ENSDF Schema

Adam Hayes, B. Shu, E.A. McCutchan (lead-PI), A. Sonzogni, S. Zhu, A. Mattera, L. Krejci (NNDC, BNL)
S. Yoo (co-PI), C. Soto, S. McCorkle (CSI, BNL)
F. Kondev (ANL)
C. Mattoon (LLNL)

Nuclear Data Week (USNDP)
2020

Outline

- ENSDF development status and plans
- Database technology changes
- Schema under development
- What we can get out of a new database?
- Machine learning for table extraction

ENSDF upgrade status & plan

- Finished first version of “WalletCraft” Nuclear Wallet Cards database, jumping-off point (earlier talk)
- July 2020: received funding for 3 years from Nuclear Data Interagency Working Group FOA LAB 19-2114
  - Develop new ENSDF database
  - Develop machine learning (ML) for table comprehension
  - E.A. McCutchan, S. Yoo (Co-PI’s, BNL); A. Mattera, S. McCorkle, B. Shu, A. Sonzogni, C. Soto, S. Zhu (BNL); F. Kondev (ANL); C. Mattoon (LLNL)
- Computer Science Initiative (CSI) at BNL leading machine learning component
- Major database design work in FY21:
  - Finish new database design
  - Copy full ENSDF database into new database
  - Validate 100% of records against existing ENSDF

Present state of ENSDF and XUNDL

rec_id	dsid	ds_type	nucid
238092001	ADOPTED LEVELS, GAMMAS	ADOPTED	238U
238092002	238PA B- DECAY	DECAY	238U
238092003	238U IT DECAY (280 NS)	DECAY	238U
238092004	242PU A DECAY	DECAY	238U
238092005	236U(T,P)	REACTI	238U

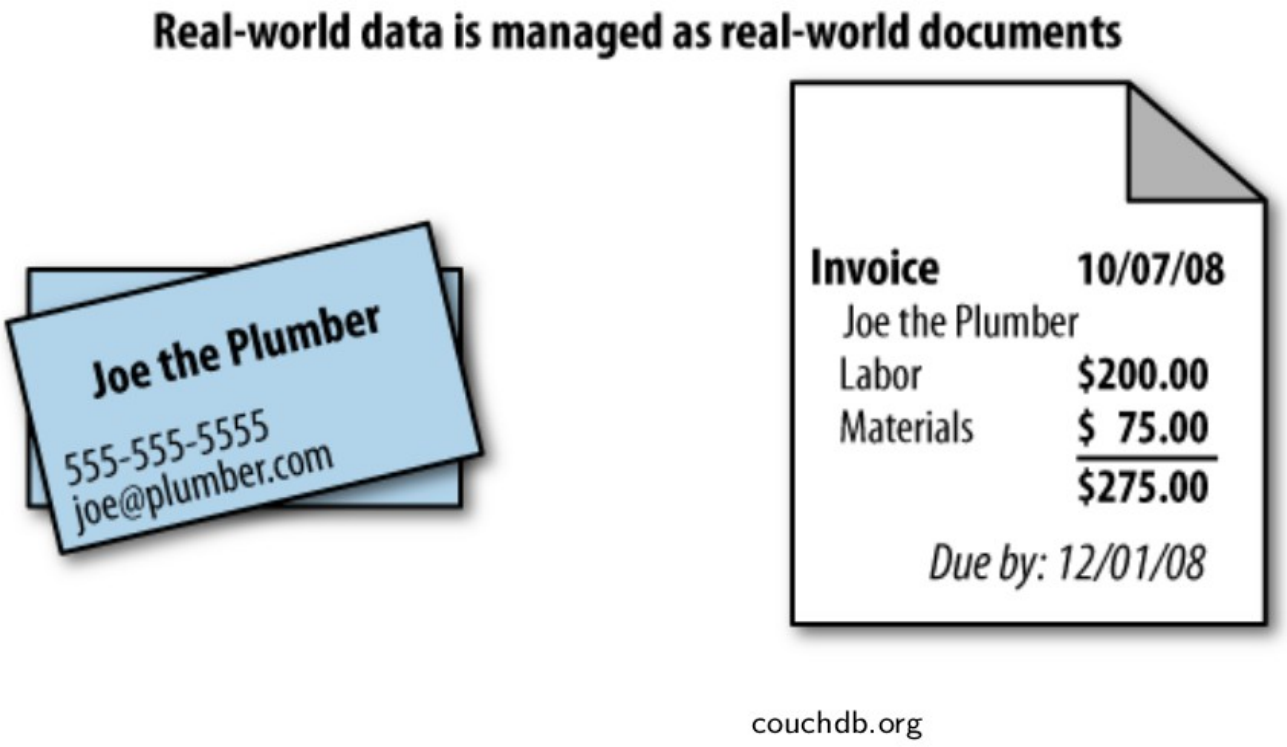
rec_id	line_text
238092002	238U 238PA B- DECAY 15NDS 2
238092002	238U H TYP=FUL$AUT=E. BROWNE, J. K. TULI$CIT=NDS 127, 191 (2015)$
238092002	238U 2 H CUT=1-Jun-2014$
238092002	238U D Slightly modified by E. Browne (7/24, 2014).
238092002	238PA P 0.0 (3-) 2.28 M 10 3585 16
238092002	238U N 1.0
238092002	238U c The Gamow-Teller |b-decay strength function has been calculated in
238092002	238U 2c 1978Iz04 and 1979KlZT.

- Each line is one string, comments wrapped into next record
- Stored in relational (MySQL) database
- One text line, not in fields
- Requires parsing each string column-by-column
- Difficult due to heavy comment use

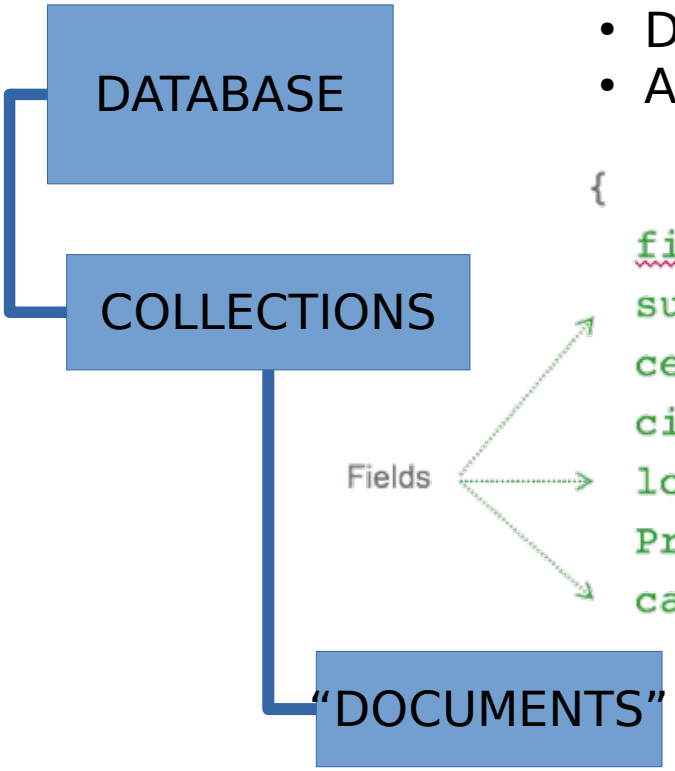
Move to object-oriented database

Relational databases: rows and columns

Object-oriented databases: “documents”



Object-oriented database “documents”



- Document structure; **NOT** stored as a file
- Accessed via database query / search

```
{
  first_name: ‘Paul’,
  surname: ‘Miller’,
  cell: 447557505611,
  city: ‘London’,
  location: [45.123,47.232],
  Profession: [‘banking’, ‘finance’, ‘trader’],
  cars: [
    { model: ‘Bentley’,
      year: 1973,
      value: 100000, … },
    { model: ‘Rolls Royce’,
      year: 1965,
      value: 330000, … }
```

Fields

String

Number

Geo-Coordinates

Typed field values

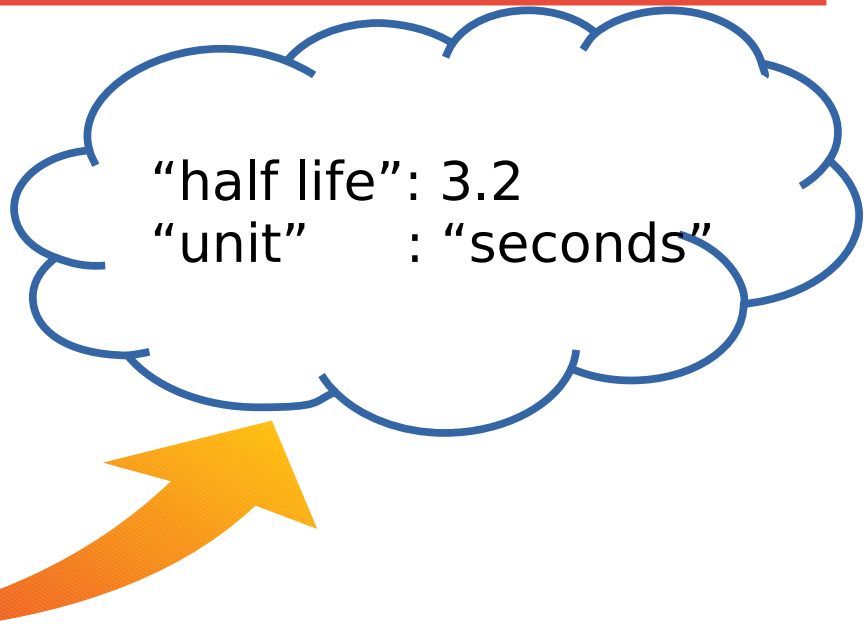
Fields can contain arrays

Fields can contain an array of sub-documents

MongoDB.com

Why object-oriented database?

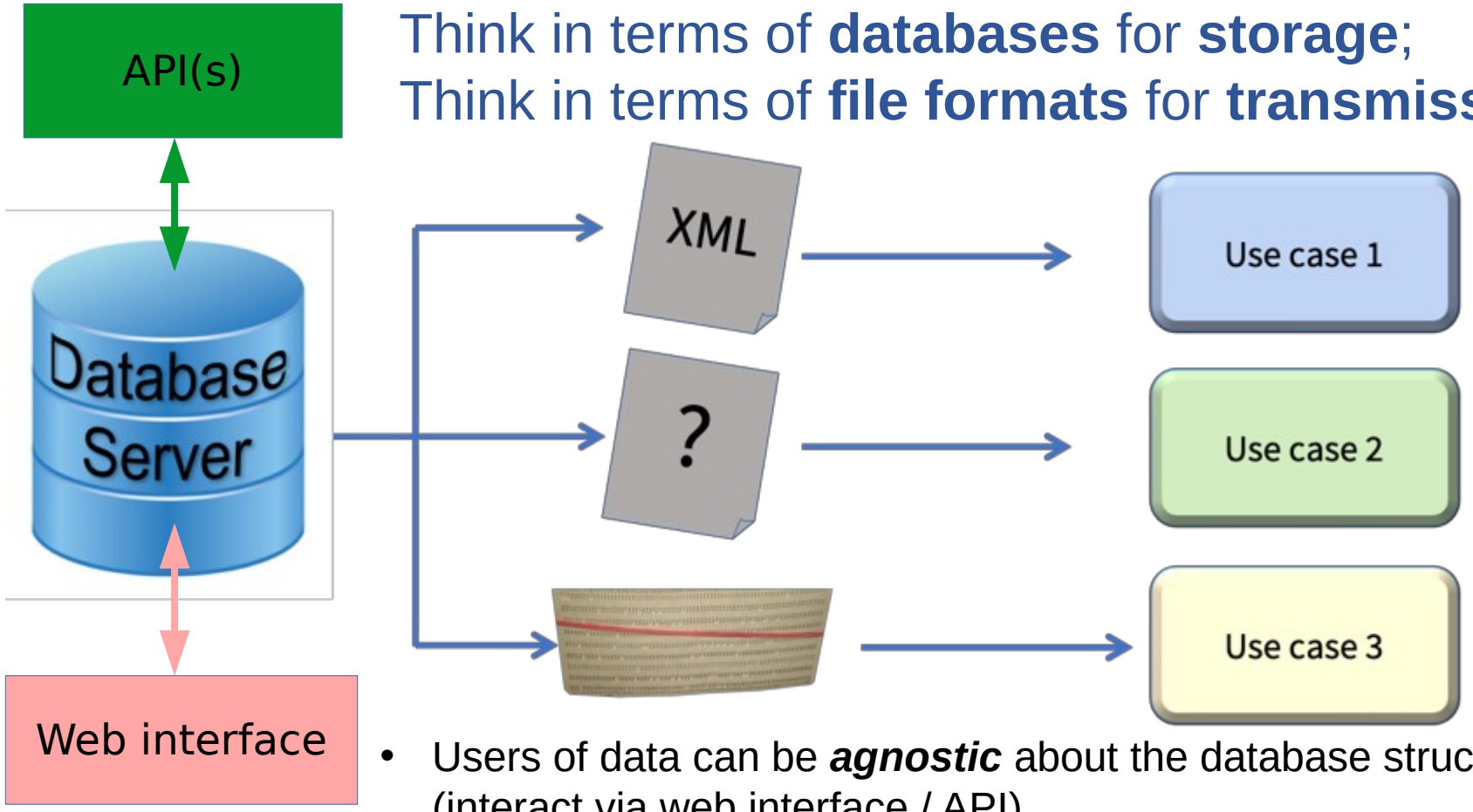
- Variety of data types
  - Numerical, text
  - Arrays (changeable size)
  - Documents
  - Images (e.g. plots)
  - Binary data
- Easily made “human-readable”
- Heterogeneous data (ideal for open data)
- Hierarchical records with fewer cross-references
- Expandible without disruptive changes to codes & users
- Simplicity paradigm: “Store together what you access together”
- Less work loading object-oriented *code* objects



Broad adoption of object-oriented databases (OODB)

1960s-1970s	advent of relational databases, origins of modern types	
1980s	• SQL (Dbase III) • Experiments with Object-oriented Databases	
1990s	Remote servers more common	
1995-96	MySQL, PostgreSQL (relational)	
1997	Caché hybrid relational / hierarchical	(Healthcare, financial systems...)
2000s	Object-oriented databases • Apache "CouchDB" ← • MongoDB	CERN, Apple, GrubHub, Credit Suisse, Motorola, Facebook Apps CERN, eBay, Google, Facebook, PayPal
2020	• Big data • Cloud, replication • Many variants of common database types...	

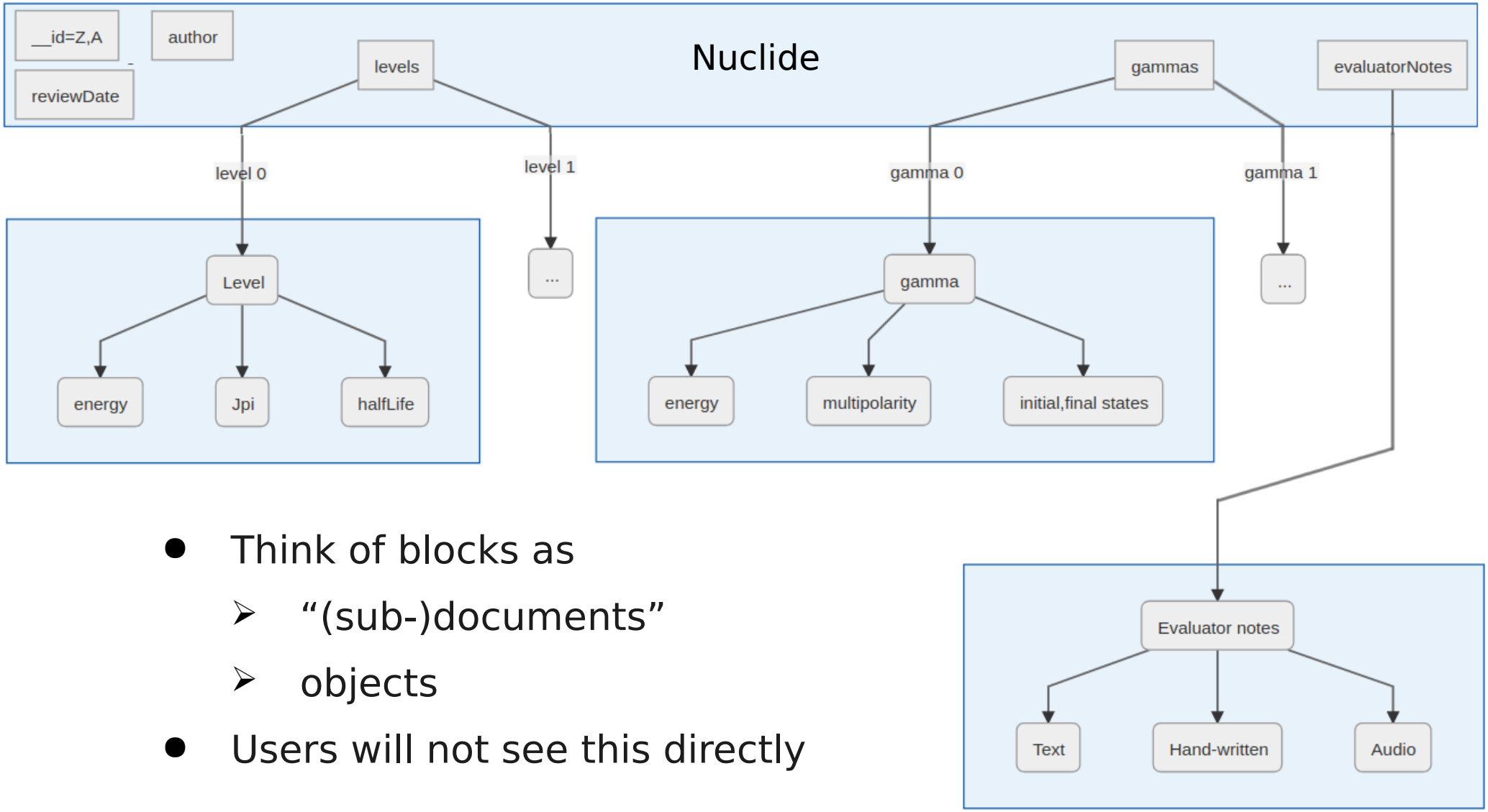
A new concept for ENSDF using OODB



Think in terms of **databases** for **storage**;
Think in terms of **file formats** for **transmission.**

- Users of data can be ***agnostic*** about the database structure (interact via web interface / API)
- Database can be changed ***independently*** of the file format
- Database can store ***many data types;*** images, pdfs, experimental data …

Database design (in progress)
Top-level Nuclide document



- Think of blocks as
  - “(sub-)documents”
  - objects
- Users will not see this directly

Database design (in progress)
Top-level Nuclide document as JSON

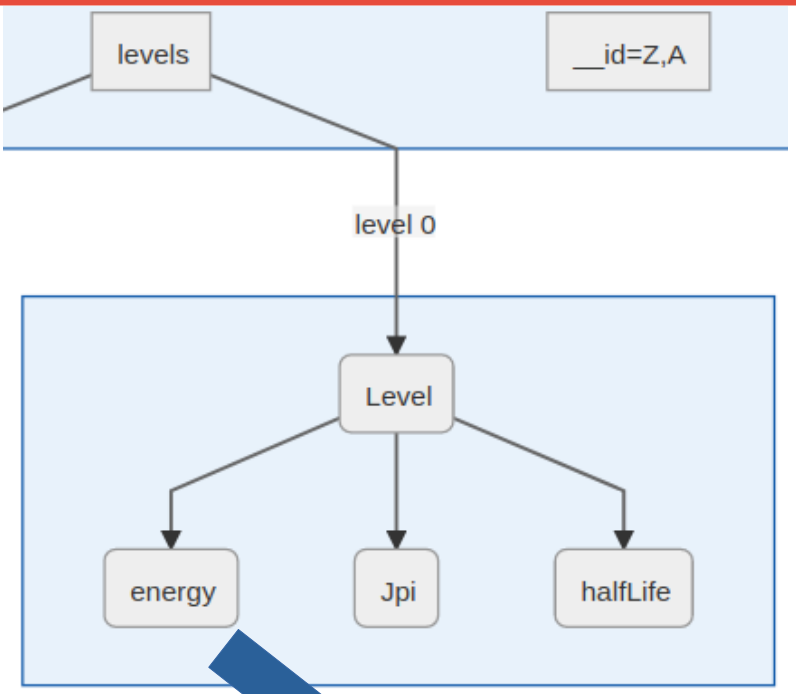
```
{
    "_id": "72,178",
    "author": {},
    "reviewDate": "1/1/1980",
    "levels": {},
    "gammas": {},
}
```

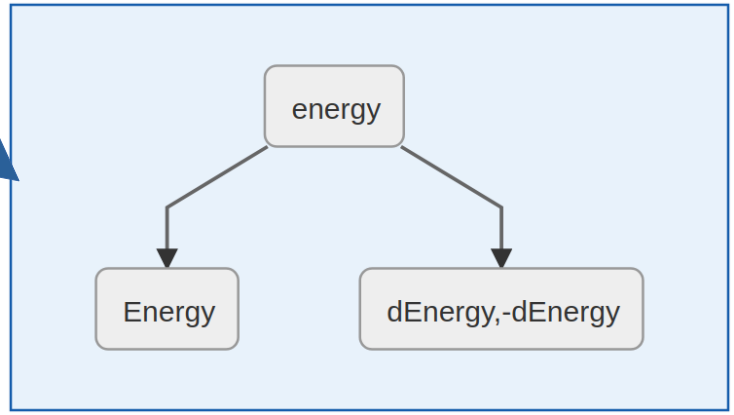
Predominantly dictionaries
→ easier to expand

{“key”:value}

```
"levels": {
    "0": {↔},
    "1": {
        "energy": {↔},
        "halfLife": {},
        "JPi": {↔}
    }
```

Hierarchy of documents (in progress)
Level, Level Energy sub-documents

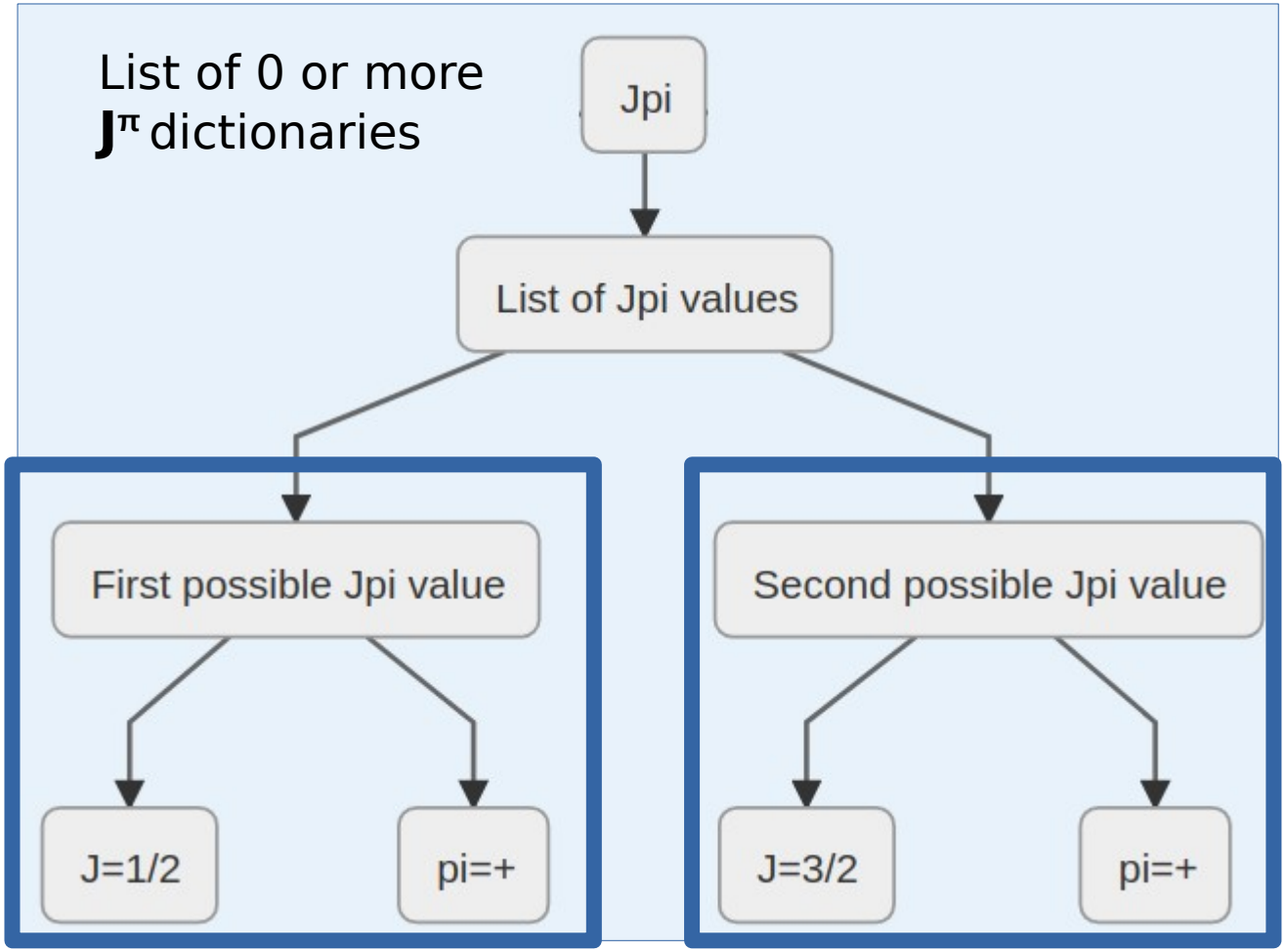




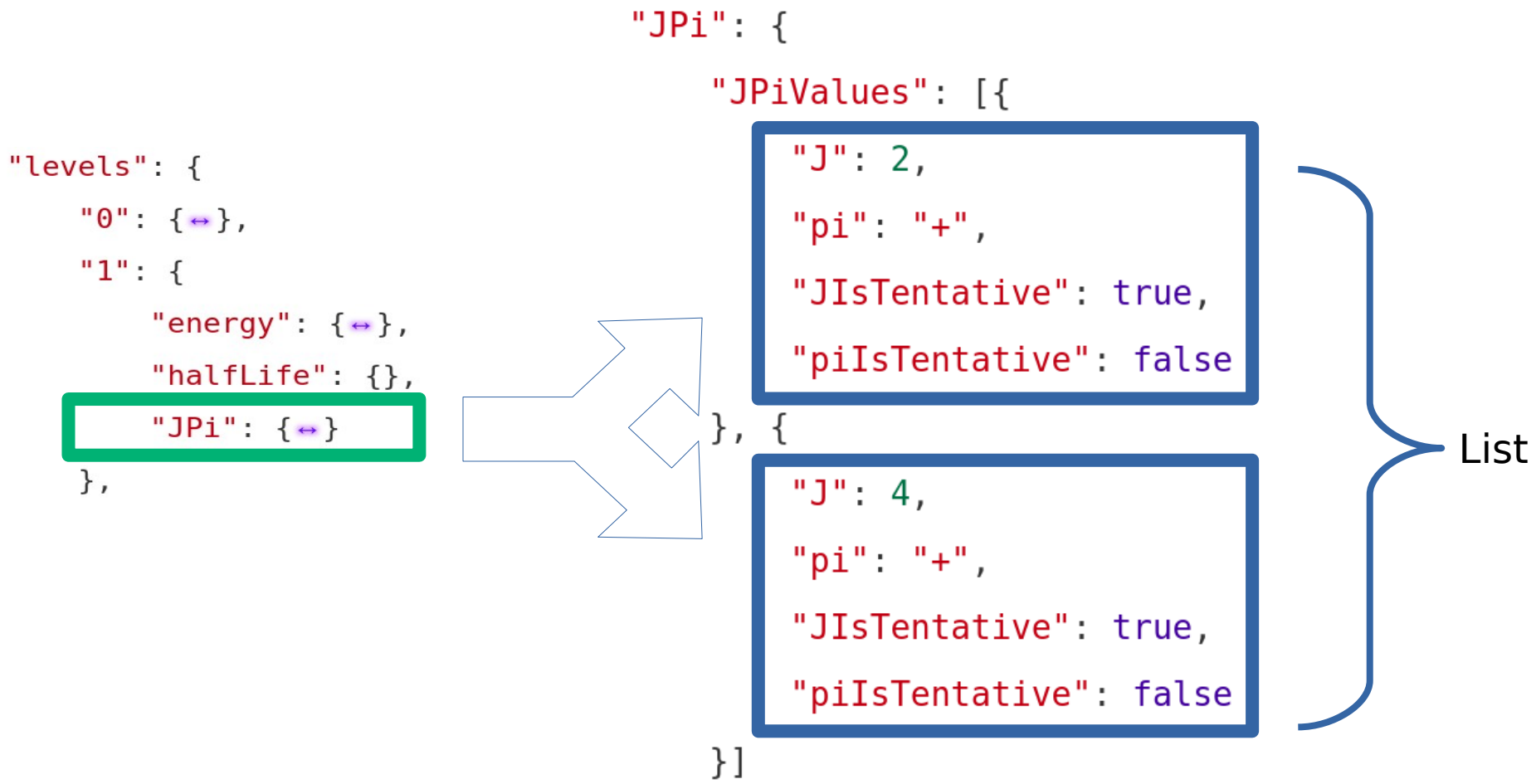

```
"levels": {
    "0": {↔},
    "1": {
        "levelEnergy": {
            "energy": 100.0,
            "dEnergy": [-0.5, 1.0]
        },
        "halfLife": {},
```

list

Hierarchy of documents (in progress)
J^{π} sub-document

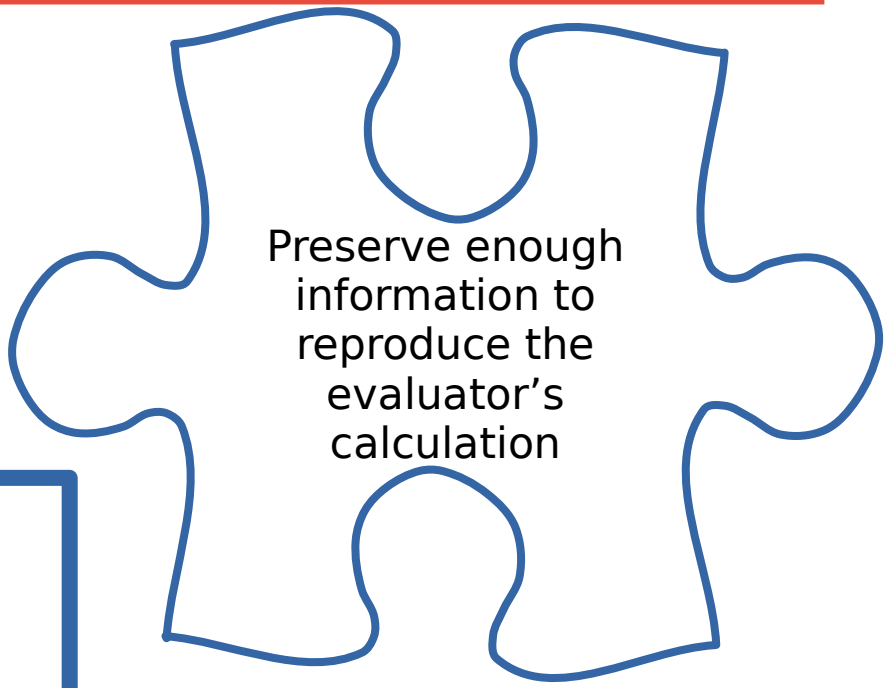


Database “schema”
J^{π} sub-document



Database “schema” $t_{1/2}$ sub-document (and others)

```
"halfLife": {
    "halfLife": 1.494E-09,
    "dHalfLife": [0.023E-09],
    "upperLimit": null,
    "lowerLimit": null,
    "measurements": {
        "1991Ab01": {
            "included": false,
            "reason": "Unresolved isomer component"
            "method": "DSAM"
        }
    },
```



Database “schema” $t_{1/2}$ sub-document (and others)

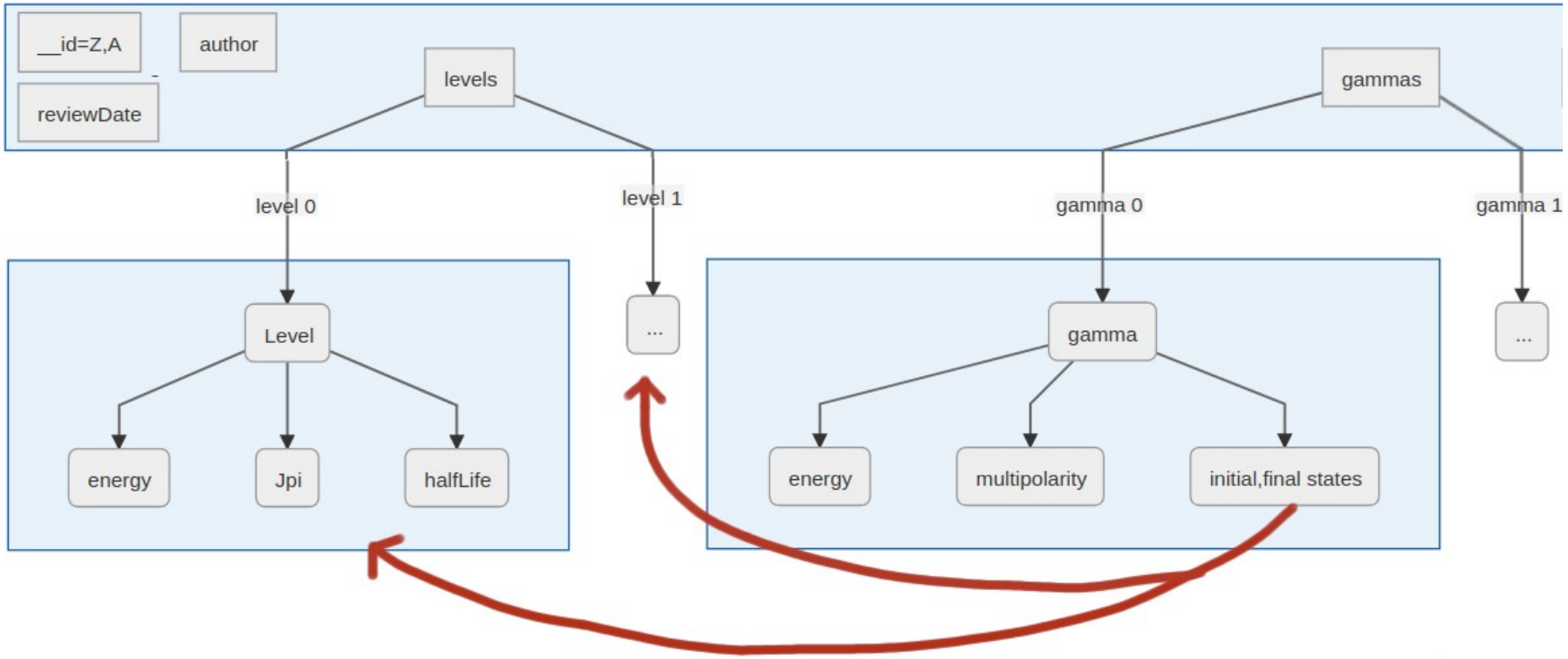
```
"halfLife": {
    "halfLife": 1.4
    "dHalfLife": [0
    "upperLimit": n
    "lowerLimit": n
    "measurements":
        "1991Ab01":
            "includ
            "reason
            "metho
        }
    },
```

And similar design considerations for

- Decay modes
- Q Values
- Abundance
- ...

Database “schema” gamma decays

- Similar to level documents
- **Refer explicitly** to initial and final levels in every case (not inferred by processing codes), except for unplaced gammas



Database “schema”
Take advantage of binary data

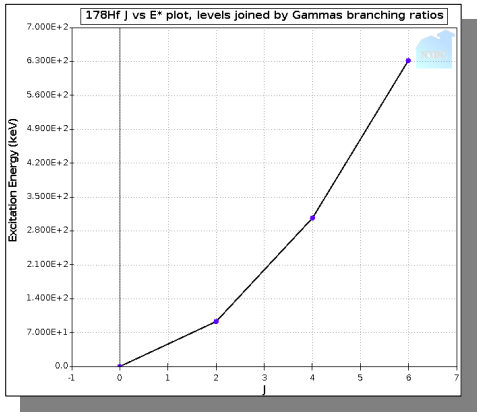
- CouchDB handles binary objects
- Potential to speed the workflow
- Example: preserve valuable history of evaluator’s notes
- Content for internal use
- **Any format → no editing needed**


```
{
    "_id": "72,178",
    "author": {},
    "reviewDate": "1/1/1980",
    "evaluatorNotes": {↔},
    "levels": {
        "0": {
            "energy": {
```

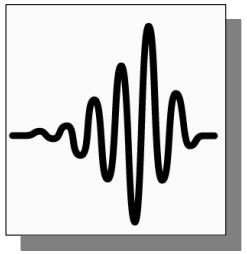

Website links

See: www.nndc....

Screenshots



Even audio



Typed notes

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Scanned handwritten

E(level)†	L‡	dσ/dΩ(25°)#
0.0	0	266
80 *1*		21
264 *1*		12
548 *1*		2.1
821 *2*		~~3.7~~ 2.9
997 *2*		12
1195 *2*		≈4.5 typo in orig.
1217 *2*	0	30
1275 *2*		2.7
1359 *3*		3.0
1411		5.4
1422	0	21

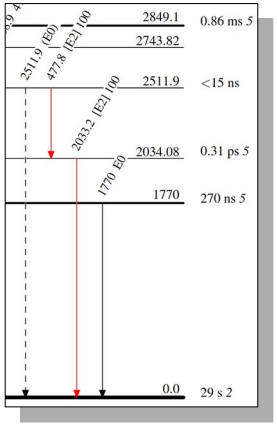
Icons: thenounproject.com/indygo/

Defining consistent standards

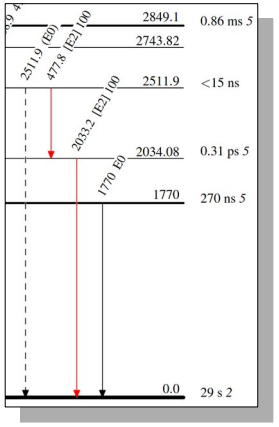
- Developing a standard for key names (names of quantities, e.g. energy, halflife...)
  - Consistent camel-case “levelEnergy”
  - Always full words (predictable): “levelEnergy,” not “levE”
  - “d” for uncertainty: “levelEnergy” → “dLevelEnergy”
  - “Is” for true / false flag: “parityIsTentative = true”

Taking advantage of a well-planned database

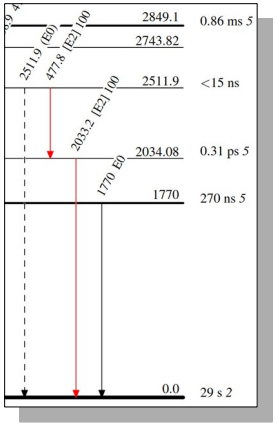
- Proper versioning *within the database*—evaluators have instant access to history
- Allows for
  - Highlighting changes
  - Fast cross-checking



2015 evaluation

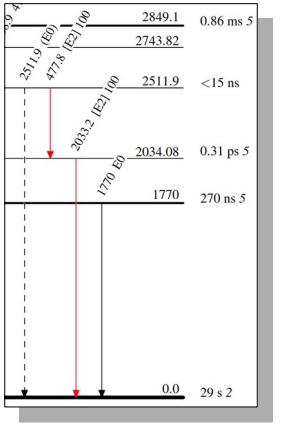


2017 evaluation

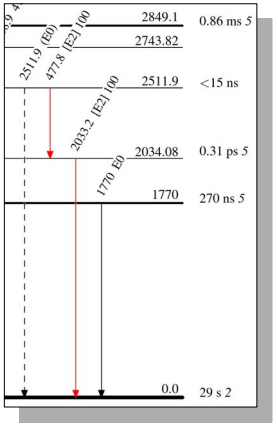


2019 evaluation

. . .



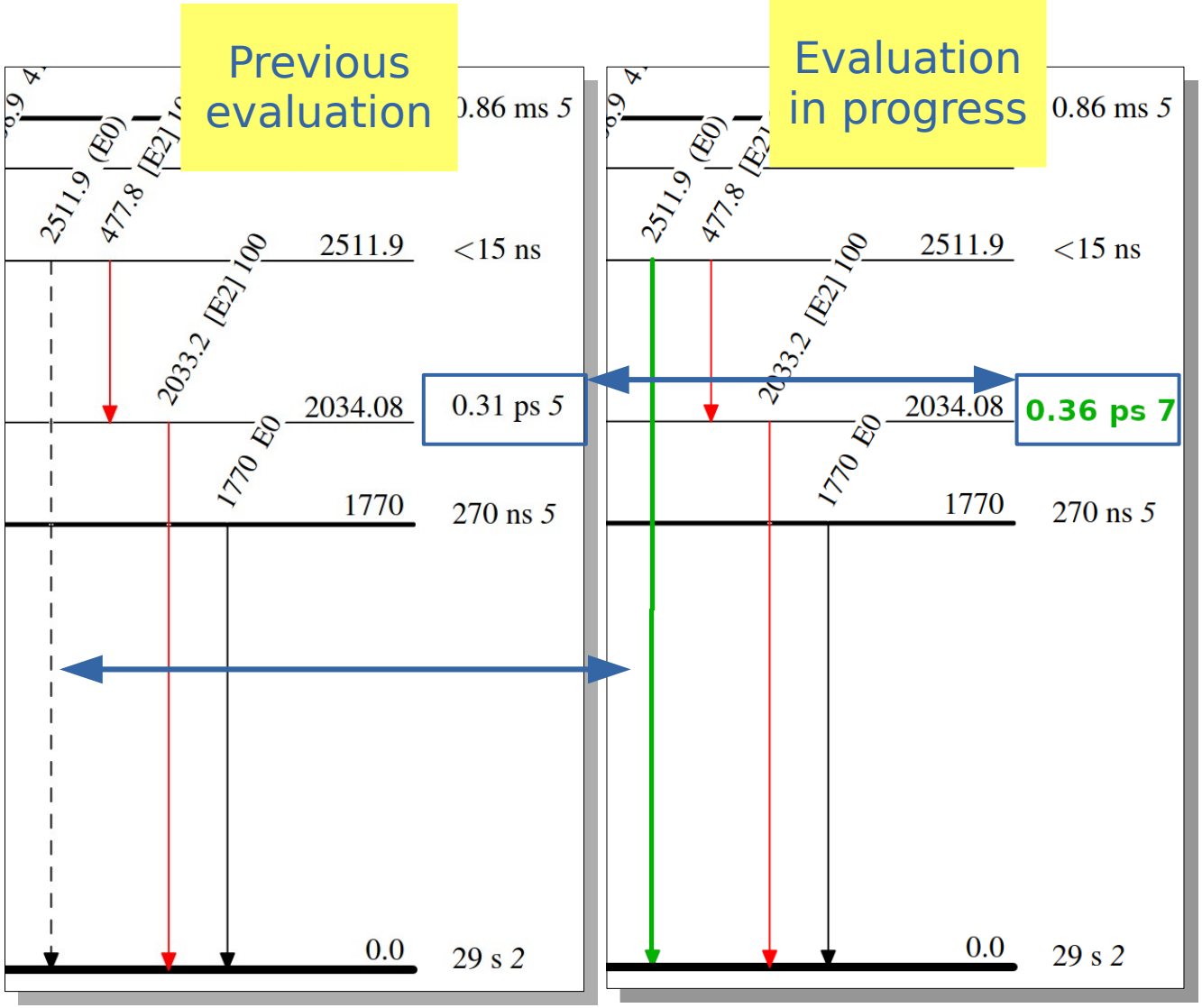
Published ”public” version



Unpublished evaluation in progress?

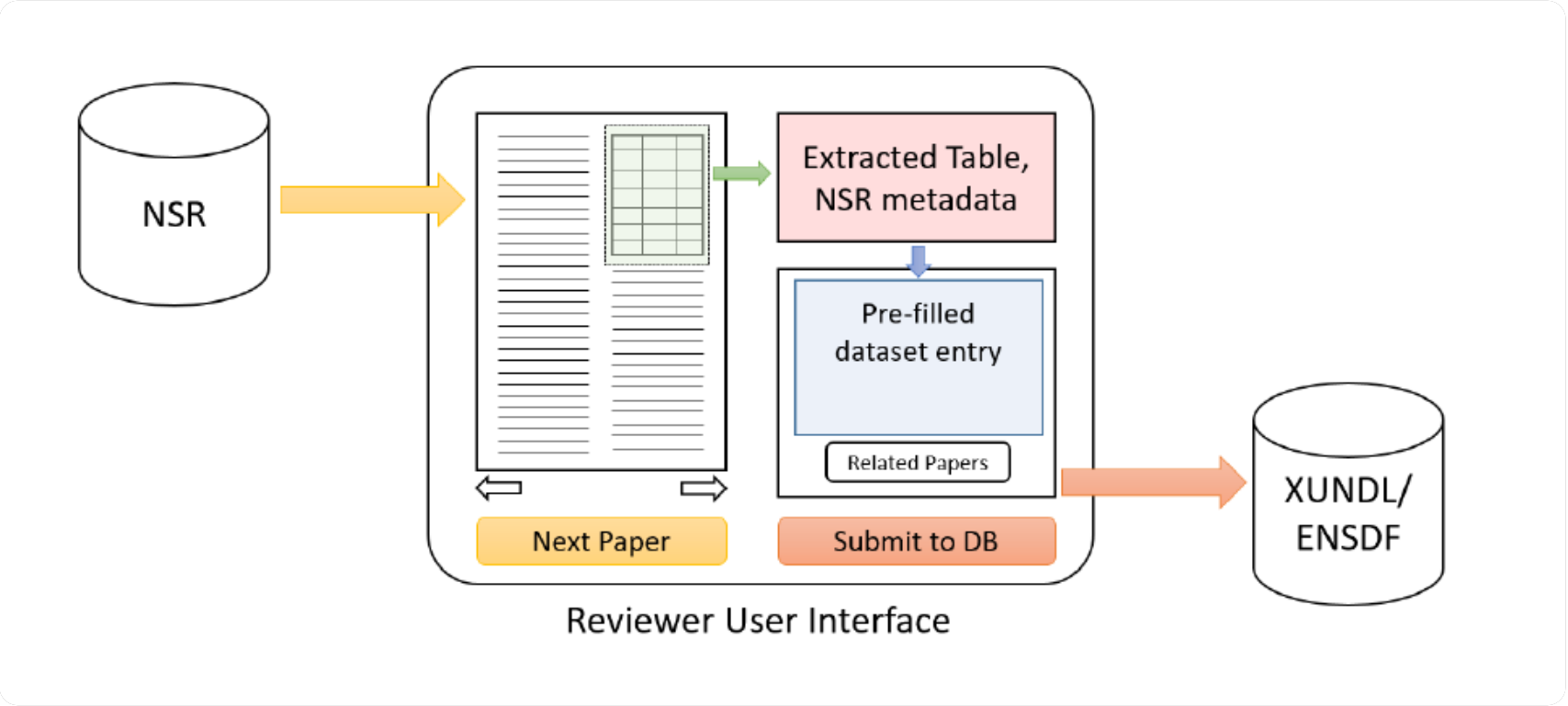
Taking advantage of a well-planned database

- Can we improve database currency with database tools?
- What tools would improve evaluator efficiency?
  - “diff” tool?
  - …?



National Nuclear Data Center

Funded machine learning project: Make the most of the evaluator's effort

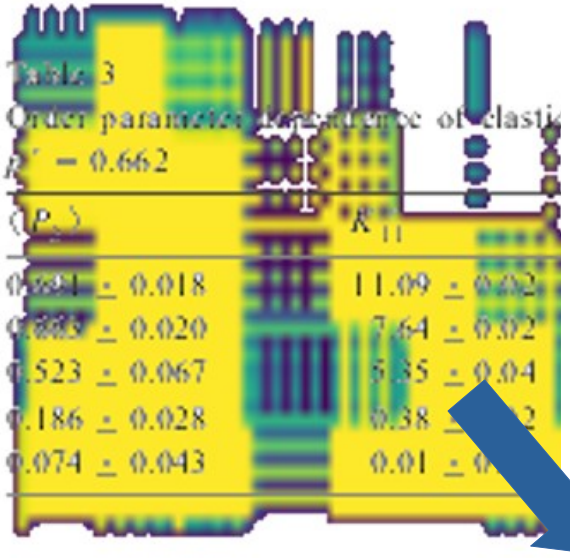


- OCR for extraction of metadata and numerical data
- Automatic population of database

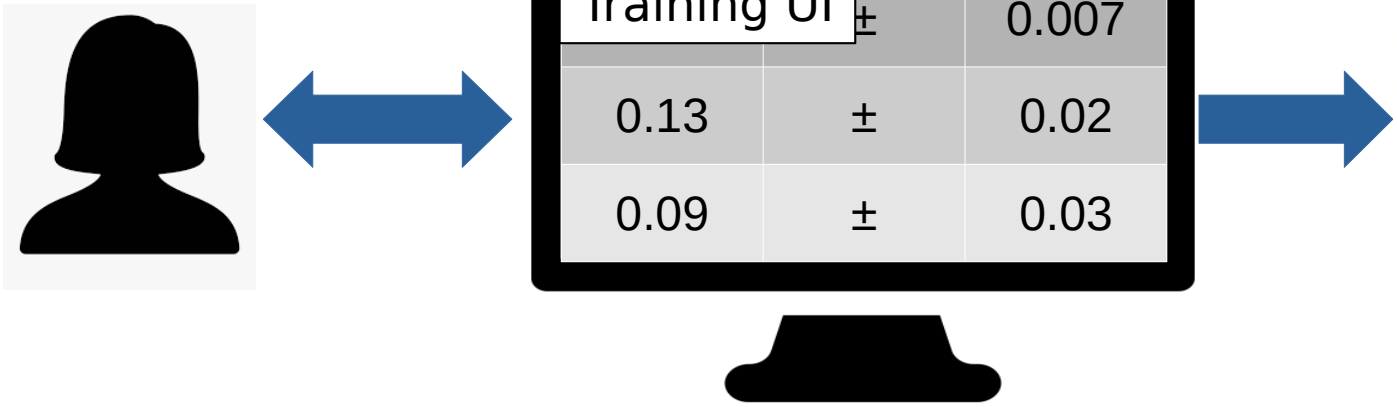
- Reviewer interface for side-by-side comparison
- Consideration by scientist still essential

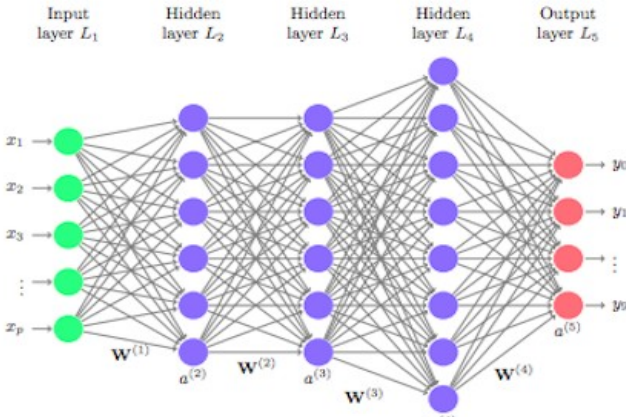
Slide: E.A.McCutchan

Brookhaven CSI group's Table extraction using Machine Learning



- CSI evaluating deep-learning models, including TableNet https://arxiv.org/abs/2001.01469
  - Example of segmentation before tuning thresholding
  - Operating on a table image
- Preparing for training on both PDF and LaTeX





UC Business Analytics R Programming Guide

Image: Calos Soto, BNL's Computer Science Initiative

Icons: thenounproject.com/nhy7565/

Brookhaven CSI group's Table extraction using Machine Learning

Image: Calos Soto, BNL's Computer Science Initiative

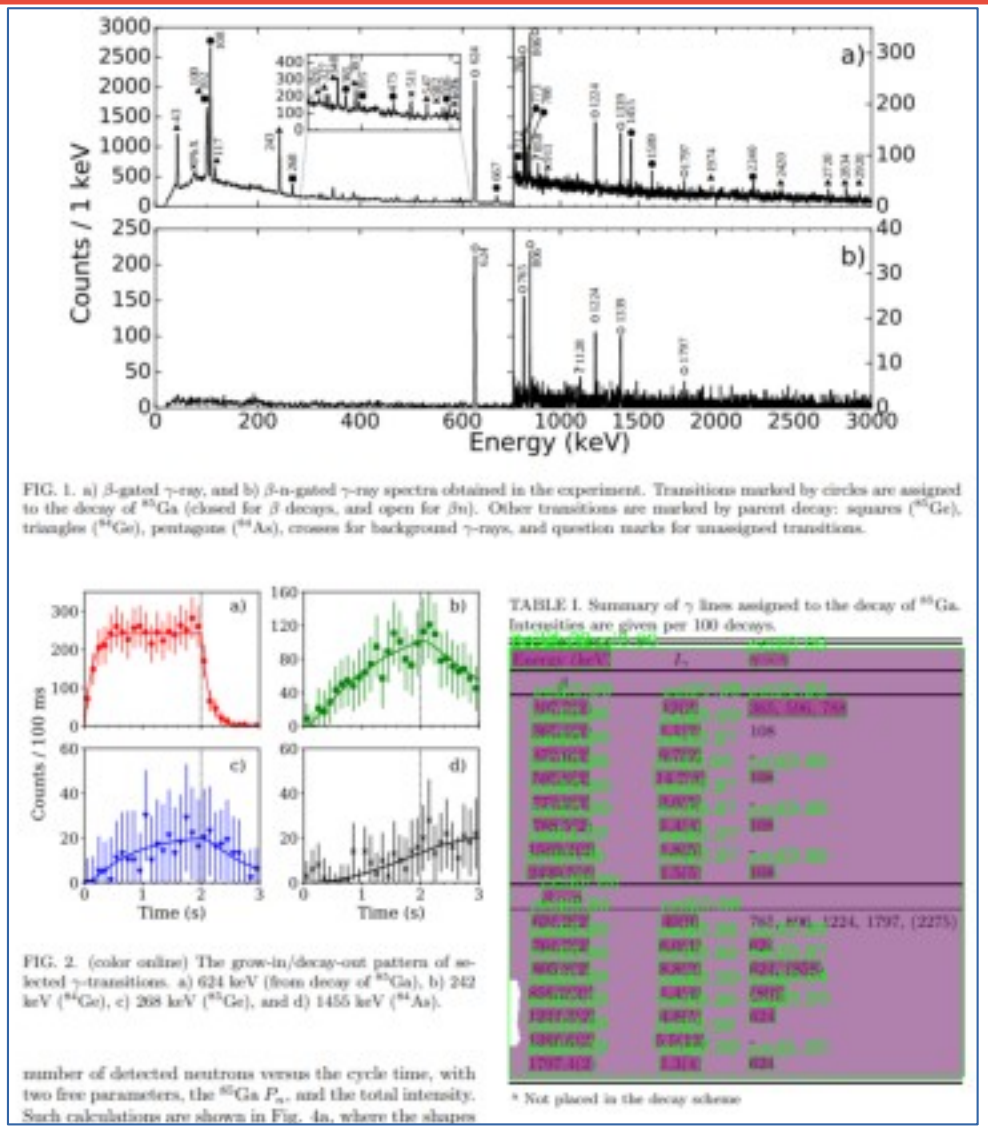


FIG. 1. a) β-gated γ-ray, and b) β-n-gated γ-ray spectra obtained in the experiment. Transitions marked by circles are assigned to the decay of ^{85}Ga (closed for β decays, and open for βn). Other transitions are marked by parent decay: squares (^{85}Ge), triangles (^{84}Ge), pentagons (^{84}As), crosses for background γ-rays, and question marks for unassigned transitions.

FIG. 2. (color online) The grow-in/decay-out pattern of selected γ-transitions. a) 624 keV (from decay of ^{85}Ga), b) 242 keV (^{84}Ge), c) 268 keV (^{85}Ge), and d) 1455 keV (^{84}As).

number of detected neutrons versus the cycle time, with two free parameters, the ^{85}Ga P_n, and the total intensity. Such calculations are shown in Fig. 4a, where the shapes

TABLE I. Summary of γ lines assigned to the decay of ^{85}Ga. Intensities are given per 100 decays.

* Not placed in the decay scheme

- Further tuning: identifies table in a complex image
- Developed a preliminary schema for "raw" table data
- Two major aspects
  - Identify rows, columns
  - Comprehend meaning of values

```
{"levelEnergy" : 171.7
{"initialJpi"  : "3/2-"
{"gammaEnergy" : 171.7
{"A0"          : 6
{"RDCO"        : 0.65
```

Input needed

- In what ways can we exploit new **database technology** and related **software & APIs** to
  - Streamline evaluator's work
  - Improve database currency
  - Add valuable new content (open data, supplemental data)
- Committees?
  - New data
  - Schema design
  - Codes / APIs
  - Others?

Summary

- Moving from 50-year old nuclear data storage format to
- Object-oriented database technology supports
  - Supplemental data for evaluators
  - Open data
  - Heterogeneous data
- Preparing for unknown future needs
- Test case: Nuclear Wallet Cards (my earlier talk)
- ENSDF modernization in progress using Wallet Card upgrade as a starting point
- Machine learning
  - Funded ENSDF machine learning project underway led by E. McCutchan (BNL,ENSDF) and S. Yoo (BNL, CSI)
  - Improve efficiency of evaluators pulling data from journals

END

END

END

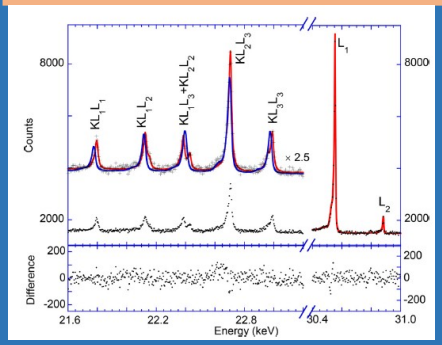
Input needed

- Rules and standards, *e.g.*
  - What Q-values do we want to store? Sn, Sp, Qα, Qβ...
  - Do we allow things like J^{π} = 1+,(2+), or other combinations where some, but not all, are tentative?
- What tools in a new interface would help to streamline evaluation process (*e.g.* a “diff” view)?
- Are there additional types of data that would help the evaluator?

Rethinking ENSDF and XUNDL

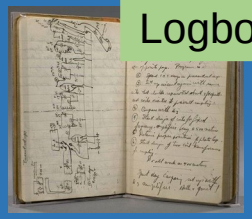
ENSDF now:
Discrete levels
Discrete radiation

X-rays and Auger electrons

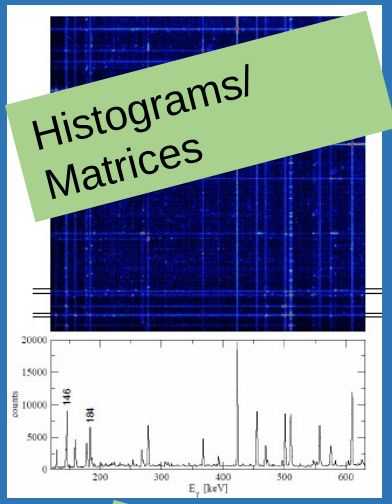


M. Alotiby et al., J. Elec. Spect. 232, 73 (2019)

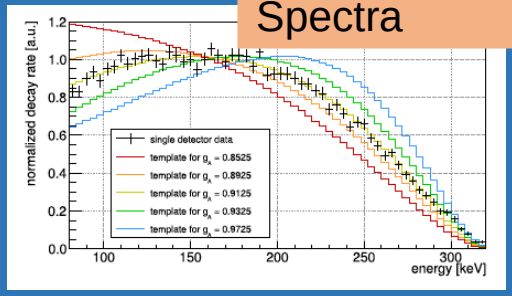
Logbook



Histograms/ Matrices



Continuous Spectra



Cobra, Phys. Lett. B 800, 135092 (2020)

Raw Event Data & Codes

```
Event 20
Det 37
Det 61
Event 20
Det 04
Det 24   E 615
Det 46   E 339
Event 20585
Det 17   E 761
Det 20   E 186
Det 59   E 615
```

Slide: E.A. McCutchan

Schema: Binary data

01 Sep 2020 | Network Updates

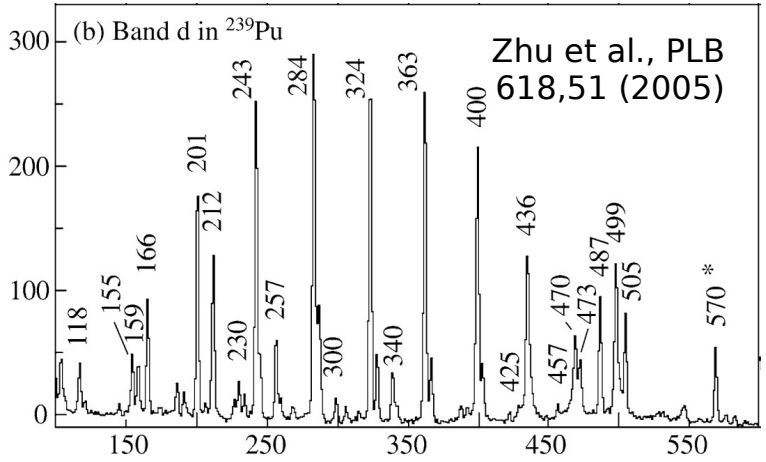
CERN releases fifth batch of open data recorded from Large Hadron Collider experiment

By Communication from CERN

All research-quality data recorded by CMS during the first two years of LHC operation are now publicly available.

- Evaluated nuclear data: “open results”

238	0+	46.166	87.7 y *1*	α, SF 1.9×10^{-7}%
239	1/2+	48.591	24110 y *30*	α, SF $3.\times10^{-10}$%
240	0+	50.128	6561 y *7*	α, SF 5.7×10^{-6}%



- Ultimate conclusion: “Open data”
  - Unpublished material
  - Preprocessed data
  - Example source code
  - Education, basic science
  - Reproducibility
  - Repeatability
  - Preservation
  - Re-use