

# Statistical Approaches to Complex Data

Kipoong Kim

Department of Statistics, Changwon National University

January 16, 2025

# 국가건강검진

## <그림 국가건강검진대상자>

짝수년도	지역가입자	세대주와 20세 이상 세대원 중 짝수년도 출생자	
	피부양자	20세 이상 짝수년도 출생자	
	직장가입자	비사무직 전체, 격년제 실시에 따른 사무직대상자(2018년도부터 사무직 격년제는 출생년(짝,홀) 기준 적용)	사업장으로 검진대상자 명부가 송부됩니다.
	의료급여 수급자	20세~64세 짝수년도 출생자	
홀수년도	지역가입자	세대주와 20세 이상 세대원 중 홀수년도 출생자	
	피부양자	20세 이상 홀수년도 출생자	
	직장가입자	비사무직 전체, 격년제 실시에 따른 사무직대상자(2018년도부터 사무직 격년제는 출생년(짝,홀) 기준 적용)	사업장으로 검진대상자 명부가 송부됩니다.
	의료급여 수급자	20세~64세 홀수년도 출생자	

# 검사항목

## <그림 일반건강검진 공통 검사 항목>

공통 검사 항목	대상질환	검사 항목
	비만	신장, 체중, 허리둘레, 체질량지수
	시각, 청각이상	시력, 청력
	고혈압	혈압
	신장질환	요단백, 혈청크레아티닌, e-GFR
	빈혈증	혈색소
	당뇨병	공복혈당
	간장질환	AST, ALT, r-GTP
	폐결핵 / 흉부질환	흉부방사선촬영
	구강질환	구강검진

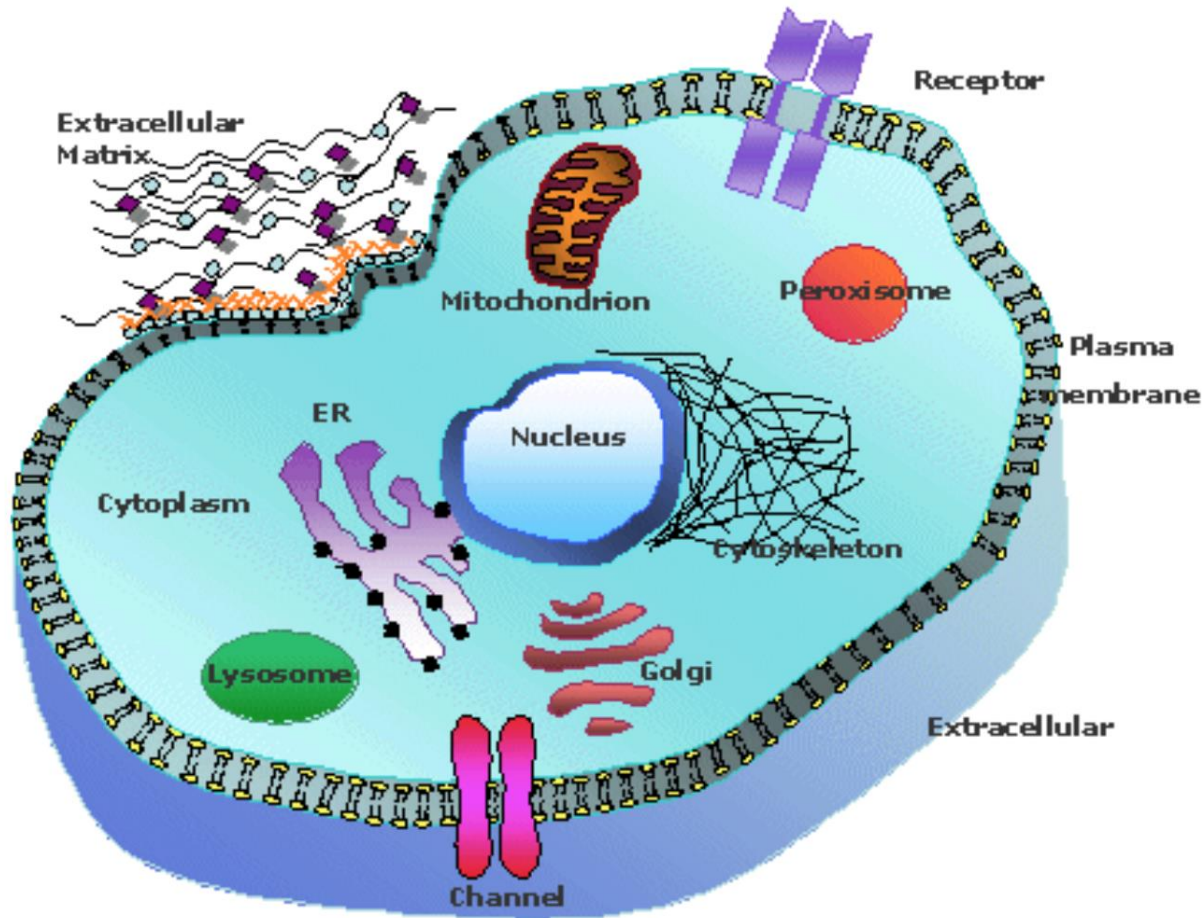
※ 시력, 청력은 운전면허 신체검사 수치로 활용되므로 안경 착용자는 교정시력으로 검사를 받으시기 바랍니다.

# 추가검사 데이터: 유전자 검사



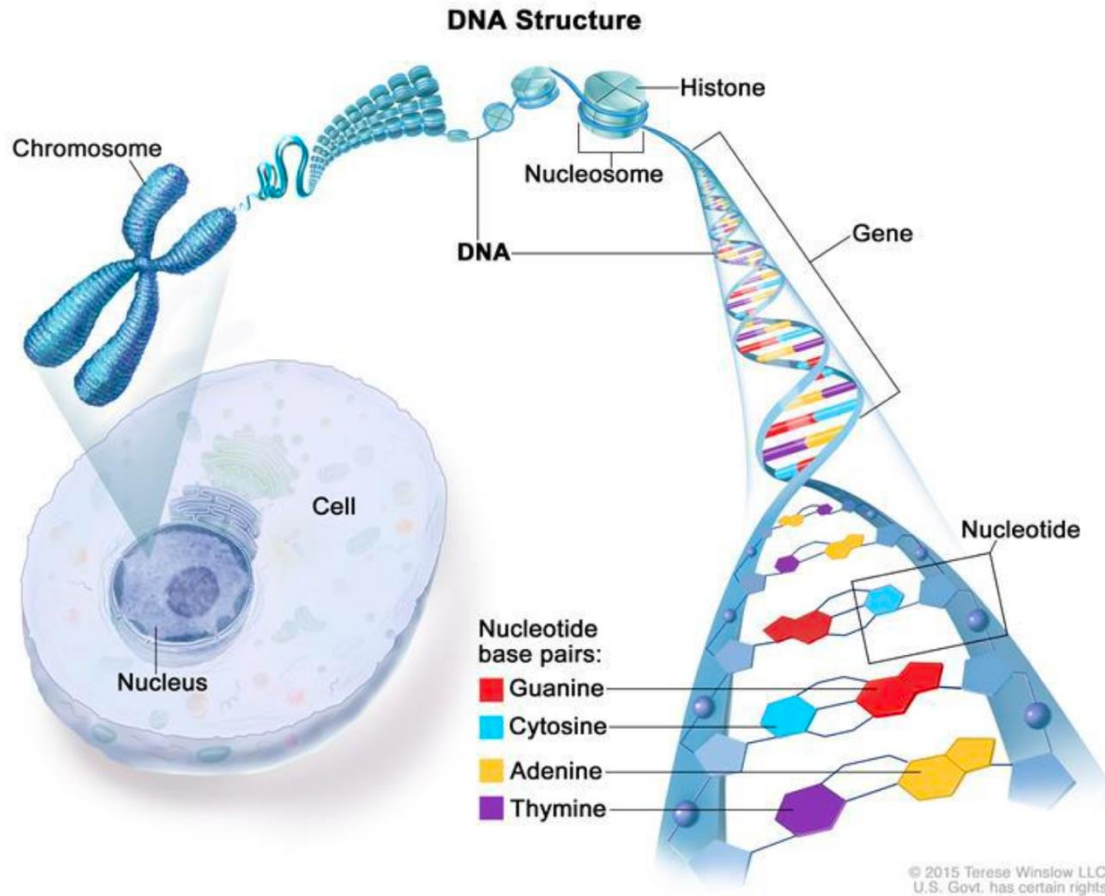
# 추가검사 데이터: 유전자 검사

## □ 세포 구조



# 추가검사 데이터: 유전자 검사

## □ DNA 구조





# 추가검사 데이터: 유전자 검사

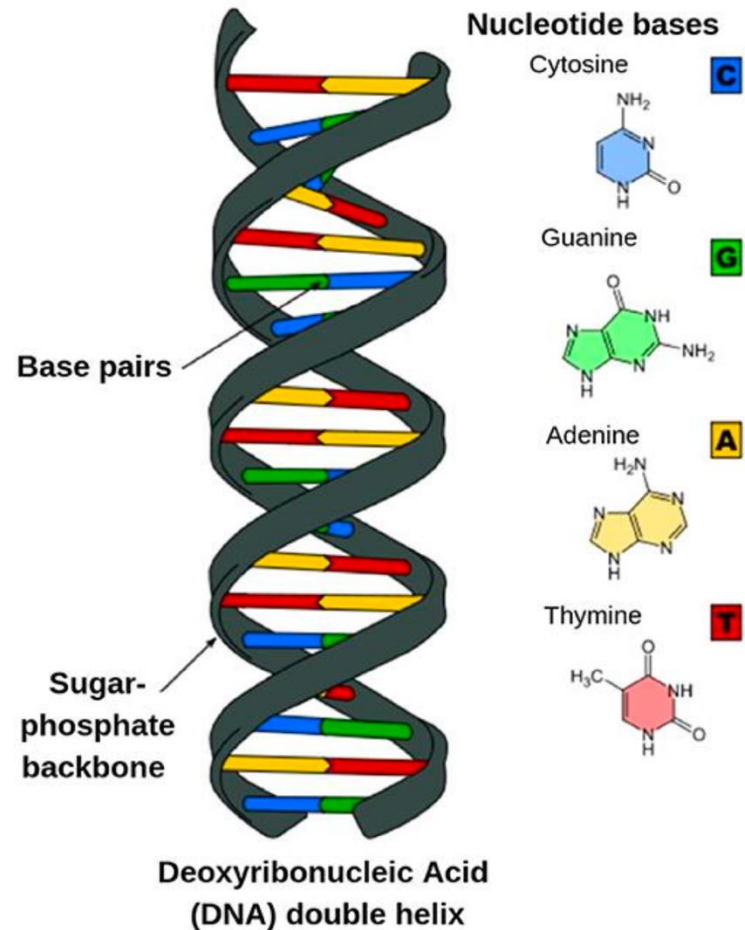
## □ DNA = 염기구조

### ➤ 첫번째 지역 (base)에서의 염기

- 사람 A: CC >> '0' 코딩
- 사람 B: GC >> '1' 코딩
- 사람 C: GG >> '2' 코딩
- .....

## □ 간단하게 말하면,

DNA 데이터 = {0, 1, 2}로 구성



# 추가검사 데이터: 유전자 검사

## □ 인간 유전체

- 인간 유전자: ~40,000개  
DNA: ~5,000,000개
- Whole Genome Sequencing (WGS) 비용: 130만원 (2020. 9 기준)

## □ 유전체 데이터 특징

- 고차원 저표본 (Low sample high-dimensional)
- DNA 간 강한 상관관계 (High correlation)
- Genetic network
- Population structure (white vs asian)



# 추가검사 데이터: 유전자 검사

□ 무엇을 할 수 있을까? 질병 예측



## 암 유전자

- EGFR 돌연변이 분석
- BRAF 돌연변이 분석
- ERBB2 (HER2) 발현 분석
- PIK3CA 돌연변이 분석
- KRAS/NRAS 돌연변이 분석
- BRCA1/2 돌연변이 분석
- FGFR2 돌연변이 분석
- MLH1 돌연변이 분석 외 암 관련 유전자 바이오마커 분석

# 추가검사 데이터: 유전자 검사

□ 무엇을 할 수 있을까? 용의자 혈흔으로부터 DNA phenotyping

## Florida police used a smidgen of DNA to try to fully reconstruct an alleged criminal's face

### Snapshot Prediction Results Phenotype Report



Case #150124-024343



**Contact:** Miami-Dade Crime Stoppers  
(305) 471-TIPS (8477)

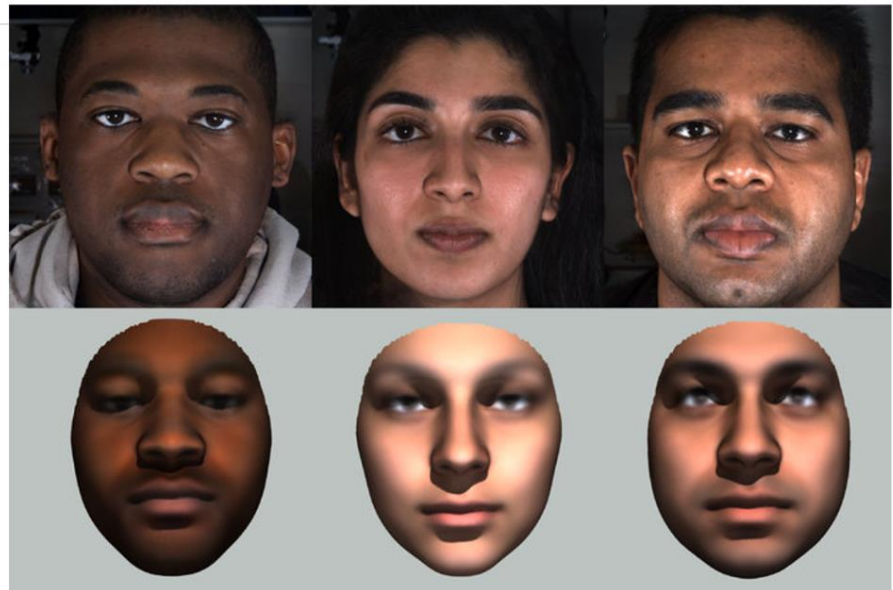
**Sex:** Male ♂

**Age:** Unknown  
(Shown at age 25)

**Ancestry:** Latino



Region	Percent
Europe – Southwest	19.91%
Americas – South	16.59%
Americas – Central	12.50%
Africa – West	11.66%
Middle East – NW Africa	8.99%
Europe – Central West	8.97%
Europe – Caucasus	6.91%



# 추가검사 데이터: CT / MRI

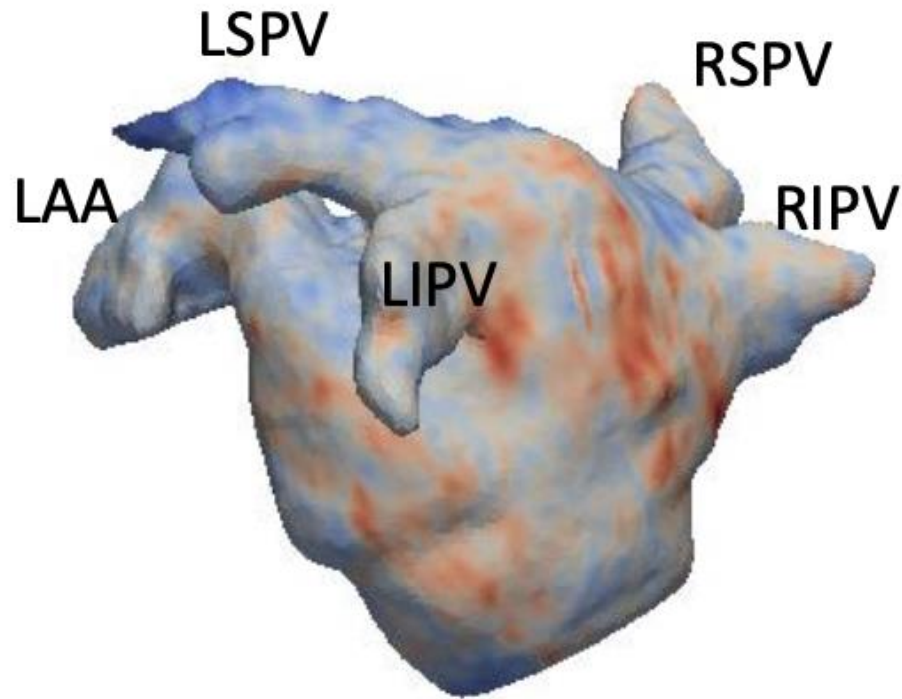
□ CT: x선 흡수율 기반

□ MRI: 자기장으로 인한 수소 원자핵 정렬신호 기반



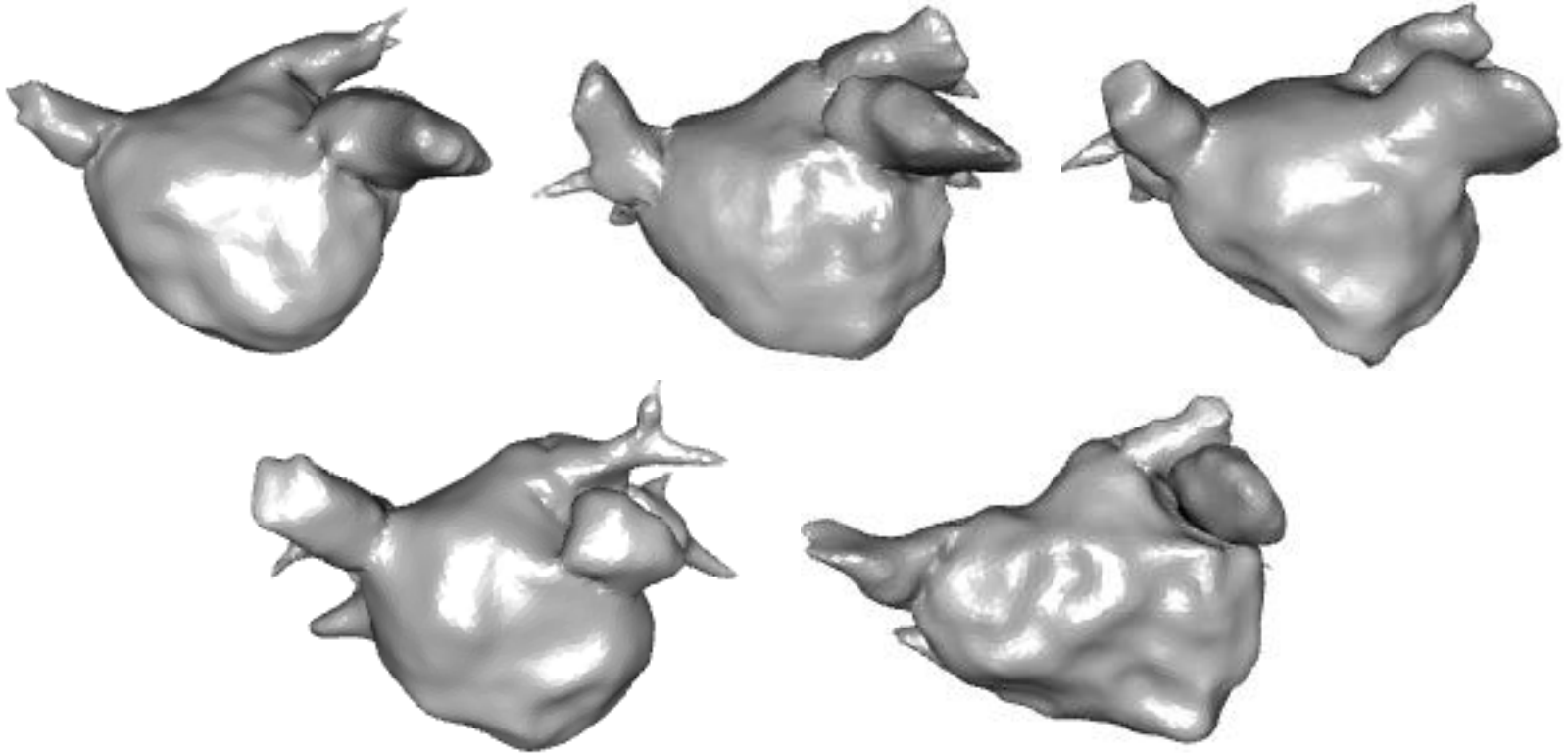
# 추가검사 데이터: CT / MRI

- 3차원 포인트 클라우드 데이터 (심장) from CT / MRI
  - 추가적으로, 각 지점에서의 심장 벽 두께, 전압 등을 측정



# 추가검사 데이터: CT / MRI

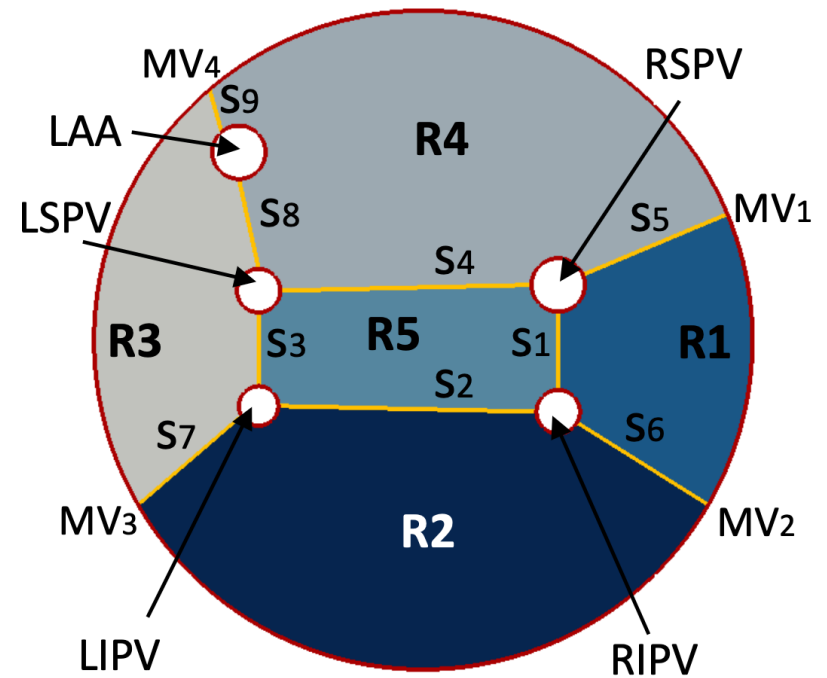
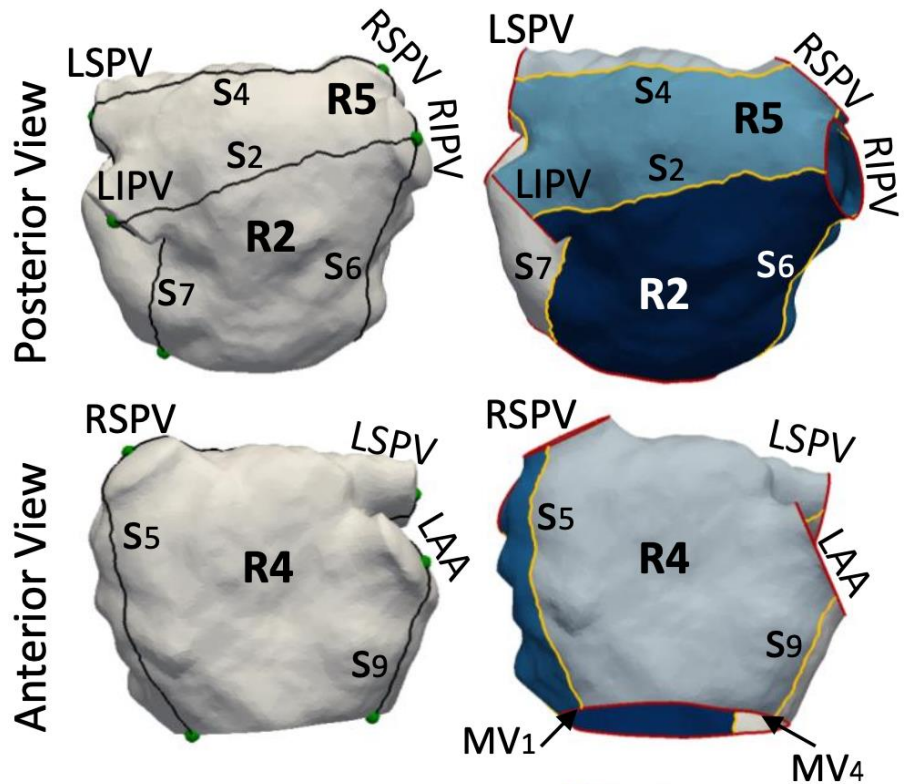
□ 문제점: 각 사람마다의 심장 모양이 다르다.





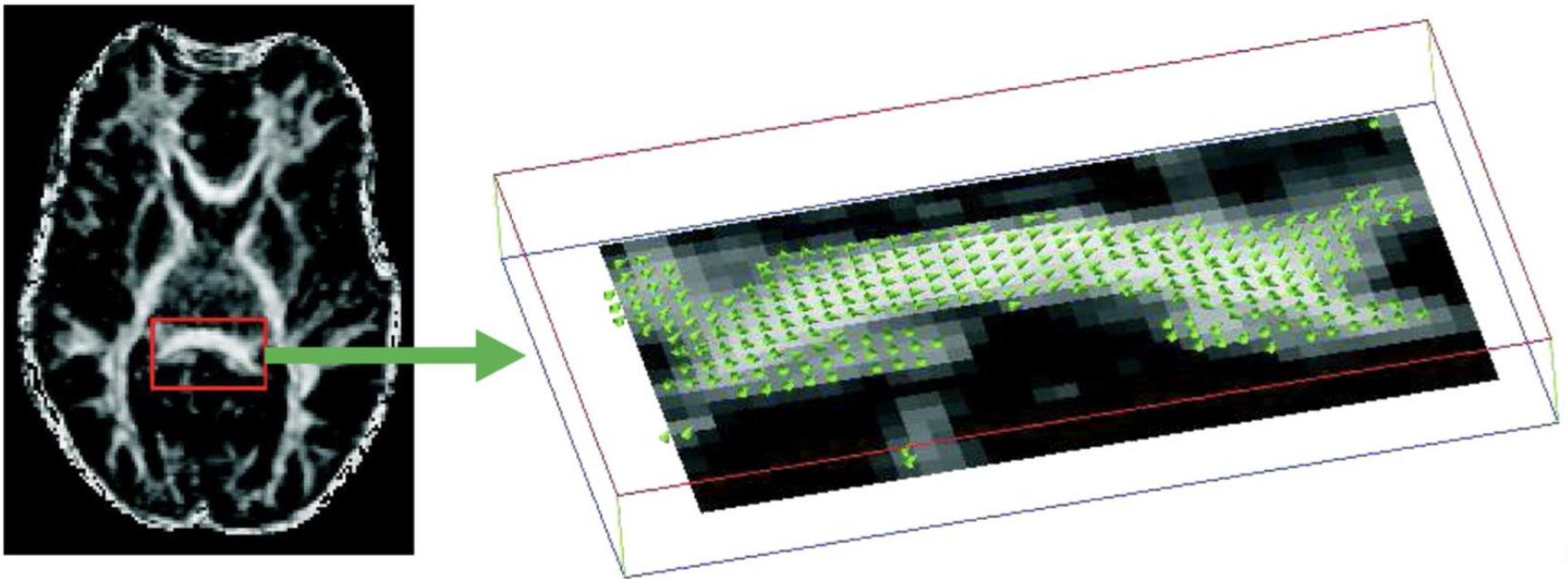
# 추가검사 데이터: CT / MRI

□ Flattening을 이용한 registration



# 추가검사 데이터: CT / MRI

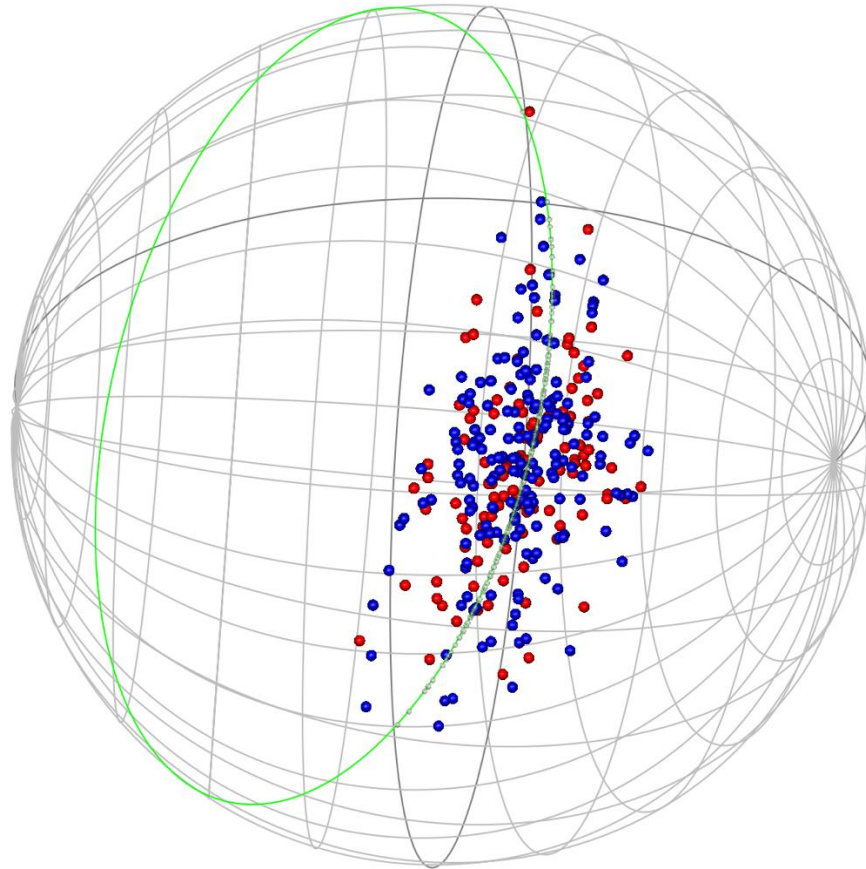
- Diffusion Tensor Imaging (DTI) data from MRI





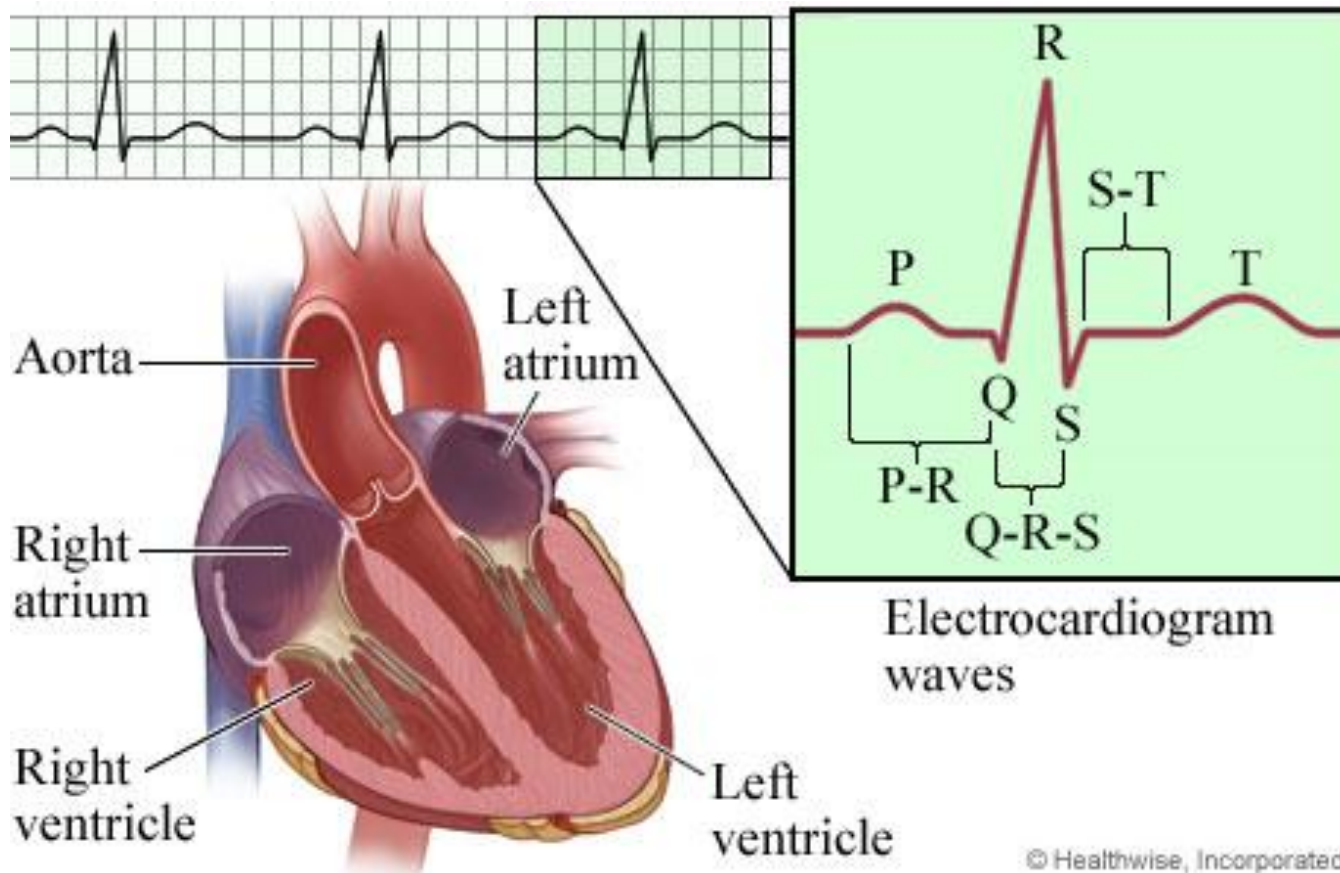
# 추가검사 데이터: CT / MRI

□ Spherical data



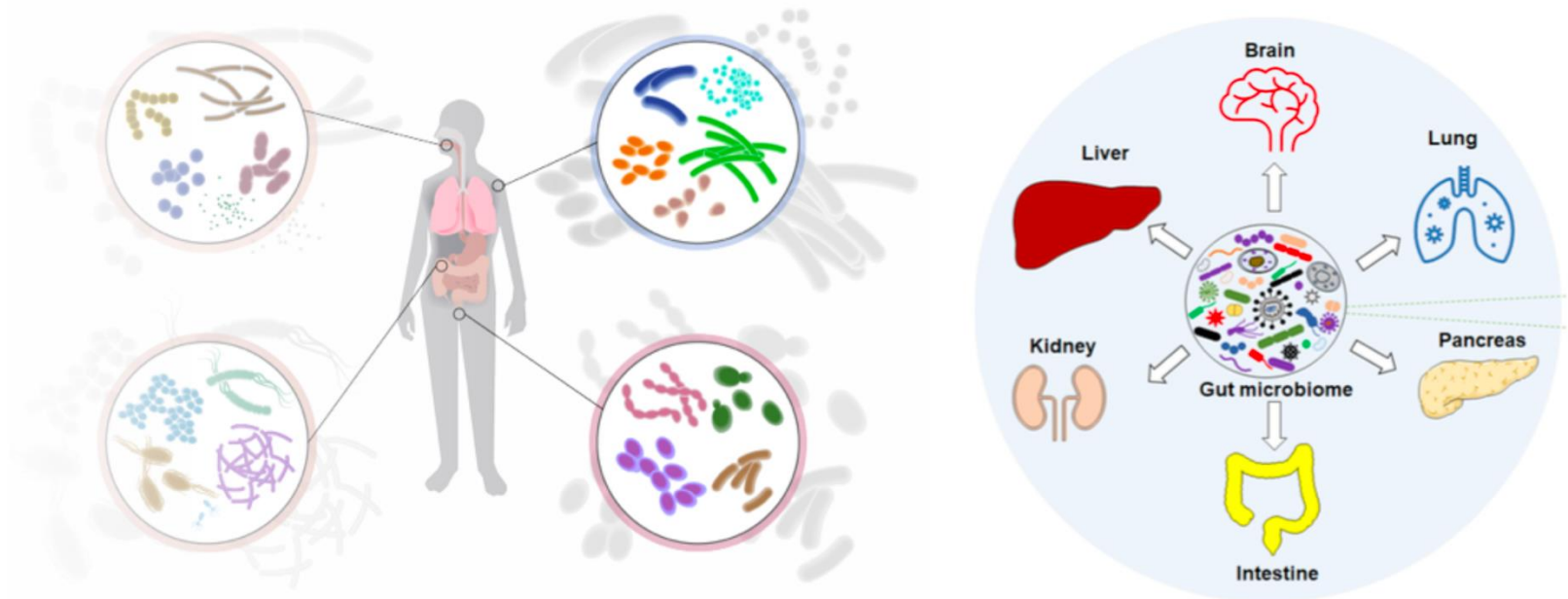
# 추가검사 데이터: Electrocardiogram (ECG)

## ❑ Functional data



# 추가검사 데이터: 장내미생물 (Microbiome)

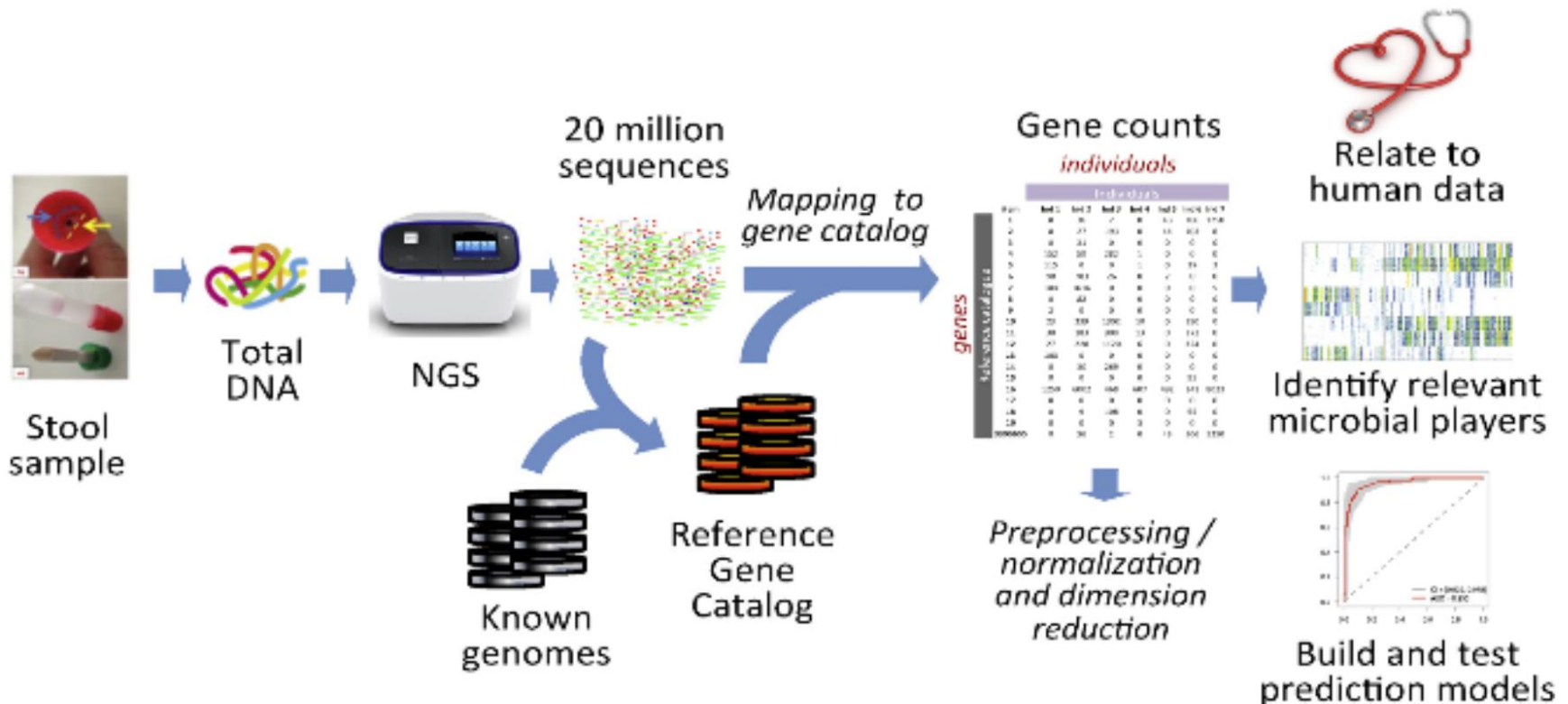
- ❑ 출생 방법(자연분만/제왕절개)과 수유 형태(모유/분유)에 따라 영아의 초기 장내미생물군 형성에 차이가 발생함.
- ❑ 이후 약 2-3년(약 1,000일) 동안 식이패턴과 환경요인에 의해 개인의 고유한 장내미생물군이 안정적으로 형성됨.
- ❑ 형성된 장내미생물이 다양한 인체내 물질대사와 연관있다는 보고가 있음.



# 추가검사 데이터: 장내미생물 (Microbiome)

## □ 데이터 추출 방법

- 샘플링 → DNA 추출 → PCR 증폭 + 라이브러리 준비  
→ 시퀀싱 & 라이브러리 매핑 → Microbiome count data



# 추가검사 데이터: Connectome (뇌 연결체)

- ❑ 뇌의 모든 신경 연결을 총체적으로 표현한 지도
- ❑ Levels
  - 개별 뉴런간 연결
  - 신경세포간 연결
  - 뇌 영역간 연결





# 추가검사 데이터: 의사진단서

## ❑ Doctor – Patient Question data

➤ 레딧에 현재 증상 말해주면 의사분들이 답변해주는 게시판 데이터

➤ 예: The risk of dying following swallowing and ingesting a toothpick.

이쑤시개를 삼켰을 때의 위험도는?

Table. Example Questions with Physician and Chatbot Response<sup>a</sup>

Question <sup>b</sup>	Physician response	Chatbot response <sup>c</sup>	Evaluation scores
Question about risk of dying following swallowing and ingesting a toothpick.	<p>If you've surpassed 2-6 h, chances are they've passed into your intestines. Which means it can't be retrieved easily.</p> <p>Many people can swallow razorblades and toothpicks without issue. In case you develop stomach ache, then don't hesitate seeking out an emergency room, and remember to point out the swallowed tooth pick.</p> <p>For now, they won't be doing much.</p>	<p>It's natural to be concerned if you have ingested a foreign object, but in this case, it is highly unlikely that the toothpick you swallowed will cause you any serious harm.</p> <p>Toothpicks are made of wood, which is not toxic and will not cause poisoning. However, it is possible for a toothpick to cause injury if it becomes stuck in the throat or if it punctures the digestive tract. However, this is unlikely to happen with a dull, cooked toothpick that is only 2 cm long.</p> <p>If you are experiencing any discomfort or symptoms such as abdominal pain, difficulty swallowing, or vomiting, it is important to contact a health care provider for further evaluation. However, if you are not experiencing any symptoms, it is safe to assume that the toothpick has passed through your digestive system and you do not need to be concerned.</p> <p>It's always a good idea to be cautious when consuming food and to remove any foreign objects before eating. It's understandable that you may be feeling paranoid, but try not to worry too much. It is highly unlikely that the toothpick will cause you any serious harm.</p>	<ul style="list-style-type: none"><li>• 100% preferred the chatbot</li><li>• 4.67 mean quality score (chatbot)</li><li>• 3.33 mean quality score (physician)</li><li>• 4.33 mean empathy score (chatbot)</li><li>• 2.00 mean empathy score (physician)</li></ul>

# 다양한 특성을 갖는 데이터들

Clinical Outcomes

	A	B	C	D	E	F	G	H	I
1	DATE	DAY	BRAIN	LUNGS	HEART	SYSTOLIC	DIASTOLIC	CELSIUS	PULSE
1	11/1/2020	Sunday	5	5	5	123	82	36.6	172
2	11/2/2020	Monday	5	5	5	119	78	36.6	179
3	11/3/2020	Tuesday	5	5	5	111	80	36.6	84
4	11/4/2020	Wednesday	5	5	5	120	80	36.6	162
5	11/5/2020	Thursday	5	4	5	120	80	36.6	52
6	11/6/2020	Friday	5	5	5	125	81	36.6	80
7	11/7/2020	Saturday	2	4	5	90	56	37.2	95
8	11/8/2020	Sunday	2	2	3	101	68	37.4	171
9	11/9/2020	Monday	5	4	4	147	95	37.6	76
10	11/10/2020	Tuesday	5	3	4	199	133	37.7	151
11	11/11/2020	Wednesday	4	2	3	97	70	37.8	154
12	11/12/2020	Thursday	4	3	4	193	125	38.3	140
13	11/13/2020	Friday	2	1	2	114	74	38.4	134
14	11/14/2020	Saturday	2	1	3	207	151	38.5	102

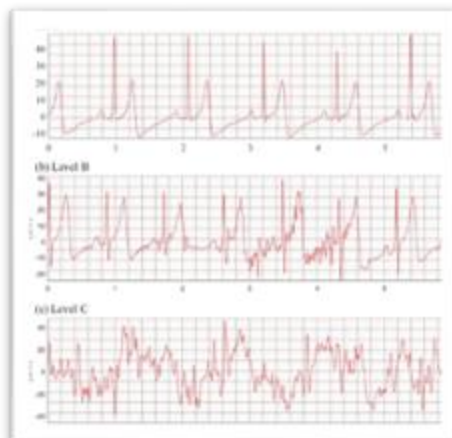
Electrical Health Records



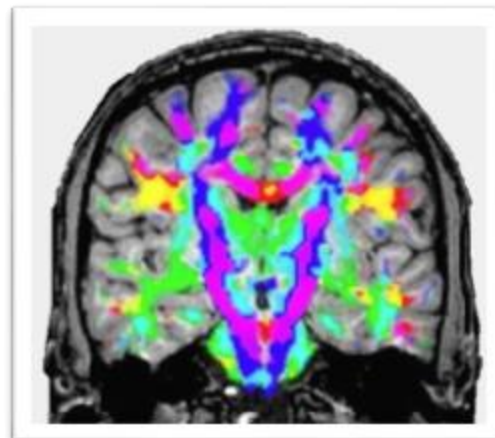
Sequencing data



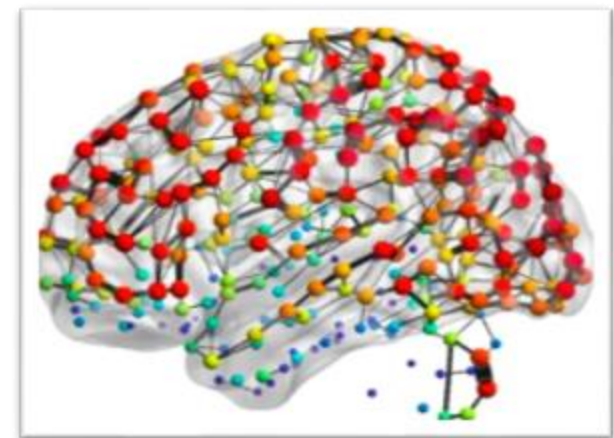
Electrocardiogram (ECG)



Medical Imaging data



Connectome





# 비유클리드 데이터 (Non-Euclidean data)

- 비유클리드 데이터: 내재된 구조가 유클리드가 아닌 데이터
- $k$ 차원 유클리드 공간 ( $R^k$ )의 특성:
  - 벡터 덧셈(+)과 스칼라 곱(·)에 대해 닫혀있는 벡터 공간
  - 유한 차원 공간
- 비유클리드 데이터의 특성:
  - 벡터 덧셈(+)과 스칼라 곱(·)에 대해 닫혀있지 않은 유한 차원 벡터 공간
  - 유한 차원 비벡터 공간
  - 무한 차원 공간

# Non-Euclidean data: Image data (pixel)

□ A: 고양이 사진

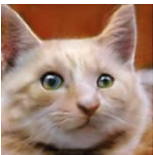
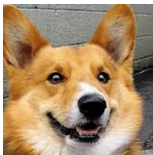
$$\triangleright \begin{bmatrix} 100 & 150 \\ 120 & 160 \end{bmatrix}$$

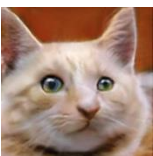


□ B: 강아지 사진

$$\triangleright \begin{bmatrix} 80 & 120 \\ 90 & 140 \end{bmatrix}$$



□  +  =  $\begin{bmatrix} 180 & 270 \\ 210 & 300 \end{bmatrix}$  ?

□   $\times 2 = \begin{bmatrix} 200 & 300 \\ 240 & 320 \end{bmatrix}$  ?

# Non-Euclidean data: Image data (pixel)

□ 유클리드 거리가 실제 유사도를 대표하지 않음.



# Non-Euclidean data: Text data

## ❑ 거리 개념이 잘 정의되지 않음: (1)

- A: ChatGPT는 텍스트를 생성한다.
- B: ChatGPT는 대화를 한다.
- C: AI는 텍스트를 만든다.

## ❑ 거리 개념이 잘 정의되지 않음: (2)

- A: ChatGPT는 대화를 잘한다.
- B: 인공지능은 대화를 한다.
- C: 로봇은 움직인다.

- $d(A, C) \gg d(A, B) + d(B, C)$

# Non-Euclidean data: Network data

## ❑ Research topics of interest

- [가설검정]: 두 네트워크의 구조는 동일하다.
- [회귀분석]: 시간에 따른 네트워크 구조의 변화

## ❑ Key questions:

- [가설검정]: 네트워크 구조 간 차이를 어떻게 정의할까?
- [회귀분석]: 평균 네트워크가 어떻게 정의되는가?

# Non-Euclidean data: Compositional data

## □ Examples

- 토양 구성 비율
  - 모래 (45.3%), 점토 (30.2%), 유기물 (15.8%), 미네랄 (8.7%)
- 하루 시간 사용 분석 (전체 24시간)
  - 수면 (8시간), 업무 (9.5시간), 여가 (3.5시간), 기타 (3.0시간)
- 주식 포트폴리오 구성
  - 엔비디아 (45.9%), 테슬라 (7.2%), TQQQ (34.7%), 비트코인 (12.2%)
- 문서 내 단어 빈도 (배민리뷰 감성분석)
  - 긍정표현 (65%), 부정표현 (10%), 중립표현 (25%)
- 장내 미생물 비율
  - Firmicutes (59.5%), Proteobacteria (35.5%), Bacteroidetes (5%)

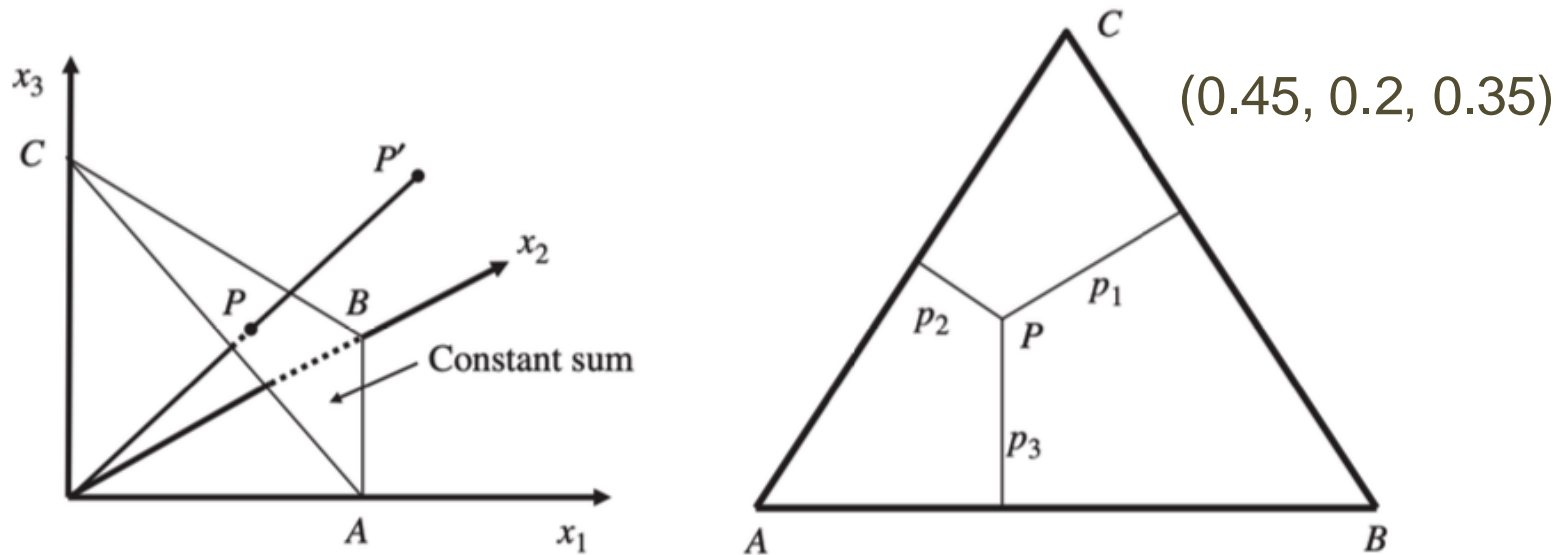
# Non-Euclidean Data Analysis Example (1)

□ Compositional data = 합계가 1인 데이터들

$$C^{p-1} = \{x = [x_1, \dots, x_p] \in \mathbb{R}^p : x_1 + \dots + x_p = 1, x_j > 0 \forall j\}$$

: Compositional space

□ 예를 들어,  $p = 3$ 일 때  $C^2$ 는 어떤 공간일까?



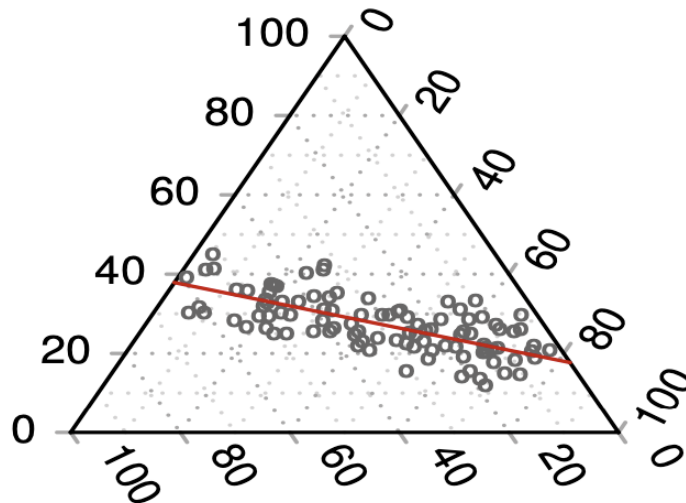
Convex hull embedded in  $\mathbb{R}^3$



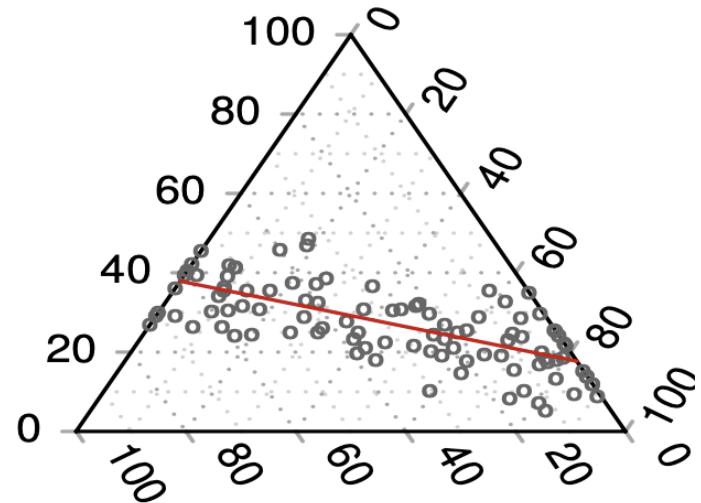
# Non-Euclidean Data Analysis Example (1)

□  $p = 3$ 일 때, 데이터의 분포는?

High signal-to-noise ratio



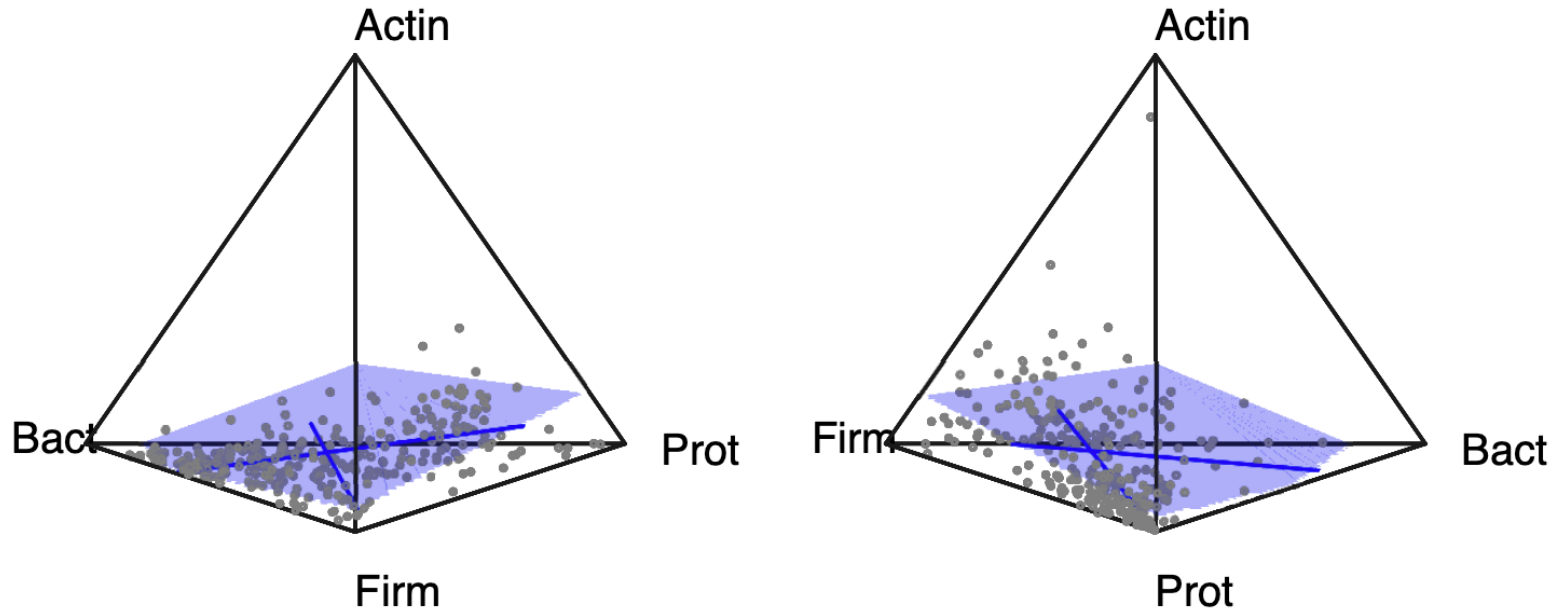
Low signal-to-noise ratio



# Non-Euclidean Data Analysis Example (1)

□ 예를 들어,  $p = 4$  일 때는 ?

➤ Tetrahedron



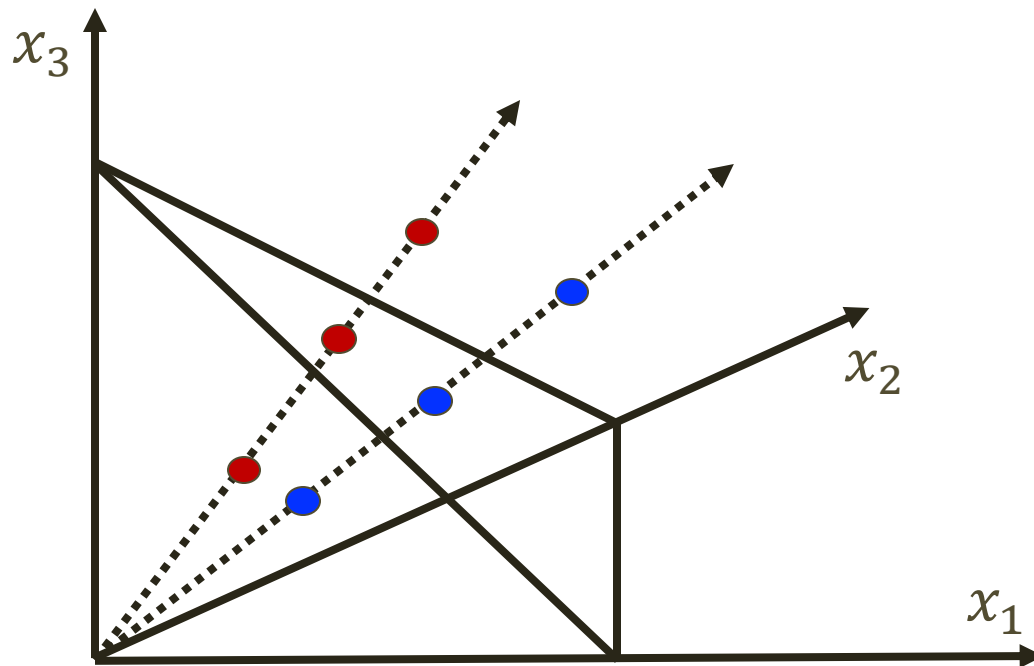
\*Prot: Proteobacteria; Firm: Firmicutes; Actin: Actinobacteria; Bact: Bacteroidetes

# Compositional data

□ Compositional data로의 변환

$$Cls(x_1, \dots, x_p) = \left[ \frac{x_1}{\sum x_j}, \dots, \frac{x_p}{\sum x_j} \right]$$

: Closure operation (like projection?)



# Compositional data is a vector space?

□ For  $\mathbf{x}, \mathbf{y} \in C^{p-1}$  and  $1 \neq c \in \mathbb{R}$

➤  $\mathbf{x} + \mathbf{y} \notin C^{p-1}$

➤  $c \cdot \mathbf{x} \notin C^{p-1}$

➤  $\mathbf{x} - \mathbf{y} \notin C^{p-1}$

# New vector operations for compositional data

□ Compositional data에 대한 new vector operations (덧셈, 곱셈, 뺄셈 등)

➤ Perturbation

$$\mathbf{x} \oplus \mathbf{y} = \left[ \frac{x_1 y_1}{\sum x_j y_j}, \dots, \frac{x_p y_p}{\sum x_j y_j} \right] = \text{Cls}[x_1 y_1, \dots, x_p y_p]$$

➤ Powering

$$\alpha \odot \mathbf{y} = \left[ \frac{x_1^\alpha}{\sum x_j^\alpha}, \dots, \frac{x_p^\alpha}{\sum x_j^\alpha} \right] = \text{Cls}[x_1^\alpha, \dots, x_p^\alpha]$$

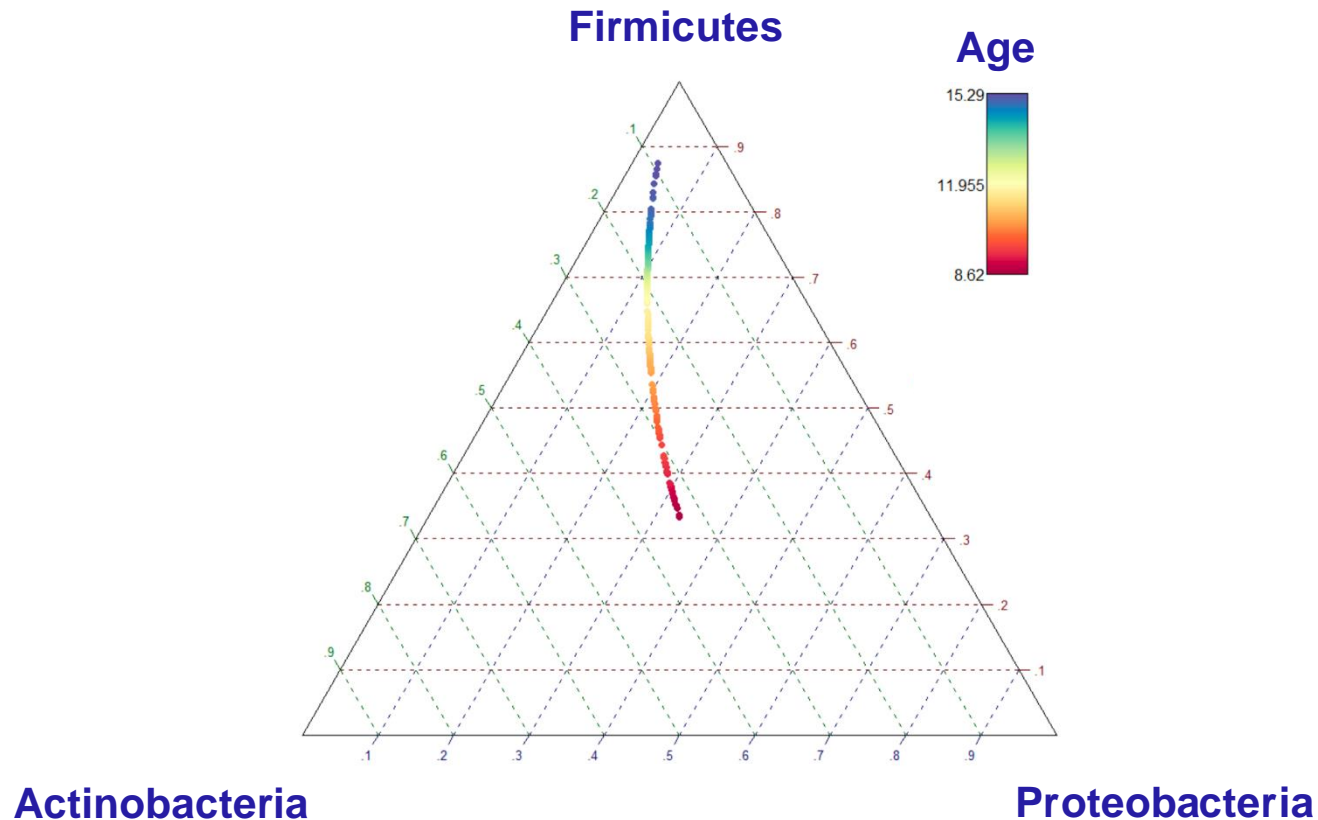
➤ Subtraction

$$\mathbf{x} \ominus \mathbf{y} = \mathbf{x} \oplus (-1 \odot \mathbf{y})$$

# Regression on compositional data

## □ Compositional data에 대한 회귀모형

➤  $E(Y | X) = \beta_0 \oplus (X_1 \odot \beta_1) \oplus \dots \oplus (X_d \odot \beta_d)$

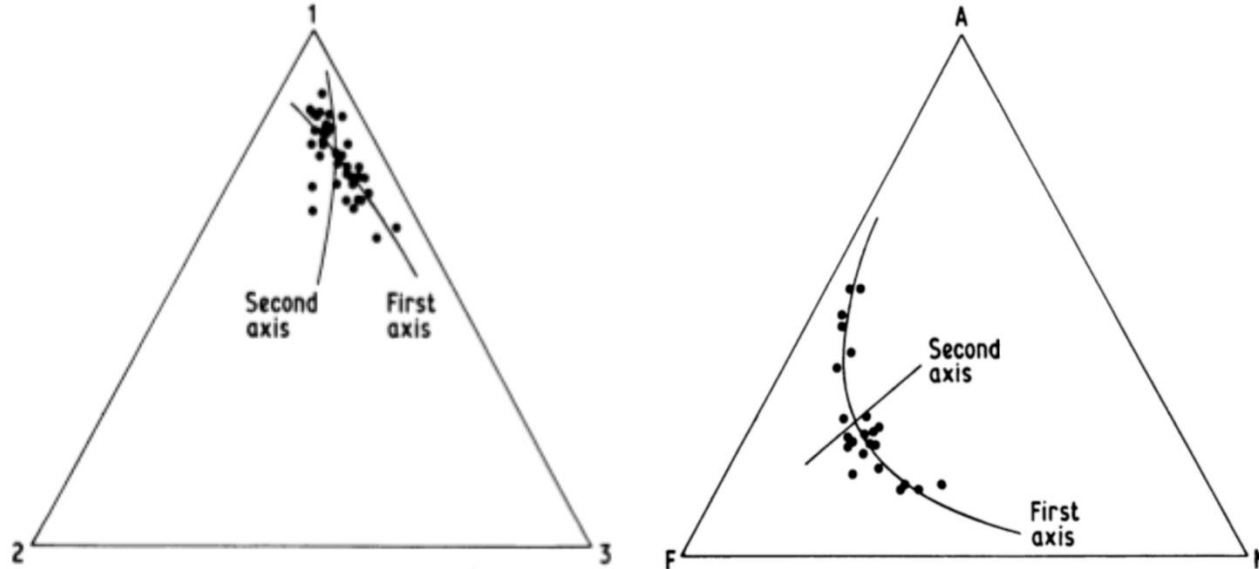


# Log-ratio transformation for compositional data

❑ Centered log-ratio (CLR) 변환: Compositional → Euclidean

$$\text{➤ } clr(x) = \log x_j - \frac{1}{p} \sum_{j=1}^p \log x_j$$

❑ Log-ratio PCA (Aitchison, 1983)



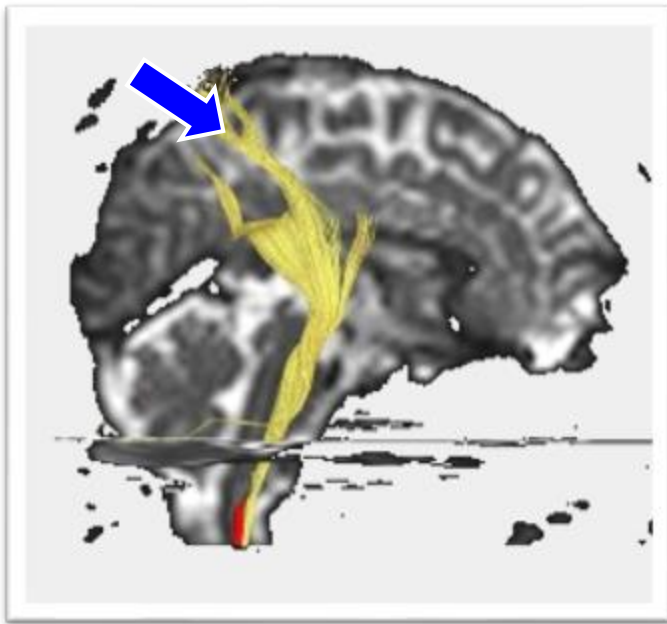


# Non-Euclidean Data Analysis Example (2)

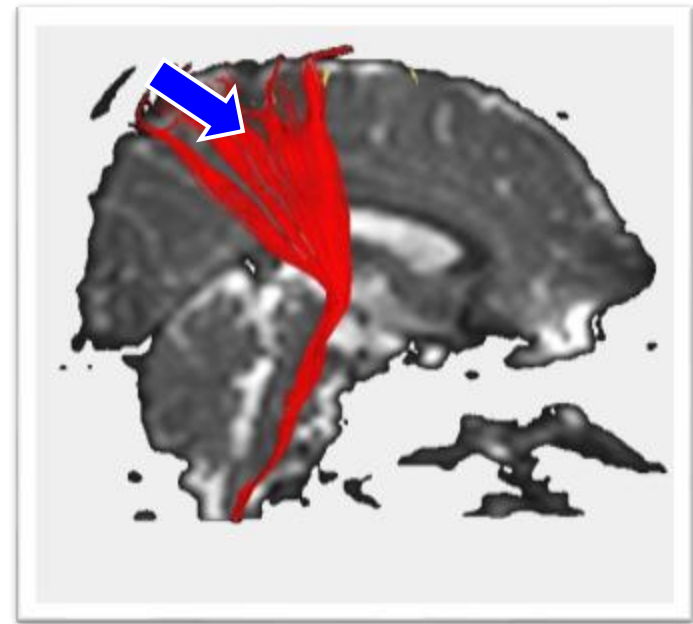
## □ Diffusion Tensor Imaging (DTI) data

- Two-sample testing problem: 두 그룹 간 피질척수경로 비교

환자



정상인



# Spherical data and its distribution

□  $S^{q-1} := \{\mathbf{x} \in \mathbb{R}^q : x_1^2 + \dots + x_q^2 = 1\}$

□ von Mises-Fisher (vMF) distribution

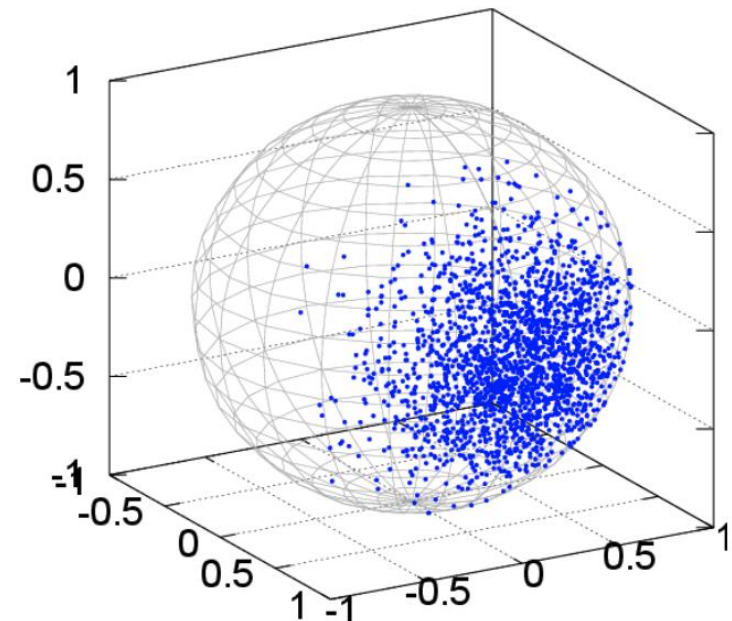
$$f_{vMF}(\mathbf{y}; \boldsymbol{\zeta}, \kappa) = C_q(\kappa) \cdot \exp(\kappa \cdot \boldsymbol{\zeta}^T \mathbf{y}),$$

where  $C_q(\kappa) = \frac{\kappa^{q/2-1}}{(2\pi)^{q/2} \cdot I_{q/2-1}(\kappa)}$  and  $I_\nu(\cdot)$  is the modified Bessel function of the first kind at order  $\nu$ .

➤  $\boldsymbol{\zeta}$  : mean direction

➤  $\kappa$  : *concentration parameter*

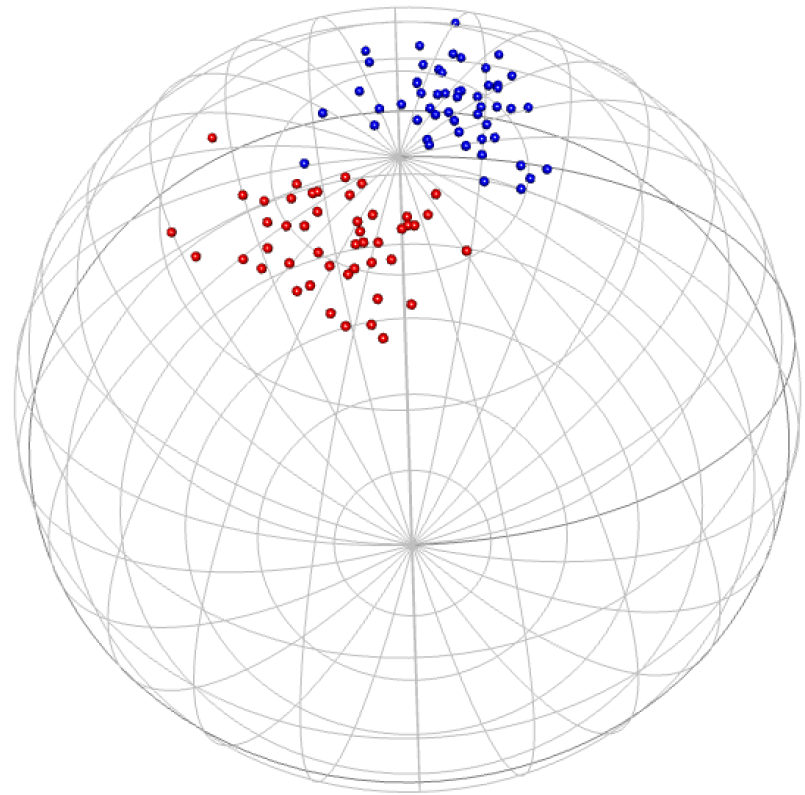
□  $q = 3, \boldsymbol{\zeta} = (1, 0, 0)^T, \kappa = 10 \Rightarrow \Rightarrow$



# Two-sample testing problem in spherical data

## □ (예제) 시뮬레이션 데이터

- 빨강: 정상군; 파랑: 환자군
- 두 그룹 간 평균 위치 차이 존재
- 분산 차이는 없어보임.



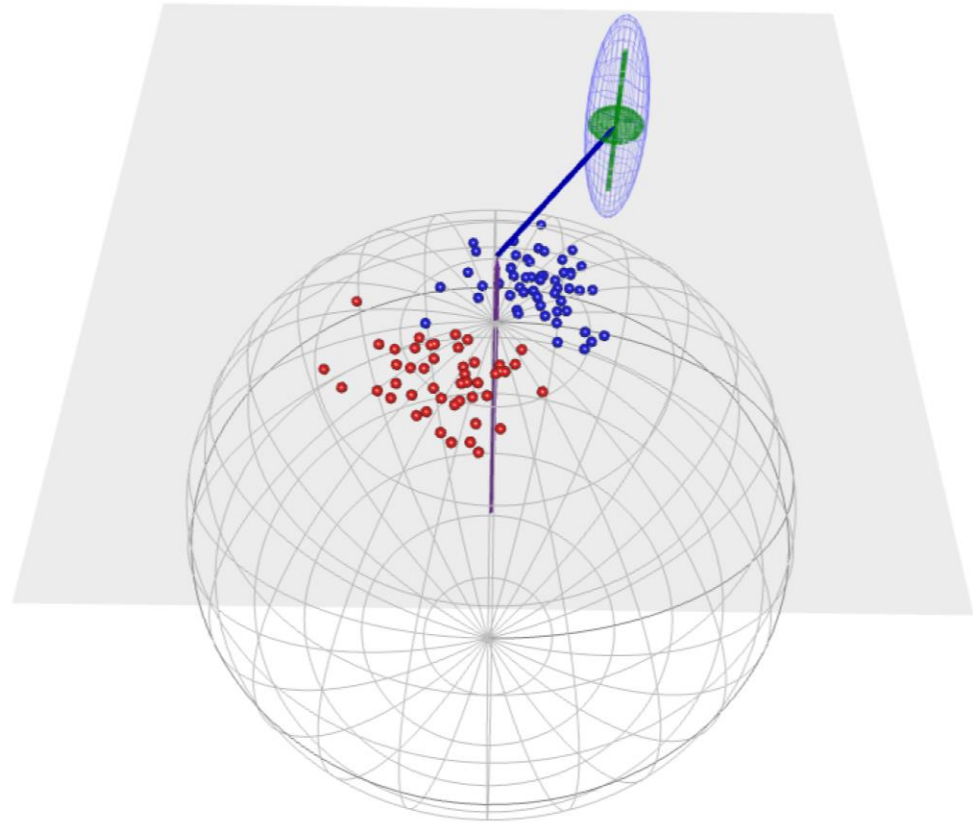
# Two-sample testing problem in spherical data

□ 이에 대한 회귀모형:

$$g( E(\mathbf{y}_i | x_i) ) = \boldsymbol{\theta}_i = \boldsymbol{\mu} + x_i \boldsymbol{\beta},$$

$$\mathbf{y}_i | x_i \sim vMF(\cdot | \boldsymbol{\theta}_i)$$

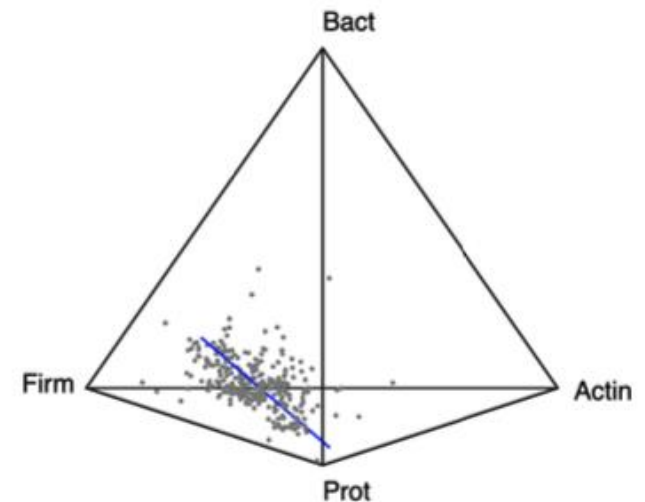
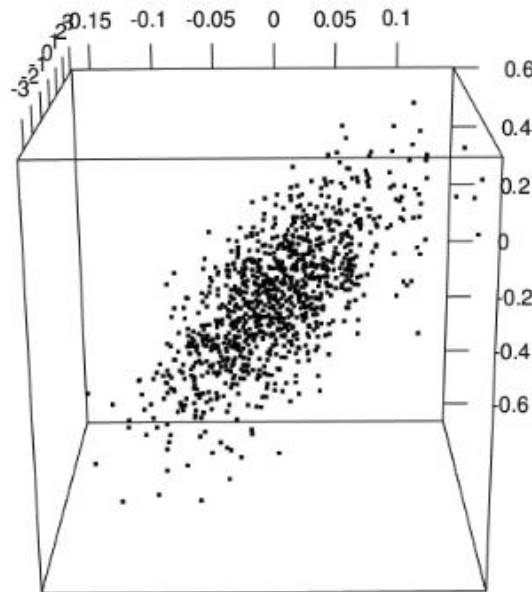
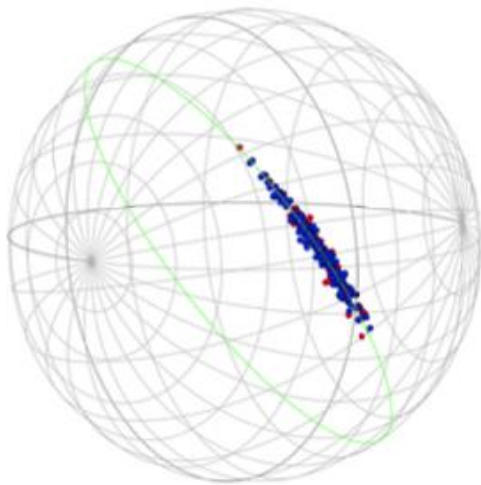
□ 빨강에서 파랑으로의 변화를  
나타내는 회귀계수벡터 및  
점근적 신뢰영역에 대한 추정



# Our future work of interest

## ❑ Non-Euclidean data integration

➤ e.g. structural decomposition for multiple (non-)Euclidean datasets:



Thank you for your attention !