KAIST CHEMISTRY

Lecture 01
Introduction

Prof. Ph.D. Woo Youn Kim

KAIST

About me

- Ph.D. in quantum chemistry
- Postdoc in theoretical physics
- Professor at KAIST since 2011
- Leader of ACE Team: wooyoun.kaist.ac.kr
  - Development of quantum chemistry software
  - Development of automated chemical reaction prediction software
  - Deep learning for drug discovery

Course Description

AI has become a big social issue as it spreads rapidly to science, industry, and even daily life. Chemical research based on AI using big data in chemistry has been reexamined. In this course, we will discuss the role of AI in modern chemistry and investigate the latest trends in this field. It aims to learn practical knowledge that can be used in actual research field through theory and practice focused on deep learning.

Prerequisite

- python language
- introductory course of linear algebra
- physical chemistry I and II

AI in the News

Quiz show, 2011

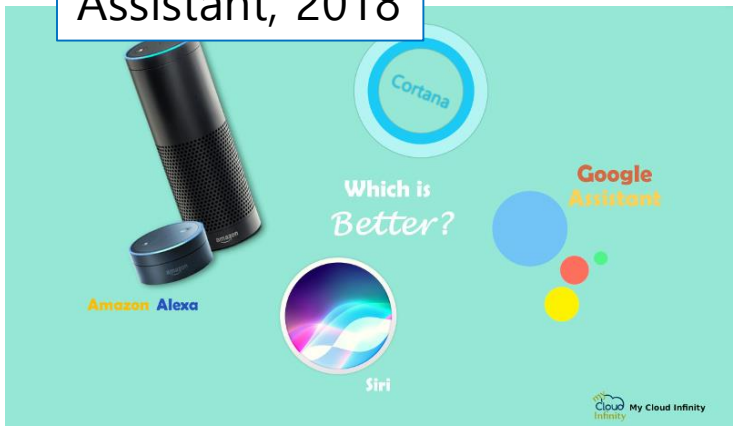
Go game, 2016

Drug discovery, 2018. June ~100 startup

Autonomous vehicle

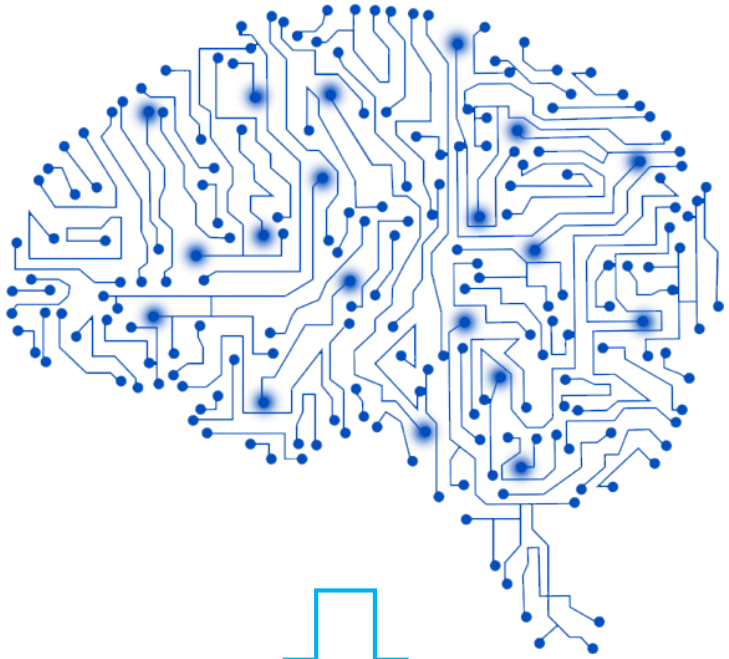
Healthcare

Assistant, 2018

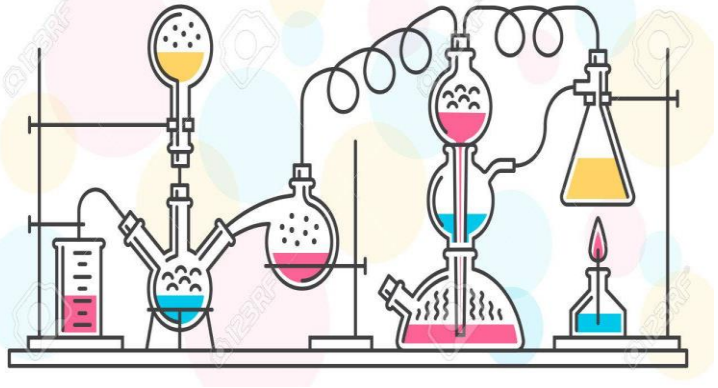


https://www.youtube.com/watch?v=ADI_mjhxvgs

AI in Scientific Journals



Materials discovery

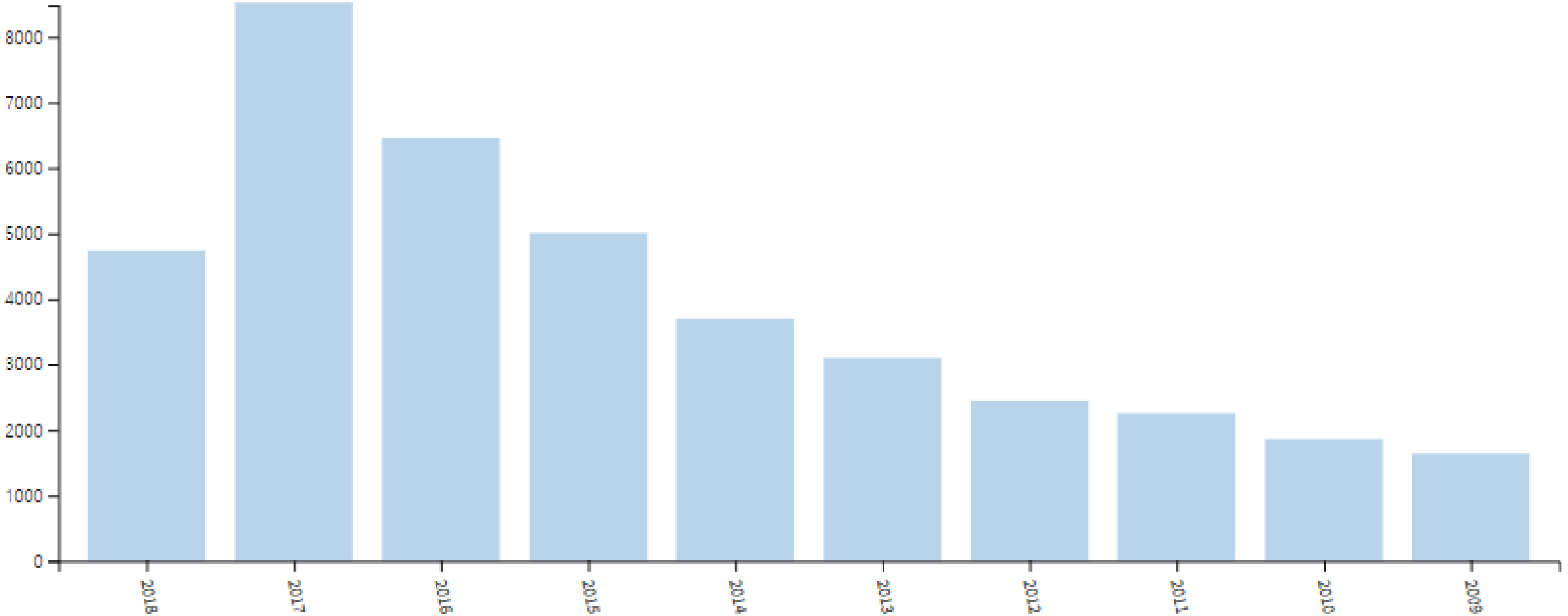


Prediction and control of chemical reactions

Drug discovery

AI in Scientific Journals

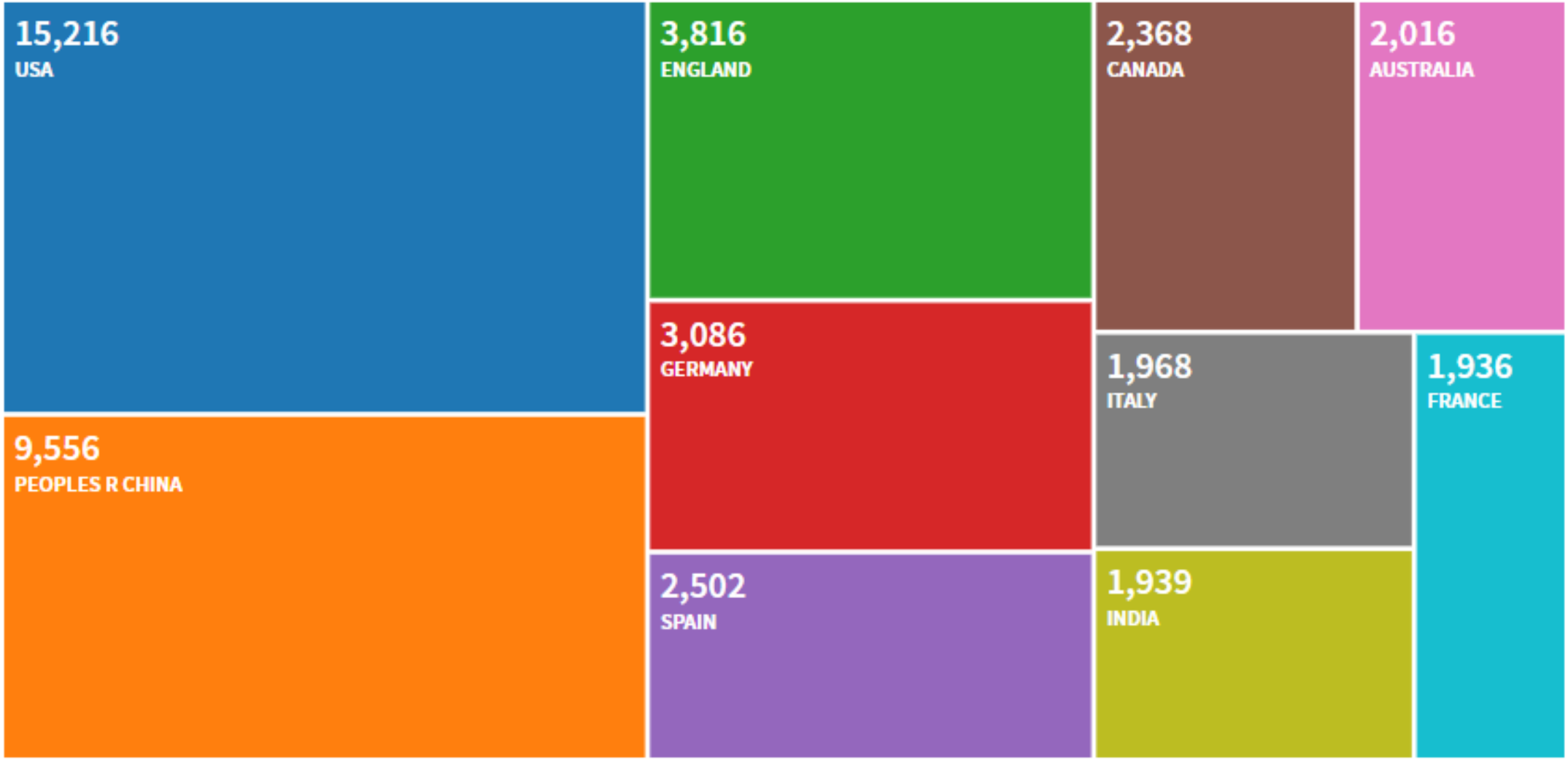
http://wcs.webofknowledge.com/
Keyword: machine learning, 2018. June 16



AI in Scientific Journals

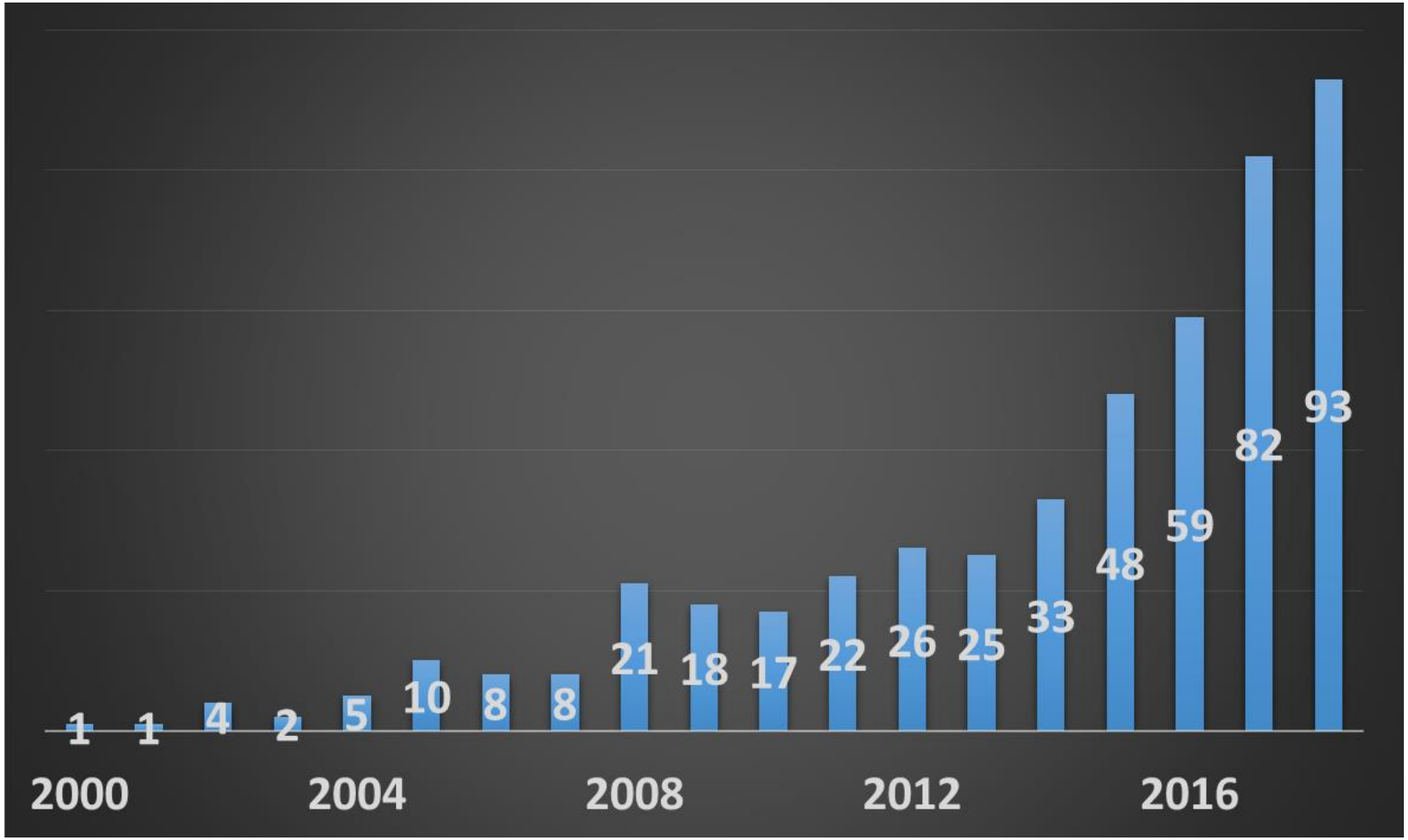
http://wcs.webofknowledge.com/
Keyword: machine learning, 2018. June 16

Korea: 1490 (12th, 3 %)



AI in Scientific Journals

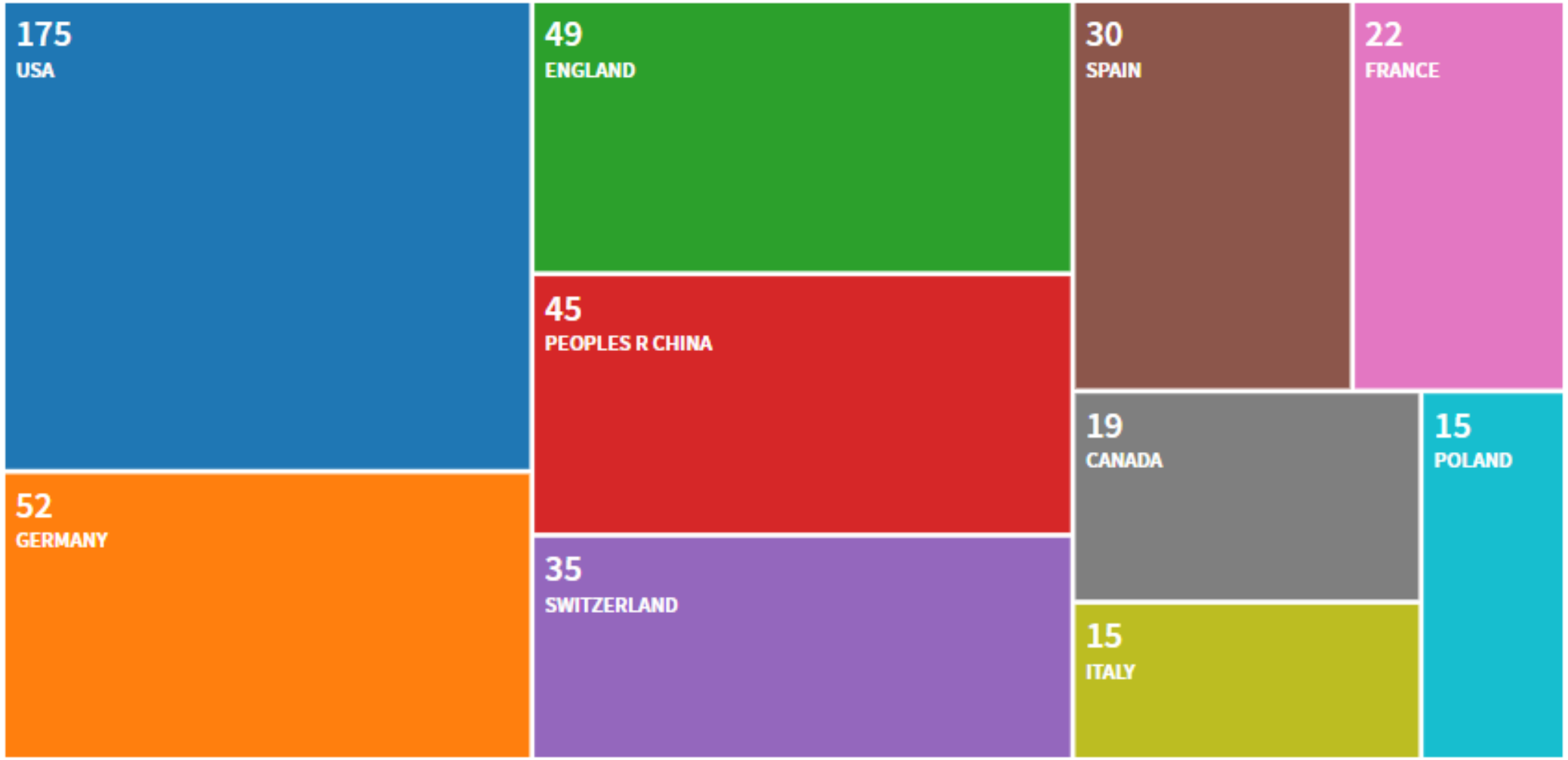
Keyword: machine learning and chemistry
2018. June 16



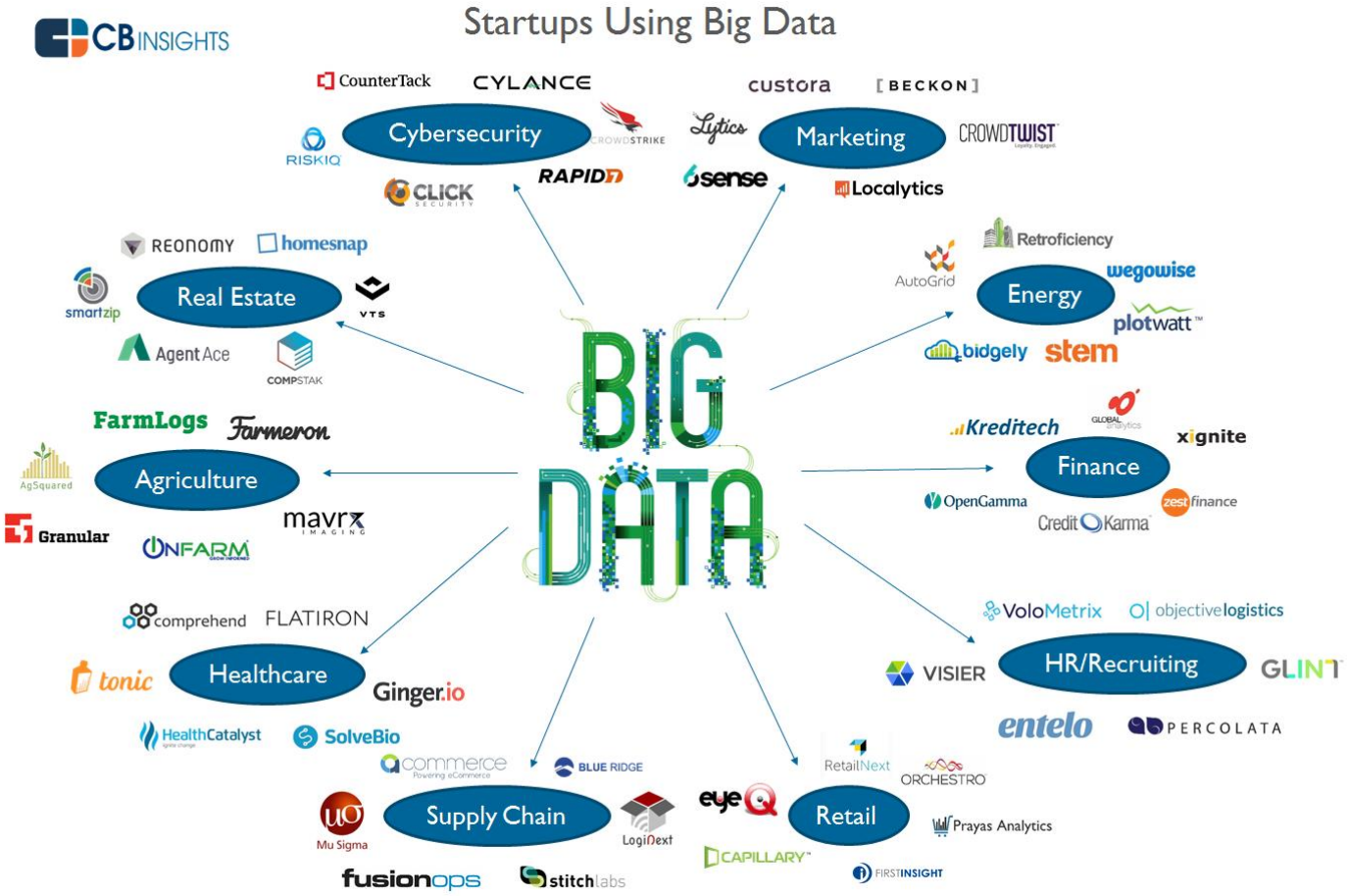
AI in Scientific Journals

Keyword: machine learning and chemistry
2018. June 16

Korea: 7 (22th, 1.50 %)



Big Data


CB INSIGHTS
Startups Using Big Data
BIG DATA
Cybersecurity
CounterTack
CYLANCE
RISKIQ
CROWDSTRIKE
CLICK
RAPID7
Marketing
custora
BECKON
Lytics
6sense
Localytics
CROWDTWIST
Real Estate
REONOMY
homesnap
smartzip
VTS
Agent Ace
COMPSTAK
Energy
Retroficiency
AutoGrid
wegowise
plotwatt
bidgely
stem
Agriculture
FarmLogs
Farmeron
AgSquared
Granular
mavrx
IMAGING
ONFARM
Finance
Kreditech
xignite
OpenGamma
zest finance
Credit Karma
Healthcare
comprehend
FLATIRON
tonic
Ginger.io
HealthCatalyst
SolveBio
HR/Recruiting
VoloMetrix
objectivelogistics
VISIER
GLINT
entelo
PERCOLATA
Supply Chain
commerce
BLUE RIDGE
Mu Sigma
LogiNext
fusionops
stitchlabs
Retail
RetailNext
ORCHESTRO
eyeQ
Prayas Analytics
CAPILLARY
FIRSTINSIGHT

Healthcare Big Data

Big Data in Chemistry

✓ Free big data for drug design

Zinc: http://zinc.docking.org/
14 million purchasable compounds

ChEMBL: https://www.ebi.ac.uk/chembl/
bioactivity outcomes across thousands of protein targe
1,828,820 compounds

PubChem:
http://pubchem.ncbi.nlm.nih.gov/
Information on the biological activities of 26 million small molecules

ChemSpider
60 million chemical structures

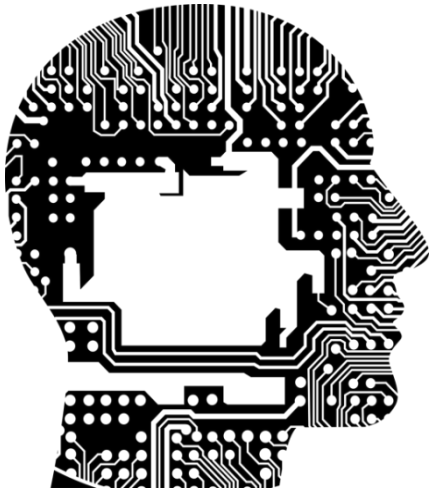
The DrugBank database:
http://www.drugbank.ca/
6712 drugs entries including
1441 FDA-approved small molecule drugs
134 FDA-approved biotech drugs
83 nutraceuticals
5086 experimental drugs

Crystallography Open Database
Cambridge Structural Database
large repositories of organic and inorganic compounds.
The protein data bank a repository of experimentally resolved three dimensional protein structures.

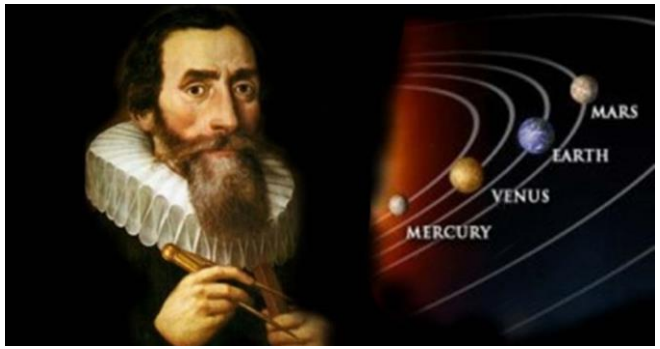
Reactions: Raxys and
US Patent: 112 만개 화학 반응 정보 공개

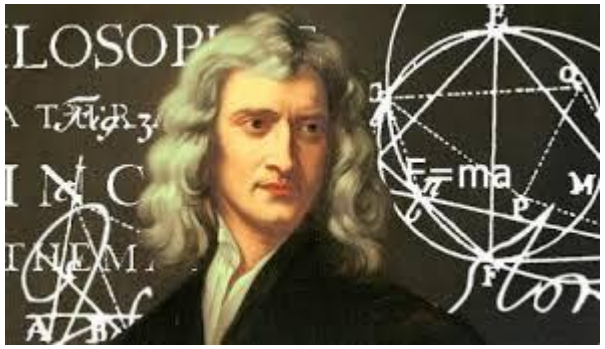
AI + Big Data = Data-driven Science

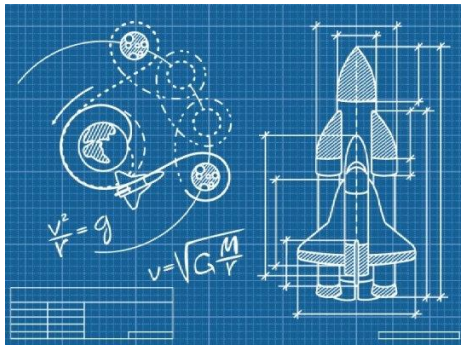
AI

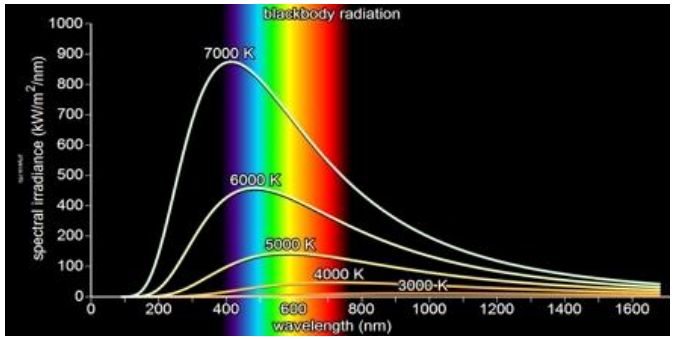


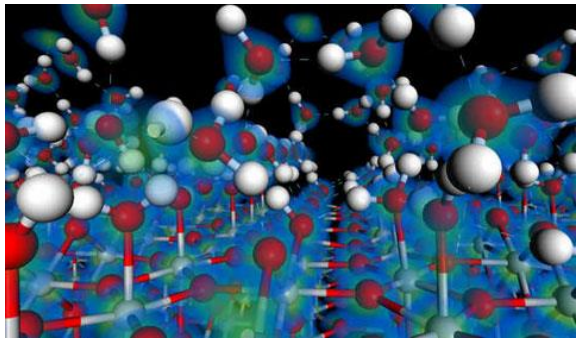
Science Evolution











Data Scientist

Data Scientist: The Sexiest Job of the 21st Century

https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

What data scientists do is make discoveries while swimming in data

Hal Varian, the chief economist at Google, is known to have said, “The sexy job in the next 10 years will be statisticians.

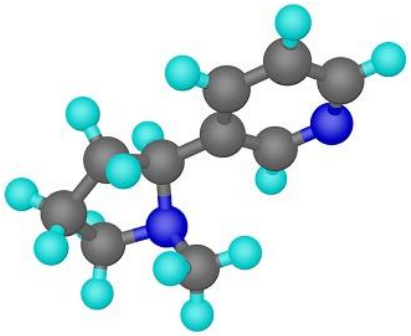
Best job in the U.S in 2015 [Forbes, LinkedIn].

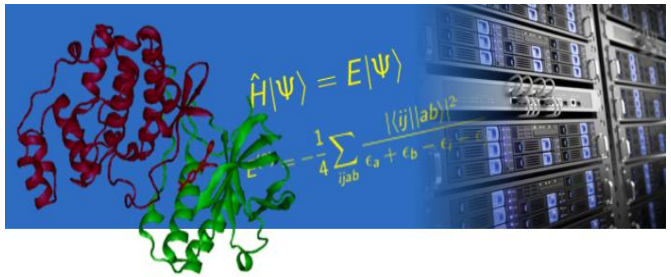
Salary has jumped from $125,000 to $200,000+ [Glassdoor].

McKinsey projects that “by 2018, the U.S. alone may face a 50 percent to 60 percent gap between supply and requisite demand of deep analytic talent.”

Learning molecular structure-property relationship

Conventional approach





Solubility
Toxicity
Drug efficacy
.
.
.

Input data
Structure (X)

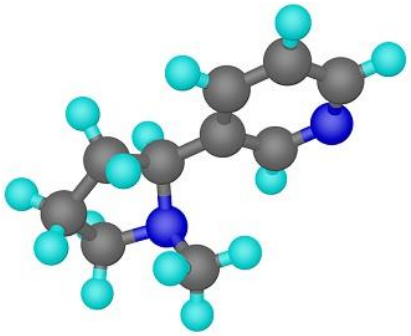
Quantum chemistry
& related methods

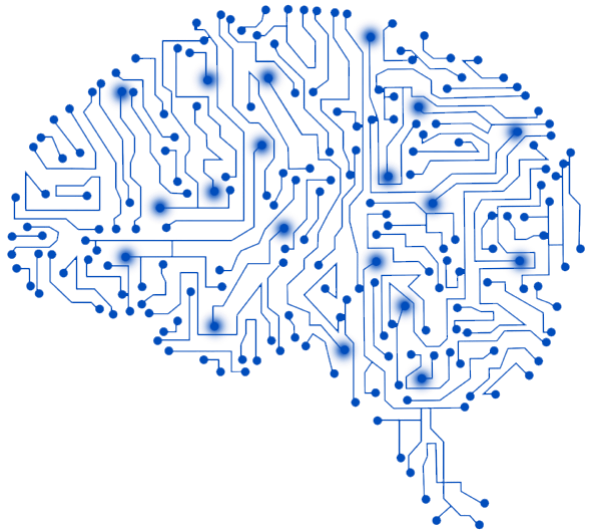
Output data
Property (Y)

Y = f(X); f = Hamiltonian

Learning molecular structure-property relationship

Deep learning approach





Solubility
Toxicity
Drug efficacy
.
.
.

Input data
Structure (X)

Feature extraction (**L**)

Output data
Property (Y)

Y = f(X); f = neural networks

ACS central science

Research Article

Cite This: ACS Cent. Sci. 2018, 4, 268–276

Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules

Rafael Gómez-Bombarelli,[†,#] Jennifer N. Wei,[‡,#] David Duvenaud,[¶,#] José Miguel Hernández-Lobato,[§,#] Benjamín Sánchez-Lengeling,[‡] Dennis Sheberla,[‡] Jorge Aguilera-Iparraguirre,[†] Timothy D. Hirzel,[†] Ryan P. Adams,[∇,∥] and Alán Aspuru-Guzik*[,‡,⊥]

[†]Kyulux North America Inc., 10 Post Office Square, Suite 800, Boston, Massachusetts 02109, United States

[‡]Department of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts 02138, United States

[¶]Department of Computer Science, University of Toronto, 6 King's College Road, Toronto, Ontario M5S 3H5, Canada

[§]Department of Engineering, University of Cambridge, Trumpington Street, Cambridge CB2 1PZ, U.K.

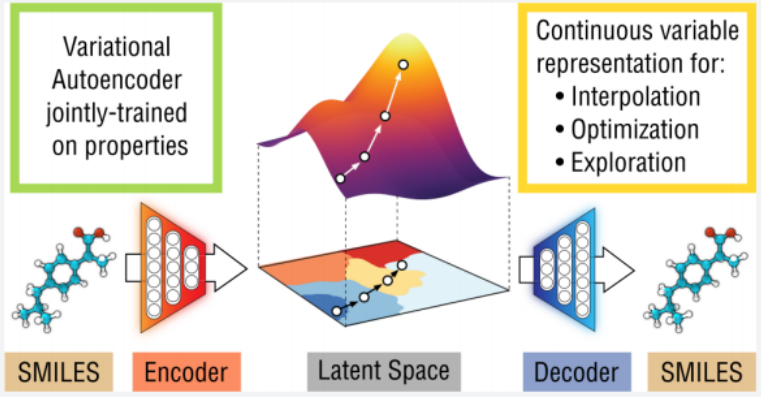
[∇]Google Brain, Mountain View, California, United States

[∥]Princeton University, Princeton, New Jersey, United States

[⊥]Biologically-Inspired Solar Energy Program, Canadian Institute for Advanced Research (CIFAR), Toronto, Ontario M5S 1M1, Canada

Supporting Information

ABSTRACT: We report a method to convert discrete representations of molecules to and from a multidimensional continuous representation. This model allows us to generate new molecules for efficient exploration and optimization through open-ended spaces of chemical compounds. A deep neural network was trained on hundreds of thousands of existing chemical structures to construct three coupled functions: an encoder, a decoder, and a predictor. The encoder converts the discrete representation of a molecule into a real-valued continuous vector, and the decoder converts these continuous vectors back to discrete molecular representations. The predictor estimates chemical properties from the latent continuous vector



Chemical Science

ROYAL SOCIETY OF CHEMISTRY

EDGE ARTICLE

View Article Online
View Journal | View Issue

Cite this: *Chem. Sci.*, 2018, **9**, 513

MoleculeNet: a benchmark for molecular machine learning†

Zhenqin Wu, ‡[a] Bharath Ramsundar,‡[b] Evan N. Feinberg,§[c] Joseph Gomes, §[a] Caleb Geniesse,[c] Aneesh S. Pappu,[b] Karl Leswing[d] and Vijay Pande*[a]

Molecular machine learning has been maturing rapidly over the last few years. Improved methods and the presence of larger datasets have enabled machine learning algorithms to make increasingly accurate predictions about molecular properties. However, algorithmic progress has been limited due to the lack of a standard benchmark to compare the efficacy of proposed methods; most new algorithms are benchmarked on different datasets making it challenging to gauge the quality of proposed methods. This work introduces MoleculeNet, a large scale benchmark for molecular machine learning. MoleculeNet curates multiple public datasets, establishes metrics for evaluation, and offers high quality open-source implementations of multiple previously proposed molecular featurization and learning algorithms (released as part of the DeepChem open source library). MoleculeNet benchmarks demonstrate that

ACS central science

ACS AuthorChoice

Research Article

Cite This: ACS Cent. Sci. 2018, 4, 120−131

Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks

Marwin H. S. Segler,*[,†] Thierry Kogej,[‡] Christian Tyrchan,[§] and Mark P. Waller*[,∥]

[†]Institute of Organic Chemistry & Center for Multiscale Theory and Computation, Westfälische Wilhelms-Universität Münster, 48149 Münster, Germany

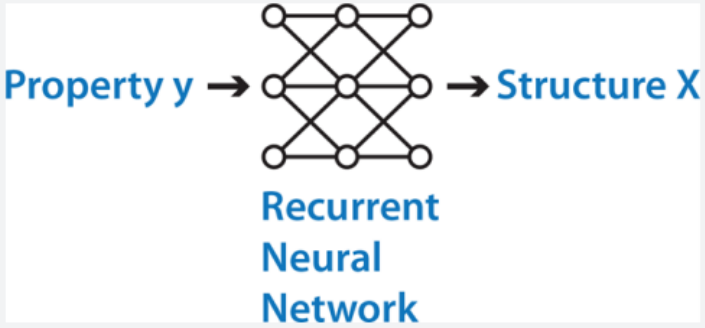
[‡]Hit Discovery, Discovery Sciences, AstraZeneca R&D, Gothenburg, Sweden

[§]Department of Medicinal Chemistry, IMED RIA, AstraZeneca R&D, Gothenburg, Sweden

[∥]Department of Physics & International Centre for Quantum and Molecular Structures, Shanghai University, Shanghai, China

Supporting Information

ABSTRACT: In *de novo* drug design, computational strategies are used to generate novel molecules with good affinity to the desired biological target. In this work, we show that recurrent neural networks can be trained as generative models for molecular structures, similar to statistical language models in natural language processing. We demonstrate that the properties of the generated molecules correlate very well with the properties of the molecules used to train the model. In order to enrich libraries with molecules active toward a given biological target, we propose to



References

(1) Gómez-Bombarelli, Rafael, et al. "Automatic chemical design using a data-driven continuous representation of molecules." ACS central science 4.2 (2018): 268-276.
(2) Duvenaud, David K., et al. "Convolutional networks on graphs for learning molecular fingerprints." Advances in neural information processing systems. 2015.
(3) Ryu, Seongok, Jaechang Lim, and Woo Youn Kim. "Deeply learning molecular structure-property relationships using graph attention neural network." arXiv preprint arXiv:1805.10988 (2018).
(4) Wu, Zhenqin, et al. "MoleculeNet: a benchmark for molecular machine learning." Chemical science 9.2 (2018): 513-530.
(5) Segler, Marwin HS, et al. "Generating focused molecule libraries for drug discovery with recurrent neural networks." ACS central science 4.1 (2017): 120-131.
(6) Gilmer, Justin, et al. "Neural message passing for quantum chemistry." arXiv preprint arXiv:1704.01212 (2017).
(7) Lim, Jaechang, et al. "Molecular generative model based on conditional variational autoencoder for de novo molecular design." arXiv preprint arXiv:1806.05805 (2018).
(8) Goodfellow, Ian, et al. "Generative adversarial nets." Advances in neural information processing systems. 2014.
(9) Zhao, Junbo Jake, et al. "Adversarially Regularized Autoencoders." (2018).
(10) Li, Yujia, et al. "Learning deep generative models of graphs." arXiv preprint arXiv:1803.03324 (2018).
(11) Jin, Wengong, Regina Barzilay, and Tommi Jaakkola. "Junction Tree Variational Autoencoder for Molecular Graph Generation." arXiv preprint arXiv:1802.04364 (2018).
(12) De Cao, Nicola, and Thomas Kipf. "MolGAN: An implicit generative model for small molecular graphs." arXiv preprint arXiv:1805.11973 (2018).

Goal of the Course

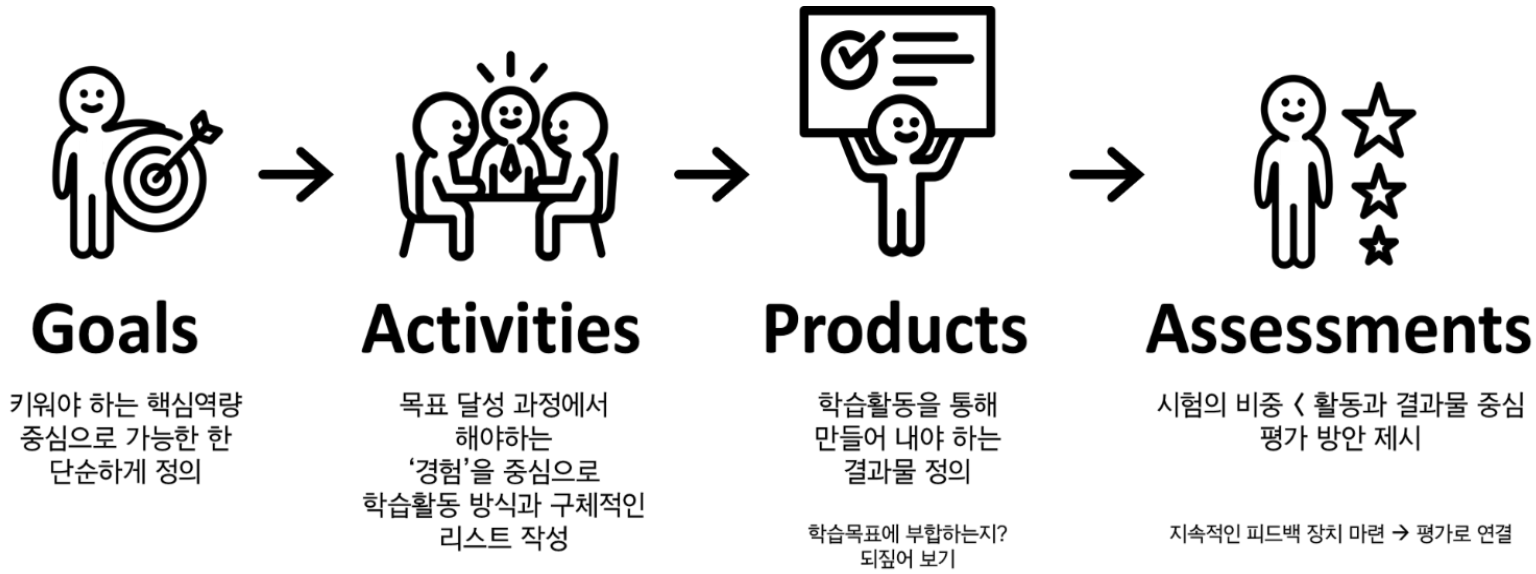
딥러닝 중심의 인공지능에 대한 이해 향상

인공지능과 화학 시스템을 결합한 문제 발굴 능력 개발

실습을 통한 문제 해결 능력 배양

How works?

Education 4.0 Program: **active Learning**



프로젝트 중심 학습(PjBL): 10 assignments and 1 final project

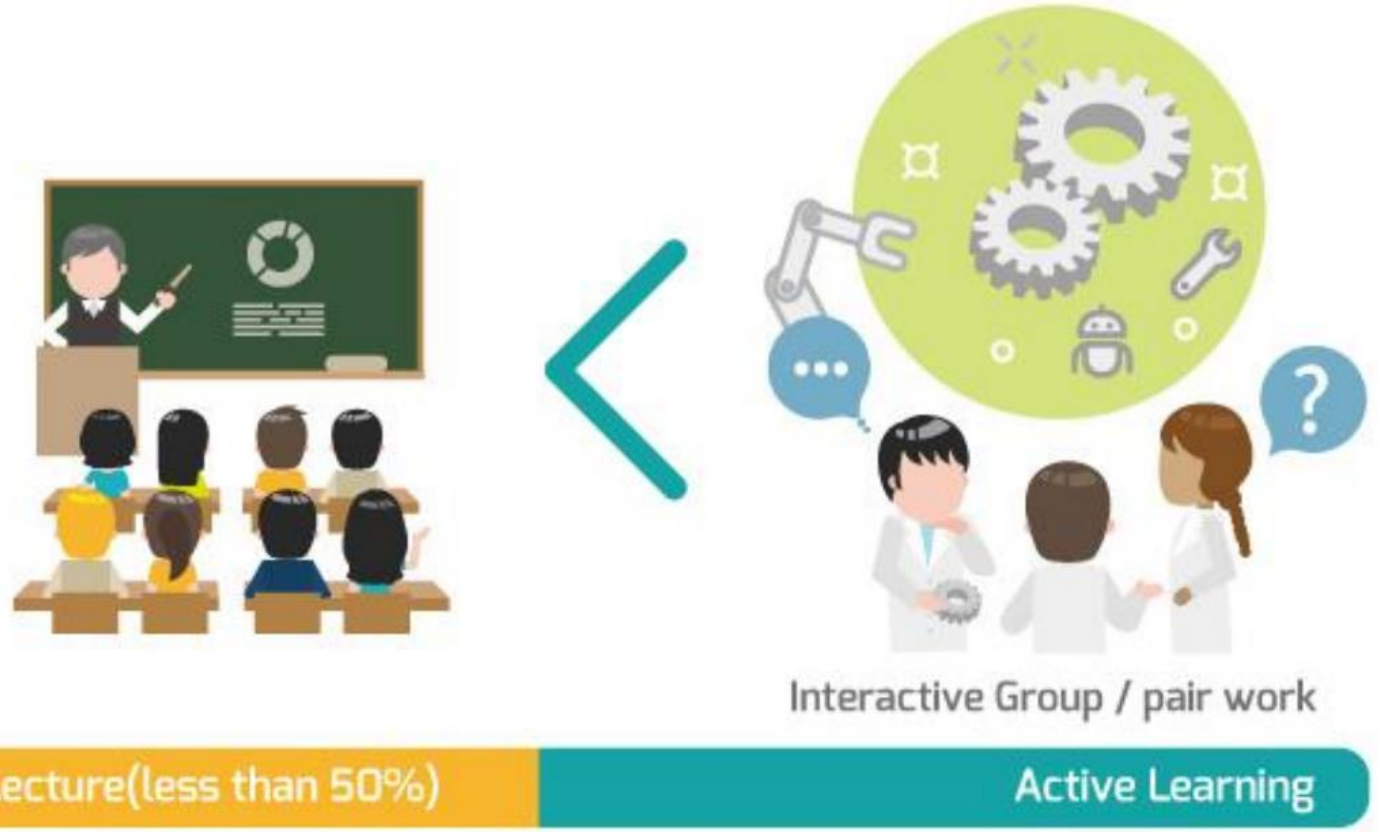
토의, 토론

동료학습: 2인 1조, assignments와 project는 각 개인 제출

Track 2: Reducing one-way lecture + Encouraging students to achieve learning goals by participating in learning activities

Edu 4.0

- Track 1 (Edu 3.0)
- Track 2
- Track 3



Course Schedule

기간	분류	주제	학습활동	결과물
1주	주제	Introduction & math review		
	목표	Review of fundamental mathematics required to follow the course works.		
	내용	Data science, Linear algebra, Probability		
2주	주제	Machine learning fundamentals	Installation of anaconda, python3, numpy, scikit-learn, RDkit, tensorflow, exercising linear regression	Assign #1: linear regression
	목표	Understanding the basic principle of machine learning such as cost function and gradient descent.		
	내용	Linear regression, logistic classification		
3주	주제	Support vector machine (SVM) & summary	Exercising SVM for classification problem	Assign #2: regression using SVM
	목표	Understanding a key idea of SVM		
	내용	SVM, Regression and classification		
4주	주제	Deep learning & multilayer perceptron (MLP)	Applying MLP for classification problem and comparison between ReLU and sigmoid functions	
	목표	Understanding the perceptron concept and a basic principle of deep learning		
	내용	Universal approximation theorem backpropagation, vanishing gradient, activation function, ReLU		

Course Schedule

기간	분류	주제	학습활동	결과물
5주	주제	Multilayer perceptron 2	Exercising MLP for supervised learning	Assign #3: supervised learning with MLP and comparison with SVM
	목표	Knowing various issues on MLP and techniques to resolve them		
	내용	Overfitting, regularization, dropout, batch normalization, cross validation		
6주	주제	Convolutional Neural Network (CNN) & SMILES	Exercising CNN with SMILES for supervised learning of Log P and TPSA Ref. (1)	Assign #4: supervised learning of various molecular properties with CNN
	목표	Understanding CNN and molecular representation with SMILES		
	내용	Convolution, receptive field, stride, pooling Supervised learning of Log P and TPSA		
7주	주제	Molecular graphs & Graph Neural Network (GNN)	Exercising GCN with molecular graphs for supervised learning of Log P and TPSA Ref. (2), (3), (4)	Assign #5: improvement of vanilla GCN early-feedback(CELT)

Course Schedule

기간	분류	주제	학습활동	결과물
9주	주제	Recurrent neural network (RNN)	Exercising RNN with SMILES for supervised learning of Log P and TPSA Ref. (5)	Assign #6: supervised learning with RNN and comparison to GCN and SVM
	목표	Understanding RNN and molecular representations with SMILES		
	내용	RNN, LSTM, GRU, Feature extraction of molecules using RNN		
10주	주제	Message Passing Neural Network (MPNN)	Exercising GGNN with molecular graphs for supervised learning of Log P and TPSA Ref. (4), (6)	Assign #7: supervised learning with GGNN and comparison to GCN, GAT, RNN
	목표	Understanding the most general expression of graph neural network		
	내용	MPNN, molecular graph representation, GGNN, supervised learning of logP and TPSA		
11주	주제	Molecular generative model 1	Exercising VAE and CVAE for molecular design Ref. (7)	Assign #8: Optimization of molecular properties on latent space
	목표	Understanding the principle of autoencoder and unsupervised learning		
	내용	Molecular autoencoder, VAE, CVAE, de novo molecular design		
12주	주제	Molecular generative model 2	Molecular design from continuous latent space Ref. (8), (9)	Assign #9: comparison to the result of assign #8
	목표	Understanding difference between GAN and VAE		
	내용	GAN, ARAE ARAE: conditional molecular design		

Course Schedule

기간	분류	주제	학습활동	결과물
13주	주제	Molecular generative model 3	Molecular design with graph generative models Ref. (10), (11), (12)	Assign #10: scaffold-based molecular design
	목표	Understanding and graph structure based generative models		
	내용	Graph generative model, MolGAN, JTVAE		
14주	주제	No lecture (entrance interview)		
	목표			
	내용			
15주	주제	Term project presentation	Student presentation for the results of their own term project	final feedback(CELT)
	목표			
	내용			
16주	주제	Final exam	Student presentation for the results of their own term project	
13주	목표			
	내용			

Grade

Evaluation: Each assignment (7 %), Final project (30 %), Bonus (10 %)
No attendance score, but if missing more than 7 times, F will be given regardless of the other records.

Grade: Absolute evaluation.
90-100pts: A+, 80-90pts: A0, 70-80pts: A-, 60-70pts: B+, 50-60pts: B0, 40-50pts: B-, 30-40pts: C+, 25-30pts: C0, 20-25pts: C-, 15-20pts: D+, 10-15pts: D0, <10pts: F
If not finishing the final project, F will be given regardless of the other records

Bonus Points: Bonus points will be given to those who actively participate in the class (ex. Questions). Either in class or after class, you may raise questions. A Q&A board is given through the KLMS site for the after-class participation.

Course Website: KLMS and CLASSUM App will be utilized. Important notices including assignments release will be posted on this website. https://www.welcome.classum.kr/
Assignments will be posted on KLMS and submit your assignments to KLMS.

Bring your laptop

TA

god_seongok@kaist.ac.kr
T. 350-2855
E6-4, 3120

Seongok Ryu (5th year)

- **Graph Neural Network**
- **Generative Model**
- **Quantum Chemistry**