Studying Temporal Phenomenon through Maintenance, Abstraction and Analysis of Temporal Data:
Case Study of HIV Drug Resistance Research

Amar K. Das, MD, PhD

Assistant Professor

Departments of Medicine and of Psychiatry and Behavioral Sciences

Stanford University

Stanford Center for Biomedical Informatics Research

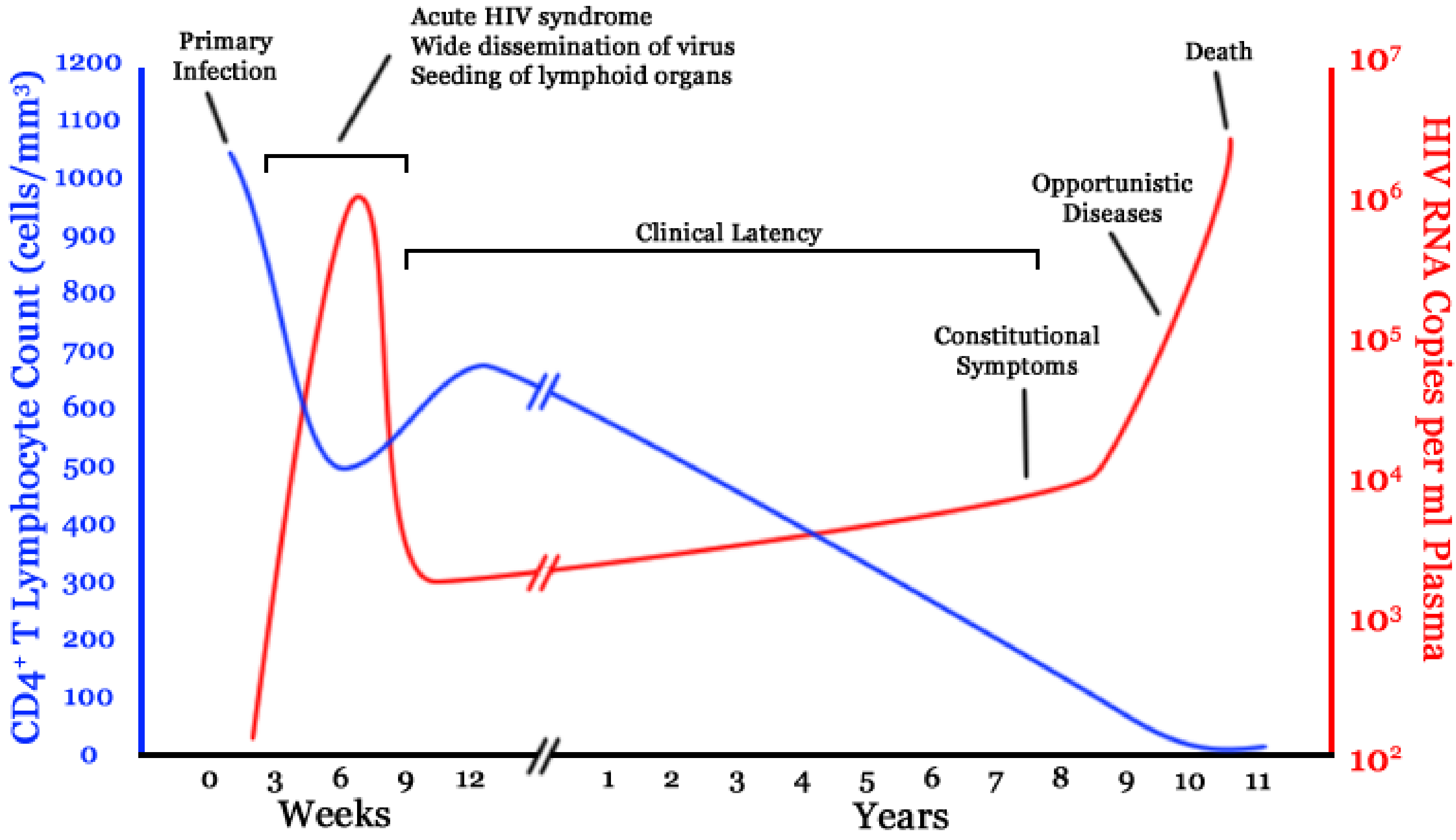
Goals of Presentation

- To understand how to best model data and knowledge for studying temporal phenomenon
- To understand how such temporal information can be used to support the maintenance, abstraction and analysis of research data

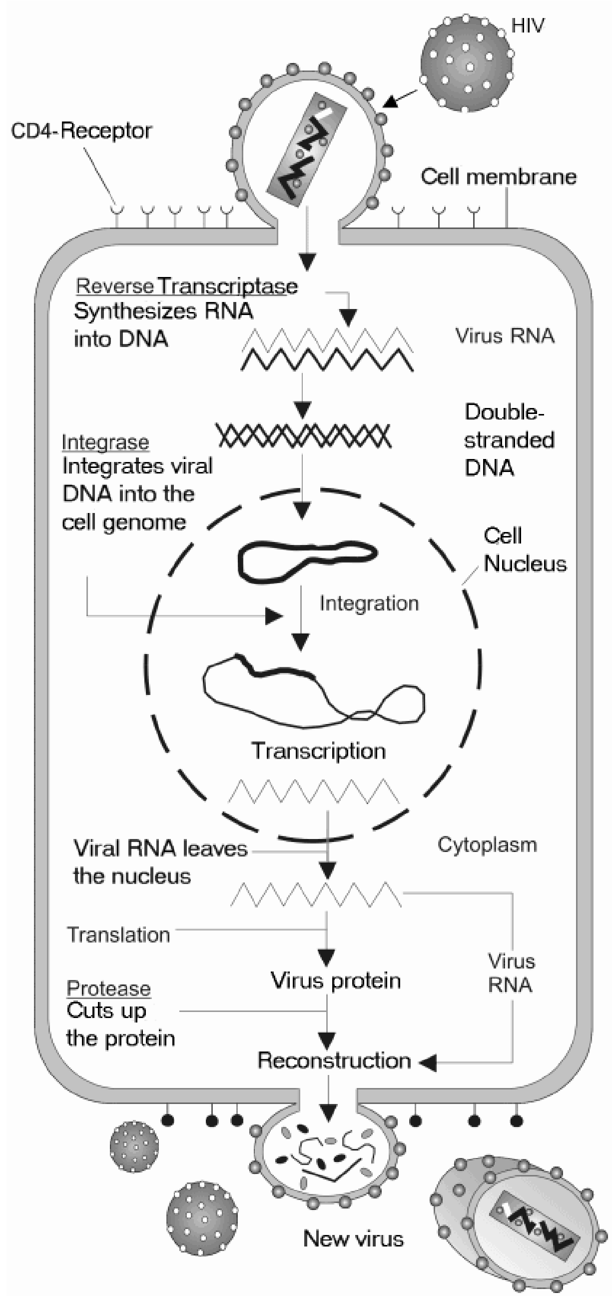
Case Study

- Drug-resistant HIV mutations are a major obstacle to successful antiretroviral treatment
- Routine genotype testing is recommended before selecting a treatment regimen
- Test-result interpretation is challenging

Time Course of HIV Infection

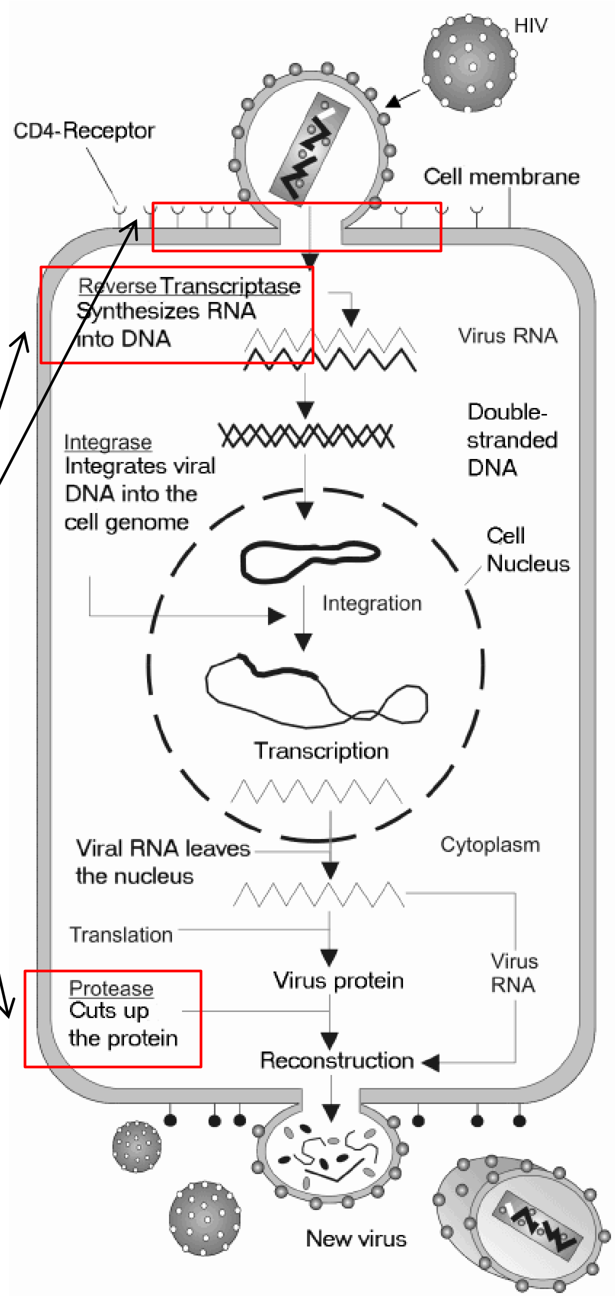


Stanford Center for
Biomedical Informatics Research



Stanford Center for
Biomedical Informatics Research

Current targets of antiretroviral therapy

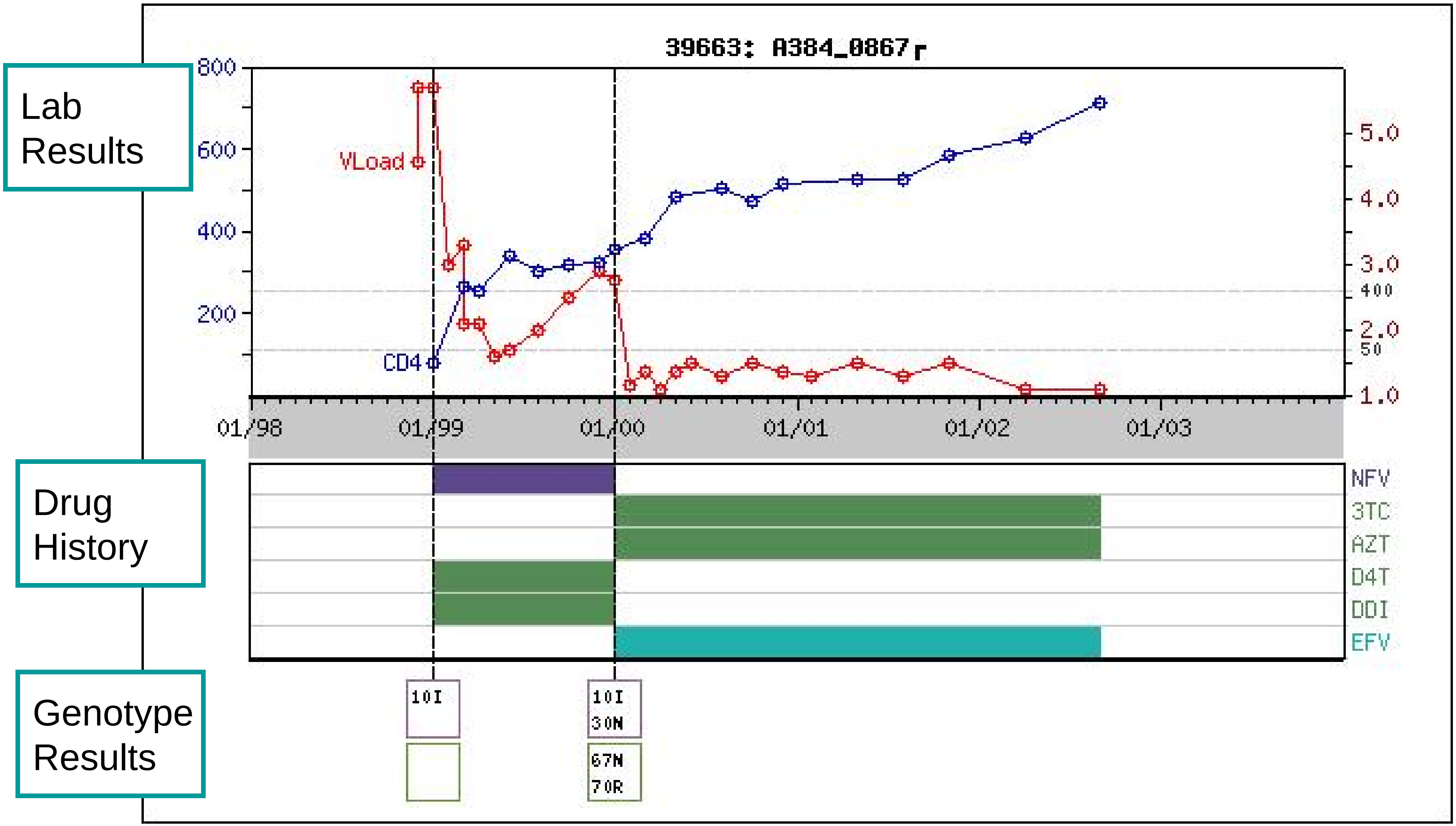


Stanford Center for
Biomedical Informatics Research

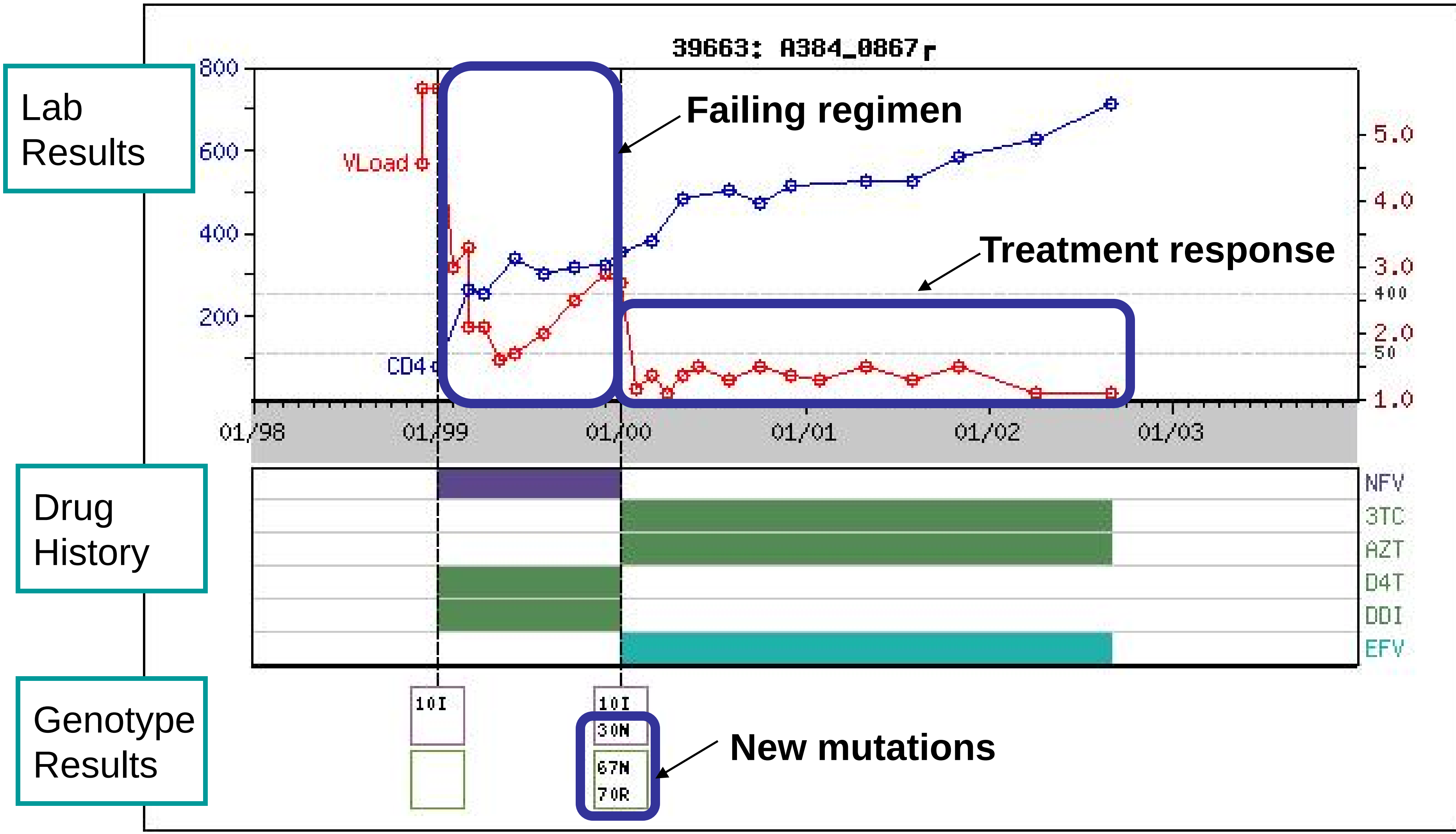
Antiretroviral Therapy

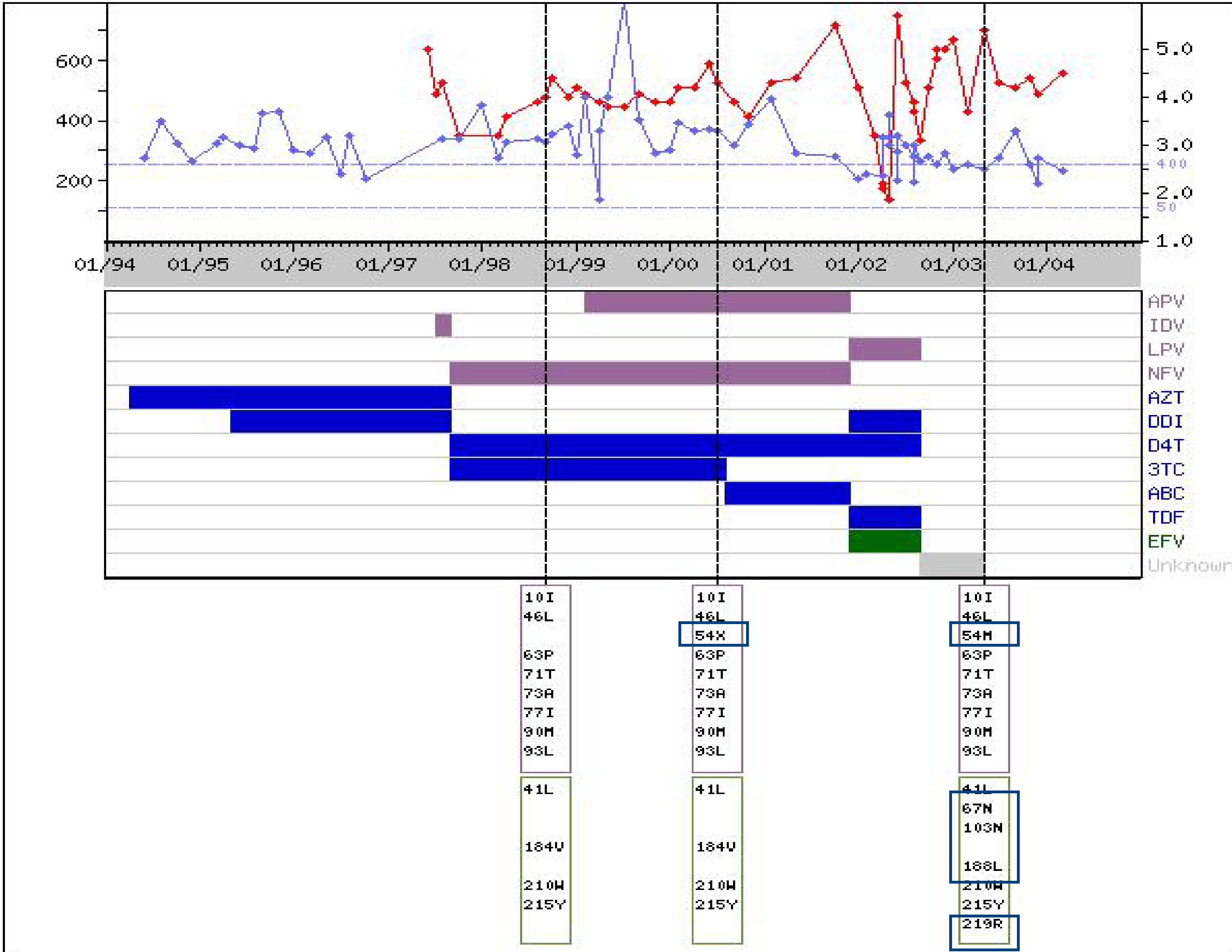
- Protease Inhibitors (PI)
  - Atazanavir (ATV) · Darunavir (DRV) · Fosamprenavir (fAPV) · Indinavir (IDV) · Lopinavir (LPV) · Nelfinavir (NFV) · Saquinavir (SQV) · Tipranavir (TPV)
- Nucleoside Reverse Transcriptase Inhibitors (NRTI)
  - Abacavir (ABC) · Didanosine (ddI) · Emtricitabine (FTC) · Lamivudine (3TC) · Stavudine (d4T) · Tenofovir (TDF) · Zidovudine (AZT)
- Non-Nucleoside Reverse Transcriptase Inhibitors (NNRTI)
  - Delavirdine (DLV) · Efavirenz (EFV) · Etravirine (ETR) · Nevirapine (NVP)
- Fusion Inhibitor (FI)
  - Enfuvirtide (ENF)

Data on Drug Resistance



Patterns of Drug Resistance




600
400
200
5.0
4.0
3.0
400
2.0
50
1.0
01/94
01/95
01/96
01/97
01/98
01/99
01/00
01/01
01/02
01/03
01/04
APV
IDV
LPV
NFV
AZT
DDI
D4T
3TC
ABC
TDF
EFV
Unknown
10I 46L 63P 71T 73A 77I 90M 93L
41L 184V 210W 215Y
10I 46L 54X 63P 71T 73A 77I 90M 93L
41L 184V 210W 215Y
10I 46L 54M 63P 71T 73A 77I 90M 93L
41L 67N 103N 188L 210W 215Y 219R

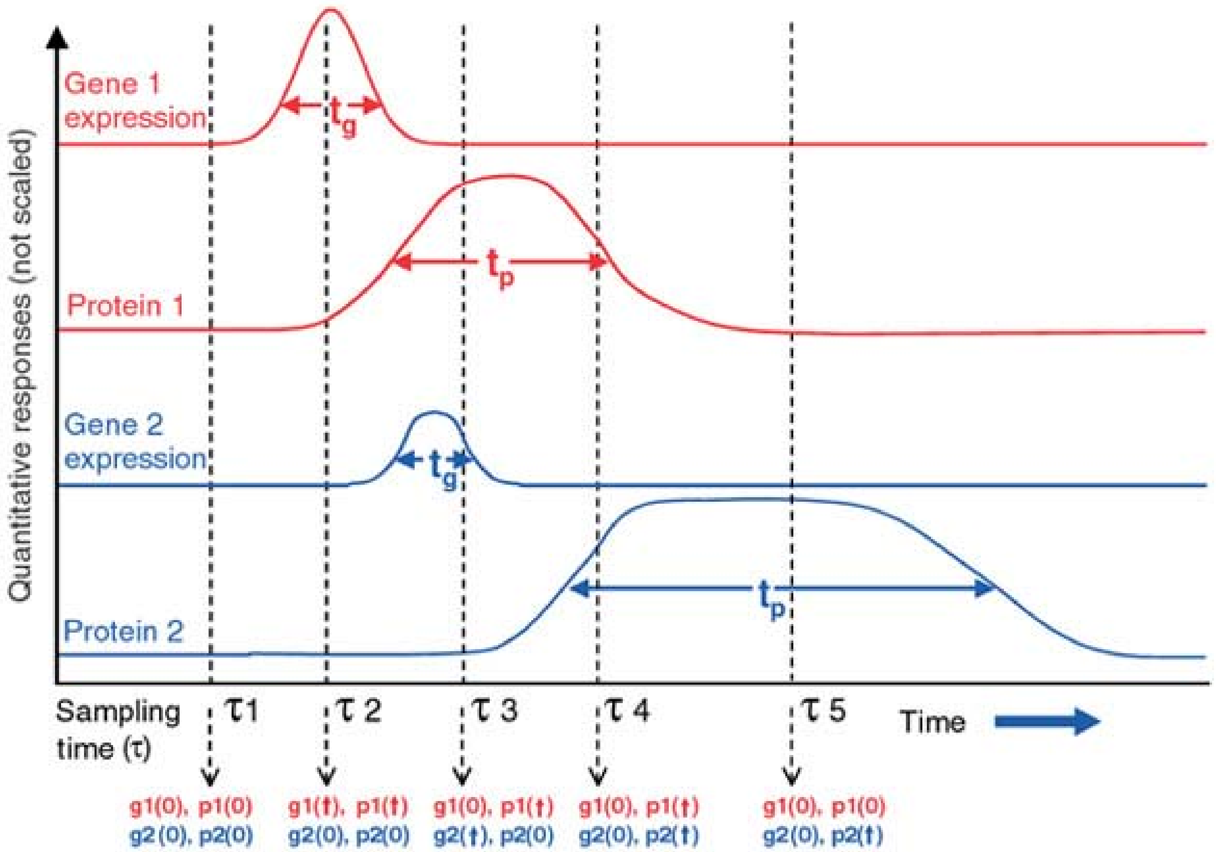
Stanford HIV Database

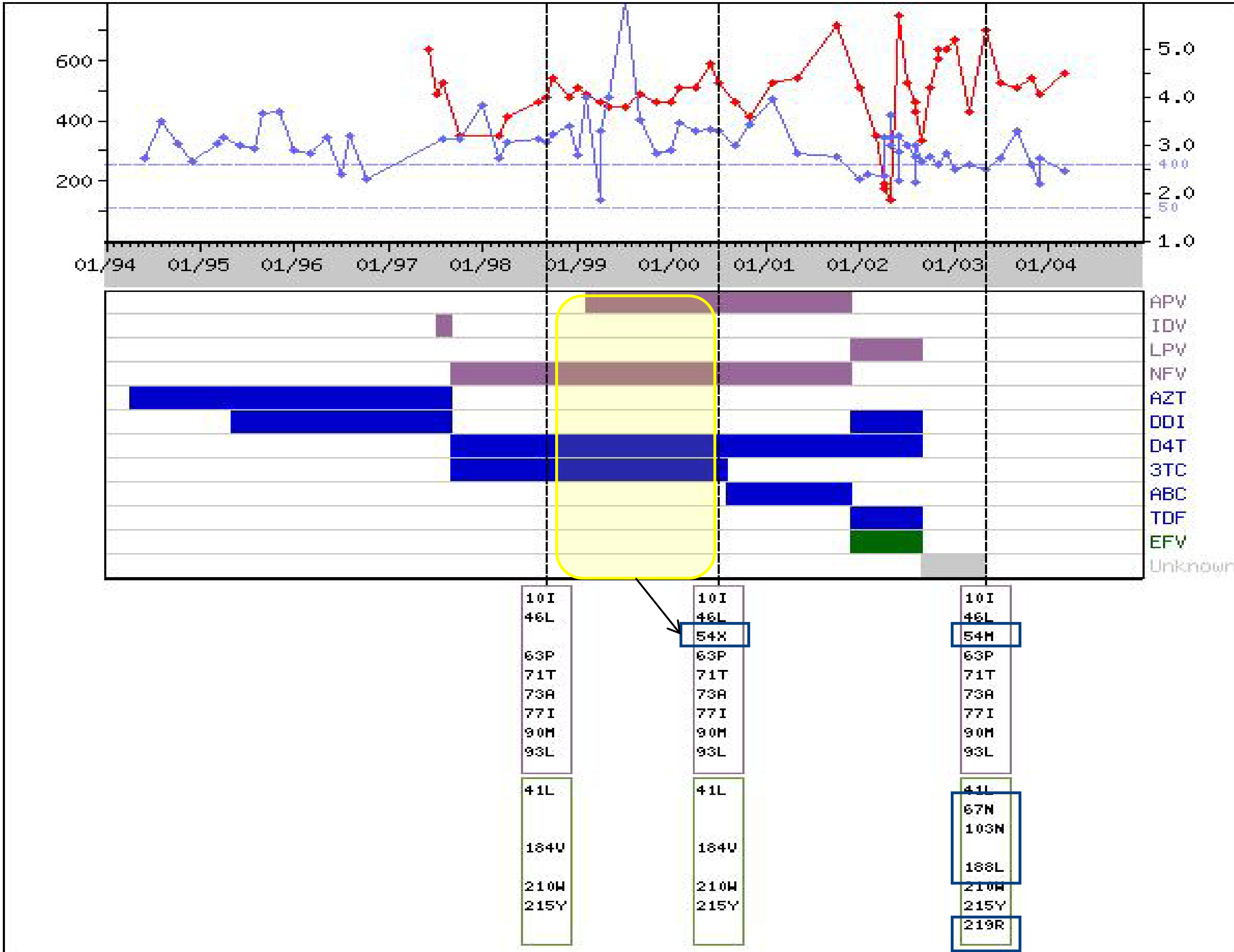
- Publicly available repository on HIV drug resistance data about patients in clinical practice
- Research records link treatment histories with outcome data and genotype-test results
  - ~18,000 subjects
  - ~42,000 genotype test results
  - ~850,000 mutations

Temporal Phenomena in HIV Drug Resistance

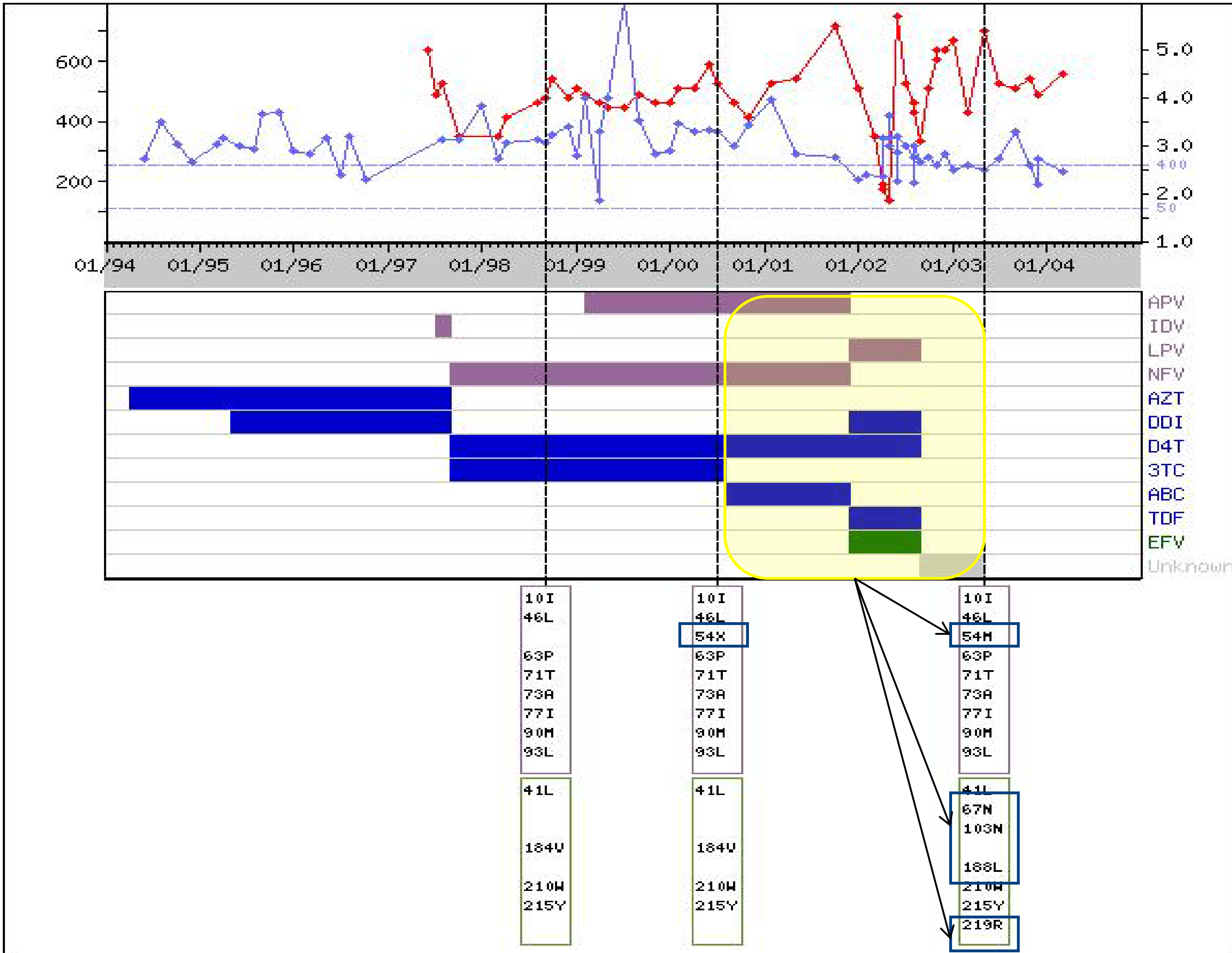
- Sequence of treatment regimens
- Patterns of drug resistance
- Appearance of new mutations

Sampling of Temporal Phenomena





600
400
200
5.0
4.0
3.0
400
2.0
50
1.0
01/94
01/95
01/96
01/97
01/98
01/99
01/00
01/01
01/02
01/03
01/04
APV
IDV
LPV
NFV
AZT
DDI
D4T
3TC
ABC
TDF
EFV
Unknown
10I
46L
63P
71T
73A
77I
90M
93L
41L
184V
210W
215Y
10I
46L
54X
63P
71T
73A
77I
90M
93L
41L
184V
210W
215Y
10I
46L
54M
63P
71T
73A
77I
90M
93L
41L
67N
103N
188L
210W
215Y
219R



600
400
200
5.0
4.0
3.0
400
2.0
50
1.0
01/94
01/95
01/96
01/97
01/98
01/99
01/00
01/01
01/02
01/03
01/04
APV
IDV
LPV
NFV
AZT
DDI
D4T
3TC
ABC
TDF
EFV
Unknown
10I
46L
63P
71T
73A
77I
90M
93L
41L
184V
210W
215Y
10I
46L
54X
63P
71T
73A
77I
90M
93L
41L
184V
210W
215Y
10I
46L
54M
63P
71T
73A
77I
90M
93L
41L
67N
103N
188L
210W
215Y
219R

Modeling Temporal Data and Knowledge

- Persistence
  - – Instance-based vs. interval-based data
- Composition
  - – Regimens consist of individual drugs
- Classification
  - – Patterns of resistance

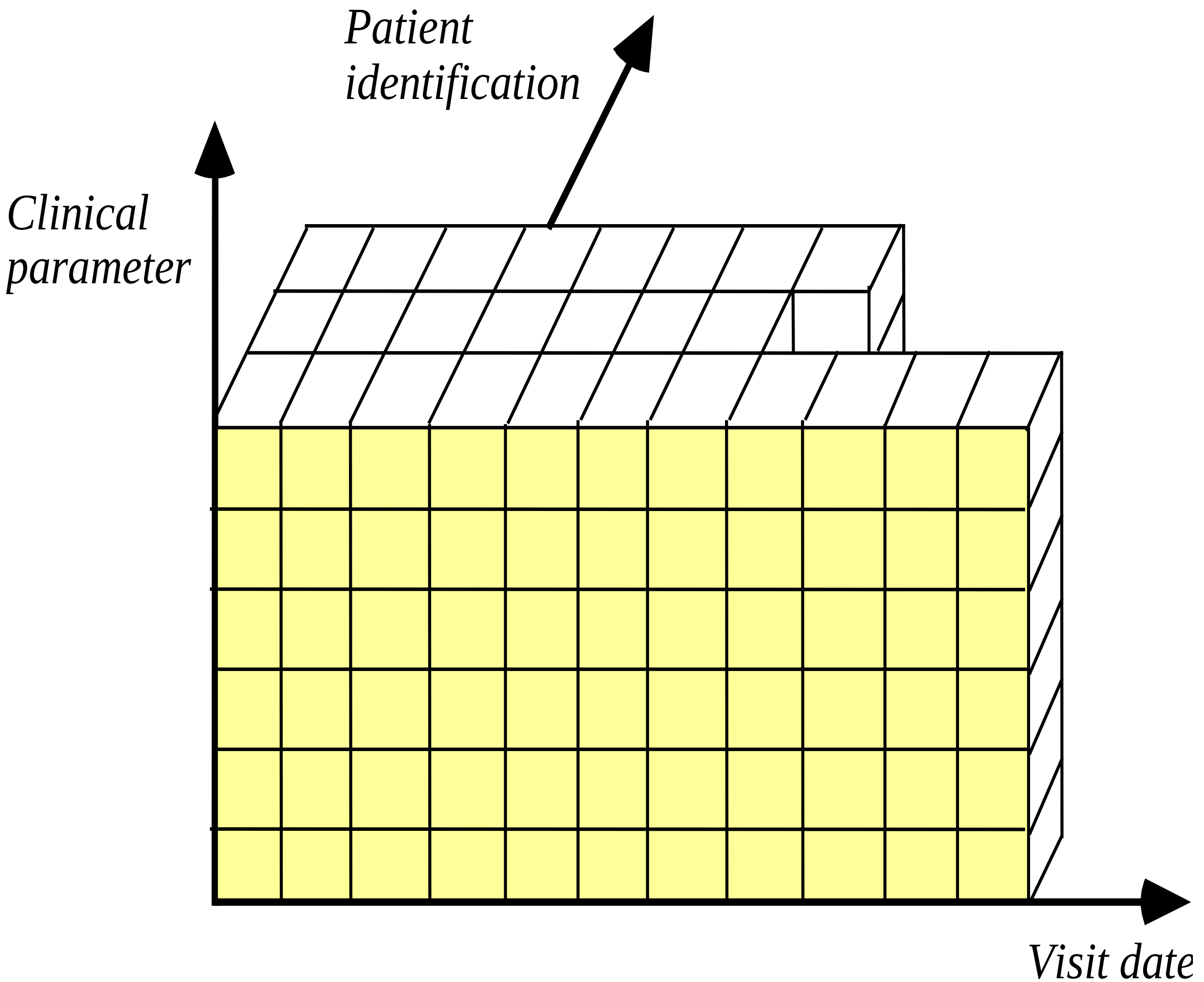
Temporal Persistence

- Instant-based vs. interval-based representation
- Temporal relational model

Instant-Based Model

- **Time-Oriented Database Model**
  - Support queries on longitudinal clinical research data
  - Uses a three-dimensional view of medical data (patient identification, clinical parameter, and date of visit)

Instant-Based Model



Interval-Based Model

- **Temporal Relational Model**
  - Supports instant-based and interval-based temporal representations
  - Allows temporal querying using arbitrarily complex patterns

Standard Relational Model

Time Drawn	Time Entered	MRNO	VL
Jun 12, 2009 6:50:23AM	Jun 12, 2009 9:25:47AM	123-4567	2.5
Sep 13, 2009 7:33:54AM	Sep 13, 2009 9:41:12AM	123-4567	2.7

Standard Relational Model

Time Started	Duration	MRNO	Drug Name	Drug Dose
Aug 1, 2009 12:00AM	31	123-4567	AZT	600
Sep 1, 2009 12:00AM	30	123-4567	AZT	800

Standard Relational Model

Time Drawn	Time Entered	MRNO	VL
Jun 12, 2009 6:50:23AM	Jun 12, 2009 9:25:47AM	123-4567	2.5
Sep 13, 2009 7:33:54AM	Sep 13, 2009 9:41:12AM	123-4567	2.7

Time Started	Duration	MRNO	Drug Name	Drug Dose
Aug 1, 2009 12:00AM	31	123-4567	AZT	600
Sep 1, 2009 12:00AM	30	123-4567	AZT	800

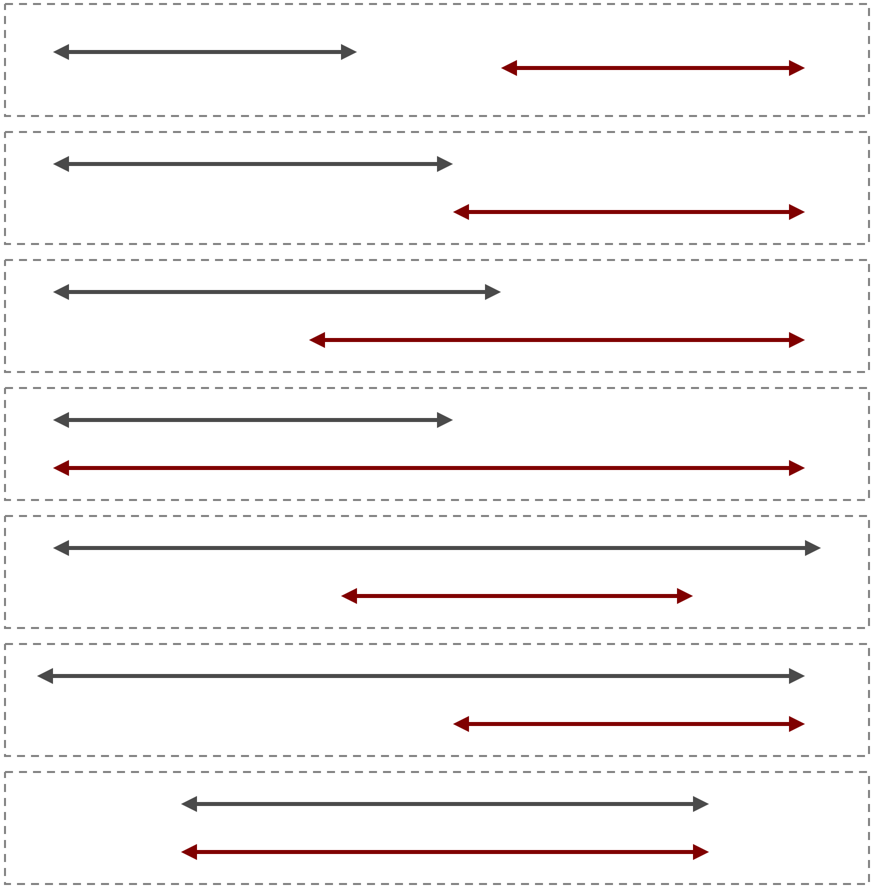
Temporal Relational Model

Time Started	Duration	MRNO	Drug Name	Drug Dose
Aug 1, 2009 12:00AM	31	123-4567	AZT	600
Sep 1, 2009 12:00AM	30	123-4567	AZT	800

Start Time	End Time	MRNO	Drug Name	Drug Dose
Aug 1, 2009	Aug 31, 2009	123-4567	AZT	600
Sep 1, 2009	Sep 30, 2009	123-4567	AZT	800

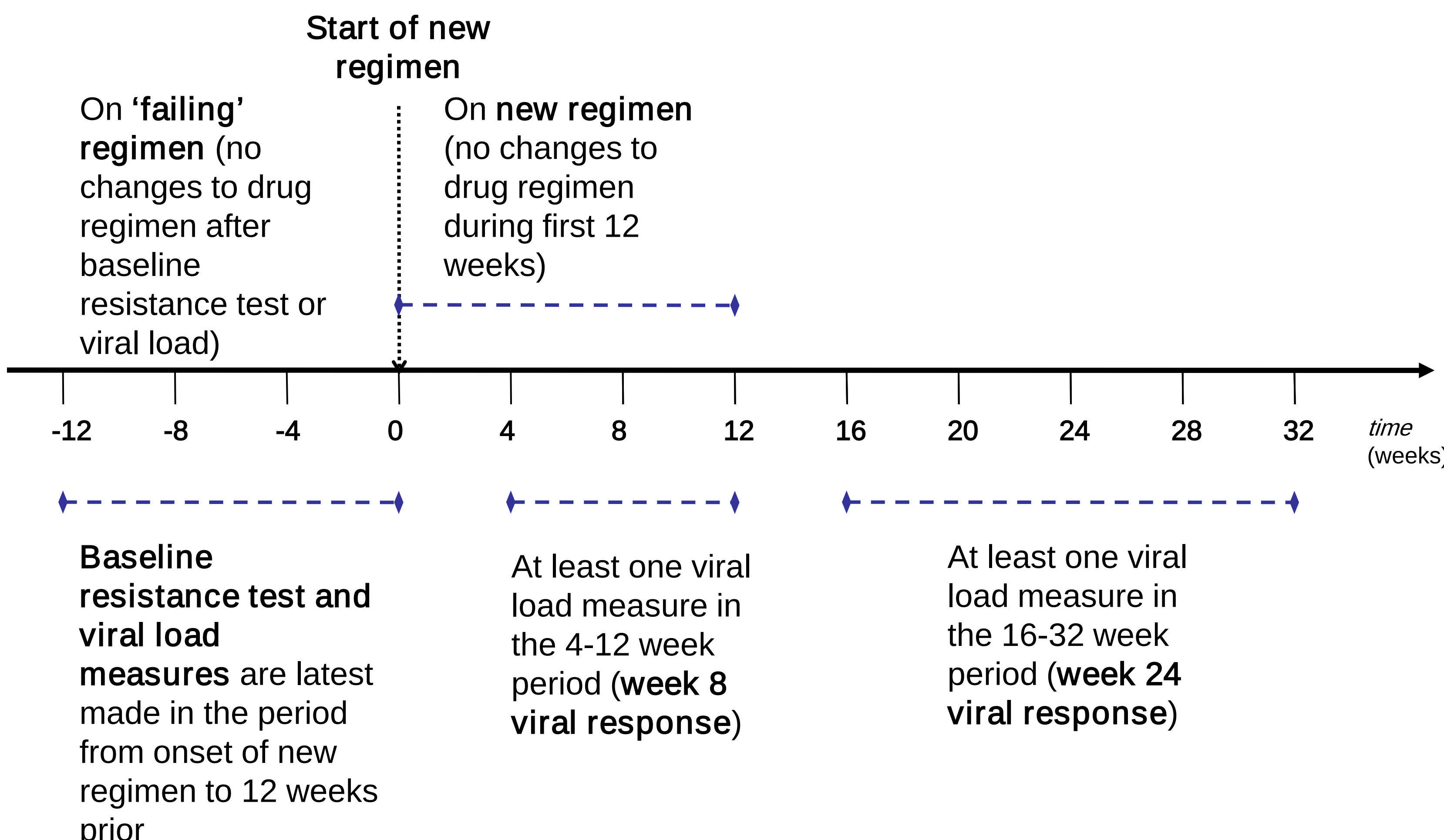
Start Time	End Time	MRNO	Drug Name
Aug 1, 2009	Sep 30, 2009	123-4567	AZT

Allen’s Interval Logic



Stanford Center for
Biomedical Informatics Research

Example Pattern: TCE



(adapted from Phillips et al., 2005)

Temporal Extensions to SQL

```
TEMPORAL SELECT     T1.ID, T2.Regimen, LAST(START(T2)) AS newRegimenStart,
        LAST(FINISH(S)) AS baselineSequenceDate, LAST(VALID(V1)) AS baselineVLoad,
        V2.VLoad, V3.VLoad
FROM        Treatment AS T1, // Will contain pre baseline non-DDI regimen
    Treatment AS T2, // Will contain the new DDI regimen lasting at least 12 weeks
    Sequen AS S, // Will contain sequence data within 12 weeks of new regimen start
    RNA AS V1,    // Will contain viral loads within 12 weeks of new regimen start
    RNA AS V2,   // Will contain viral loads between 4 and 12 weeks after start of new regimen
    RNA AS V3   // Will contain viral loads between 16 and 32 weeks after start of new regimen
    WHERE   S.ID = T1.ID AND S.ID = T2.ID AND S.ID = V1.ID AND
    T1.Regimen NOT LIKE '%DDI%' AND T2.Regimen LIKE '%DDI%' AND V1.VLoad >= 500
WHEN        DURATION(T2, 'weeks') >= 12 AND
    BEFORE (S1, T2) AND DURATION(FINISH(S1), START(T2), 'weeks') < 12 AND
    BEFORE(V1, T2) AND DURATION(V1, START(T2), 'weeks') < 12 AND
    CONTAINS(PERIOD(START(T2) + ‘weeks(4)’, START(T2)+’weeks(12)’), V2) AND
    CONTAINS(PERIOD(START(T2) + ‘weeks(16)’, START(T2)+’weeks(32)’), V3)
GROUP BY            ID; X, Y
SUCH THAT           X .ID = ID, Y .ID = ID
HAVING      MIN(DURATION (X .VALID (V2) , X.FIRST(START(T2)) + ‘weeks(8)’)) OR
            MIN(DURATION (Y .VALID (V3) , Y.FIRST(START(T2)) + ‘weeks(24)’))
```

Temporal Composition

7655	3TC+D4T+SQV	1996-07-01 00:00:00	1997-01-01 00:00:00
7655	3TC+D4T+RTV	1997-01-01 00:00:00	1997-06-01 00:00:00
7655	3TC+D4T+IDV	1997-06-01 00:00:00	1998-01-01 00:00:00
7656	3TC+D4T+RTV	1995-07-01 00:00:00	1996-07-01 00:00:00
7656	3TC+D4T+IDV	1996-07-01 00:00:00	1998-08-01 00:00:00
7657	AZT+DDC+SQV	1996-07-01 00:00:00	1996-09-01 00:00:00
7657	DDI+D4T+SQV	1996-12-01 00:00:00	1997-02-01 00:00:00
7657	D4T+3TC+IDV	1997-02-01 00:00:00	1997-11-01 00:00:00
7660	3TC+AZT+D4T+DDI+IDV+RTV+SQV	1996-09-11 00:00:00	1997-09-11 00:00:00
7660	NVP+ABC+D4T+NFV	1997-09-11 00:00:00	1998-01-01 00:00:00
7661	3TC+AZT+D4T+DDC+RTV+SQV	1996-11-06 00:00:00	1997-11-06 00:00:00
7661	NVP+ABC+NFV	1997-11-06 00:00:00	1998-01-01 00:00:00
7663	3TC+AZT+D4T+DDI+IDV+SQV	1996-11-06 00:00:00	1997-11-06 00:00:00
7663	NVP+ABC+D4T+RTV+SQV	1997-11-06 00:00:00	1998-01-01 00:00:00
7664	3TC+AZT+IDV+RTV	1996-11-06 00:00:00	1997-11-06 00:00:00

Temporal Composition

Start Time	End Time	MRNO	Drug Name
Aug 1, 2009	Sep 30, 2009	123-4567	AZT
Aug 1, 2009	Sep 30, 2009	123-4567	3TC
Aug 1, 2009	Sep 30, 2009	123-4567	NFV

Temporal Composition

Start Time	End Time	MRNO	AZ T	3TC	…	NFV
Aug 1, 2009	Sep 30, 2009	123-4567	1	1		1
Oct 1, 2009	Dec 31, 2009	123-4567	0	1		1
Jan 1, 2010	Apr 30, 2010	123-4567	0	0		1

Temporal Classification

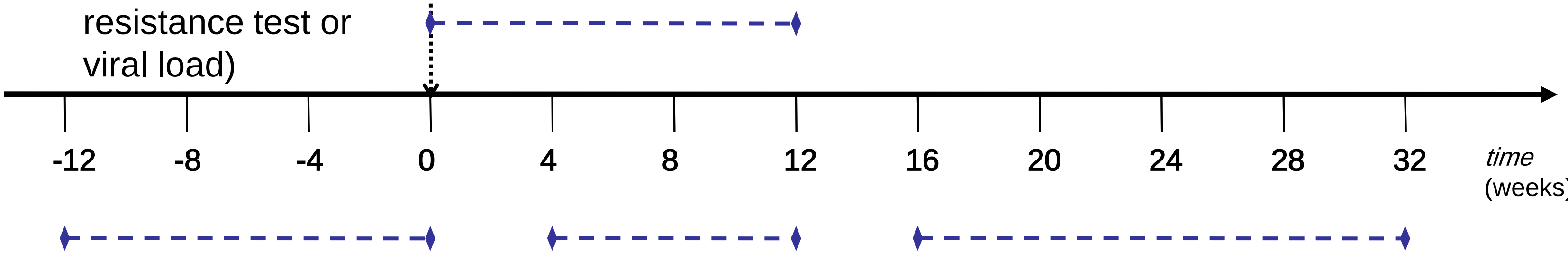
- **Knowledge-Based Temporal Abstraction**
  - Uses interval-based representation and reasoning
  - Creates high-level abstractions from primary data

Example Abstractions

Start of new regimen

On **'failing' regimen** (no changes to drug regimen after baseline resistance test or viral load)

On **new regimen** (no changes to drug regimen during first 12 weeks)



Baseline resistance test and viral load measures are latest made in the period from onset of new regimen to 12 weeks prior

At least one viral load measure in the 4-12 week period (**week 8 viral response**)

At least one viral load measure in the 16-32 week period (**week 24 viral response**)

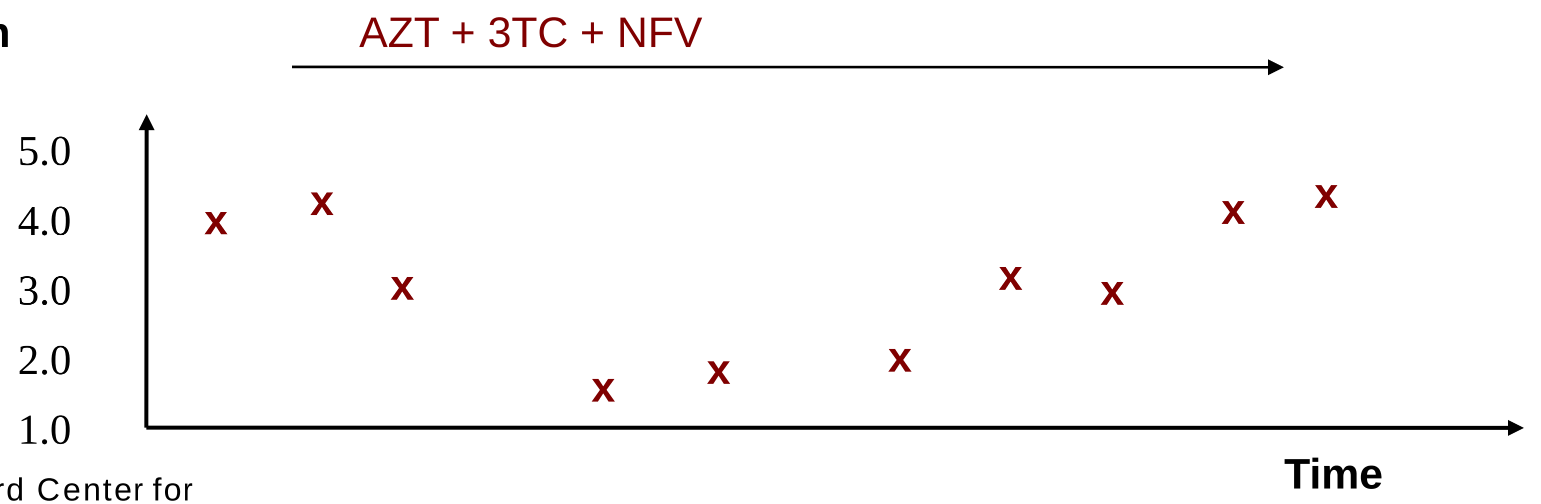
Stanford Center for Biomedical Informatics Research

(adapted from Phillips et al., 2005)

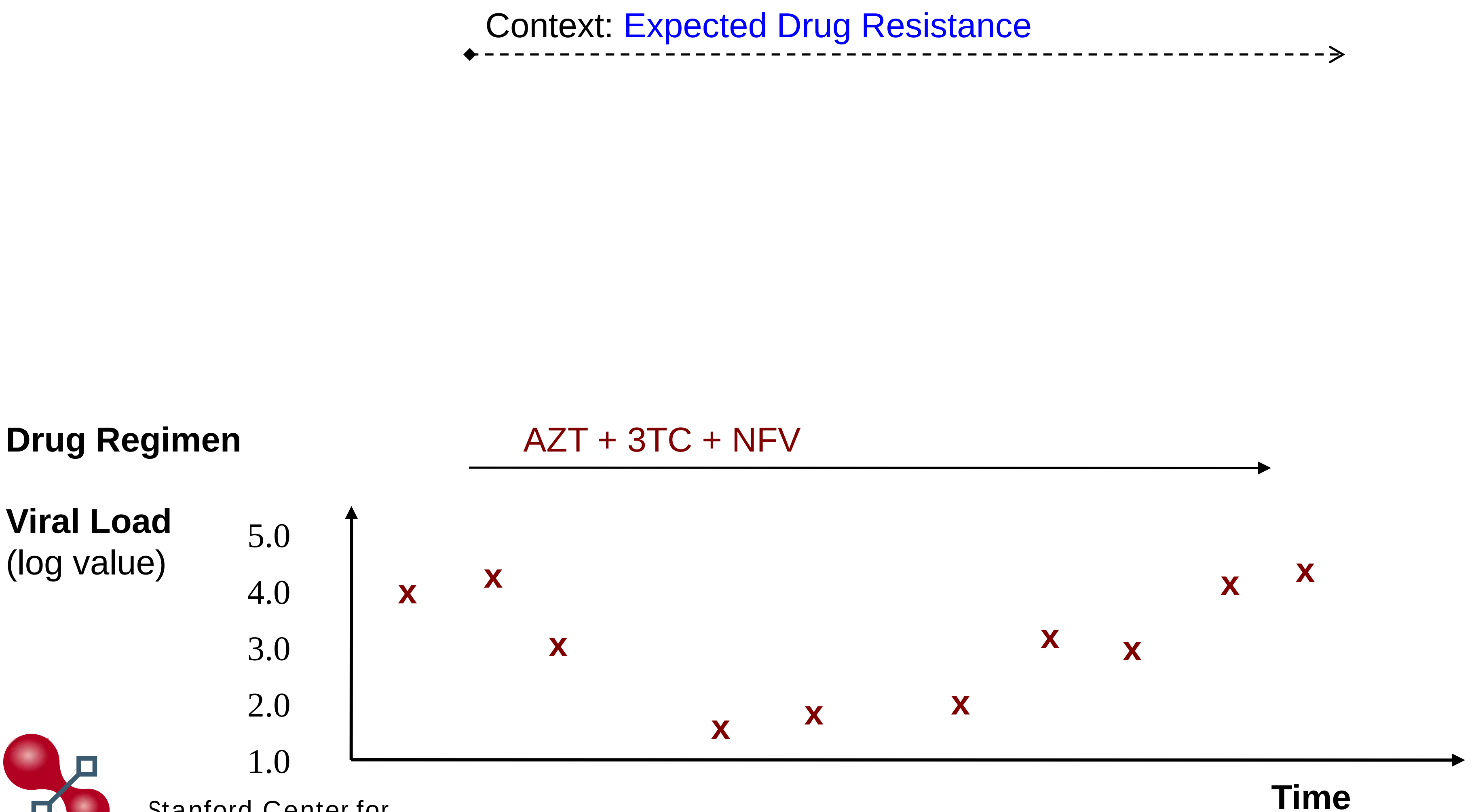
Abstraction Process

Drug Regimen

Viral Load
(log value)

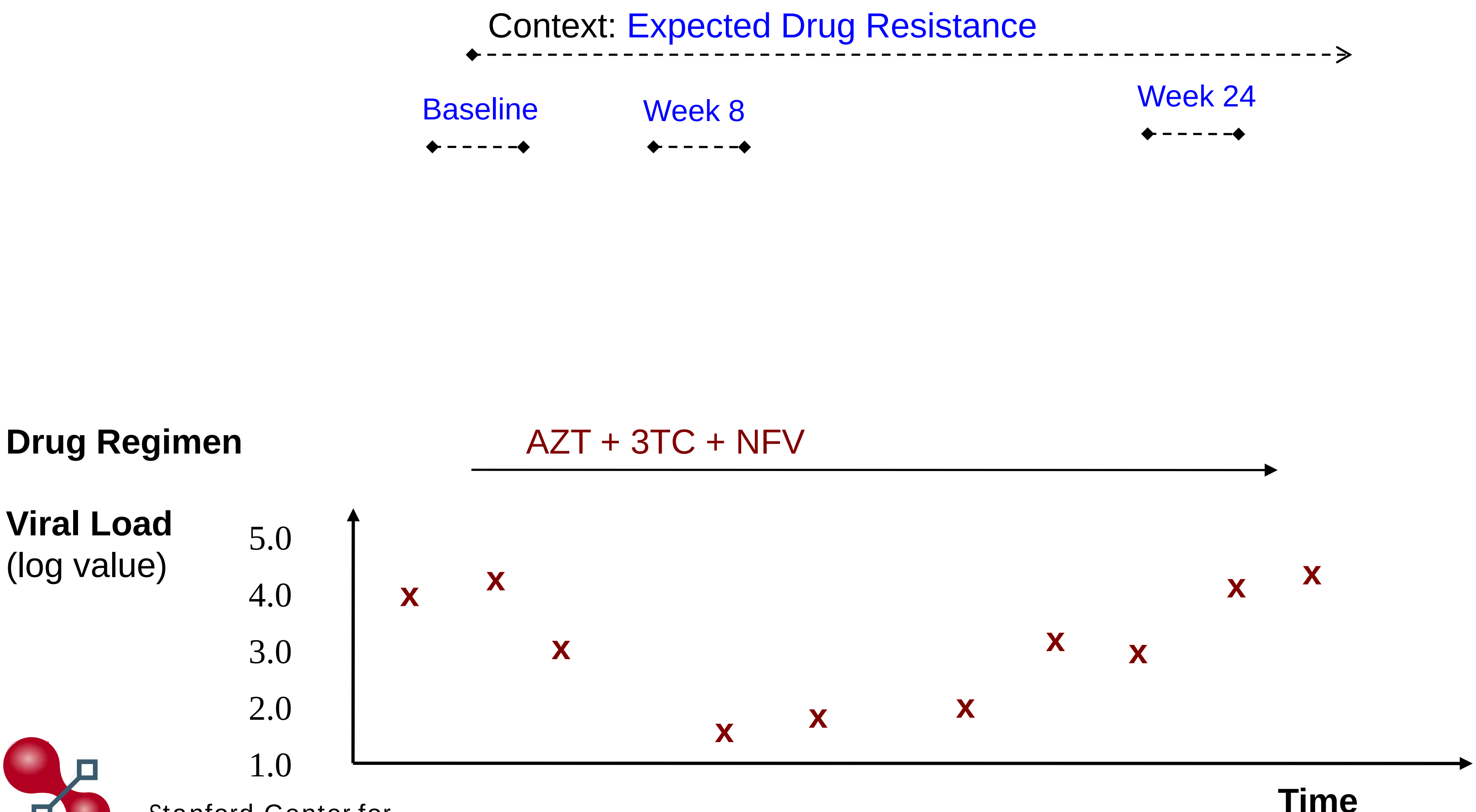


Abstraction Process



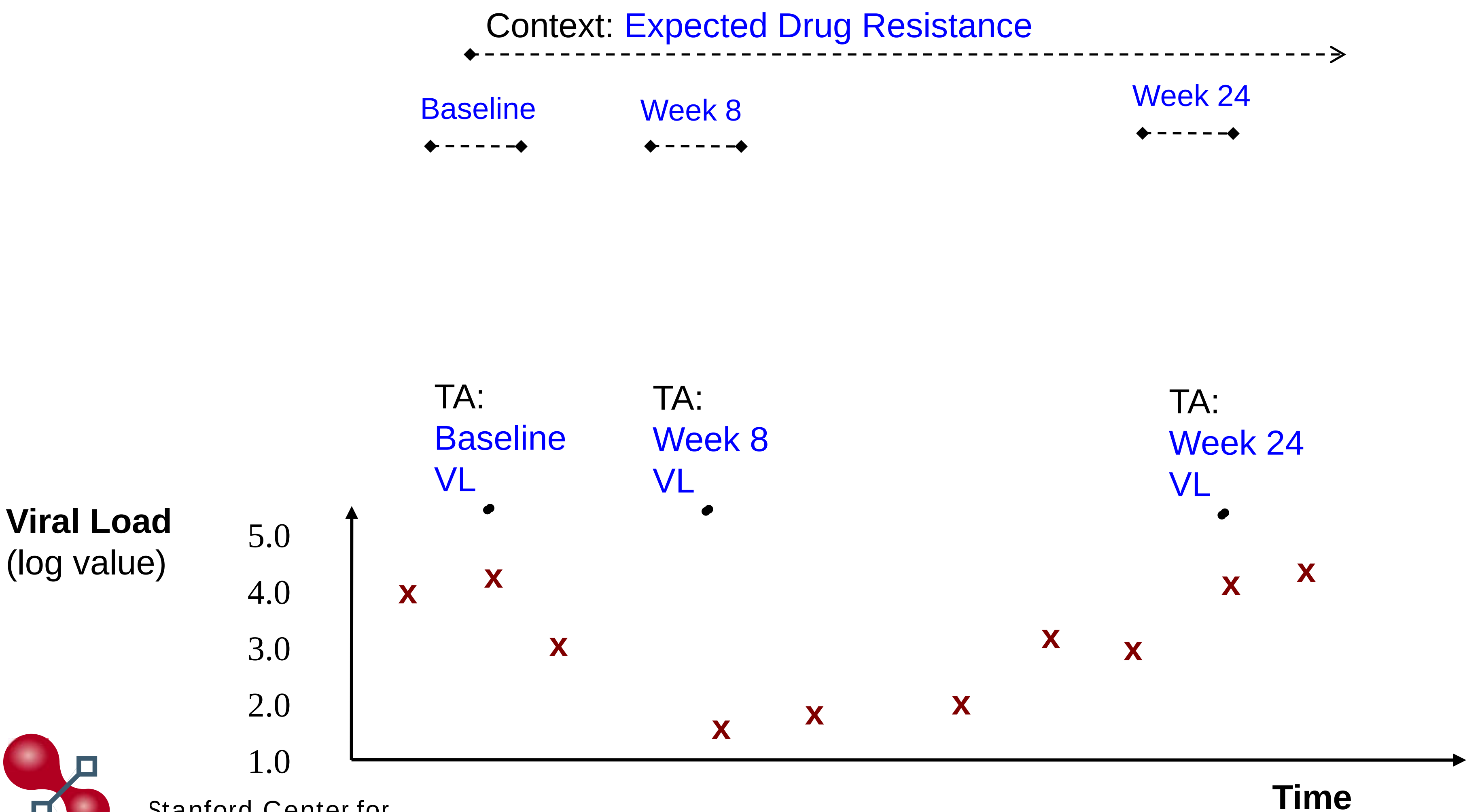
Stanford Center for
Biomedical Informatics Research

Abstraction Process



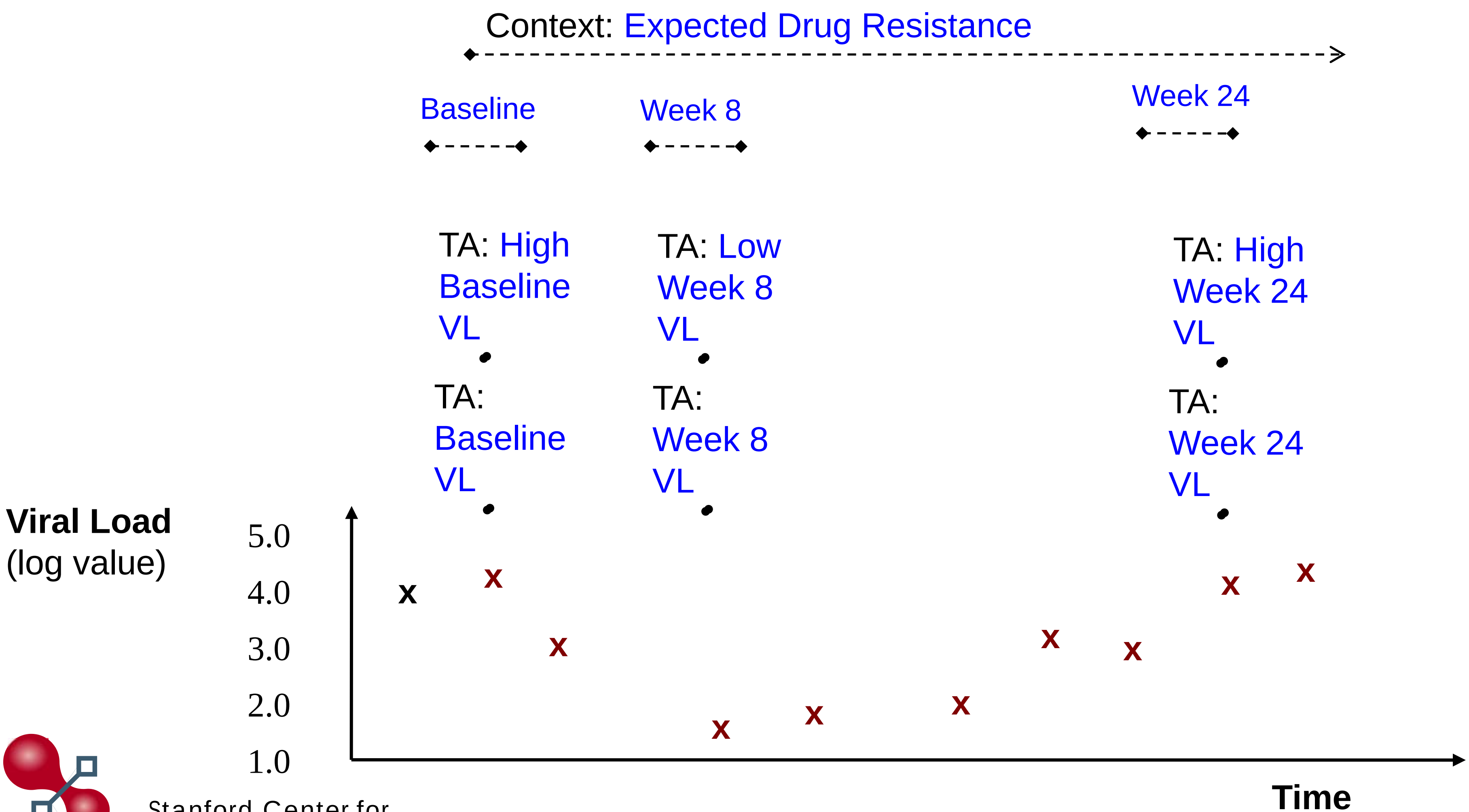
Stanford Center for
Biomedical Informatics Research

Abstraction Process



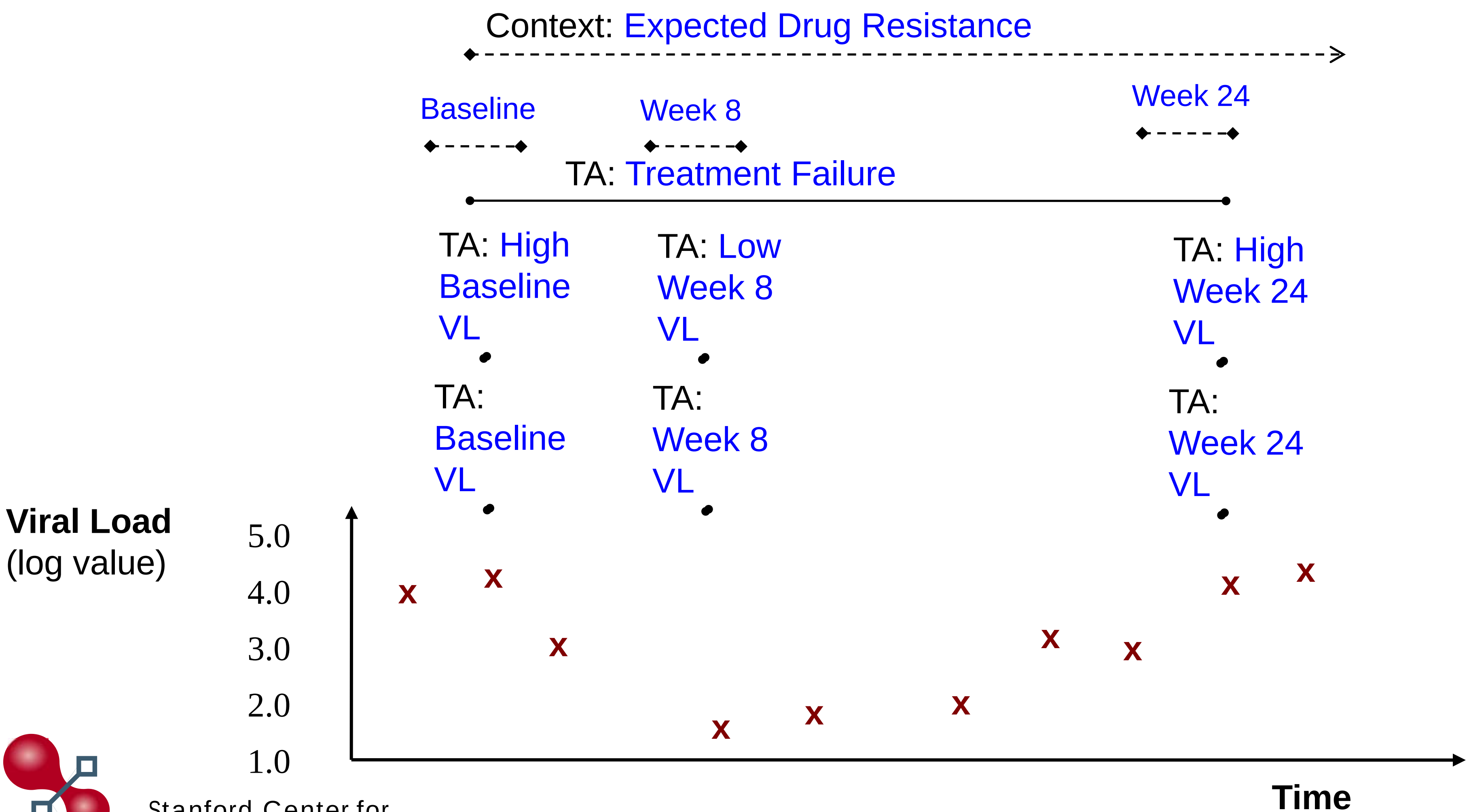
Stanford Center for
Biomedical Informatics Research

Abstraction Process



Stanford Center for
Biomedical Informatics Research

Abstraction Process



Stanford Center for
Biomedical Informatics Research

Temporal Abstraction Knowledge

- Structural knowledge

  AZT is a type of antiretroviral therapy

- Classification knowledge

  A ‘low’ viral load level is between 1.0 and 2.5

- Temporal semantic knowledge

  Two adjacent periods of ‘treatment response’ can be catenated into one period

HIV Ontology

Model concepts as classes:

- Patient

  hasVL

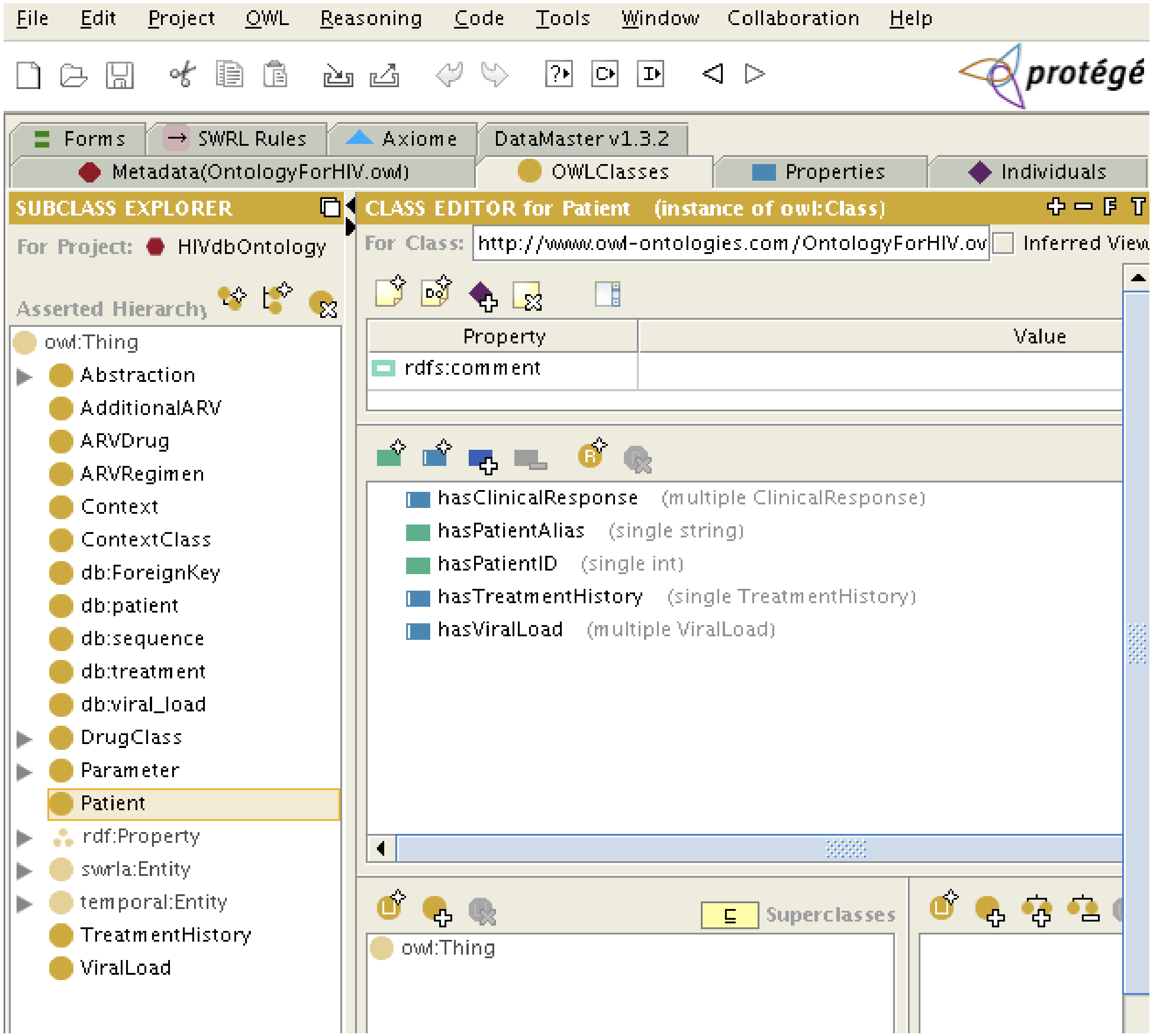
  hasTreatmentHistory

  hasClinicalResponse

- TreatmentHistory

  hasARVRegimen

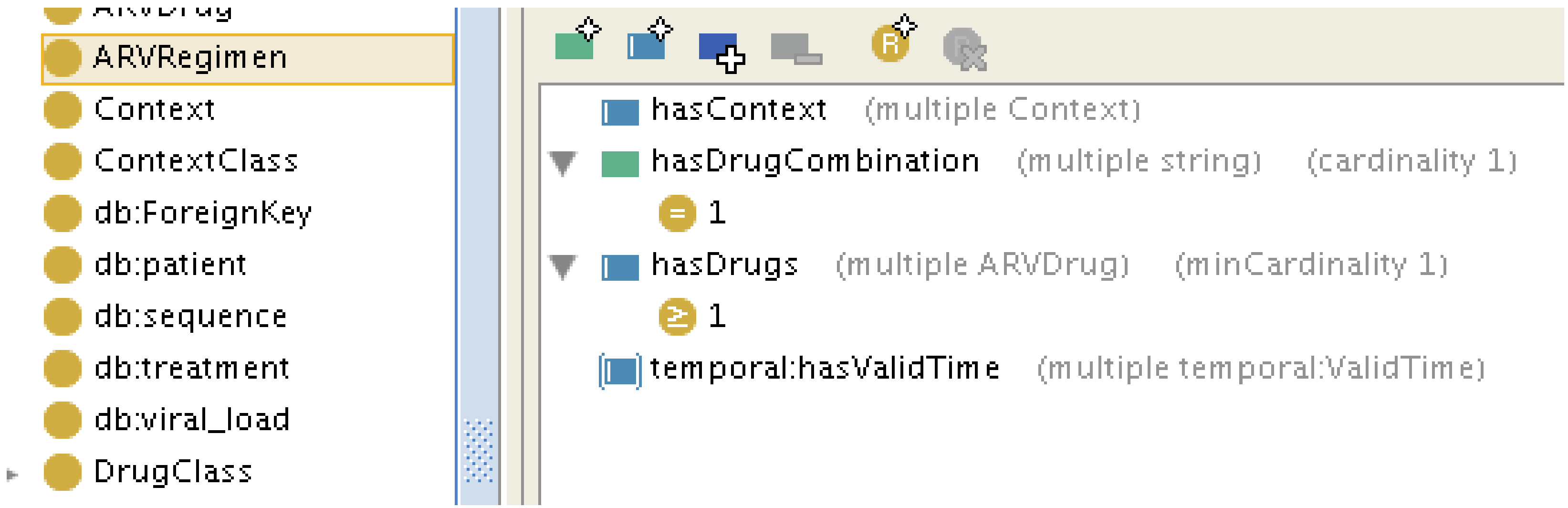
- Abstraction
  - PatternAbstraction
    - ClinicalResponse


File Edit Project OWL Reasoning Code Tools Window Collaboration Help
protégé
Forms
SWRL Rules
Axiome
DataMaster v1.3.2
Metadata(OntologyForHIV.owl)
OWLClasses
Properties
Individuals
SUBCLASS EXPLORER
For Project: HIVdbOntology
Asserted Hierarchy
owl:Thing
Abstraction
AdditionalARV
ARVDrug
ARVRegimen
Context
ContextClass
db:ForeignKey
db:patient
db:sequence
db:treatment
db:viral_load
DrugClass
Parameter
Patient
rdf:Property
swrla:Entity
temporal:Entity
TreatmentHistory
ViralLoad
CLASS EDITOR for Patient (instance of owl:Class)
For Class: http://www.owl-ontologies.com/OntologyForHIV.ov
Inferred View
Property
Value
rdfs:comment
hasClinicalResponse (multiple ClinicalResponse)
hasPatientAlias (single string)
hasPatientID (single int)
hasTreatmentHistory (single TreatmentHistory)
hasViralLoad (multiple ViralLoad)
Superclasses
owl:Thing

Modeling Time in Ontology

- temporal:Entity
  - temporal:Granularity
  - temporal:Proposition
    - temporal:ExtendedProposition
      - ExtendedAbstractProposition
      - ExtendedPrimitiveProposition
  - temporal:ValidTime
    - temporal:ValidInstant
    - temporal:ValidPeriod

Adding Temporal Properties



Temporal Abstraction Rule

SWRL Rule

Patient(?p) ∧ hasTreatmentHistory(?p, ?th) ∧ hasARVRegimens(?th, ?arv) ∧
hasContext(?arv, ?context) ∧ hasContextClass(?context, ?contextClass) ∧
hasContextClassName(?contextClass, "AtLeastThirtyTwoWk") ∧
temporal:hasValidTime(?context, ?cVt) ∧ hasViralLoad(?p, ?rna) ∧
temporal:hasValidTime(?rna, ?rnaVt) ∧
temporal:contains(?cVt, ?rnaVt, temporal:Days) ∧ hasValue(?rna, ?rnaValue) ∧
swrlb:greaterThanOrEqual(?rnaValue, 1.0) ∧ swrlb:lessThan(?rnaValue, 2.5) ∧
swrlx:createOWLThing(?rnaSA, ?rna) →
RNAStateAbstraction(?rnaSA) ∧ hasStateAbstraction(?rna, ?rnaSA) ∧
hasName(?rnaSA, "Low")

Rule Paraphrasing

```
- Rule RNAStateAbstractionLow:

IF
        "p" IS A Patient
        AND "p" HAS TreatmentHistory "th"
                WHERE "th" HAS ARVRegimens "arv"
                        WHERE "arv" HAS Context "context"
                                WHERE "context" HAS ContextClass "contextClass"
                                        WHERE "contextClass" HAS ContextClassName "AtLeastThirtyTwoWk"
                                AND "context" HAS ValidTime "cVt" WHERE "cVt" contains "rnaVt" AND "Days"
        AND "p" HAS ViralLoad "rna"
                WHERE "rna" HAS ValidTime "rnaVt"
                AND "rna" HAS Value "rnaValue" WHERE "rnaValue" IS GREATER THAN OR EQUAL TO 1.0
                AND "rnaValue" IS LESS THAN 2.5

THEN
        FOR EACH "rna" THERE IS A "rnaSA" SUCH THAT
                "rnaSA" IS A RNAStateAbstraction
                AND "rnaSA" HAS Name "Low"

        "rna" HAS StateAbstraction "rnaSA"
```

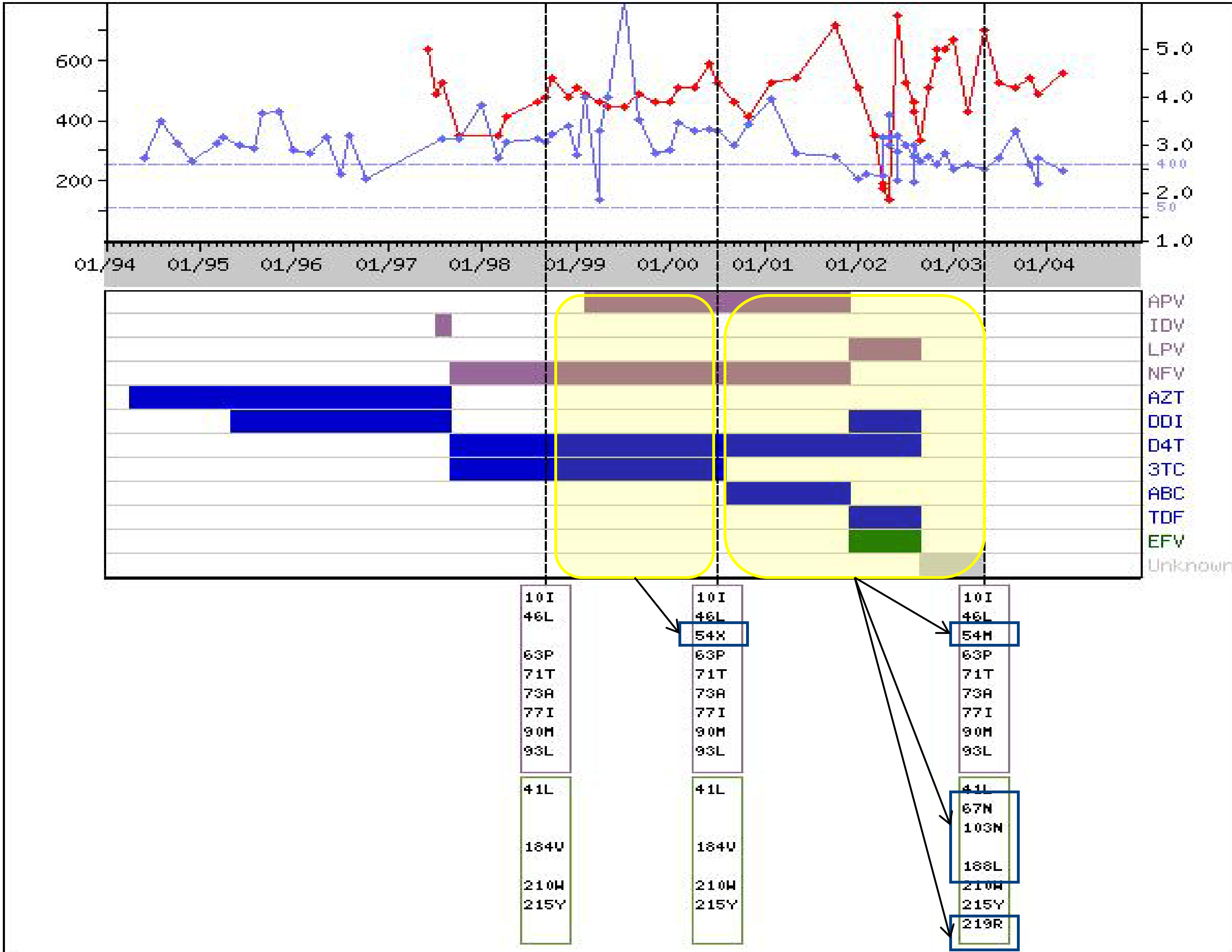
Stanford Center for
Biomedical Informatics Research

Using Temporal Information in Data Analysis

- Temporal associations
  - Between treatment regimen and new mutation
  - Between new mutation and drug resistance
- Temporal similarity
  - Between treatment histories and recommended treatments
  - Between treatment histories of patients

Temporal Associations

- Which new mutations are associated with which prior treatment regimens?



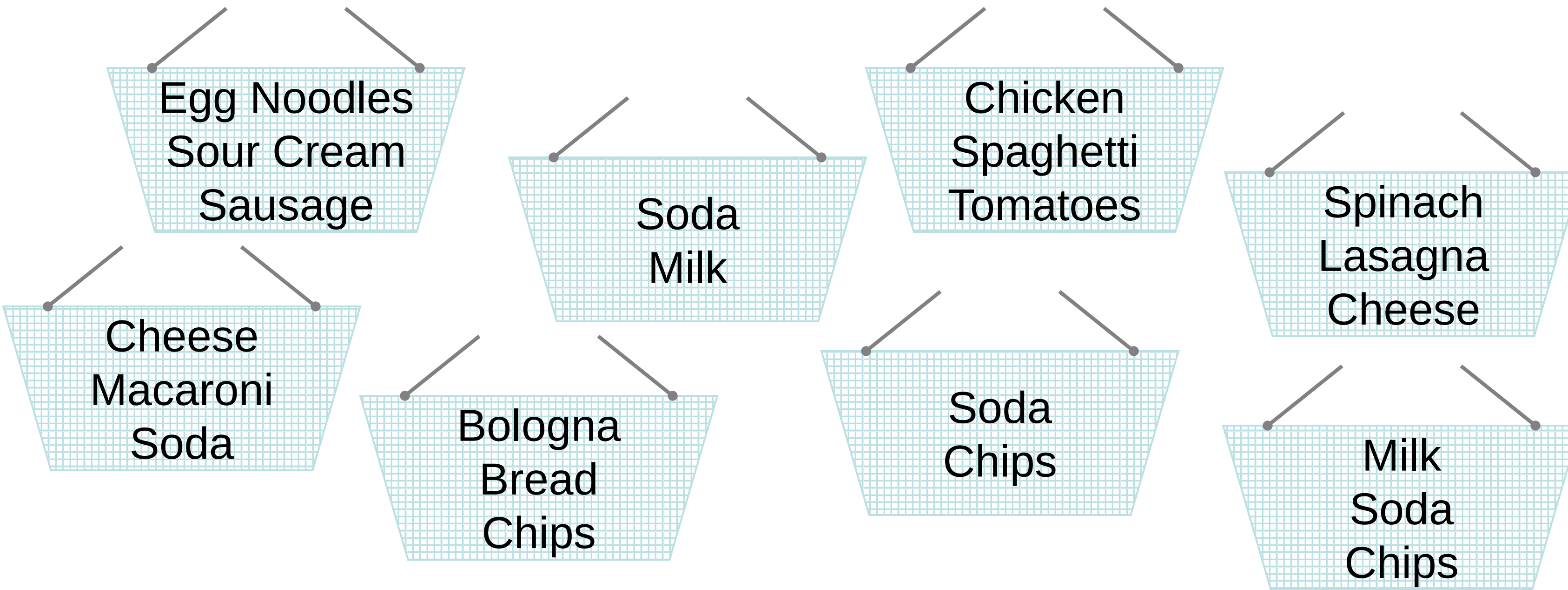
600
400
200
5.0
4.0
3.0
400
2.0
50
1.0
01/94
01/95
01/96
01/97
01/98
01/99
01/00
01/01
01/02
01/03
01/04
APV
IDV
LPV
NFV
AZT
DDI
D4T
3TC
ABC
TDF
EFV
Unknown
10I
46L
63P
71T
73A
77I
90M
93L
41L
184V
210W
215Y
10I
46L
54X
63P
71T
73A
77I
90M
93L
41L
184V
210W
215Y
10I
46L
54M
63P
71T
73A
77I
90M
93L
41L
67N
103N
188L
210W
215Y
219R

Method

Association Rule Mining

- Unsupervised learning method
- Market Basket Analysis: commonly associated sets of items that appear together in market baskets

Association Rule Mining



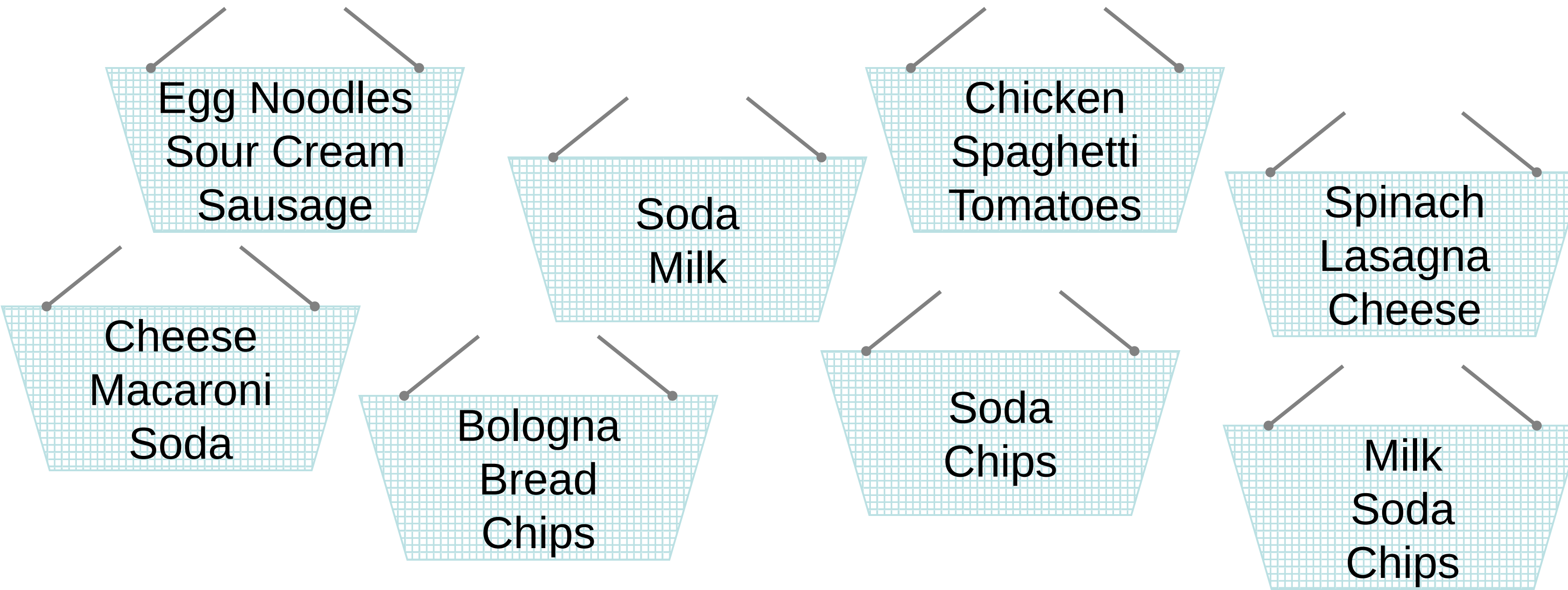
{LHS} => {RHS}

Confidence: Prob(RHS | LHS)

Support: Prob(LHS $\cup$ RHS)

Stanford Center for
Biomedical Informatics Research

Association Rule Mining



{Soda} => {Chips} Confidence: 0.50

Support: 0.25

Stanford Center for
Biomedical Informatics Research

Extracted Temporal Data Set

- 4,907 records ("time windows" associating regimens with new mutations)
- 2,681 unique patients
- Variables
  - Treatments
    - 7 PI (APV, IDV, NFV, RTV, SQV, LPV, ATV)
    - NRTI
    - nNRTI
  - Protease gene sequence
    - 99 positions (P1 – P99)

Association Rule Mining

- Apriori algorithm
- High sensitivity
  - Support = 0.002
  - Confidence = 0.1
  - Max number of items = 10
- 449,077 rules mined in total
  - 1,406 rules: treatment $\rightarrow$ new mutation

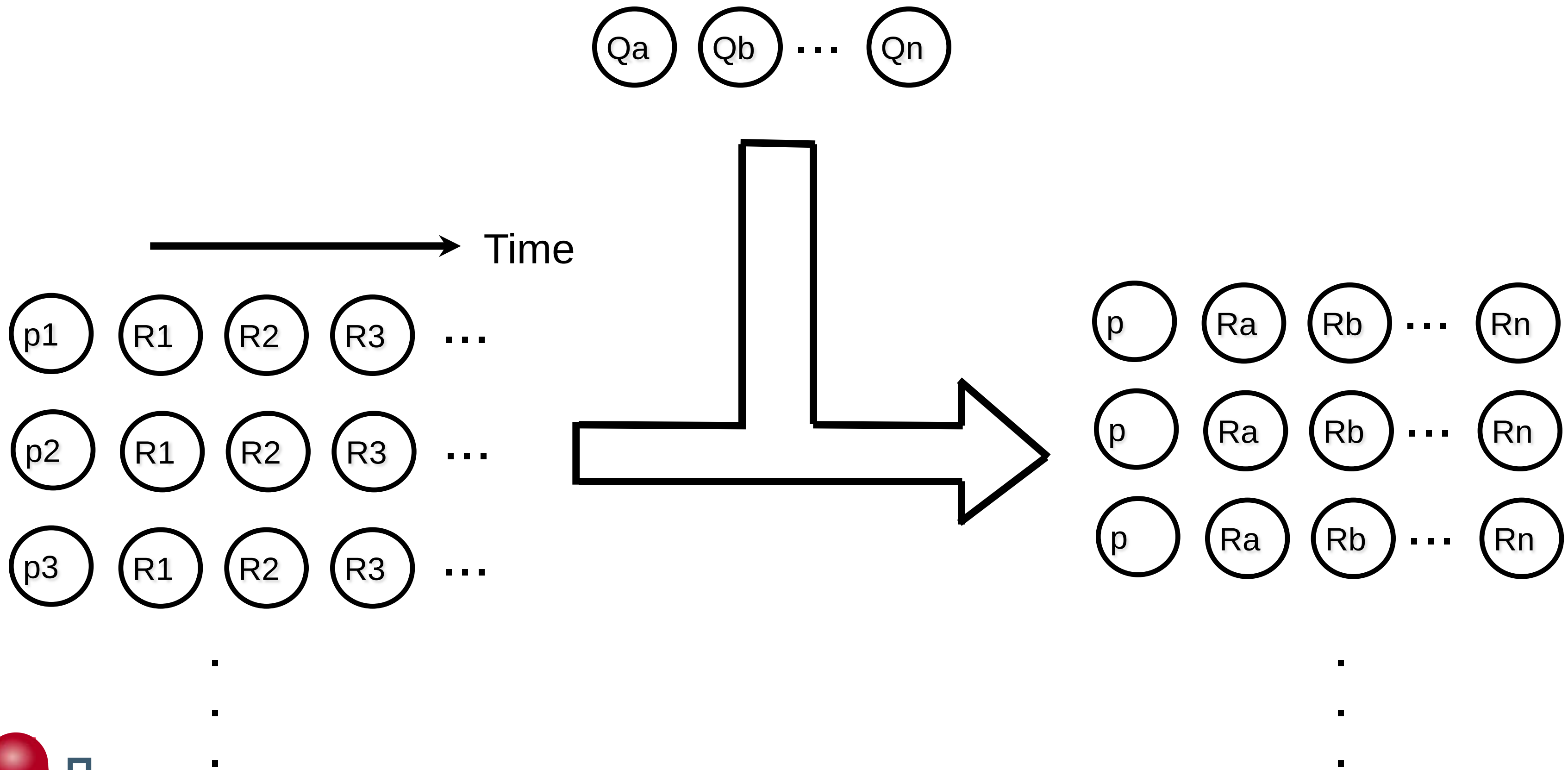
Top 10 Confident Rules

LHS	RHS	Support	Confidence
{APV, SQV, LPV}	{P10}	0.0022	0.50
{APV, SQV, LPV, NRTI}	{P10}	0.0022	0.50
{APV, RTV, LPV, nNRTI}	{P10}	0.0029	0.47
{APV, SQV, LPV}	{P54}	0.0020	0.45
{APV, SQV, LPV, NRTI}	{P54}	0.0020	0.45
{APV, LPV, nNRTI}	{P10}	0.0049	0.44
{APV, LPV, NRTI, nNRTI}	{P10}	0.0047	0.43
{APV, LPV, nNRTI}	{P54}	0.0047	0.43
{APV, RTV, LPV}	{P13}	0.0035	0.43
{APV, LPV, NRTI, nNRTI}	{P54}	0.0044	0.42

Temporal Similarity

- How do I find patients who have a clinical history *similar* to recommended treatment patterns?

Local Alignment Tool for Clinical Histories



Stanford Center for
Biomedical Informatics Research

BLAST

Query sequence: PQGEFG

Word 1: PQG

Word 2: QGE

Word 3: GEF

Word 4: EFG

Scoring Substitution Matrix

High Scoring Segment Pair

Query sequence: R P P Q G L F

Database sequence: D P **P E G** V V

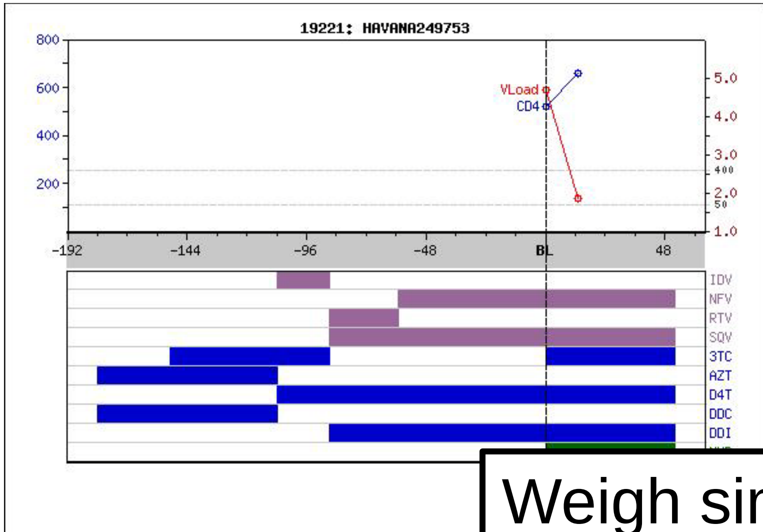
Exact match is scanned.

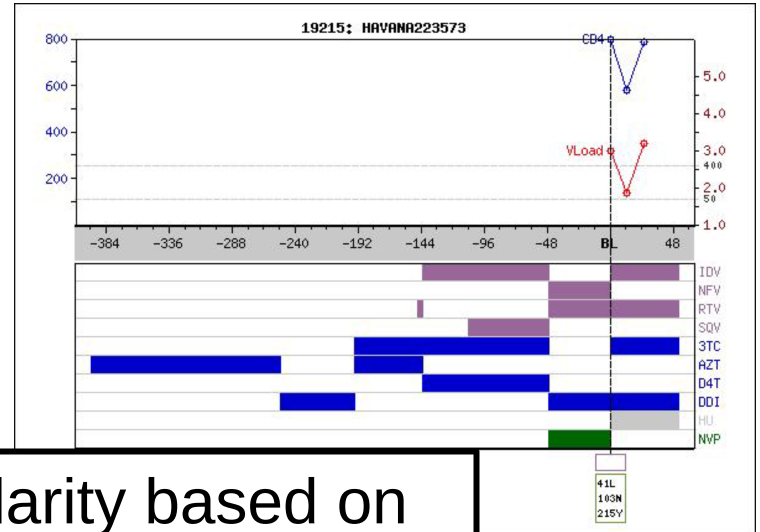
Score: -2 **7 7 2 6 1** -1

HSP

Optimal accumulated score = 7+7+2+6+1 = 23

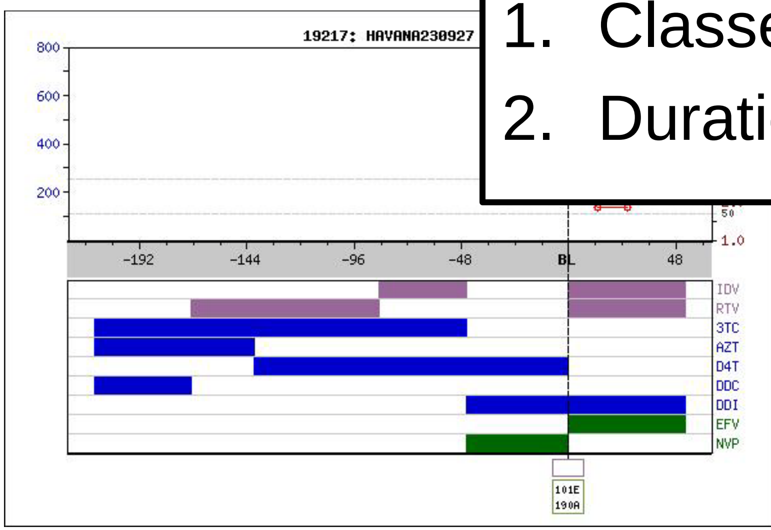
Stanford Center for
Biomedical Informatics Research

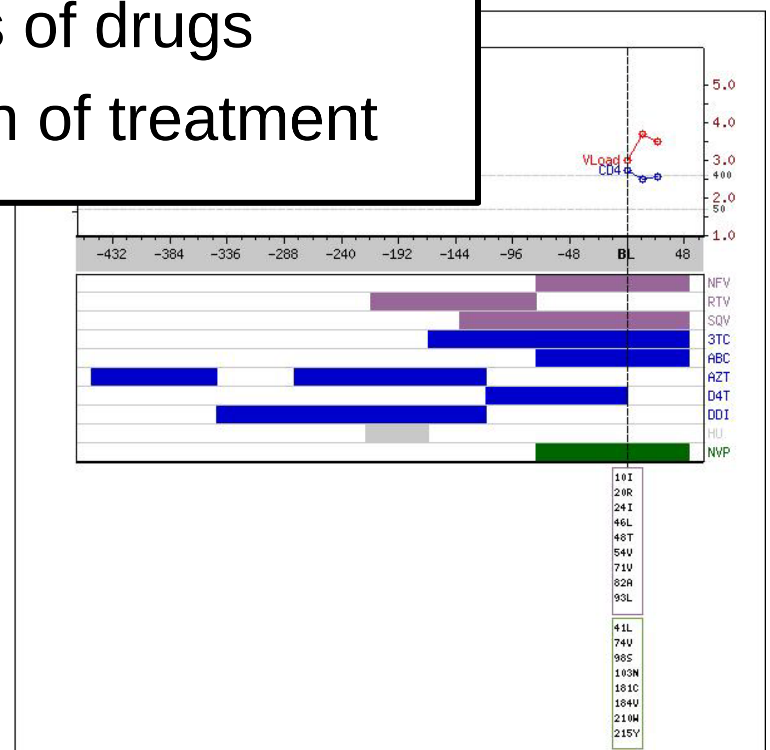




Weigh similarity based on

1. Classes of drugs
2. Duration of treatment





Stanford Center for
Biomedical Informatics Research

Method

- Input
  - Interval-based query pattern
  - Database of sequences
  - A minimum threshold score from 0 to 1
- Output
  - Sequences with matching score greater than or equal to threshold

Scoring Function

Query	Sequence	
$[qD_1, qD_2, \ldots, qD_n]$	$[sD_1, sD_2, \ldots, sD_n]$	$\text{drugScore} = \dfrac{(nD_q+nD_S)*\text{match}}{nD_q^{\,2} + nD_S^{2}}$
$[qC_1, qC_2, \ldots, qC_n]$	$[sC_1, sC_2, \ldots, sC_n]$	$\text{classScore} = \dfrac{(nC_q+nC_s)*\text{match}}{2(nC_q^{\,2}+nC_S^{2})}$
Time_Q	Time_S	$\text{overlapScore} = \dfrac{\text{Time}_Q}{\text{Time}_S}$ or $\dfrac{\text{Time}_S}{\text{Time}_Q}$

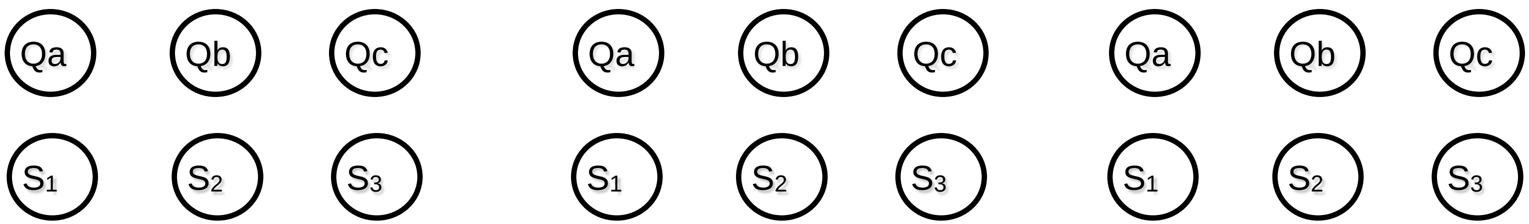
Score = overlapScore*(drugScore+classScore)/1.5

0 ≤ Score ≤ 1

Stanford Center for
Biomedical Informatics Research

Example Scoring

threshold = 0.7



0.8	0.9	0.7	0.4	1.0	0.7	0.2	0.7	1.0

Score = 0.8 Score = 0.7 Score = 0.63

Reducing the Search Space

threshold = 0.7

First match		
Qa	Qb	Qc
S_1	S_2	S_3
x	[1.0]	[1.0]

Second match		
Qa	Qb	Qc
S_1	S_2	S_3
0.4	x	[1.0]

Minimum Qualifying Score

$x = 0.1$

$x = 0.7$

$$\frac{x + \text{cumulative score} + \text{ideal score}}{\text{length of query}} \geq \text{threshold}$$

Query: Q_1(EFV,FTC,TDF), Q_2(SQV,DDI,AZT)
Threshold = 0.5

Results: 944 matches identified
Top 20 matches

Match	Regimen 1	Regimen 2	Start Index	End Index	Score
1	EFV,3TC,TDF	LPV,AZT,3TC	0	1	0.67
2	EFV,DDI,TDF	LPV,SQV,DDI, TDF,3TC	8	9	0.66
3	EFV,DDI,NFV	SQV,RTV,DDI, AZT	13	14	0.54
4	NVP,DDI,AZT, NFV	SQV,DDI,AZT	6	7	0.54
5 - 20	DDI,AZT	SQV,DDI,AZT	1	2	0.52

Lessons Learned

- Understand what temporal phenomena need to be studied
- Model temporal data and knowledge as explicitly as possible
- Reframe questions about temporal phenomena into existing data-analytic frameworks

Acknowledgments

- Funding
  - NIH Funding Support: RC1LM010583, R01LM009607, R01AI068581
  - PhRMA Research Starter Award in Informatics
  - California HIV/AIDS Research Program
  - Iowa Venture Funds
- Collaborators
  - Bob Shafer (Stanford) and George Towfic (Clarke)
- Students and Research Staff
  - Genaro Hernandez, Ray Lin, Rashmi Raj, Sarim Baig, Weech Lee, Martin O’Connor, Mike Bingen, Samson Tu

Stanford Center for
Biomedical Informatics Research