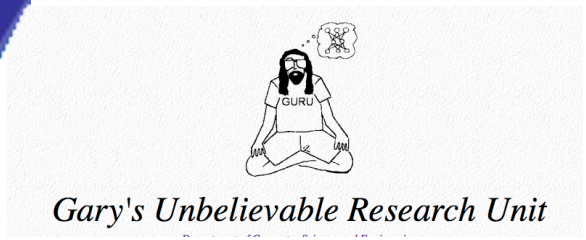
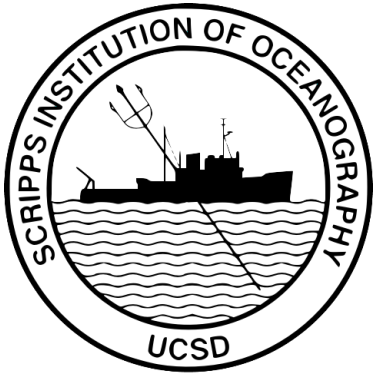




Using Deep Siamese Neural Networks to Speed Up Natural Products Research

Garrison W. Cottrell

CSE Department

Gary's Unbelievable Research Unit (GURU)

UC San Diego

Collaborators

Yerlan Idelbayev

UC Merced

Chen Zhang

Scripps Instution of Oceanography

Bill Gerwick

Scripps Instution of Oceanography

Yash_Nannapaneni

UCSD

Nick Roberts, UCSD

Poornav Sargur Purushothama

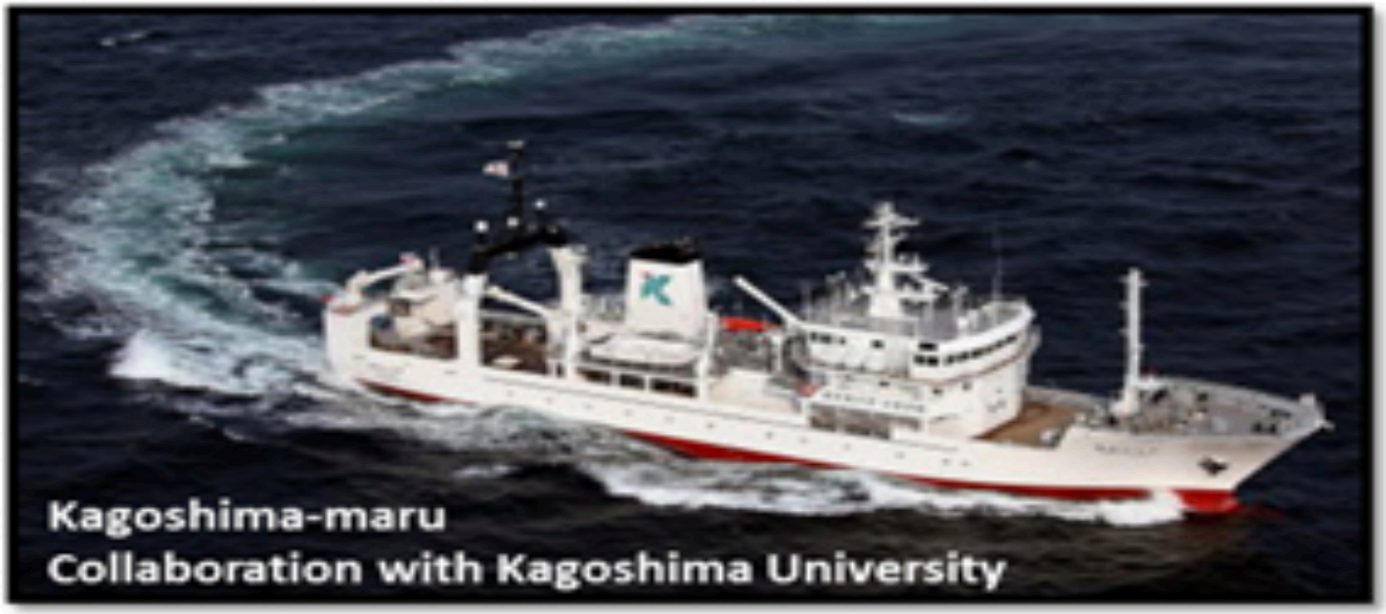
UCSD

Outline

- Natural products research
  - What is it
  - Why is it important
- Deep Siamese Convolutional Networks
  - What are they
  - Why are they cool
- SMART
  - Methods
  - Results

Natural Products Research

- 50% the approved drugs during the last 30 years are derived from natural products (Newman and Cragg 2012)



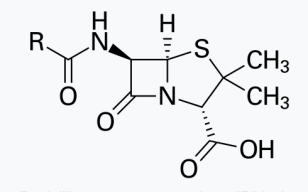
Natural Products Research

Bill Gerwick
Scripps Instution
of Oceanography

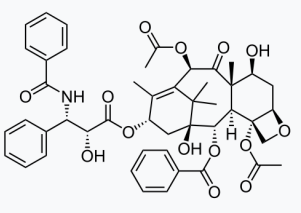
Here's Bill himself collecting samples in the Palmyra Atoll

Natural Products Research

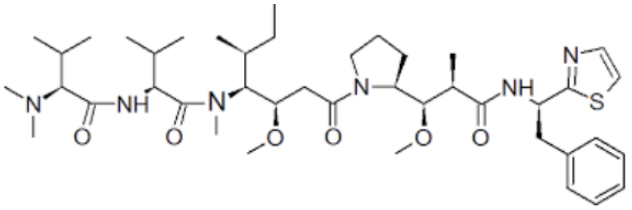
- 50% the approved drugs during the last 30 years are derived from natural products (Newman and Cragg 2012)
- Penicillin (1928)



- Taxol (1971) an anti-cancer agent derived from Yew tree bark



- Dolastatin 10 (2001) another anti-cancer agent derived from from a sea hare that feeds on tropical cyanobacteria in the Indian Ocean, and later traced back to the cyanobacterium itself.



Natural Products Research is ongoing

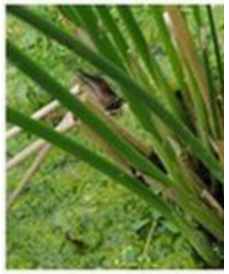
-and a lot of it is about structure discovery

Article

Five macrocyclic glycosides from *Schoenoplectus tabernaemontani* ›

Dian Peng, Xiaolin Li

Published online: 30



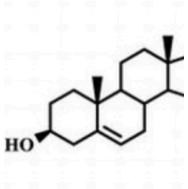
Schoenoplectus taber

Article

Synthesis, characterization, and biological evaluations of some steryl 2-methoxybenzoates as anticancer agents ›

Yanmin Huang, Haiy
Xiaolan Liu, Junyan Z

Published online: 30



Article

A new flavonoid from the leaves of *Atalantia monophylla* (L.) DC ›

Priyapan Posri, Jittra Suthiwong
Chindawadee Chuenban, Chant

Published online: 30 Mar 2018

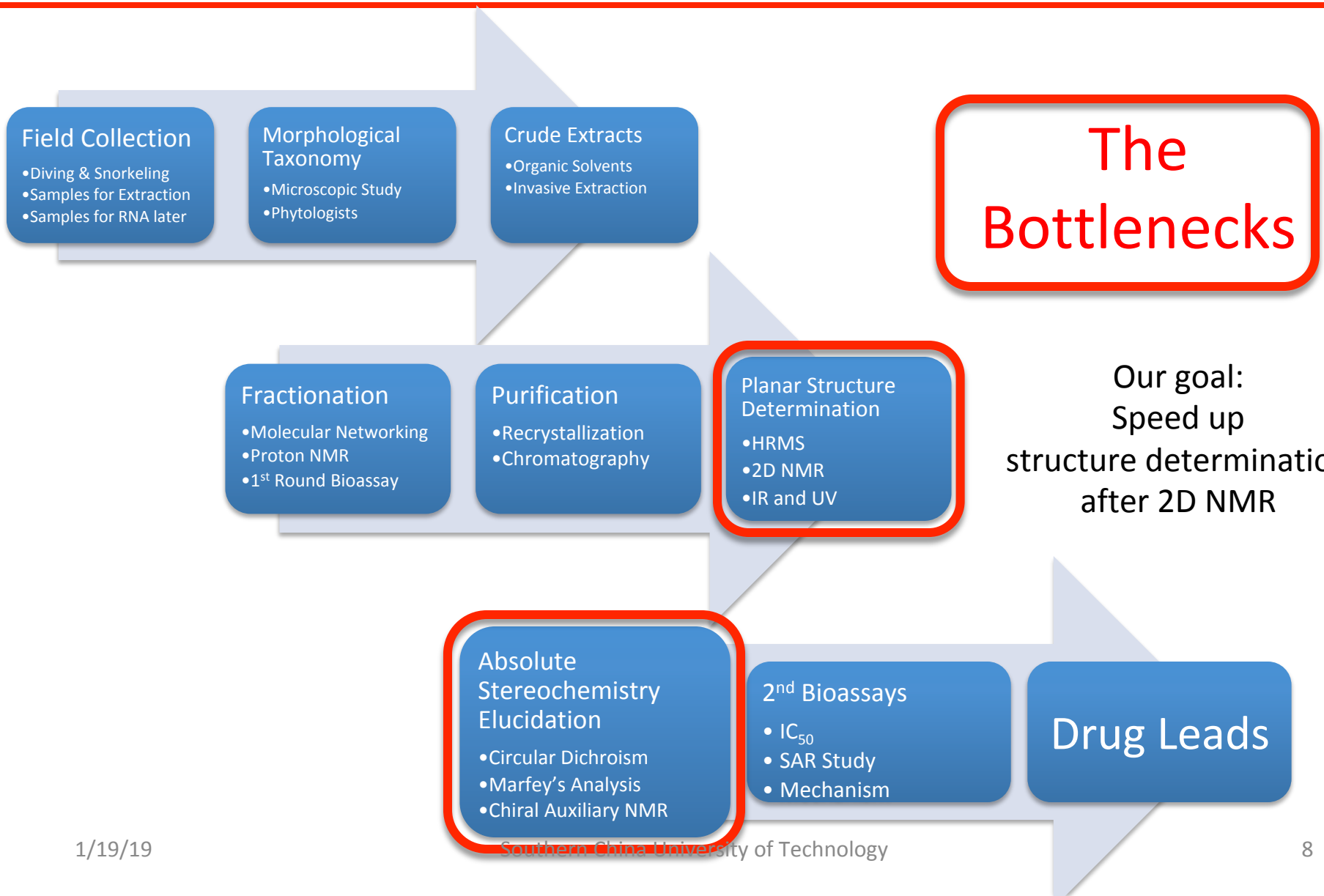
Article

Synthesis and antiproliferative activities of thioxoflavonoids on three human cancer cells ›

Wei Li, Peipei Han, Shuanglian Cai & Qiuan Wang

Published online: 27 Mar 2018

Current Natural Products Discovery is a Slow Process



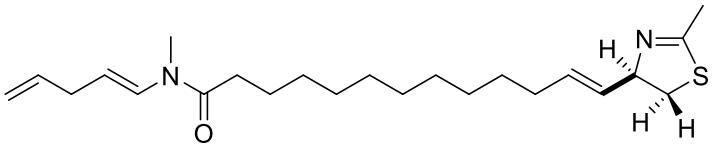
Natural Products Research

- Figuring out the *structure* of any new compound starts with the 2D NMR spectrum of the compound – the "fingerprint" of the compound
- Using a deep, convolutional siamese neural network, we have developed a system called SMART that *clusters* compounds directly from their NMR spectrum.
- This allows new compounds to be mapped into the cluster space, giving an idea of what compounds it is similar to.
- This can speed up structure discovery

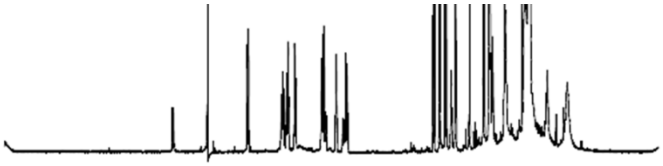
2D Nuclear Magnetic Resonance Imaging (2D NMR)

- 2D NMR is a technique that gives indications of bonds between atoms in a molecule
- We think of it as the "fingerprint" of the molecule
- And now I've told you everything I know about NMR!!

2D NMR is an Indicator of Compound Structure

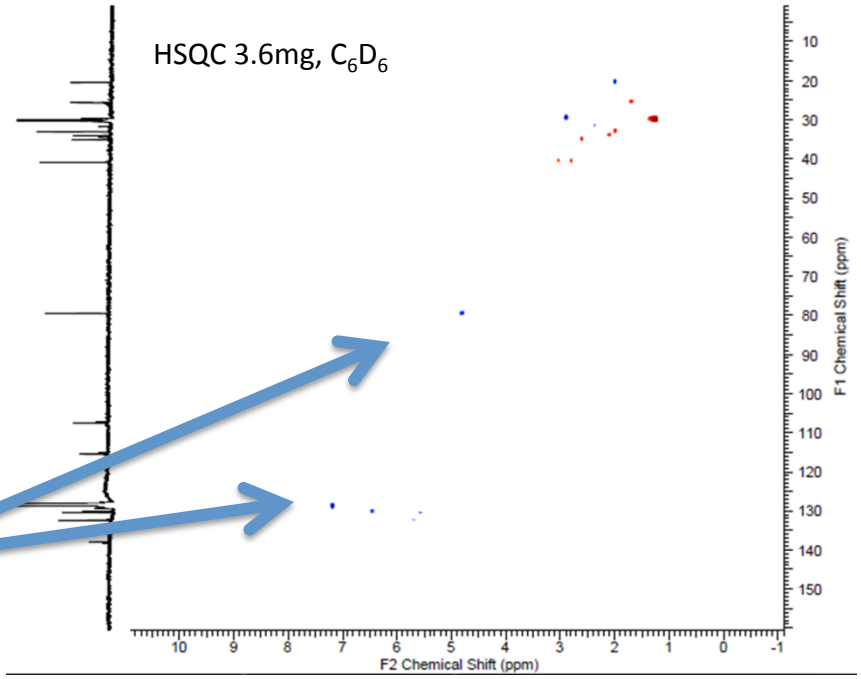


laucysteinamide A



HSQC 3.6mg, C_6D_6

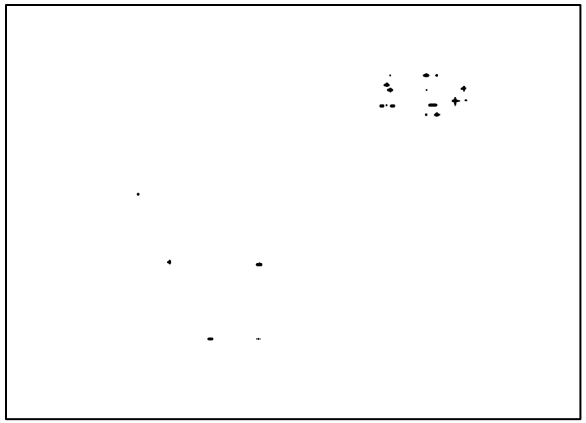
Caldora penicillata



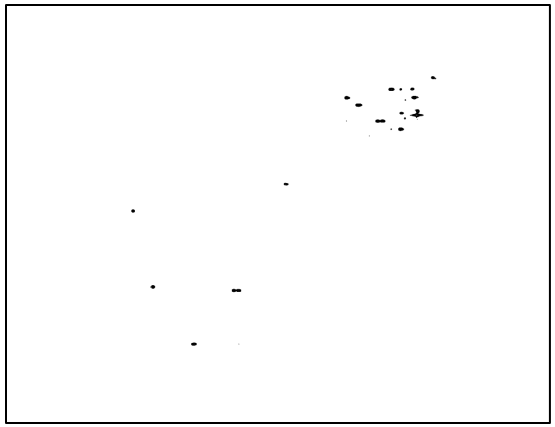
Each "dot" here corresponds to a bond

NMR structure analysis is subtle

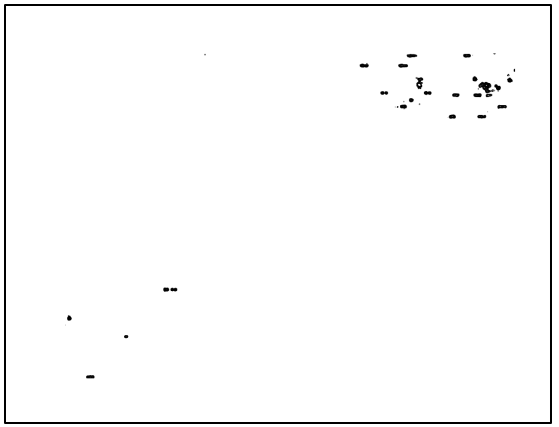
Aspewentin A



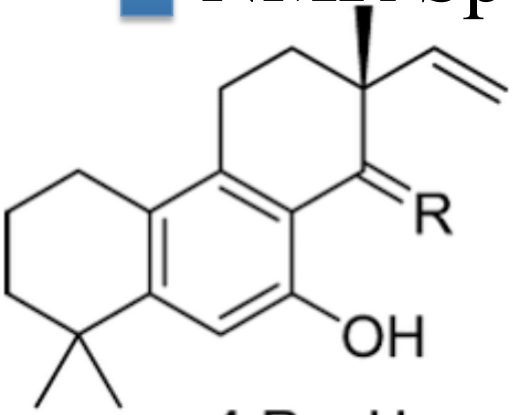
Aspewentin B



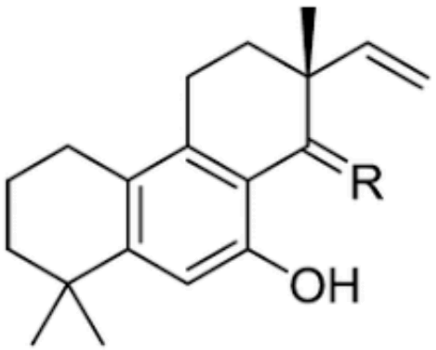
Aspewentin C



NMR Spectrum

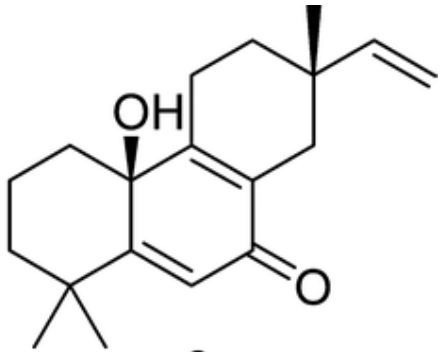


1 R = H_2



2 R= O

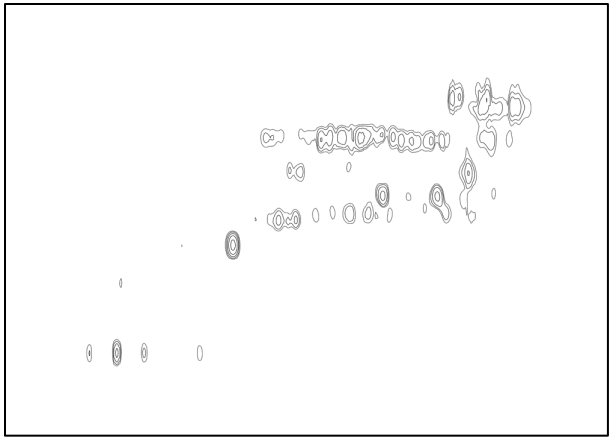
Actual Structure



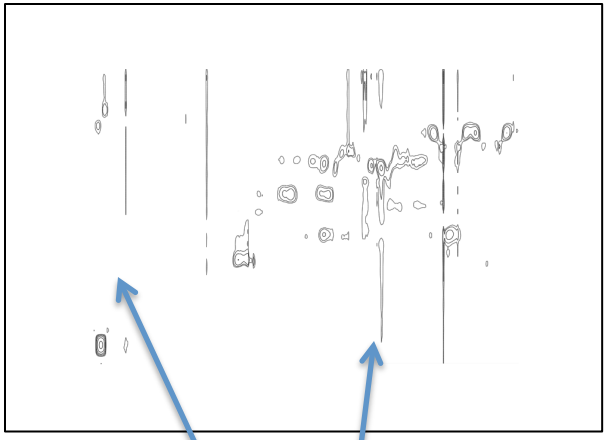
3

And Noisy!

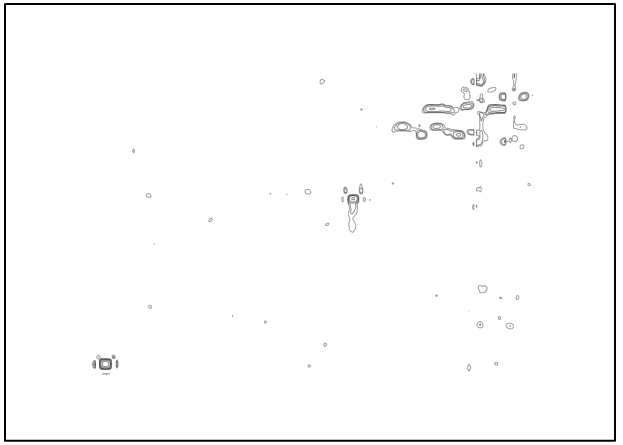
Chandonanone A



Chandonanone B



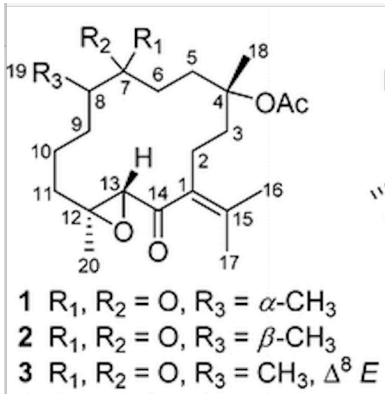
Chandonanone C



NMR Spectrum

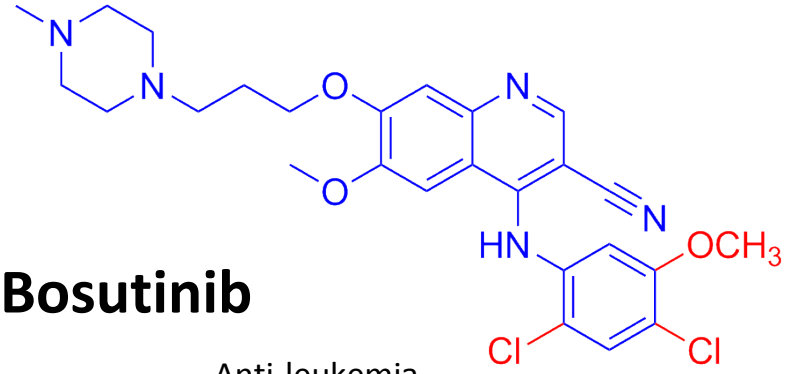
Actual Structure

These are artifacts of the solvent used...



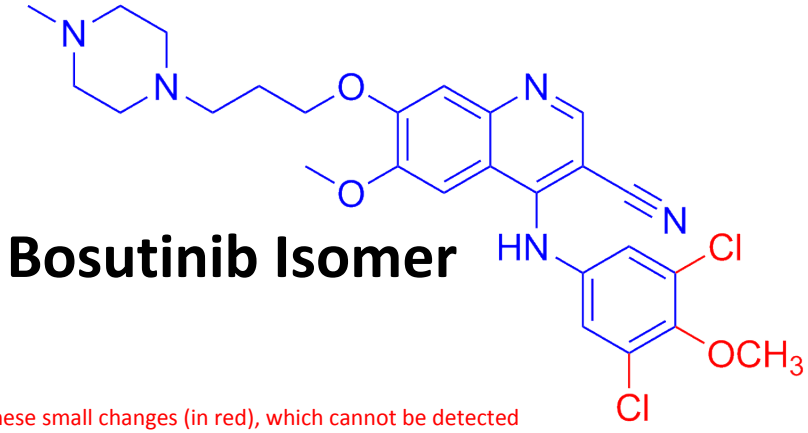
1 $R_1, R_2 = O, R_3 = \alpha\text{-}CH_3$
2 $R_1, R_2 = O, R_3 = \beta\text{-}CH_3$
3 $R_1, R_2 = O, R_3 = CH_3, \Delta^8 E$

And Mistakes can be Costly...



Bosutinib

Anti-leukemia



Bosutinib Isomer

These small changes (in red), which cannot be detected by techniques such as mass spectrometry or elemental analysis, affect the biological activity of the compounds

- Same Mass
- Same Chemical Composition
- Similar Chemical Structures
- NMR Misinterpreted by Human
- Very Different Biological Function!

"I still wonder what we could have done better. Clearly, not every lab can run and analyze an NMR spectrum of every chemical compound ordered."

— Dr. Ulrich Schweizer, Researcher at Charité-Universitätsmedizin Berlin, Germany

"We had wasted a huge amount of time and money on the wrong isomer."

— Dr. Steven G. Boxer, Distinguished Professor of Chemistry at Stanford University

Levinson, et al. *PLOS ONE* **2012,** 7, (4), e29828

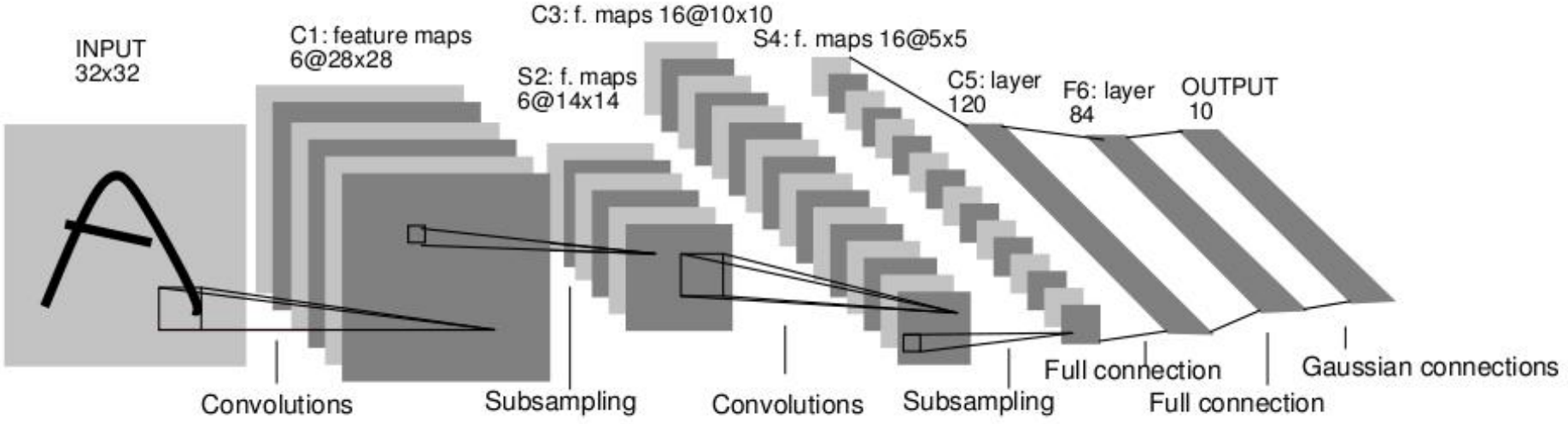
Enter Deep Learning

- Deep learning:
  - Any neural network with more than one hidden layer – usually many more!
- Convolutional Networks:
  - Have revolutionized computer vision since 2012
  - Learn many features
  - But apply the *same* feature across the image

Convolutional Neural Networks

Began with LeCun et al. 1989

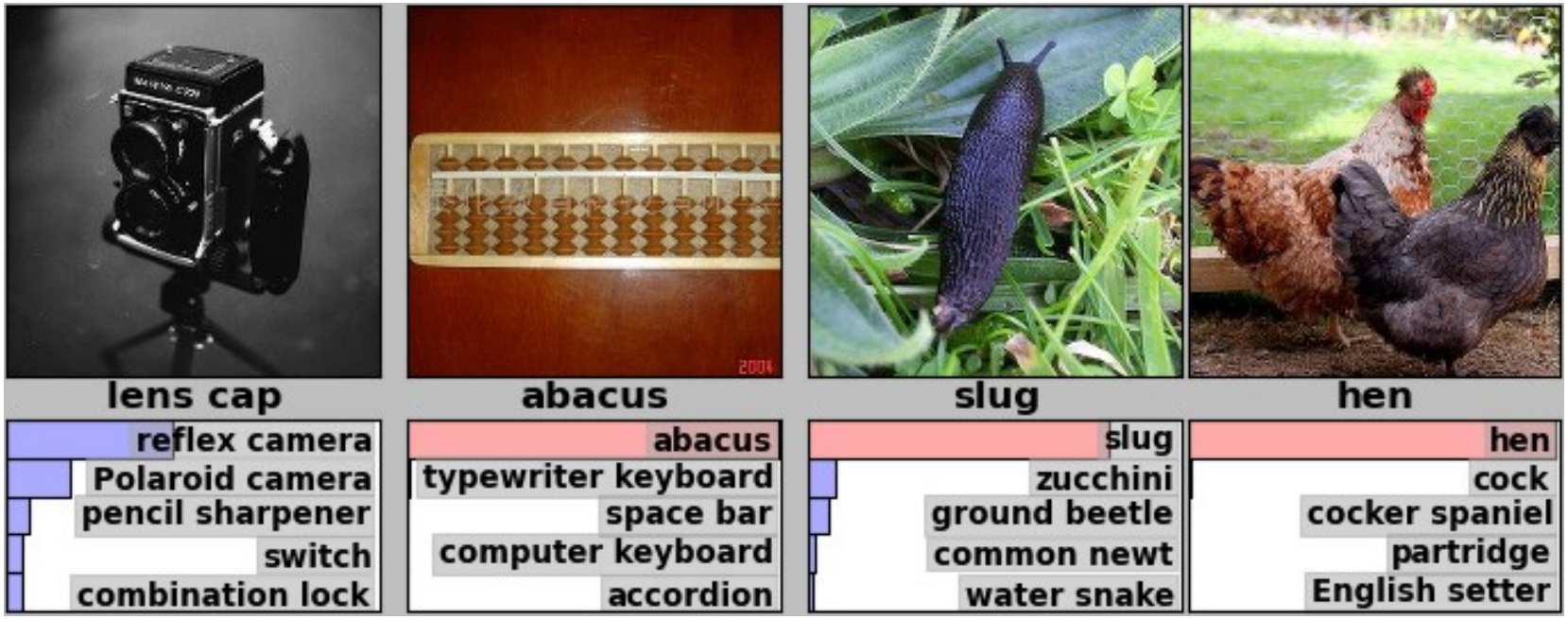
1. Small, local receptive fields and learned features (kernels): *locality*
2. These are *replicated* across the image: *stationary statistics*
3. Spatial pooling: *translation invariance*



Convolutional Networks -- What they're good at: The Imagenet Large Scale Visual Recognition Challenge (ILSVRC)

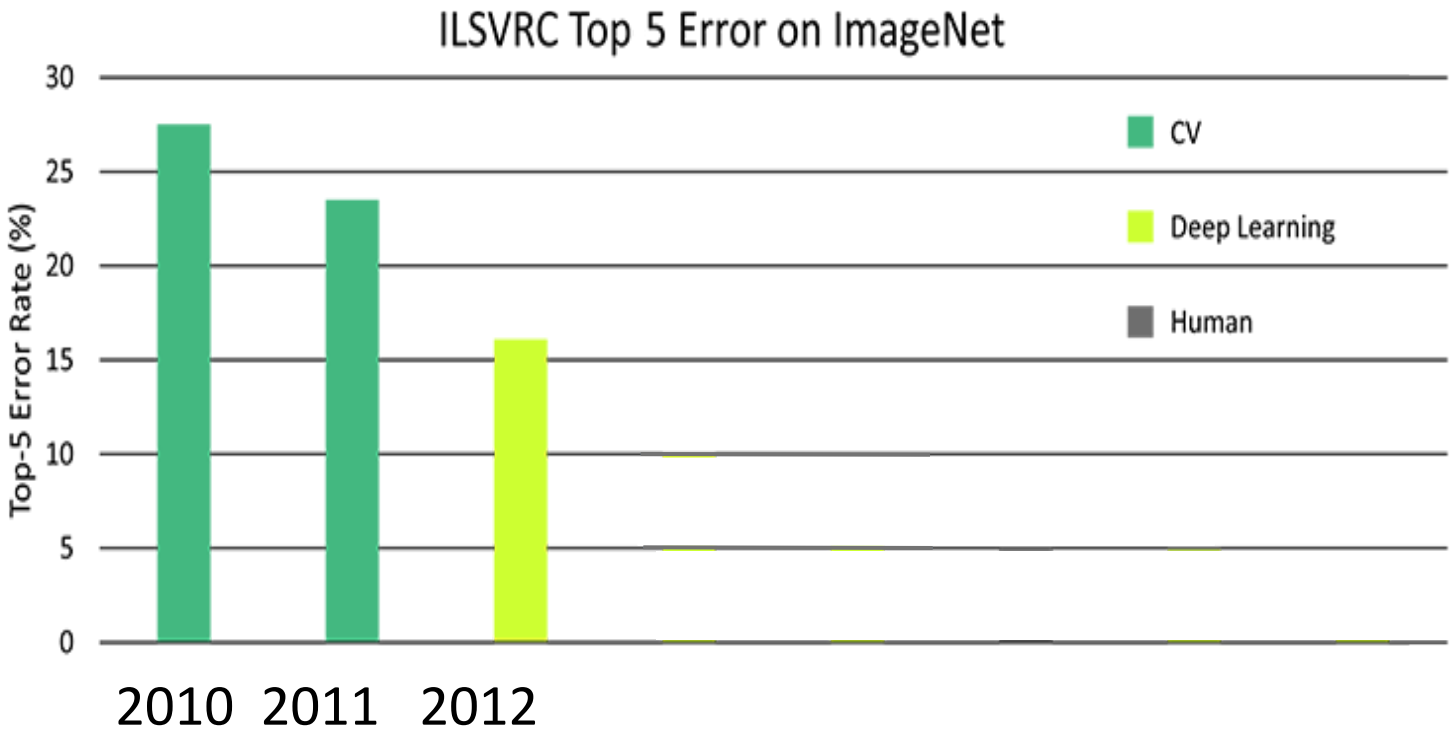
- 1.2 Million training images
- 1000 categories (732-1300 training images per class)
- 50,000 test images
- Large variation in images
- Many fine-scale categories (120 dog breeds, not just "dog")

Imagenet challenge goal: Classification



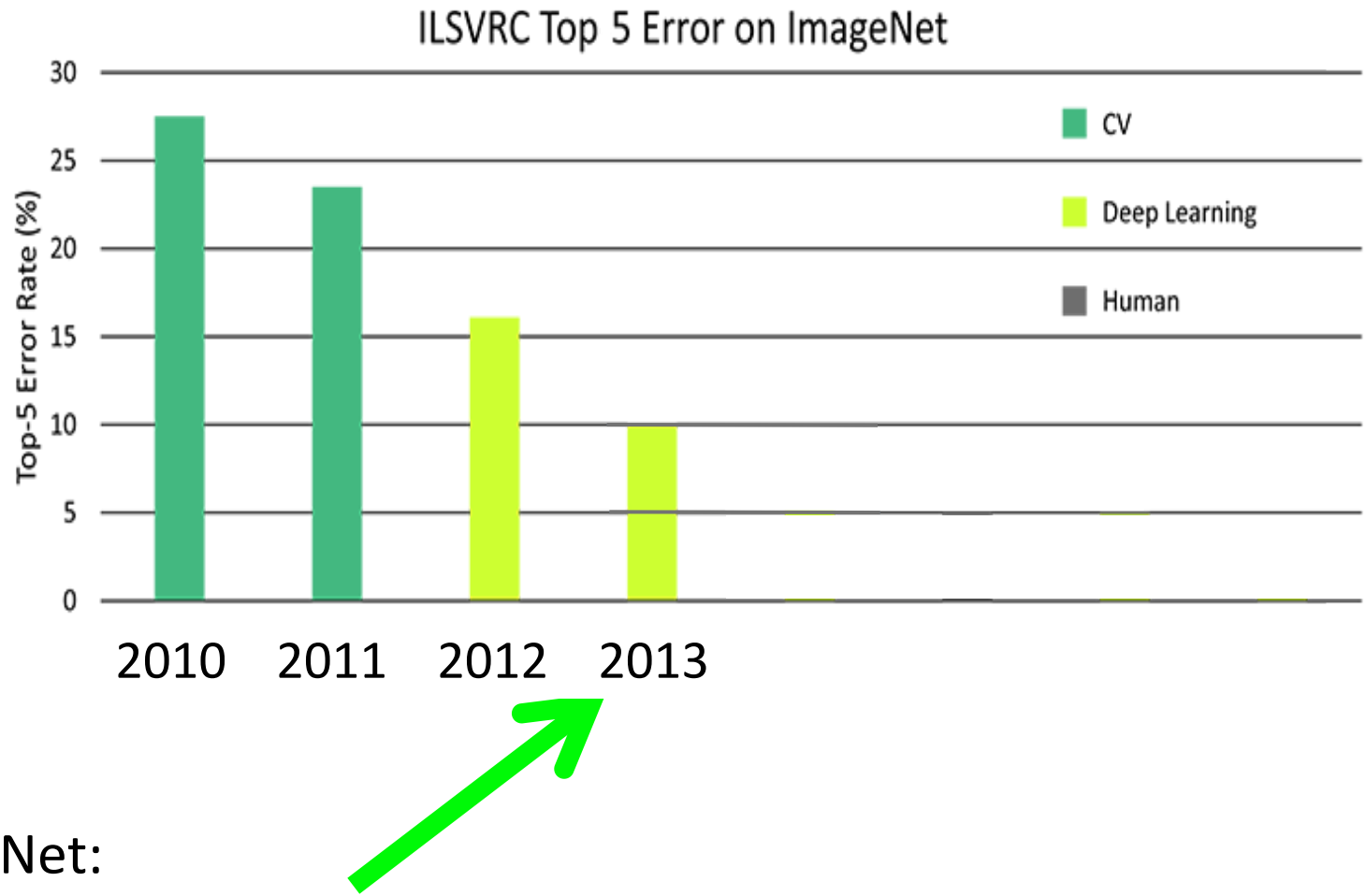
[Krizhevsky et al. NIPS 2012]

Why Deep Learning



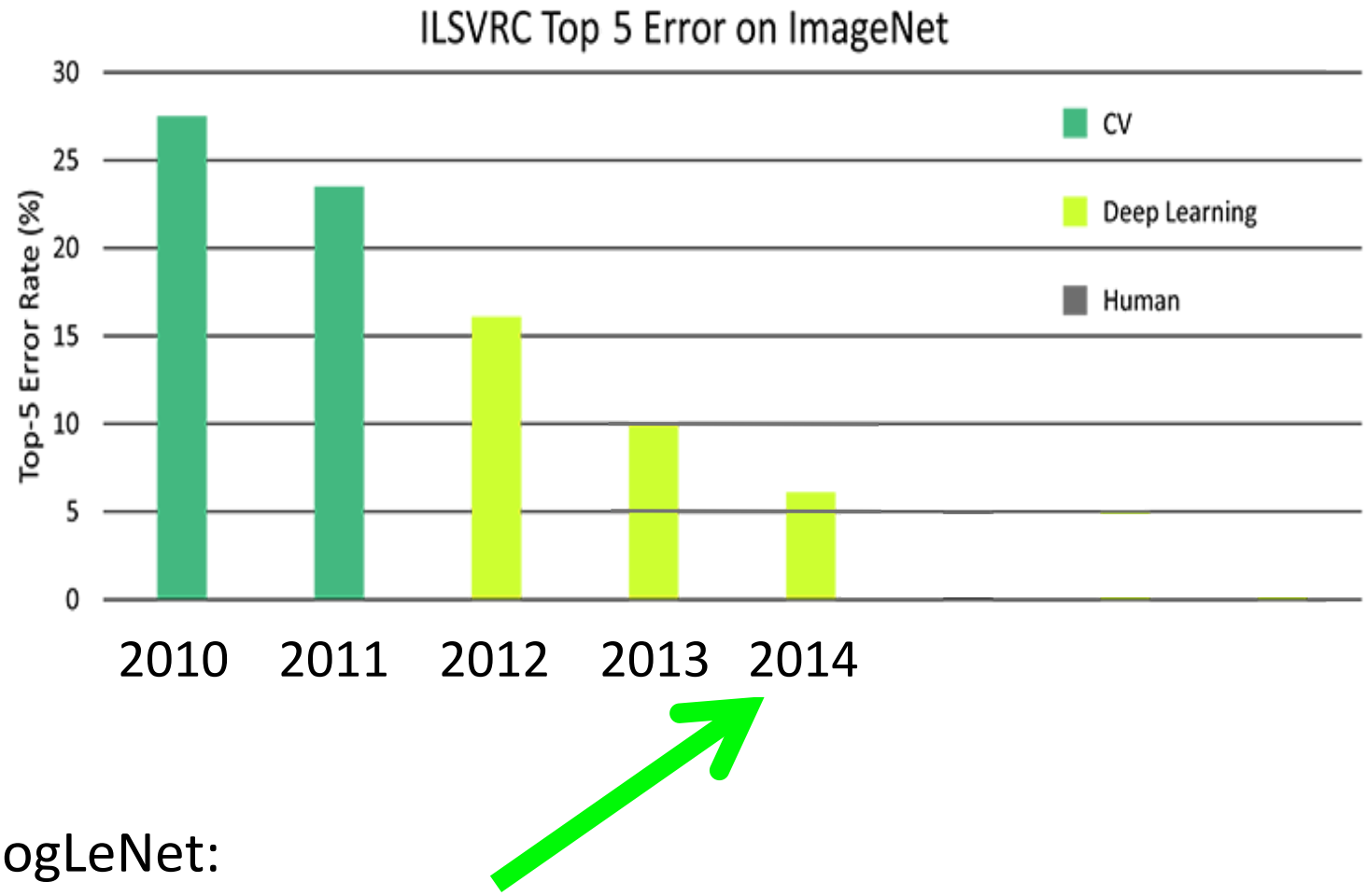
Alexnet

Why Deep Learning



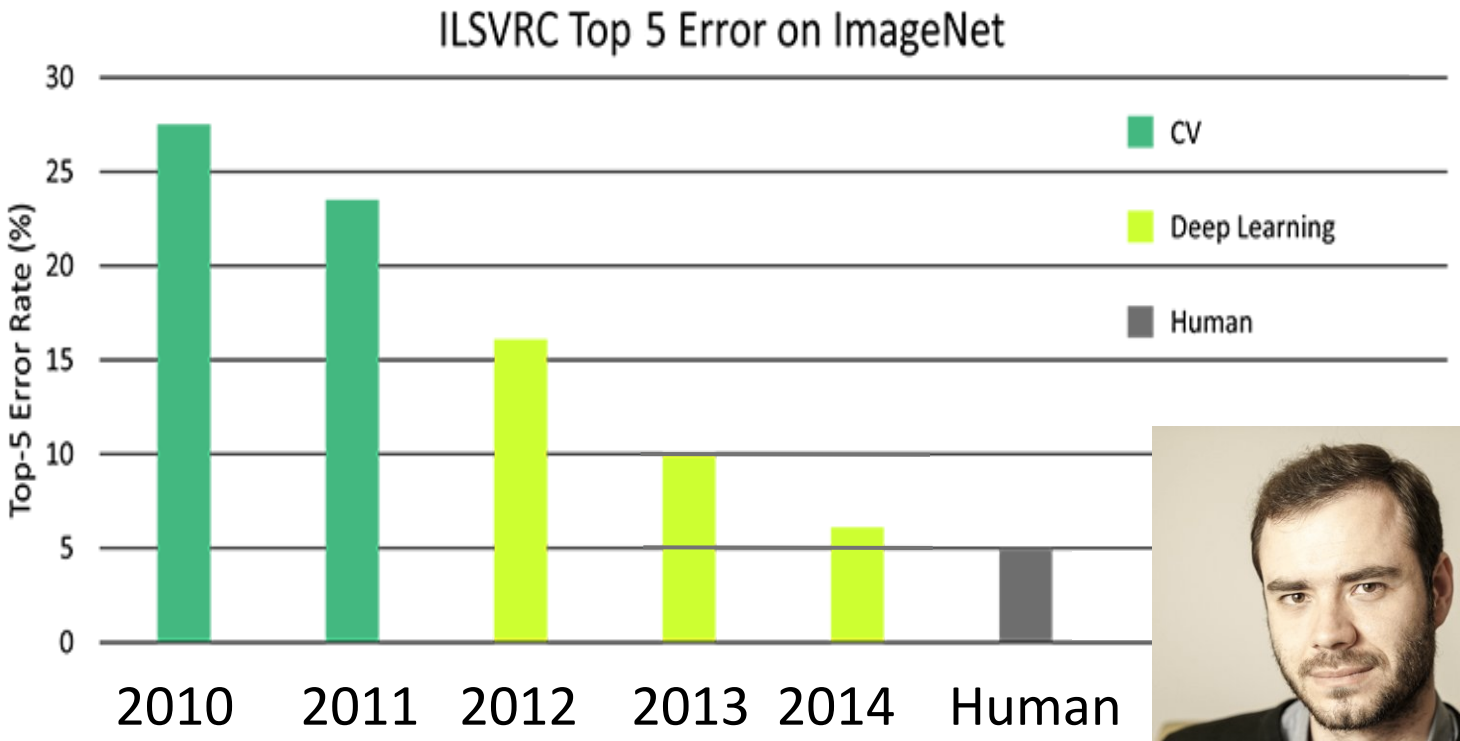
ZF Net:
Tuned Alexnet

Why Deep Learning



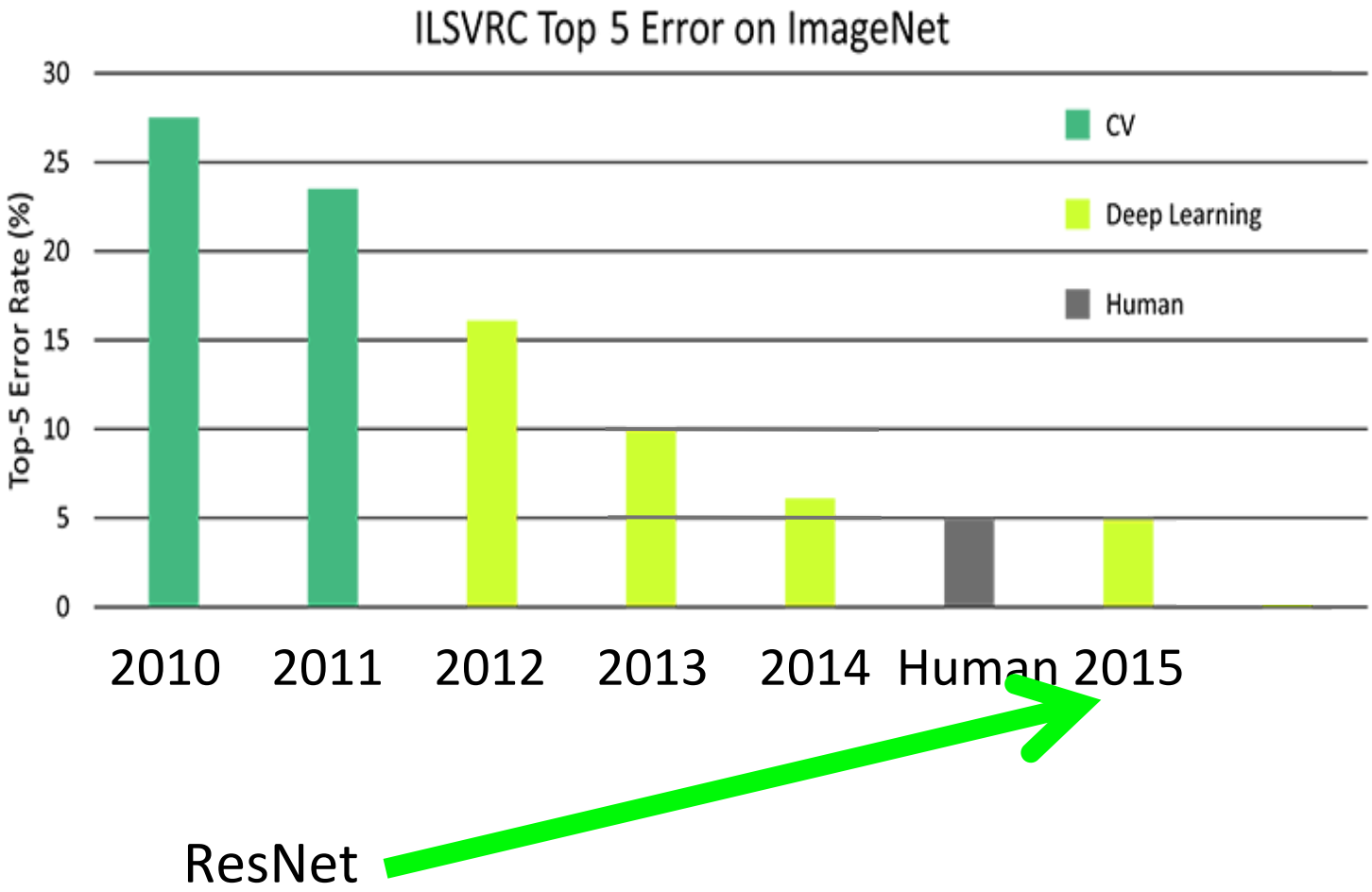
GoogLeNet:
AKA “Inception”

Why Deep Learning

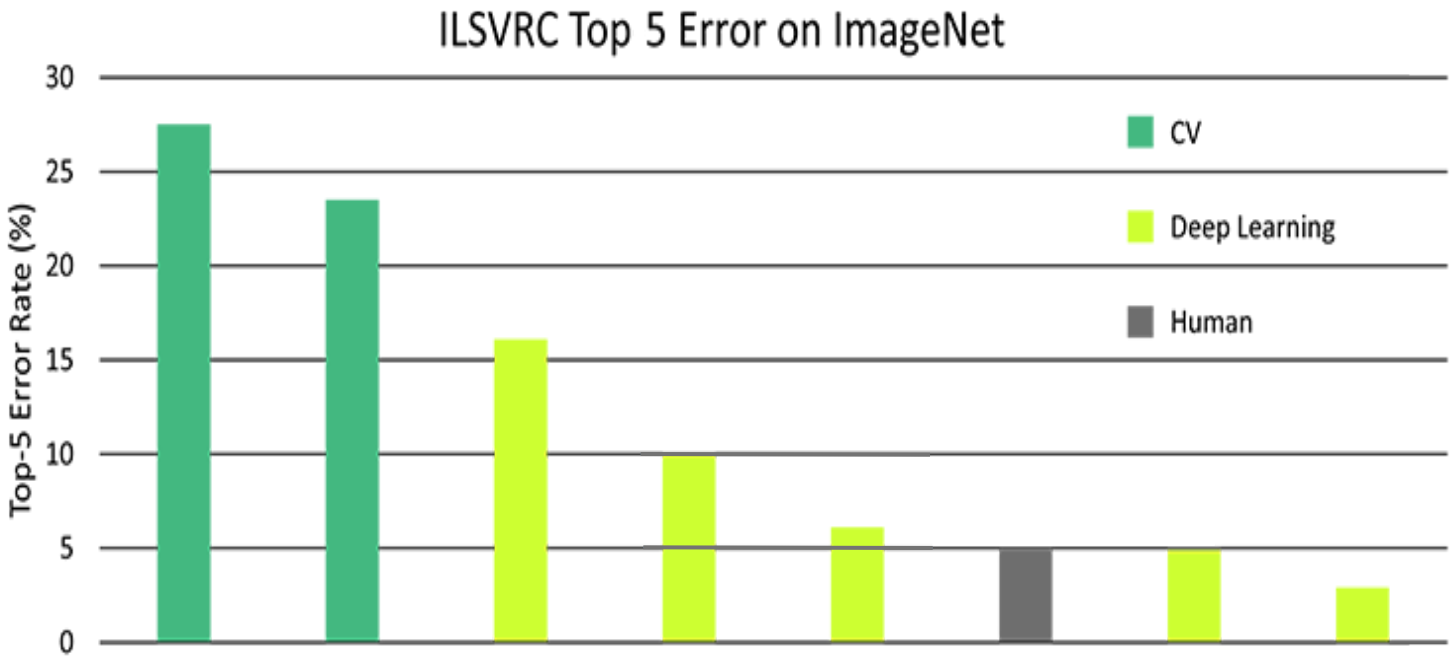


Andrej Karpathy after three days of training!

Why Deep Learning



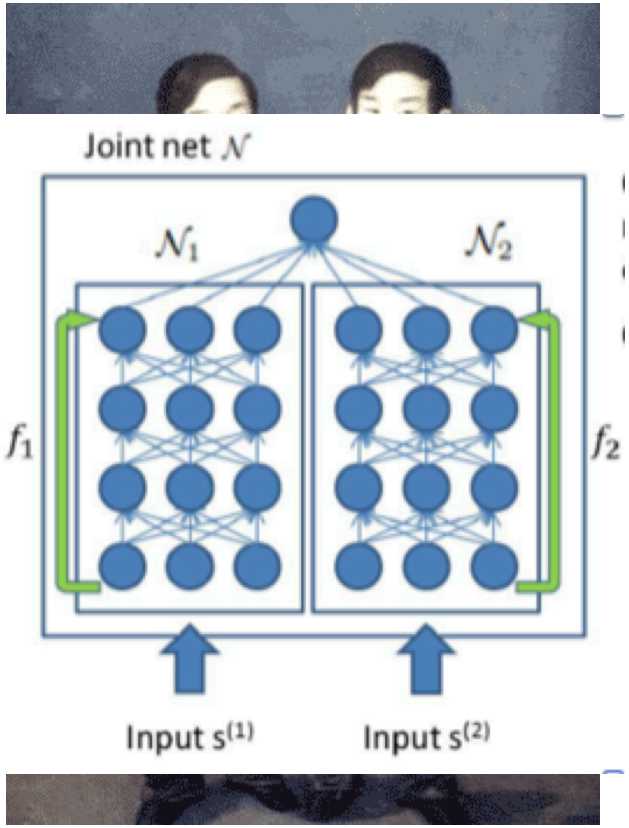
Why Deep Learning



Inception-ResNet-v2

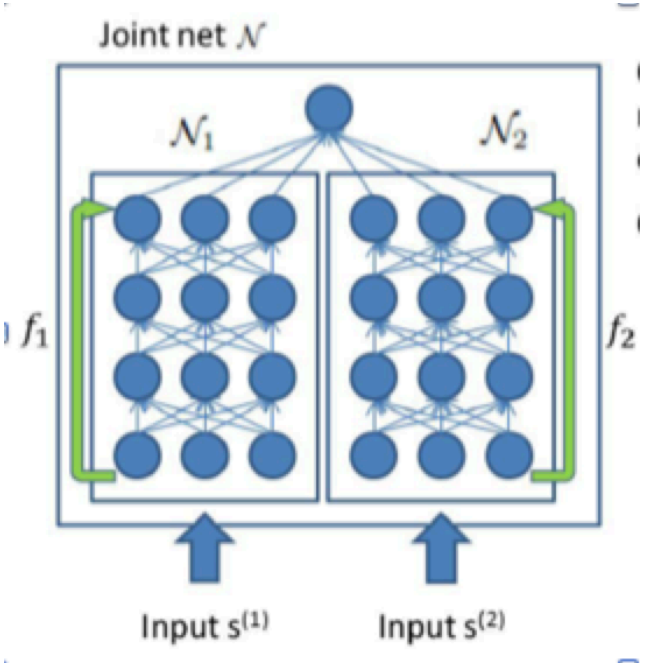
Siamese Neural Networks

- "Siamese twins" is a dated term for Conjoined Twins - twins that are physically joined at birth, sometimes sharing organs
- SNNs are the Neural Network analog for Conjoined Twins
- SNNs have two identical neural networks (like twins that are physically joined at birth)
- SNNs share their weights, both networks have the same weights (like twins sharing organs)
- They take two inputs and share a symmetric loss function (more about the loss function later)

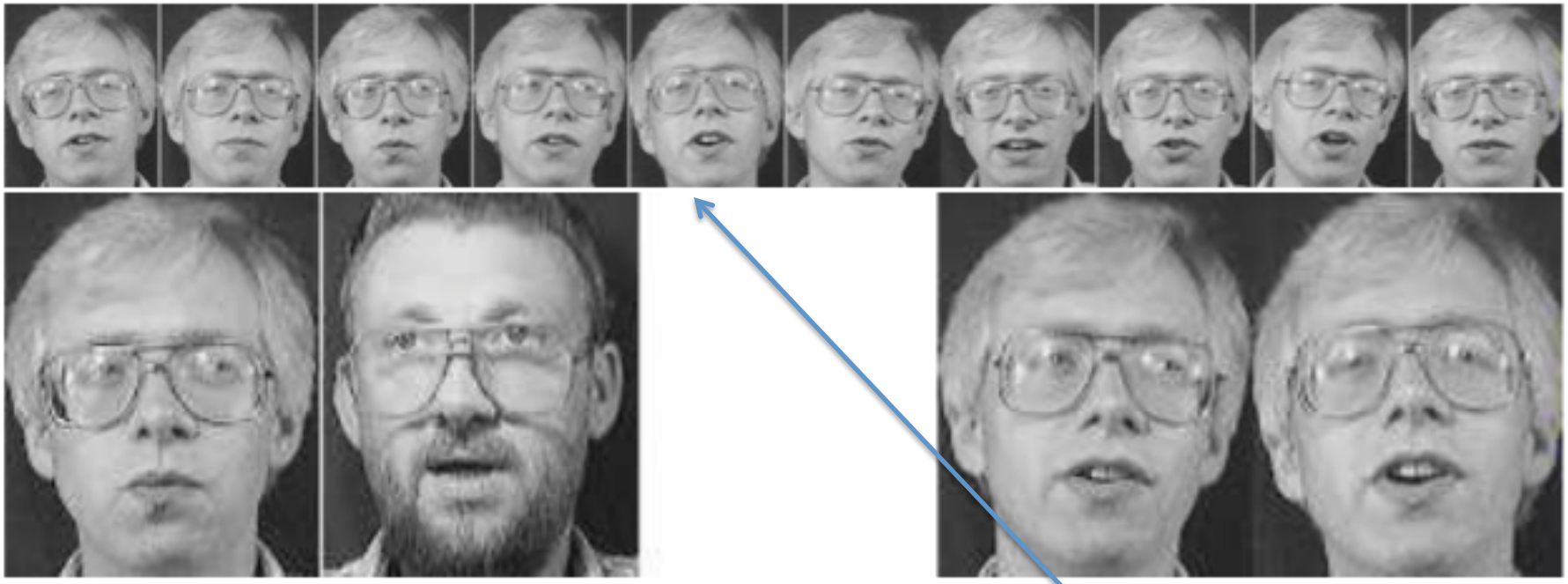


Siamese Neural Networks

- Siamese Networks can be convolutional..
- They have linear output units, but they are not trained to any particular output – instead, they are trained to move outputs closer or farther apart.
- If two inputs are from the same category, minimize the distance between them
- If two inputs are from different categories, maximize the distance between them – up to some margin



Same/Different Categories



Different:
Move the output vectors apart

Same:
Move the output vectors closer

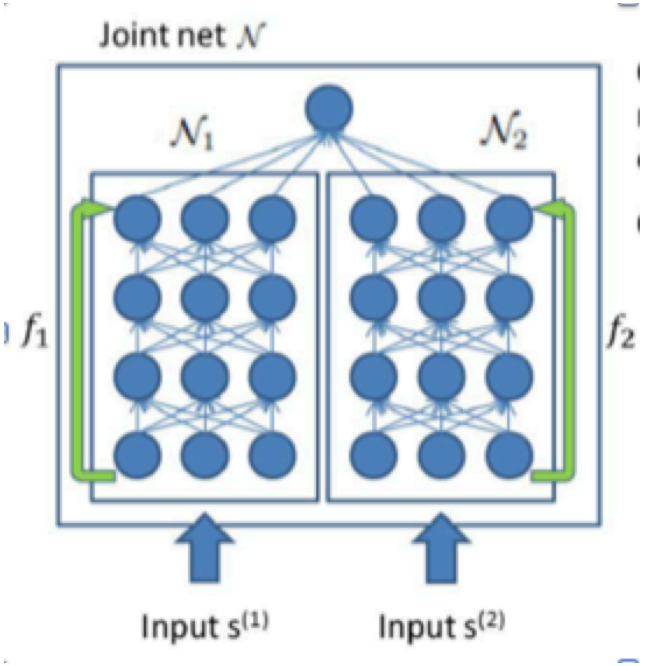
Same/Different Categories

Note that putting all of these into the same category means the network has to learn to be *invariant* to the differences in these images

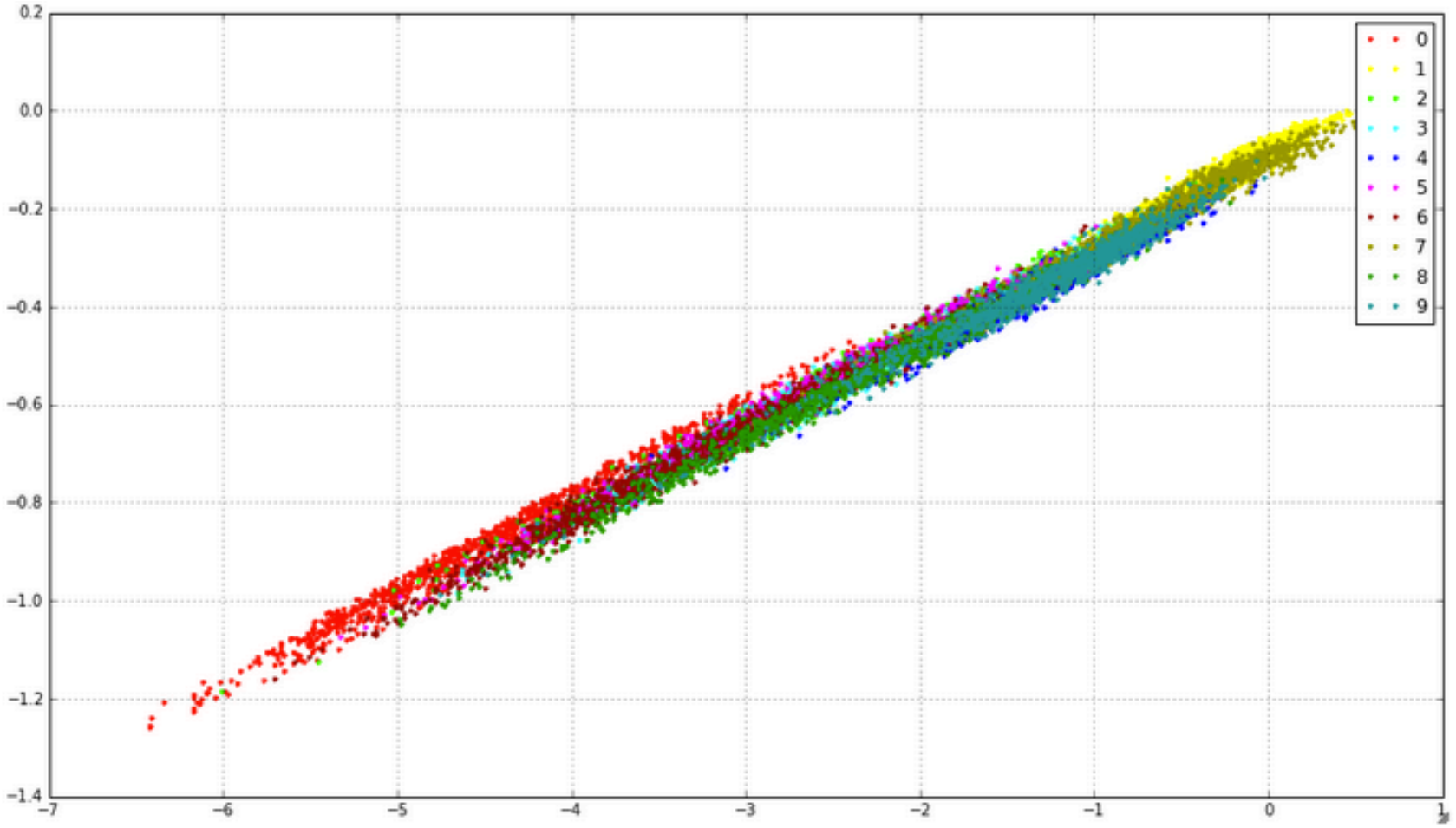
This transformation should *generalize* to new faces

Siamese Neural Networks

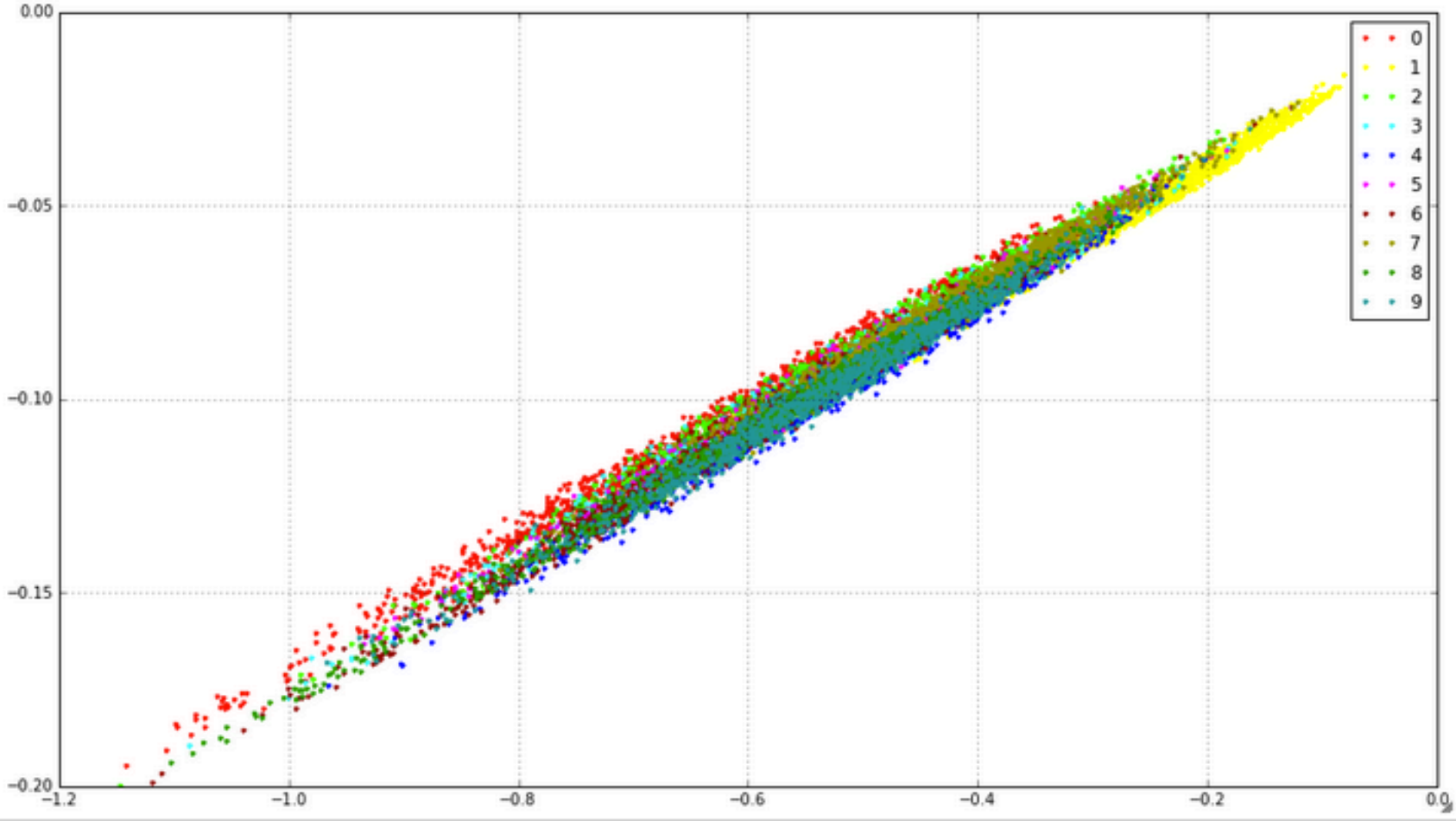
Next: MNIST Example using two output units (so you can see what they do



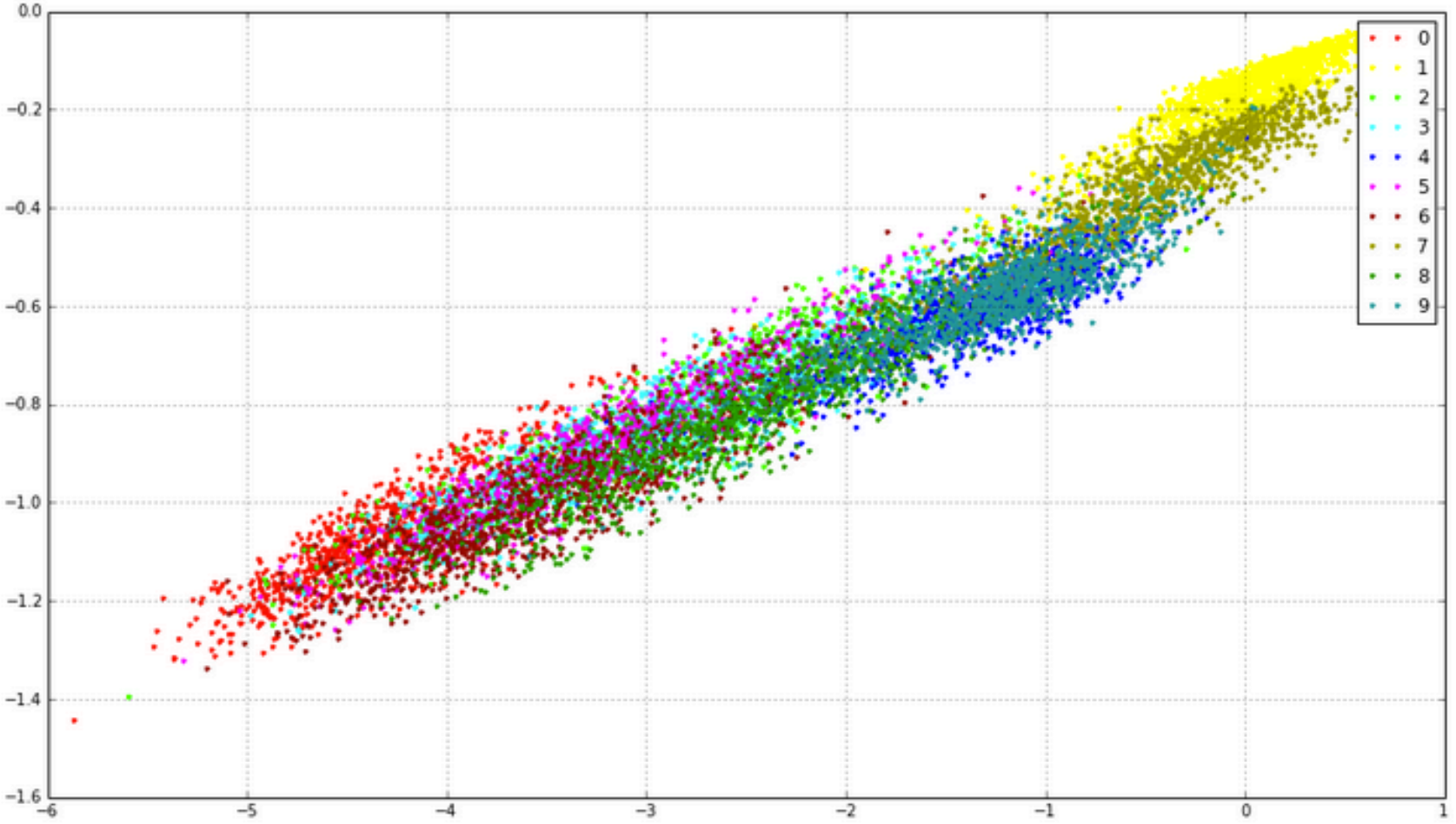
MNIST 100 iterations



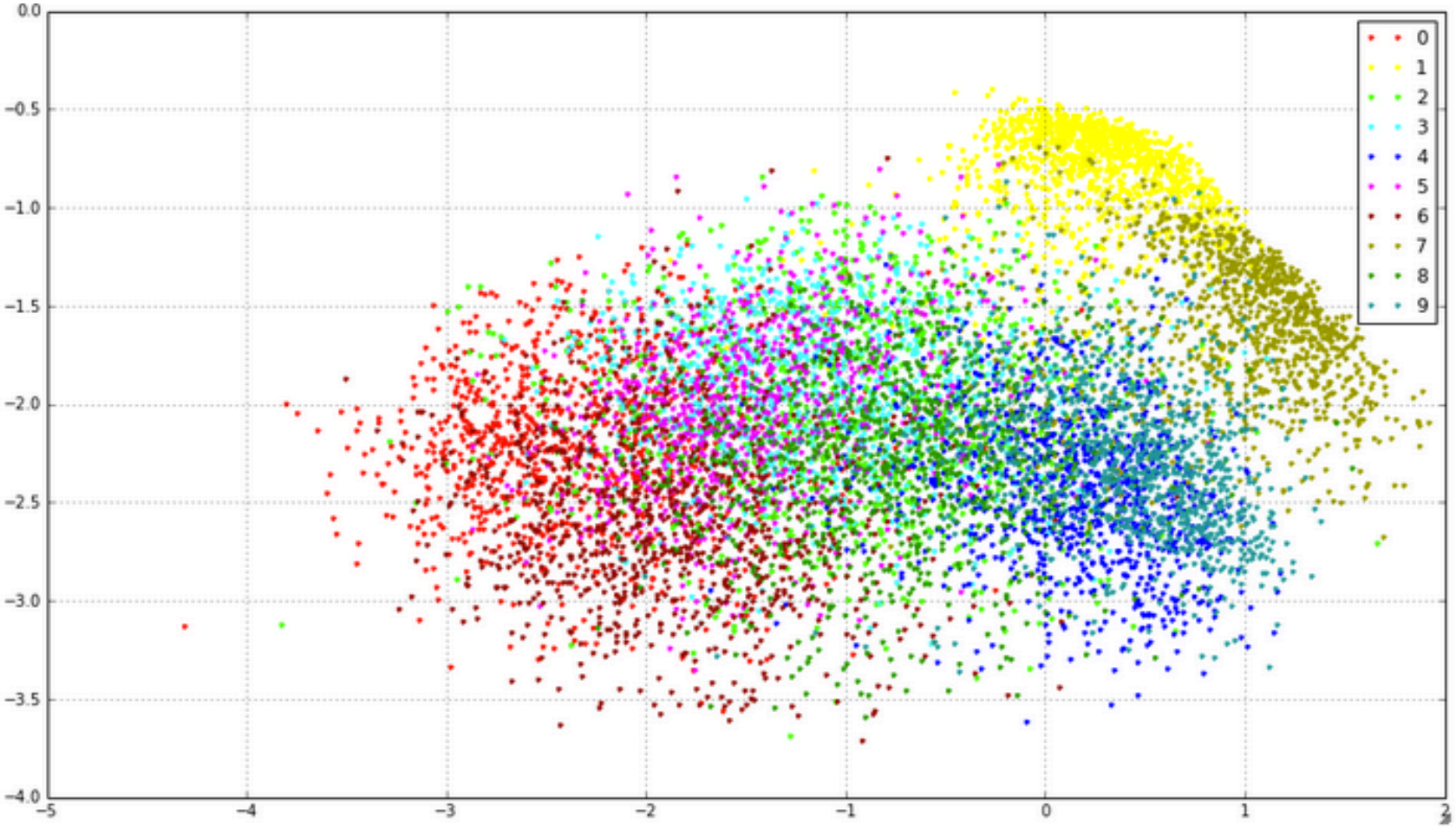
MNIST 200 iterations



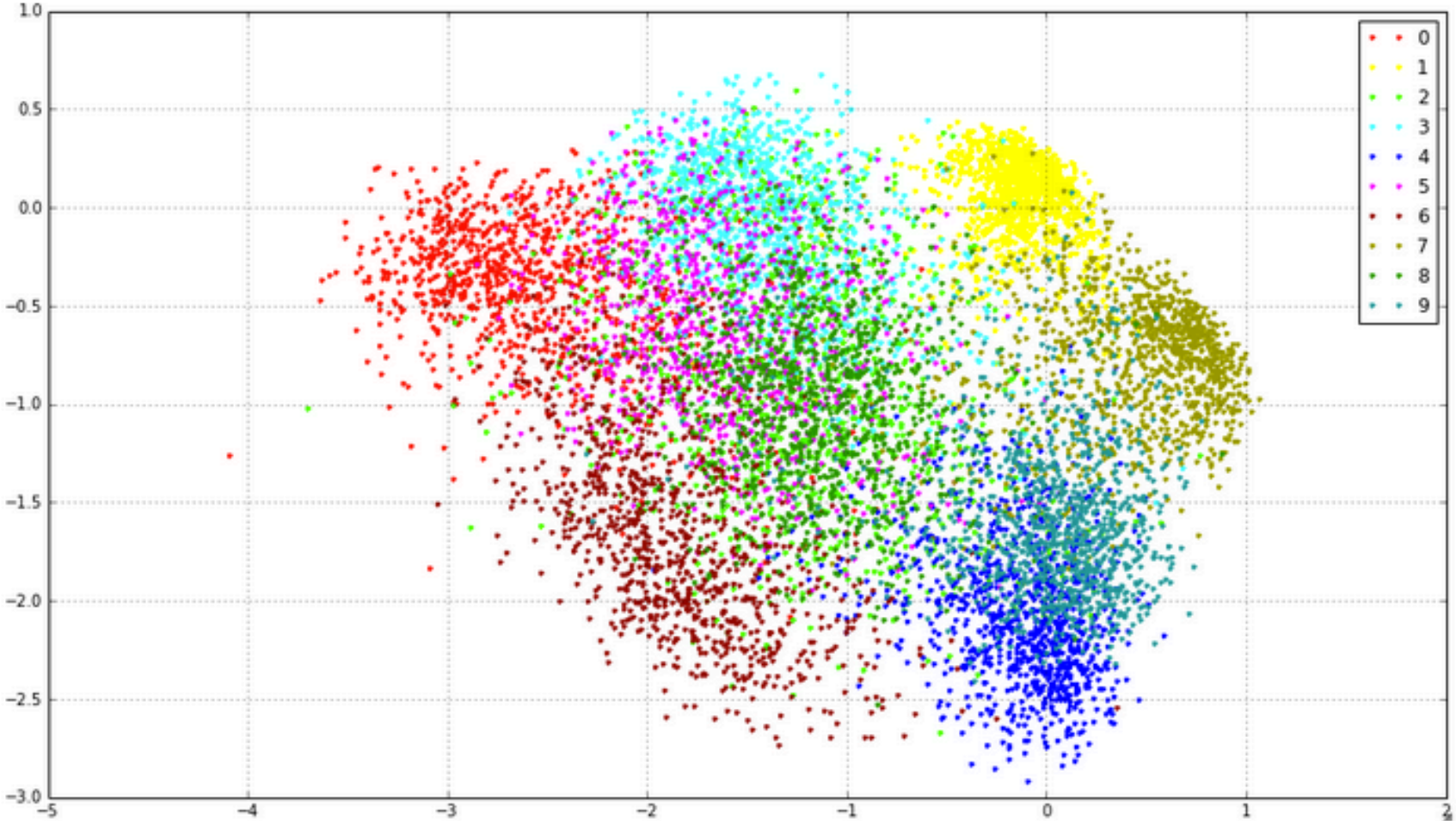
MNIST 300 iterations



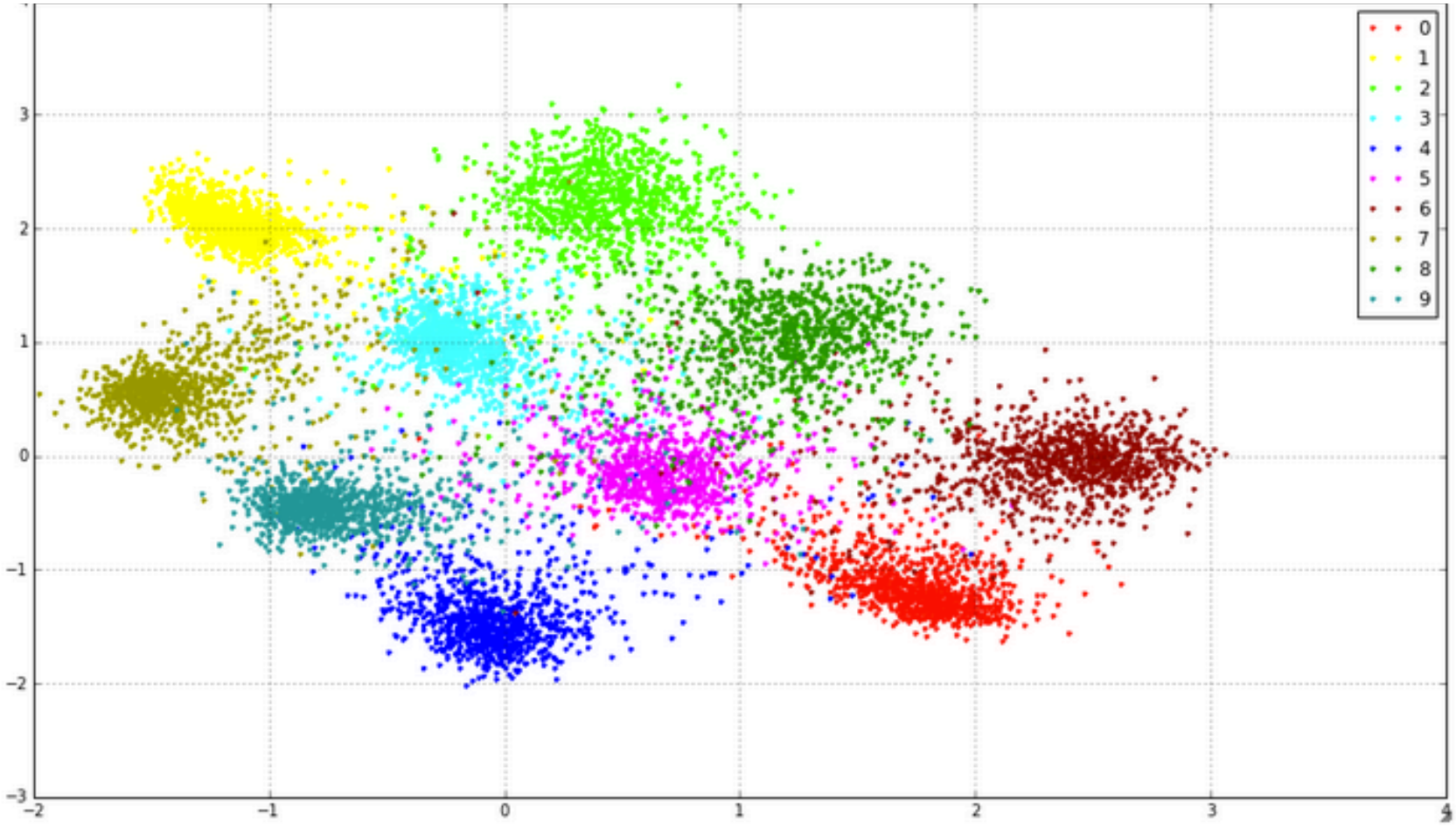
MNIST 400 iterations



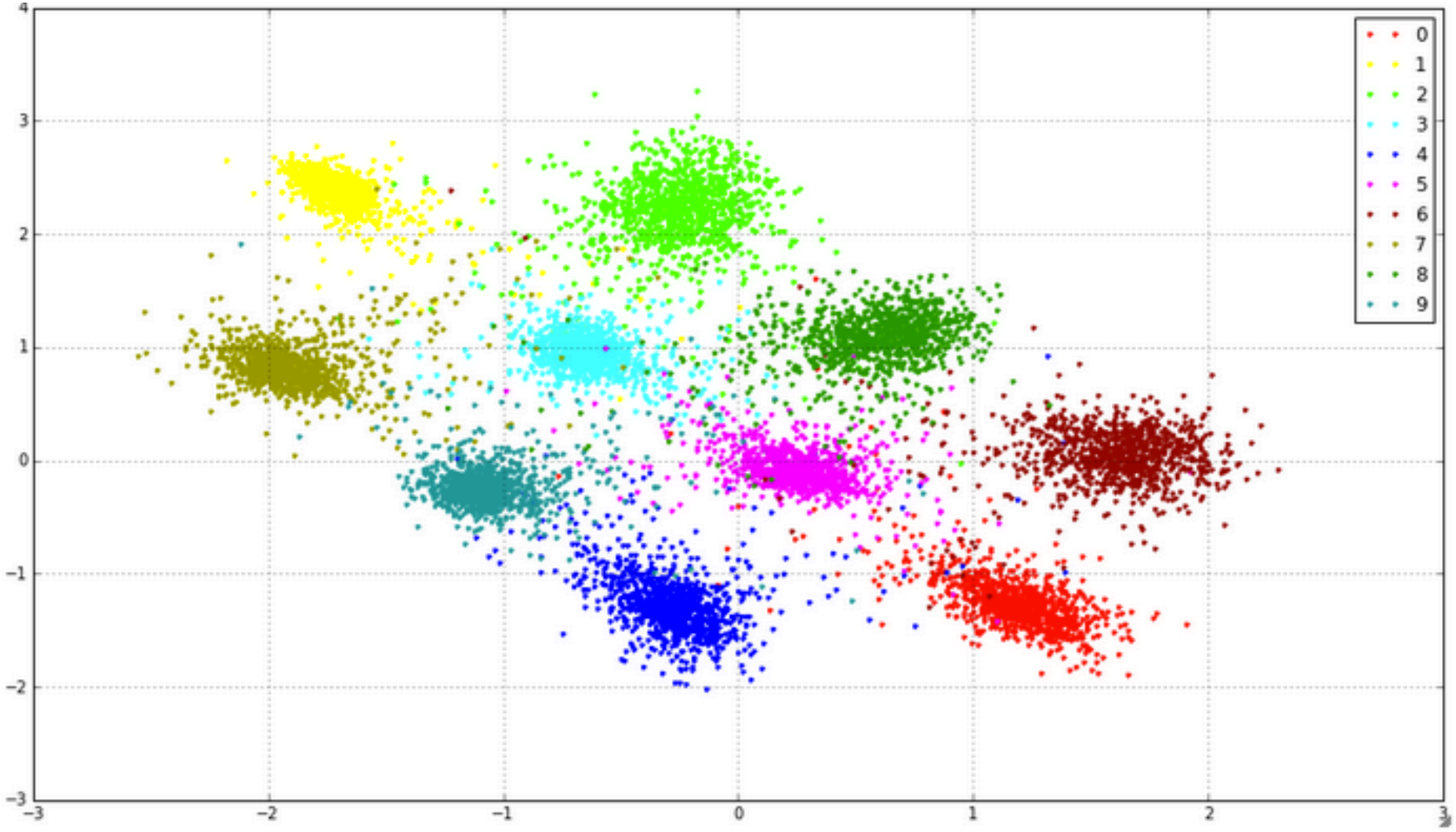
MNIST 500 iterations



MNIST 5000 iterations



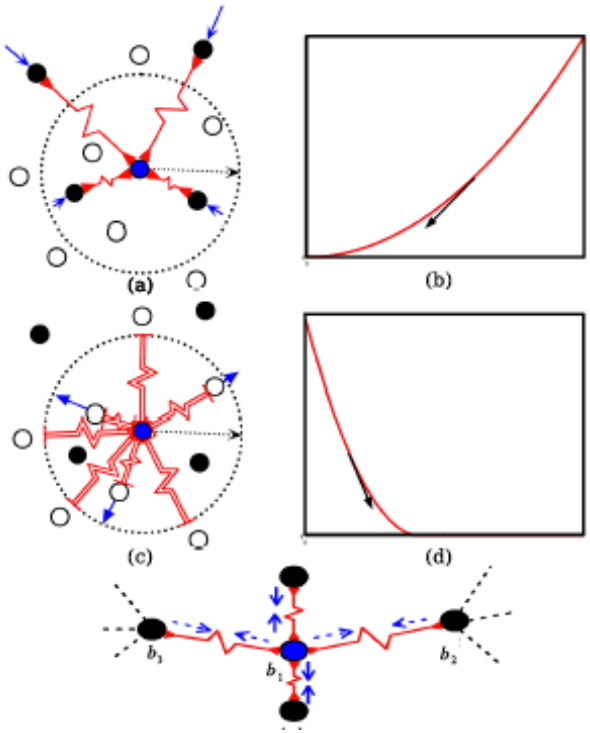
MNIST 50000 iterations



Advantages of Siamese Neural Networks

- Because you train on pairs of data points, they *amplify your training set*
- You don't need to know the number of categories in advance
- You don't even have to have examples for all categories!
- They are particularly suitable for domains where you have a large number of categories with a small number of examples

The loss function – The Spring Analogy



$$L(W,Y,\vec{X}_1,\vec{X}_2) =$$

$$(1-Y)\frac{1}{2}(D_W)^2 + (Y)\frac{1}{2}\{max(0,m-D_W)\}^2$$

Where

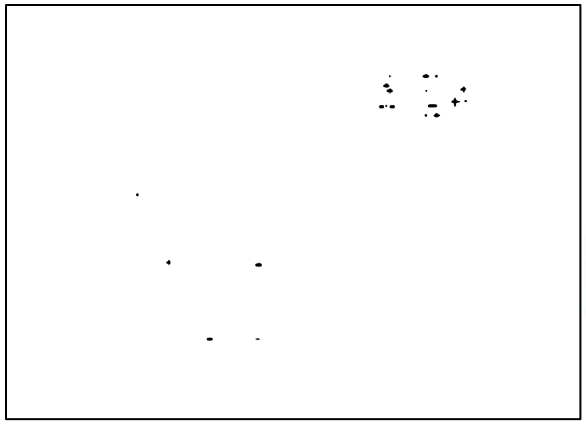
- D is the distance between the outputs of the two networks,
- m is a margin
- Y is 1 for a "different" pair and 0 for "same" pair

$$D_W(\vec{X}_1,\vec{X}_2) = \|G_W(\vec{X}_1) - G_W(\vec{X}_2)\|_2$$

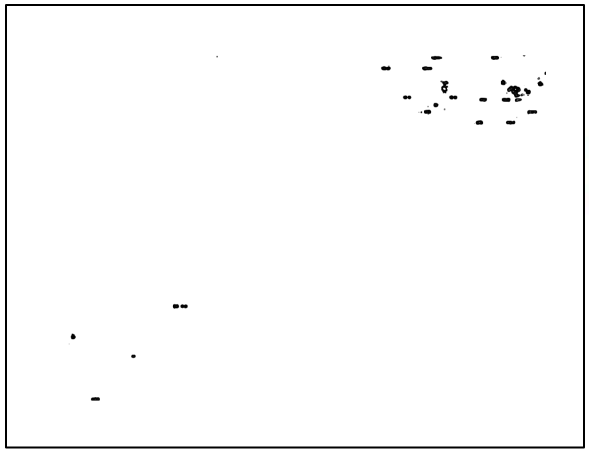
$G_W(X)$ is the output vector of the network on stimulus X.

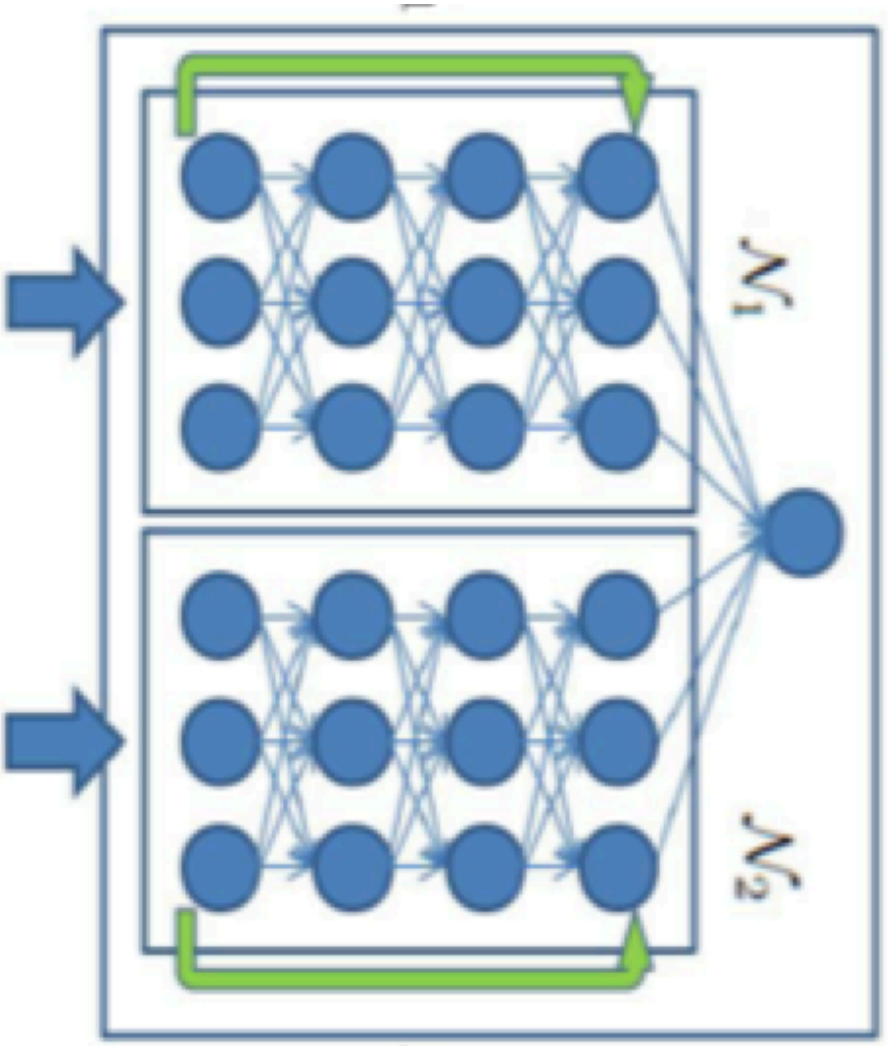
The basic idea

Aspewentin A



Aspewentin C

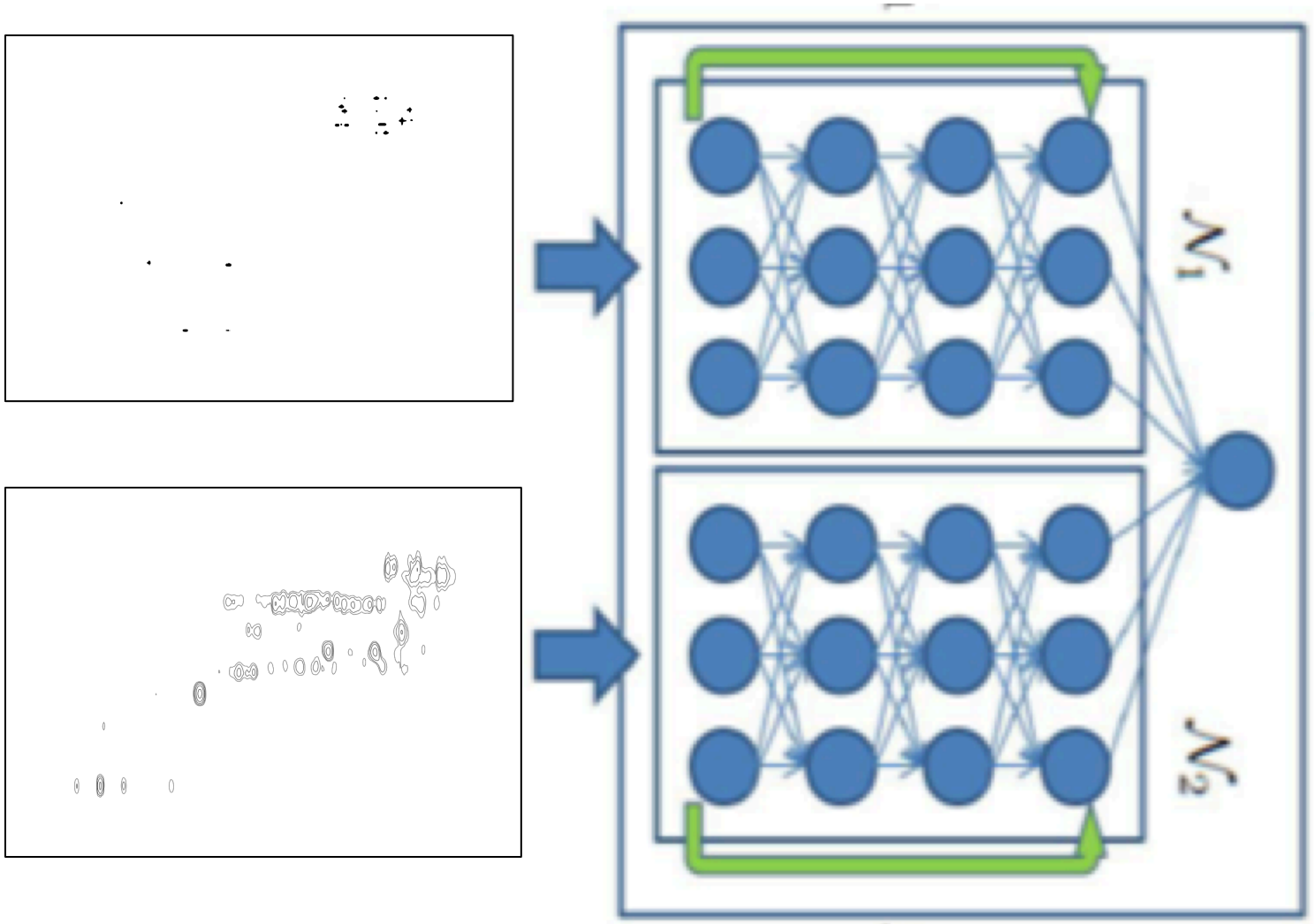




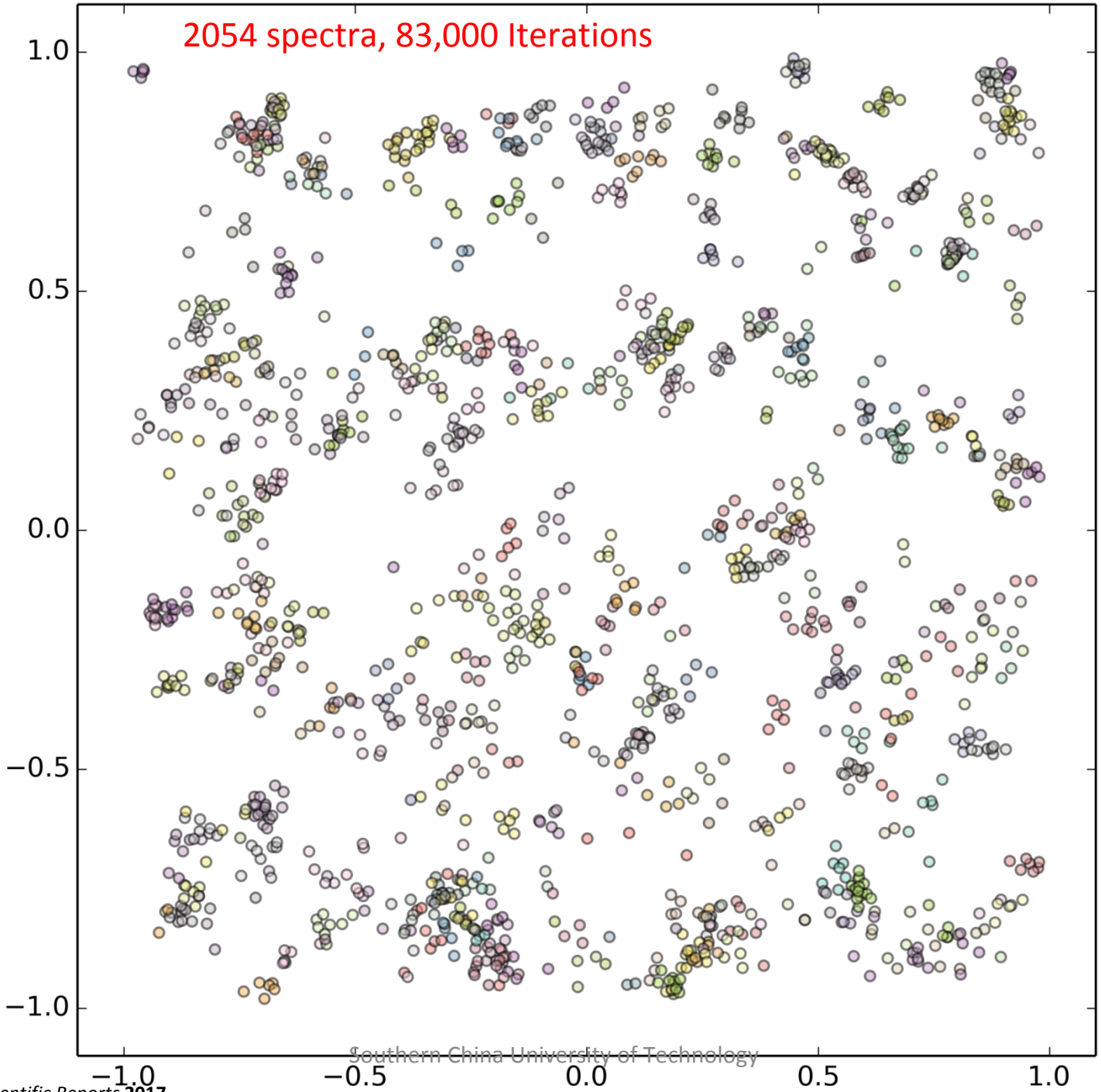
"SAME"

The basic idea

Aspewentin A Chandonanone A



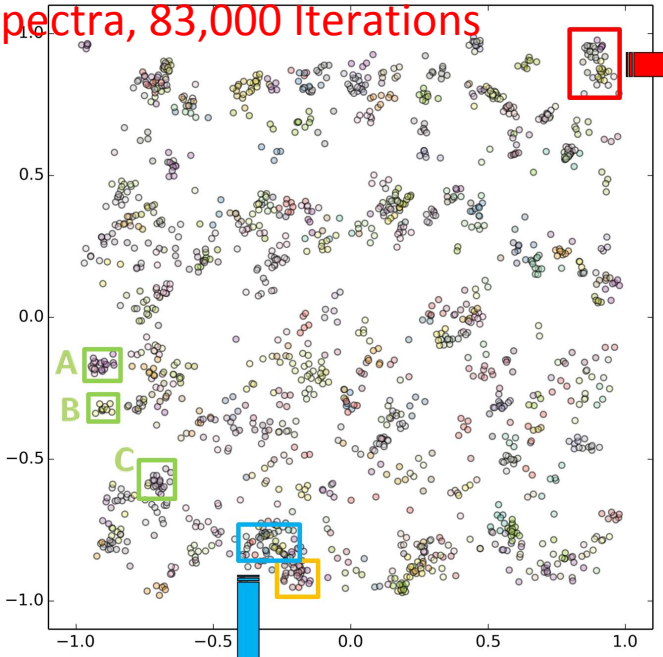
"DIFFERENT"

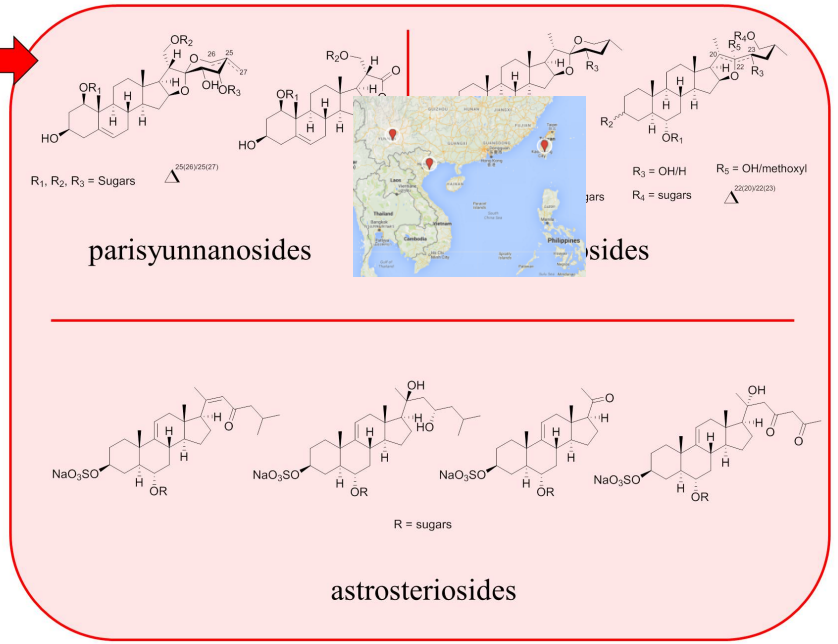


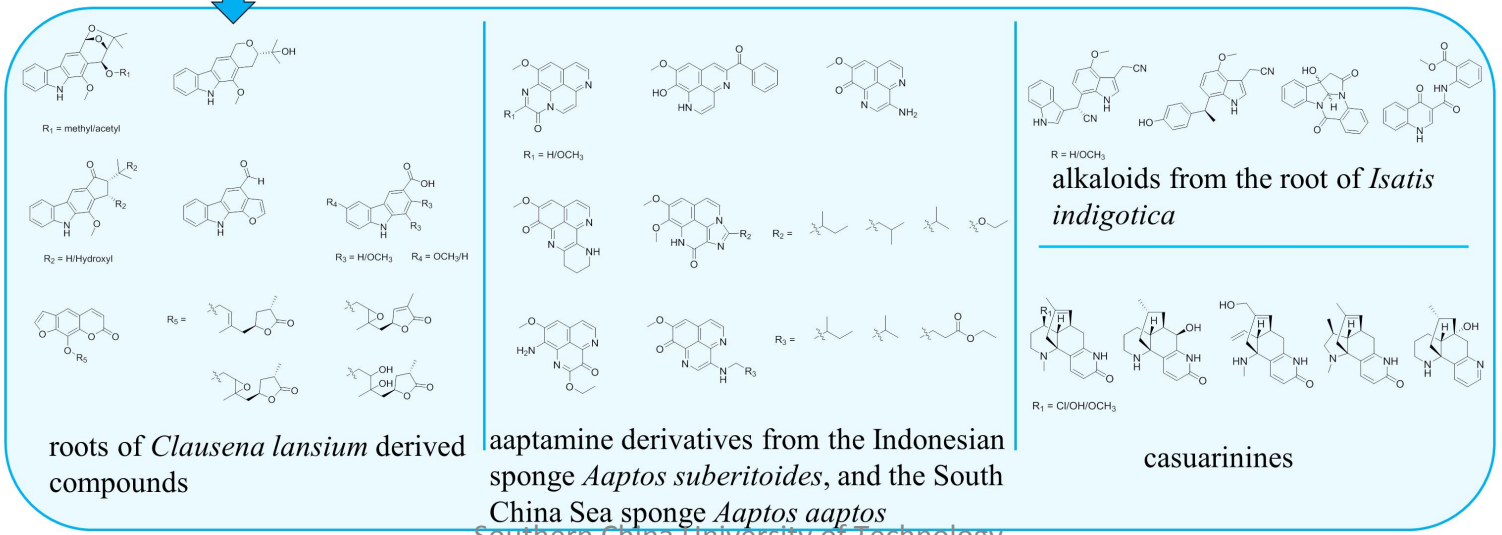
Chen Zhang et al., *Scientific Reports* **2017**

Qualitative Results

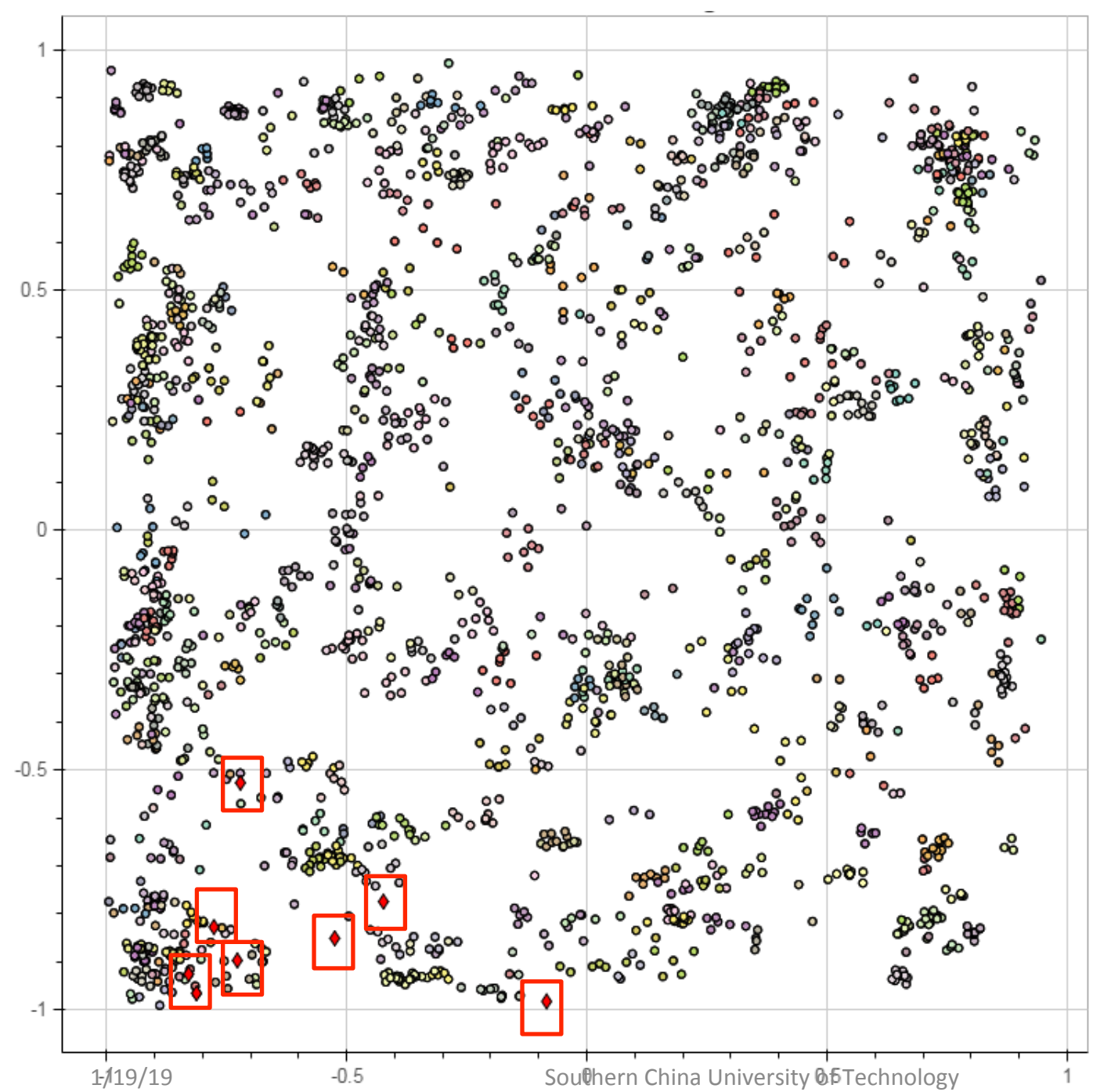
2054 spectra, 83,000 Iterations







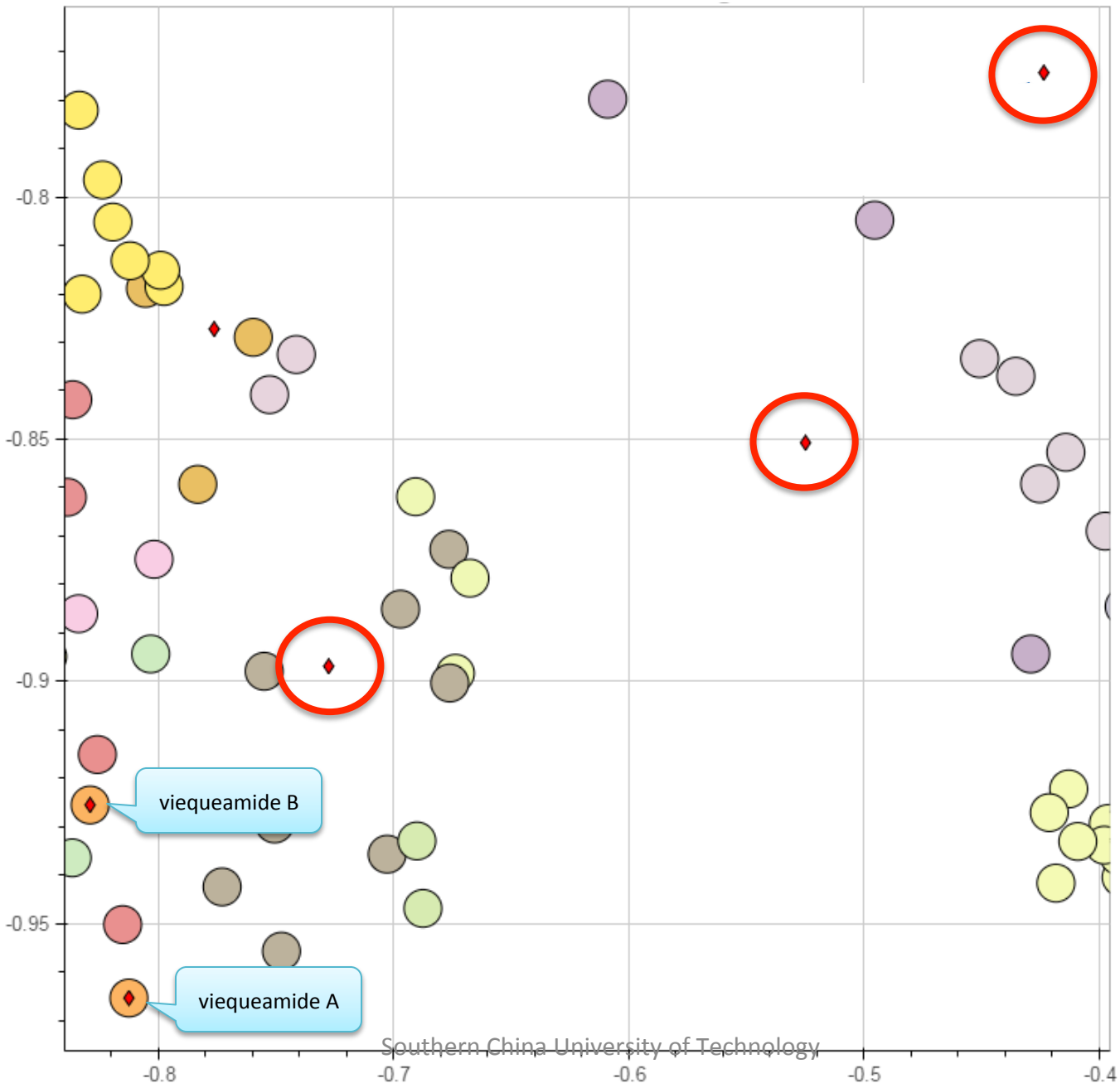
Chen Zhang et al., *Scientific Reports* **2017**



Red Diamonds: Embedding of unknown 2D NMR spectra into the Cluster Map

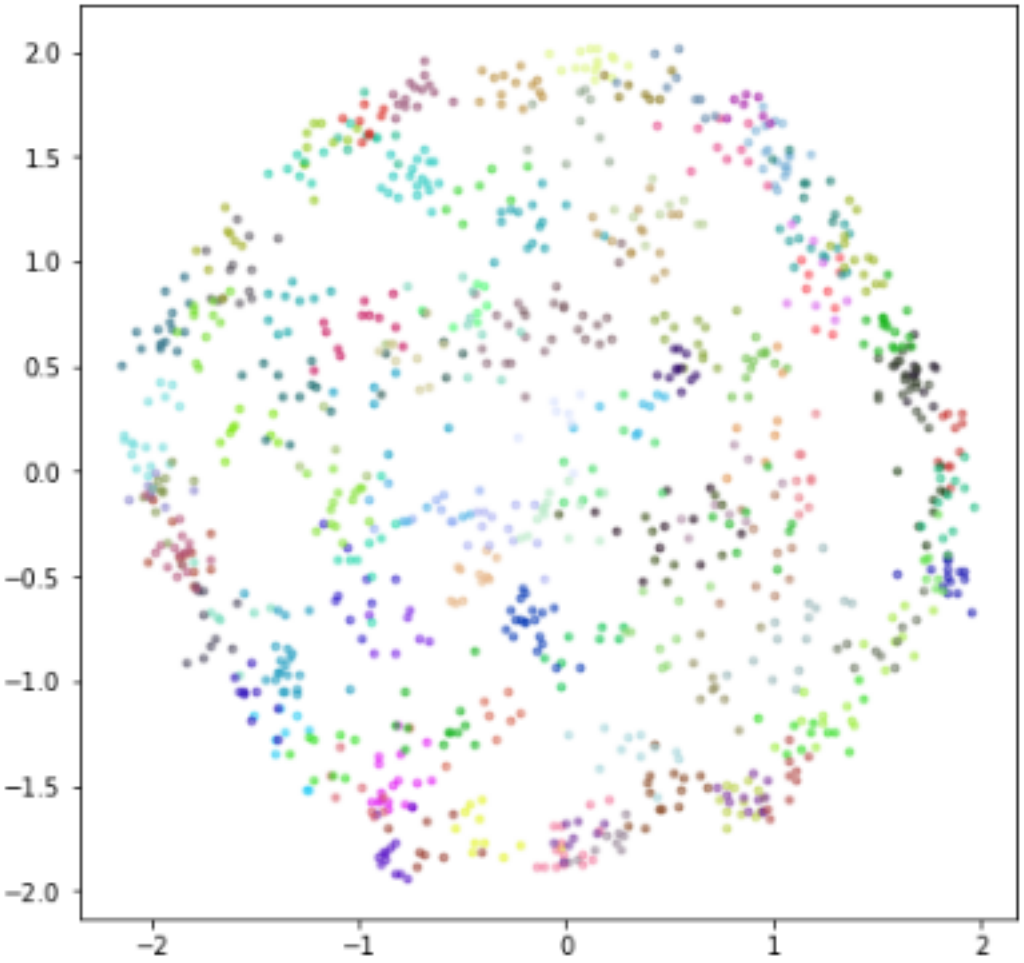
Manuscript in prep, 2017

Enlargement of the local area of the embedding map

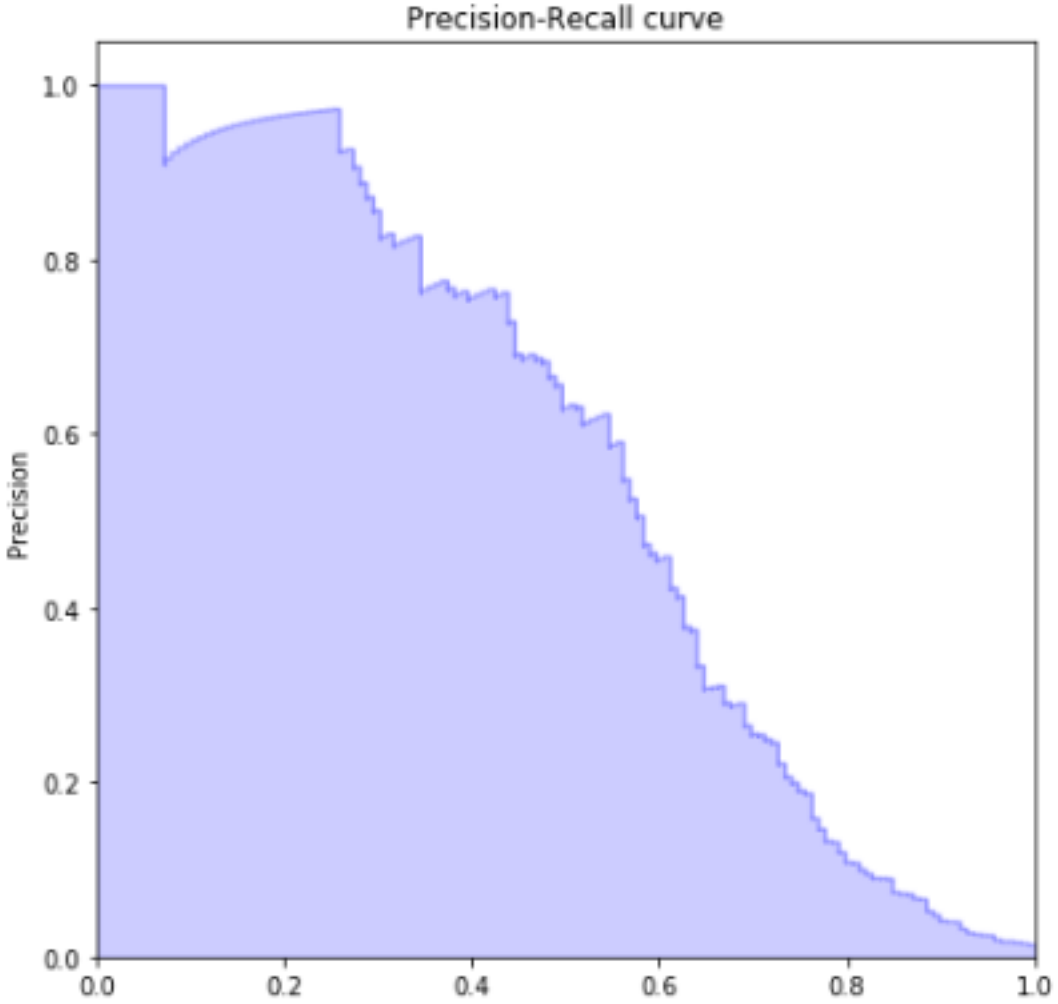


Manuscript in prep, 2017

More recent results: just 300 epochs!



Quantitative Results: 10D outputs



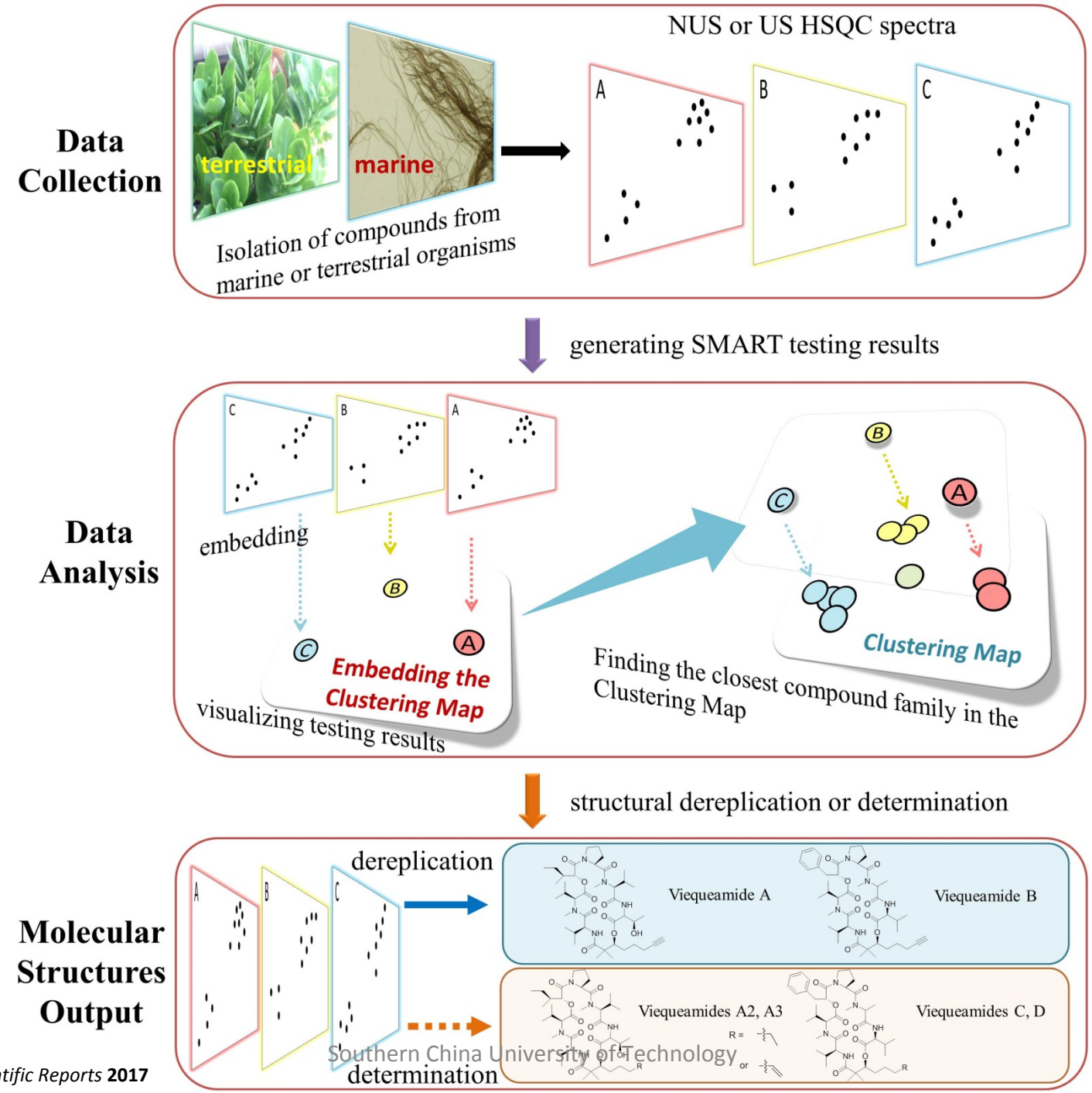
Precision

Recall

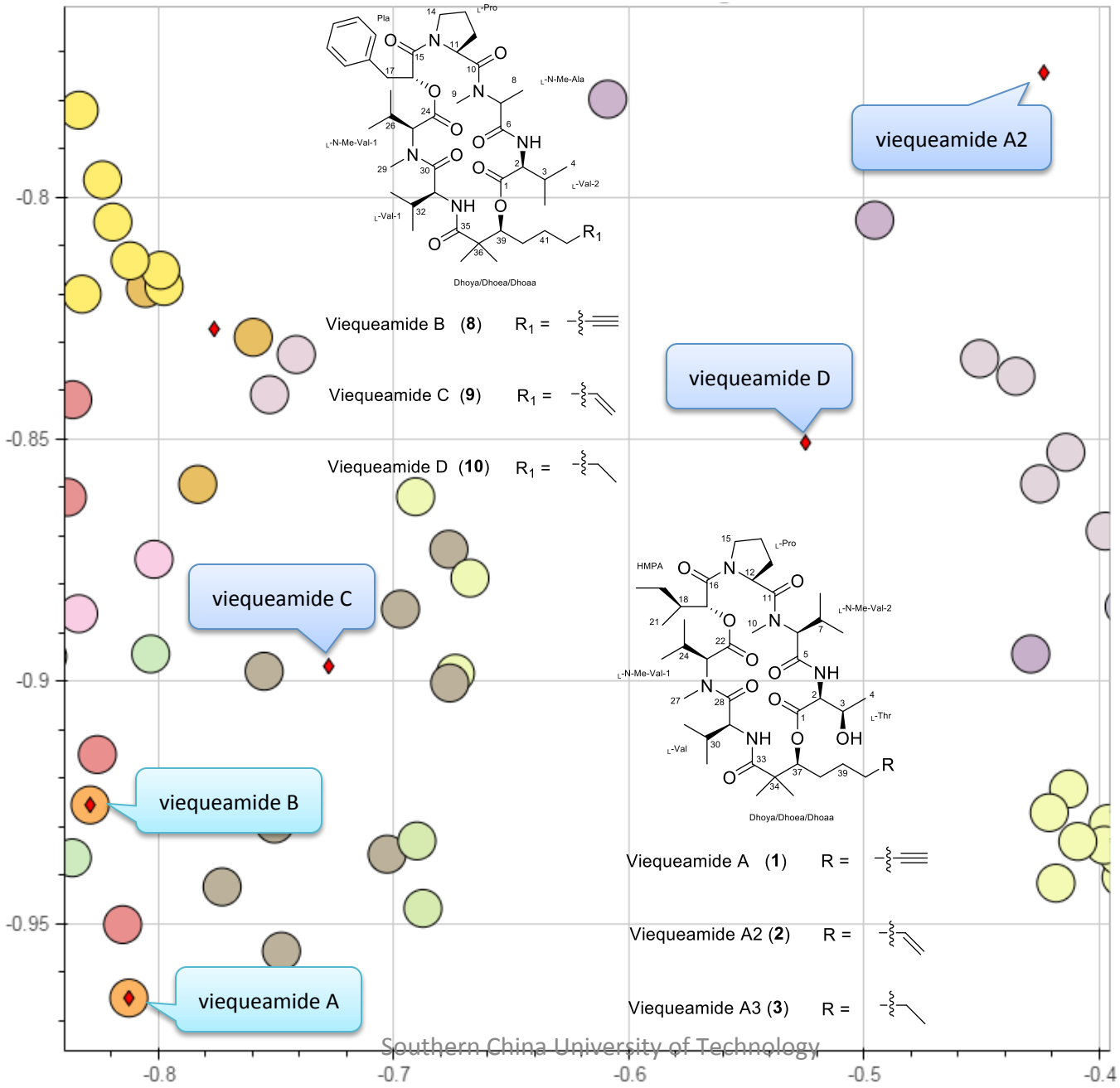
Demo

- What you will see is the result of a network with 3 outputs

- So, a 3D clustering

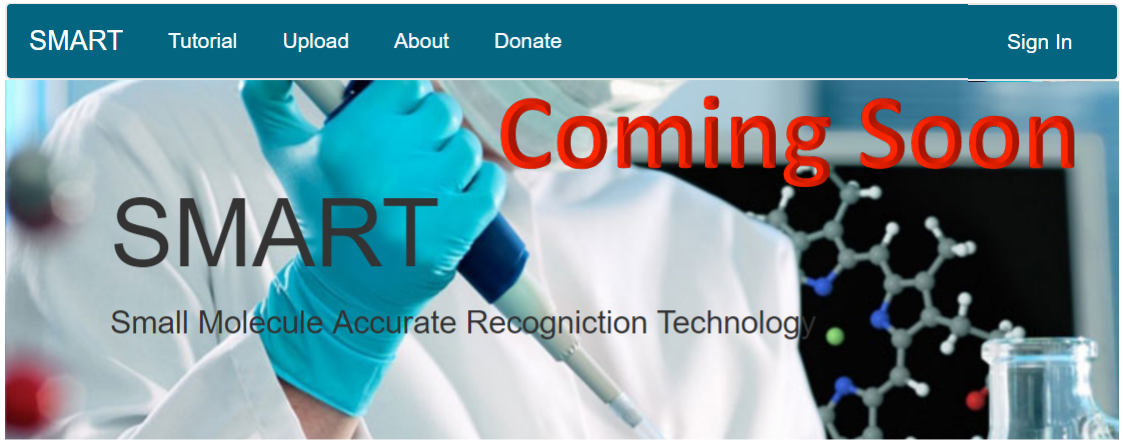

Data Collection
terrestrial
marine
Isolation of compounds from marine or terrestrial organisms
NUS or US HSQC spectra
A
B
C
generating SMART testing results
Data Analysis
embedding
Embedding the Clustering Map
visualizing testing results
Clustering Map
Finding the closest compound family in the Clustering Map
structural dereplication or determination
Molecular Structures Output
dereplication
determination
Viequeamide A
Viequeamide B
Viequeamides A2, A3
Viequeamides C, D

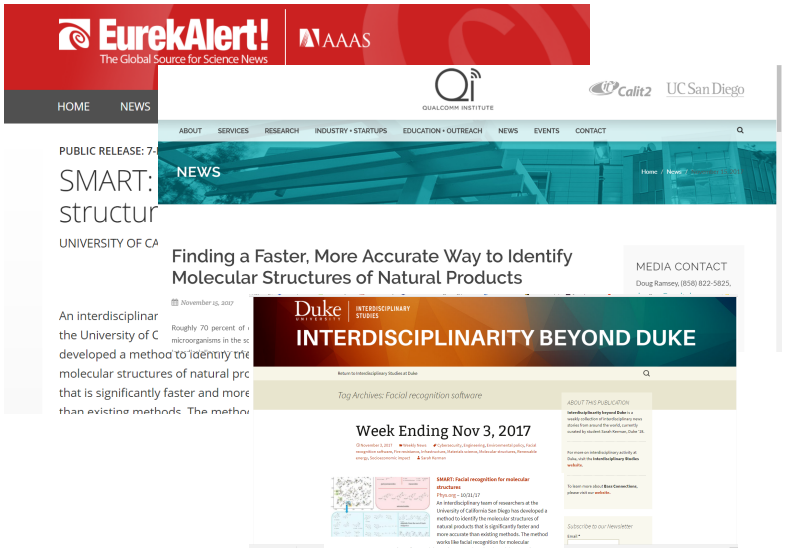
Enlargement of the local area of the embedding map



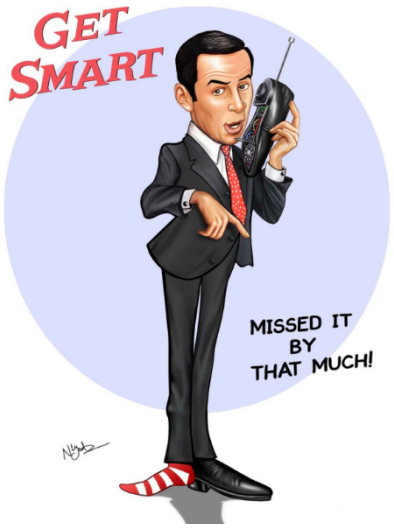
Manuscript in prep, 2017

Conclusions and Future Work





- A lot of inquiries and requests from researchers worldwide
  - Pharmaceuticals
  - Perfume and cosmetics
  - Agricultural chemistry
  - Environmental sciences
  - NMR industries
- Soon: A user-friendly website with enhanced data security features



Future Work

- These results are using a "legacy" neural network – not optimized for the problem
  - Better network architecture
  - Using an autoencoding layer is supposed to improve results (Yann LeCun – personal communication)
- We are working on a new network objective function based on "purity" – a measure of cluster purity
  - Purity is non-differentiable
  - We have created a "soft", differentiable version

Acknowledgements

The NMR:

- Dr. Brendan Duggan
- Dr. Anthony Mrse
- Dr. Eugene Lin

Physics:

- Dr. Preston Landon
- Dr. Jie Min

GURU members:

Yash Nannapaneni

Nick Roberts

Poornav Sargur Purushothama

Yerlan Idelbayev

UC Merced

The Gerwick Lab

Acknowledgements

The NMR:

- Dr. Brendan Duggan
- Dr. Anthony Mrse
- Dr. Eugene Lin

Physics:

- Dr. Preston Landon
- Dr. Jie Min

GURU members:

Yash Nannapaneni

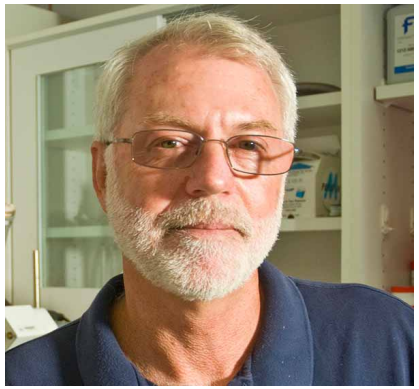
Nick Roberts

Poornav Sargur Purushothama

Yerlan Idelbayev

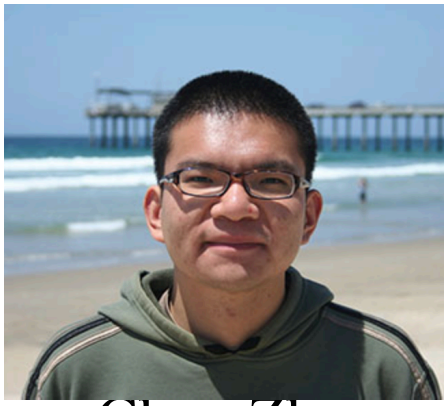
UC Merced

The Gerwick Lab



Bill Gerwick

Scripps Instution of Oceanography



Chen Zhang

Scripps Instution of Oceanography

Acknowledgements: Funders

Questions?