



Cyberspace and data mining: new paradigms for modeling a complex world

Christopher R. Stephens

*C3-Centro de Ciencias de la Complejidad y
Instituto de Ciencias Nucleares
Universidad Nacional Autónoma de México*

*Nuevos Paradigmas y Desafíos en las Ciencias y Artes del Ciberespacio, UAM,
Xochimilco* 19/10/2012



Cyberspace: What is it?

The global network of interdependent information technology infrastructures, telecommunications networks and computer processing systems.

As a social experience, individuals can interact, exchange ideas, share information, provide social support, conduct business, direct actions, create artistic media, play games, engage in political discussion, and so on, using this global network.

<http://en.wikipedia.org/wiki/Cyberspace>

A source and method of gathering data with which to model a complex world CRS

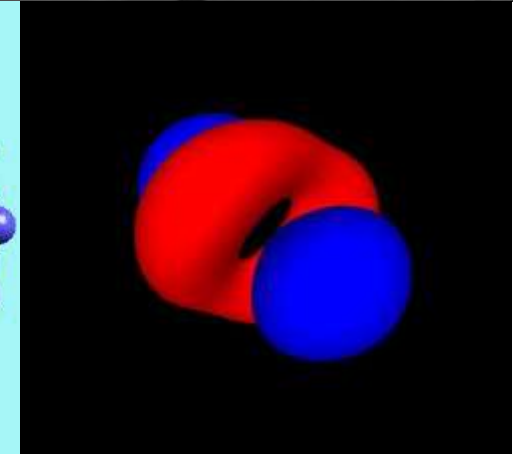
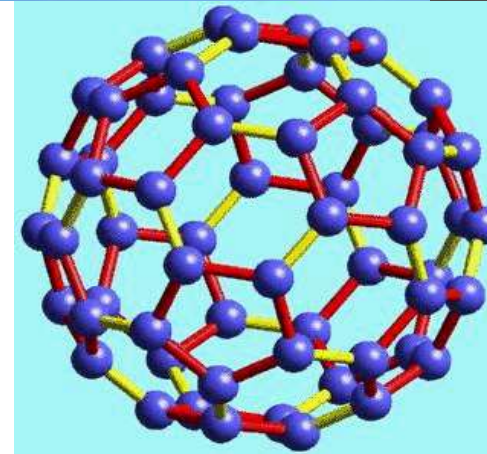
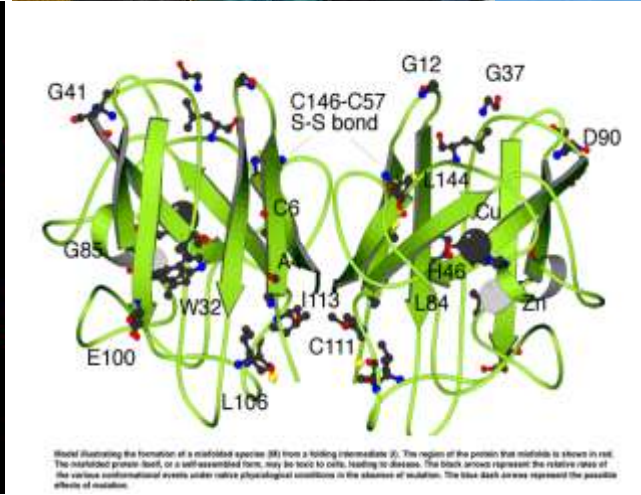
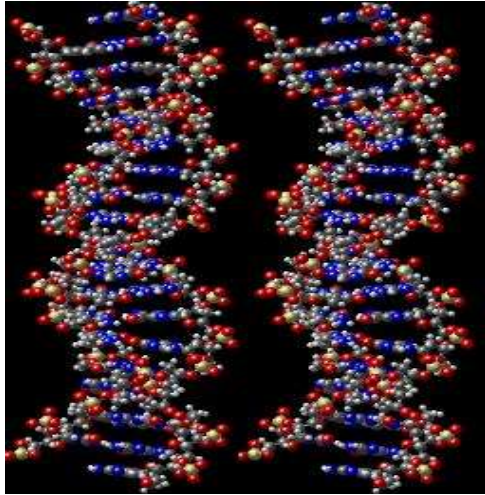


Data mining: What is it?

- "... the exploration and analysis of data in order to discover patterns, correlations and other regularities."
- There are two main data mining tasks
 - Predicting – "pattern" identification
Establishes "causal" statistical relations
 - Profiling - "pattern" description
Identifies what are the key drivers associated with a pattern
- There are three main requirements
 - Data; data processors; inference algorithms

Complexity: What is it?

Desde lo complejo...



Model illustrating the formation of a misfolded species (B) from a folding intermediate (I). The region of the protein that misfolds is shown in red. The misfolded protein (B), or a self-assembled form, may be toxic to cells, leading to disease. The black arrows represent the relative rates of the various conformational events under various physiological conditions in the absence of mutation. The blue dash arrows represent the possible effects of mutation.

...hasta lo simple



Complexity: What is it?

Propiedades estructurales

Una “jerarquía” de muchas diferentes escalas

Grados de libertad efectivos (“colectividad”) cualitativamente diferentes a escalas diferentes

Jerarquía de **bloques constructores** (modularidad)

Interacciones “intra-bloque” más fuerte que “inter-bloque”

El micro se une con el macro (adecuación, semántica,...)

Propiedades funcionales

Sistemas que son **adaptativos**

Una evolución dinámica que depende de muchas **reglas/estrategias** diferentes.

Sistemas que “**aprenden**” (retroalimentación del ambiente al sistema que se usa para actualizar las reglas)

El micro se une con el macro (adecuación, semántica,...)

Comportamiento más complejo (el “fenotipo”)

Mejor descritos por lo que hacen más que en lo que son

An example of datamining...

Who will buy an insurance policy?

IDENTIFIER	MOSTYPE	MAANTHUI	MGEMOMV	MGEMLEEF	MOSHOOFD	MGODRK	MGODPR	MGODOV	MGODGE
1	33	1	3	2	8	0	5	1	3
2	37	1	2	2	8	1	4	1	4
3	37	1	2	2	8	0	4	2	4
4	9	1	3	3	3	2	3	2	4
5	40	1	4	2	10	1	4	1	4
6	23	1	2	1	5	0	5	0	5
7	39	2	3	2	9	2	2	0	5
8	33	1	2	3	8	0	7	0	2
9	33	1	2	4	8	0	1	3	6
10	11	2	3	3	3	3	5	0	2
11	10	1	4	3	3	1	4	1	4
12	9	1	3	3	3	1	3	2	4
13	33	1	2	3	8	1	4	1	4
14	41	1	3	3	10	0	5	0	4
15	23	1	1	2	5	0	6	1	2
16	33	1	2	3	8	0	7	0	2
17	38	1	2	3	9	0	6	0	3
18	22	2	3	3	5	0	5	0	4
19	13	1	4	2	3	2	4	0	3
20	31	1	2	4	7	0	2	0	7
21	33	1	4	3	8	0	6	0	3
22	33	2	3	3	8	0	4	2	3
23	13	1	3	2	3	1	7	0	2
24	34	2	3	2	8	0	7	0	2

The data



A description of the data



TRAIN DATA					TEST DATA			
Number of fields:	86				Number of fields:	85		
Number of records:	5822				Number of records:	4000		
	Minimum	Maximum	Mean	Std	Minimum	Maximum	Mean	Std
MOSTYPE	1	41	24.25	12.85	1	41	24.25	13.02
MAANTHUI	1	10	1.11	0.41	1	10	1.11	0.42
MGEMOMV	1	5	2.68	0.79	1	6	2.68	0.77
MGEMLEEF	1	6	2.99	0.81	1	6	3.00	0.79
MOSHOOFD	1	10	5.77	2.86	1	10	5.79	2.90
MGODRK	0	9	0.70	1.00	0	9	0.71	1.03
MGODPR	0	9	4.63	1.72	0	9	4.65	1.73
MGODOV	0	5	1.07	1.02	0	5	1.02	1.00
MGODGE	0	9	3.26	1.60	0	9	3.27	1.62
MRELGE	0	9	6.18	1.91	0	9	6.20	1.88
MRELSA	0	7	0.88	0.97	0	7	0.86	0.96
MRELOV	0	9	2.29	1.72	0	9	2.28	1.69
MFALLEEN	0	9	1.89	1.80	0	9	1.89	1.75
MFGEKIND	0	9	3.23	1.62	0	9	3.25	1.59
MFWEKIND	0	9	4.30	2.01	0	9	4.31	1.95
MOPLHOOG	0	9	1.46	1.62	0	9	1.52	1.68
MOPLMIDD	0	9	3.35	1.76	0	9	3.24	1.67
MOPLLAAG	0	9	4.57	2.30	0	9	4.62	2.25
MBERHOOG	0	9	1.90	1.80	0	9	1.90	1.84
MBERZELF	0	5	0.40	0.78	0	5	0.41	0.80
MBERBOER	0	9	0.52	1.06	0	9	0.58	1.17

DESCRIPTION	USED	CARDINALITY	MIN_VALUES	MAX_VALUES
Customer Subtype	1	41	1	41
Number of houses	1	10	1	10
Avg size household	1	6	1	6
Avg age	1	6	1	6
Customer main type	1	10	1	10
Roman catholic	1	10	0	9
Protestant	1	10	0	9
Other religion	1	10	0	9
No religion	1	10	0	9
Married	1	10	0	9
Living together	1	10	0	9
Other relation	1	10	0	9
Singles	1	10	0	9
Household without children	1	10	0	9
Household with children	1	10	0	9
High level education	1	10	0	9
Medium level education	1	10	0	9
Lower level education	1	10	0	9
High status	1	10	0	9
Entrepreneur	1	10	0	9
Farmer	1	10	0	9
Middle management	1	10	0	9
Skilled labourers	1	10	0	9
Unskilled labourers	1	10	0	9

Table 22.1 Most Important Uniperspective Drivers for Mobile Home Insurance Purchasers

Profile Driver	ϵ	Number of Customers With Driver	Number of Purchasers With Driver	%
Auto insurance contribution \$500–\$2,500	10.81	2,319	262	11.3
Fire insurance contribution \$100–\$250	9.36	1,226	151	12.3
Boat insurance policy	7.69	31	12	38.7
High purchasing power	7.49	474	67	14.1
Middle-class families	7.04	339	51	15.0
Driven growers	5.02	66	13.1	6.78
Two auto insurance policies	6.27	246	38	15.4
One auto insurance policy	6.07	2,712	237	8.7
Third-party insurance contribution \$25–\$50	5.83	2,128	191	9.0
One third-party insurance policy	5.37	2,334	201	8.6
Social Security insurance contribution	5.23	81	16	19.8

Feature selection

Profile

Model construction

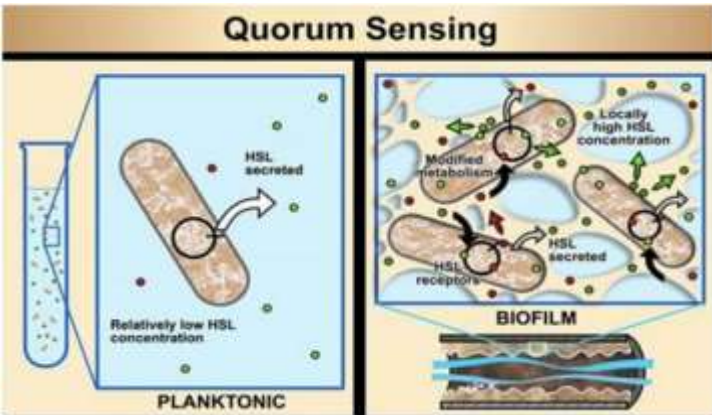
Field	Value	(Field = Value Class?)	(Field = Value Class0)	(Field = Value Class1)	(Field = Value No Class)	(Field = Value No Class)
		N (Class) =	348		N (No Class) =	5474
		P (Class) =	0.059773		P (No Class) =	0.940227
MOSTYPE	1	13	0.037356	-1.427636	111	0.020278
MOSTYPE	2	6	0.017241	-1.763428	76	0.013884
MOSTYPE	3	25	0.071839	-1.143639	224	0.040921
MOSTYPE	4	2	0.005747	-2.240549	50	0.009134
MOSTYPE	5	2	0.005747	-2.240549	43	0.007855
MOSTYPE	6	12	0.034483	-1.462398	107	0.019547
MOSTYPE	7	3	0.008621	-2.064458	41	0.007490
MOSTYPE	8	51	0.146552	-0.834009	288	0.052612
MOSTYPE	9	12	0.034483	-1.462398	266	0.048593
MOSTYPE	10	9	0.025862	-1.587337	156	0.028498
MOSTYPE	11	9	0.025862	-1.587337	144	0.026306
MOSTYPE	12	16	0.045977	-1.337459	95	0.017355
MOSTYPE	13	13	0.037356	-1.427636	166	0.030325
MOSTYPE	14	0	0.000000	1.000000	0	0.000000
MOSTYPE	15	0	0.000000	1.000000	5	0.000913
MOSTYPE	16	0	0.000000	1.000000	16	0.002923
MOSTYPE	17	0	0.000000	1.000000	9	0.001644
MOSTYPE	18	0	0.000000	1.000000	19	0.003471
MOSTYPE	19	0	0.000000	1.000000	3	0.000548
MOSTYPE	20	2	0.005747	-2.240549	23	0.004202
MOSTYPE	21	0	0.000000	1.000000	15	0.002740
MOSTYPE	22	4	0.011494	-1.939519	94	0.017172

Predictions

What next...?

IDENTIFIER	SCORE
1	-2.09349915
2	4.68377695
3	1.63224903
4	0.15334467
5	-3.61928271
6	-4.09623639
7	-1.35255102
8	3.30637666
9	0.31675739
10	1.92161957
11	-4.87274355
12	2.50316284
13	-0.402142
14	-3.02995004
15	-3.78862209
16	-4.80597027
17	-0.69363628
18	0.42011252
19	-5.77513435
20	0.51878013
21	-2.81516504

Some other examples of datamining...



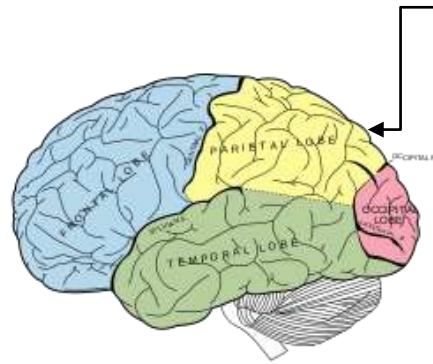
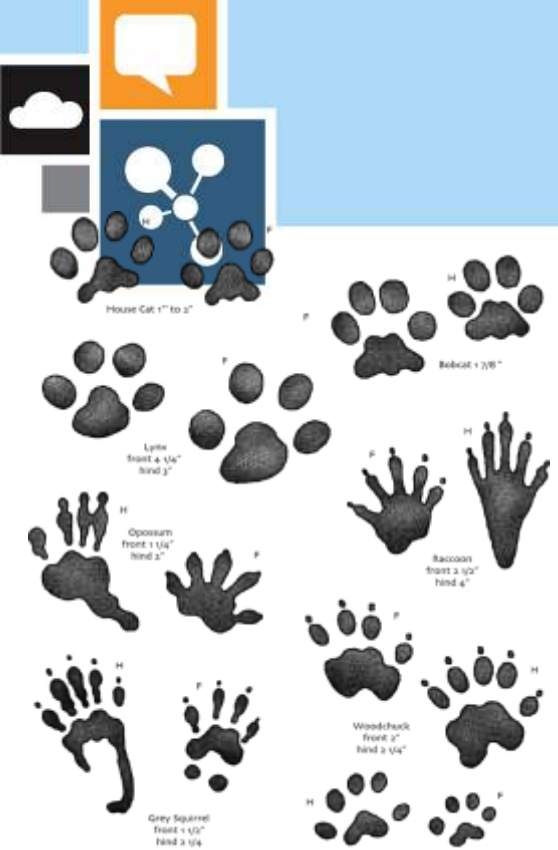
What's the difference/same between these and predicting who's going to buy an insurance policy?



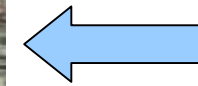
Data mining in cyberspace versus Euclidean space



Different data, different data processors and different inference algorithms



Same goal of modeling a *complex* world for *decision making* and "*optimization*"





Evolution produces dataminers

Evolution has created some incredible “dataminers”, with the capacity to “sense” and “process” data from an immensely complex environment, “infer” what is going to happen and take corresponding decisions.

But...

There are biases in the datamining...

Each organism is adapted to sense only certain data types – electromagnetic, acoustic, chemical etc.; inferences are heavily biased to avoiding bad things (false positives much less important than false negatives); feature selection and the restrictions of dimensionality (we live in a Euclidean world)



Three important points to remember about datamining in the real world...

- 1) Data (biological or non-biological) is a “coarse grained” proxy for the real world
- 2) The real world is complex, incorporates factors and variables from the micro to the macro
- 3) The real world is adaptive, what we do in it changes it



Datamining in the XXIst century

What's so special now? We data mine in cyberspace!

For the first time in history the data captured and stored in non-biological substrates is much larger than that in biological substrates

~ 10^{20} - 10^{23} bits produced per year; about ten thousand DVDs for every person on the planet

It is data that can be a direct or indirect representation of the "environment" and its adaptive dynamics

Examples: data from all transactions in financial markets; customer purchasing data (Walmart); satellite image data...

Concept of a "genotype-phenotype" map for data

But...

Data is biased, not objective enough, not comprehensive enough, not integrated enough,...



And making inferences...?

The Predictability Landscape

probability
versus
regression

$$P(C|X(t))$$

causation versus
correlation

What we want to
know – our goals



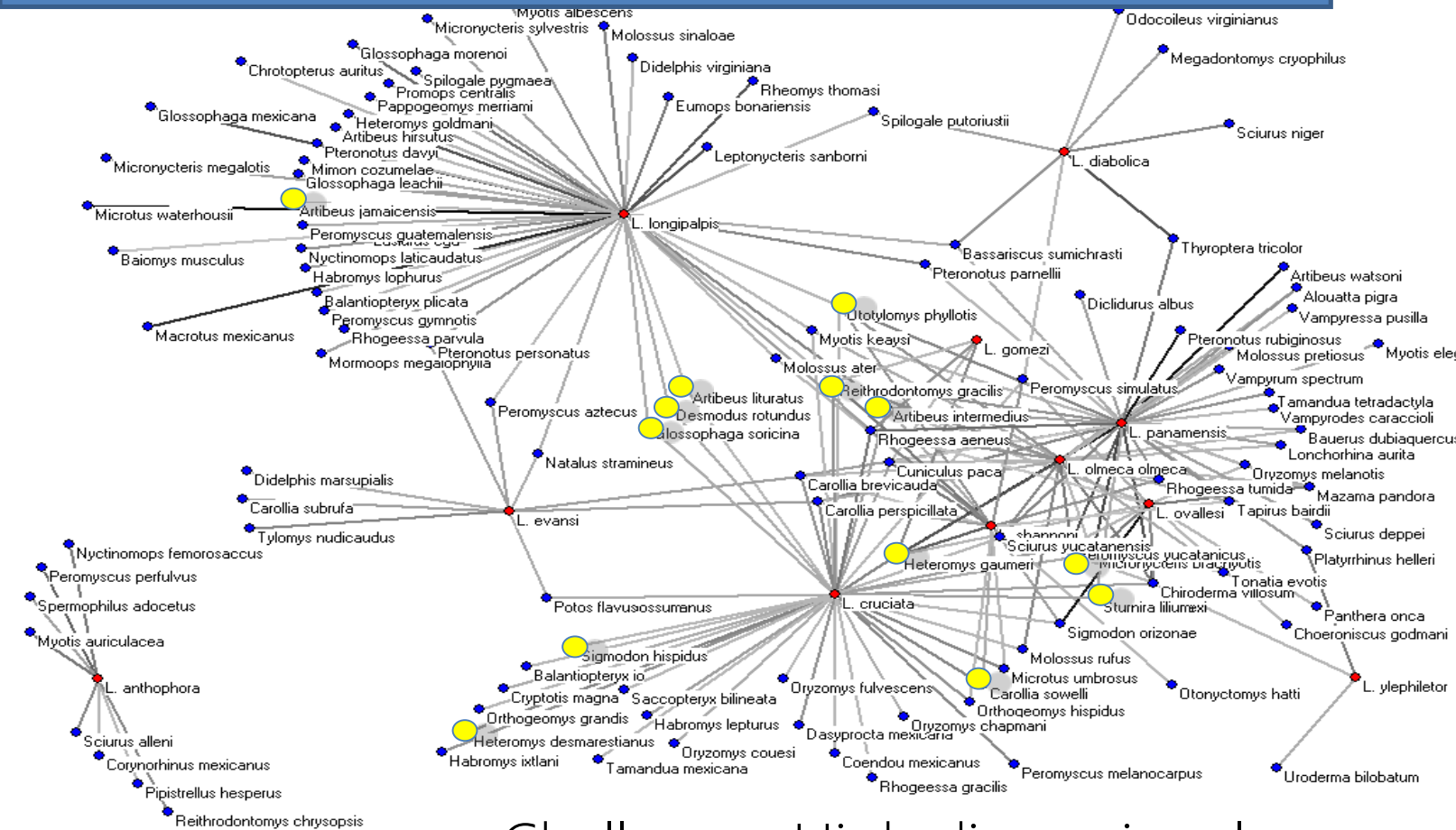
The complex
dynamic, adaptive
"environment"



How to model a complex world?

Inferring ecological interactions for emerging diseases

A network of the most important potential interactions between vector and mammal



By (spatial) datamining species distribution data have identified 21 new species of reservoir of the emerging disease Leishmaniasis

Converted 440 discrete spatial distributions to a easily interpretable complex network

Challenge: High dimensional spatial data mining

Stephens CR, Heau JG, González C, Ibarra-Cerdeña CN, Sánchez-Cordero V, et al. ... **PLoS ONE** 4(5): e5725 (2009)

R. Sierra and C.R. Stephens, International Journal of Geographical Information Systems 26, 4

How to model a complex world?

Predicting the dynamics of high school dropouts



District Level - Student List

May 21, 2012

Home Dropouts School List Program Details Historical Dropout Rates **Student List**

School Year: 2011 Student Number: Last Name: LEEP Available: Show All Regardless of LEEP

Current Year Probability: Minimum 0 % Maximum 0 % School Search: -- All Schools -- Grade Level: 0

Clear Filter Search

School Year	Student Number	Last Name	First Name	School Name	Current Year Probability %	Last Year Probability %	GPA	YTD Absences	In LEEP	Grade Level
2011	136925		Samantha	ACKERLY/BINGHAM HIGH	92.41 %	9.27 %	0.69	38	<input type="checkbox"/>	9
2011	141016		Jason	ACKERLY/BINGHAM HIGH	92.34 %	12.67 %	N/A		<input type="checkbox"/>	9
2011	68124		Aireal	VALLEY TRADITIONAL HIGH	87.07 %	20.01 %	1.38	47	<input type="checkbox"/>	11
2011	109292		Bobby	SHAWNEE HIGH	86.84 %	40.56 %	0.55	44	<input type="checkbox"/>	9
2011	72261		Stephen	VALLEY TRADITIONAL HIGH	86.28 %	41.64 %	0.00		<input type="checkbox"/>	11
2011	75139		Brittany	LIBERTY HIGH	84.71 %	48.37 %	0.00	12	<input type="checkbox"/>	11
2011	71717		Lajuanta	LIBERTY HIGH	83.76 %	49.20 %	0.20	4	<input type="checkbox"/>	12
2011	89388		Demontez	LOUISVILLE METRO YOUTH CENTER	83.06 %	37.16 %	1.12	46	<input type="checkbox"/>	10

Challenge: How do we model the effects of interventions?



www.adaptiveinc.com

District Level - Dropouts

May 21, 2012

Home **Dropouts** School List Program Details Historical Dropout Rates Student List

Dropout Statistics Export to PDF Export to XLS Export to CSV

District Forecast

School Year	Total Population	Actual Dropouts	Predicted Dropouts	Actual Dropout %	Dropout Probability %
2011	99533	1176	1187	1.18 %	1.19 %
2010	104680	1467	1432	1.40 %	1.37 %
2009	104441	2045	2122	1.96 %	2.03 %
2008	104436	2145	2265	2.05 %	2.17 %

District Level - LEEP Intervention Alert

May 21, 2012

Home Dropouts School List Program Details Historical Dropout Rates Student List

Program Impact for: LEEP

Reduction in Dropouts

Last Year's Intervention value	Current Year Intervention Value	Intervention Potential
5%	4%	18%

Show participating Students Show Potential Students Ranked by Highest Impact

Student Number	Last Name	First Name	School Name	Probability without Intervention	Probability with Intervention	Change in Probability with Intervention	Currently in Program
87721	Mustafa	Raymeen	DOSS HIGH	74	16	58	YES
62372		Akeem	SOUTHERN HIGH	69	40	29	YES
96225		Christopher	SHAWNEE HIGH	39	12	27	YES
68939		Courtney	FAIRDALE HIGH	34	12	22	YES
82741		Iyah	DOSS HIGH	40	21	19	YES
82348		Tomiono	DOSS HIGH	30	14	16	YES
106238		Hila	DOSS HIGH	23	9	14	YES
98417		Marshaan	JEFFERSONTOWN HIGH	26	12	14	YES



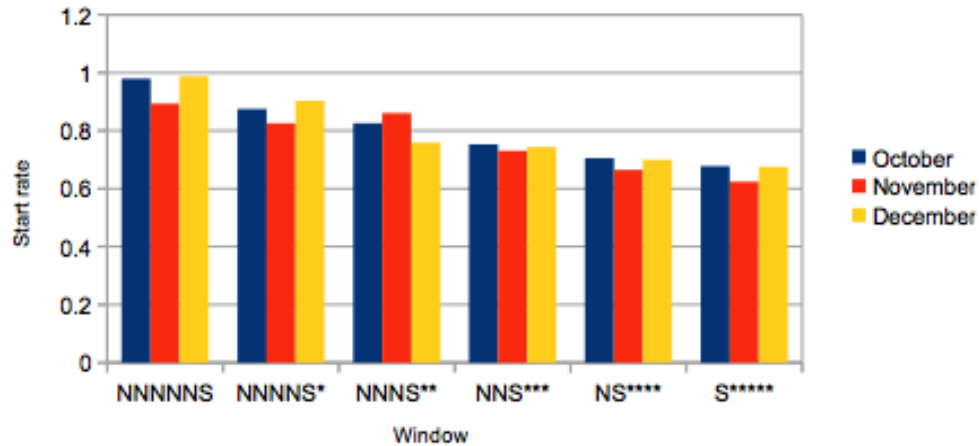
How to model a complex world?

How do you model nothing?

Problem: Predict the probability that someone will start classes at a university.

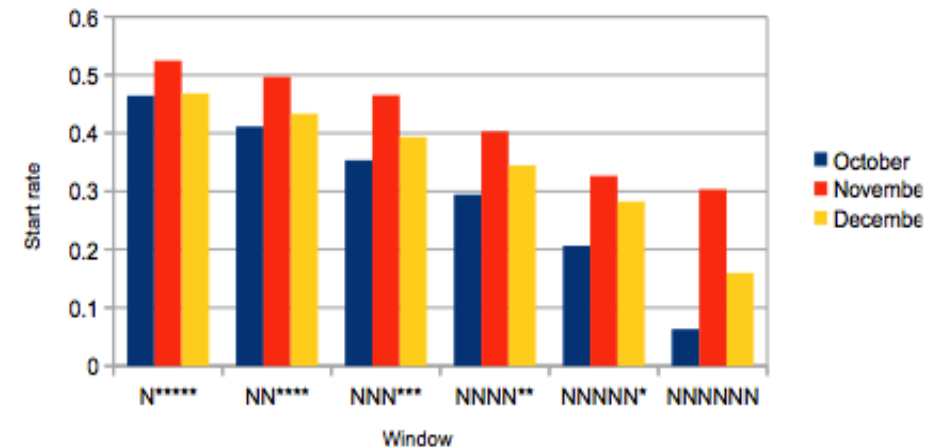
What is the role of **when** an event occurs and...

Start rate for different YES windows for package



...the role of **when** an event doesn't occur, e.g., financial packaging

Graph of start rate for different NO windows



Challenge: How do we model histories?



Conclusions

Much can be learned by considering datamining not as a branch of computer science but as a fundamental task of all evolving systems.

But...

There are biases in the datamining...

Each organism is adapted to sense only certain data types – electromagnetic, acoustic, chemical etc.; inferences are heavily biased to avoiding bad things (false positives much less important than false negatives); feature selection and the restrictions of dimensionality (we live in a Euclidean world); data, although integrated, is generally at the individual level, i.e., not distributed

There is now available vast quantities of data in non-biological substrates



Conclusions

Challenges

1) How do we model complex systems?

- How to incorporate factors from the micro to the macro?
 - How much data? What types? Data integration (connectivity of cyberspace)
- Biological systems already successfully model complex systems (but unconsciously)
 - How do they do it?

i) How do we model complex systems that are represented by large numbers of related spatio-temporal distributions?

ii) How do we model the effect of interventions?

iii) How do we model histories?



Prediction

The XXIst century will be dominated by complexity science

Datamining will be the most important paradigm for modelling complexity and a complex world

Almost all datamining will be in cyberspace