

# 機器學習於材料資訊的應用

## Machine Learning on Material Informatics

---

陳南佑(NAN-YOW CHEN)

[nanyow@narlabs.org.tw](mailto:nanyow@narlabs.org.tw)

楊安正(AN-CHENG YANG)

[acyang@narlabs.org.tw](mailto:acyang@narlabs.org.tw)

# Data comes from Computational Materials Science

A new powerful approach to  
discover novel material

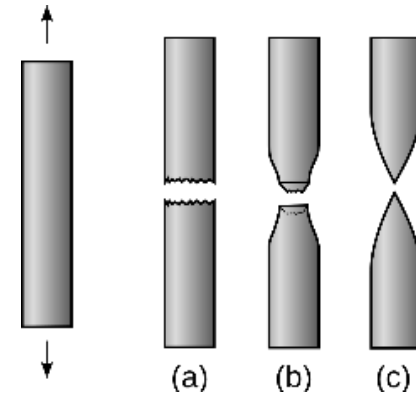
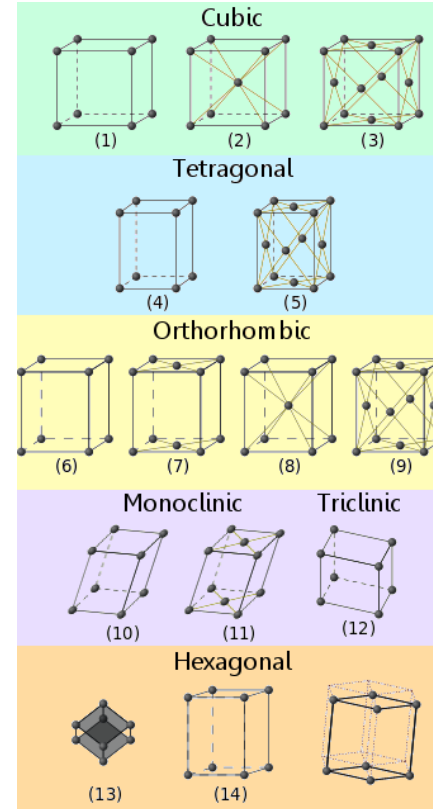


# 材料基因

## Human Genome V.S. Material Genome

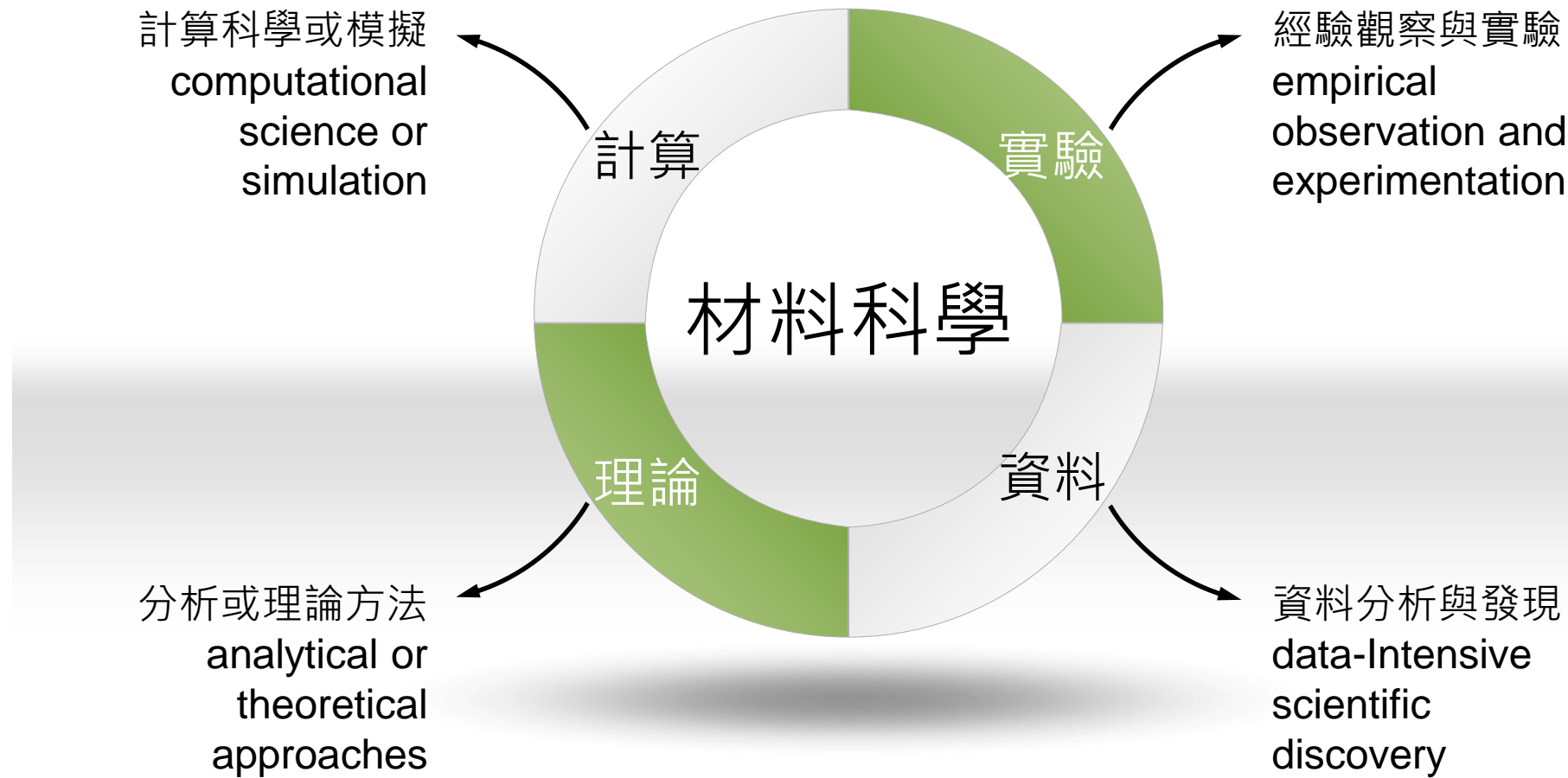


生物個體的外在表現叫作「表現型 (phenotype)」，由內在的「基因型 (genotype)」決定。改變表現型，首先要改變基因型。

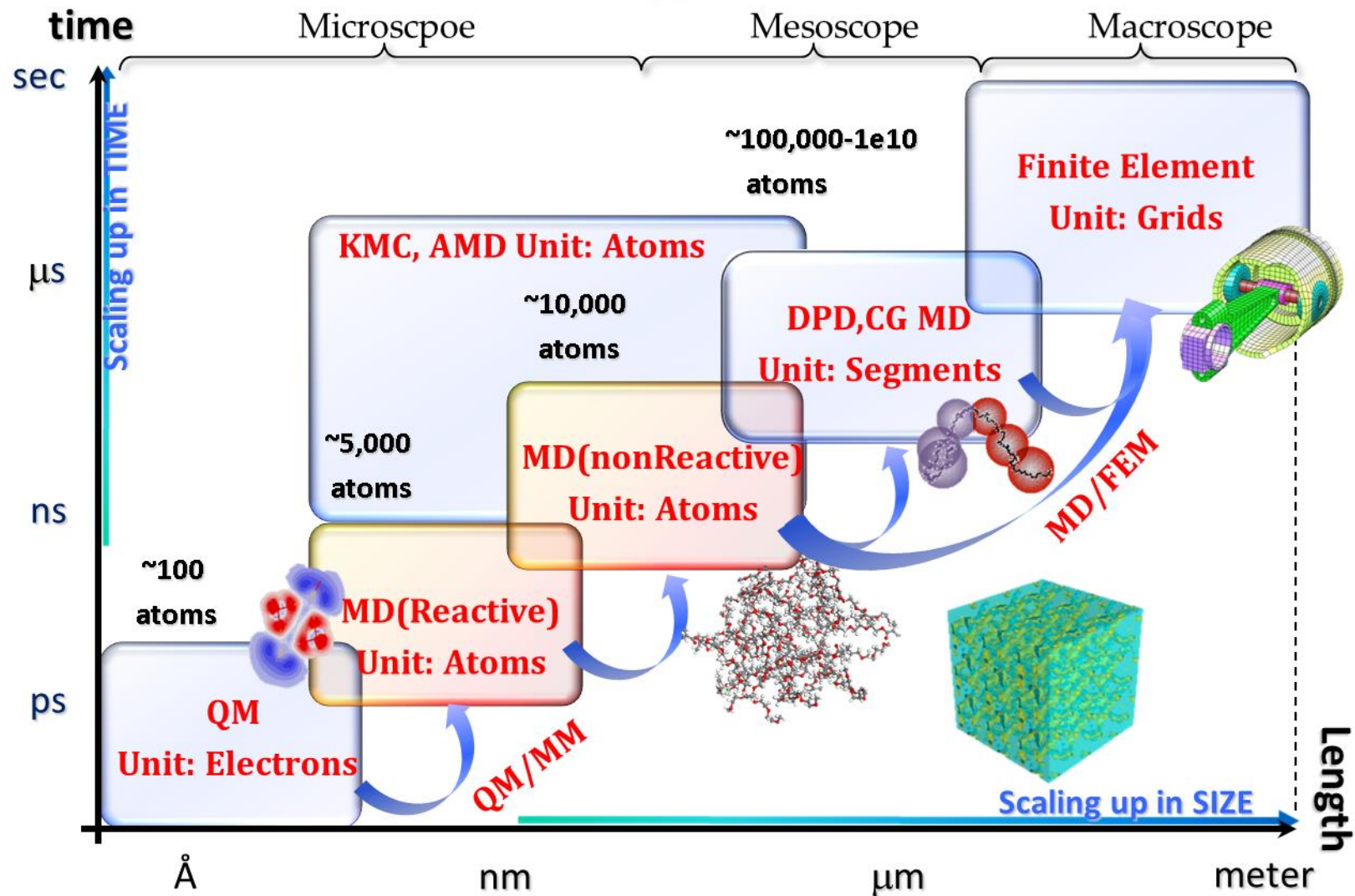


材料在巨觀尺度表現出來的材料特性是由微觀尺度下的原子組態所決定的。要改變材料特性，首先要改變微觀的原子組態。

# 材料科學的分支

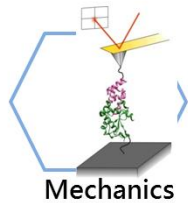


# Multi-Scale Modeling for Materials Simulation

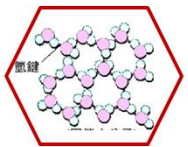




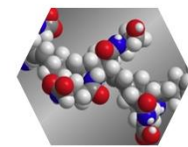
# Properties you can get from calculation



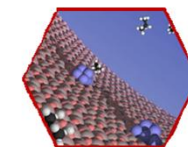
Elastic constant  
Young's modulus  
Stress-strain  
Hardness



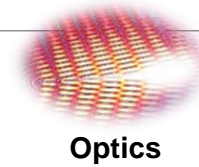
Glass transition temperature  
Coefficient of Thermal expansion  
Thermal conductivity(Green Kubo)  
Phonon density  
Solubility Parameters



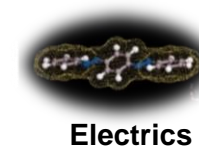
Radius distribution function  
Bond length, angle, dihedral angle  
Gyration Radius  
Atomic shear strain



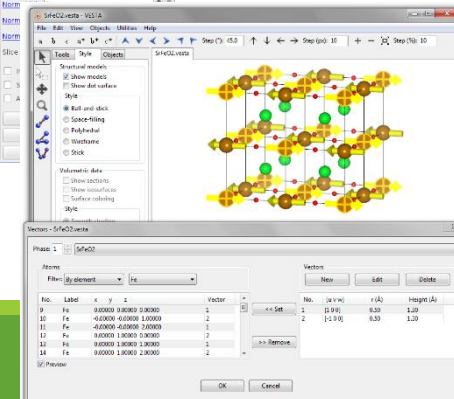
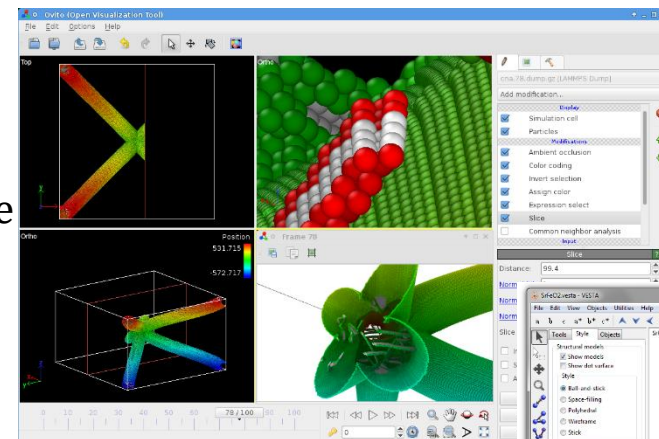
Diffusion coefficient  
Viscosity (Green Kubo)  
Auto-correlation function



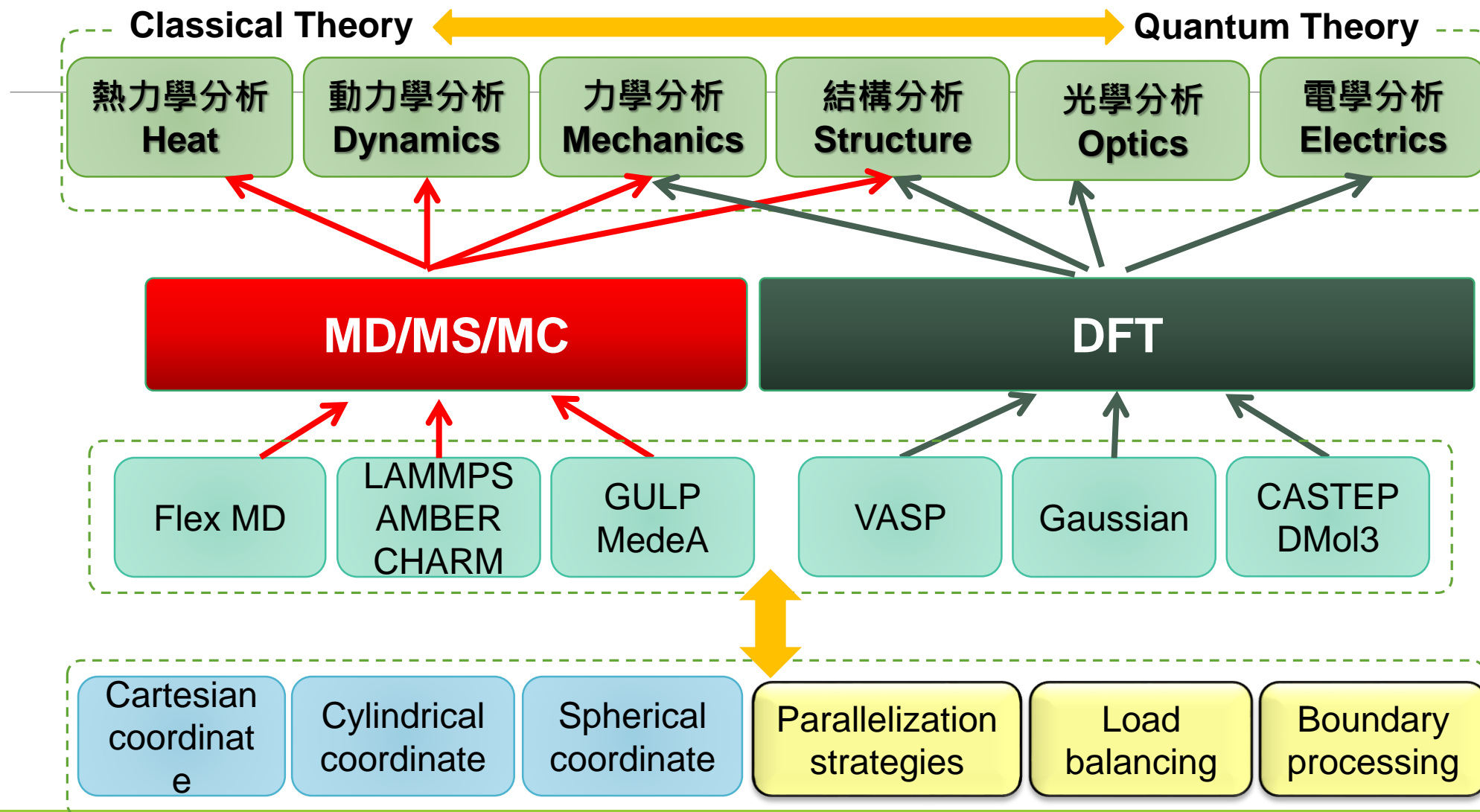
Birefringence  
Dielectric constant  
Abbe Number  
Reflective Index



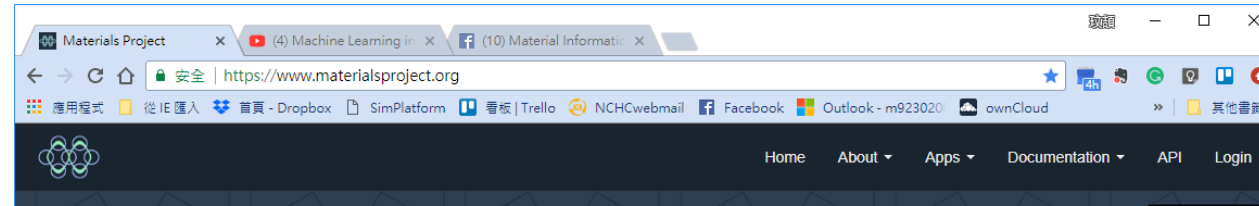
Band structure/gap  
Density of state



# Molecular Modeling



# Database Material Project



## The Materials Project

Harnessing the power of supercomputing and state of the art electronic structure methods, the Materials Project provides open web-based access to computed information on known and predicted materials as well as powerful analysis tools to inspire and design novel materials.

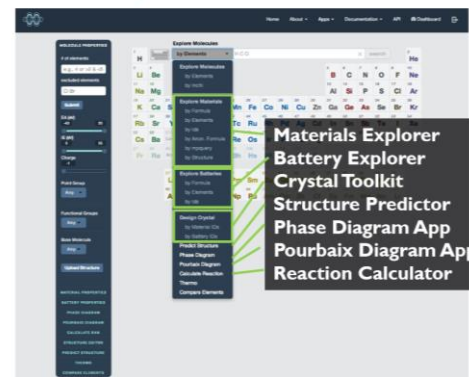
Learn more

YouTube Tutorials

Sign In or Register

to start using

User-friendly web interface  
(but unfriendly for advanced  
users requiring large quantities  
of data!)



Materials Explorer  
Battery Explorer  
Crystal Toolkit  
Structure Predictor  
Phase Diagram App  
Pourbaix Diagram App  
Reaction Calculator



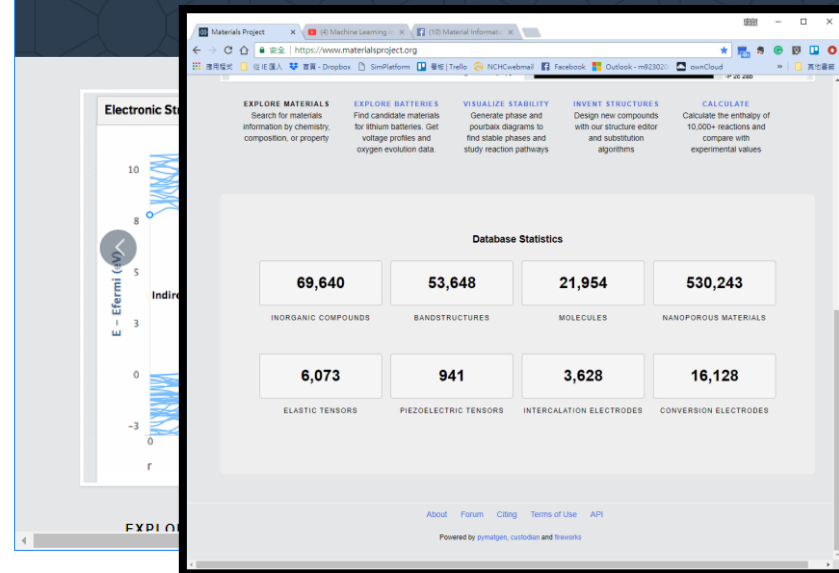
Energetic properties

Structure

Electronic Structure

XRD

Elastic properties



Development Jain et al. API Mater. 2013, 1, 11002

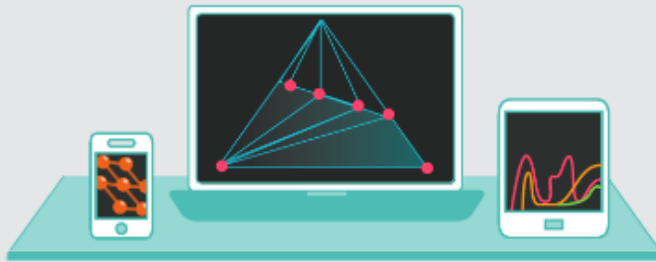
Energy Sciences (BES) and Advanced Scientific Computing Research (ASCR) programs, and through its Office of Energy Efficiency and Renewable Energy (EERE), via the Battery Materials Research (BMR, formerly BATT) program. A notable source of support within DOE-BES is the Joint Center for Energy Storage Research (JCESR).

The Materials Project is also supported by a Laboratory Directed Research and Development grant from LBNL. Disseminated science is supported by DOE (BES and BMR), the National Science Foundation (NSF), Gillette, Volkswagen, Umicore, and Bosch.





# About Material Project



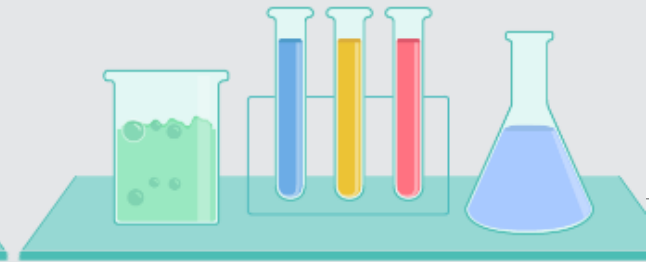
## software

By computing properties of all known materials, the Materials Project aims to remove guesswork from materials design in a variety of applications. Experimental research can be targeted to the most promising compounds from computational data sets. Researchers will be able to data-mine scientific trends in materials properties. By providing materials researchers with the information they need to design better, the Materials Project aims to accelerate innovation in materials research.



## supercomputers

Supercomputing clusters at national laboratories provide the infrastructure that enables our computations, data, and algorithms to run at unparalleled speed. We principally use the Lawrence Berkeley National Laboratory's [NERSC](#) Scientific Computing Center and Computational Research Division, but we are also active with Oak Ridge's [OLCF](#), Argonne's [ALCF](#), and San Diego's [SDSC](#).



## screening

Computational materials science is now powerful enough that it can predict many properties of materials before those materials are ever synthesized in the lab. By scaling materials computations over supercomputing clusters, we have predicted several new battery materials which were made and tested in the lab. Recently, we have also identified new transparent conducting oxides and thermoelectric materials using this approach.



Search for materials information by chemistry, composition, or property

### Explore Materials

Advanced Search Syntax

1 H	<div>by Elements</div> <div>Na-O</div> <div>search</div>																2 He	
3 Li	4 Be																	10 Ne
11 Na	12 Mg																	18 Ar
19 K	20 Ca	21 Sc	22 Ti	23 V	24 Cr	25 Mn	26 Fe	27 Co	28 Ni	29 Cu	30 Zn	31 Ga	32 Ge	33 As	34 Se	35 Br	36 Kr	
37 Rb	38 Sr	39 Y	40 Zr	41 Nb	42 Mo	43 Tc	44 Ru	45 Rh	46 Pd	47 Ag	48 Cd	49 In	50 Sn	51 Sb	52 Te	53 I	54 Xe	
55 Cs	56 Ba	57-71 La-Lu	72 Hf	73 Ta	74 W	75 Re	76 Os	77 Ir	78 Pt	79 Au	80 Hg	81 Tl	82 Pb	83 Bi	84 Po	85 At	86 Rn	
87 Fr	88 Ra	89-103 Ac-Lr	104 Rf	105 Db	106 Sg	107 Bh	108 Hs	109 Mt	110 Ds	111 Rg	112 Cn							
57 La	58 Ce	59 Pr	60 Nd	61 Pm	62 Sm	63 Eu	64 Gd	65 Tb	66 Dy	67 Ho	68 Er	69 Tm	70 Yb	71 Lu				
89 Ac	90 Th	91 Pa	92 U	93 Np	94 Pu	95 Am	96 Cm	97 Bk	98 Cf	99 Es	100 Fm	101 Md	102 No	103 Lr				

輸入搜尋條件

也可以直接點  
週期表元素

搜尋結果的  
FILTER

主要搜尋結果會呈現在這邊

# of elements

e.g., 4 or >2 & <6

excluded elements

Cl Br

Submit

External Provenance

ICSD

Exptl. ICSD

Material Tags

imgreite

Band Gap (eV)

0 10

Energy Above Hull

0 6

Formation Energy

-4 4

# unit cell sites

1 298

Density

0 24.6

Volume

7 7697

Nelements		Elements							
100 records per page		Batch Structures		Edit Structures		Show / hide columns		Print	Export
Materials Id	Formula	Spacegroup	Formation Energy (eV)	E Above Hull (eV)	Band Gap (eV)	Volume	Nsites	Density (gm/cc)	
mp-1027525	MoS <sub>2</sub>	P $\bar{3}$ m1	-1.306	0	1.746	350.447	12	3.034	<input type="checkbox"/>
mp-1025874	MoS <sub>2</sub>	P $\bar{6}$ m2	-1.306	0	1.783	284.872	9	2.799	<input type="checkbox"/>
mp-1023939	MoS <sub>2</sub>	P $\bar{3}$ m1	-1.306	0	1.551	219.296	6	2.424	<input type="checkbox"/>
mp-1018809	MoS <sub>2</sub>	P6 <sub>3</sub> /mmc	-1.305	0.001	1.338	123.449	6	4.306	<input type="checkbox"/>
mp-1023924	MoS <sub>2</sub>	P $\bar{6}$ m2	-1.305	0.001	1.658	153.721	3	1.729	<input type="checkbox"/>
mp-2815	MoS <sub>2</sub>	P6 <sub>3</sub> /mmc	-1.304	0.002	1.229	118.065	6	4.503	<input type="checkbox"/>
mp-1434	MoS <sub>2</sub>	R3m	-1.303	0.003	1.228	116.637	6	4.558	<input type="checkbox"/>
mp-2164	Mo <sub>3</sub> S <sub>4</sub>	R $\bar{3}$	-1.052	0.068	0.000	272.96	14	5.062	<input type="checkbox"/>
mp-31257	Mo <sub>15</sub> S <sub>19</sub>	P6 <sub>3</sub> /m	-1.011	0.084	0.000	1436.887	68	4.734	<input type="checkbox"/>
mp-1627	Mo <sub>2</sub> S <sub>3</sub>	P2 <sub>1</sub> /m	-1.08	0.096	0.000	167.797	10	5.702	<input type="checkbox"/>
mp-1104577	Mo <sub>3</sub> S <sub>4</sub>	P $\bar{1}$	-1.006	0.113	0.000	229.105	14	6.031	<input type="checkbox"/>
mp-673645	Mo <sub>7</sub> S <sub>8</sub>	P1	-0.919	0.126	0.000	543.214	30	5.674	<input type="checkbox"/>
mp-990083	MoS <sub>2</sub>	Pmmn	-1.142	0.164	0.000	1790.583	54	2.672	<input type="checkbox"/>
mp-1210708	Mo <sub>21</sub> S <sub>8</sub>	I4/m	-0.289	0.251	0.000	431.006	29	8.75	<input type="checkbox"/>
mp-1239169	MoS <sub>3</sub>	P2 <sub>1</sub> /m	-0.719	0.26	0.791	209.34	8	3.048	<input type="checkbox"/>
mvc-11780	MoS <sub>2</sub>	F $\bar{4}$ 3m	-1.041	0.265	0.000	221.127	12	4.808	<input type="checkbox"/>
mp-558544	MoS <sub>2</sub>	R $\bar{3}$ m	-1.029	0.277	0.000	111.639	6	4.762	<input type="checkbox"/>

Formation Energy

-4 4

# unit cell sites

1 296

Density

0 24.6

Volume

7 7697

Crystal Systems

Any

Spacegroup Number

Any

Spacegroup Symbol

Any

Has properties →

Elasticity →

Piezoelectricity →

Dielectricity →

可以直接點  
進去看細節

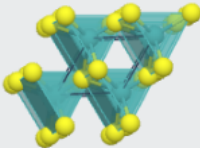
也可以批次  
操作

將結構視覺化  
的區域

MATERIAL ID: DOI: [Show Help Guides](#)

**MoS<sub>2</sub>** **mp-1434** **10.17188/1190621**

[Electronic Structure](#) [X-Ray Diffraction](#) [X-Ray Absorption](#) [Substrates](#) [Elasticity](#) [Piezoelectricity](#) [Dielectric Properties](#)  
[Similar Structures](#) [Synthesis Descriptions](#) [Calculation Summary](#) [User Contributions](#) [Provenance/Citation](#)



**Material Details**

Final Magnetic Moment  
0.000  $\mu_B$

Magnetic Ordering  
NM

Formation Energy / Atom  
-1.303 eV

Energy Above Hull / Atom  
0.003 eV

Density  
4.56 g/cm<sup>3</sup>

Decomposes To  
[MoS<sub>2</sub>](#)

Band Gap  
1.228 eV

**Lattice Parameters**

a 3.194 Å  $\alpha$  97.687°  
b 5.530 Å  $\beta$  76.529°  
c 6.855 Å  $\gamma$  90.000°  
Volume 116.637 Å<sup>3</sup>

**Final Structure** [CIF](#)  
Fractional Coordinates

Mo		
a	b	c
0	0	0.0001
0.5	0.5	0.0001

S		
a	b	c
0.1187	0.6272	0.7626
0.3813	0.2061	0.2374
0.6187	0.1272	0.7626
0.8813	0.7061	0.2374

**Space Group**

Hermann Mauguin  
R3m [160]

Hall  
R 3  $\bar{2}$

Point Group  
m

Crystal System  
trigonal

**Mo** **S** ☒ Atoms Unit Cell ☒ Bonds Polyhedra [CIF](#)

Zoom in/out Rotate along the center axis Shift + Drag cursor Option + Drag cursor

[Edit Crystal](#) [Generate Phase Diagram](#)

Tags: [Molybdenite 3R](#) [Molybdenum\(IV\) disulfide](#) [Molybdenum\(IV\) sulfide](#)

[File Formats](#) [Download](#)

下載資料  
的地方

CIF:晶體結構的標準檔案交換格式

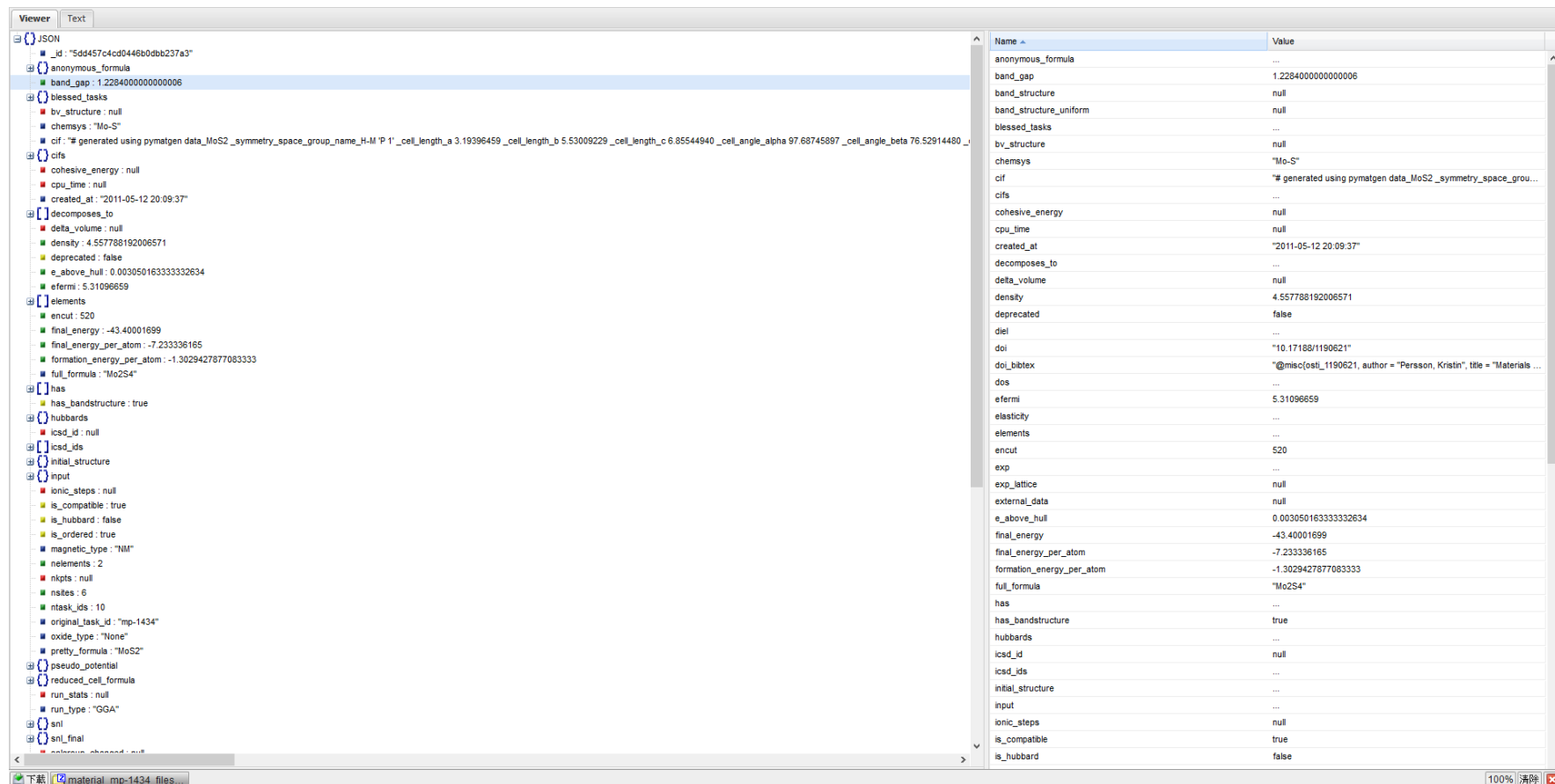
VASP:執行VASP計算的設定檔案(不包含最重要的POTCAR)

POSCAR:原子的座標輸入

CSSR: unit cell 定義和原子的fractional coordinates.

JSON:JSON格式的所有資訊

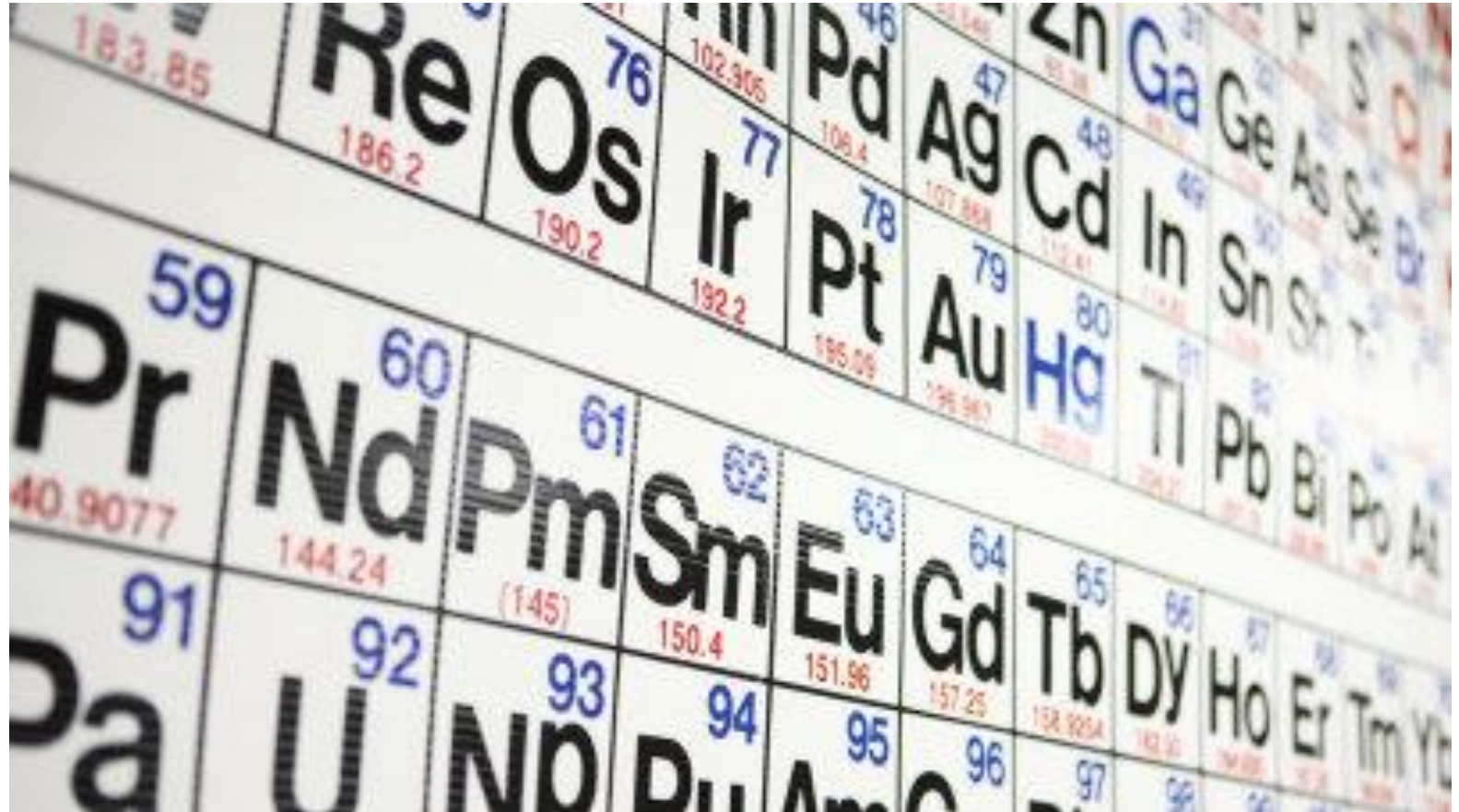
<http://jsonviewer.stack.hu/>





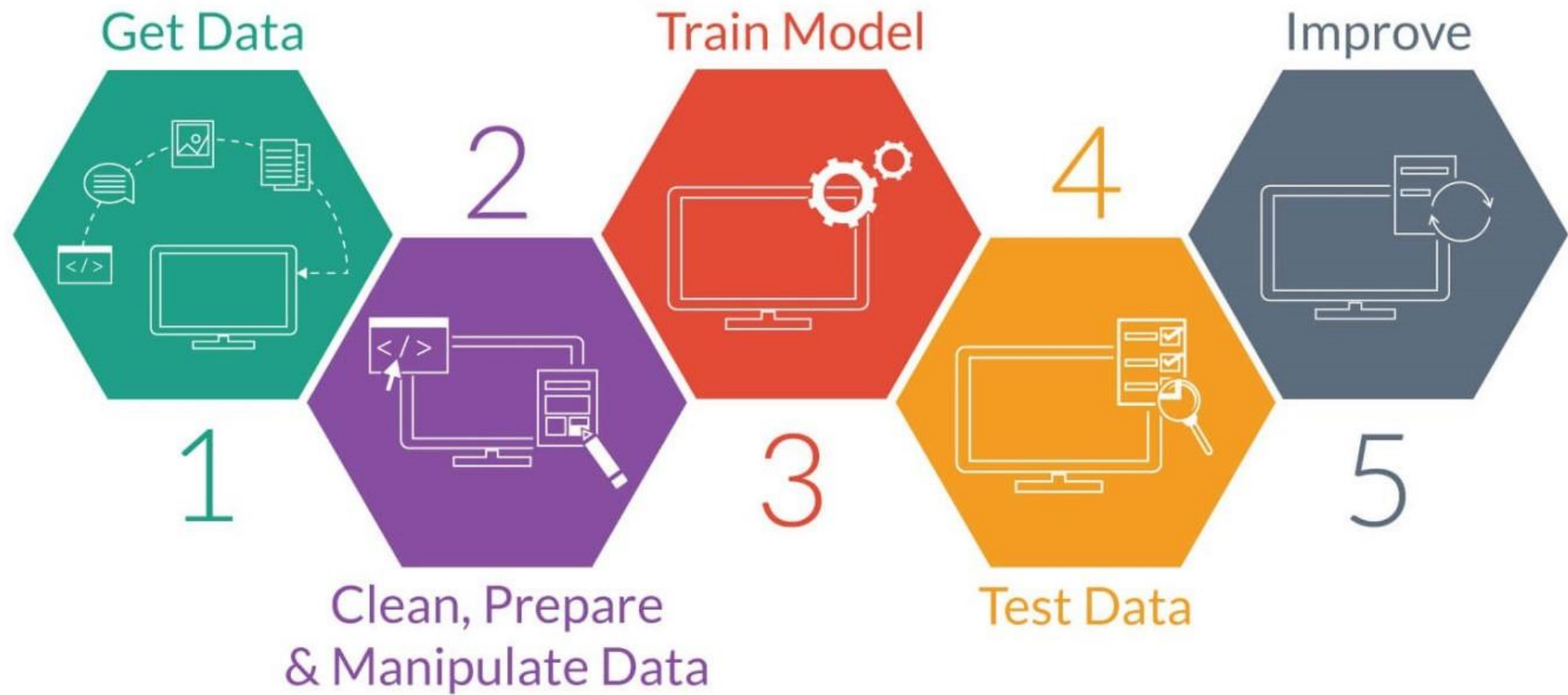
# Material Properties Prediction

digging into the periodic table



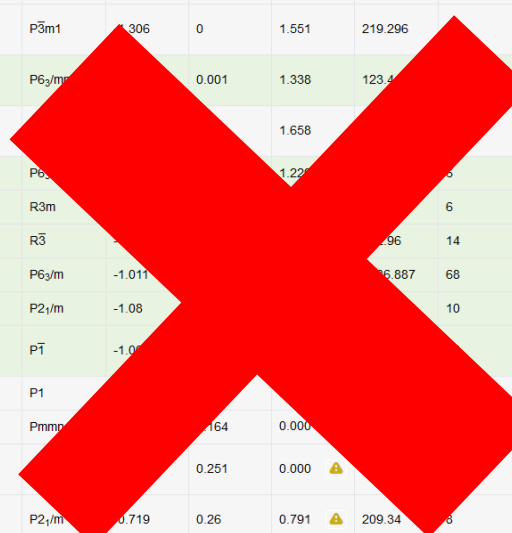
# Problem Define

- 在早期，元素被發現的種類不多，化學家只能局部的對某些性質相近的元素進行歸類整理，例如1865年英國化學家伍德林（W. Woodling）按原子量排列元素順序，初步排出今日元素週期表中的鹵族、氮族、氧族。
- 俄國化學家門得列夫（Dmitri Ivanovich Mendeleev, 1834 – 1907）全面考慮了元素的各種性質，不僅根據元素的原子量，而且很重視元素的性質及其與其他元素的關係，他依原子量遞增的順序把元素排列成幾行，同時把各行中性質相似的元素左右對齊，這樣使得每一橫排化學元素的性質相近，每一縱列化學元素性質的變化也呈現着規律性，整個元素系列呈現出周期性變化。  
資料的整理歸納
- 1869年2月，門得列夫發表了《元素性質和原子量的關係》論文，同時公布了他的第一張化學元素週期表，周期表中留下了四個空位，空位上沒有元素名稱，只有預計的原子量，表示尚待發現的元素。  
新元素的預測
- 那化合物的特性能不能找出週期性？
- 可不可以從既有的材料資料庫預測出新材料的性質？



# Get Data

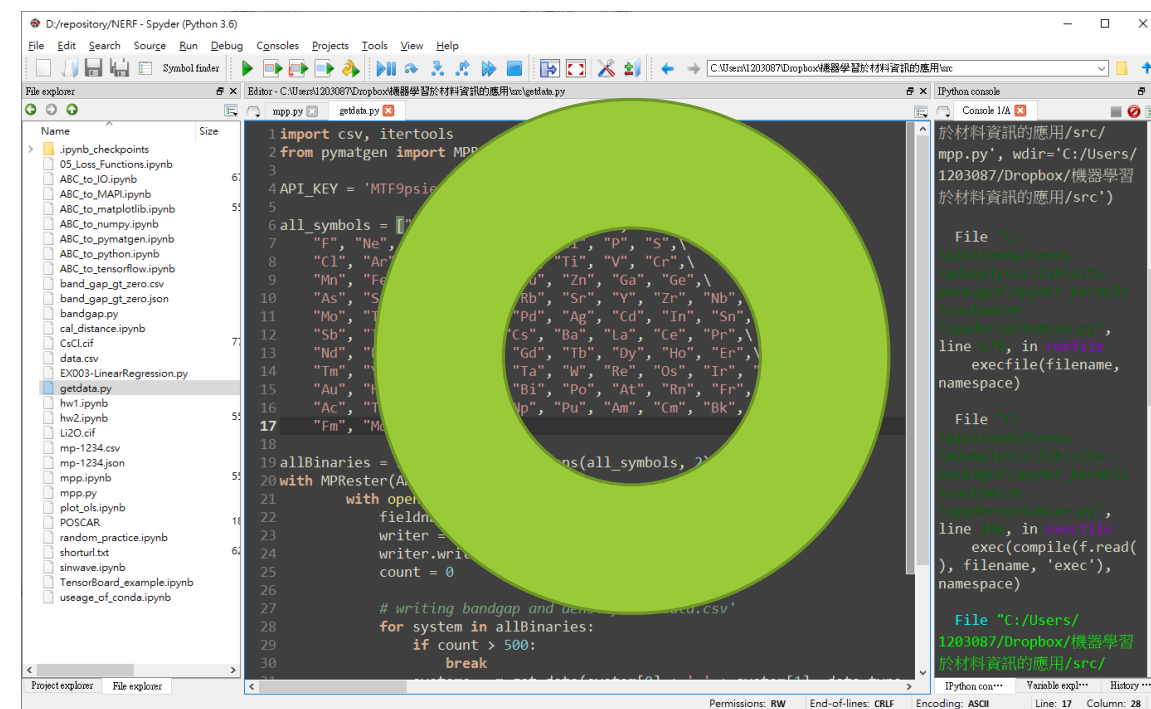
## 手動下載



Materials Id	Formula	Spacegroup	Formation Energy (eV)	E Above Hull (eV)	Band Gap (eV)	Volume	Nsites	Density (gm/cc)	
mp-1027525	MoS <sub>2</sub>	P3m1	-1.306	0	1.746	350.447	12	3.034	
mp-1025874	MoS <sub>2</sub>	P6m2	-1.306	0	1.783	284.872	9	2.799	
mp-1023939	MoS <sub>2</sub>	P3m1	-1.306	0	1.551	219.296		2.424	
mp-1018809	MoS <sub>2</sub>	P6 <sub>3</sub> /mm2	-1.306	0.001	1.338	123.4		4.306	
mp-1023924	MoS <sub>2</sub>				1.658			1.729	
mp-2815	MoS <sub>2</sub>	P6 <sub>3</sub> /mm2			1.22			4.503	
mp-1434	MoS <sub>2</sub>	R3m					6	4.558	
mp-2164	Mo <sub>3</sub> S <sub>4</sub>	R3				96	14	5.062	
mp-31257	Mo <sub>15</sub> S <sub>19</sub>	P6 <sub>3</sub> /m	-1.011			6.887	68	4.734	
mp-1627	Mo <sub>2</sub> S <sub>3</sub>	P2 <sub>1</sub> /m	-1.08				10	5.702	
mp-1104577	Mo <sub>3</sub> S <sub>4</sub>	P1	-1.0					6.031	
mp-673645	Mo <sub>7</sub> S <sub>9</sub>	P1						5.674	
mp-990083	MoS <sub>2</sub>	Pmmn		0.164	0.000			2.672	
mp-1210708	Mo <sub>21</sub> S <sub>8</sub>			0.251	0.000			8.75	
mp-1239169	MoS <sub>3</sub>	P2 <sub>1</sub> /m	0.719	0.26	0.791	209.34	8	3.048	
mp-11780	MoS <sub>2</sub>	F43m	-1.041	0.265	0.000	221.127	12	4.808	
mp-558544	MoS <sub>2</sub>	R3m	-1.029	0.277	0.000	111.639	6	4.762	

## 自動化下載

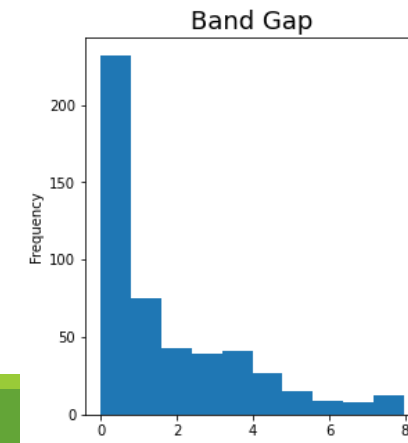
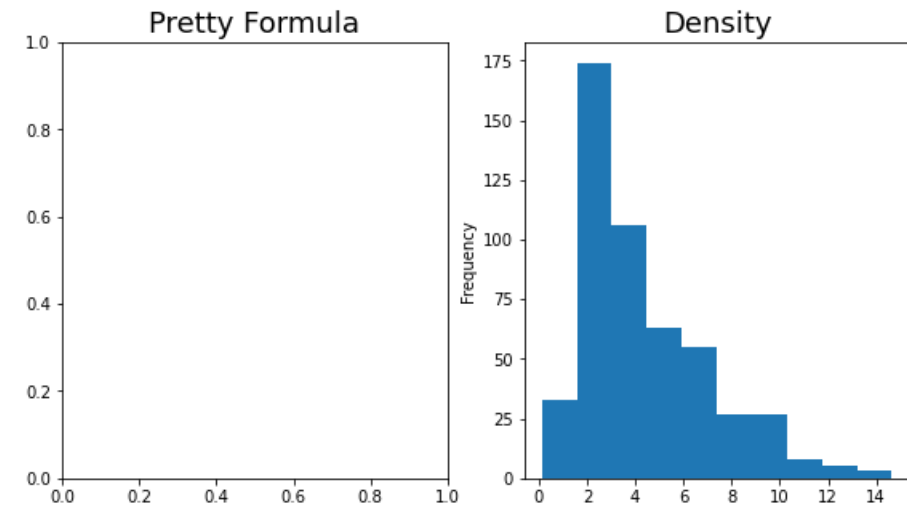
open Materials Application Programming Interface (API)



```
1 import csv, itertools
2 from pymatgen import MP
3
4 API_KEY = 'MTF9psie'
5
6 all_symbols = [
7     "F", "Ne", "p", "S",
8     "Cl", "Ar", "Ti", "V", "Cr",
9     "Mn", "Fe", "Co", "Ni", "Cu", "Zn", "Ga", "Ge",
10    "As", "Se", "Br", "Kr", "Rb", "Sr", "Y", "Zr", "Nb",
11    "Mo", "Tc", "Pd", "Ag", "Cd", "In", "Sn",
12    "Sb", "Te", "Gd", "Tb", "Dy", "Ho", "Er",
13    "Nd", "Pm", "Ta", "W", "Re", "Os", "Ir",
14    "Au", "Hg", "Pt", "Bi", "Po", "At", "Rn", "Fr",
15    "Ac", "Th", "Pa", "U", "Np", "Pu", "Am", "Cm", "Bk",
16    "Fm", "Md", "No", "Lr"
17 ]
18
19 allBinaries = []
20 with MPRester(API_KEY) as r:
21     with open('data.csv', 'w') as writer:
22         writer.write('system,formation_energy_per_atom,band_gap,elastic_modulus\n')
23         for system in allBinaries:
24             if count > 500:
25                 break
26             # writing bandgap and density to data.csv
27
28 # writing bandgap and density to data.csv
29 for system in allBinaries:
30     if count > 500:
31         break
```

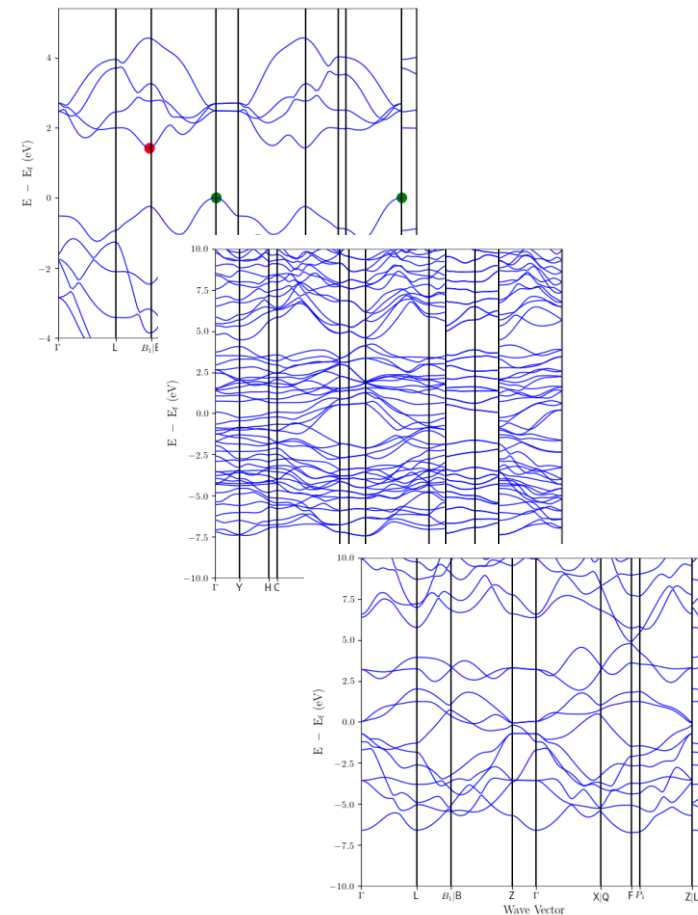
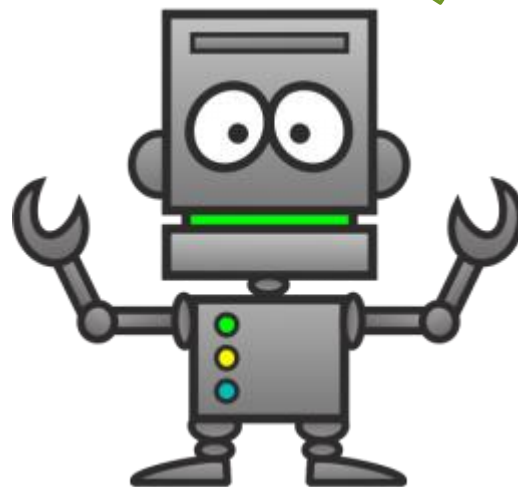
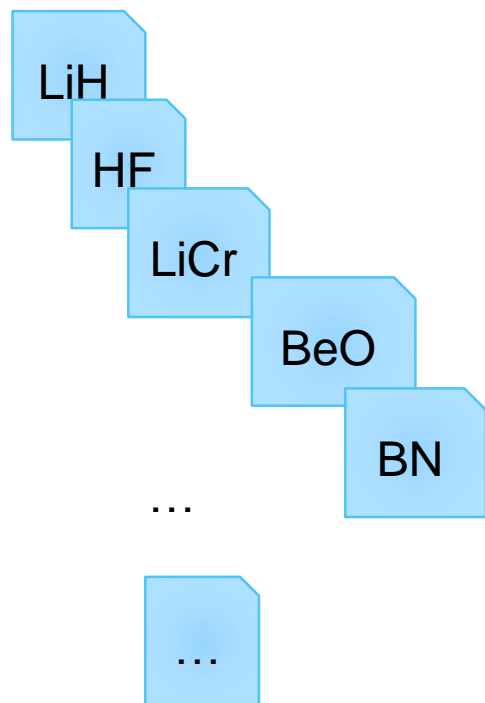
# Clean, Prepare, Manipulate Data

Pretty Formula	Density	Band Gap
LiH	0.814395727	3.018
BeH2	0.807534506	5.3418
B3H5	0.811598971	3.6983
HC	1.409269759	3.091
HN	1.336811255	4.1598
H2O2	1.834767366	4.1525
HF	1.721540912	6.7187
NaH	1.394380186	3.7974
MgH2	1.450102639	3.6284
AlH3	1.459511166	2.1855
...	...	...





# Train Model



這是你寫的門得烈夫機器人

# Model

## Finding a function from data

$f_1(\text{LiH}) = \text{aaaaaa}$      $f_2(\text{LiH}) = \text{eeeeee}$   
 $f_1(\text{HF}) = \text{bbbbbb}$      $f_2(\text{HF}) = \text{ffffff}$   
 $f_1(\text{LiCr}) = \text{cccccc}$      $f_2(\text{LiCr}) = \text{gggggg}$     .....  
 $f_1(\text{BeO}) = \text{dddddd}$      $f_2(\text{BeO}) = \text{hhhhh}$

訓練的過程說穿了就是找出一個合適  
的function來描述輸入和輸出的關係

# Scikit-Learn Regression algorithm

---

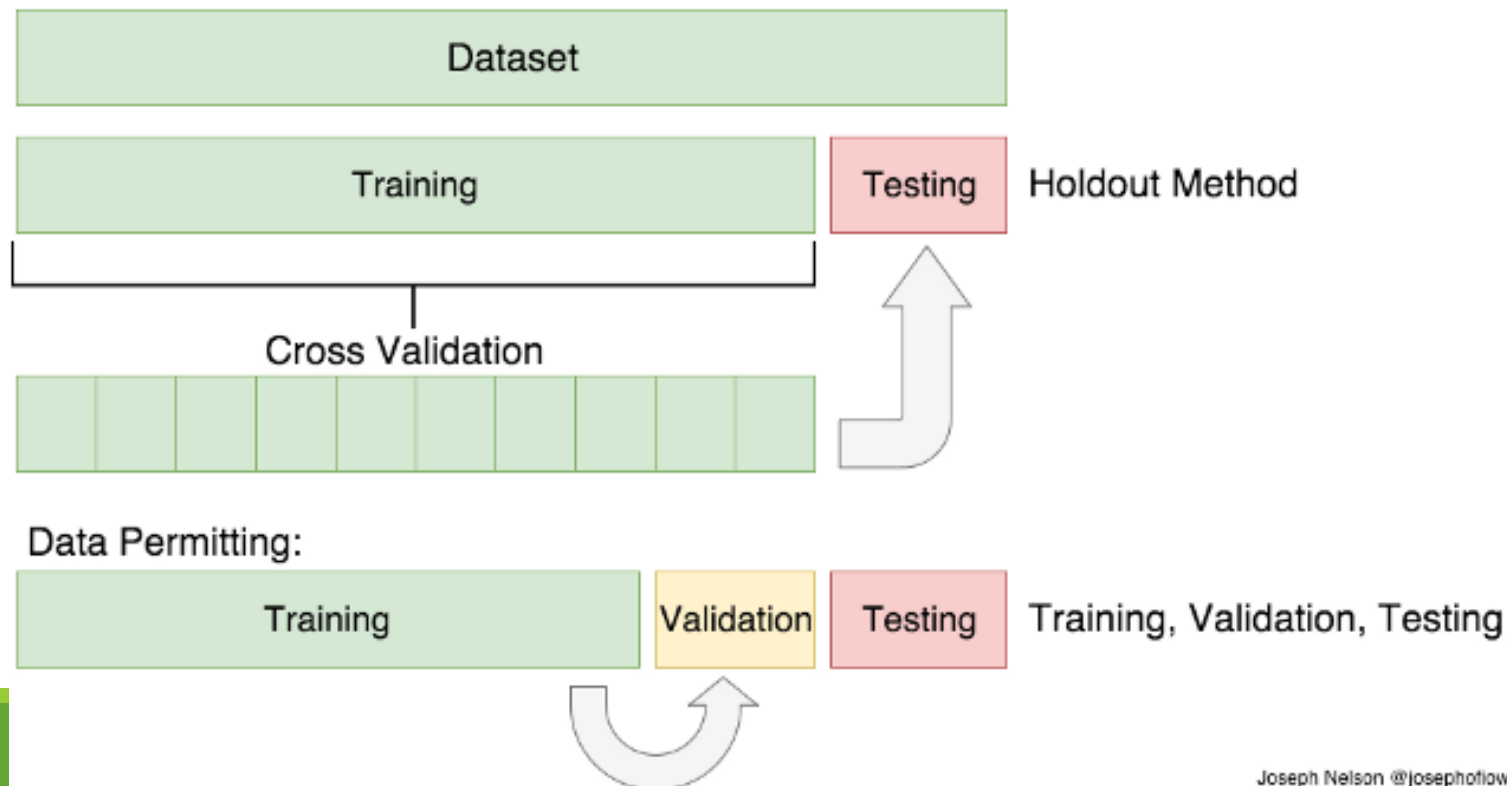


- Linear Models
- Kernel ridge regression
- Support Vector Machines
- Gaussian Processes
- Decision Trees
- Ensemble methods
- ...

[https://scikit-learn.org/stable/supervised\\_learning.html#supervised-learning](https://scikit-learn.org/stable/supervised_learning.html#supervised-learning)

# Test Data (Test Model)

- 不要把所有所有的資料都餵進去給model，只要把一部分的資料餵進去(Training Dataset)訓練模型，需要保留一些資料拿來檢驗模型(Testing Dataset)。
- Cross Validation(交叉驗證)的部份之後會再講。



# Train/Test Split

---

□ sklearn.model\_selection.train\_test\_split

```
1 from sklearn.model_selection import train_test_split
2 X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_state = 0)
```

[https://scikit-learn.org/stable/modules/generated/sklearn.model\\_selection.train\\_test\\_split.html](https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html)