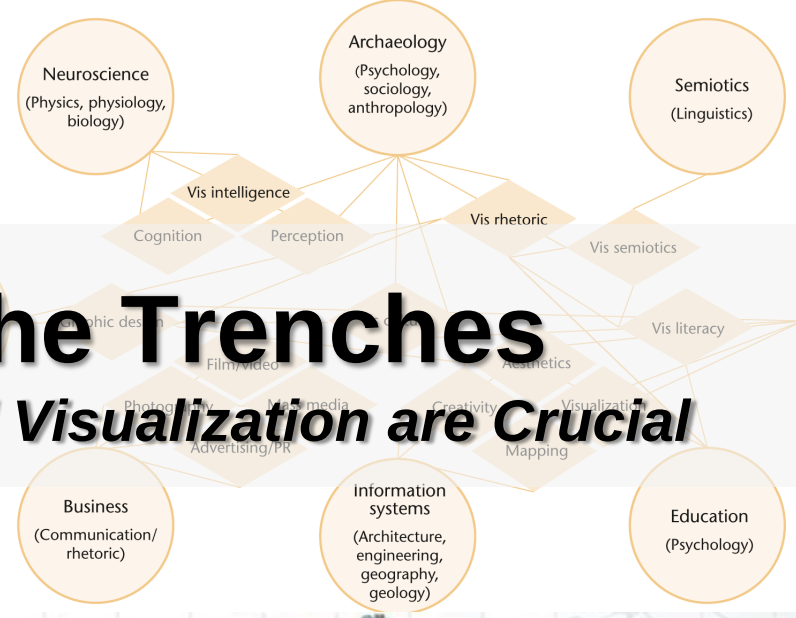


Science in the Trenches

Why Data Wrangling and Visualization are Crucial

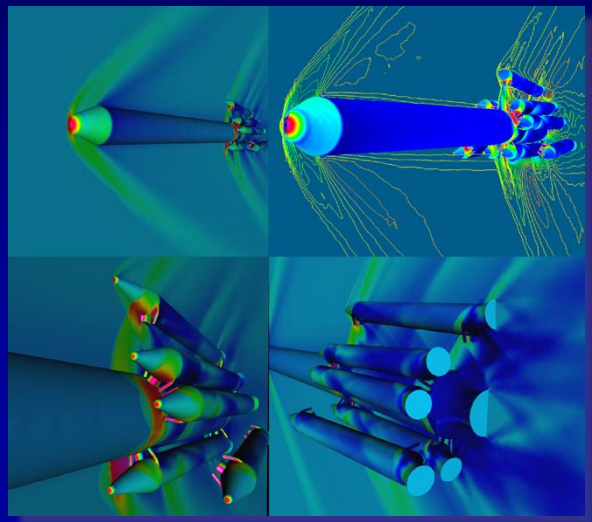
Kelly Gaither

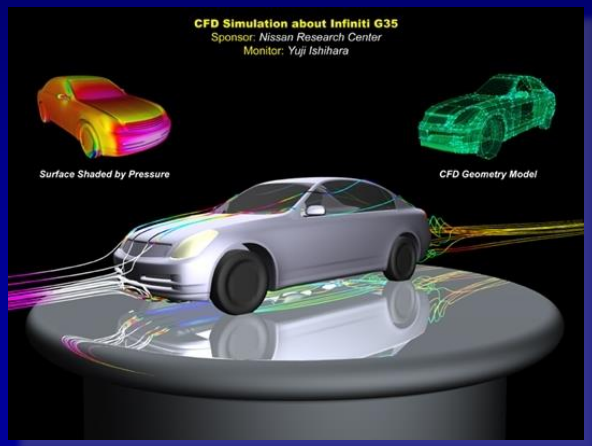
Director of Visualization/Interim Director of Education & Outreach,
Senior Research Scientist, Texas Advanced Computing Center

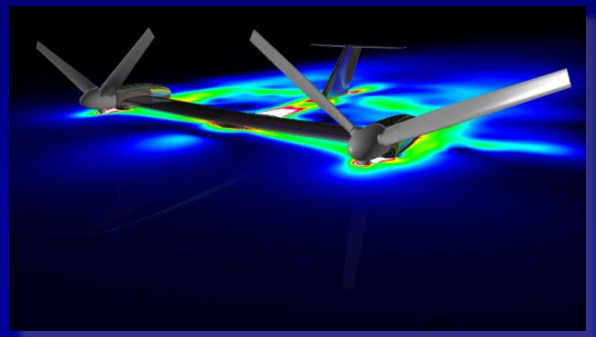
Associate Professor, Women's Health, Dell Medical School

The University of Texas at Austin

Visualizing Science Over 20+ Years





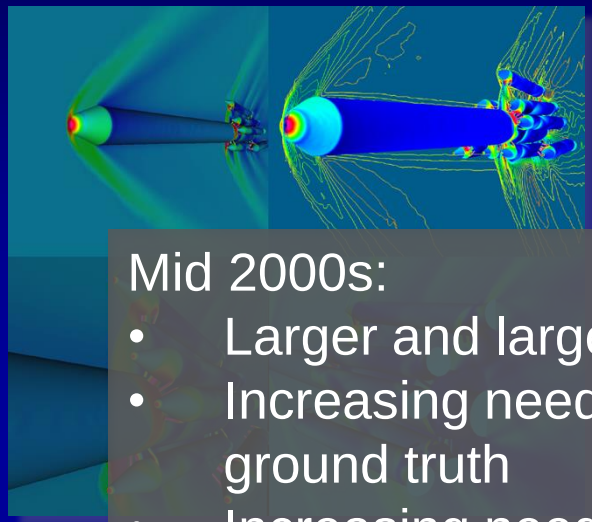


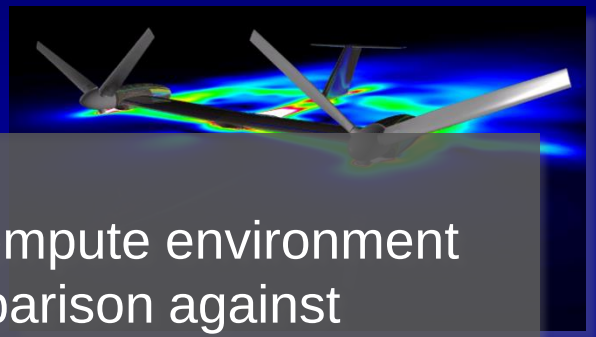
Mid to late 1990s:

- Simulation data (time dependent)
- More memory wrangling rather than data wrangling

TACC

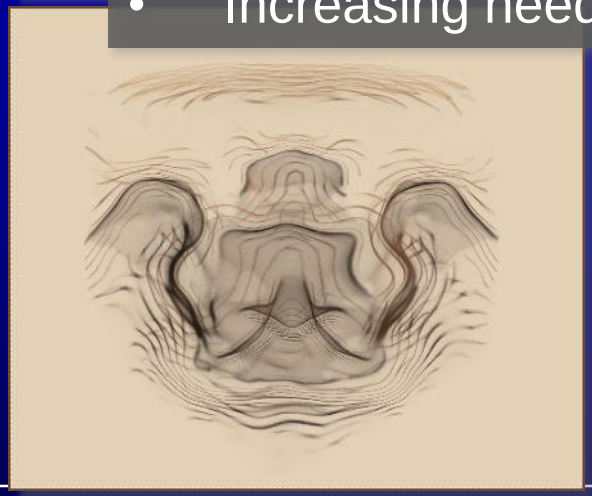
Visualizing Science Over 20+ Years

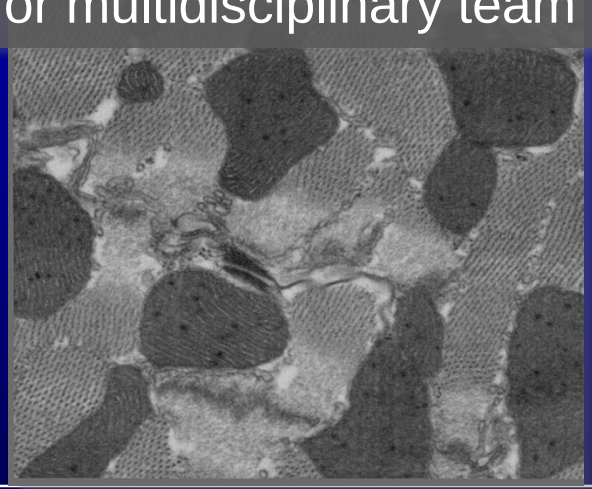




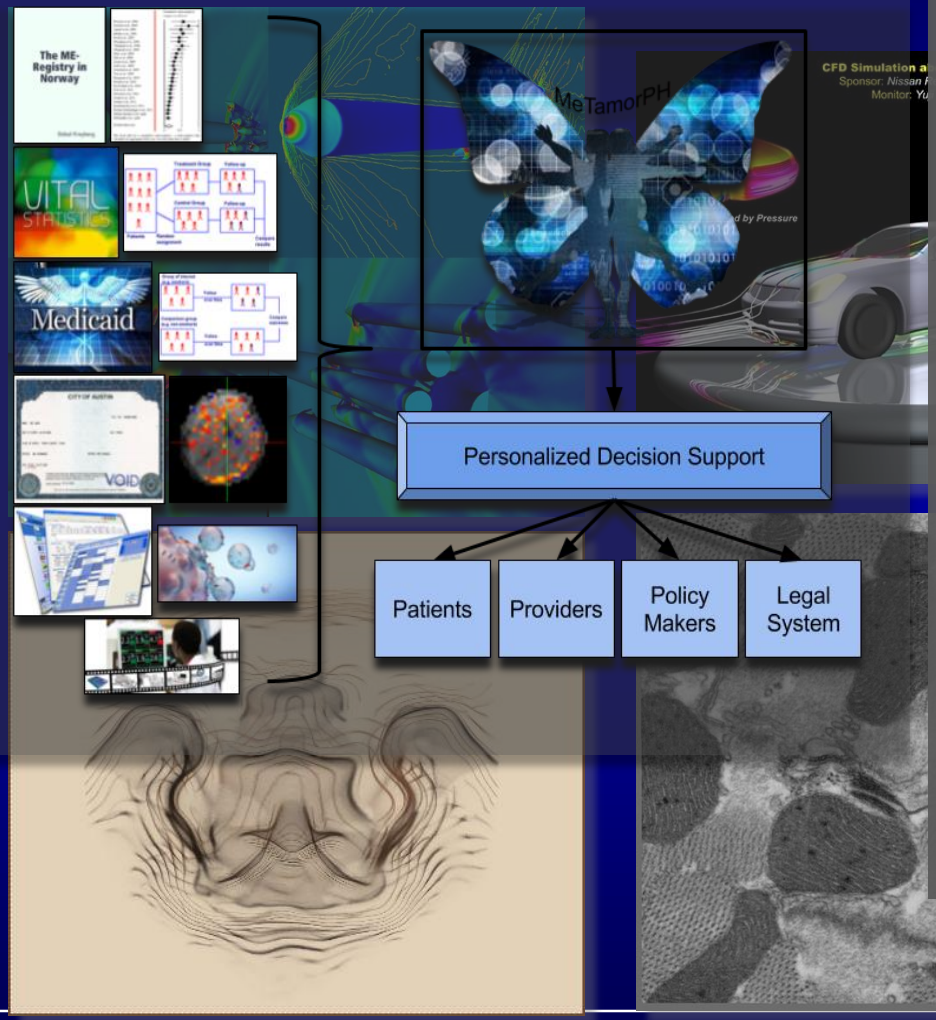
Mid 2000s:

- Larger and larger data sets in distributed compute environment
- Increasing need for true analytics and comparison against ground truth
- Increasing need for multidisciplinary team approach





Visualizing Science Over 20+ Years

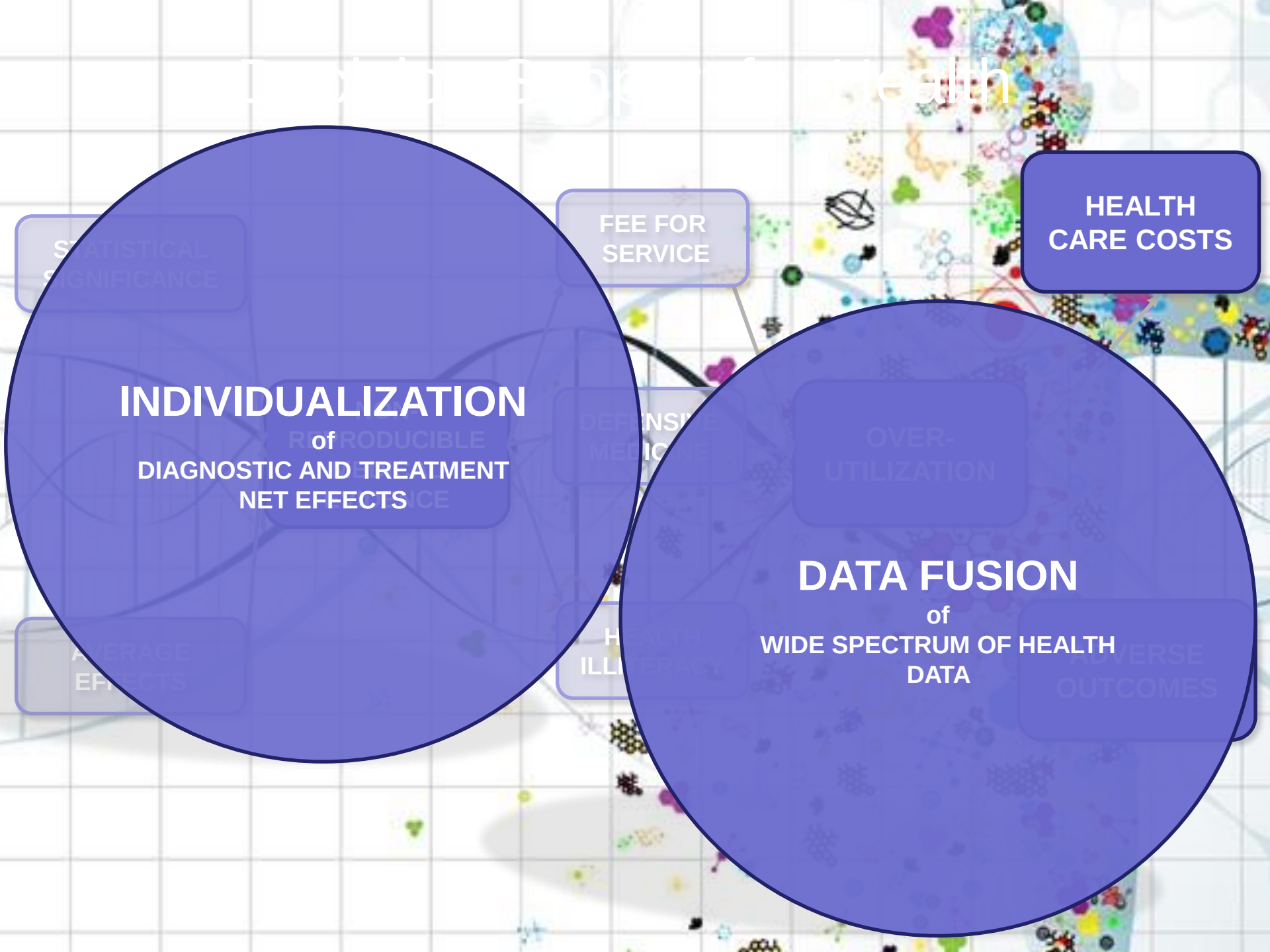


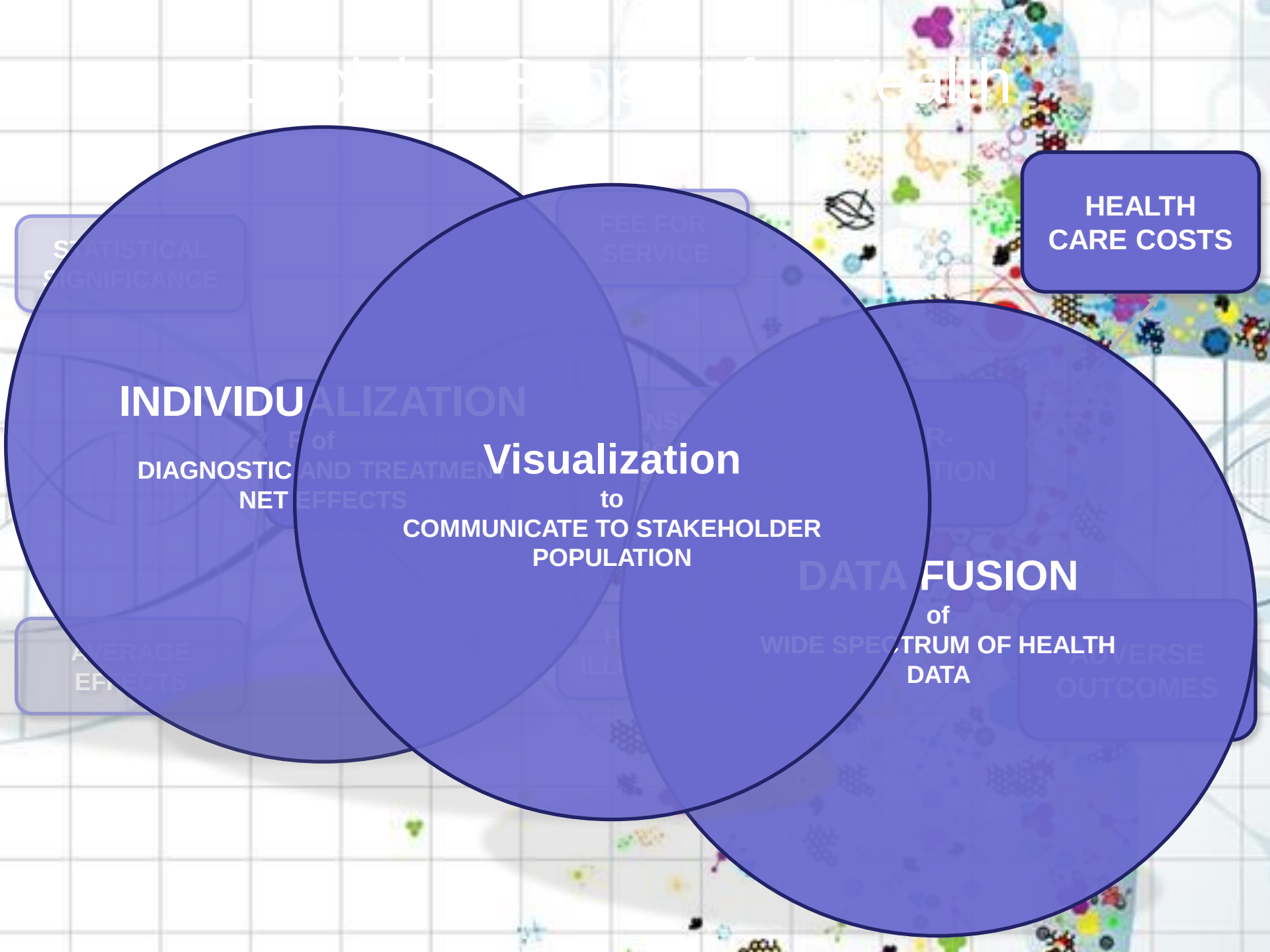
Current:

- Heterogeneous sources of data needed; located in geographically disparate locations
- Combination of highly secure and publicly available data
- Need for cleaning/matching/verifying each source 100% of the time before visualizing
- Data generated with almost no thought of scalability or technically advanced access mechanisms
- No structure for longitudinal tracking
- No provenance

TACC


STATISTICAL SIGNIFICANCE
AVERAGE EFFECTS
NON-REPRODUCIBLE MEDICAL EVIDENCE
FEE FOR SERVICE
DEFENSIVE MEDICINE
HEALTH ILLITERACY
OVER-UTILIZATION
HEALTH CARE COSTS
ADVERSE OUTCOMES


FEE FOR SERVICE
HEALTH CARE COSTS
INDIVIDUALIZATION
of
DIAGNOSTIC AND TREATMENT NET EFFECTS
DATA FUSION
of
WIDE SPECTRUM OF HEALTH DATA


HEALTH
CARE COSTS
INDIVIDU
DIAGNOSTIC
NET
Visualization
to
COMMUNICATE TO STAKEHOLDER
POPULATION
FUSION
of
TRUM OF HEALTH
DATA

Data Wrangling Constitutes ~95% of the Process

- Sources:
  - Privately insured patients – inpatient/outpatient
  - Medicaid – inpatient/outpatient
  - Uninsured – inpatient/outpatient
  - Medicare
- Social Determinants?
- Interoperability?
- Standards?
- Matching?
- Missing/Incomplete/Inconsistent Data!

Data Wrangling Constitutes ~95% of the Process

- Sources:
  - Privately insured patients – inpatient/outpatient
  - Medicaid – inpatient/outpatient
  - Uninsured – inpatient/outpatient
  - Medicare – inpatient/outpatient
- Social Determinants?
- Interoperability?
- Standards?
- Matching?
- Missing/Incomplete/Inconsistent Data!

Visualization is key to data wrangling to help us mine through the issues!

How Much Data Are We Looking At in the Future?

~30M People

- Individual Genetic Code
  - 3M Variants == 125MB
- EMR/EHR/HIE per year
  - < 1MB healthy adult
  - 40MB w/o images for unhealthy adult
  - 300MB w images for unhealthy adult
- Life History Trails per year
  - ~50TB/year

How Much Data Are We Looking At in the Future?

~30M People

- Individual Genetic Code
  - 3M Variants == 125MB
- EMR/EHR/HIE per year
  - < 1MB healthy adult
  - 40MB w/o images for unhealthy adult
  - 300MB w images for unhealthy adult
- Life History Trails per year
  - ~50TB/year

Data for population the size of Texas: 1.59ZB/year

TACC

Decision Support

- Requires Data (Models)
  - Reliable
  - Reproducible
  - Robust
  - Accessible
  - Interoperable
  - Analyzable for multiple types of analysis

  - Intelligence
  - Flexibility
  - Adaptability


"The Data Says WHAT?"
TRUSTING DATA
& how to make it relevant to you
activitycloud.com/blog

For more information, contact:

Kelly Gaither
kelly@tacc.utexas.edu

Questions?