Chemical Engineering

Knowledge and solutions
for a changing world

Be boundless

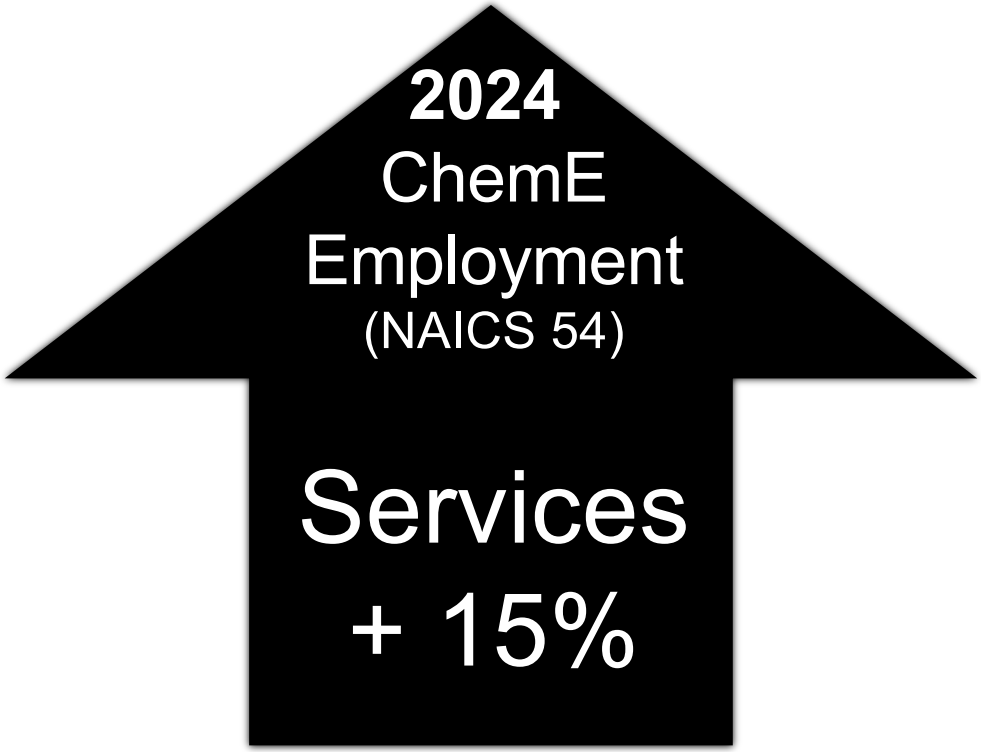
eScience Institute

Advancing data-intensive
discovery in all fields

Graduate Research & Training in Molecular Data Science through a University Wide Approach to Integrative Data Science Education

David A. C. Beck, dacb@uw.edu
Chemical Engineering & eScience Institute
University of Washington

Why data science for ChemE?



#1 Basic Chemicals Man.
#2 Polymer Man.

#1 Engineering Services
#2 R&D Services

U.S. Bureau of Labor Statistics, http://data.bls.gov/projections/nationalMatrixHome?ioType=o

Data Science for ChemE!

How do we bring data science to ChemEs?

Education

AIChE

Perspective

Data Science: Accelerating Innovation and Discovery in Chemical Engineering

David A. C. Beck
Department of Chemical Engineering, University of Washington, Seattle, WA

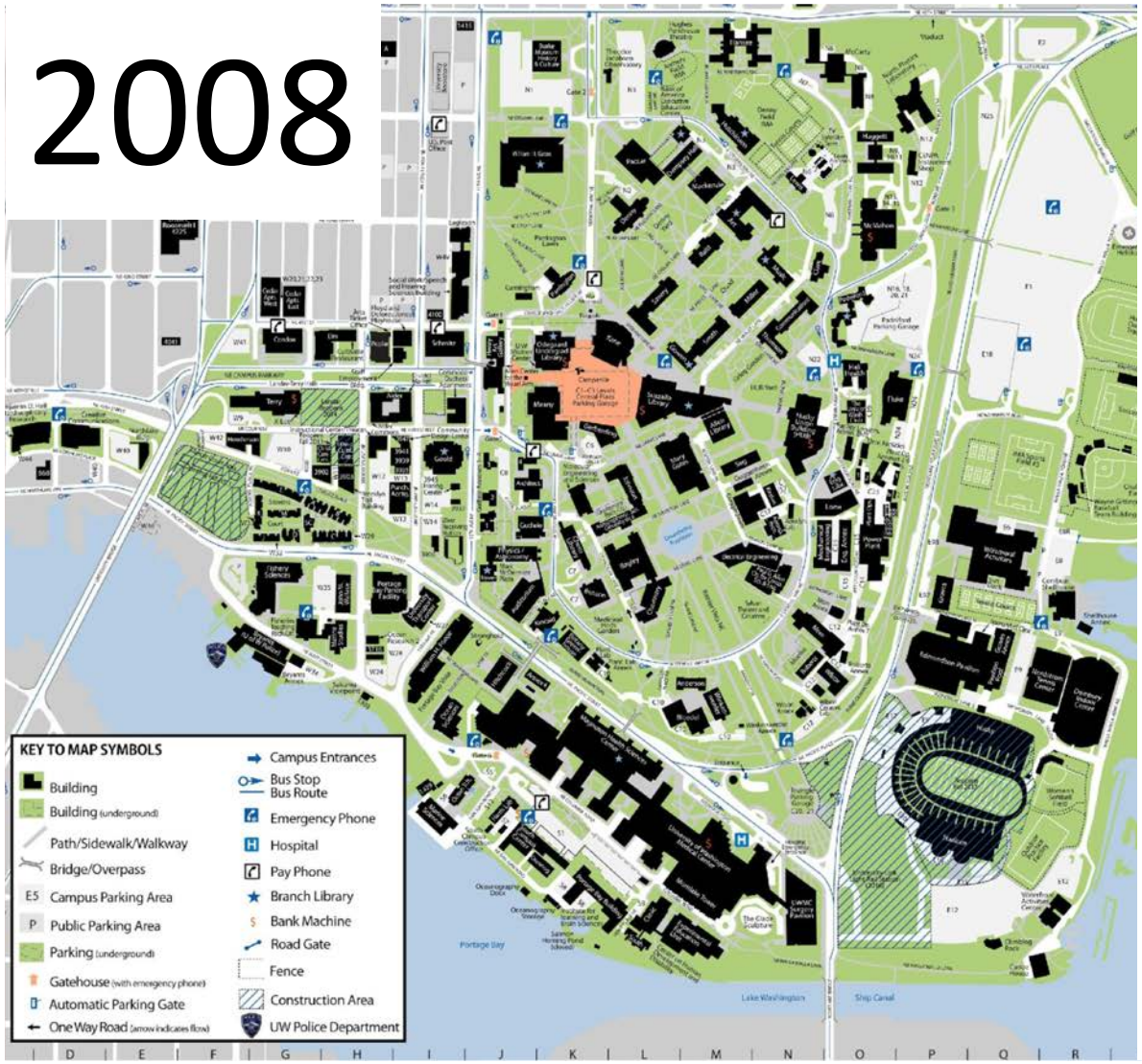
eScience Institute, University of Washington, Seattle, WA

James M. Carothers, Venkat R. Subramanian, and Jim Pfaendtner
Department of Chemical Engineering, University of Washington, Seattle, WA

DOI 10.1002/aic.15192
Published online February 28, 2016 in Wiley Online Library (wileyonlinelibrary.com)

History of data science @ UW

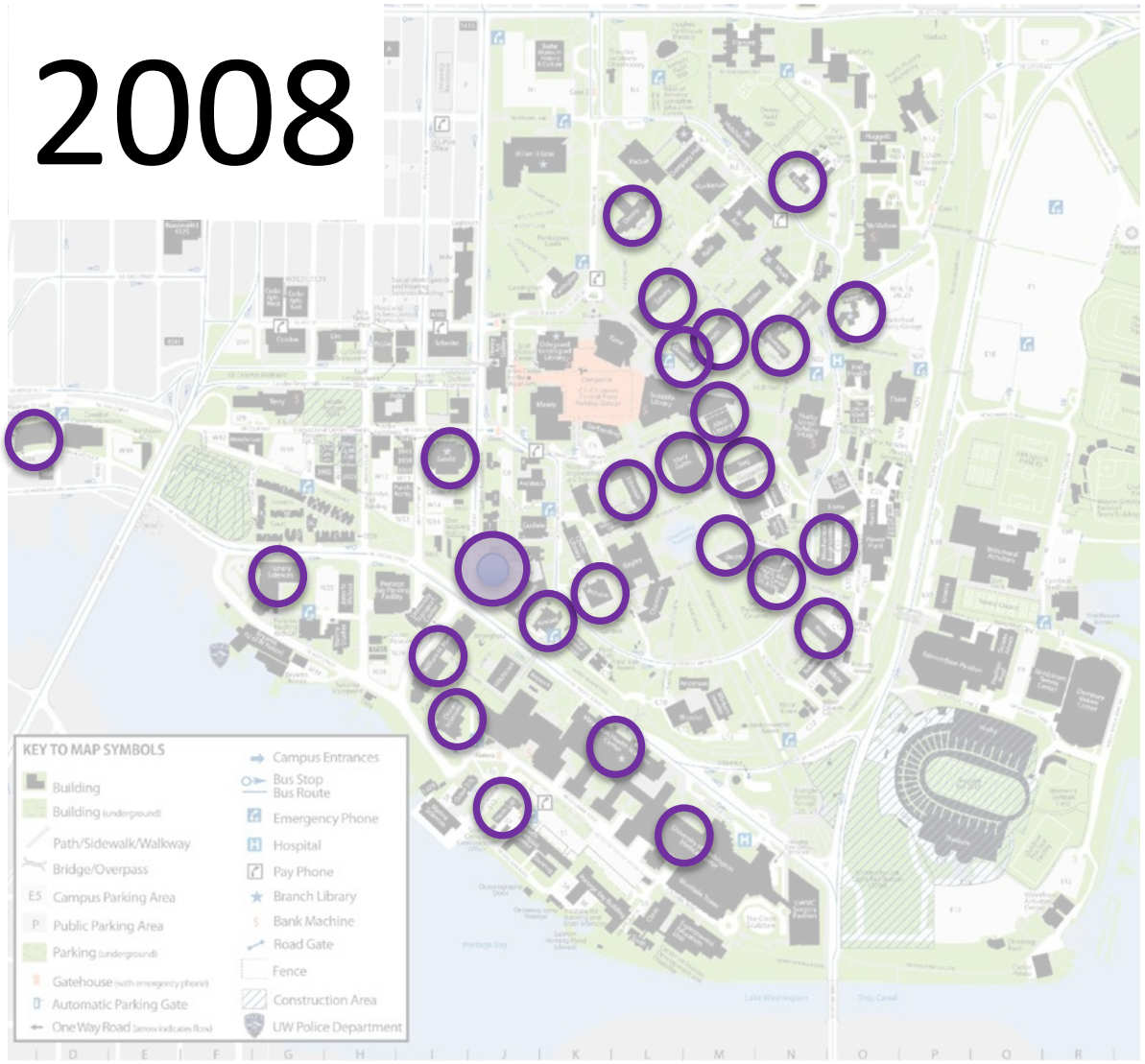
2008



- "Big Data" is a common buzzword
- Beyond "big" data
  - Complex data sets
  - High 'velocity' data
  - Veracity
- Something was missing…
  - What to do with the data?

History of data science @ UW

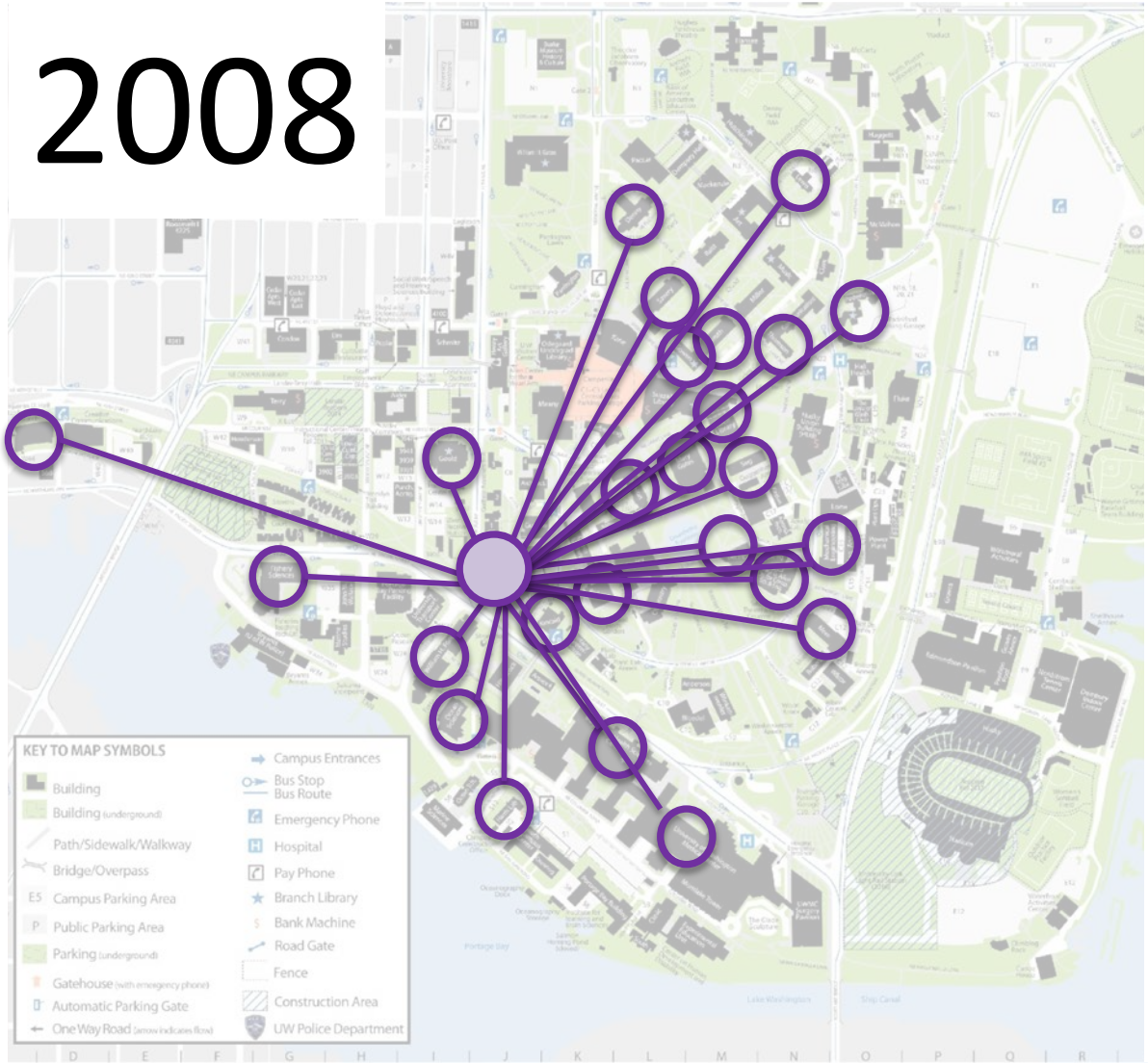
2008



- Many groups on the UW campus had data sets that were posing challenges
- **How to extract knowledge and actionable information from these data?**

History of data science @ UW

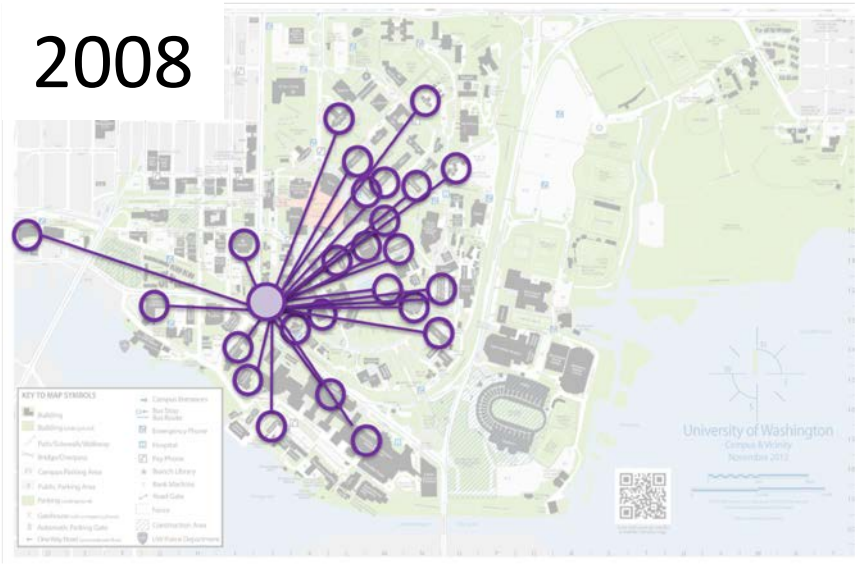
2008



- Convened groups from across campus to move beyond “big data” and into data driven modes of thinking (i.e. data science)
  - **Education**
  - Research
  - Community of practice

History of data science @ UW

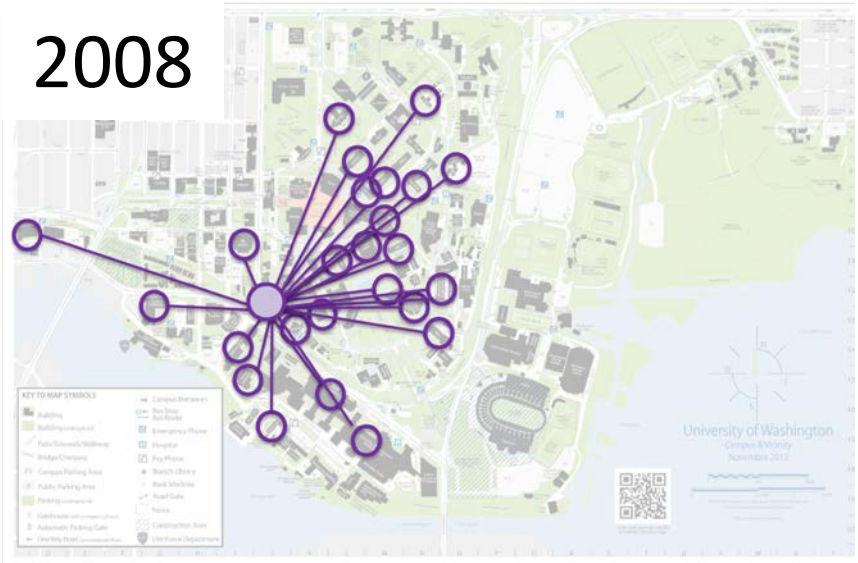
2008



- Core skills (i.e. data science):

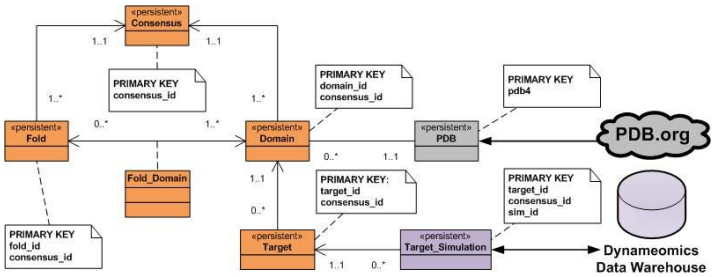
History of data science @ UW

2008



- Core skills (i.e. data science):
  - Data management

Relational databases (SQL)



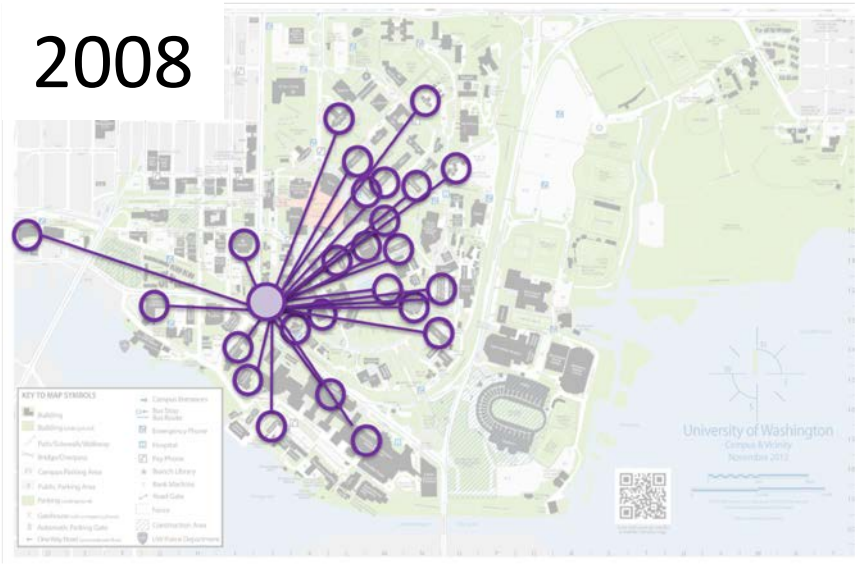
Data parallel computation

APACHE Spark™

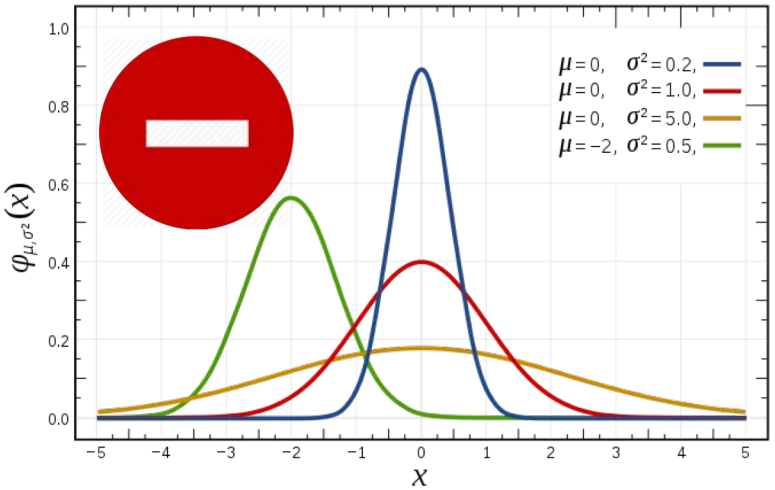
High volume streaming data

History of data science @ UW

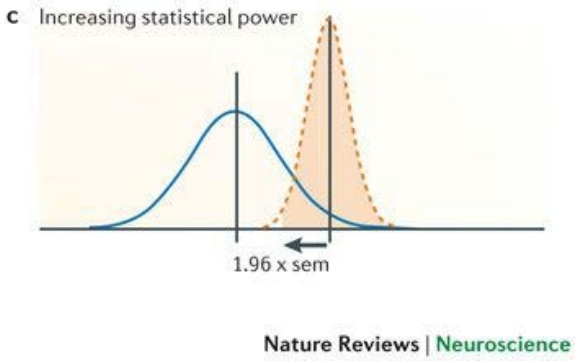
2008



- Core skills (i.e. data science):
  - Data management
  - Statistics

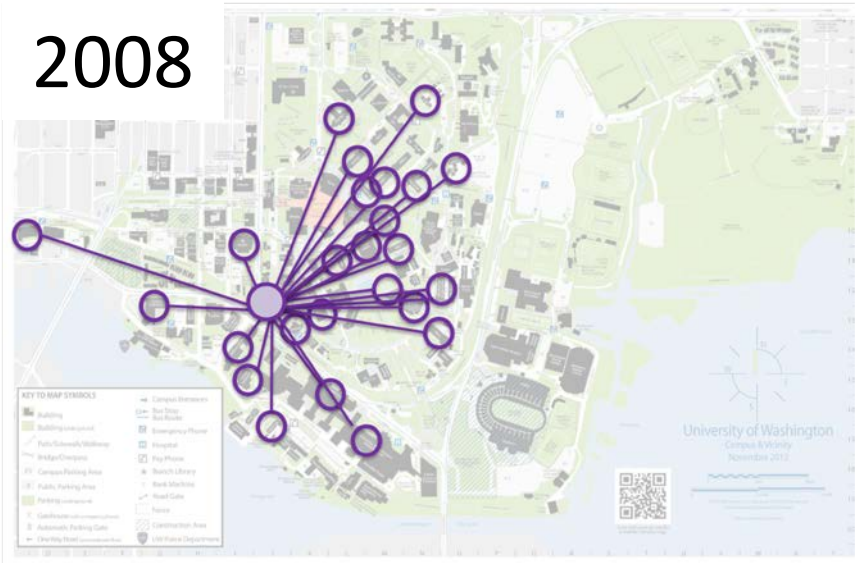


Wikipedia



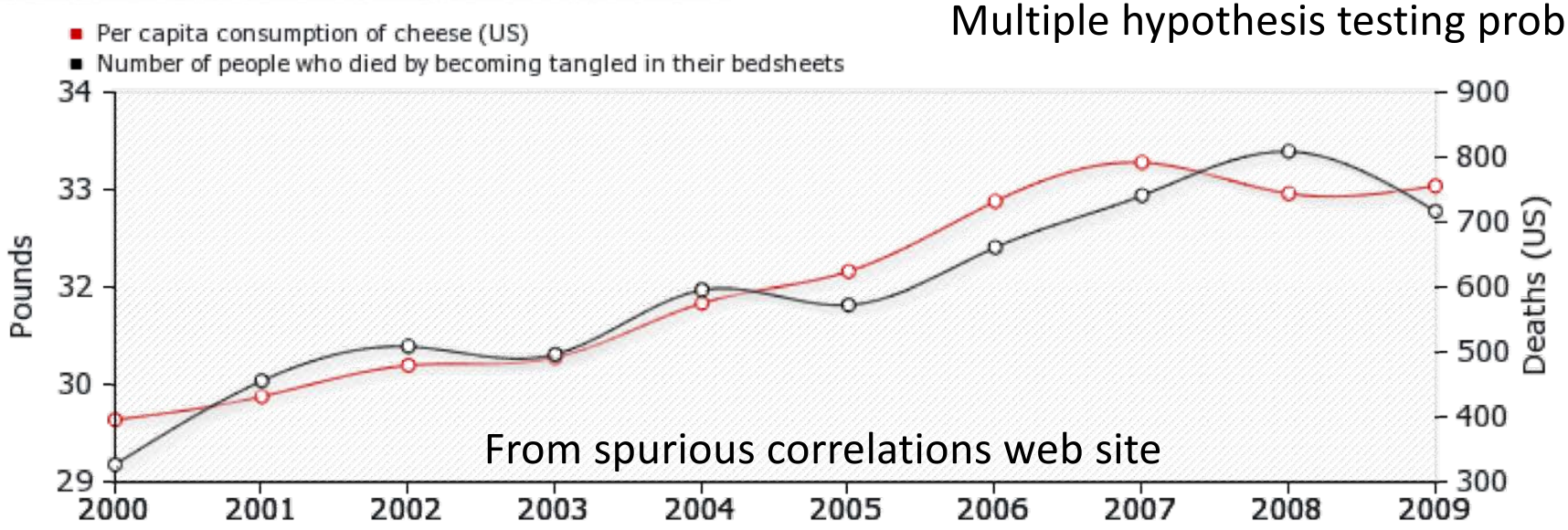
History of data science @ UW

2008



- Core skills (i.e. data science):
  - Data management
  - Statistics

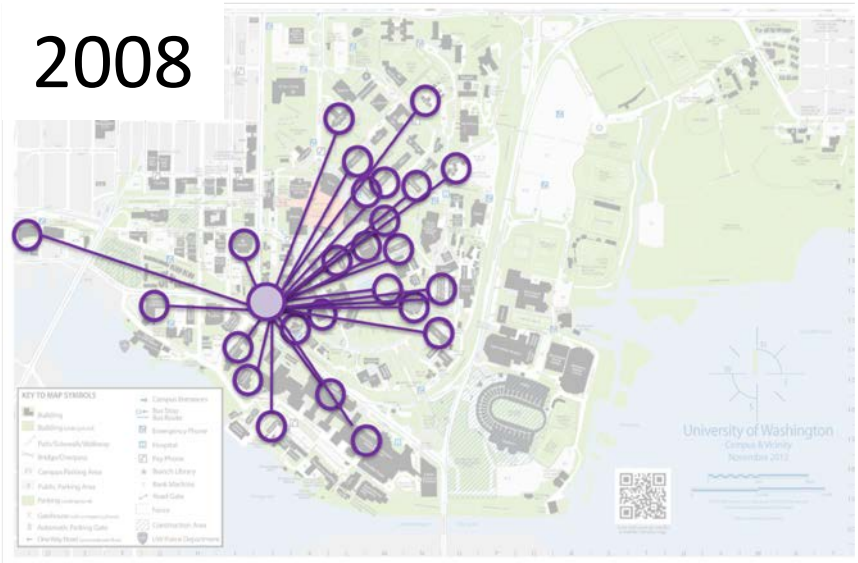
Multiple hypothesis testing problem



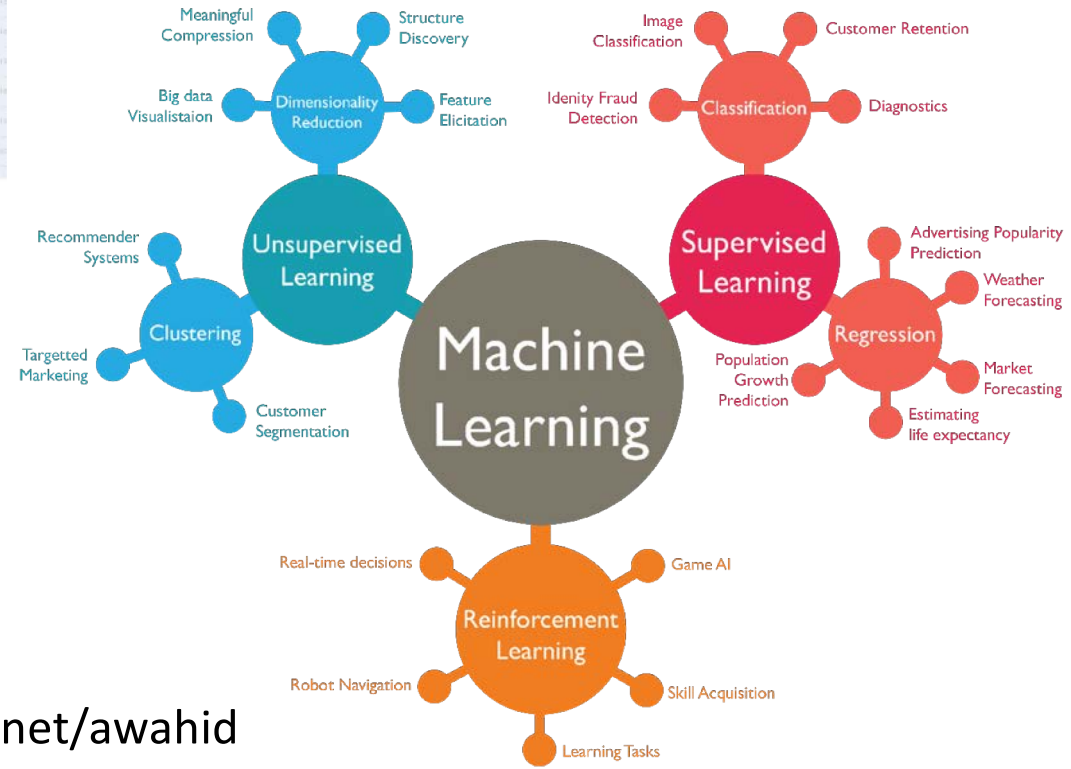
From spurious correlations web site

History of data science @ UW

2008



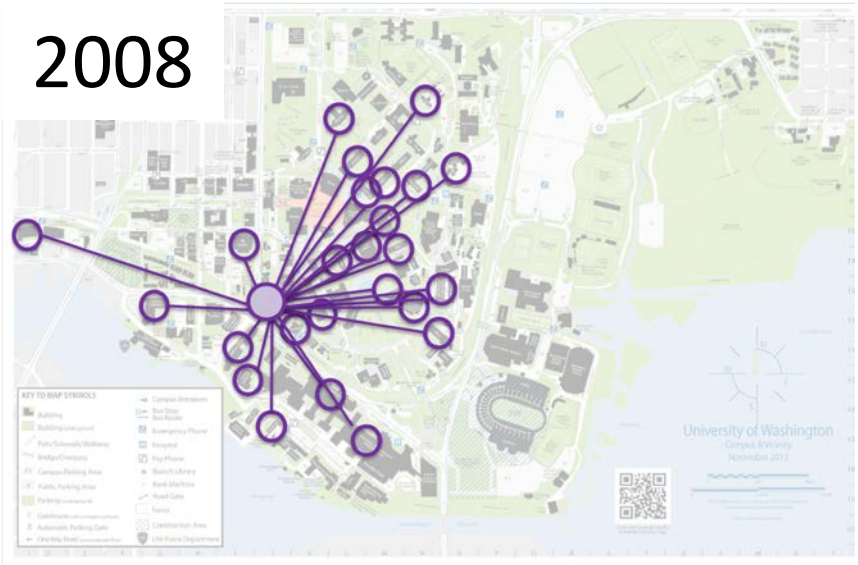
- Core skills (i.e. data science):
  - Data management
  - Statistics
  - Machine Learning



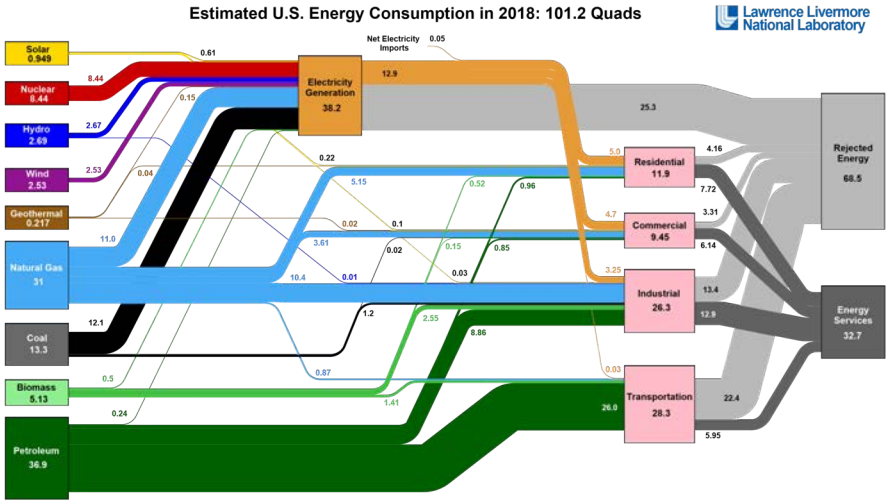
www.slideshare.net/awahid

History of data science @ UW

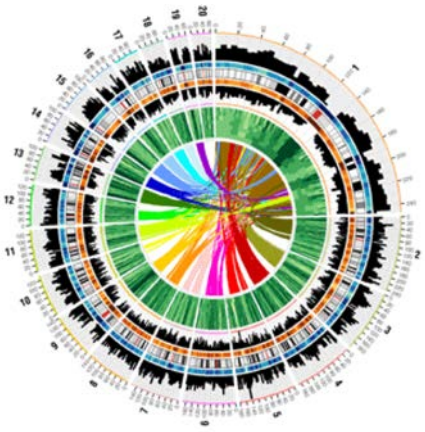
2008



- Core skills (i.e. data science):
  - Data management
  - Statistics
  - Machine Learning
  - Visualization



LLNL

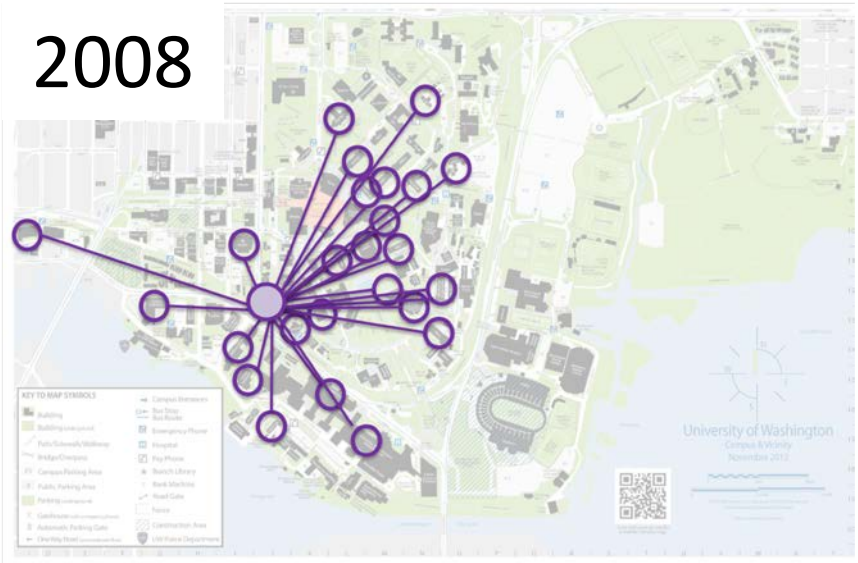


circos.ca

forbes.com

History of data science @ UW

2008

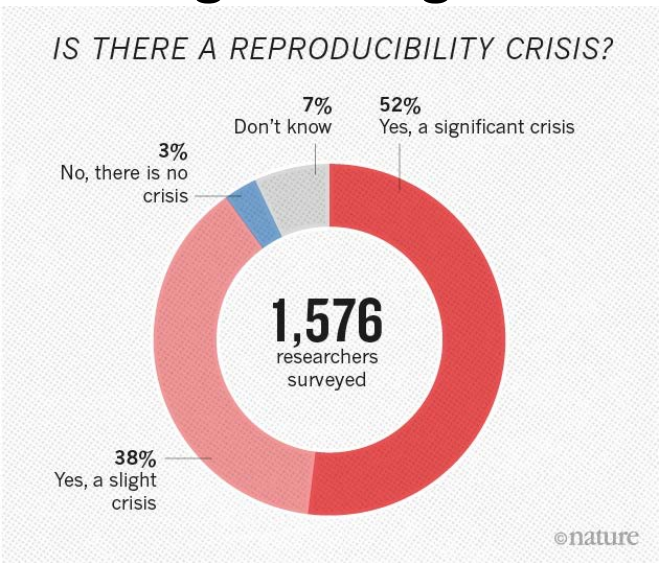


- Core skills (i.e. data science):
  - Data management
  - Statistics
  - Machine Learning
  - Visualization
  - Software Engineering

Version control

GitHub

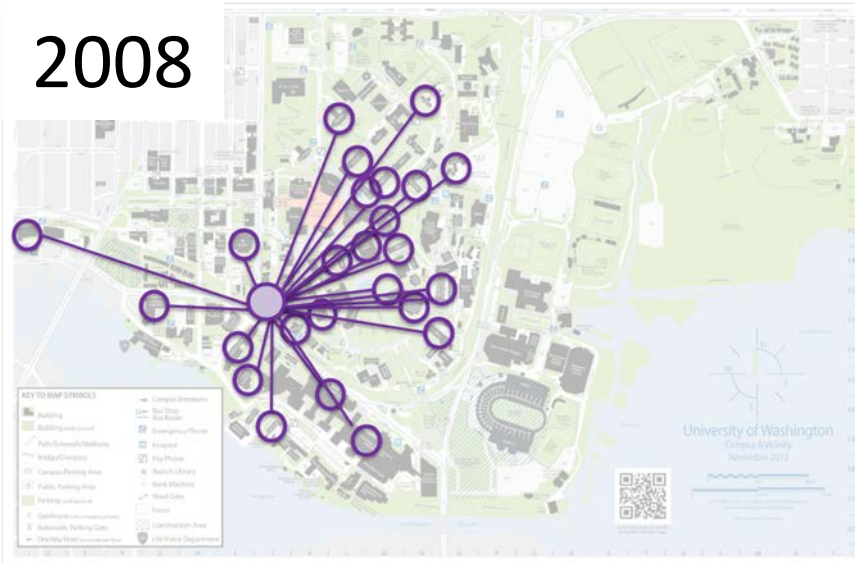
- Use cases & design
- Testing & verification
- Programming style
- Documentation



Manuscript = Patent = Software = Dataset

History of data science @ UW

2008



- Core skills (i.e. data science):
  - Data management
  - Statistics
  - Machine Learning
  - Visualization
  - Software Engineering

UNIVERSITY *of* WASHINGTON

eScience Institute

ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS

ChemE @ UW gets into data science education

- 2013 IGERT-CIF21: Big Data U: A Program for Integrated Multidisciplinary Education and Research for Big Data Science, #1258485
  - Astronomy, Biology, **Chemical Engineering**, Computer Science, Oceanography, Statistics
  - IGERTs directly funded graduate student trainees
  - Goal: Create π shaped scientists and engineers

π

Chemical Engineering

Data science

ChemE @ UW gets into data science education

- 2013 IGERT
- Students need to take 3 out of 4 core courses in methods
  - Statistics (STAT 509 or STAT 512-513)
  - Machine learning (CSE 546 or STAT 535)
  - Data management (CSE 544)
  - Data visualization (CSE 512)
- Participate in Data Science Community Seminar (CHEME 599)
- Data science lunch program

***500 level classes are graduate classes @ UW**

Cohort building activities

ChemE @ UW gets into data science education

- 2013 IGERT
- Departments can offer "**transcriptable options**" that add on to their degrees, e.g.

  PhD in Chemical Engineering with Advanced Data Science Option

- Process:
  - Departmental faculty vote to add an option
  - Graduate school reviews option with input from campus, e.g. eScience Institute

ChemE @ UW gets into data science education

- 2013 IGERT
- 2015 UW graduated the first Advanced Data Science Option (ADSO) student…

Chemical Engineering!

- Signals from this first student:
  - She completed the ADSO without receiving IGERT funding
  - She landed her “dream job” as a data scientist in a synthetic biology company
  - San Francisco company created an office in Seattle specifically to build a data science team

ChemE @ UW gets into data science education

- Lessons learned from the Advanced Data Science Option in ChemE
  - – Statistics grad level classes are **HARD**
  - – CSE grad level machine learning is even **HARDER**
  - – Not really available to our MS students
  - – 14+ ADSO offering units on campus (now), classes are highly subscribed and often wait listed
  - \+ There is demand from students for these skills
  - \+ They want to participate even without fellowships
  - \+ Employers really want our grads with data science skills
  - \+ Cohort based learning works for data science

ChemE data science for all

- 2016 NRT-DESE: Data Intensive Research Enabling Clean Technologies (**DIRECT**), #1633216
  - PI: Jim Pfaendtner (ChemE)
  - **ChemE**, Chemistry, Materials Science & Engineering, Molecular Science & Engineering program (2019)
  - NSF NRT is less about directly funding students and more about creating programs
  - Build a broadly accessible graduate data science education environment with a focus on clean energy / clean tech

CLEAN ENERGY INSTITUTE

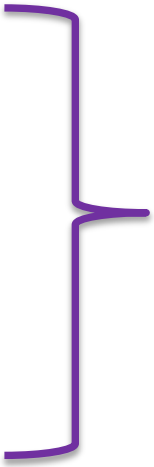
UNIVERSITY of WASHINGTON eScience Institute
ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS

Goals of DIRECT

- Students should be fluent in data science methods, best practices and tool development
  - E.g. they need to know how to choose a neural network architecture (FF ANN, RNN, CNN) but not how to derive a variable learning rate optimizer
  - E.g. they need to know how to use database query languages, but now how to build query planners
  - E.g. they need to know how to write software, use test-driven development, and perform code reviews but now how to build a compiler

Goals of DIRECT

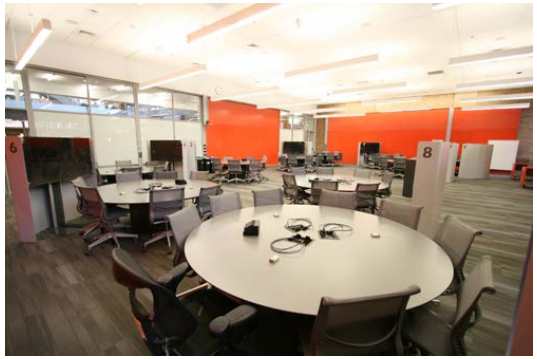
- Students should be fluent in data science methods, best practices and tool development
- No prerequisites
- 6 month intensive experience, 3 courses, 2 qtrs.
- Use project based learning to teach:
  - Software Engineering
  - Statistics
  - Machine Learning
  - Data Management
  - Visualization



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

Points of leverage in DIRECT

- Participating departments are molecularly focused
  - Contextualize data science in the language of molecules
  - E.g. talk about predicting molecular properties, not frenemy networks in twitter
- Use '**cohort effect**' to enhance learning experience
- Active learning classrooms



Points of leverage in DIRECT

- Participating departments are molecularly focused
  - Contextualize data science in the language of molecules
  - E.g. talk about predicting molecular properties, not frenemy networks in twitter
- Use '**cohort effect**' to enhance learning experience
- Active learning classrooms
- UW's 'transcriptable options'
- Huge set online learning resources for **Python**

DIRECT course overview

- Three courses make up the 'Data Science Option'
  - CHEME 546: Software Engineering for Molecular Data Scientists (SEMDS)
    - Winter quarter (10 weeks)
  - CHEME 545: Data Science Methods for Clean Energy Research (DSMCR)
    - Winter quarter (10 weeks)
  - CHEME 547: Molecular Data Science Capstone
    - Spring quarter + Summer A-term (14 weeks)

Cohort from colearning

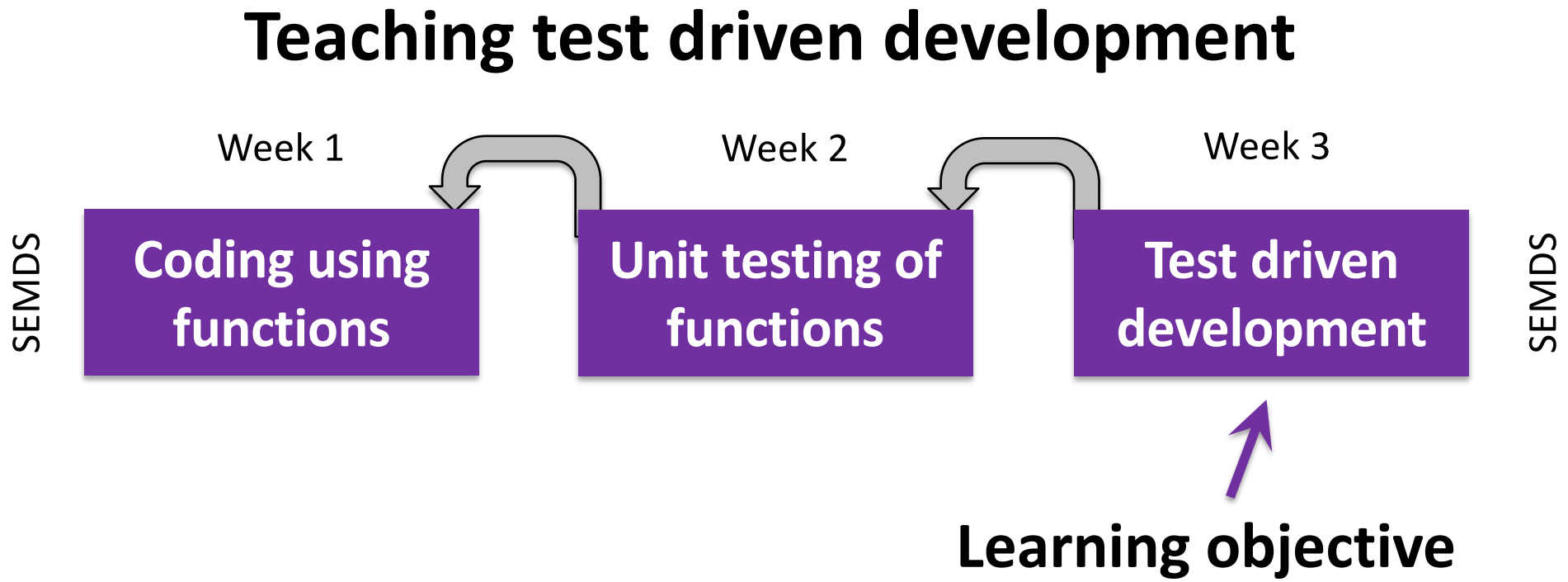
- SEMDS (Soft. Eng.) & DSMCER (DS Methods)
  - Run concurrently
  - All students take both courses at the same time
  - 6 hours a week contact time with instructor
  - +2 hours a week office hours with instructor & TAs
  - Slack organization & channels for questions and student help, peer support, “clicker” in class, study groups, even socialization

Cohort from co-learning

- SEMDS (Soft. Eng.) & DSMCER (DS Methods)
  - Run concurrently
  - All students take both courses at the same time
  - 6 hours a week contact time with instructor
  - +2 hours a week office hours with instructor & TAs
  - Slack organization & channels for questions and student help
  - **Group based project is shared across both courses**

Why concurrent courses?

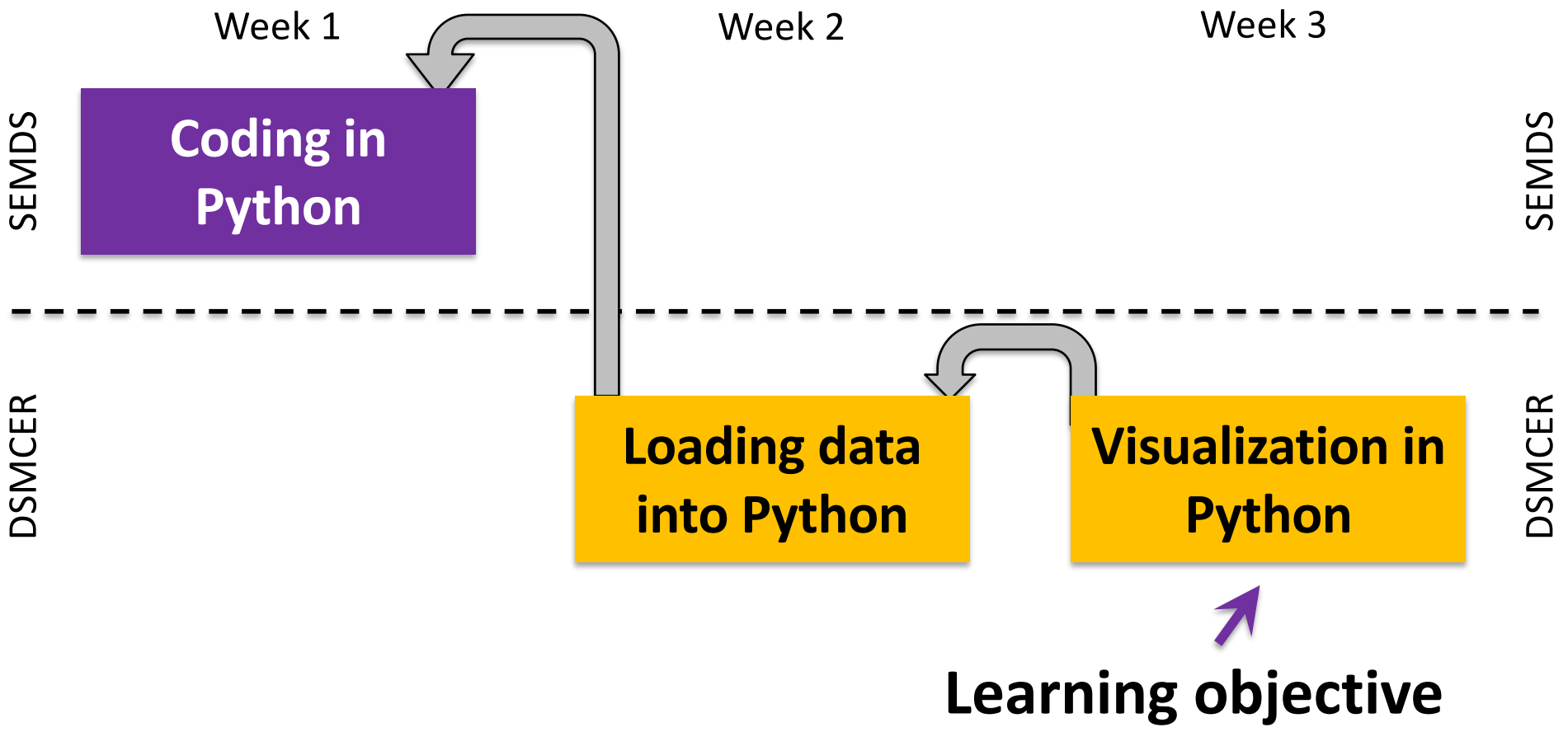
- Dependency graph of learning objectives



Why concurrent courses?

- Dependency graph of learning objectives

Teaching visualization in Python w/ no prereqs



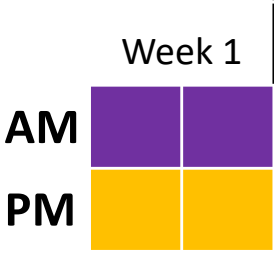
Why concurrent courses?

- Dependency graph of learning objectives can cross course boundaries
- Shortens the length of the dependency in real world time vs. sequential courses
  - Higher retention of concepts
  - Immediate practice of concept in application
- Why not one large course?
  - Students arrive with different baselines
  - Different courses enable differential feedback on strengths and weaknesses

Syllabus for SEMDS & DSMCER

SEMDS

DSMCER

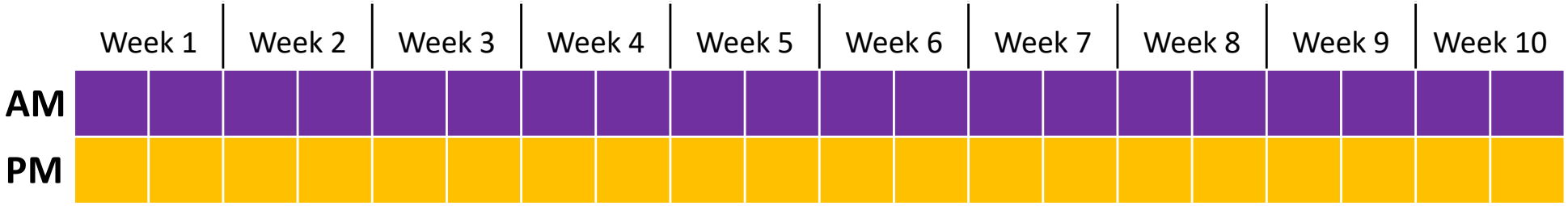


Two courses × two 1.5 hr classes / week = 6 hrs / wk

Syllabus for SEMDS & DSMCER

SEMDS

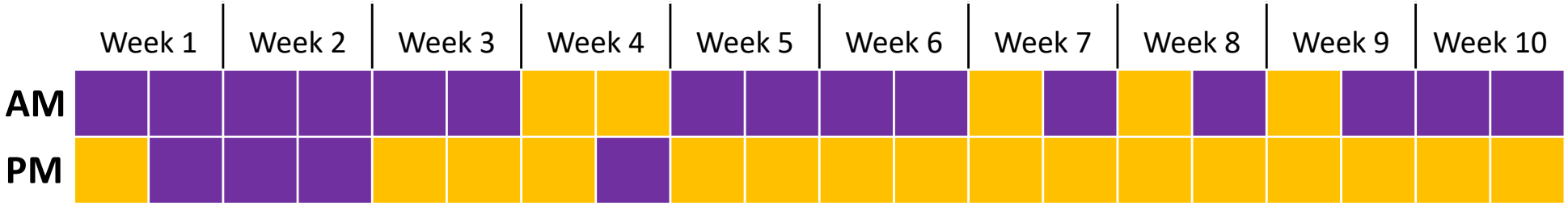
DSMCER



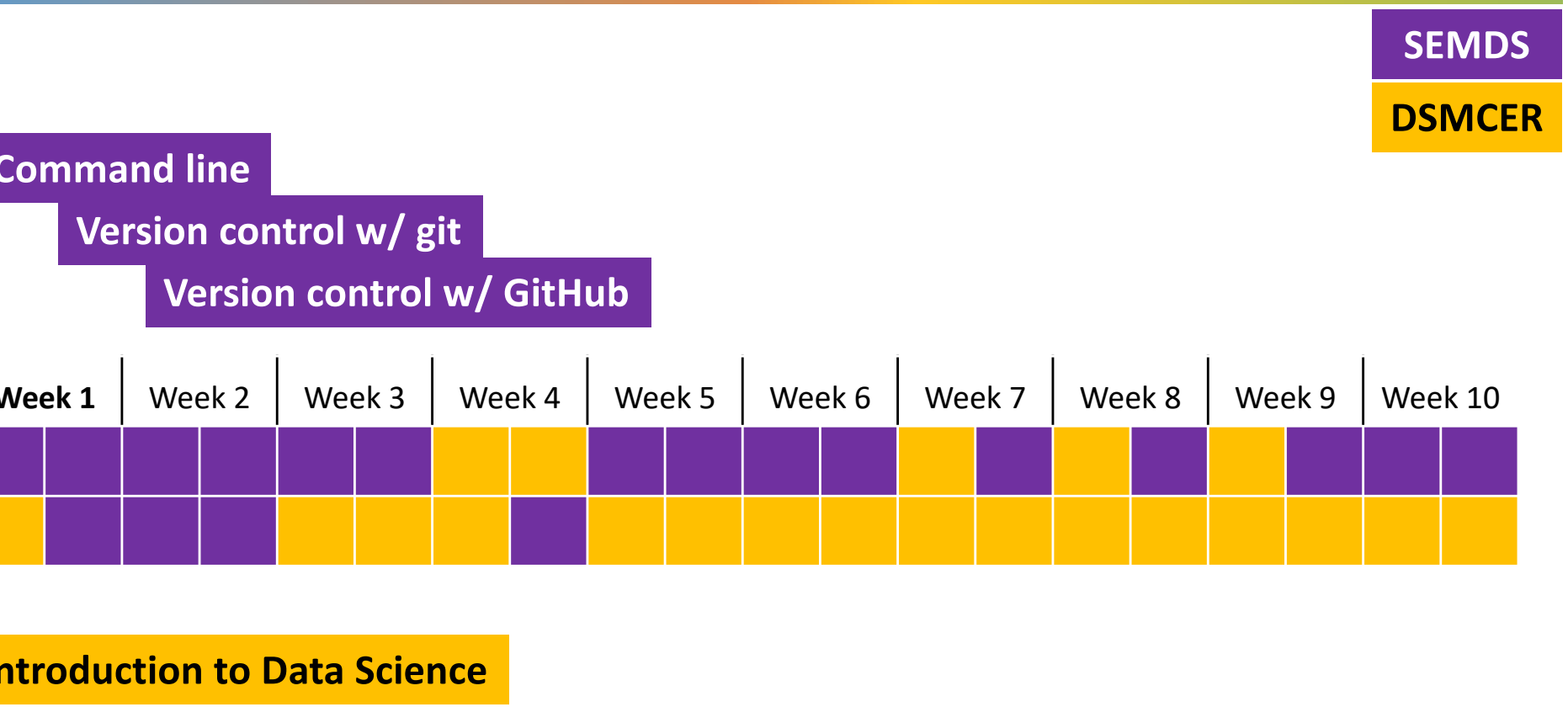
Syllabus for SEMDS & DSMCER

SEMDS

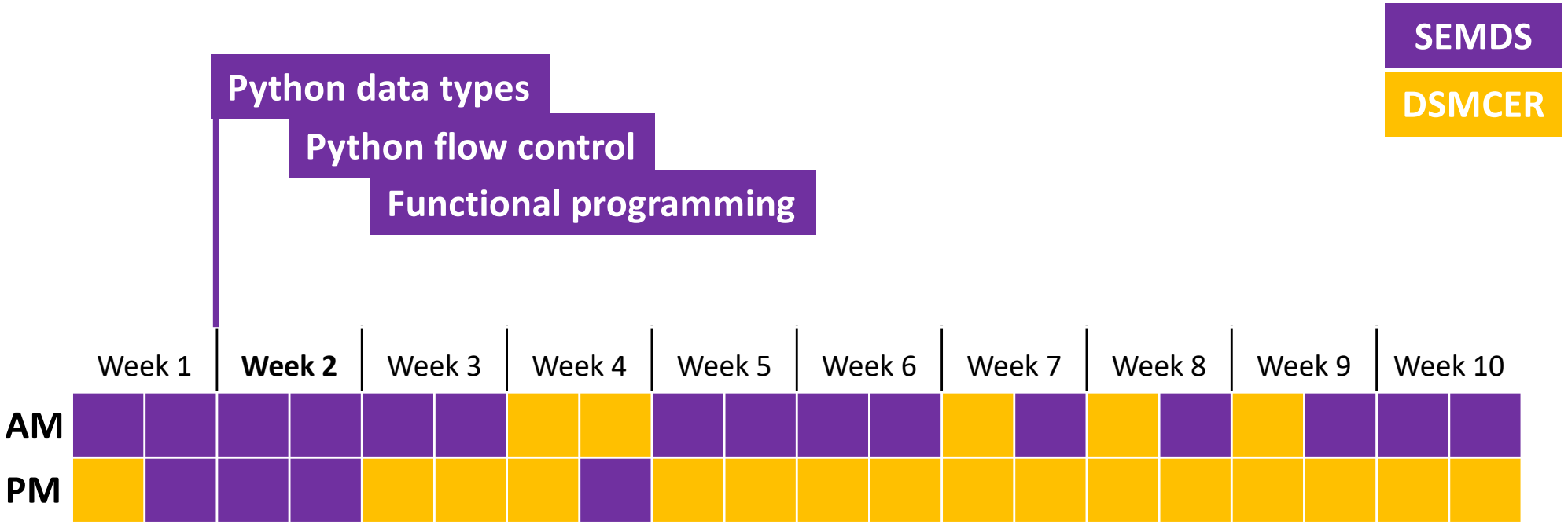
DSMCER



Syllabus for SEMDS & DSMCER



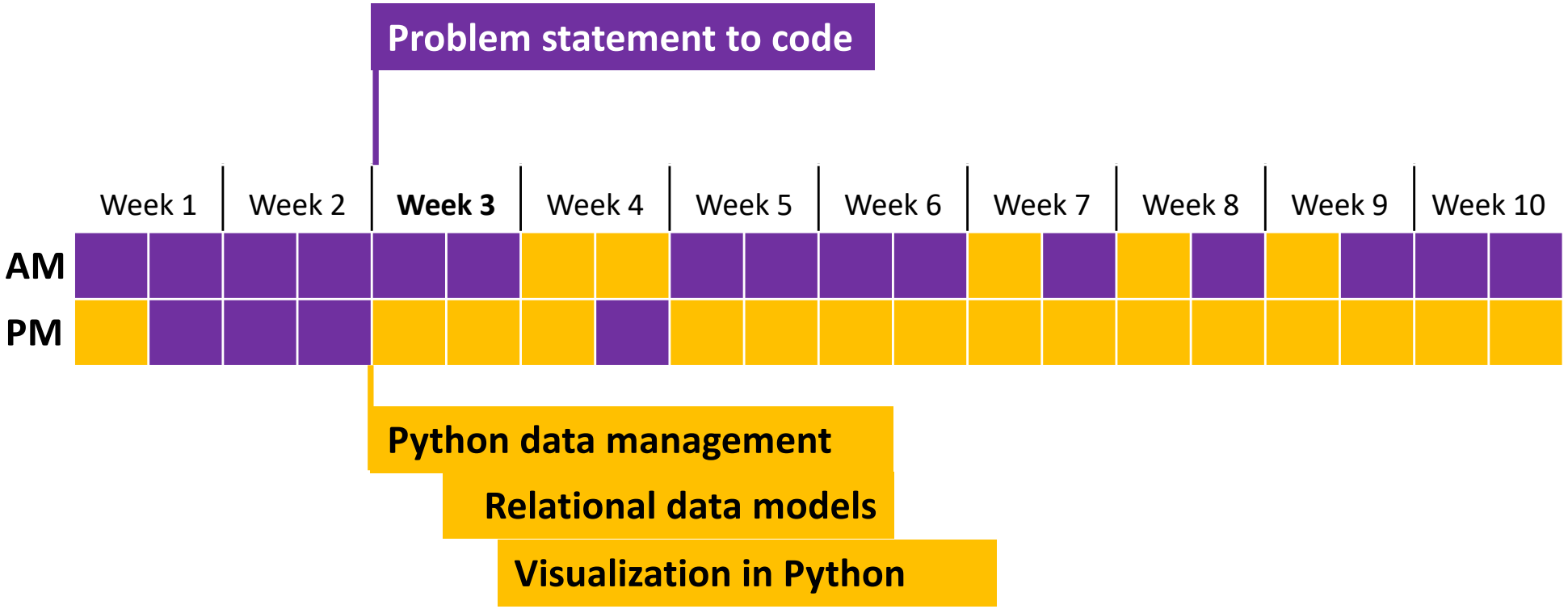
Syllabus for SEMDS & DSMCER



Syllabus for SEMDS & DSMCER

SEMDS

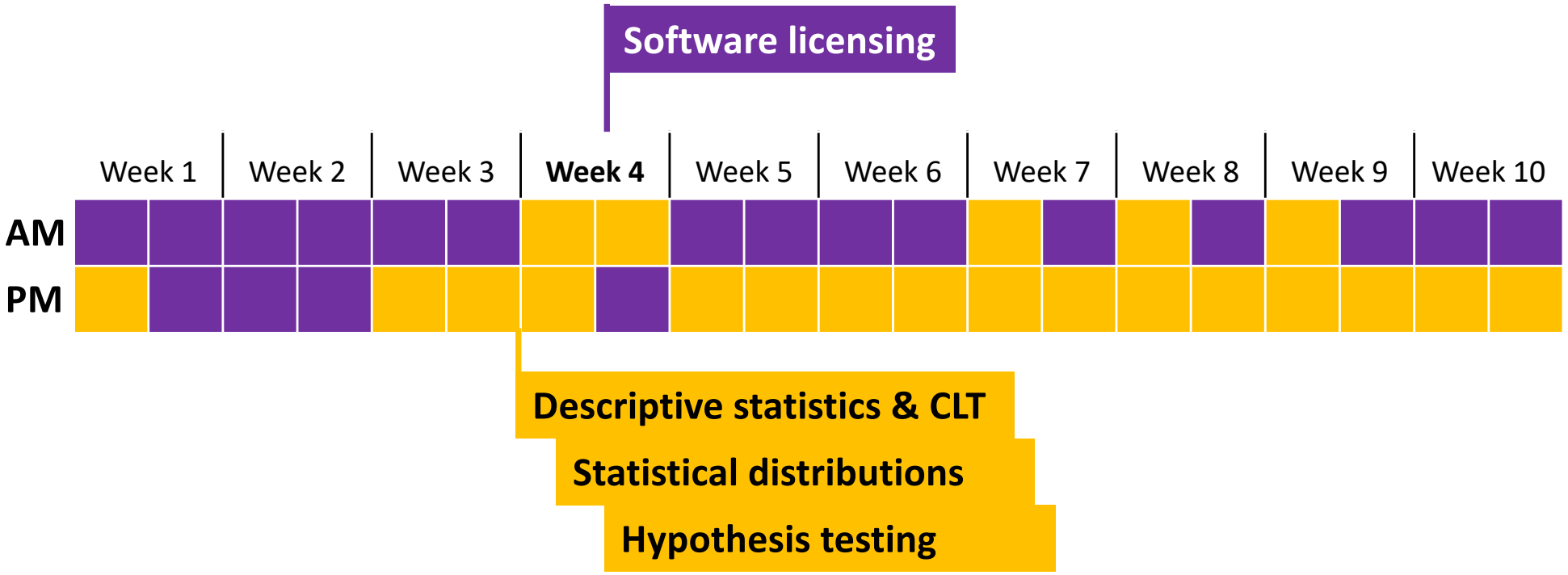
DSMCER



Syllabus for SEMDS & DSMCER

SEMDS

DSMCER



Syllabus for SEMDS & DSMCER

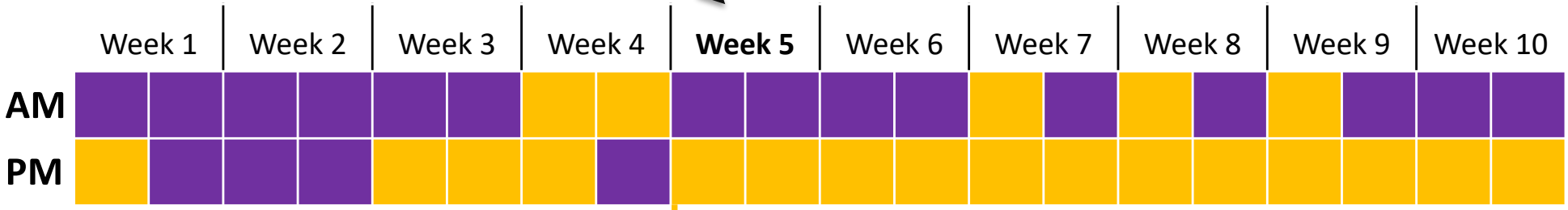
SEMDS

DSMCER

Project team formation proceeded by Slack discussions and in class student presentations

Unit testing & continuous integration

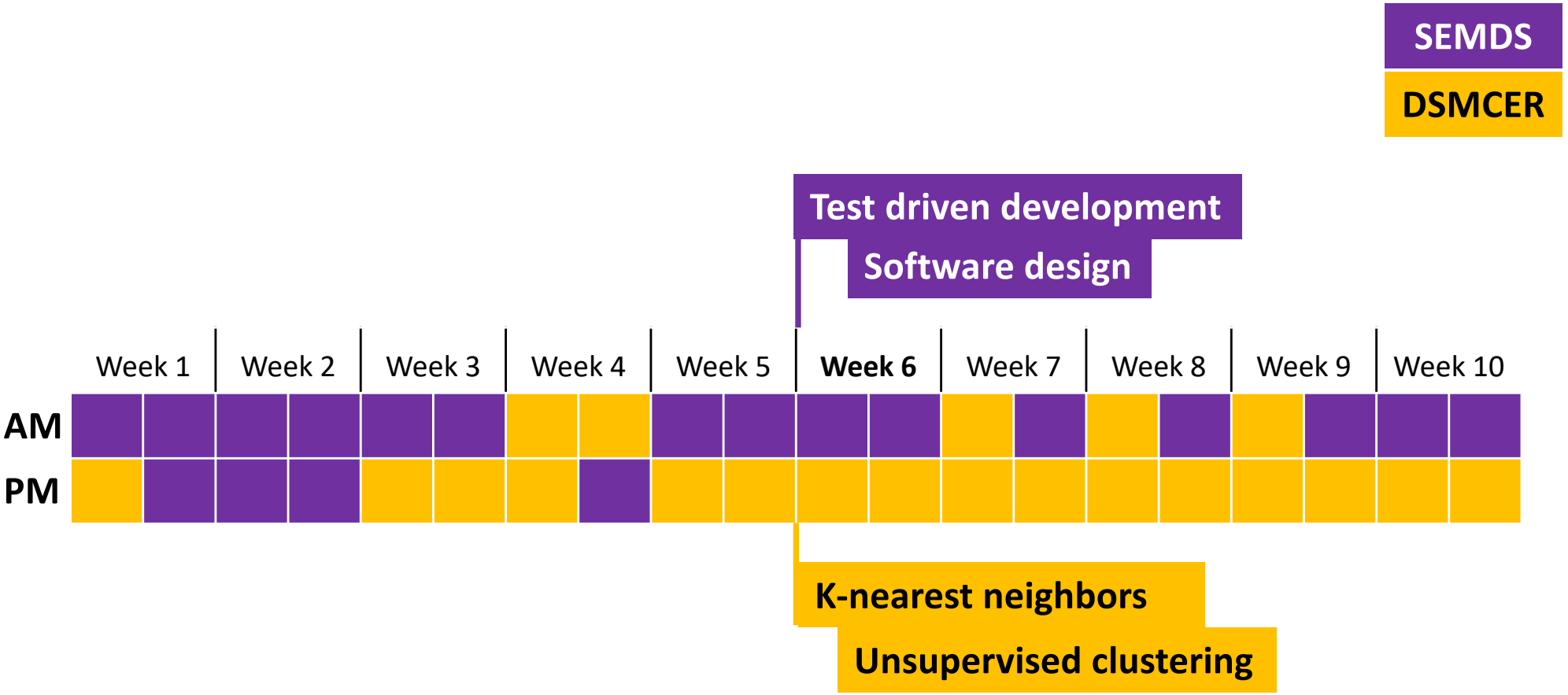
Documentation & programming style



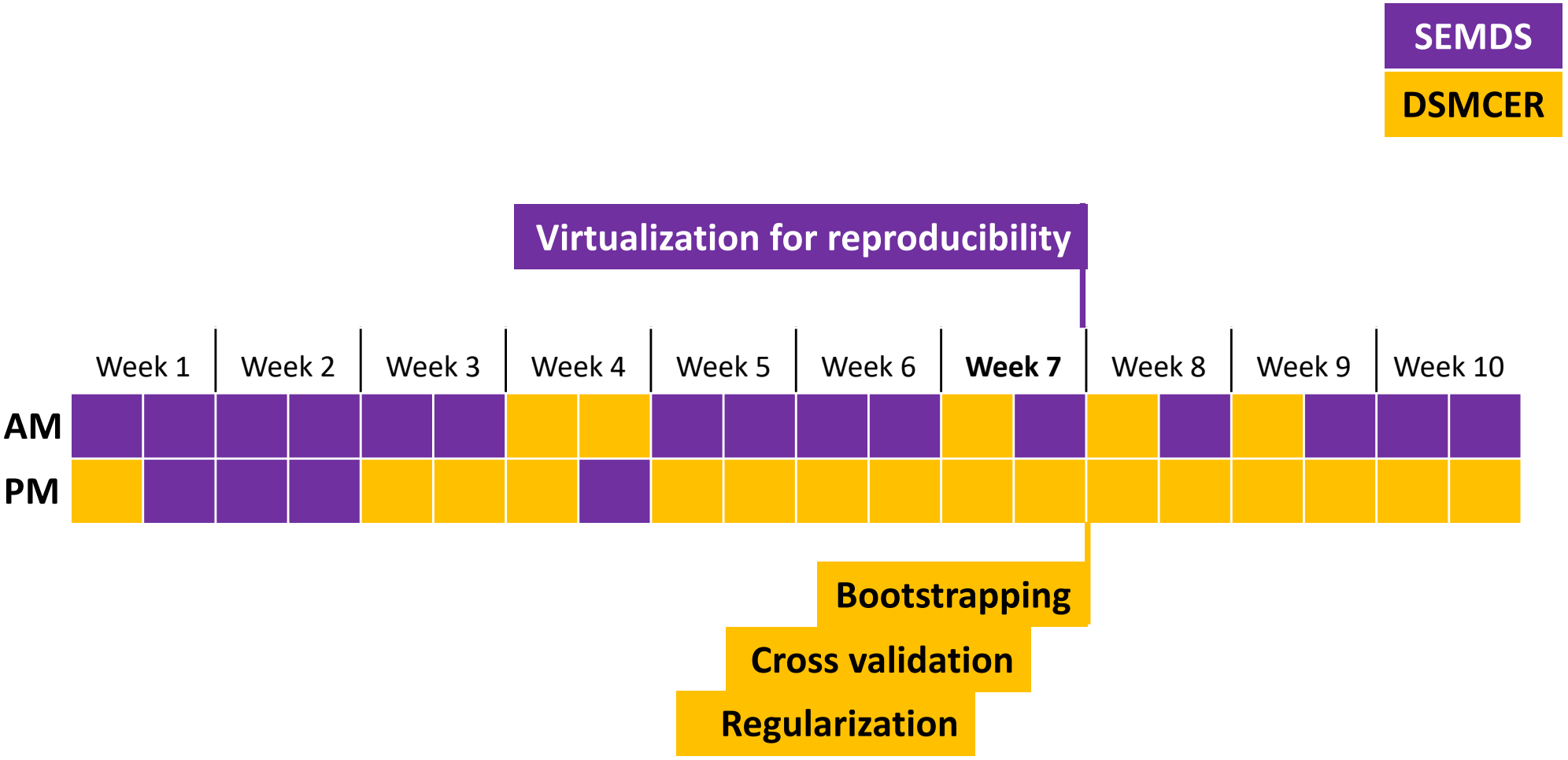
Linear regression

Bias variance tradeoff

Syllabus for SEMDS & DSMCER



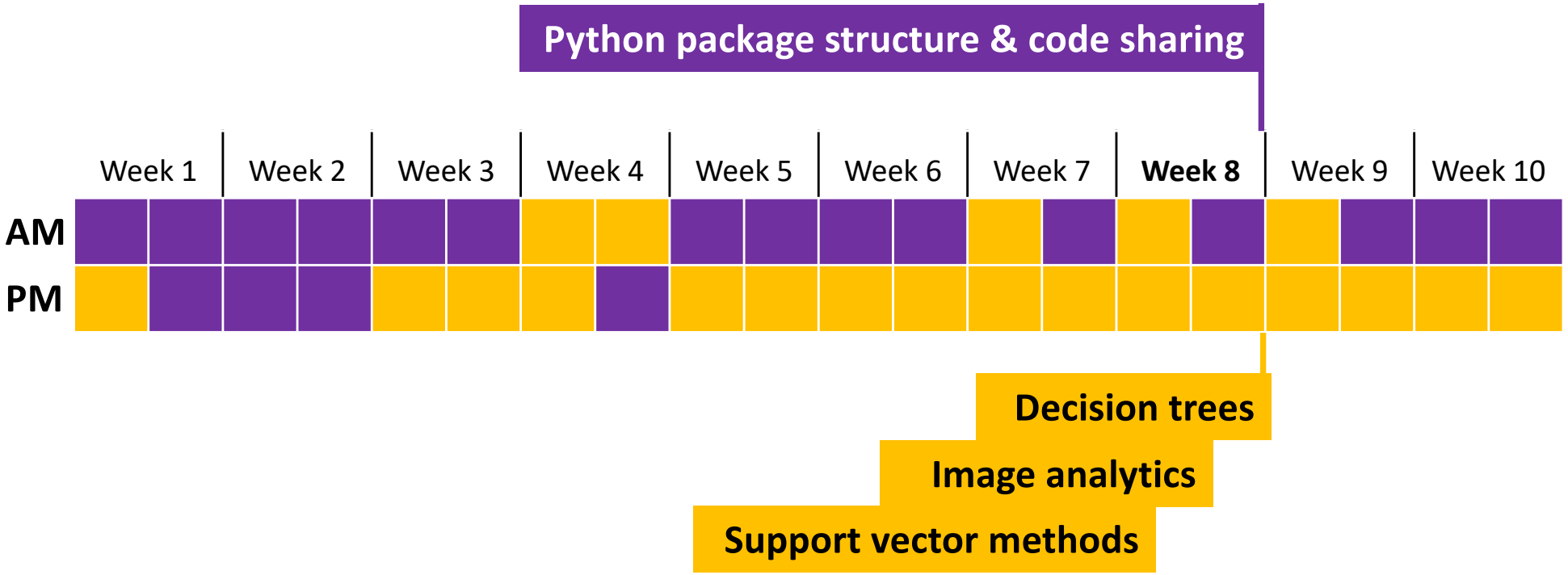
Syllabus for SEMDS & DSMCER



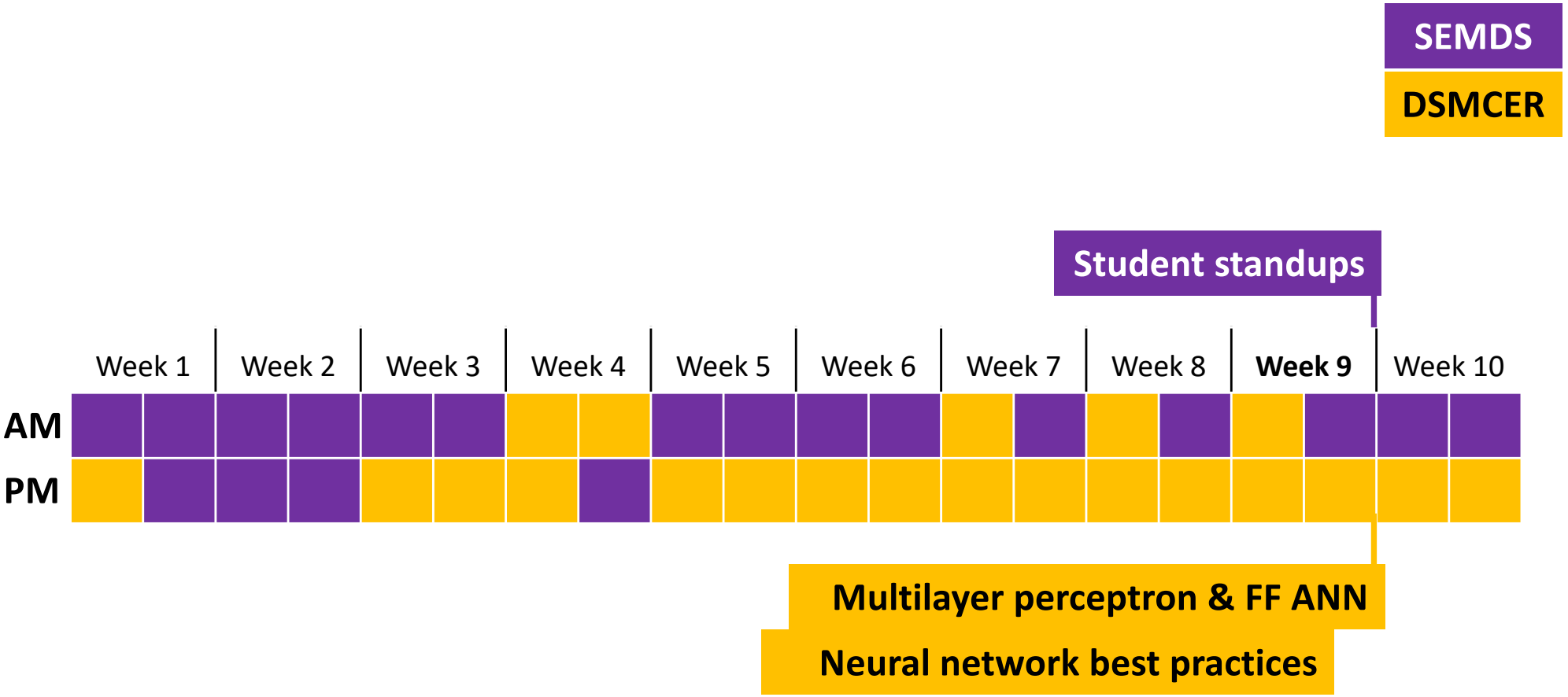
Syllabus for SEMDS & DSMCER

SEMDS

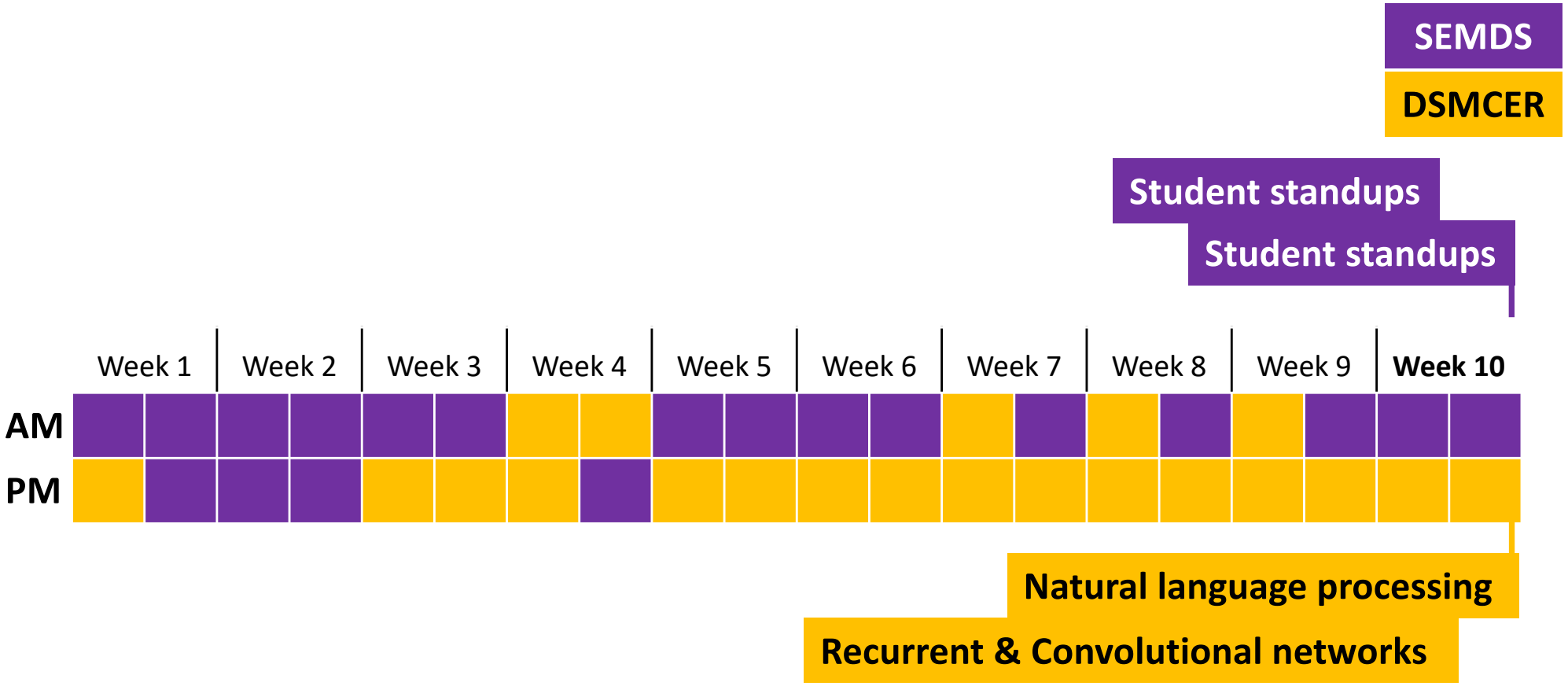
DSMCER



Syllabus for SEMDS & DSMCER



Syllabus for SEMDS & DSMCER

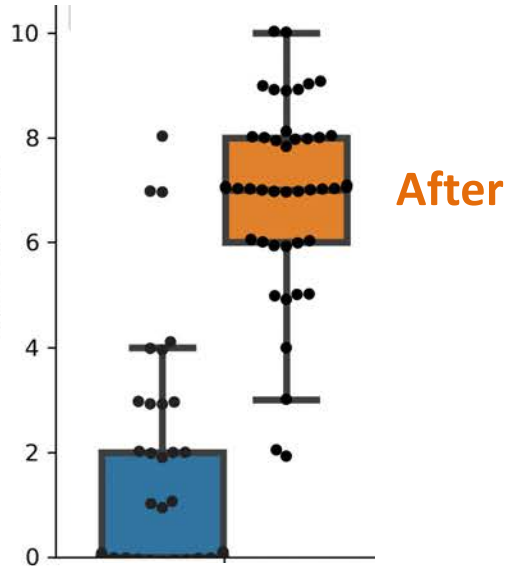


Homework: dual task learning

- Five homework assignments for each course
  - E.g. write a K-nearest neighbor classifier for selecting force field parameters
- All homework is submitted using GitHub
  - Reinforces the usefulness of version control
  - Reinforces the technical knowledge

Students can revise homework source code for up to two weeks after submission

Confidence level 0-10 using version control



Projects for SEMDS & DSMCER

- Same project for both courses
  - Differential grading rubrics
  - Double student hour efforts enabling adv. projects
- Criteria for the student project
  - Topic should be molecular* or clean tech focused
  - Must utilize two or more non-trivial data sets
  - Teams should be 4 members (3 & 5 discouraged)
  - Students must use Python & software design, test driven development, documentation, style
  - Students must use best practices for DS methods

Projects for SEMDS & DSMCER

- Projects are presented at poster session where all faculty, chairs, and previous student cohorts are invited

- E.g. posters…

In-Situ Raman Spectroscopy Component Identification for Machine Learning Based Decomposition Analysis

DIRECT
Data Intensive Research Enabling Clean Technologies

CLEAN ENERGY INSTITUTE
UNIVERSITY of WASHINGTON

eScience Institute
ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS

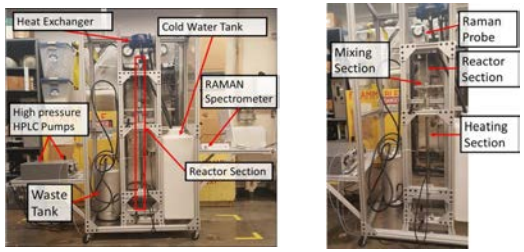
Introduction

- Using data from a custom built supercritical gasification reactor on campus to analyze formic acid Raman spectra.
- Decomposition of formic acid constitutes the combination of two pathways:

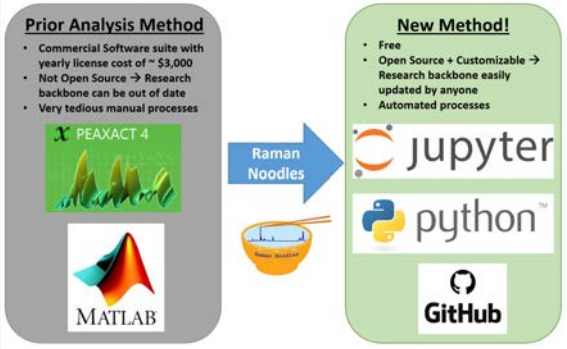
$$HCOOH \rightarrow H_2 + CO_2$$

$$HCOOH \rightarrow H_2O + CO$$

Photographs of UW Gasification Reactor with In-Situ Raman Probe



Motivation



Goals

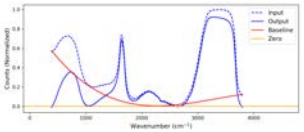
1. Data Mining and Baseline Subtraction
   - Importing open source data sets, create a library of spectra, uniformly format data for analysis
2. Data Visualization
   - Outputting plots of baseline subtraction and peak identification
3. Machine Learning
   - Prepare least squares regression model for calculating kinetic rate decomposition at different resonance times and temperatures

Materials and Methods

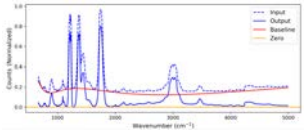
Baseline Subtraction of Raman Signal

Used PeakUtils built in baseline function to perform polynomial fit baseline subtraction on NIST database spectra. Examples of this functionality are shown below.

Water Baseline Subtraction Output



Acetone Baseline Subtraction Output



Component Analysis in a Mixture Raman Signal

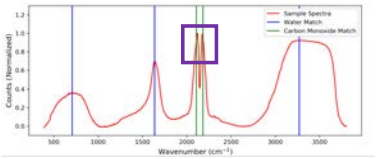
The LMFit package was utilized to identify 5 descriptors per peak in a mixture's Raman signal including: location of the peak, peak height, and peak width. As a mixture has more peaks (components from decomposition) the amount of descriptors increases.

Least Squares Model Equation

$$f(x) = \sum_{i=1}^{n} f_i(x, A, \mu, \sigma)$$

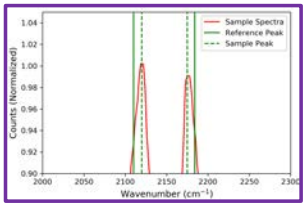
Lorentzian Equation for a Single Peak

$$f_i(x, A, \mu, \sigma) = \frac{A}{\pi}\left[\frac{\mu}{(x-\mu)^2+\mu^2}\right]$$



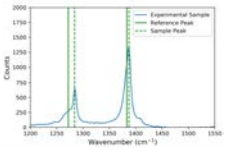
Component Confidence based on Euclidean Distance

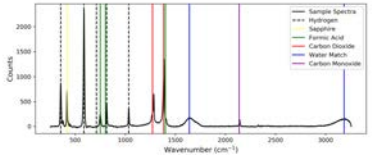
Once components in a 'testing' dataset mixture are identified the next step is to take a 'training' dataset location and share with the user the Euclidean distance between the two datasets. For this software if a peak location is more then ± 10 cm^{-1} from the literature values the confidence that the peak represents the compound is zero. This range was set from experimental considerations.



Results Using Experimental Data

After creating a theoretical NIST 'training' datasets and proving the functionality of the code the final step was to test it on experimental data sets taken in the lab.



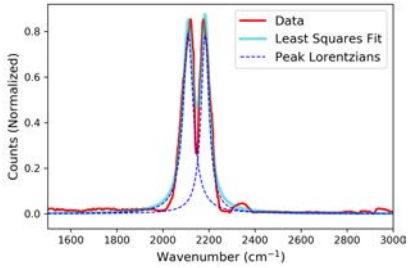


Continuing Work

Machine Learning for Material Decomposition

- We implemented functionality to run a least squares regression that fits Lorentzian curves to the data. The function is given the peak locations determined using scipy.signal.find_peaks.
  1. Expand software to be able to compute decomposition rates across varying parameters such as temperature, resonance time, possibly pressure.
  2. From the defined decomposition rate the software can predict the decomposition rates using machine learning beyond the known data set limits

Carbon Monoxide Spectra Lorentzian Output



Conclusions and Future Work

In conclusion our team successfully created a fast functioning open source code base that saves hours of research time in data cleaning and analysis of Raman Spectra. We have also set a strong base for the next step of our focus which is on calculating decomposition of substances using Lorentzian peak information that will be applied to machine learning optimum temperatures and pressures in a gasification reactor system.

Future work

This work sets up a free and user friendly platform for researchers to be able to analyze their own Raman Spectra.

Acknowledgements

- Dave Beck, Chad Curtis, and Kelly Thornton
- Data sets were taken from publicly available from the NIST WebBook Database and Mendeley Data, "Raman Spectra of Formic Acid Gasification Products in Subcritical and Supercritical Water"
- Only open source packages were used in this work, documentation of all packages used can be found at our GitHub at:
  - https://github.com/raman-noodles/Raman-noodles

Spectra Prediction for the Excitation and Emission of Dyes and other Conjugated Organic Molecules

UW MEM·C

UNIVERSITY of WASHINGTON

Overview

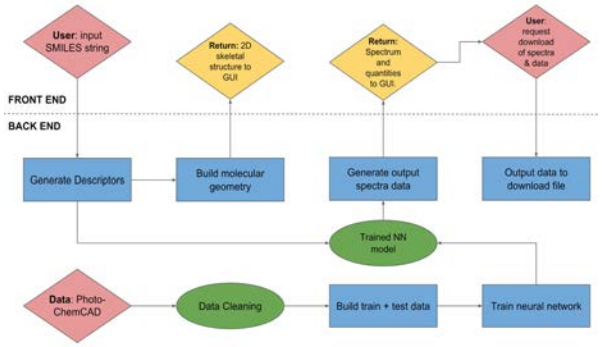
SPEEDCOM is an open-source python package that uses deep learning methods to predict the absorption and emission spectra of small organic molecules.

GitHub: https://github.com/emissible/SPEEDCOM

Motivations

The use of *ab initio* methods to calculate molecular spectra is usually lengthy, expensive, and may even be inaccurate depending on the choices for the level of theory. As such, a fast, experimental, data-derived method for predicting excitation and emission spectra for organic species is proposed to aid in rapid prediction of spectral features. This has potential uses in applications such as fluorophore-design.

Use Cases



***Via* a GUI, users can...**

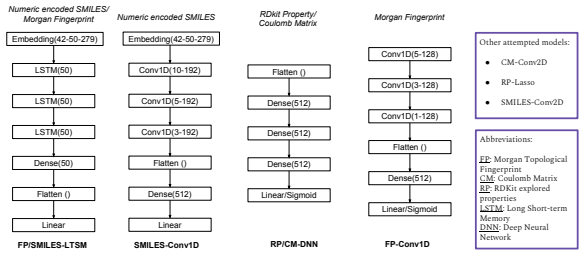
- Input the SMILES string of a given molecule
- Visualize the 2D skeletal structure of this molecule
- Visualize and download predicted spectra and associated characteristics such as the quantum yield and molar extinction coefficient.

Data Cleaning

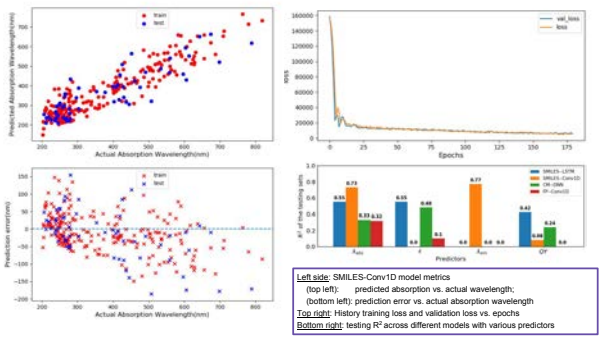
To obtain the dataset used for training, the files from the database were parsed to:

- Obtain the absorption and emission spectra;
- Obtain the smiles strings for molecules using *pubchempy* package;
- Removing extraneous counter ions from generated SMILES strings;
- Generating descriptors using *RDkit* package:
  - Coulomb Matrix of nuclei
  - Morgan Topological Fingerprint
  - Molecular Properties

Model Architectures



Metrics

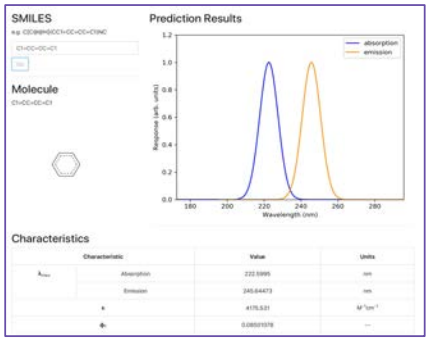


Left side: SMILES-Conv1D model metrics
(top left): predicted absorption vs. actual wavelength;
(bottom left): prediction error vs. actual absorption wavelength
Top right: History training loss and validation loss vs. epochs
Bottom right: testing R^2 across different models with various predictors

Results & Discussion

- The proposed package frameworks were built successfully with intended functionalities and have achieved $R^2 > 0.7$ for wavelength prediction with validation data
- Multidimensional property exploration was performed for the molecules included in the database
- The structural information encoded in SMILES/ connectivity fingerprint/ Coulomb matrix can be used to calculate spectroscopic properties
- The accuracies of our models are largely limited by the small size of the dataset, and the complexity of the problems.
- With the pre-trained models weights, the prediction speed can be guaranteed, while the fine-tuned accurate models still remain as biggest challenges.

Example GUI:



Future Work

- Sanitize SMILES input; add alternative input options
- Expand database and tune parameters for more accurate models
- Include multiple features in predicted absorption/emission spectra
- Allow users to train models with their own data
- Add a feature for pipelining predictions

References

Data: PhotoChemCAD (http://www.photochemcad.com/PhotochemCAD.html)
Dependencies: Keras, TensorFlow, RDKit, Pandas, Numpy, PubChemPy (all open-source).
Publication: Garrett B. Goh et al. 2018. SMILES2vec. In Proceedings of ACM SIGKDD Conference, London, UK, Aug, 2018 (KDD 2018), 8 pages

SPEEDCOM would like the thank UW DIRECT and the Clean Energy Institute at the University of Washington.

CLEAN ENERGY INSTITUTE
UNIVERSITY of WASHINGTON

SPEEDCOM

DIRECT
Data Intensive Research Enabling Clean Technologies

MetaMolES : A Biochemical Retrosynthesis Tool Using Logistic Regression Model Based on Enzyme Promiscuity

Introduction

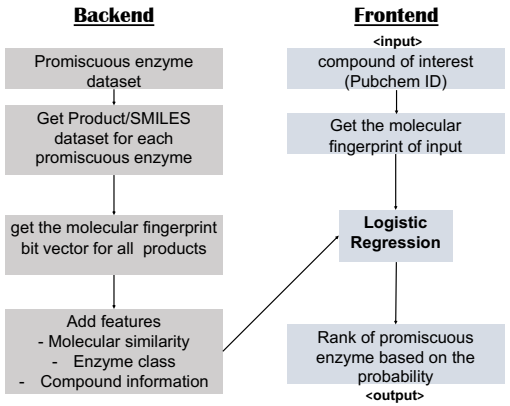
Background

Substrate promiscuous enzymes can biochemically transform several substrates. The use of promiscuous enzymes in metabolic engineering is particularly advantageous because it relieves the burden on the engineered organism. However, most enzyme databases only link one main substrate to each enzyme instead of a repertoire of suitable compounds, which limits the chance for researchers to utilize promiscuous enzymes. Also, another limitation for metabolic engineers comes from the limited access to closed-source biochemical retrosynthesis tools. To address these limitations, herein, we present a user-friendly open-source tool that helps curate the most plausible enzymatic transformation based on substrate promiscuity.

Goal

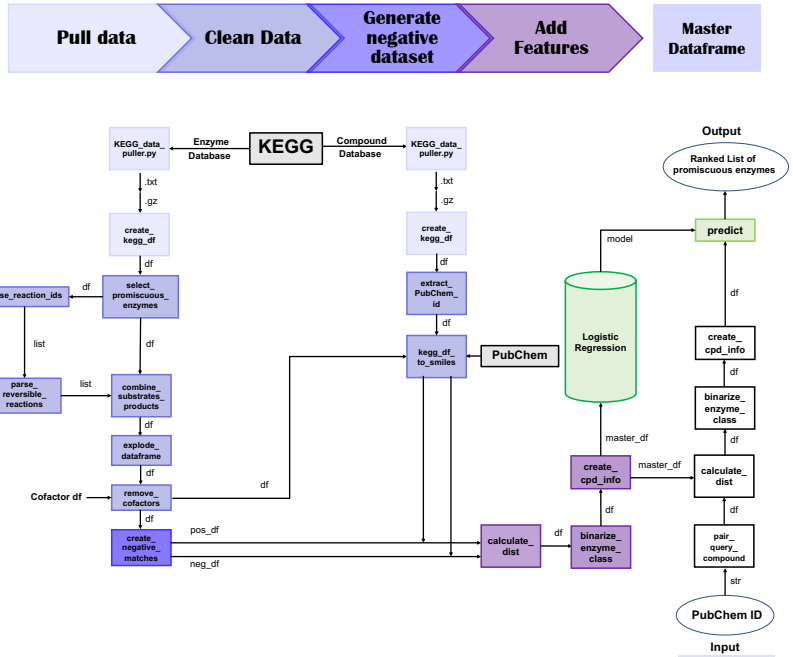
- Aim to utilize data science and software engineering intuition to find, and predict, a plausible metabolic pathway for production of a given molecule with retrosynthetic analysis approach.
- Find a novel promiscuous substrate for enzymatic transformation

Workflow

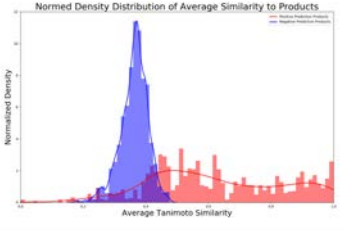


Method

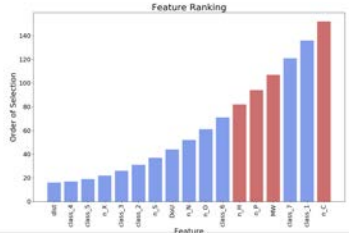
Data Curation



Logistic Regression



normalized counts of calculated average Tanimoto similarity for positive and negative predictions.

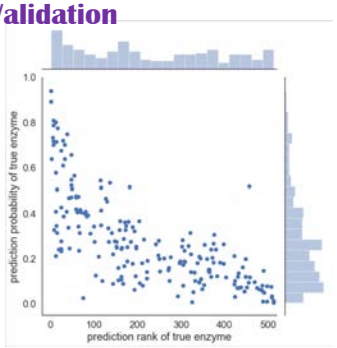


Out of 17 features, 13 features were chosen for logistic regression based on the ranking in their significance.

We selected logistic regression over an SVM to perform enzyme/compound reaction pairs because of our relatively small dataset and our desire to select and rank outputs based on predicted likelihood of reaction. The model was generated with sci-kit learn from 13 features with balanced feature weights and a liblinear solver. The output, after organizing and sorting, is the likelihood of reaction with a specific enzyme. The enzyme can then be looked up in the KEGG database.

Result

Model Validation



We set aside 207 (10%) known promiscuous enzyme-product pairs from our dataset prior to training and testing the logistic regression model. As an evaluation of predictive power, we fed each of the products from this set into the model and recorded the prediction probability for the known true enzyme, as well as the relative rank of the enzyme suggestion out of the master set of 516 promiscuous enzymes. We found a negative correlation between probability and rank (slope=-9.21e-4, p=1.70e-35, R^2=0.54). This suggests a bias away from false positives and towards false negatives. Top and left marginal figures are marginal histograms of rank and probability, respectively.

Conclusion

- The logistic regression model is ~90% predictive when 20% of the data was reserved for testing
- The model correctly predicted reactivity (P>0.5) for 14% of the validation reactions, and among these positive categorizations, the median prediction rank was 28
- Results suggest that using distance of chemical similarity is a reasonable approach
- Validation testing revealed a negative correlation between prediction rank and prediction probability, suggesting a bias away from false positives, and targets for model improvement

Future Work

- Add compounds with canonical SMILES string but no isomeric SMILES string
- Test inclusion of additional features, such as full chemical fingerprints, and enzyme descriptors
- Explore alternative models, such as SVMs, neural networks, decision trees/random forests, and ensemble methods.
- Extend approach to include non-promiscuous enzymes
- Include simple chemical transformation for biocatalysis application

Github repo: https://github.com/theicechol/metamoles

Dependency : Rdkit, bioPython, Pubchempy, scipy, sklearn, pandas, numpy

References

Kanehisa, M., Sato, Y., Furumichi, M., Morishima, K., and Tanabe, M.: New approach for understanding genome variations in KEGG. Nucleic Acids Res. 47, D590-D595 (2019).

Kanehisa, Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K.; KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 45, D353-D361 (2017).

Kanehisa, M. and Goto, S.; KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 28, 27-30 (2000).

Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, Li Q, Shoemaker BA, Thiessen PA, Yu B, Zaslavsky L, Zhang J, Bolton EE. PubChem 2019 update: improved access to chemical data. Nucleic Acids Res. 2019 Jan 8; 47(D1):D1102-1109. doi:10.1093/nar/gky1033.

Pertusi, Dante A., et al. "Predicting novel substrates for enzymes with minimal experimental effort with active learning." Metabolic engineering 44 (2017): 171-181.

Capstone Project (CHEME 547)

- Holistic integration of previous courses in real world project setting enabling skill mastery
- Spring quarter, 14 weeks (10 + 4)

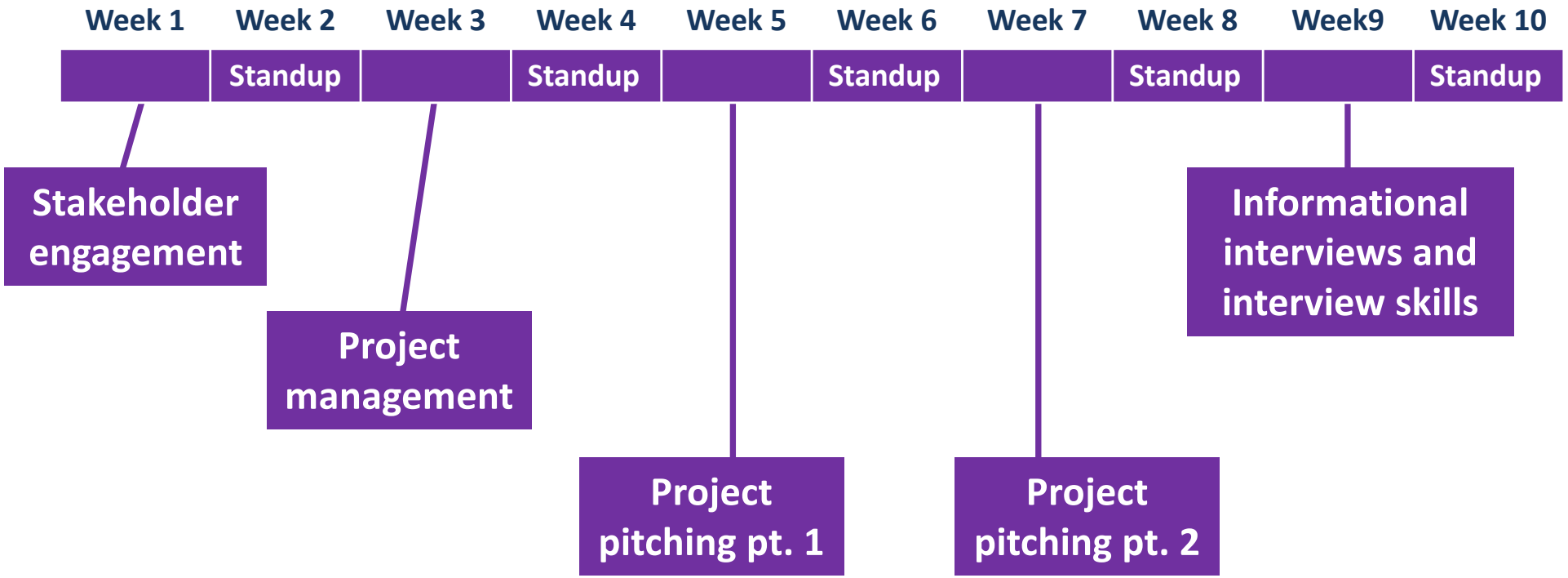
- Students build professional skills
  - Project management, communication
- Students build soft-skills through practice
- Students build professional networks
  - Internships, sponsored research, etc.

Capstone Project (CHEME 547)

- Capstone projects are supplied by
  - University of Washington faculty
  - National labs, e.g. PNNL
  - NGOs & government agencies, e.g.
    - Alaska Center for Energy and Power
    - Metro Transit of King County
  - Companies, e.g.
    - Optimum Energy
    - KPMG
    - Novo Nordisk

Capstone timeline

- Alternating
  - Student standups
  - Professional development enrichment sessions



Capstone timeline

- 4 additional weeks after week 10
  - Intended to build independence from instructor
  - No standups
- Capstone showcase including
  - Project ‘elevator’ pitches
  - Poster session
- Invitees include faculty, sponsors, community

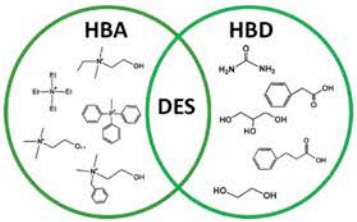
High-Throughput Measurement of Deep Eutectic Solvent Melting Points using IR Bolometry

Project Sponsor: Dr. Lilo D. Pozzo (*Chemical Engineering*)

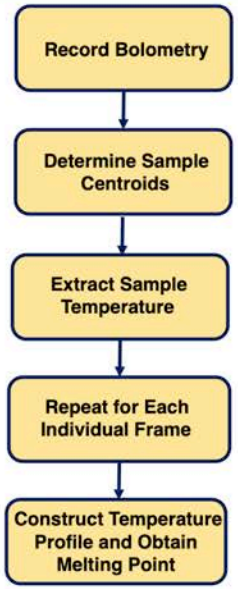
Team Members: Shrilakshmi Bonageri (*Chemical Engineering*), Jaime Rodriguez (*Chemical Engineering*), Sage Scheiwiller (*Chemical Engineering*)

Background and Motivation

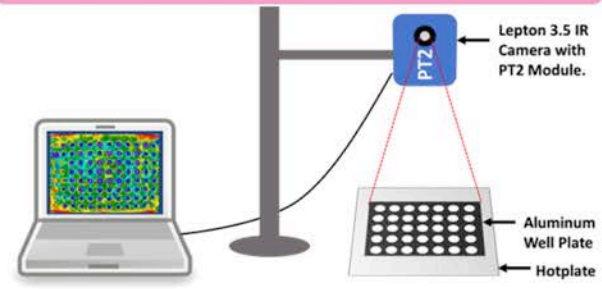
Deep eutectic solvents (DES) are novel solvents formed between organic hydrogen bond acceptors and donors. DES can be formed at low-cost for several important applications, such as chemical synthesis, extractions, electrochemistry, and even drug delivery. However, the design space for DES is enormous and high-throughput measurement of melting points is required to rapidly identify DES with melting points that are feasible for their intended application.



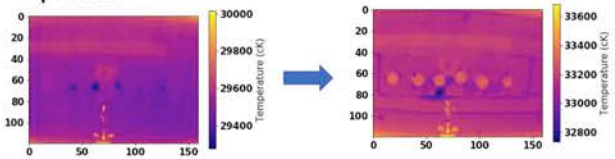
Workflow



Bolometry Setup

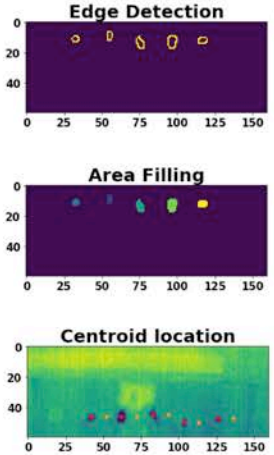


- Sample temperature is monitored by an IR camera.
- Melting points are detectable from a sudden increase in sample temperature due to an increased thermal conductivity in the liquid phase.

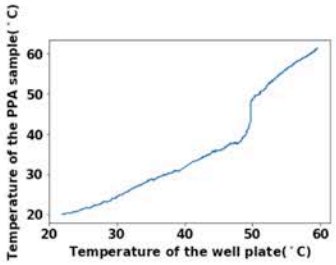


Solid Sample **Melted Sample**

Temperature Profile via Edge Detection

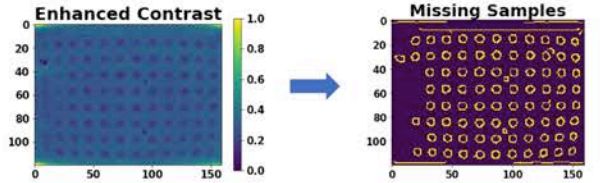


- The temperature profile of the samples and plate is determined by detecting the edges, filling and labeling them, and monitoring the temperature at their centroids.

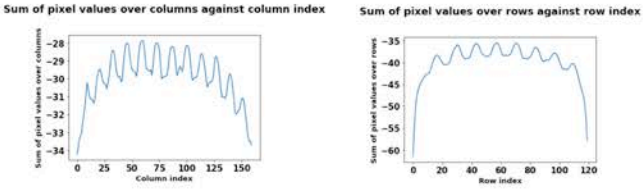


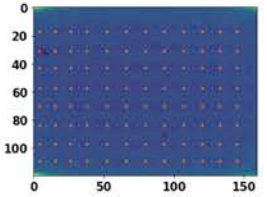
Alternative Method for Low-Contrast Samples

- In some situations, the contrast between the image and sample may be too low for edge detection, even with contrast enhancement.



- Alternatively, centroid locations for each sample can be found by summing pixel values over individual rows and columns of the sample holder (well plate).





- Using this alternative method, centroids were located for each sample in a 96 well microplate.

Conclusions and Future Work

Using IR Bolometry, melting points can be determined for multiple samples at once in a matter of minutes, as opposed to standard techniques which may take up to an hour for a single sample. Once sufficient data is collected, future work may include the development of a machine learning model to predict the melting points of DES based on their composition.

DIRECT Data Intensive Research Enabling Clean Technologies

CLEAN ENERGY INSTITUTE UNIVERSITY of WASHINGTON

DopeDefects

Predicting impurity energy levels of semiconductors using machine learning

Ryan Beck, Lauren Koulias, Linnette Teo

Project Mentors: Argonne National Lab – Maria K. Chan, Arun Kumar Mannodi Kanakkithodi

Overview

Overview: DopeDefects is an open source python package that aims to predict the enthalpy of formations, as well as the charge transition levels, of various defects embedded in Cd/chalcogenide crystals.

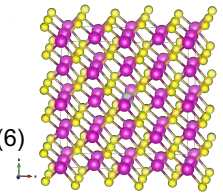
Available on GitHub:

https://github.com/dopedefects/dopedefects.git

Motivation

- Chemical space for potential solar cell materials is large
- Use Density Functional Theory (DFT) computations, an ab inito method for calculating chemical properties
- DFT requires significant computational resources both in time and energy costs
- Number of calculations required to explore entire space is unfeasible
- Possible solution: predictive models trained on small subset of calculated properties

About the Data



Properties to predict

- Supercell enthalpies of formation (3)
- Supercell energies of charged states (6)

Descriptors of defect system (109 in total)

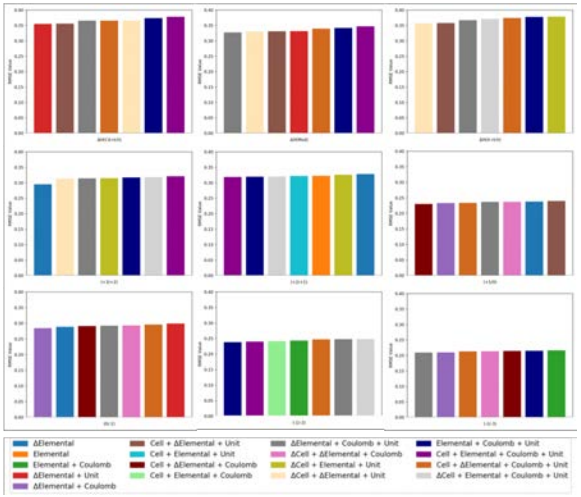
- ***Elemental:*** elemental properties of the dopants such as group, period, ionic, atomic, & covalent radii, boiling point, atomic weight, etc.
- ***ΔElemental:*** change in elemental properties between the dopant atom and the atom that it replaced
- ***Unit:*** ΔH values from the unit cell calculation, conduction and valence band edges
- ***Cell:*** bond angles and bond lengths for all atoms in the unit cell
- ***ΔCell:*** change in the bond angles and bond lengths for all atoms between the doped and undoped cell
- ***Coulomb:*** coulomb matrix for the unit cell

Data Cleaning Functionalities

- Scan through the provided directory for VASP (Vienna Ab inito Simulation Package) geometry files and convert the coordinates to cartesian space
- Determine the position and type of vacancy
- Calculate the bond lengths and angles for the atoms surrounding the defect, as well as determining the change in comparison to a pure system
- Collect all the properties into a pandas dataframe, as well as save and resume the data so that data parsing does not need to be redone

Feature Selection

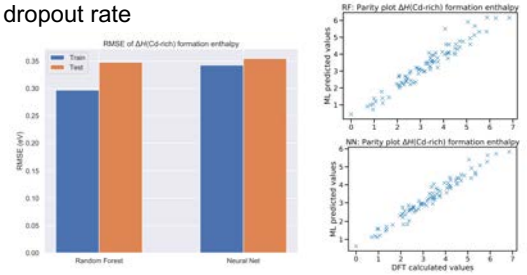
RMSE for Random Forrest Regression



- For each property being predicted, a different set of descriptors was the most accurate, the top 7 for each category are shown above
- Overall it seems that Elemental properties are always necessary, combined with other descriptors for the most accurate results

Neural networks

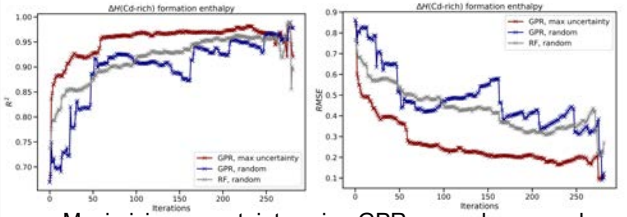
- Used elemental and unit descriptors; 425 data points
- Split data 6:2:2 (training:validation:testing)
- Hyperparameter tuning: batch size, epoch number, number of nodes in hidden layers, learning rate, dropout rate



Neural net does not perform significantly better than random forest – need further optimization, more data

Iterative Method using Gaussian Process Regression

- Start with small subset of data to fit GPR model (10% of 315 CdTe structures)
- Use model to predict mean and uncertainty (standard deviation) on remaining test points
- Choose a test point that maximizes uncertainty
- Add test point (with calculated value) to model and retrain
- Iterate - keep adding points till satisfied



Maximizing uncertainty using GPR vs random search helps reduce initial number of known points needed

Future Work

- Multiple output predictions
- Improvement of prediction for impurity transition levels
- More detailed analysis into different CdX structures

CLEAN ENERGY INSTITUTE
UNIVERSITY of WASHINGTON

Route_Dynamics: An open-source package for visualizing and ranking transit routes

Atinuke Ademola-Idowu, Erica E. Eggleton, Yohan Min, Kaiming Tao

DIRECT
Data Intensive Research Enabling Clean Technologies

Purpose

Driving up and down hills is a source of stress for batteries in electrified vehicles due to high charge/discharge rates. This software uses geographic information systems (GIS) data to elevation profiles for King County ranks the difficulty based on road also be used as a component for degradation model.

Wor

Read Data

Bus route coordinates from King County Metro

Elevation data for the King County area from Washington Department of Natural Resources

Process D

Re-trace route extract only on bus routes

Convert eleva data to coordi system used b routes.

- Geopandas
- Rasterio

- ArcGIS/ArcMa
- Geopandas
- Rasterio

Input Da

King County METRO

Bus routes: Shapefile (.shp) [1]

Results

Data Analysis Output

Visualization Output

Road Grade 0.00 0.08 0.16 0.24 0.32 0.40

ap, either within a notebook or saved lected routes and road grade.

nking

umber

Route Dynamics

235th ECS Meeting

May 26-30, 2019

Dallas, TX

Sheraton Dallas

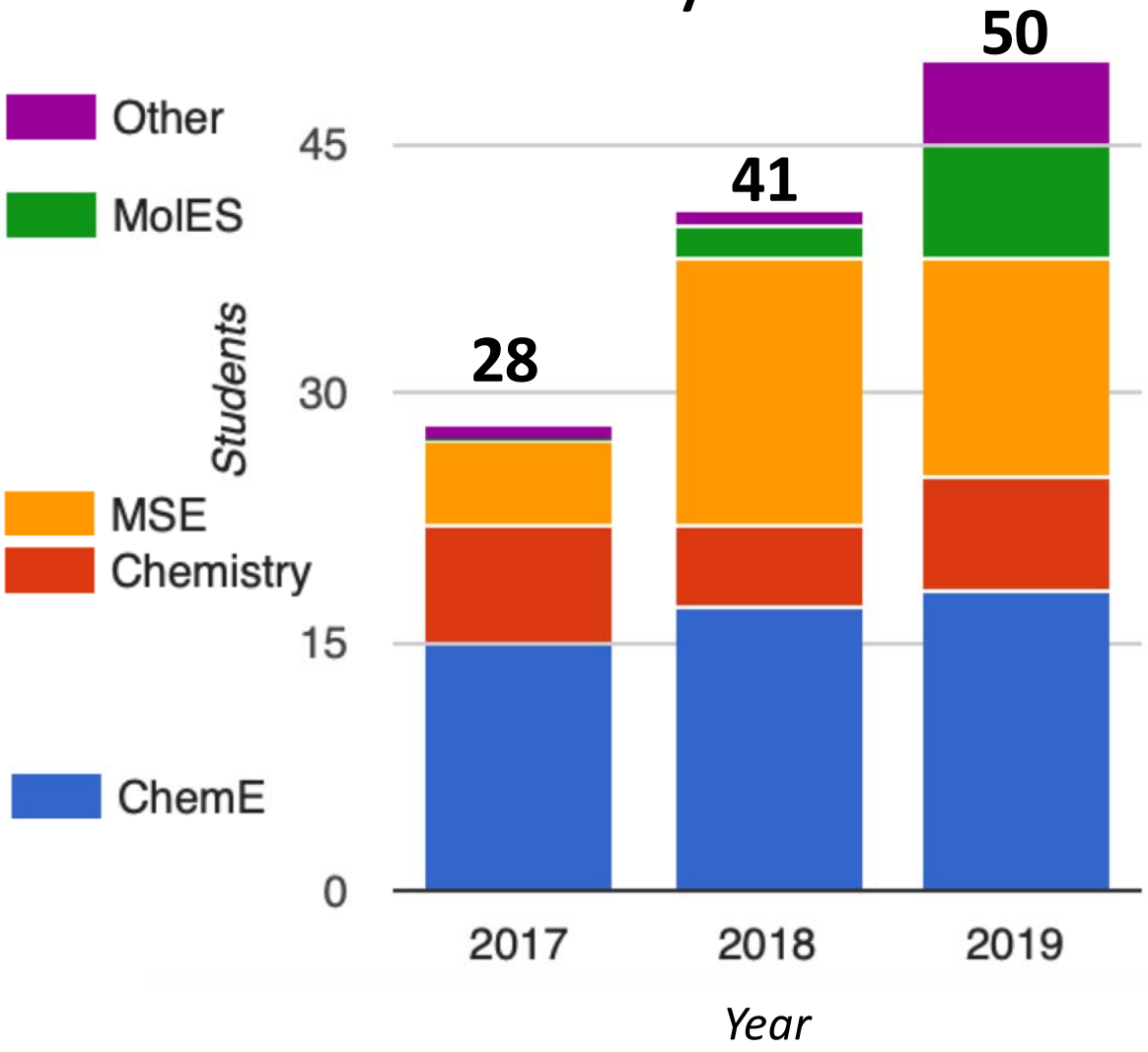
Multiscale Modeling, Simulation and Design 3: Enhancing Understanding, and Extracting Knowledge from Data - Session 5 – 08:00 – 11:40
Chair(s): Ankur Jain

•••

09:20 **1155** **Incorporating GIS Data into a Battery Performance Model for Electric and Hybrid-Electric Buses** – E. E. Eggleton and D. T. Schwartz (University of Washington)

Enrollment

Enrollment by unit

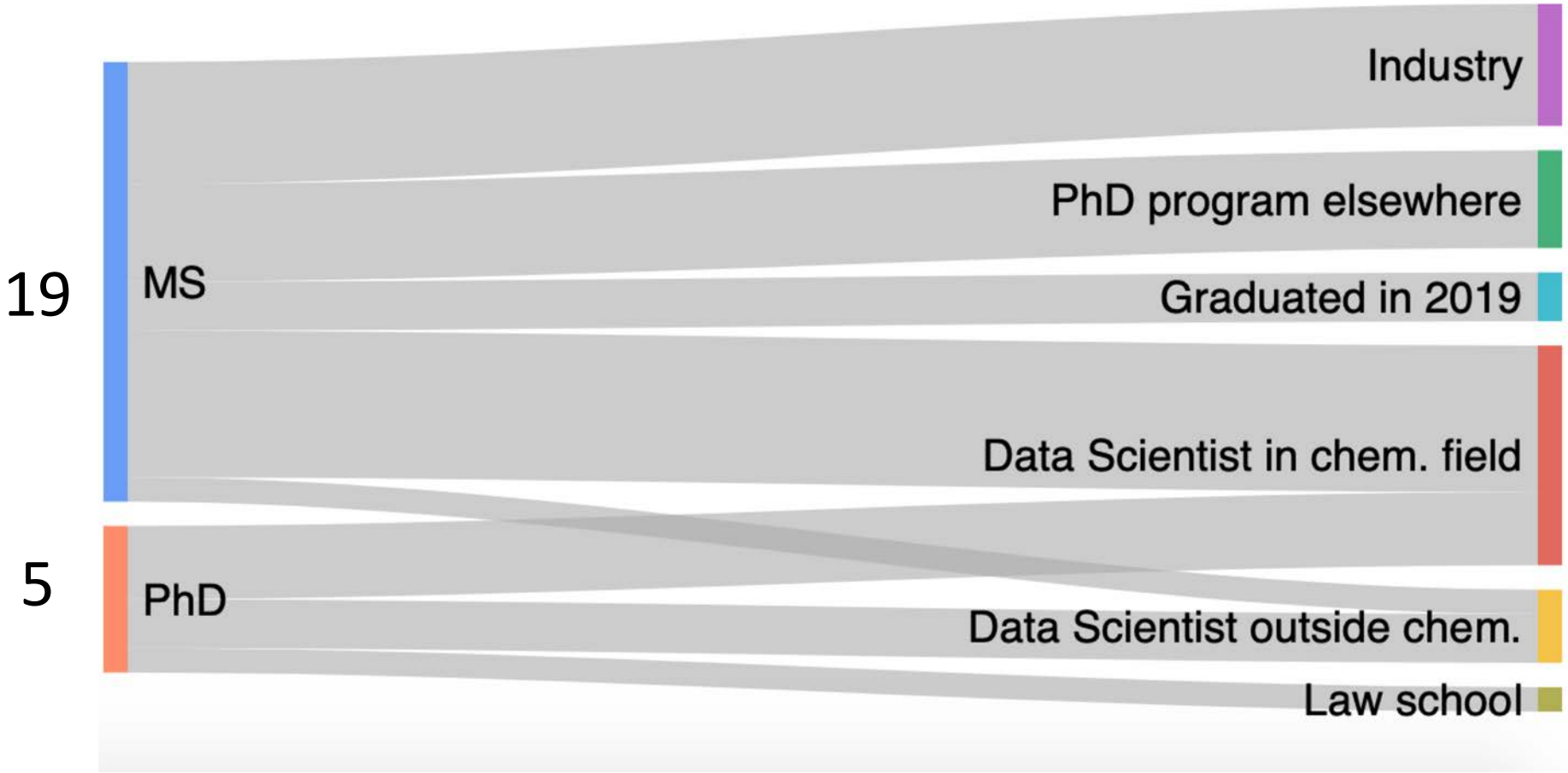


CHEME Enrollment

Year	Students
2017	15
2018	17
2019	18

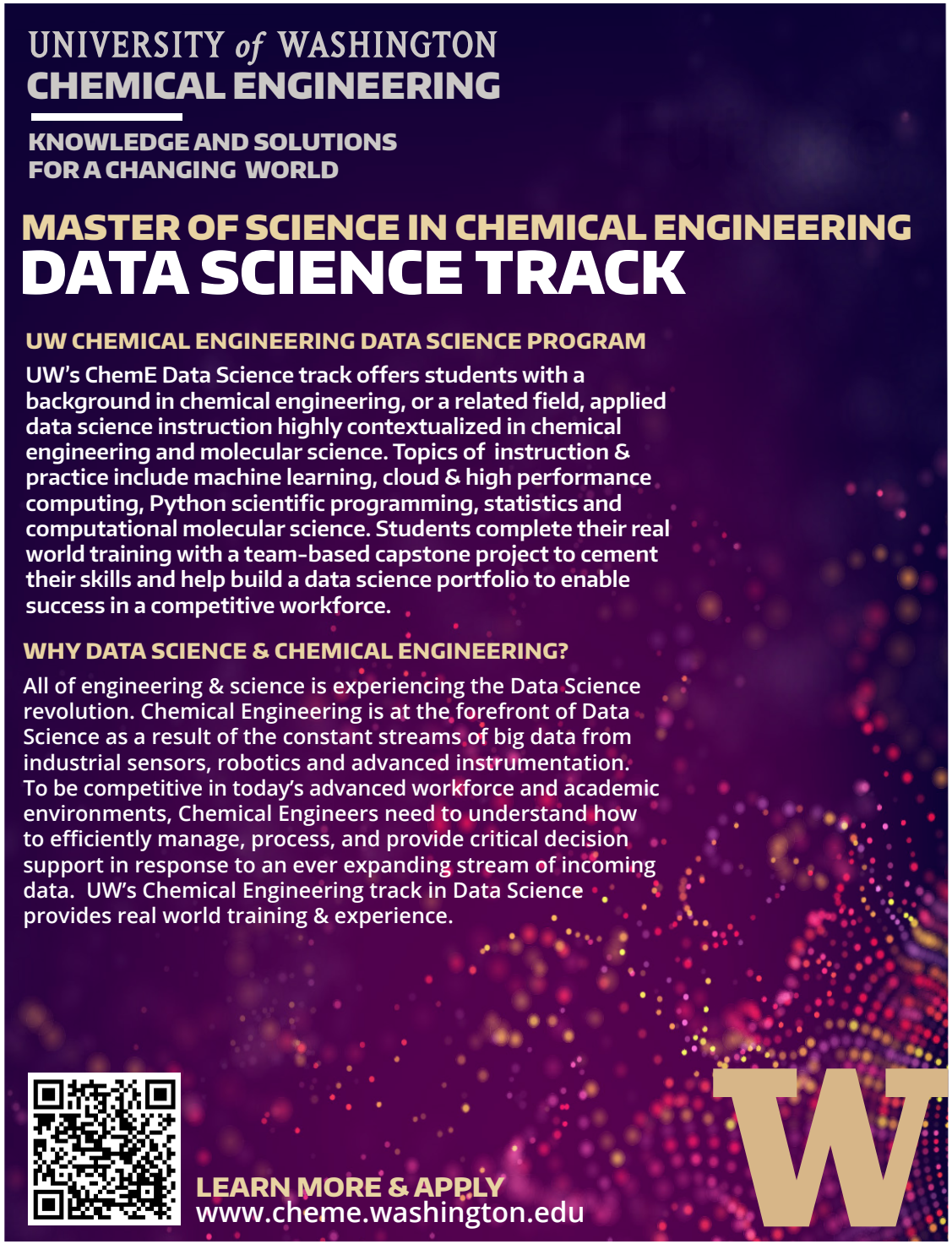
Options student outcomes

- Outcomes for degrees granted (CHEME)
  - Most MS are two years (thesis) so limited data



Future

- Diversification from molecules
  - Controls, operations, process analytics
  - Industry 4.0, IoT, online decision making
- Challenges
  - How to structure sequencing of new courses
  - How to offer this to our undergrads (4+1?)
- Create Data Science Options in
  - MolES (done, May 2019)
  - Chemistry (in progress)
  - MSE (in progress)



2020 class will be our first year of this non-thesis data science track

More info & future directions

- The course materials are all open source and available online in our GitHub repository (BSD)
  - https://github.com/UWDIRECT/UWDIRECT.github.io
  - Lectures, homework assignments, some videos
- We need more ChemE Data Scientists!

ChemE Data Scientist
Knows transport, thermodynamics **and** machine learning

Acknowledgements

- NSF
  - IGERT-CIF21: Big Data U: A Program for Integrated Multidisciplinary Education and Research for Big Data Science, #1258485
  - NRT-DESE: Data Intensive Research Enabling Clean Technologies (DIRECT), #1633216
- Moore & Sloan Foundations
  - Moore & Sloan Data Science Environments
- UW eScience Institute, Clean Energy Institute

Student thoughts

- “DIRECT is a multidimensional and interdisciplinary training program that teaches students effective communication, good coding practices, team building, and the core fundamentals of data science. With multiple opportunities for students to engage in data science related research outside of their field of expertise, students are forced to tackle difficult, real-world problems in team settings with industry, national labs, and academia.”