Integrative modeling for prediction

Alfred Hero

University of Michigan – Ann Arbor

www.eecs.umich.edu/~hero

Outline

- **Integrative**: handle vast amounts of diverse information
- **Modeling**: know what we know
- **Prediction**: know what we don't know
- **Uncertainty assessment**: don't know what we don't know

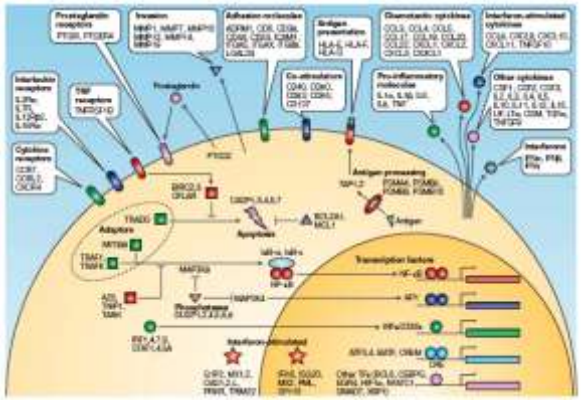
- Motivating example**: predictive health and disease**

Goal of PHD (WSJ Nov 3)

Presymptomatic detection and diagnosis of illness resulting from infectious pathogens performed by a blackberry-sized POC device.

We have a diverse number of inputs…

Functional pathway model



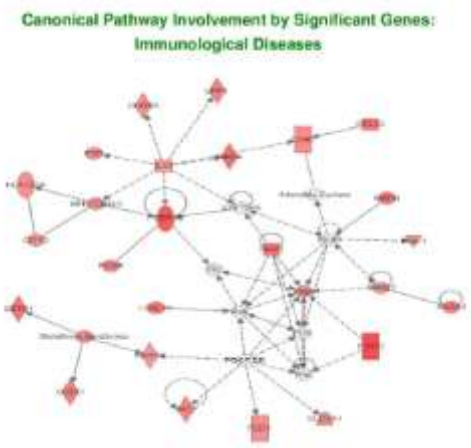
Jenner and Young, Nature Reviews 2005

Symptom model

Epidemic model

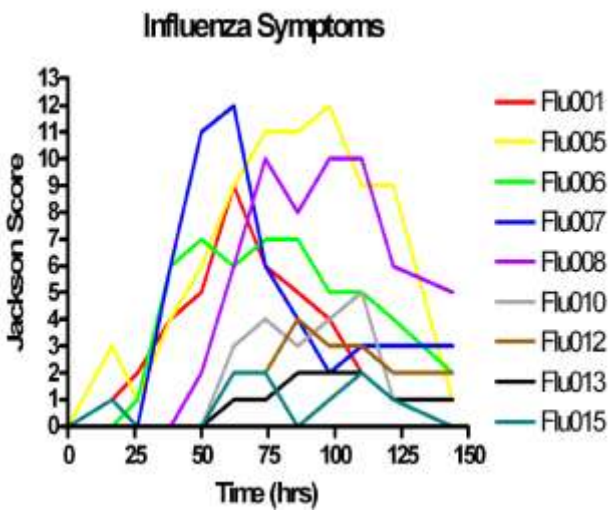
Google Health Map

Metabolo/Geno/Proteo-mics



Ingenuity pathway analysis software

Clinical data



Close contact data

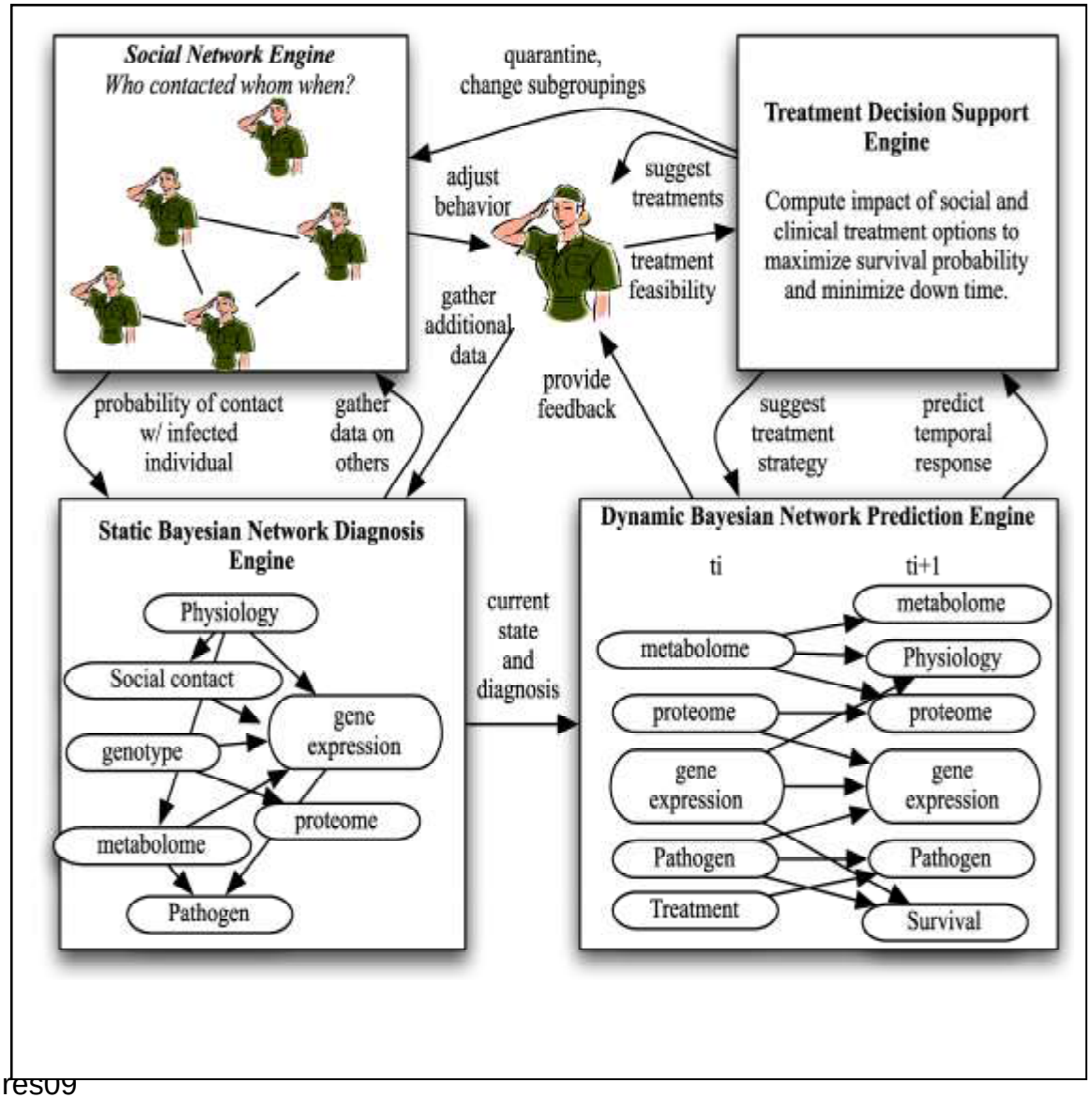
Aiello Research Group, UM SPH

…and a diverse number of outputs

- Individual predictions: risk, individualized diagnosis and prognosis
- Drug efficacity predictions: required dose, effectiveness
- Epidemiologic predictions: communicability, attack rate, spread rate

- Integrative prediction engine is required

Integrative prediction engine



Are present tools adequate?

- Linear/additive model fitting with soft constraints/penalties
  - Potentially huge state space, not scalable to large diverse data sets
- Parametric, semiparametric, non-parametric uncertainty
  - Overfitting issues, parsimony is not modular or decomposable
- Robustification and adaptation of simple predictors
  - Commission of errors of over-conservativism or over-optimism
- Asymptotic performance prediction
  - Limited to large N, modeled uncertainty, i.i.d. observations, CLT

Tools of the future?

- Scalable integrated phenomenological and statistical models
  - Hybrid discrete/cts time/value models beyond state quantization
  - Homogenization/decomposition that avoids embedded simulation
  - Seamless combination of parametric and geometric methods
- Integrated information-geometric uncertainty quantification
  - The most natural domain for aggregation of diverse information
  - Estimation and reduction of data dimensionality
  - Geodesic estimation with uncertainty quantification
- Actionable online performance prediction
  - Tight performance approxs that adapt to tasks and interventions
  - Must account for interactions between info fusion, models (gm, mm), active learning

Conclusions

- Present tools inadequate to deal with future demands in health prediction
- Other burgeoning areas: tomographic information security, multimodal database indexing and retrieval, quantum molecular imaging…
- Problems of the future will involve
  - Massive diversity of data with (hopefully) low latent dimension
  - Physical models that are better characterized than today
  - More stringent requirements for performance guarantees
  - Human in-the-loop required at many more places in the system
  - Flexible information visualization without information overload
- Will need to exploit more domain knowledge, e.g., biological mechanistic and kinetic models