Microsoft Research

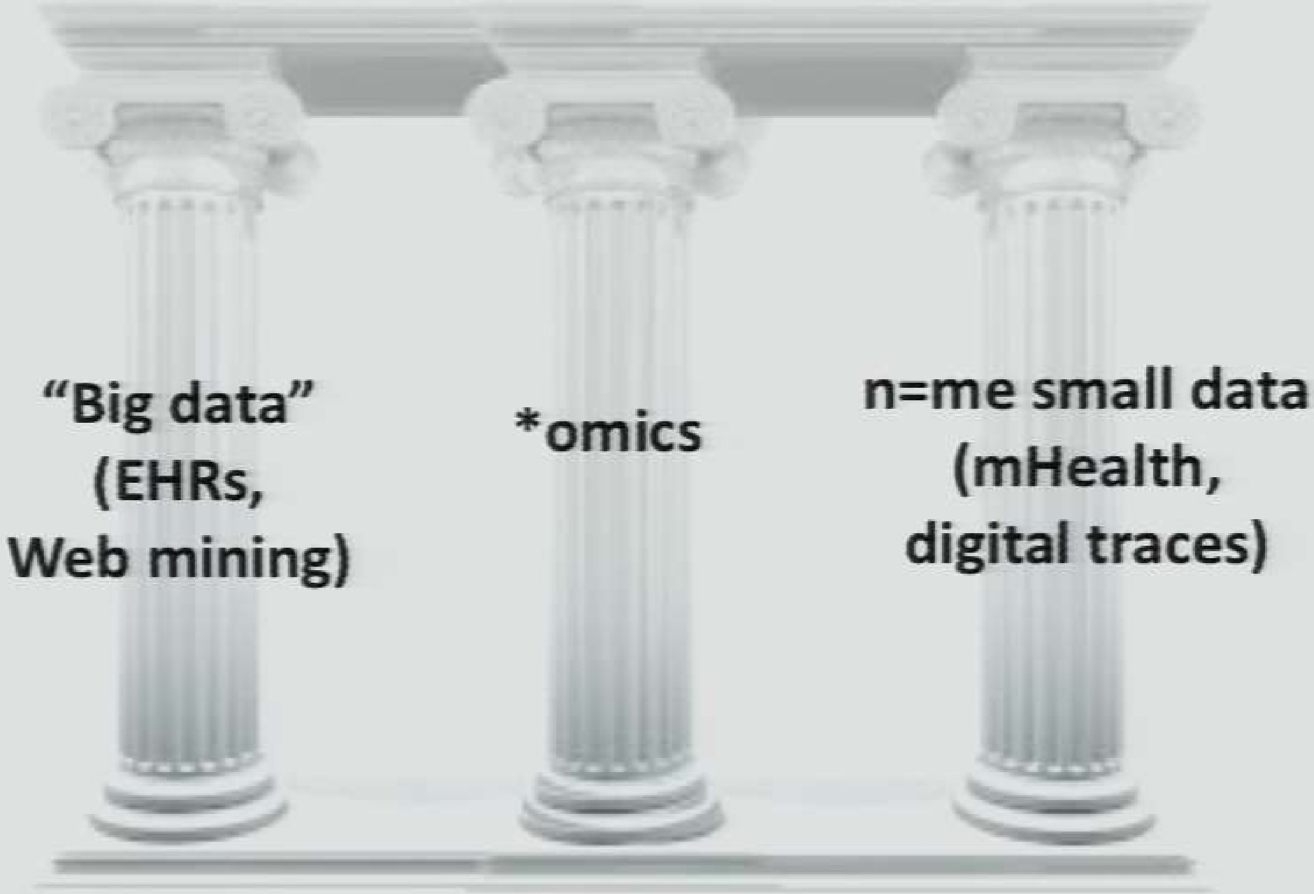
Each year Microsoft Research hosts hundreds of influential speakers from around the world including leading scientists, renowned experts in technology, book authors, and leading academics, and makes videos of these lectures freely available.

Small, n=me, data

Deborah Estrin
Professor, Computer Science, Cornell NYC Tech
Professor, Public Health, Weill Cornell Medical College
Co-founder, Open mHealth

destrin@cs.cornell.edu
work done with collaborators from
Cornell, UCLA, openmhealth.org, ...

Personalized, precision, medicine



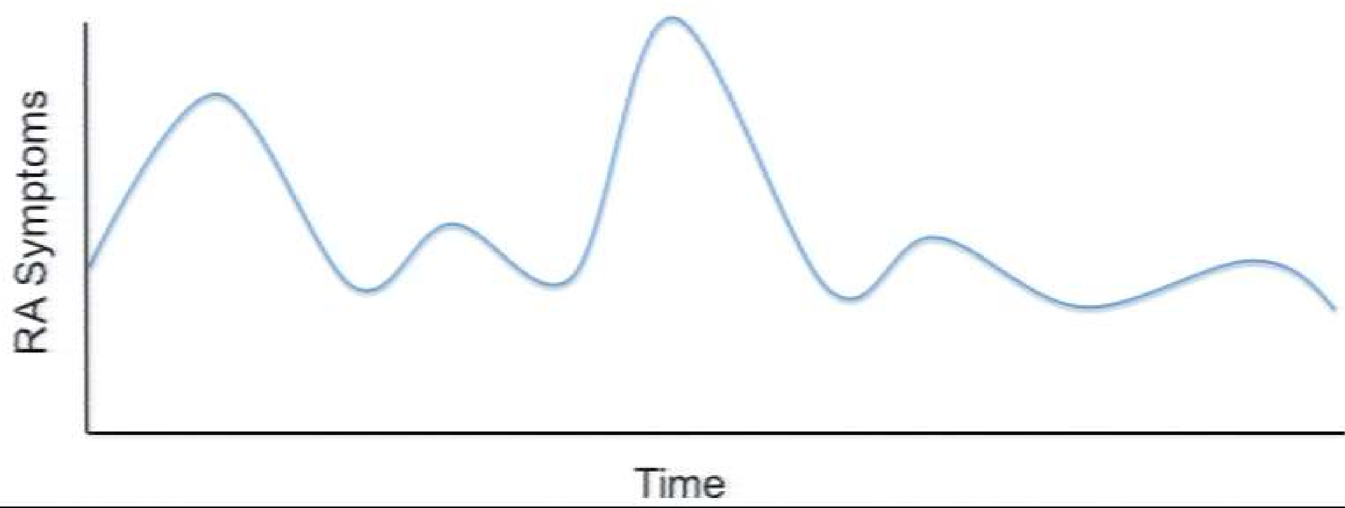
harness previously-unmeasured function and behavior
to fuel personalized and evidence-producing care

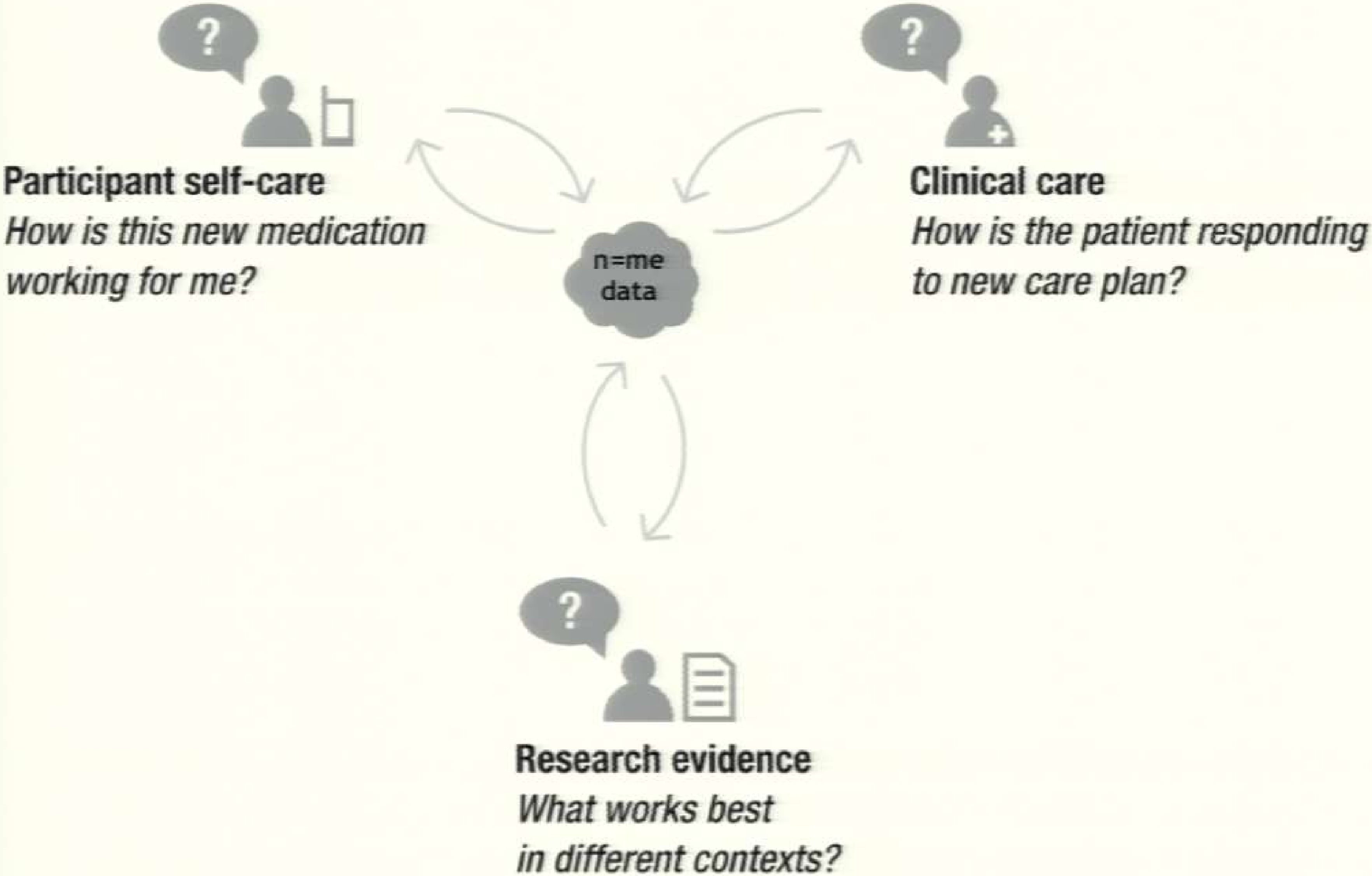
A patient with arthritis

Understanding fluctuations in RA disease activity

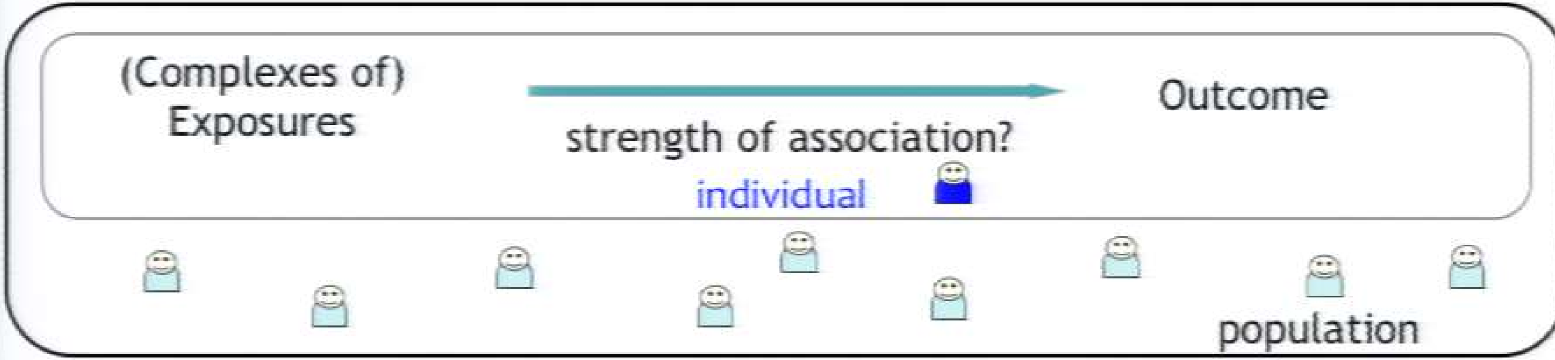
(D. Orange, R. Darnell, et al)

- 80% of RA patients experience relapsing-remitting course
    - unpredictable flares limit patient function, work productivity
- Treatment often has unwanted side effects (short courses of steroids: insomnia, hypertension, glucose intolerance, ..)
- "biologic agents" (TNF inhibitors) only make 60% of patients >20% better; 40%, >50% better; 15%, >70% better
- Early diagnosis and treatment leads to improved outcomes




Participant self-care
How is this new medication working for me?
n=me data
Clinical care
How is the patient responding to new care plan?
Research evidence
What works best in different contexts?

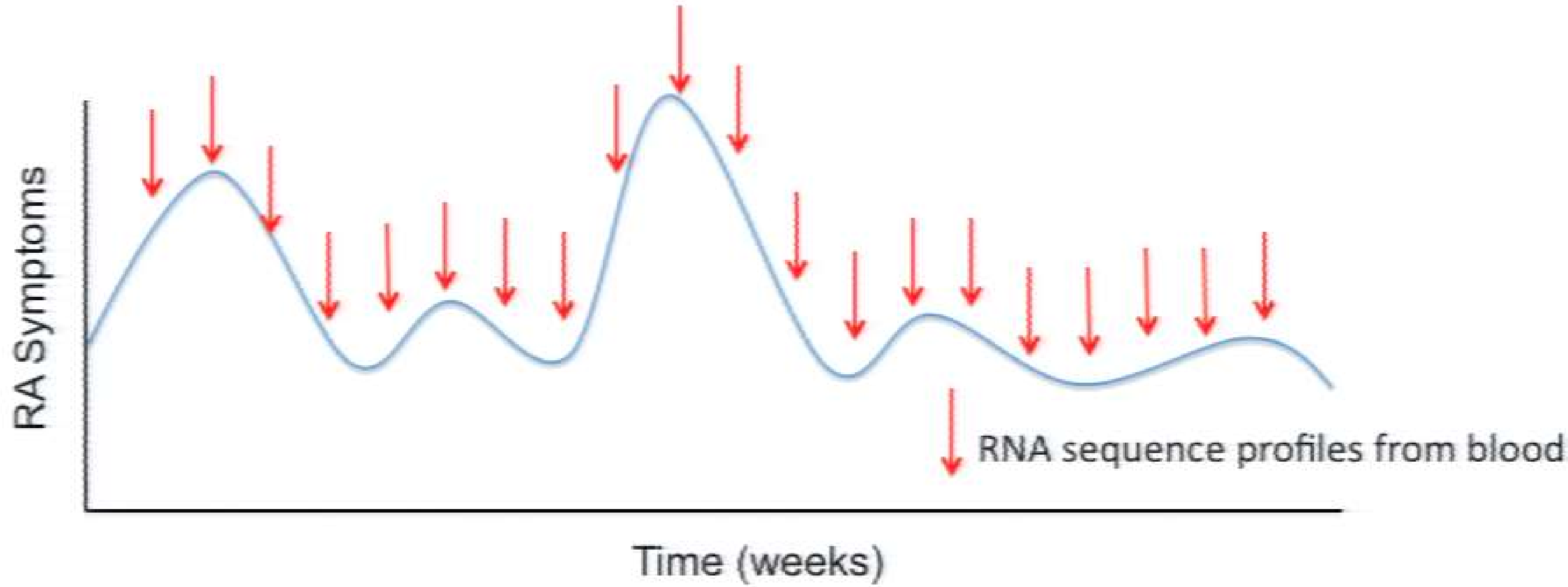
Profound potential to rephrase 'does it work?'

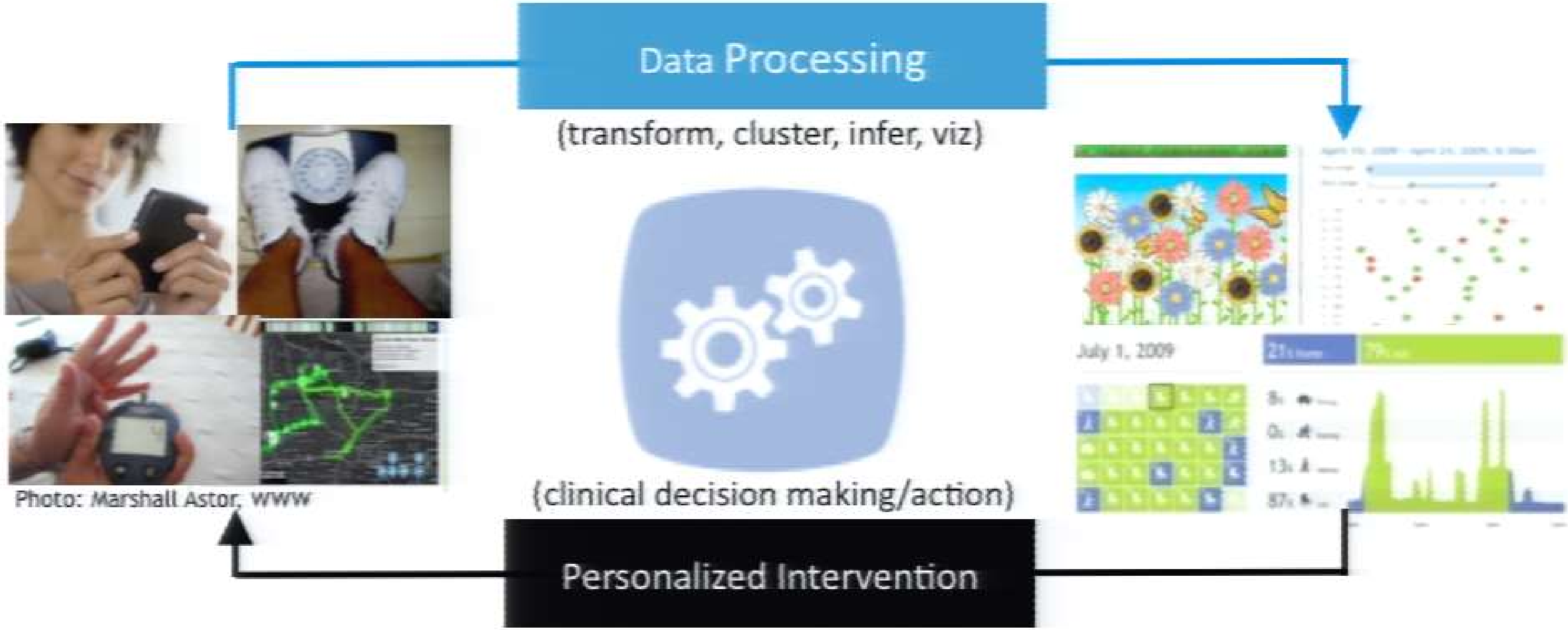


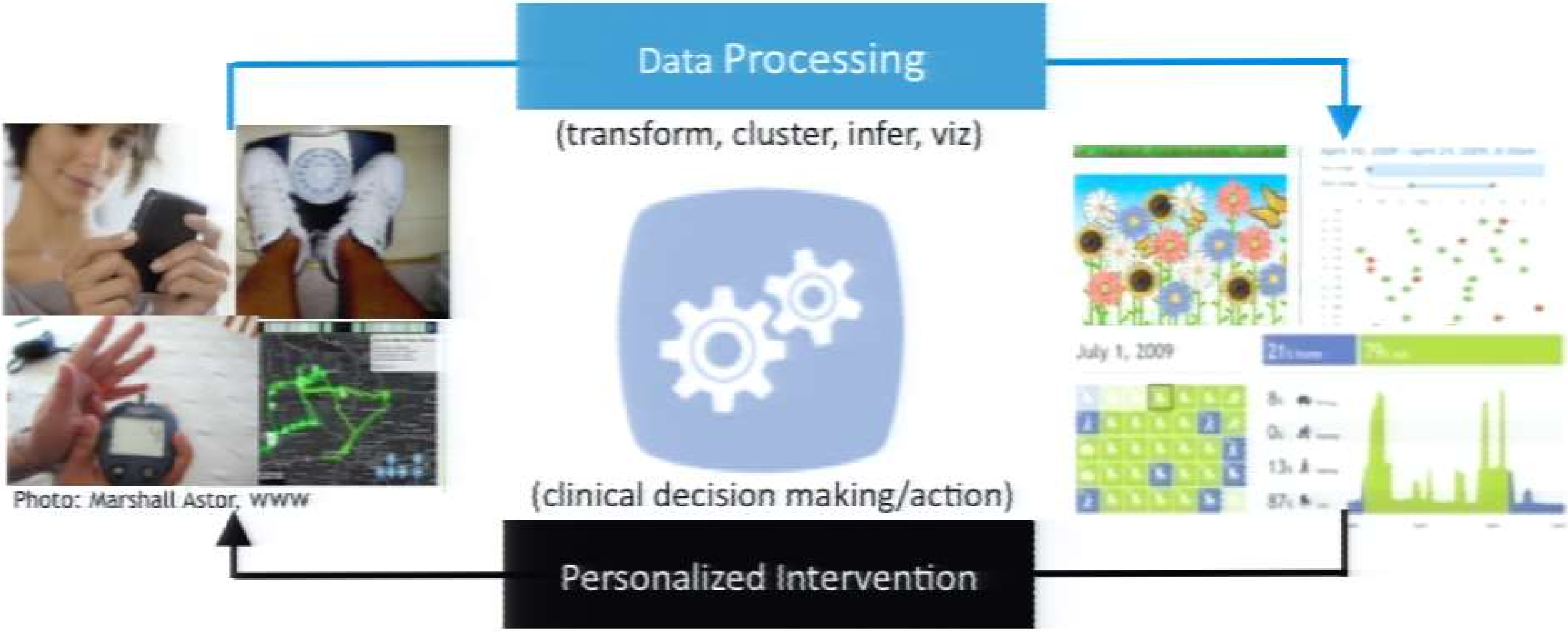
'does it work on average?' (RCT)

Using bioinformatics to understand fluctuations in RA disease activity (Orange, Darnell, et al)

- Frequent blood samples taken and mailed in to support analysis pre, during and after flare
- mHealth:
  - make such studies practical
  - support adaptive sampling (sample more when signs of flare begin)?
  - support personalized management based on mechanism discovery



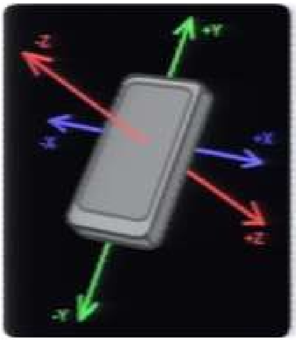

Data Processing
{transform, cluster, infer, viz}
{clinical decision making/action}
Personalized Intervention
Photo: Marshall Astor, www
July 1, 2009

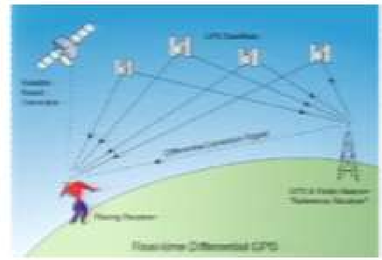


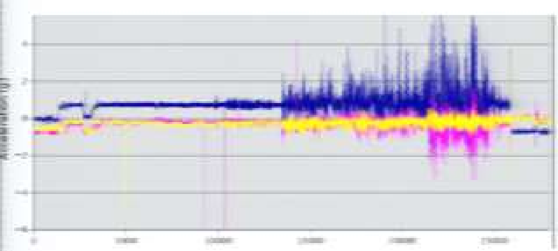
ultimate goal: create scalable tools for PCP and clinic based care that supports precision, personalization, and continuity of care in:

- RA, Lupus, Crohns, Asthma, MS
- Hospital discharge, Surgical recovery
- Pain management, Chronic fatigue, Migraines
- Depression, ADHD, insomnia, post-traumatic stress disorder
- Integrative medicine effectiveness
- Behavior change (individual, family , community)

Passively-recorded activity and location traces

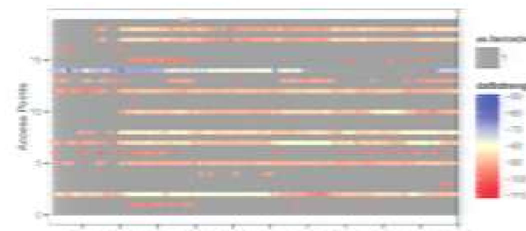




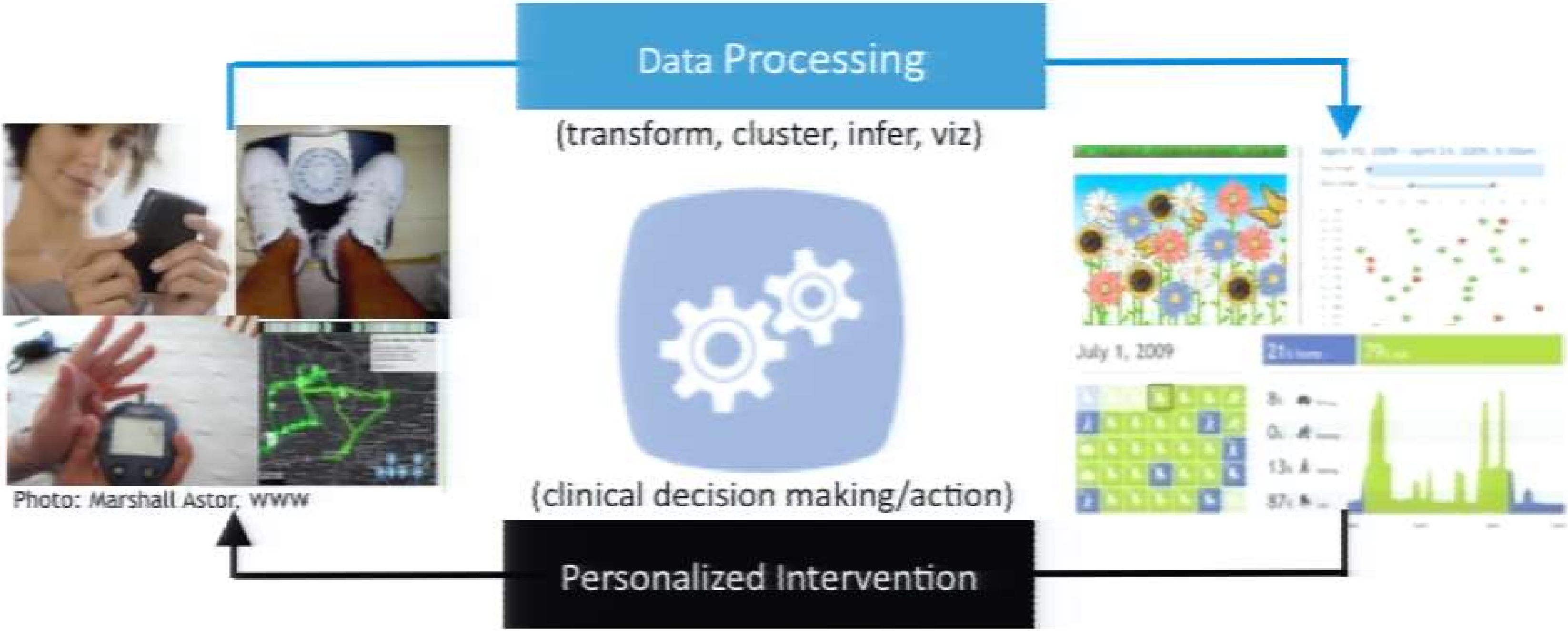


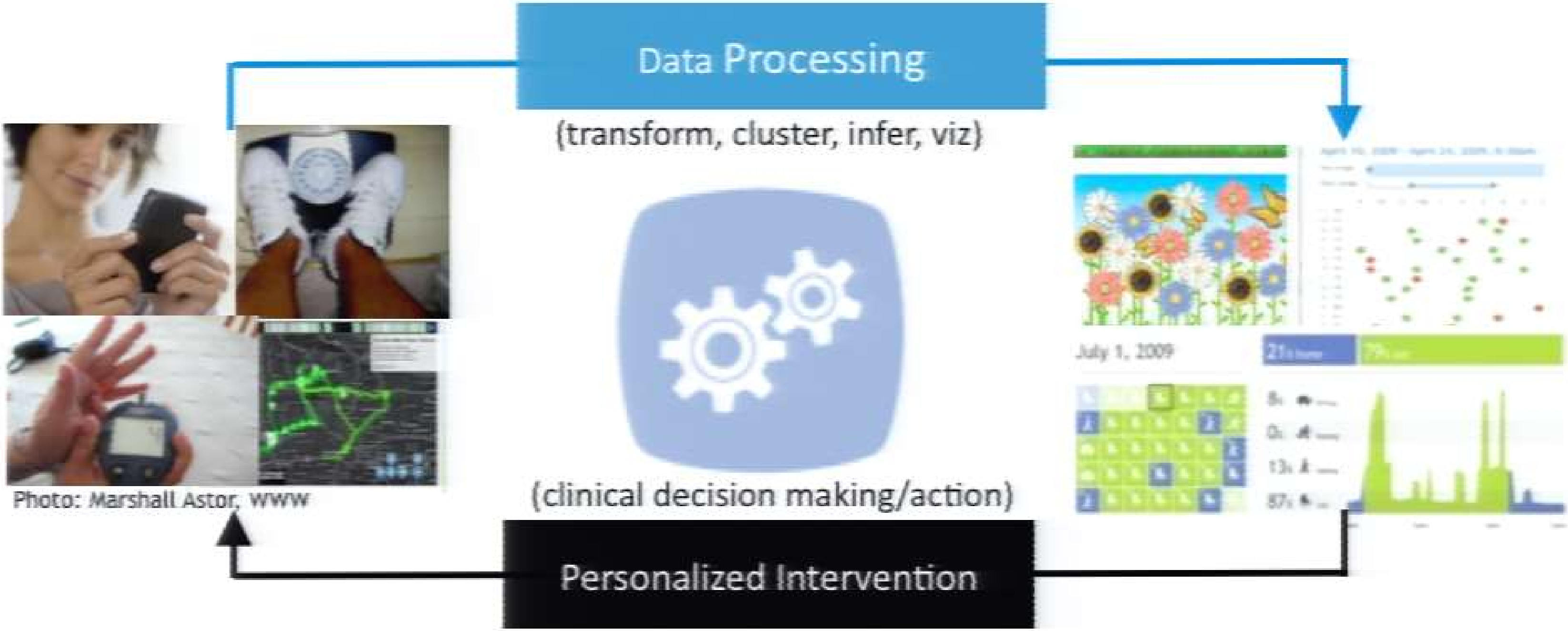
Accelerometry

GPS Data



Ambient Wi-Fi Signals

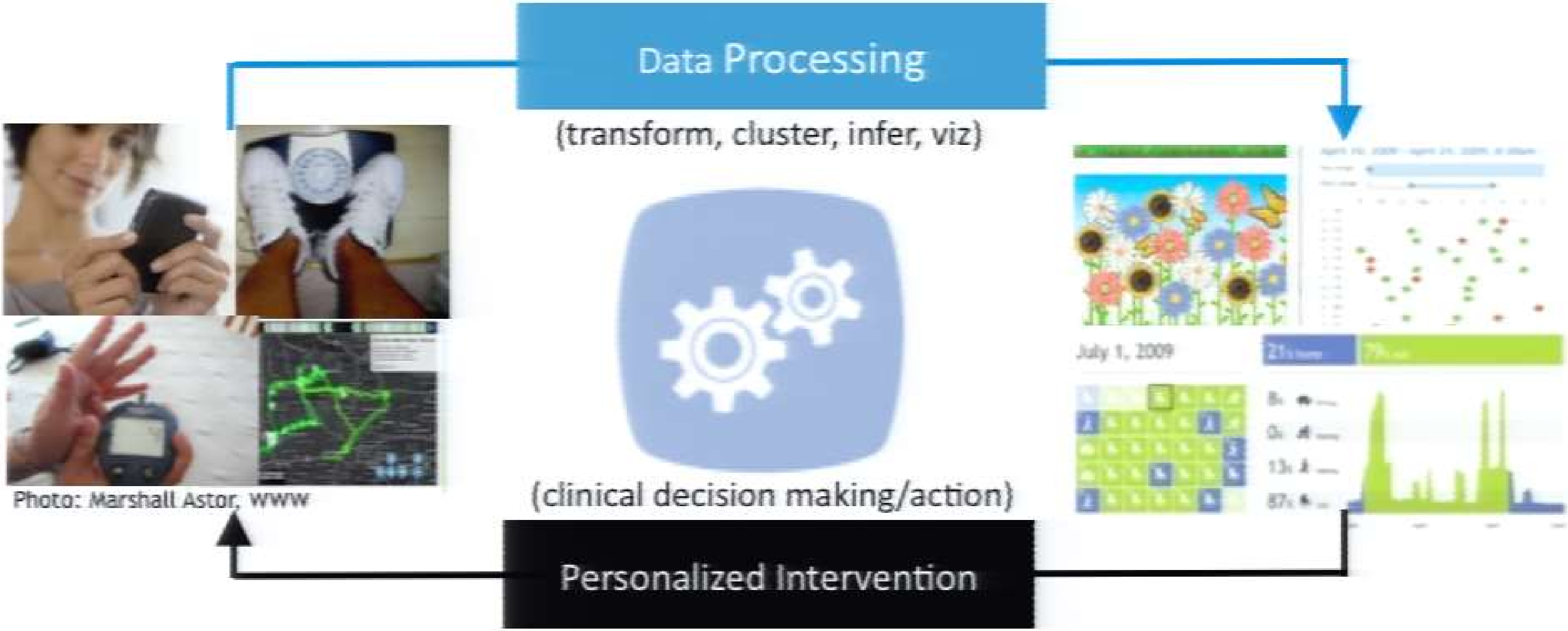

Data Processing
(transform, cluster, infer, viz)
(clinical decision making/action)
Personalized Intervention
July 1, 2009
Photo: Marshall Astor, www

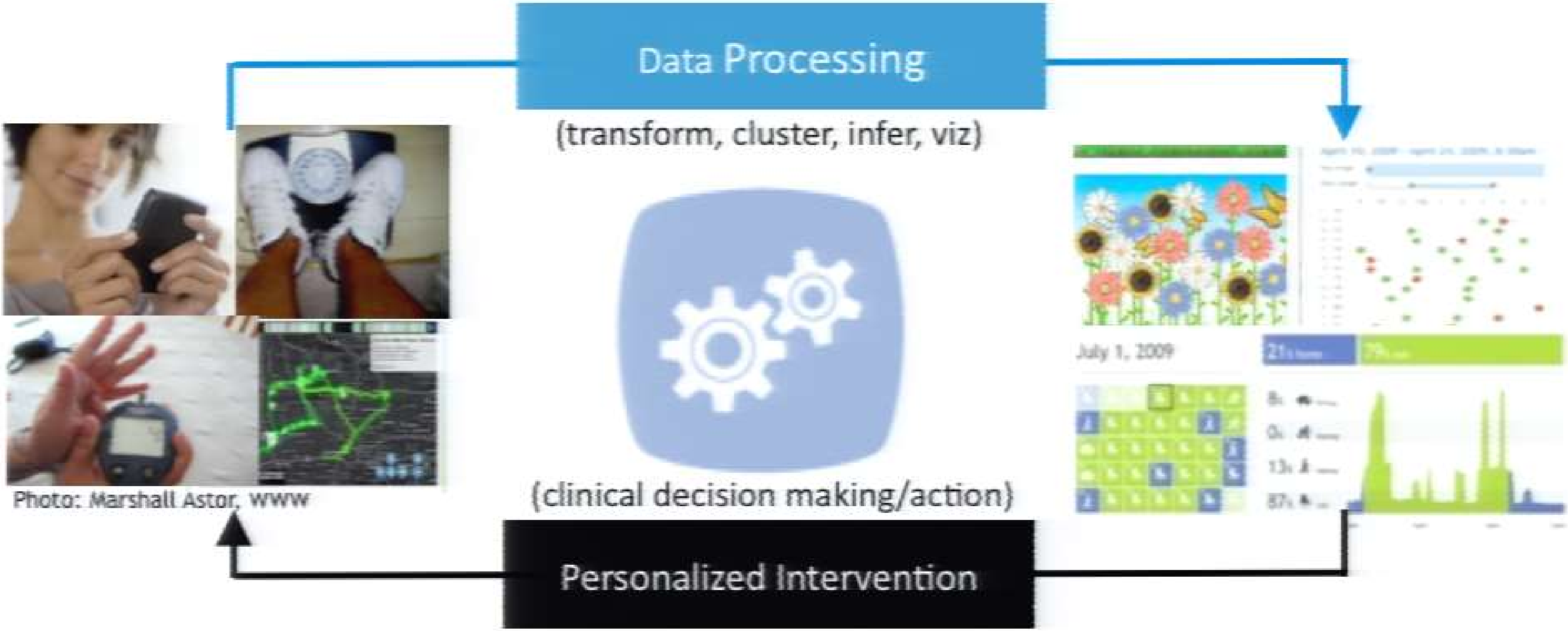


ultimate goal: create scalable tools for PCP and clinic based care that supports precision, personalization, and continuity of care in:

- RA, Lupus, Crohns, Asthma, MS
- Hospital discharge, Surgical recovery
- Pain management, Chronic fatigue, Migraines
- Depression, ADHD, insomnia, post-traumatic stress disorder
- Integrative medicine effectiveness
- Behavior change (individual, family , community)


Keynote File Edit Insert Slide Format Arrange View Play Window Share Help
Deborah Estrin
KaurKremerMullainathan SelfControl.pdf
registry.openmealth.org.documentation.wheel.graffle
KaurKremerMullainathan AERP&P2010.pdf
mPire One Page-092813-2.doc
AnnualReviewEdits4.pdf
mHealth-Greenhouse...813V2.docx
148 KB
Precommitment and Flexibility 1992.pdf
mPire One Page-092813.doc
99034103.pdf
20130913_173401 copy.pdf
13312.pdf
ML01-Introduction.pdf
PDF14978959.pdf
347.pdf
AJMC June 2013 UHG format.ppt.pptx
680 KB
raymerieee2003.pdf
Alshaban and Taher.pdf
clockwise-aug-29-2013.pdf


Data Processing
{transform, cluster, infer, viz}
Photo: Marshall Astor, WWW
July 1, 2009
{clinical decision making/action}
Personalized Intervention

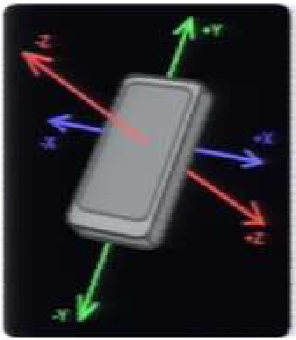


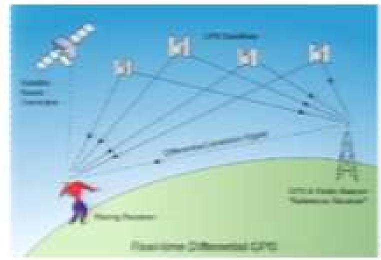
ultimate goal: create scalable tools for PCP and clinic based care that supports precision, personalization, and continuity of care in:

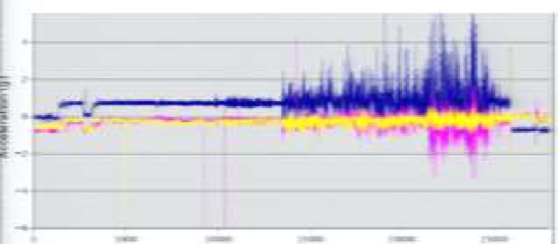
- RA, Lupus, Crohns, Asthma, MS
- Hospital discharge, Surgical recovery
- Pain management, Chronic fatigue, Migraines
- Depression, ADHD, insomnia, post-traumatic stress disorder
- Integrative medicine effectiveness
- Behavior change (individual, family , community)


Keynote File Edit Insert Slide Format Arrange View Play Window Share Help
Deborah Estrin
KaurKremerMullainathan SelfControl.pdf
registry.openmhealth.org.documentation.wheel.graffle
KaurKremerMullainathan AERP&P2010.pdf
mPire One Page-092813-2.doc
AnnualReviewEdits4.pdf
mHealth-Greenhouse...813V2.docx
148 KB
Precommitment and Flexibility 1992.pdf
mPire One Page-092813.doc
99034103.pdf
20130913_173401 copy.pdf
13312.pdf
ML01-Introduction.pdf
PDF14978959.pdf
347.pdf
AJMC June 2013 UHG format.ppt.pptx
680 KB
raymerieee2003.pdf
Alshaban and Taher.pdf
clockwise-aug-29-2013.pdf

Passively-recorded activity and location traces

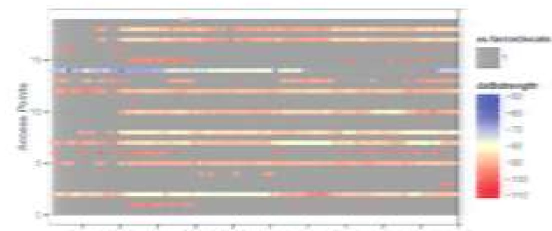






Accelerometry

GPS Data

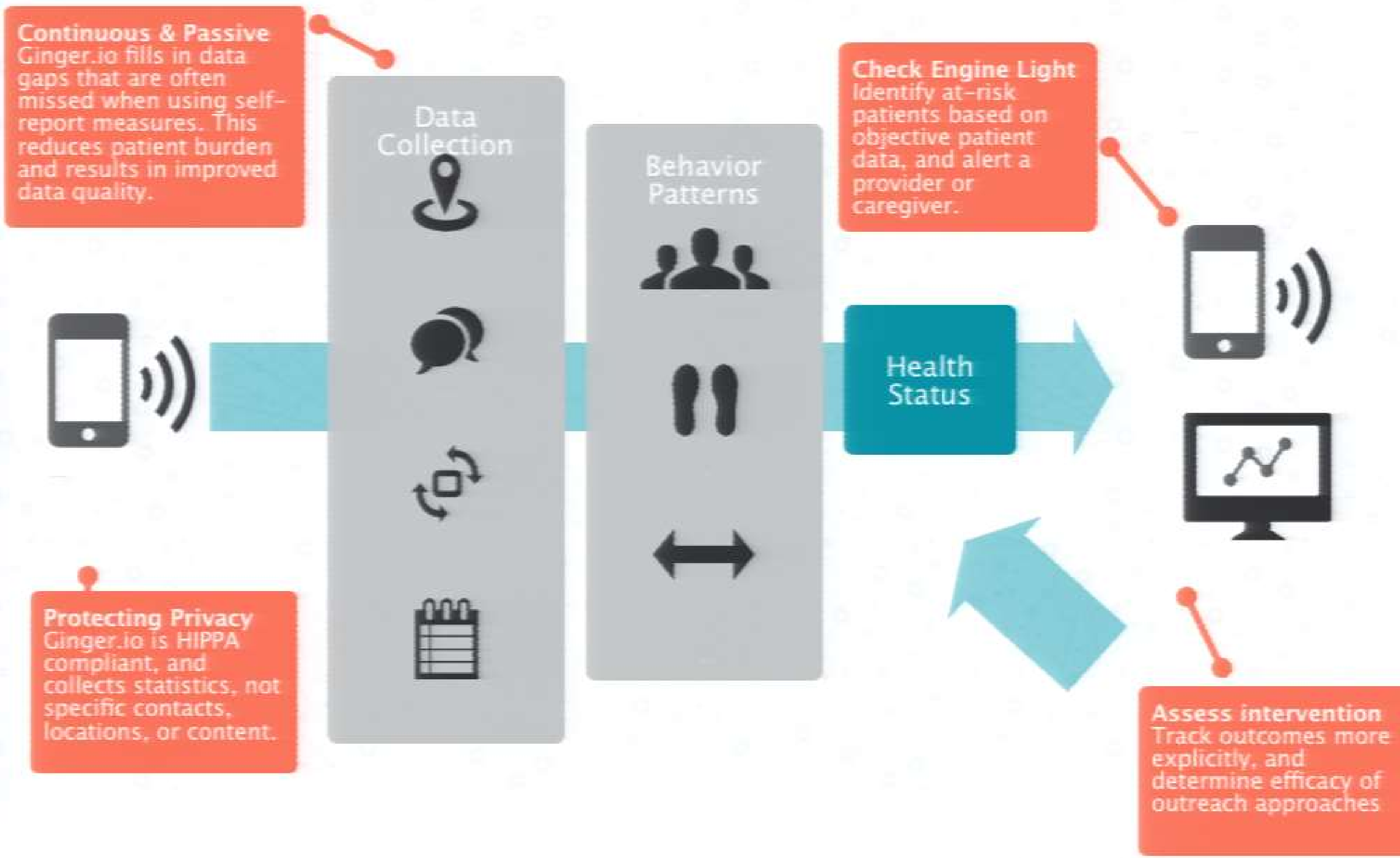


Ambient Wi-Fi Signals

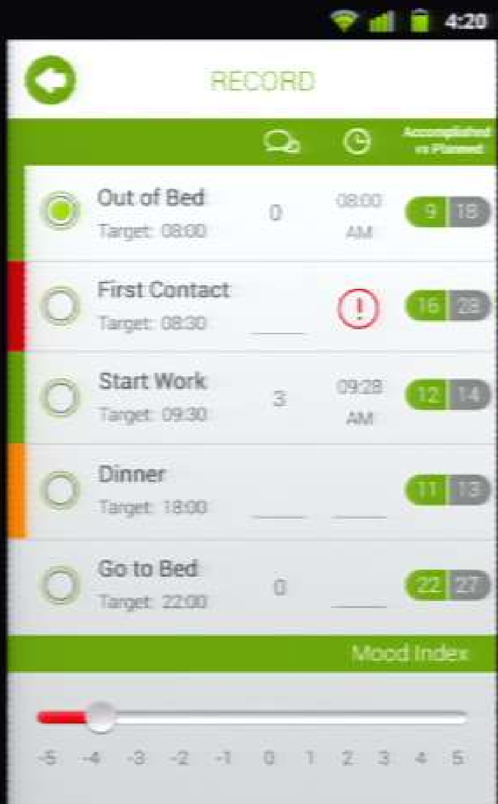
mobile apps data

Smart self report/EMA:
Photographic Affect Meter, PAM (Pollak et al)

Rich communication and activity data: The Ginger.io Platform



Data driven tools for patients: MoodRhytm(TM) (Choudhury et al)

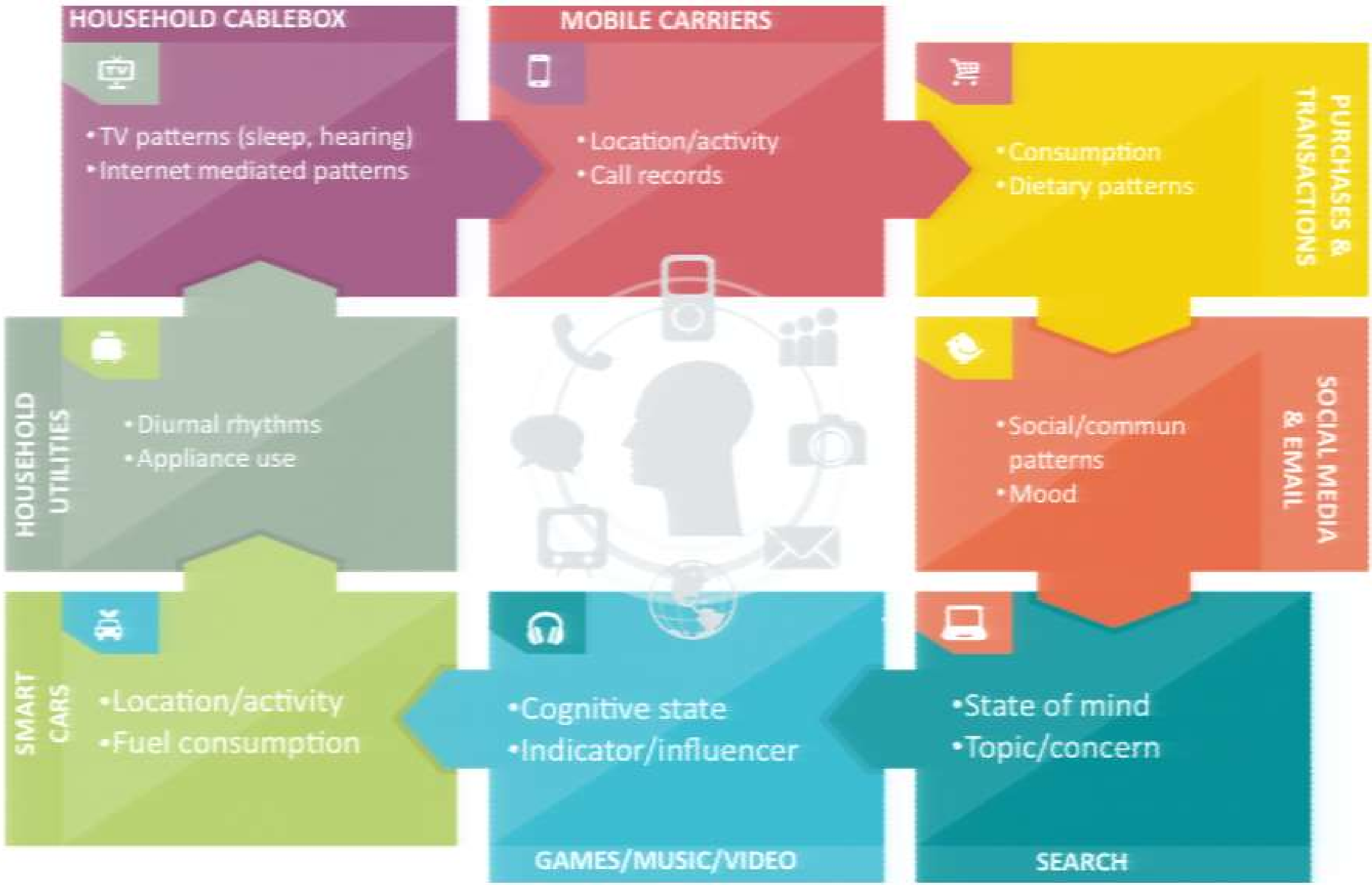


And yes…"Real Sensor" streams too

Beyond mobile...

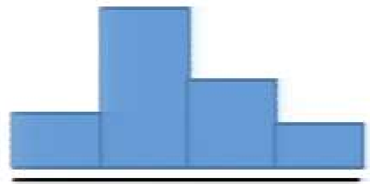
Leveraging digital traces from diverse consumer services

Your rows of their matrices...



Example: consumer transaction patterns as small data

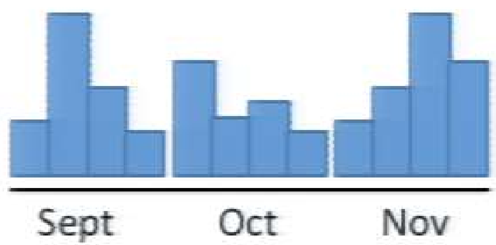
The goal is to transform purchasing patterns…



…into a consumption model (a categorical distribution of restaurant, fast food, drug store, etc. purchases)…

…on which descriptive statistics can be computed.

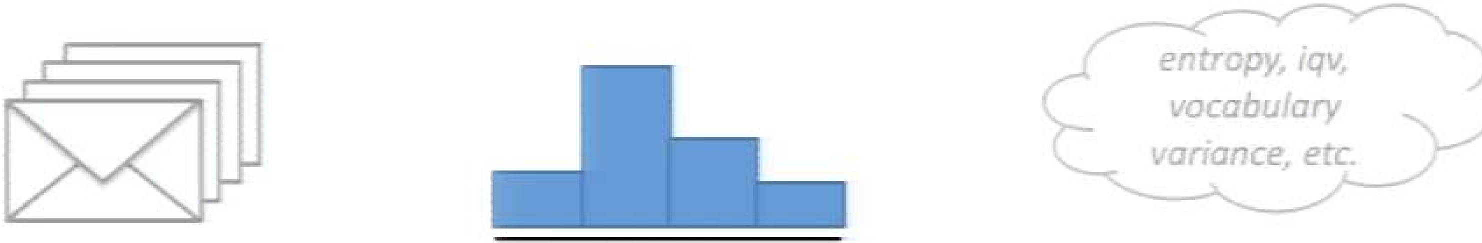
Since we have a progression of spending patterns over time…



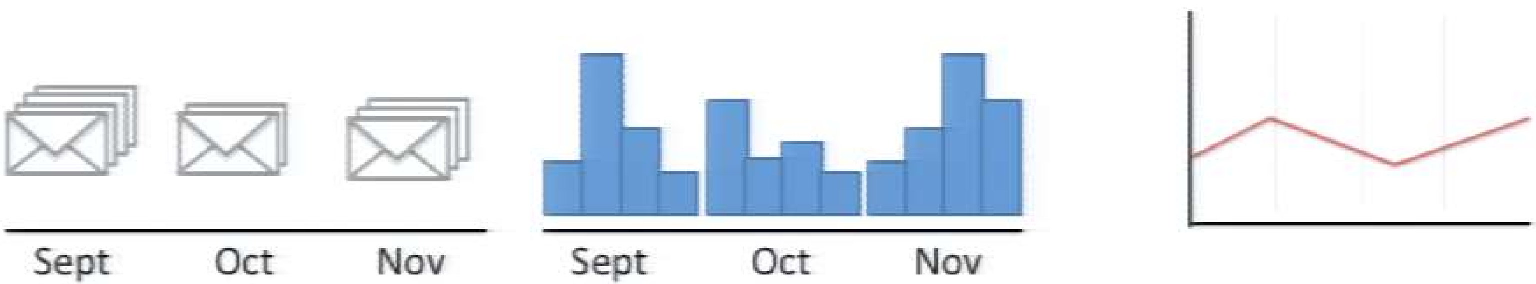
…we can examine how the resulting distributions change…

…and plot the statistics for each time frame, resulting in a time series which can then be correlated against other time series.

Example: communication language patterns as small data



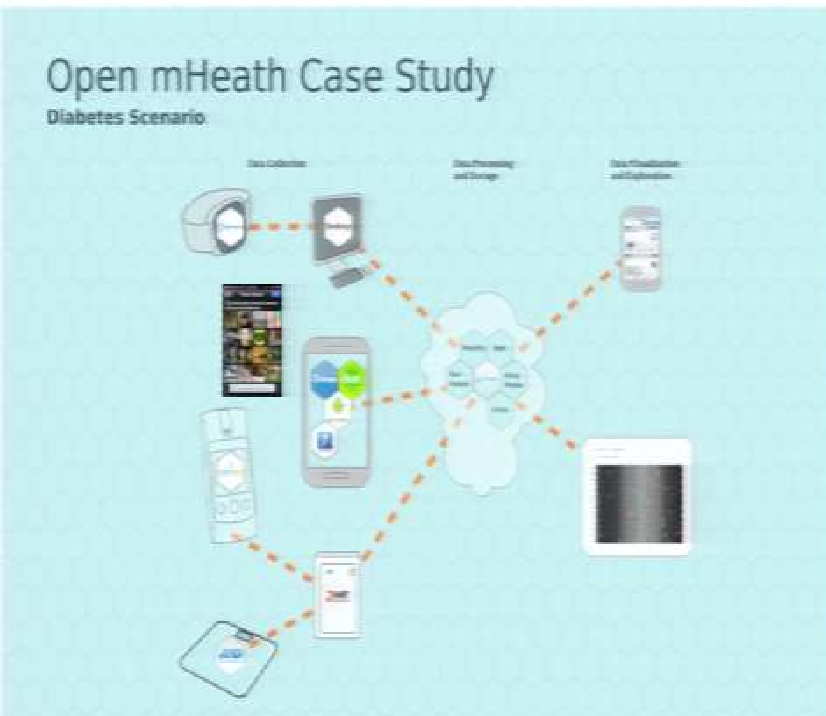
The goal is to transform text communication… …into a "bag of words" model (a categorical distribution)… …on which descriptive statistics can be computed.

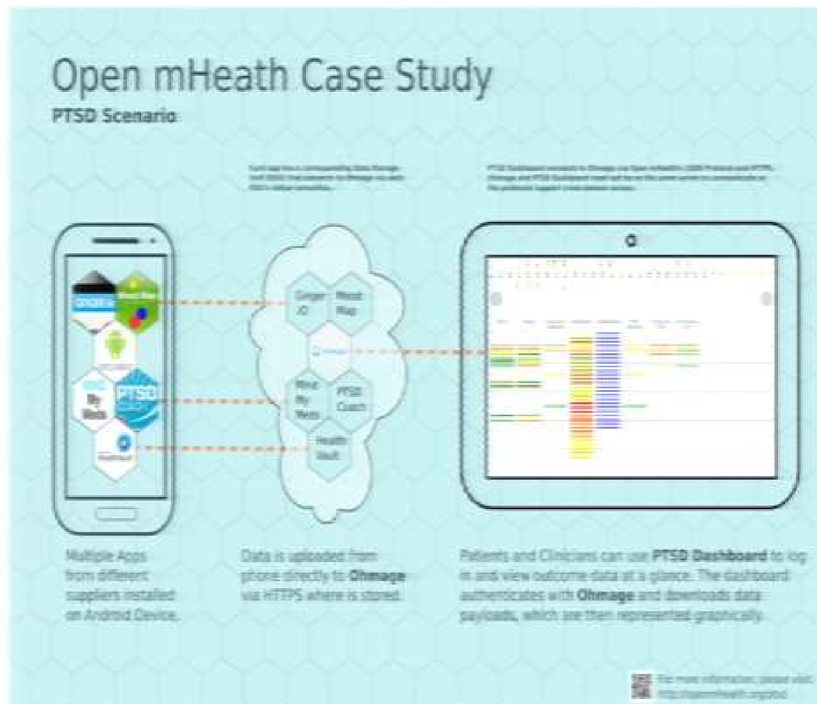


Since we have a progression of emails over time… …we can examine how the resulting distributions change… …and plot the statistics for each time frame, resulting in a time series which can then be correlated against other time series.

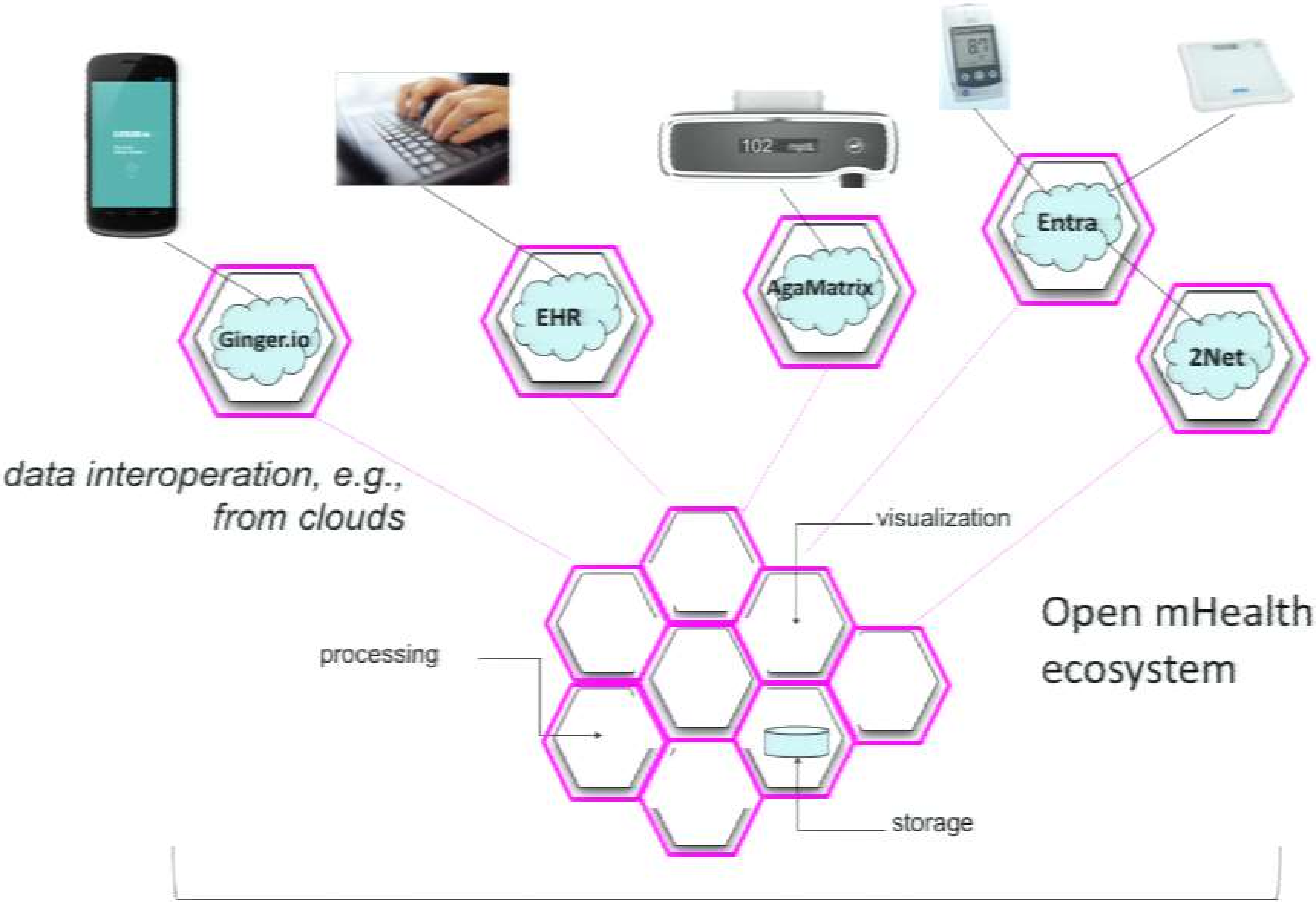
no one data stream tells the story

its about integration, fusion, and sense-making





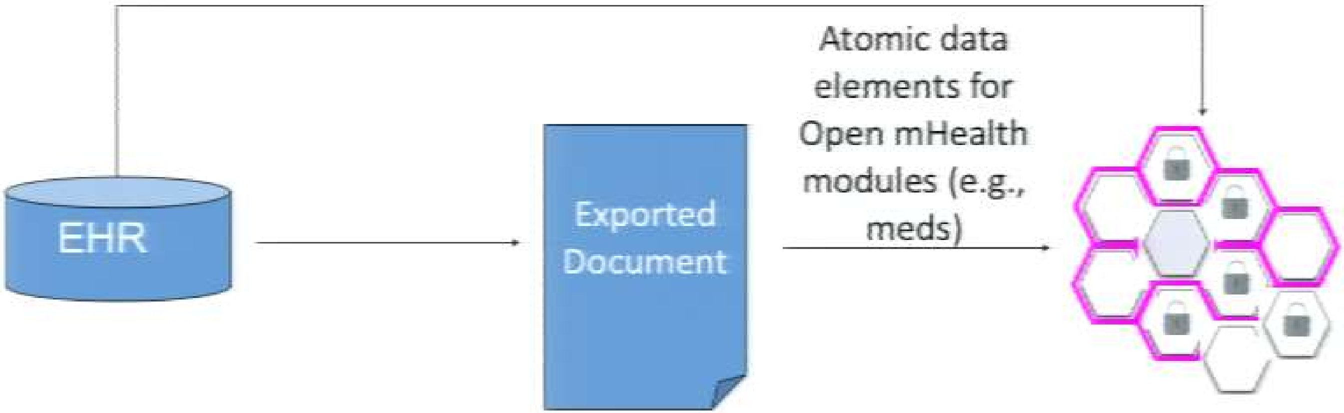
promote modularity, data integration, and software reusability



mobile integrated 'solutions/systems'

Eventual integration w/ clinical data sources and workflows via EHR

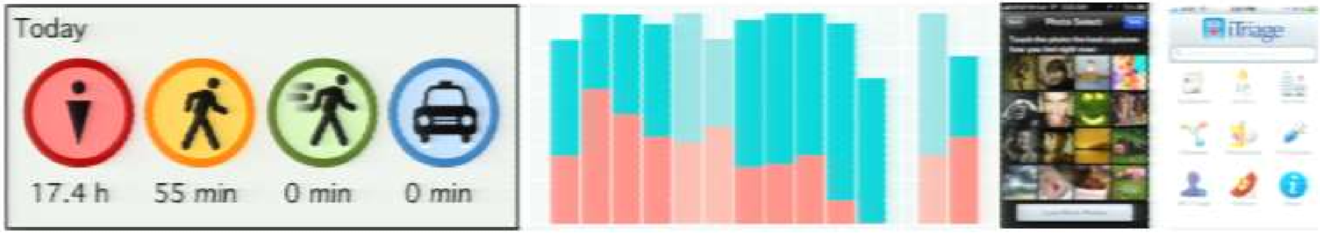
data *from* the EHR



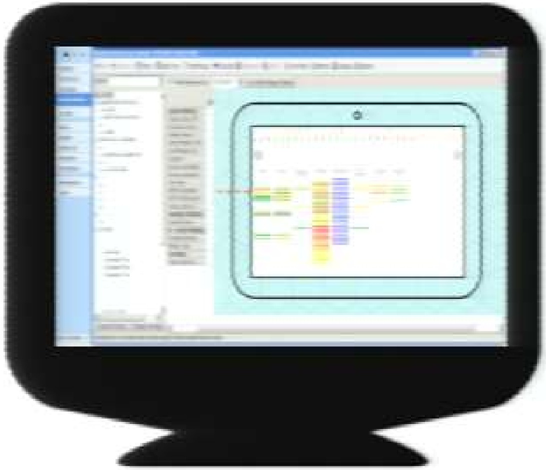
data *into* the EHR

Ida Sim, MD, PhD, UCSF

Key challenge 1: making "clinical sense" out of raw n=me data



???



Transform raw data streams into ***behavioral biomarkers***:

specific behavioral traits to measure progress of disease and treatment

state classification

- sedentary/ambulatory
- at home/work
- app analytics (games, media...)
- communication

Transform raw data streams into ***behavioral biomarkers**:*

specific behavioral traits to measure progress of disease and treatment

summarization

- ambulatory/sedentary cumulative and durations,walking speed
- sleep times, meal times
- time spent key locations, diameter of day
- social interaction

↑

state classification

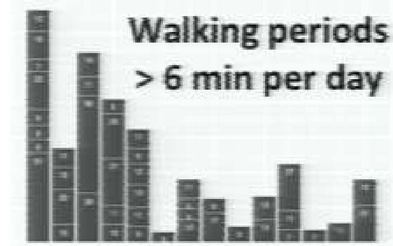
- sedentary/ambulatory
- at home/work
- app analytics (games, media...)
- communication

Transform raw data streams into ***behavioral biomarkers***:

specific behavioral traits to measure progress of disease and treatment

behavioral biomarker

- individual's patterns; relevance is symptom and condition dependent
- 'function, fatigue, pain, depression, insomnia, cognition, self-medication...



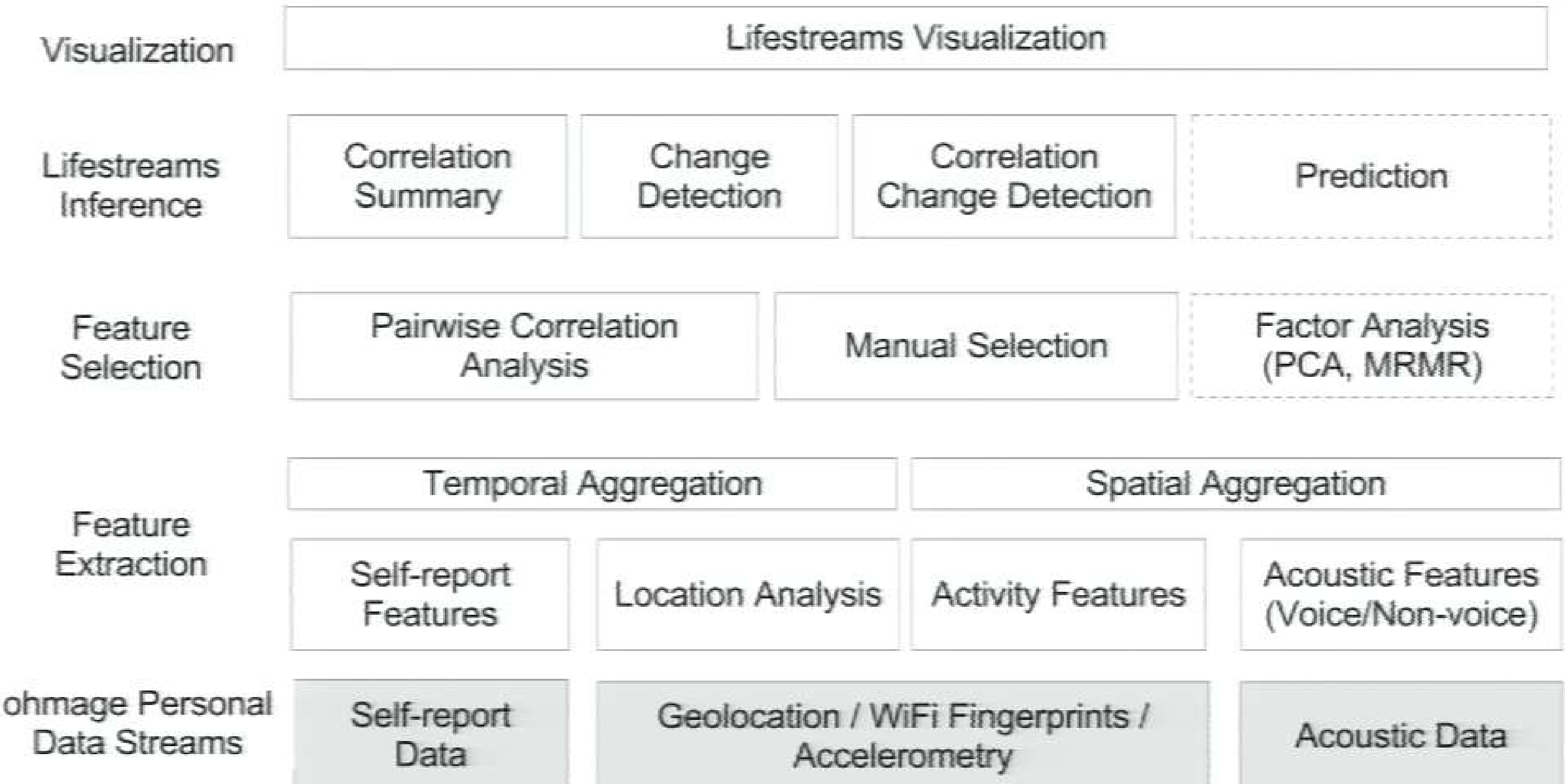
summarization

- ambulatory/sedentary cumulative and durations,walking speed
- sleep times, meal times
- time spent key locations, diameter of day
- social interaction

state classification

- sedentary/ambulatory
- at home/work
- app analytics (games, media...)
- communication

Modular tools to identify, iterate, share building blocks for behavioral biomarkers, sensemaking



(*Lifestreams, Hsieh et al, Sensys 2013*)

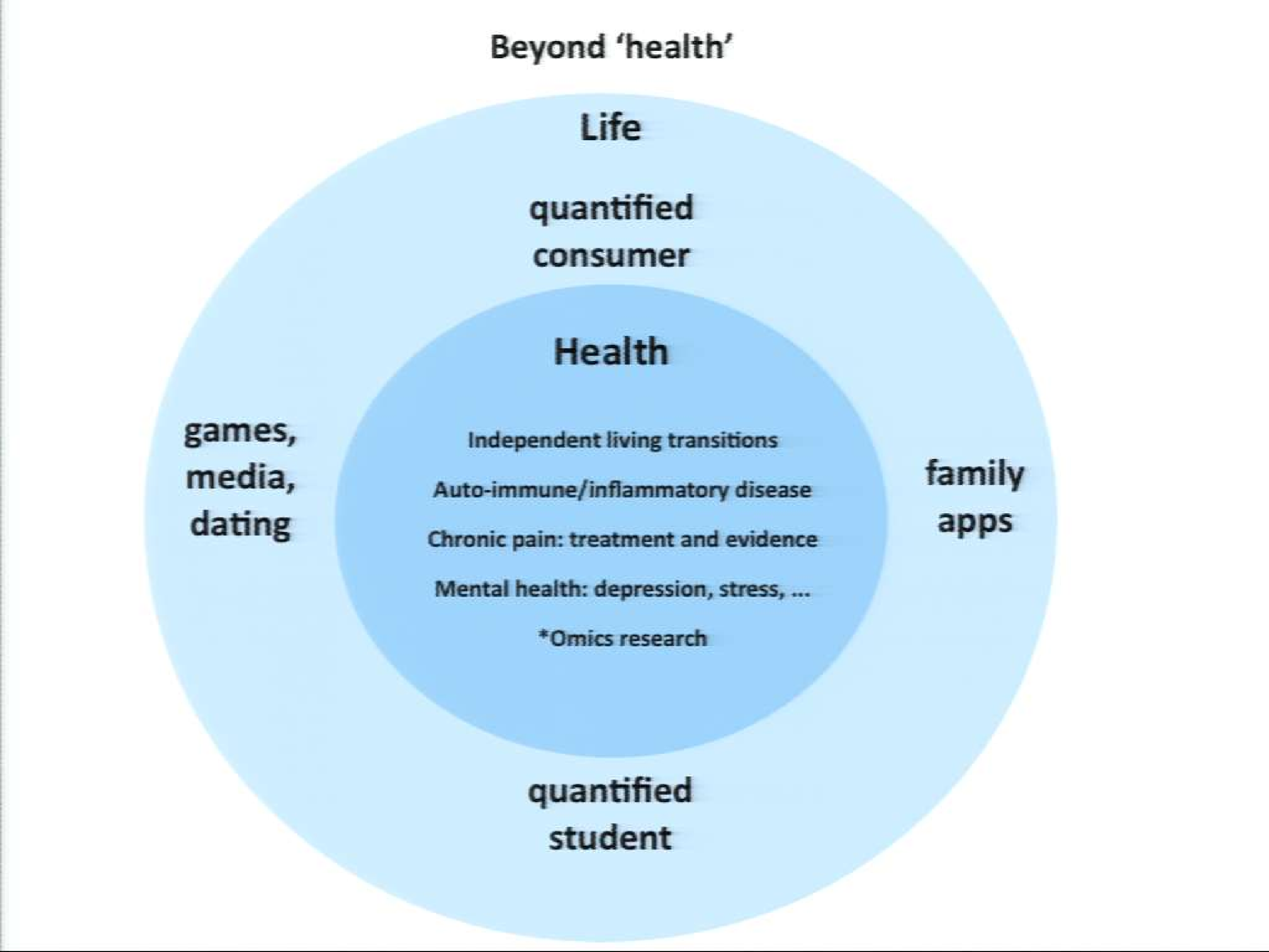
Hsieh, Tangmunarunkit, Ramanathan, et al

Key challenge 2: small data governance

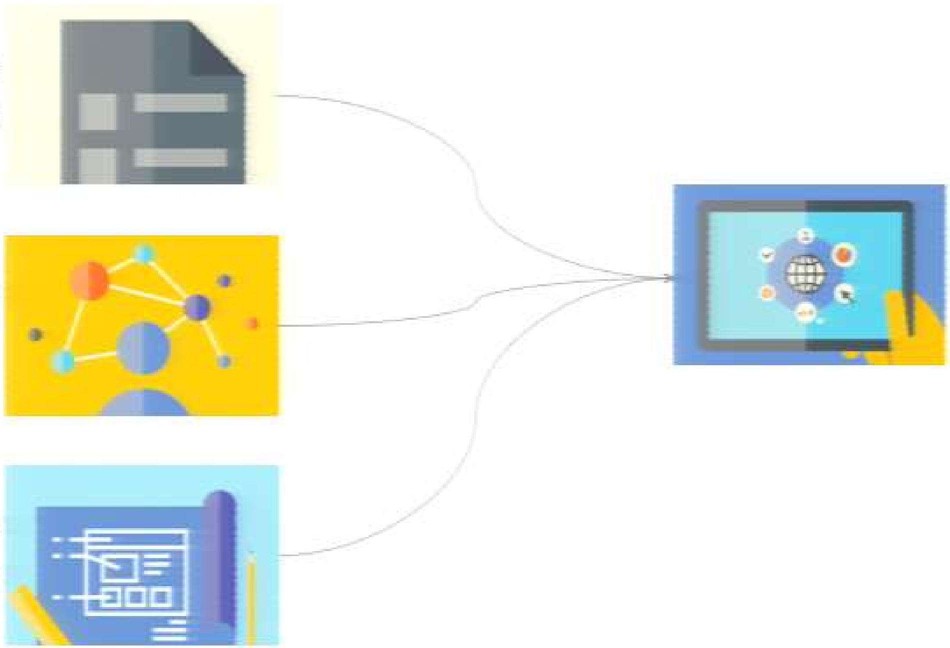
- Each data source has shared/other origins
- Individual has control over their corpus of data streams to correlate, fuse
- App/service utility derives from lack of anonymity
- Selective sharing embodied in apps--TMI works both ways in clinical domains

Socio-technical architecture for small data:

individual as nexus of control


Beyond 'health'
Life
quantified consumer
Health
games, media, dating
Independent living transitions
Auto-immune/inflammatory disease
Chronic pain: treatment and evidence
Mental health: depression, stress, ...
*Omics research
family apps
quantified student

mpire: test-bed for small data and personal informatics



Liberate the Data

Build a community

Catalyze app development

Start doing – real people creating & using micro-apps

Partners

TIME WARNER CABLE

at&t

verizon wireless

intel

AMERICAN EXPRESS

IBM Research

It’s not rocket science....

It’s pocket science

A Tech campus for the 21st century

A campus that embraces and embeds external engagement

- Between technology user and technology creator: co-innovation
    - Connective media, Healthier life, Built environment
- Between the academic and the non-academic worlds
    - Large and small businesses, Startups, Government, Non-profits

NIPS

Scalable Influence Estimation in Continuous-Time Diffusion Networks

Nan Du [1]

Joint work with Le Song [1], Manuel Gomez Rodriguez[2] and Hongyuan Zha[1]

[1]Georgia Institute of Technology

[2]Max Planck Institute for Intelligent Systems

Motivation

- Diffusions of news, events, and virus, take place over social and information networks.

Page 2 of 49

Motivation

- Diffusions of news, events, and virus, take place over social and information networks.
- Influence Estimation : how to predict how many people will follow the fashion lead by the influential users ?

Motivation

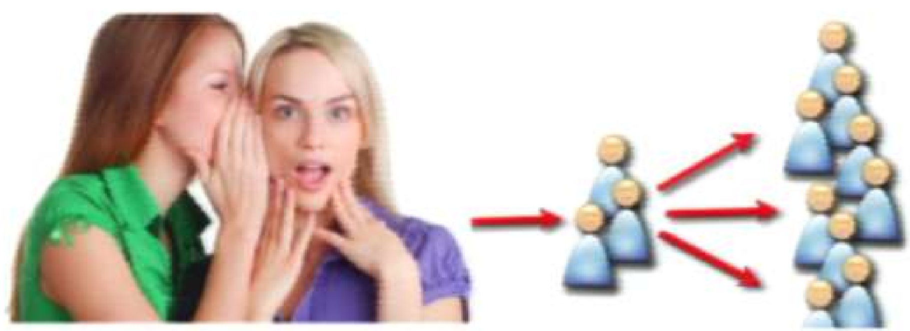
- Diffusions of news, events, and virus, take place over social and information networks.
- Influence Estimation : how to predict how many people will follow the fashion lead by the influential users ?
- Influence Maximization : how to identify such influential users to trigger the largest expected number of follow-ups ?

Motivation

- Diffusions of news, events, and virus, take place over social and information networks.
- Influence Estimation : how to predict how many people will follow the fashion lead by the influential users ?
- Influence Maximization : how to identify such influential users to trigger the largest expected number of follow-ups ?
- Time-Sensitive : influence most users before time T.

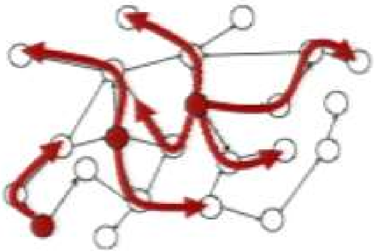
Motivation

- Diffusions of news, events, and virus, take place over social and information networks.
- Influence Estimation : how to predict how many people will follow the fashion lead by the influential users ?
- Influence Maximization : how to identify such influential users to trigger the largest expected number of follow-ups ?
- Time-Sensitive : influence most users before time T.
- Scalability : deal with large networks (millions of nodes) in practice.



time-sensitive viral marketing

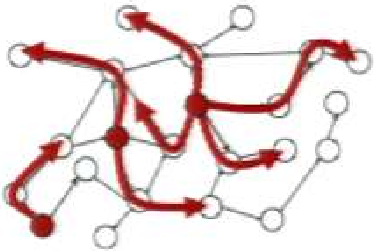
1 Continuous-time diffusion process.



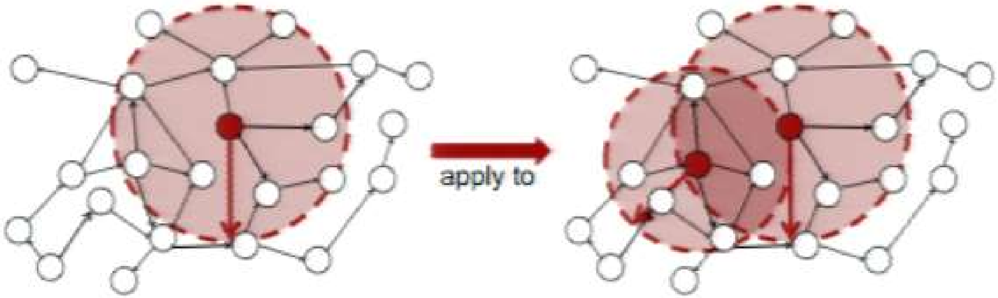
Page 7 of 49

Outline

1 Continuous-time diffusion process.



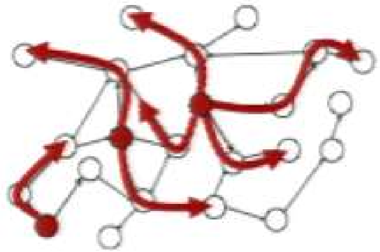
2 Efficient influence estimation and maximization.



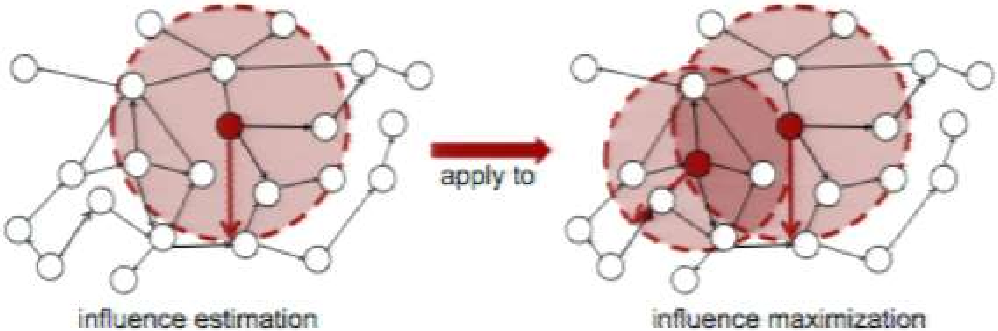
Page 8 of 49

OUTLINE

1 Continuous-time diffusion process.



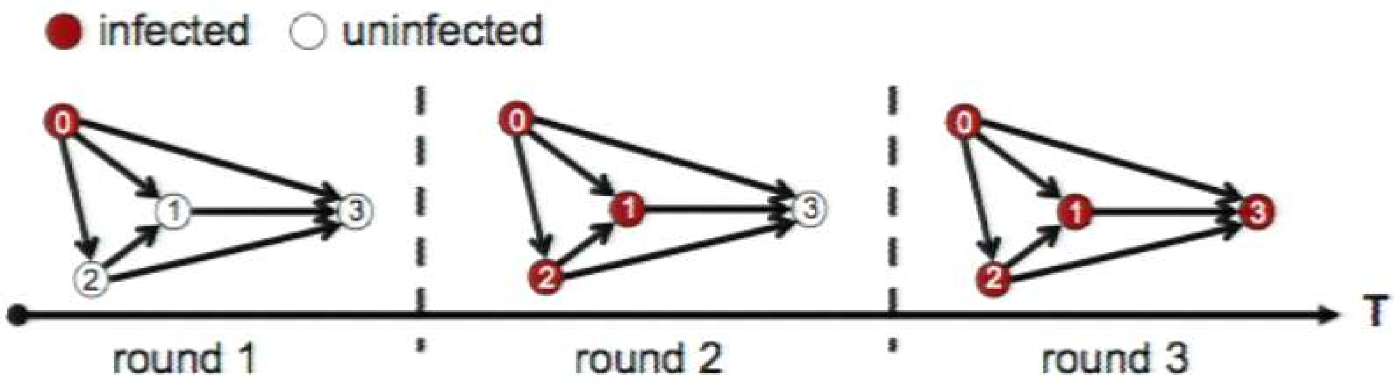
2 Efficient influence estimation and maximization.



3 Experimental evaluation with synthetic and real diffusion data.

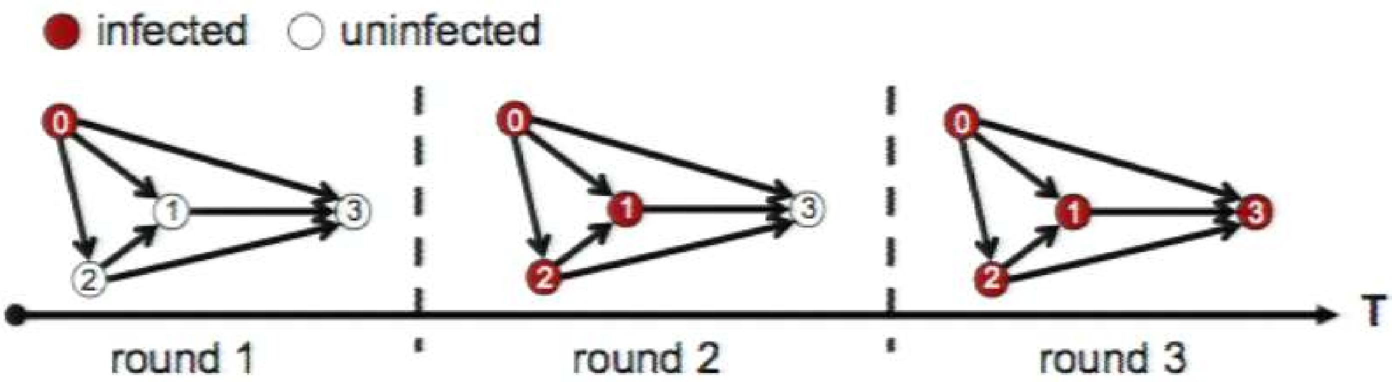
Continuous vs. Discrete Time Diffusion Model

- Traditionally, diffusion has been modeled as discrete steps (or rounds).



Continuous vs. Discrete Time Diffusion Model

- Traditionally, diffusion has been modeled as discrete steps (or rounds).



- In reality, propagation does not go in rounds !
  - how long is each round ?
  - how many rounds do we need ?

Continuous-time Independent Cascade Model

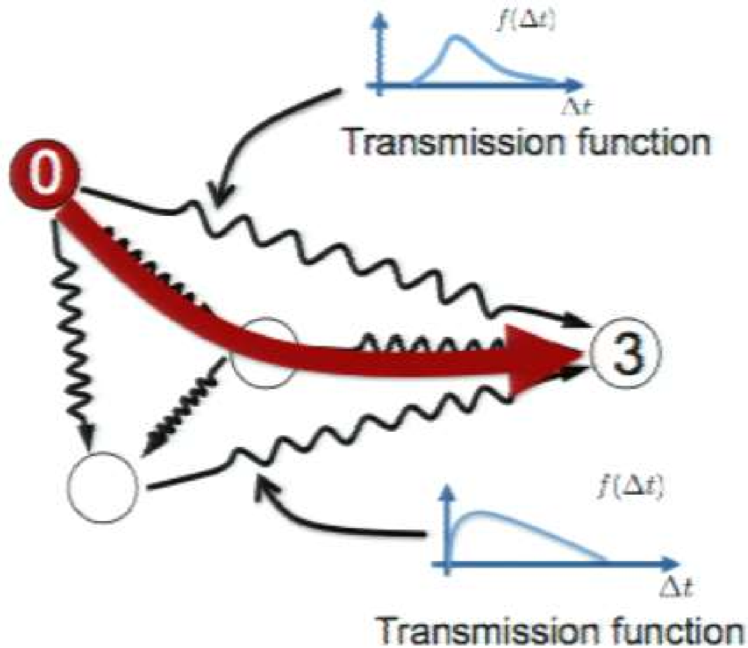
- Model mutually independent transmission time

$$\tau_{ji} = t_i - t_j.$$

- Pairwise conditional density (transmission function)

$$f_{ji}(t_i | t_j) = f_{ji}(t_i - t_j).$$

- A network with stochastic edge weights.
- Infection time t_i = length of shortest path.



Absolute Infection Time View

- The influence of sources $\mathcal{A}$ by time T is

$$\sigma(\mathcal{A}, T) = \mathbb{E}\left[\sum\nolimits_{i \in \mathcal{V}} \mathbb{I}\{t_i \leq T\}\right] = \sum\nolimits_{i \in \mathcal{V}} \Pr\{t_i \leq T\}$$

Absolute Infection Time View

- The influence of sources $\mathcal{A}$ by time T is

$$\sigma(\mathcal{A}, T) = \mathbb{E}\left[\sum_{i\in\mathcal{V}} \mathbb{I}\{t_i \leq T\}\right] = \sum_{i\in\mathcal{V}} \Pr\{t_i \leq T\}$$

- Infection probability

$$\Pr\{t_i \leq T\} = \int_0^\infty \cdots \int_{t_i=0}^T \cdots \int_0^\infty \left(\prod_{j\in\mathcal{V}} p\left(t_j | \{t_l\}_{l\in\pi_j}\right)\right)\left(\prod_{j\in\mathcal{V}} dt_j\right)$$

Absolute Infection Time View

- The influence of sources $\mathcal{A}$ by time T is

$$\sigma(\mathcal{A}, T) = \mathbb{E}\left[\sum\nolimits_{i\in\mathcal{V}} \mathbb{I}\{t_i \leq T\}\right] = \sum\nolimits_{i\in\mathcal{V}} \Pr\{t_i \leq T\}$$

- Infection probability

$$\Pr\{t_i \leq T\} = \int_0^\infty \cdots \int_{t_i=0}^{T} \cdots \int_0^\infty \left(\prod\nolimits_{j\in\mathcal{V}} p\left(t_j | \{t_l\}_{l\in\pi_j}\right)\right) \left(\prod\nolimits_{j\in\mathcal{V}} dt_j\right)$$

  - Need to integrate all possible configurations of cascades where $t_i < T$.
  - No closed form solution for general heterogeneous transmission function.
  - Hard to approximate.

Neighborhood Size Estimation

- Influence function

$$\sigma(\mathcal{A}, T) = \sum\nolimits_{i \in \mathcal{V}} \Pr\{t_i \leq T\}$$

- No need to calculate $\Pr\{t_i \leq T\}$ individually.

Page 16 of 49

Neighborhood Size Estimation

- Influence function

$$\sigma(\mathcal{A}, T) = \sum\nolimits_{i \in \mathcal{V}} \Pr\{t_i \leq T\}$$

- No need to calculate $\Pr\{t_i \leq T\}$ individually.

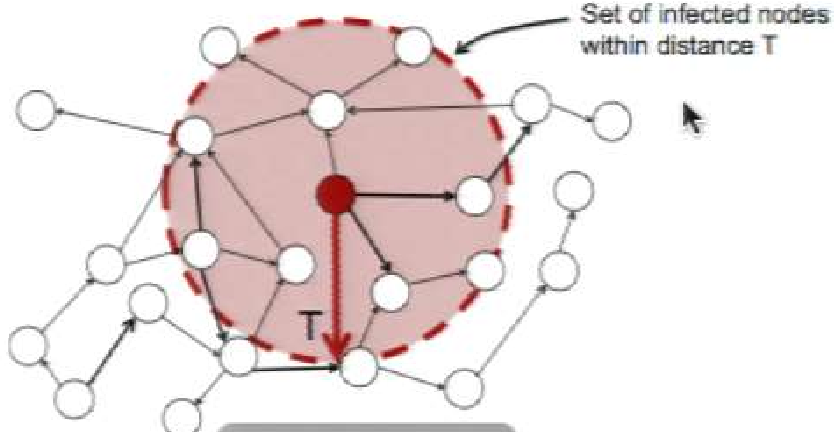
Neighborhood Size Estimation

- Influence function

$$\sigma(\mathcal{A}, T) = \sum\nolimits_{i\in\mathcal{V}} \Pr\{t_i \le T\}$$

- No need to calculate $\Pr\{t_i \le T\}$ individually.
- Given a set of $\{\tau_{ji}^{l}\}_{(j,i)\in\mathcal{E}}$, only care about

$$\sum\nolimits_{i\in\mathcal{V}} \mathbb{I}\{t_i \le T\} = |\mathcal{N}(\{j\}, T)| = |\{i : t_i \leqslant T\}|$$



influence estimation

Neighborhood Size Estimation

- Sample n sets of $G_l := \{\tau_{ji}^l\}_{(j,i)\in\mathcal{E}} \sim \prod_{(j,i)\in\mathcal{E}} f_{ji}(\tau_{ji})$

Neighborhood Size Estimation

- Sample n sets of $G_l := \{\tau_{ji}^l\}_{(j,i)\in\mathcal{E}} \sim \prod_{(j,i)\in\mathcal{E}} f_{ji}(\tau_{ji})$
- Average the counts across n samples.

$$\begin{aligned}\sigma(\mathcal{A}, T) &= \mathbb{E}\left[\sum\nolimits_{i\in\mathcal{V}} \mathbb{I}\{t_i \le T\}\right] \\ &\approx \frac{1}{n}\left(\sum_{i\in\mathcal{V}} \mathbb{I}\{t_i \le T | G_1\} +, \dots, + \sum_{i\in\mathcal{V}} \mathbb{I}\{t_i \le T | G_n\}\right)\end{aligned}$$

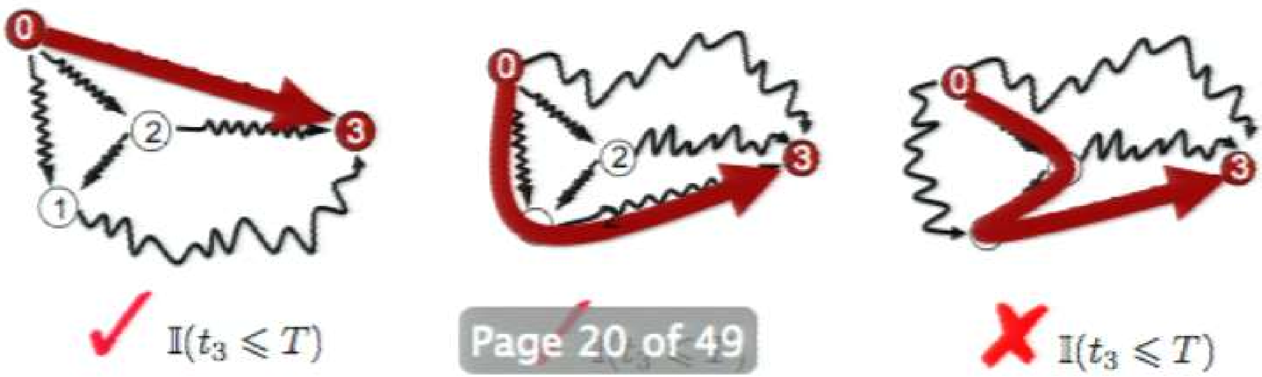
Page 19 of 49

Neighborhood Size Estimation

- Sample n sets of $G_l := \{\tau_{ji}^l\}_{(j,i)\in\mathcal{E}} \sim \prod_{(j,i)\in\mathcal{E}} f_{ji}(\tau_{ji})$
- Average the counts across n samples.

$$\sigma(\mathcal{A}, T) = \mathbb{E}\left[\sum\nolimits_{i\in\mathcal{V}} \mathbb{I}\{t_i \leq T\}\right]$$
$$\approx \frac{1}{n}\left(\sum_{i\in\mathcal{V}} \mathbb{I}\{t_i \leq T | G_1\} +, \ldots, + \sum_{i\in\mathcal{V}} \mathbb{I}\{t_i \leq T | G_n\}\right)$$

- To calculate $\mathbb{I}\{t_i \leq T | G_l\}$, check whether length of shortest path $\leqslant T$ on each sampled network.

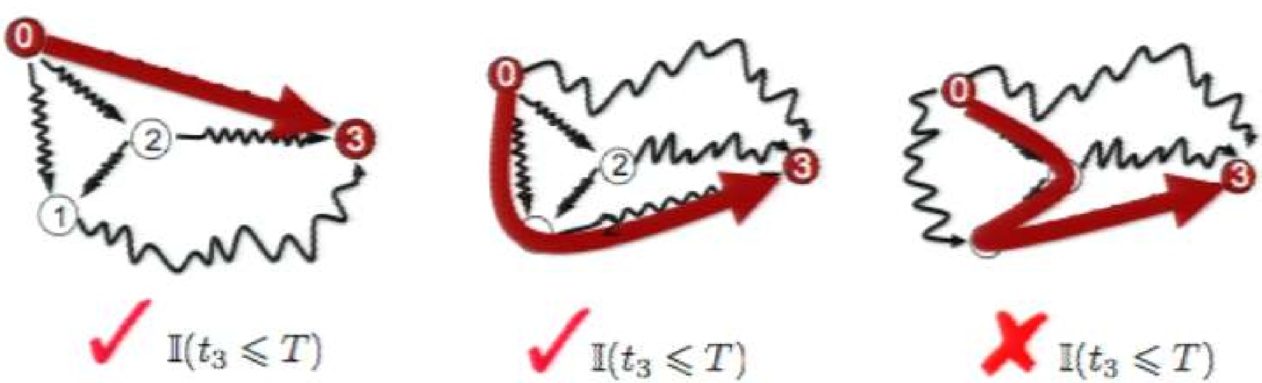


Neighborhood Size Estimation

- Sample n sets of $G_l := \{\tau_{ji}^l\}_{(j,i)\in\mathcal{E}} \sim \prod_{(j,i)\in\mathcal{E}} f_{ji}(\tau_{ji})$
- Average the counts across n samples.

$$\sigma(\mathcal{A}, T) = \mathbb{E}\left[\sum\nolimits_{i\in\mathcal{V}} \mathbb{I}\{t_i \leq T\}\right]$$
$$\approx \frac{1}{n}\left(\sum_{i\in\mathcal{V}} \mathbb{I}\{t_i \leq T | G_1\} +, \ldots, + \sum_{i\in\mathcal{V}} \mathbb{I}\{t_i \leq T | G_n\}\right)$$

- To calculate $\mathbb{I}\{t_i \leq T | G_l\}$, check whether length of shortest path $\leqslant T$ on each sampled network.



Neighborhood Size Estimation

- Using shortest path is not scalable.

Page 21 of 49

Neighborhood Size Estimation

- Using shortest path is not scalable.
- Influence Estimation of a single source j
  - $\sigma(\{j\}, T)$
  - Compute all shortest paths from j to the other nodes.

Neighborhood Size Estimation

- Using shortest path is not scalable.
- Influence Estimation of a single source j
  - $\sigma(\{j\}, T)$
  - Compute all shortest paths from j to the other nodes.
- Which source is the best ?
  - Chose j with the largest $\sigma(\{j\}, T)$
  - Try source $j = 0, \ldots, |\mathcal{V}| - 1$, $O(|\mathcal{V}|^2)$
- Quadratic in network size

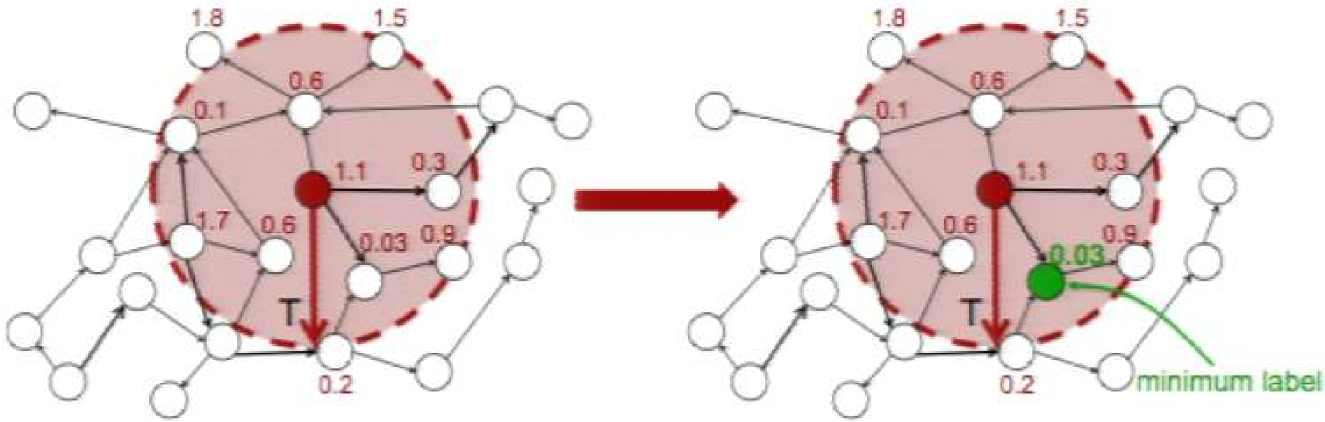
  Can not deal with large networks !

Neighborhood Size Estimation

- Directly estimate the neighborhood size by Cohen's algorithm !

Neighborhood Size Estimation

- Directly estimate the neighborhood size by Cohen's algorithm !

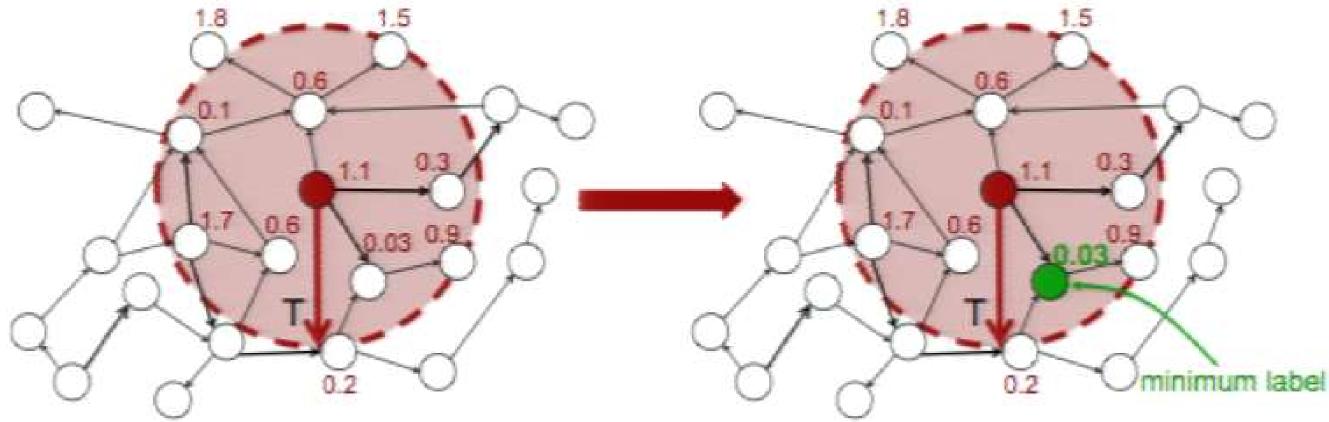


assign exponential random label r_i find the minimum label r_*

- Draw m sets of i.i.d random labels $\{r_i^u\}_{u=1}^m \backsim e^{-r_i}$.

Neighborhood Size Estimation

- Directly estimate the neighborhood size by Cohen's algorithm !

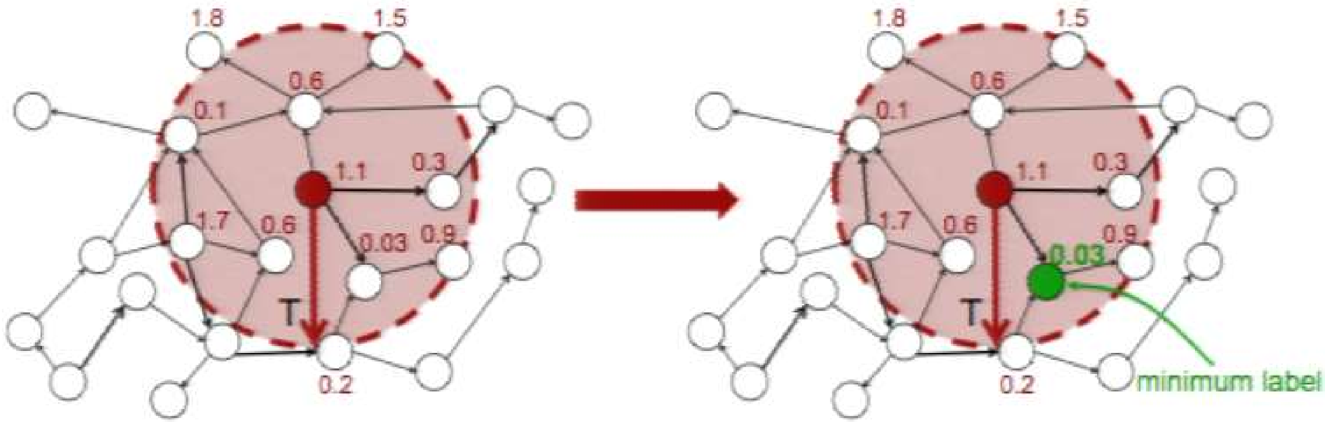


assign exponential random label r_i find the minimum label r_*

- Draw m sets of i.i.d random labels $\{r_i^u\}_{u=1}^m \backsim e^{-r_i}$.
- Find the minimum label $\{r_*^u\}_{u=1}^m$ within distance T by Cohen's algorithm in $\tilde{O}(|\mathcal{E}|)$.

Neighborhood Size Estimation

- Directly estimate the neighborhood size by Cohen's algorithm !



assign exponential random label r_i find the minimum label r_*

- Draw m sets of i.i.d random labels $\{r_i^u\}_{u=1}^m \backsim e^{-r_i}$.
- Find the minimum label $\{r_*^u\}_{u=1}^m$ within distance T by Cohen's algorithm in $\tilde{O}(|\mathcal{E}|)$.
- Estimate $|\mathcal{N}(\{j\}, T)| \approx \frac{m-1}{\sum_{u=1}^m r_*^u}$ using the property of exponential distribution.

Page 27 of 49

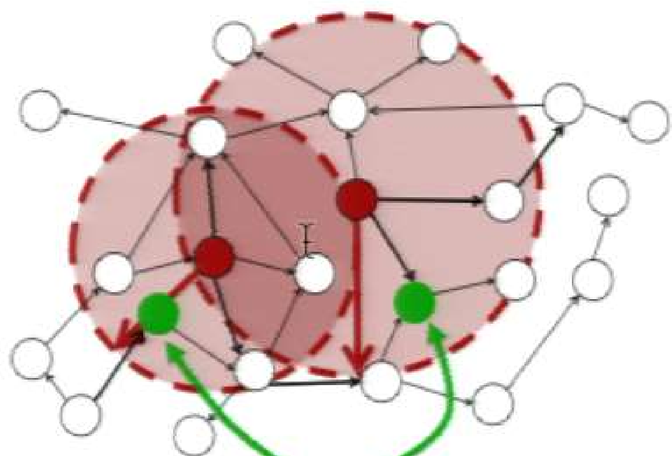
Multiple Sources

- Multiple sources $\mathcal{A}$

$$\mathcal{N}(\mathcal{A}, T) = \bigcup_{s \in \mathcal{A}} \mathcal{N}(s, T).$$

- The overall least label

$$r_* = \min_{i \in \mathcal{A}} \ \min_{j \in \mathcal{N}(i,T)} r_j$$



mini[illegible] least labels

Overall Algorithm ConTinEst

1. Sample n sets of random transmission times

$$\{\tau_{ji}^{l}\}_{(j,i)\in\mathcal{E}} \sim \prod_{(j,i)\in\mathcal{E}} f_{ji}(\tau_{ji})$$

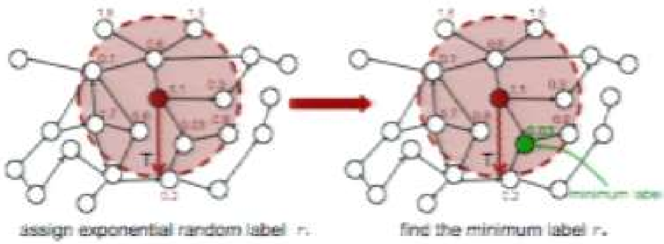
Page 29 of 49

Overall Algorithm ConTinEst

1. Sample n sets of random transmission times

$$\{\tau_{ji}^{l}\}_{(j,i)\in\mathcal{E}} \sim \prod_{(j,i)\in\mathcal{E}} f_{ji}(\tau_{ji})$$

2. Given a set of $\{\tau_{ji}^{l}\}_{(j,i)\in\mathcal{E}}$, sample m sets of random labels



$$\{r_i^u\}_{i\in\mathcal{V}} \sim \prod_{i\in\mathcal{V}} \exp(-r_i)$$

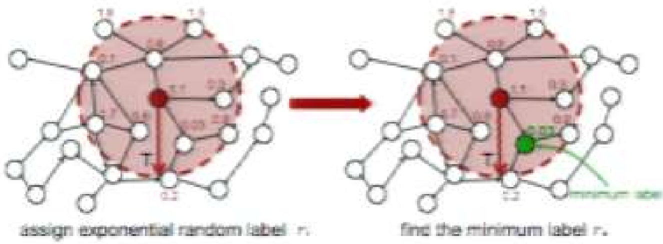
Page 30 of 49

Overall Algorithm ConTinEst

1. Sample n sets of random transmission times

$$\{\tau_{ji}^{l}\}_{(j,i)\in\mathcal{E}} \sim \prod_{(j,i)\in\mathcal{E}} f_{ji}(\tau_{ji})$$

2. Given a set of $\{\tau_{ji}^{l}\}_{(j,i)\in\mathcal{E}}$, sample m sets of random labels



$$\{r_i^u\}_{i\in\mathcal{V}} \sim \prod_{i\in\mathcal{V}} \exp(-r_i)$$

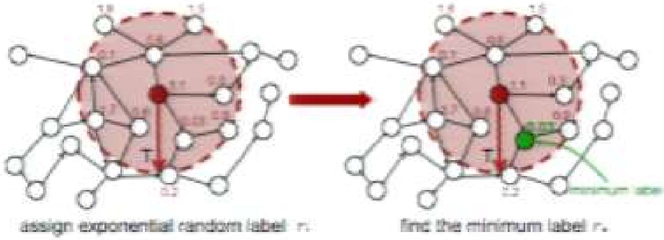
3. Find the minimum label $\{r_*^u\}_{u=1}^{m}$ within T using Cohen's algorithm.

Overall Algorithm ConTinEst

1. Sample n sets of random transmission times

$$\{\tau_{ji}^{l}\}_{(j,i)\in\mathcal{E}} \sim \prod_{(j,i)\in\mathcal{E}} f_{ji}(\tau_{ji})$$

2. Given a set of $\{\tau_{ji}^{l}\}_{(j,i)\in\mathcal{E}}$, sample m sets of random labels



$$\{r_i^u\}_{i\in\mathcal{V}} \sim \prod_{i\in\mathcal{V}} \exp(-r_i)$$

3. Find the minimum label $\{r_*^u\}_{u=1}^m$ within T using Cohen's algorithm.

4. Estimate $\sigma(\mathcal{A}, T)$ by sample averages

$$\sigma(\mathcal{A}, T) \approx \frac{1}{n}\sum_{l=1}^{n}\left((m-1)/\sum_{u_l=1}^{m} r_*^{u_l}\right)$$

Overall Algorithm ConTinEst

Theorem

Draw the following number of samples for the set of random transmission times

$$n \geqslant \frac{C\Lambda(T, 1/m)}{\epsilon^2} \log\left(\frac{2|\mathcal{V}|}{\delta}\right),$$

and for each set of random transmission times, draw m set of random labels. Then $|\widehat{\sigma}(\mathcal{A}, T) - \sigma(\mathcal{A}, T)| \leqslant \epsilon$ uniformly for all $\mathcal{A}$ with $|\mathcal{A}| \leqslant C$, with probability at least $1 - \delta$.

Page 33 of 49

Overall Algorithm ConTinEst

Theorem

Draw the following number of samples for the set of random transmission times

$$n \geqslant \frac{C\Lambda(T, 1/m)}{\epsilon^2} \log\left(\frac{2|\mathcal{V}|}{\delta}\right),$$

and for each set of random transmission times, draw m set of random labels. Then $|\widehat{\sigma}(\mathcal{A}, T) - \sigma(\mathcal{A}, T)| \leqslant \epsilon$ uniformly for all $\mathcal{A}$ with $|\mathcal{A}| \leqslant C$, with probability at least $1 - \delta$.

- Implications : influence at the longer time window T requires more samples.
- In practice : large $n = 10K$ allows small $m = 5$ to achieve good performance.

Influence Maximization

- We seek to solve

$$\mathcal{A}^* = \operatorname{argmax}_{|\mathcal{A}| \leqslant C} \ \sigma(\mathcal{A}, T)$$

which is NP-hard in general.

Page 36 of 49

Influence Maximization

- We seek to solve

$$\mathcal{A}^* = \operatorname{argmax}_{|\mathcal{A}| \leqslant C} \sigma(\mathcal{A}, T)$$

which is NP-hard in general.

- $\sigma(\mathcal{A}, T)$ is a non-negative, monotonic, submodular function.

Influence Maximization

- We seek to solve

$$\mathcal{A}^* = \operatorname{argmax}_{|\mathcal{A}| \leqslant C} \sigma(\mathcal{A}, T)$$

  which is NP-hard in general.
- $\sigma(\mathcal{A}, T)$ is a non-negative, monotonic, submodular function.
- Greedy algorithm achieves at least a fraction $(1 - 1/e)$ of the optimal value (OPT)

Influence Maximization

- We seek to solve

$$\mathcal{A}^* = \operatorname{argmax}_{|\mathcal{A}| \leqslant C} \sigma(\mathcal{A}, T)$$

  which is NP-hard in general.
- $\sigma(\mathcal{A}, T)$ is a non-negative, monotonic, submodular function.
- Greedy algorithm achieves at least a fraction $(1 - 1/e)$ of the optimal value (OPT)

Theorem

Suppose the influence $\sigma(\mathcal{A}, T)$ for all $\mathcal{A}$ with $|\mathcal{A}| \leq C$ are estimated uniformly with error ϵ and confidence $1 - \delta$, the greedy algorithm returns a set of sources $\widehat{\mathcal{A}}$ such that

$$\sigma(\widehat{\mathcal{A}}, T) \geq (1 - 1/e)OPT - 2C\epsilon$$

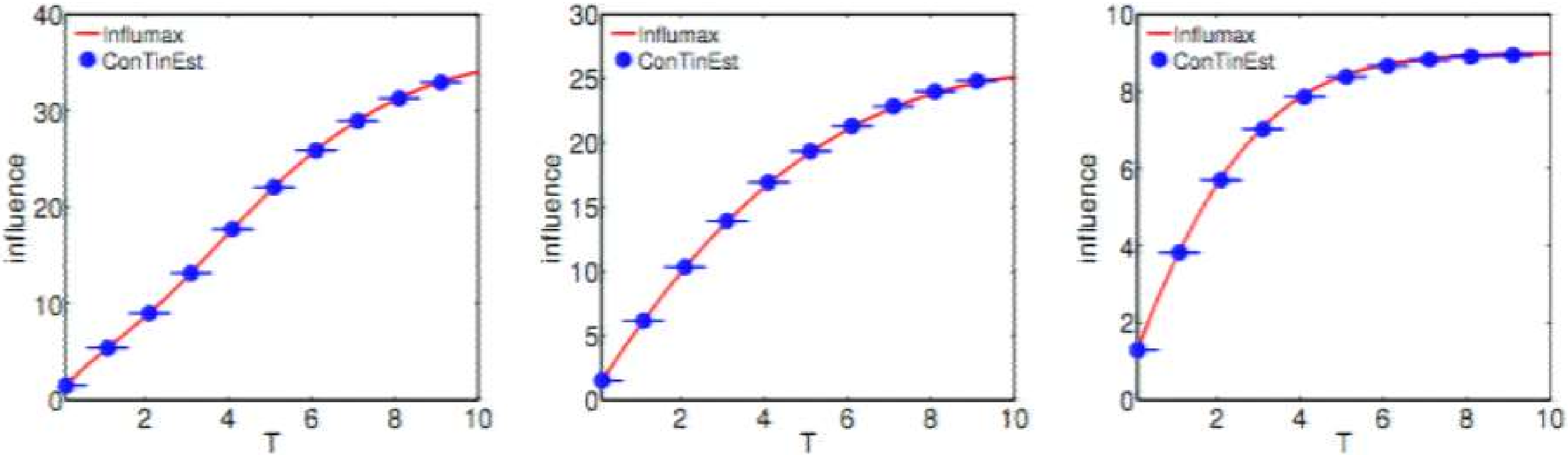
with probability at least $1 - \delta$.

Experimental Evaluation

- Synthetic dataset
  - Generate network structure.
  - Weibull pairwise transmission function.
- Real dataset
  - MemeTracker data (172m news articles 08/2009-09/2009).
- Evaluation
  - Accuracy of estimated influence.
  - Quality of selected sources.
  - Scalability.

Page 40 of 49

Accuracy of the estimated influence (highest out-degree node)

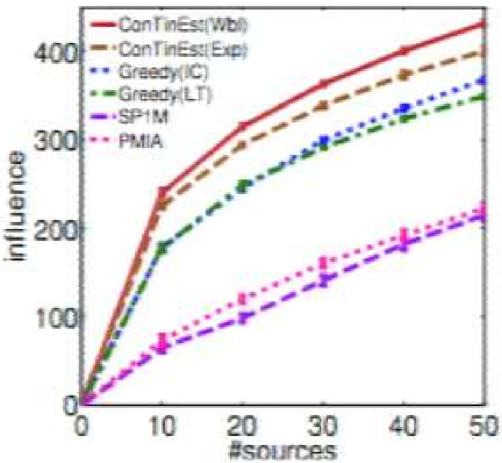


(a) Core-periphery (b) Random (c) Hierarchical

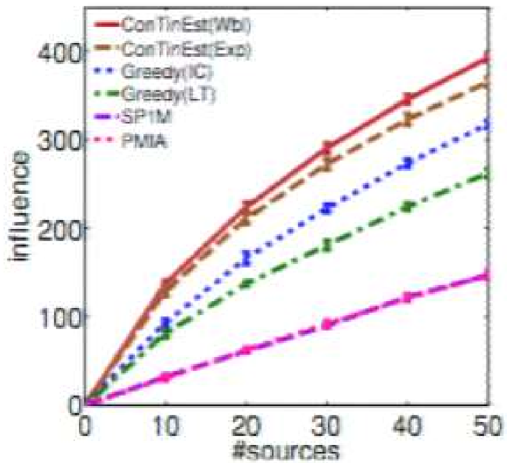
- CONTINEST is close to INFLUMAX (sparse small networks, exponential transmission functions).
- accuracy does not depend on network structure (128 nodes, 141 edges).

Synthetic Dataset

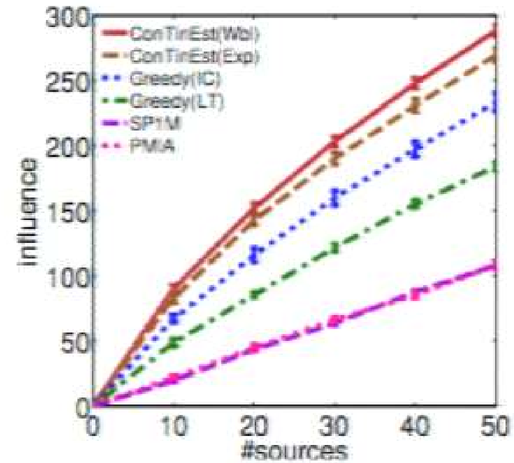
Quality of the selected nodes for influence maximization



(a) Core-periphery



(b) Random

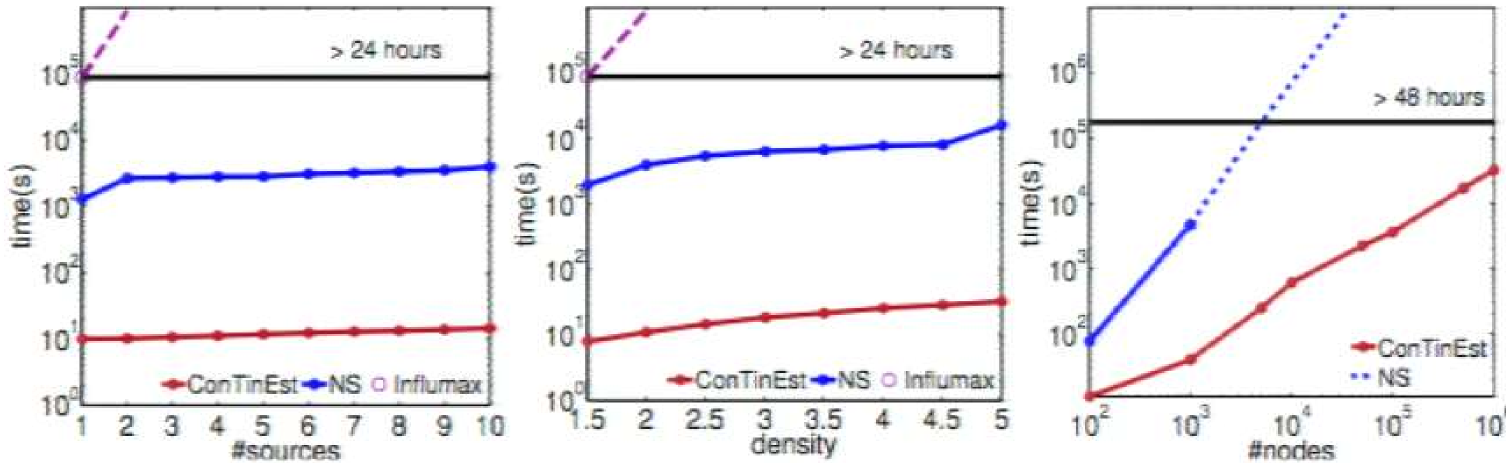


(c) Hierarchical

- ConTinEst typically outperforms competitive methods by 20%.
- Performance does not depend on network structure (1024 nodes, 2048 edges).

Synthetic Dataset

Scalability of influence maximization



(a) # sources
Small network

(b) network density
Small network

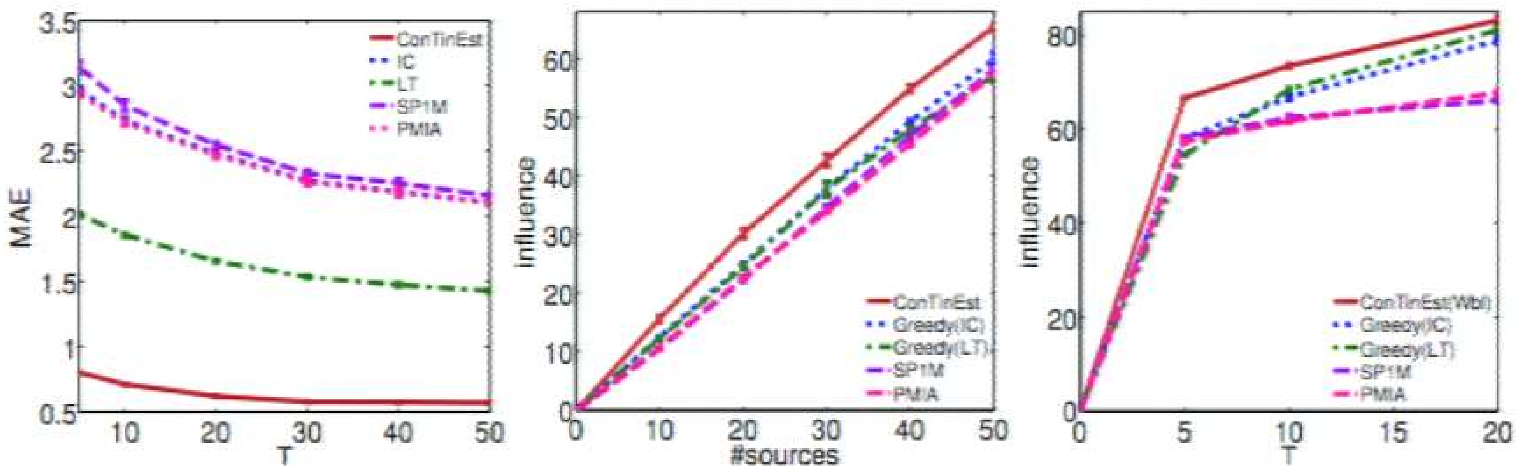
(c) network size
Up to one million nodes

- Small network : 128 nodes.
- Large network : up to 1 million nodes, with density 1.5.
- Our algorithm : sample 10K networks, 5 random labels.

Real Dataset

- 10,967 cascades.
- Use 80% cascades for learning continuous-time diffusion model.
- Select sources based on the learnt model.
- Evaluate influence of the sources using 20% test cascades.
- Compared to discrete-time diffusion models and scalable heuristics.

Page 44 of 49



(a) Estimation error (b) #sources (c) observation window

- CONTINEST achieves the lowest MAE error.

Page 45 of 49

Conclusion

- A randomized algorithm achieving :
  - the lowest estimation error in real data.
  - the largest influence within short time period.
  - the scaling up to millions of nodes in practice.

Page 47 of 49

Conclusion

- A randomized algorithm achieving :
  - the lowest estimation error in real data.
  - the largest influence within short time period.
  - the scaling up to millions of nodes in practice.
- Future work :
  - User engagement maximization of online systems.
  - Influence minimization and manipulation.
  - More general continuous-time diffusion model.

Conclusion

- A randomized algorithm achieving :
  - the lowest estimation error in real data.
  - the largest influence within short time period.
  - the scaling up to millions of nodes in practice.
- Future work :
  - User engagement maximization of online systems.
  - Influence minimization and manipulation.
  - More general continuous-time diffusion model.
- Welcome to our poster F41 for detailed discussion.

N I
P S