Data Driven Computational Science (DDCS) in Drug Development

Robert A. Lodder
University of Kentucky

CIC Pharmaceutical Sciences workshop during the 2021 10th International Conference on Software and Computer Applications (ICSCA 2021), Kuala Lumpur, Malaysia February 2021

OUR DDCS OBJECTIVE

Keep 4 novel Investigational New Drug (IND) programs running at FDA.

The United States Food and Drug Administration's **Investigational New Drug (IND)** program is the means by which a pharmaceutical company obtains permission to start human clinical trials and to ship an experimental drug across state lines (usually to clinical investigators) before a marketing application for the drug has been approved.

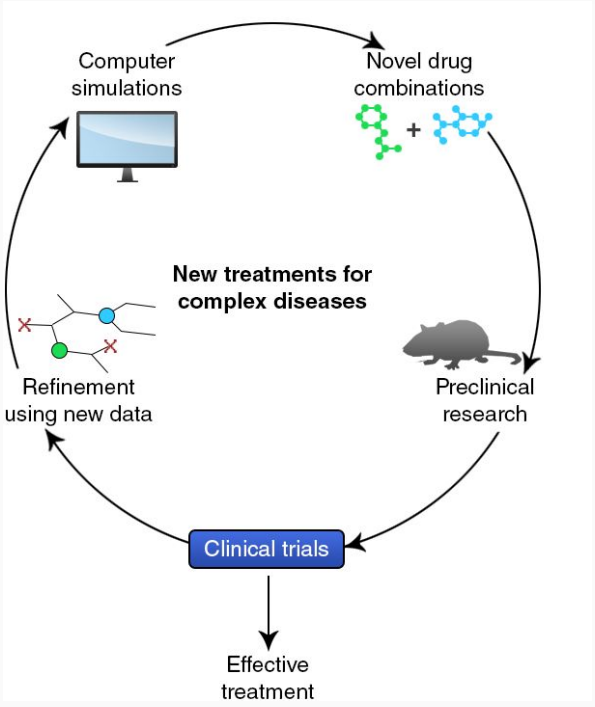
DDCS Mission:

To develop combination drugs for orphan tropical and pediatric diseases, testing potential binary and ternary therapies in dynamic data-driven simulations at various combinations of two or three points in the pathways, choosing to bring forward the most valuable combinations that show the lowest probability of side effects.

The problem

Developing novel therapies for new diseases is risky and expensive.

Not just for Lambert Eaton Myasthenic Syndrome and Menkes Syndrome, but also for other rare diseases, like Ebola, Chikungunya, Prader-Willi Syndrome, and ADHD in FXS.



The solution

Artificial Intelligence

Simulations and AI speed the process of development and reduce the risk.

AI System

AI must look at more than just science to solve the problem.

Market

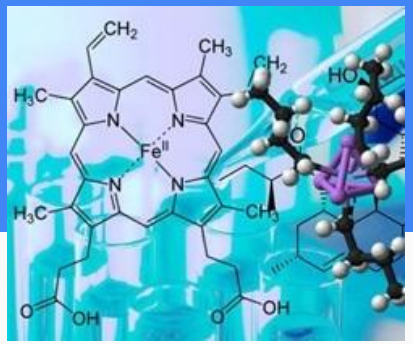
Disease prevalence, reimbursement, competition, pricing, capital requirements and capital availability

Intellectual Property

Patentability, contracts, portfolios, encumbrances, white space, landscape, valuation, quality, monetization, licensing

Regulatory

Pediatric rare disease, tropical disease, fast track, breakthrough status, jurisdictions, guidances, quality management, outcomes



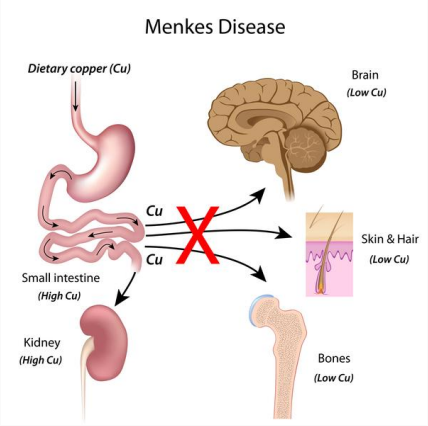
Science

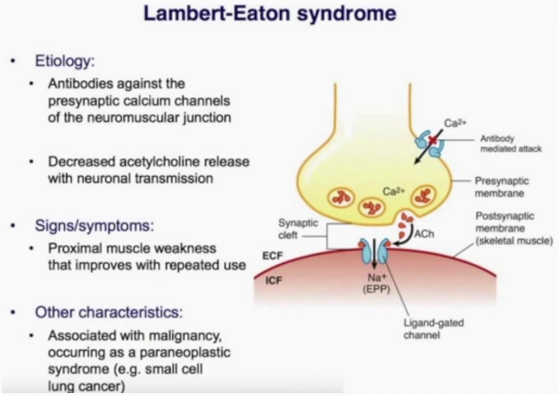
Pathways, signaling, receptors, genes, RNAi, cells, immunotherapy, populations, druggability, manufacturability

AI Approach Leads to New Therapies

AND a 50% increase in INDs!

1. *Methylpropanedihydrazide Compounds for Menkes Syndrome*, US patent application number 62732350, filed Sep 17 2018
2. *Nanotube Delivery Device for MPDH Compounds for Menkes Disease*, US patent application number 62732379, filed Sep 17 2018
3. *Biodegradable Nanoparticles for Controlled Release of ICP4 Compounds for LEMS*, US patent application number 62722146, filed Aug 23 2018
4. *ICP4 Compounds for Treatment of Lambert Lambert-Eaton Myasthenic Syndrome*. US patent application number 62690557, filed Jun 27, 2018
5. *Preparation of ICP4 Compounds.* US patent application number 62690606, filed Jun 27, 2018





Two General Objectives for New AI

1. Explore and develop novel new AI algorithms and approaches for discovery of scientific laws and governing equations for complex physical phenomena
2. Explore new approaches to assess where data are too sparse, noisy, or are otherwise inadequate to build predictive models; to generate testable hypotheses; to identify high value experiments that could alleviate the problems of data shortfalls; and to *quantify the confidence of predictions outside of the training space*. (more on this later)

Generations of AI

AI still falls short of being a trusted and valued collaborator in scientific discovery and technology development.

“**First Wave**” (rule based) (and this is still very fast, where it exists.)

“**Second Wave**” (statistical learning based) AI technologies. Mostly the current state.

“**Third wave**” AI theory and applications that address current limitations by enabling machines to contextually incorporate available data, facts, models, heuristics, and additional information to achieve greater robustness, adaptability and generalizability.

AI Challenges

A key challenge is how well statistical AI methodologies can generalize beyond the narrow set of questions they are initially trained on. As AI and deep learning are increasingly being applied to predict the behavior of complex nonlinear dynamic systems where the system state is not fully observable and the data lack coverage, this question of generalizability becomes even more important. Generalizability in today's state-of-the-art approaches remains poorly understood and quantified. This leads to wasted time and money in new drug development.

AI is not yet a valued partner in scientific discovery and the scientific process

Current “second wave” AI systems are in general incapable of “understanding” whether the questions they have been trained on can be usefully answered, or whether they even make sense given the provided set of input data.

Today’s AI, if incorporated at all in a process, is applied in Yes-No and regression types of analyses of big data. Human scientists handle most of the knowledge-centric aspects of the process, often relying on human experience, heuristics, ad hoc and domain-specific methodologies.

What would we like AI to do for us?

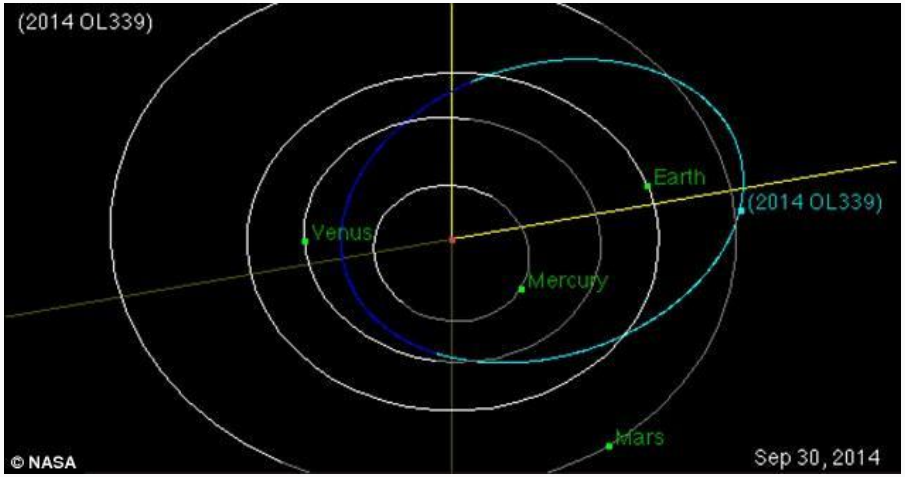
Assess the adequacy of the data, formulate questions, generate hypotheses and testable predictions, identify additional high value experiments that could alleviate data shortfalls, and quantify confidence of predictions outside of the training space.

Currently these all lie outside the capabilities of most state-of-the-art AI approaches and fall into the domain of human experts.

Example for AI: Tycho Brahe's data (along with some supplemental data) mapping the trajectories of planets in the solar system as seen from an observatory on earth.

The trajectories, although complicated, clearly correspond to reduced dimensional manifolds.

So, could an AI Research Associate accomplish or assist in the following:



- Recognize that the coordinate system of the observed data is non-ideal and that a better coordinate system might lie two or three unitary transformations removed?
- Identify areas where data are inadequate, e.g., due to ambiguities, gaps, inadequate coverage, inadequate resolution, inadequate signal-to-noise, insufficient modalities, confounding variables, measurement errors, or outlier results, and provide guidance for obtaining the highest-value augmenting data?
- Arrive at a parsimonious, predictive model (e.g., uncover Kepler's laws)?
- Arrive at a parsimonious, predictive model with the potential to generalize across multiple domains (e.g., uncover Newton's laws and a 1/r gravitational potential model for gravity)?

- Deduce conservation of energy, momentum and angular momentum?
- Propose testable experiments to validate hypotheses and new models (e.g., predict the next return of Halley’s comet)?
- Uncover a problem with Mercury’s orbit and deduce a better abstraction such as Einstein’s general relativity?
- Test for consistency across multiple datasets and observations?

Example 2. Multiscale structured materials.
Density Functional Theory and other ab initio modeling methods at the primitive level are computationally intensive and don't scale well beyond a small number of atoms.

Modeling materials properties across scales generally requires a hand-tweaked stitching together of a patchwork of different models at different scales.

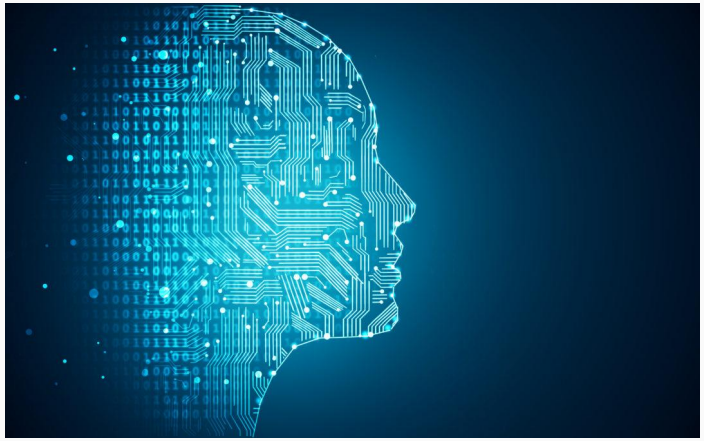
What happens if one changes constituents or adds different dopants?

How do defects and microstructures affect properties?

Could AI derive deep insights from existing experimental and model data obtained for similar material structures?

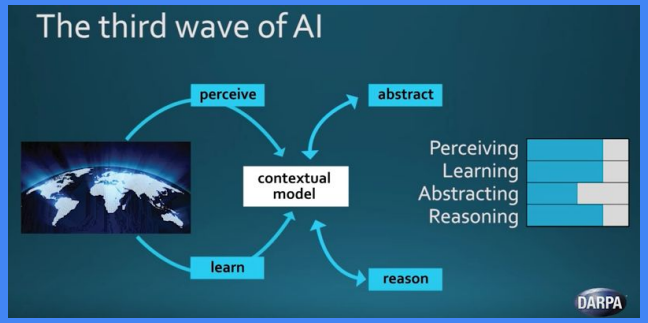
- Could it generate useful multiscale representations and models that generalize gracefully across different constituents?
- Could it guide the discovery of new materials with desired properties?
- And importantly, could it provide feedback regarding the adequacy of the data, where additional data would help, and how confident the predictions are in regimes not covered in the training database?

Humans, while capable of generating deep insights, lack the bandwidth to process the high throughput data of complex, high dimensional systems.



For AI to collaboratively support humans, AI technologies must be able to process and distill big data into physically-relevant representations comprehensible and suitable for collaboration.

Pathway to Third Wave AI



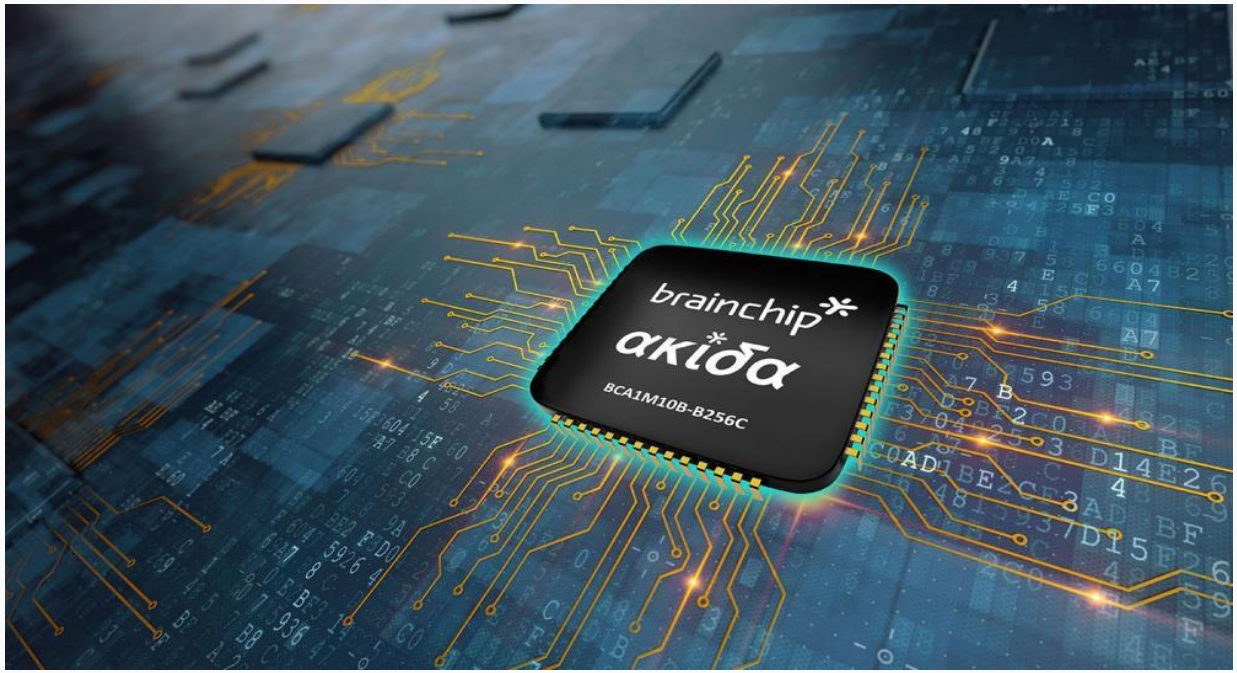
1. AI should assess both accuracy and generalizability of models. How well does the AI predict evolutionary trajectories, dependencies on input or control parameters, and AI response to perturbations?

2. Find and resolve inconsistencies and ambiguities, either in the data or in model predictions. Can the AI answer new questions of the system for which it has not been trained? AI should quantify or provide bounds for model performance for both in-training and out-of training input space.

3. Quantify the extent to which additional data, higher resolution data or new data modalities can help. Can the AI identify high value experiments to better inform and constrain its models?

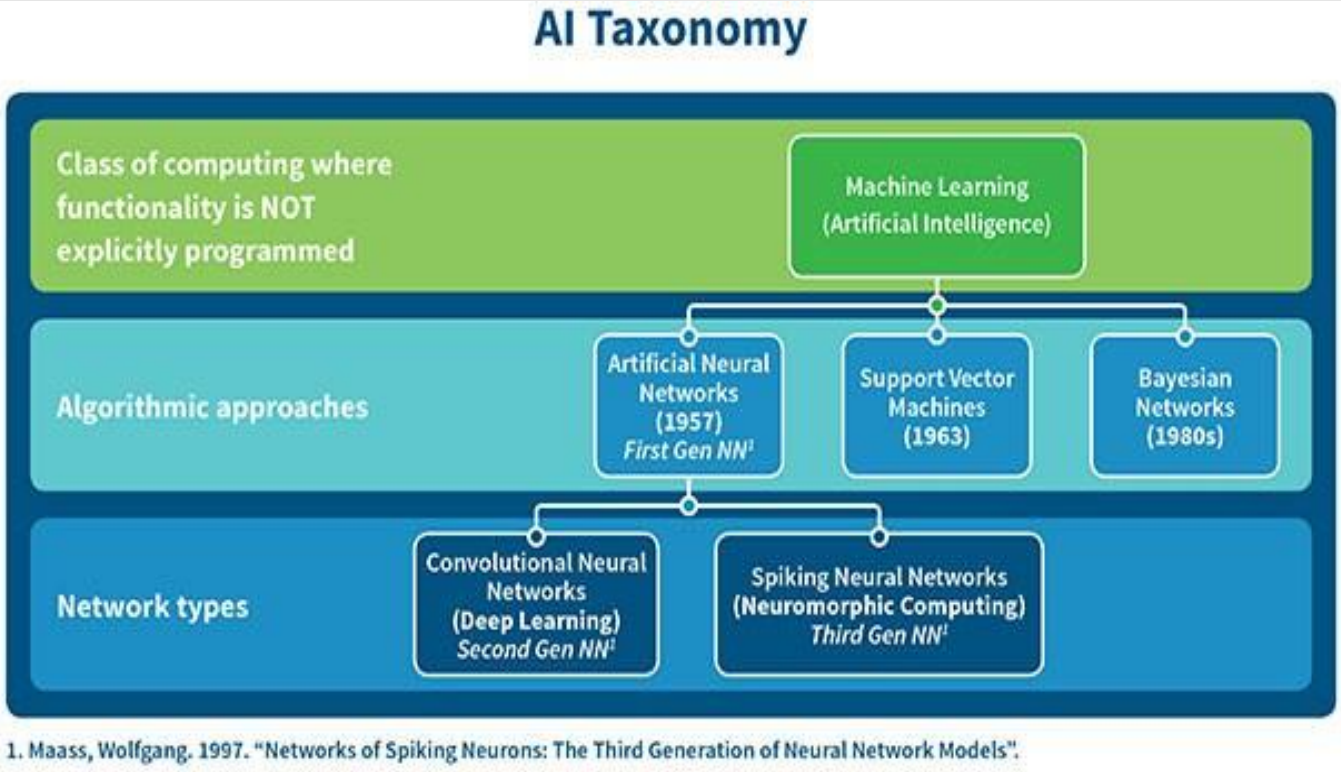
Hardware for Third Wave AI

New Hardware for Third Wave: BrainChip Uses Akida Architecture



Neuromorphic System-on-Chip (NSoC) is based on spiking-neural-network (SNN) technology. Armed with approximately 1.2 million neurons and 10 billion synapses, the Akida NSoC spiking-neural-network chip takes on training and inference tasks.

Spiking neural networks offer an alternative to the convolutional neural networks that have become very popular.

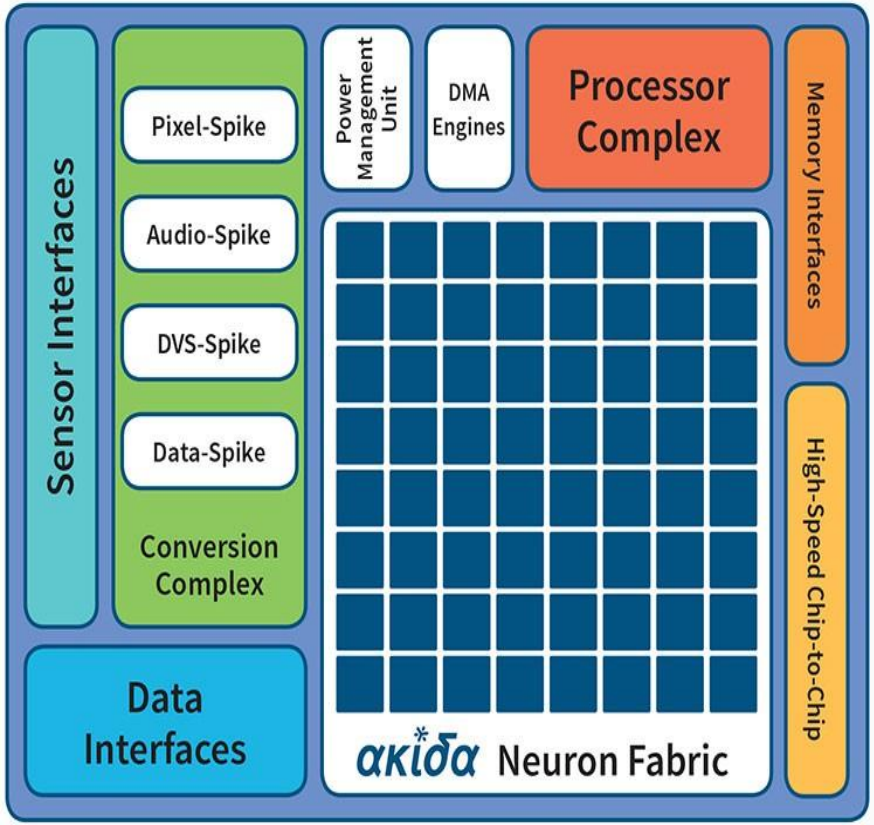


SNNs translate data into a stream of spikes that also flow through the neural network. These are discrete events rather than the CNN's array of values.

BrainChip's Akida NSoC is a self-contained chip with a conventional processor plus the Akida neuron fabric.

sensory input including Dynamic Vision Sensor

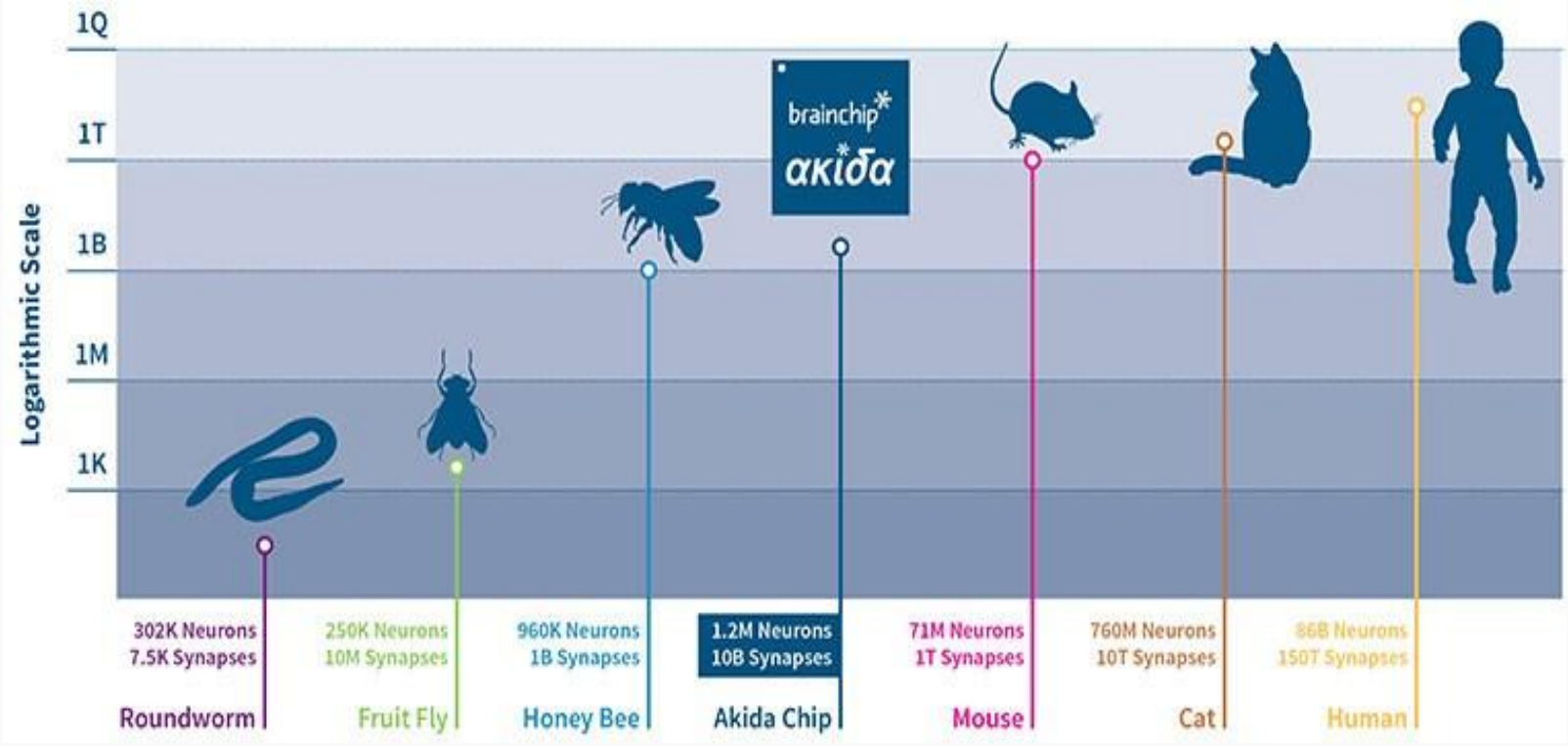
for market, IP, regulatory and scientific data



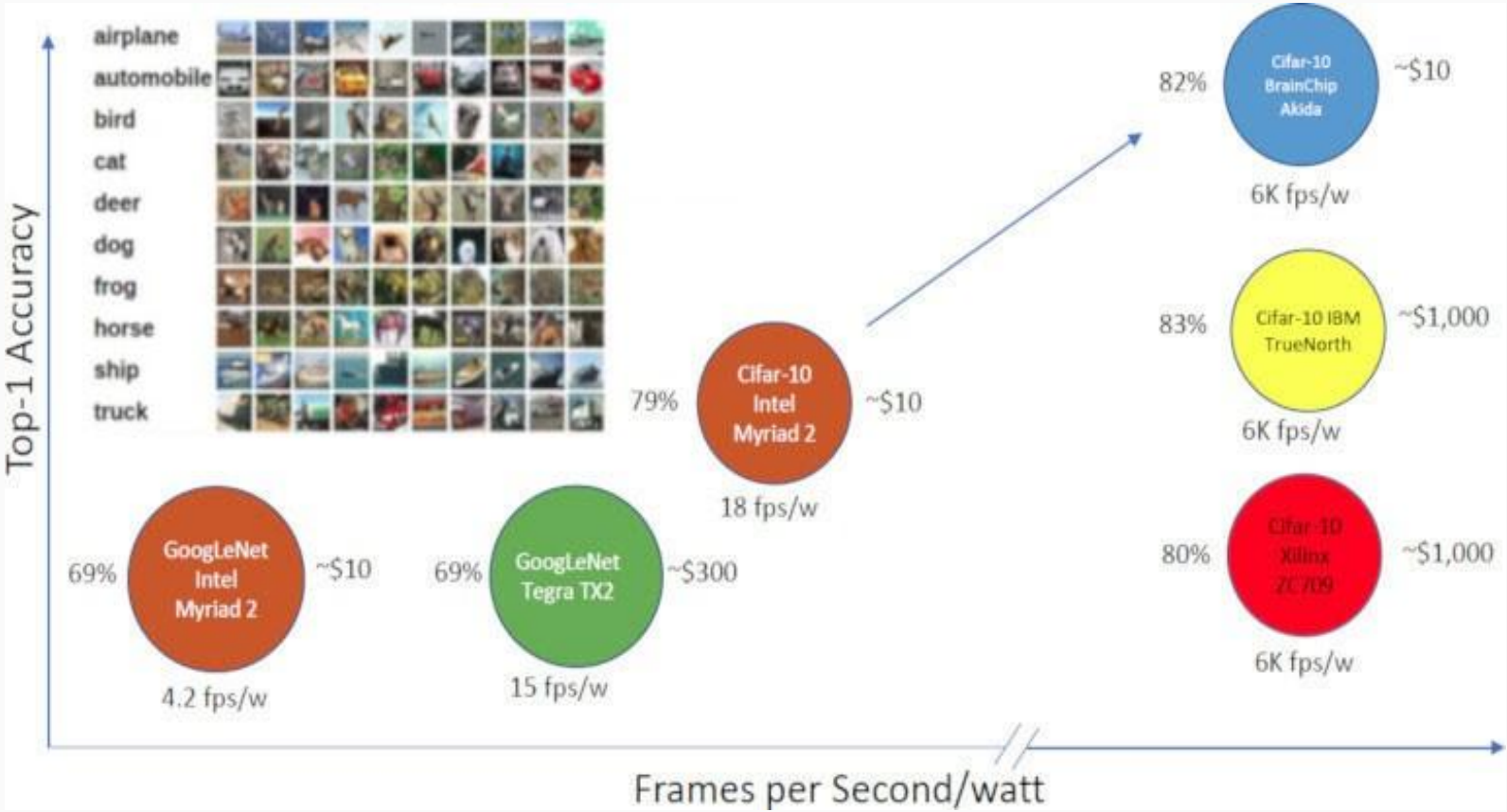
for data storage

for massively parallel machines

The Akida NSoC is moving up the ladder in terms of complexity with 1.2 million neurons and 10 billion synapses in a multichip system.



The Akida does very well with the popular Cifar-10 dataset. The Cifar-10, which identifies 10 common objects, uses less power and is significantly less costly.



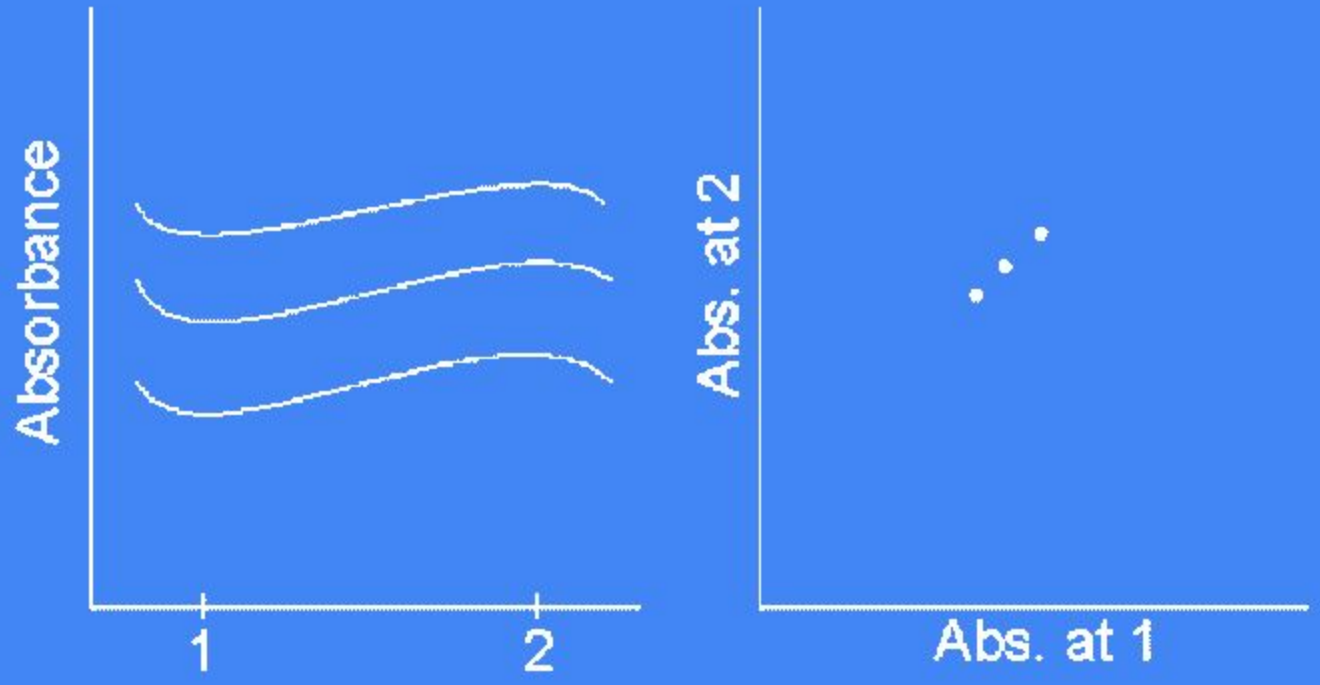
BEST Engine for Adaptive Resonance

Adaptive Resonance Theory in Neural Networks

ART works in a series of levels. Over time, images that are learned are represented in categories on upper levels of the system.

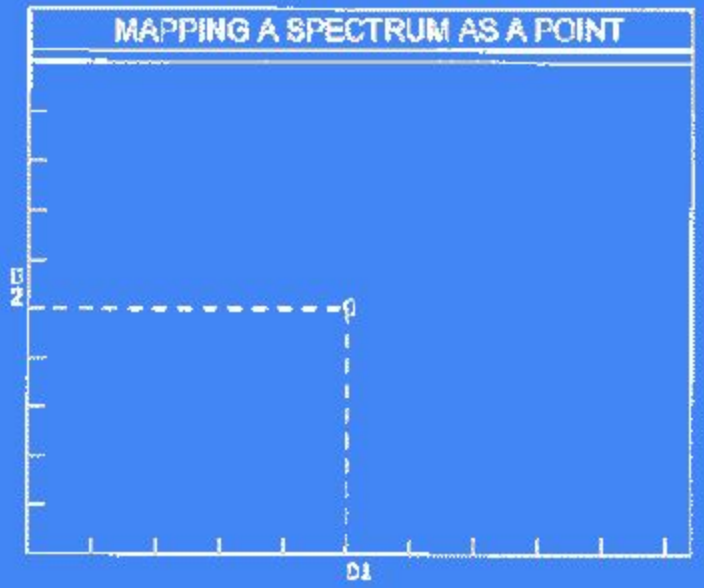
When a new image of a **truck** is fed into the first level, for example, it activates one of the memory categories and is sent up to be matched with that category. This "bottom-up" process is a common property among adaptive systems. What makes an ART system different is that at the same time the signal is going up, the receiving category is sending a signal "top-down" to the first level to make sure an adequate match exists. A category looks back down and says, “What am I learning; what category should this be in?" If the match isn't close enough, an orienting subsystem is activated, which automatically closes off the activated category and searches for a category that would give an **adequate match**. If no adequate match is found, the system creates a **new category**.

Projecting Data into Hyperspace



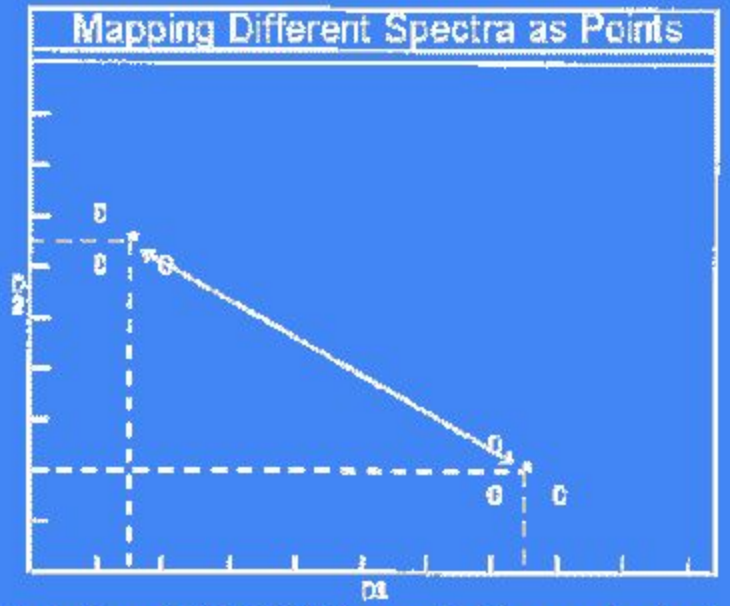
Provides a means of measuring increased analytical power

Projecting Spectra into Hyperspace



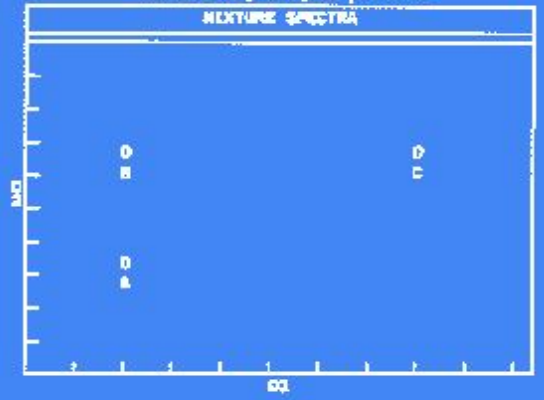
Displacement on each axis represents signal intensity each information vector.

Similar spectra cluster in similar regions of hyperspace.

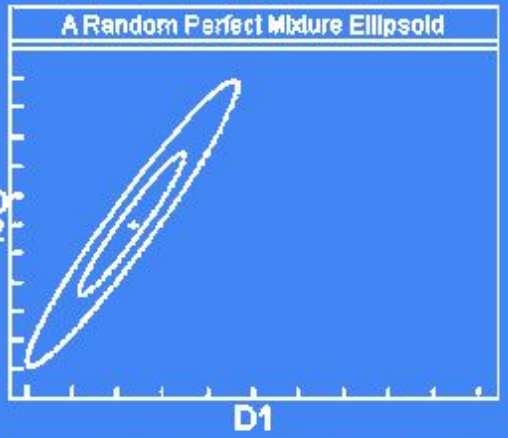


The best technique is the one that separates the samples best (smallest group size and largest distance between groups).

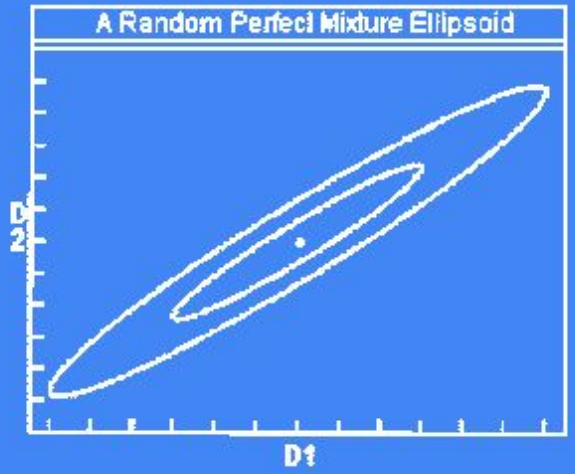
The shape, size and skew of such clusters is not necessarily easy to predict.



Make random mixtures of three components whose spectra are known perfectly.

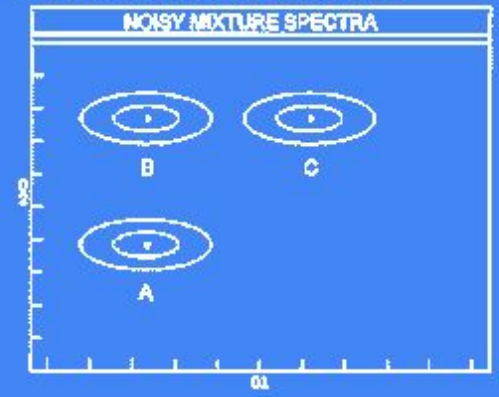


d-variate normal clusters result from the random mixing.

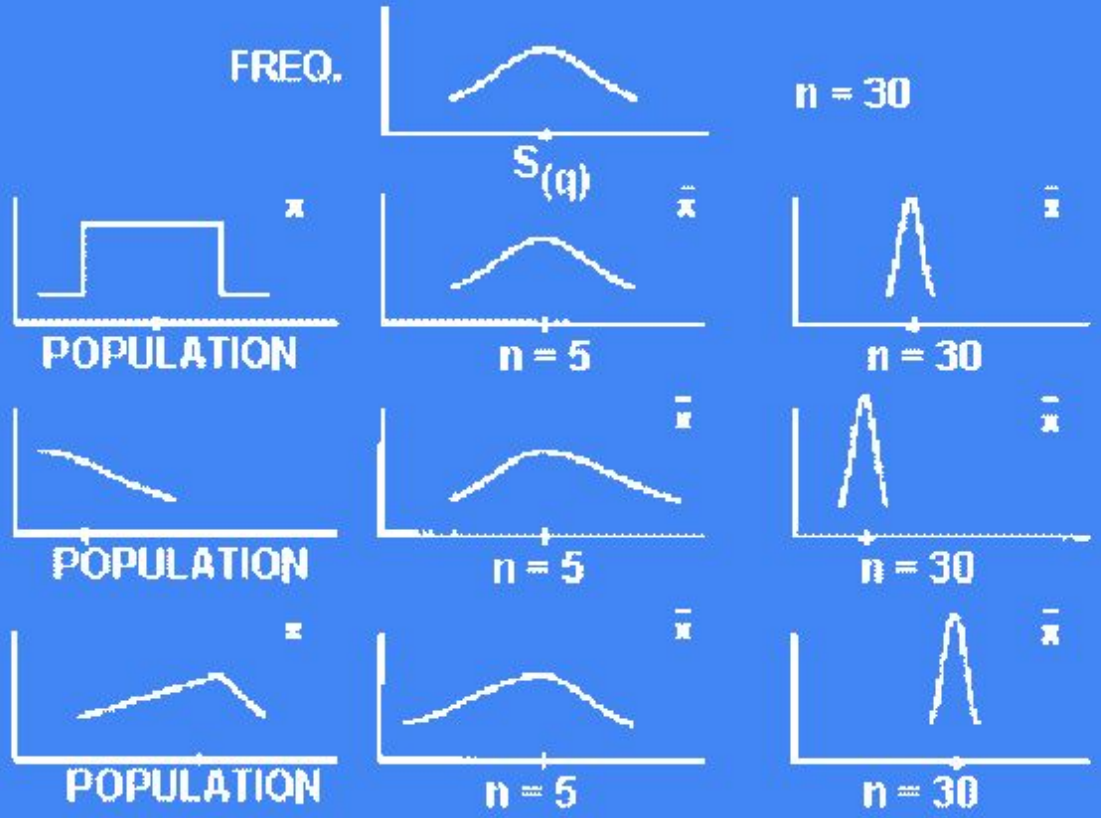


Moving one of the constituent points merely changes the size and orientation of the cluster of mixtures.

When error is allowed to enter in the locations of the components, a more complex picture emerges.

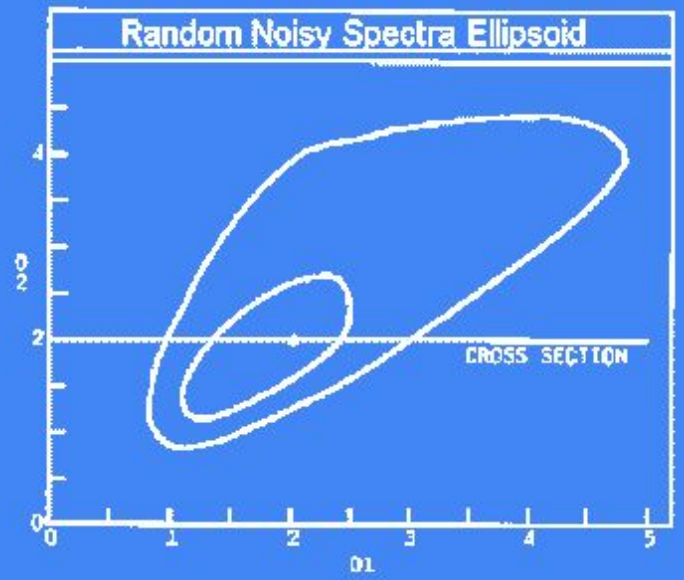


Central Limit Theorem Frequency Distribution



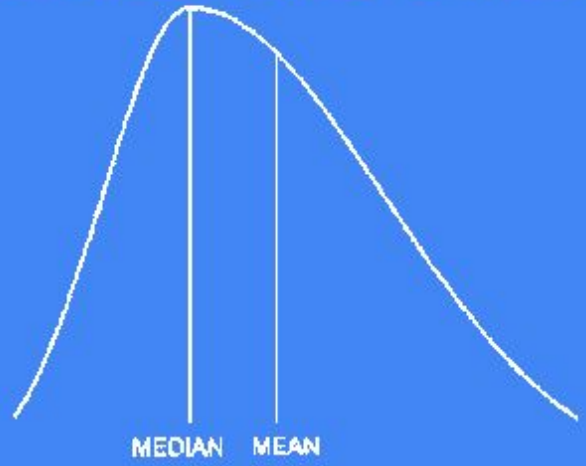
The central limit theorem explains why elliptical clusters result.

Skew appears in the multivariate distribution.



Skew in the distributions of the underlying constituents, such as in cholesterol in man, also is reflected in spectral data clusters.

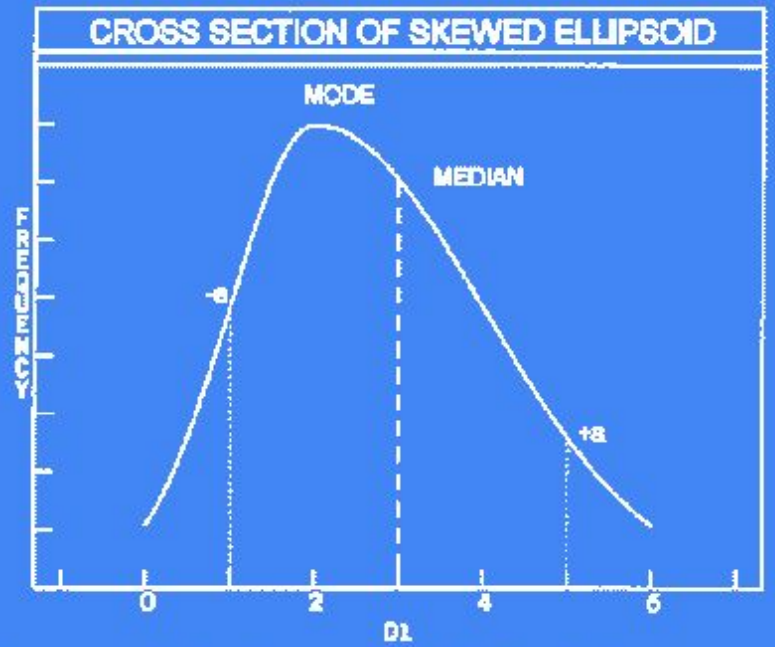
Balancing Property of Mean



The mean shifts in the direction of the skew, while the median does not.

Balancing Property of Mean

The movement of the mean also occurs in multiple dimensions.



Confidence limits also change when a distribution is skewed.

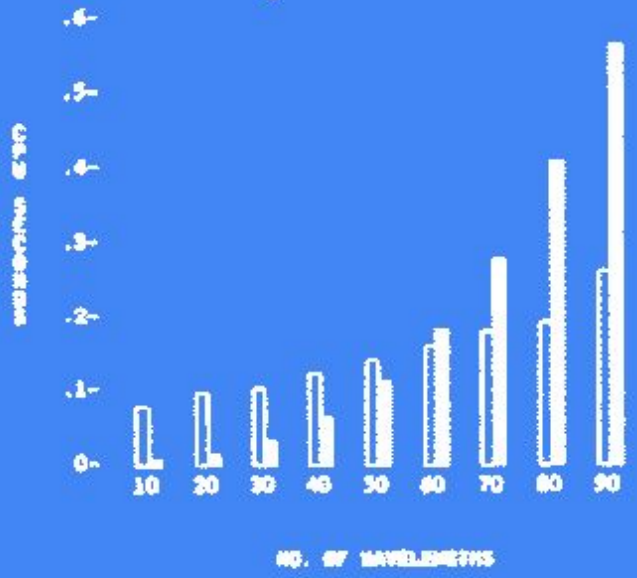
RADIALLY SYMMETRIC CLUSTERS

Have radially symmetric confidence limits

For the normal distribution one SD is a central 68% confidence interval

The Mahalanobis distance ($D = SQRT(T^{-1} * M2 * T^{-1})$, where M2 the inverse covariance matrix) is a multidimensional SD for normal distributions.

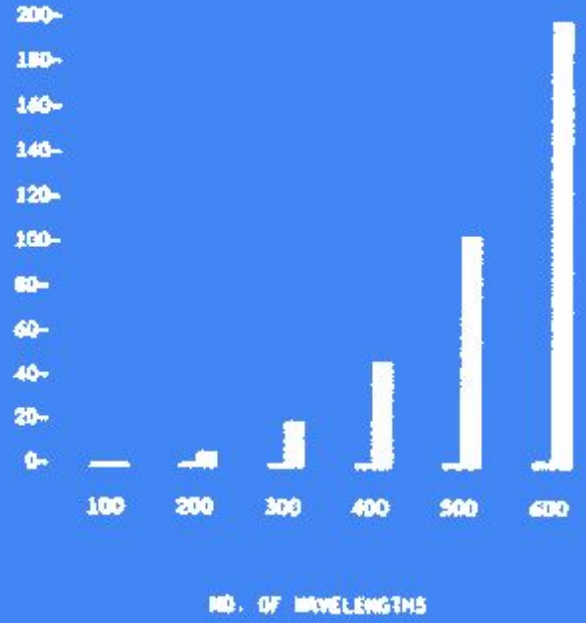
Efficiency of Mahalanobis and BEST Metric Calculations



1 million

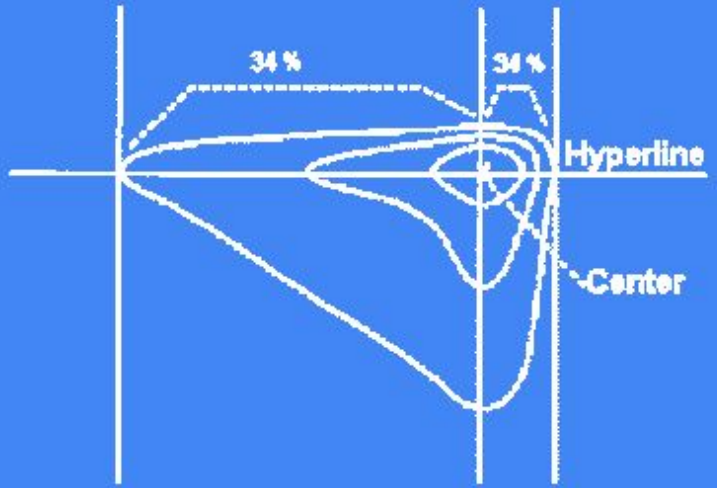
Mahalanobis = $O(n^3)$ 57,000 years

BEST = $O(n)$ 7.5 hours



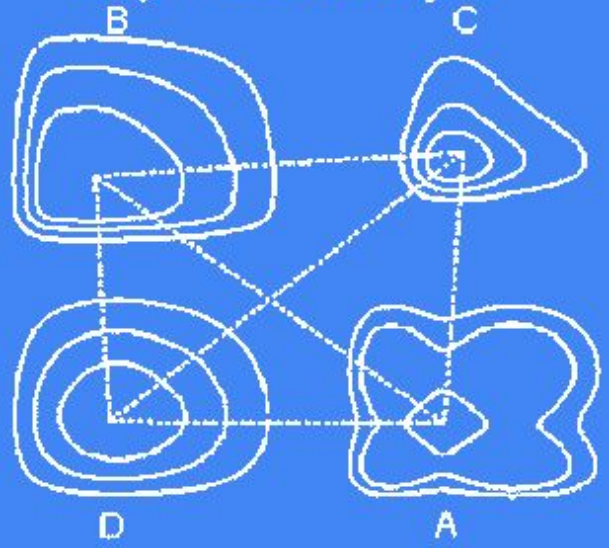
BEST can be used with full spectra without preprocessing by transformation to principal axes, fewer misclassified samples.

Skewed Cluster & Rubber Yardstick



The same definition of one SD can also be applied nonparametrically to skewed distributions.

Nonparametric Analysis



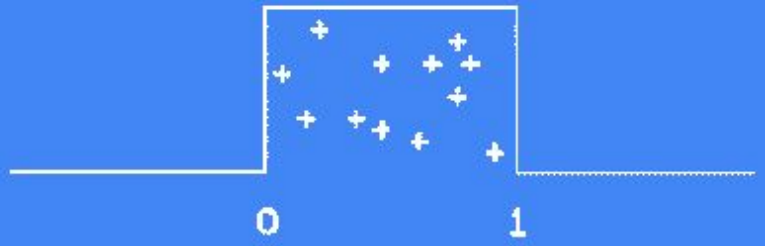
Distances in SDs depend not only on direction in space, but also on the cluster you select as your metric.

How do you calculate a multidimensional SD for a skewed distribution? One way is the BEST method (Bootstrap Error-adjusted Single-sample Technique). Like all statistics, you have to start with an observed data set (*calibration* set or *training* set).

Hypothetical Training Set Coverage

Suppose you start with an observed data set T, and you would like to use T to estimate the variability of the population from which it was drawn. The bootstrap lets you make the estimate.

Monte Carlo Integration

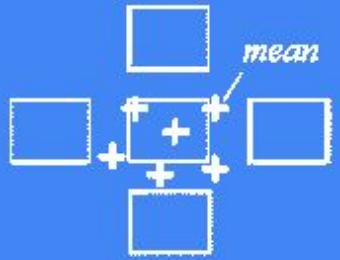


$P = p_i = r$

of the bootstrap distribution

THE BOOTSTRAP DISTRIBUTION

B becomes:



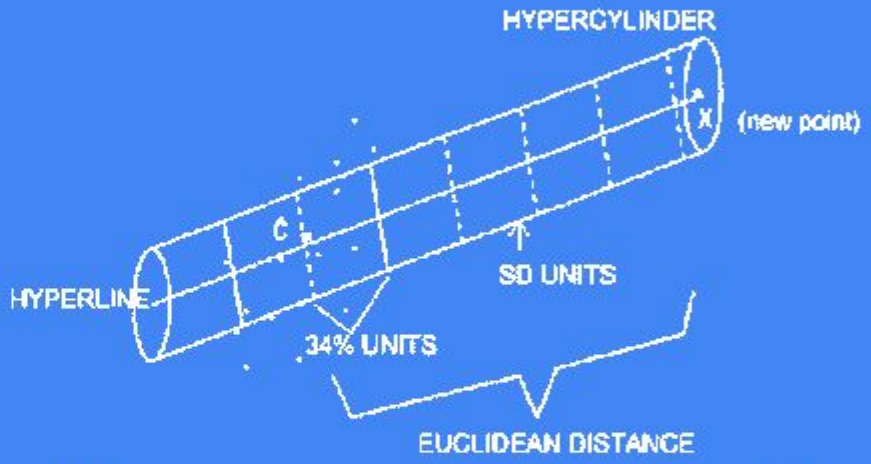
$P = (nP + 1)$

Monte Carlo of Bootstrap Distribution

B =  etc.

$$B_{(s)} - t_{kj}$$

Formation of Hypercylinder

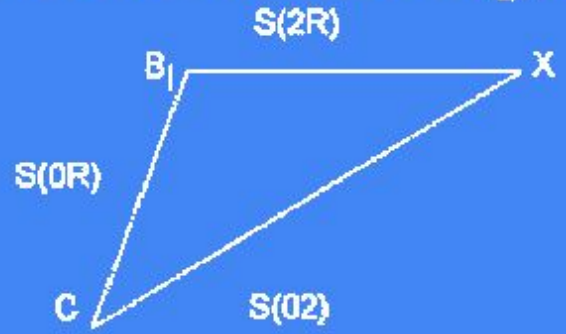


$$b_{oj} - \sum_{i=1}^{n} b_{(s)ij} / n$$

$$c_j - \sum_{i=j}^{m} b_{ij} / m$$

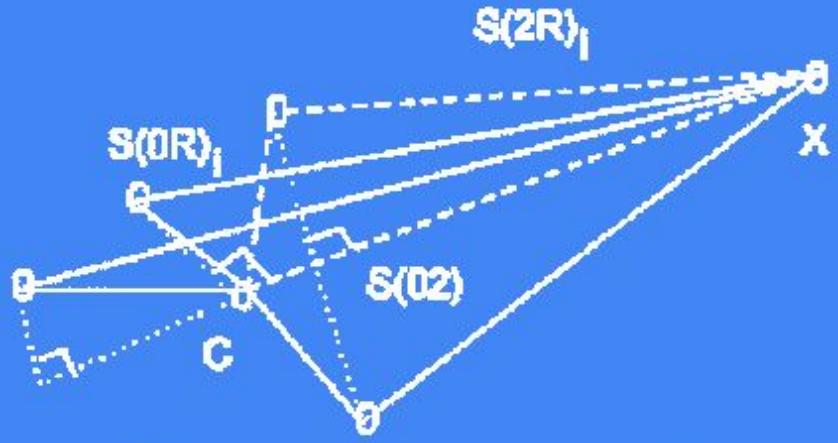
Strategy: project replicates onto hyperline and estimate SD

Area of Triangle



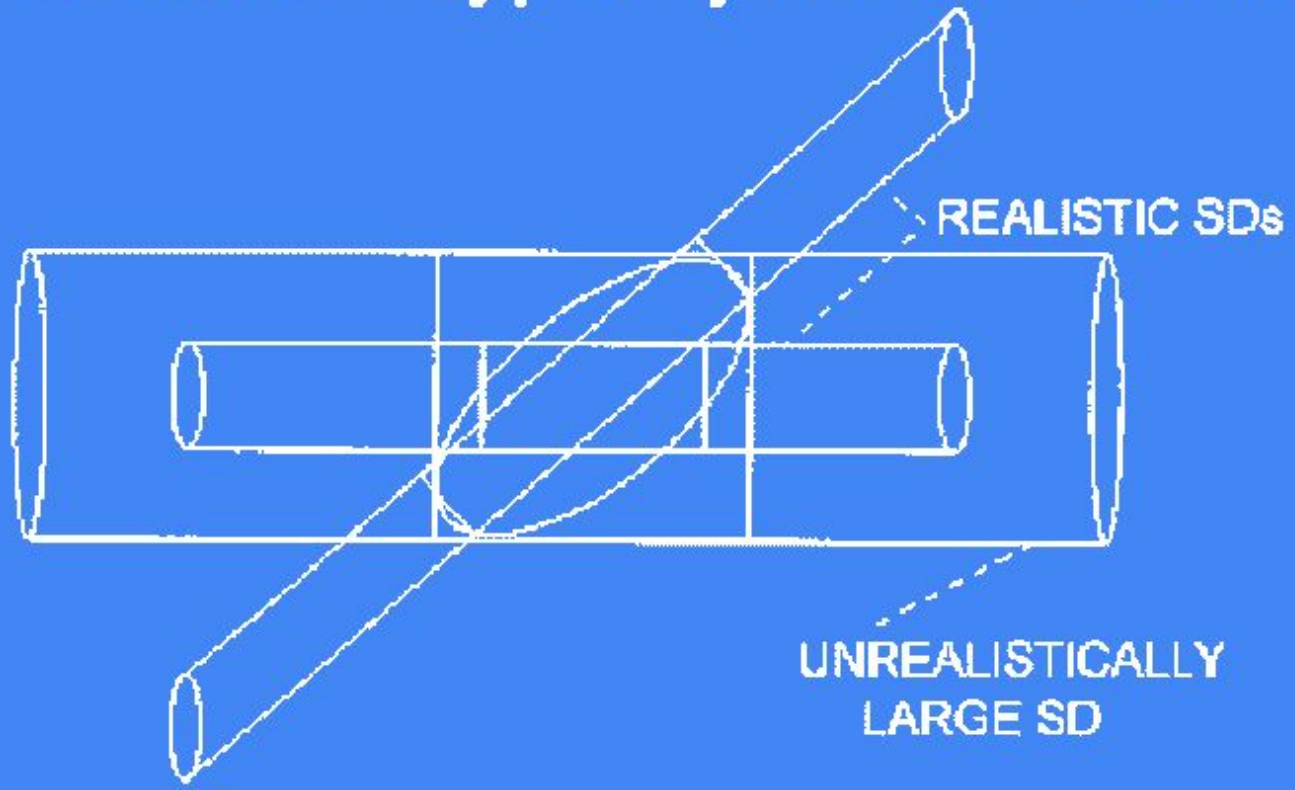
$$S_{(02)} = \left(\sum_{j=1}^{d} (x_j - c_j)^2 \right)^{1/2}$$

Hypercylinder Formation from Triangles



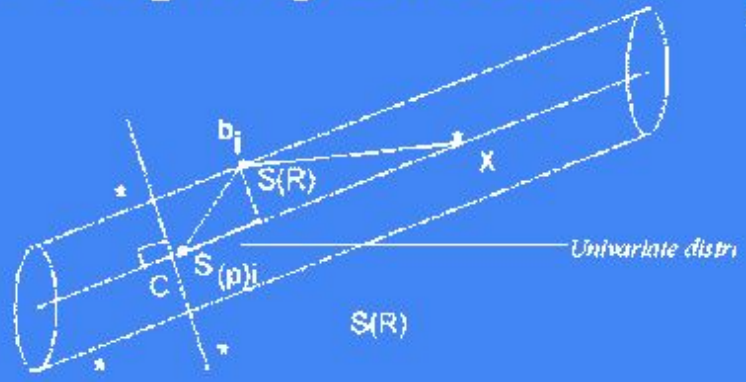
$$s_{(0R)l} = \left(\sum_{j=1}^{d} (b_{lj} - c_j)^2 \right)^{1/2} \qquad s_{(2R)l} = \left(\sum_{j=1}^{d} (b_{lj} - x_j)^2 \right)^{1/2}$$

Effect of Hypercylinder Radius



When projecting distances, hypercylinder radius important off the principal axes

Assigning Directions



$$s_{(p)i} = (s_{(0R)i}^{2} - s_{(R)i}^{2})^{1/2}$$

n_h = m(% of B spectra desired in hypercylinder) / 100

Radius is set as a fraction of the total replicates

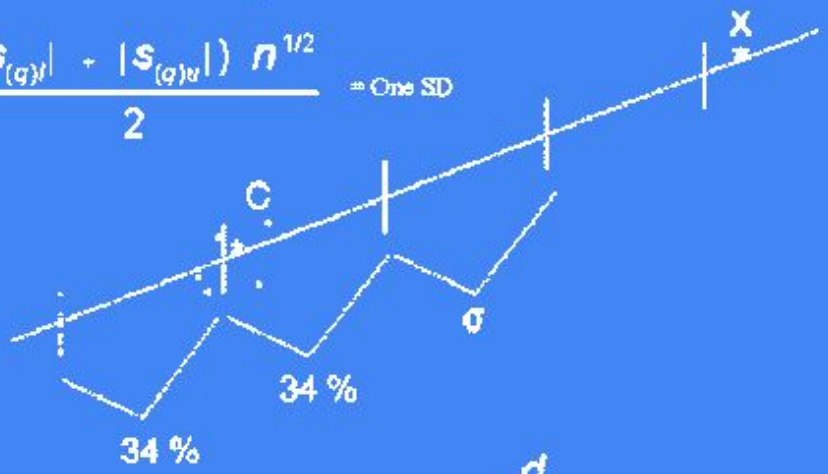
$$r_h = O(S_{(R)})_{n_h}$$

$$S_{(q)} = O(\{s_{(p)i} \mid s_{(R)i} < r_h\})$$

Only projected distances from replicates inside the hypercylinder contribute to the probability determination

Counting Distances in SDs

$$\frac{(|s_{(q)l}| + |s_{(q)u}|)\, n^{1/2}}{2} = \text{One SD}$$



$$\text{Distance in SDs} = \left(\sum_{j=1}^{d} (c_j - x_j)^2\right)^{1/2} / \sigma$$

However, this SD is symmetric like the Mahalanobis distance. The difference between the median and the mean can be used as the basis for a skew correction.

$$S_{(C0R)} = \left(\sum_{j=1}^{d} (c_{(T)j} - c_j)^2\right)^{1/2}$$

$$S_{(C2R)} = \left(\sum_{j=1}^{d} (c_{(T)j} - x_j)^2\right)^{1/2}$$

S_{02} was calculated previously.

$$S_{(CUB)} = (S_{(02)} + S_{(C0R)} + S_{(C2R)}) \ / \ 2$$

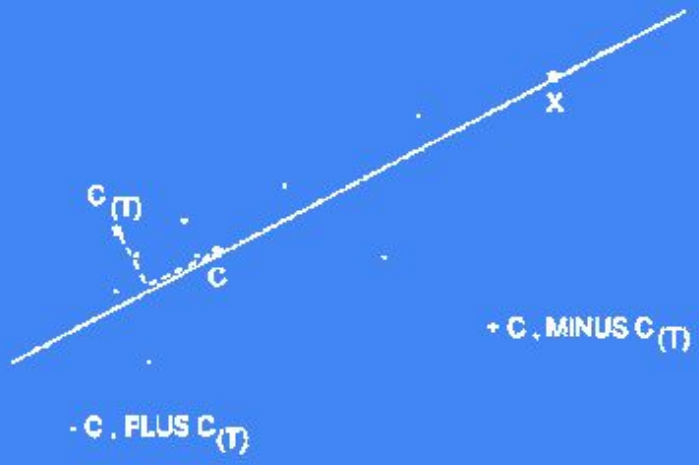
$$A_c = (S_{(CUB)}(S_{(CUB)} - S_{(02)})(S_{(CUB)} - S_{(C0R)})(S_{(CUB)} - S_{(C2R)}))^{1/2}$$

$$S_{(CR)} = 2(A_c) \ / \ S_{(02)}$$

$$S_{(CP)} = (S_{(C0R)}^{\ 2} - S_{(CR)}^{\ 2})^{1/2}$$

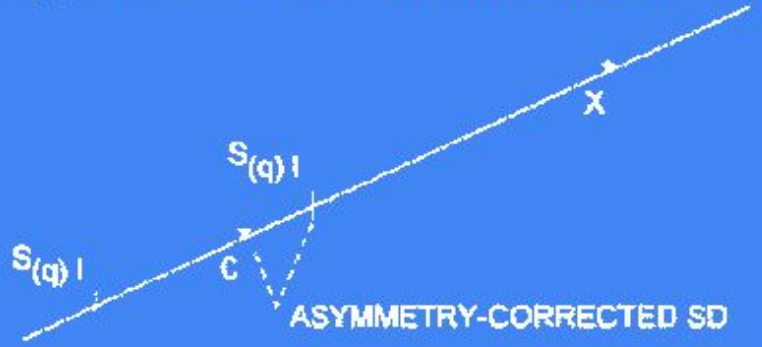
This gives the difference between the mean and median projected on the hyperline connecting C and X.

Direction of Correction



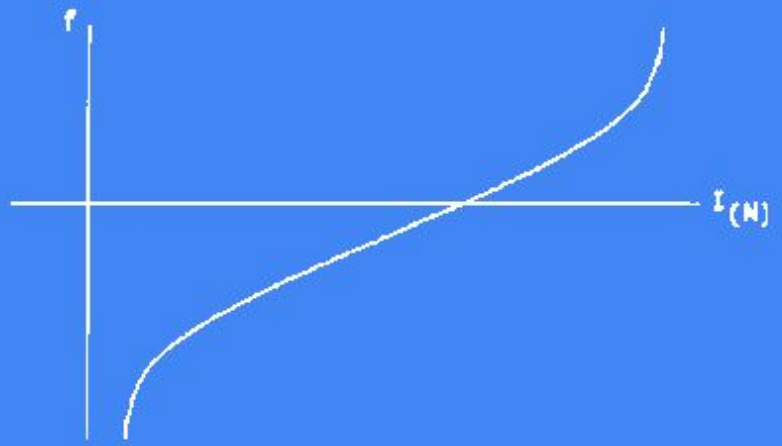
If $\{S_{(c2)}^2 + S_{(c2R)}^2 > S_{(c2R)}\}$
then S is multiplied by -1

Upper and Lower Confidence Limits



$$f_i \cdot s_{(q)i} \cdot S_{(CP)}$$

Skew-Corrected Distance Function



$$I_{(N)} = \{1, 2, 3, \ldots, n_h\}$$

$$z_e = (R(F(I_{(N)})))$$

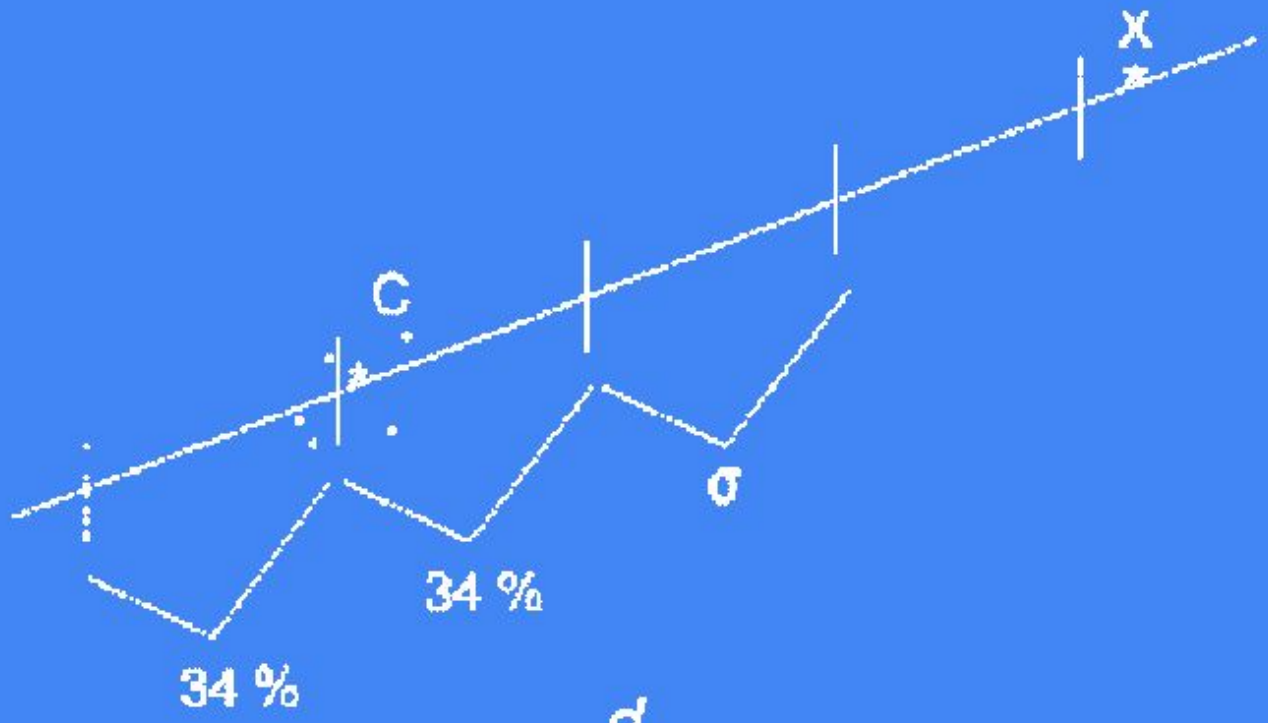
$$z_o = \Phi^{-1}(z_e / n_h)$$

$$l = (\Phi(2z_o + z_\alpha) n_h)$$

$$u = (\Phi(2z_o - z_\alpha) n_h)$$

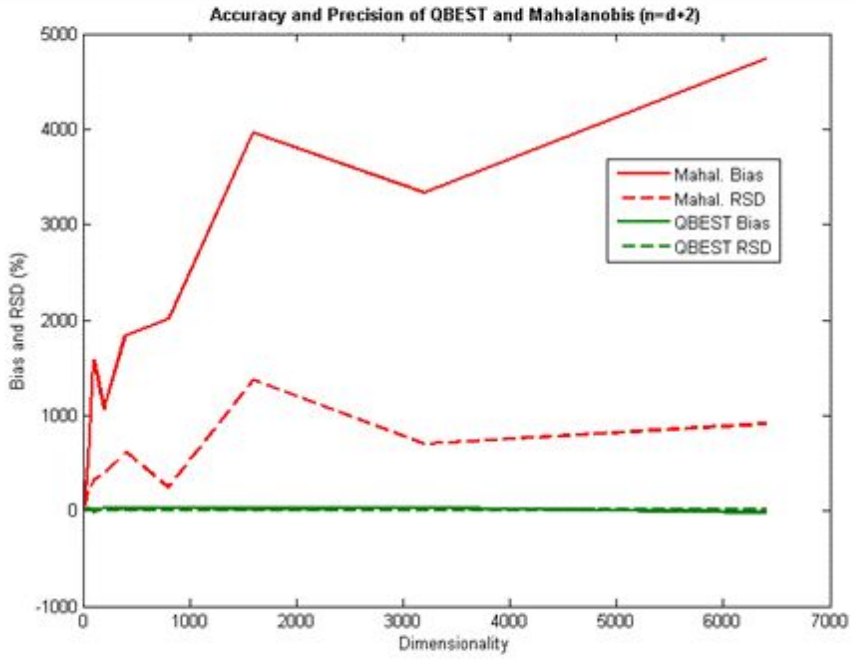
Efron's lower (l) and upper (u) skew-corrected confidence limits based on normal quantiles

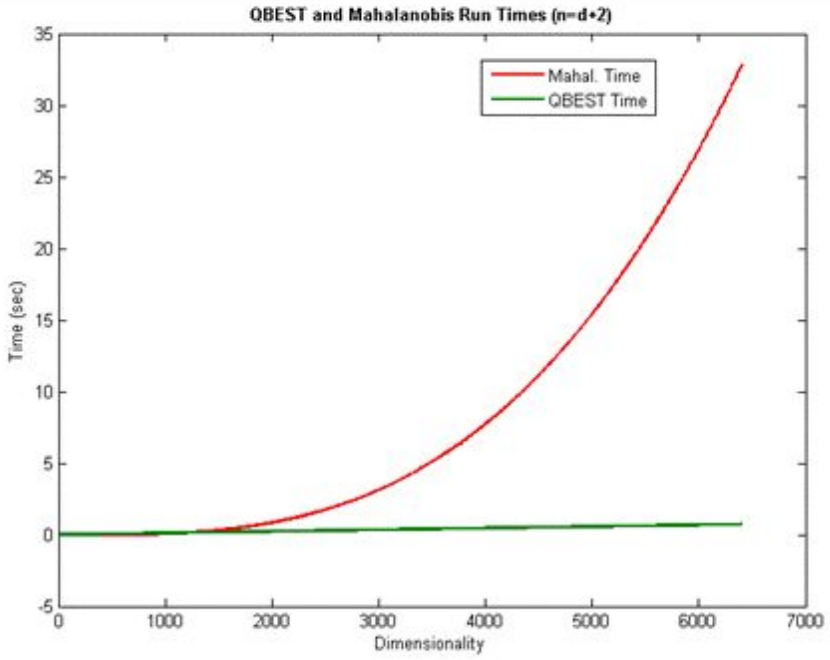
Counting Distances in SDs



$$\text{Distance in SDs} = \left(\sum_{j=1}^{d} (c_j - x_j)^2\right)^{1/2} \Big/ \left((\sigma_c / |z_\alpha|) n^{1/2}\right)$$

BEST Third Wave Outperforms Second Wave Statistics



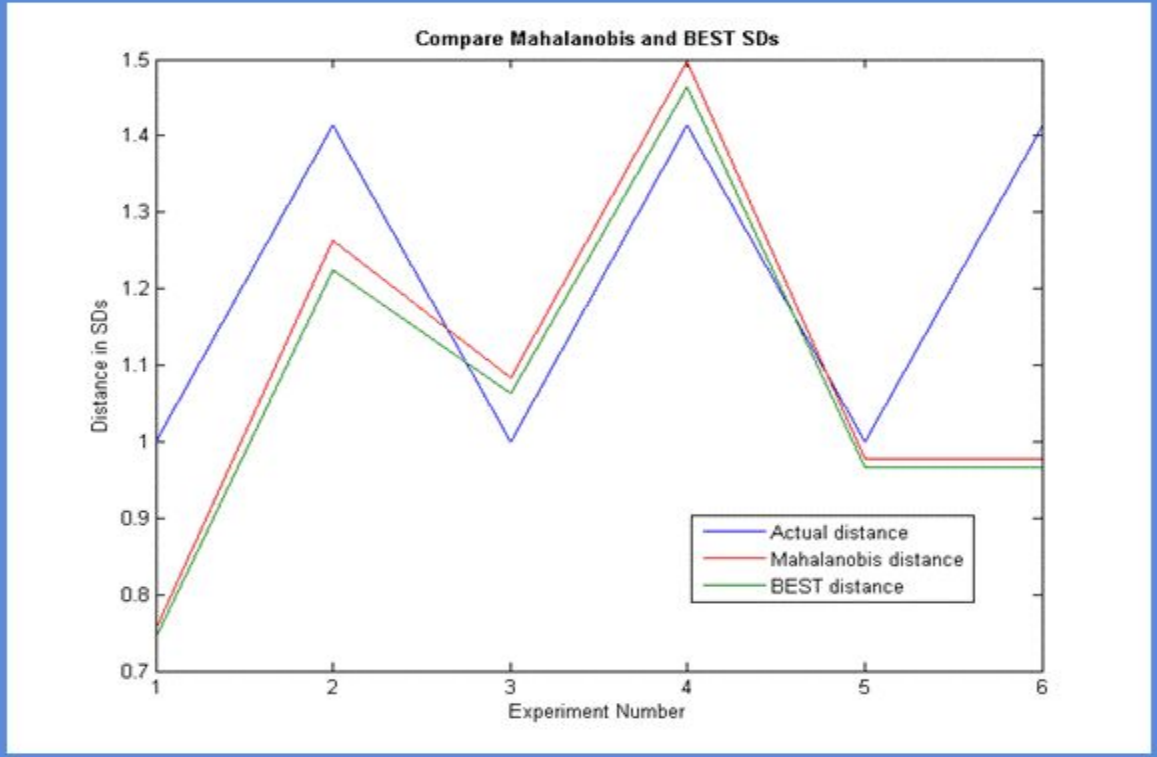


Accuracy and precision of the BEST and Mahalanobis metrics using an N(0,1) synthetic data set

The running time of the BEST calculation is n^2 faster than the Mahalanobis even without parallelization

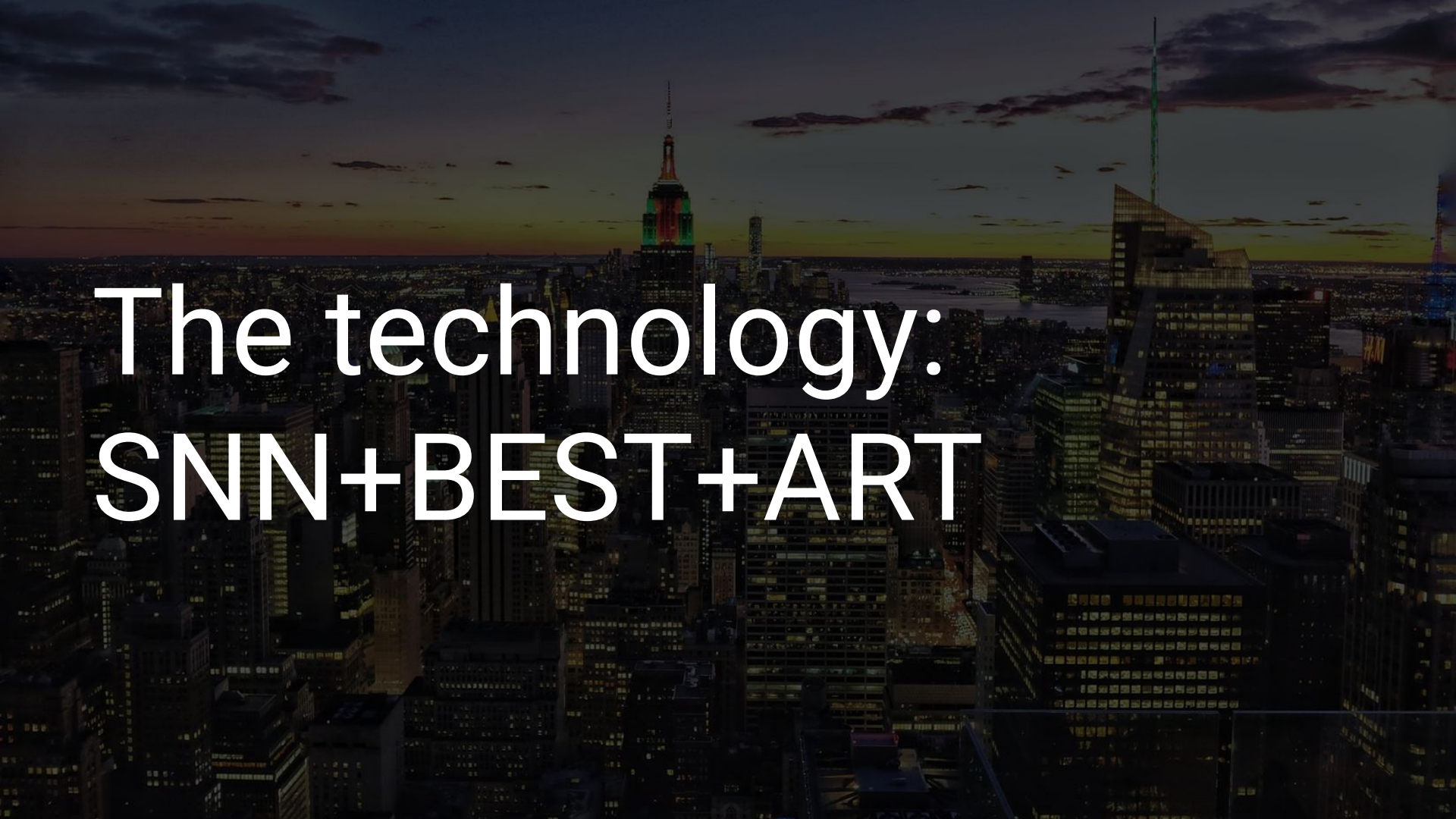
BEST Third Wave Matches Performance of Second Wave for Parametric Problems

n=d*50, a best case scenario for Mahalanobis



n=d*50 is not a realistic scenario for most pharma development

With only 2 independent variables and 100 samples from a synthetic N(0,1) distribution, the accuracy of the BEST metric is about the same as the Mahalanobis metric. The error comes from the 100 randomly selected samples poorly representing the N(0,1) distribution.


The technology:
SNN+BEST+ART

AI System

AI must look at more than science to solve the problem.

Market

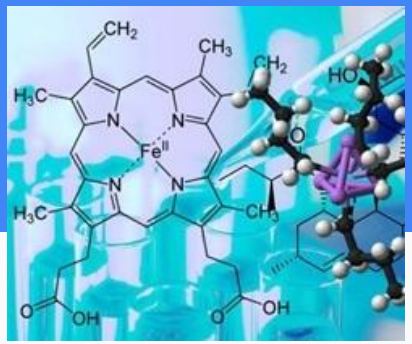
Disease prevalence, reimbursement, competition, pricing, capital requirements and capital availability

Intellectual Property

Patentability, contracts, portfolios, encumbrances, white space, landscape, valuation, quality, monetization, licensing

Regulatory

Pediatric rare disease, tropical disease, fast track, breakthrough status, jurisdictions, guidances, quality management, outcomes



Science

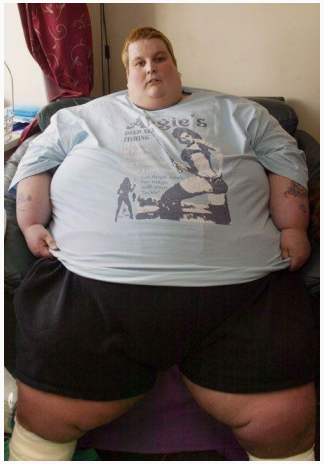
Pathways, signaling, receptors, genes, RNAi, cells, immunotherapy, populations, druggability, manufacturability

Drugs in Development

We now have animal and human data for four new molecular entities (NMEs) to treat: Chikungunya and Ebola virus infections, and Prader Willi Syndrome as well as Attention Deficit Hyperactivity Disorder (ADHD) in patients with Fragile X Syndrome (FXS).

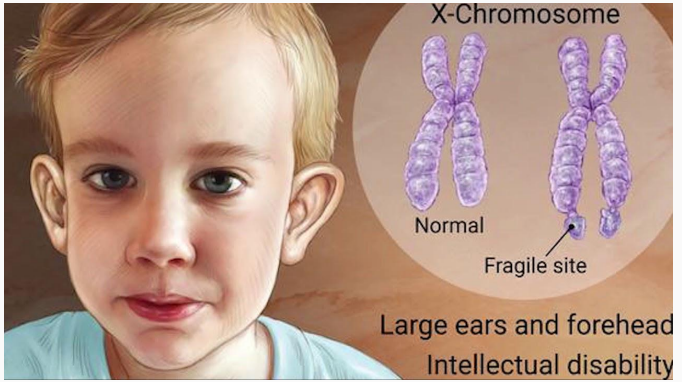
We have FDA Orphan Drug Designations for Ebola virus and Prader Willi Syndrome

Chikungunya



Prader Willi

Ebola

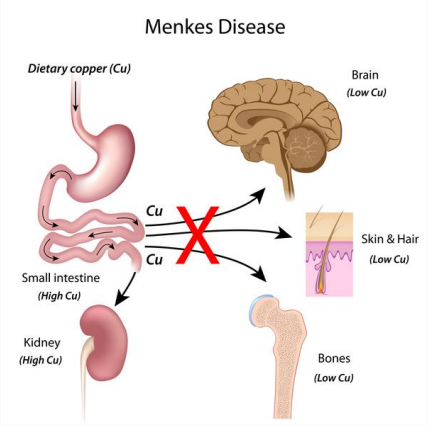


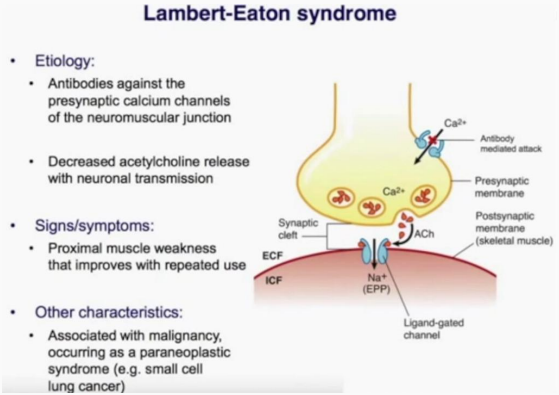
ADHD in FXS

AI Approach Leads to New Therapies

AND a 50% increase in INDs!

1. *Methylpropanedihydrazide Compounds for Menkes Syndrome*, US patent application number 62732350, filed Sep 17 2018
2. *Nanotube Delivery Device for MPDH Compounds for Menkes Disease*, US patent application number 62732379, filed Sep 17 2018
3. *Biodegradable Nanoparticles for Controlled Release of ICP4 Compounds for LEMS*, US patent application number 62722146, filed Aug 23 2018
4. *ICP4 Compounds for Treatment of Lambert Lambert-Eaton Myasthenic Syndrome*. US patent application number 62690557, filed Jun 27, 2018
5. *Preparation of ICP4 Compounds.* US patent application number 62690606, filed Jun 27, 2018





Lodder Lab

Students and Postdocs

Alyson Ackerman
Andy Du
Anne Brooks
Braden Adams
Christopher Maynard
Cynthia Dickerson
Mark Ensor
Markus Ville Tiitto
Mayte Hernandez-Murillo
Sam Hacker
Stephen Yin

Other Universities

Faculty

Prof. Craig Douglas (U. Wyoming)

Credit

"*Artificial neural networks for modeling and simulation*", Google project 815523038794

"*Computer Simulations of Glucose-Insulin Interaction*", National Science Foundation ACI-1053575 number BIO170011

"*Applications of Parallel Computing Course*", National Science Foundation CCR140008