

# SEINE: There and Back Again

**Dan Connolly**

Biomedical informatics software engineer

Division of Medical Informatics



# KUMC Medical Informatics

- Russ **Waitman**, Director of Medical Informatics
- Software Engineers: Dan **Connolly**, Nathan Graham, Bhargav **Adagarla**, Matt **Hoag**, Mike Prittie, Lav Patel, Nazma Kotcherla
- Analysts, Honest Brokers: Tamara **McMahon**, Sravani Chandaka, Li Huang, Rachel Gyore, Maren Wennberg
- Project Management: Steve Fennel, Brittany Zschoche, Hillary Sandoval

**Your researchers are fisherman: wanting to land data to answer their research hypothesis**

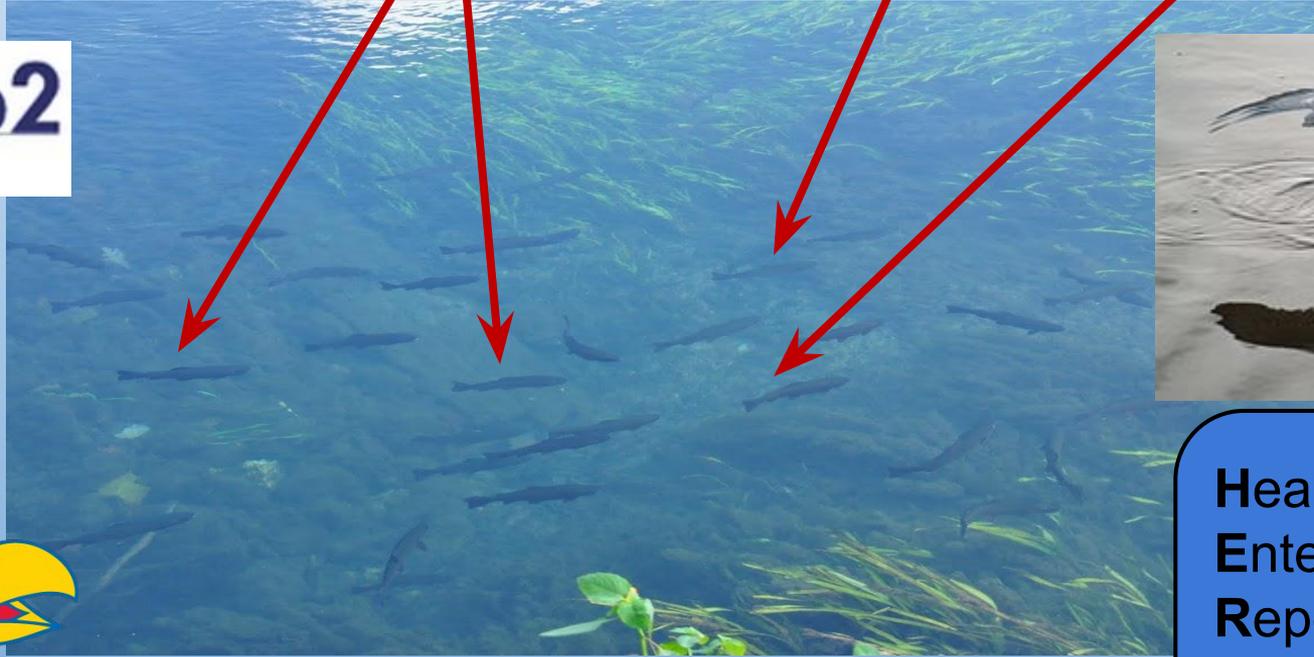


Bennett Spring Trout Park, Lebanon Missouri  
<http://mdc.mo.gov/regions/southwest/bennett-spring>

# The Fish: Diagnoses, Demographics, Observations, Treatments



# The Fish: Diagnoses, Demographics, Observations, Treatments



Healthcare  
Enterprise  
Repository for  
Ontological  
Narration

# SEINE = EDC <-> IDR



Synthesizing  
EDC  
IDR  
Network  
Exchange

# REDCap Forms, Fields + Data -> i2b2 Folders, Concepts + Facts

Pilot project:  
Triple  
Negative  
Breast  
Cancer  
(TNBC)  
Registry

The image displays a REDCap form for a 'Registration' event and its corresponding i2b2 folder structure. The form on the left contains the following data:

Form Field	Value
Event Name	Registration
Baseline Demographics CRF	123456
Baseline Tumor Characteristics CRF	(H)
Date of Biopsy	2013-10-0
Location of Biopsy	KU
Surgical Pathology Number	0000000
Pathological Type	Ductal
TNM Staging: T	Medullary
TNM Staging: N	Ductal
TNM Staging: M	Other
Stage	I
Triple Negative?	Yes
ER%	22
PR%	24
HER2: IHC	2+
HER2: FISH	22
EGFR%	32
KI67%	46
P53%	34
Others	
Form Status	

The i2b2 folder structure on the right shows the following hierarchy:

- REDCap [265,590 facts; 4,510 patients]
  - Distress2 [3,675 facts; 3,675 patients]
  - Triple Negative BRCA Registry- TNBC only [37,336 facts]
    - 66-120 month follow ups [467 facts; 15 patients]
    - BRCA Status [1,684 facts; 186 patients]
    - Baseline Demographics [1,302 facts; 186 patients]
    - Baseline Medical History [11,307 facts; 186 patients]
    - Baseline Tumor Characteristics [1,849 facts; 186 patients]
      - Complete? [186 facts; 186 patients]
      - Date of Biopsy [184 facts; 184 patients]
      - HER2: IHC [182 facts; 182 patients]
      - Location of Biopsy [186 facts; 186 patients]
        - KU [56 facts; 56 patients]
        - Other [130 facts; 130 patients]
      - Pathological Type [181 facts; 181 patients]
        - Ductal [176 facts; 176 patients]
        - Lobular [<10 facts]
        - Medullary [<10 facts]
        - Other [<10 facts]
      - Stage [186 facts; 186 patients]
      - TNM Staging: M [186 facts; 186 patients]
      - TNM Staging: N [186 facts; 186 patients]
      - TNM Staging: T [186 facts; 186 patients]
      - Triple Negative? [186 facts; 186 patients]
        - No
        - Yes [186 facts; 186 patients]

# REDCap Forms, Fields + Data -> i2b2 Folders, Concepts + Facts

Pilot project:  
Triple  
Negative  
Breast  
Cancer  
(TNBC)  
Registry

The image shows a screenshot of a REDCap form titled "Registration". The form contains several data entry fields, many with dropdown menus. A summary table at the bottom right shows the distribution of data for several key fields.

Field	Value
Event Name	Registration
Baseline Demographics CRF	123456
Baseline Tumor Characteristics CRF	(H)
Date of Biopsy	2013-10-0
Location of Biopsy	KU
Surgical Pathology Number	0000000
Pathological Type	Ductal
TNM Staging: T	Medullary
TNM Staging: N	Ductal
TNM Staging: M	Lobular
Stage	I
Triple Negative?	Yes
ER%	22
PR%	24
HER2: IHC	2+
HER2: FISH	22
EGFR%	32
Ki67%	46
P53%	34
Others	(H)
Form Status	

Field	Count	Patients
TNM Staging: T	186	186 patients
TNM Staging: N	186	186 patients
TNM Staging: M	186	186 patients
Triple Negative? (Yes)	186	186 patients
Triple Negative? (No)	0	0 patients

Pilot experience:

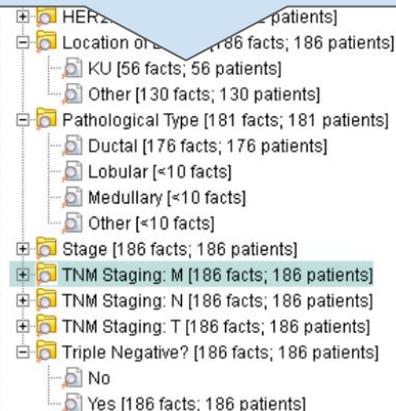
- Instrument Validation
  - Age as drop-down vs. text with number validation
  - Repetition vs. longitudinal
- Data Integration (overlap)
  - Biospecimen Repository: 61%
  - NAACCR tumor registry: 56%
  - Social Security Death Index: 2%

# REDCap Forms, Fields + Data -> i2b2 Folders, Concepts + Facts

Event Name: **Registration**

Baseline Demographics CRF	123456
Baseline Tumor Characteristics CRF	(H)
Date of Biopsy	(H) 2013-10-0
Location of Biopsy	(H) KU
Surgical Pathology Number	(H) 0000000
Pathological Type	(H) Ductal
TNM Staging: T	(H) Medullary
TNM Staging: N	(H) Ductal
TNM Staging: M	(H) Lobular
Stage	(H) I
Triple Negative?	(H) Yes
ER%	(H) 22
PR%	(H) 24
HER2: IHC	(H) 2+
HER2: FISH	(H) 22
EGFR%	(H) 32
Ki67%	(H) 46
P53%	(H) 34
Others	(H)
Form Status	

- ETL code developed as part of HERON ETL
- For access, see [MultiSiteDev](http://MultiSiteDev.informatics.gpcnetwork.org) in [informatics.gpcnetwork.org](http://informatics.gpcnetwork.org)



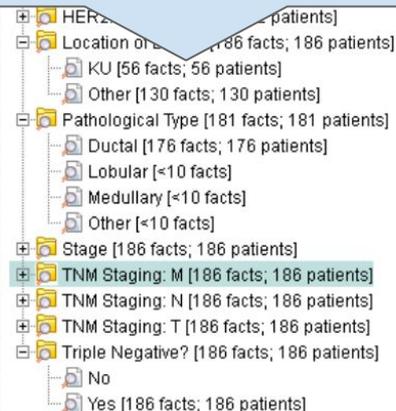
# REDCap Forms, Fields + Data -> i2b2 Folders, Concepts + Facts

Event Name: **Registration**

Baseline Demographics CRF	123456
Baseline Tumor Characteristics CRF	(H)
Date of Biopsy	(H) 2013-10-0
Location of Biopsy	(H) KU
Surgical Pathology Number	(H) 0000000
Pathological Type	(H) Ductal
TNM Staging: T	(H) Medullary
TNM Staging: N	(H) Ductal
TNM Staging: M	(H) Lobular
Stage	(H) I
Triple Negative?	(H) Yes
ER%	(H) 22
PR%	(H) 24
HER2: IHC	(H) 2+
HER2: FISH	(H) 22
EGFR%	(H) 32
KI67%	(H) 46
P53%	(H) 34
Others	(H)
Form Status	

Road Not Taken: ODM

Poor fit for KUMC's automated ETL process.



# Data Privacy

- **HERON is fully de-identified.**
- **ETL strips:**
  - **Free text**
  - **REDCap identifier fields**

# Access Control

- **Norm: i2b2 UI respects REDCap access control**
  - **i2b2 projects configured real-time at i2b2 log-in**
- **Special case: REDCap project open to all i2b2 users**

# Access Control

- **Norm: i2b2 UI respects REDCap access control**
  - **i2b2 projects configured real-time at i2b2 log-in**
- **Special case: REDCap project open to all i2b2 users**

Monthly transition not smooth.  
Design improvements pending

# SEINE DataBuilder: i2b2 -> REDCap

- Data Delivery for HERON i2b2 user community at KUMC
  - Pilot studies in 2012
- Federated Data collection in GPC
  - Breast Cancer
  - ALS



# Data Delivery via REDCap

- REDCap output project has
  - Demographics CRF (age, sex, race, etc.)  
From i2b2 patient dimension, plus
    - Diagnoses CRF if anything from i2b2 Diagnoses selected
    - Medications CRF
    - Lab Results CRF
    - ...
- The project can be organized either
  - by-patient or
  - by-encounter.

# SEINE DataBuilder: Diabetes + Vertigo Study

Pilot Experience:

- REDCap **Graphical View & Stats** sufficient for preliminary analysis
- Export to stats program



D'Silva et. al J Vestib Res. 2016  
PMCID:[PMC4791946](https://pubmed.ncbi.nlm.nih.gov/3141444/)

# Identified, de-identified data fulfillment

- ~40 deliveries/month
- A common case: an identified cohort
  - Variables: just MRN
  - the customer then adds CRFs

# Data Builder

1. Governance Committee OKs data request
2. Honest Broker uses Data Builder plug-in to specify data

The screenshot shows the 'Data Builder' interface within a 'DataFrameBuilder' window. At the top, there is a title bar and a note: 'Note: Access to [RStudio Server](#) is currently limited to members of the HERON study team.' Below the note, the 'Patient Set' is defined as 'Dementia Timeline [@12:39:37 [3-24-2017] [dconnolly] [PATIENTSET\_95579]'. A link to 'Running a Query' is provided. The 'Patient Data' section lists 'Birthdate, sex, vital status, race, etc. are automatically included.' The 'Other Observations' section lists several data points with fact and patient counts: '005- #12654 Verbal Fluency [2,034 facts; 1,610 patients]', '290.4 Vascular dementia [5,198 facts; 1,114 patients]', '331.0 Alzheimer's disease [49,384 facts; 5,525 patients]', 'Body Mass Index [4,211,058 facts; 518,384 patients]', 'Ethnicity [2,246,187 facts; 2,246,184 patients]', 'HEMOGLOBIN A1C(#2034) [427,215 facts; 90,820 patients]', and 'Marital Status [2,246,187 facts; 2,246,184 patients]'. The 'Modifiers to exclude' section contains a search box with the text '...finding relevant modifiers...'. The 'Date Range' section has two empty input fields separated by 'to'. There are two checkboxes: 'Include MRN?' and 'Include Contact Info?', both with '(only available in identified databuilder requests)' as a note. At the bottom, the 'Observations from Query' section shows the same patient set as the top: 'Dementia Timeline [@12:39:37 [3-24-2017] [dconnolly]'. The interface is clean with a light blue header and a white main area.

DataFrameBuilder

## Data Builder

Note: Access to [RStudio Server](#) is currently limited to members of the HERON study team.

Patient Set: Dementia Timeline [@12:39:37 [3-24-2017] [dconnolly] [PATIENTSET\_95579]

The [Running a Query](#) section shows how to create one.

Patient Data: Birthdate, sex, vital status, race, etc. are automatically included.

Other Observations:

- 005- #12654 Verbal Fluency [2,034 facts; 1,610 patients]
- 290.4 Vascular dementia [5,198 facts; 1,114 patients]
- 331.0 Alzheimer's disease [49,384 facts; 5,525 patients]
- Body Mass Index [4,211,058 facts; 518,384 patients]
- Ethnicity [2,246,187 facts; 2,246,184 patients]
- HEMOGLOBIN A1C(#2034) [427,215 facts; 90,820 patients]
- Marital Status [2,246,187 facts; 2,246,184 patients]

Modifiers to exclude: ...finding relevant modifiers...

Date Range: [ ] to [ ]

Include MRN?  (only available in identified databuilder requests)

Include Contact Info?  (only available in identified databuilder requests)

Observations from Query: Dementia Timeline [@12:39:37 [3-24-2017] [dconnolly]

# i2b2 Timeline: Specify Data

- Patient Set
  - Self-service cohort query
- Concepts

Timeline

Specify Data View Results Plugin Help

Drop a Patient Set and one or more Concepts (Ontology Terms) into the input boxes below, and the timeline showing when those concepts were observed in the selected patient set.

Patient Set: Patient Set for "Dementia Timeline [@12:39:37"

Concept(s):

003- #15324 Verbal Fluency [2,303 facts; 760 patients]
003- #95012446 Mini-Mental Status Examination [5,381 f
005- #12654 Verbal Fluency [2,034 facts; 1,610 patients]
006- #12655 Trailmaking A Time (secs) [1,680 facts; 1,44
007- #12658 Trailmaking A Errors [1,665 facts; 1,429 pati
008- #12657 Trailmaking B Time (secs) [1,610 facts; 1,37
008- #15319 MOCA Score (out of 30) [3,479 facts; 2,419 p
009- #12659 Trailmaking B Errors [1,592 facts; 1,361 pati
290.4 Vascular dementia [5,429 facts; 1,166 patients]
331.0 Alzheimer's disease [50,936 facts; 5,690 patients]

# i2b2 Timeline: View Results

LDS access role

The screenshot displays the i2b2 Timeline application interface. At the top, there is a 'Timeline' header with navigation icons. Below it are three tabs: 'Specify Data', 'View Results' (which is selected), and 'Plugin Help'. A search bar contains the text '<<< start: 1 size: 10 go >>>' and a date range from 12/14/1923 to 1/9/2017. A zoom control shows '- + pan: < >'. The main content area lists search results for 'Person #44206 m 67yroid White'. An orange arrow points from the 'LDS access role' text in the top right to the 'Person #44206' entry. The results for Person #44206 include:

- 003- #15324 Verbal Fluency [2,303 facts; 760 patients]
- 008- #15319 MOCA Score (out of 30) [3,479 facts; 2,419 patients]
- 332 Parkinson's disease [165,266 facts; 8,114 patients]
- Age [2,222,790 facts; 2,222,787 patients]
- Body Mass Index [4,211,058 facts; 518,384 patients]
- Ethnicity [2,246,187 facts; 2,246,184 patients]
- G20 Parkinson's disease [86,381 facts; 4,624 patients]
- Gender [2,246,187 facts; 2,246,184 patients]
- Marital Status [2,246,187 facts; 2,246,184 patients]
- Race [2,246,560 facts; 2,246,184 patients]

Below this, the start of results for 'Person #526833 f 82yroid White' is visible:

- 003- #95012446 Mini-Mental Status Examination [5,381 facts; 2,096 patients]
- 005- #12654 Verbal Fluency [2,034 facts; 1,610 patients]
- 006- #12655 Trailmaking A Time (secs) [1,680 facts; 1,442 patients]
- 007- #12658 Trailmaking A Errors [1,665 facts; 1,429 patients]
- 008- #12657 Trailmaking B Time (secs) [1,610 facts; 1,377 patients]
- 009- #12659 Trailmaking B Errors [1,592 facts; 1,361 patients]
- 331 0 Alzheimer's disease [50,936 facts; 5,690 patients]

# REDCap records $\sim$ spreadsheet rows

Subject	Gender	Age	...	
4206	m	67		
833	f	82		

# Many i2b2 facts -> 1 spreadsheet cell

Subject	Gender	Age	Verbal Fluency	BMI
4206	m	67	???	???
833	f	82	???	???

**Timeline**

Specify Data | View Results | Plugin Help

<<< start: 1 size: 10 go >>> zoom: - + pan: < >

12/14/1923 6/28/1970 1/9/2017

**Person #44206 m 67yroid White**

- 003- #15324 Verbal Fluency [2,303 facts; 760 patients]
- 008- #15319 MOCA Score (out of 30) [3,479 facts; 2,419 patients]
- 332 Parkinson's disease [165,266 facts; 8,114 patients]
- Age [2,222,790 facts; 2,222,787 patients]
- Body Mass Index [4,211,058 facts; 518,384 patients]
- Ethnicity [2,246,187 facts; 2,246,184 patients]
- G20 Parkinson's disease [86,381 facts; 4,624 patients]
- Gender [2,246,187 facts; 2,246,184 patients]
- Marital Status [2,246,187 facts; 2,246,184 patients]
- Race [2,246,560 facts; 2,246,184 patients]

**Person #526833 f 82yroid White**

- 003- #95012446 Mini-Mental Status Examination [5,381 facts; 2,096 patients]
- 005- #12654 Verbal Fluency [2,034 facts; 1,610 patients]
- 006- #12655 Trailmaking A Time (secs) [1,680 facts; 1,442 patients]
- 007- #12658 Trailmaking A Errors [1,665 facts; 1,429 patients]
- 008- #12657 Trailmaking B Time (secs) [1,610 facts; 1,377 patients]
- 009- #12659 Trailmaking B Errors [1,592 facts; 1,361 patients]

# Technique 1: Aggregation

patient_num	...	BMI: count	BMI: first	BMI: last	BMI: median	...
4206		19	23.5	24.2		
833		4	25.1	24.8		

The screenshot shows a software interface with a 'Timeline' tab. Below the tab are buttons for 'Specify Data', 'View Results', and 'Plugin Help'. A navigation bar includes 'start: 1', 'size: 10', 'go', '>>>', 'zoom: -', 'par: <', and '>'. A timeline bar shows dates from 12/14/1923 to 12/9/2011. The main area displays patient data for two individuals:

**Person #44206 m 67yroid White**

- 003- #15324 Verbal Fluency [2,303 facts; 760 patients]
- 008- #15319 MOCA Score (out of 30) [3,479 facts; 2,419 patients]
- 332 Parkinson's disease [165,266 facts; 8,114 patients]
- Age [2,222,790 facts; 2,222,787 patients]
- Body Mass Index [4,211,058 facts; 518,384 patients]
- Ethnicity [2,246,187 facts; 2,246,184 patients]
- G20 Parkinson's disease [86,381 facts; 4,624 patients]
- Gender [2,246,187 facts; 2,246,184 patients]
- Marital Status [2,246,187 facts; 2,246,184 patients]
- Race [2,246,560 facts; 2,246,184 patients]

**Person #526833 f 82yroid White**

- 003- #95012446 Mini-Mental Status Examination [5,381 facts; 2,096 patients]
- 005- #12654 Verbal Fluency [2,034 facts; 1,610 patients]
- 006- #12655 Trailmaking A Time (secs) [1,680 facts; 1,442 patients]
- 007- #12658 Trailmaking A Errors [1,665 facts; 1,429 patients]
- 008- #12657 Trailmaking B Time (secs) [1,610 facts; 1,377 patients]
- 009- #12659 Trailmaking B Errors [1,592 facts; 1,361 patients]

# Technique 2: Record-per-encounter

encounter_num	patient_num	Age	start_date	...
934875	4206	67	2017-01-02	
374830	4206	67	2017-01-09	
302975	833	82	2017-01-08	

**Timeline**

Specify Data | View Results | Plugin Help

<<< start: 1 size: 10 go >>> zoom: - + pan: < >

12/14/1923 6/28/1970 1/9/2017

**Person #44206 m 67yroid\_White**

- 003- #15324 Verbal Fluency [2,303 facts; 760 patients]
- 008- #15319 MOCA Score (out of 30) [3,479 facts; 2,419 patients]
- 332 Parkinson's disease [165,266 facts; 8,114 patients]
- Age [2,222,790 facts; 2,222,787 patients]
- Body Mass Index [4,211,058 facts; 518,384 patients]
- Ethnicity [2,246,187 facts; 2,246,184 patients]
- G20 Parkinson's disease [86,381 facts; 4,624 patients]
- Gender [2,246,187 facts; 2,246,184 patients]
- Marital Status [2,246,187 facts; 2,246,184 patients]
- Race [2,246,560 facts; 2,246,184 patients]

**Person #526833 f 82yroid\_White**

- 003- #95012446 Mini-Mental Status Examination [5,381 facts; 2,096 patients]
- 005- #12654 Verbal Fluency [2,034 facts; 1,610 patients]
- 006- #12655 Trailmaking A Time (secs) [1,680 facts; 1,442 patients]
- 007- #12658 Trailmaking A Errors [1,665 facts; 1,429 patients]
- 008- #12657 Trailmaking B Time (secs) [1,610 facts; 1,377 patients]
- 009- #12659 Trailmaking B Errors [1,592 facts; 1,361 patients]

# GPC Federated Query: Breast Cancer Cohort

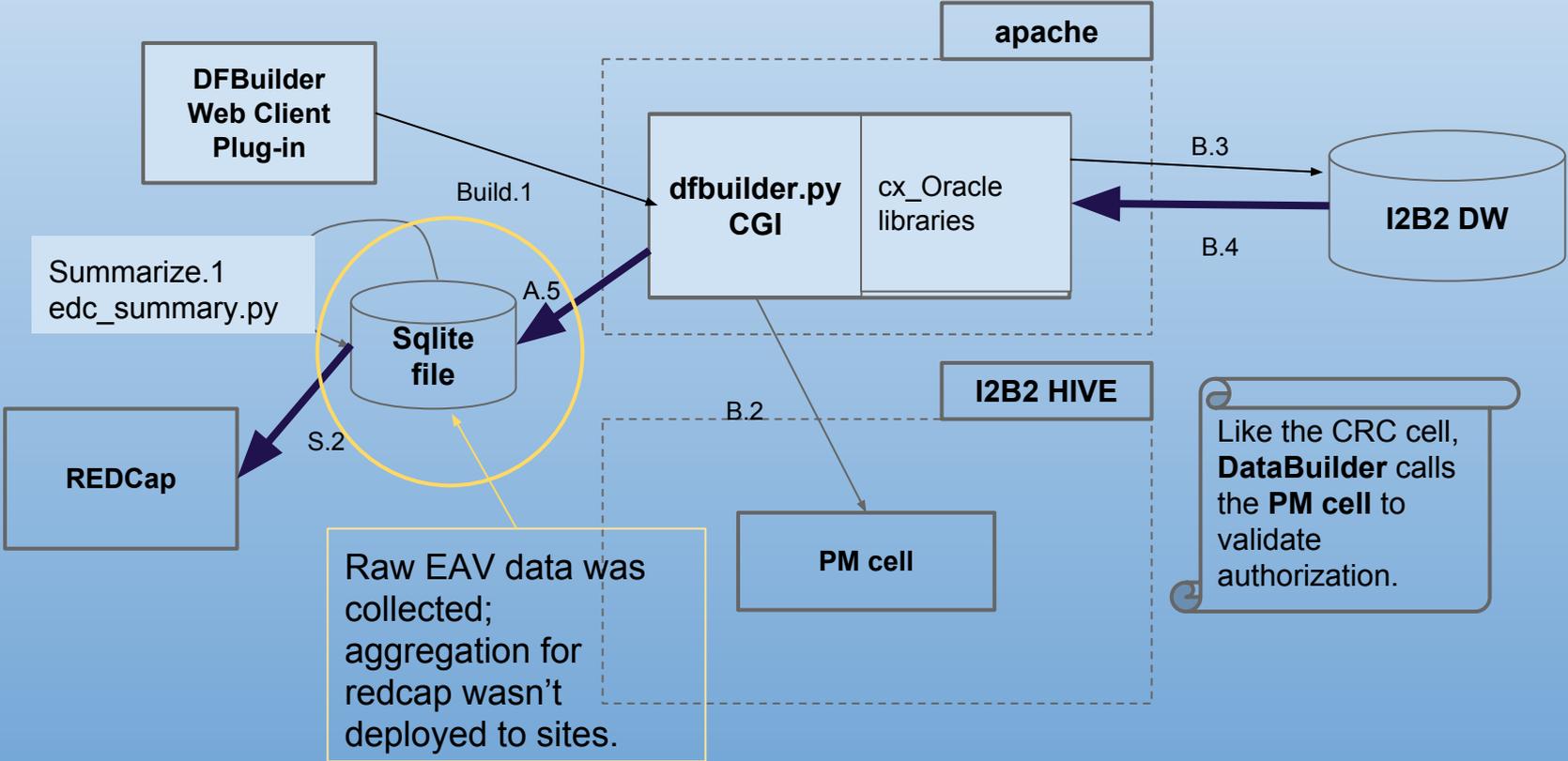
- GPC common condition = Breast Cancer
- 2015: Federated query to define survey cohort
  - Putting RC11 from the [GPC Proposal](#) into practice
- SEINE DataBuilder method used at 8 sites
  - ~6 sites used common python code
  - ~2 sites developed work-alikes based on specs in GPC Wiki



# Federating raw EAV data

- Initial BC query:
  - raw files sent to KUMC for aggregation into a spreadsheet
  - QC was labor intensive
    - Lots of round-trips

# DataBuilder Architecture: 2012-2013



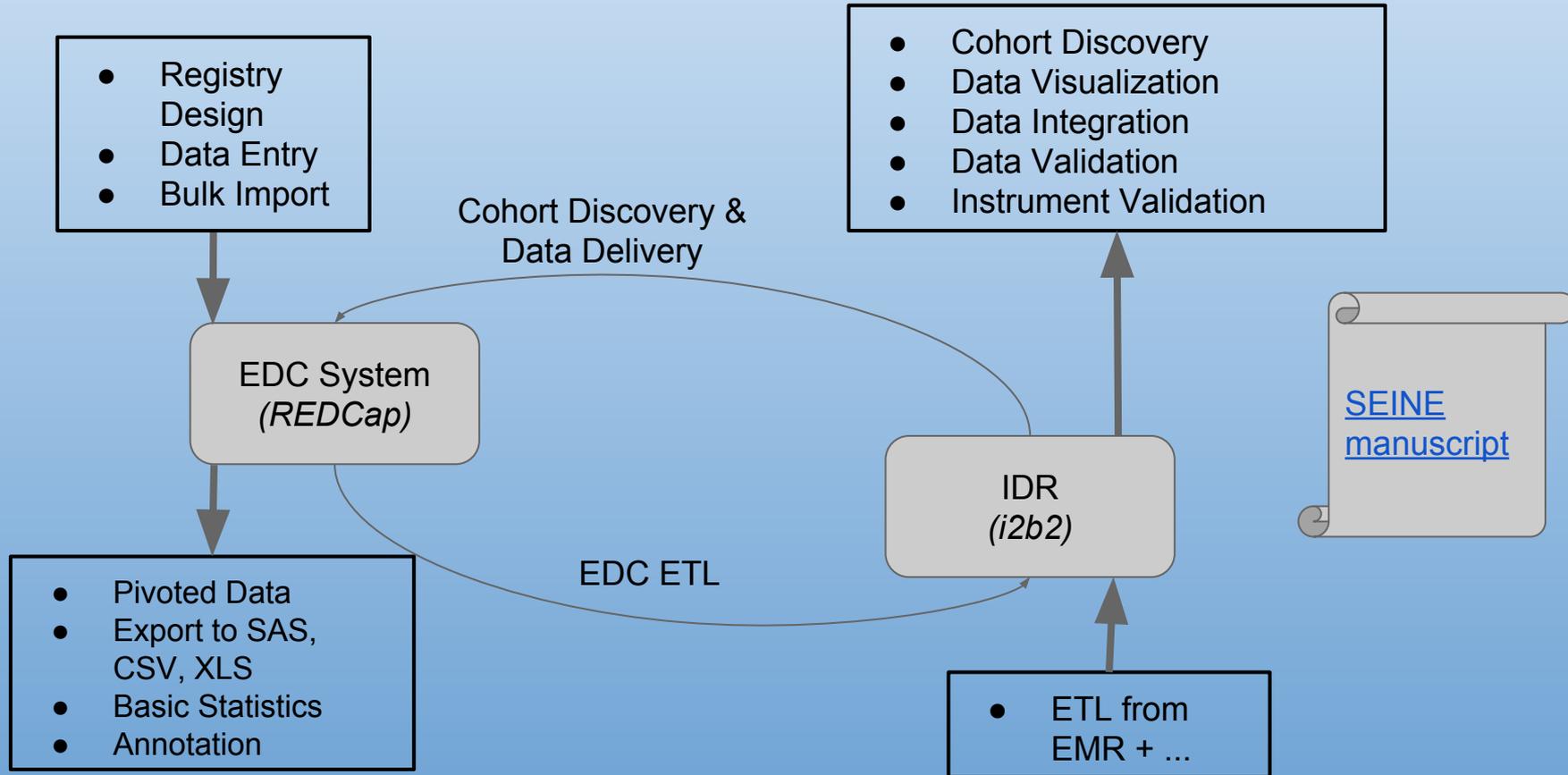
# REDCap as gatekeeper

- Initial BC query:
  - raw files sent to KUMC for aggregation into a spreadsheet
  - QC was labor intensive
    - Lots of round-trips
- Later queries (ALS, 2nd BC):
  - raw files were pivoted on-site before submission
  - REDCap data dictionary constraints provided much/most of the QC

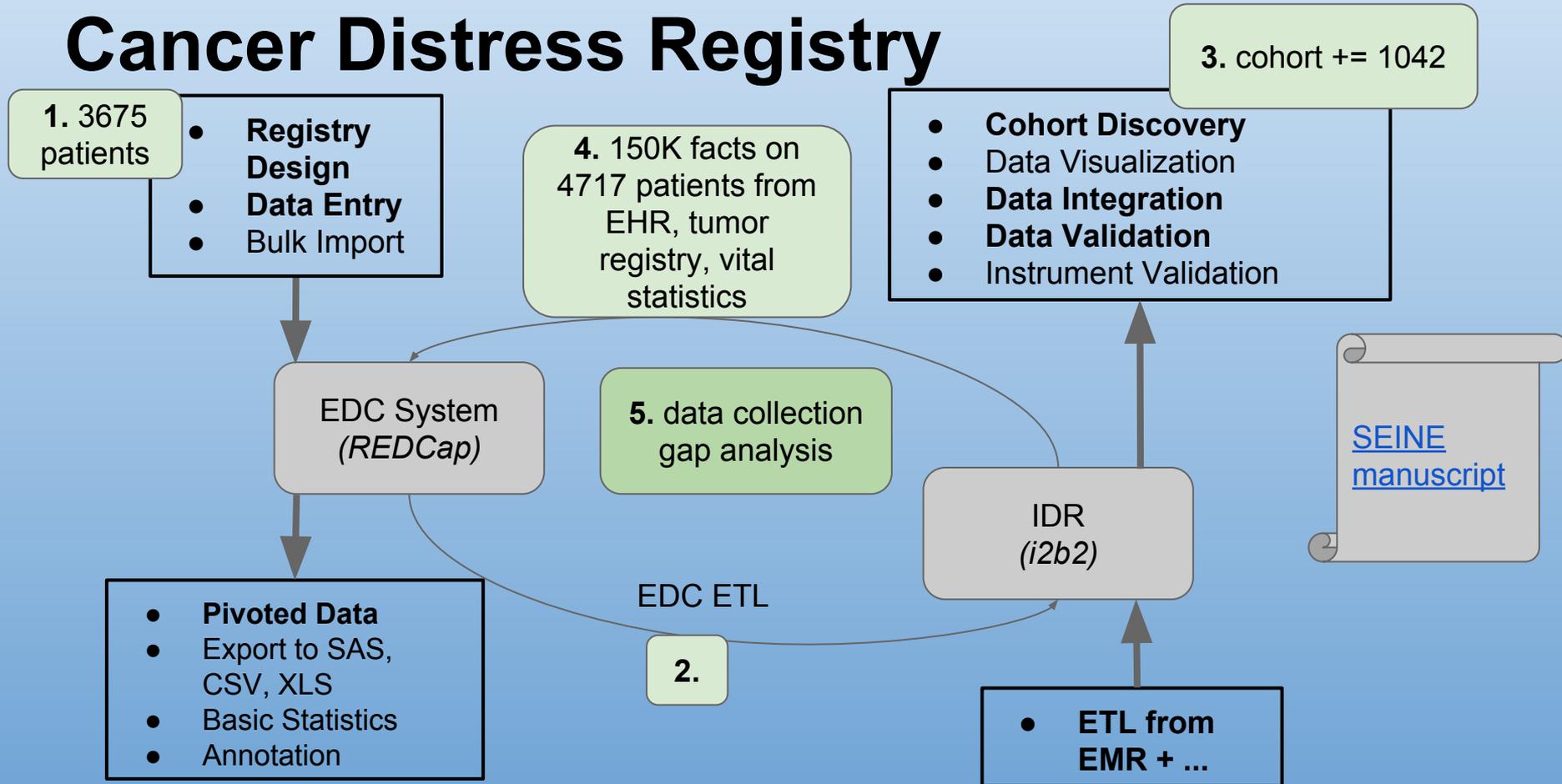
Custom  
EAV->spreadsheet  
pivoting code;  
not generic SEINE  
edc\_summary.py code



# SEINE = REDCap<-i2b2 + REDCap->i2b2



# SEINE Full Cycle Case: Cancer Distress Registry



# DataBuilder scalability

- 10 or 20 minutes is typical at KUMC
  - i2b2 star schema on solid-state storage
- $\geq 30K$  patients stresses the system
  - But does not break it

# Support: open gpc-dev process



- [DataBuilder](#) in [informatics.gpcnetwork.org](http://informatics.gpcnetwork.org) wiki ->
  - SEINE Manuscript
  - **heron\_extract** hg repository

- “if it breaks, you get to keep both pieces” ;)
- gpc-dev mailing list is open
- weekly teleconferences, annual meetings accommodate occasional guests



# UTSW Success Story

“Huge time saver”  
— T. Bosler

- Used i2b2 to select a cohort of 2000 patients (out of 5M from i2b2 repository).
  - 20 additional observations along with Diagnoses, Procedures
  - Took ~3 hrs to run the DataBuilder
- Some post-processing for MRI etc. into REDCap.
- Just 4hrs total.

