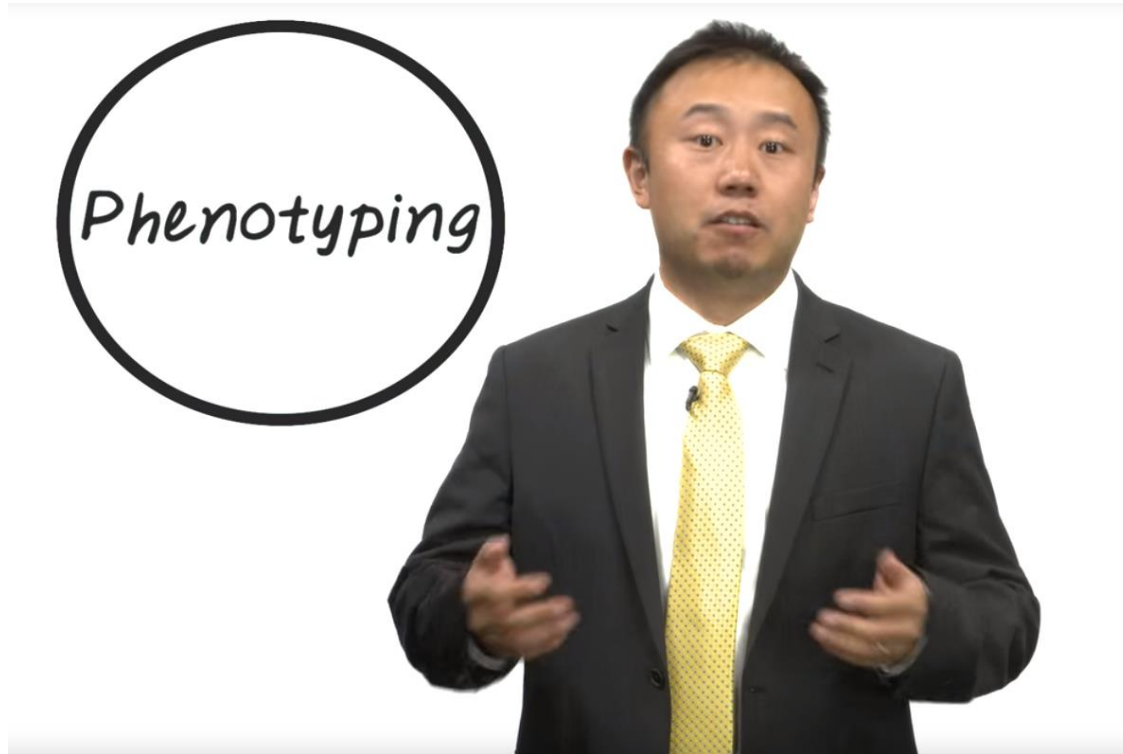


Big Data Analytics in Healthcare- Lesson 8: Computational Phenotyping

출처 : <https://www.udacity.com/>

1. Introduction to Phenotyping



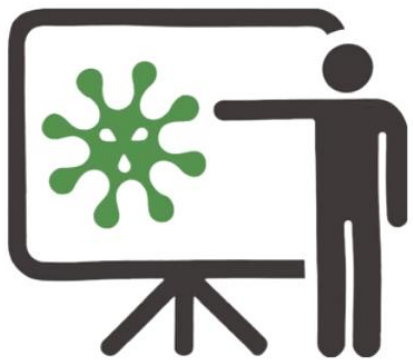
- 클러스터링의 헬스케어 응용에서 사용했던 Phenotyping에 대해서 논의해보자.

1. Introduction to Phenotyping



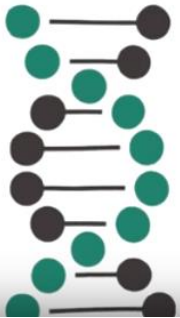
- Phenotyping은 질병 또는 병상과 같은 의학적인 개념.

1. Introduction to Phenotyping



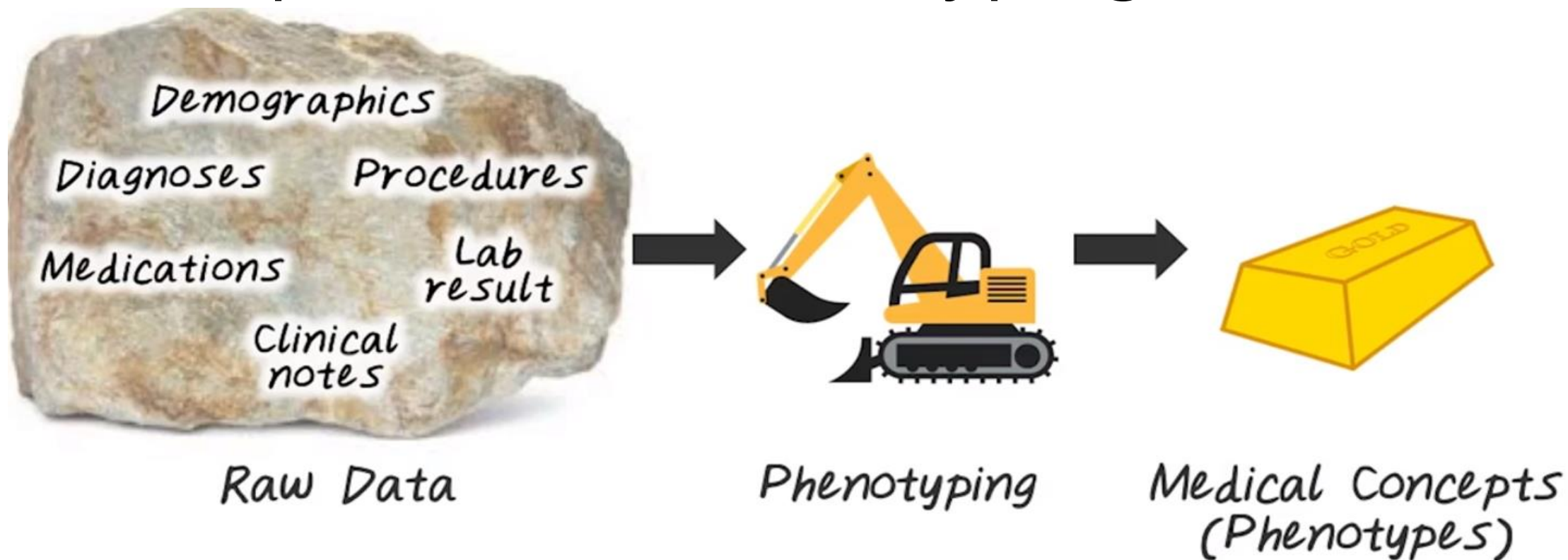
- 대부분의 질병에 대해서 현재의 의료 지식을 기반한 환자의 많은 표현형을 알고 있음.
- 그러나 더 많은 표현형과 세부 표현형이 아직 발현지지 않음.
- Computational Phenotyping은 새로운 표현형을 발견하기 위한 데이터를 활용하기 위한 방법임.

1. Introduction to Phenotyping



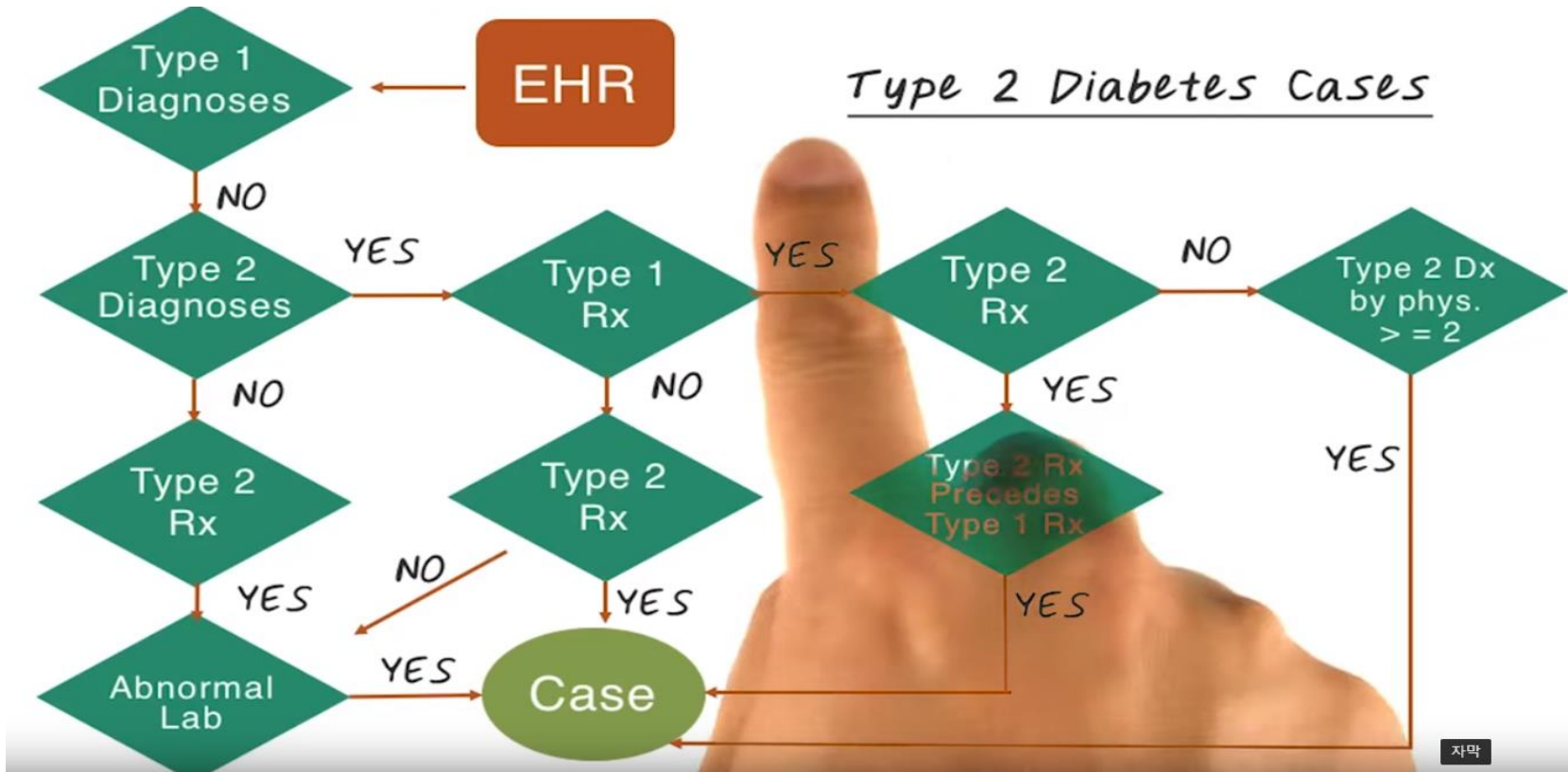
- 표현형은 질병 진단만을 위한 것은 아니고, 의료비 예측, 재입원 위험와 유전체 연구를 지원에도 사용될 수 있음

2. Computational Phenotyping



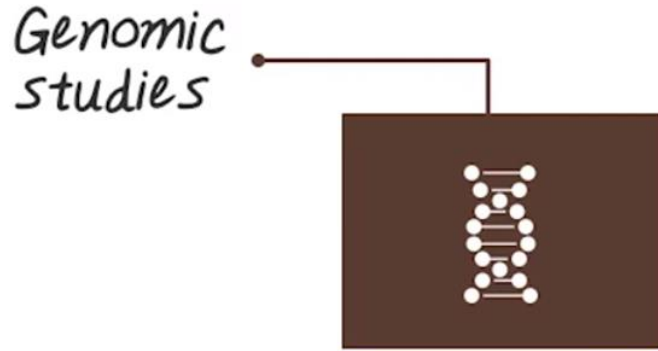
- Computational 표현형에 대해서 처음에 배운것을 상기해보자.
- Phenotyping 알고리즘을 통해서 원천 전자의료레코드에서 변환한 가치있는 의료적인 개념 또는 표현형.
- 예를 들면, 타임2 당뇨병 같이 여러 다른 소스로부터 원천데이터를 구성.
- 원천데이터로부터 표현형을 잘 설명하지 못하는 이유는
 - 1) data noisy – missing data.
 - 2) 데이터들이 주로 비용청구와 같은 병원 운영을 지원하기 위해서 사용 -> 연구목적으로 설계되지 않음.
 - 3) 중복정보 – 진단정보는 diagnosis code로 나타내지고, clinical note에도 같은 정보가 포함됨.
- Phenotyping은 RawData부터 컴퓨팅 알고리즘을 사용해서 연구 등급 표현형을 도출하는 과정이다.

3. Phenotyping Algorithm



- 이 알고리즘의 목적은 전자의료기록을 기반해서 type2 당뇨병인 환자 여부를 판단.
- 이것은 의료전문가에 의해서 수작업으로 개발됨.
- 알고리즘에 대해서 더 자세히 배워보자.

4. Applications of Phenotyping



- Phenotyping은 많은 여러 응용에 사용되어짐.
- 첫번째로 유전정보와 표현형 데이터 사이의 관계를 찾는 유전학 연구에 활용됨.

4. Applications of Phenotyping



- 의료예측모델링 : 질병발병과 다른 관련있는 목표를 정확하고 건고하고 해석가능한 예측 모형

4. Applications of Phenotyping



- 실용적인 임상 시험 : 전자의료데이터 같은 관찰데이터 사용해서 실제 의료 환경에서 치료 효과 비교

4. Applications of Phenotyping



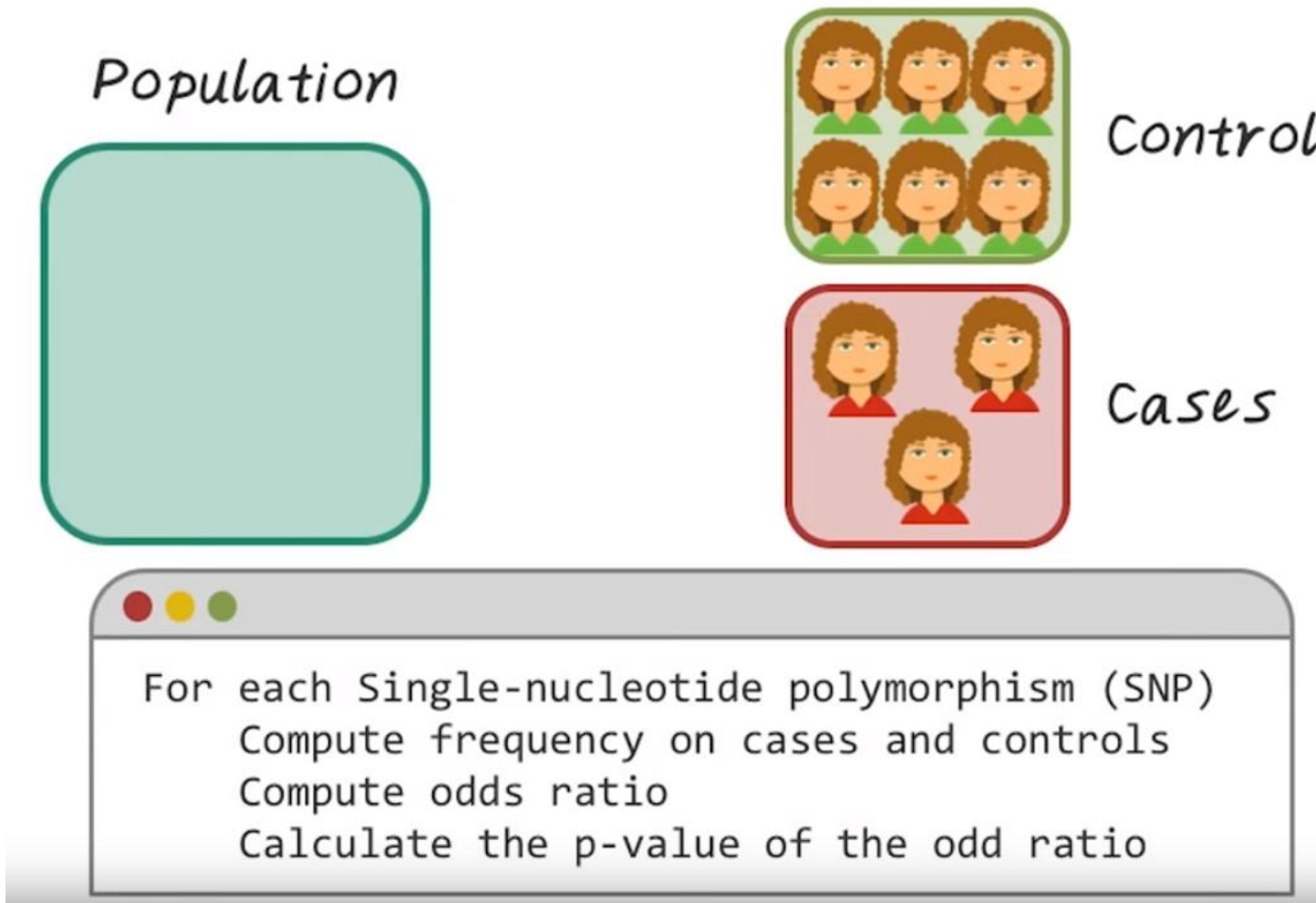
- 건강 관리 품질 측정 : 서로 다른 병원 사이 치료의 효율성과 품질을 측정

5. Genomic Wide Association Study





- Phenotyping algorithm은 GWAS에서 중요함.
- GWAS는 Single nucleotide polymorphism(SNP)과 같은 유전자 마커와 특별한 질병의 표현형과 연관성을 찾는 연구임.
- 새로운 유전적인 연관성은 식별하고 질병은 진단/치료/예방을 위한 더 좋은 전략을 개발하는데 활용.

5. Genomic Wide Association Study



- GWAS의 첫번째는 질병 표현형을 식별
- 질병 표현형이 있는 사람을 case, 질병 표현형이 없는 사람을 control
- 모든 대상자에게서 DNA샘플을 채취하고 이 샘플로부터 유전적인 변이를 조사 => SNP
- 특정 유전적인 변이가 case 그룹에서 중요하게 더 자주 발생함을 찾음.
- 이런 변이가 질병과 연관되었다고 함.
- Case그룹과 Control그룹간의 SNP의 빈도를 계산=> 오즈비
- 이 오즈비의 p값을 계산해서 작은 것을 유의한 변이로 결론 내림.

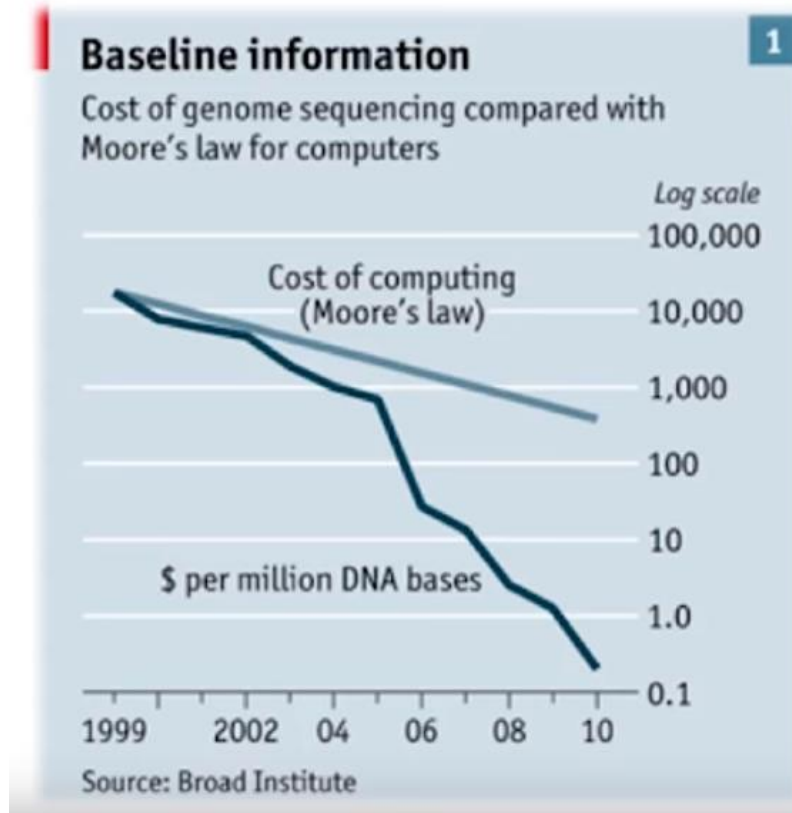
5. Genomic Wide Association Study

	SNP1	SNP2	SNP...
Control	 Count of G: 2676 of 6000 Frequency of G: 44.6%	Count of G: 2532 of 6000 Frequency of G: 42.2%	Repeat for all SNPs
Cases	 Count of G: 2104 of 4000 Frequency of G: 52.6%	Count of G: 1648 of 4000 Frequency of G: 41.2%	
	P-value: $5.0 \cdot 10^{-15}$	P-value: 0.33	

- GWAS에서는 case와 control이 잘구분되는 매우 품질이 좋은 표현형이 필수임.

6. Why Do We Care About Phenotyping

We need rich and deep phenotypic data in order to analyze genomic data.



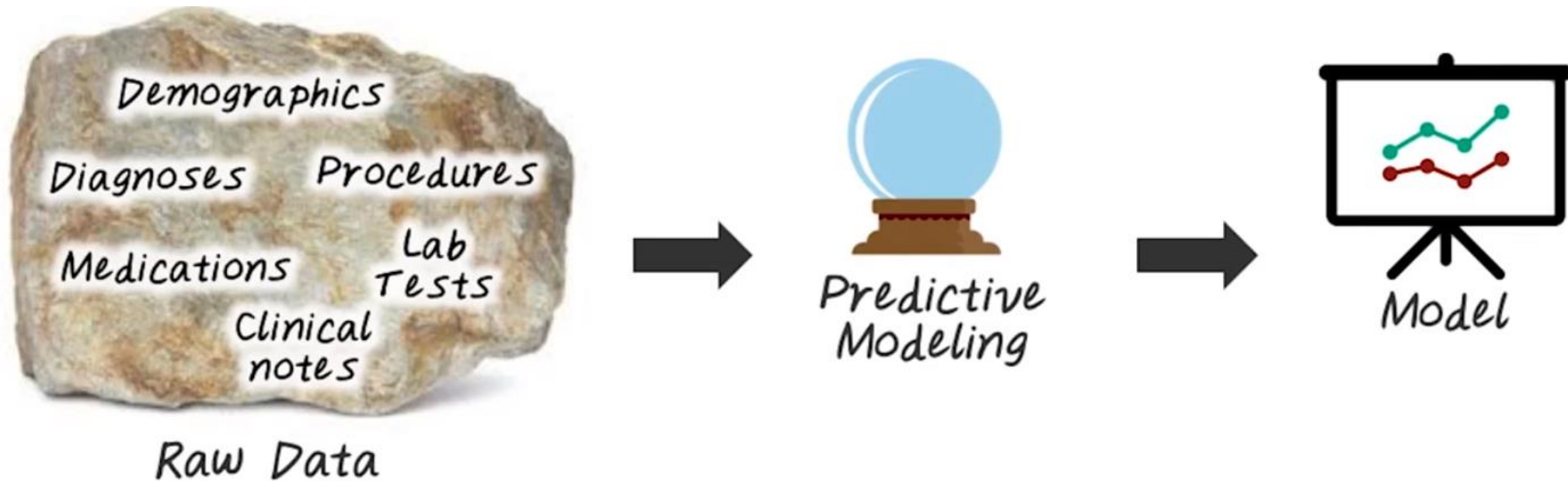
Cost of phenotypic data

Cost of genomic data



- GWAS 비용이 상대적으로 저렴해져서 유전정보를 활용한 phenotyping이 중요해짐.

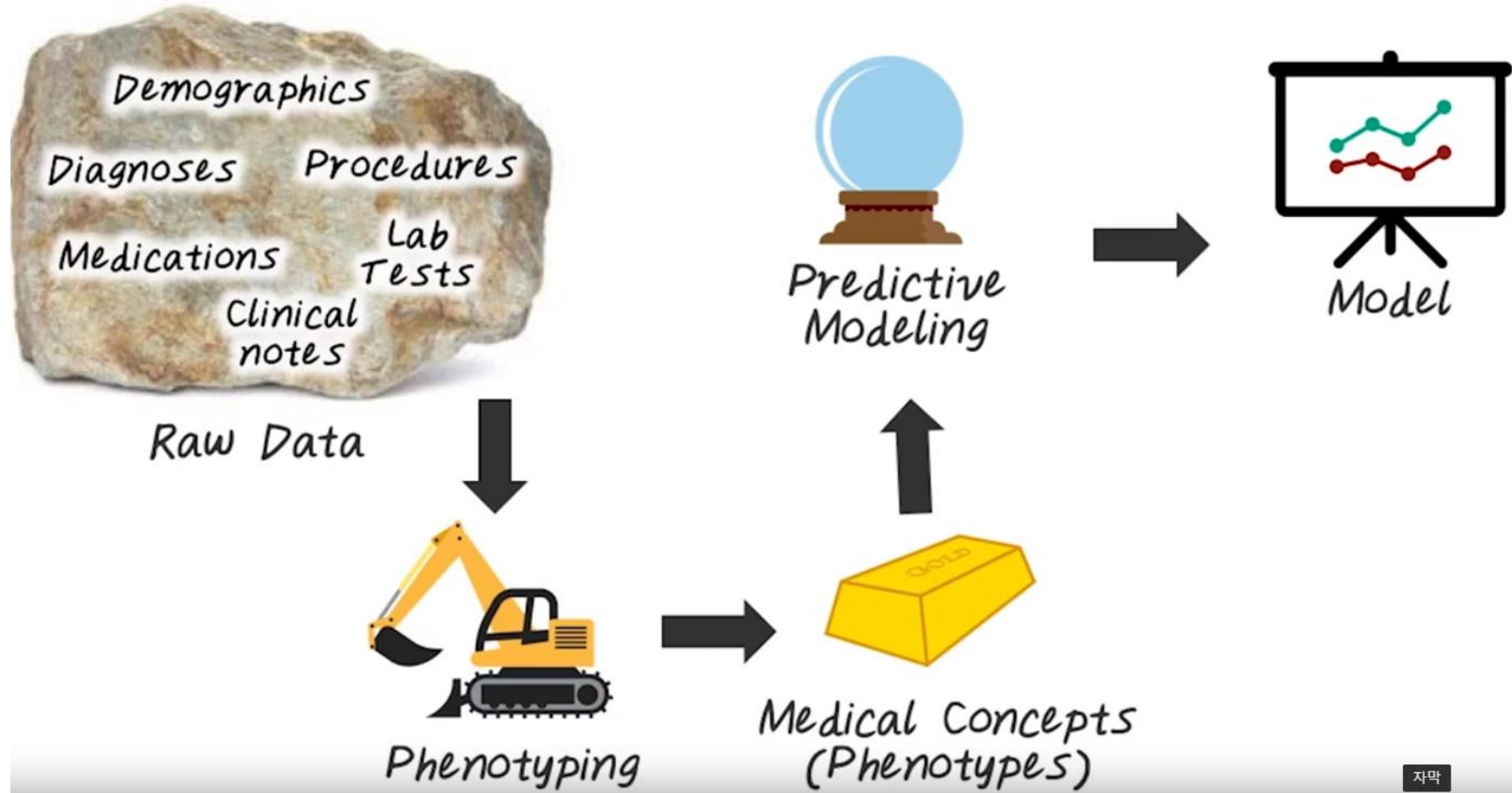
7. Clinical Predictive Modeling



전통적인 의료 예측 모델을 잘 맞지 않음.

- 1) Rawdata 가 좋은 않음. - 중복이 많고 복잡
- 2) 각각의 병원별로 구조가 다름.

7. Clinical Predictive Modeling



자막

- EHR에서부터 phenotyping 과정을 거쳐서 중요한 새로운 표현형을 만들고 예측 모형을 만듦.

8. Pragmatic Clinical Trials



TRADITIONAL

- One condition
- One drug
- Must randomize
- Careful selection
- Carefully controlled



PRAGMATIC

- Multiple conditions
- Potentially multiple drugs
- No randomization
- Any patient
- Real-world environment

- 고성능 phenotyping algorithm이 매우 중요.

9. Healthcare Quality Measurement



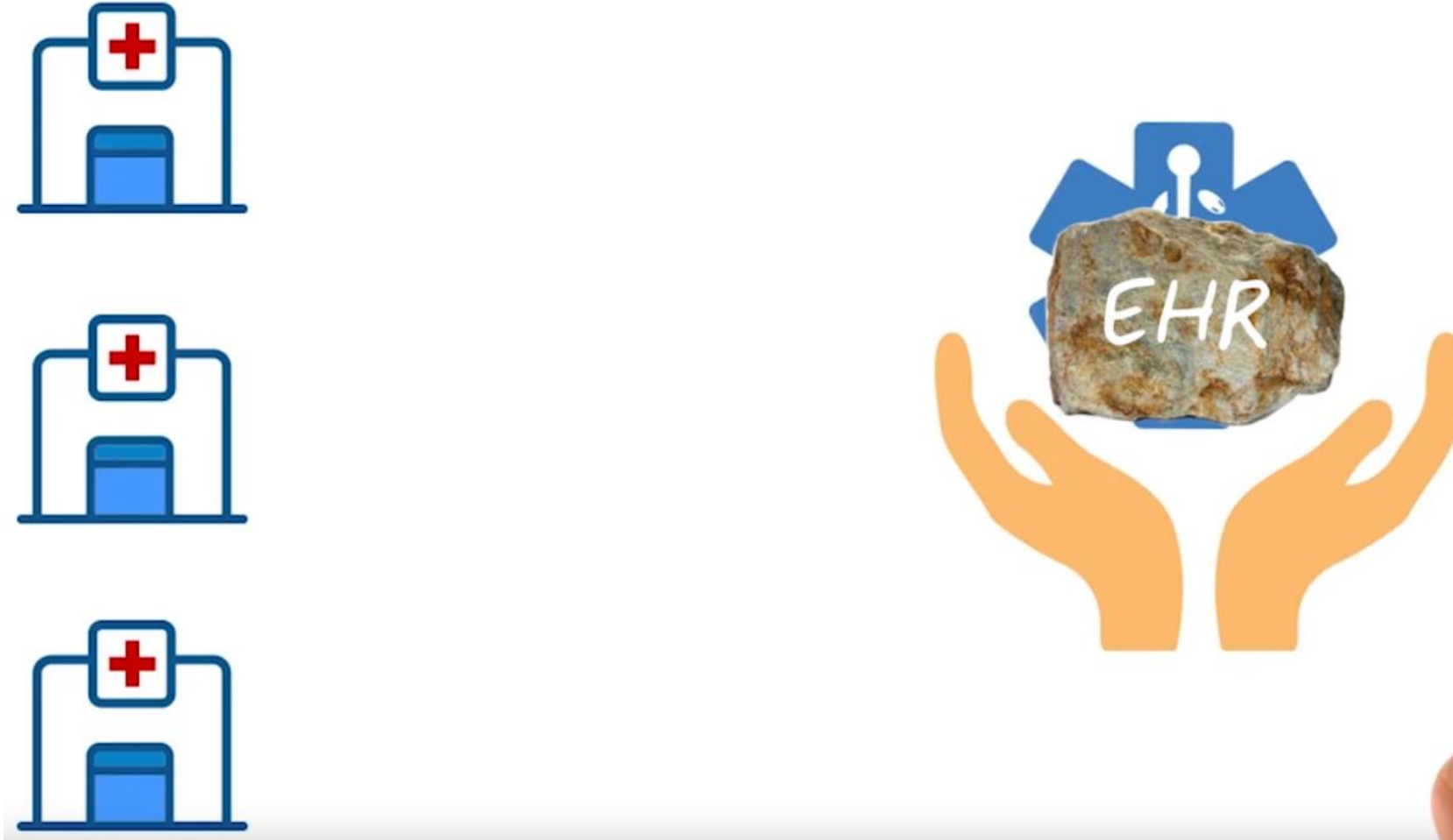
- phenotyping algorithm이 헬스케어 품질 측정에 중요함.
- 모든 병원의 EHR 데이터를 중앙으로 전송하고, 중앙은 보험회사 또는 질병관리본부임.

9. Healthcare Quality Measurement



- 모든 병원의 EHR 데이터를 중앙으로 전송하고, 중앙은 보험회사 또는 질병관리본부임.

9. Healthcare Quality Measurement



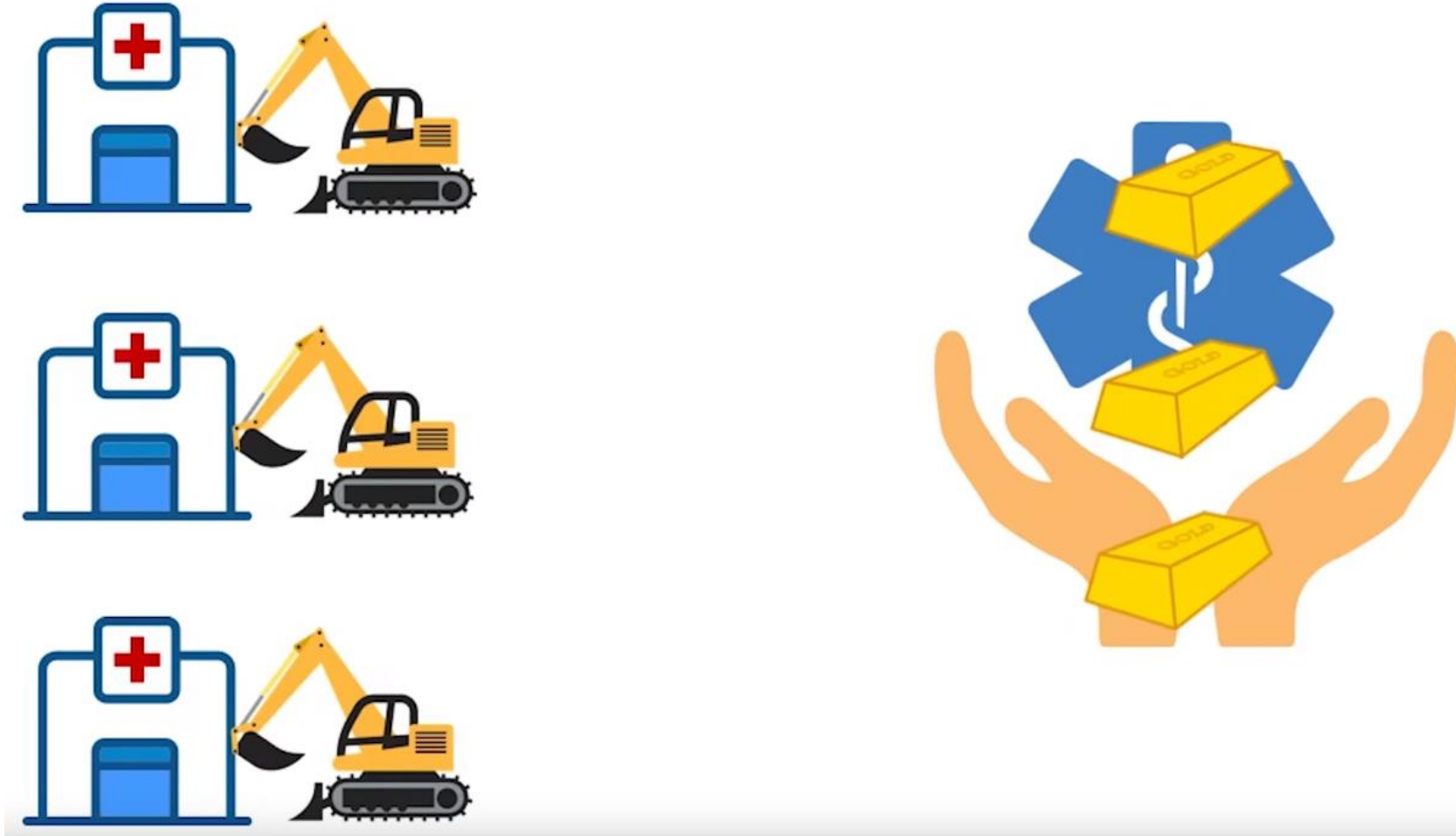
- 중앙에서는 모든 진료품질측정을 위해서 모든 원천 데이터를 취합
- 이 과정은 각각의 병원마다 다른 포맷을 사용해서 매우 어려운 과정임.

9. Healthcare Quality Measurement



- 이문제를 다룰 수 있는 확장성 있는 방법으로 phenotyping통해서 원천 EHR데이터를 처리
- 여기에서 얻은 고품질 phenotypic 정보를 중앙에서 공유

9. Healthcare Quality Measurement



- 여기에서 얻은 고품질 phenotypic 정보를 중앙에서 공유
- 이 정보를 가지고 병원들간의 의료품질 비교가 가능해짐.

10. Phenotyping Methods Part

SUPERVISED LEARNING

UNSUPERVISED LEARNING

Three minicourses

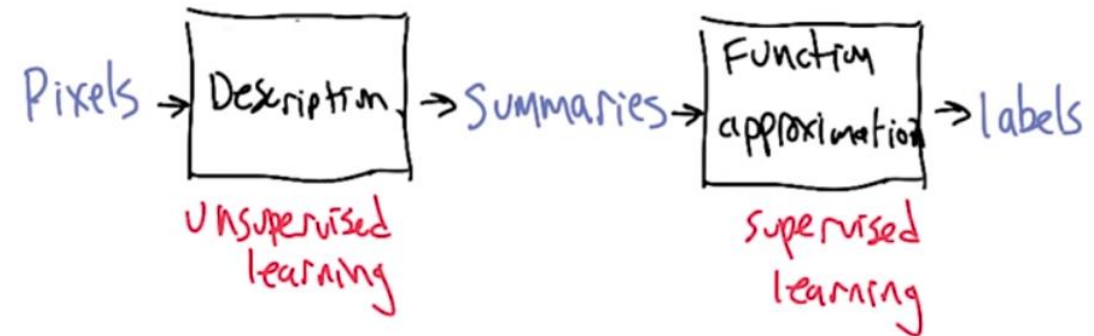
- supervised learning
- unsupervised learning
- reinforcement learning

Supervised learning \equiv Approximation

Unsupervised learning \equiv Description

input	1	2	3	4	5	6	7	10
output	1	4	9	16	25	36	49	100

$$\text{output} \leftarrow \text{input}^2$$



10. Phenotyping Methods Part

SUPERVISED LEARNING

- *Expert-defined rules*
- *Classification*

UNSUPERVISED LEARNING

- *Dimensionality Reduction*
- *Tensor factorization*