THE EHR LANGUAGE GARDEN

Leveraging Variability in Health Documentation

Denis Newman-Griffis

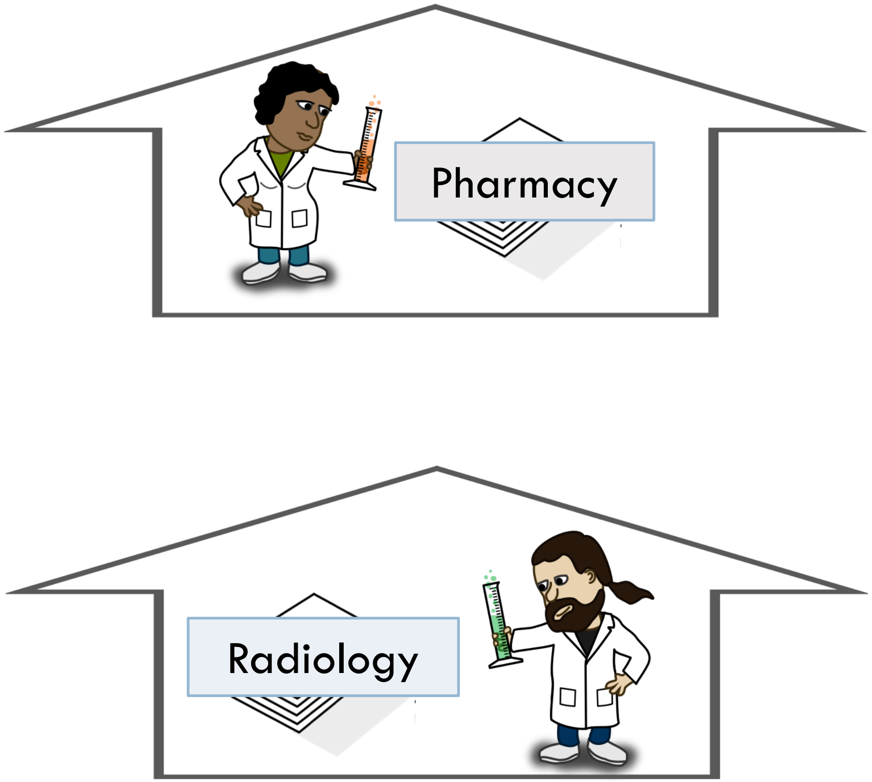
NIH Clinical Center / University of Pittsburgh

September 1, 2020

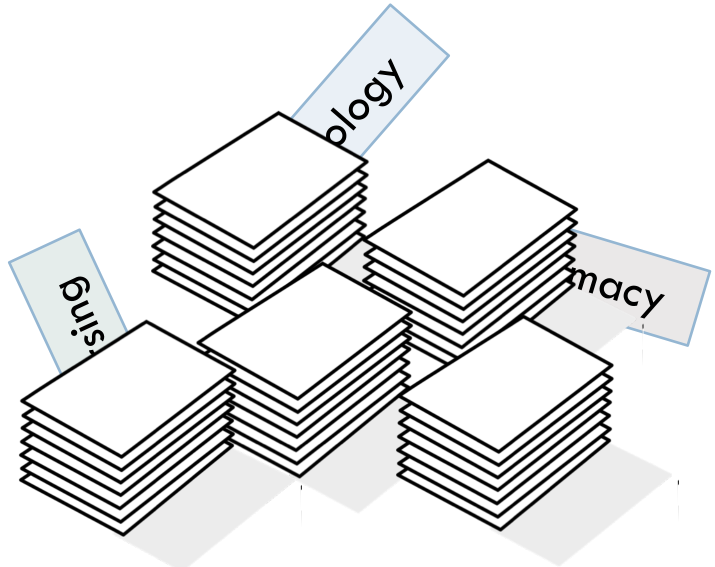
National Institutes of Health
Clinical Center


Radiology
Pharmacy
Nursing
Discharge Summaries

Research → Practice

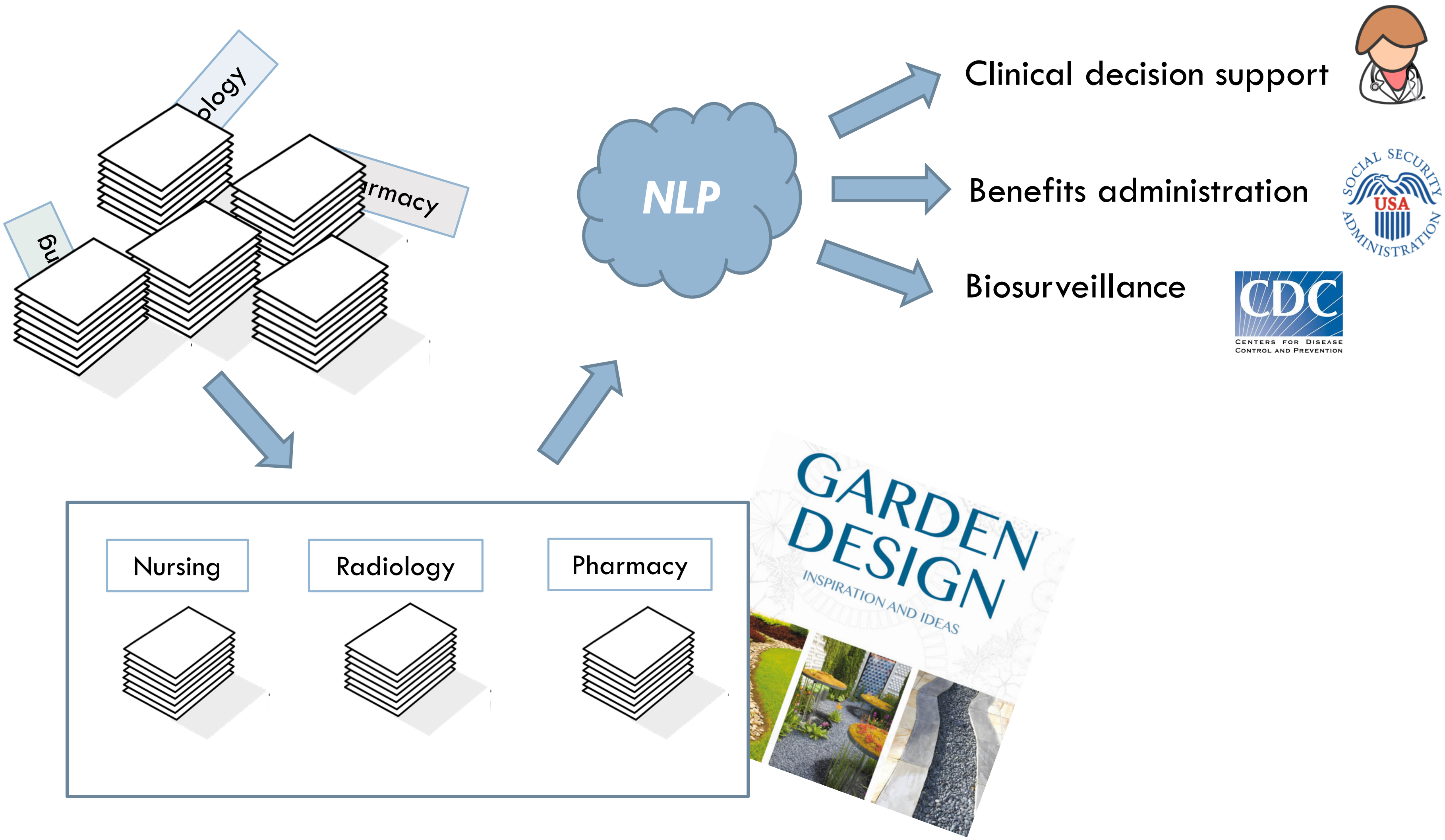


Research → Practice



SOCIAL SECURITY ADMINISTRATION
USA

Sublanguage: the secret sauce



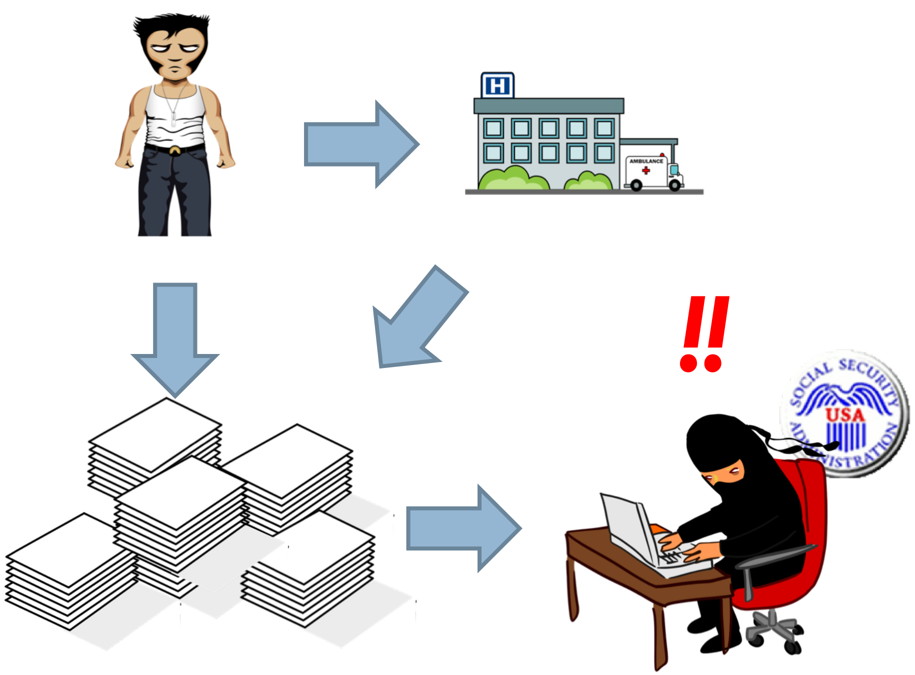
ERH data variability at scale

Content – what do the records say?

Form – how do they say it?

Structure – what are the pieces?

Context: SSA disability programs



- ✓ National data
- ✓ All providers/EHRs
- ✓ Unreliable metadata

“Defining” disability

Medical conditions

- High mortality conditions
- Medical listings (business rules)

Functional limitations

- Ability to perform work-related activities
- Substantial Gainful Employment

Need NLP that can handle both!

Planting the garden: findings

Content – Rehabilitation medicine as a sublanguage

Form

Structure

Health strategies – Rehabilitation

- Preventive
- Curative
- Rehabilitative
- Supportive

Goal: Restore/optimize function

- Adapt to health condition (e.g., chronic or incurable)
- Interactions with world

Under-studied domain

Health strategies – Curative

- Preventive
- Curative
- Rehabilitative
- Supportive

Goal: Cure health conditions

- Diagnosis
- Treatment
- Physiological/internal

Most of clinical NLP!

Multi-institution data

BTRIS

- 155K records
- Research patients
- 130 doctypes

OSUMC

- 418K records
- Chronic diseases
- 43 doctypes

MIMIC-III

- 2M records
- ICU admissions
- 25 doctypes*

Data classifications

Domain

Discipline

Functional Area

Diagnostic – most concerned with diagnosis and treatment

Functioning – most concerned with functioning status

Data classifications

Domain

Discipline

Functional Area

Medical – curative documents from ”standard” specialties

Therapy – therapeutic specialties (PT, OT, RT, etc.)

Ancillary – non-physiological (mental health, social work)

Other – administrative documents

Data classifications

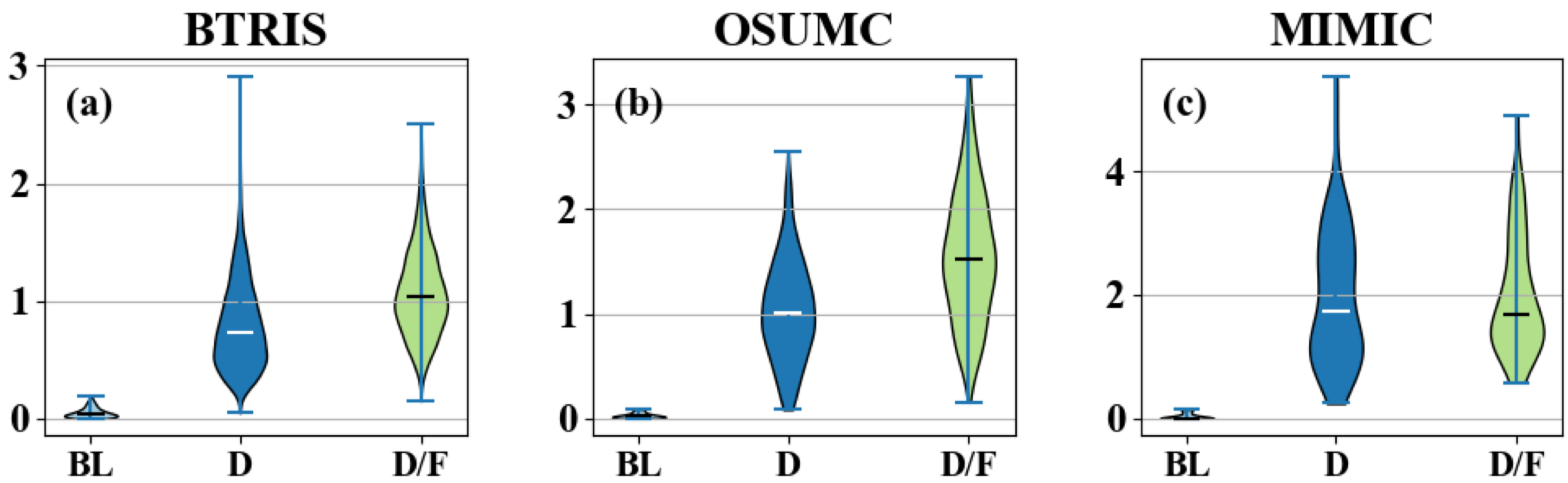
Domain

Discipline

Functional Area

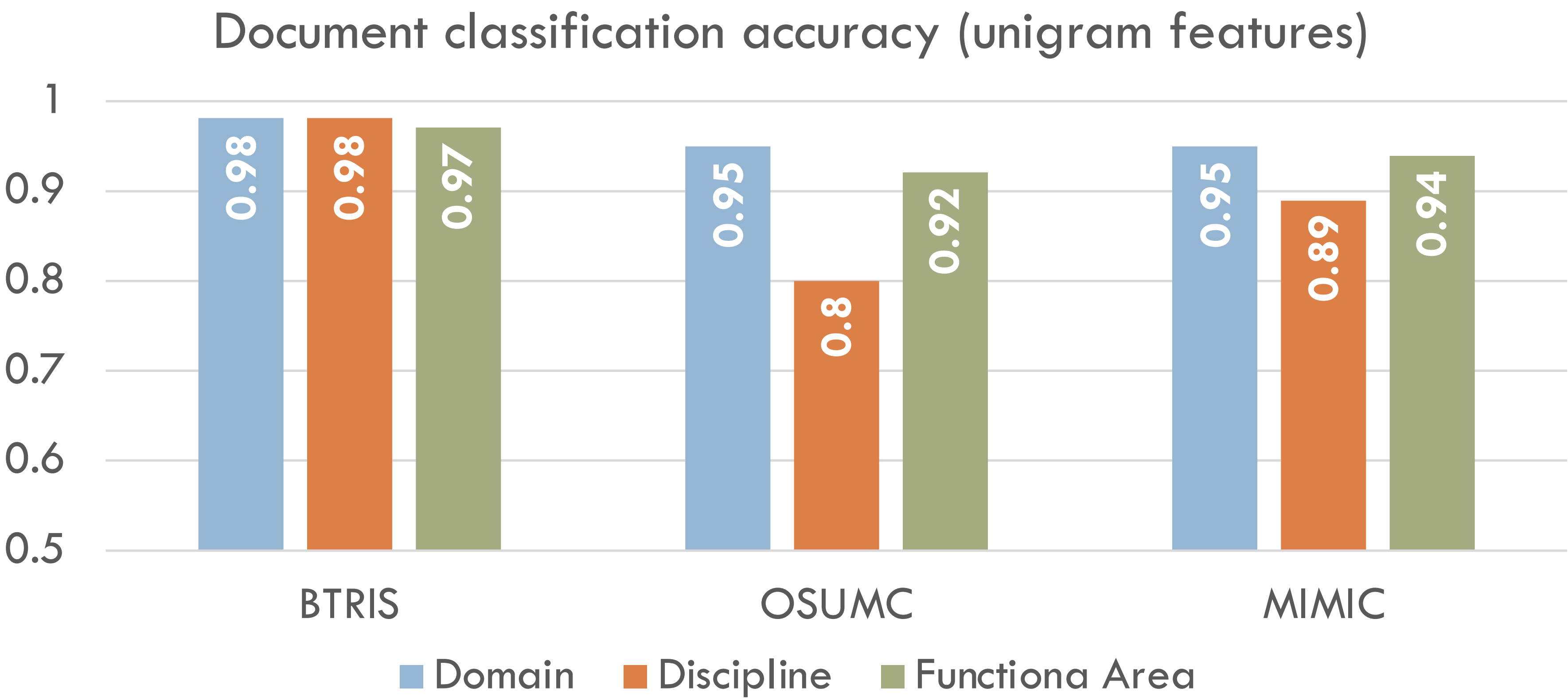
PT – physical therapy
OT – occupational therapy
RT – recreational therapy
SLP – speech/language pathology
Psych – psychological/iatric
Neuro – neurological
SW – social work
General – catchall bucket

Rehab medicine vocabulary is distinct

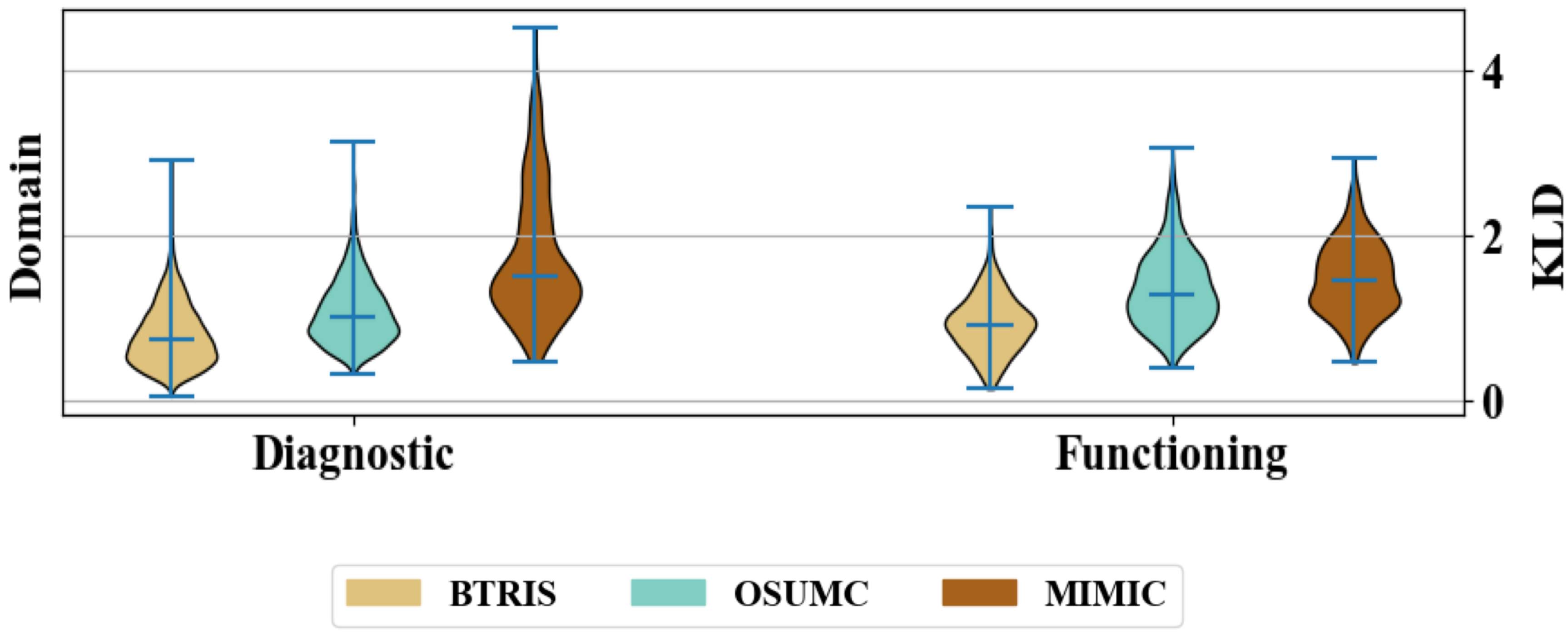


BL = Variance within document types
D = Variance between doctypes in Diagnostic
D/F = Variance between Diagnostic doctypes and Functioning doctypes

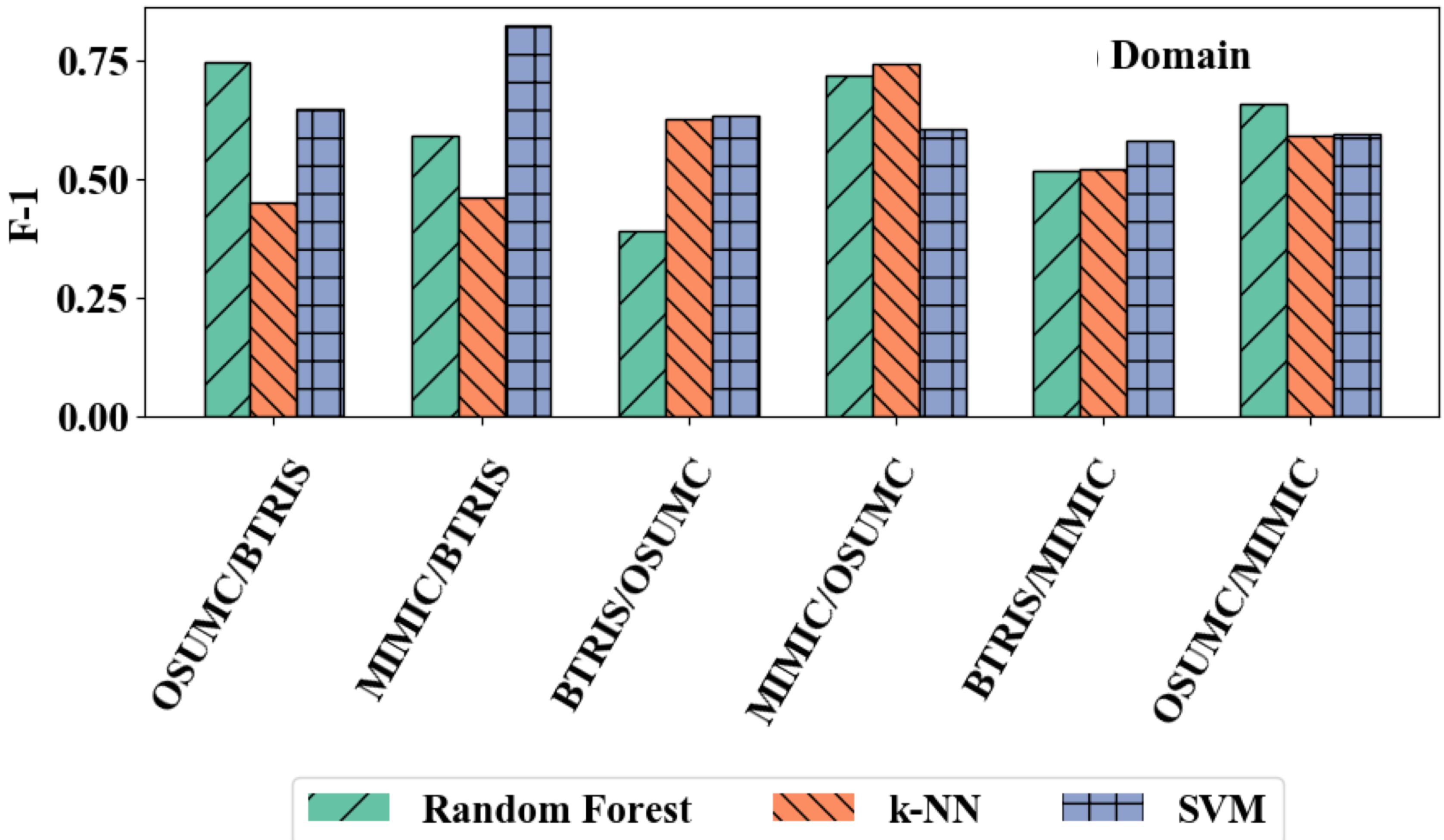
Rehab medicine vocabulary is distinct



Significant differences across institutions



Significant differences across institutions



Different structure of information

i2b2

```
Ejection fraction: 90%
Lab creatinine: 3 mg/dL

There has been removal
of [a swan-ganz
catheter]Treatment and
placement of [a right
internal jugular
vascular
catheter]Treatment.
```

Rehab data

```
Pt 45 yr old tech worker,
sedentary activity but
hikes on weekends.

[Ambulation: 4]Mobility
Observations:
Pt is weight bearing: [she
ambulates independently
w/o use of assistive
device]Mobility. Limited to
very brief examination.
```

Planting the garden: findings

Content

Form – Differences in clinical concept usage

Structure

D N-G, E Fosler-Lussier. “Writing habits and telltale neighbors: analyzing clinical concept usage patterns with sublanguage embeddings.” *LOUHI*, 2019.

Characterizing document types

Document/section structural patterns inform meaning

- Field names *vs* observations
- Temporality (future/past/recurrent)
- Perceived importance (e.g. Chief Complaint)

Document types change priors for disambiguation

- "Depression" in Psychiatric Consult vs GE Exam

Discharge summaries
!=
Nursing notes

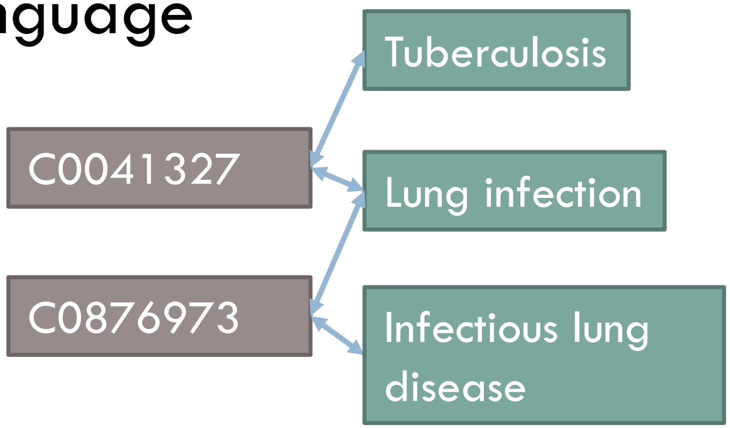
Conceptual vs lexical analysis

Prior work used lexical content to describe clinical sublanguages

- Feldman et al, 2016
- Grön et al, 2019

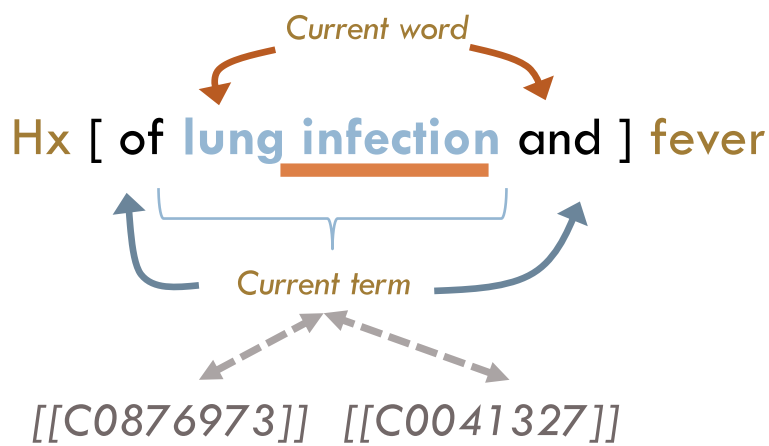
Concepts (symptoms, diseases, procedures, etc) are stock in trade of clinical language

- Multiple surface forms
- Ambiguity ("Cold")



Learning concept embeddings: JET

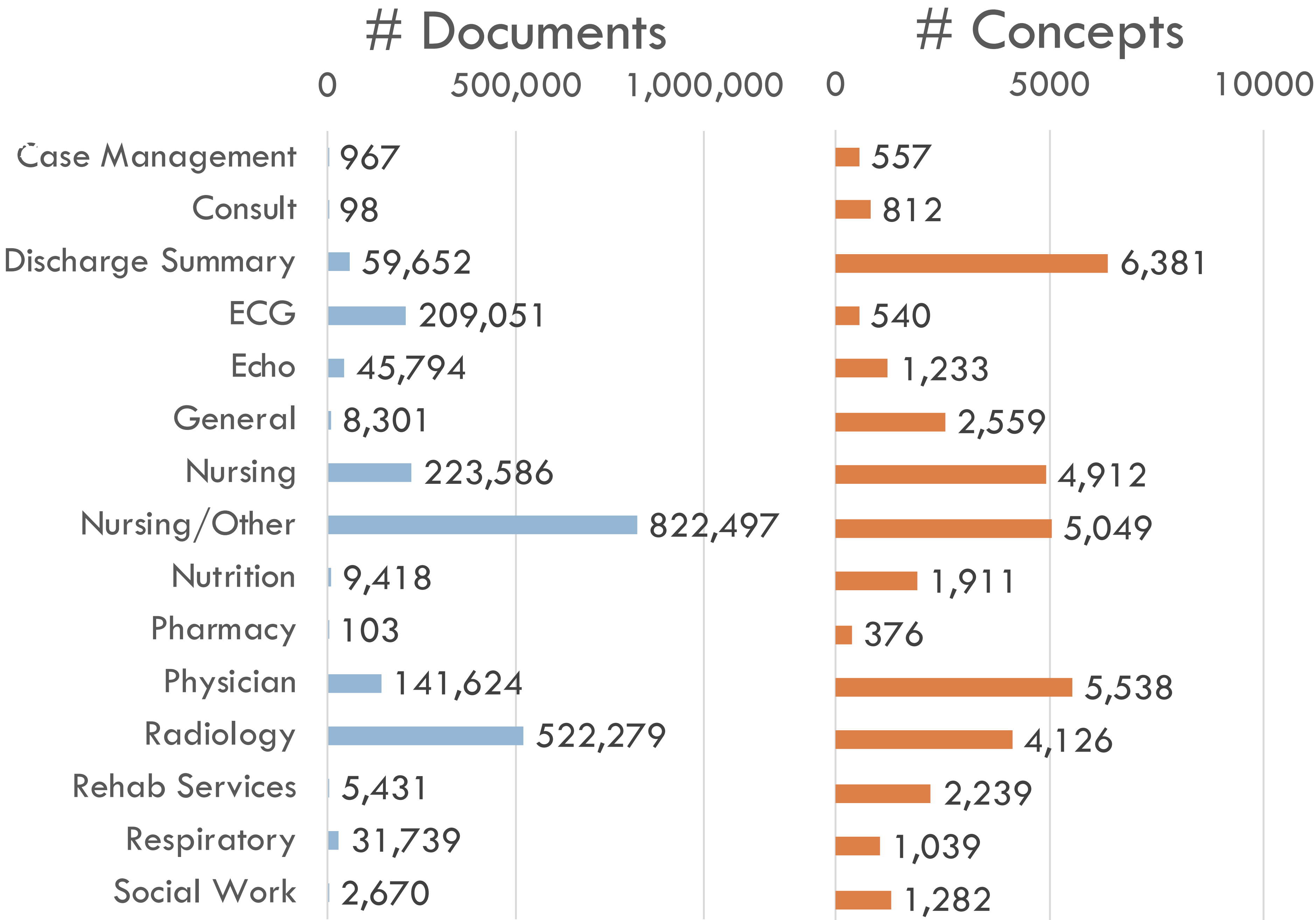
- Train word/term/concept embeddings jointly
- Distant supervision using known terminology
- Noisy, but good quality



Terminologies: SNOMED CT, LOINC

Data: MIMIC-III

D N-G et al. "Jointly embedding entities and text." Repl4NLP, 2018.



Documents
Concepts
0
500,000
1,000,000
0
5000
10000
Case Management
Consult
Discharge Summary
ECG
Echo
General
Nursing
Nursing/Other
Nutrition
Pharmacy
Physician
Radiology
Rehab Services
Respiratory
Social Work
967
98
59,652
209,051
45,794
8,301
223,586
822,497
9,418
103
141,624
522,279
5,431
31,739
2,670
557
812
6,381
540
1,233
2,559
4,912
5,049
1,911
376
5,538
4,126
2,239
1,039
1,282

Measuring concept usage similarity

- Measured by overlap of nearest neighbor sets
- Similarity metric in [0,1]
- Compare inter-type overlaps to intra-type overlaps

Neighbors of Onion

Set A	Set B
Cucumber	Squash
Squash	Pumpkin
Beans	Pasta
Green	Beans
Pasta	Cheese

Inter-type similarity is significantly lower than intra-type

	Case Management	Discharge Summary	Echo	Nursing/Other	Nutrition	Physician	Radiology
Case Management	0.75	0.01	0.01	0.00	0.00	0.00	0.01
Discharge Summary	0.01	0.67	0.24	0.32	0.00	0.34	0.33
Echo	0.01	0.24	0.65	0.13	0.00	0.36	0.40
Nursing/Other	0.00	0.32	0.13	0.60	0.00	0.27	0.31
Nutrition	0.00	0.00	0.00	0.00	0.73	0.01	0.00
Physician	0.00	0.34	0.36	0.27	0.01	0.57	0.26
Radiology	0.01	0.33	0.40	0.31	0.00	0.26	0.63

Nearest neighbors:
Diabetes Mellitus (C0011849)

Discharge Summary	Nursing/Other	Radiology
Diabetes (C0011847)	Gestational Diabetes (C0085207)	Poorly controlled (C3853134)
Type 2 (C0441730)	A2 immunologic symbol (C1443036)	Insulin (C0021641)
Type 1 (C0441729)	Diabetes Mellitus, Insulin-Dependent (C0011854)	Diabetes Mellitus, Insulin-Dependent (C0011854)
Gestational Diabetes (C0085207)	Factor V (C0015498)	Diabetes Mellitus, Non-Insulin-Dependent (C0011860)
Diabetes Mellitus, Insulin-Dependent (C0011854)	A1 immunologic symbol (C1443035)	Stage level 5 (C0441777)

Strings: "diabetes mellitus", "diabetes mellitus dm"

Nearest neighbors:
Mental state (C0278060)

Discharge Summary	Echo	Radiology
Coherent (C4068804)	Donor [LOINC] (C3263710)	Mental status changes (C0856054)
Confusion (C0009676)	Donor person (C0013018)	Abnormal mental state (C0278061)
Respiratory status [LOINC] (C2598168)	Respiratory arrest (C0162297)	Level of consciousness (C0234425)
Respiratory status (C1998827)	Organ donor [LOINC] (C1716004)	Level of consciousness [LOINC] (C4050479)
Abnormal mental state (C0278061)	Swallowing (C4281783)	Mississippi (C0026221)

Strings: "mental status", "mental state"

Embeddings pick up template patterns

Mental status in Echo notes

```
PATIENT/TEST INFORMATION
Indication: Pt presents with
reduced mental status
```

```
PATIENT/TEST INFORMATION
Indication: Pt presents in
vegetative state, consider for
organ donation
```

Planting the garden: findings

Content

Form

Structure – Structural text features capture format and content variation

Work by Bart Desmet, Guy Divita, and Aya Zirikly

Sources of variability in SSA data

Document Source

- SSA Consultative Exams, CCDA documents from EHR, VA data, scanned notes

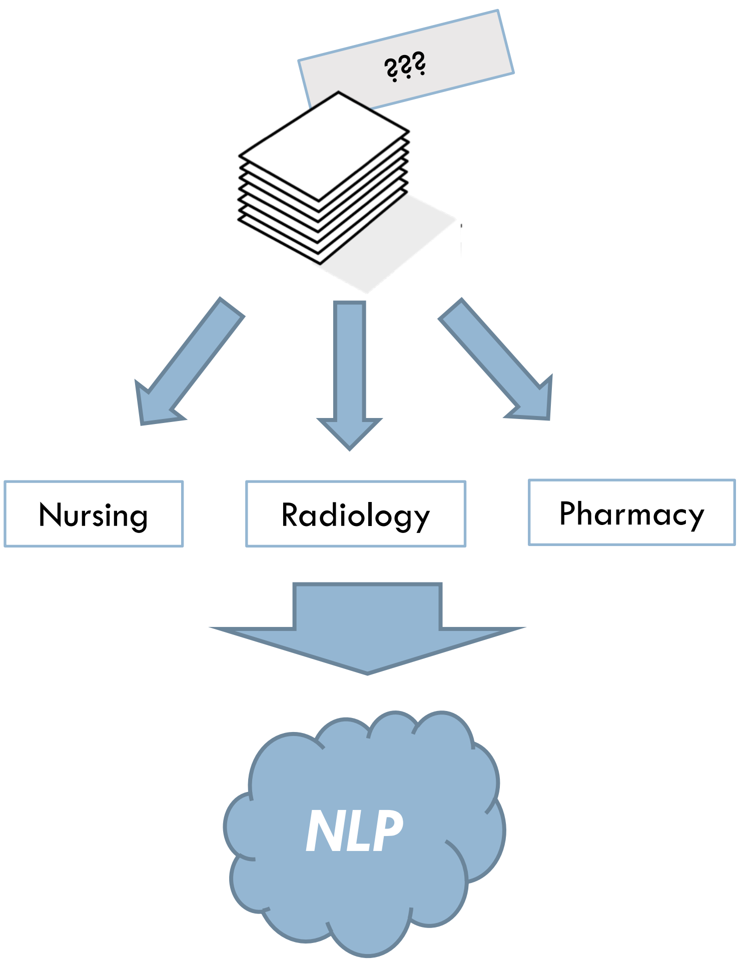
Content Types

- SOAP notes, radiology reports, labs, surveys

Formatted Structure

- Headers/footers, columns, section names, checkboxes

Classify early, process better



- 70K documents
- Disability claimants from 5 states
- Unreliable doctypes

Page-level Features

- Number of Characters, Words, Lines, Sentences
- Number of Punctuation, Delimiters
- Number of Section Names, Section Zones, Nested Sections
- Number of Slot Values, Slot Names, Slot Value Values
- Number of Check Boxes (this wasn't actually working as it turns out)
- Number of Tables
- Number of Lists, List Elements
- Number of Questions
- "Text Tiling" Vector fingerprint (2 numbers)

Related Work: Text Tiling

Marti Hearst (1994): Using word sequences to build a signal to indicate topic/paragraph shifts.

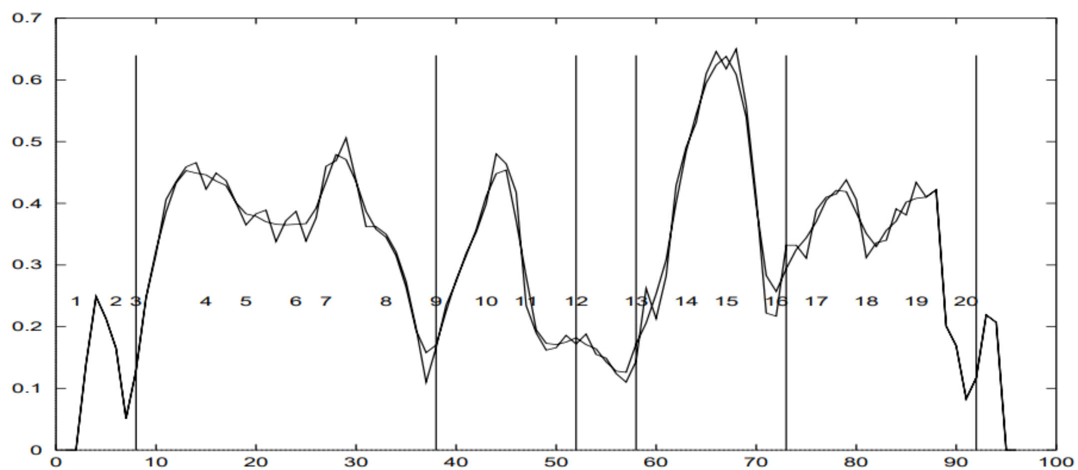


Figure 6
Results of the block similarity algorithm on the *Stargazer* text with k set to 10 and the loose boundary cutoff limit. Both the smoothed and unsmoothed plot are shown. Internal numbers indicate paragraph numbers, x-axis indicates token-sequence gap number, y-axis indicates similarity between blocks centered at the corresponding token-sequence gap. Vertical lines indicate boundaries chosen by the algorithm; for example, the leftmost vertical line represents a boundary after paragraph 3. Note how these align with the boundary gaps of Figure 5 above.

Marti A. Hearst, Multi-Paragraph Segmentation of Expository Text. *Proceedings of the 32nd Meeting of the Association for Computational Linguistics*, Los Cruces, NM, June, 1994.

Page-level PCA

Input

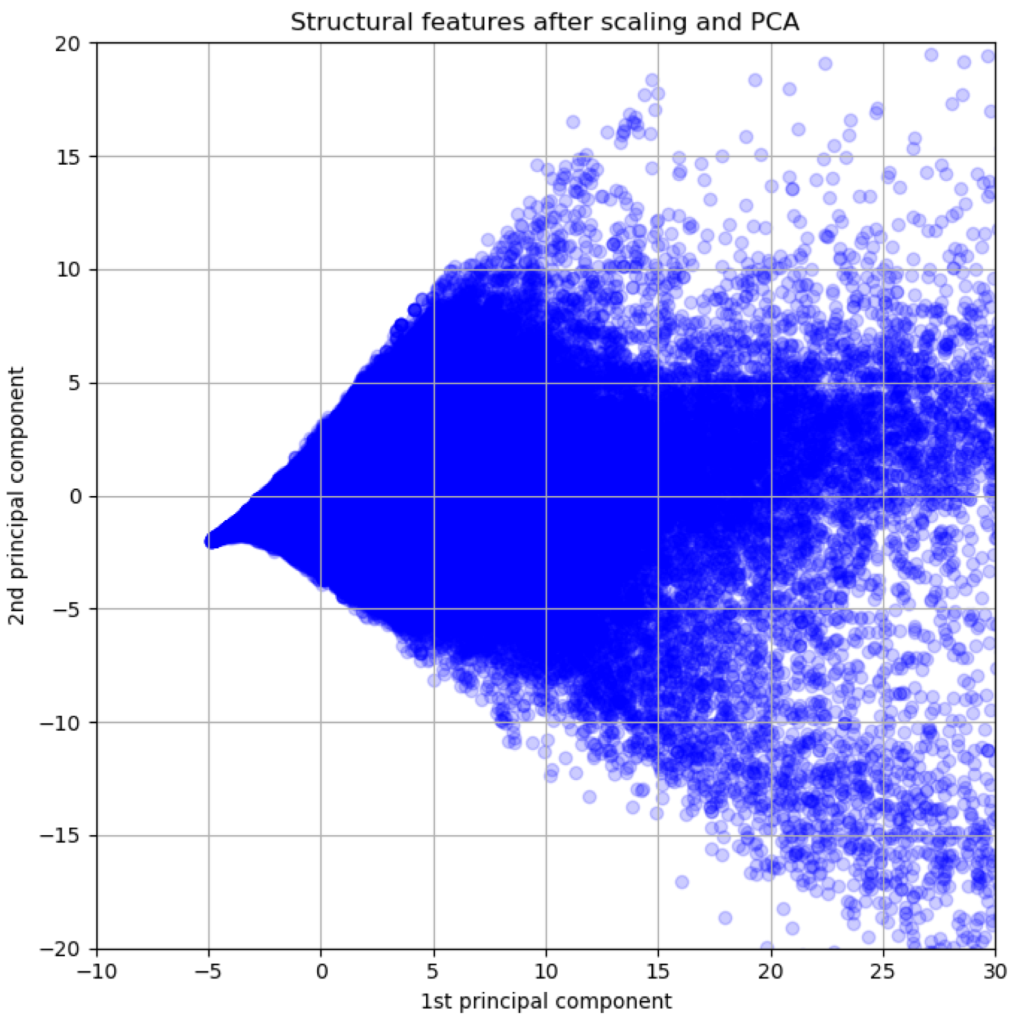
- Structural features for each page

Output

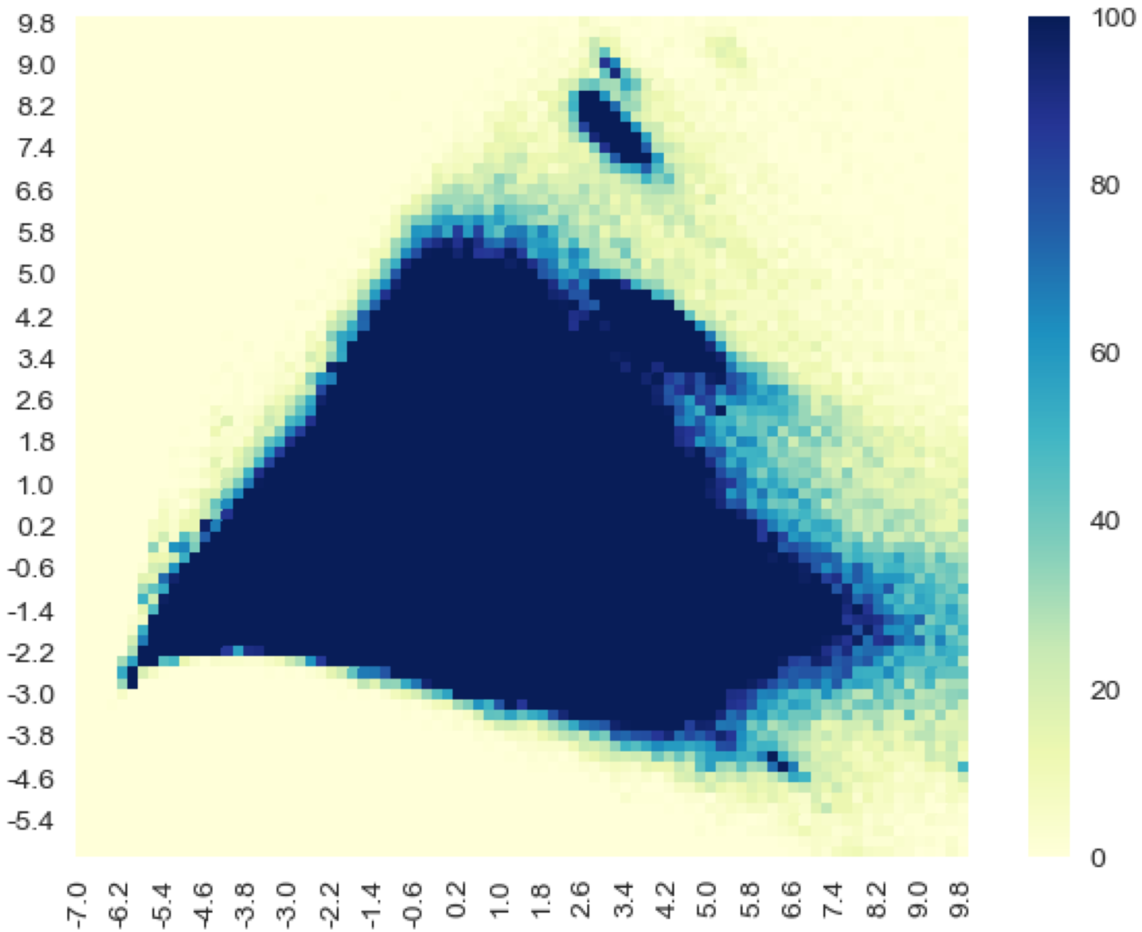
- Orthogonal transform into a set of principal components
- Dimensionality reduction and variance identification

PCA "Angelfish" Plot

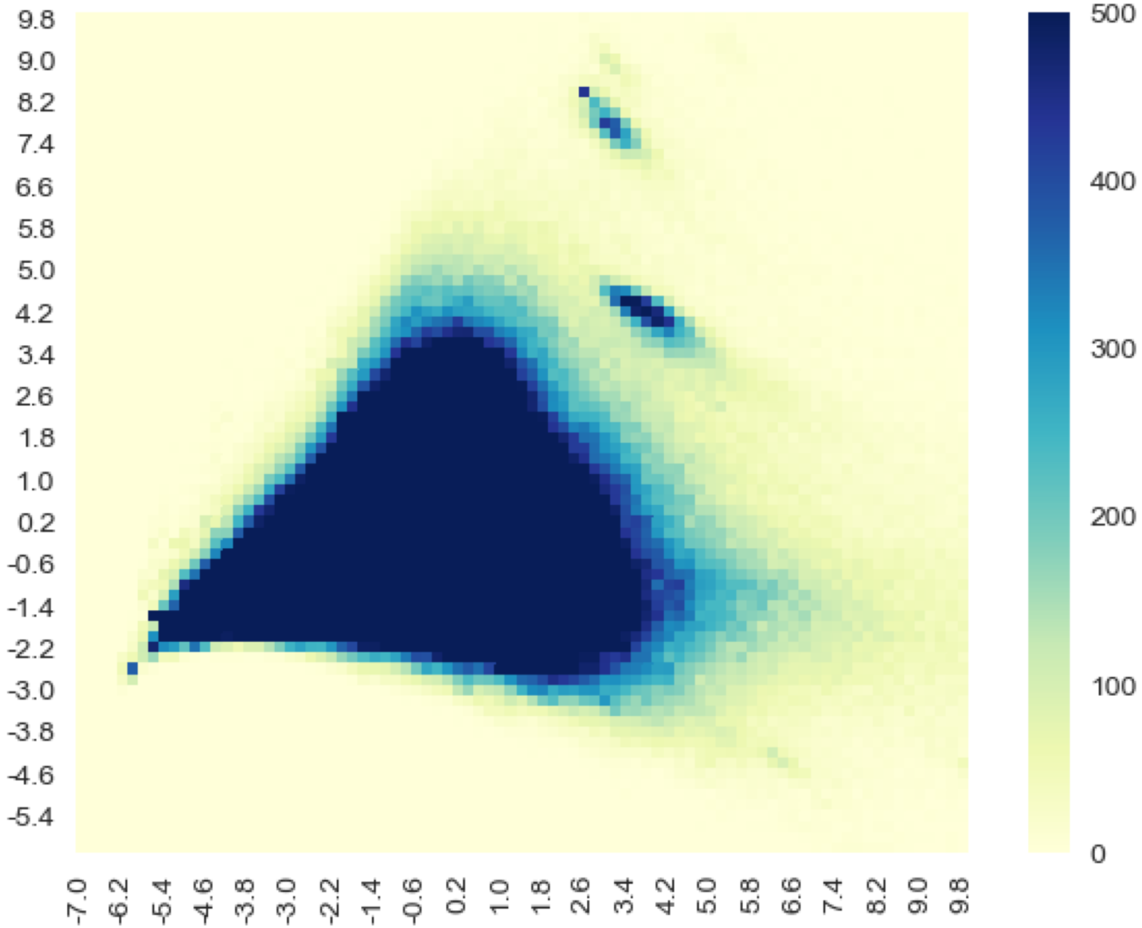
PCA "Angelfish" Plot



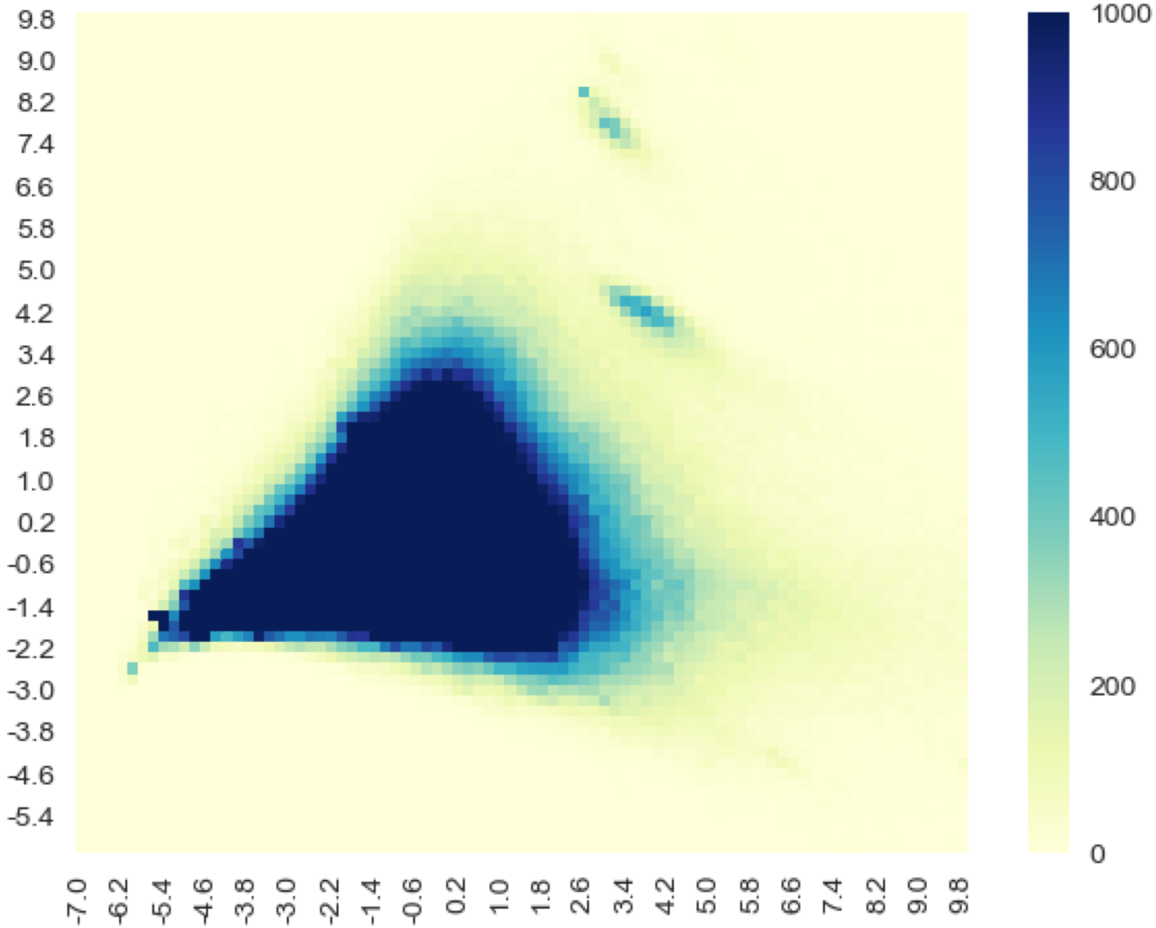
Density estimation (dot=100pgs)



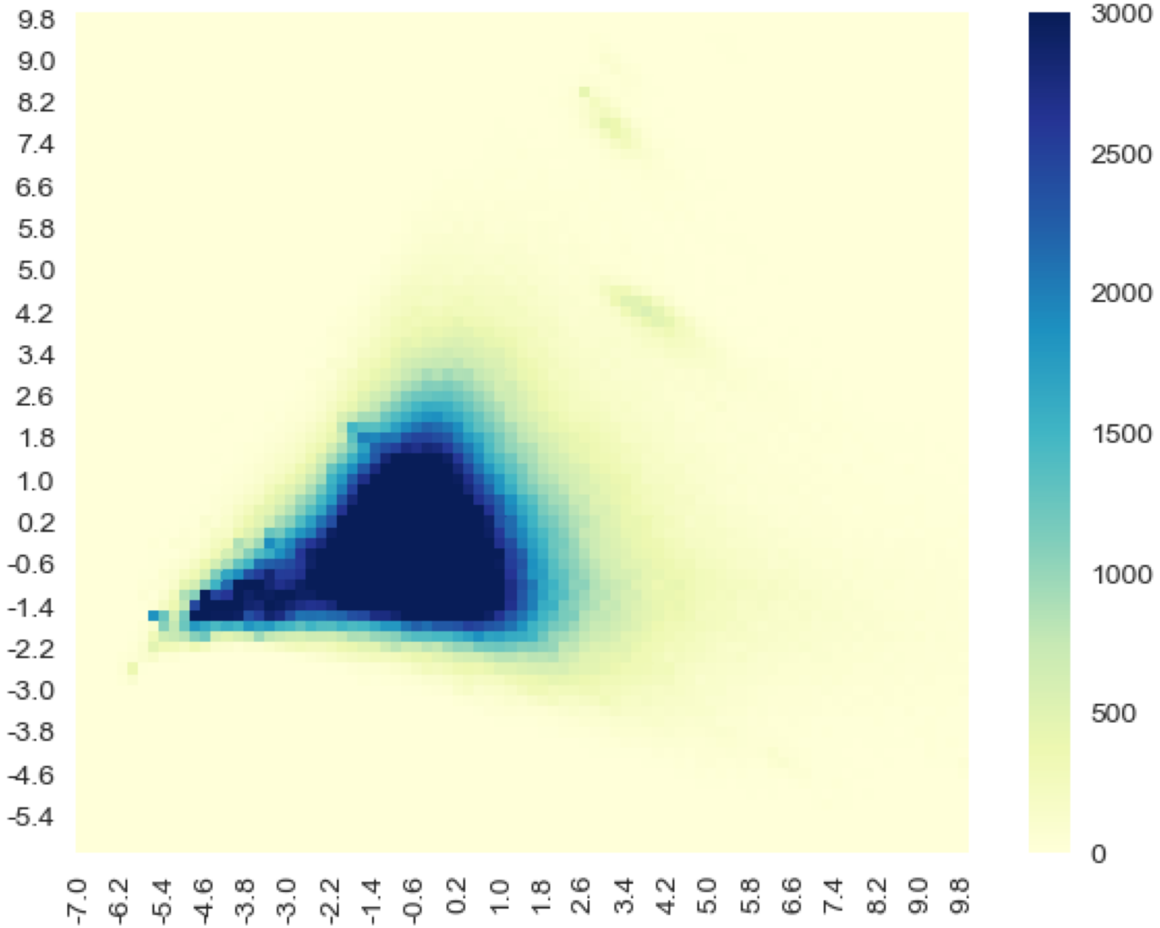
Density estimation (dot=500pgs)



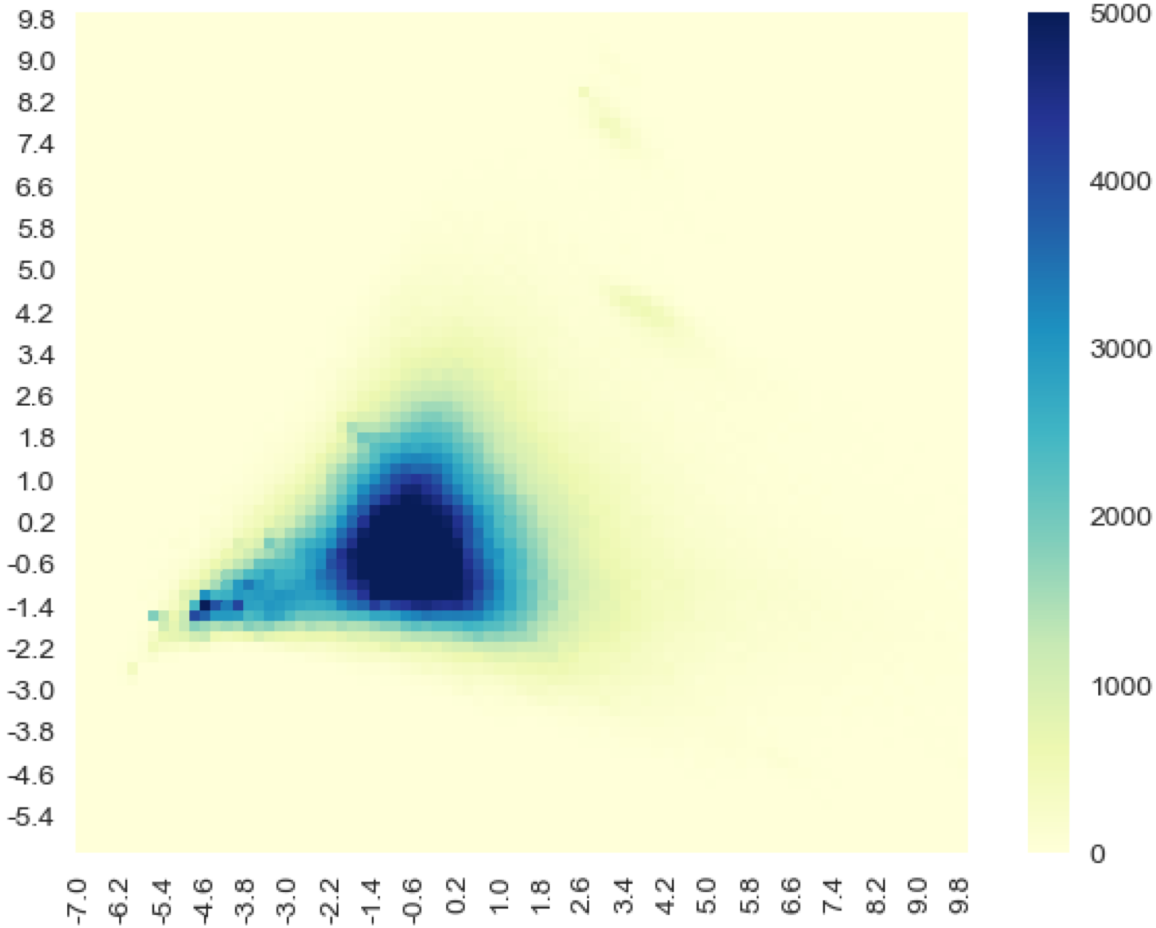
Density estimation (dot=1000pgs)



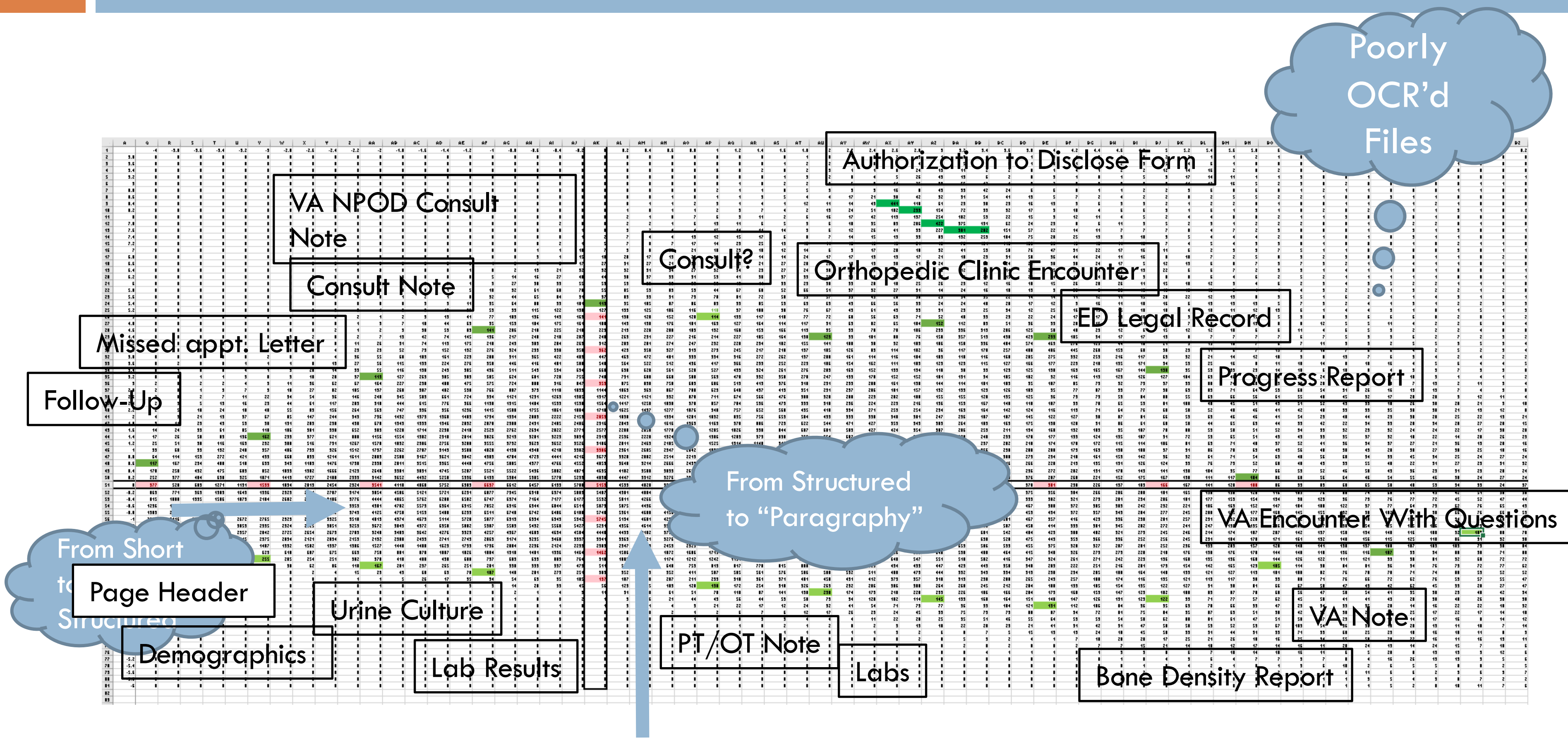
Density estimation (dot=3000pgs)



Density estimation (dot=5000pgs)



Observations from the Density Plot



Topic analysis → semantic correlations

Topic	# non-relevant pages	#relevant pages
Social/family history for mental disorder	11358	975
mental status evaluation -risk of suicide	14239	122
Mental disorder symptoms and treatment history	9900	3613
Impression of mental disorder	15500	100
Lab test results [Topic 6]	11959	2406

Planting the garden: findings

Content

Form

Structure

Technology development

Whither sublanguage analysis?

EpiBio: Heterogeneous SSA data

- Geographic variation in mental health-related documentation
  - Stigma
  - Lack of details / re-coding
- Format → structure
  - Sectionizing
  - Semi-structured forms

National Institutes of Health
Clinical Center

SOCIAL SECURITY ADMINISTRATION USA

Pitt: EHR language and health equity

- Documentation differences for patients of different races
  - What is recorded?
  - How is it recorded?
- Integrating patient-generated language with clinical observations
  - Self-reported functional status
- Ambiguity in health language

University of Pittsburgh

VA: Knowledge exchange

- Challenges shared by national health systems
  - Geographic and institutional variation
  - Large portion of SSA medical evidence comes from VA
- Cerner transition
  - Changes in documentation practice
  - Effect on NLP pipelines

VA | U.S. Department of Veterans Affairs

Acknowledgments

- Aya Zirikly (NIH)
- Guy Divita (NIH)
- Bart Desmet (NIH)
- Jona Camacho Maldonado (NIH)
- Pei-Shu Ho (NIH)
- Beth Rasch (NIH)
- Eric Fosler-Lussier (Ohio State)
- Albert Lai (WashU)

Funding support from NIH Intramural Research Program and the US Social Security Administration.

Thank you!

denis.griffis@nih.gov

dnewmangriffis@pitt.edu