IBM Research: Mobile, Solutions, and Mathematical Sciences

Data Analytics in Industry Research: A Personal Perspective*

Naoki Abe
Senior Manager, Data Analytics
Mathematical Sciences and Analytics
IBM T. J. Watson Research Center

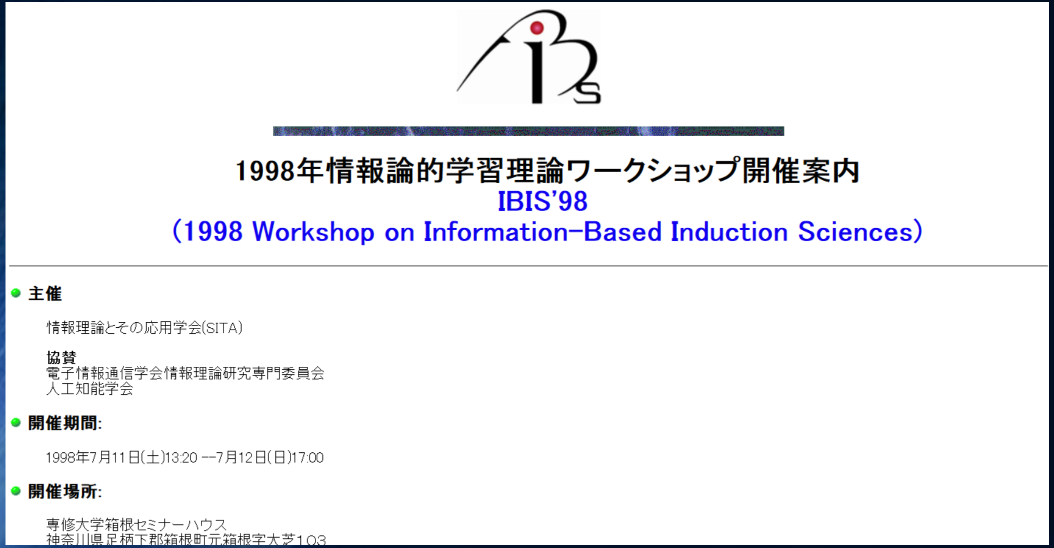
*Based on joint work with H. Mamitsuka, A. Nakamura, H. Li, J. Takeuchi, E. Pednault, P. Melville, S. Rosset, Y. Liu, A. Lozano, R. Luss P. Olsen, E. Yang, K. Ramamurthy, M. Kshirsagar, et al

Agenda

- NEC days (1990’s)
  - Stochastic tree grammars and Genetic Information Processing
  - On-line binary relation learning and Natural Language Processing
  - …
- IBM days (2000’s - present)
  - Smarter Government
    - Markov Decision Process and Tax Collections
  - Smarter Planet , Enterprise & Cloud
    - Temporal Causal Modeling and Applications
  - Smarter Agriculture
    - Heterogeneous Data Analytics and Accelerated Plant Breeding
- Summary and Retrospective

NEC and History of IBIS…

- Founding members of “AI Basic Research Group” (Katsuhiro Nakamura, Kenji Yamanishi, Junichi Takeuchi) had a great idea
- “IBIS” = Endangered Species



1998年情報論的学習理論ワークショップ開催案内
IBIS’98
(1998 Workshop on Information-Based Induction Sciences)

- **主催**

情報理論とその応用学会(SITA)

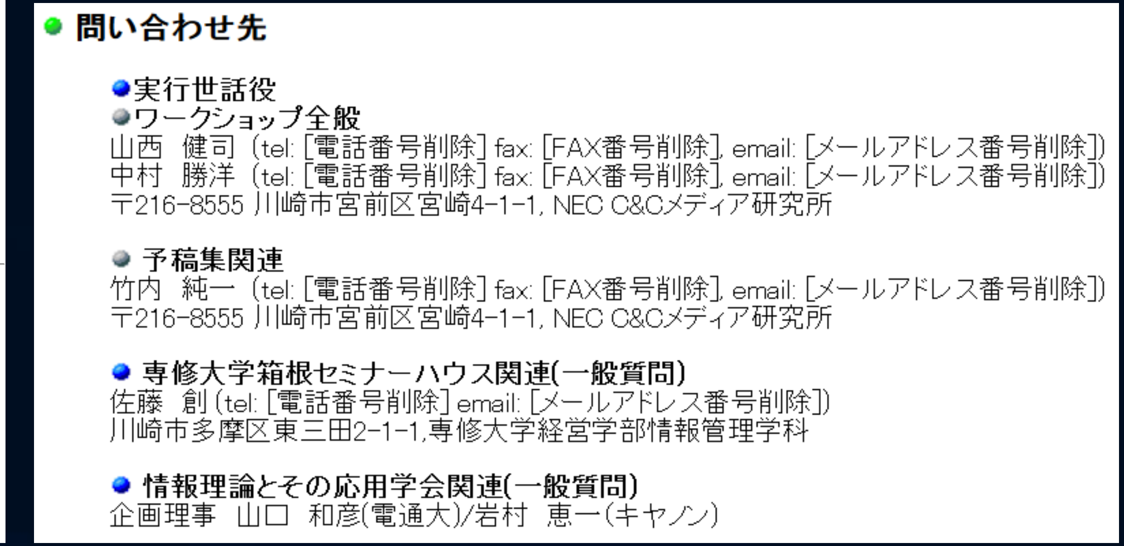
協賛
電子情報通信学会情報理論研究専門委員会
人工知能学会

- **開催期間:**

1998年7月11日(土)13:20 --7月12日(日)17:00

- **開催場所:**

専修大学箱根セミナーハウス
神奈川県足柄下郡箱根町元箱根字大芝103



- **問い合わせ先**

  - 実行世話役
  - ワークショップ全般

山西　健司（tel: [電話番号削除] fax: [FAX番号削除], email: [メールアドレス番号削除]）
中村　勝洋（tel: [電話番号削除] fax: [FAX番号削除], email: [メールアドレス番号削除]）
〒216-8555 川崎市宮前区宮崎4-1-1, NEC C&Cメディア研究所

  - 予稿集関連

竹内　純一（tel: [電話番号削除] fax: [FAX番号削除], email: [メールアドレス番号削除]）
〒216-8555 川崎市宮前区宮崎4-1-1, NEC C&Cメディア研究所

  - 専修大学箱根セミナーハウス関連(一般質問)

佐藤　創（tel: [電話番号削除] email: [メールアドレス番号削除]）
川崎市多摩区東三田2-1-1,専修大学経営学部情報管理学科

  - 情報理論とその応用学会関連(一般質問)

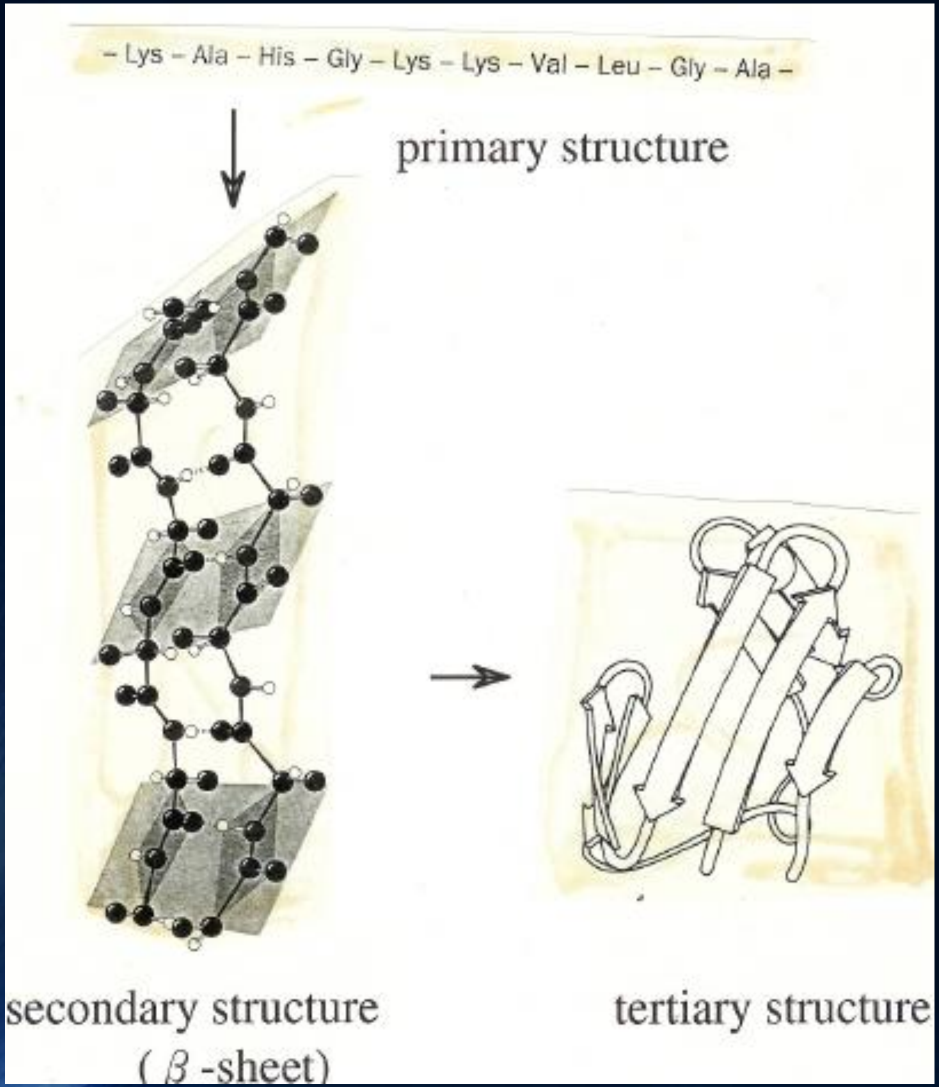
企画理事　山口　和彦(電通大)/岩村　恵一(キヤノン)

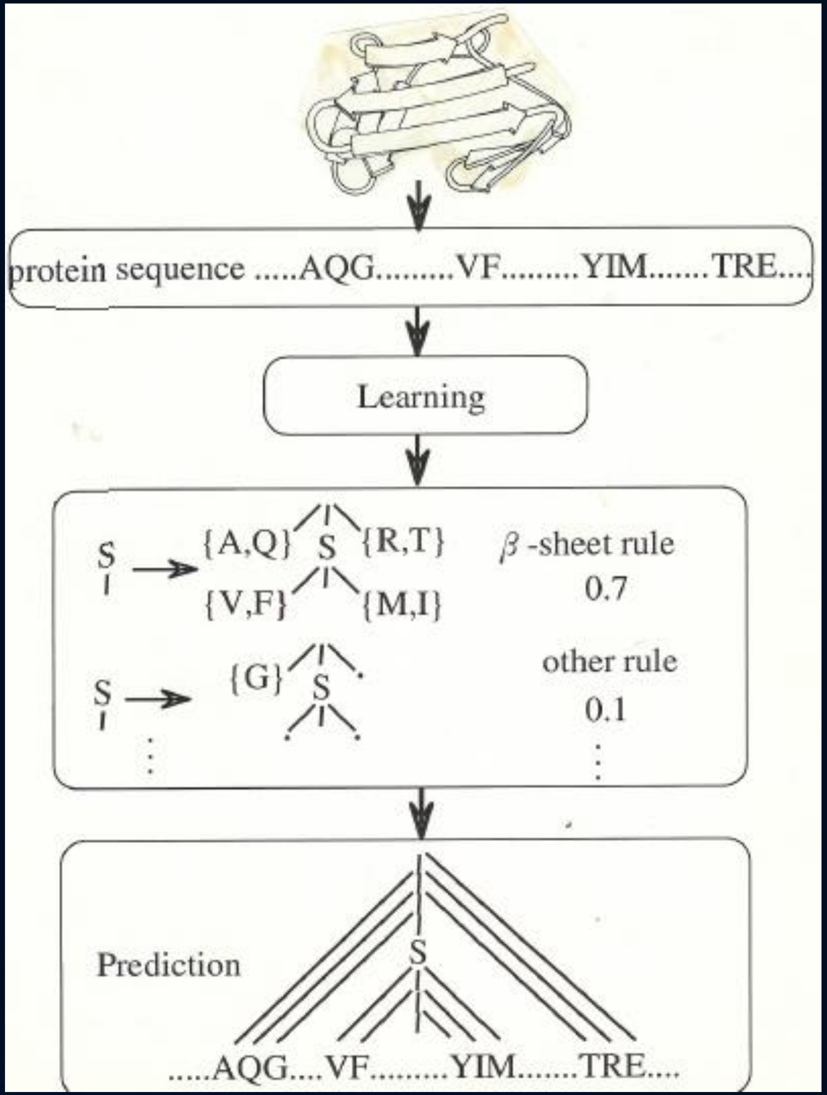
NEC Days: 1990's

- ***Stochastic Tree Grammars and Genetic Information Processing*** (H. Mamitsuka)
- On-line Binary Relation Learning and Natural Language Processing (with A. Nakamura, H. Li)
- On-line Active Learning and Rational Choice Theory (with J. Takeuchi, S. Amari)
- On-line Active Learning and Internet Advertisement Optimization (with A. Nakamura, et al)
- Active Learning and Immunological Experimental Design (H. Mamitsuka, K. Udaka, et al)
- …

Predicting Protein Secondary Structure (Beta-sheet Regions) with Stochastic Tree Grammars

H. Mamitsuka & N. Abe

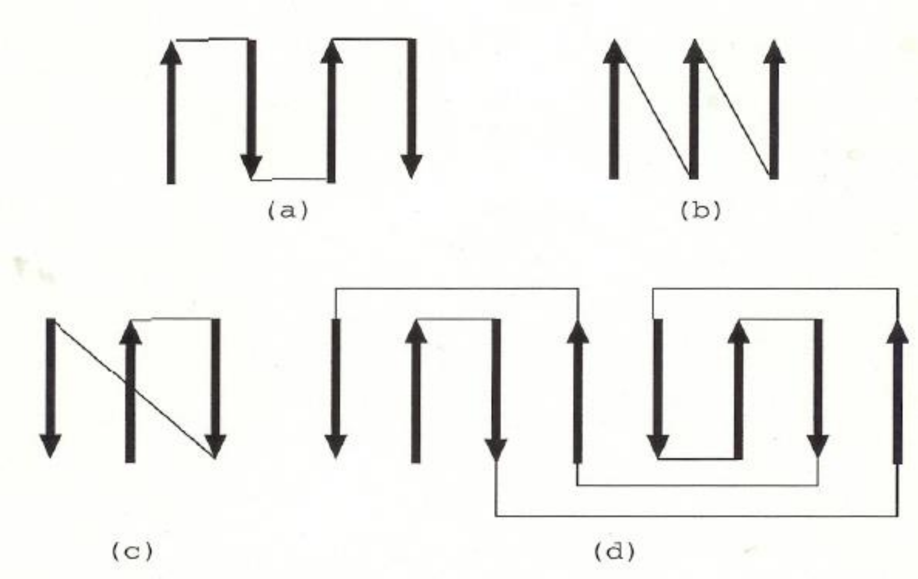


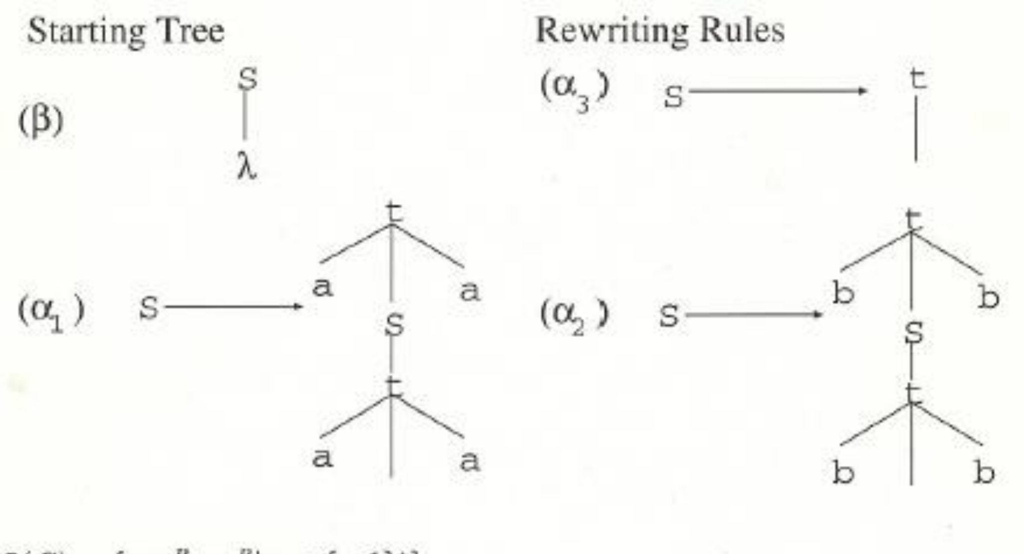


protein sequence .....AQG.........VF.........YIM.......TRE....
Learning
S → {A,Q} S {R,T} {V,F} {M,I}
β -sheet rule
0.7
S → {G} S
other rule
0.1
Prediction
S
.....AQG....VF.........YIM.......TRE....

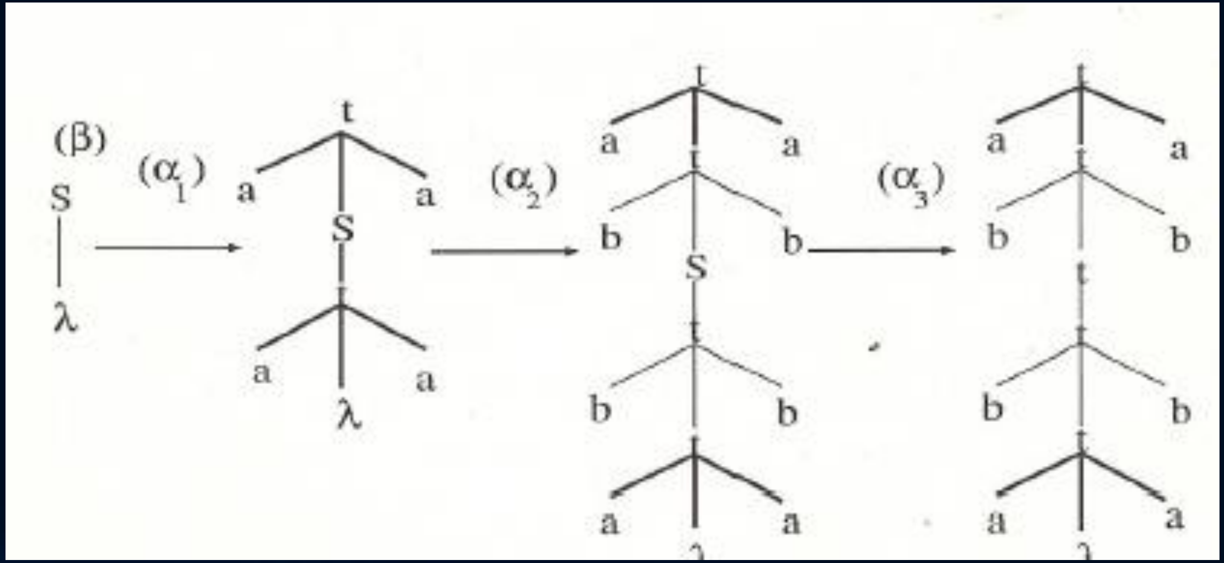
Beta-sheet Patterns and Tree Grammars

- Beta sheet patterns exhibit so-called "unbounded dependencies"
- Ranked Node Rewriting Grammar RNRG inspired by (and generalizes) Tree Adjoining Grammars (Joshi et al)



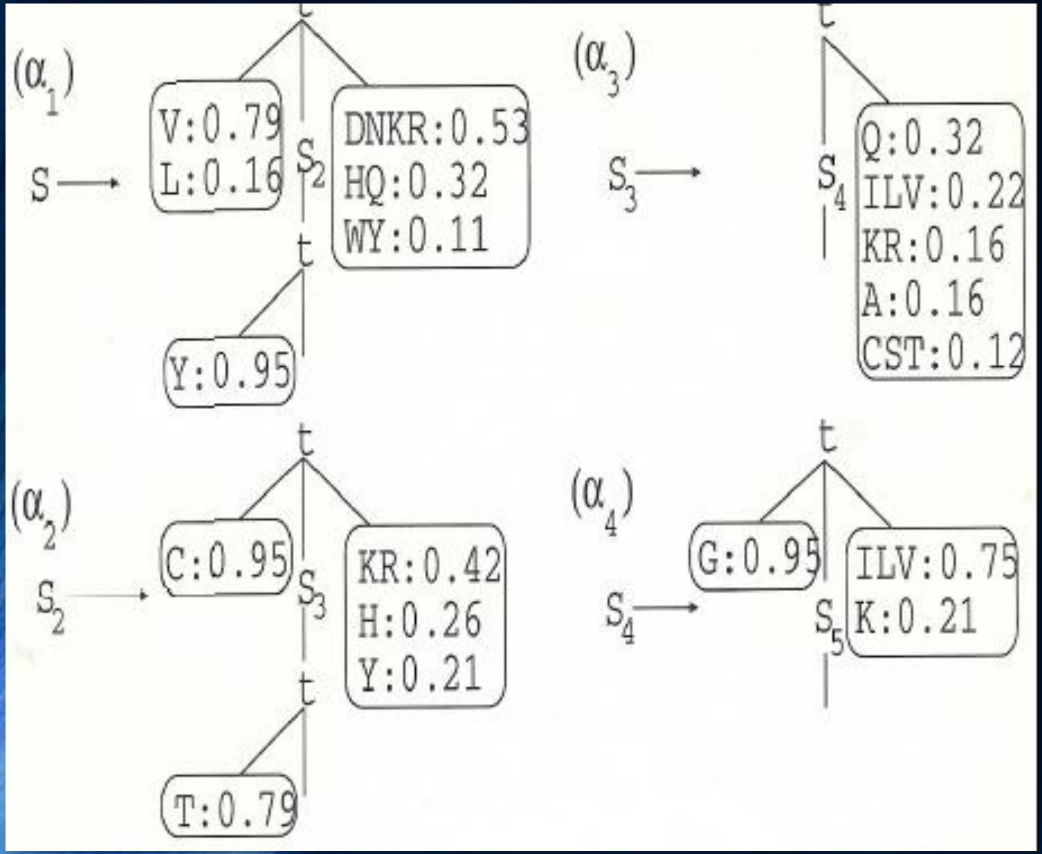


$L(G) = \{ww^R ww^R | w \in \{a, b\}^*\}$

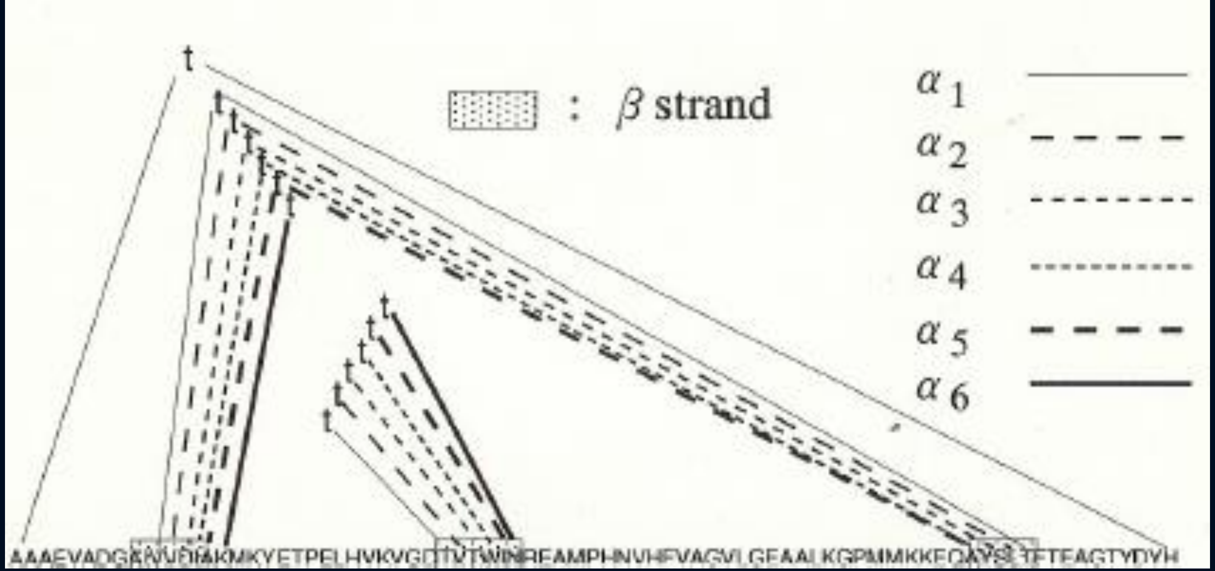


Sample results of predicting beta-strand in a Toxyn

- Learned grammar



- Most likely parse



t
▨ : β strand
α 1
α 2
α 3
α 4
α 5
α 6
AAAEVADGAVVDIAKMKYETPELHVKVGDTVTWINREAMPHNVHFVAGVLGEAALKGPMMKKEQAYSLTFTEAGTYDYH

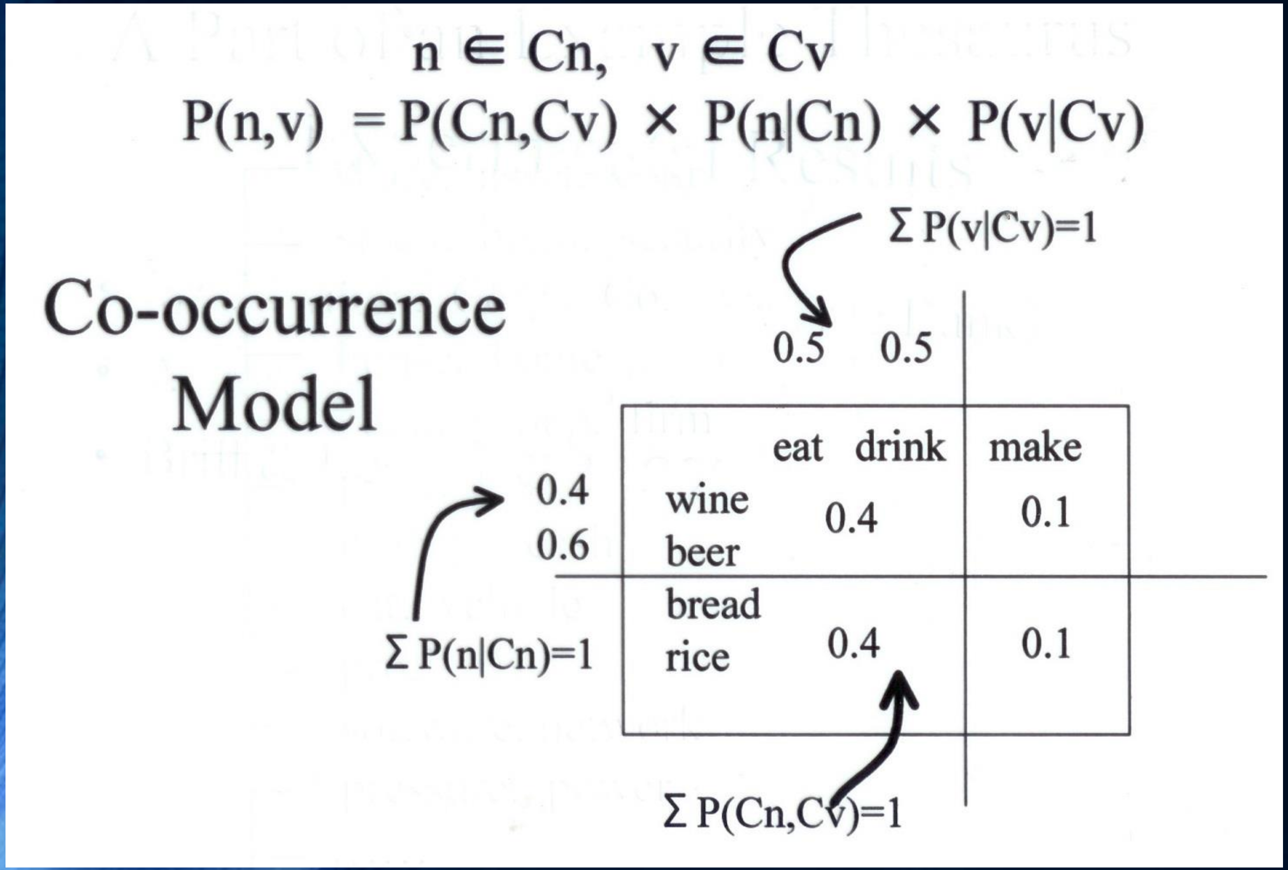
NEC Days: 1990's

- Stochastic Tree Grammars and Genetic Information Processing (H. Mamitsuka)
- ***On-line Binary Relation Learning and Natural Language Processing*** (with A. Nakamura, H. Li)
- On-line Active Learning and Rational Choice Theory (with J. Takeuchi)
- On-line Active Learning and Internet Advertisement Optimization (with A. Nakamura)
- Active Learning and Immunological Experimental Design (H. Mamitsuka)
- …

On-line binary relation learning and word clustering

A. Nakamura, H. Li & N. Abe

- Sub-categorization in NLP: naturally viewed as clustering via density estimation



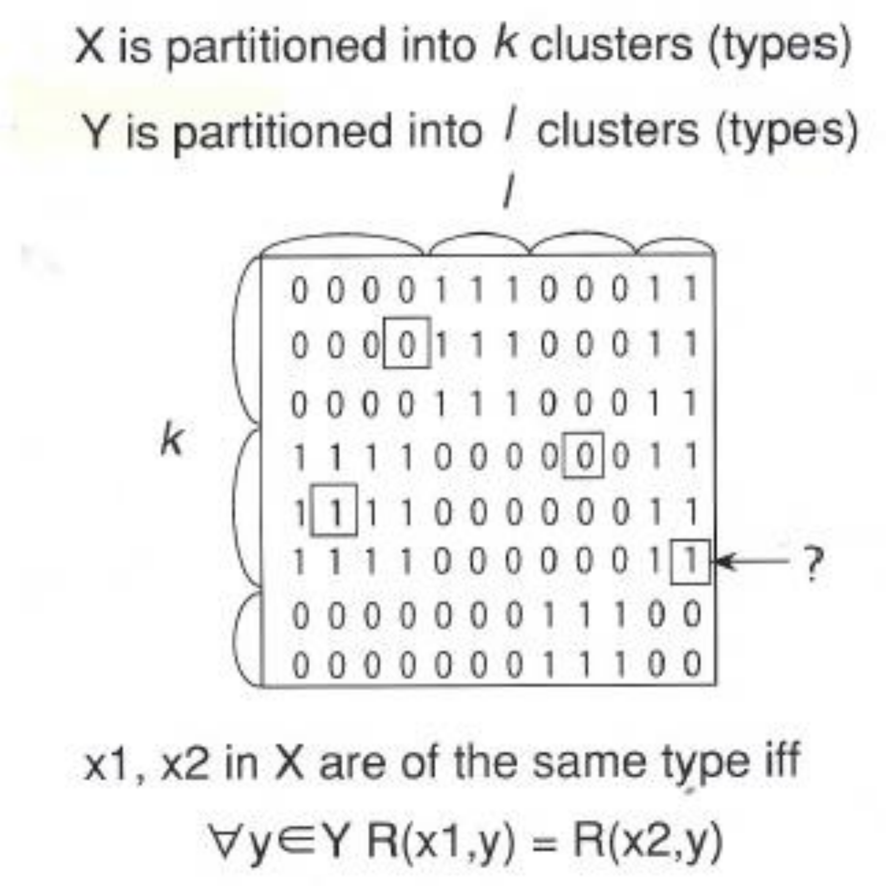
$n \in Cn, \quad v \in Cv$

$$P(n,v) = P(Cn,Cv) \times P(n|Cn) \times P(v|Cv)$$

Co-occurrence Model

	eat drink	make
wine beer	0.4	0.1
bread rice	0.4	0.1

- An Alternative Formulation: On-line Learning of Binary Relations



X is partitioned into k clusters (types)

Y is partitioned into l clusters (types)

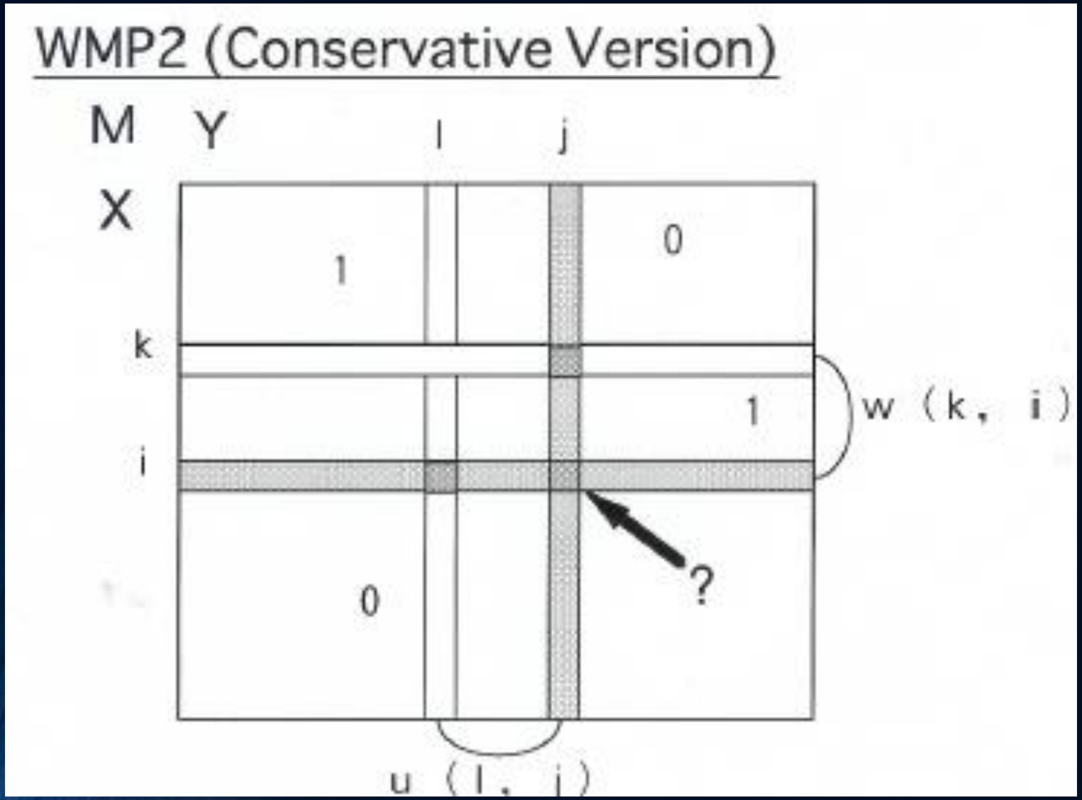
$$\begin{array}{cccccccccccc}
0&0&0&0&1&1&1&0&0&0&1&1\\
0&0&0&\boxed{0}&1&1&1&0&0&0&1&1\\
0&0&0&0&1&1&1&0&0&0&1&1\\
1&1&1&1&0&0&0&0&\boxed{0}&0&1&1\\
1&\boxed{1}&1&1&0&0&0&0&0&0&1&1\\
1&1&1&1&0&0&0&0&0&0&1&\boxed{1}\\
0&0&0&0&0&0&0&1&1&1&0&0\\
0&0&0&0&0&0&0&1&1&1&0&0
\end{array}$$

x1, x2 in X are of the same type iff

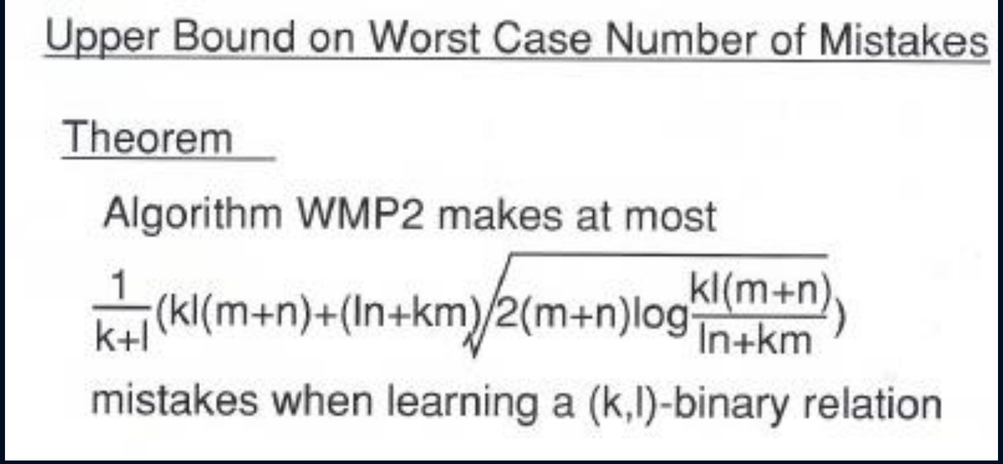
$\forall y \in Y \; R(x1,y) = R(x2,y)$

On-line Learning: Minimize number of prediction errors!

Two-dimensional Weighted Majority Algorithm Out-Performs Heuristic Methods



WMP2 (Conservative Version)



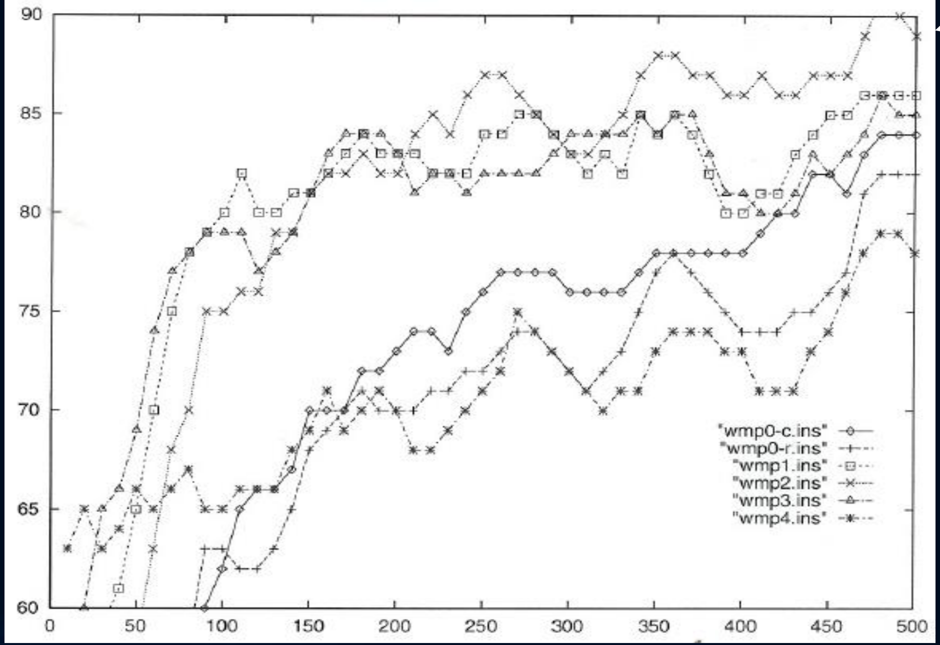
Upper Bound on Worst Case Number of Mistakes

Theorem

Algorithm WMP2 makes at most

$$\frac{1}{k+l}(kl(m+n)+(ln+km)\sqrt{2(m+n)\log\frac{kl(m+n)}{ln+km}})$$

mistakes when learning a (k,l)-binary relation



WMP2

2-dimensional WMPs

1-dimensional WMPs

- 2-dimensional WMP algorithms outperform 1-dimensional ones
- Theoretically sound algorithm (WMP2) outperforms heuristic ones…

IBM Days (2000's – present)

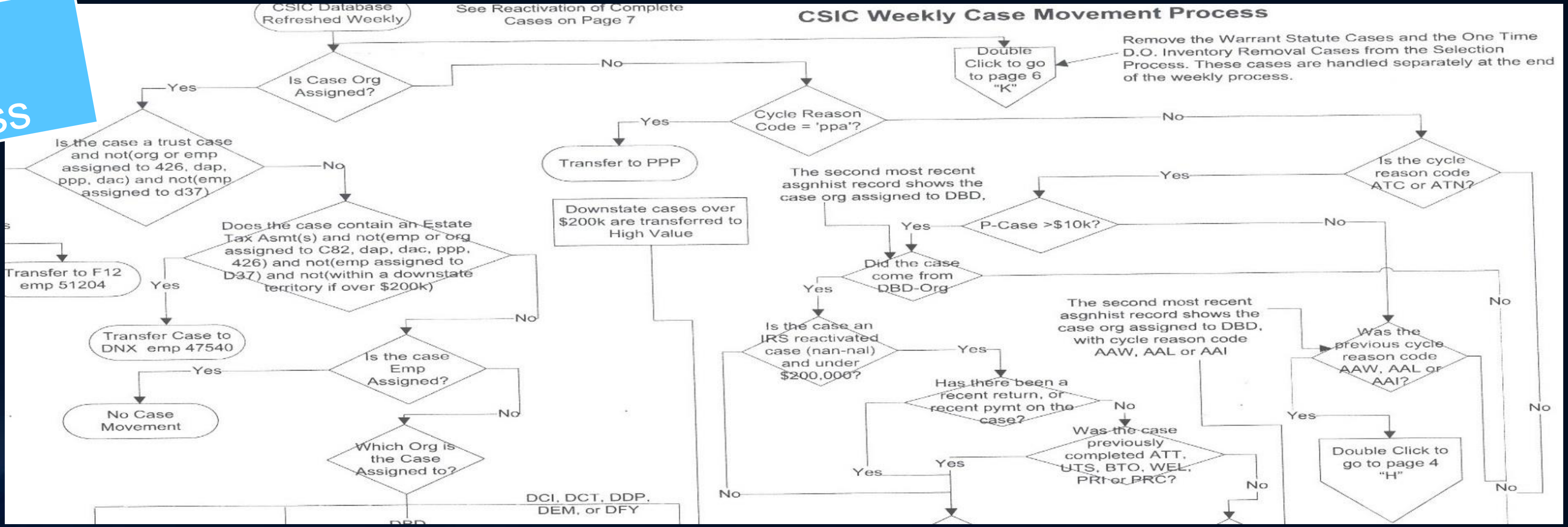
- ***Smarter Government***
  - ***Markov Decision Process and Tax Collections***
- Smarter Planet , Enterprise & Cloud
  - Temporal Causal Modeling and Applications
- Smarter Agriculture
  - Heterogeneous Data Analytics and Accelerated Plant Breeding

Smarter Government

- Tax Collections Optimization System (TACOS) video
  - https://www.youtube.com/watch?v=VGKp13APsBg

Tax Collections Optimization for NYS: Technical Challenge

- Tax collections process is a complex process involving various legal/business constraints
- Most existing approaches rely on rigid, manual rules, including NYS legacy system
- Goal: take this rigid procedure apart, leaving fragments of it intact wherever necessary, and automatically configure the rest, based on analytics and optimization

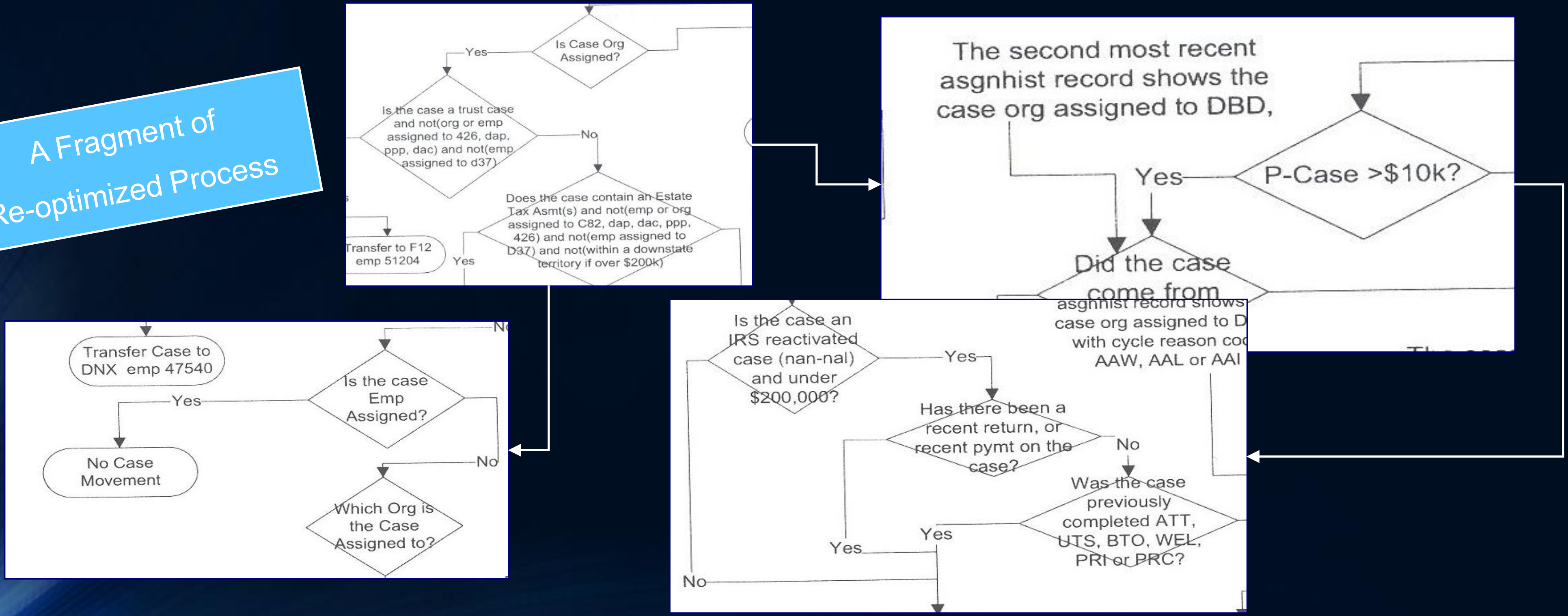

CSIC Weekly Case Movement Process
CSIC Database Refreshed Weekly
See Reactivation of Complete Cases on Page 7
Double Click to go to page 6 "K"
Remove the Warrant Statute Cases and the One Time D.O. Inventory Removal Cases from the Selection Process. These cases are handled separately at the end of the weekly process.
Is Case Org Assigned?
Yes
No
Is the case a trust case and not(org or emp assigned to 426, dap, ppp, dac) and not(emp assigned to d37)
Cycle Reason Code = 'ppa'?
Transfer to PPP
Is the cycle reason code ATC or ATN?
The second most recent asgnhist record shows the case org assigned to DBD,
P-Case >$10k?
Downstate cases over $200k are transferred to High Value
Does the case contain an Estate Tax Asmt(s) and not(emp or org assigned to C82, dap, dac, ppp, 426) and not(emp assigned to D37) and not(within a downstate territory if over $200k)
Transfer to F12 emp 51204
Did the case come from DBD-Org
Transfer Case to DNX emp 47540
Is the case an IRS reactivated case (nan-nal) and under $200,000?
The second most recent asgnhist record shows the case org assigned to DBD, with cycle reason code AAW, AAL or AAI
Was the previous cycle reason code AAW, AAL or AAI?
Is the case Emp Assigned?
Has there been a recent return, or recent pymt on the case?
No Case Movement
Was the case previously completed ATT, UTS, BTO, WEL, PRI or PRC?
Which Org is the Case Assigned to?
Double Click to go to page 4 "H"
DCI, DCT, DDP, DEM, or DFY
DBD

A Fragment of Legacy Process

Tax Collections Optimization for NYS: Technical Challenge

- Tax collections process is a complex process involving various legal/business constraints
- Most existing approaches rely on rigid, manual rules, including NYS legacy system
- Goal: take this rigid procedure apart, leaving fragments of it intact wherever necessary, and automatically configure the rest, based on analytics and optimization

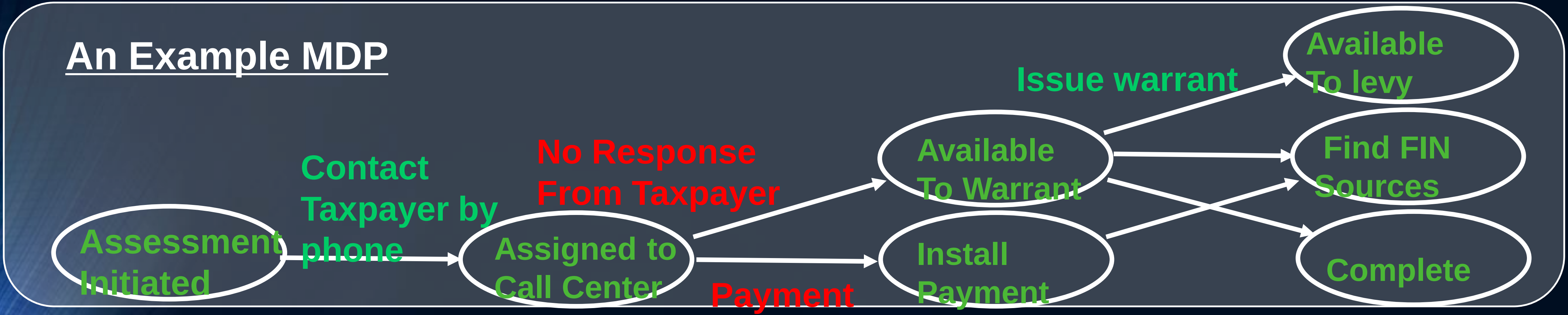
A Fragment of Re-optimized Process



Is Case Org Assigned?
Yes
Is the case a trust case and not(org or emp assigned to 426, dap, ppp, dac) and not(emp assigned to d37)
No
Does the case contain an Estate Tax Asmt(s) and not(emp or org assigned to C82, dap, dac, ppp, 426) and not(emp assigned to D37) and not(within a downstate territory if over $200k)
Transfer to F12 emp 51204
Yes
The second most recent asgnhist record shows the case org assigned to DBD,
Yes
P-Case >$10k?
Did the case come from
asgnhist record shows case org assigned to D with cycle reason co AAW, AAL or AAI
Transfer Case to DNX emp 47540
Yes
Is the case Emp Assigned?
No Case Movement
No
Which Org is the Case Assigned to?
Is the case an IRS reactivated case (nan-nal) and under $200,000?
Yes
Has there been a recent return, or recent pymt on the case?
No
Was the case previously completed ATT, UTS, BTO, WEL, PRI or PRC?
Yes
Yes
No

Tax Collections Optimization for NYS: The Framework

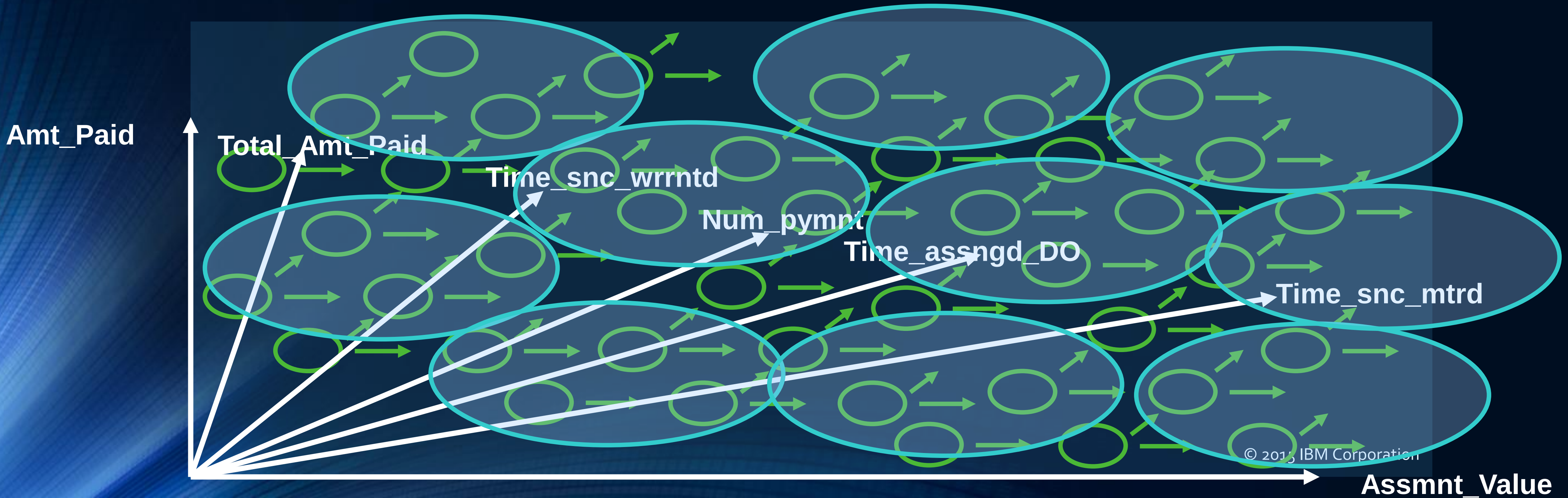
- **Markov Decision Process (MDP) formulation provides an advanced framework for modeling tax collection process**
  - "States", *s,* summarize information on a taxpayer's stage in collection process
  - "Action", *a,* is a collection action (e.g. phone call, warrant, levy)
  - "Reward", *r,* is the tax collected for the taxpayer in question



- The goal in MDP is formulated as outputting a policy which maps TP's states to collection actions so as to maximize the long term cumulative rewards
- **Constrained MDP** requires additionally that output policy belongs to a constrained class adhering to certain constraints

Methodology: Constrained Reinforcement Learning (C-RL)

- **Business requirement to customize collections actions depending on detailed taxpayer characteristics**
  - Use of a high-dimensional state space necessary
- **With high-dimensional state space**
  - Estimating the structure of MDP is extremely challenging
  - Reinforcement Learning (RL) solves MDP with access only to data (not MDP itself)
  - We develop constrained-RL (C-RL) methods for high dimensional state space



Tax Collections Optimization for NYS: Example Segment

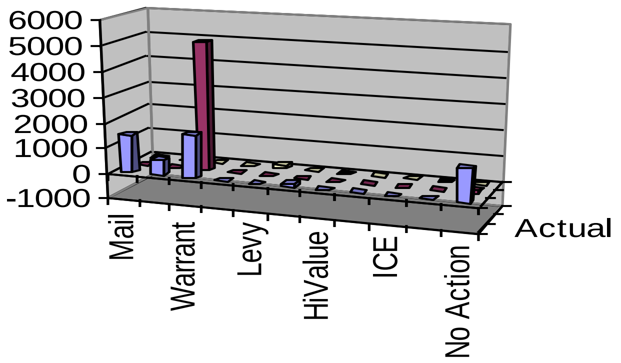
Segment Definition

state = Call-Center-Not-Warranted
and tax_pd_lst_yr < $X
and 1 <= num_st_ff_cllct_asmts
and 1 <= num_non_rstrctd_fin_srcs
and 1 <= st_inactv_ind
and num_pymnts_snc_lst_actn < 1

- **Interpretation**
  - Case is in Call Center and has not been warranted
  - There is at least one non restricted fin source identified
  - Sales tax inactive indicator is on
  - There was no payment in the last period
  - Tax paid last year is less than 1000 dollars
- **Create warrant recommended**

Segment size 1.0%

Action Distributions Segment 212



	Mail	Phone	Warrant	FS IE	Levy	DO	HiValue	CVS	ICE	Field Visit	No Action
🟦 Actual	1522	612	1702	0	3	126	2	0	0	3	1293
🟥 Allocated	0	0	5103	0	0	0	0	0	0	0	152
🟨 Coefficients	-0.208	0.821	1.934	0	130.22	0.662	-0.109	0	0	-0.139	0

Deployment Results: Tax Collection Optimization for NYS

- 2010 realized an 8.22% revenue increase over 2009 - $+83M

Statistics	
Average age of cases	9.3% decrease
Dollar per staff day for field agents	15% Increase
Collection by field staff	12% increase
Dollar per warrant	22% increase
Dollar per levy	11% increase
Number of warrants filed	9% decrease
Number of levies served	3% decrease
35,000 less taxpayers had serious actions taken against them	

Many taxpayers that would have had action taken in the past had no action taken yet debt was still collected

Edelman Presentation NYS DTF Final updated 4-11-2011

THE FRANZ EDELMAN AWARD
Achievement in Operations Research

Recognition in Academia:

- **Awarded "best government & industry paper" at KDD 2010**
- **Finalist for Edelman prize of INFORMS 2011**

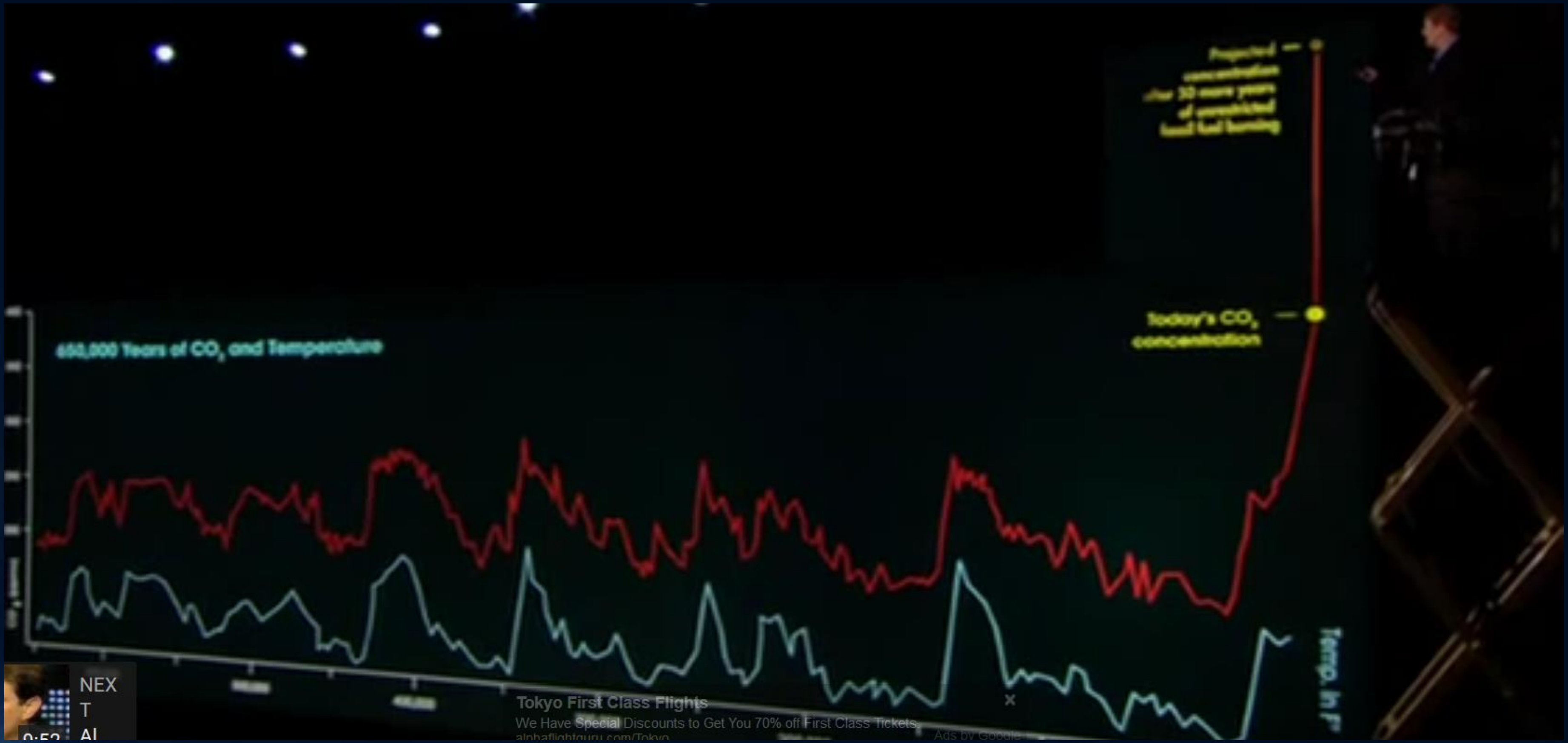
KDD-2010
The 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining
WASHINGTON, DC • JULY 25-28, 2010

IBM Days (2000’s – present)

- Smarter Government
  - Markov Decision Process and Tax Collections
- ***Smarter Planet , Enterprise & Cloud***
  - ***Temporal Causal Modeling and Applications***
- Smarter Agriculture
  - Heterogeneous Data Analytics and Accelerated Plant Breeding

Smarter Planet (and Enterprise and Cloud…)

- “The most ridiculous thing I’ve ever heard…”


Projected concentration after 50 more years of unrestricted fossil fuel burning
Today's CO_2 concentration
650,000 Years of CO_2 and Temperature
Temp. in F
NEX T AI
0:52
Tokyo First Class Flights
We Have Special Discounts to Get You 70% off First Class Tickets
alphaflightguru.com/Tokyo
Ads by Google

Source: “An Inconvenient Truth” the movie

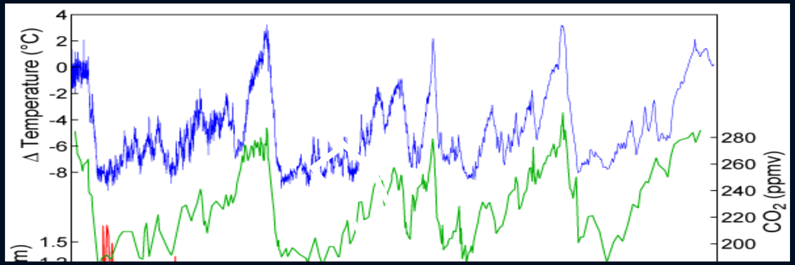
Projected gap between demand and supply for major row crops

Temporal Causal Modeling by Graphical Granger Modeling

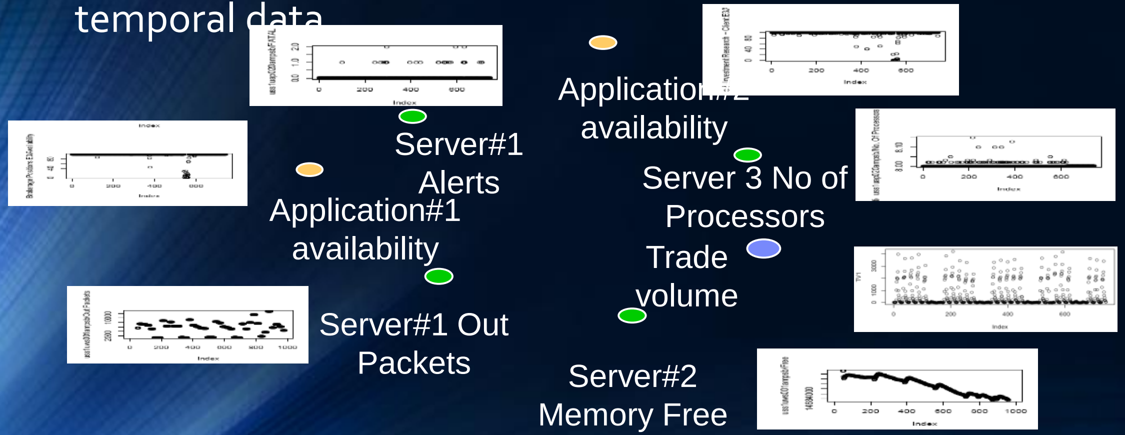
- Granger causality
  - First introduced by the Nobel prize winning economist, Clive Granger
- Definition: a time series x is said to "Granger cause" another time series y, if and only if regressing for y in terms of both past values of y and x is statically significantly better than that of regressing in terms of past values of y only

$$y_t \approx A \cdot \vec{y_{t-1}} + B \cdot \vec{x_{t-1}} \qquad (1)$$

$$y_t \approx A \cdot \vec{y_{t-1}} \qquad (2)$$



- Combination of Granger Causality and cutting-edge graphical modeling techniques provides efficient and effective methodology for graphical causal modeling of temporal data



Server#1 Alerts

Application#1 availability

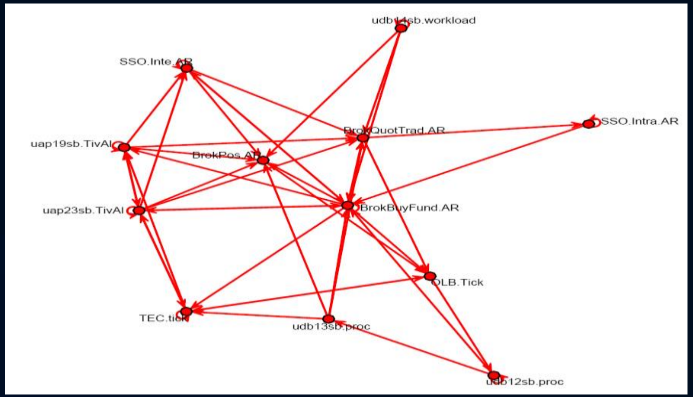
Server#1 Out Packets

Application#2 availability

Server 3 No of Processors

Trade volume

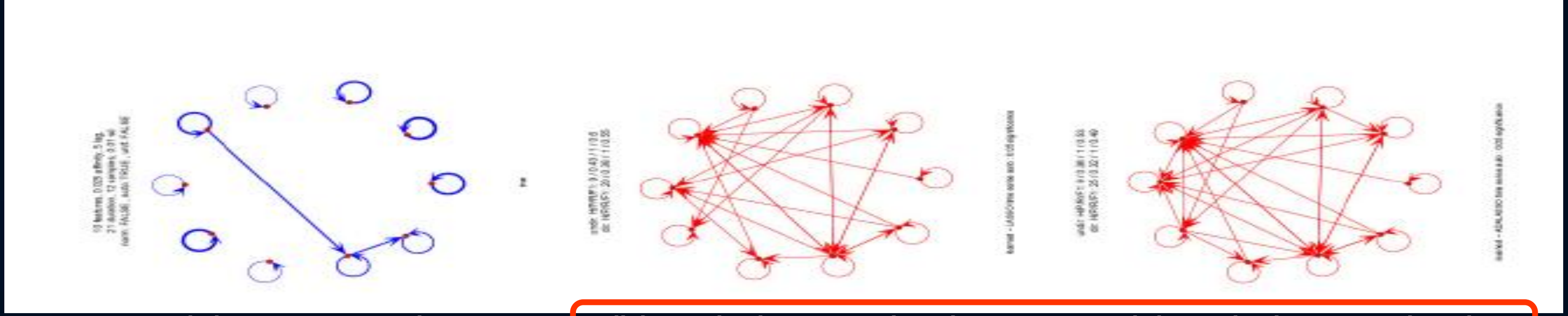
Server#2 Memory Free



Novel Graphical Granger Modeling Methods

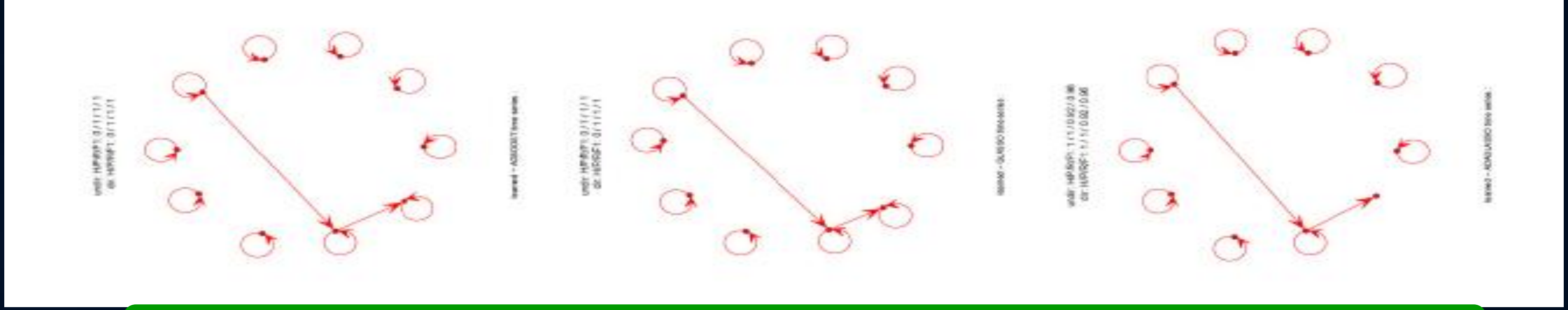
- Our methodology leverages temporal constraints in graphical Granger modeling by treating lagged variables of the same feature as a group, and invokes ***"structured sparse modeling"*** technique

Example Outputs



Existing methods

(a) True graph (b) Existing Method 1 (c) Existing Method 2



Our methods

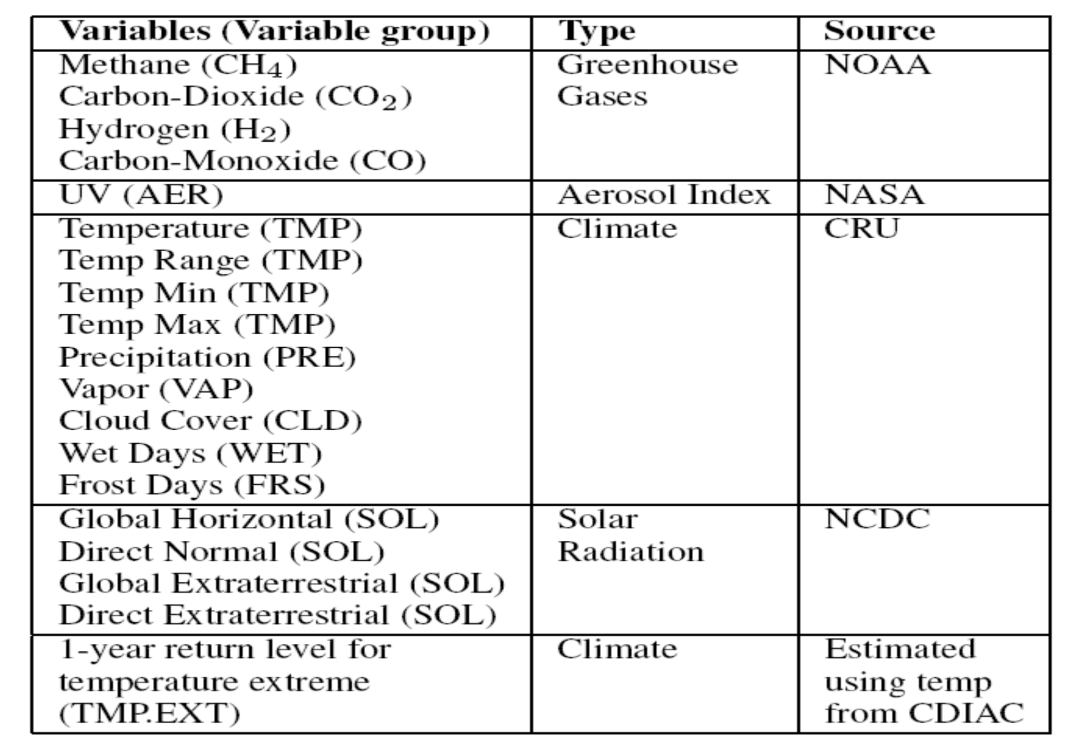
(d) Our Method 1 (e) Our Method 2 (f) Our Method 3

Accuracy Comparison

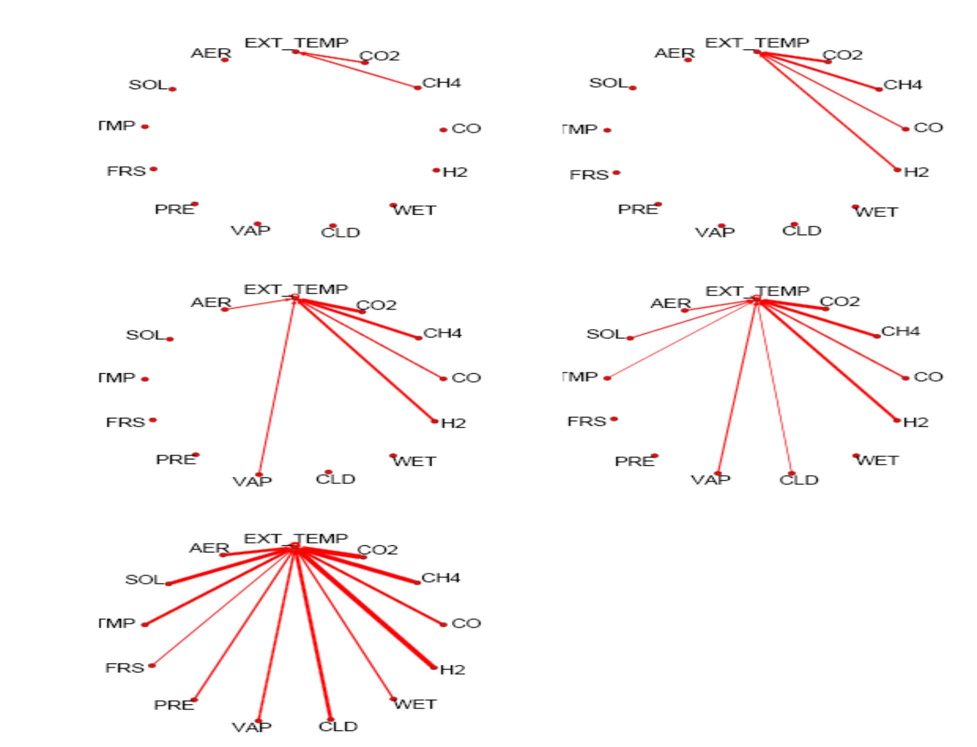
Method	Existing 1	Existing 2	Our 1	Our 2	Our 3
Accuracy(%)	62	65	92	87	91

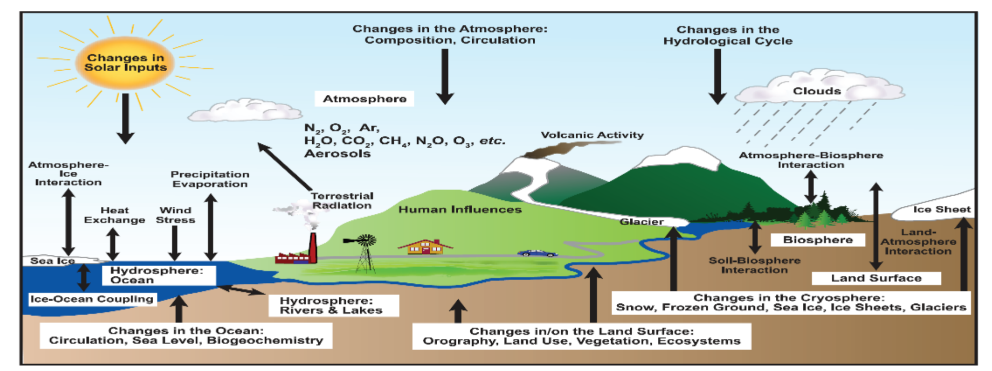
Climate change attribution by spatial-temporal causal modeling

- Identify causal relationships between natural/anthropogenic forcing factors and climate variables, based on spatio-temporal observations.



Variables (Variable group)	Type	Source
Methane (CH_4) Carbon-Dioxide (CO_2) Hydrogen (H_2) Carbon-Monoxide (CO)	Greenhouse Gases	NOAA
UV (AER)	Aerosol Index	NASA
Temperature (TMP) Temp Range (TMP) Temp Min (TMP) Temp Max (TMP) Precipitation (PRE) Vapor (VAP) Cloud Cover (CLD) Wet Days (WET) Frost Days (FRS)	Climate	CRU
Global Horizontal (SOL) Direct Normal (SOL) Global Extraterrestrial (SOL) Direct Extraterrestrial (SOL)	Solar Radiation	NCDC
1-year return level for temperature extreme (TMP.EXT)	Climate	Estimated using temp from CDIAC

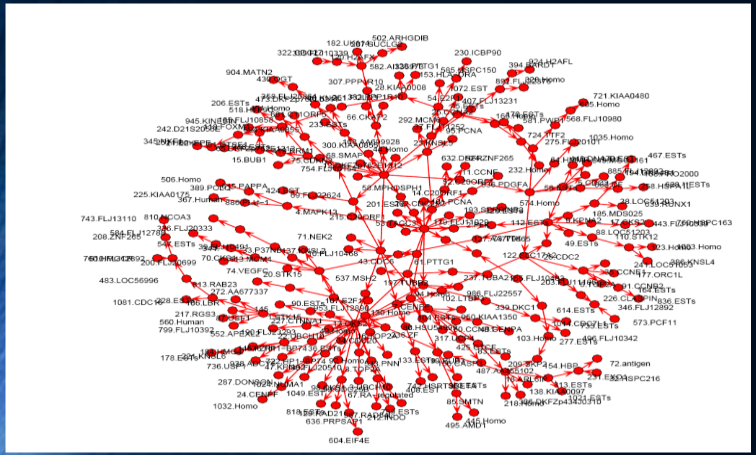



Changes in Solar Inputs
Changes in the Atmosphere: Composition, Circulation
Changes in the Hydrological Cycle
Clouds
Atmosphere
N$_2$, O$_2$, Ar, H$_2$O, CO$_2$, CH$_4$, N$_2$O, O$_3$, etc. Aerosols
Volcanic Activity
Atmosphere-Biosphere Interaction
Atmosphere-Ice Interaction
Precipitation Evaporation
Terrestrial Radiation
Heat Exchange
Wind Stress
Human Influences
Glacier
Ice Sheet
Biosphere
Land-Atmosphere Interaction
Sea Ice
Hydrosphere: Ocean
Soil-Biosphere Interaction
Land Surface
Ice-Ocean Coupling
Hydrosphere: Rivers & Lakes
Changes in the Cryosphere: Snow, Frozen Ground, Sea Ice, Ice Sheets, Glaciers
Changes in the Ocean: Circulation, Sea Level, Biogeochemistry
Changes in/on the Land Surface: Orography, Land Use, Vegetation, Ecosystems

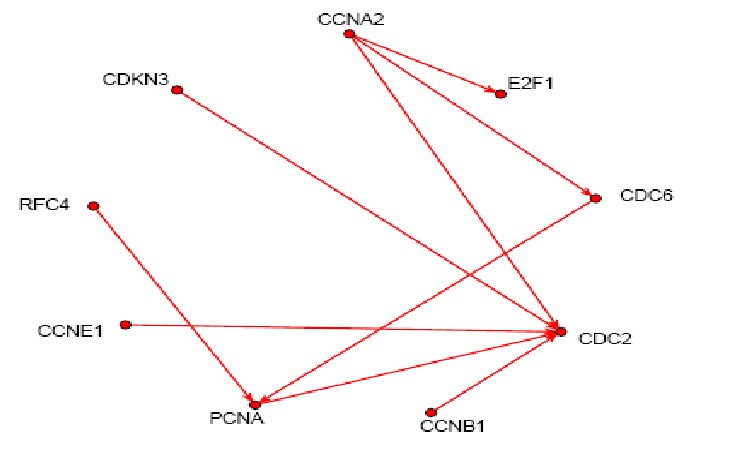
IPCC – Climate Change 2001, the scientific basis

Modeling Gene Regulations in Micro-array Data

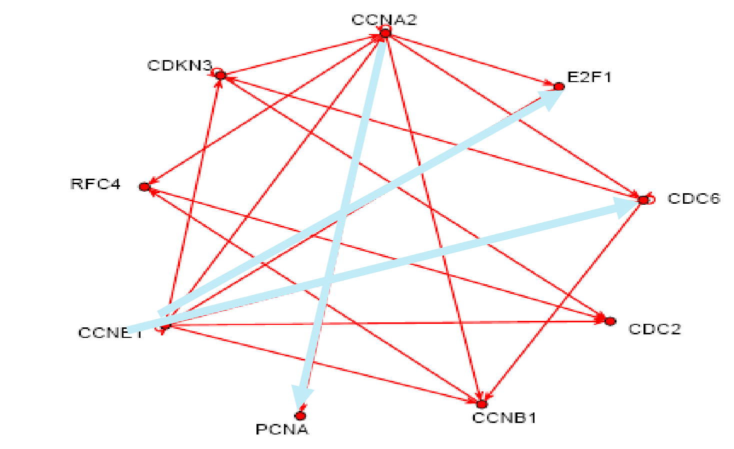
- We applied our Grouped Graphical Granger Modeling technique to discover gene expression regulatory networks for the human cancer cell HeLa S3
- Our method achieved higher accuracy, and uncovered previously unknown relationships, given data involving 1134 genes
    - CCNA2 -> PCNA verified in [Liu, et al 2007]
    - CCNE1 -> ETF1 verified in [Merdzhanova, et al 2007]
    - CCNE1 -> CDC6 verified in [Furstenthal, et al 2001]



Full output network



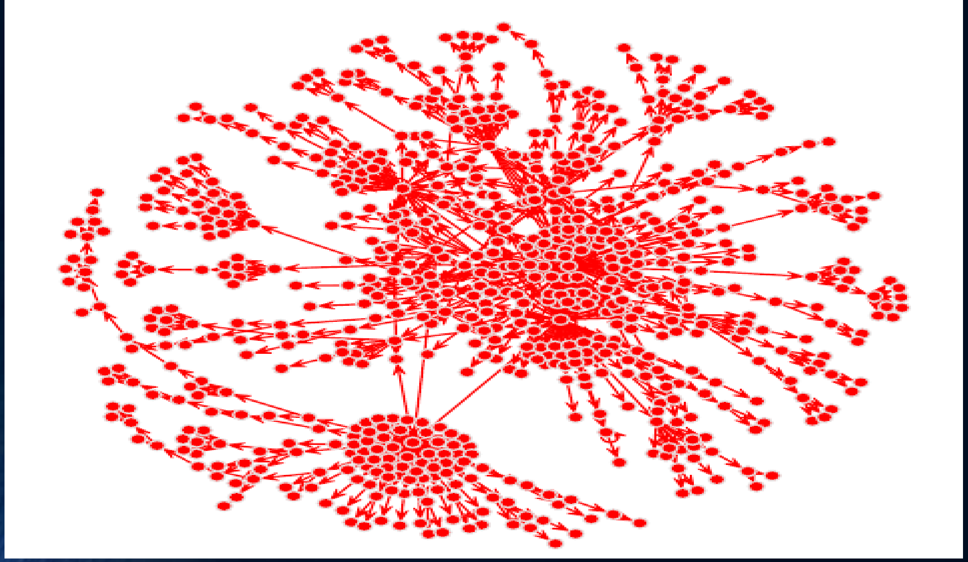
Known links in database (BioGRID)

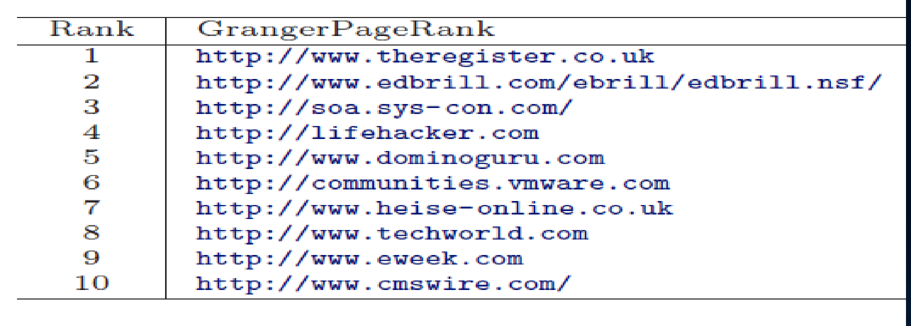


Causal links discovered by our method

Social Media Analytics: Key Influencer Identification

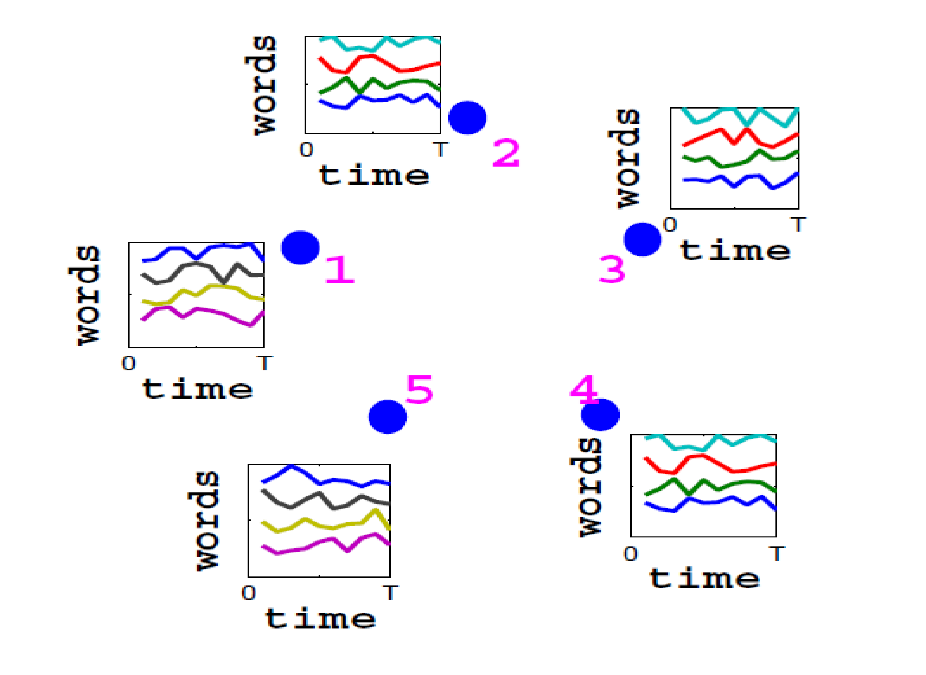
Question 1: Who is influencing whom?
Question 2: What is the context of influence?
Question 3: Who is most influential on a given topic?





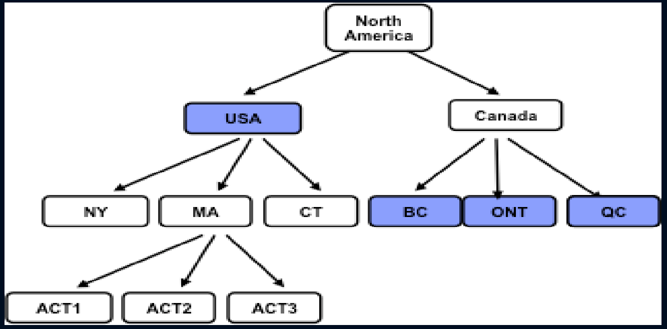
Rank	GrangerPageRank
1	http://www.theregister.co.uk
2	http://www.edbrill.com/ebrill/edbrill.nsf/
3	http://soa.sys-con.com/
4	http://lifehacker.com
5	http://www.dominoguru.com
6	http://communities.vmware.com
7	http://www.heise-online.co.uk
8	http://www.techworld.com
9	http://www.eweek.com
10	http://www.cmswire.com/

Variables are Bloggers represented in terms of word usage across time

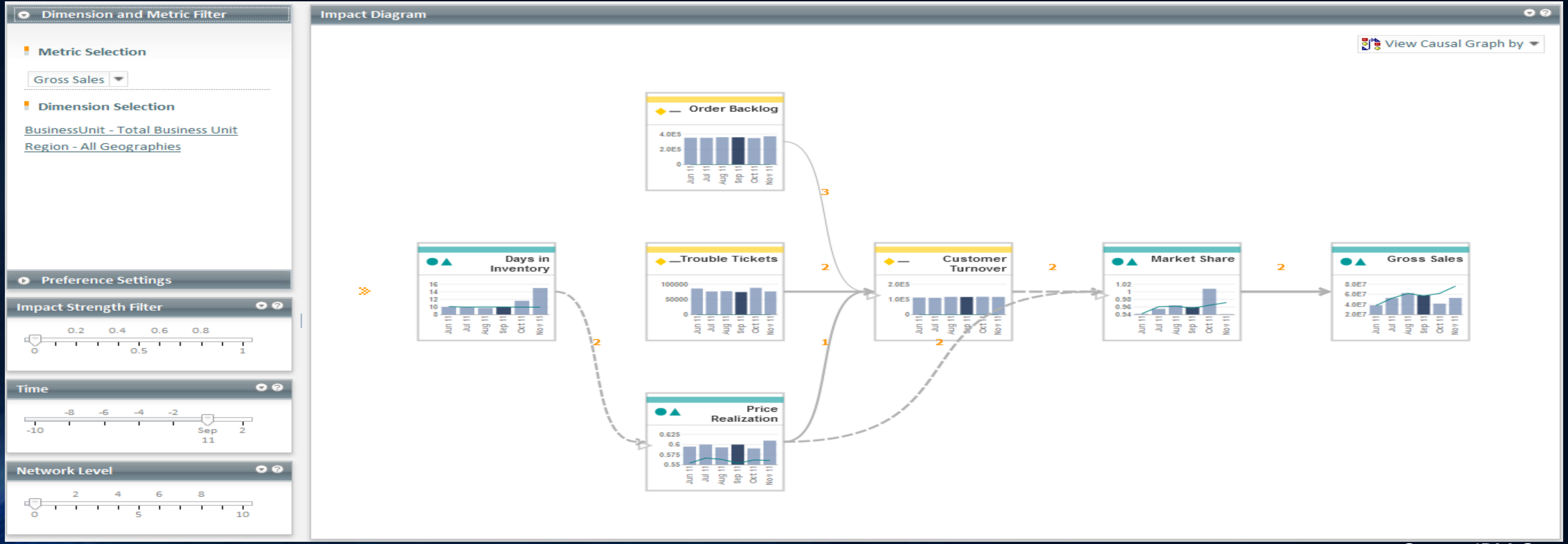


Business KPI Monitoring and the CFO Performance Insights

- Captures significant, meaningful relationships between risk and performance metrics;
- Operationalizes the derived relationships to support various types of analysis
  - What-if scenario analysis
  - Goal seek
  - Root cause analysis



- Leverages multi-dimensional data for hierarchical modeling at various granularities



Dimension and Metric Filter

Metric Selection

Gross Sales

Dimension Selection

BusinessUnit - Total Business Unit
Region - All Geographies

Preference Settings

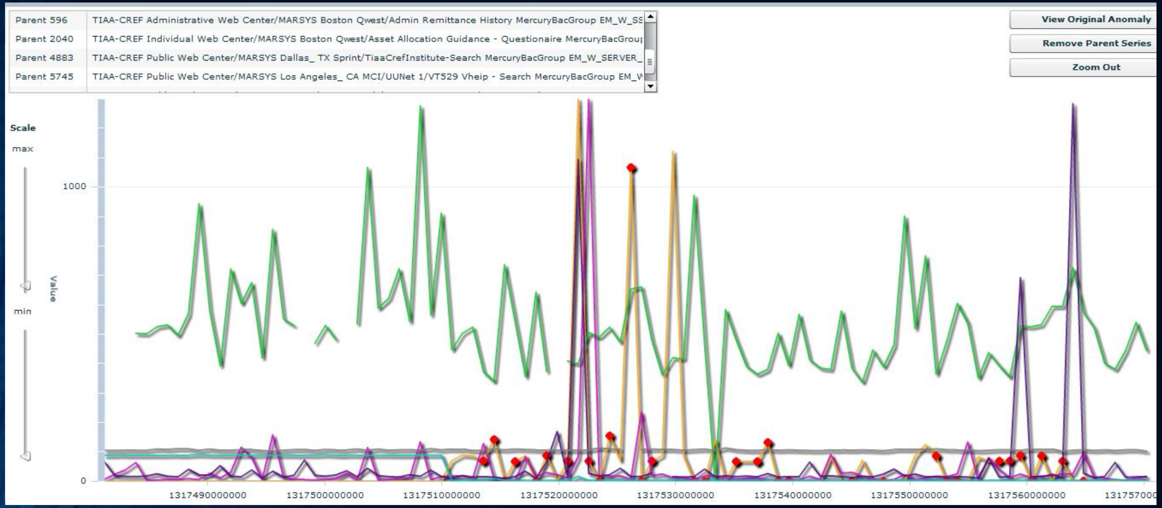
Impact Strength Filter

Time

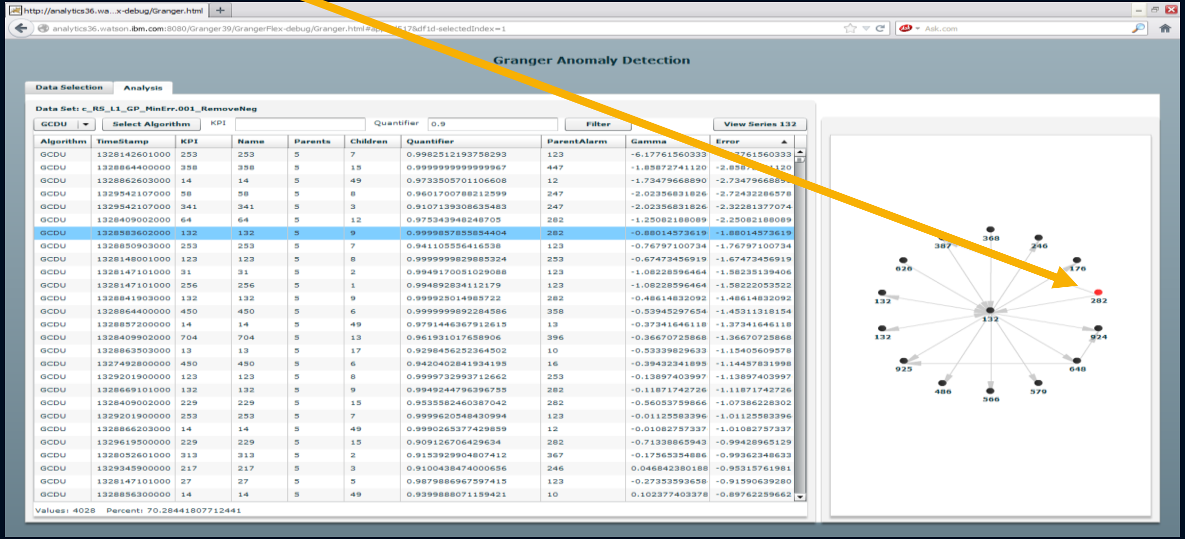
Network Level

IT Network Monitoring: Smart Cloud Predictive Insights

- Along with an “anomaly alarm”, associated causal structure is presented, as well as “root cause” variables identified via alarm consolidation



Parent 596	TIAA-CREF Administrative Web Center/MARSYS Boston Qwest/Admin Remittance History MercuryBacGroup EM_W_SS
Parent 2040	TIAA-CREF Individual Web Center/MARSYS Boston Qwest/Asset Allocation Guidance - Questionaire MercuryBacGrou
Parent 4883	TIAA-CREF Public Web Center/MARSYS Dallas_ TX Sprint/TiaaCrefInstitute-Search MercuryBacGroup EM_W_SERVER_
Parent 5745	TIAA-CREF Public Web Center/MARSYS Los Angeles_ CA MCI/UUNet 1/VT529 Vheip - Search MercuryBacGroup EM_W



http://analytics36.wa...x-debug/Granger.html
analytics36.watson.ibm.com:8080/Granger39/GrangerFlex-debug/Granger.html#a...517&df1d-selectedIndex=1
Ask.com
Data Selection
Analysis
GCDU
Select Algorithm
Filter
View Series 132
368
387
246
626
176
132
282
132
132
924
925
648
486
566
579

Granger Anomaly Detection

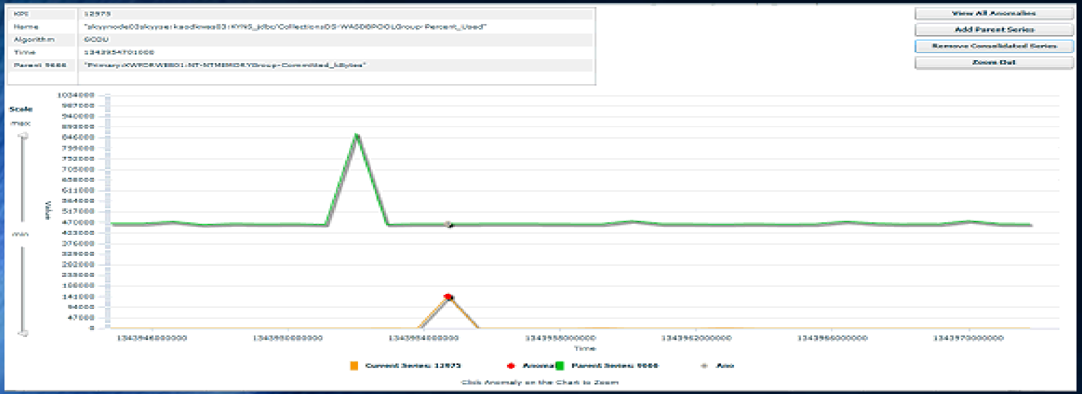
Data Set: c_RS_L1_GP_MinErr.001_RemoveNeg

KPI

Quantifier 0.9

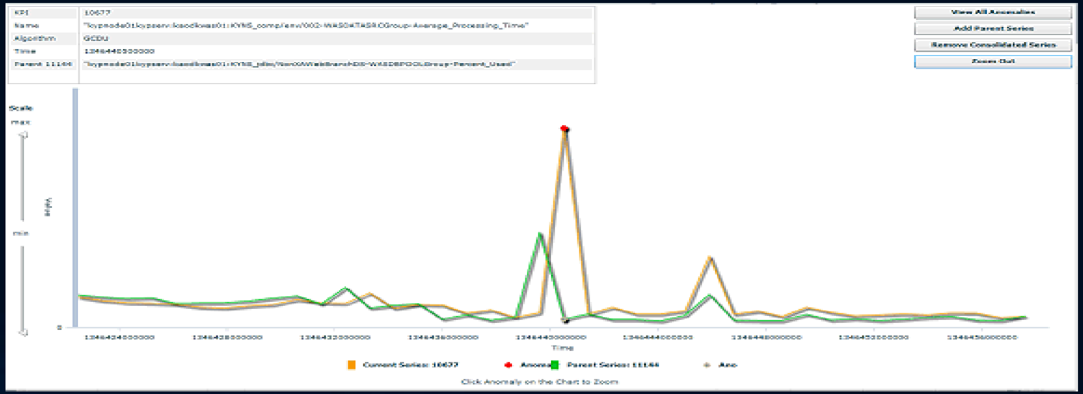
Algorithm	TimeStamp	KPI	Name	Parents	Children	Quantifier	ParentAlarm	Gamma	Error
GCDU	1328142601000	253	253	5	7	0.9982512193758293	123	-6.17761560333	-6.17761560333
GCDU	1328864400000	358	358	5	15	0.9999999999999967	447	-1.85872741120	-2.85872741120
GCDU	1328862603000	14	14	5	49	0.9733505701106608	12	-1.73479668890	-2.73479668890
GCDU	1329542107000	58	58	5	8	0.9601700788212599	247	-2.02356831826	-2.72432286578
GCDU	1329542107000	341	341	5	3	0.9107139308635483	247	-2.02356831826	-2.32281377074
GCDU	1328409002000	64	64	5	12	0.975343948248705	282	-1.25082188089	-2.25082188089
GCDU	1328583602000	132	132	5	9	0.9999857855854404	282	-0.88014573619	-1.88014573619
GCDU	1328850903000	253	253	5	7	0.941105556416538	123	-0.76797100734	-1.76797100734
GCDU	1328148001000	123	123	5	8	0.9999999829885324	253	-0.67473456919	-1.67473456919
GCDU	1328147101000	31	31	5	2	0.9949170051029088	123	-1.08228596464	-1.58235139406
GCDU	1328147101000	256	256	5	1	0.994892834112179	123	-1.08228596464	-1.58222053522
GCDU	1328841903000	132	132	5	9	0.999925014985722	282	-0.48614832092	-1.48614832092
GCDU	1328864400000	450	450	5	6	0.9999999892284586	358	-0.53945297654	-1.45311318154
GCDU	1328857200000	14	14	5	49	0.9791446367912615	13	-0.37341646118	-1.37341646118
GCDU	1328409902000	704	704	5	13	0.961931017658906	396	-0.36670725868	-1.36670725868
GCDU	1328863503000	13	13	5	17	0.9298456252364502	10	-0.53339829633	-1.15405609578
GCDU	1327492800000	450	450	5	6	0.9420402841934195	16	-0.39432341895	-1.14457831998
GCDU	1329201900000	123	123	5	8	0.9999732993712662	253	-0.13897403997	-1.13897403997
GCDU	1328669101000	132	132	5	9	0.9949244796396755	282	-0.11871742726	-1.11871742726
GCDU	1328409002000	229	229	5	15	0.9535582460387042	282	-0.56053759866	-1.07386228302
GCDU	1329201900000	253	253	5	7	0.9999620548430994	123	-0.01125583396	-1.01125583396
GCDU	1328866203000	14	14	5	49	0.9990265377429859	12	-0.01082757337	-1.01082757337
GCDU	1329619500000	229	229	5	15	0.909126706429634	282	-0.71338865943	-0.99428965129
GCDU	1328052601000	313	313	5	2	0.9153929904807412	367	-0.17565354886	-0.99362348633
GCDU	1329345900000	217	217	5	3	0.9100438474000656	246	0.046842380188	-0.95315761981
GCDU	1328147101000	27	27	5	5	0.9879886967597415	123	-0.27353593658	-0.91590639280
GCDU	1328856300000	14	14	5	49	0.9399888071159421	10	0.102377403378	-0.89762259662

Values: 4028 Percent: 70.28441807712441



KPI	12975
Name	"skypnode03:skypse:kaodkwas03:KYNS_jdbc/CollectionsDS-WASDBPOOLGroup-Percent_Used"
Algorithm	GCDU
Time	1343954701000
Parent 9666	"Primary:KWRDRWEB01:NT-NTMEMORYGroup-Committed_kBytes"

Click Anomaly on the Chart to Zoom



KPI	10677
Name	"kypnode01:kypserv:kaodkwas01:KYNS_comp/srv/002-WAS04TAS0Group-Average_Processing_Time"
Algorithm	GCDU
Time	1346440300000
Parent 11144	"kypnode01:kypserv:kaodkwas01:KYNS_jdbc/NonXAWebBanxNDS-WASDBPOOLGroup-Percent_Used"

Click Anomaly on the Chart to Zoom

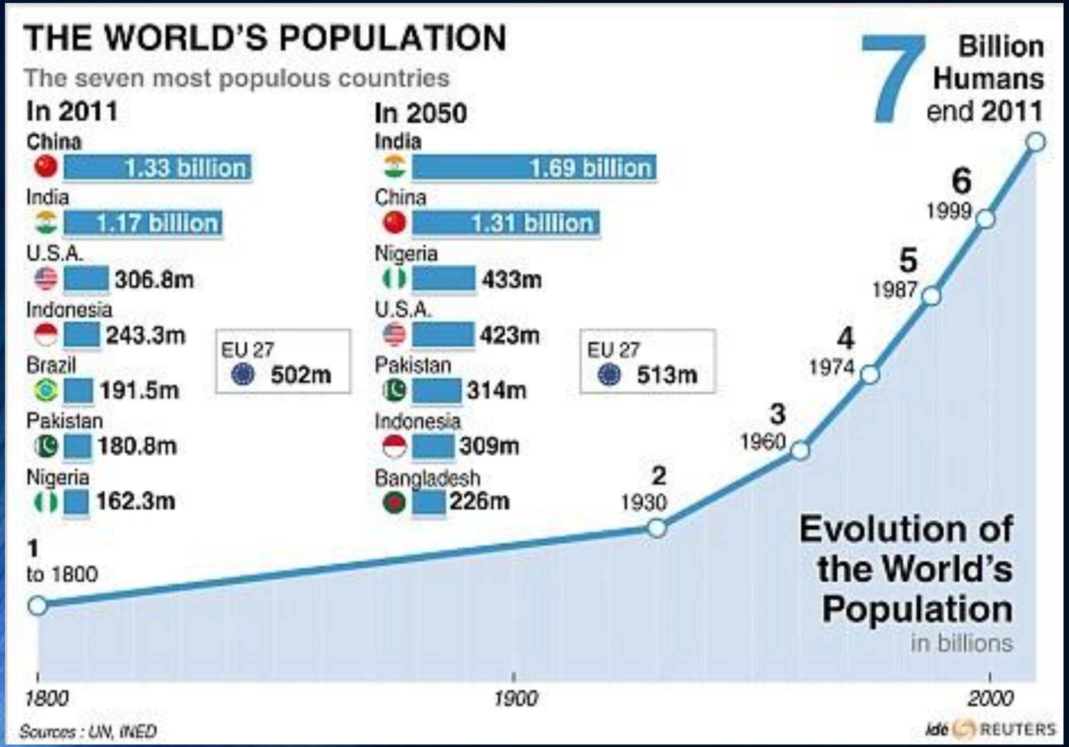
© 2015 IBM Corporation

IBM Days (2000's – present)

- Smarter Government
  - Markov Decision Process and Tax Collections
- Smarter Planet , Enterprise & Cloud
  - Temporal Causal Modeling and Applications
- ***Smarter Agriculture***
  - ***Heterogeneous Data Analytics and Accelerated Plant Breeding***

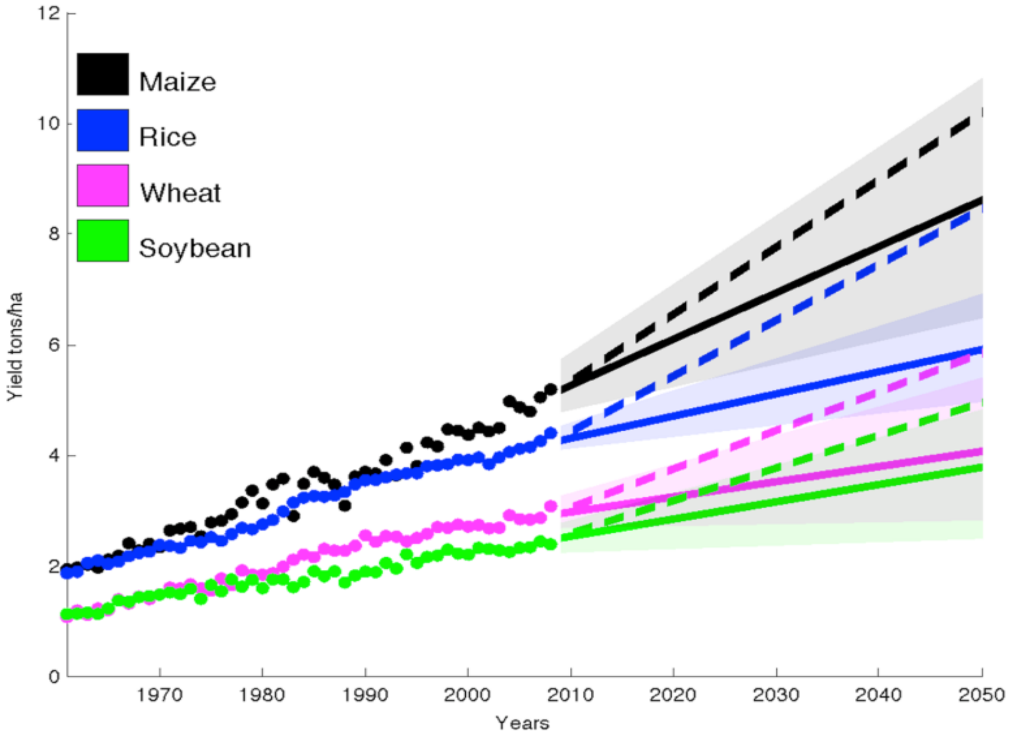
Smarter Agriculture

- Food production needs to be nearly doubled by 2050, to keep pace with population growth
- Challenged by limited resources, e.g. water and arable land, need to increase production efficiency



Source: : Rueters

Food Gap

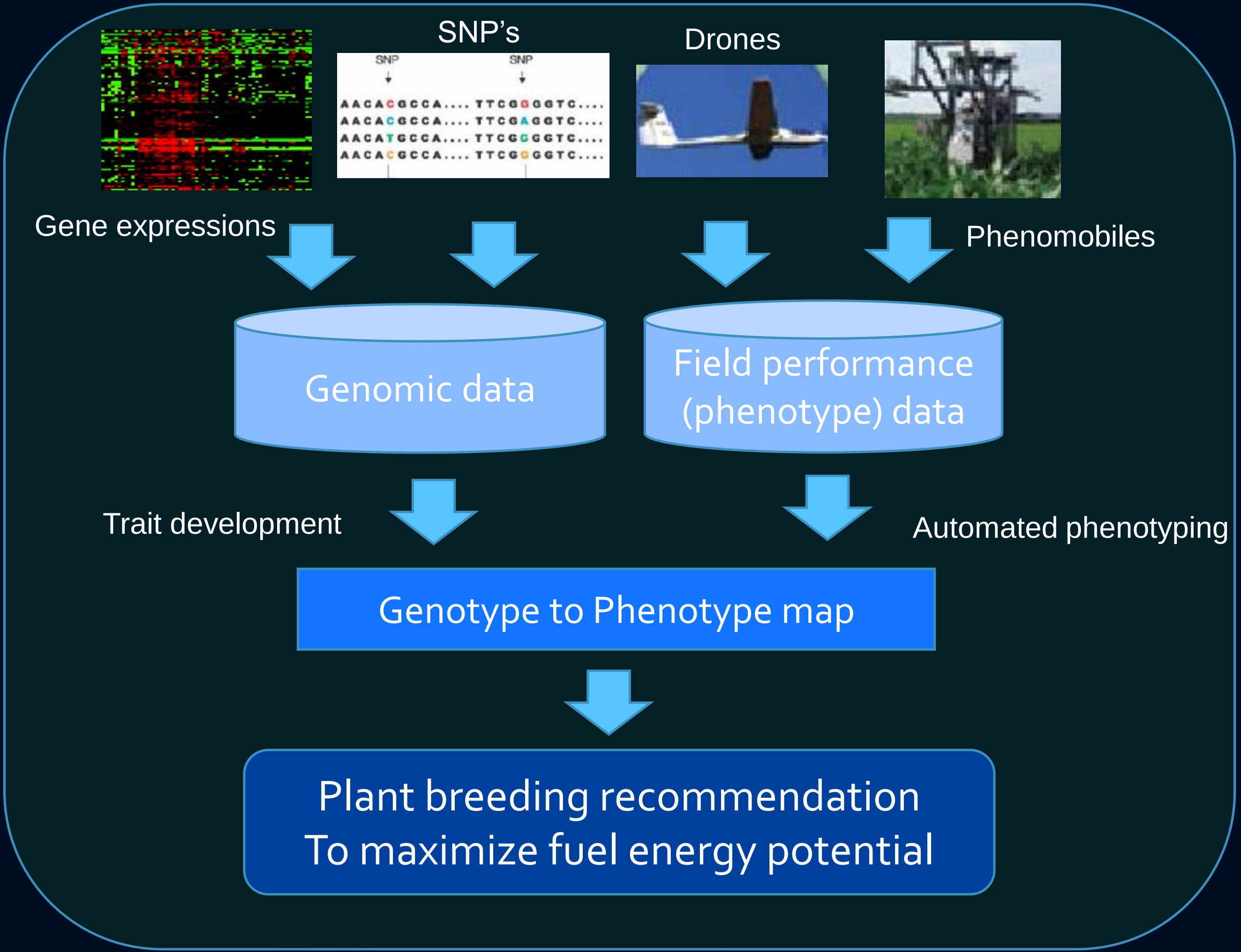


Source: United Nation's Food and Agriculture Organization (FAO), The State of Food and Agriculture, June 2013

Projected gap between demand and supply for major row crops

Department of Energy (DOE) ARPA-E Funded Project
“Transportation Energy Resources from Renewable Agriculture (TERRA)”

- Project duration
  - 3 years (Sep. 2015 to Aug. 2018)
- Project Goals
  - To develop “**Automated Sorghum Phenotyping and Trait Development Platform**”
    - An automated high-throughput system for determining how variations in the sorghum genome impact field performance and agricultural productivity
    - Capacity to use sensing data from ground-based mobile and airborne platforms for automated phenotyping will advance plant breeding to maximize energy potential for transportation fuel
- Partnership
  - Purdue University, IBM Research, CSIRO


SNP’s
Drones
SNP
SNP
AACACGCCA.... TTCGGGGTC....
AACACGCCA.... TTCGAGGTC....
AACATGCCA.... TTCGGGGTC....
AACACGCCA.... TTCGGGGTC....
Gene expressions
Phenomobiles

Genomic data

Field performance (phenotype) data

Trait development

Automated phenotyping

Genotype to Phenotype map

Plant breeding recommendation
To maximize fuel energy potential

Images from the Purdue Fields: Sorghum

Images from the Purdue Fields: Drones and Phenomobiles

Images from the Purdue Fields (videos)

- The IBM team (minus one) is here in the Purdue experimental field
- Here's the phenomobile in action
  - https://youtu.be/adCJDrB4Sus

Project Scope

- Overall Project: Complete Integrated Phenotyping Systems Solutions
  - High Throughput Automated Hardware & Sensing Technologies (Purdue)
    - Optimize high-throughput remote-sensing technologies to acquire relevant data on sorghum plant phenotypes
  - Image/Sensor Data Analytics (Purdue & IBM)
    - Develop data analytics algorithms for segmentation and feature extraction
  - Genetics, Genomics and Bioinformatics (IBM)
    - Develop sophisticated genetic analysis pipelines to identify genes controlling sorghum performance

Overall deliverable will be “Automated Sorghum Phenotyping and Trait Development Platform”

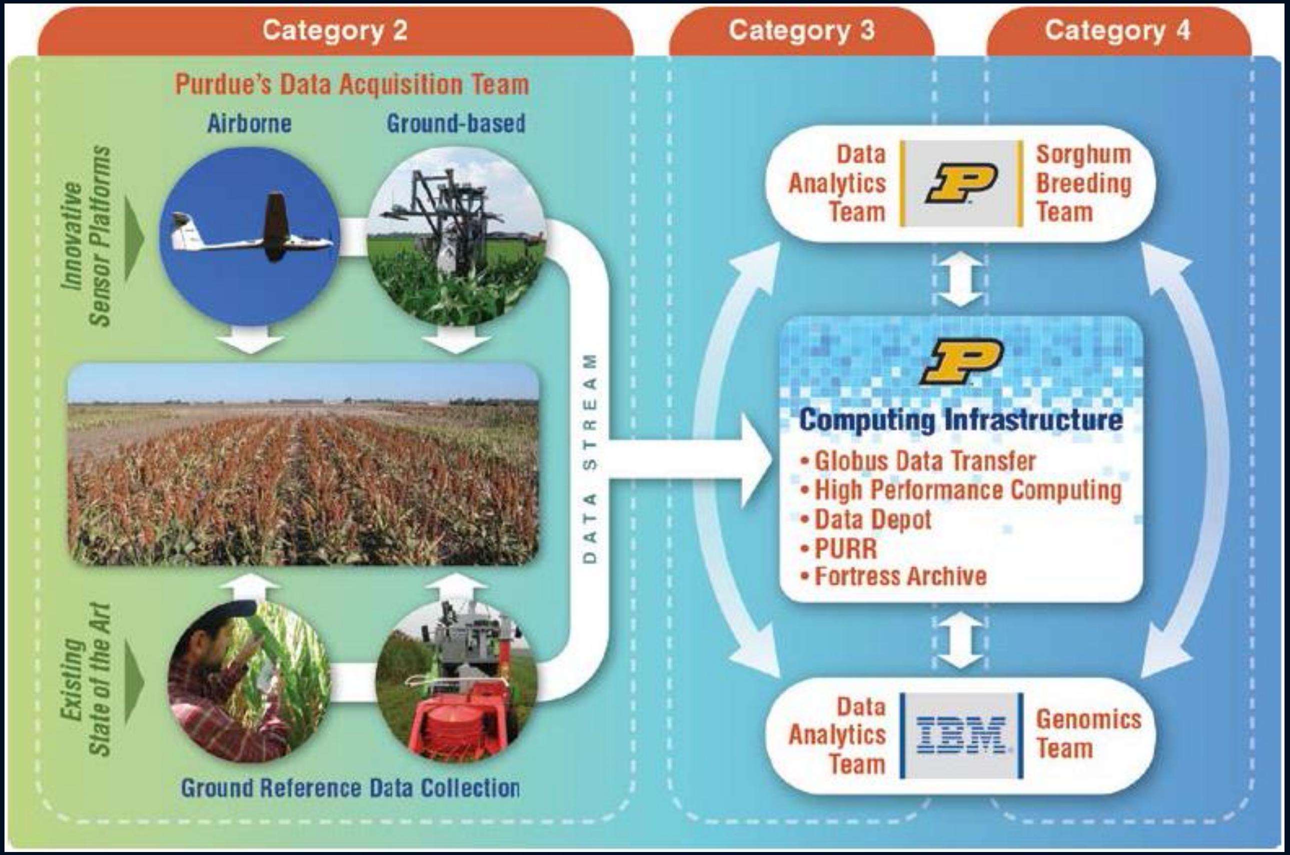
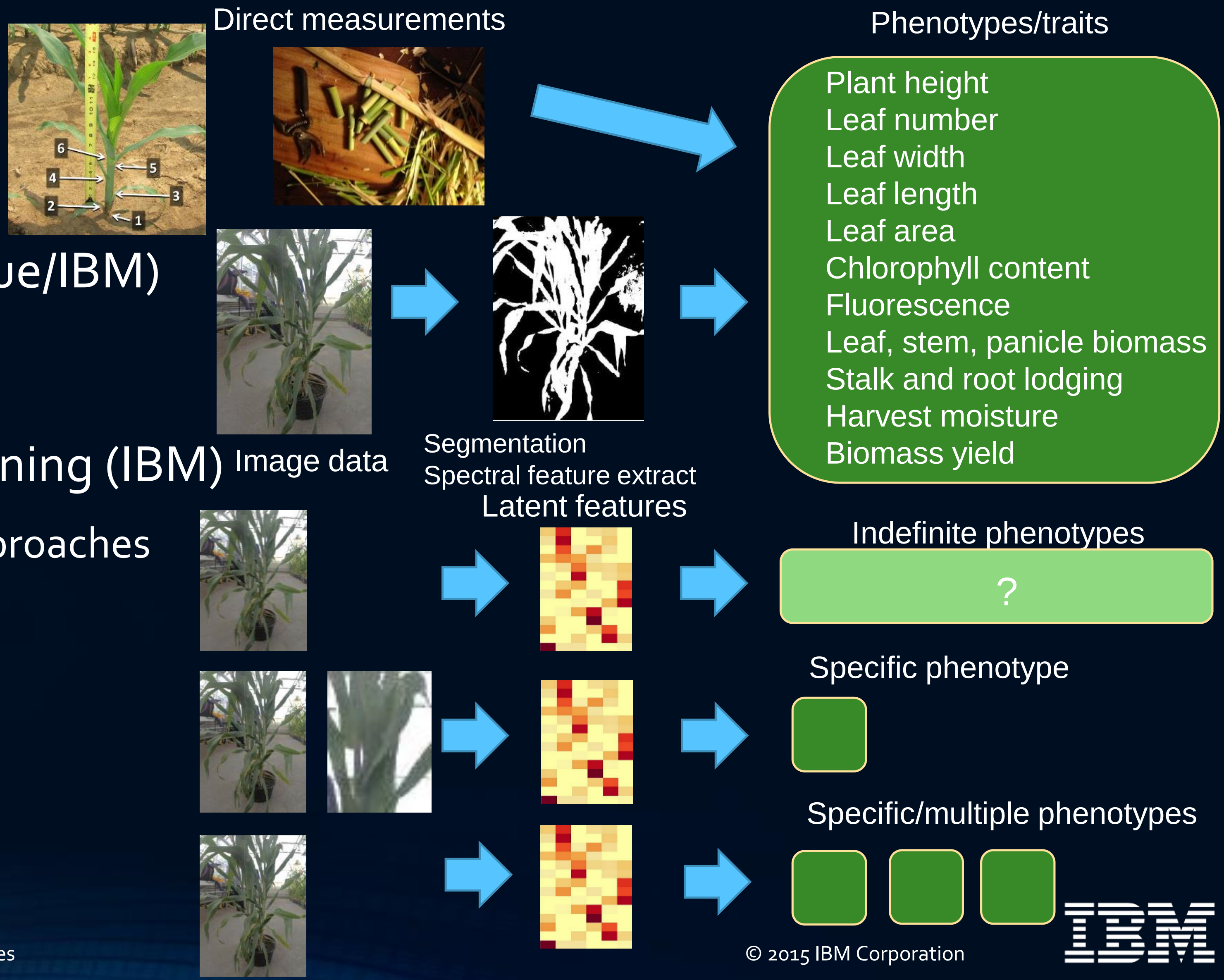


Image Data Feature Extraction by Machine Learning

- Manual Phenotyping (Purdue)
- Automated Phenotyping (Purdue/IBM)
- Latent Phenotype Feature Learning (IBM)
  - Unsupervised Feature Learning Approaches
  - Multi-scale Modeling Approach
  - Multi-task Feature Learning



Direct measurements

Phenotypes/traits

Plant height
Leaf number
Leaf width
Leaf length
Leaf area
Chlorophyll content
Fluorescence
Leaf, stem, panicle biomass
Stalk and root lodging
Harvest moisture
Biomass yield

Image data

Segmentation
Spectral feature extract

Latent features

Indefinite phenotypes

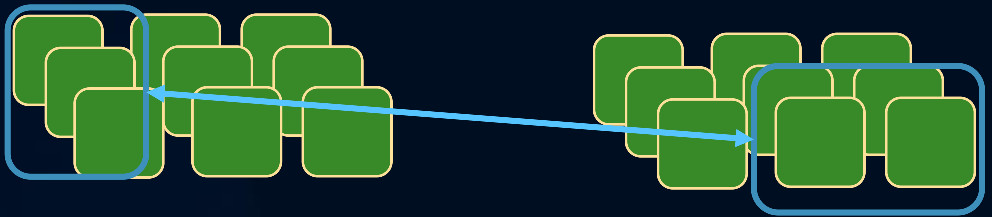
?

Specific phenotype

Specific/multiple phenotypes

Unified Genotype Phenotype Association Analysis

- Discovering underlying group structure in genotype phenotype association

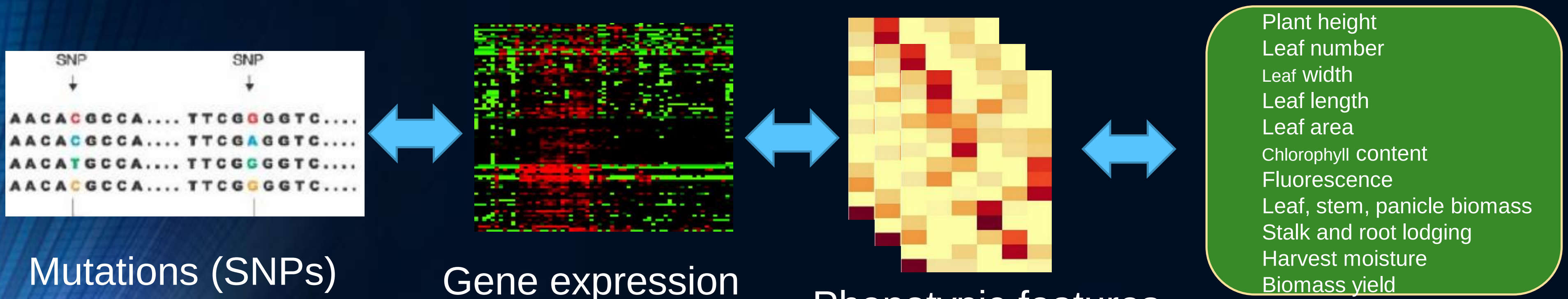


SNPs

Phenotypes

structured sparse models !

- Discovering links between heterogeneous data (genotype, phenotype, etc)


SNP
SNP
AACACGCCA.... TTCGGGGTC....
AACACGCCA.... TTCGAGGTC....
AACATGCCA.... TTCGGGGTC....
AACACGCCA.... TTCGGGGTC....

Plant height
Leaf number
Leaf width
Leaf length
Leaf area
Chlorophyll content
Fluorescence
Leaf, stem, panicle biomass
Stalk and root lodging
Harvest moisture
Biomass yield

Mutations (SNPs)

Gene expression

Phenotypic features

Phenotypes

Mixed graphical models !

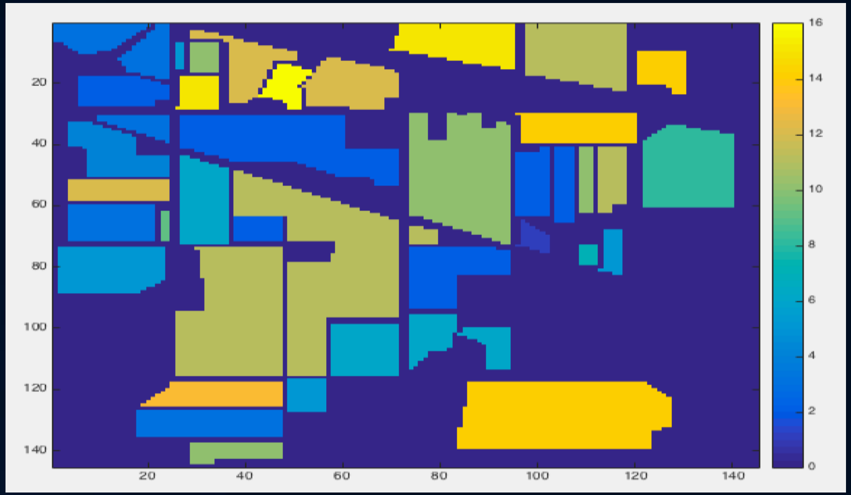
- Discovering indirect links with less reliable phenotype measurement data

robust modeling methods !

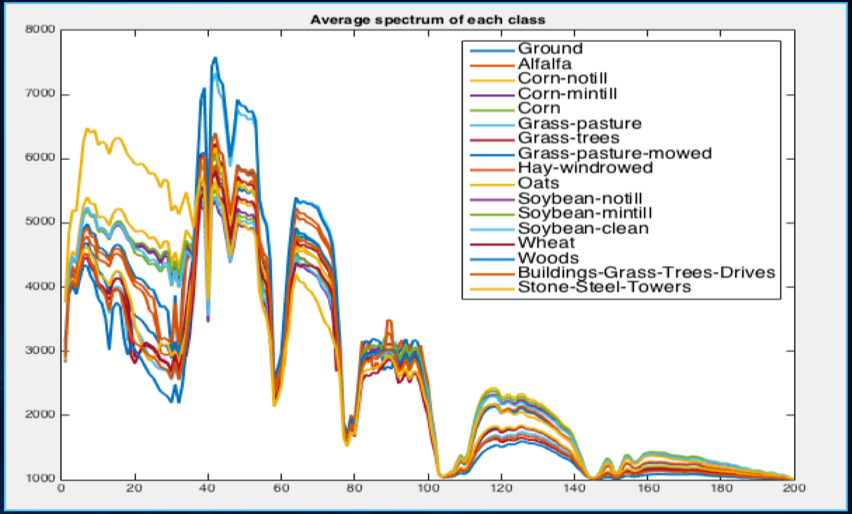
Hyperspectral Data Analysis (Preliminary Results)

- Extracting hyperspectral signatures for forests, crops, soil, etc

Example Data
Indian Pine Ground truth

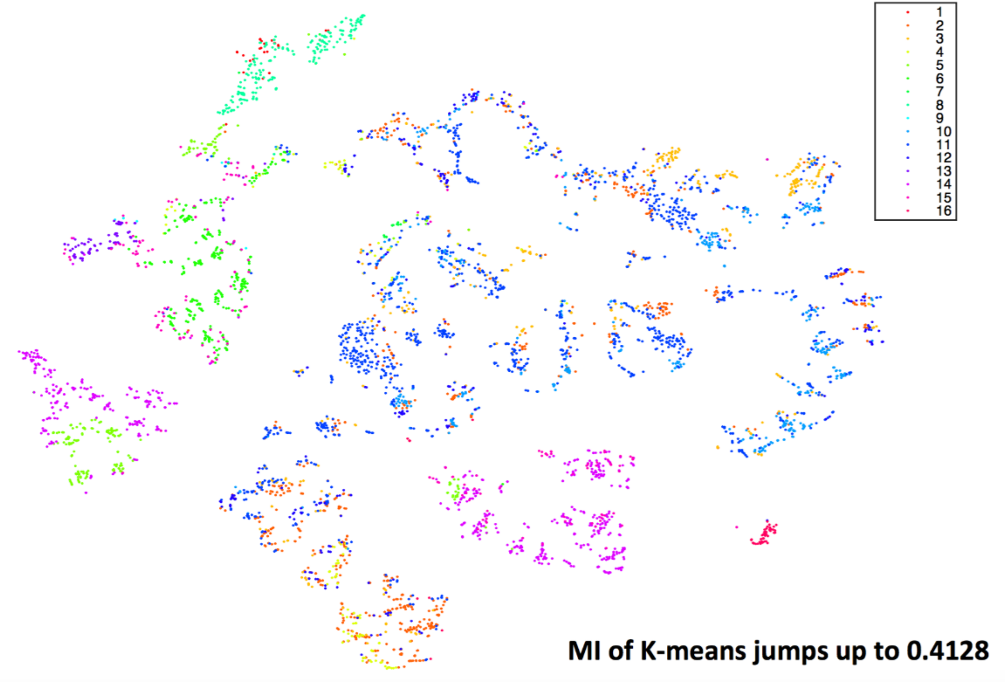


Average spectra
Per class



Wavelengths
0.4 to 2.5 10^(-6) m

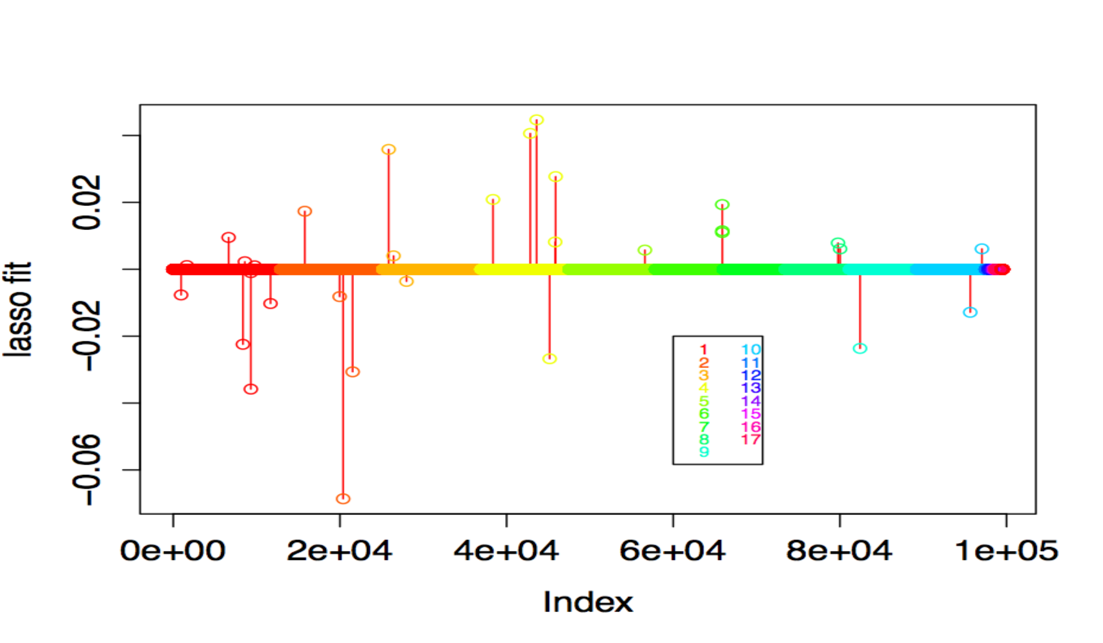
Non-linear embedding & clustering



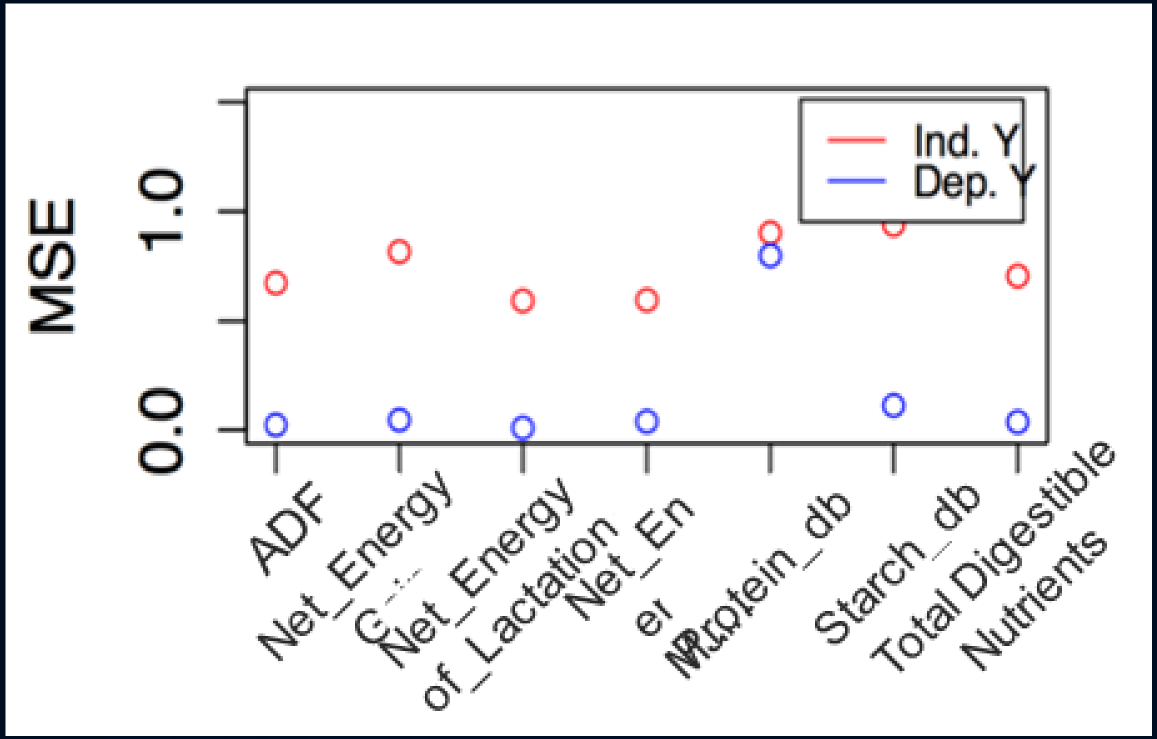
MI of K-means jumps up to 0.4128

Genotype Network & Phenotype Prediction (Preliminary Results)

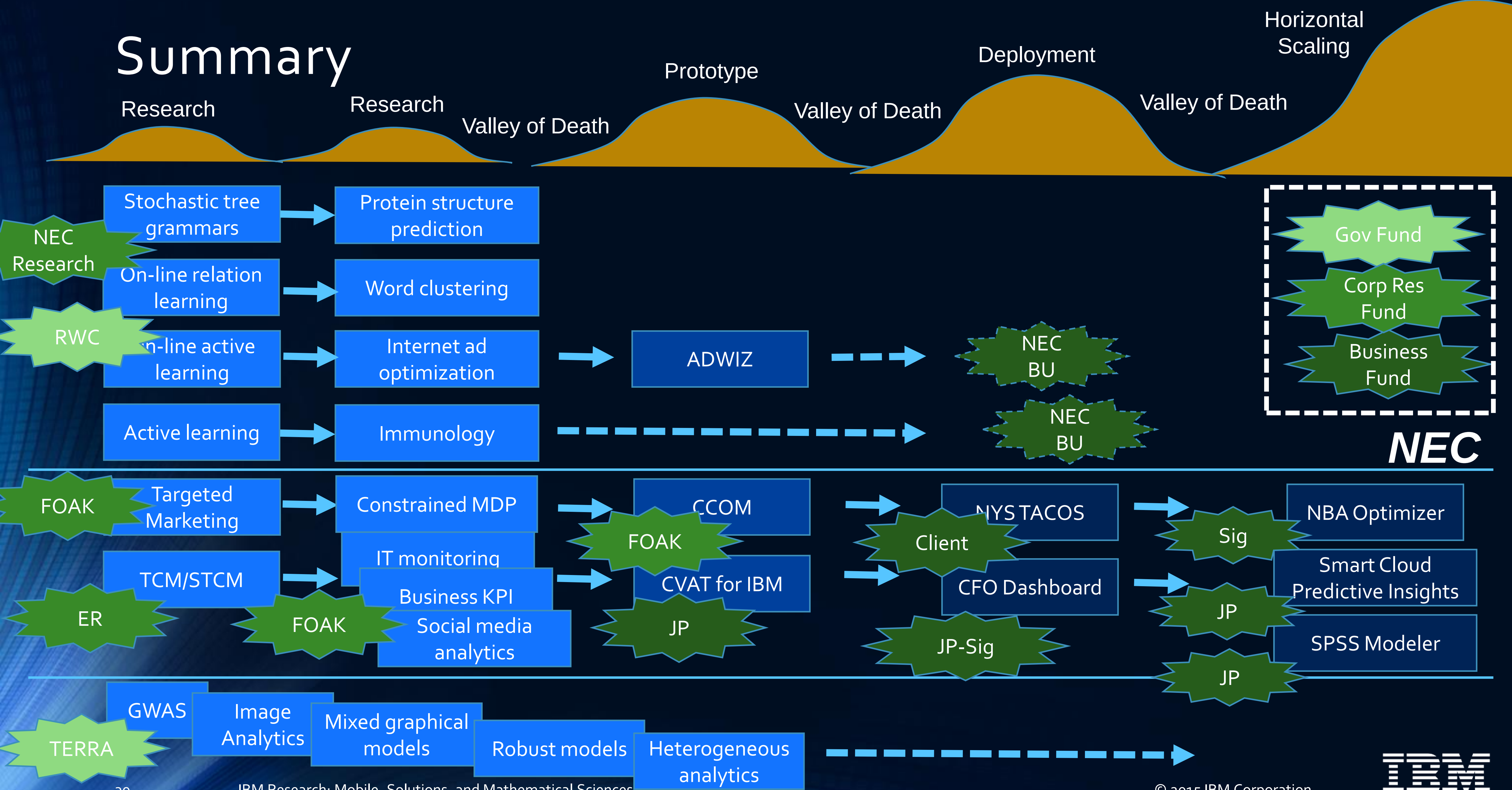
Examples of selected SNPs and chromosome positions



Phenotype prediction accuracy improves by modeling multiple phenotypes



Summary



Research

Research

Valley of Death

Prototype

Valley of Death

Deployment

Valley of Death

Horizontal Scaling

NEC Research

RWC

Stochastic tree grammars → Protein structure prediction

On-line relation learning → Word clustering

On-line active learning → Internet ad optimization → ADWIZ → NEC BU

Active learning → Immunology → NEC BU

Gov Fund

Corp Res Fund

Business Fund

NEC

FOAK

Targeted Marketing → Constrained MDP → CCOM → NYS TACOS → NBA Optimizer

TCM/STCM → IT monitoring → CVAT for IBM → CFO Dashboard → Smart Cloud Predictive Insights

Business KPI

Social media analytics

SPSS Modeler

ER

FOAK

FOAK

JP

Client

JP-Sig

Sig

JP

JP

TERRA

GWAS

Image Analytics

Mixed graphical models

Robust models

Heterogeneous analytics

Some Take-Home Messages for Industrial Researchers

- ***Think long term, but be flexible !***
- **Decades:** Think of your career in phases (in the unit of decades)
- **Bottom-Up:** Long term strategic thinking is important, but they don't always come from top down strategy
  - Indulge yourself in basic research while you are young…
- **Breath-First:** Explore breath-first first and then go in depth
  - Focus on today's research and your path tomorrow will present itself
- **Challenge:** Do dare to cross the "valleys of death" sometime during your career
- **Adaptive funding:** Use a mixture of different funding types wisely

Some Relevant Publications

NEC days

1. Predicting Protein Secondary Structure Using Stochastic Tree Grammars. Naoki Abe, Hiroshi Mamitsuka. *Machine Learning*, November 1997, Volume 29, Issue 2, pp 275-301

2. On-line Learning of Binary Lexical Relations Using Two-dimensional Weighted Majority Algorithms. Naoki Abe, Hang Li and Atsuyoshi Nakamura. Proceedings of The Twelfth International Conference on Machine Learning, July 1995.

3. The Lob-Pass Problem. Jun'ichi Takeuchi, Naoki Abe and Shun'ichi Amari. *Journal of Computer and System Sciences, 61(3), 2000.*

4. Learning to Optimally Schedule Internet Banner Advertisements. Naoki Abe and Atsuyoshi Nakamura. Proceedings of The Sixteenth International Conference on Machine Learning, July 1999.

5. Prediction of MHC Class I Binding Peptides by Dynamic Experiment Design based on Query Learning with Hidden Markov Models. Keiko Udaka, Hiroshi Mamitsuka, Yukinobu Nakaseko and Naoki Abe. *Journal of Immunology*, 169(10), 5744-5753, 2002.

Some Relevant Publications (Cont'd)

Constrained Markov Decision Process

1. Optimizing debt collections using constrained reinforcement learning. Naoki Abe, Prem Melville, Cezar Pendus, Chandan K. Reddy, David L. Jensen, Vince P. Thomas, James J. Bennett, Gary F. Anderson, Brent R. Cooley, Melissa Kowalczyk, Mark Domick, Timothy Gardinier. KDD 2010: 75-84

2. Tax Collections Optimization for New York State. Gerard Miller, Melissa Weatherwax, Timothy Gardinier, Naoki Abe, Prem Melville, Cezar Pendus, David L. Jensen, Chandan K. Reddy, Vince P. Thomas, James J. Bennett, Gary F. Anderson, Brent R. Cooley. Interfaces 42(1): 74-84 (2012)

Temporal Causal Modeling

1. Spatial-temporal causal modeling for climate change attribution. Aurelie C. Lozano, Hongfei Li, Alexandru Niculescu-Mizil, Yan Liu, Claudia Perlich, Jonathan R. M. Hosking, Naoki Abe. KDD 2009: 587-596

2. Grouped graphical Granger modeling methods for temporal causal modeling. Aurelie C Lozano, Naoki Abe, Yan Liu, Saharon Rosset, Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2009.

3. Grouped graphical Granger modeling for gene expression regulatory networks discovery. Aurelie C Lozano, Naoki Abe, Yan Liu, Saharon Rosset, Bioinformatics, Oxford Univ Press, 2009.

TERRA Project

http://www.purdue.edu/newsroom/releases/2015/Q2/purdue-leading-research-using-advanced-technologies-to-better-grow-sorghum-as-biofuel.html