Synthesizing data for simulating the epidemic

Stephen Eubank

IeDEA Global Consortium Annual Meeting
Rockville, MD June 23, 2016

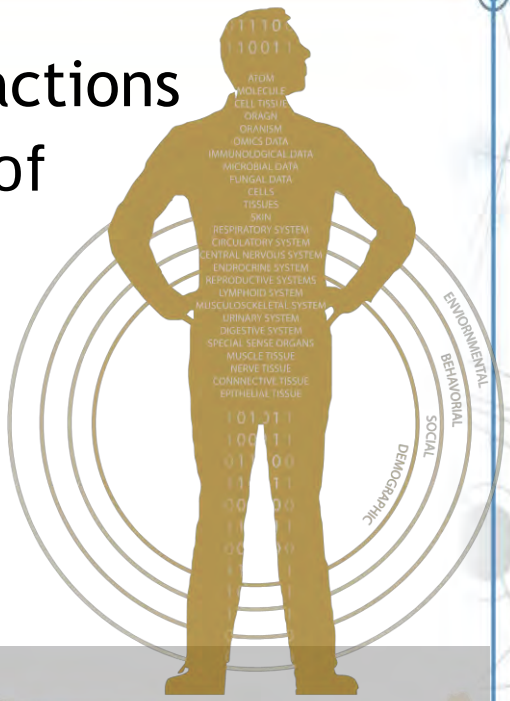
NDSSL TR-16-042

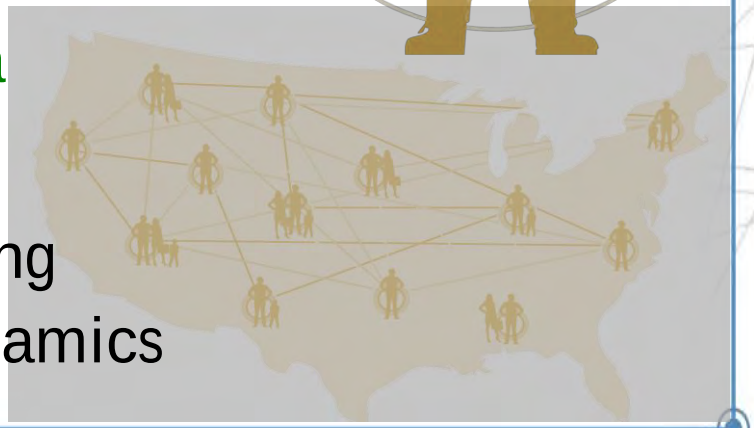
VirginiaTech.
Biocomplexity Institute

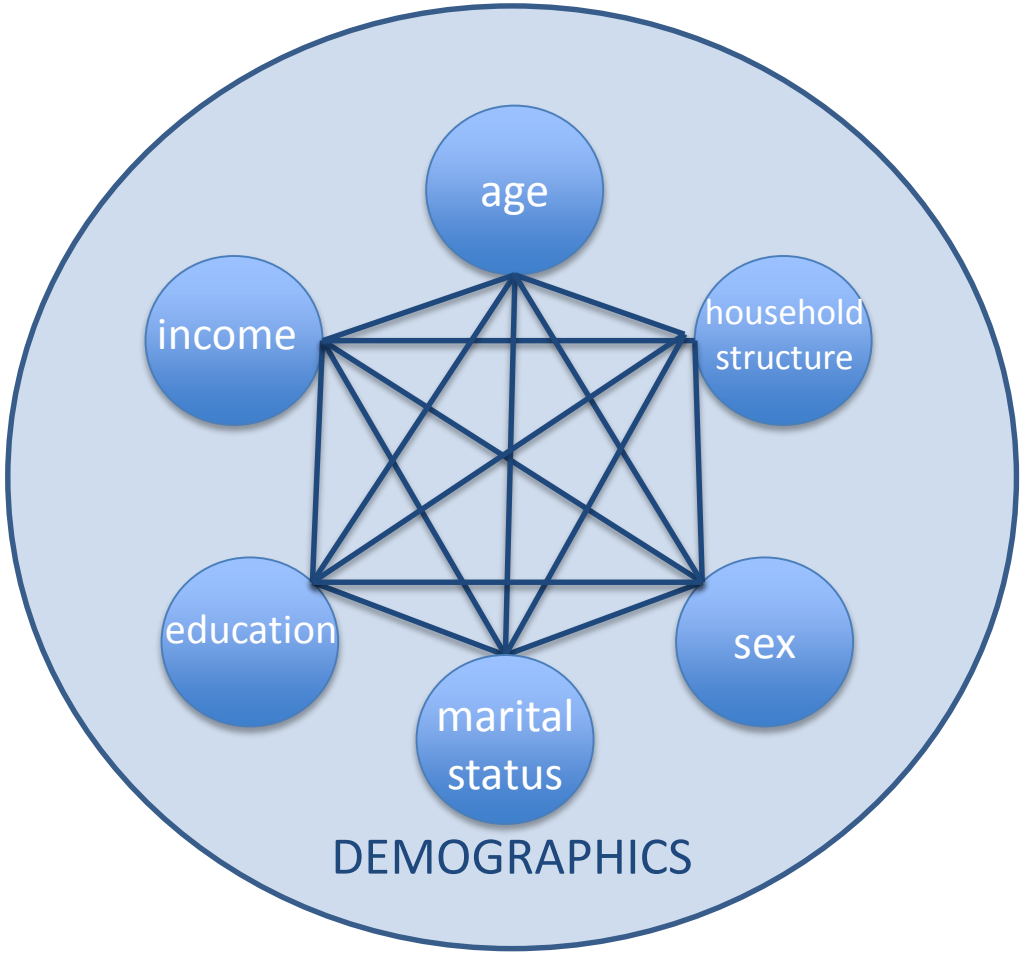
NDSSL NETWORK DYNAMICS & SIMULATION SCIENCE LABORATORY

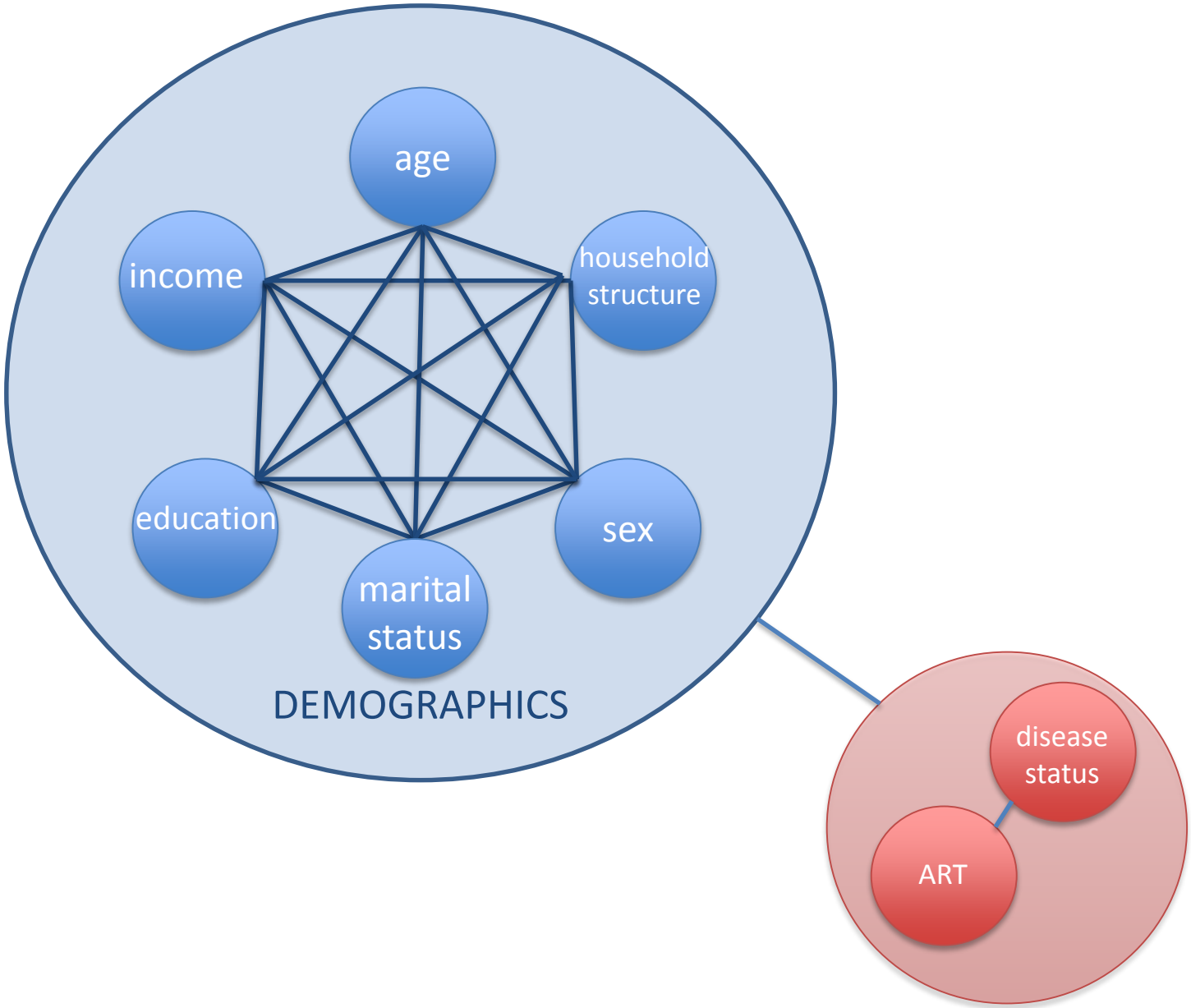
What is synthetic data?

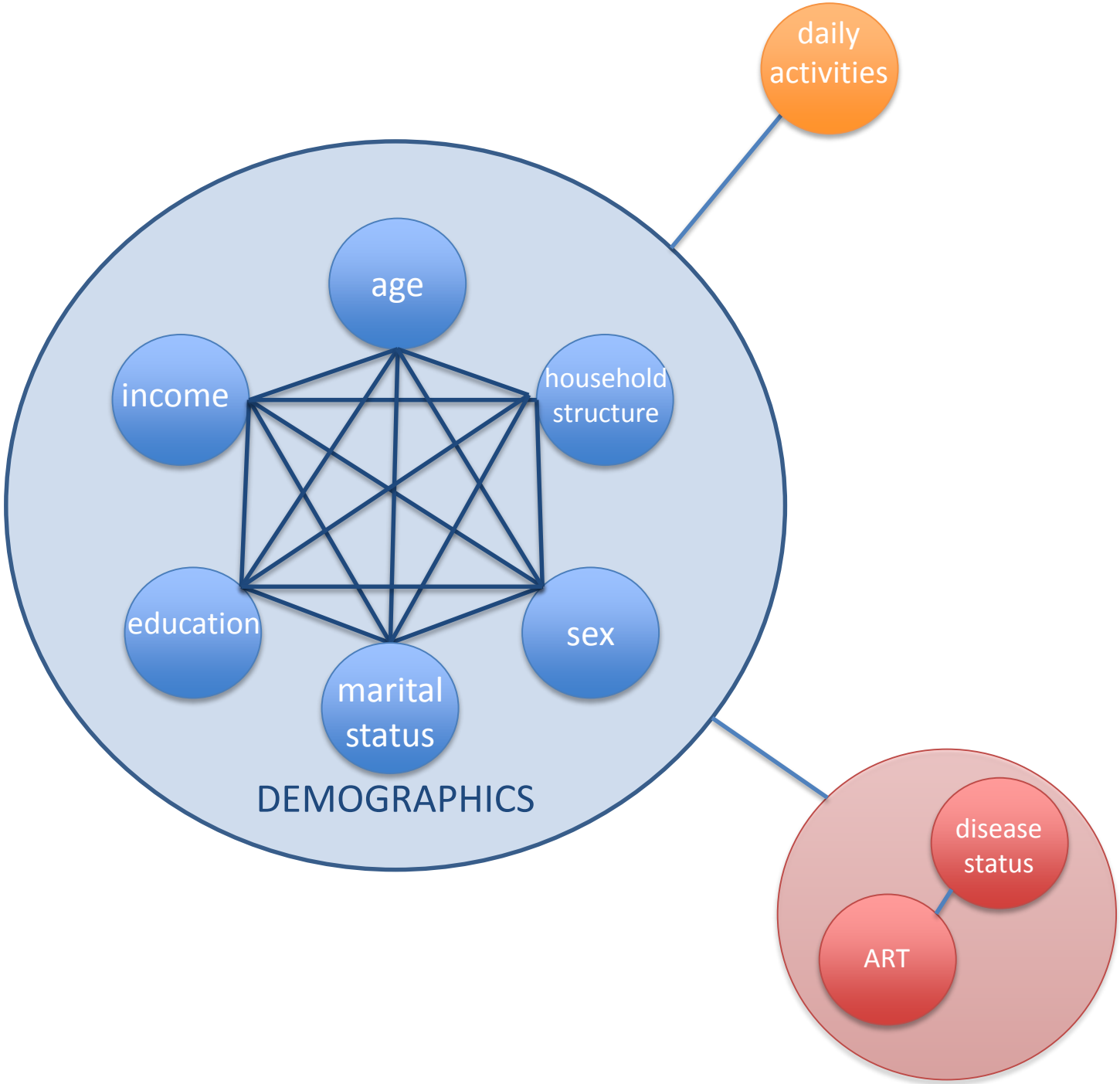
- A *representation* of things' states and interactions that is not intended to match any snapshot of the system, but to provide a statistically accurate overall picture:
  - people, places, things
  - cells, cytokines, organs
  - and their relationships
- A *synthesis* of incommensurate data and models from many domains
- A *coordinate system* for incorporating new information and describing dynamics

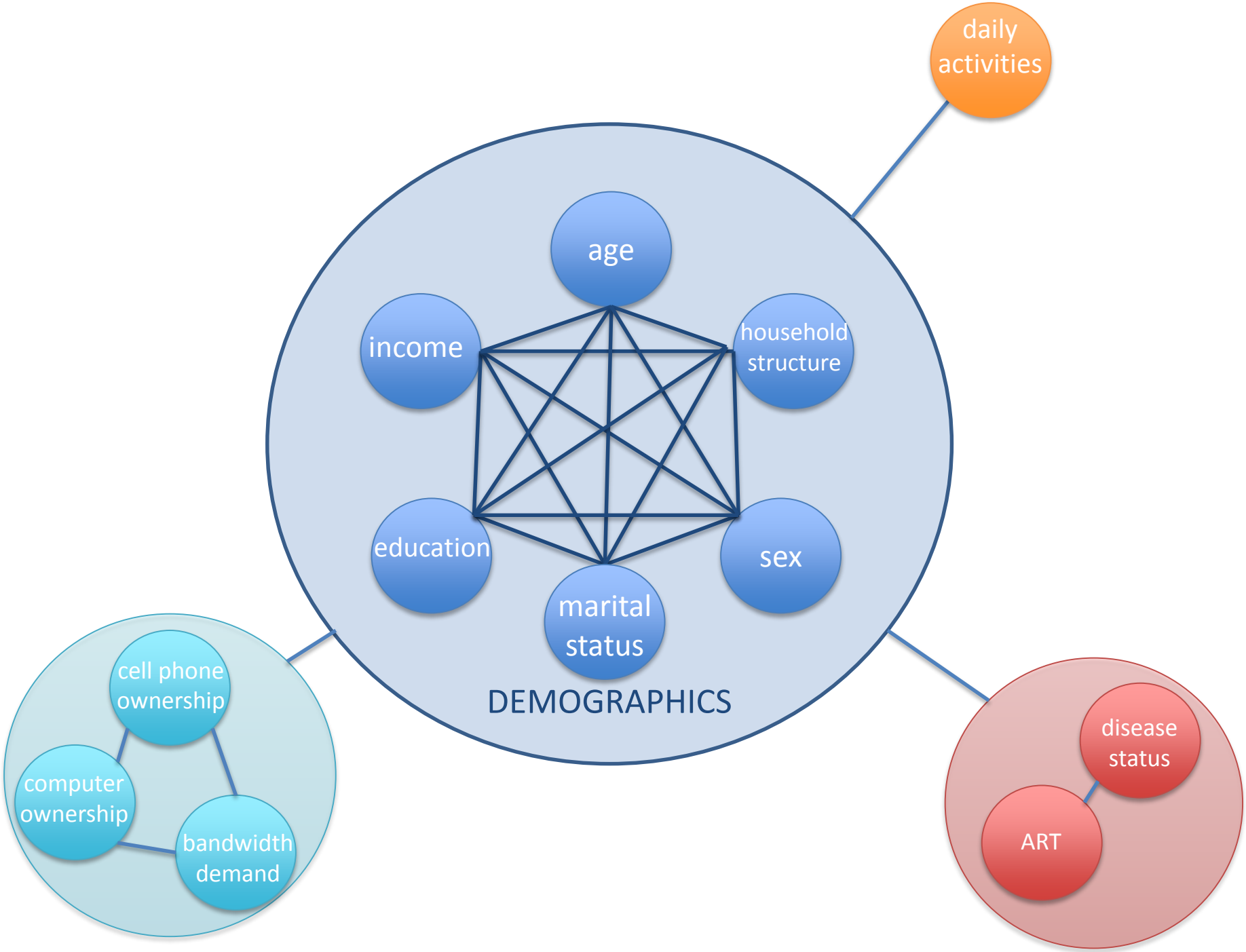


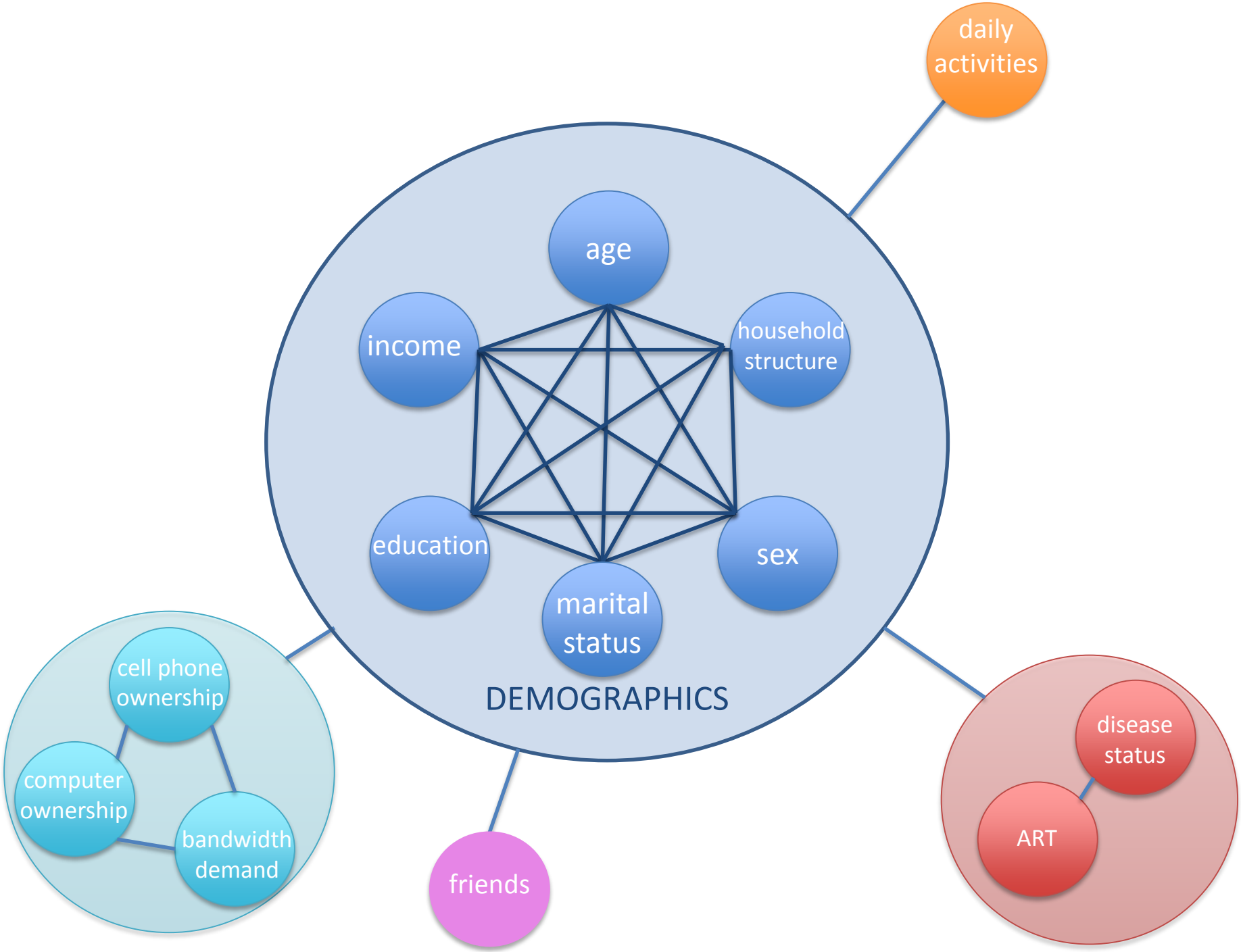


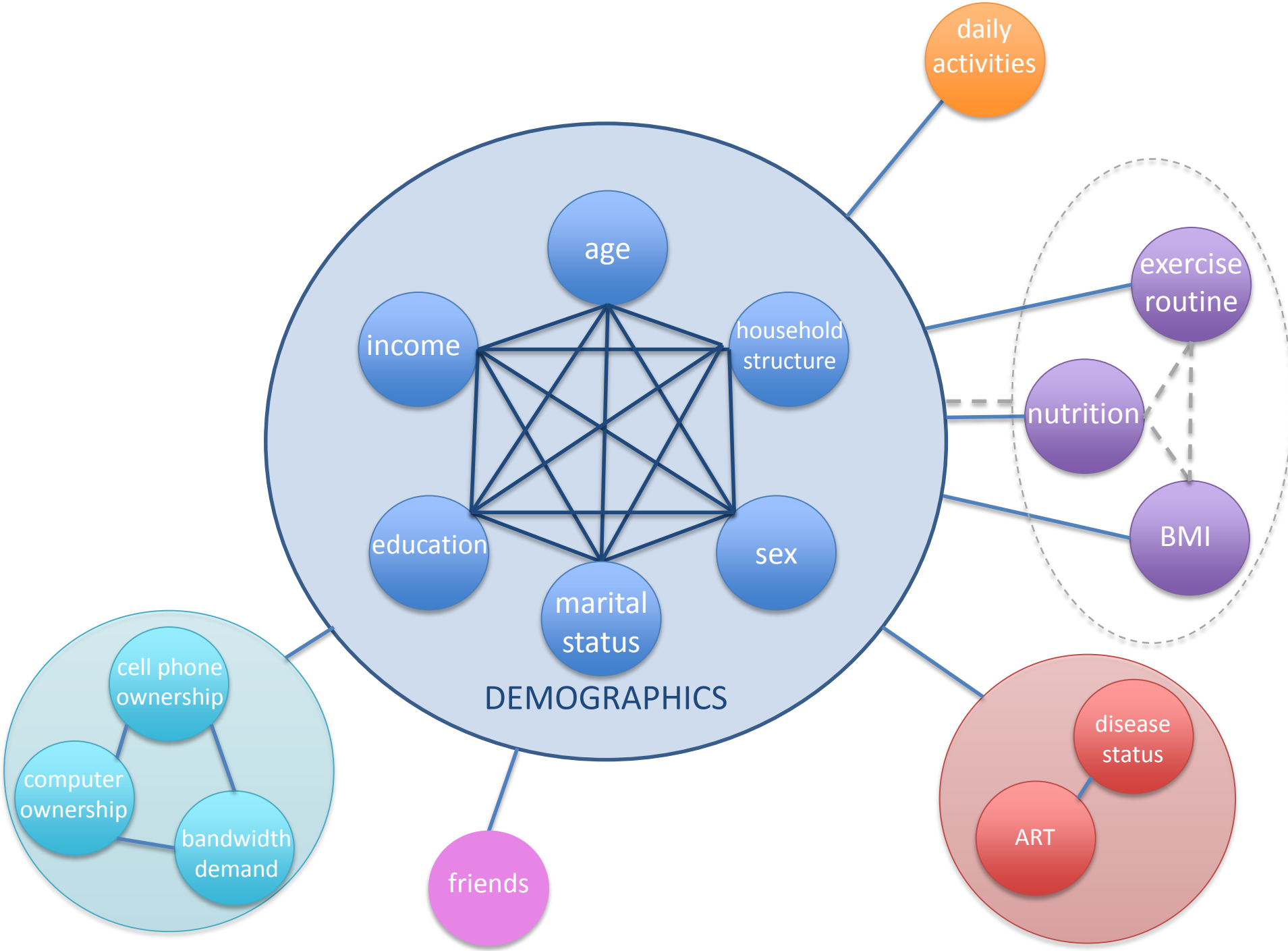

age
income
household structure
education
sex
marital status
DEMOGRAPHICS

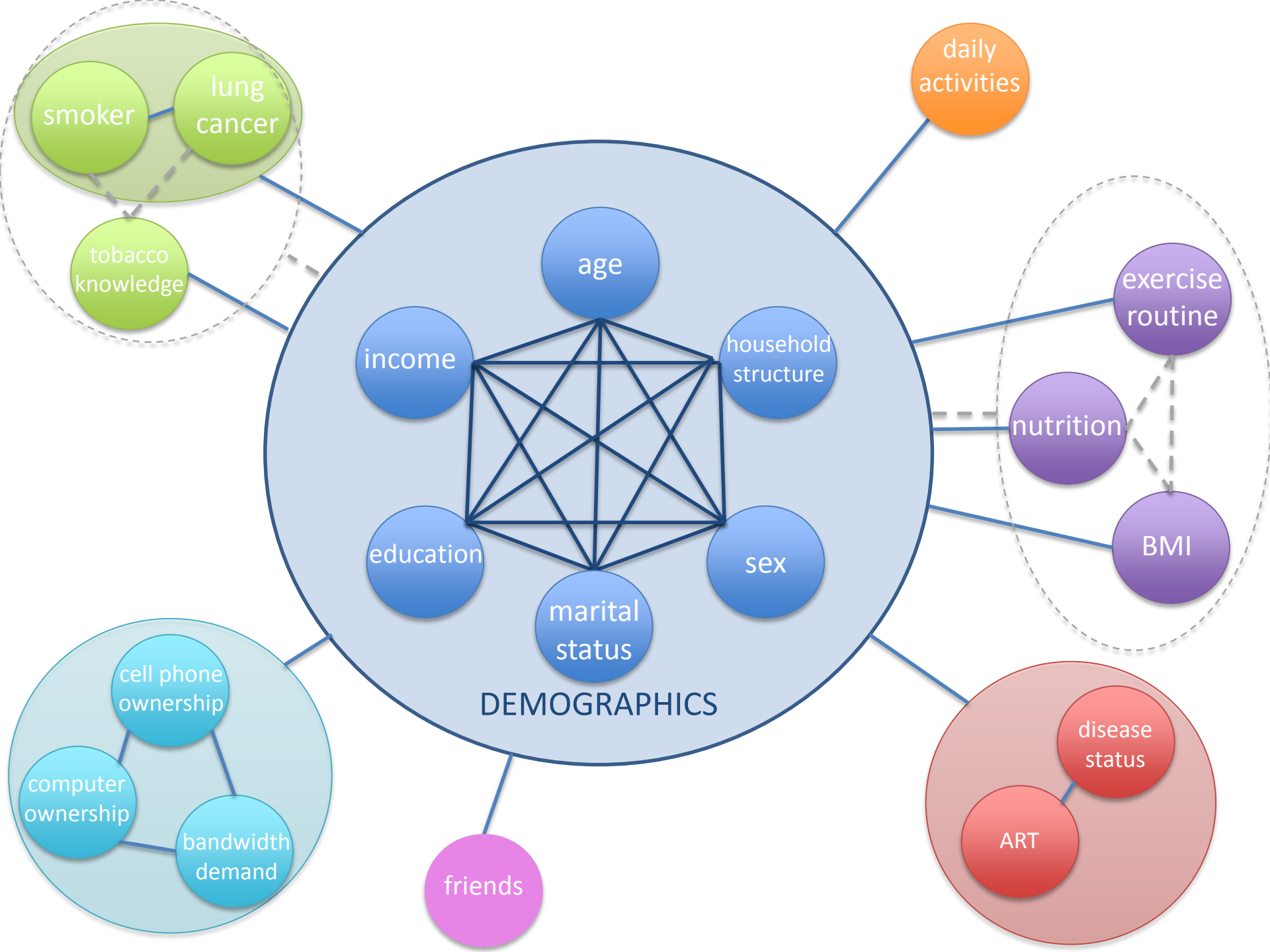

age
income
household structure
education
sex
marital status
DEMOGRAPHICS
disease status
ART

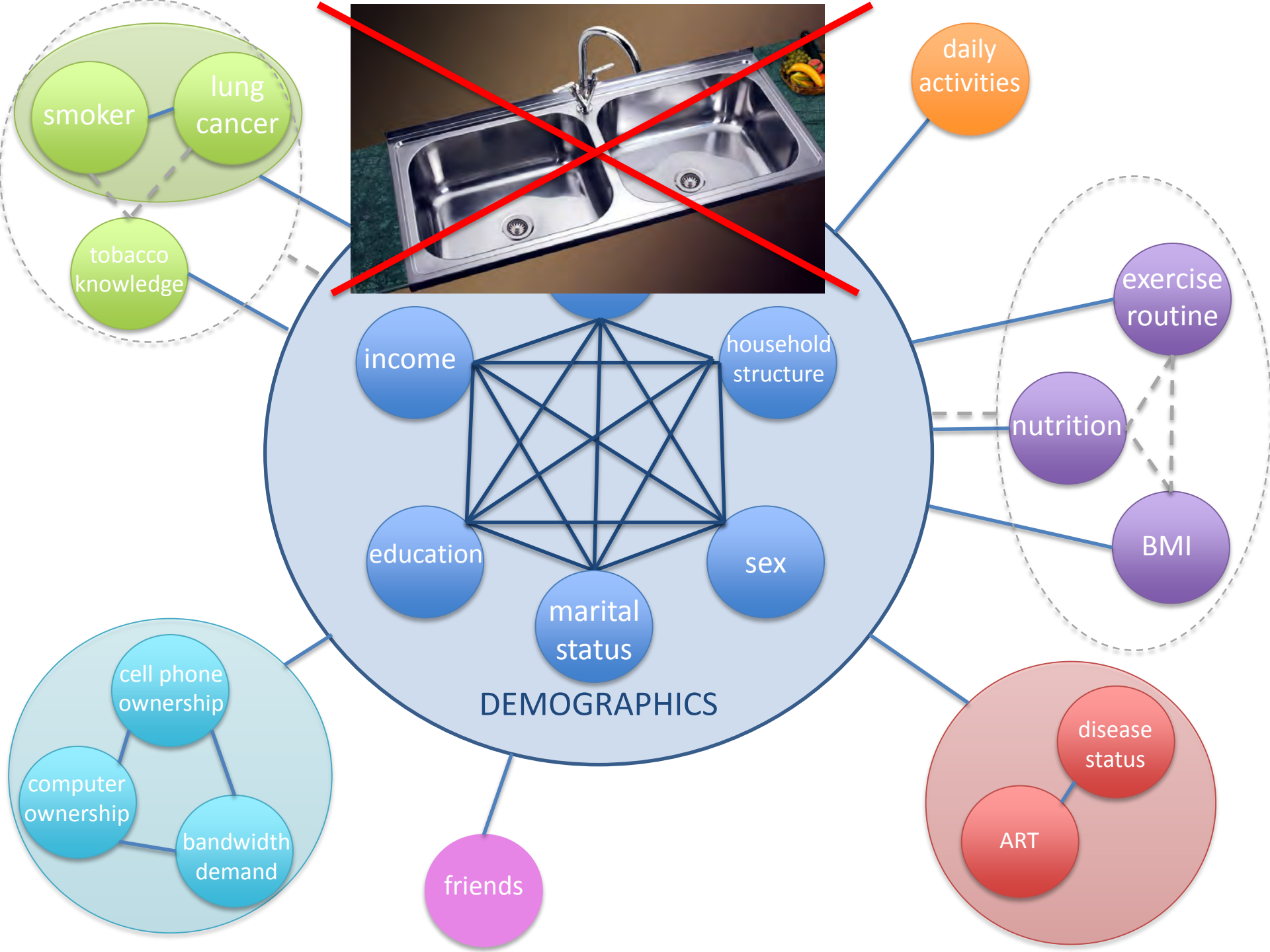

daily activities
age
income
household structure
education
sex
marital status
DEMOGRAPHICS
disease status
ART


daily activities
age
income
household structure
education
sex
marital status
DEMOGRAPHICS
cell phone ownership
computer ownership
bandwidth demand
disease status
ART


daily activities
age
income
household structure
education
sex
marital status
DEMOGRAPHICS
cell phone ownership
computer ownership
bandwidth demand
friends
disease status
ART


daily activities
age
income
household structure
education
sex
marital status
DEMOGRAPHICS
exercise routine
nutrition
BMI
cell phone ownership
computer ownership
bandwidth demand
friends
disease status
ART


smoker
lung cancer
tobacco knowledge
daily activities
age
income
household structure
education
sex
marital status
DEMOGRAPHICS
exercise routine
nutrition
BMI
cell phone ownership
computer ownership
bandwidth demand
friends
disease status
ART


smoker
lung cancer
tobacco knowledge
daily activities
exercise routine
nutrition
BMI
income
household structure
education
sex
marital status
DEMOGRAPHICS
cell phone ownership
computer ownership
bandwidth demand
friends
disease status
ART

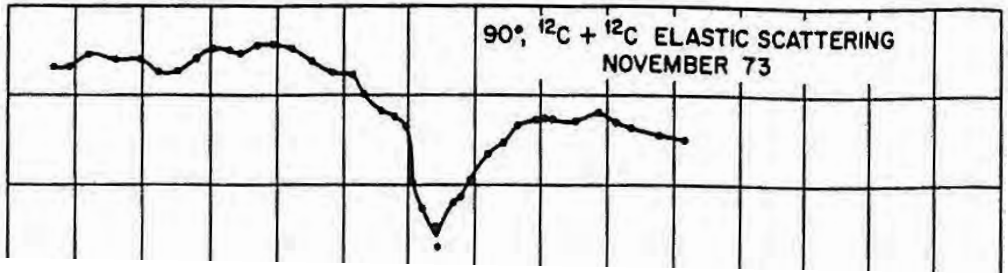
Why speak to IeDEA about *synthetic* data?

- "Real" system is one instance of a time-varying, random process in a huge space only partly observed.
- Planning relies on counterfactual analysis.
- Most alternatives sweep known errors under the rug.
- Synthesis provides a way to leverage data collection and harmonization efforts.

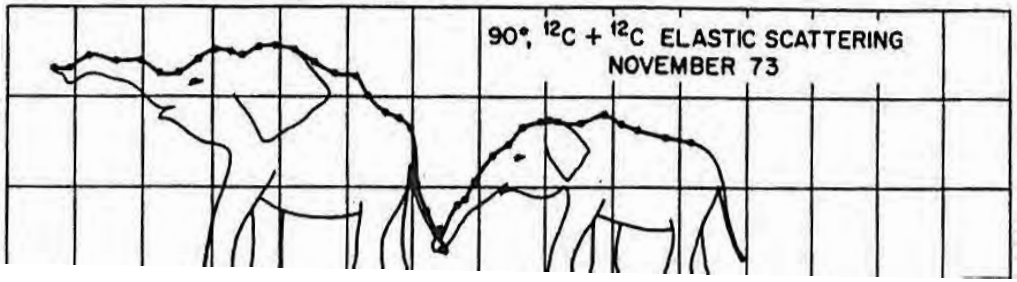
The real world is complicated

- "estimate engagement in care … in resource limited setting, … the outcome itself was missing due to loss to follow-up."
  Rebora, Paola et al. *BMC Medical Research Methodology* 16 (2016): 5
- "genetic …, viral … and environmental … factors influence CD4 recovery. These factors differ markedly around the world …"
  Geng, Elvin H et al. *International Journal of Epidemiology* 44.1 (2015): 251–263
- "…many practical studies assume away interference for the sake of simplicity. … ignoring interference can lead to completely wrong conclusions." Van der Laan, Mark J. *Journal of causal inference* 2.1 (2014): 13-74. *PMC*.
- "[W]e re-estimate the duration and relative infectivity of the acute phase, while accounting for … the retrospective cohort exclusion criteria and unmeasured heterogeneity in risk."
  Bellan SE, et al. (2015) PLoS Med 12(3): e1001801.
- "a significant proportion of seroconversion events in serodiscordant couples may be unlinked." Eshleman et al. J Infect Dis. (2011)

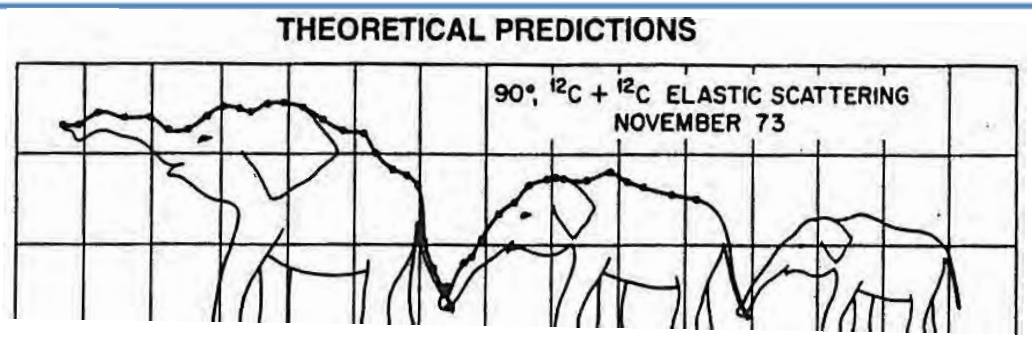
EXPERIMENTAL DATA



THEORETICAL FITS



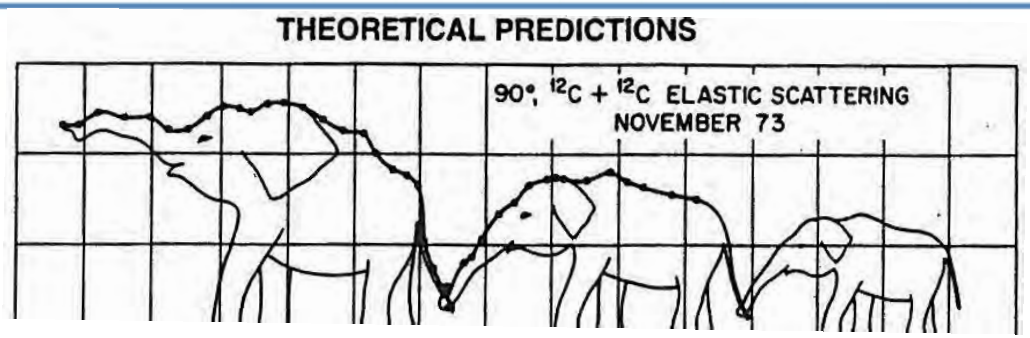
An overly complex model has no predictive power!



Science with a Smile,
Subramanian Raman

- Usual Lessons:
  - Occam's razor
  - Information criteria for complexity control

An overly complex model has no predictive power!



Science with a Smile,
Subramanian Raman

- Alternative Lesson:
  - Use causal / generative / mechanistic models
  - Representing arguably relevant mechanisms
  - To simulate alternatives and compare to data

An overly complex model has no predictive power!

Unless it’s the correct model…

The spherical elephant

several parameters

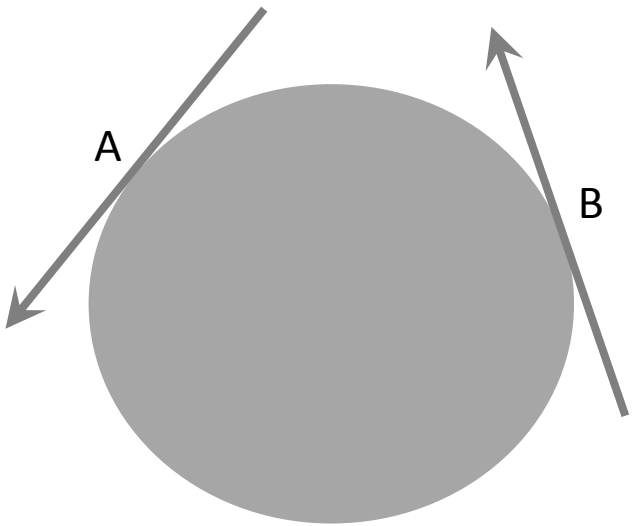
The spherical elephant and an even simpler, linear model

one parameter: slope

several parameters

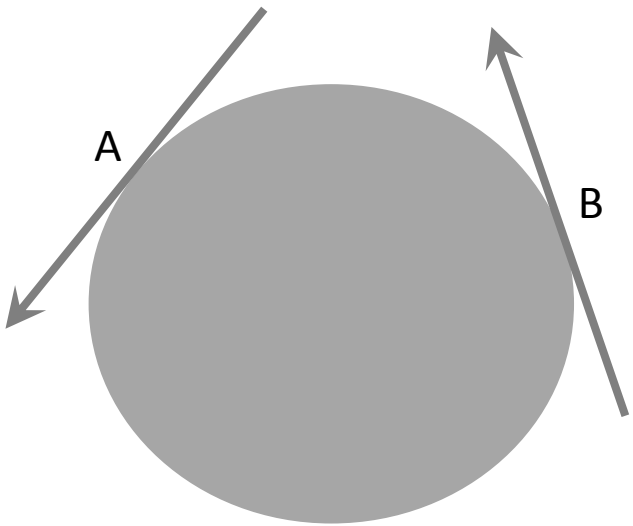
The spherical elephant and an even simpler, linear model

one parameter: slope can be calibrated to data at A or B



The spherical elephant and an even simpler, linear model

one parameter: slope can be calibrated to data at A or B



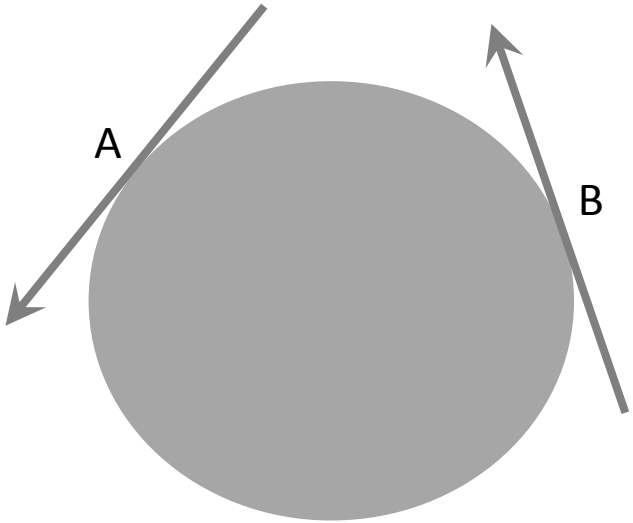
… with diametrically opposed implications for interventions

Is moving left a good way to get off the elephant?
That depends on where you are.

The spherical elephant and an even simpler, linear model

Is PrEP a good way to reduce morbidity?
That, too, depends on where you are.

one parameter: slope can be calibrated to data at A or B

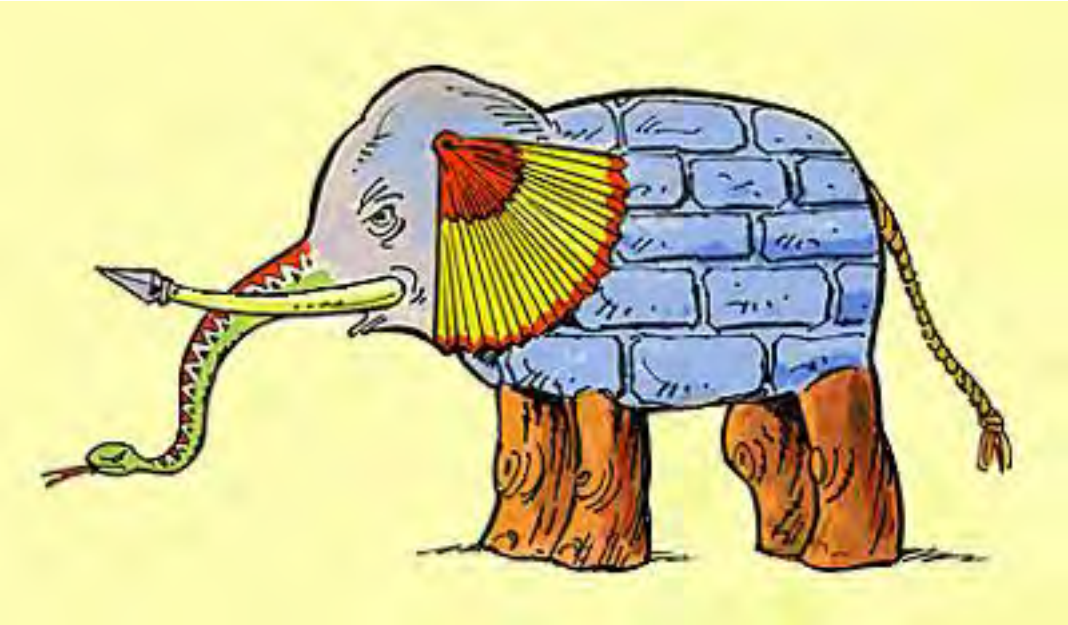


… with diametrically opposed implications for interventions

Is moving left a good way to get off the elephant?
That depends on where you are.

Overly simple models …

… lead to incorrect conclusions
even when they are locally correct.



Symptom: long-running disagreements,
with all sides supported by “the evidence”

Overly simple models …

… are *guaranteed* to mis-specify complex systems.
But there's not enough data for a complex model.
So we'll be satisfied with a mis-specified one.

Making effective use of data

To understand a property of the system:

- Sample the property
- Model its distribution
- Reason about the implications of the distribution

Making effective use of data

To understand a property of the system:

- Sample the property
- Model its distribution
- Reason about the implications of the distribution

In data-poor contexts, prior knowledge is key

e.g. parameterized distributions

Making effective use of data

To understand a property of the system:

- Sample the property
- Model its distribution
- Reason about the implications of the distribution

In data-poor contexts, prior knowledge is key
e.g. parameterized distributions

Worry about confounding, independence:
interactions among properties

Making effective use of data

To understand interactions among properties:

- Sample them
- Model their joint distribution
- Reason about the causal dependencies

In data-poor contexts, prior knowledge is key
e.g. synthetic data from simulated interactions

Simulation: approximate answers to real questions,
not exact analysis for the wrong questions

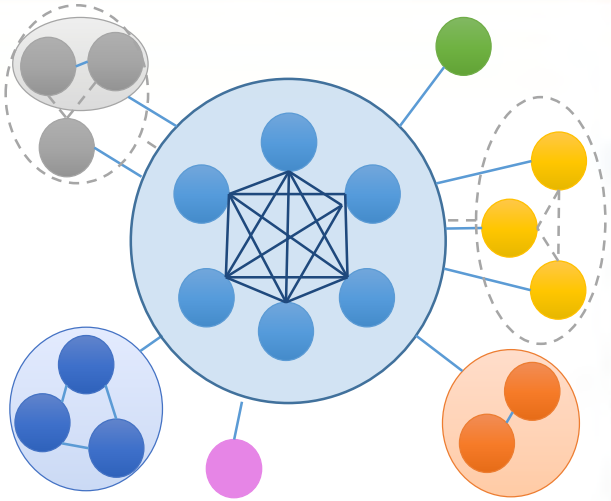
Why treat synthetic information as data?

With trust, provenance, dynamic updates, simulation, …

- Assumptions are on the table, subject to tinkering
- Users need not understand models to use data
- Results are not "thrown over the fence"
- Experts in specific domains embed context in data that is transparent to experts in other domains:

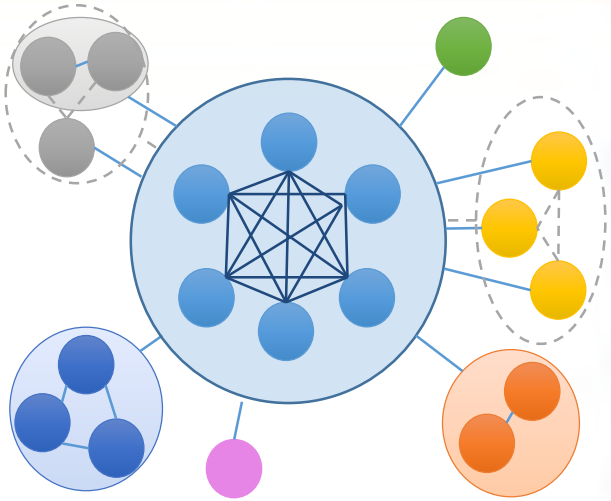
demography → trial design
pharmacodynamics + social psych → implementation

Characteristics of an ideal information resource



- Natural representation encourages transdisciplinary team science & communities of practice
- Attribution is easy, even automatic
- Models and results are reproducible
- Assumptions & limits on appropriate use are transparent
- Alternative hypotheses (models) are easily tested
- Supports an *abductive* reasoning loop: hypothesize, test, correct

Characteristics of an ideal information resource



- Natural representation encourages transdisciplinary team science & communities of practice
- Attribution is easy, even automatic
- Models and results are reproducible
- Assumptions & limits on appropriate use are transparent
- Alternative hypotheses (models) are easily tested
- Supports an *abductive* reasoning loop: hypothesize, test, correct

Building such a thing is still a research problem!

Why haven't we done this already?

Computational issues are as hard as statistical ones:

- scalable digital library
- efficient algorithms to access & process Big Data
- data sleuthing & licensing / Data Use Agreements
- testing & maintaining provenance at all stages
- designing systems for non-computing-experts

Conclusions

- $N \to \infty$, p-values $\to 0$, anything is publishable!!

Conclusions

- ~~$N \to \infty$, p values $\to 0$, anything is publishable!!~~
- IeDEA provides unique opportunities to study demographically, geographically targeted interventions.
- Synthetic information sets, e.g. populations, can complement the data collected in IeDEA.
- Fully realizing the promise of IeDEA will require new kinds of data management platforms.