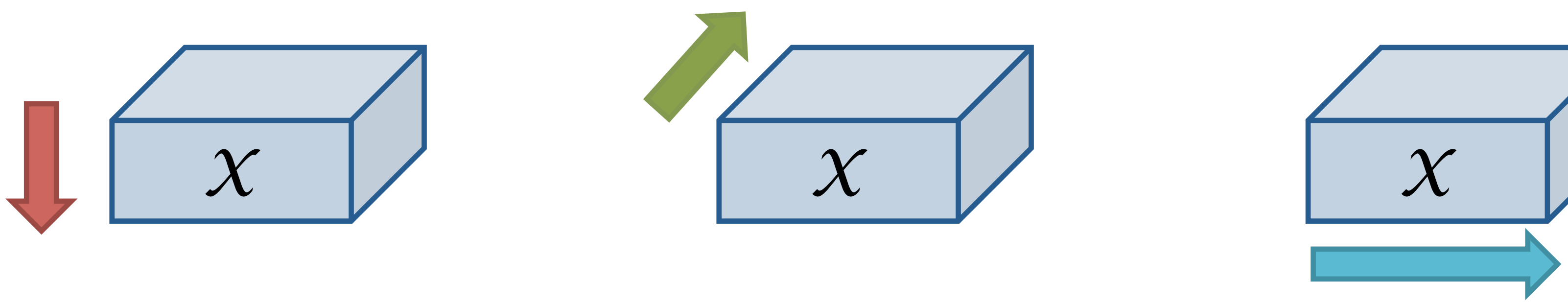
$$\{\boldsymbol{x}_{ij}(t)\}_{i,j,t}^{d,l,n}$$

Q3. Scalability

Wall clock time vs. diseases , states , Time



(a) Diseases (d) (b) States (l) (c) Duration (n)



FunnelFit is linear w.r.t. data size : O(dln)