Knowledge and solutions for a changing world

Be boundless

Advancing data-intensive discovery in all fields

Data Science Methods for Clean Energy Research (DSMCER)

&

Software Engineering for (SEMDS) Molecular Data Scientists

UW DIRECT

(Data Intensive Research Enabling Cutting-edge Tech)

https://uwdirect.github.io

Dave Beck (dacb)

eScience/Chemical Engineering

Evan Komp

Chemical Engineering

What is this class about?

Overview of **Data Science methods**

- Basics of statistics
- Tool selection
- Best practices
- **Not** about designing new algorithms

Group project using these methods

Provide you with computational tools and a **software mindset** to tackle science problems

Course website & setup

Equipment / work platforms

– Laptops: use them !

– Canvas: use it!

– Software: install it!

Class website: https://uwdirect.github.io

Code of Conduct

Please **read and make sure to respect the code of conduct**

https://uwdirect.github.io/code_of_conduct.html

If you find cases where the code of conduct is not being respected, let us know!

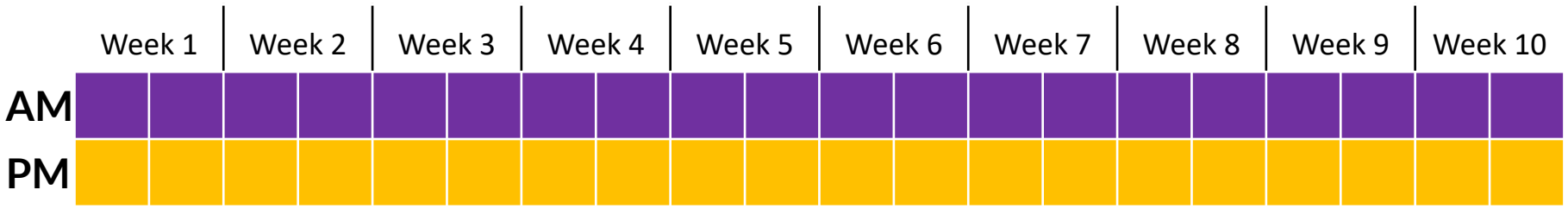
If you do not respect the code of conduct, we may refer to the university for action.

Two quarters of courses in one!

- Two quarter courses make up training
  - CHEME 546: Software Engineering for Molecular Data Scientists (SEMDS)
    - One quarter (10 weeks)
  - CHEME 545: Data Science Methods for Chemical Engineering Research (DSMCER)
    - One quarter (10 weeks)
- UW runs these concurrently

SEMDS

DSMCER

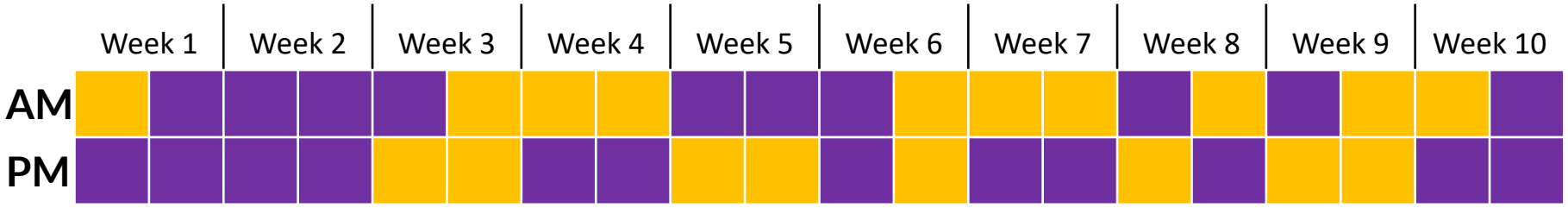


Two quarters of courses in one!

- Two quarter courses make up training
  - CHEME 546: Software Engineering for Molecular Data Scientists (SEMDS)
    - One quarter (10 weeks)
  - CHEME 545: Data Science Methods for Chemical Engineering Research (DSMCER)
    - One quarter (10 weeks)
- We run these concurrently

SEMDS

DSMCER



About Evan

Office: BNS 236
Email: evankomp+dsmcer@uw.edu
Office Hours: M 11:00 @ BNS 241

- **Background** – ChE (BS), PhD student in ChemE
- **Research** – Conducting and Teaching at the intersection of Machine Learning and Chemical/Biological sciences, Reaction Kinetics, Protein Stability

About Dave Beck

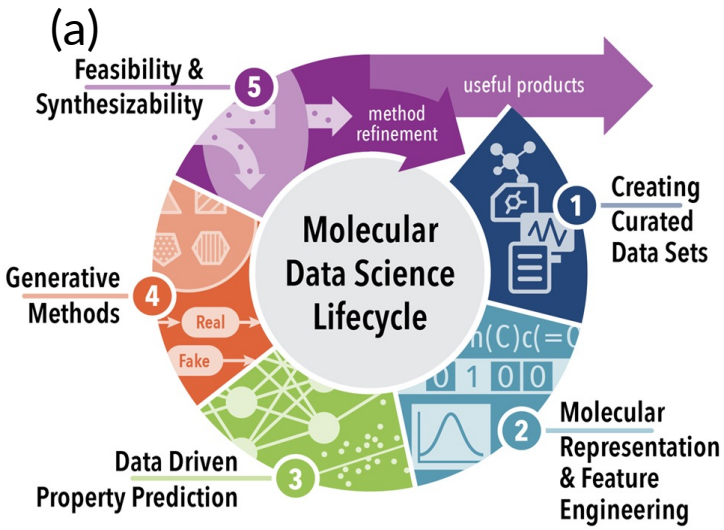
Office: Zoom and Slack

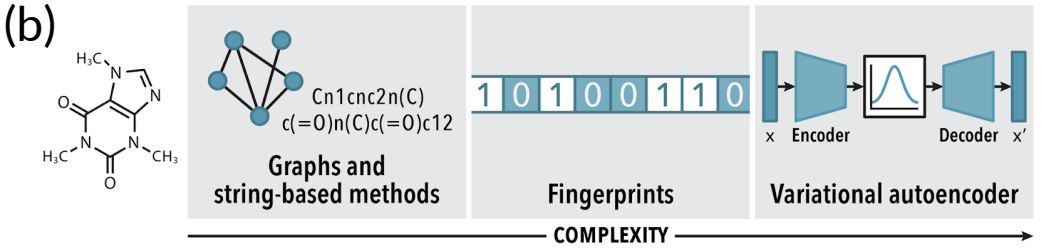
Email: dacb@uw.edu

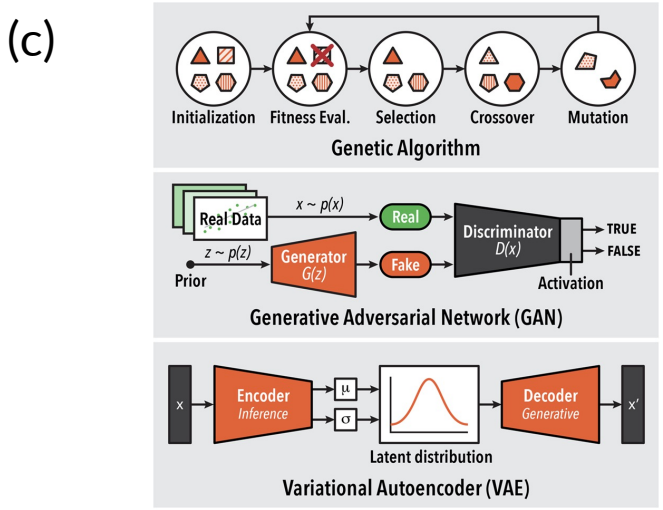
- **Background** – Computer science, biology, chemistry (BS); Medicinal Chemistry (PhD); BioE (post-doc)
- **Research** – Molecular data science with applications in energy, environment and health. Scientific software engineering for reproducibility.
  - Molecular mechanics
  - *omics
  - ML in molecular settings

Molecular Data Science

Combining (a) diverse data sources and unique (b) molecular representations and (c) generative models to design new molecules.







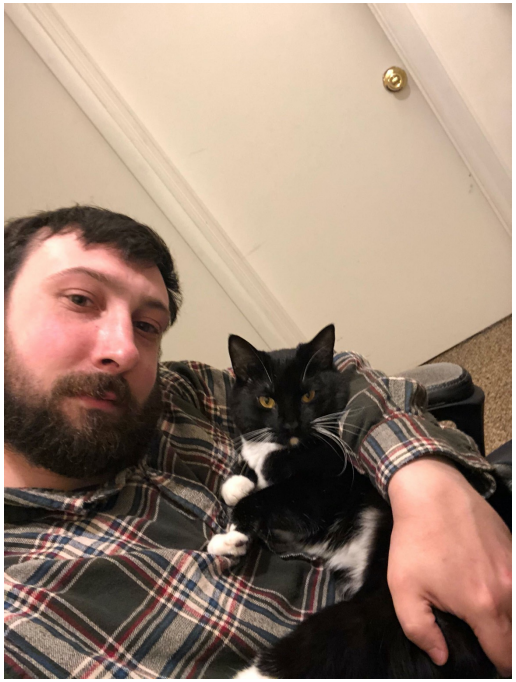
[1] Ashraf, Joshi, Beck, Pfaendtner. *Ann. Rev. Bio. Chem. Eng.* https://doi.org/10.1146/annurev-chembioeng-101220-102232

Teaching Assistants

Orion

Email: odollar@uw.edu

Nels



Email: nlsschim@uw.edu

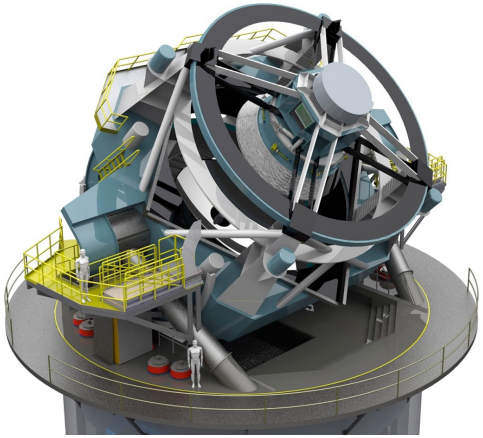
Data + science?

Science, Art & Engineering

- **Physical Sciences**: Physics, Chemistry, Biology, Molecular science
- **Engineering**: Environment, Materials, Civil, Electrical
- **Social Sciences**: Psychology, Economics, Sociology, Public Policy

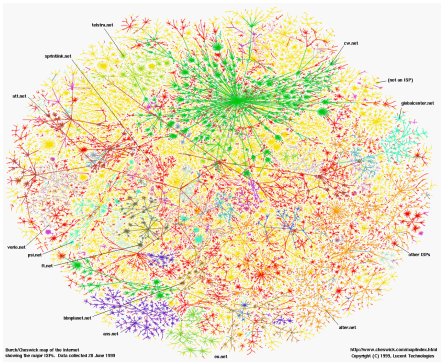
OMG, so much data!

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



Astronomy: LSST

Physics: LHC



Sociology: Social networks

Biology: Sequencing

Neuroscience: EEG, fMRI

Economics: POS terminals

Need time & energy to compute

Modeling data – the larger the data the more time needed to carry out an operation

- Scalar multiplication - $1.0 \cdot 2.0$
- Vector multiplication – $[1.0\ 2.0\ 3.0\ \ldots\ ] \cdot [2.3\ 4.5\ 5.2\ \ldots]$

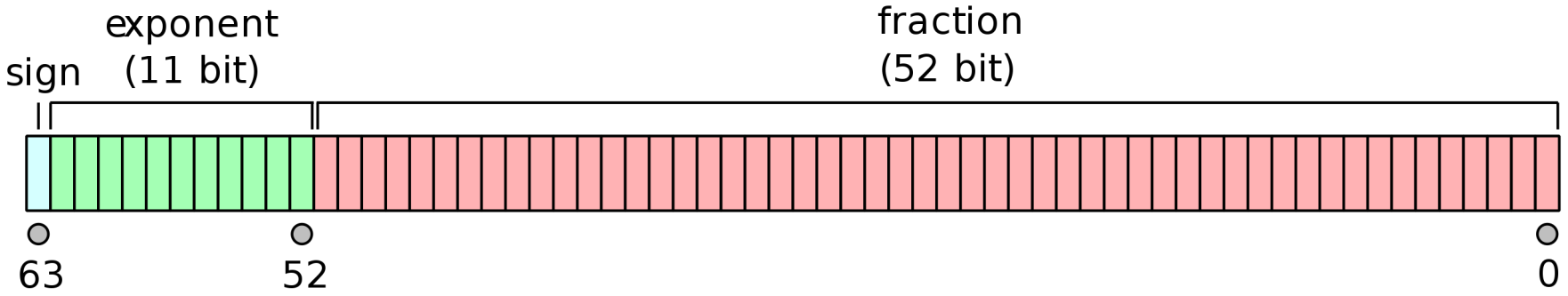
Time 1 $\ll$ Time 2

- For vector of length N, when carrying out the dot product we make $N \cdot N$ multiplications and sum N numbers afterwards

Need memory to store data

One double precision number (16 significant figures) needs 64 bits – 8 bytes

3.14159265358979311599796346844185161590576171875



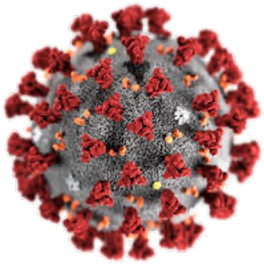
One color image from your phone is **~24 million numbers**

-> A dataset of images can be many terabytes in size, even modern computers cannot work on the whole dataset simultaneously

This cost can be prohibitive

Extracting properties from/using large dataset is **computationally demanding: high memory cost** & **time demand**

We need answers now for important problems



- Drug/antiviral design
- Global warming - ecological forecasting
- Real-time sensing, learning, decision-making
- Detecting hazardous weather

Ref: The National Science Foundation's (NSF) Harnessing the Data Revolution (HDR) Big Idea

Computers then and now

1943

ENIAC - Electronic Numerical Integrator and Computer

- High-speed memory - 20 words (equivalent to about **80 bytes**)
- simple calculations involving 20 numbers of ten decimal digits
- add and subtract **5000 times per second**

https://en.wikipedia.org/wiki/Computer#First_computing_device

2018

SUMMIT - 2nd fastest in the world

- Non-volatile memory per node **1600 gigabytes GB**
- mathematical calculations at the rate of **200 quadrillion per second**, or 200 petaflops

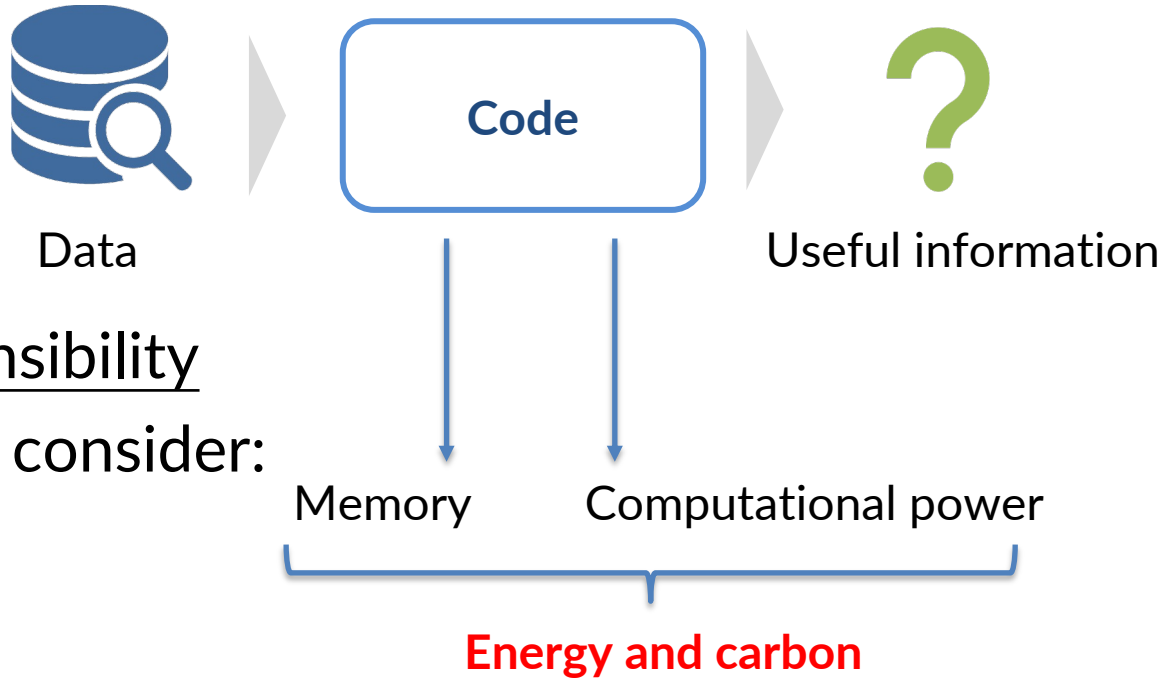
https://www.olcf.ornl.gov/summit/

Big Data + Big Compute =

Big Utility…

We will explore in this course:

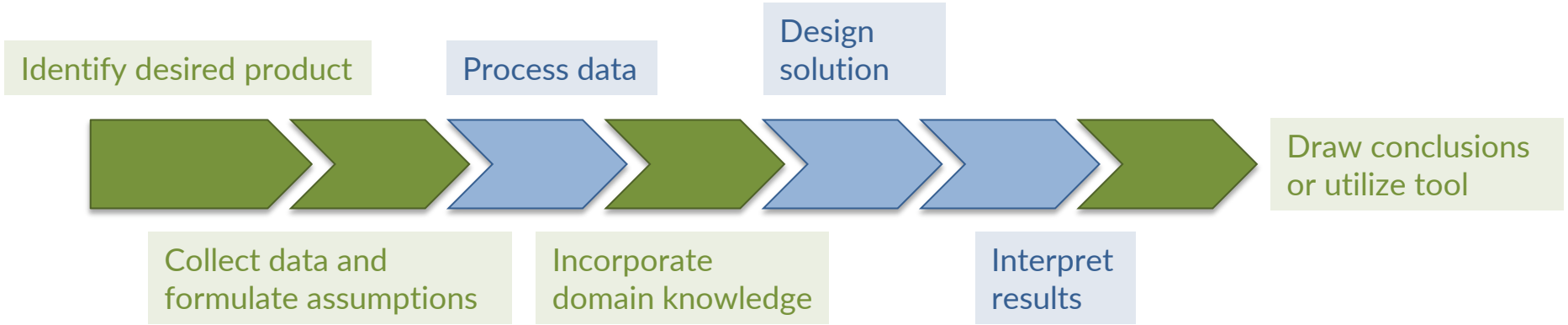
- what kind of **data** is out there, how can we operate on it, and what **information** we can extract



+ Big responsibility

And how to consider:

Your skillset is critical!



Software engineers make sure all of this can be used by others!

New paradigm of data-intensive discovery and innovation

Deep domain knowledge

Data Science

Data management

- Databases, scalable data handling, data curation

Statistics

Visualization

Machine learning

- Supervised & unsupervised
- Utility cost tradeoff

Software engineering

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)

Scalable data processing systems

New paradigm of data-intensive discovery and innovation

Deep domain knowledge

Data Science

Data management

- Databases, scalable data handling, data curation

Statistics

Visualization

Machine learning

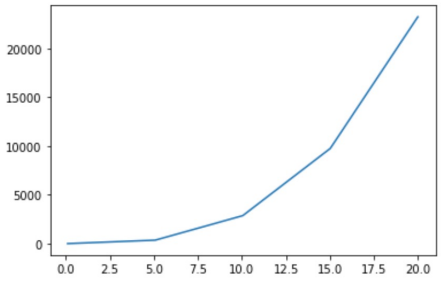
- Supervised & unsupervised
- Utility cost tradeoff

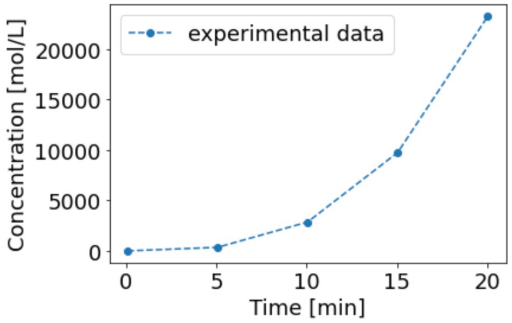
Software engineering

Visualization

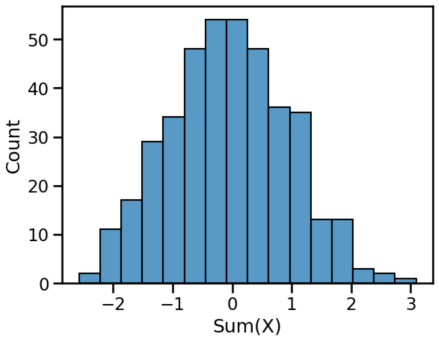
How can visually display data in a useful way?

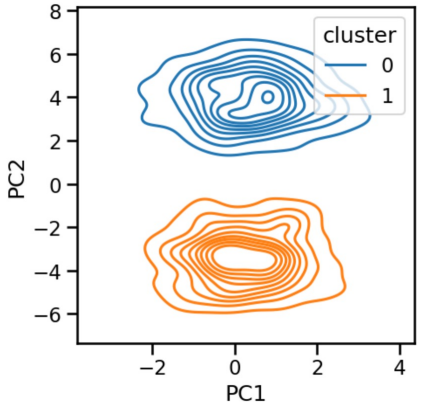
Effectiveness





Expressiveness





New paradigm of data-intensive discovery and innovation

Deep domain knowledge

Data Science

Data management

- Databases, scalable data handling, data curation

Statistics

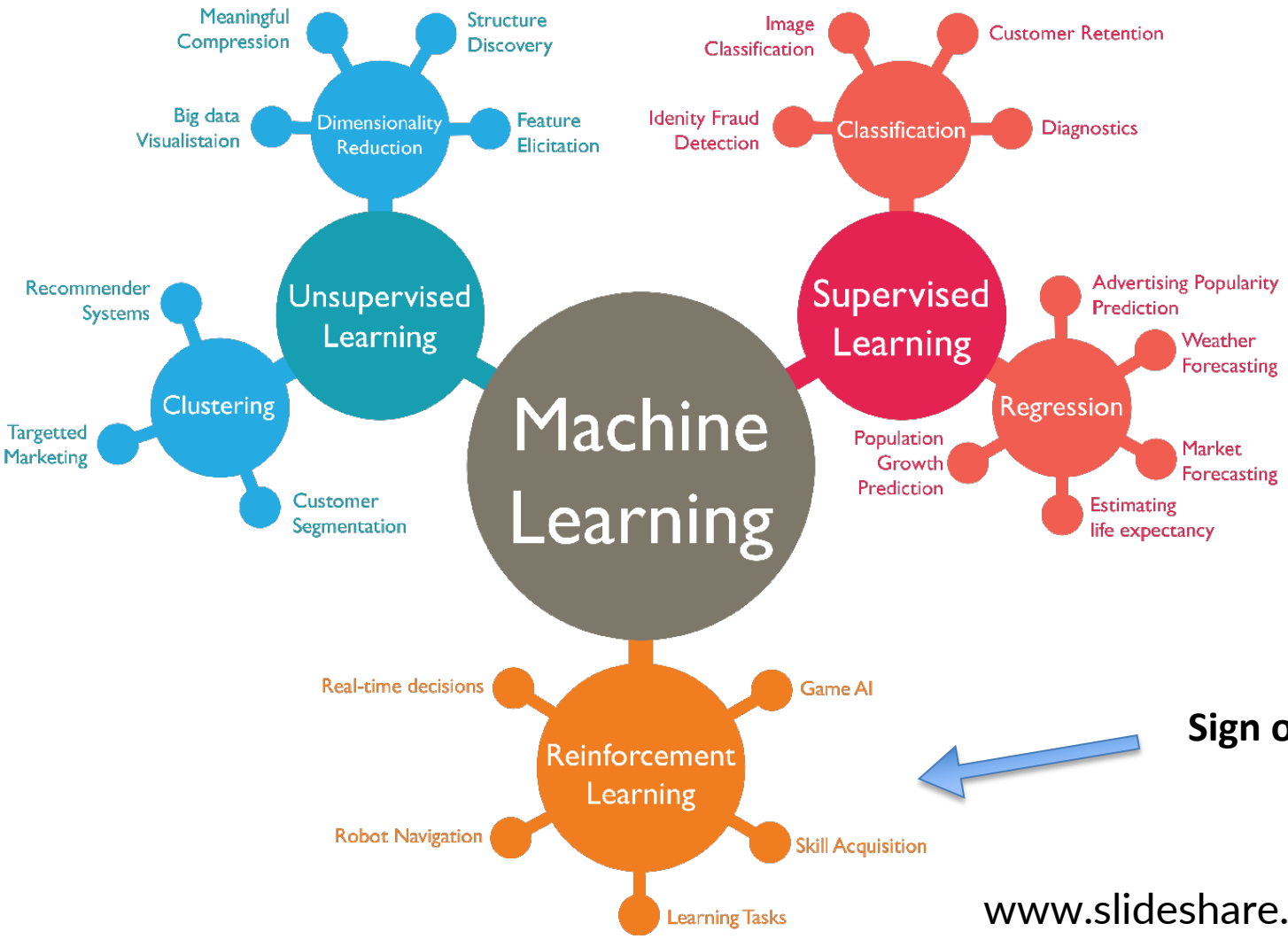
Visualization

Machine learning

- Supervised & unsupervised
- Utility cost tradeoff

Software engineering

Machine learning



Sign on Bonus!

www.slideshare.net/awahid

Supervised ML helps us accelerate

Supervised: Predict a measured, expensive label using low cost features

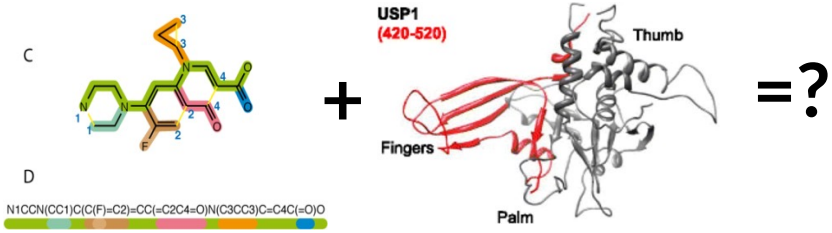
Example: develop structure-activity-relationship eg. binding affinity to cancer target

Database of 400,000 drug like molecules

Experimental inhibition activities against cancer drug target

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

For a new small molecule, predict the Kd?[1,2]



Pearl Philip & Rahul Avadhoot

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

SML requires labeled data

Supervised

Features or predictors

Outcome

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

SML requires labeled data

Supervised

Features or predictors

Outcome

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
New!	6	121	1	0	0	???????

Unsupervised ML can help us interpret

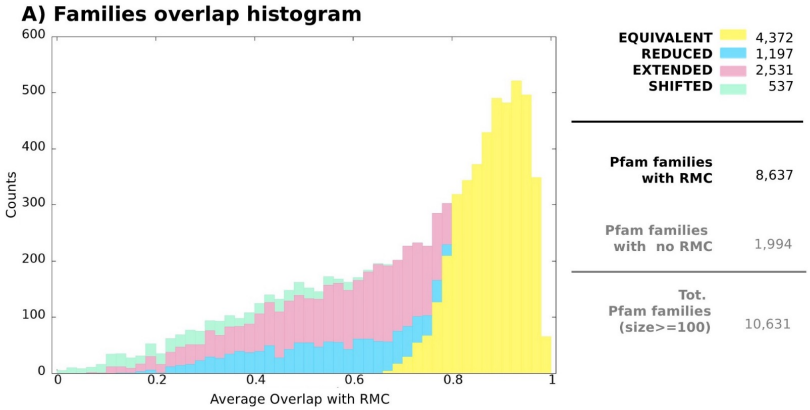
Unsupervised: Intuit patterns in data to increase information or utility

Example: cluster protein alignments into families that conduct similar functions

Database of thousands of protein sequences and alignments to other proteins

Build a model that groups together protein sequences by evolutionary similarity

Explored groups that have not been tested before or guess the family of a new protein



Russo, E. T.; Barone, F.; Bateman, A.; Cozzini, S.; Punta, M.; Laio, A. DPCfam: Unsupervised Protein Family Classification by Density Peak Clustering of Large Sequence Datasets. *PLOS Computational Biology* **2022**, *18* (10), e1010610. https://doi.org/10.1371/journal.pcbi.1010610.

USML does not need labels

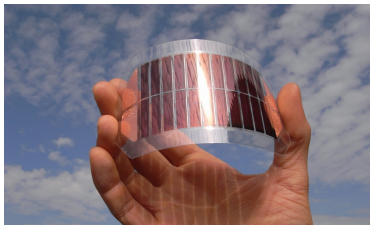
Unsupervised

Features

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

Dataset 1: HCEPDB

Source: Harvard Clean Energy Project – DOI 10.1021/jz200866s



Data

Code

?

Useful information

Data Description

- **2.3mil photovoltaic materials**
- Each has a **chemical identity**
- Experimental **measurements**:
  - Open circuit voltage
  - Energy gap
  - Etc.
- Synthesizing them is **expensive**

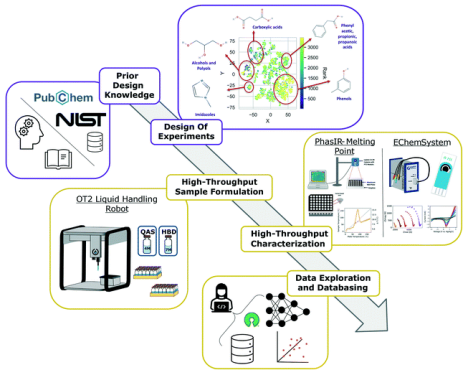
Interesting tasks

- Can we **predict** solar cell capability without having to synthesize a new material?
- How do the measurements taken **correlate** to each other?
- Are there types of materials we should **explore**?

https://www.ise.fraunhofer.de/en/business-areas/photovoltaics/perovskite-and-organic-photovoltaics/organic-solar-cells-and-modules.html

Dataset 2: High Throughput DES

Source: Pozzo Lab (Politi), UW ChemE - DOI 10.1039/D2ME00050D



Data

Code

?

Useful information

Data Description

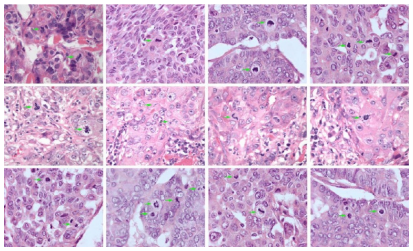
- Data for **mixtures** of 5 quaternary ammonium salts and 10 hydrogen bond donors
- Experimental **measurements** for each:
  - Liquid at room temp
  - Electrolytes potentials

Interesting tasks

- Can we **interpolate** the concentration a mixture will be solid?
- Can we **predict** electrolyte potentials for new mixtures?
- Can we **suggest** new mixtures to try to learn more?

Dataset 3: Breast Cancer Cells

Source: Mittal Lab, Chemical Engineering



Data

Code

Useful information

Data Description

- An H&E stained image of breast tissue
- 11 Infra red fields for each pixel in the image

Interesting tasks

- Can we **identify different cell types?**
- Can we **diagnose** cancerous regions?
- Can we identify **spatial patterns?**

New paradigm of data-intensive discovery and innovation

Deep domain knowledge

Data Science

Data management

- Databases, scalable data handling, data curation

Statistics

Visualization

Machine learning

- Supervised & unsupervised
- Utility cost tradeoff

Software engineering

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(){$,O*=27840;O--;)O[(I*)k->pixels]=-!!(1<<7-O%8&r[O/2880*90+O%720/8+
(88+952[l]/128*4+O/720%4<<13)]);SDL_Flip(k);}main(BX,nE)n**nE;{9[i=E=r+P]=P>>4;$;q;)j[--q]=*++nE?
open(*nE,32898):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,l]=0,t=(c=y)&7,a=c/8&7,Y]>>6,g=~-T?y:
(n)y,d=BX=y,l],!T*t-6&&T-2?T-1?d=g:0:(d=y),Q&&Q--,R&&R--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<2?M(U,+=1-2*a+,P+24),v(f=1),G(S+1-a==1<<C-1),u=u&4?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(I 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-
=~!o,u=17+(m=a)_(a=m B L(+=),F(N<S)B L(|=)B e(+)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 B
V<C||F(A),G(0),R(h[r],+=,A*=~((1<<C)-1>>V)))_(V=!!--1[a=X,i]B V&=!m[E]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[l]^=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(),l[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)?
l[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),memcpy(r+u(8,3,),localtime(j+3),m)),a<2?*E=~lseek(O=4[E]
[j],a(I)5[i]<<9,0)?((I(*)())(a?write:read))(O,r+u(8,3,),*i):0:0),O=u,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,*l?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!

We are all in this for your learning!

Notice the classroom?

- Let's make these ~~lectures~~ **conversations**
- Ask questions or interrupt at **any time**
- Periodic exercises and polls – **discuss with your group**
  - One person per exercise called to share what their group did

Meet each other!

The best way to learn is from each other!

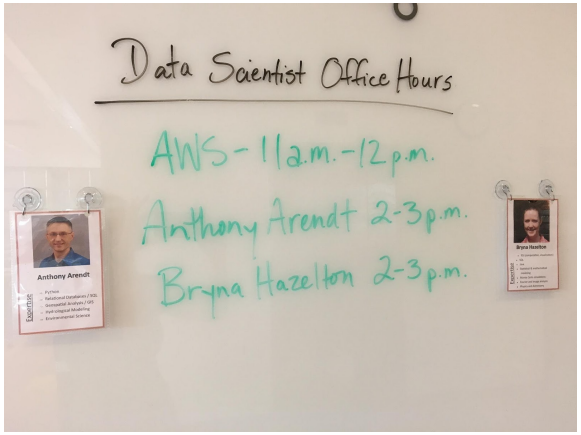
- Introduce yourselves (1st names and departments)
- Go around the table, come to a consensus
    1. *What is Data Science?*
    2. *What/who is a data scientist?*
    3. *Why is Data Science a thing all of a sudden?*
    4. *Why does Data Science matter, broadly, in my field of [insert]?*
    5. *Why does Data Science matter, specifically, in my sub-discipline of [insert]?*

eScience to the rescue!

Data Scientist Office Hours

Get help with your data science questions

- Data management
- Machine learning
- Statistics
- Visualization
- Software engineering



UNIVERSITY of WASHINGTON eScience Institute
ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS

UW-IT

amazon web services

CENTER for STATISTICS and the SOCIAL SCIENCES

MathWorks

tableau

UNIVERSITY LIBRARIES

Questions?