Disclosures

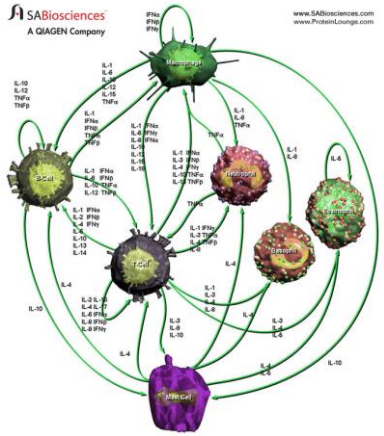
- Torque Therapeutics – board of directors
- Boehringer-Ingelheim Pharmaceuticals – research collaboration
- Janssen Pharmaceuticals – research collaboration
- Immuneering – scientific advisory board
- Applied BioMath – scientific advisory board
- BerGenBio – consultant
- Genentech -- consultant

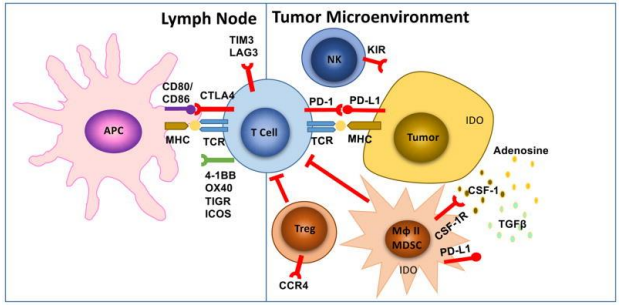
MIT BE
BIOLOGICAL ENGINEERING

A quick introductory tutorial on Artificial Intelligence, Machine Learning, and 'Big Data'

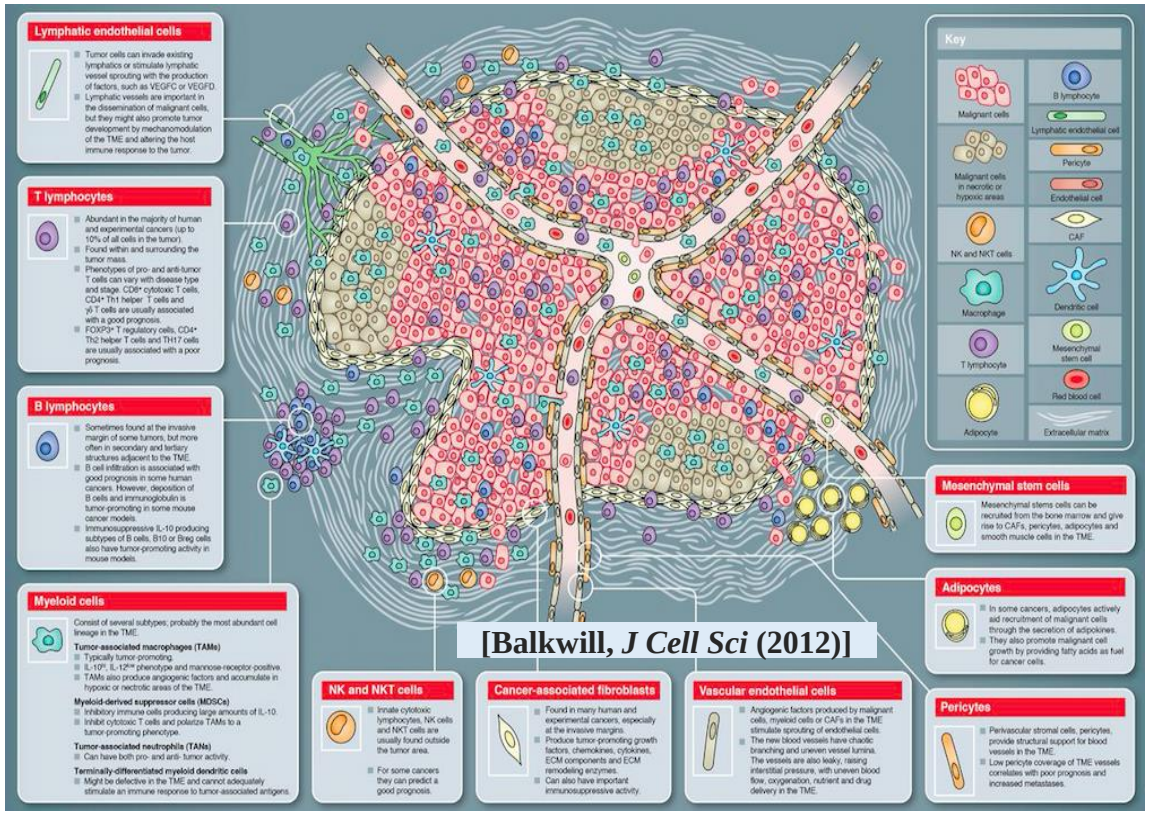
MIT BE
BIOLOGICAL ENGINEERING

Motivation: Enhance capabilities for prediction and insight in immune system-related pathologies and therapeutics based on intensive and extensive molecular / cellular experimental interrogation

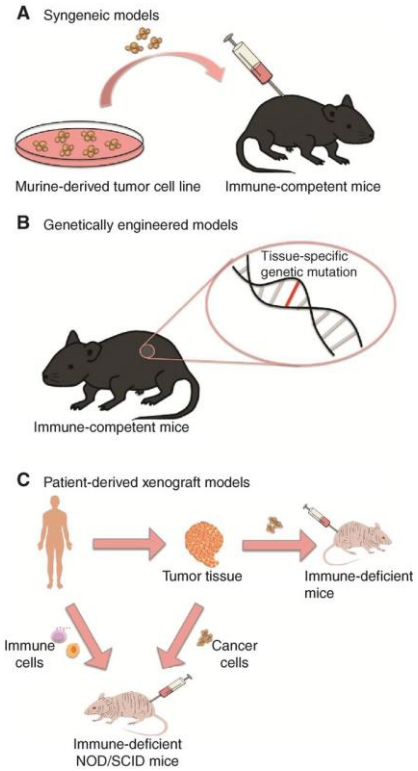




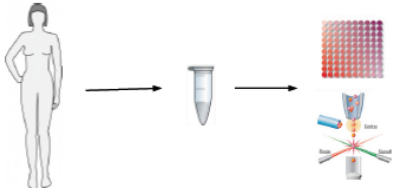
[Sharma, *Cell* (2017)]



[Balkwill, *J Cell Sci* (2012)]



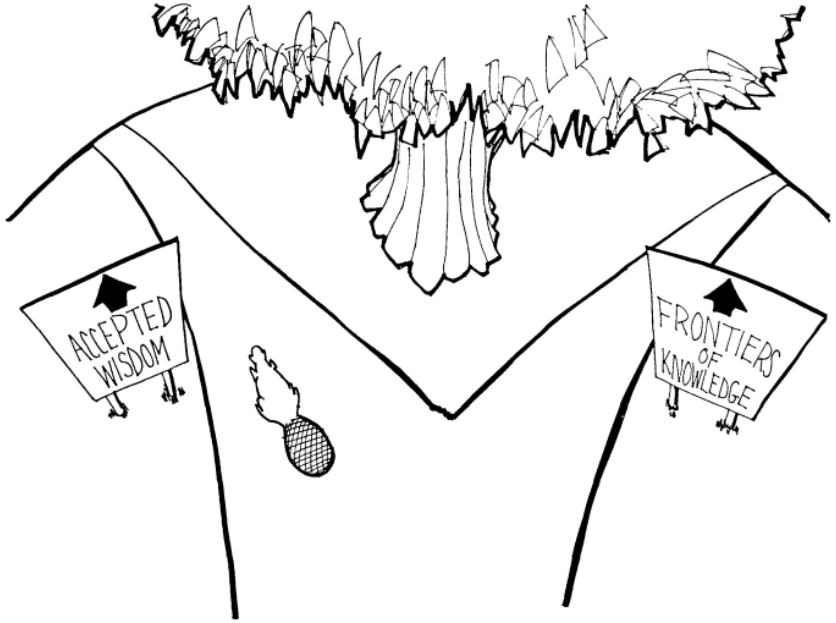
[Olsen, *Canc Disc* (2018)]



MIT BE
BIOLOGICAL ENGINEERING

Artificial Intelligence – categories

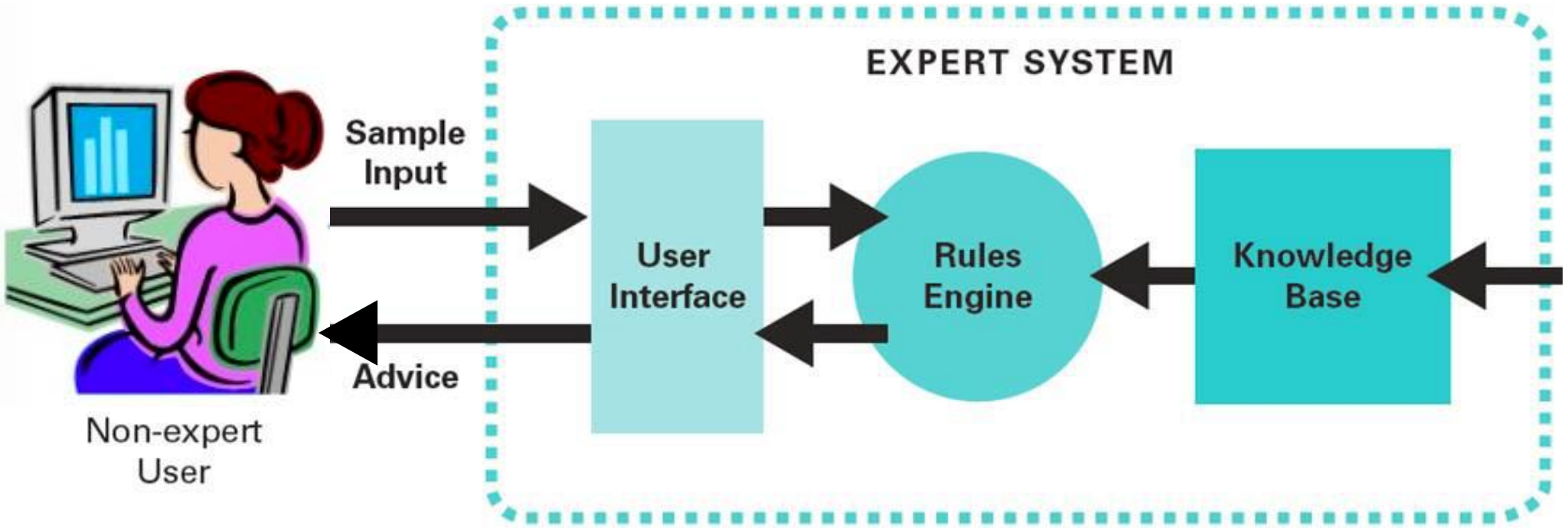
- **Expert Systems**
  - **‘forward engineering’**
  - **’rules-based’**
  - **‘knowledge-based’**
  - **‘hypothesis-based’**
- **Machine Learning**
  - **‘reverse engineering’**
  - **‘data-driven’**
  - **‘hypothesis-generating’**



[Davis, *AAAI AI Magazine* (1982)]

MIT BE BIOLOGICAL ENGINEERING

Expert Systems



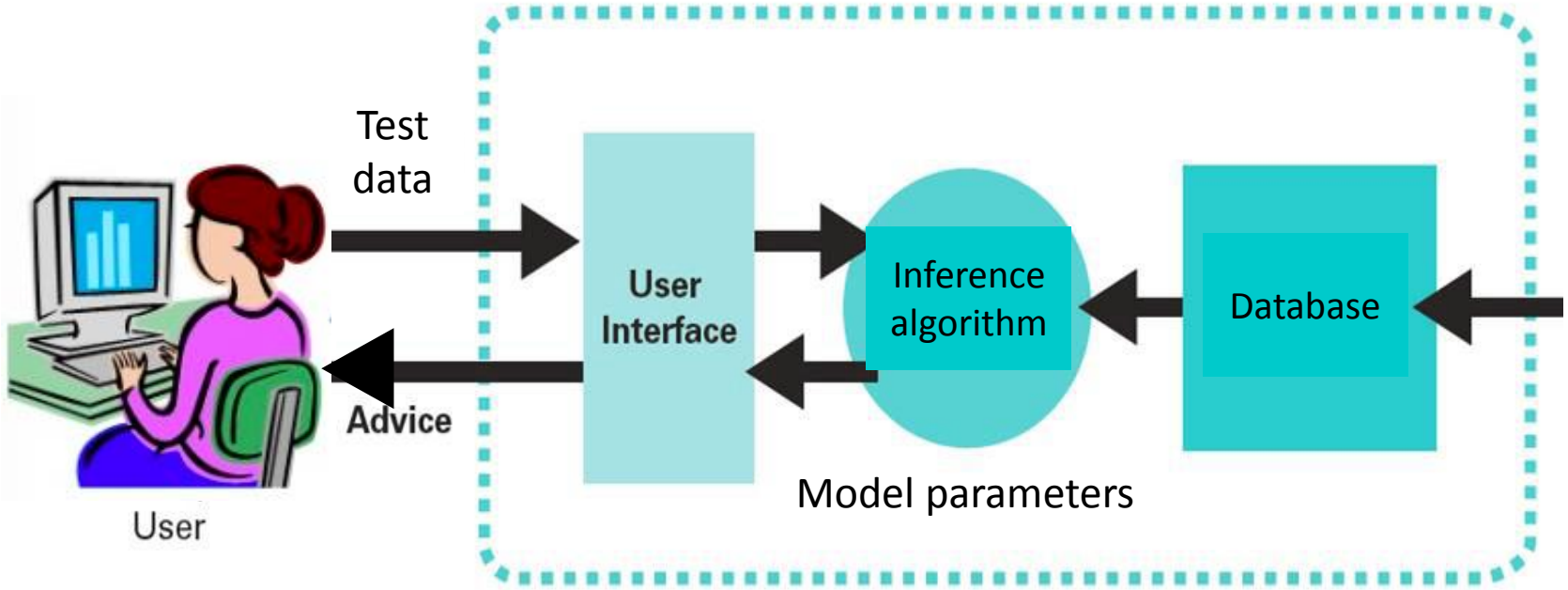
Knowledge from an Expert

-- possibly including Natural Language Processing for archival journal information

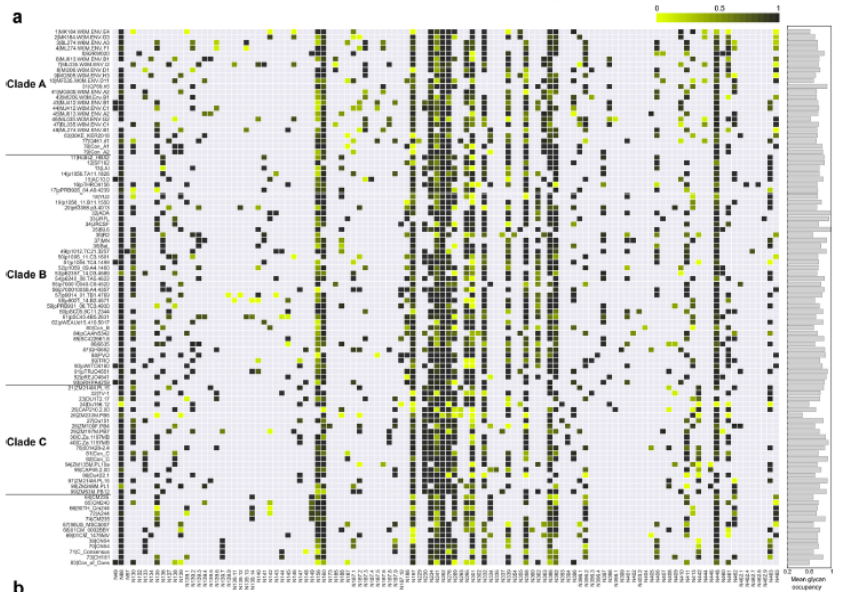
Model is based on prior knowledge / theory

MIT BE
BIOLOGICAL ENGINEERING

Machine Learning



*



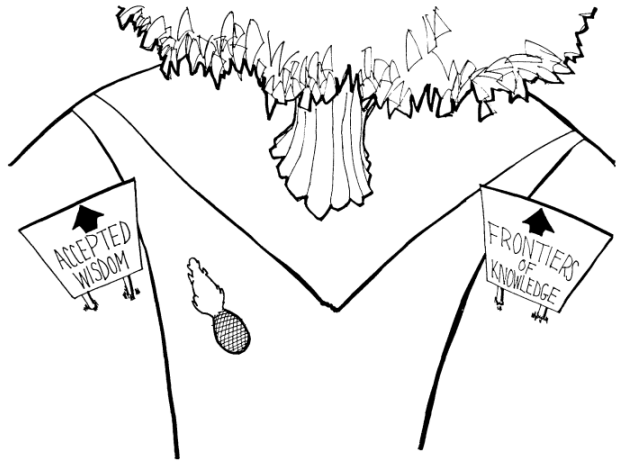
Training data

Model is derived from empirical data

MIT BE
BIOLOGICAL ENGINEERING

Artificial Intelligence – categories

- **Expert Systems**
  - **‘forward engineering’**
  - **’rules-based’**
  - **‘knowledge-based’**
  - **‘hypothesis-based’**
- **Machine Learning**
  - **‘reverse engineering’**
  - **‘data-driven’**
  - **‘hypothesis-generating’**



[Davis, *AAAI AI Magazine* (1982)]

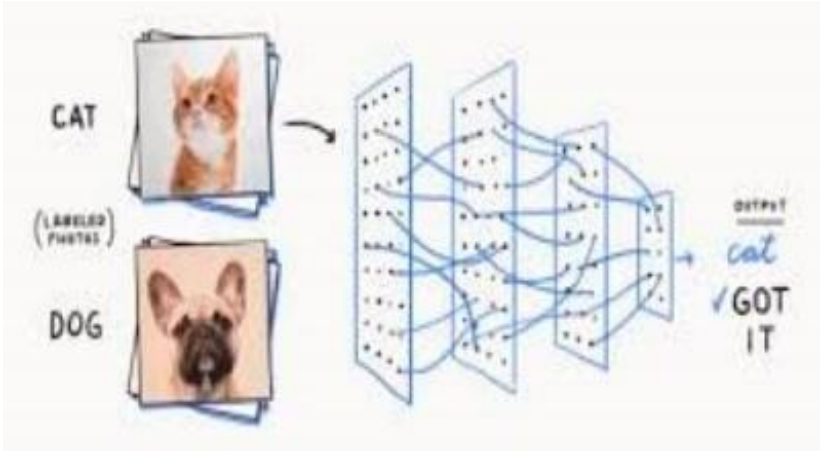
For most problems rooted in bioscience Expert Systems approaches do not work well, because the Rules are inadequately grounded in either theory or knowledge

Thus, Machine Learning is generally the more useful approach (although hybrid frameworks may be helpful)

MIT BE BIOLOGICAL ENGINEERING

Machine Learning Example

- **Example: Image Recognition**
  - **reverse engineering = determining how pixel data are related to object categories**
  - **data-driven = large number of training images to develop pixel-object relationships**
  - **Inputs = image pixels**
  - **Outputs = likelihood of image showing a particular category of object**
  - **Model = parameters quantifying relationships among measured and derived features**



stuffed toy	0.799
plush	0.763
snout	0.693
textile	0.682
fur	0.678
animal	0.612
dog like mammal	0.504

MIT BE
BIOLOGICAL ENGINEERING

Differences from Traditional Biostatistics

- **traditional Biostatistics focus is on Yes-vs-No hypothesis testing, whereas Machine Learning focus is on input-output relationship modeling**

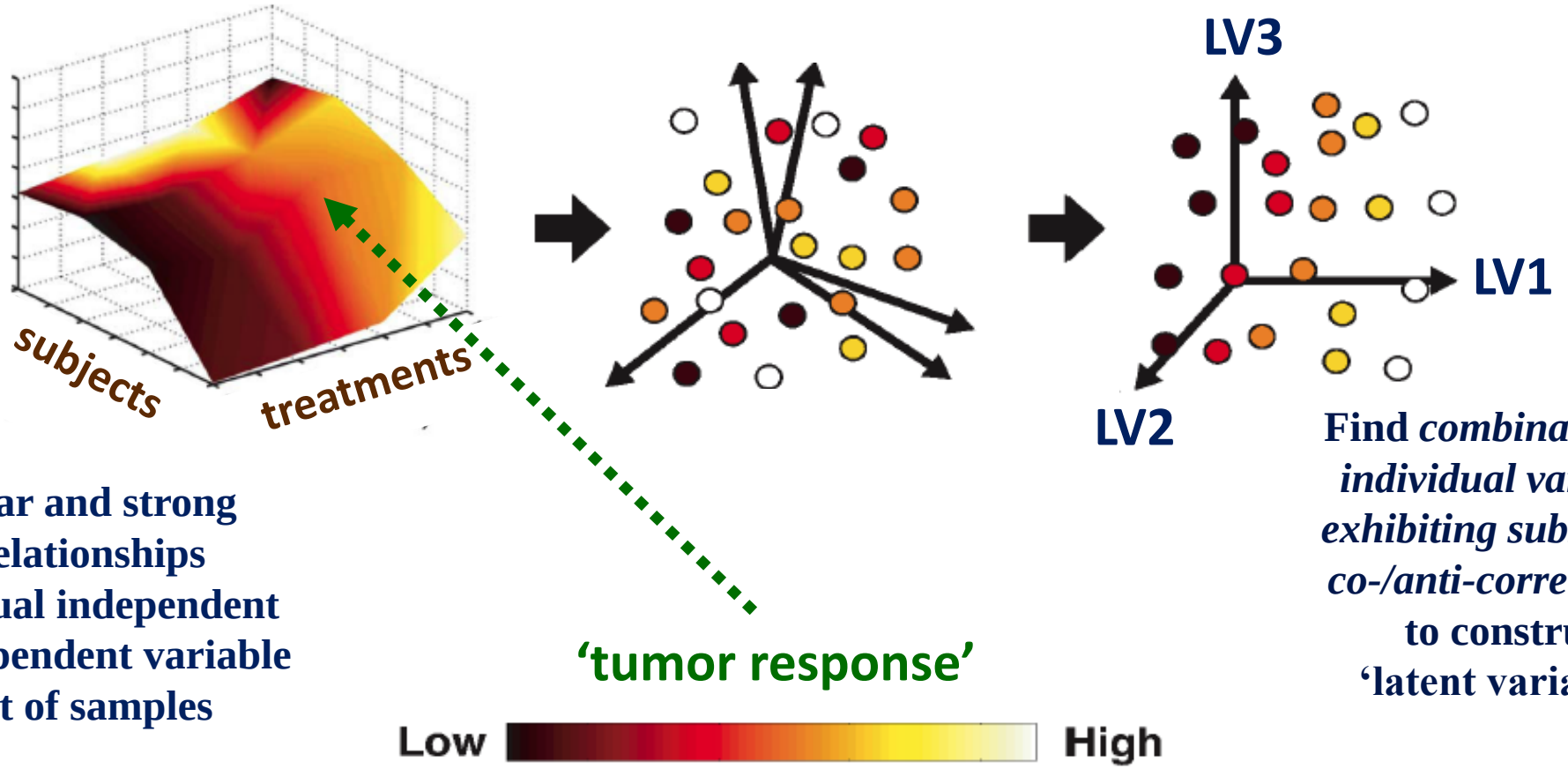
- **traditional Biostatistics permits ’power calculations’ (concerning Yes-vs-No questions), whereas Machine Learning instead employs cross-validation and randomization techniques for ascertaining confidence in and significance of results**

Differences from Traditional Multi-Variate Analyses

- **traditional multi-variate analyses generally assume that input variables are independent – *i.e.*, no co- or anti-correlations among them**

- **Machine Learning methods generally accommodate co-variation among input variables – indeed, they view this is as a typical characteristic feature of system to be discerned**
  - **‘dimensionality reduction’ is sought, in which co-correlated variables provide ‘latent variables’, or “features”**

Dimensionality Reduction to generate 'Latent Variables'



Often, few clear and strong univariate relationships between individual independent variables and dependent variable across full set of samples

Find *combinations of individual variables exhibiting substantive co-/anti-correlations,* to construct 'latent variables'

MIT BE
BIOLOGICAL ENGINEERING

Data Characteristics

- **Types**
  - **Molecular – genomic, transcriptomic, proteomic, …**
  - **Cellular – types, phenotypic functions, …**
  - **Physiological – sensitive-vs-resistant to treatment, …**
- **Amounts**
  - **Measurements – *O*(10^1) cell types / functions, *O*(10^2) proteomic, *O*(10^3) transcriptomic, *O*(10^4) genomic [SNPs]**
  - **Samples – *O*(10^1)-*O*(10^2) lab / clinical studies, *O*(10^3)-*O*(10^4) genomic data-bases**
- **Geometry – relationship between Measurement amounts and Sample amounts // *the most important characteristic***

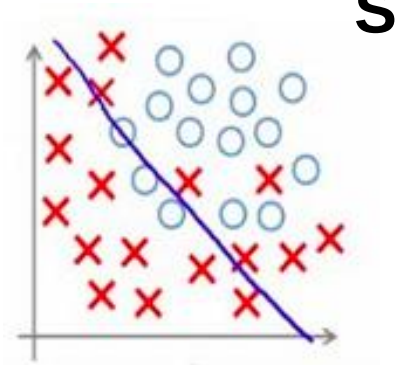
'Big Data' is generally viewed as ~ O(10^6) in samples and > O(10^3) in measurements

MIT BE
BIOLOGICAL ENGINEERING

Data -- Geometry

V

S



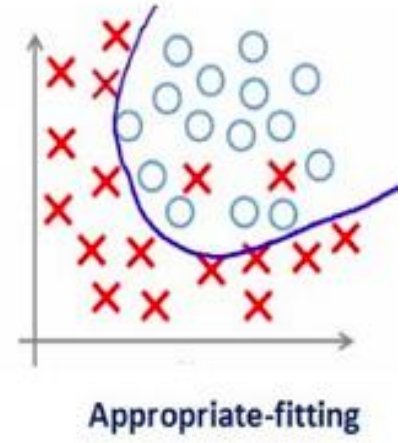
Under-fitting

(too simple to explain the variance)

- **# Samples >> # Independent Variables – "fat" data; can result in 'under-fitting'**
- **# Samples << # Independent Variables – "thin" data; can result in 'over-fitting'**
- **# Samples ~ # Independent Variables – "just right"; generally desirable**

V

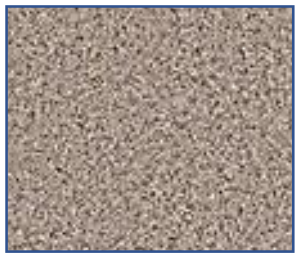
S

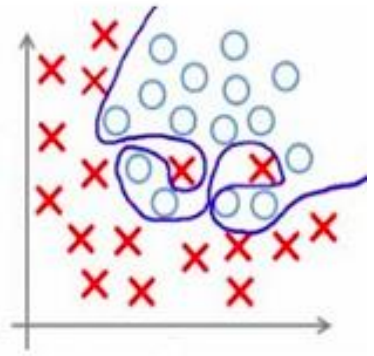


Appropriate-fitting

V

S





Over-fitting

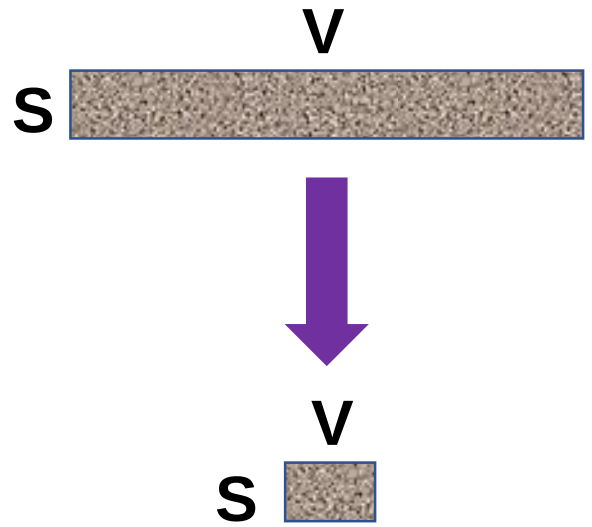
(forcefitting -- too good to be true)

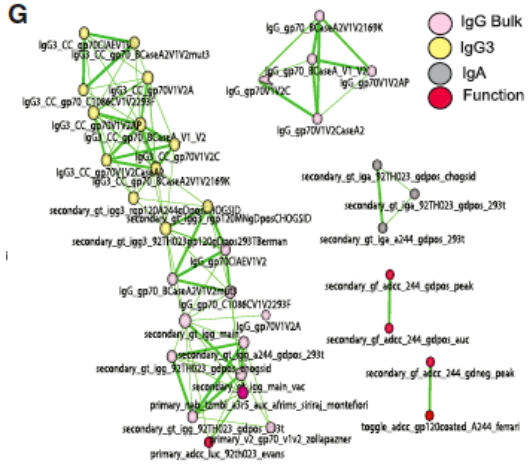
'fitting' = model parameter determination

MIT BE
BIOLOGICAL ENGINEERING

"Thin" Data Geometry is Most Common

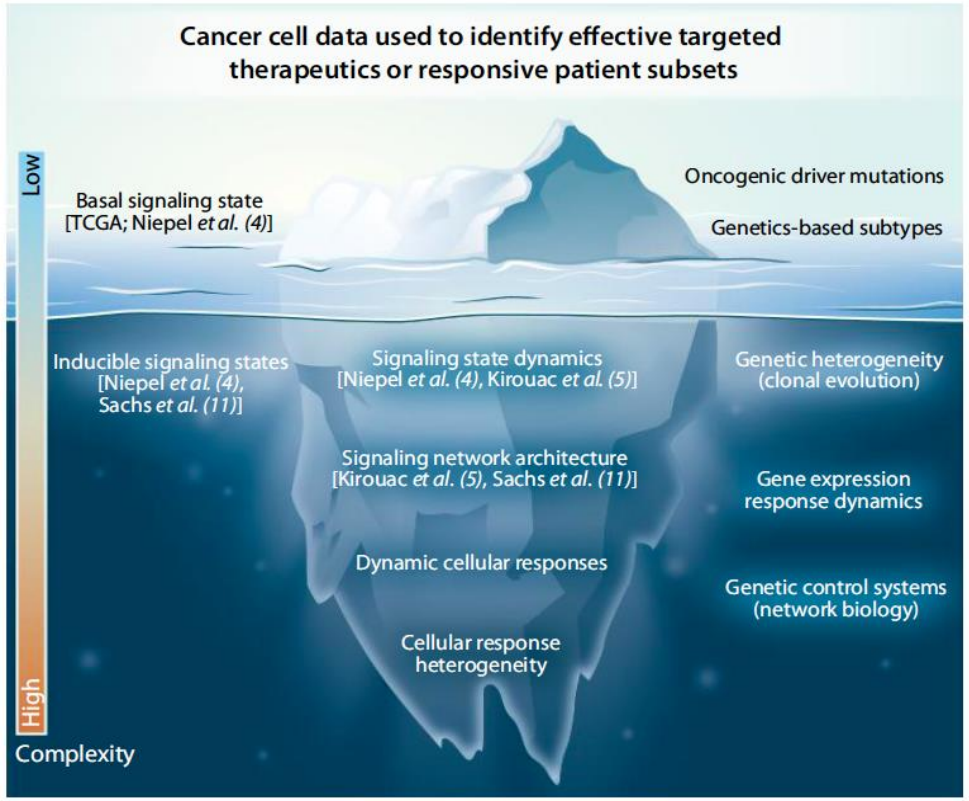
- **Consequence is requirement for 'feature selection', or regularization**
  - **Regularization reduces number of variables used in model**
    - **Robustness of prediction is enhanced, by preventing 'over-fitting'**
    - **Insight is impaired, because relatively small number of variables renders mechanistic interactions difficult to ascertain**
    - **Computational pipeline often then includes an ensuing algorithm to identify pathways / processes enriched in the selected variables**



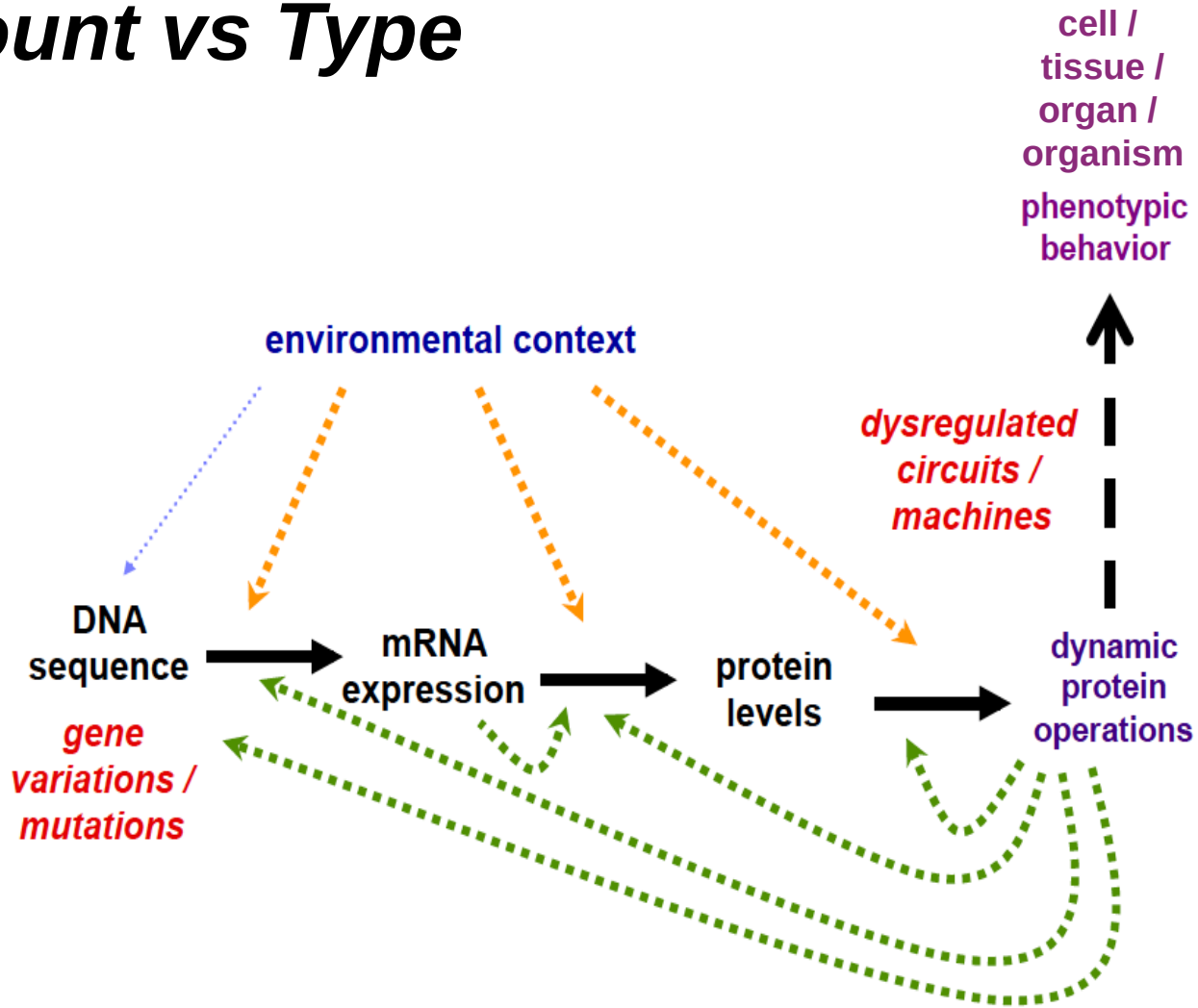


MIT BE BIOLOGICAL ENGINEERING

Data – Amount vs Type



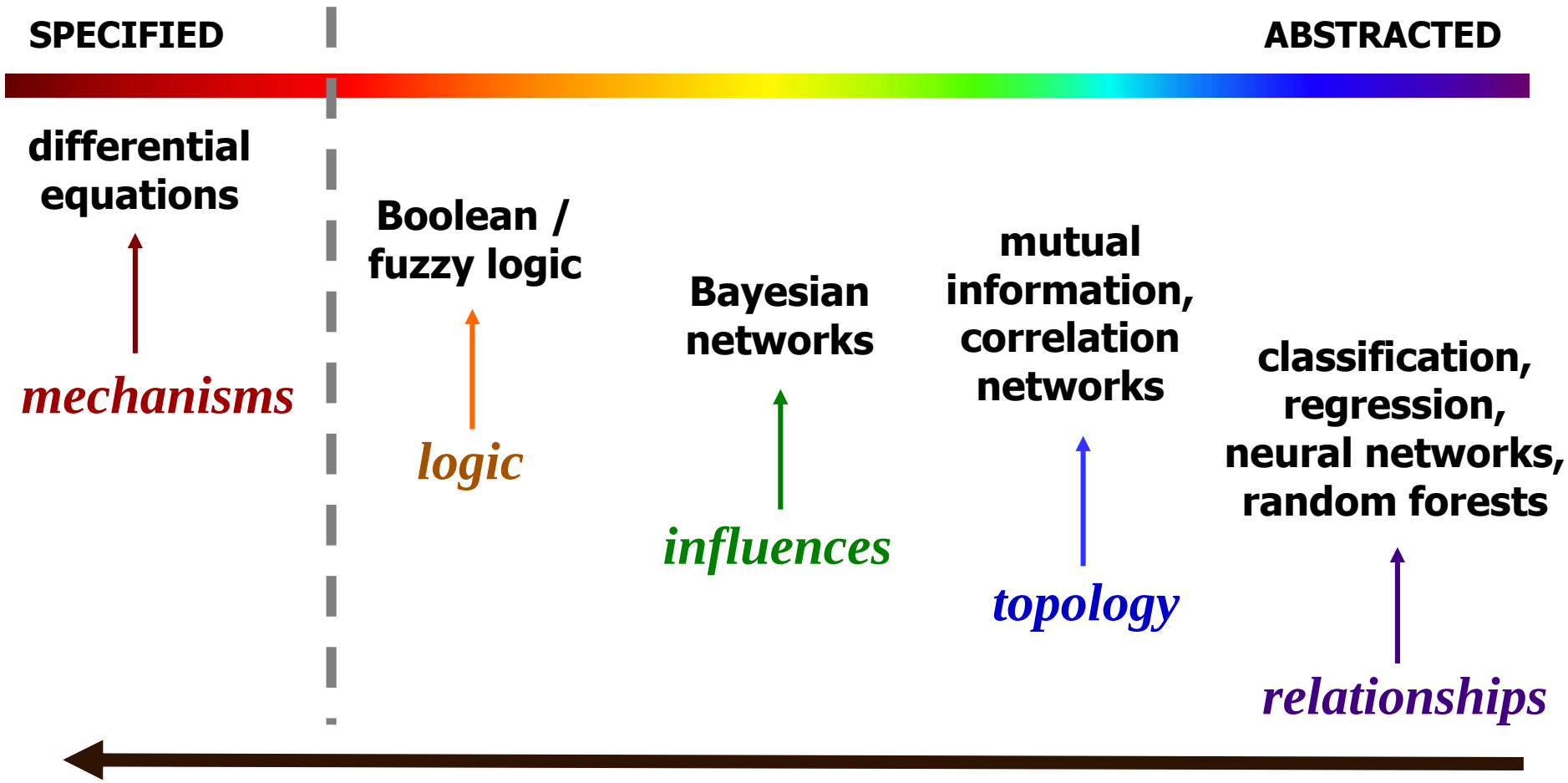
[Quaranta & Tyson, *Science Signaling* (2013)]



The more proximal data-type is to biology / physiology, the more meaningful (and actionable) information it represents

MIT BE
BIOLOGICAL ENGINEERING

Landscape of computational methods that can be pursed in machine learning manner – preference depends on data and goals



adapted from Ideker & Lauffenburger, *Trends in Biotechnology* [2003]

MIT BE
BIOLOGICAL ENGINEERING

Example Immune System Applications
-- Lauffenburger laboratory research

- **Chung et al, *Cell* [2015] – identification of HIV immune response correlates in clinical trials**
- **Moynihan et al, *Nature Medicine* [2016] – analysis of processes underlying effects of diverse tumor immunotherapy approaches**
- **Ackerman et al, *Nature Medicine* [2018] – elucidation of mechanism differences between HIV vaccine administration routes**
- **Kumar et al, *Cell Reports* [2018] – modeling single-cell RNAseq data in tumor microenvironment**
- **Brubaker et al, *PLoS Computational Biology* [2019] – principled framework for translation between mouse and human pathophysiology**
- **Yu et al, *JCI Insight* [2019] – prediction of most effective antibody combination treatments for given HIV reservoir sequence distribution**