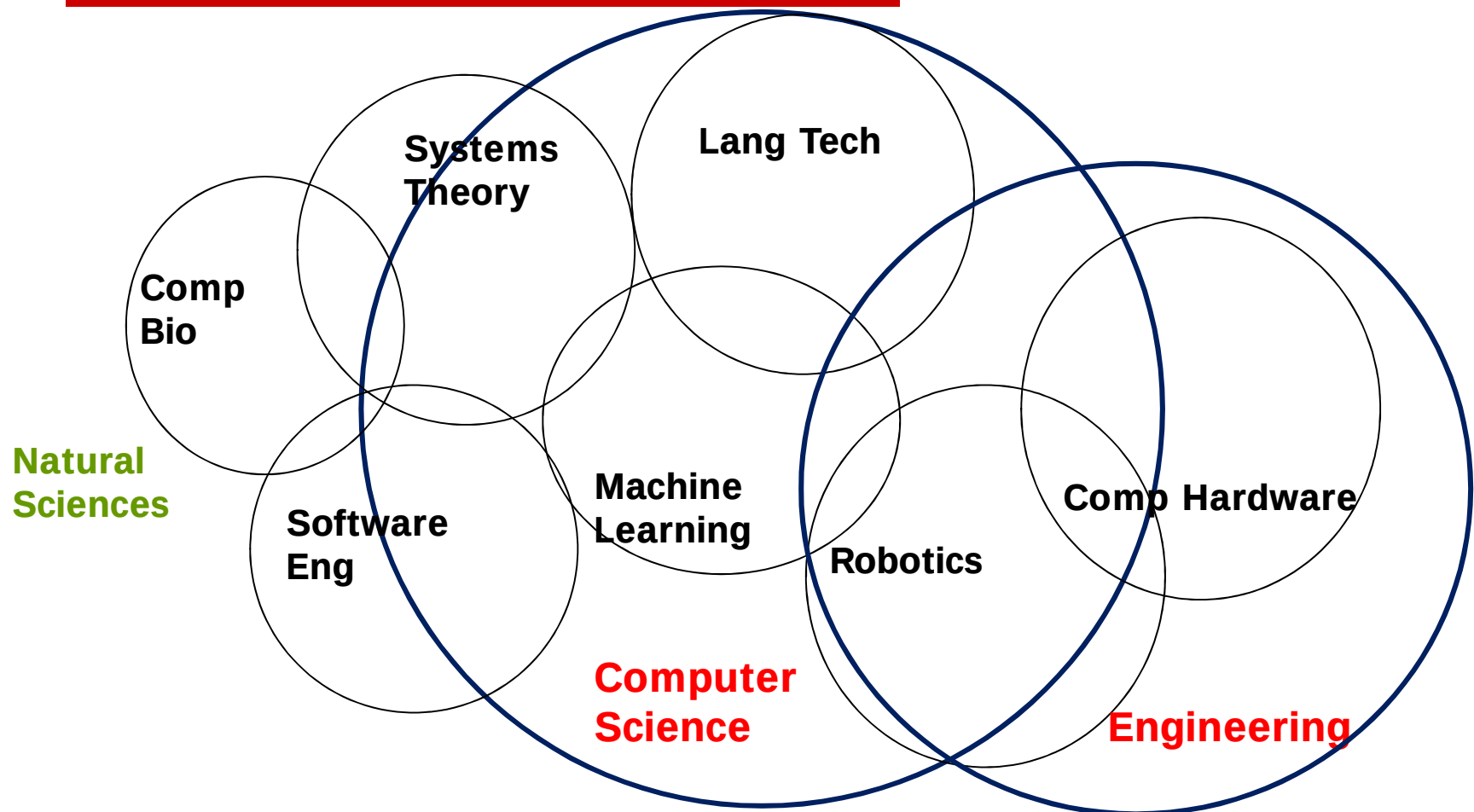


Data Sciences = Big Data + Machine Learning + Domain Expertise (NLP, Bio, ...)

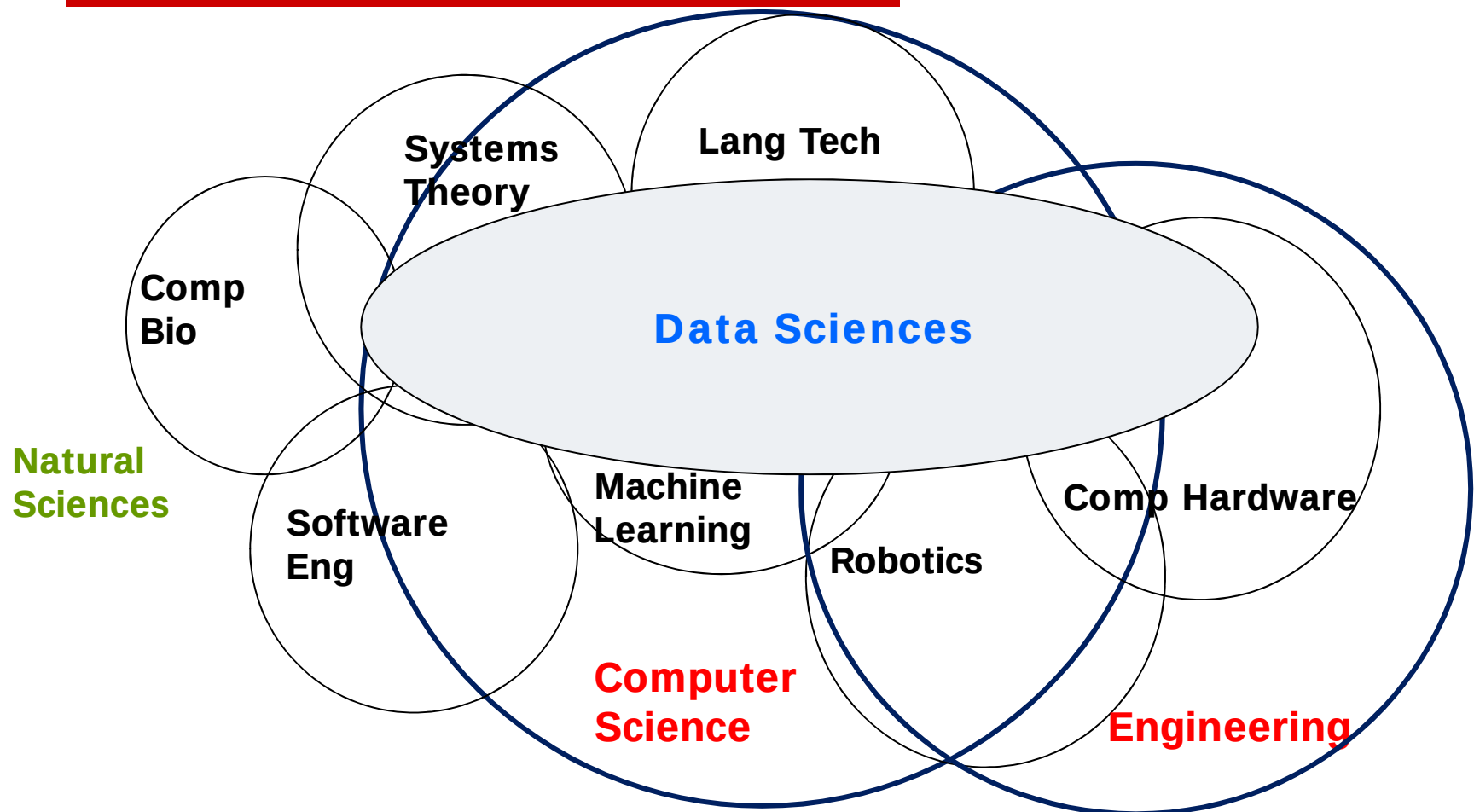
NAS – March, 2015

**Jaime Carbonell & Colleagues
Language Technologies Institute
School of Computer Science
Carnegie Mellon University
www.cs.cmu.edu/~jgc**

Computing Departments at CMU



Computing Departments at CMU



How Big is Big?

Dimensions of Big Data Analytics

LARGE-SCALE : TERABYTES → PETABYTES → EXOBYTES



**Billions++ of entries:
Terabytes/Petabytes of data**

HIGH-COMPLEXITY



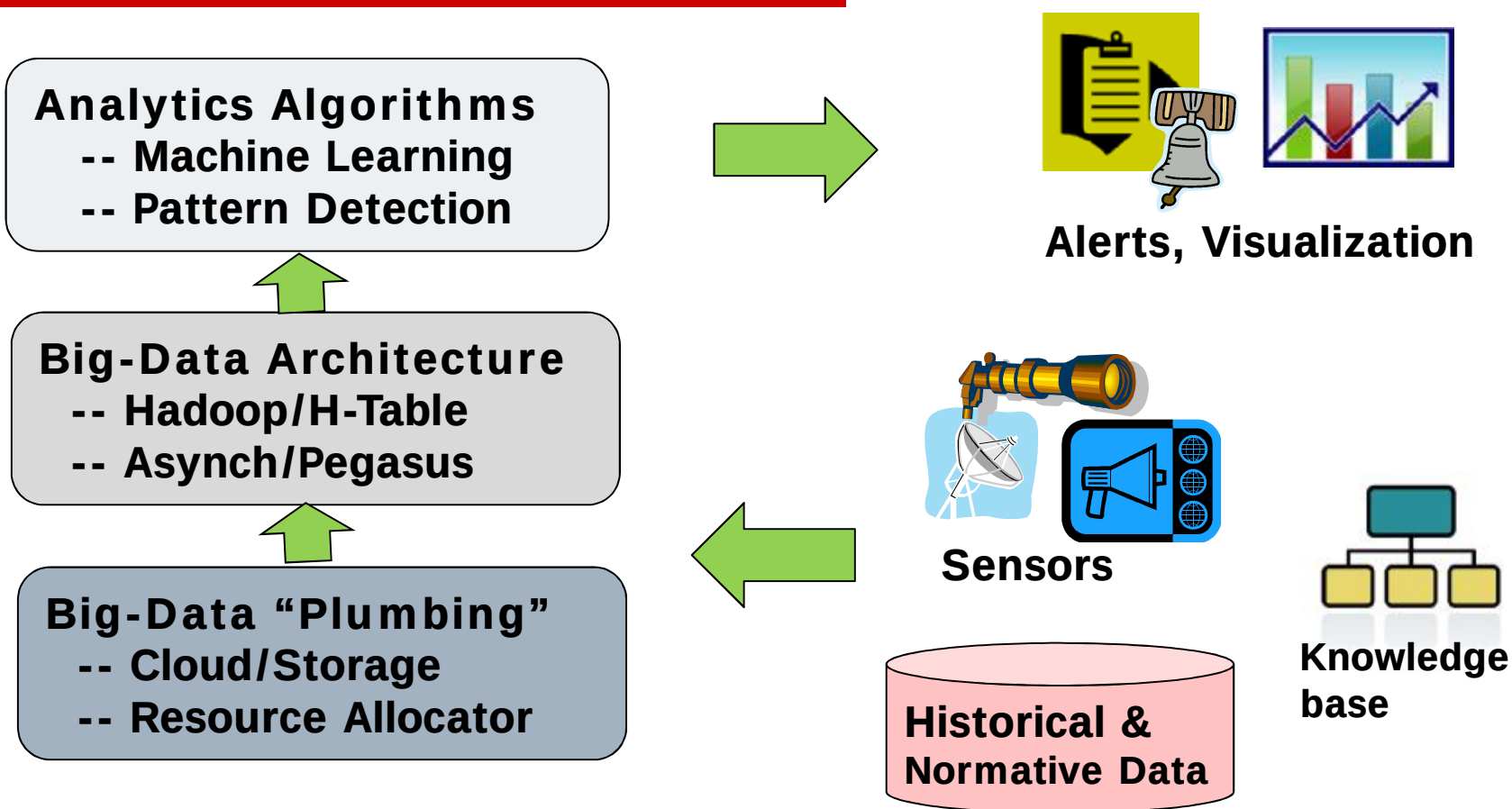
**Trillions of potential relations
among entries (graphs)**

HIGH-DIMENSIONAL

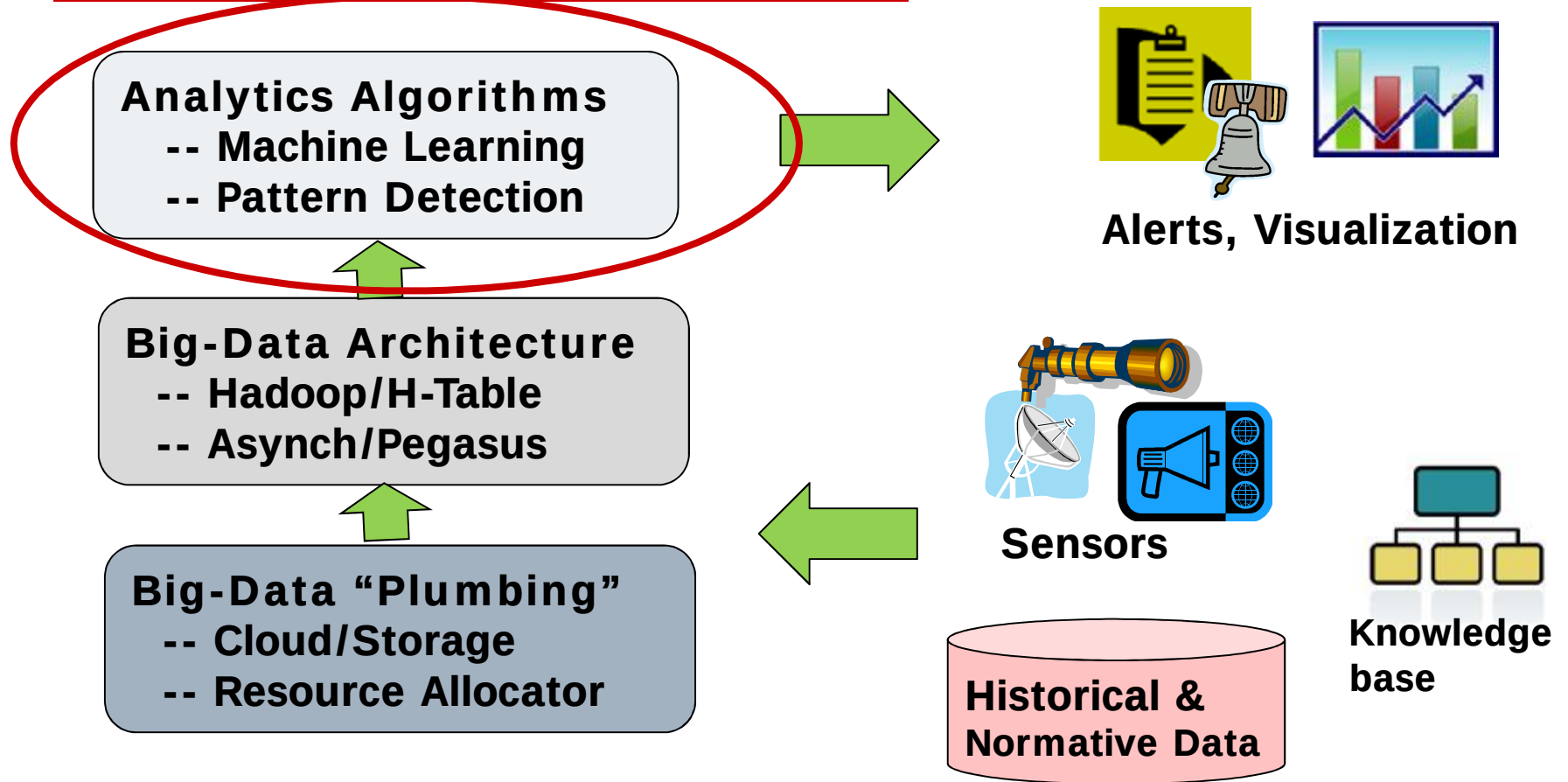


**Millions of attributes per entry
(but typically sparse encoding)**

The Big-Data “Stack”



The Big-Data “Stack”



Machine Learning

- Objectives
 - Discover patterns in data à make useful predictions
- General approach
 - Select & train functional models on historical data
 - Apply trained models to make predictions on new data
- Applications
 - Robotics, speech, medicine, biology, finance, cyber, energy, ...
- Hot topics
 - Deep neural networks
 - Transfer and proactive learning

Machine Learning in a Nutshell

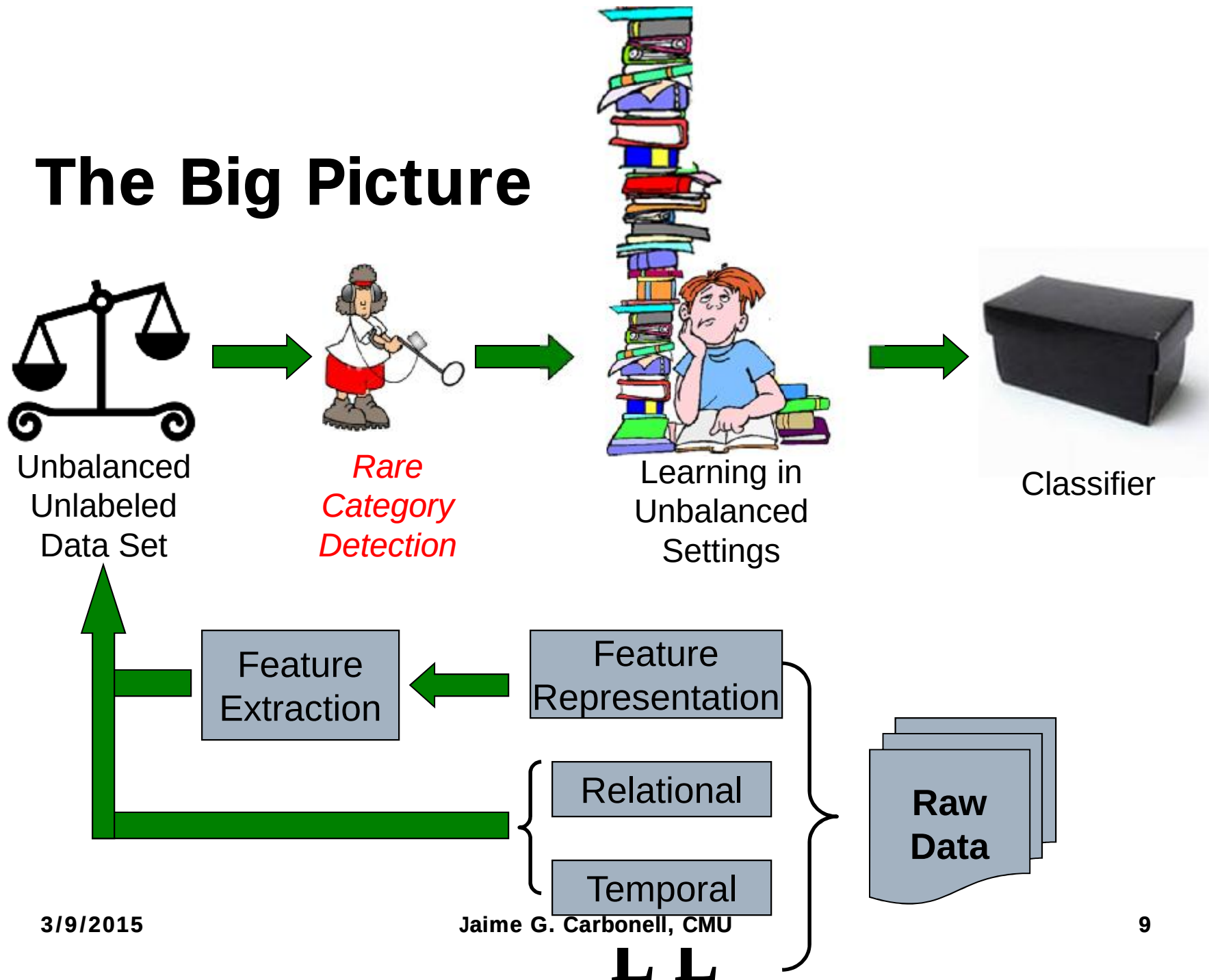
- o **Training data:** $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1,\dots,k} + \{\mathbf{x}_i\}_{i=k+1,\dots,n} + O : \mathbf{x}_i \rightarrow y_i$
 - n **Special case:** $k = 0$
- o **Functional space:** $\mathfrak{S} = \{f_j \times p_l\}$
- o **Fitness Criterion:**
 - n **a.k.a. loss function**

$$\arg \min_{j,l} \left[\left(\sum_i \|y_i - f_{j,p_l}(\mathbf{x}_i)\| \right) + b(f_j, p_l) \right]$$

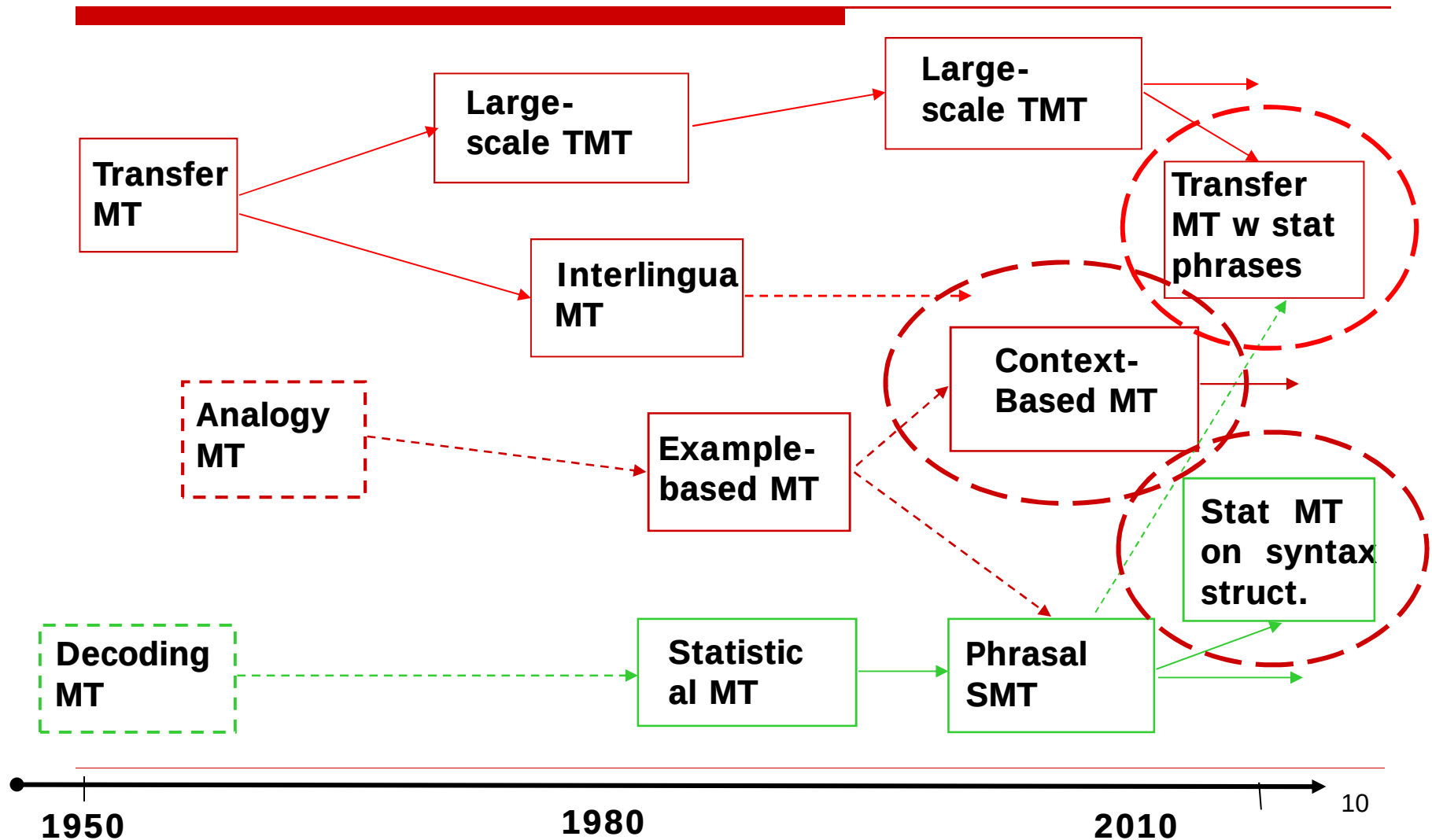
- o **Active Sampling Strategy:**

$$\arg \min_{\mathbf{x}_i \in \{\mathbf{x}_{k+1}, \dots, \mathbf{x}_n\}} \left(L(\hat{f}(\mathbf{x}_{test}, \hat{y}_{test})) \mid \mathbf{x}_i \cup \{\mathbf{x}_1, \dots, \mathbf{x}_k\} \right)$$

The Big Picture



Evolutionary Tree of MT Paradigms



Morpho-Syntactics & Multi-Morphemics

o Iñupiaq (North Slope Alaska, Lori Levin)

n Tauqsigñiaḡvinmunñianitchugut.

n 'We won't go to the store.'



o Kalaallisut (Greenlandic, Per Langaard)

n Pittsburghimukarthussaqaarnaviannngilaq

n Pittsburgh + PROP + Trim + SG + kar + tuq + ssaq + qar + n
aviar + nngit + v + IND + 3SG

n "It is not likely that anyone is going to Pittsburgh"

Context Needed to Resolve Ambiguity

Example: English à Japanese Machine Translation

Power line – densen (電線)

Subway line – chikatetsu (地下鉄)

(Be) on line – onrain (オンライン)

(Be) on the line – denwachuu (電話中)

Line up – narabu (並ぶ)

Line one's pockets – kanemochi ni naru (金持ちになる)

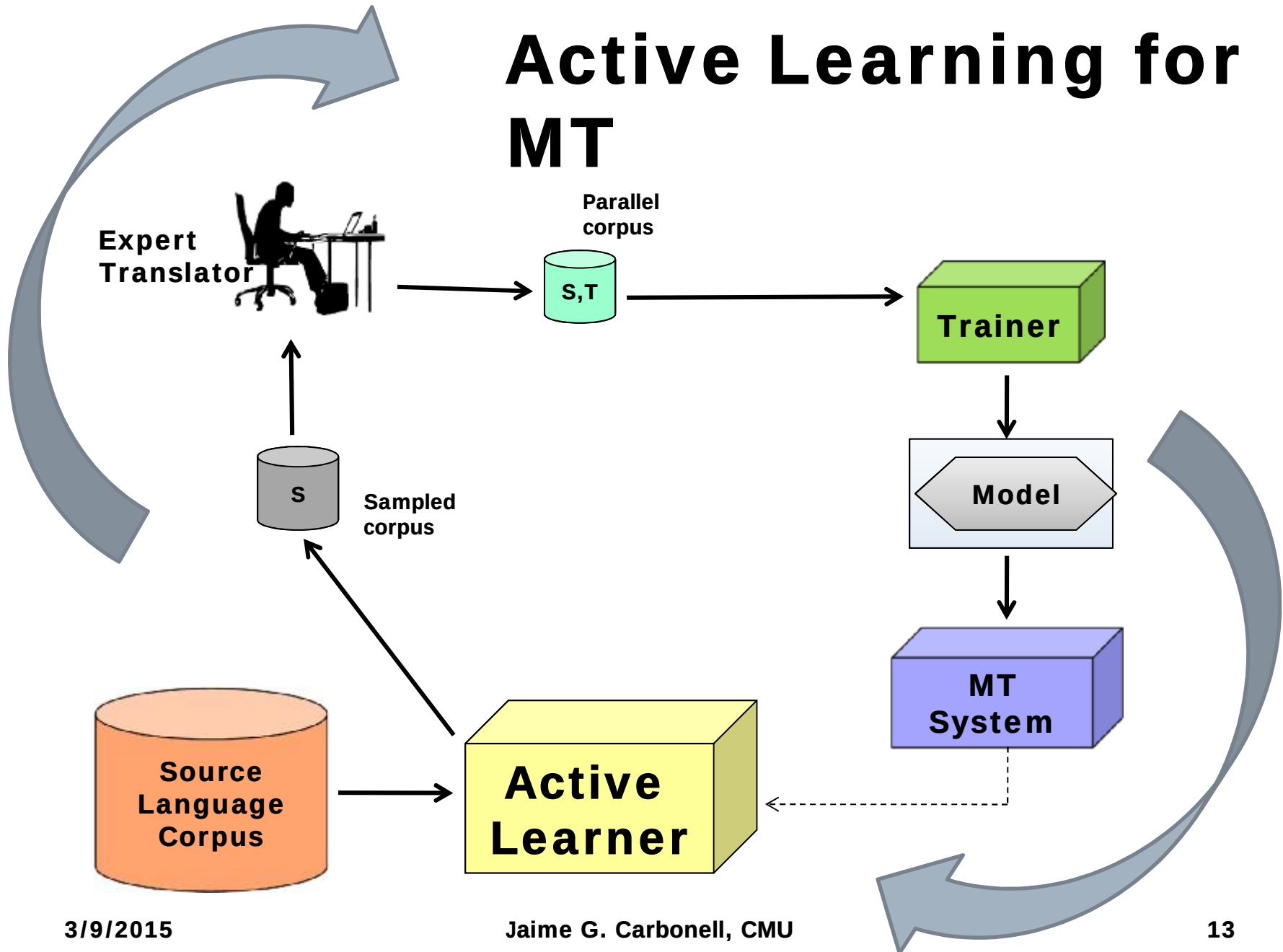
Line one's jacket – uwagi o nijuu ni suru (上着を二重にする)

Actor's line – serifu (セリフ)

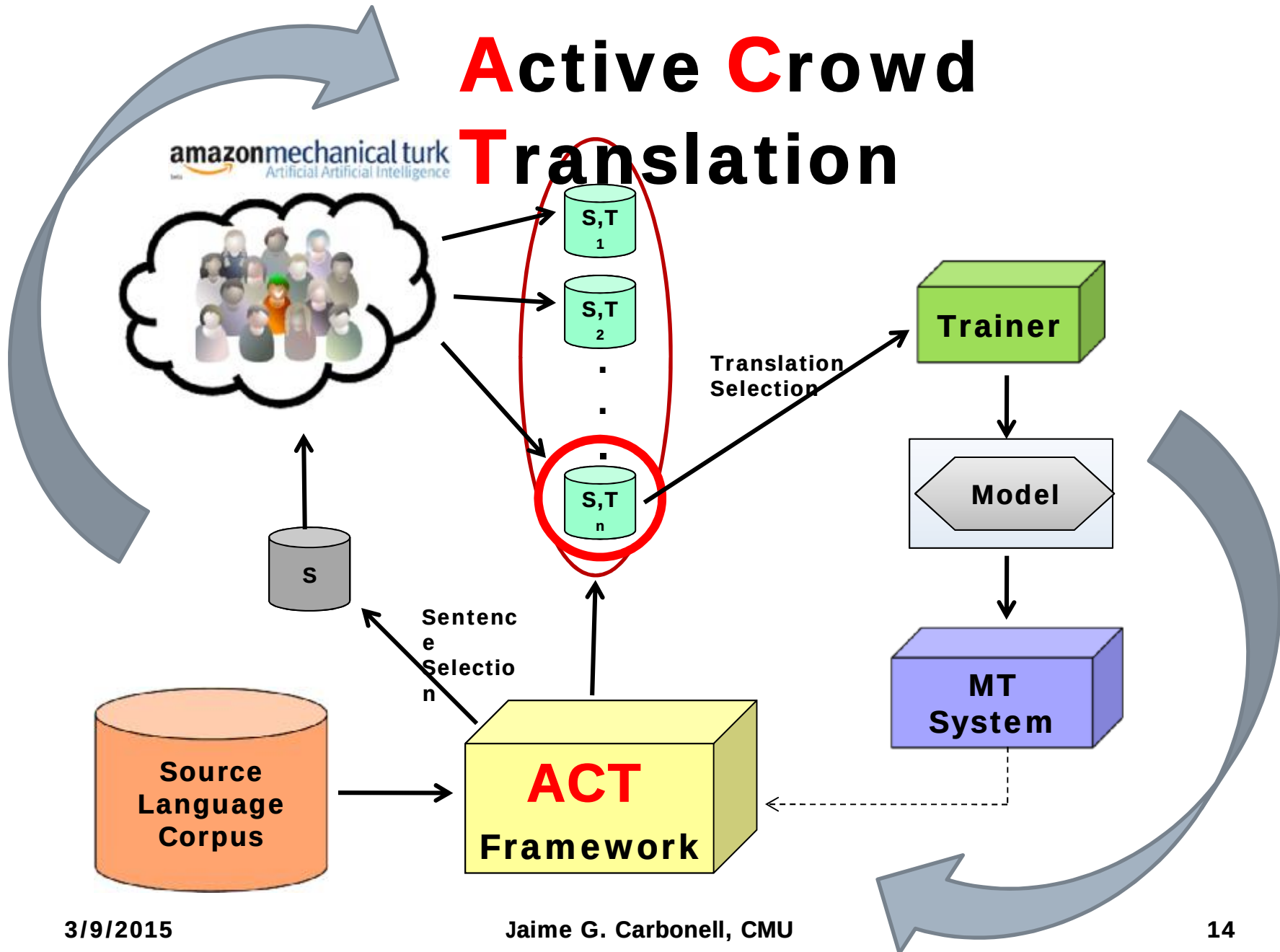
Get a line on – joho o eru (情報を得る)

Sometimes local context suffices (as above)

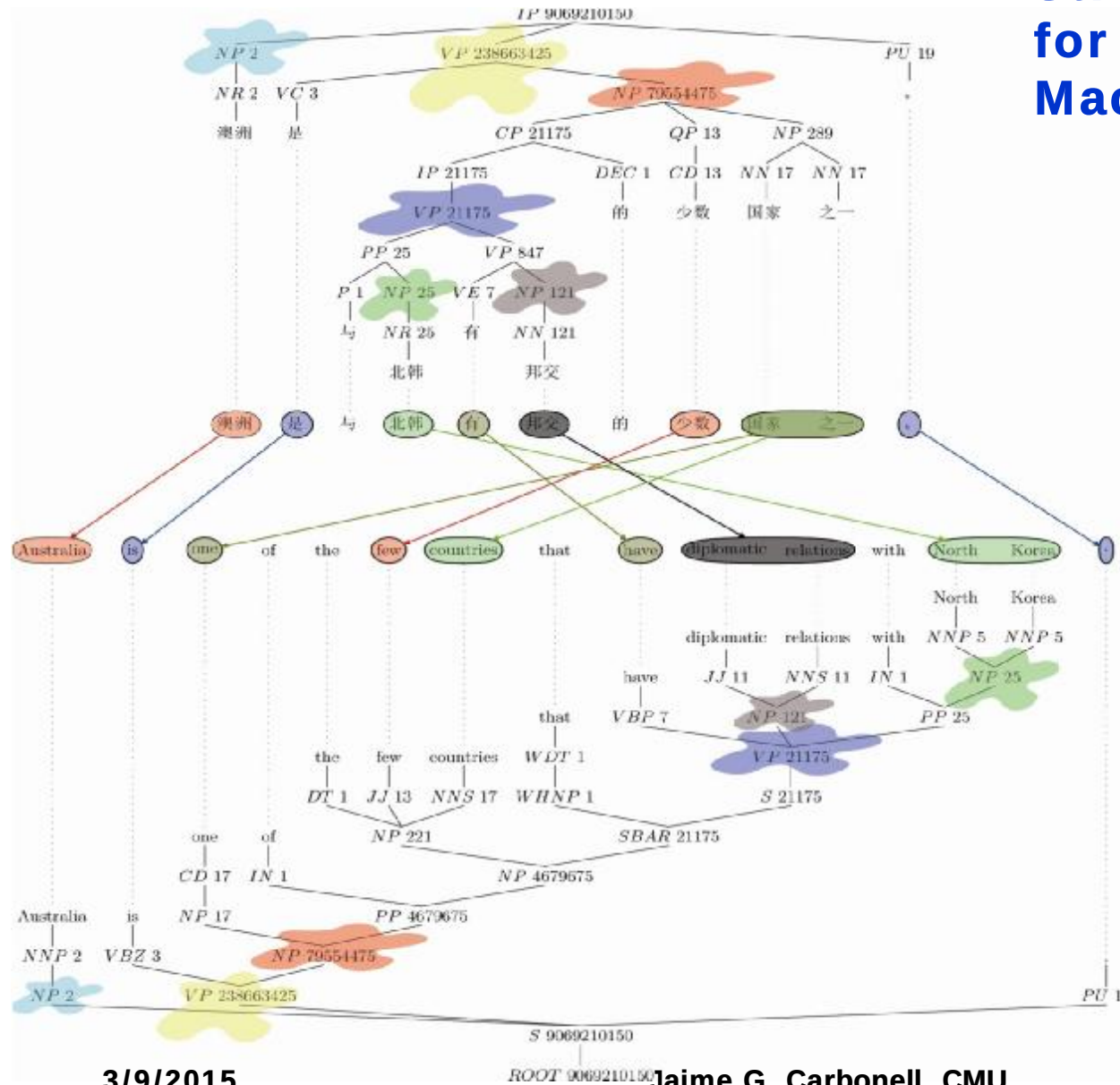
Active Learning for MT



Active Crowd Translation



Structure Learning for Next-Generation Machine Translation



- Tree-tree aligner enforces equivalence constraints and optimizes over terminal alignment scores (words/phrases)

- Resulting aligned nodes are highlighted in figure

- Transfer rules are partially lexicalized and read off tree.

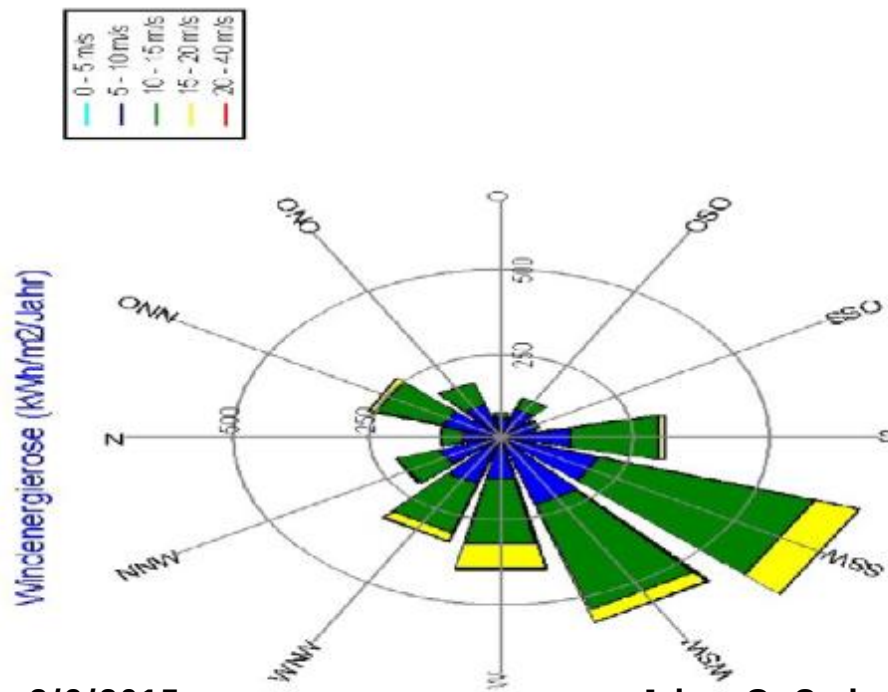
Active vs Proactive Learning

	Active Learning	Proactive Learning
Number of Oracles	Individual (only one)	Multiple, with different capabilities, costs and areas of expertise
Reliability	Infallible (100% right)	Variable across oracles and queries, depending on difficulty, expertise, ...
Reluctance	Indefatigable (always answers)	Variable across oracles and queries, depending on workload, certainty, ...
Cost per query	Invariant (free or constant)	Variable across oracles and queries, depending on workload, difficulty, ...

Note: “Oracle” $\in$ {expert, experiment, computation, ...}

Proactive Learning: Wind Sampling

- Predict: Prevalent Direction, Speed, seasonality
- Measurement towers: Expensive – where, when?



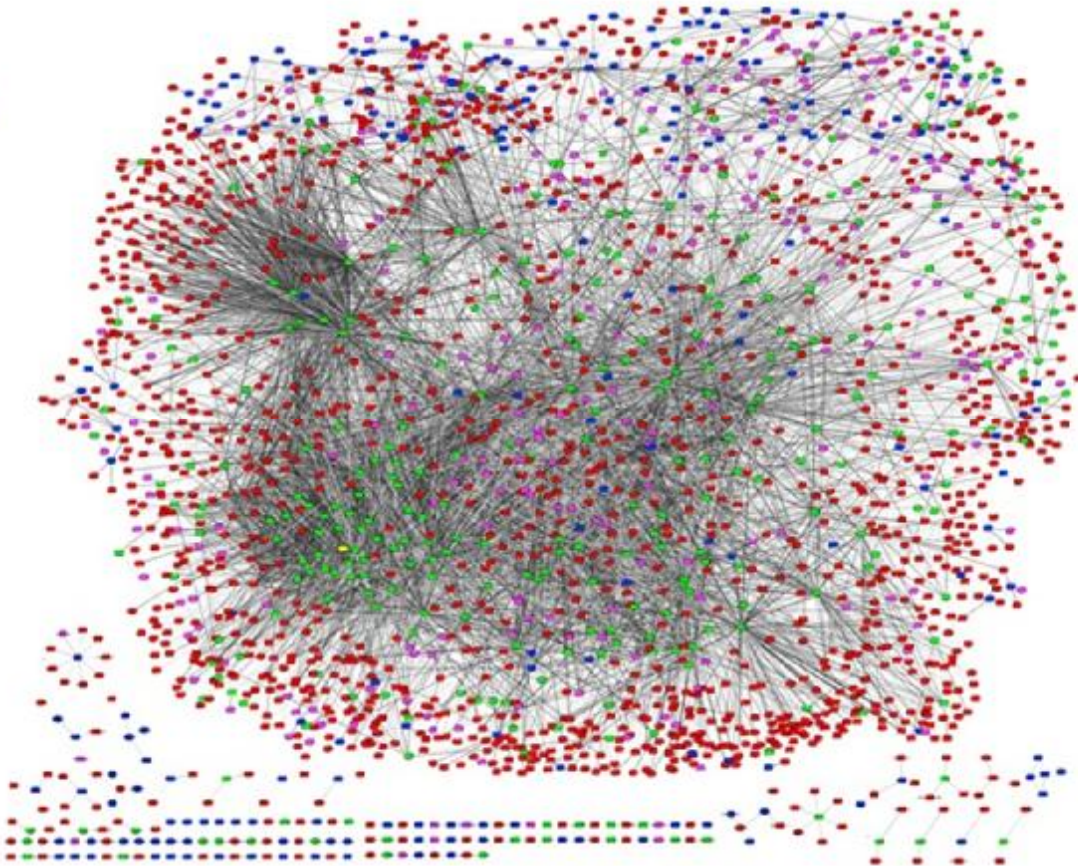
3/9/2015

Jaime G. Carbonell, CMU



17

Inferring the Protein Interactome

- 
- Type I Receptor
 - GPCR
 - Ligand
 - Other

Proteins

Receptor	551
Other	1752

Interactions

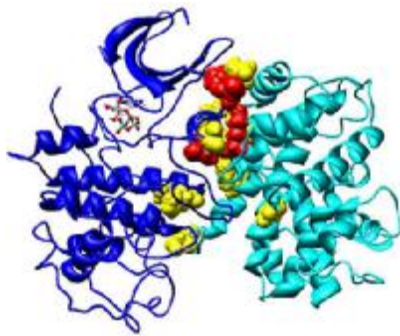
Total	9144
HPRD Known	1462

- Degree distribution / Hub analysis / Disease checking
- Graph modules analysis (from bi-clustering study)
- Protein-family based graph patterns (receptors / ligands / etc)

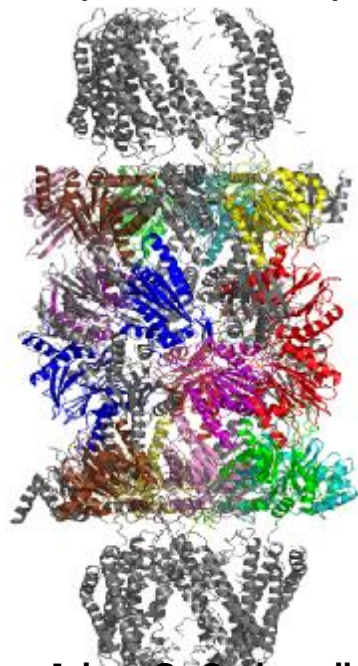
PPIs: Protein-Protein Interactions

- o The cell machinery is run by the proteins
 - n Enzymatic activities, replication, translation, transport, signaling, structural
- o Proteins interact with each other to perform these functions

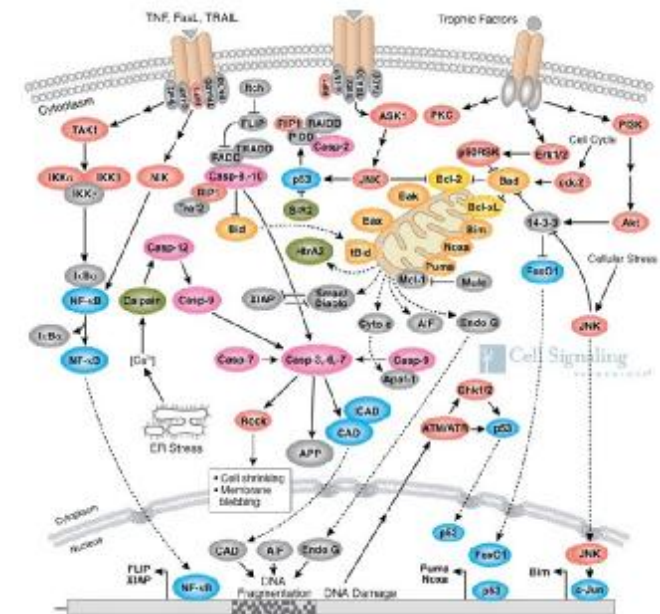
Through physical contact



Indirectly
in a protein complex



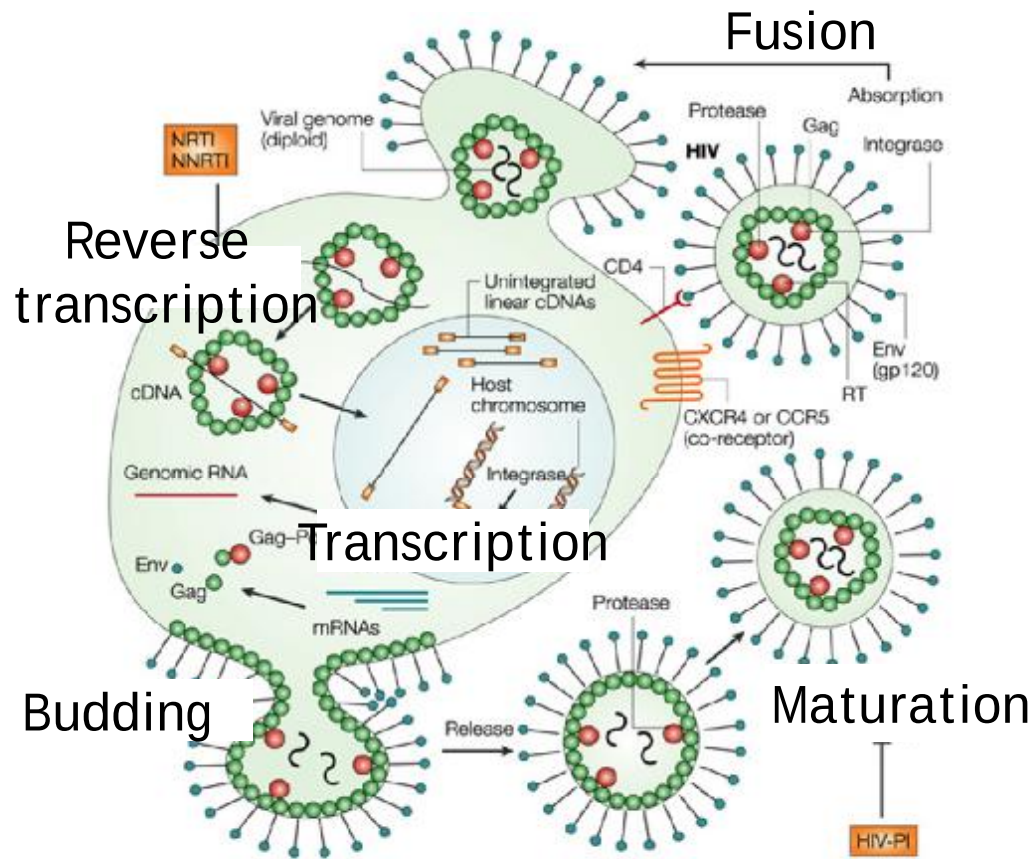
Indirectly in pathway



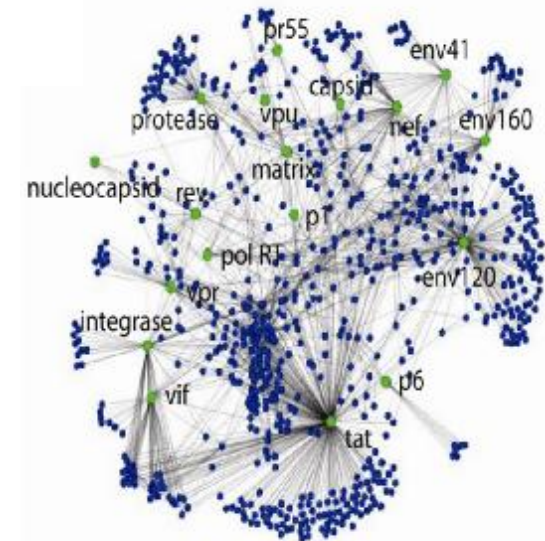
3/9/2015

Jaime G. Carbonell, CMU http://www.cellsignal.com/reference/pathway/Apoptosis_Overview.html

Host Pathogen Interactomes



HIV-1 depends on the cellular machinery in every aspect of its life cycle.



Peterlin and Torono, *Nature Rev Immu* 2003.

PROTEINS

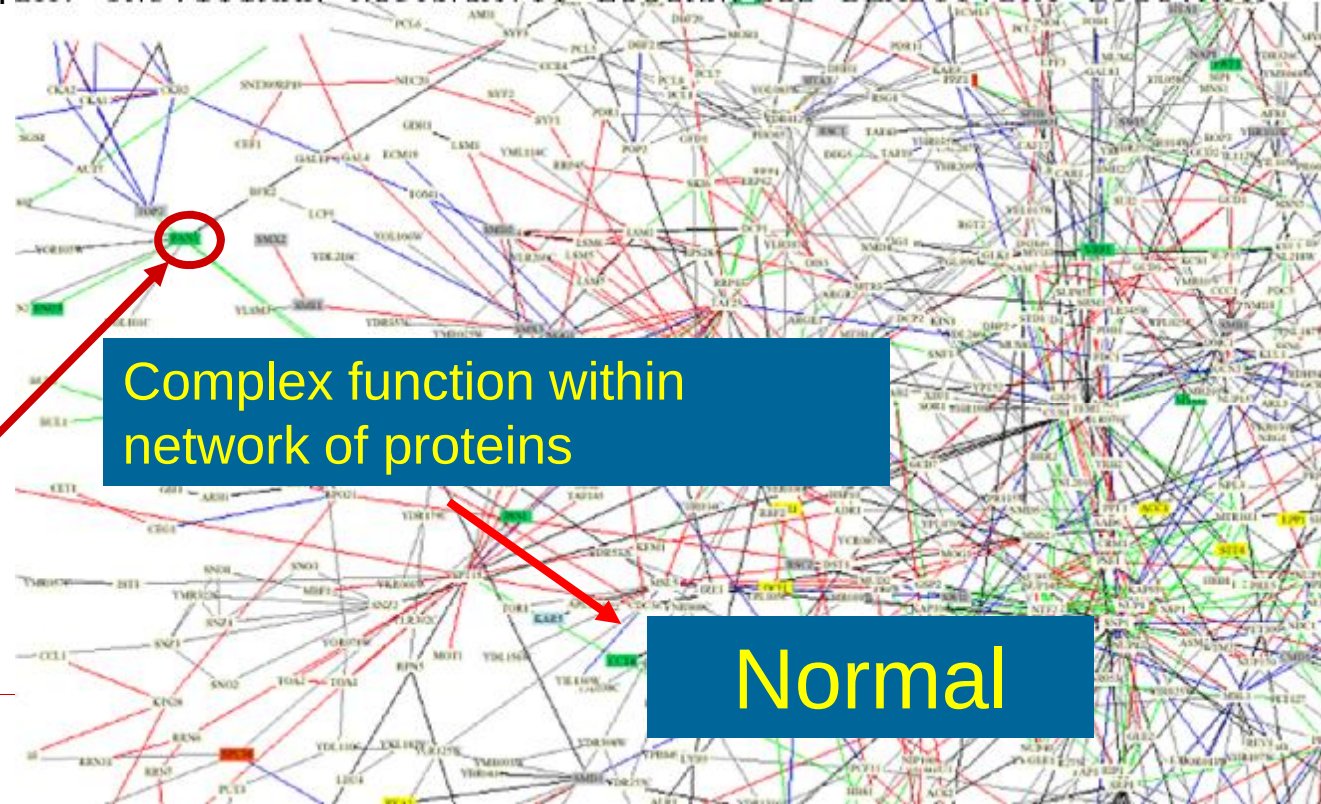
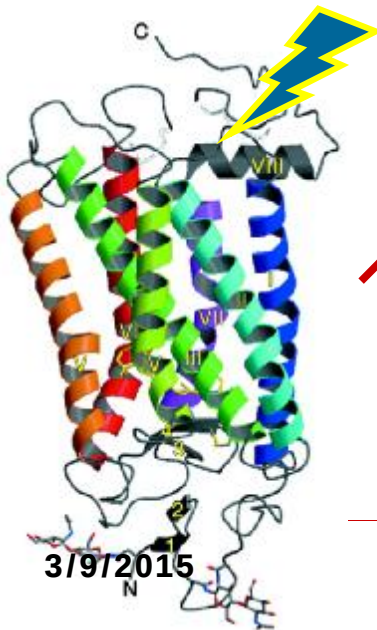
(Borrowed from: Judith Klein-Seetharaman)

Sequence à Structure à Function

Primary Sequence

MNGTEGPNFY VPFSNKTGVV RSPFEAPQYY LAEPWQFSML AAYMFLLIML GFPINFLTLY VTVQHKLRRT
PLNYILLNLA VADLFMVFGG FTTTLYTSLH GYFVFGPTGC NLEGFFATLG GEIALWSLVV LAIERYVVVC
KPMNSNRFGE NHAIMGVAFT WMALACAAP PLVGWSRYIP EGMQCSCGID YYTPHEETNN ESFVIYMFVV
HFIIPLIVIF FCYQQLVFTV KEAAAQQQES ATTQKAEKEV TRMVIIMVIA FLICWLPHYAG VAFYIFTHQG
SDFGPIMFTI PAFFAKTSAV YNPVIYIMMN KOERNMVTI LCCGKNPLGD DEASTTVSKT ETSOVAPA

↓ Folding
3D Structure



PROTEINS

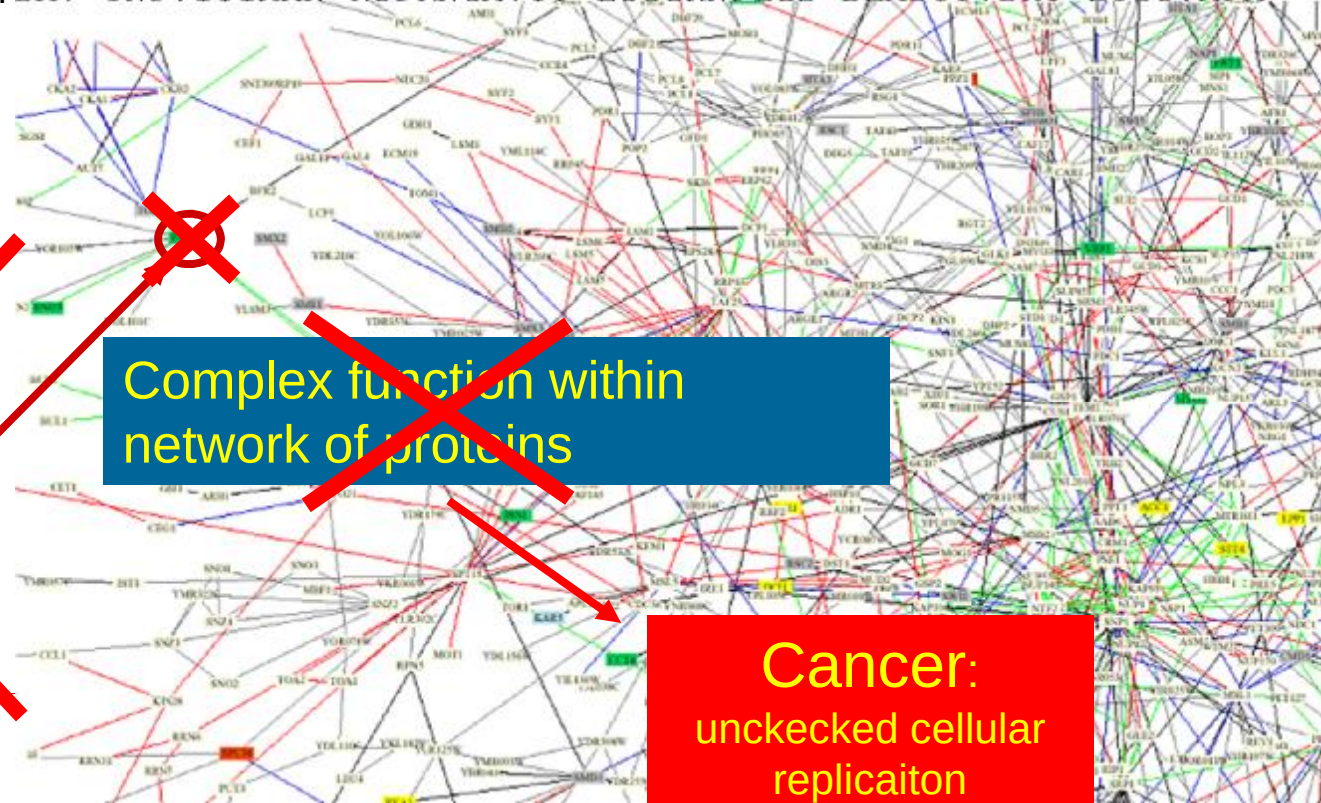
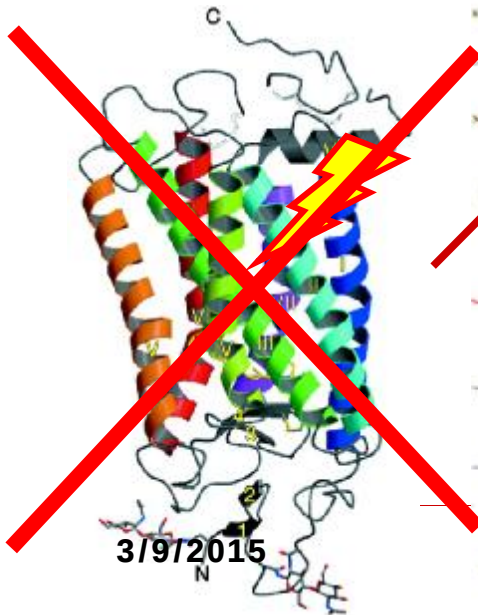
Sequence à Structure à Function

Primary Sequence

MNGTEGPNFY VPFSNKTGVV RSPFEAPQYY LAEPWQFSML AAYMFLLIML GFPINFLY VTVQHKKLRT
PLNYILLNLA VADLFMVFGG FTTTLYTSLH GYFVFGPTGC NLEGFFATLG GEIALWGLVV LAIERYVVVC
KPMNSNFRFGE NHAIMGVAFT WMALACAAP PLVGWSRYIP EGMQCSCGID YYTHIETNN ESFVIYMFVV
HFIIPLIVIF FCYQQLVFTV KEAAAQQQES ATTQKAEKEV TRMVIIMVIA FLICWLPHYAG VAFYIFTHQG
SDFGPIMFTI PAFFAKTSAY YNPVIYIMMN KOERNMVTI LCCGKNPLGD DEASTTVSKT ETSOVAPA

↓ Folding

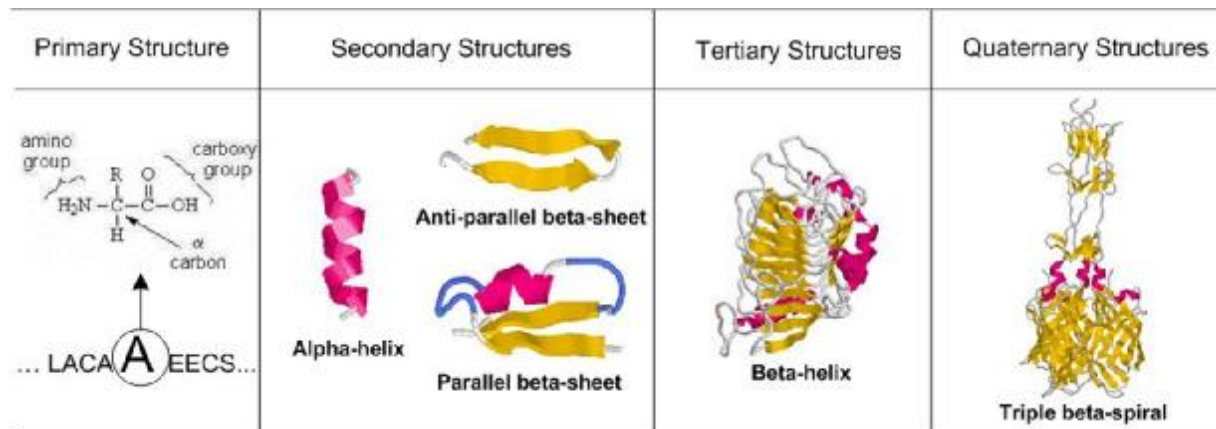
3D Structure



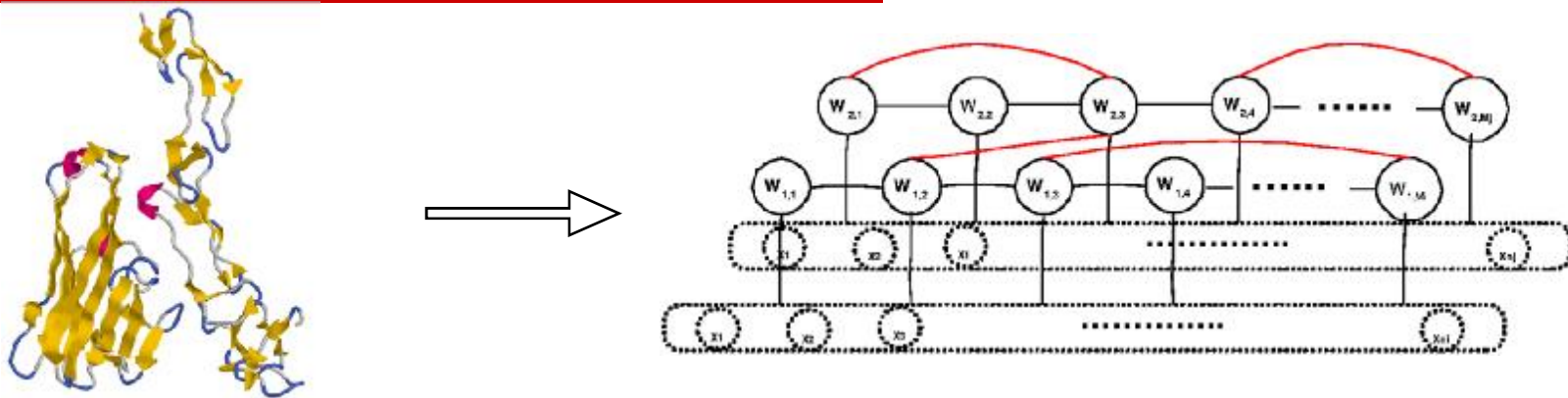
Cancer:
unchecked cellular
replicaition

The “Folding Problem:” 3D Protein Structure Determination In Silico

- o In Vitro Lab experiments: time, cost, uncertainty, ...
 - n X-ray crystallography (months to crystalize, uncertain outcome)
Nobel Prize, Kendrew & Perutz, 1962
 - n NMR spectroscopy (only works for small proteins or domains)
Nobel Prize, Kurt Wuthrich, 2002
- o The gap between sequence and structure necessitates computational methods of protein structure determination
 - n 3,023,461 sequences v.s. 36,247 resolved structures (1.2%)



Linked Segmentation CRF



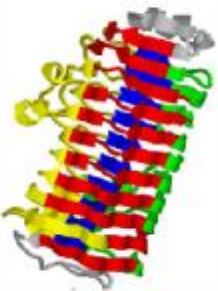
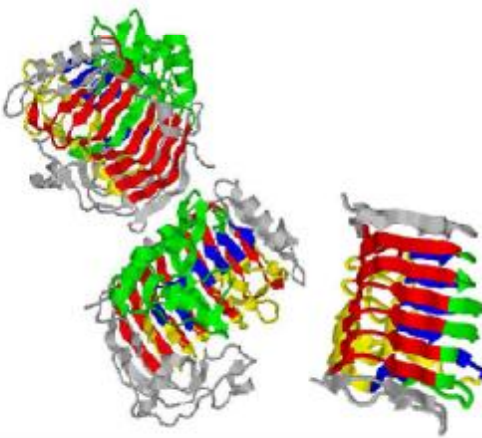
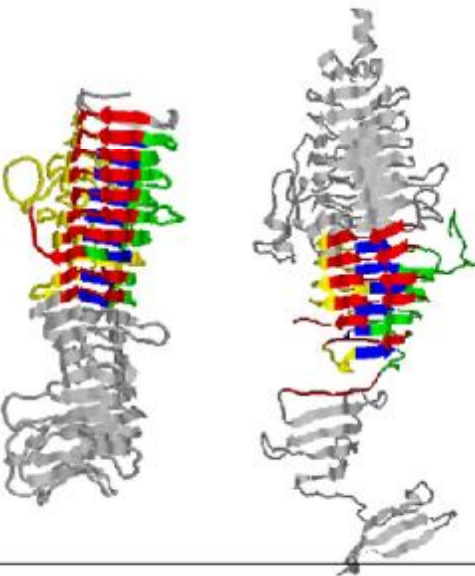
- **Node:** secondary structure elements and/or simple fold
- **Edges:** Local interactions and long-range *inter-chain* and *intra-chain* interactions
- **L-SCRF:** conditional probability of $\mathbf{y}$ given $\mathbf{x}$ is defined as

$$P(\mathbf{y}_1, \dots, \mathbf{y}_R | \mathbf{x}_1, \dots, \mathbf{x}_R) = \frac{1}{Z} \prod_{\mathbf{y}_{i,j} \in V_G} \exp\left(\sum_k l_k f_k(\mathbf{x}_i, \mathbf{y}_{i,j})\right) \prod_{\langle \mathbf{y}_{i,j}, \mathbf{y}_{a,b} \rangle \in E_G} \exp\left(\sum_l m_l g_l(\mathbf{x}_i, \mathbf{x}_a, \mathbf{y}_{i,j}, \mathbf{y}_{a,b})\right)$$

$\uparrow$
 Joint Labels

Fold Alignment Prediction: β -Helix

- Predicted alignment for known β -helices on cross-family validation

Group	Perfect match	Good match	OK match
Missing rungs	0	1-2	3 or more
PDB-ID	1czf	1air, 1bhe, 1bn8, 1dbg, 1ee6(right), 1idj, 1ktw(left), 1qcx, 1qjv, 1rmg	1dab(left), 1ea0, 1tyu(right)
			

Parting Thoughts

- o **Data Science = Big Data + ML + Domains**
 - n Big data requires big (scalable) ML
 - n Leverage scarce human expertise
- o **Machine Learning is Engine for AI**
 - n Heretofore it was hand-coded knowledge
 - n Applicable to virtually any domain
- o **Big Impact Everywhere**
 - n **Bio**: genetic, molecular, healthcare
 - n **Language**: speech, MT, text mining
 - n **Education**: student models, transfer
 - n **Energy**: optimization, production, conservation
 - n **Finance**: investment, fraud detection, risk analysis

THANK YOU!

