

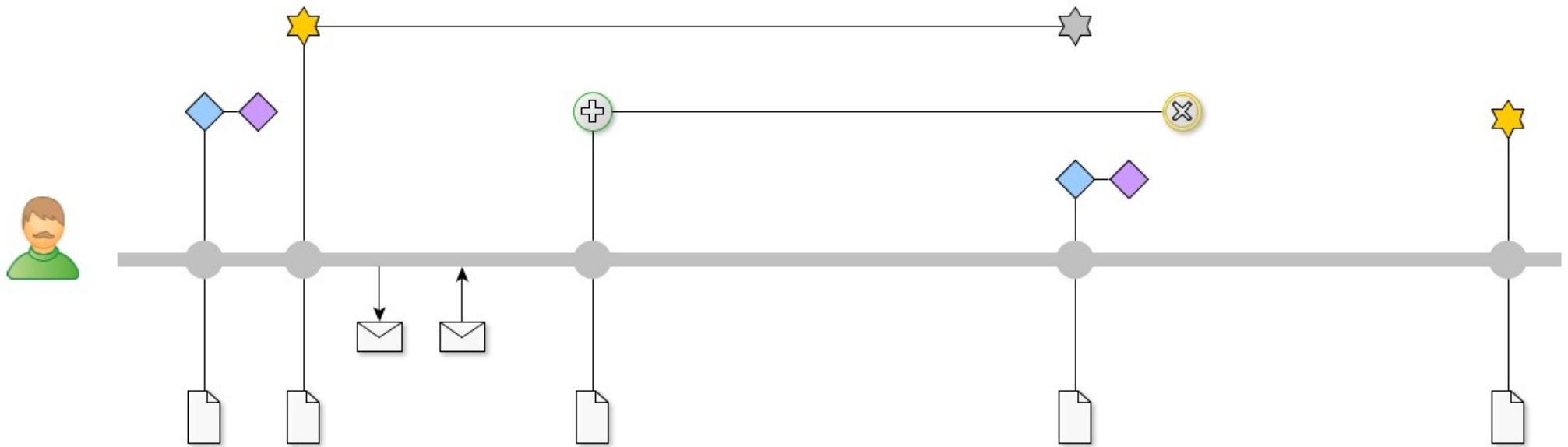
A vibrant cosmic background featuring a large, detailed galaxy in the upper left quadrant, a bright, colorful nebula in the center, and various stars and smaller galaxies scattered across the field. The colors range from deep blues and purples to bright oranges and yellows.

USING WORD EMBEDDINGS TO REPRESENT DIFFERENT TYPES OF CLINICAL DATA

MERIJN BEEKSMA (MERIJNBEEKSMA@GMAIL.COM)

I MEDICAL RECORDS

ELECTRONIC MEDICAL RECORDS



ELECTRONIC MEDICAL RECORDS

MORE OR LESS...

SIMILAR PROPERTIES

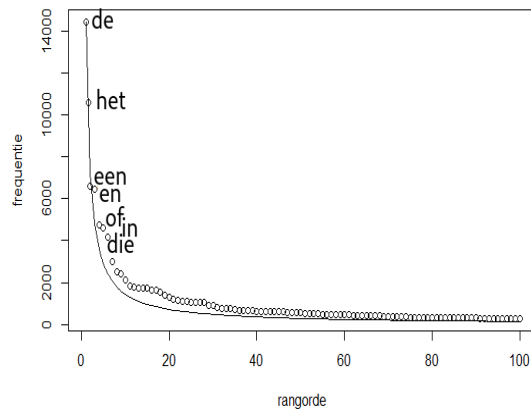
MANY FEATURES

SPARSE FEATURES

ZIPFIAN DISTRIBUTIONS

SIMILAR INFORMATION

TEXT



ICD-10



1M PATIENTS: 5862 UNIQUE CODES, 50% FREQ ≤ 5

GENERIC SOLUTIONS

LANGUAGE-INDEPENDENT
APPLICABLE TO MULTIPLE DATA TYPES
ABLE TO HANDLE UNSEEN INPUT
ROBUST TO NEW DEVELOPMENTS

SIMPLE SOLUTIONS

MINIMAL PREPROCESSING
RETAIN IDIOSYNCRACIES

SHAREABLE SOLUTIONS

"DATA CANNOT LEAVE THE BUILDING"
HANDLE DISTRIBUTED DATA SOURCES

P R O S A N D C O N S

PROS

MINIMAL PREPROCESSING

RETAIN/DETECT IDIOSYNCRACIES

CAPTURE SIMILARITY

DENSE REPRESENTATION

SMALL AMOUNT OF FEATURES

CONS

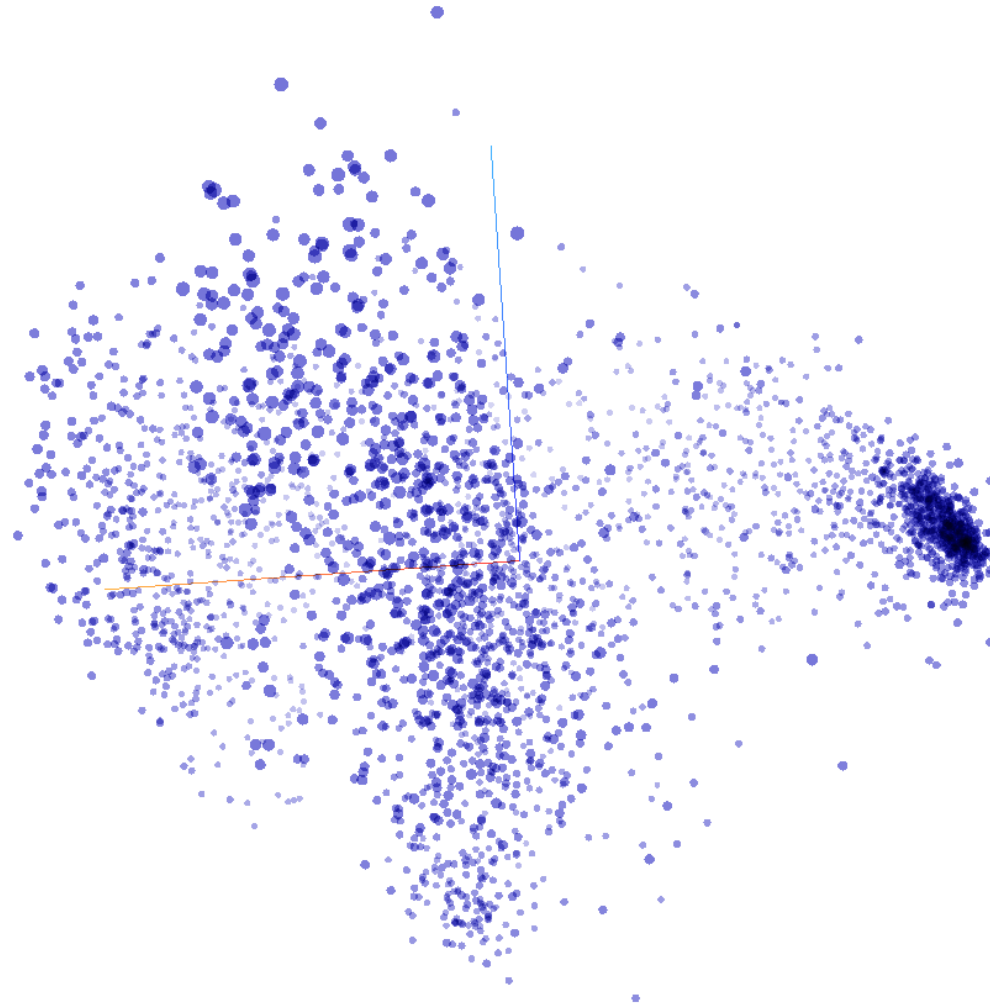
EVALUATION

OTHER DATA TYPES

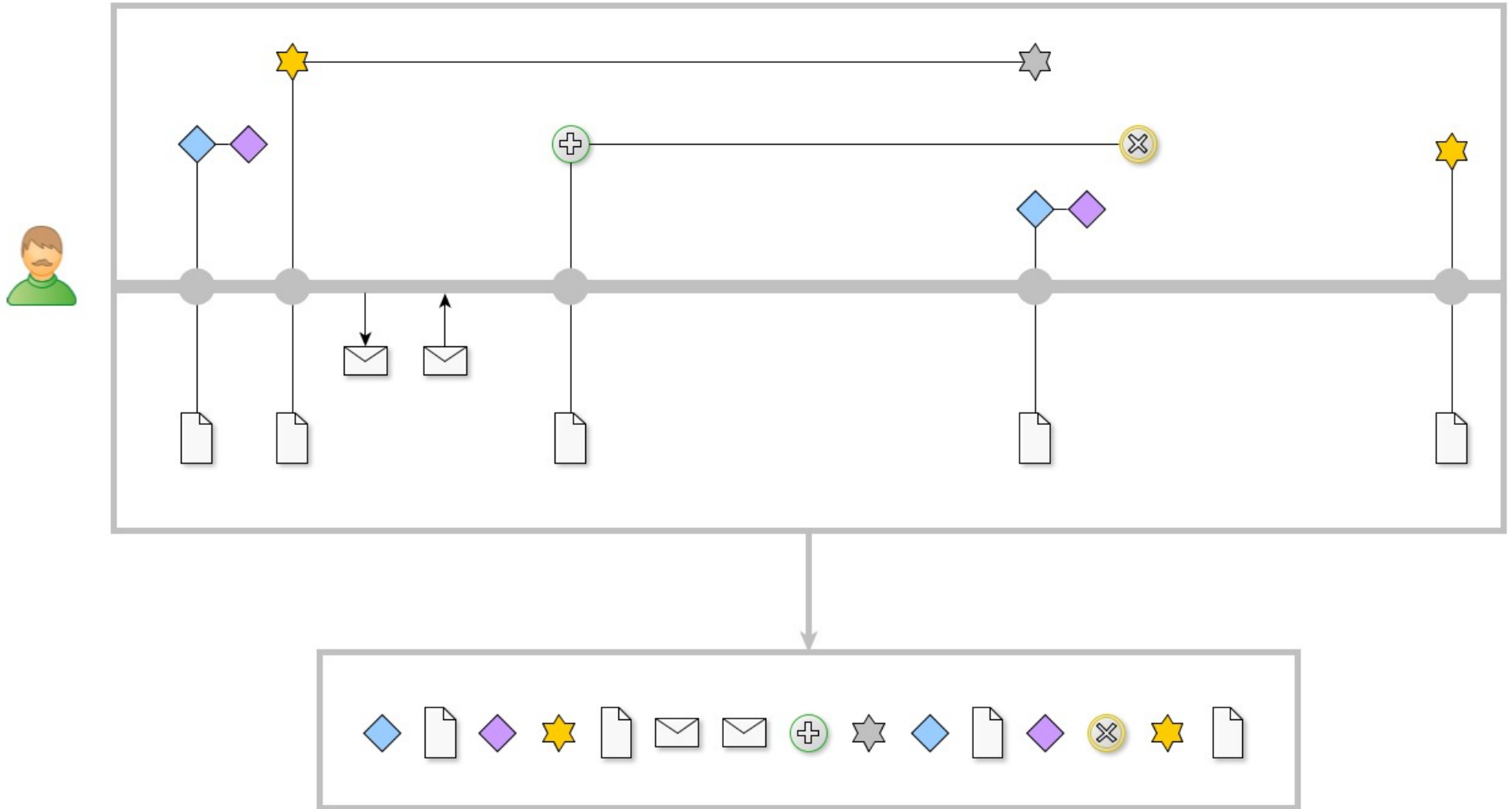
REDUNDANCY WITH OTHER DATA

FREQUENCY IMPACTS STABILITY

EMBEDDED ICPC-1 CODES



TIMELINE TO SENTENCE



' P I L E ' O F D A T A



' P I L E ' O F D A T A

HOWEVER...

- CAN'T INCLUDE TEXT DATA
- SOME DATA TYPES CLUSTER TOGETHER
- COMPARED TO DATA-TYPE SPECIFIC SPACES,
SIMILARITIES WITHIN DATA TYPES ARE DISTURBED
- MANY UNSTABLE DATA POINTS

II STABILITY

WHY MEASURE STABILITY?

OPTIMIZE PARAMETER SETTINGS

DETERMINE IMPACT FREQUENCY

LEVERAGE STABLE POINTS

TO STABILIZE UNSTABLE POINTS

INTRINSIC MEASUREMENT OF QUALITY

WHY NOT JUST DOWNSTREAM TASK?

OVERFITTING

HOW TO MEASURE STABILITY?

- I EMBED SAME DATA MULTIPLE TIMES WITH DIFFERENT INITIALIZATION*
- II FOR EACH ITEM:
 - DETERMINE SIMILARITY BETWEEN THE VECTORS OF THIS ITEM IN DIFFERENT SPACES
- III DO SOMETHING USEFUL WITH IT, SUCH AS:
 - CALCULATE AVERAGE STABILITY
 - RANK THE ITEMS BY STABILITY

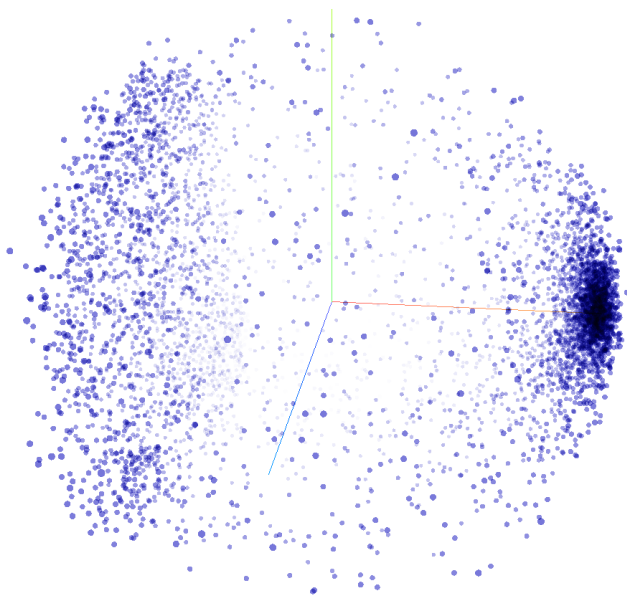
*NB: WANT TO MAKE A FULLY REPRODUCIBLE RUN? (YES!)

- FIX ALGORITHM PARAMETER "SEED"
- USE 1 CPU
- FIX ENVIRONMENTAL VARIABLE "PYTHONHASHSEED"

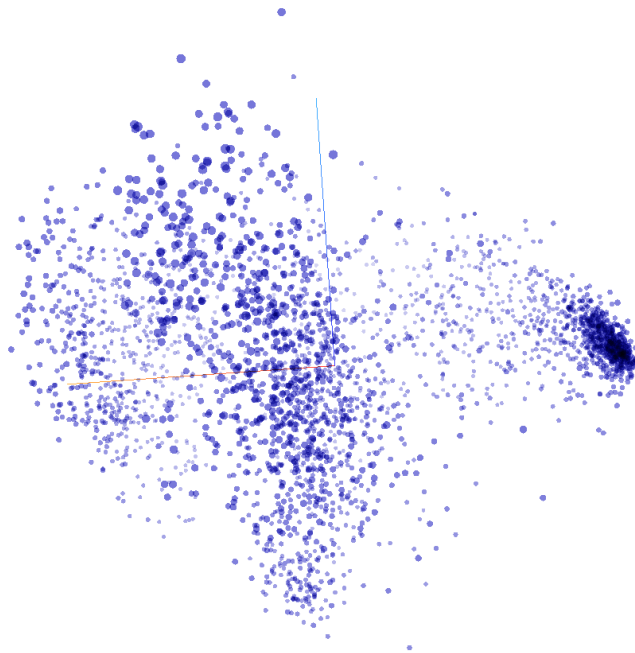
WHEN WORKING WITH PYTHON AND GENSIM

II MAPPING SPACES

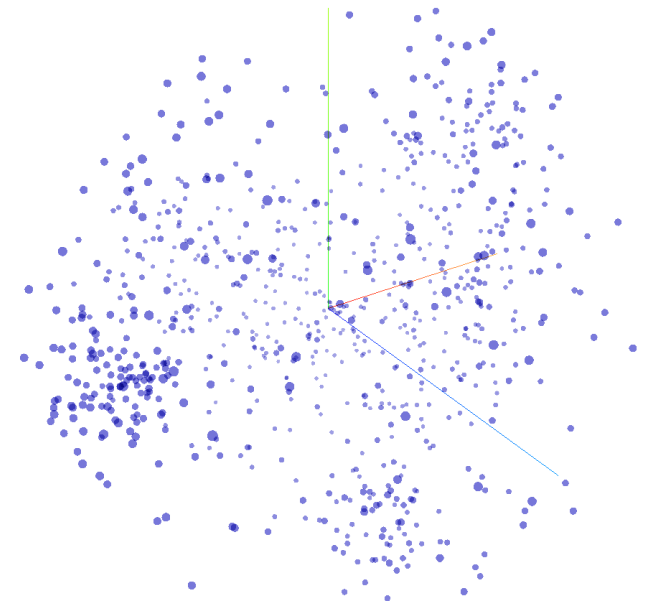
MAPPING BETWEEN CODEBOOKS



ICD-10



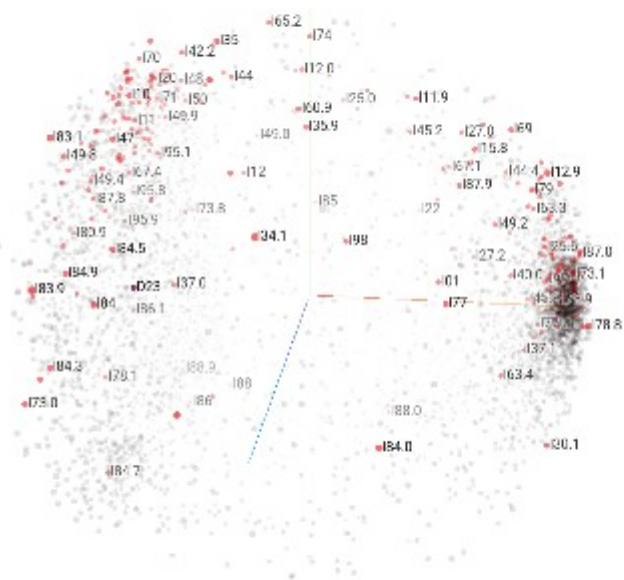
ICPC-1



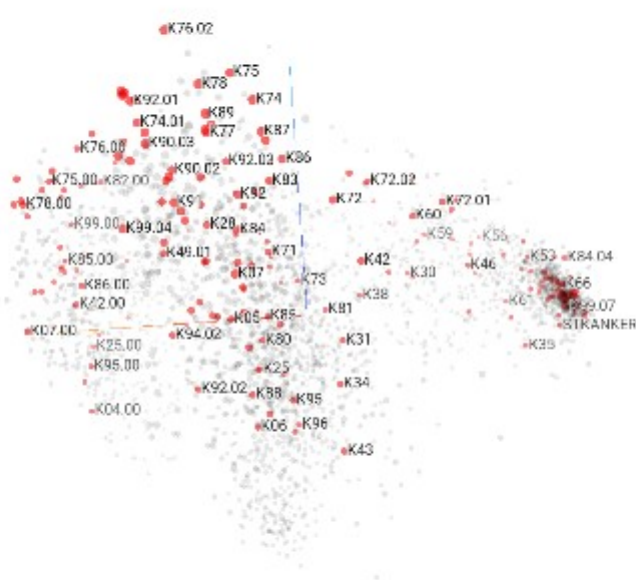
ICPC-2

MAPPING BETWEEN CODEBOOKS

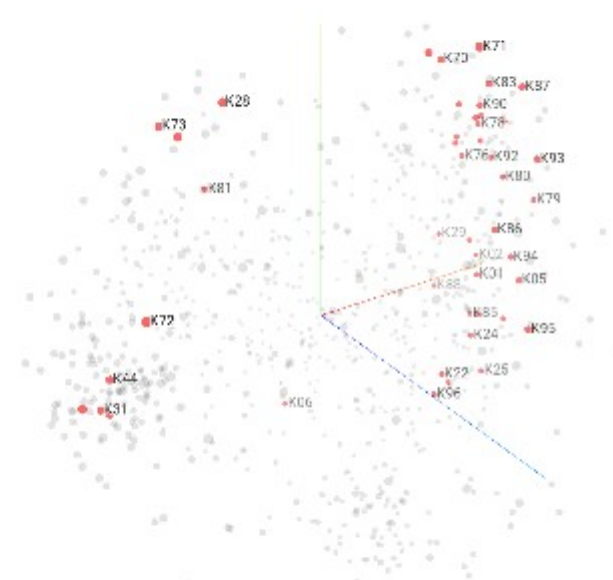
CIRCULATORY SYSTEM



ICD-10



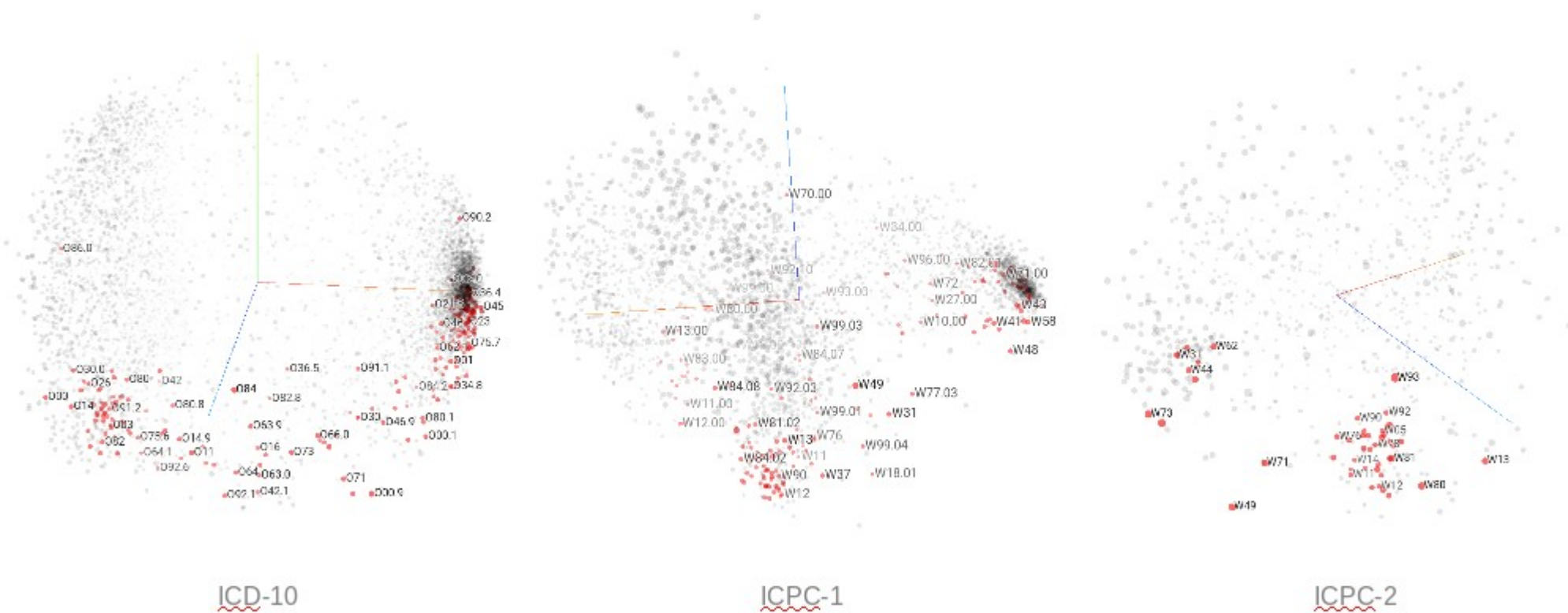
ICPC-1



ICPC-2

MAPPING BETWEEN CODEBOOKS

PREGNANCY



PROJECT ONTO SAME SPACE

HOW?

- I DETERMINE ANCHOR POINTS
- II RANK BY STABILITY
- III ROTATE SPACE A ONTO SPACE B
(E.G. WITH LEAST SQUARED ERROR METRIC)

WHY?

AUTOMATIC MAPPING

SIMILAR DATA, SIMILAR REPRESENTATION

→ MINIMIZES AMOUNT OF FEATURES

ORIGINAL SPACES ARE NOT ALTERED

HMM...

WILL IT WORK FOR MORE DIVERSE DATA TYPES TOO?