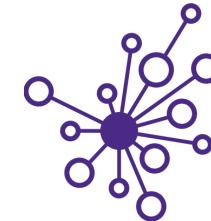




Knowledge and solutions
for a changing world



Be boundless



Advancing data-intensive
discovery in all fields

Data Science Methods for Clean Energy Research (DSMCER)

UW DIRECT

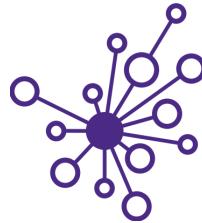
(Data Intensive Research Enabling Cutting-edge Tech)

<https://uwdirect.github.io>

David A. C. Beck (dacb)
Chemical Engineering & eScience Institute



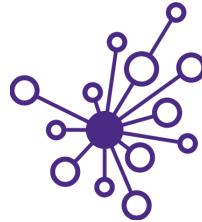
Who is that guy Dave?



- Computer Science, BS
 - Accounting and contact management software
 - Biomolecular Structure & Design / Medicinal Chemistry, PhD
 - Scalable parallel software for molecular sims
 - Director of Research, eScience Institute
 - Manage data science research programs
 - Research Associate Professor, ChemE
 - Software and Data Science methods at the intersection of chemistry, biology, energy, health & environment
- Not open source! ☹*



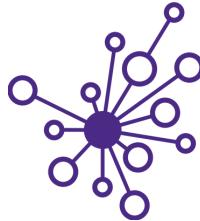
What is this class about?



- Survey of Data Science methods
 - Tool selection
 - Best practices
 - Not about designing new algorithms
- Group project using these methods



I need more details!



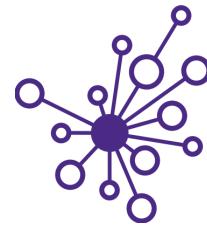
- Class website:
 - <https://uwdirect.github.io>
 - Let's go there now!
- Logistical items
 - Laptops – bring 'em!
 - Slack – use it!
 - Software – install it!

What's this all about?



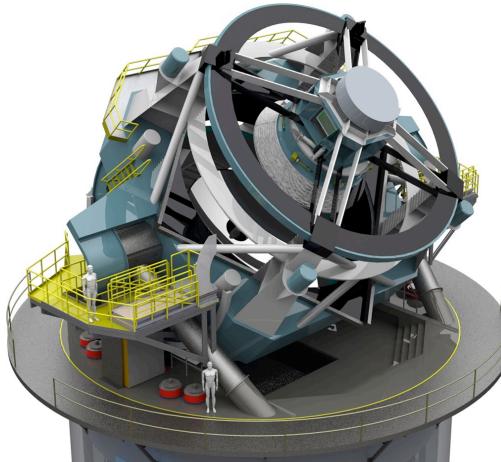
- Data
- Science & engineering
 - Chemical engineering
 - Science & engineering at large
 - Environment, energy, health, urban





OMG, so much data!

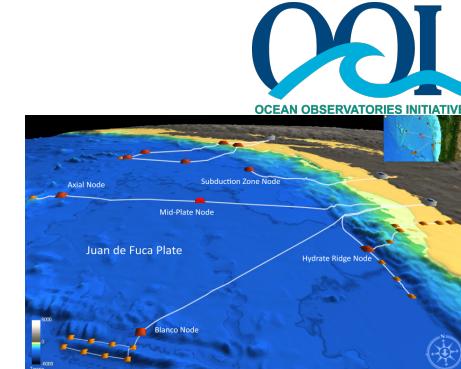
- All fields of science: “data poor” → “data rich”



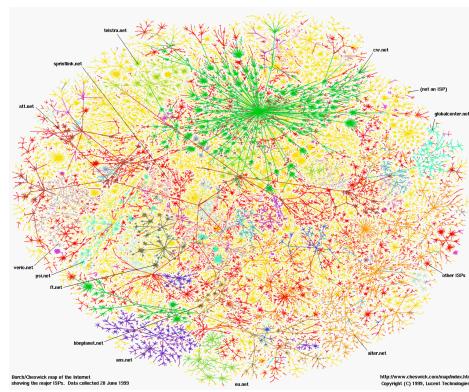
Astronomy: LSST



Physics: LHC



Oceanography: OOI



Sociology: Social networks



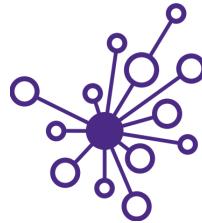
Biology: Sequencing



Economics: POS terminals



Neuroscience: EEG, fMRI

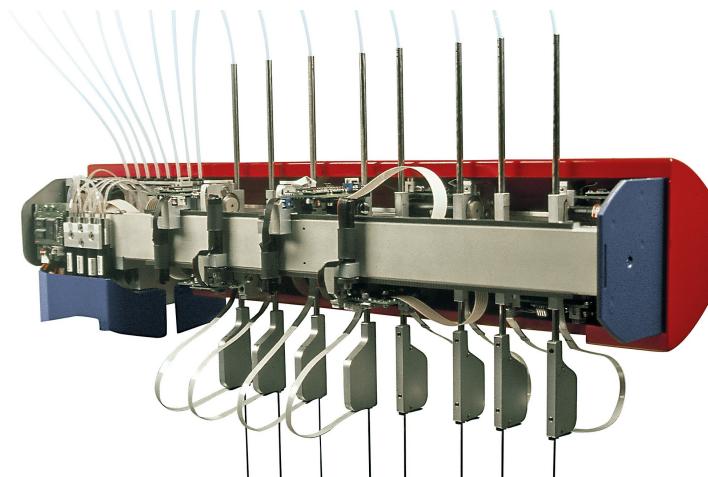


On the origins of data

- Chemical engineering is not an exception
 - Robotic high-throughput instrumentation
 - E.g. Parallel high-throughput solution phase synthesis for combinatorial chemistry and characterization



Tecan Xantus



Modular plug and play arms & GUI for experiment configuration

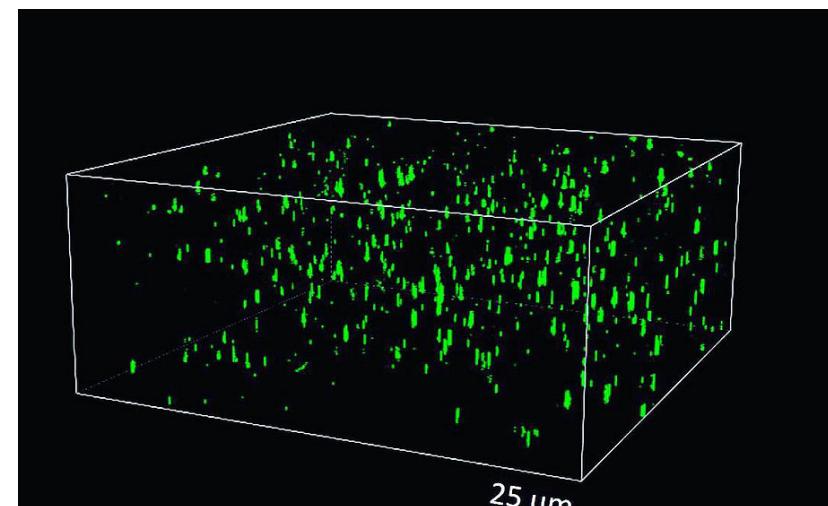
On the origins of data



- Chemical engineering is not an exception
 - High resolution imaging for particle tracking & 3D reconstruction



<http://www.nancelab.com>

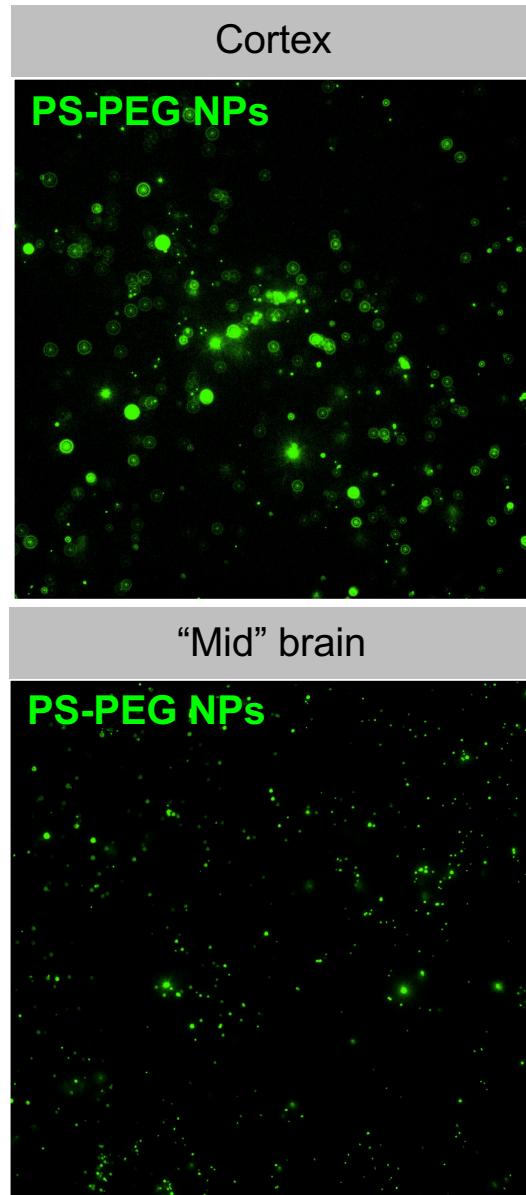
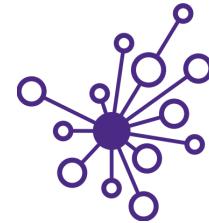


**ELIZABETH
NANCE**
Luce Assistant Professor

Understanding nanoparticle behavior in physiological environments with implications for therapeutic delivery

W

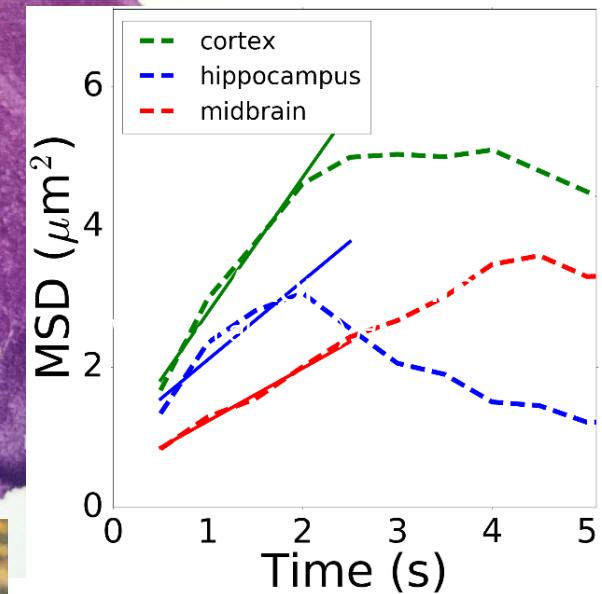
On the origins of data

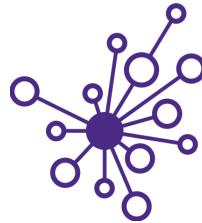


Not just static images!



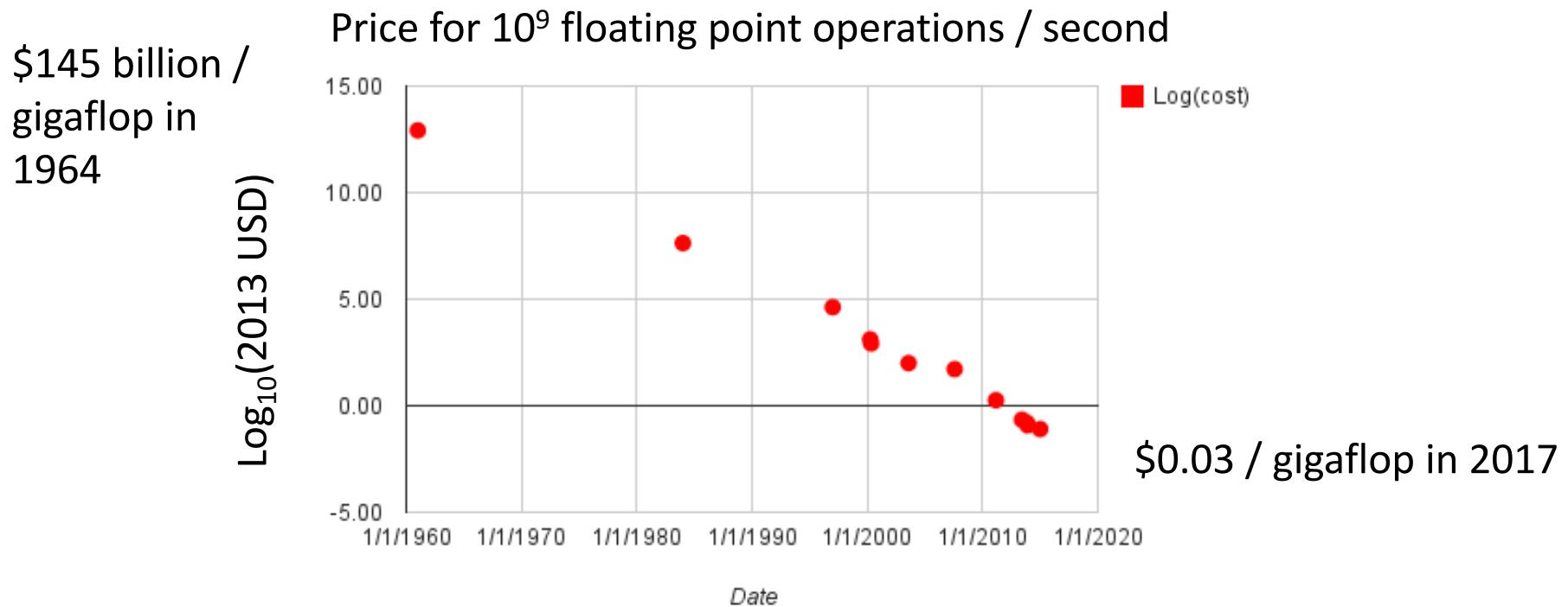
<http://www.nancelab.com>





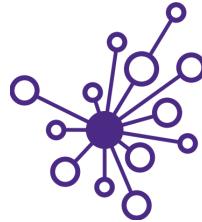
On the origins of data

- Chemical engineering is not an exception
 - Exponential decline in computing cost

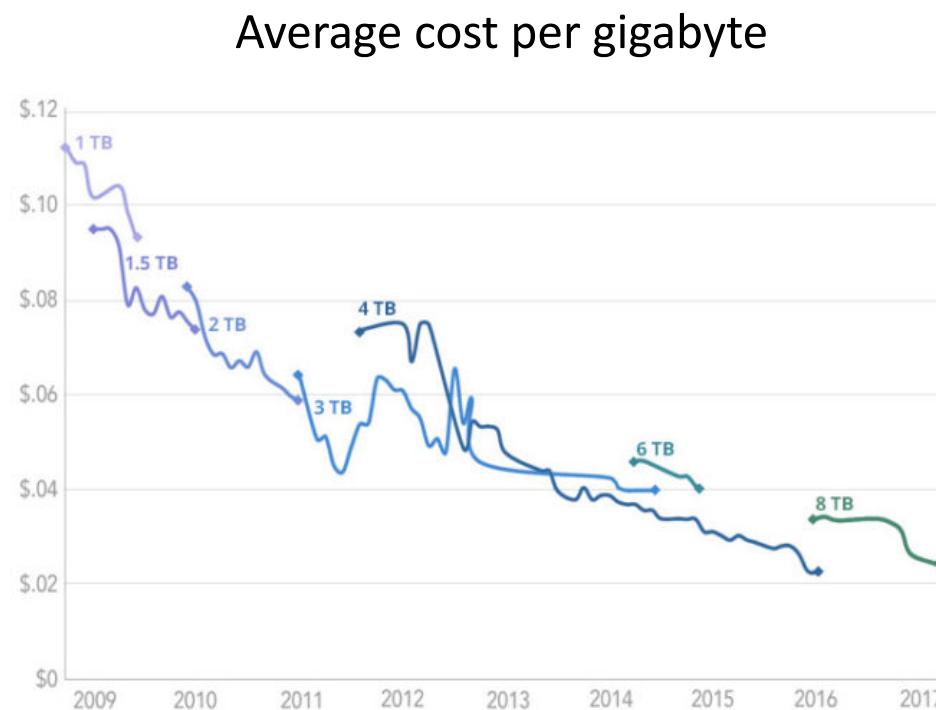


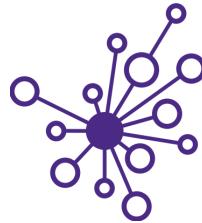
<https://aiimpacts.org/trends-in-the-cost-of-computing/>

On the origins of data



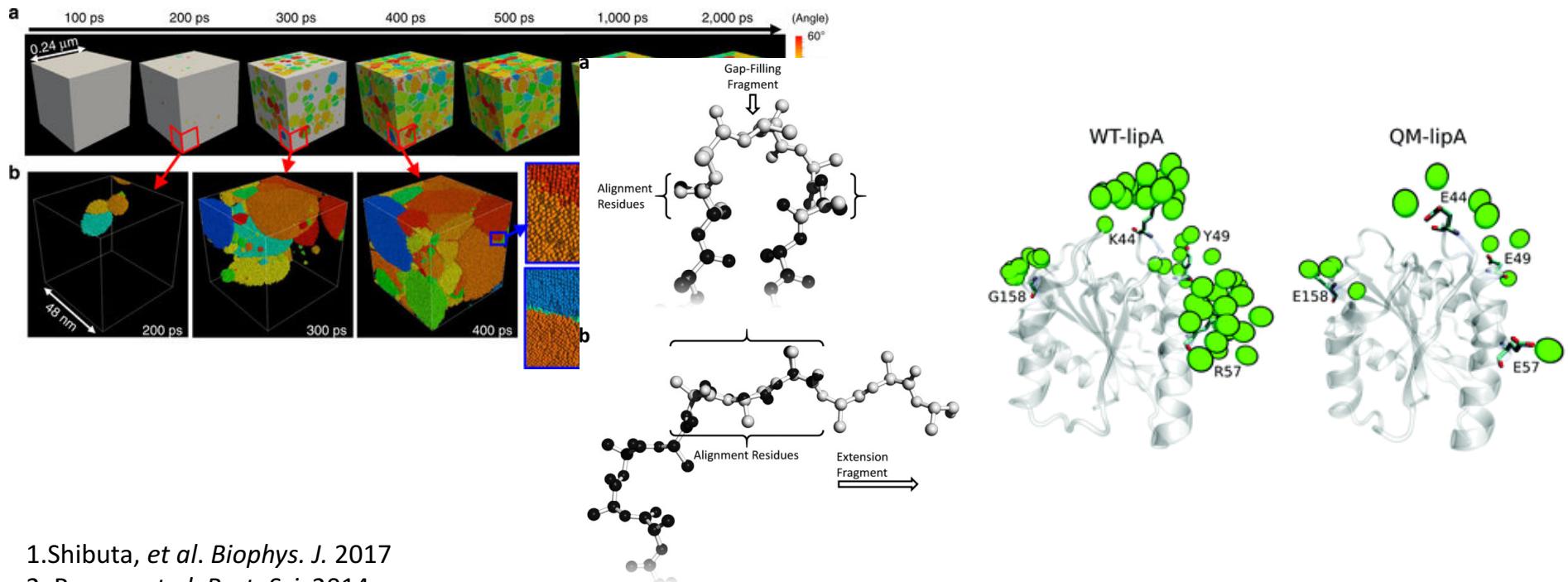
- Chemical engineering is not an exception
 - Substantial decline in storage cost





On the origins of data

- Chemical engineering is not an exception
 - Cheaper faster compute and storage resources
 - E.g. bigger¹, more², longer³ molecular simulations



1. Shibuta, et al. *Biophys. J.* 2017

2. Rysavy, et al. *Prot. Sci.* 2014.

3. Sprenger, et al. *Roy. Soc. Chem.* 2017.



On the origins of data

- Chemical engineering is not an exception
 - Synthetic and systems biology
 - E.g. high throughput gene sequencers¹, long read gene sequencers², ultra-cheap gene sequencers³



1. Illumina HiSeq
 10^{10} bases / day

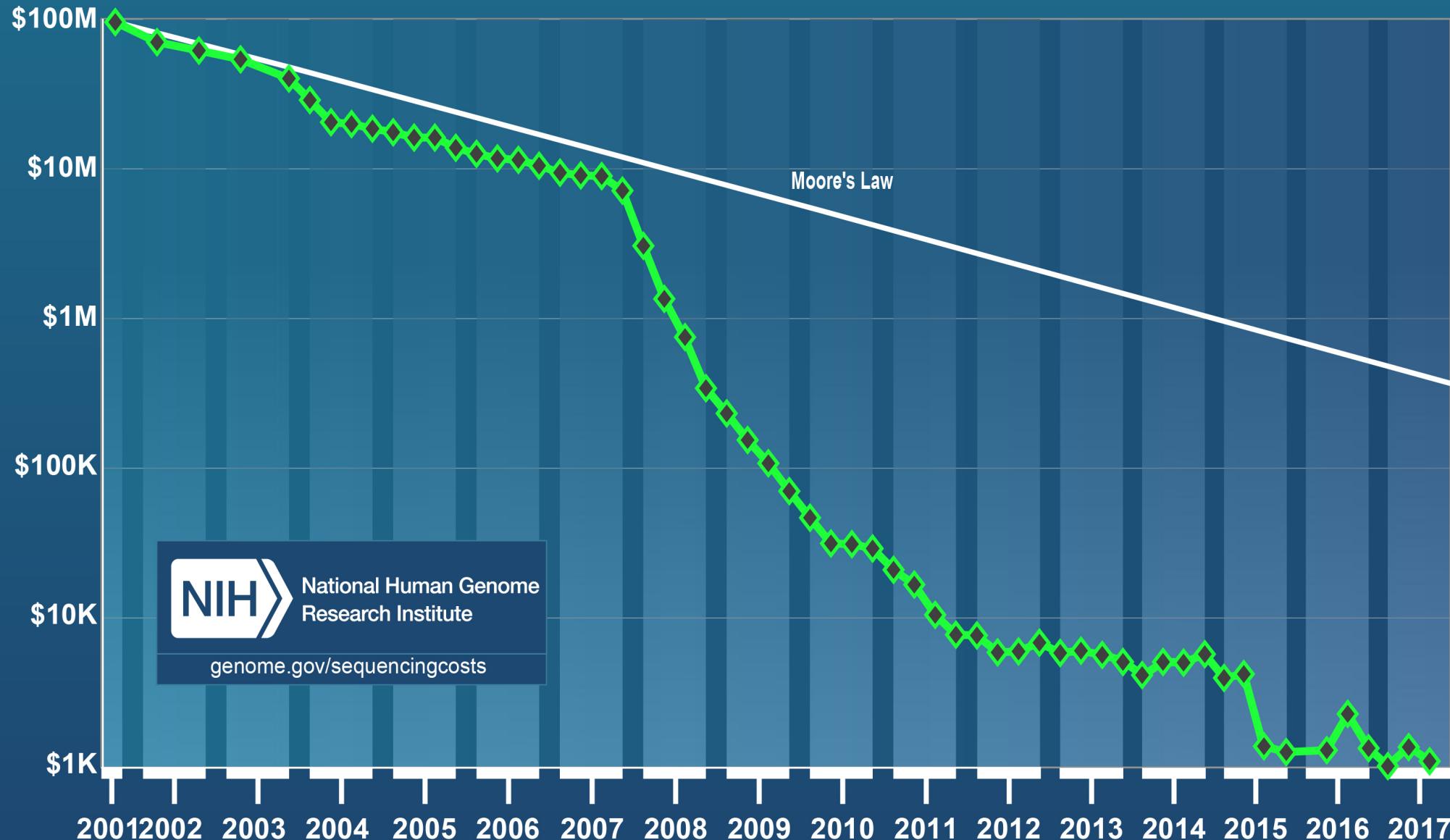


2. PacBio Sequel
 10^9 bases / day

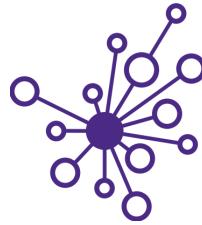


3. Oxford Nanopore MinION
~\$1000 USB powered

Cost per Genome



On the origins of data

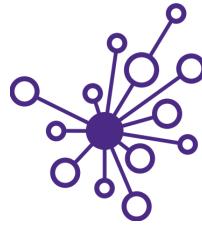


- Chemical engineering is not an exception
 - Industrial sensor networks & internet of things (IoT)
 - E.g. EU's public-private partnership RECOBA (BASF led)

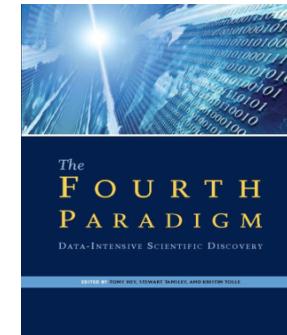


- Massive coordinated sensor networks with high volume data streams
- Online model predictive control
- Process optimization and cost reduction

Evolution of discovery



- Paradigm shifts in discovery
 - Empirical & experimental



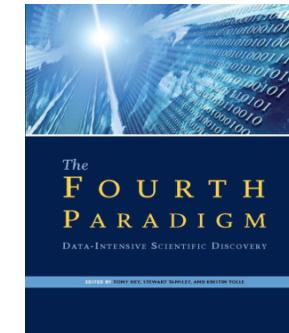
2009, MS



Evolution of discovery



- Paradigm shifts in discovery
 - Empirical & experimental



2009, MS

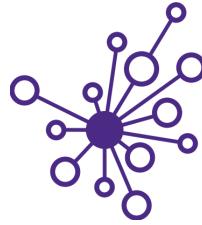


Dave w/ out
Van de Graff

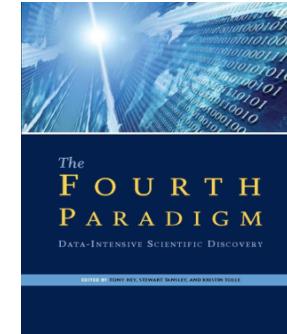
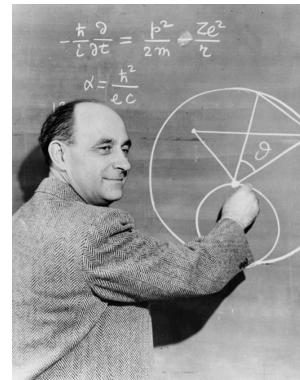


Dave w/
Van de Graff

Evolution of discovery

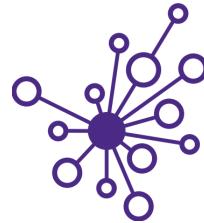


- Paradigm shifts in discovery
 - Empirical & experimental
 - Theoretical
 - Computational
 - Data-intensive



2009, MS

What makes this possible?



- What the new paradigm of data-intensive discovery and innovation?
 - Deep domain knowledge
 - Data Science
 - Data management
 - Databases, scalable data handling, data curation
 - Machine learning
 - Regression & classification
 - Supervised & unsupervised
 - Statistics
 - Visualization
 - Software engineering



Molecular Data Scientist

Knows thermodynamics **and** machine learning

Data management

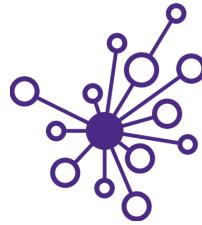


- Data management
 - You've got lots of data, how do you manage it?
 - Not about lab notebooks anymore!
 - Databases
 - Structure and store large, heterogeneous data
 - Slice, subset, and retrieve it efficiently
 - Track provenance and metadata of your data
 - Relational databases
 - Structured Query Language (SQL)

```
SELECT experiment FROM experiments WHERE  
experimenter = "Dave 'No Lab Skills' Beck";
```



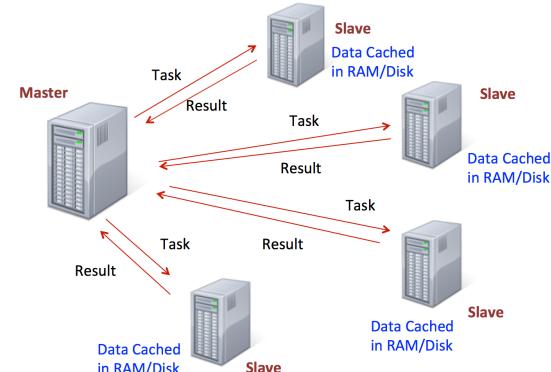
Data management



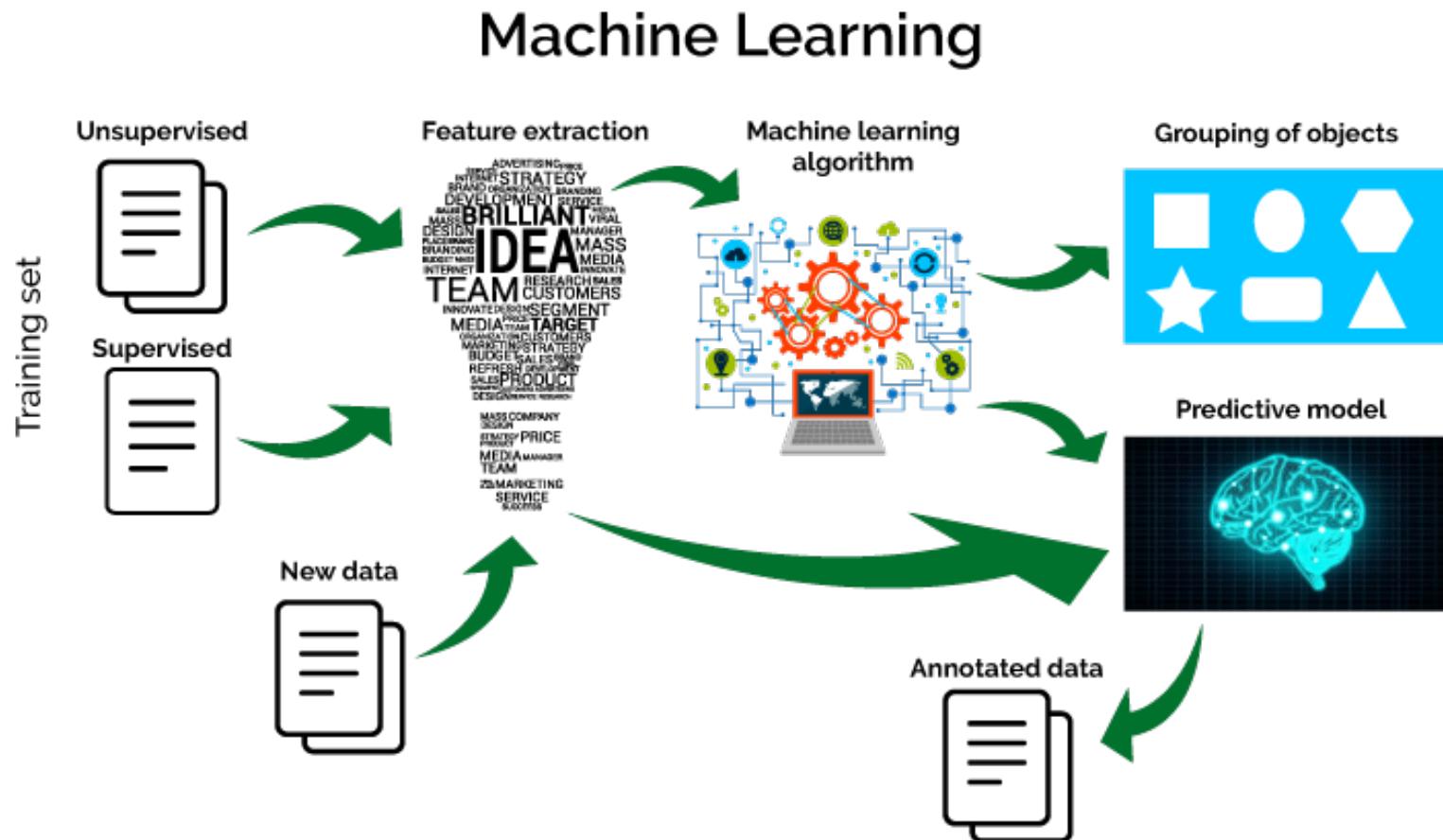
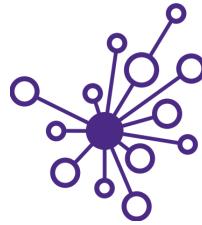
- Data management
 - Scalable data processing systems
 - Hadoop & MapReduce
 - Apache Spark
 - SQL libraries
 - Machine learning libraries
 - Distribute your data and workload over lots of machines
 - Fault tolerance, scalability

40,000 + nodes (Hadoop)

8,000 + nodes (Spark)

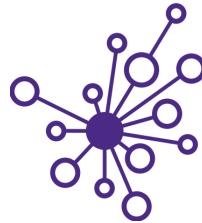


Machine learning





Machine learning

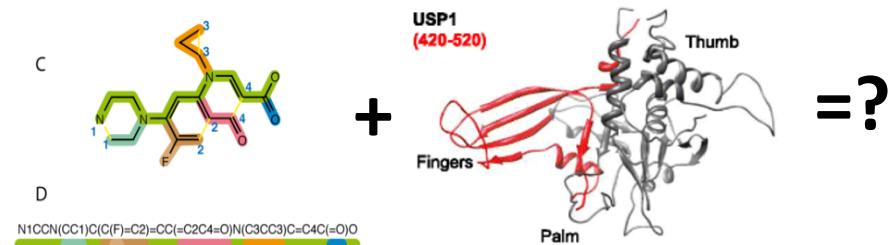


- Regression
 - Predict a numerical response from input features

Machine learning



- Regression
 - E.g. predict binding affinity of small molecules to cancer drug target



Database of 400,000 drug like molecules

Experimental inhibition activities against cancer drug target

Build a regression model that relates molecular features to a numerical measure of inhibition, dissociation constant (Kd)

For a new small molecule, predict the Kd^{1,2}

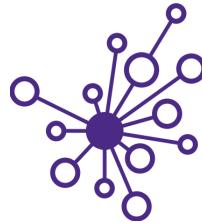
1. <https://github.com/BeckResearchLab/USP-inhibition>
2. <https://github.com/BeckResearchLab/small-molecule-design-toolkit>



Pearl Philip & Rahul Avadhoot



Machine learning



- Regression
 - Predict molecular property or activity from molecular structure

Quantitative Structure Property Relationship (QSPR)

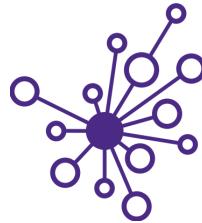
Quantitative Structure Activity Relationship (QSAR)



Machine learning

- Regression

Molecule	Features or predictors					Outcome
	Diffusion (nm/s)	Molecular Weight	C=C bonds	H-bond donors	H-bond acceptors	
1	8	66	1	0	0	1
2	5	44	2	1	0	12
3	6	95	0	2	0	14
4	7	63	4	0	0	1
5	8	65	1	0	1	4
6	3	91	1	1	1	3
7	6	94	2	0	1	2
8	3	96	1	2	1	2
9	7	57	3	1	1	6
10	6	121	1	0	0	8



Machine learning

- Regression

Molecule	Features or predictors					Outcome
	Diffusion (nm/s)	Molecular Weight	C=C bonds	H-bond donors	H-bond acceptors	
1	8	66	1	0	0	1
2	5	44	2	1	0	12
3	6	95	0	2	0	14
4	7	63	4	0	0	1
5	8	65	1	0	1	4
6	3	91	1	1	1	3
7	6	94	2	0	1	2
8	3	96	1	2	1	2
9	7	57	3	1	1	6
10	6	121	1	0	0	8
11	4	98	2	1	1	?????????

New!

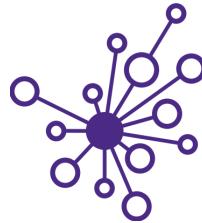


Machine learning

- Regression
 - Linear regression
 - LASSO regression (least absolute shrinkage and selection operator)
 - Variable selection (which features are actually useful)
 - Regularization (avoid overfitting to your training data)

$$f(X) = \beta_0 + \sum_{j=1}^p X_j \beta_j.$$

Molecule	Diffusion (nm/s)	Molecular Weight	C=C bonds	H-bond donors	H-bond acceptors	Kd (fM)
1	8	66	1	0	0	1
2	5	44	2	1	0	12
3	6	95	0	2	0	14
4	7	63	4	0	0	1
5	8	65	1	0	1	4
6	3	91	1	1	1	3
7	6	94	2	0	1	2
8	3	96	1	2	1	2
9	7	57	3	1	1	6
10	6	121	1	0	0	8

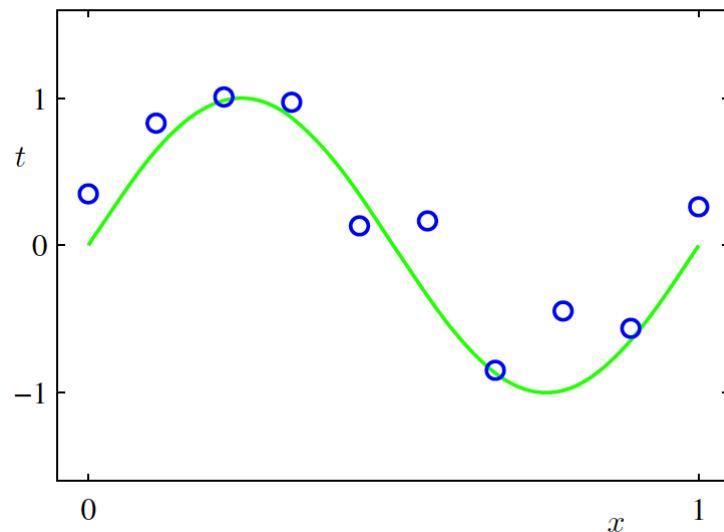


Machine learning

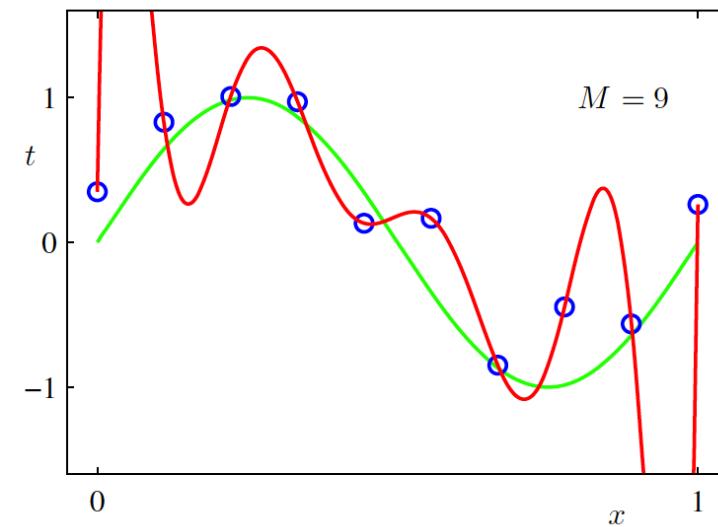
- Overfitting

“With four parameters I can fit an elephant, and with five I can make him wiggle his trunk.”

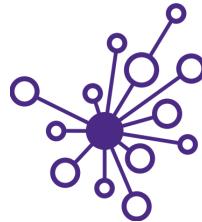
- John von Neumann



Points generated from green
function $f(x) = \sin(2\pi x) + \text{noise}$

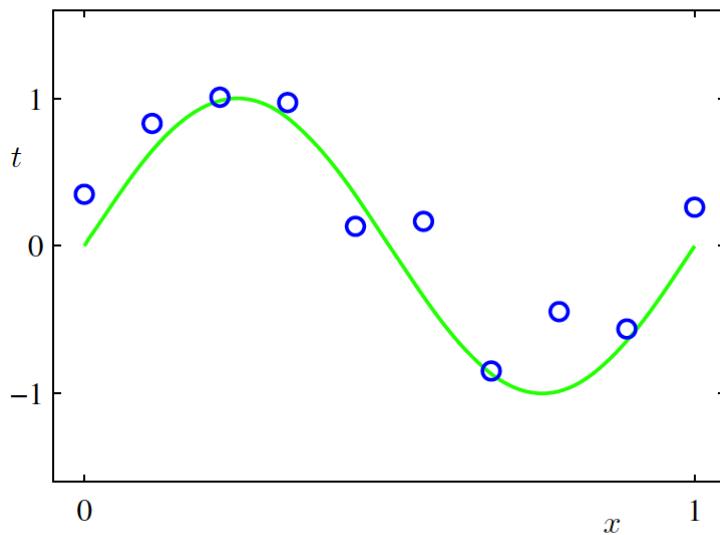


Fitting to points with a
polynomial with order $M = 9$

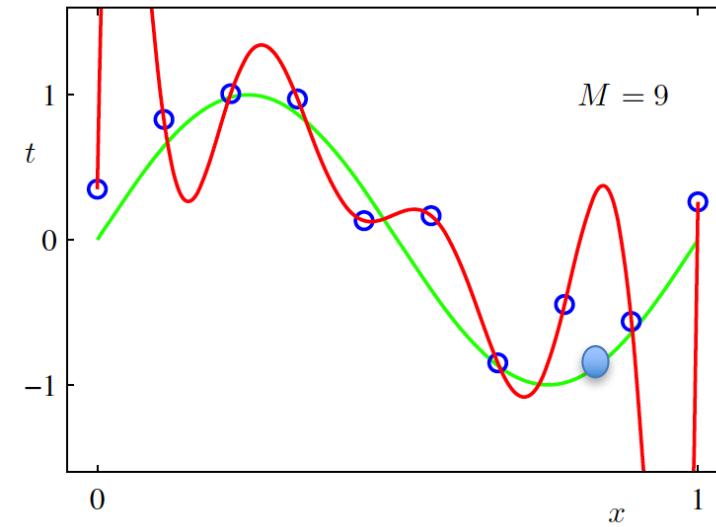


Machine learning

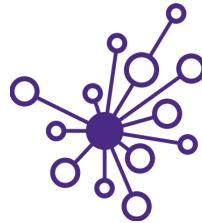
- Overfitting
 - Making your model too specific to training data
 - Performs poorly on new data relative to "truth"



Points generated from green
function $f(x) = \sin(2\pi x) + \text{noise}$

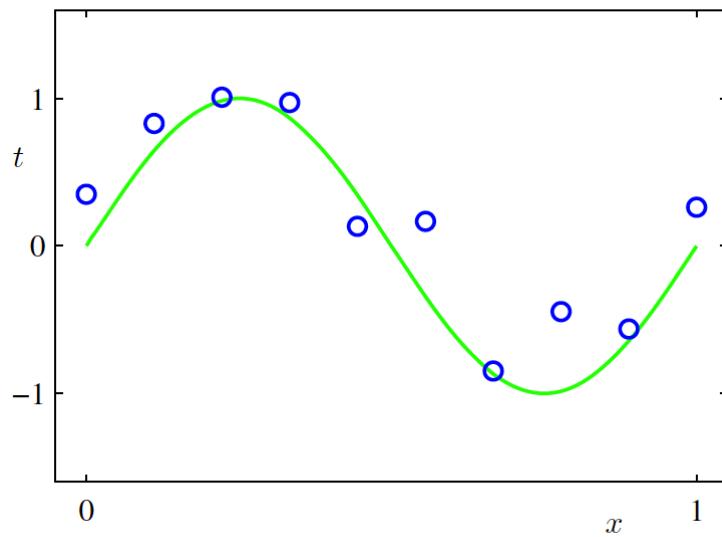
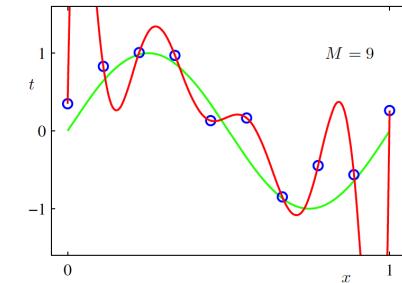


Fitting to points with a
polynomial with order $M = 9$

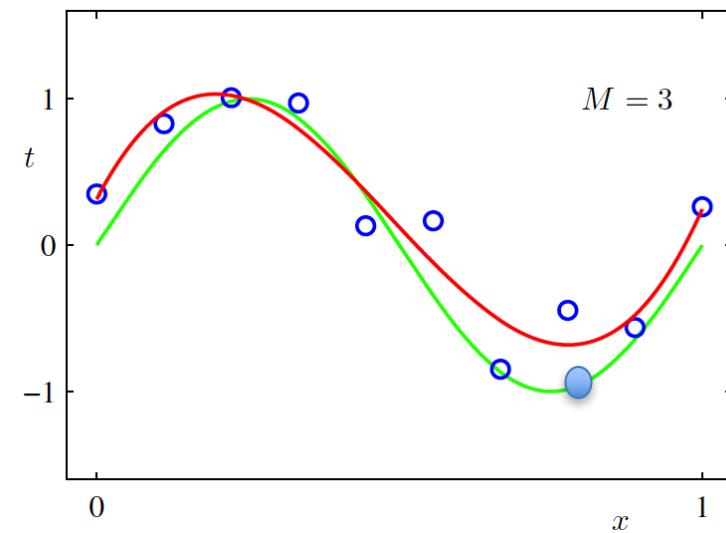


Machine learning

- Proper fitting
 - Still fits the training data
 - Performs better on new data

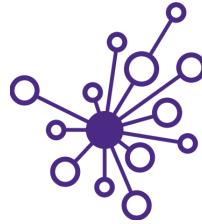


Points generated from green function $f(x) = \sin(2\pi x) + \text{noise}$

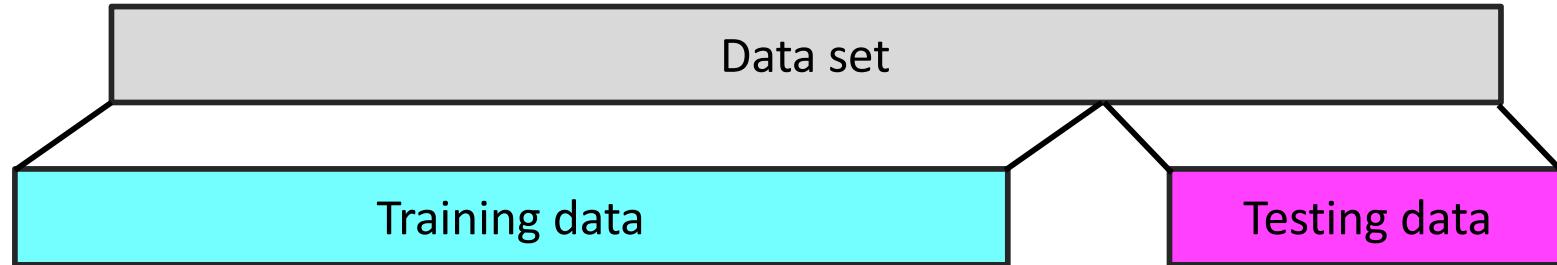


Fitting to points with a polynomial with order $M = 3$

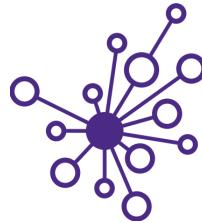
Machine learning



- Train / test split
 - How can you identify overfitting?
 - Partition your input data into
 - Training set (e.g. 80%) used to build the ML model
 - Test set (e.g. 20%) used to validate and characterize the error in the model



- **Never ever ever ever ever** contaminate your model training with data from the test set



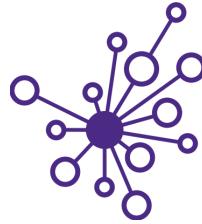
Machine learning

- Regression
 - Linear regression
 - LASSO regression (least absolute shrinkage and selection operator)
 - Variable selection (which features are actually useful)
 - Regularization (avoid overfitting to your training data)

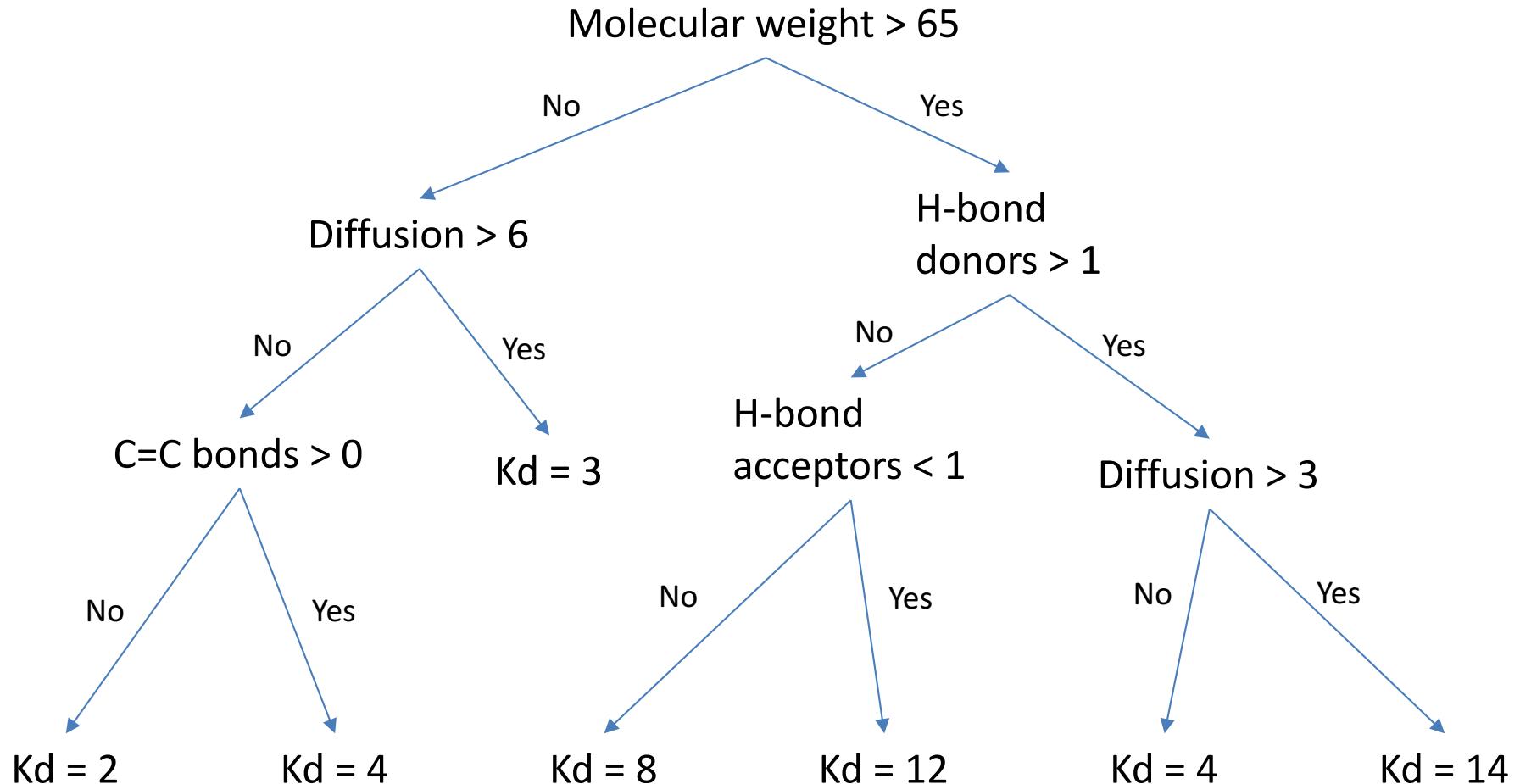
$$f(X) = \beta_0 + \sum_{j=1}^p X_j \beta_j.$$

Molecule	Diffusion (nm/s)	Molecular Weight	C=C bonds	H-bond donors	H-bond acceptors	Kd (fM)
1	8	66	1	0	0	1
2	5	44	2	1	0	12
3	6	95	0	2	0	14
4	7	63	4	0	0	1
5	8	65	1	0	1	4
6	3	91	1	1	1	3
7	6	94	2	0	1	2
8	3	96	1	2	1	2
9	7	57	3	1	1	6
10	6	121	1	0	0	8

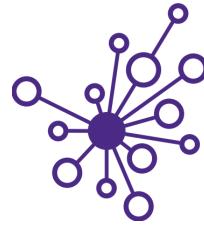
Machine learning



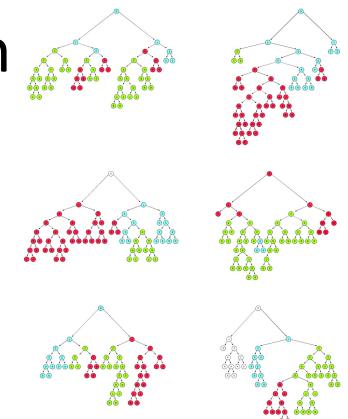
- Regression
 - Decision trees

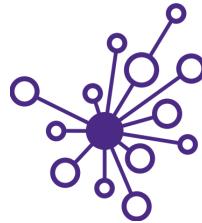


Machine learning



- Regression
 - Multiple trees can describe the data equally well
 - A single deep tree can overfit to training data
 - Ensemble learning
 - Combining several weak learners to get a strong learner
 - Random forests (lots of different decisions trees)
 - Predicted value is mean or mode or median

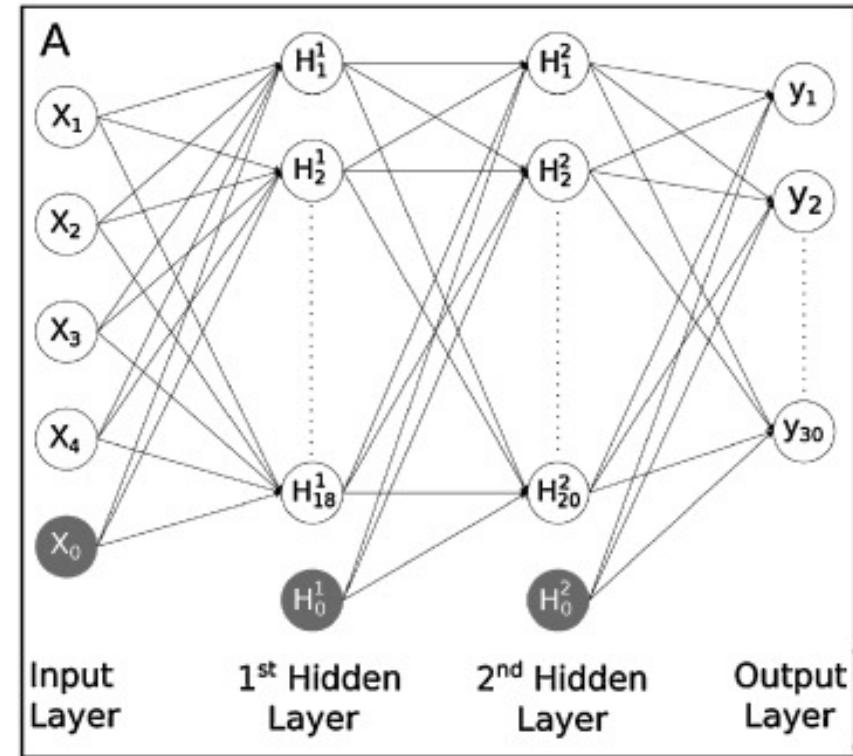
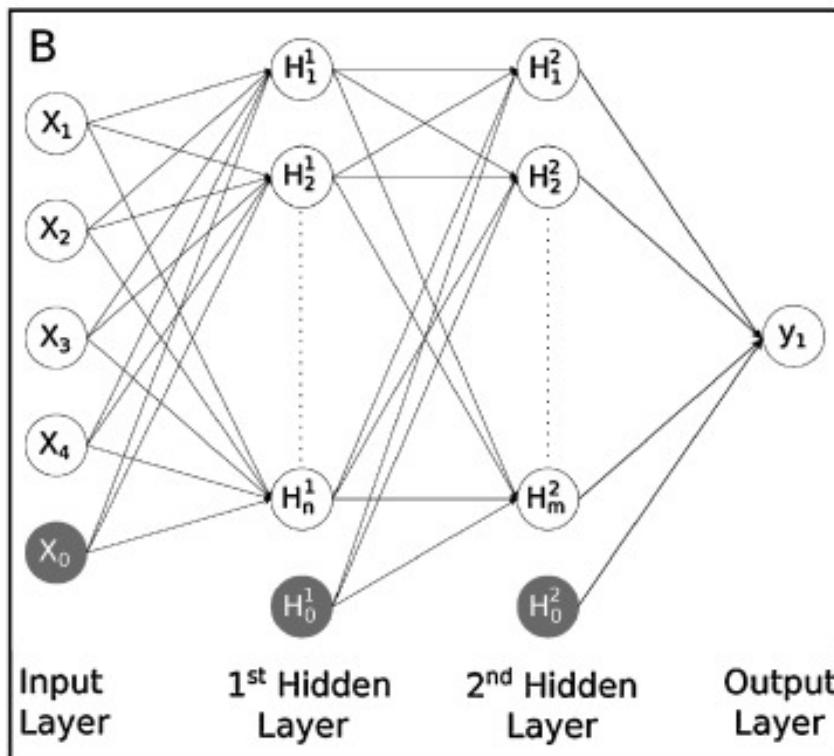




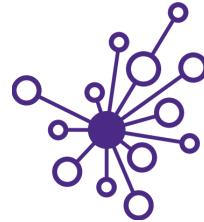
Machine learning

- Regression
 - Artificial neural networks

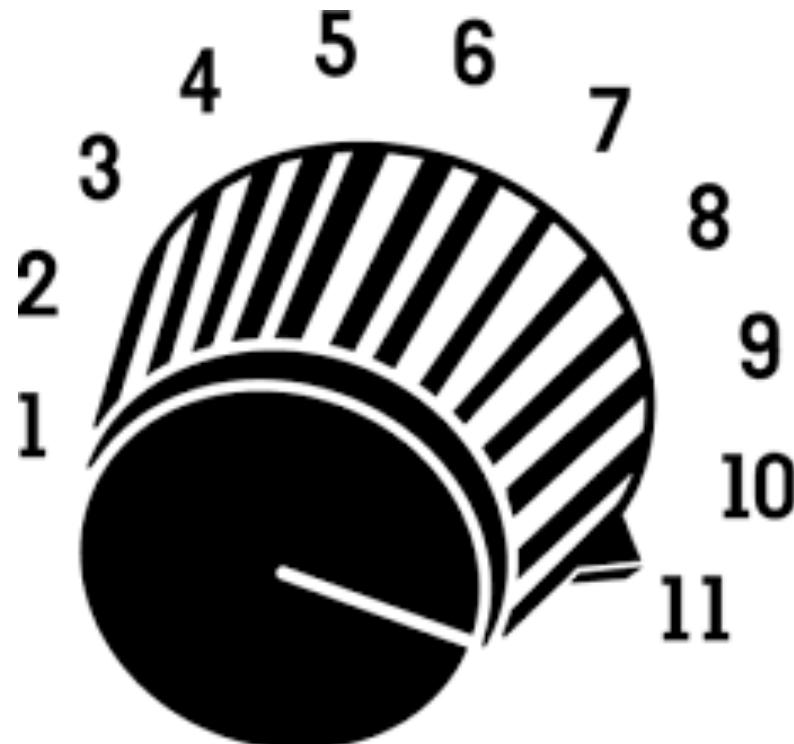
$$y=f(x)$$



Machine learning



- Regression
 - Artificial neural networks

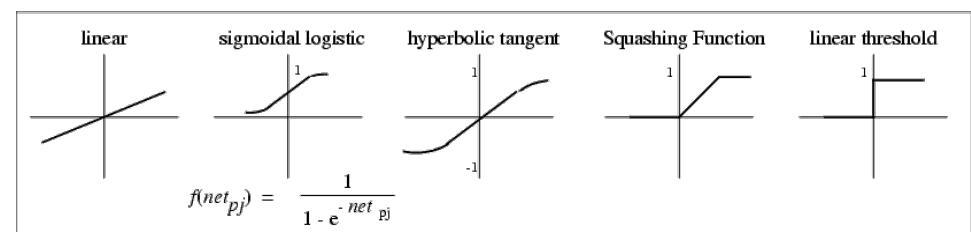


Nodes connect to successive layers via weighted transfer or activation function

Learning occurs by tuning weights (w) and parameters of transfer functions

Lots of parameters in neural networks:

- Number of layers
- Number of nodes in each layer
- Transfer function of each layer



Machine learning



- Regression
 - Artificial neural networks
 - E.g. Reducing solution time for lignocellulosic biomass pyrolysis kinetic model

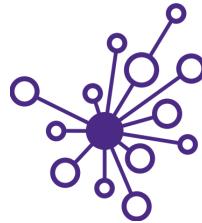
Original kinetic model had

- Inputs like max pyrolysis T, heating rate, mass fractions of C & H in feedstock
- ~100 chemical species
- ~400 reactions
- 30 real valued outputs, e.g.
 - Yields of light & heavy oil
 - Distribution of C functional groups in heavy oil fraction



Solving complete set of stiff ODEs takes nearly 5 seconds

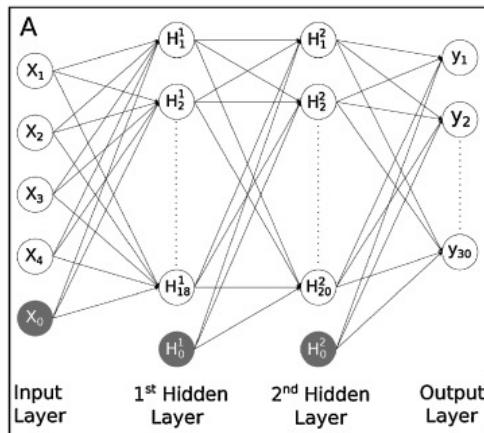
Blake Hough
UW ChemE PhD
Data Scientist,
EnergySavvy



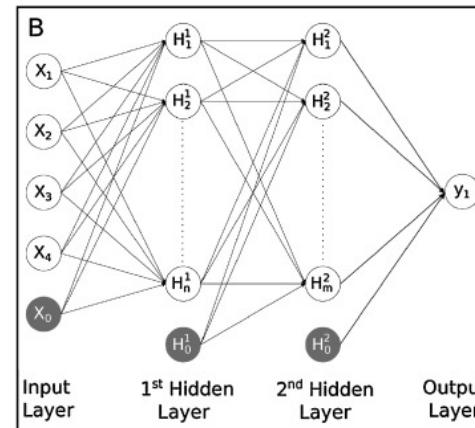
Machine learning

- Regression
 - Artificial neural networks
 - E.g. Reducing solution time for lignocellulosic biomass pyrolysis kinetic model
 - Ran the kinetic model 250,000 times varying inputs across their valid range
 - Train/test split: 200,000 for training, 50,000 for testing

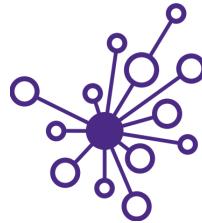
One model w/ 30 output



30 models w/ 1 output



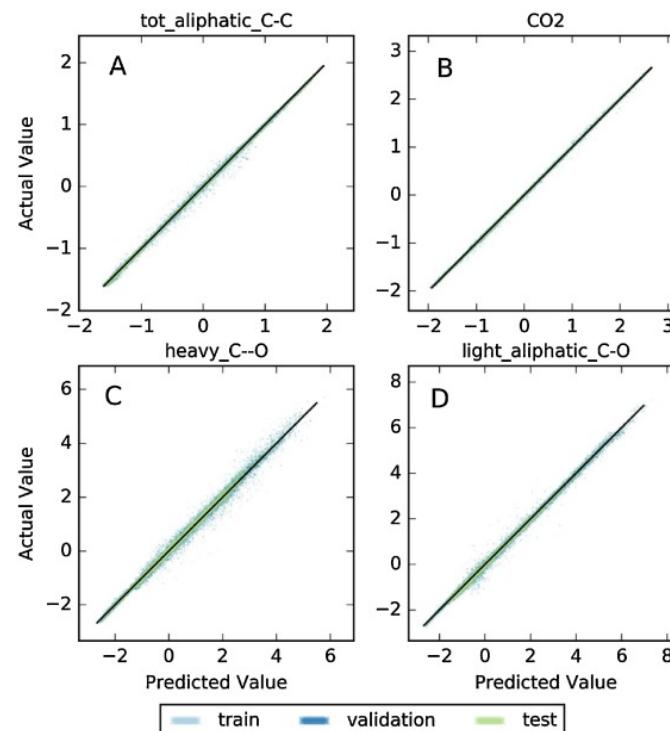
Blake Hough
UW ChemE PhD
Data Scientist,
EnergySavvy



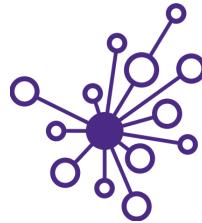
Machine learning

- Regression
 - Artificial neural networks
 - E.g. Reducing solution time for lignocellulosic biomass pyrolysis kinetic model

$R^2 > .98$
across all
outputs



Blake Hough
UW ChemE PhD
Data Scientist,
EnergySavvy



Machine learning

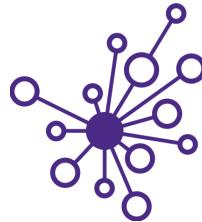
- Regression
 - Artificial neural networks
 - E.g. Reducing solution time for lignocellulosic biomass pyrolysis kinetic model

Table 2. Benchmarking results for solving the kinetic model. The time reported is the average of 1000 model calls.

Kinetic model format	Average code execution time (s)
Decision tree	1.106×10^{-4}
Full net	1.690×10^{-4}
Single net	1.747×10^{-4}
30 single nets (run serially)	4.236×10^{-3}
Complete ODE model(B. R. Hough et al., 2016)	4.725



Blake Hough
UW ChemE PhD
Data Scientist,
EnergySavvy



Machine learning

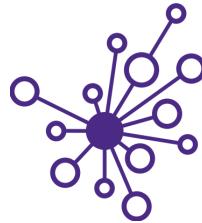
- Regression
 - Artificial neural networks
 - E.g. Reducing solution time for lignocellulosic biomass pyrolysis kinetic model

Table 2. Benchmarking results for solving the kinetic model. The time reported is the average of 1000 model calls.

Kinetic model format	Average code execution time (s)
Decision tree	1.106×10^{-4}
Full net	1.690×10^{-4}
Single net	1.747×10^{-4}
30 single nets (run serially)	4.236×10^{-3}
Complete ODE model(B. R. Hough et al., 2016)	4.725



Blake Hough
UW ChemE PhD
Data Scientist,
EnergySavvy



Machine learning

- Regression
 - Artificial neural networks
 - E.g. Reducing solution time for lignocellulosic biomass pyrolysis kinetic model

Table 2. Benchmarking results for solving the kinetic model. The time reported is the average of 1000 model calls.

Kinetic model format	Average code execution time (s)
Decision tree	1.106×10^{-4}
Full net	1.690×10^{-4}
Single net	1.747×10^{-4}
30 single nets (run serially)	4.236×10^{-3}
Complete ODE model(B. R. Hough et al., 2016)	4.725



Blake Hough
UW ChemE PhD
Data Scientist,
EnergySavvy

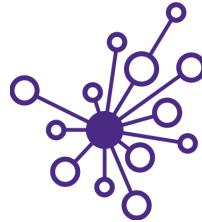


Machine learning



- Regression
 - Artificial neural networks
 - Keras
 - Tensorflow
 - Theano

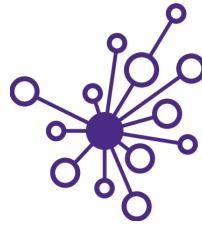
Machine learning



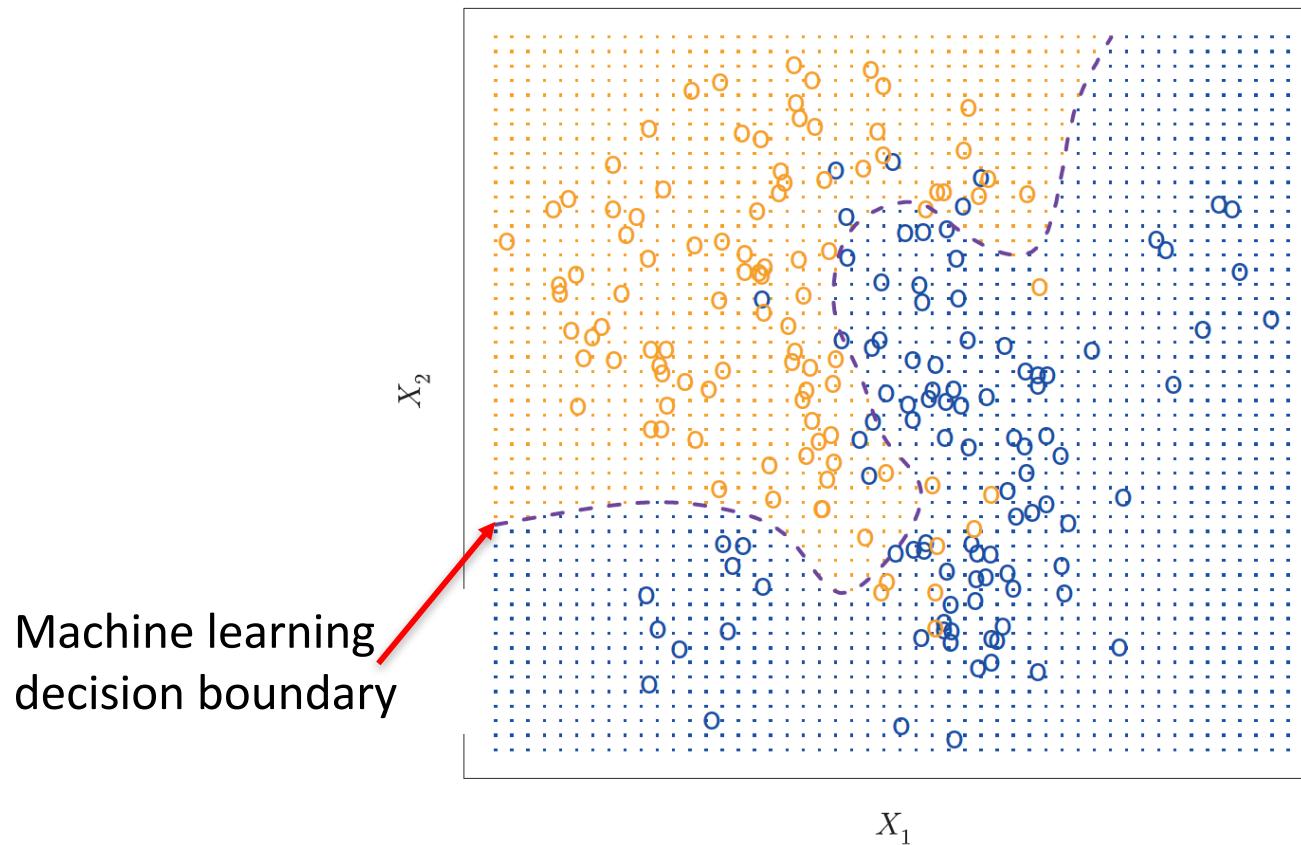
- Regression
 - Ridge regression
 - Support vector regression
 - ElasticNet
 - Least-angle regression (LARS)
 - Bayesian Regression
 - ...
 - An Introduction to Statistical Learning (free PDF)
 - Pattern Recognition and Machine Learning



Machine learning



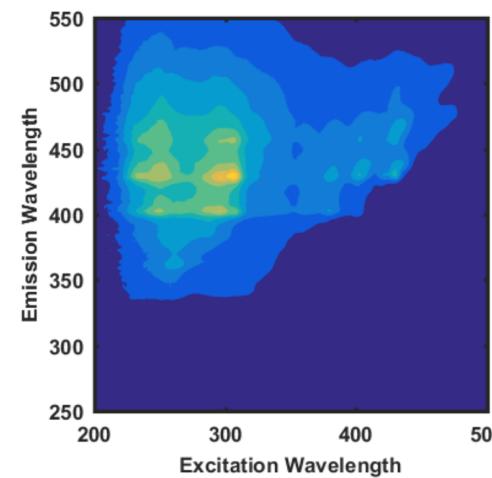
- Classification
 - What state or class does a sample belong to?





Machine learning

- Classification
 - What state or class does a sample belong to?
 - E.g. Source identification or atmospheric particulate matter (smoke)
- Filter air
- Extract the residue from the filter
- Fluorescent Excitation Emission Spectroscopy (EEM)



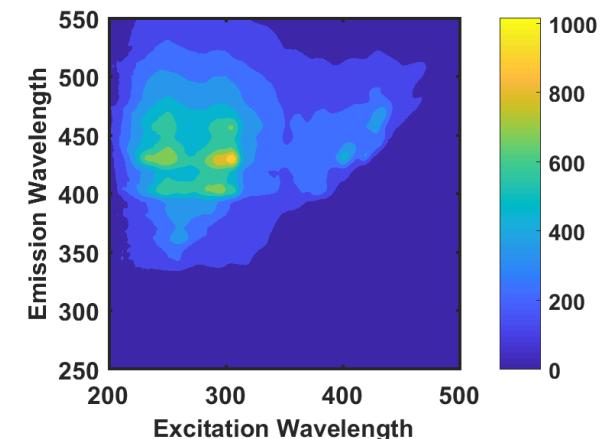
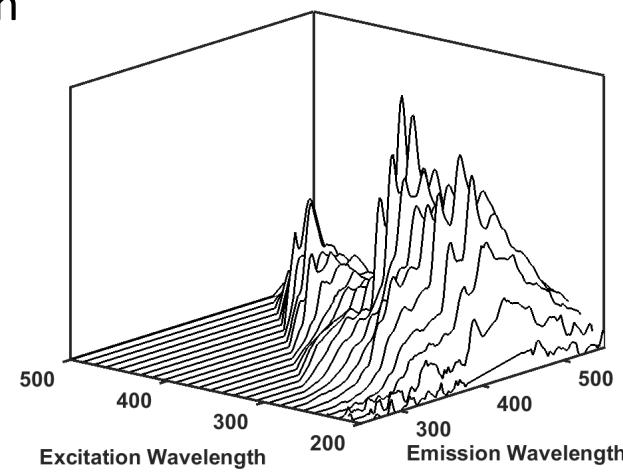
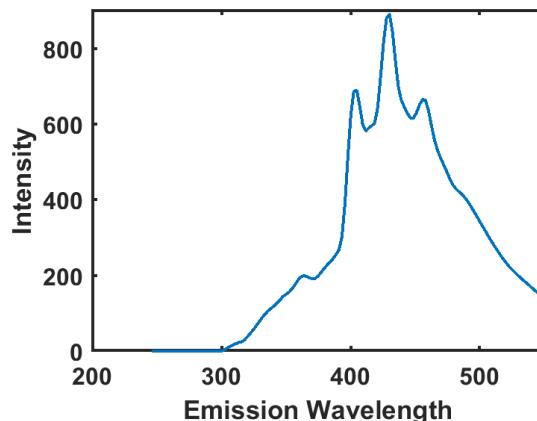
Jay Rutherford
(Posner Lab)



Machine learning

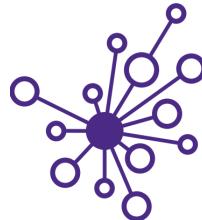
- Fluorescent Excitation Emission Spectroscopy (EEM)
 - Spectra collected at multiple excitation wavelengths
 - Combined into an “Excitation Emission Matrix”

Single excitation wavelength



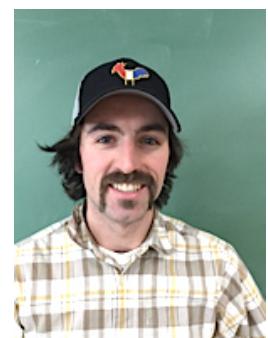
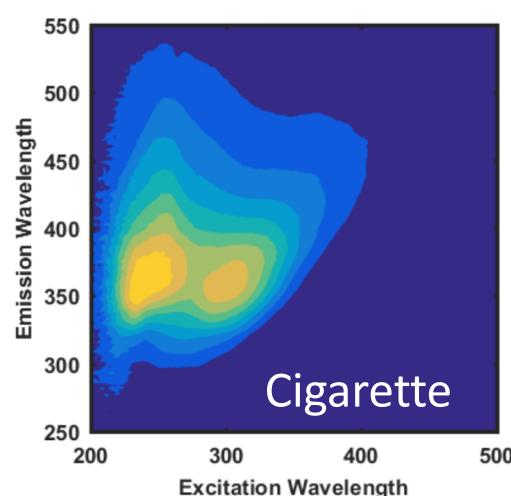
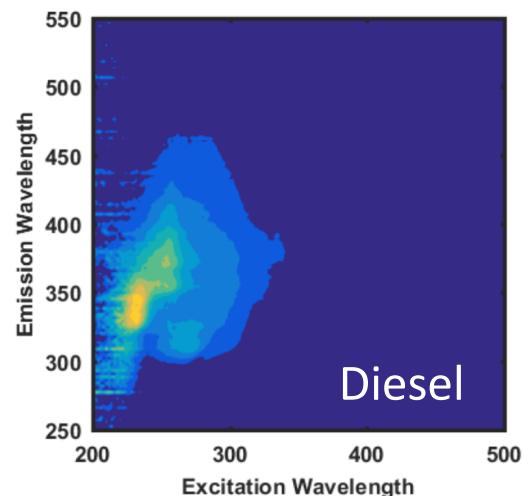
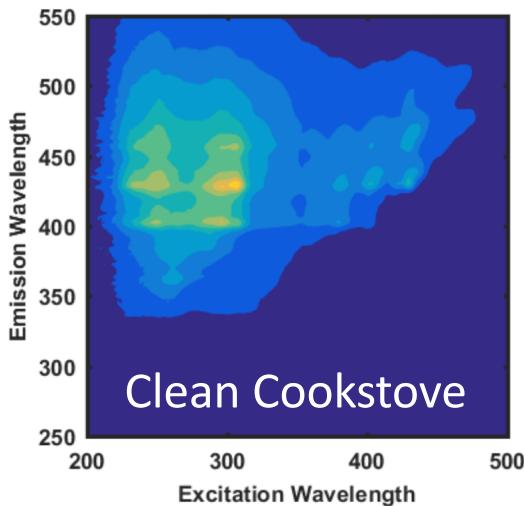
W

Ma

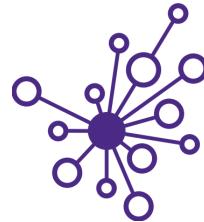


uq

- Classification
 - What state or class does a sample belong to?

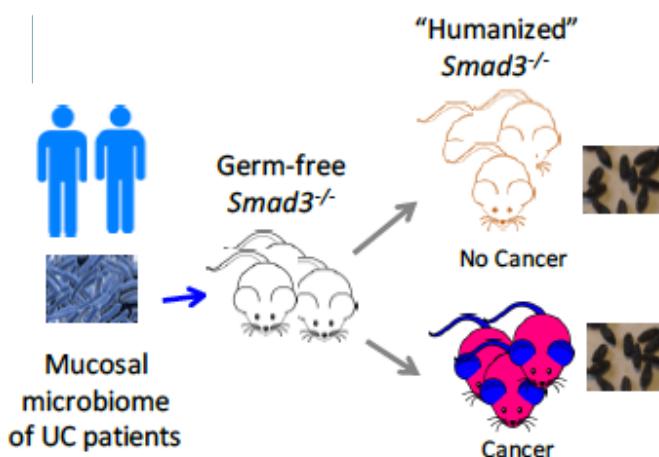


Jay Rutherford
(Posner Lab)

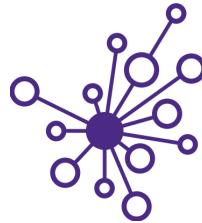


Machine learning

- Classification
 - What state or class does a sample belong to?
 - Ulcerative colitis (UC) & colon cancer diagnostic
 - Take gut microbiome sample from UC patient
 - Inoculate gnotobiotic mice with bacteria
 - See if mice get cancer



16+ weeks!

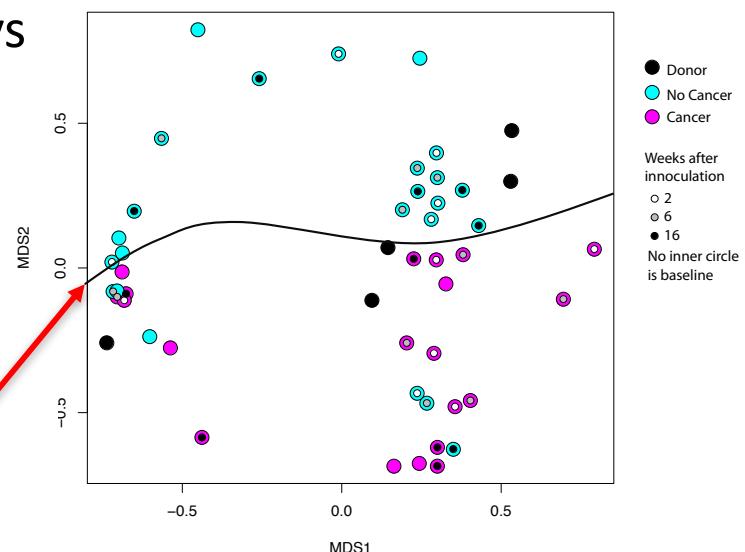


Machine learning

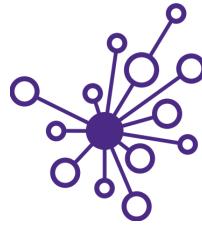
- Classification
 - What state or class does a sample belong to?
 - Rapid ulcerative colitis (UC) & colon cancer
 - Build a model of that relates the gut microbiome structure to cancer likelihood using the gnotobiotic mice data
 - Examine the microbiome of the mucosal sample directly
 - Make cancer assessment in days



Machine learning
decision boundary



Machine learning



- Classification
 - What state or class does a sample belong to?
 - Rapid ulcerative colitis (UC) & colon cancer

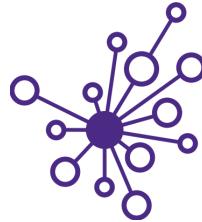


Ultra-rapid, super cheap, metagenome sequencing + Data Science =





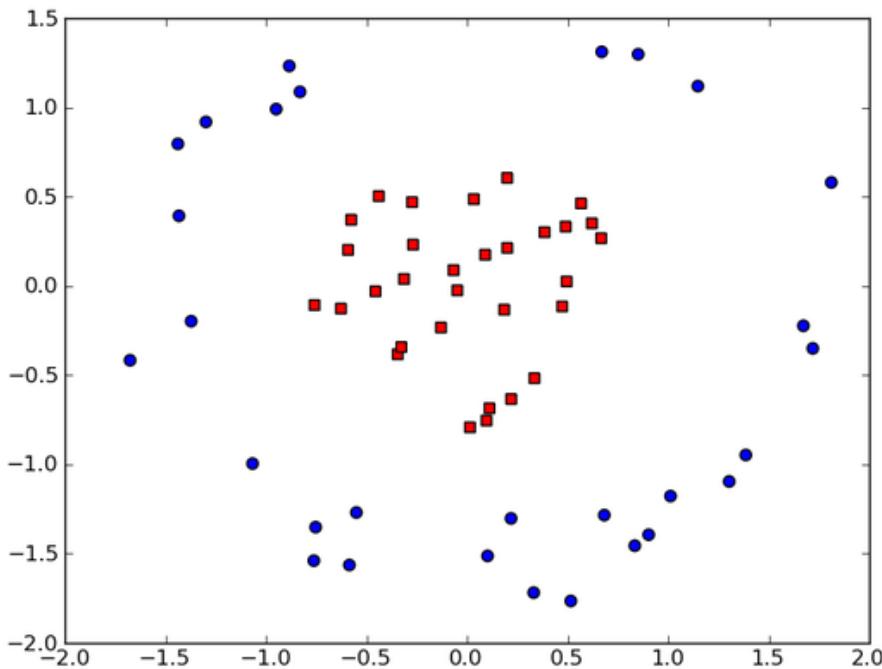
Machine learning



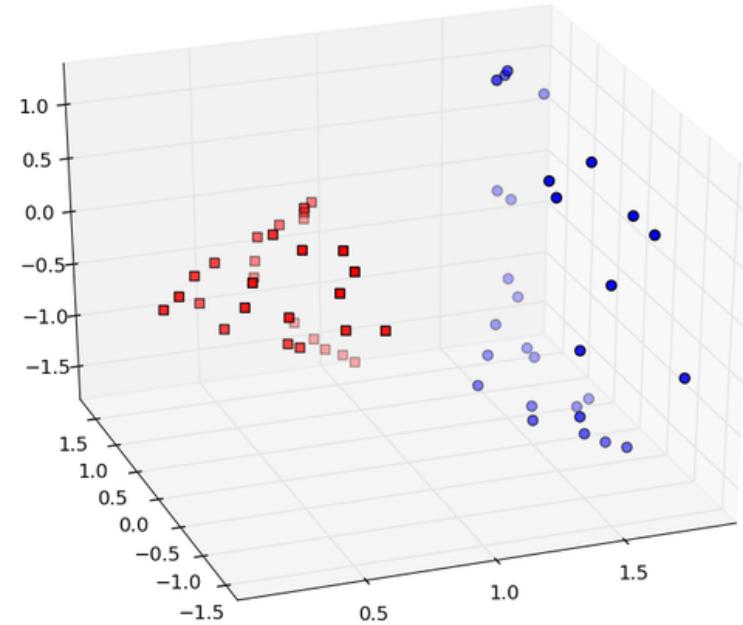
- Classification
 - What state or class does a sample belong to?
 - Methodological approaches:
 - Naïve Bayes Classifier
 - Probabilistic classifier that assumes conditional independence
 - Support vector machines
 - Linear discriminant analysis
 - Neural networks



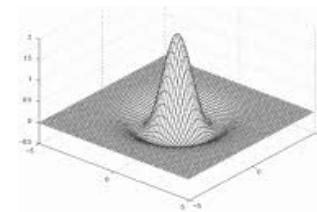
Machine learning



Original feature space

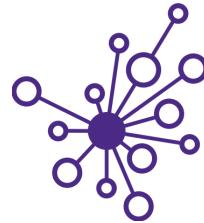


Transformed feature space

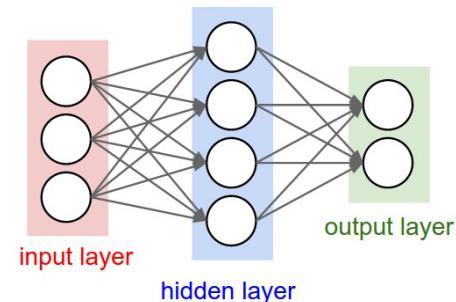


Now we can easily find a plane that separates the points!

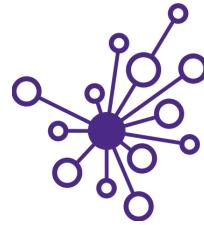
Machine learning



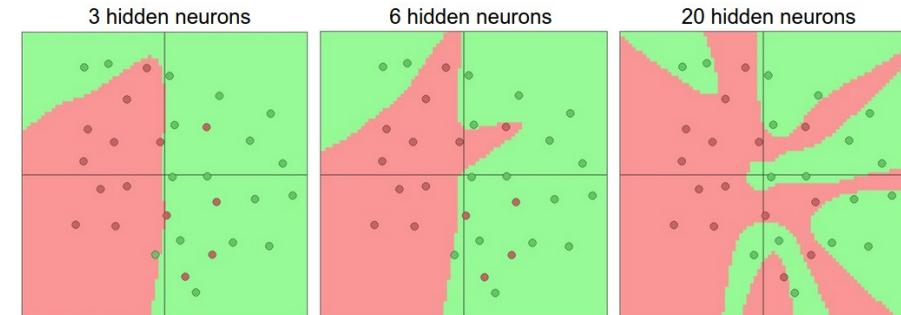
- Classification
 - What state or class does a sample belong to?
 - Methodological approaches:
 - Naïve Bayes Classifier
 - Probabilistic classifier that assumes conditional independence
 - Logistic regression
 - Support vector machines
 - Linear discriminant analysis
 - Neural networks

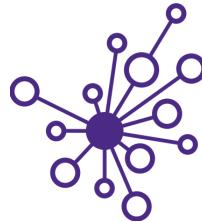


Machine learning



- Classification
 - What state or class does a sample belong to?
 - Methodological approaches:
 - Naïve Bayes Classifier
 - Probabilistic classifier that assumes conditional independence
 - Logistic regression
 - Support vector machines
 - Linear discriminant analysis
 - Neural networks



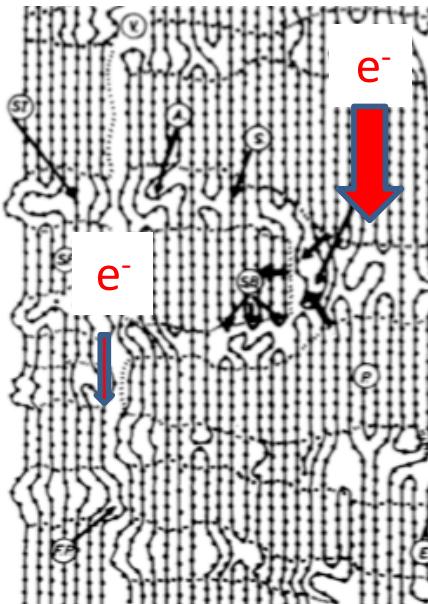


Regression + classification

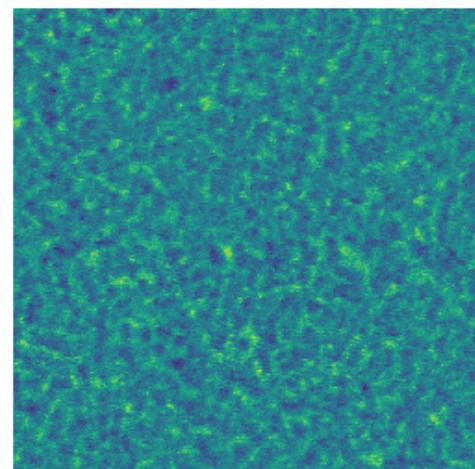
- Sometimes you really need it all
 - E.g. Optimize annealing parameters in creation of polymer thin films

Morphological Analysis of Nanostructured Thin-films ([MANA-T](#))

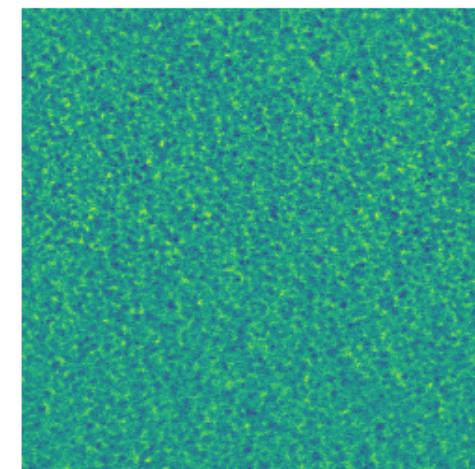
Morphology of
Polymer Thin Films



Atomic Force Microscopy



Control



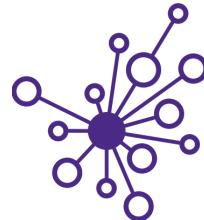
Annealed



Wes Tatum
(Luscombe Lab)



Regression + classification



Adhesion

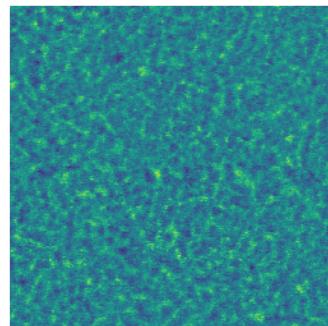
PixelClassifier

EuclideanClassifier

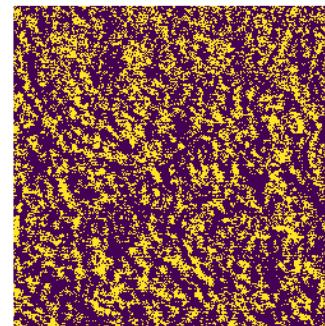
- Sometimes you really need it all
 - E.g. Optimize annealing parameters in creation of polymer thin films

Control
Sample

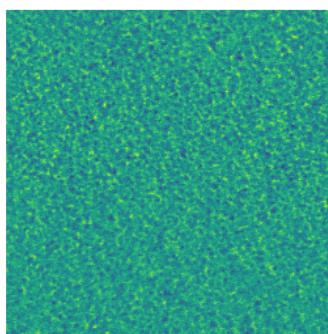
AFM image



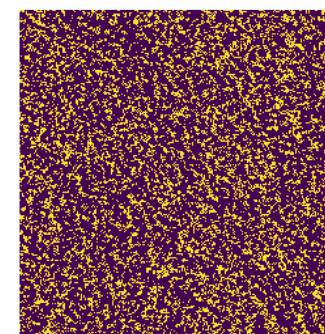
PixelClassifier



Annealed
Sample



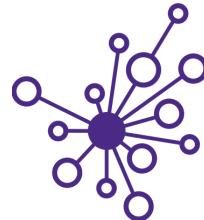
Yellow = Crystalline
Purple = Amorphous



Wes Tatum
(Luscombe Lab)



Regression + classification



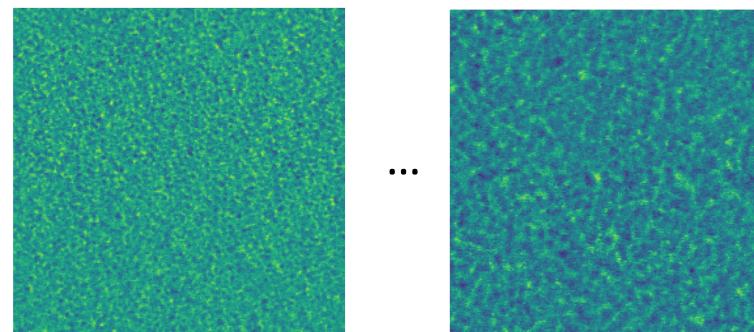
Adhesion

PixelClassifier

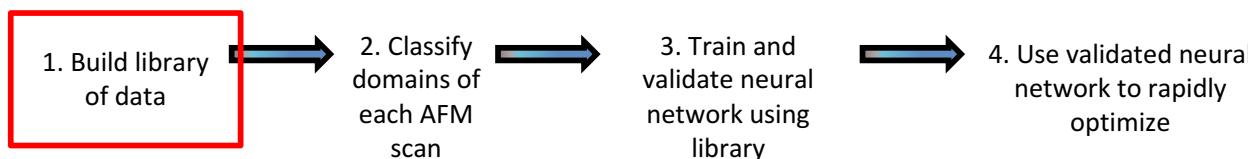
EuclideanClassifier

- Sometimes you really need it all
 - E.g. Optimize annealing parameters in creation of polymer thin films

Create a library of images
with films with different
annealing parameters

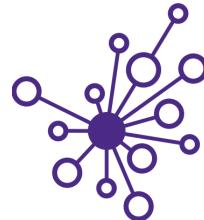


Many thin polymer files
characterized by AFM



Wes Tatum
(Luscombe Lab)

Regression + classification



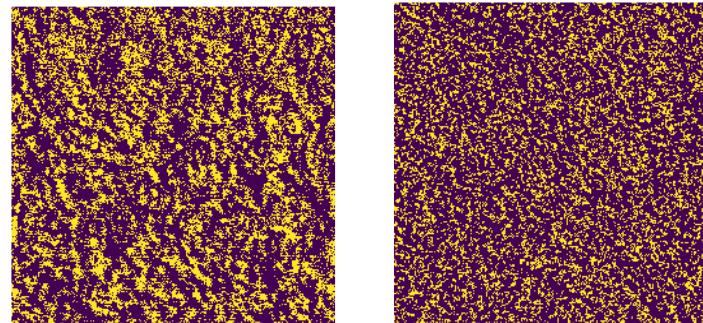
Adhesion

PixelClassifier

EuclideanClassifier

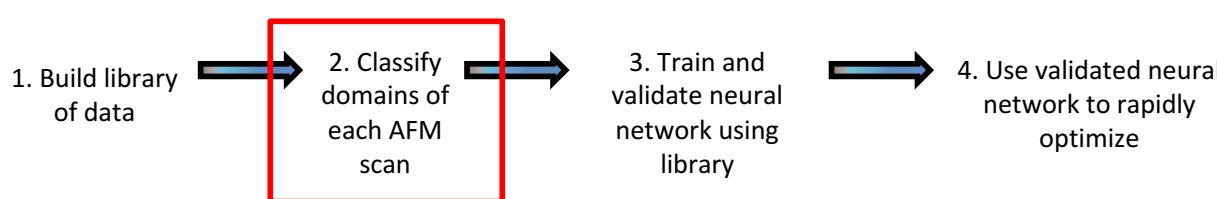
- Sometimes you really need it all
 - E.g. Optimize annealing parameters in creation of polymer thin films

Classify the domains in the images as crystalline or amorphous



Yellow = Crystalline
Purple = Amorphous

For each material can compute ratio



Wes Tatum
(Luscombe Lab)

Regression + classification



Adhesion

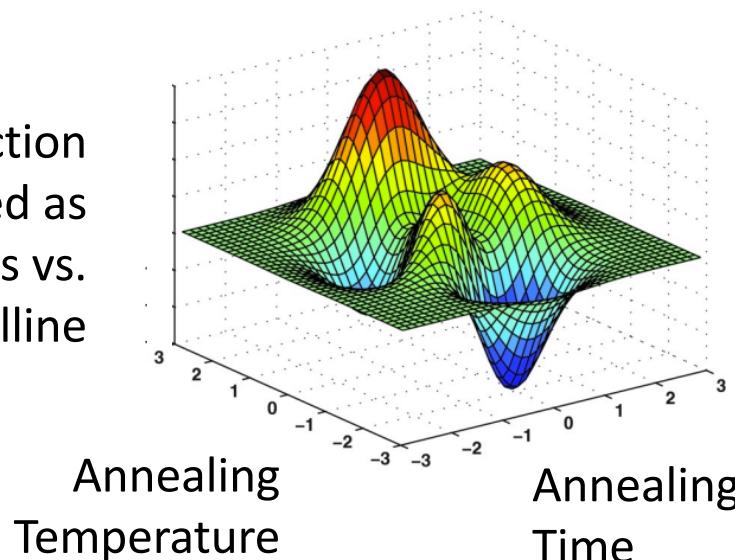
PixelClassifier

EuclideanClassifier

- Sometimes you really need it all
 - E.g. Optimize annealing parameters in creation of polymer thin films

Train neural network to predict morphology based on annealing parameters

Fraction classified as amorphous vs. crystalline



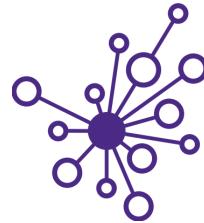
1. Build library of data
2. Classify domains of each AFM scan
3. Train and validate neural network using library
4. Use validated neural network to rapidly optimize



Wes Tatum
(Luscombe Lab)



Regression + classification



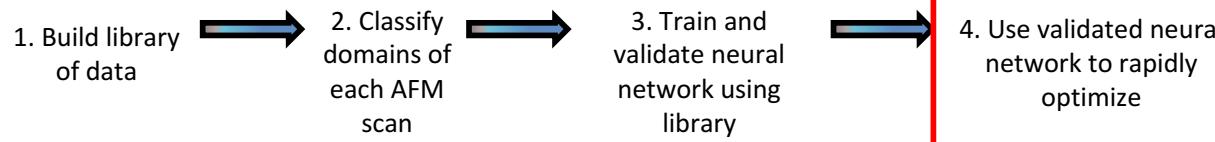
Adhesion

PixelClassifier

EuclideanClassifier

- Sometimes you really need it all
 - E.g. Optimize annealing parameters in creation of polymer thin films

Use the neural network to rapidly optimize the process for idealized functionalization



DIRECTee
Wes Tatum
(Luscombe Lab)

W

Supervised vs. unsupervised

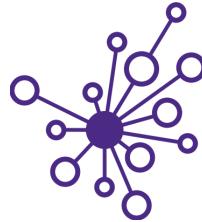


vs.

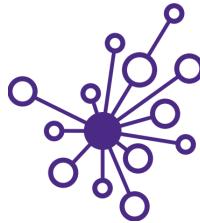




Supervised vs. unsupervised

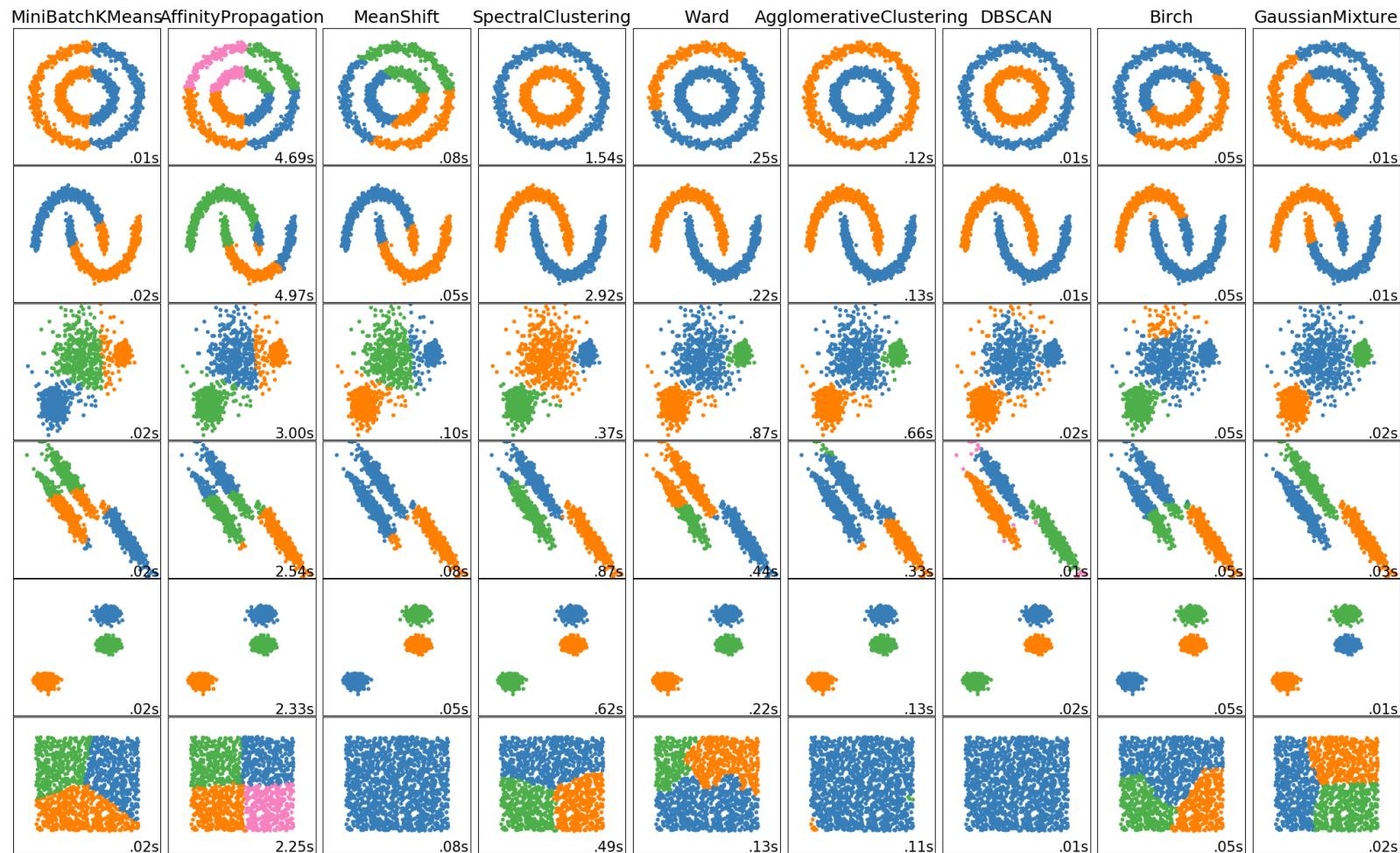


- Supervised
 - Input data is “labeled”
 - Know the Kd
 - Know the source of the atmospheric particulates
- Unsupervised
 - No labels associated with the data
 - Discovering structure in the data



Discovering hidden structure

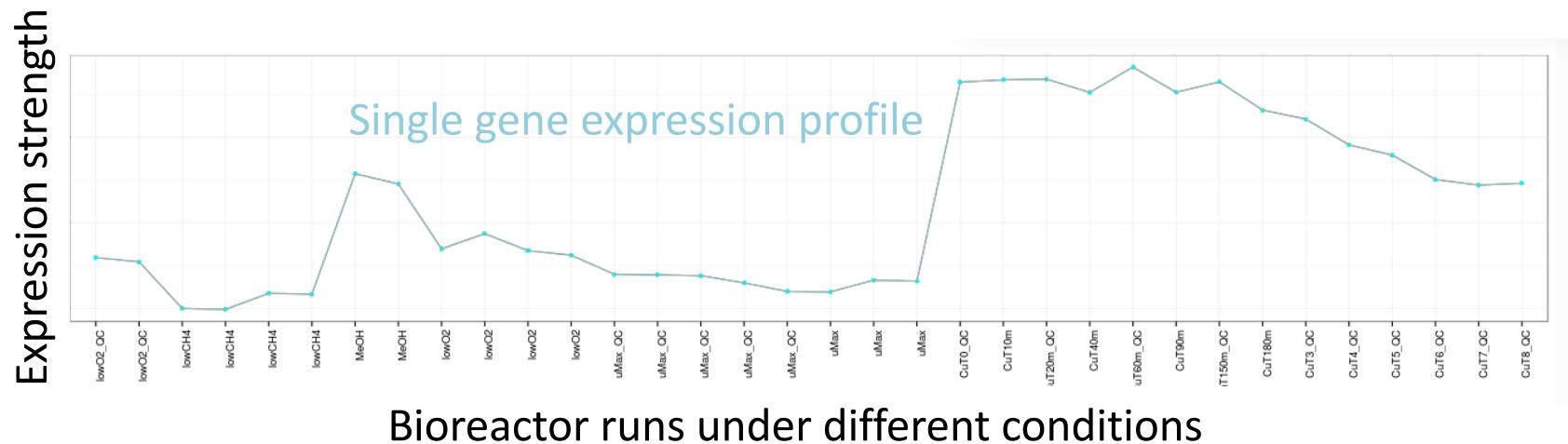
- Clustering



Discovering hidden structure



- Clustering
 - E.g. In an industrially relevant bacterium, identify genes that are co-regulated across conditions



Find all the genes with similar expression profiles (clustering)

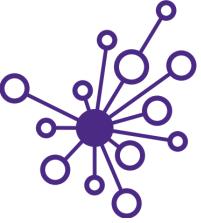
Search their upstream regions for regulatory motifs



Alexey Gilman



Jiayuan Guo



ture

Expression strength
of genes in cluster

Predicted regulatory
sequences

Location of sequences
relative to gene

DIRECTee

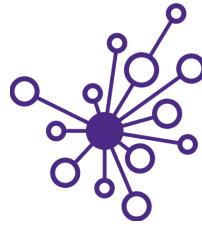


Alexey Gilman

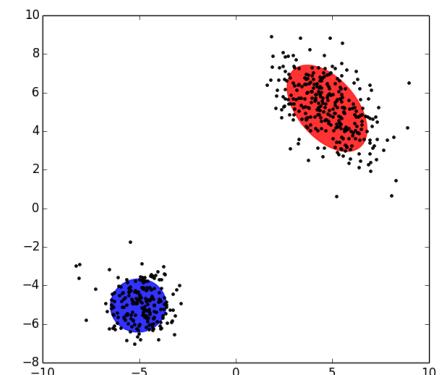
Jiayuan Guo



Discovering hidden structure

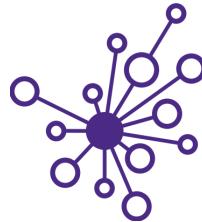


- Clustering
 - K-means
 - Find k clusters in the data
 - gdbSCAN
 - Find clusters with a given local density
 - DPGMM
 - Dirichlet Process Gaussian Mixture Model
 - Grows Gaussians across your data





Visualization

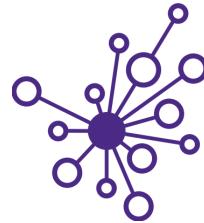


- How can you convey a complex multivariate data set to stakeholders & peers?
 - E.g. how does water order around amino acids?

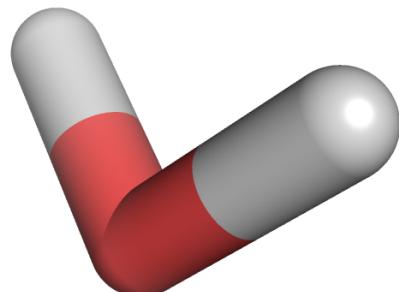
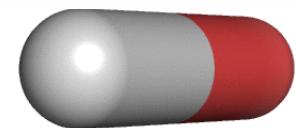
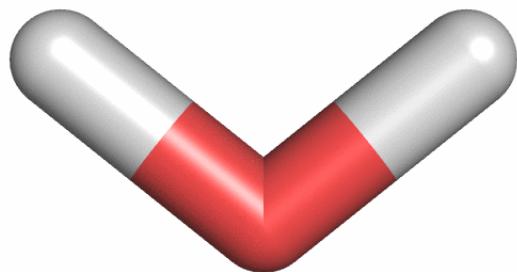
Run 100ns+ molecular dynamics simulations of G-G-X-G-G where X is each of 20 amino acids

Build a model of the occupancy and orientation at a grid of sites around amino acids

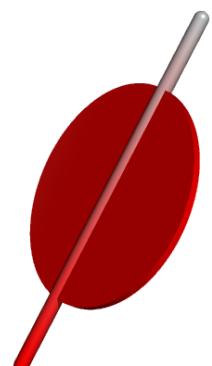
Visualization



Dynamics



Representation

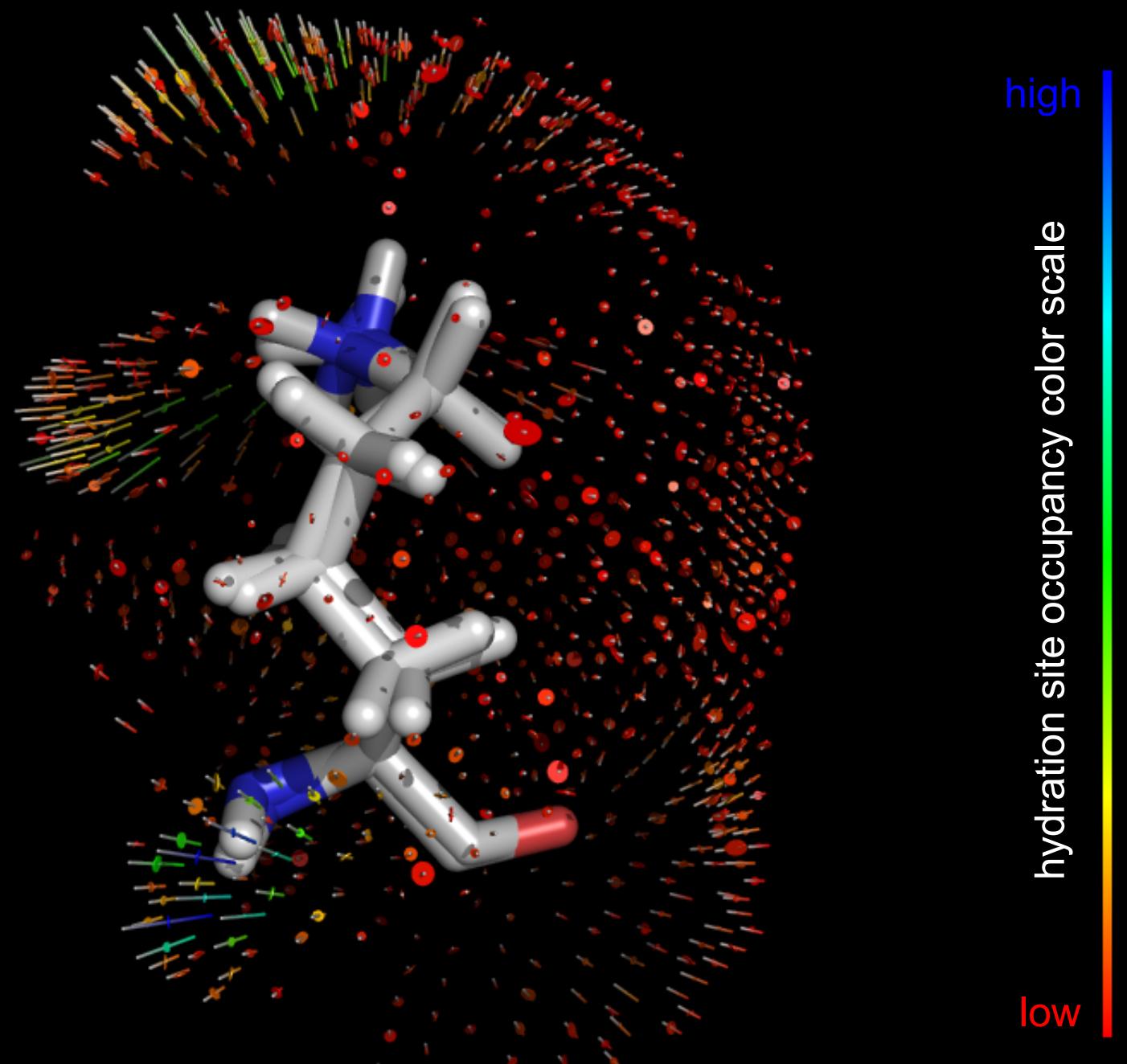


H_2O
Ordering

Dipole
(length)

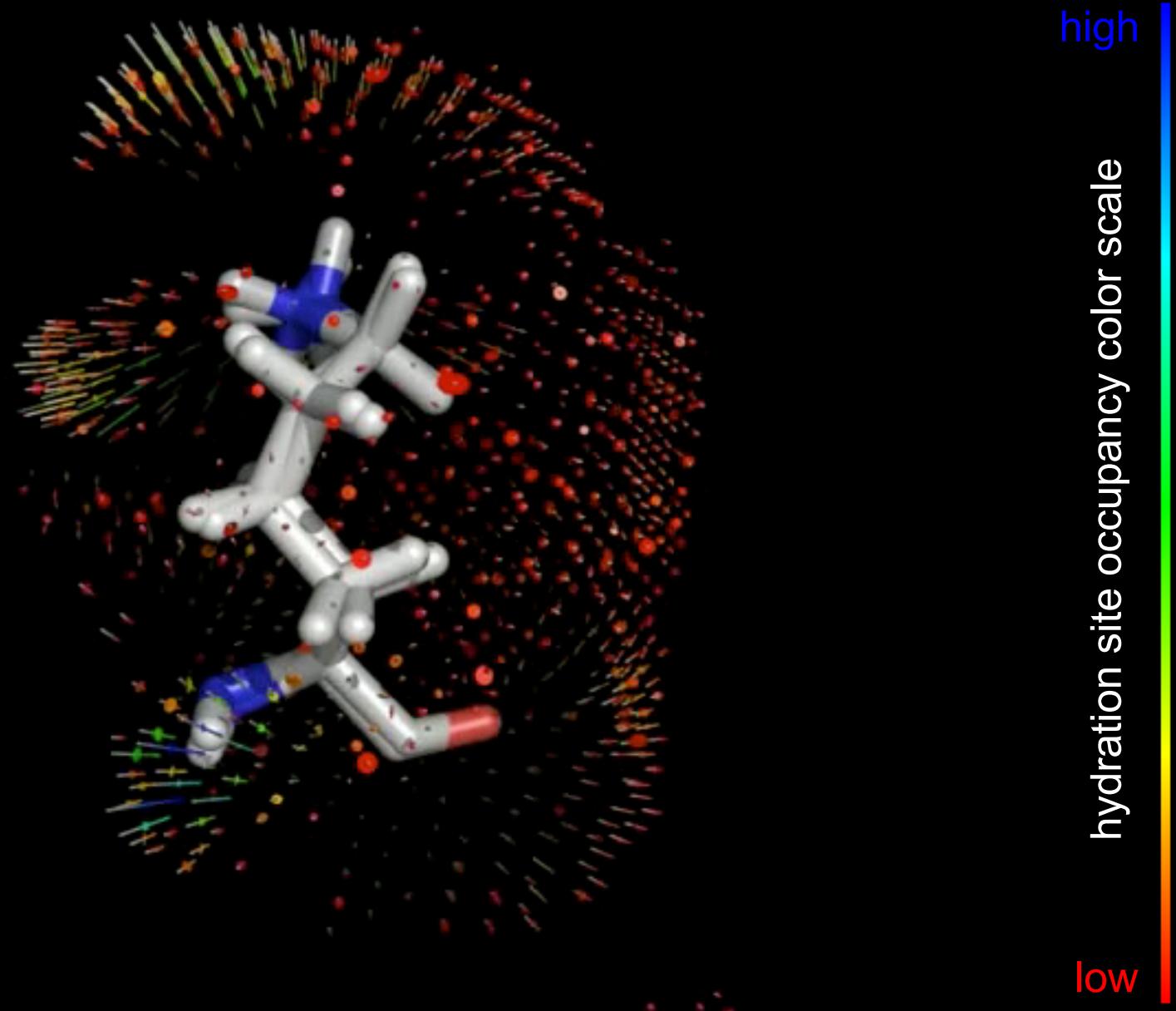
Plane
(disc radius)

Dipole & Plane

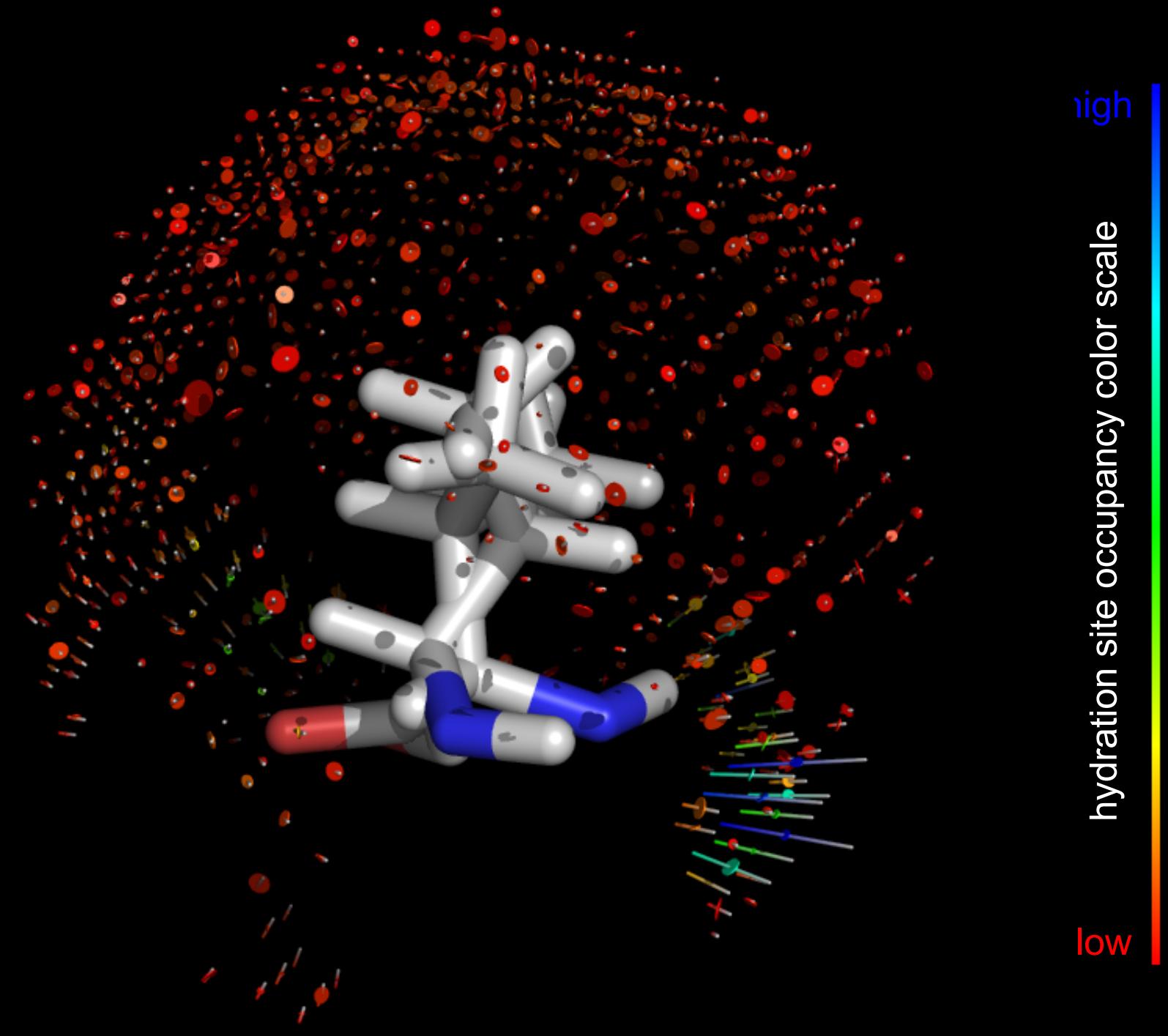


Ace-GG**K**GG-Amd

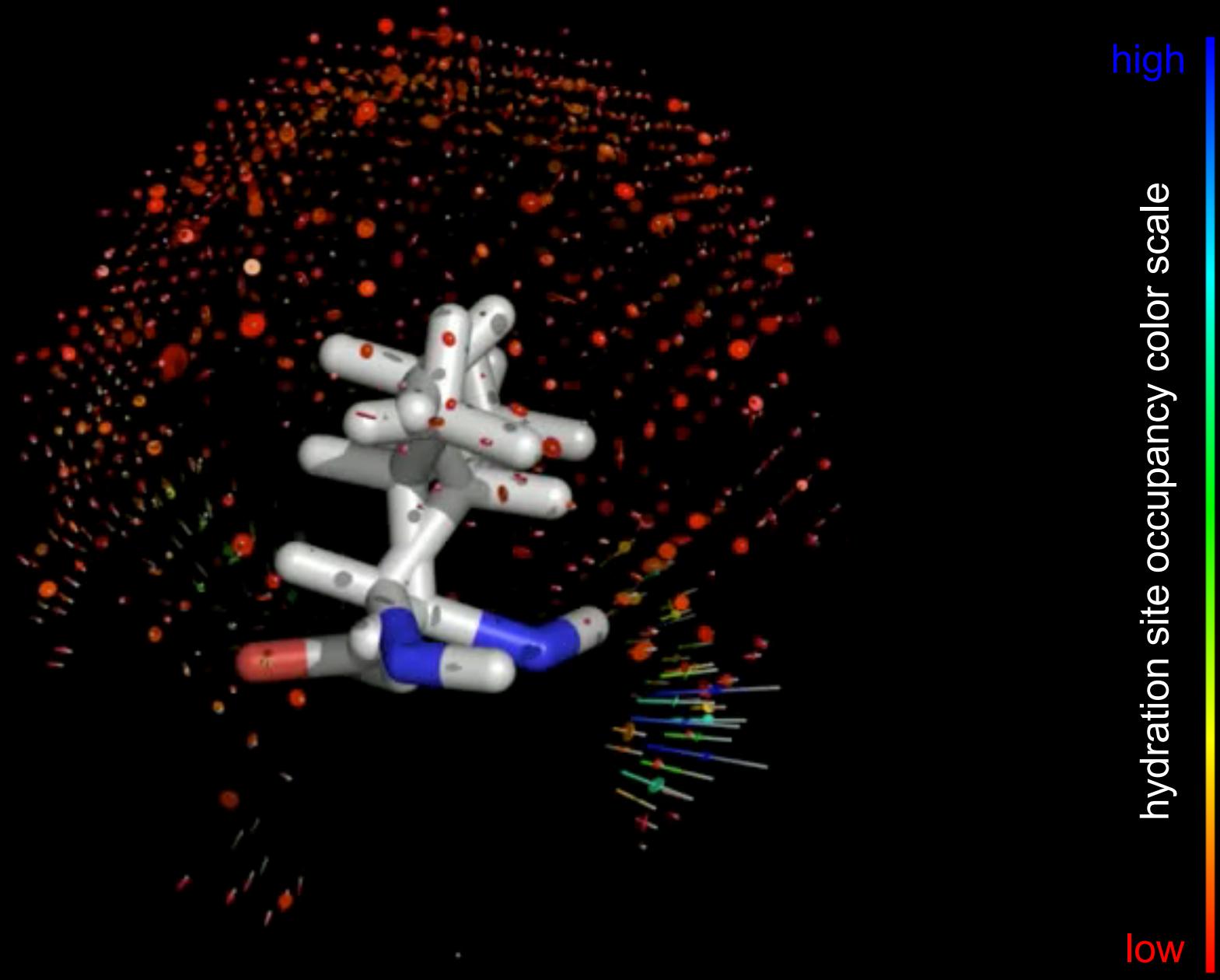
Ace-GG**K**GG-Amd



Ace-GG V GG-Amd

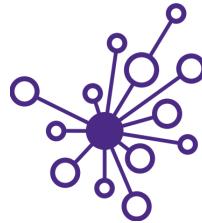


Ace-GG V GG-Amd





Statistics



- Experimental design
 - Power analysis, sample size, effect size
 - Reproducible results
- Correct statistical test
 - Not everything follows a normal distribution!

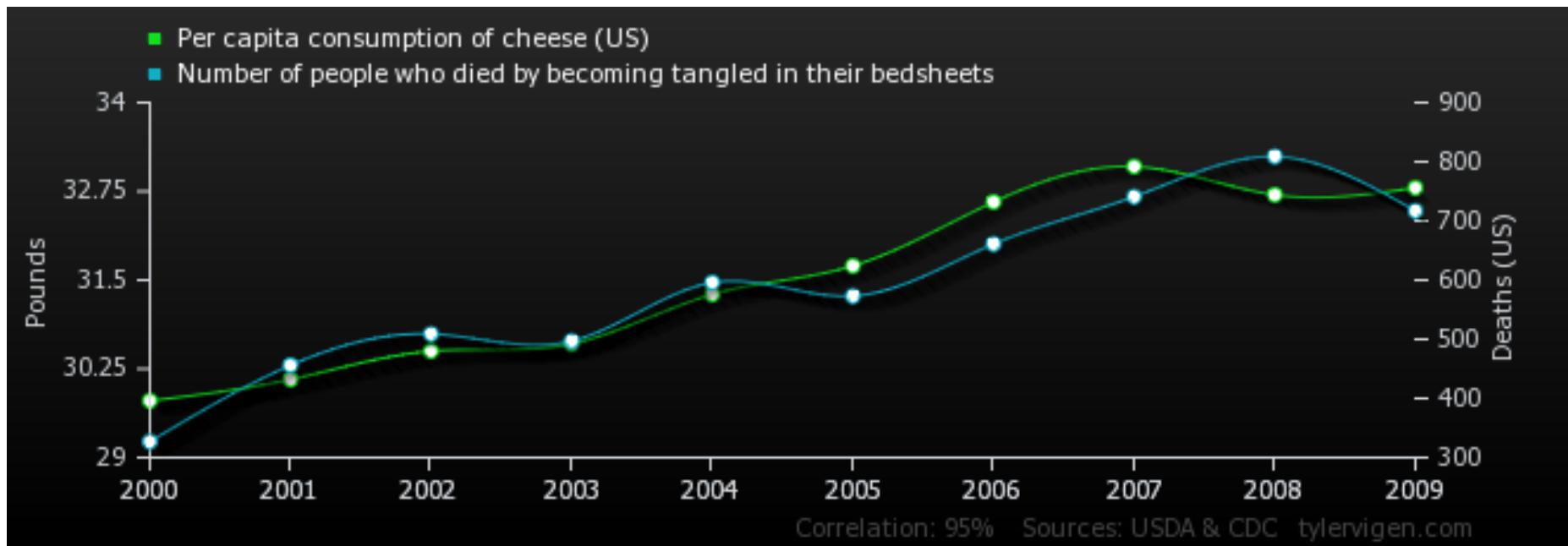


William Sealy Gosset
“Student”



Statistics

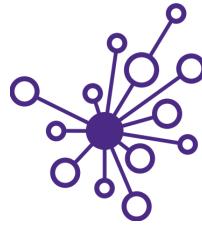
- Multiple hypothesis testing & p -hacking



Correlation: 0.947091



Software engineering

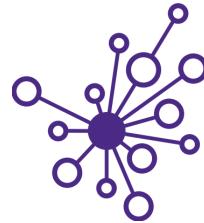


- Your code should not look like this:

```
I n t,e,l[80186],*E,m,u,L,a,T,o,r[1<<21],X,*Y,b,Q,R;I
Z*i,M,p,q=3;I!=localtime(),f,S,kb,h,W,U,c,g,d,V,A;N,O,P=983040,j[5];SDL_Surface*k;i(F,40[E]==!o)i(
z,42[E]==!o)i(D,r[a(I)E[259+4*o]+O])i(w,i[o]==(-2*47[E])*~L)i(v,G(N-S&&1&
(40[z((f^=S'N)&16),E]^f>>C-1)))J(){V=61442;$;O--;}V+=40[E+O]<<D(25);}i(H,
(46[u=76,J(),T(V),T(9[i]),T(M),M(P+18,=,4*o+2),R(M,=,r[4*o]),E]=0))s(o){$;O--
;)40[E+O]=1&&1<<D(25)&o;}i(BP,(*i+=262*o*z(F((*E&15)>9|42[E])),*E&=15))i(SP,(w(7),R&&--1[i]&&o?
R++,Q&Q++,M--:0))DX(){$,O*=27840;O--;}O[(I*)k->pixels]=-!!(1<<7-0&8&r[0/2880*90+0%720/8+
(88+952[1]/128*4+0/720%4<<13]);}SDL_Flip(k);}main(BX,nE)n**nE;{9[i=E=r+P]=P>>4;$;q;)j[--q]=*++nE?
open(*nE,32998):0;read(2[a(I)*i==j]?lseek(*j,0,2)>9:0,j,E+(M=256),P);$;Y=r+16*9[i]+M,Y-
r;Q|R||kb&46[E]&&KB)--64[T=1|O=32[L=(X==Y&7)&1,o=X/2&1,1]=0,t=(c=y)&7,a=c/8&7,Y]>>6,g==~T?y:
(n)y,d=Bx=y,1,!T*t-6&T-2?T-1?d=g:0:(d=y),Q&Q--,R&&L--x(O==Y,O=u=D(51),e=D(8),m=D(14)_
O==Y/2&7,M=(n)c*(L^(D(m)[E]|D(22)[E]|D(23)[E])^D(24)[E]))_L==Y&8,R(K(X)[r],=,c)_L=e+=3,o=0,a=X
x a=m _ T(X[i])_ A(X[i])_ a<22M(U,+=1-2*a+,P+24),v(f=1),G(S+1-a==1<<C-1),u=u&?19:57:a-6?CX+2,a-
3||T(9[i]),a&2&T(M),a&1&M(P+18,=,U+2),R(M,=,U[r]),u=67:T(h[r])_(W= U B u=m,M==~L,R(W[r],&,d)B 0
B L(=~-B L(=--),S=0,u=22,F(N>S)B L?c(Z,i):c(I,n,E)B/**/L?c(Z,i):c(n,E)B L?V(I Z,I,i):V(I n,I
Z,E)B L?V(Z,int,i):V(n,Z,E))_+e,h=P,d=c,T=3,a=m,M--_++e,13[W=h,i]=(o)==!L)?(n)d:d,U=P+26,M-
=~1o,u=17+(m=a)_ (a=m B L(=+),F(N<S)B L(|)=)B e(+)+B L(&)=)B L(=--),F(N>S)B L(^)=)B L(=),F(N>S)B
L(=))_!L?L=a+=8 x L(=):!o?Q=1,R(r[p=m x V],=,h):A(h[r])_T=a=0,t=6,g=c x M(U,=,W)_ (A=h(h[r]),V=m?
++M,(n)g:o?31&2[E]:1)&&(a<4?V%a/2+C,R(A,=,h[r]):0,a&1?R(h[r],>>,V):R(h[r],<=,V),a>3?
u=19:0,a<5?0:F(S>>V-1&1)B R(h[r],+=,A>C-V),G(h[N])^F(N&1))B A&=(1<<V)-1,R(h[r],+=,A<<C-
V),G(h(N*2)^F(h(N)))B R(h[r],+=(40[E]<<V-1)+,A>>1+C-V),G(h[N])^F(A&1<<C-V))B R(h[r],+=(40[E]<<C-
V)+,A<<1+C-V),F(A&1<<V-1),G(h[N]^h(N*2))B G(h(N)^F(h(S<<V-1)))B G(h(S))B 0
V<C| F(A),G(0),R(h[r],+=,A*==((1<<C)-1>>V)))_(V==!!--1[a=X,i]B V=&!m[E]B 0 B
V=!++1[i]),M+=V*(n)c _ M+=3-o,L?0:o?9[M=0,i]=BX:T(M),M+=o*L?(n)c:c _ M(U,&,W)_
L=e+=8,W=P,U=K(X)_!R|_1[i]?M(m<2?u(8,7,:P,=,m&1?P:u(Q?p:11,6,)),m&1|_w(6),m&2|_SP(1):0
_!R|_1[i]?M(m?P:u(Q?p:11,6,=,u(8,7,)),43[u=92,E]=!N,F(N>S),m|_w(6),SP(!N==b):0
_o=L,A(M),m&&A(9[i]),m&2?s(A(V)):o||(4[i]+=c)_R(U[r],=,d)_986[1]^=9,R(*E,=,l[m?2[i]:n)c])_
R(l[m?2[i]:n)c,=*E)_R=2,b=L,Q&Q++_W-U?L(^),M(U,^=,W),L(^):0 _ T(m[i])_ A(m[i])_
Q=2,p=m,R&&R++_L=0,o==*E,F(D(m+=3*42[E]+6*40[E])),z(D(1+m)),N-*E=D(m-1)_N=BP(m-1)_1[E]=-h(*E)-
2[i]==-h(*i)_9[T(9[i]),T(M+5),i]=BX,M=c _ J(),T(V)_s(A(V))_J(),s((V&-m)+1[E])_J(),1[E]=V-
L=o=1 x L(=),M(P+m,=,h+2)_+M,H(3)_M+=2,H(c&m)_+M,m[E]&&H(4)_(_c&=m)?
1[E]==*E/c,N-*E%==c:H(0)_*i=N=m&E[L=0]+c*1[E]*E=-m[E]*E=r[u(Q?p:m,3,*E+)]_m[E]^=1 _ E[m/2]=m&1 -
R(*E,&c)_(_a=c B write(1,E,1)B time(j+3),memcp(r+u(8,3,_),localtime(j+3,m))),a<2?*E=-1seek(O=4[E]
[j],a(I)5[i]<<9,0)?((I*))((a?write:read))(O,r+u(8,3,_),*i):0:0),O,D(16)?
v(0):D(17)&&G(F(0)),CX*D(20)+(D(18)-D(19))*~!L,D(15)?O=m=N,41[43[44[E]=h(N),E]=!N,E]=D(50):0,!++q?
kb=1,*1?SDL_PumpEvents(),k=k?k:SDL_SetVideoMode(720,348,32,0),DX():k?
SDL_Quit(),k=0:0:0;}i(G,48[E]=o)i(K,P+(L?2*o:2*o+o/4&7))
```



Software engineering



- Scientific and engineering software tools are first class research products!
- Programming is **not** software engineering
- This is what CHEME/CHEM/MSE 546 is about!

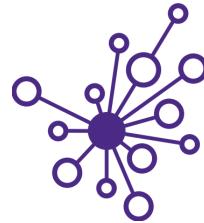
W

Data: Don't be afraid of it!





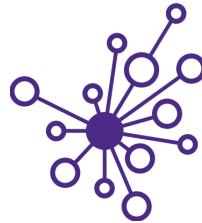
DIRECT to the rescue!



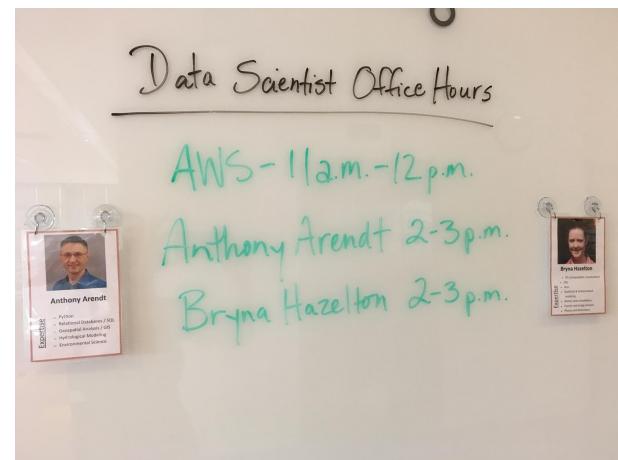
- Data Science Option
 - Data Intensive Research Enabling Clean Tech (DIRECT)
 - ChemE 545
 - Data Science Methods for Clean Energy Research
 - ChemE 546
 - Software Engineering for Molecular Data Scientists
 - ChemE 547
 - Capstone Project in Molecular Data Science



eScience to the rescue!



- Data Scientist Office Hours
 - Get help with your data science questions
 - Data management
 - Machine learning
 - Statistics
 - Visualization
 - Software engineering

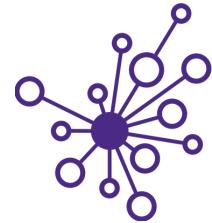


+ a b l e a u

UNIVERSITY LIBRARIES

W

Swipe right for Data Science¹

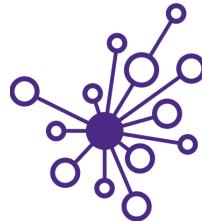


Molecular Data Scientist
Knows thermodynamics **and** machine learning



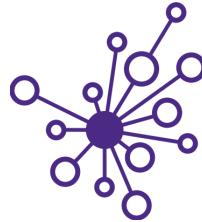


What is data science?



- Quick in class exercise
- Alone: Define Data Science by answering as many of the following questions as you can (write or type your answer)
 - What is Data Science?
 - What/who is a data scientist?
 - Why is Data Science a thing all of a sudden?
 - Why does Data Science matter, broadly, in my field of [insert]?
 - Why does Data Science matter, specifically, in my sub-discipline of [insert]?
- ~5 min working alone – take some notes!

What is data science?



- Quick in class exercise
- At your tables:
 - Introduce yourselves (1st names and departments)
 - Appoint 1 facilitator (soonest birthday, e.g. today) and 1 scribe (farthest birthday, e.g., yesterday)
 - Go around the table, each person answers each question, then move onto next question → OK if you didn't answer something
 - *What is Data Science?*
 - *What/who is a data scientist?*
 - *Why is Data Science a thing all of a sudden?*
 - *Why does Data Science matter, broadly, in my field of [insert]?*
 - *Why does Data Science matter, specifically, in my sub-discipline of [insert]?*
- ~10 min (I will flex time as needed)

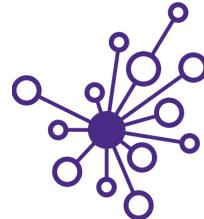


Questions?



- Questions about DIRECT or DSMCER?

Acknowledgements



Beck Research Lab

- Pearl Philip (ChemE, GSK)
- Rahul Avadhoot (ChemE)
- Jiayuan Guo (ChemE)
- Alexey Gilman (ChemE, PhD)

Gut microbiome & health

- Jisun Paik (Comparative Medicine)

Showcase examples

- Elizabeth Nance (ChemE)
 - Chad Curtis (ChemE)
 - Mike McKenna (ChemE)
- Wes Tatum (MSE)
 - Luscombe (Chemistry)
- Jay Rutherford (ChemE)
 - Posner (ChemE + MechE)
- Blake Hough (ChemE, PhD)
 - Pfaendtner & Schwartz



GORDON AND BETTY
MOORE
FOUNDATION

ALFRED P. SLOAN
FOUNDATION

FRED HUTCH
UNIVERSITY OF WASHINGTON
CANCER CONSORTIUM

eScience Institute
ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS