Diversity Minute

But first, an announcement…

March 4, 2022

Learn to Change the World

HARVARD
GRADUATE SCHOOL OF EDUCATION

Dear Haihao,

Congratulations! I am pleased to inform you that you have been admitted to the Master of Education program in Learning Design, Innovation, and Technology at the Harvard Graduate School of Education to pursue full-time study for the 2022-2023 academic year.

You will be a part of a cohort of students who bring with them not only impressive professional experience and excellent academic training, but also dedication to the profession of education. You are to be commended on the fine record of achievement by which you earned your place in the class.

About me

- From Shenzhen, China
  - and Stamford, CT, and Hong Kong
- Rice University
  - BS in Materials Science and NanoEngineering
  - BA in Mathematics
- Diverse research experiences abroad in undergrad

Climbing Mount Fuji

Skiing in the French Alps

Tomb of Confucius

Information Extraction from Materials Science Literature:
Machine Learning Tasks and Methods

Haihao Liu

March 7, 2022

Materials Informatics

- “Fourth paradigm” after experiments, theory, simulation
- Need high-quality structured data for materials informatics
- Sparsity and scarcity key issues

Information Extraction

Input → **IE Tasks** → **Output**

Input

Unstructured natural language text

IE Tasks

Machine learning and other methods

Output

Structured data for informatics

Natural Language Processing

- Empirical linguistics since 90s
- Applied fields – chemistry, polymers, biology/medical
- Contextualized word embeddings (ELMo, BERT)

Information Extraction from Materials Science Literature

Task 1:

Sample name recognition

Task 2:

Aluminum alloy data extraction

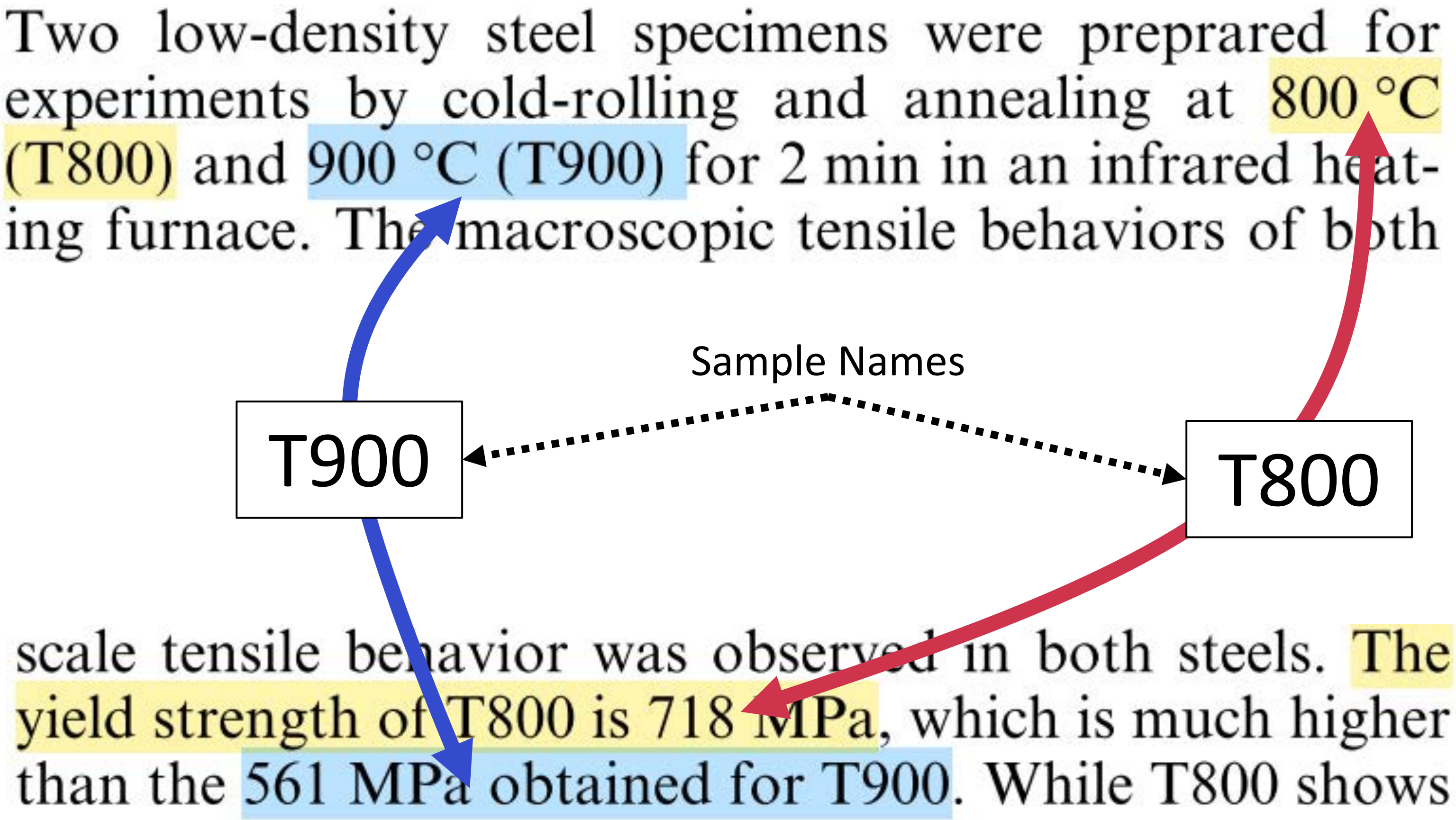
Task 3:

Key phrase ranking and analysis

Task 1:
Sample name recognition

Task Definition

Task 1



Hyperparameter Tuning

Task 1

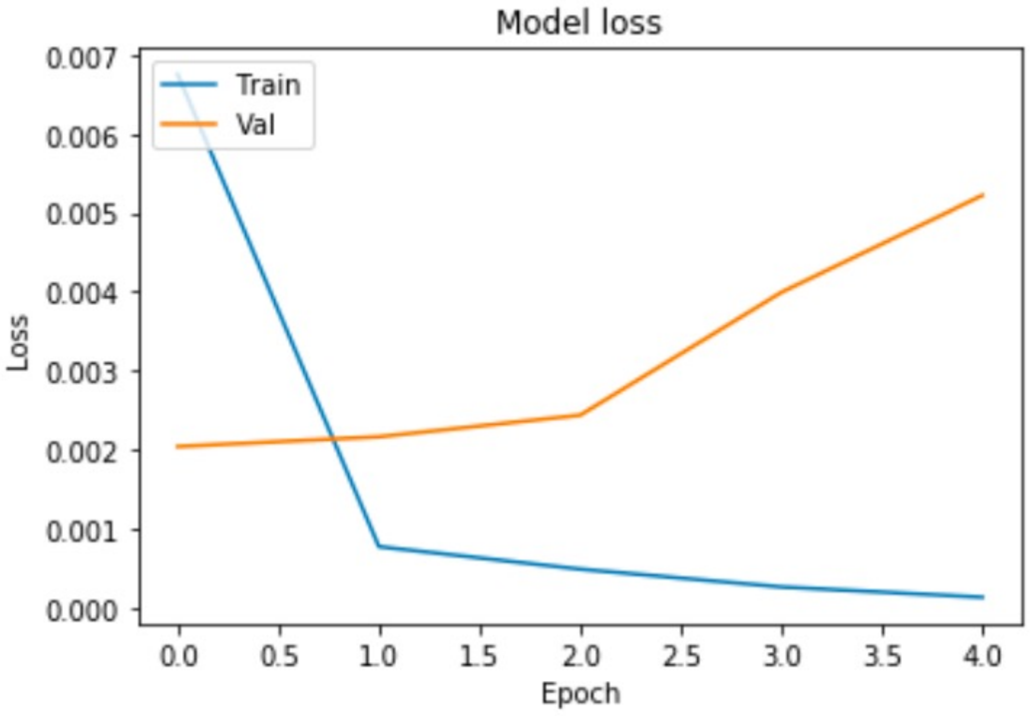
$$\mathrm{FL}(y, \hat{y}) = -\alpha \sum_{c=1}^{m} (1 - \tilde{\hat{y}}_c)^{\gamma} y_c \log(\tilde{\hat{y}}_c) \in \mathbb{R}^n$$

$\gamma \setminus lr$	10^{-7}	10^{-6}	10^{-5}	10^{-4}	10^{-3}
0.0	0.0536	**0.4857**	0.5035	0.4463	0.0
2.0	0.0929	0.4739	**0.5048**	**0.4770**	0.0
4.0	**0.2045**	0.4795	0.4926	0.4388	0.0

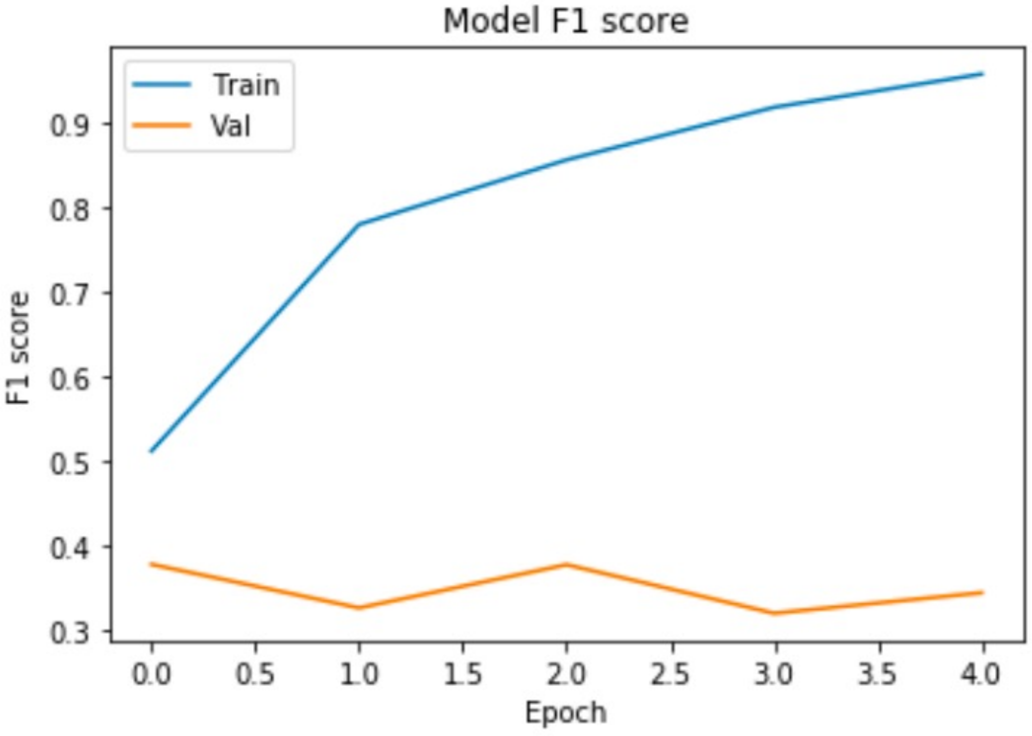
Table 3: F1 results of hyperparameter grid search

Hyperparameter Tuning

Task 1



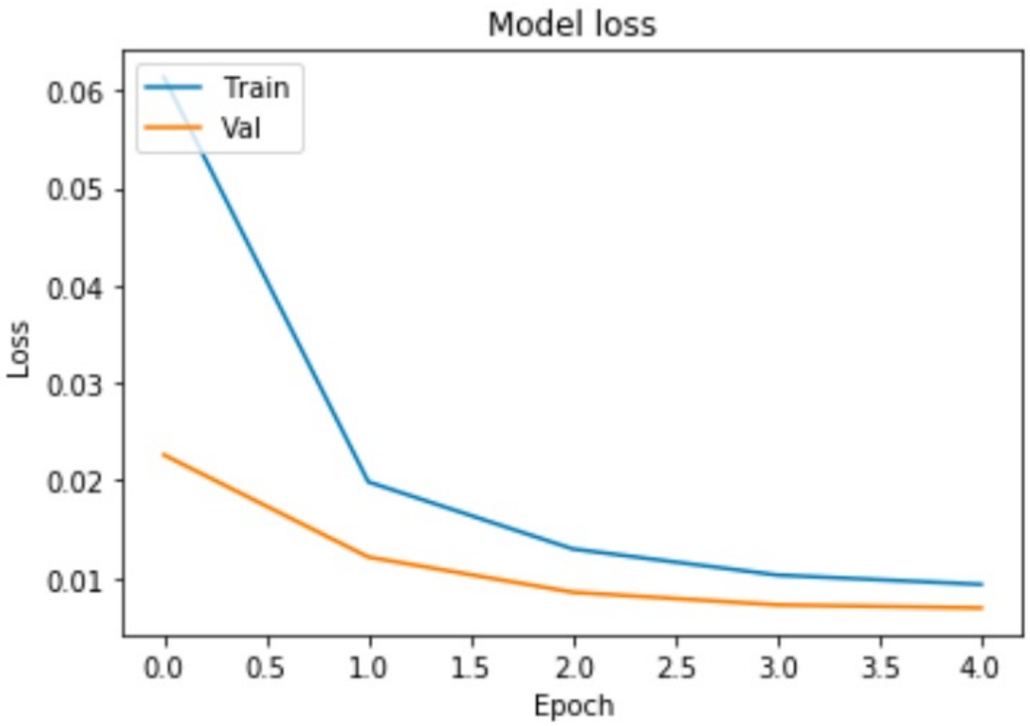
(a) Focal loss over time ($lr = 10^{-4}, \gamma = 2.0$)



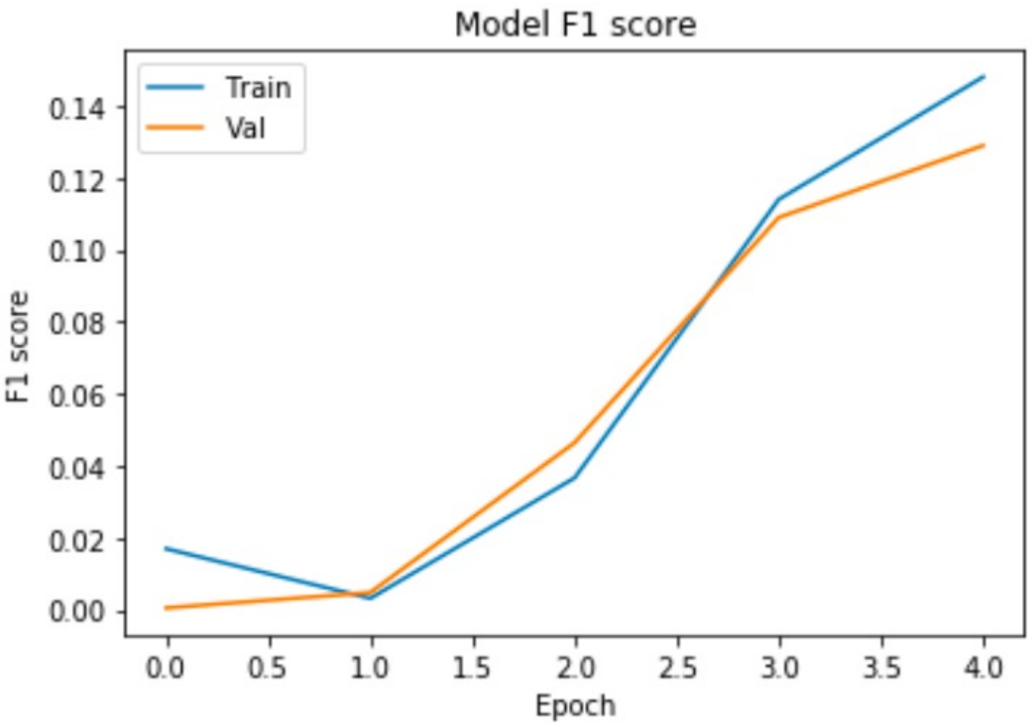
(b) F1 over time ($lr = 10^{-4}, \gamma = 2.0$)

Hyperparameter Tuning

Task 1



(c) Focal loss over time ($lr = 10^{-7}, \gamma = 4.0$)



(d) F1 over time ($lr = 10^{-7}, \gamma = 4.0$)

Model Interpretation

Task 1

y=B (probability **0.895**, score **3.429**) top features

Contribution?	Feature
+13.798	Highlighted in text (sum)
-10.369	<BIAS>

[CLS] The affinity follows the sequence : H ##60 ##0 - 0 . 5 N > H ##60 ##0 - 3 N > H ##60 ##0 - 5 N > H ##60 ##0 . [SEP]

y=B (probability **0.959**, score **4.469**) top features

Contribution?	Feature
+14.062	Highlighted in text (sum)
-9.594	<BIAS>

[CLS] All of the data were collected by adding 10 mg solid material into 10 m ##L H ##A solution , which was then kept at room temperature for 6 h . A few points worth noting are : (1) comparing Re value of PA ##C - C and PA ##C - P , it is shown that PA ##C - C exhibits a slight better performance than PA ##C - P . [SEP]

Model Interpretation

Task 1

y=O (probability **0.555**, score **0.599**) top features

Contribution?	Feature
+2.194	<BIAS>
-1.595	Highlighted in text (sum)

[CLS] After the four consecutive cat ##alytic runs , we observed a higher decrease of conversion for Amber ##ly ##st - 15 than for S ##BA - 15 - M ##w ##S and SM ##w - AG catalyst ##s , which showed just a slight conversion decrease (Fi ##g . [SEP]

y=B (probability **0.441**, score **0.052**) top features

Contribution?	Feature
+4.392	Highlighted in text (sum)
-4.340	<BIAS>

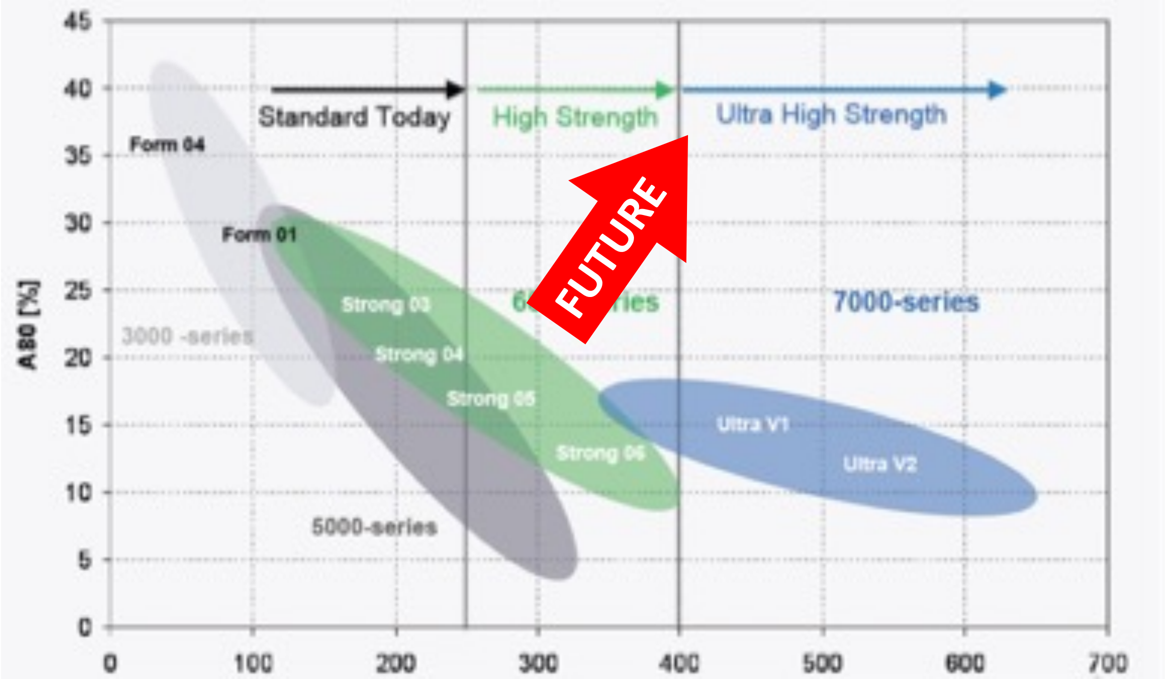
[CLS] After the four consecutive cat ##alytic runs , we observed a higher decrease of conversion for Amber ##ly ##st - 15 than for S ##BA - 15 - M ##w ##S and SM ##w - AG catalyst ##s , which showed just a slight conversion decrease (Fi ##g . [SEP]

Task 2:
Aluminum alloy data extraction

Aluminum alloy design

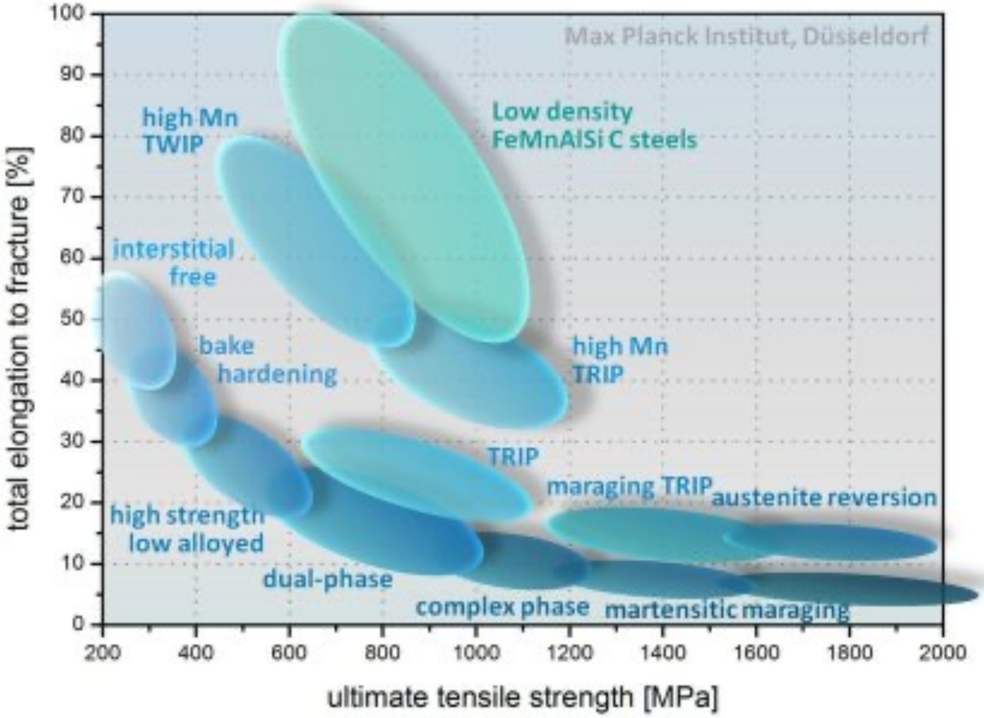
- Second-most produced metal after steel, more room to innovate

Aluminum alloys



yield strength [MPa]

Steel alloys



TWIP/TRIP: Twinning/
transformation-induced plasticity

Task Definition

Task 2

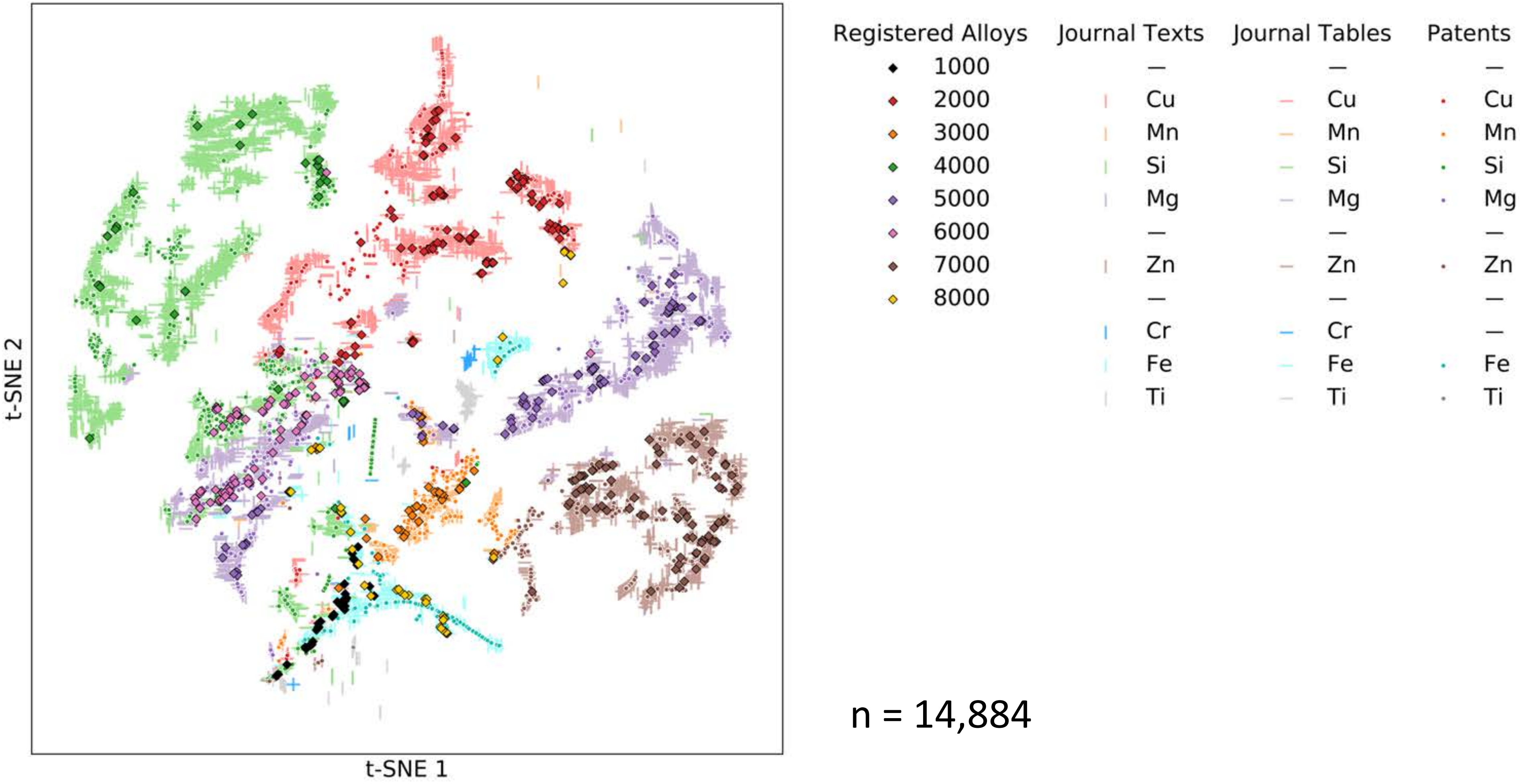
Table 1 Alloying elements in wrought aluminum and aluminum alloys

Representative list of common wrought aluminum alloys

Aluminum Association designation	Unified Numbering System (UNS) designation	Composition, maximum unless a range or minimum is specified(a), wt%							
		Si	Fe	Cu	Mn	Mg	Zn	Other specified alloying elements	Al, min or bal
1050	A91050	0.25	0.40	0.05	0.05	0.05	0.05	···	99.50
1060	A91060	0.25	0.35	0.05	0.03	0.03	0.05	···	99.60
1145	A91145	0.55 Si + Fe		0.05	0.05	0.05	0.05	···	99.45
1175	A91175	0.15 Si + Fe		0.10	0.02	0.02	0.04	···	99.75
1200	A91200	1.00 (Si + Fe)		0.05	0.05	···	0.10	···	99.0
1230	A91230	0.70 Si + Fe		0.10	0.05	0.05	0.10	···	99.30
1235	A91235	0.65 Si + Fe		0.05	0.05	0.05	0.10	···	99.35
1345	A91345	0.30	0.40	0.10	0.05	0.05	0.05	···	99.45
1350	A91350	0.10	0.40	0.05	0.01	···	···	···	99.50
2011	A92011	0.40	0.7	5.0–6.0	···	···	0.30	0.20–0.6% Bi; 0.20–0.6% Pb	bal
2014	A92014	0.50–1.2	0.7	3.9–5.0	0.40–1.2	0.20–0.8	0.25	···	bal
2017	A92017	0.20–0.8	0.7	3.5–4.5	0.40–1.0	0.40–0.8	0.25	···	bal
2018	A92018	0.9	1.0	3.5–4.5	0.20	0.45–0.9	0.25	1.7–2.3Ni	bal
2024	A92024	0.50	0.50	3.8–4.9	0.30–0.9	1.2–1.8	0.25	···	bal
2025	A92025	0.50–1.2	1.0	3.9–5.0	0.40–1.2	0.05	0.25	···	bal
2036	A92036	0.50	0.50	2.2–3.0	0.10–0.40	0.30–0.6	0.25	···	bal
2117	A92117	0.8	0.7	2.2–3.0	0.20	0.20–0.50	0.25	···	bal
2124	A92124	0.20	0.30	3.8–4.9	0.30–0.9	1.2–1.8	0.25	···	bal
2218	A92218	0.9	1.0	3.5–4.5	0.20	1.2–1.8	0.25	1.7–2.3Ni	bal

Alloy Compositions

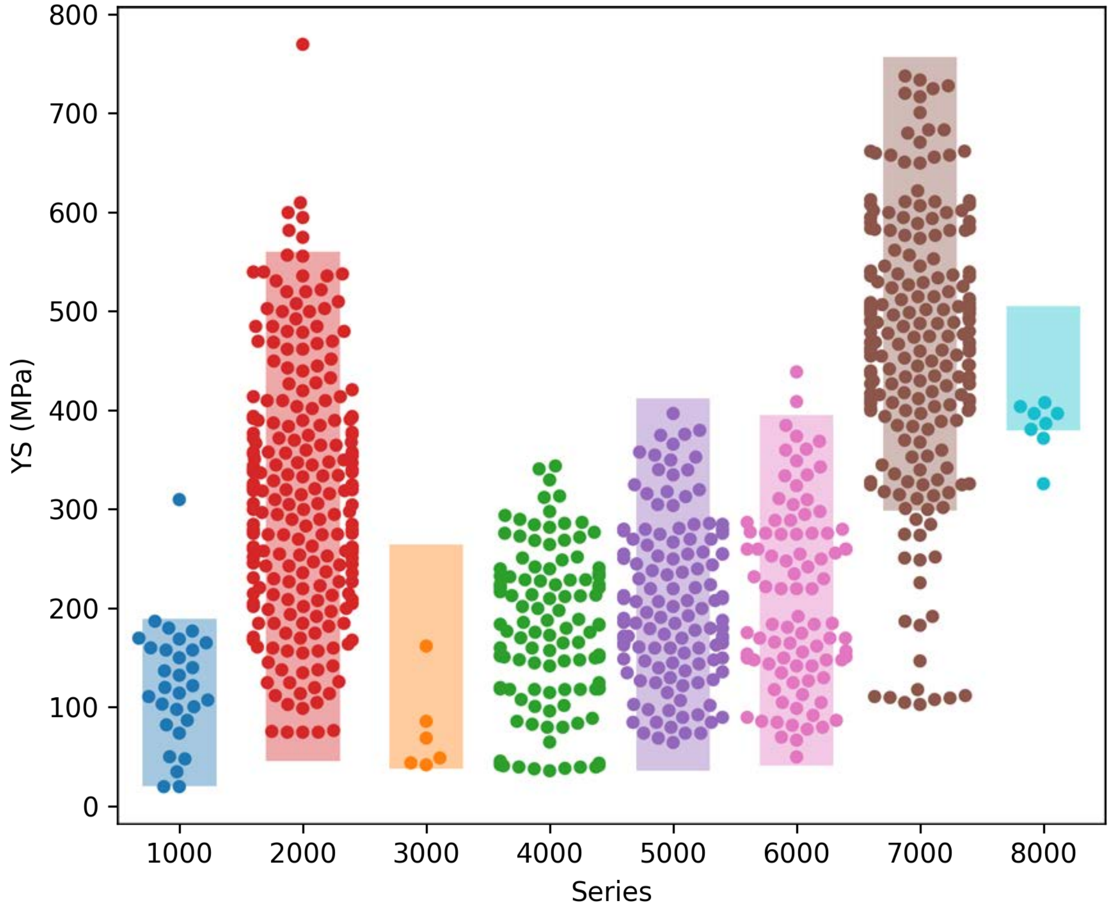
Task 2

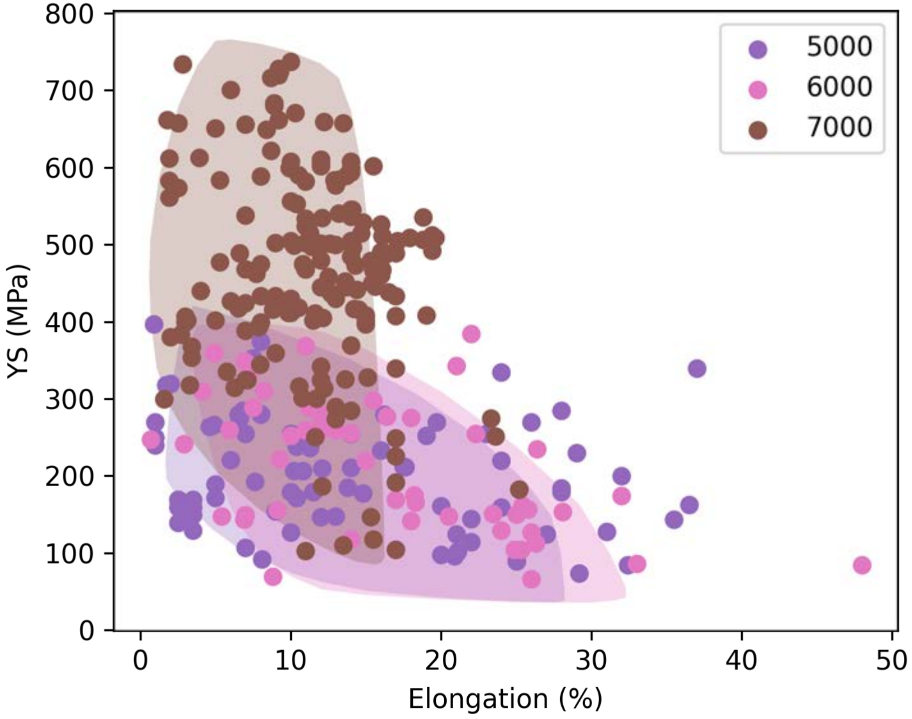


n = 14,884

Alloy Properties

Task 2





n = 1,278

Task 3:
Key phrase ranking and analysis

Top key phrases per series

Task 3

5000		6000		7000	
friction stir welding	134000	t6 aluminium alloy	86000	aluminum alloy 7075	159000
5083 aluminum alloy	113000	severe plastic deformation	85000	7050 aluminum alloy	137000
5052 aluminum alloy	95000	finite element method	82000	stress corrosion cracking	131000
ultimate tensile strength	64000	friction stir processing	65000	fatigue crack growth	88000
aluminum alloy 5083	61000	tool rotational speed	62000	ultimate tensile strength	83000
friction stir processing	53000	finite element analysis	57000	aluminium alloy 7075	79000
aluminum alloy 5052	50000	energy dispersive x	57000	aluminum alloy 7050	58000
heat affected zone	37000	6082 aluminum alloy	54000	average grain size	57000
resistance spot welding	34000	experimental results show	51000	heat affected zone	55000
finite element analysis	33000	differential scanning calorimetry	51000	t7451 aluminum alloy	53000
aluminum alloy sheet	32000	response surface methodology	48000	t6 aluminium alloy	52000
aluminium alloy 5083	32000	6061 aluminum alloys	47000	finite element method	48000
aluminum alloy sheets	30000	fatigue crack growth	44000	7075 aluminum alloys	47000
5754 aluminum alloy	30000	metal matrix composites	43000	experimental results show	43000
average grain size	29000	aa6061 aluminum alloy	40000	solution heat treatment	38000
tool rotational speed	25000	average grain size	40000	fatigue crack initiation	37000
stress corrosion cracking	24000	friction stir welded	40000	7085 aluminum alloy	36000
finite element method	24000	6061 al alloy	39000	crack growth rate	36000
strain rate sensitivity	23000	finite element model	39000	high strength aluminum alloys	35000
5182 aluminum alloy	23000	solution heat treatment	37000	strain rate range	33000
energy dispersive x	23000	aluminum alloy 6063	36000	energy dispersive x	33000
digital image correlation	23000	friction stir spot welding	35000	7075 al alloy	32000

Top studied alloys by series

Task 3

2024	977.655614
2219	201.870831
2014	87.982418
2124	41.268038
2017	37.173993
2618	32.9
2524	27.848851
2050	11
2195	9.766667
2519	9.474747
2624	8
2011	7.333333
2099	5.8
2139	5

3003	100.689273
3004	29.207618
3104	24.944444
3105	7.75

5083	427.388609
5052	316.932317
5754	149.311418
5182	101.161663
5056	20.816667
5086	16.866667
5005	15.287879
5356	11
5456	7.166667

6061	1262.18073
6063	334.403977
6082	227.730373
6016	81.378912
6060	58.515152
6111	55.923308
6013	35.199856
6022	23.55
6351	16.725
6005	15.492308
6056	14.583333
6101	11.119048

7075	1040.16121
7050	289.886416
7055	69.186334
7475	64.352704
7150	56.630952
7085	49.278571
7020	41.374747
7010	40.274359
7005	30.106061
7136	16
7175	12.271429
7449	9.733333
7003	9.383333

Scores are sums of normalized occurrences

Top concepts specific to 7000 series

Task 3

- Tempers
  - T6, T651, T7351, T7451, T76
  - cf. non-heat-treatable series e.g. 5000: H111, H116, H34, H321
- Temperatures
  - 120 °C, 470 °C, 480 °C

- Topics of study
  - stress corrosion cracking
  - fatigue crack growth
  - η′ phase
  - slow strain rate test
  - residual stress
  - creep age forming
  - high strength aluminum alloys
  - surface roughness
  - grain boundary precipitates
  - residual stress distribution

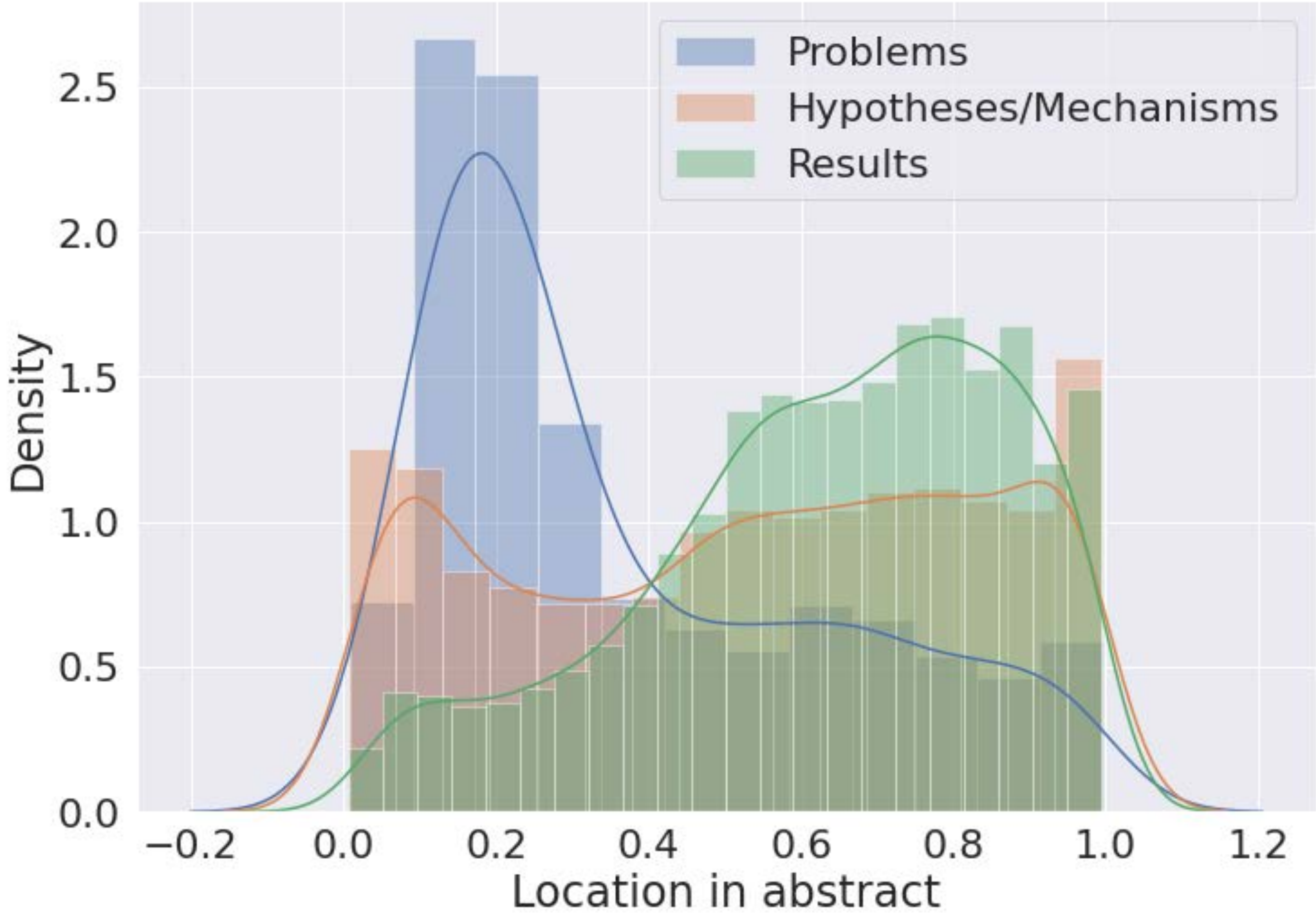
Links between keywords

Task 3

channel angular pressing	average grain size	62
laser shock peening	compressive residual stress	45
channel angular pressing	fine grain size	41
initial strain rate	grain boundary sliding	40
equal channel angular pressing	average grain size	32
channel angular pressing	initial strain rate	26
resistance spot welding	tensile shear strength	26
grain boundary sliding	high strain rate superplasticity	26
thermal expansion coefficient	metal matrix composite	26
cast aluminum alloys	dendrite arm spacing	26
initial strain rate	average grain size	23
high specific strength	metal matrix composite	23
channel angular pressing	grain boundary sliding	21
maximum tensile strength	tool rotational speed	21
response surface methodology	metal matrix composite	21
	tool rotational speed	21
channel angular pressing	high strain rate superplasticity	21
compressive residual stress	fatigue crack growth rate	20
continuous dynamic recrystallization	average grain size	19
secondary dendrite arm spacing	cast aluminum alloys	19
channel angular pressing	high angle grain boundaries	18
	continuous dynamic recrystallization	18
fatigue crack growth rate	high strength al	18

Inferring relationships between links

Task 3



Conclusions and Outlook

- Three information extraction tasks
  - Sample name recognition, Al alloy data extraction, key phrase analysis

- NLP still young in its application to mat sci, many challenges remain

- To make most use of data, need to incorporate domain knowledge

Acknowledgments

Questions?