Big Data and the search for a microbial cause for disease

1. The role of microbes in autoimmune diseases (type 1 diabetes and celiacs) as well as premature birth.
2. Understanding citrus diseases and finding solutions.
3. Understanding how land use affects soil health through microbial functions.

Eric W. Triplett, Microbiology and Cell Science Dept., IFAS

Big Data and the search for disease prevention and cure.

Type 1 diabetes – disease of the young.

Hypothesis:

An aberrant gut microbiome leads to a leaky gut which in turn initiates a cascade of immunological responses resulting in self destruction of insulin-producing cells in the pancreas.

Collect lots of data:

Bacteria in the gut across years of stool sampling, diet, antibiotic use, infectious episodes, disease progression, subject genetics, etc.

One of our goals:

Identify bacteria whose abundance can predict future disease in children.

Use those bacteria to design interventions to prevent disease.

Can these interventions reverse disease?

Machine learning – Austin Davis-Richardson

Data generation, communication, and computation:

1. Need continual investment in first iterations of new technologies.
2. Need robust computational and storage capacities.
3. Need rapid data transfer capabilities.

Our infrastructure has improved greatly but requires consistent investment.

Factors limiting Big Data at UF

- Need more intellectual capital.
- Fee-for-service informatics centers needed.
- **Training, training, training**

  At all levels, undergraduate, graduate, postdoctoral, faculty, administration.

- Are our administrative structures ideal for the future?

Big challenges addressed by Big Data

Interpretation of the data – what does it mean?

The Signal from the Noise: Why So Many Predictions Fail – and Others Don't

by Nate Silver

Bayesian Nonparametric Covariance Regression

Machine learning

More scholarship needed in these areas.

Current state of Big Data generation

- Projected growth over the next 5 years

  Decline in DNA sequencing costs has slowed recently. Niche products are now needed rather than simply higher throughput. eg. PacBio

- Anticipated infrastructure

  The largest grants are going to those with the best infrastructure and highest intellectual capital.

- Analysis challenges

  Need more intellectual capital and training

- Interpretation challenges – what does it all mean?

Last challenge – data integration

- Multi 'omics datasets
  - genomics
  - methylomics
  - transcriptomics
  - proteomics
  - metabolomics
  - metadata

From single cells to single organisms to communities and ecosystems.

How do we separate signal from noise?