eXtremeMAT National Laboratory Consortium: Data Science and Analytics

Ram Devanathan, Jennifer R. Bauer, Gary D. Black, Michael Glazoff, Turab Lookman, Pratik Ray, Vyacheslav Romanov, Kelly Rose, Michael D. Sabbatino, Arun Sathanur, Dongwon Shin, Madison Wenzlick, Yuki Yamamoto, and Jeffrey A. Hawk.

2019 Crosscutting Research Meeting, Pittsburgh, PA
April 10, 2019
https://edx.netl.doe.gov/extrememat/

Acknowledgement:

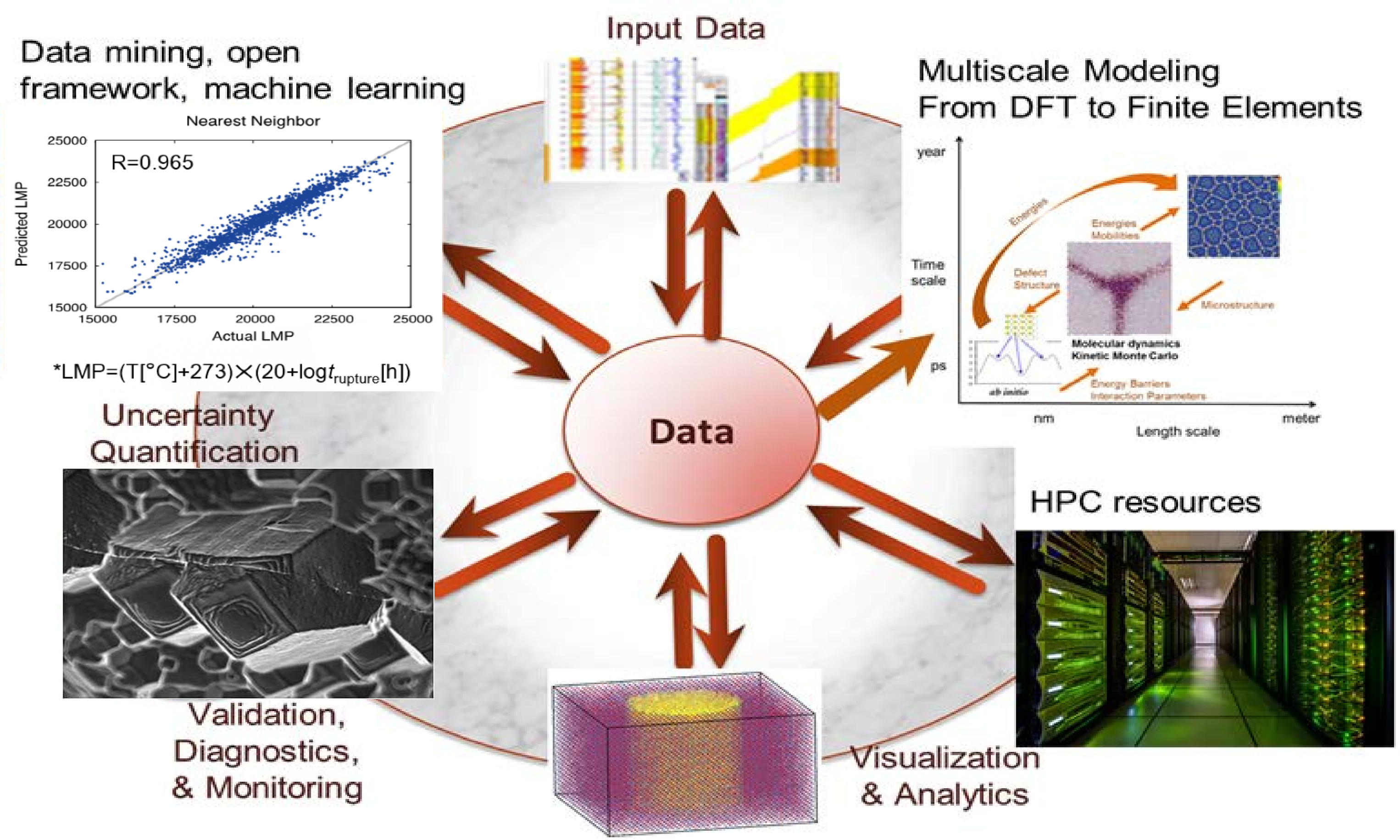
This work was supported by the NETL Crosscutting Research Program, Briggs White, NETL Technology Manager, and Regis Conrad, DOE-FE HQ Program Manager. This work was executed through the eXtremeMAT National Laboratory Field Work Proposal (NETL: FWP-1022433, LANL: FWP-FE85017FY17, ORNL: FWP-FEAA134, Ames: FWP-AL-17-510091, LLNL: FWP-FEW0234, INL: FWP-B000-17016, PNNL: FWP-7113).

Disclaimer:

Goal: Use data science to reduce the time and cost for alloy development and lifetime prediction.

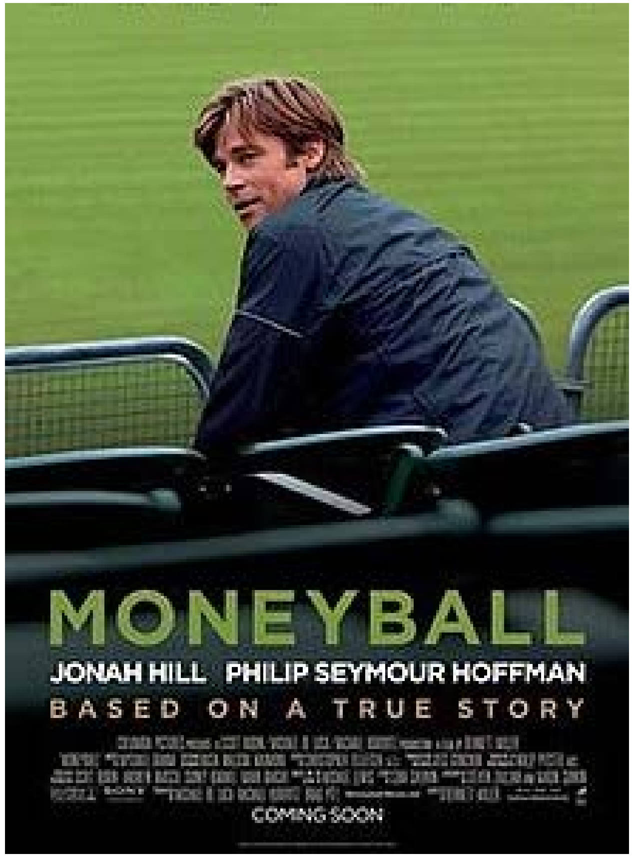
Modeling Task

Experimental Validation Task



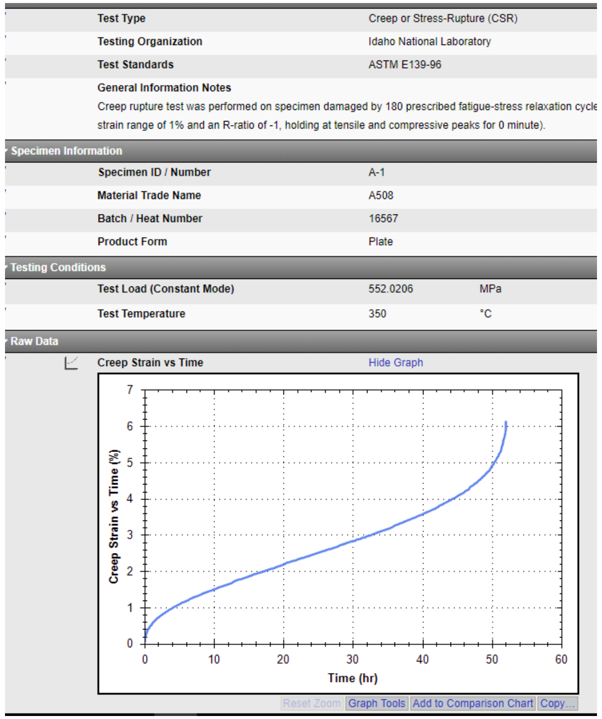
Application of data science to materials is more challenging than its use in baseball.

Issues	Baseball	Materials Science
Data availability	Reliable and readily available.	Scattered, hidden (proprietary), or unpublished.
Standards	Uniform	No data standards
Variability	Unambiguous: Hits, RBI, HR etc.	Uncertainty can be large; variability between groups; different geometries
Relations	Rules, positions etc. well defined.	Processing-microstructure-property relations nonexistent.
Stability	Rules don't change during the season	Material microstructure evolves during processing and service; surface degrades; material oxidizes or corrodes.



Where is the relevant data?

- Journal articles
- Online and printed databases
- National Institute for Materials Science (NIMS) database
- Government technical reports
- National lab databases
- Industry reports
- US Patents
- Offline: Lab notebooks and as knowledge in experts' minds



Successes are published; failures are forgotten.

Alloy development challenges

- Data is sparse and expensive to obtain.
- Metadata and provenance are missing.
- Existing frameworks focus on 0 K properties.
- Complex descriptors (e.g., microstructure)
- Going beyond confirmation of one's biases
- Interpretability of models is essential.

Alloy innovation is knowledge driven.
Knowledge is difficult to curate, collect, and contextualize for efficient use in development.
Knowledge is even less structured and often harder to obtain than data itself.

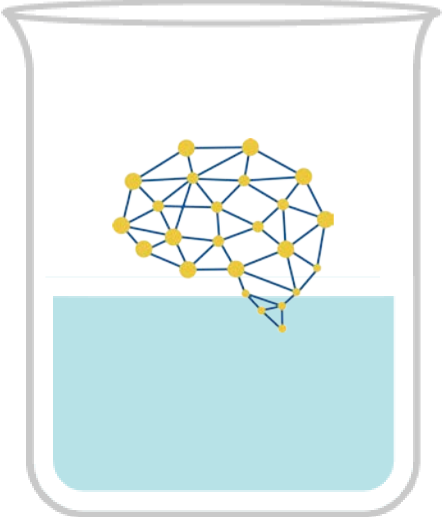
Filling the DOE FE Knowledge Container: *Using AI/ML to put knowledge to work*

Why Knowledge Management Matters:

- **Millions** in lost productivity annually because of **inefficient** knowledge sharing
- 42% of institutional knowledge is acquired specifically for an employee's current role and is **not shared by any of their coworkers**

Source: Panopto Workplace Knowledge and Productivity Report, 2018

Ongoing **development of "smart" tools** to mine and extract knowledge from users' posts, search and interaction patterns, and efficiently **connect users to personalized content**

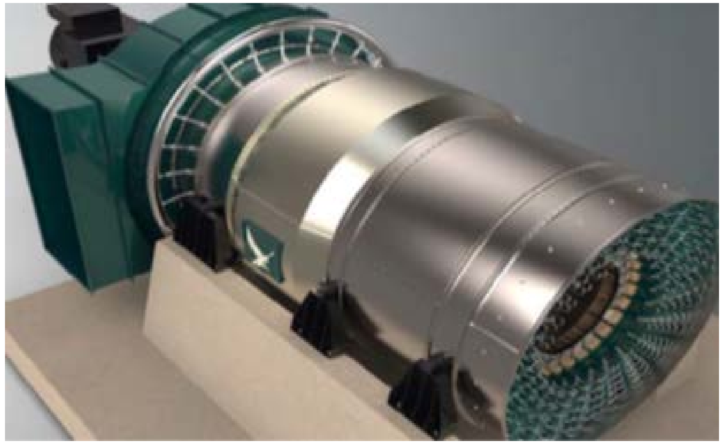


eXtremeMAT

- **Share your materials knowledge & expertise**
- For eXtremeMAT we want to jump start knowledge collection, curation and utilization
- **Wednesday & Thursday, sign up for a 15 minute slot (at registration desk)**
  - **Slots between noon and 2pm, Indiana Room**
- eXtremeMAT **experts can meet with younger scientists or program managers** in facilitated conversations
- These will serve as test-case for a new knowledge management effort for XMAT goals

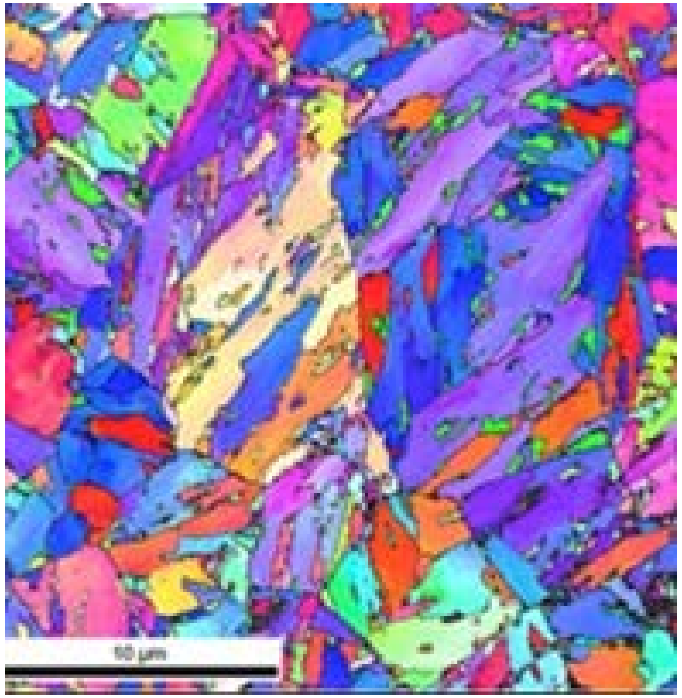
Gap analysis: Data Science for Extreme Environments

- Lack of composition-processing-structure-property relations
- Lack of validated physics-based creep and processing models
- Gaps in microstructural and thermodynamic data
- Alloy development knowledge and data are not widely shared
- Parameter space is large (minor alloying element optimization)
- Extrapolation from short-term tests to long term life estimation
- Relying on experiment alone would be very long and expensive



What is needed to advance this field?

- Processing data (time, temperature etc.)
- Characterization results (microstructure)
- Tensile, fatigue, fracture toughness, corrosion and creep results
- Guidance on features we should focus on
  - **Grain size, grain shape, voids, phase volume fraction, Cr content etc.**
- Information on what has worked well and the dead-ends
- Data from failed experiments
- Mechanism to anonymize and protect shared data.



eXtremeMAT

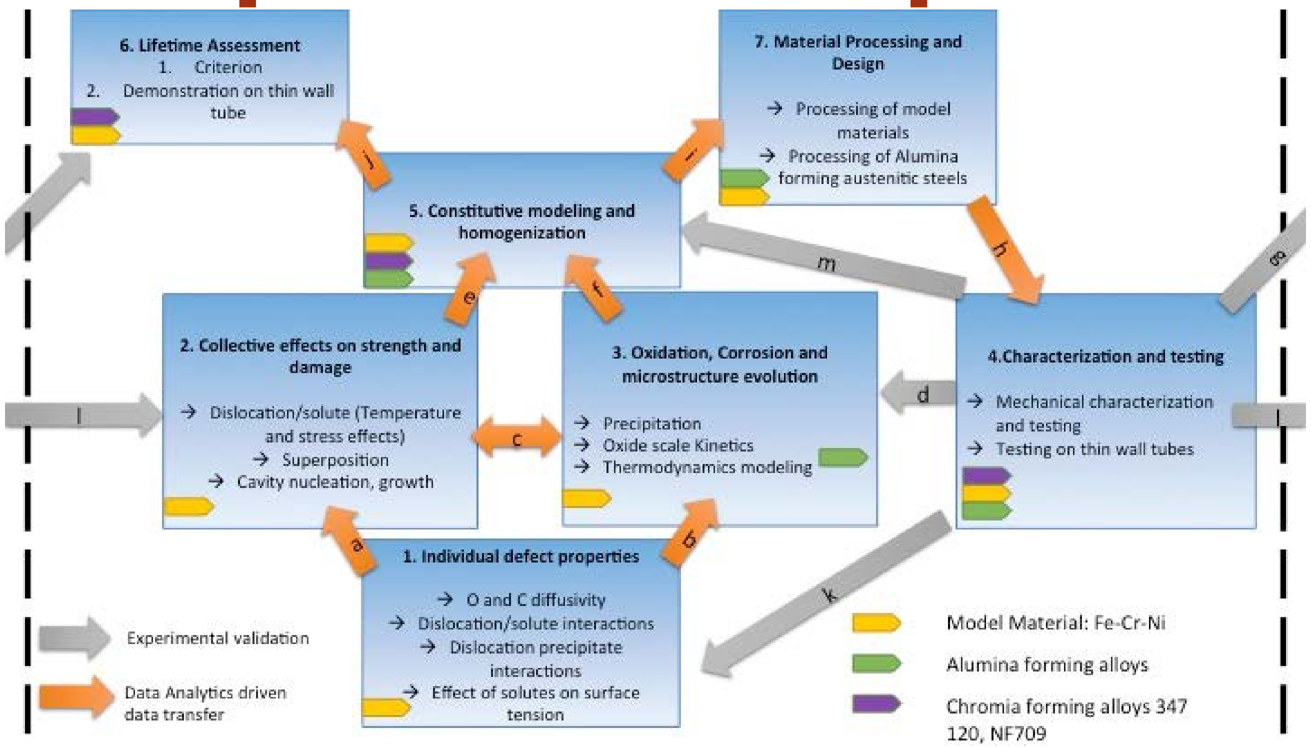
Accelerating the Development of Extreme Environment Materials

U.S. DEPARTMENT OF ENERGY

Life Time & Performance

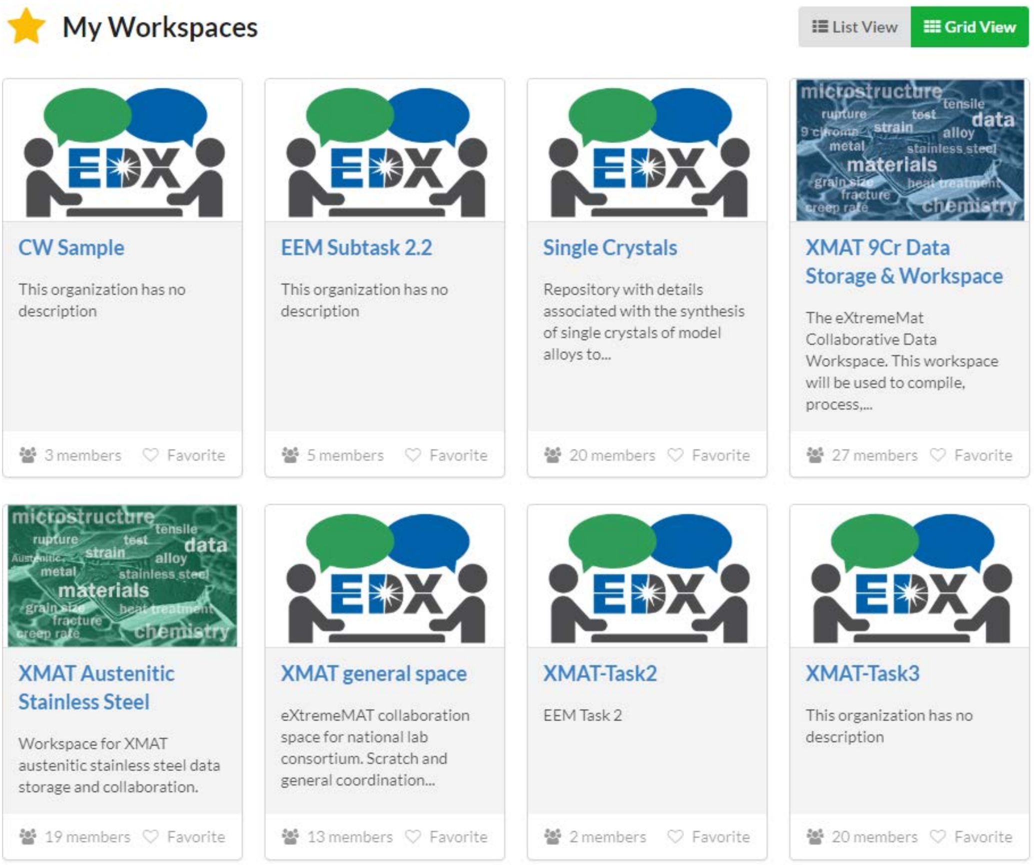
Alloy Design

Data (orange arrows) serve as the glue between model scales



Data Science includes three subtasks: data management, assessment, and analysis.

Data Management



Multiple Workspaces

- Data is being uploaded
- Metadata included
- Data assessed during upload
- Collaboration tools

EDX Workspaces

https://edx.netl.doe.gov

Example Data Set

Alloy (wt%)	N	C	Mn	Cr	Mo	Ni	Si	S	P	Fe	Fe	Grain size (micron)	Vickers Hardness (HV)
316LNSS-7N	0.07	0.027	1.7	17.53	2.49	12.2	0.22	0.0055	0.013	65.7445	Bal	87±9	155
316LNSS-11N	0.11	0.033	1.78	17.62	2.51	12.27	0.21	0.0055	0.015	65.4465	Bal	96±8	160
316LNSS-14N	0.14	0.025	1.74	17.57	2.53	12.15	0.2	0.0041	0.017	65.6239	Bal	78±8	166
316LNSS-22N	0.22	0.028	1.7	17.57	2.54	12.36	0.2	0.0055	0.018	65.3585	Bal	87±11	192

Rupture life, hrs	Applied Stress, Mpa	alloy
45.37404794	224.945114	0.07 N
121.6926012	199.8908445	0.07 N
745.4927337	175.31095	0.07 N
2732.443324	140.0087076	0.07 N
3193.802023	140.0115799	0.07 N
9155.291556	119.9832143	0.07 N
134.34722	225.0459867	0.11 N
346.2020457	200.0314365	0.11 N
1499.608062	174.8226216	0.11 N
7215.026762	140.1482481	0.11 N
158.5269254	225.0081291	0.14 N
539.503119	199.9579453	0.14 N
1422.457775	175.1887016	0.14 N
9741.009575	140.1535061	0.14 N
358.6704287	224.9410283	0.22 N
1270.865196	199.8915725	0.22 N
4426.64183	174.9234135	0.22 N
16126.39218	139.9179054	0.22 N

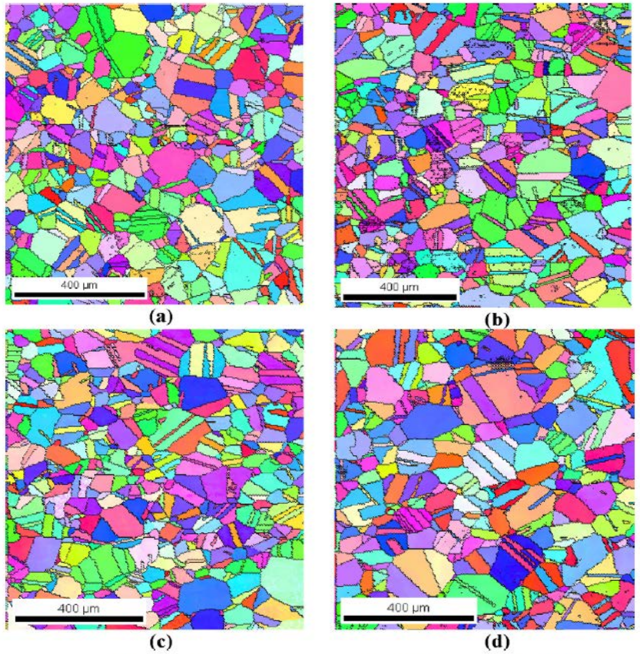


Fig. 1. Orientation imaging micrographs of solution annealed 316LN SS containing nitrogen (wt.%) of (a) 0.07, (b) 0.11, (c) 0.14 and (d) 0.22 N. Nearly equiaxed grains and annealing twins have been observed.

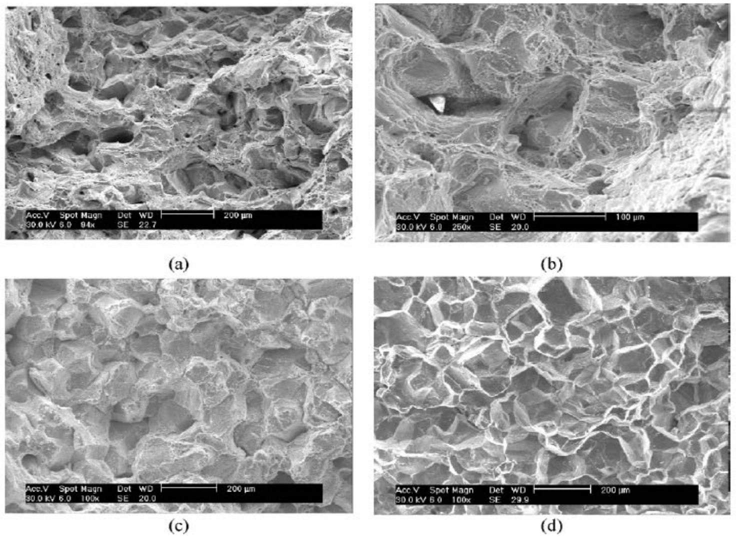
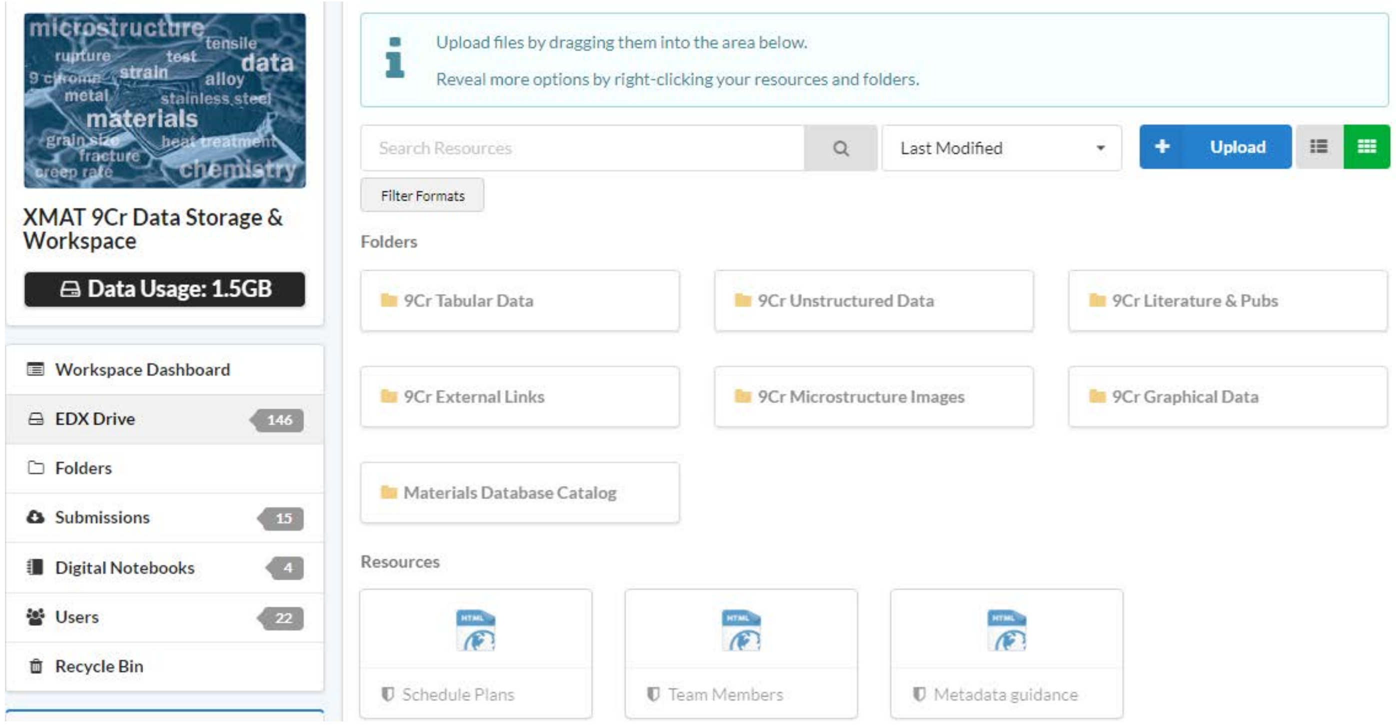


Fig. 12. SEM micrographs showing fracture surface of the creep tested specimens. (a) 0.07 wt.% nitrogen containing steel; creep rupture life = 120 h. (b) 0.11 wt.% nitrogen containing steel; creep rupture life = 344 h. (c) wt.% nitrogen containing steel; creep rupture life = 537 h. (d) 0.22 wt.% nitrogen containing steel; creep rupture life = 1262 h.

EDX: Consistent Folder Structure and Digital Notebooks



Task coordination plan for quick win, 11/9/2018 (RD)

We need to work toward a proof-of-concept "win" that can be shown by April 2019. The Quick win is **early demonstration of using data analytics to predict alloy lifetime from our compiled database.** *The caveat is that our database is not yet fully populated and may contain gaps that need to be filled with new experiments.* This prediction will improve as we obtain more data, especially high quality/reliable data. While we work to get this win, we should also devote efforts to the following:

1. Develop and populate a data infrastructure for data acquisition and storage, and to enable data analytics to lay the foundation for the data science effort.
2. Develop a data quality metric that can be given to universities funded by FE to study creep.
3. Populate creep data in the database with metadata and data quality metric.

Kelly Rose's notes from 11/16/2018

December: by end of December accomplish the following

- **Fill EDX with data sources**
- Madison and our crew will have an EDX workspace ready for 9 Chrome data ingestion by the week of 11/26. We're hoping to walk you through that space and enlist your help in sending out the request for folks to drop things into the space sometime that week.
- **Define database structure, attributes, units of measure etc.** (tasks 2 and 3 teams need to talk)
- We'll need to coordinate between PNNL (Gary et al) and our group on execution responsibilities and strategies here. I could see NETL and PNNL doing data mining and conversion in parallel or in series....there are benefits either way. We'll need a discussion to evaluate and decide how to proceed.
- We'll need to talk with Task 2 folks on what they want from their initial databases to meet their analytical goals. Laurent is sending you a note for a dedicated tasks 2-3 meeting for this asap.

January:

- Convert data using OCR, NLP, and other methods
- Integrate converted data into the database structure(s)

February:

- Modelers take over and use data
- Anticipate Task 3 data personnel will be asked to assist with any needs for updates or changes to the database

March

- more analytics, more data mining

Jen Bauer's draft on data quality, 11/27/2018

Data Quality Critical metrics

Overall Data Quality

- Overall metric, that is representative of the dataset based off the characteristics below. In the future, this could be a summary of the 4 metrics below (where a user can still modify)
- Definition: indication of whether data is complete, credible, accurate, precise, accessible and usable
- RANKINGS

RuptureLife-316SSLN-N-addition....

Edit Rate Download

Not Specified

Source: MD Mathew et al, Improving creep strength of 316L stainless steel by alloying with nitrogen, MSE A 535 (2012) 76-83.

Creep tests conducted at temperature of 923 K, 650±2 C according to ASTM E-139 standard; Solid solution strengthening observed with increasing N; No error bars in data; Composition given in wt.% is quite thorough.

Stacking fault energy was evaluated using the following relationship: SFE (from composition wt%) = 34 + 1.4Ni -1.1 Cr -77N; Increase in N decreases SFE and decreases grain boundary energy (more stable boundary; higher creep strength); Orientation imaging microscopy was used to measure grain misorientation and from that derive effective GB energy.

This is 5* data It is complete: has composition, creep life vs stress, stacking fault energy and grain size; Microstructure is available; SEM and TEM images and fracture surfaces; Processing details: Air induction melting; electroslag refining, forged, rolled into 22 mm thick plates, solution annealed at 1323 K, 1050 C. Data is from a foreign government lab;

Data extracted by Ashley Weber, America Sevilla and Ram Devanathan, 2018

Data Quality Review Results:

Audience	Overall	Completeness	Source	Accuracy	Accessibility
Owner	4.75 (4)	5.0 (1)	5.0 (1)	5.0 (1)	4.0 (1)

Inclusion of Metadata with each data set is essential.
The availability of metadata with literature data is highly variable.

XMAT has identified metadata that needs to be included with mechanical testing data.

- Composition (primary alloying elements, trace elements and impurities)
- Processing conditions
- Heat treatment conditions
- Microstructural details and phase distribution
- Testing standards followed
- Measurement conditions
- Sample Geometry
- Environmental conditions

XMAT has developed a metadata template that will be given to data generators.

Column Header	Definition	Comments & Guidelines
ID	ID according to source	Some or all information may be available. If available, then include. If not, then leave blank. This data applies to all materials and samples regardless of test type.
Alloy	Alloy type	
Heat	Heat number according to source	
Specimen No.	Specimen Number according to source	
Manufacturer	Material manufacturer/ vendor	
Fe	Element iron (Fe)	For chemistry include everything that is available, main alloys reported to 0.01 wt.%. In many instances this will only include the major elements for the alloy and some of the minor ones. Usually, for manuscripts elements like oxygen, sulfur, nitrogen as well as trace elements are omitted. If additional elements not included on this list exist then a column should be inserted appropriately for that element and noted. Include uncertainty in measurements when available. This data applies to all materials and samples regardless of test type.
C	Element carbon (C)	
Cr	Element chromium (Cr)	
Mn	Element manganese (Mn)	
Si	Element silicon (Si)	
Ni	Element nickel (Ni)	
Co	Element cobalt (Co)	
Mo	Element molybdenum (Mo)	
W	Element tungsten (W)	
Nb	Element niobium (Nb)	
Al	Element aluminum (Al)	
P	Element phosphorous (P)	
Cu	Element copper (Cu)	
Ti	Element titanium (Ti)	
Ta	Element tantalum (Ta)	
Hf	Element hafnium (Hf)	
Re	Element rhenium (Re)	
V	Element vanadium (V)	
B	Element boron (B)	
N	Element nitrogen (N)	
O	Element oxygen (O)	
S	Element sulfur (S)	
Ce	Element cerium (Ce)	
Meas. Method	Method of elemental composition measurement (such as ICP MS) and/or source of measurement (ie, vendor)	
Equip.	Equipment type used for measurement	
Standard	Standards followed, if any (ie, ASTM)	
Sample Loc.	Location from which specimen was taken from the casting or sample	
Proc. Method	Processing method: casting, rolling, forging, etc.	
Shape	Geometry of sample (bar, plate, etc)	
Notch	Notch type: V notch, rounded, etc	

Standards for assessment of data quality.

- XMAT adopted its own standard.
  - There is no universal data quality standard.
- **Completeness** of data; assess importance of missing information
  - **Ratio of available data/expected data**
  - **Identify what information is absolutely essential**
- **Source**: peer reviewed academic vs. data from unreliable source (blog). Standards?
- **Precision**: Does data have a lot of inconsistencies, gaps, obvious errors? Derivative data?
- **Accessibility**: Is the data easy to extract (bad quality PDF?)
- Tool to assess data quality is already deployed in EDX!
- **Multiple XMAT participants must assess data independently and arrive at a metric.**

Overall	Completeness	Source	Accuracy	Accessibility
4.75 (4)	5.0 (1)	5.0 (1)	5.0 (1)	4.0 (1)

Data Standards: Completeness

Level	Description
5	All material manufacturing and processing data are available • Exact chemical composition • Material production process: material manufacturer, primary melt process, deoxidation practice, secondary melt process, ingot/continuous casting, hot/cold working parameters. • Product form and characteristic dimension • Heat treatment sequence, temperature, time, and cooling medium. • Test environment for corrosion/hot oxidation studies. • Microstructure (grain size as minimum)
4	Relative to Level 5, microstructure, or the product form/dimensions are not provided
3	Relative to Level 4, material production process or heat treatment is incomplete/not provided
2	Relative to Level 3, chemical composition is not complete or given as nominal values only
1	Only indication of the name of the material is reported.

Data Standards: Precision

Level	Description
5	All data provided as exact values. All data is experimentally obtained. Document is of high resolution, so values are not required to be guessed.
4	Most data are provided in exact values, with minimal guesswork due to poor resolution.
3	Majority of data is provided in exact values with some data reported graphically.
2	Minimum of data is provided in exact values, with most data reported graphically.
1	No data provided with exact values, all data is provided in graphs. Or resolution quality is low, and most values are difficult to read. Or, all data in derived or averaged values.

Data Standards: Accessibility

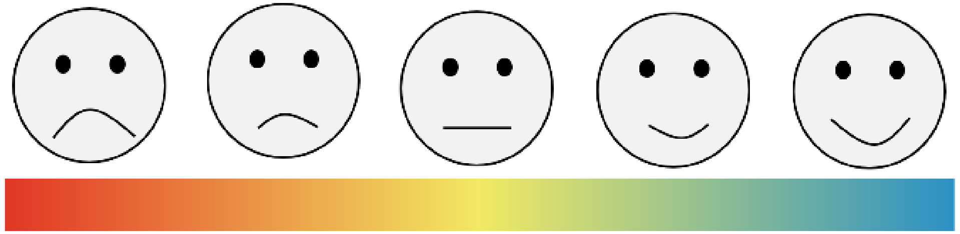
Level	Description
5	Data in original source is organized in digital, tabular form, either in text-based files, data sheets, or in a database. Location of data is intuitive. Data is easy to extract and manipulate.
4	Relative to 5, some data manipulation is needed (e.g., data in multiple files).
3	Data located in non-standard or hard to read tables. Or data requires OCR for extraction.
2	Data is located in physical copy only, requiring scanning to digitize.
1	Data requires 100% manual extraction.

Data Standards: Source

Level	Description
5	Material testing organization and/or manufacturer are both accredited. Testing and measurement performed according to an accepted national standard.
4	Relative to Level 5, accreditation is unknown.
3	Relative to Level 5, standard of testing and measurement are unknown.
2	Accreditation and standard of testing and measurement are unknown.
1	Testing and measurement are non-standard. Testing organization is not accredited.

Data Quality Assessment

Ranking system implemented in EDX



1	2	3	4	5
Red	Orange	Yellow	Green	Blue
Not confident at all	Somewhat confident	Neutral	Confident	Very confident

Poor Data — Mediocre — Good Data

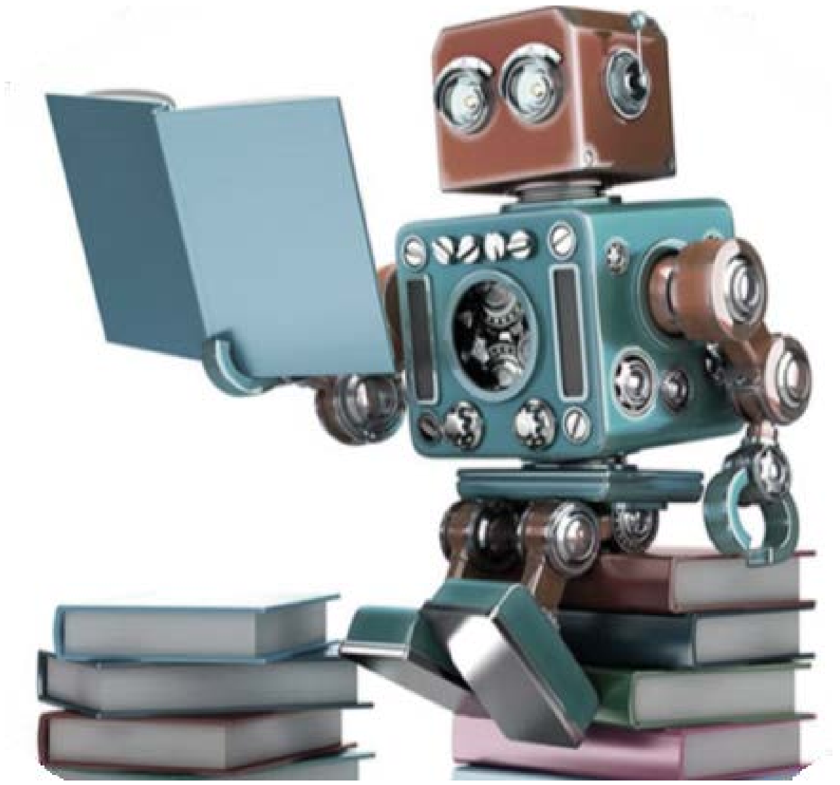
Examples of data ranking

Overall	Completeness	Source	Accuracy	Accessibility
4.75 (7)	5.0 (2)	5.0 (2)	5.0 (1)	4.0 (2)

Overall	Completeness	Source	Accuracy	Accessibility
2.0 (4)	2.0 (1)	4.0 (1)	1.0 (1)	1.0 (1)

Data Analytics/Machine Learning

- Overall Objective: Identify the main drivers of mechanical degradation in extreme environments and design new Fe-based alloys that can perform at a 50 °C higher temperature
- Desired Outcomes
  - Couple machine learning (ML) with data framework
  - Integrate data-driven and physics-based models
  - Predict the lifetime of chromia-forming alloys
  - Design new alumina-forming alloy
- Analyzed F-M 9-12% Cr steel data
  - Data from NETL, GE, NIMS and literature



Typical Creep Data Set for 9-12% Cr F-M steels

Table 1. Cleaned-up tensile & creep data set: 47 columns & 2800 rows; captures test outcomes, alloying elements, heat treatment, average grain size, and tensile or creep test outcome

ID	Fe	C	Cr	Mn	Si	Ni	Co	Mo	W	Nb	Al	P	Cu	Ta	V	B	N	O	S	…
1	85.44	0.15	9.833	0.406	0.087	0.27	1.479	1.262	0.482	0.056	0.016	0.001	0.003	0.279	0.209	0.01	0.02	0.0036	0.005	
2	85.645	0.15	9.812	0.288	0.15	0.217	1.529	1.464	0.428	0.059	0.005	0.003	0.003	0.202	0.204	0.0078	0.0252	0.0029	0.0058	
3	85.351	0.16	9.953	0.473	0.111	0.224	1.588	1.342	0.493	0.061	0.009	0.003	0.003	0.198	0.194	0.0086	0.0222	0.0034	0.0055	

Homogenization, Normalization & Tempering

Creep Properties : Rupture Time(RT), Min Creep Rate (MCR), Elongation, Reduction of Area @ Test Temp & Stress (CS)

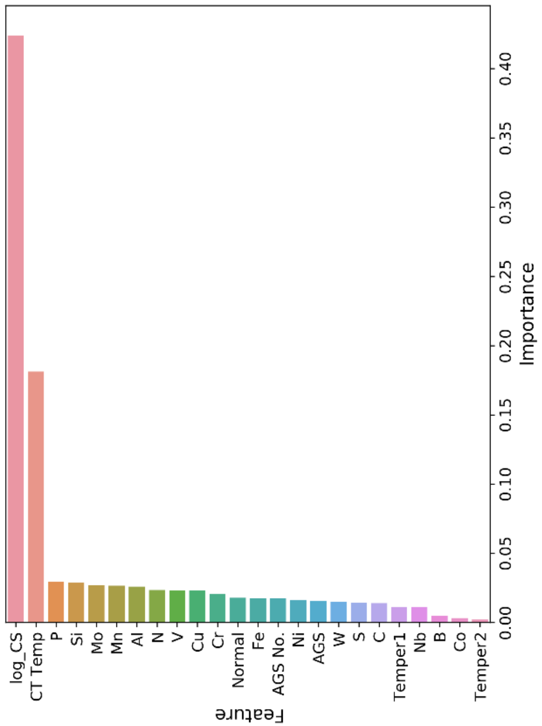
Homo	Normal	Temp1	Temp2	Temp3	AGS#	AGS	TT Temp	YS	UTS	Elong	RA	CT Temp	CS	RT	EL	RA	MCR	0.1% CS	0.2% CS
0	1045	780			10.1	8780	700	124	199	36	98								
0	1045	780			10.1	8780	750	78	141	55	99								
0	1045	780			10.1	8780						550	240	498.8	26	90	0.008	23.8	179.7
0	1045	780			10.1	8780						550	220	2635	26	90	0.002	110.9	760.3
1	1150	700					650	390	452	39	85								
1	1150	700										650	206.8	665.8		84			

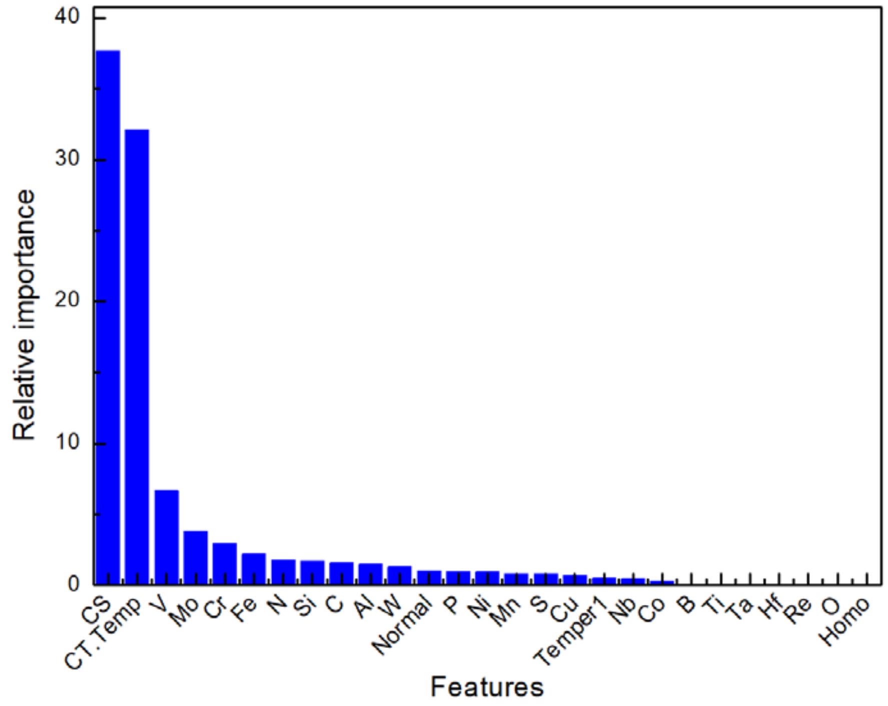
Average Prior Austenite Size estimated from Lath boundary of Martensitic Steel

Tensile Properties : Yield Strength(YS), Ultimate Yield Strength(UTS), Elongation, Reduction of Area @ Test Temp

NETL-PUB-21550

Ranking of features affecting creep rupture time of 9-12% Cr F-M steels



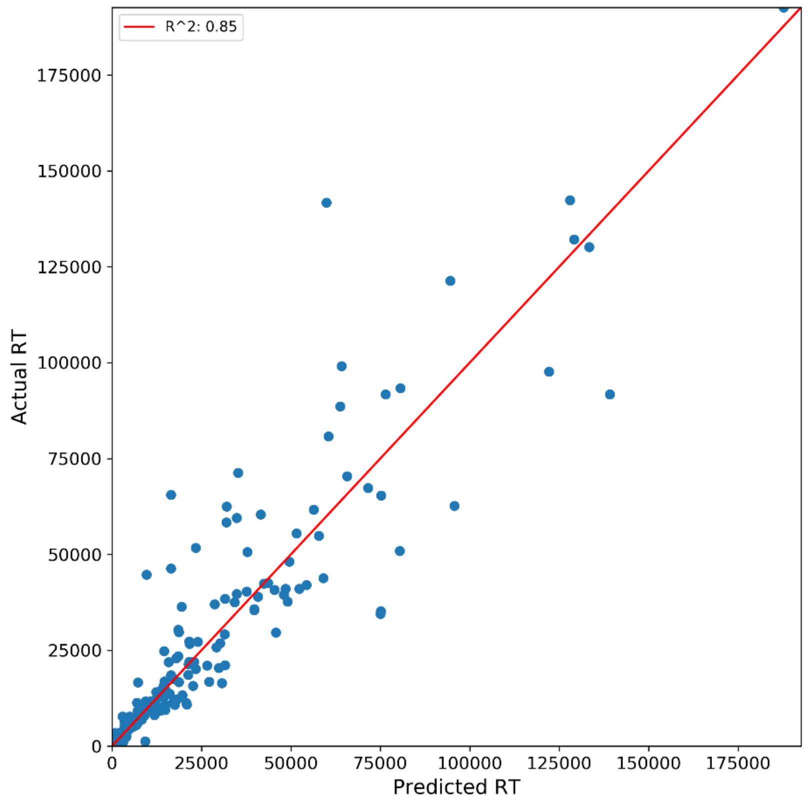


Machine learning (ML) with different methods by two XMAT groups identified similar ordering of features.
Nb and B are known to strongly influence creep life, <u>but were ranked low using ML.</u>
Physics-based models and domain expertise are needed to complement ML. Need a larger data set.

Actual Creep Rupture Time vs. Prediction for 9-12% Cr F-M steels

Gradient Boosting algorithm

- Good fit at low rupture time
- Divergence at long time due to limited data
- Shows the challenge of creep life prediction



We need larger data sets and data farther out in time--expensive and time consuming to obtain.

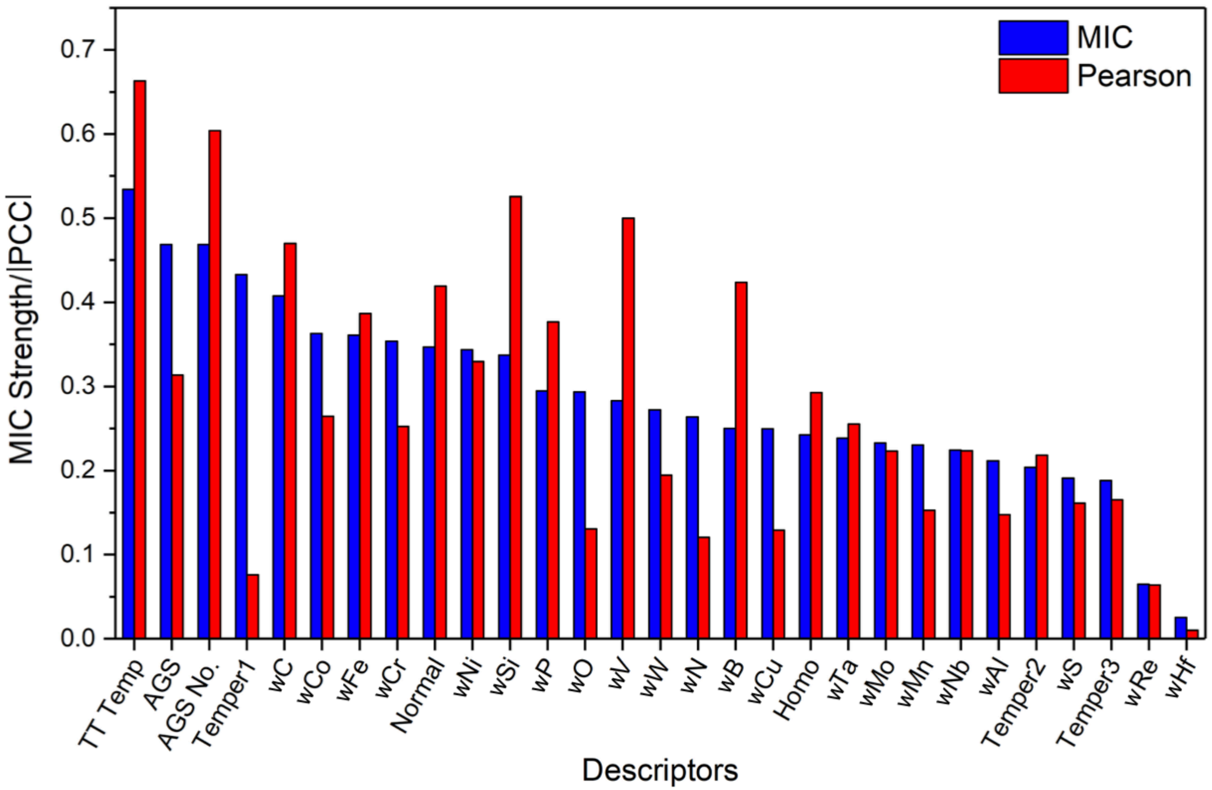
Can we use yield strength to understand creep life of 9-12% Cr steels?

Correlation analysis with > 2000 descriptors

Maximal information coefficient and Pearson correlation

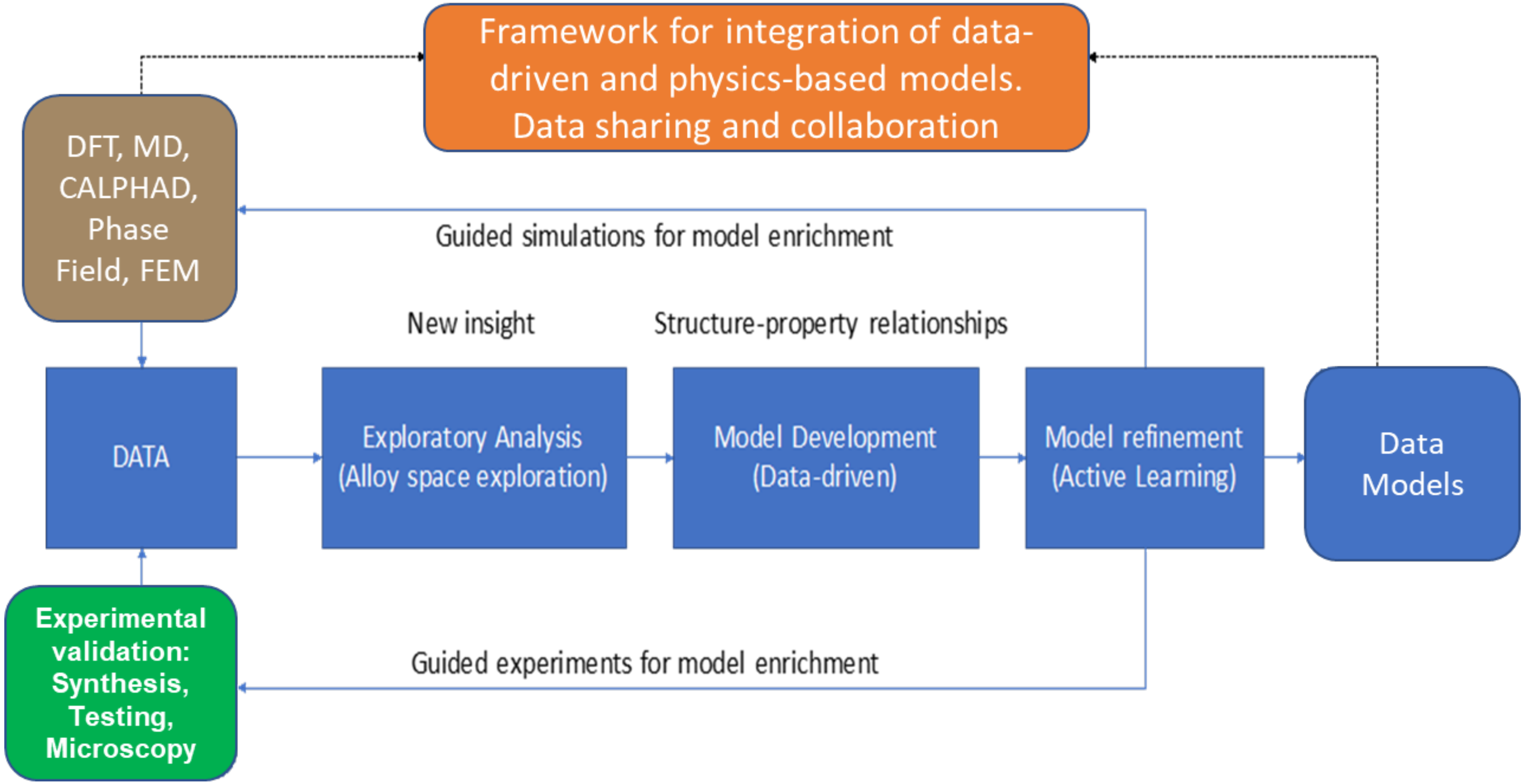
Important descriptors include:

- Heat treatment conditions
- Austenitic grain size
- Trace elements (incl. B and Nb)



Analysis identified the key role of kinetics that is often overlooked.

Includes Data from Literature



We need to integrate data models with physics models and domain science expertise.

Conclusions

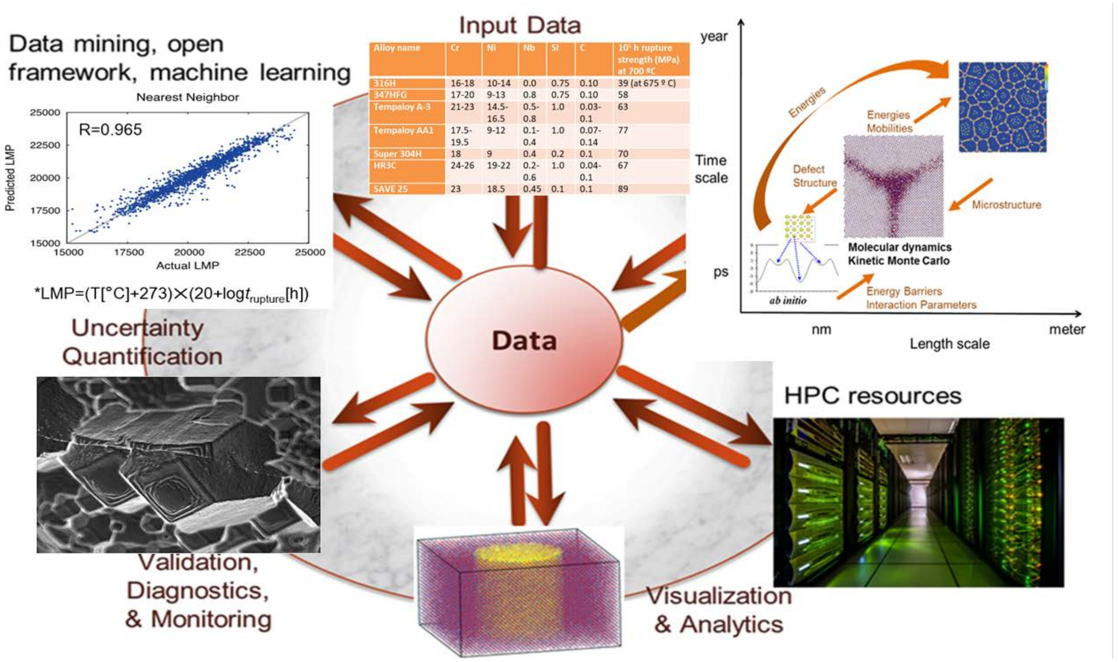
- XMAT is developing an alloy database with quality assessment.
- Framework supports adding metadata and ranking data sets.
- Serves as collaboration platform for distributed team.
- Heat treatment (temperature, hold time, rates) conditions have a key role. Kinetics cannot be overlooked!
- Complexity of phase transformations and microstructure evolution makes linear regression without physics-based insights unsuitable.
- We need to relate composition to microstructure and microstructure to properties instead of trying to link composition to properties.
- Gaps are availability of microstructure and quality of thermodynamic data.

https://edx.netl.doe.gov/

EDX
A Product of NETL

Built by researchers for research

Questions?



Alloy name	Cr	Ni	Nb	Si	C	10^5 h rupture strength (MPa) at 700 ºC
316H	16-18	10-14	0.0	0.75	0.10	39 (at 675 º C)
347HFG	17-20	9-13	0.8	0.75	0.10	58
Tempaloy A-3	21-23	14.5-16.5	0.5-0.8	1.0	0.03-0.1	63
Tempaloy AA1	17.5-19.5	9-12	0.1-0.4	1.0	0.07-0.14	77
Super 304H	18	9	0.4	0.2	0.1	70
HR3C	24-26	19-22	0.2-0.6	1.0	0.04-0.1	67
SAVE 25	23	18.5	0.45	0.1	0.1	89