ETRI Electronics and Telecommunications Research Institute

빅데이터 이야기
@ 더블에스텍

2015. 8. 4.

김 성 수

sungsoo@etri.re.kr

Data Management Research Section

ETRI

About Me

Sung-Soo Kim

Senior Researcher

Electronics and Telecommunications Research Institute (ETRI),
South Korea
sungsoo@etri.re.kr
sungsookim@kaist.ac.kr
http://sungsoo.github.com
about me

WORK EXPERIENCE

Electronics and Telecommunications Research Institute (ETRI)

September 2000 - Present
Senior Researcher

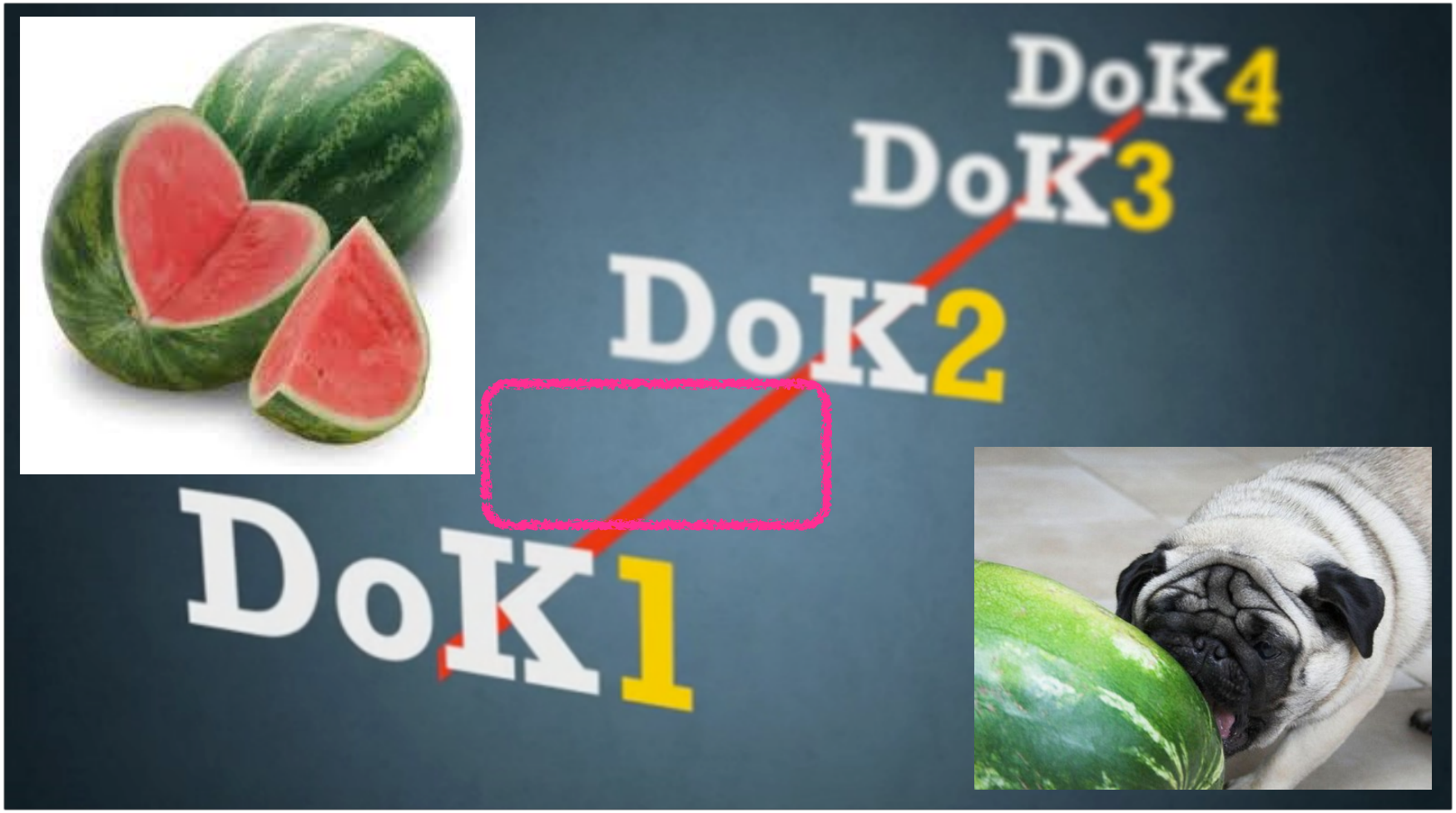
Research on multiscreen services, photo-realistic rendering (global illumination), real-time rendering, geometry compression algorithms, spatio-temporal data modeling/indexing and route determination algorithms.

Korea Advanced Institute of Science and Technology (KAIST)

- August 2000
Researcher

Research on reuse based software engineering for large information systems.

Depth of Knowledge (DoK) Level of the Seminar



Electronics and Telecommunications
Research Institute

Part I: Big Data 101

Big Data, NoSQL, MapReduce Framework and Hadoop

04 August 2015

Sung-Soo Kim

sungsoo@etri.re.kr

Data Management Research Section

ETRI


KBS1 HD
16

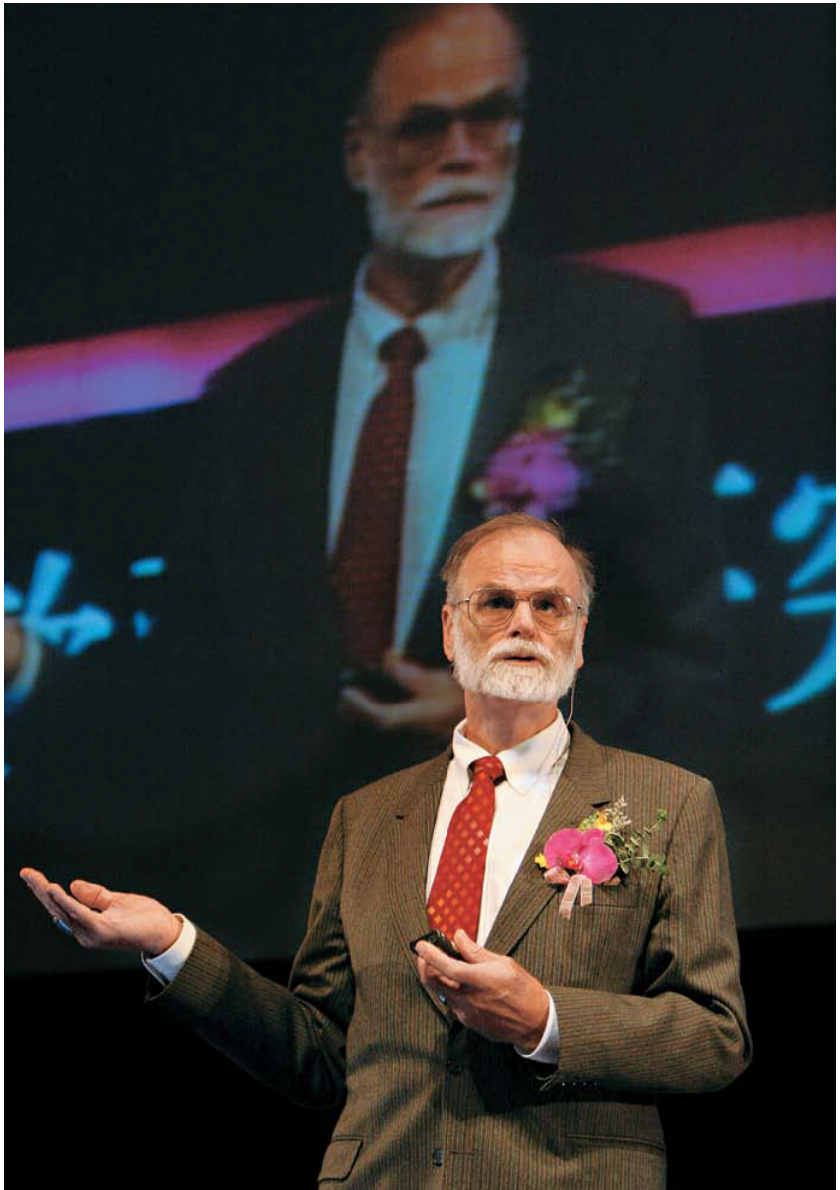
The Fourth Paradigm

The
FOURTH
PARADIGM

DATA-INTENSIVE SCIENTIFIC DISCOVERY

EDITED BY TONY HEY, STEWART TANSLEY, AND KRISTIN TOLLE

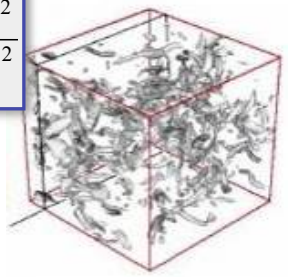
Jim Gray (Computer Scientist, Born Jan. 12. 1944 ~ Disappeared Jan. 28. 2007)

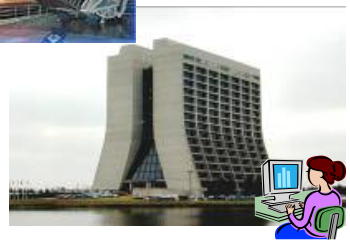


Science Paradigms

- Thousand years ago:
  science was **empirical**
  describing natural phenomena
- Last few hundred years:
  theoretical branch
  using models, generalizations
- Last few decades:
  a **computational** branch
  simulating complex phenomena
- Today:
  data exploration (eScience)
  unify theory, experiment, and simulation
  – Data captured by instruments
  Or generated by simulator
  – Processed by software
  – Information/Knowledge stored in computer
  – Scientist analyzes database / files
  using data management and statistics

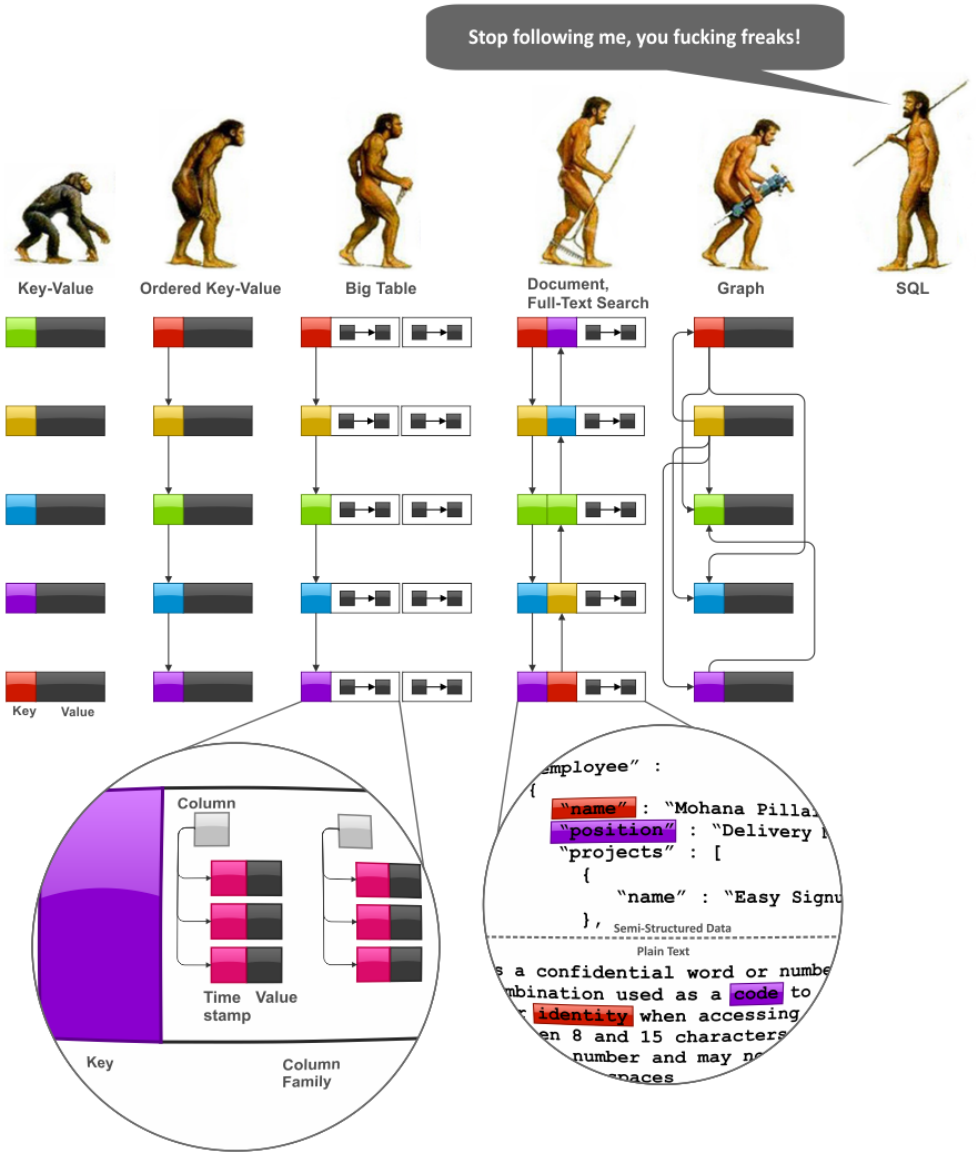
$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - \mathrm{K}\frac{c^2}{a^2}$$



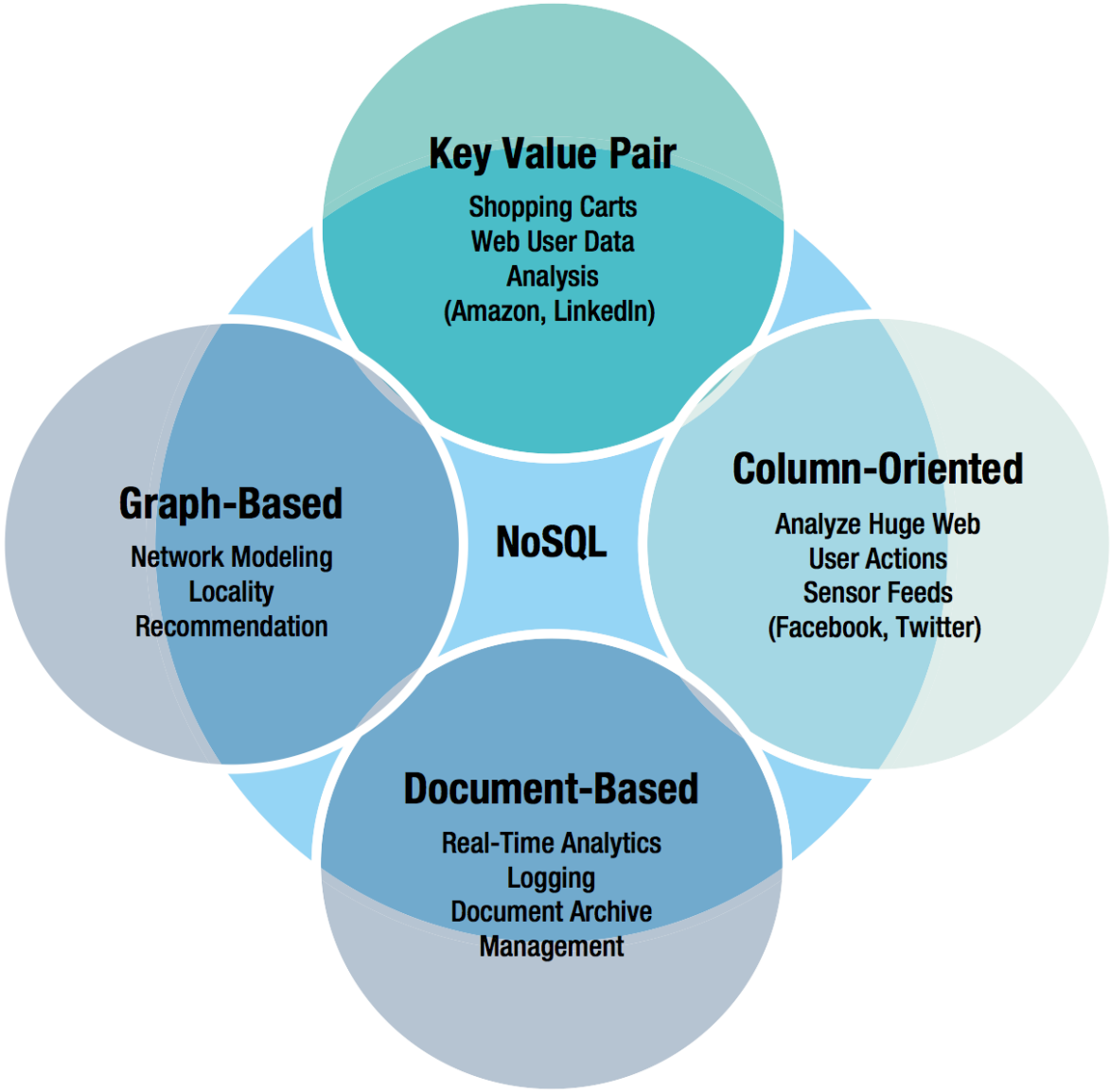


NoSQL

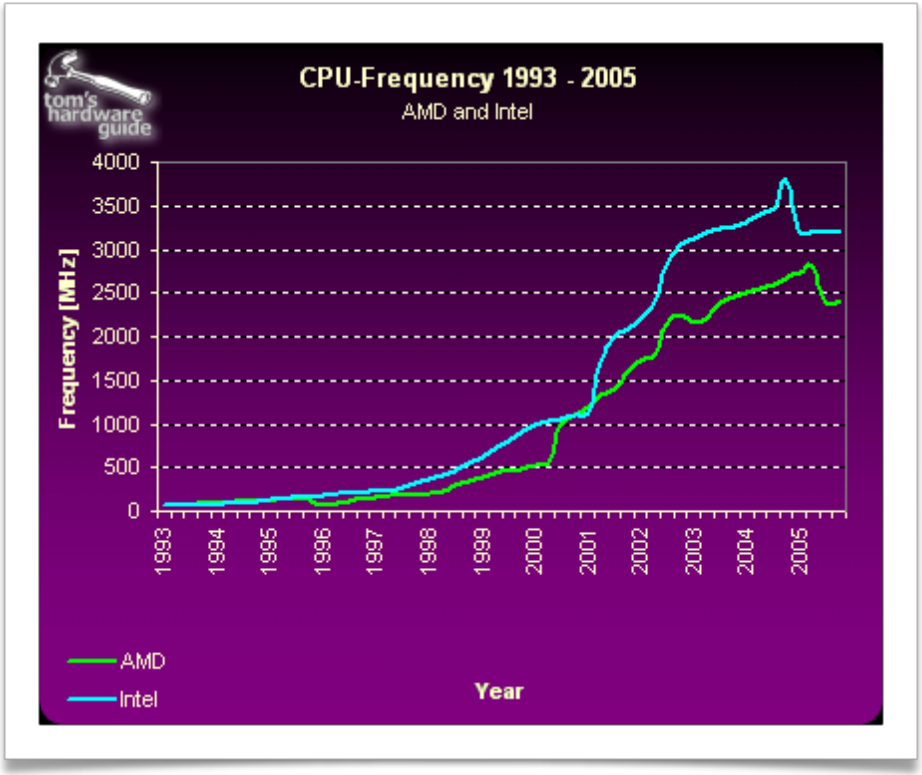
- "NoSQL" = "Not Only SQL"
- NoSQL Data Modeling Techniques
  - Key-Value
  - Ordered Key-Value
  - Big Table (Column-Oriented)
  - Document, Full-Text Search
  - Graph Database



NoSQL Database Typical Business Scenarios



Computer Speedup



Moore's Law

"The density of transistors on a chip *doubles every 18 months*, for the same cost" (1965)

Distributed Problems

- **Rendering multiple frames on high-quality animation**

Distributed Problems

- Simulating several hundred or thousand characters

Happy Feet © Kingdom Feature Productions; Lord of the Rings © New Line Cinema

Distributed Problems

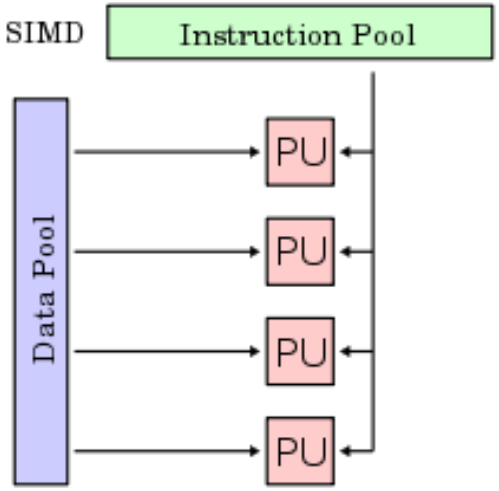
- Indexing the web (Google)
- Simulating an Internet-sized network of networking experiments (PlanetLab)
- Speeding up content delivery (Akamai)

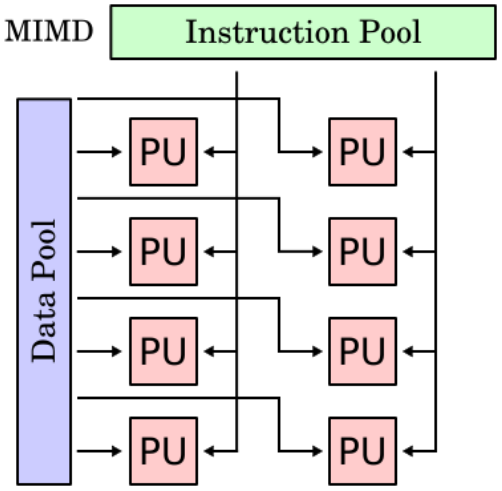
What is the *key attribute* that all these examples have in common?

Parallel vs. Distributed

- Parallel computing can mean:
  - Vector processing of data (SIMD)
  - Multiple CPUs in a single computer (MIMD)

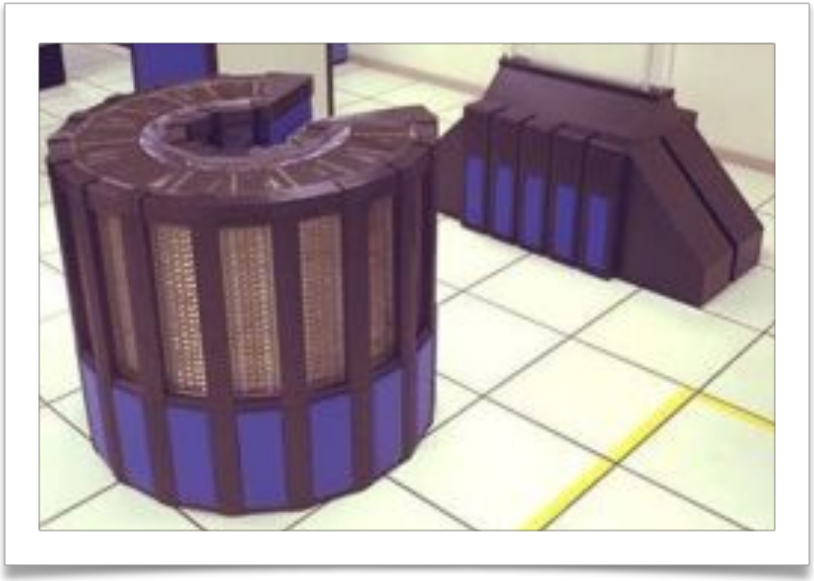
- Distributed computing is multiple CPUs across many computers (MIMD)





A Brief History … 1975 - 1985

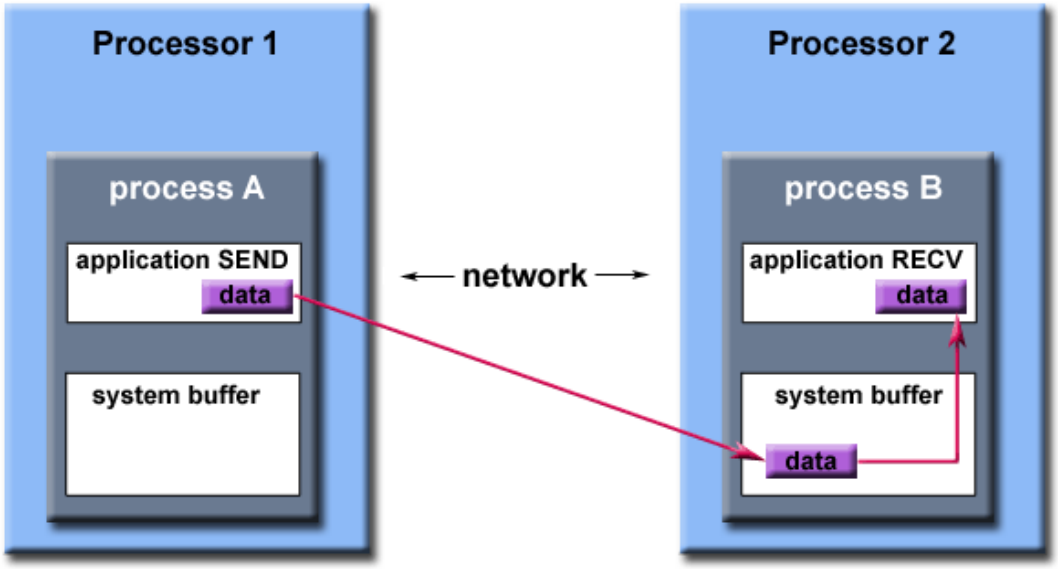
- Parallel computing was favored in the early years
- Primarily vector-based at first
- Gradually more thread-based parallelism was introduced



Cray 2 supercomputer (Wikipedia)

A Brief History … 1985 - 1995

- "Massively parallel architectures" start rising in prominence
- Message Passing Interface (MPI) and other libraries developed
- Bandwidth was a big problem



Path of a message buffered at the receiving process

A Brief History … 1995 - Today

- Cluster/grid architecture increasingly dominant
- Web-wide cluster software
- Companies like Google take this to the extreme (10,000 node clusters)

"Helps detect life threatening conditions up to 24 hours sooner"

University of Ontario Institute of Technology (UOIT)

Detect Neonatal Patient Symptoms Sooner

Up to 24 Hours

Continuously correlate data

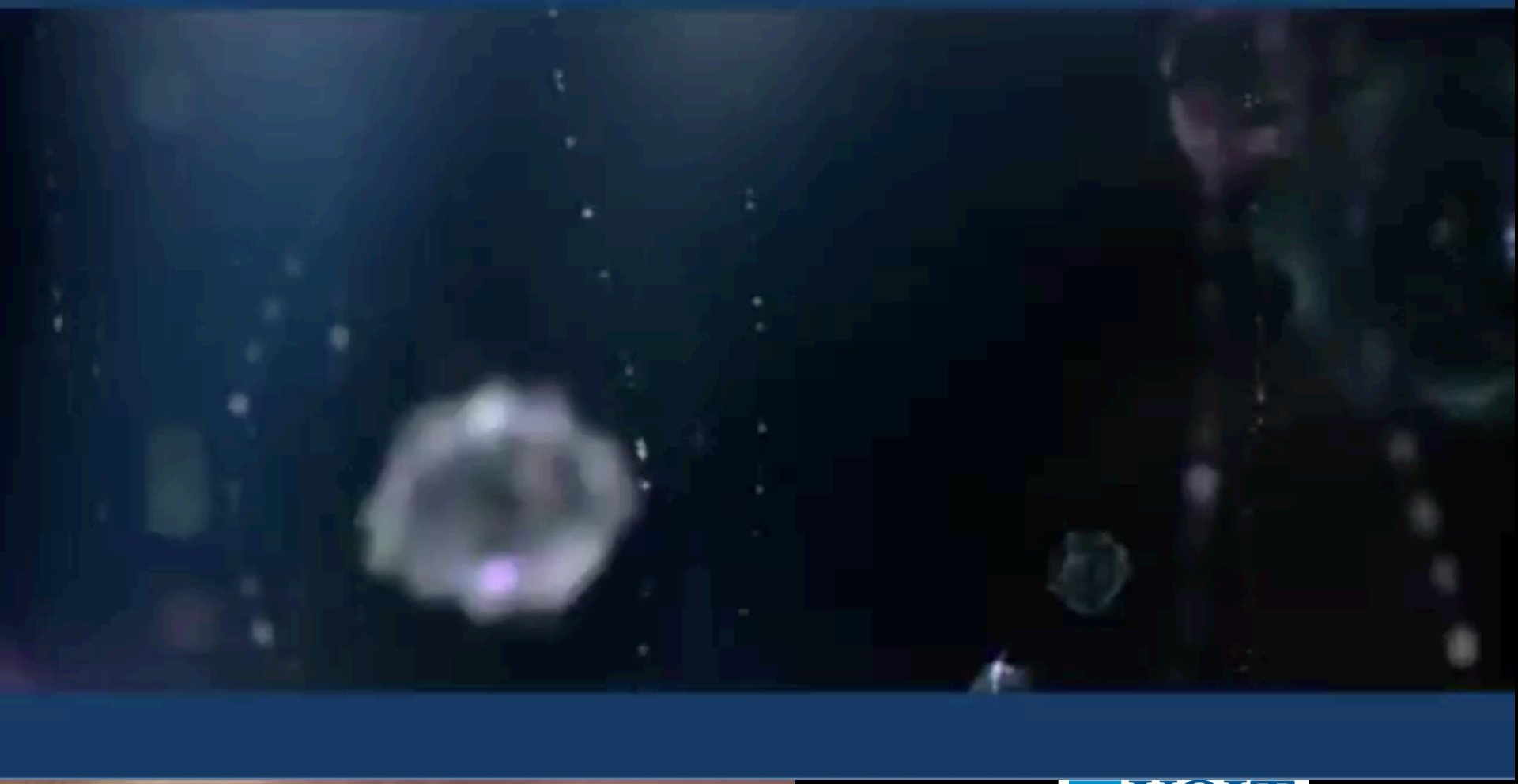
Thousands of events each second

Signal Processing and Data Cleansing

Heart Rate Variability

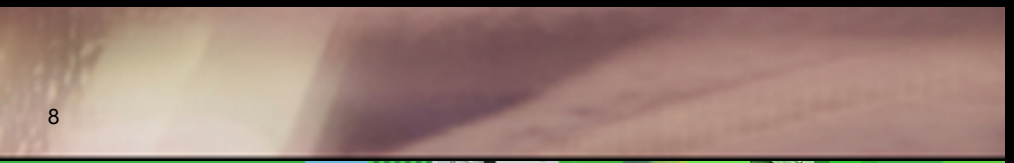
UOIT
CHALLENGE INNOVATE CONNECT

University of Ontario Institute of Technology (UOIT)



UOIT
CHALLENGE INNOVATE CONNECT

University of Ontario Institute of Technology (UOIT)



ETRI Electronics and Telecommunications Research Institute

MapReduce and Hadoop Ecosystem

Understanding MapReduce

MapReduce의 이해

- 나는 반장이다 -

Word Counting Problem

선생님 말씀

"반장. 이 책에 각 단어들이
몇 번씩 나왔는지 세어줄래?"

Oh My God!

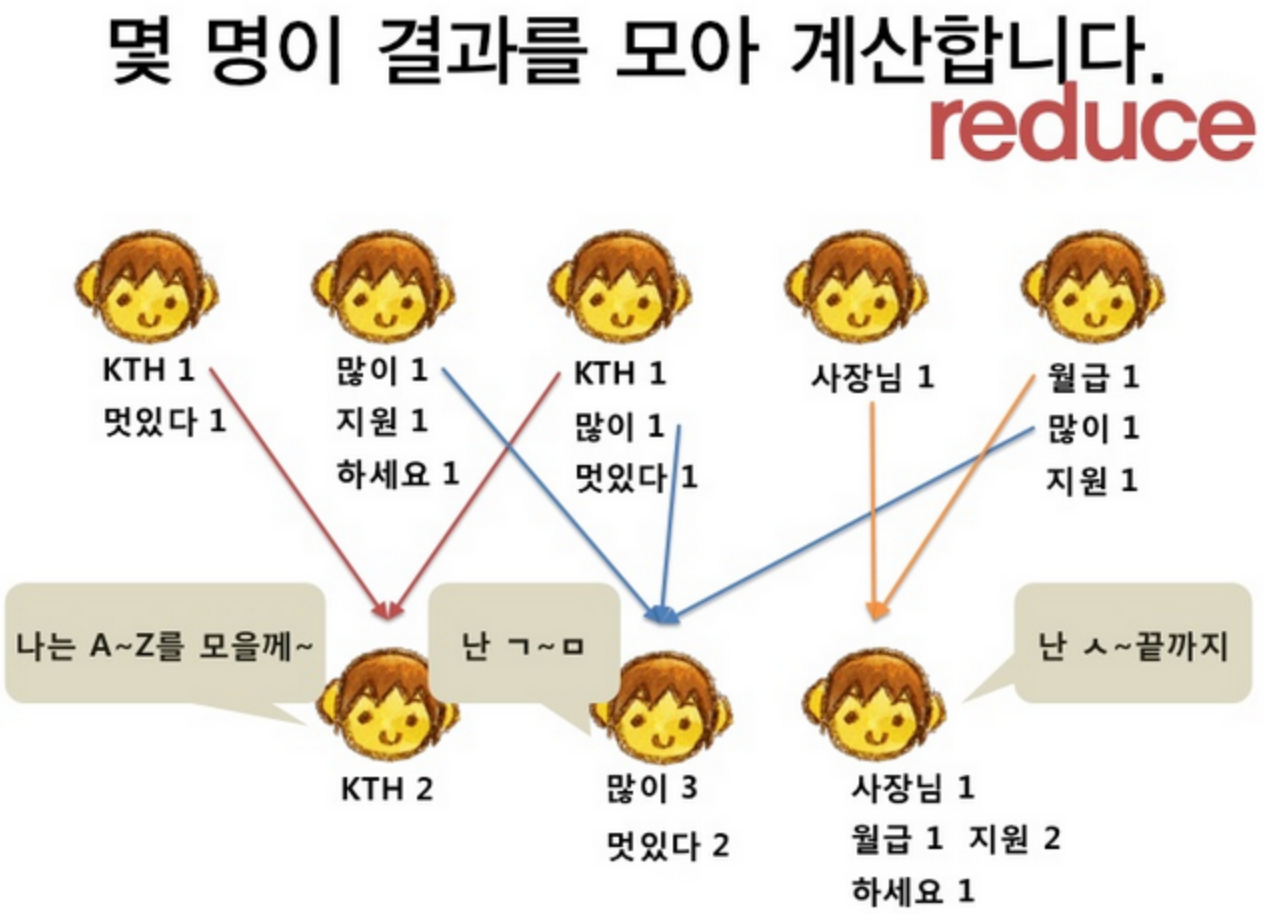
헐...

Problem Decomposition

애들을 풀어 일을 나눠 시킵니다.

map

Reduction


몇 명이 결과를 모아 계산합니다.
reduce
KTH 1
멋있다 1
많이 1
지원 1
하세요 1
KTH 1
많이 1
멋있다 1
사장님 1
월급 1
많이 1
지원 1
나는 A~Z를 모을께~
난 ㄱ~ㅁ
난 ㅅ~끝까지
KTH 2
많이 3
멋있다 2
사장님 1
월급 1 지원 2
하세요 1

What is MapReduce?

MapReduce

큰 작업을

잘게 나누기(map)

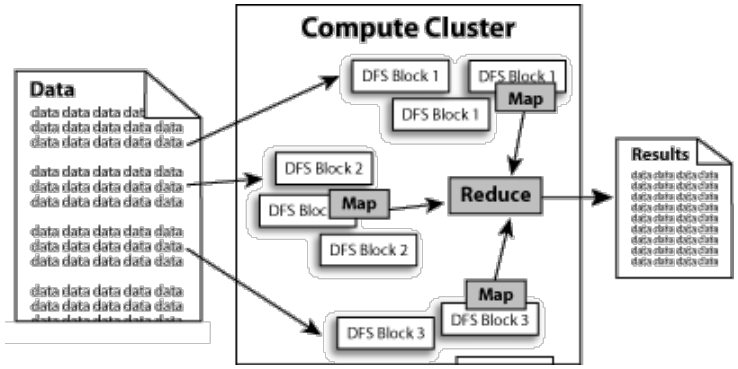
종류별로 모으기(reduce)

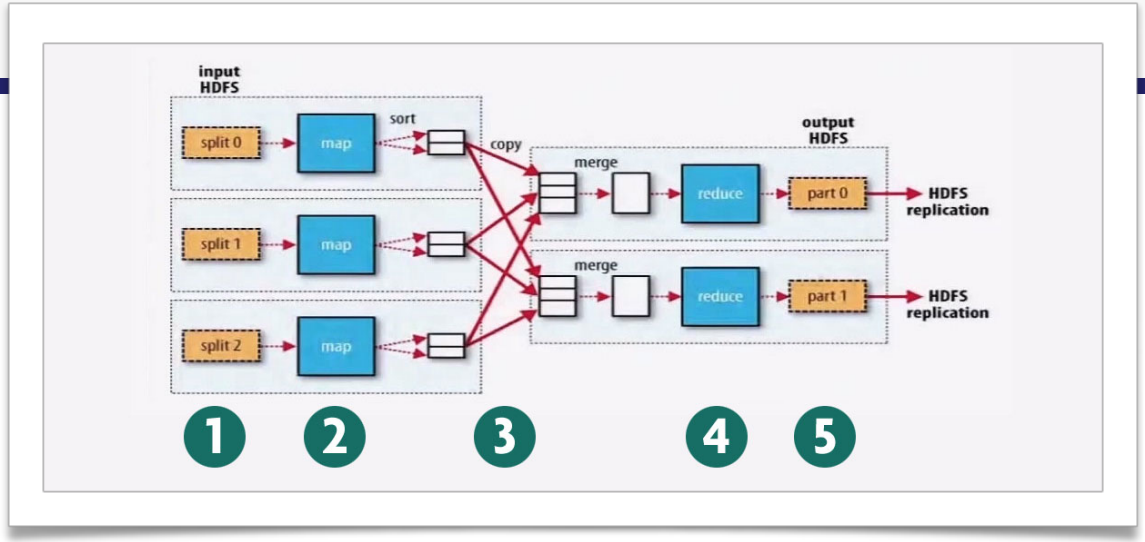
를 처리하는 방식을 말합니다.

MapReduce Framework

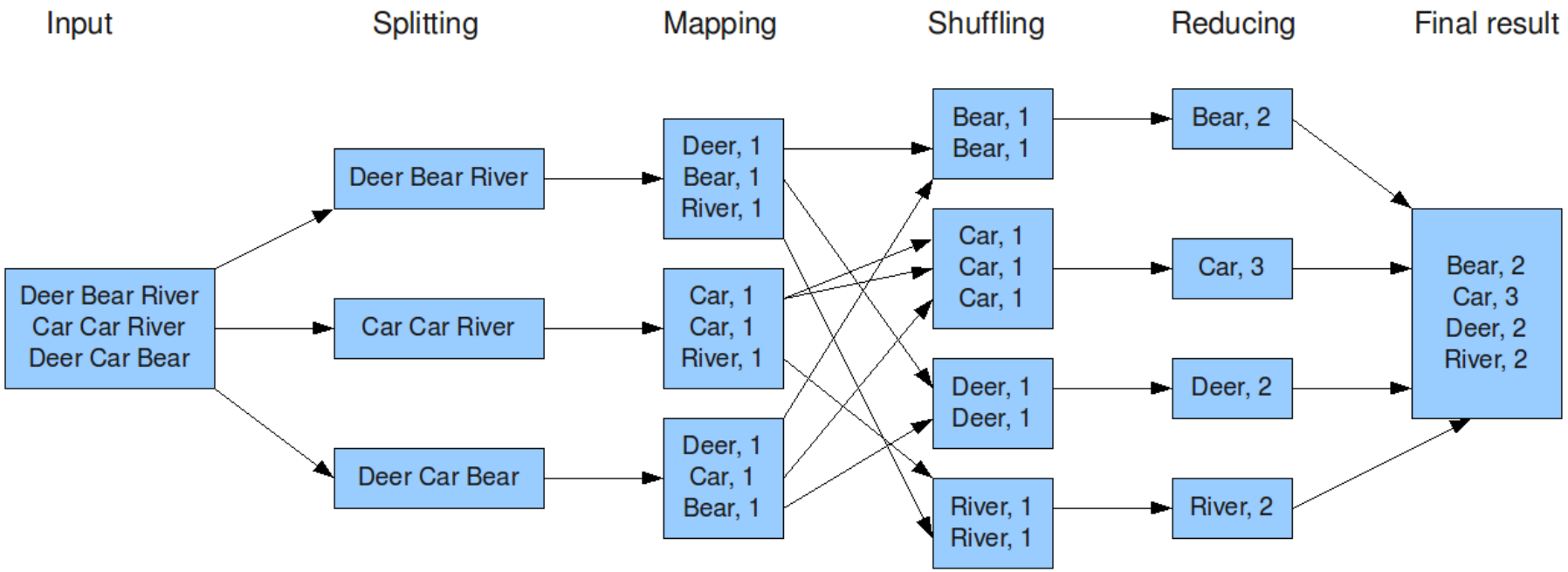
- **Originally from Google, open source Hadoop**
  - **No data model, data stored in files [GFS, HDFS; Hadoop Distributed File System]**
  - **User provides specific functions**
    - map(), reduce(), reader(), writer(), combiner()
  - **System provides data processing "glue", fault-tolerance, scalability**
- **Map and Reduce Functions**
  - **Map: *Divide* problem into *subproblems***
    - map(item) ➛ 0 or more <key, value> pairs
  - **Reduce: Do work on subproblems, *combine* results**
    - reduce(key, list-of-values) ➛ 0 or more records

WordCount Example





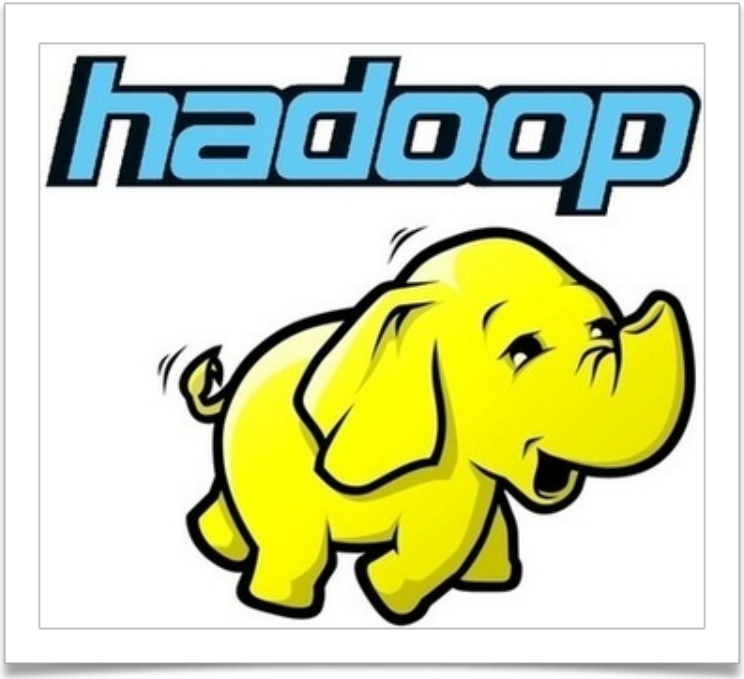
The overall MapReduce word count process



What is Apache Hadoop?

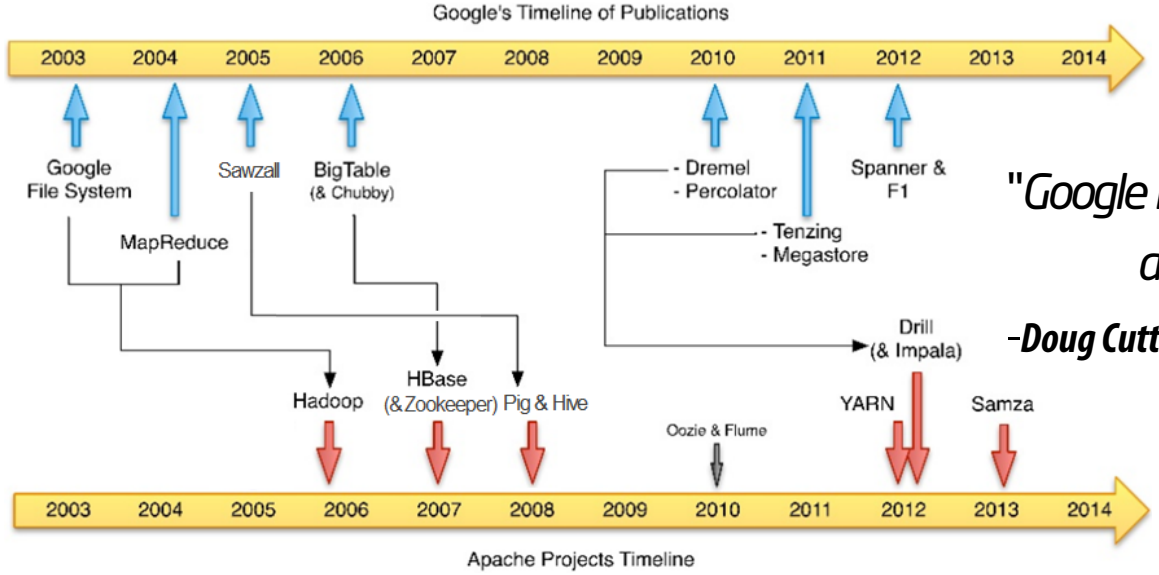
- Open source project for scalable data storage and processing
  - Inspired from Google's GFS, BigTable, MapReduce
  - Harnesses the power of *commodity servers*
  - Provides a distributed and *fault-tolerant solution*

- "Core" Hadoop consists of two main parts
  - HDFS (storage)
  - MapReduce (processing)



Brief History of Hadoop

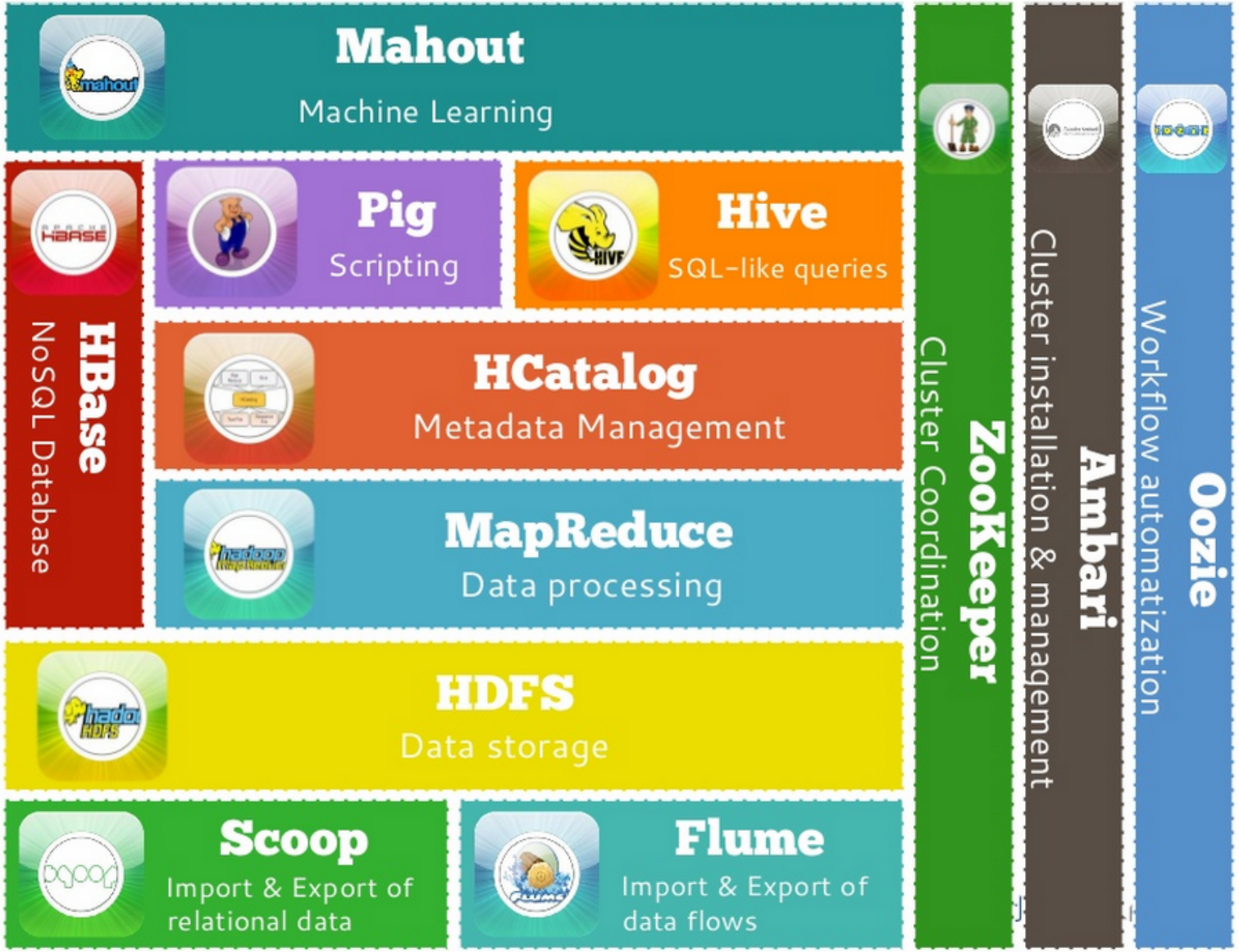
- Created by Doug Cutting
- Originated Apache Nutch [2002]
  - Nutch: web crawler engine, a part of Lucene project
  - Lucene: text search engine
- NDFS(Nutch Distributed File System, 2004)
- MapReduce (2005)
- Official start of Apache Hadoop project (Feb 2006)



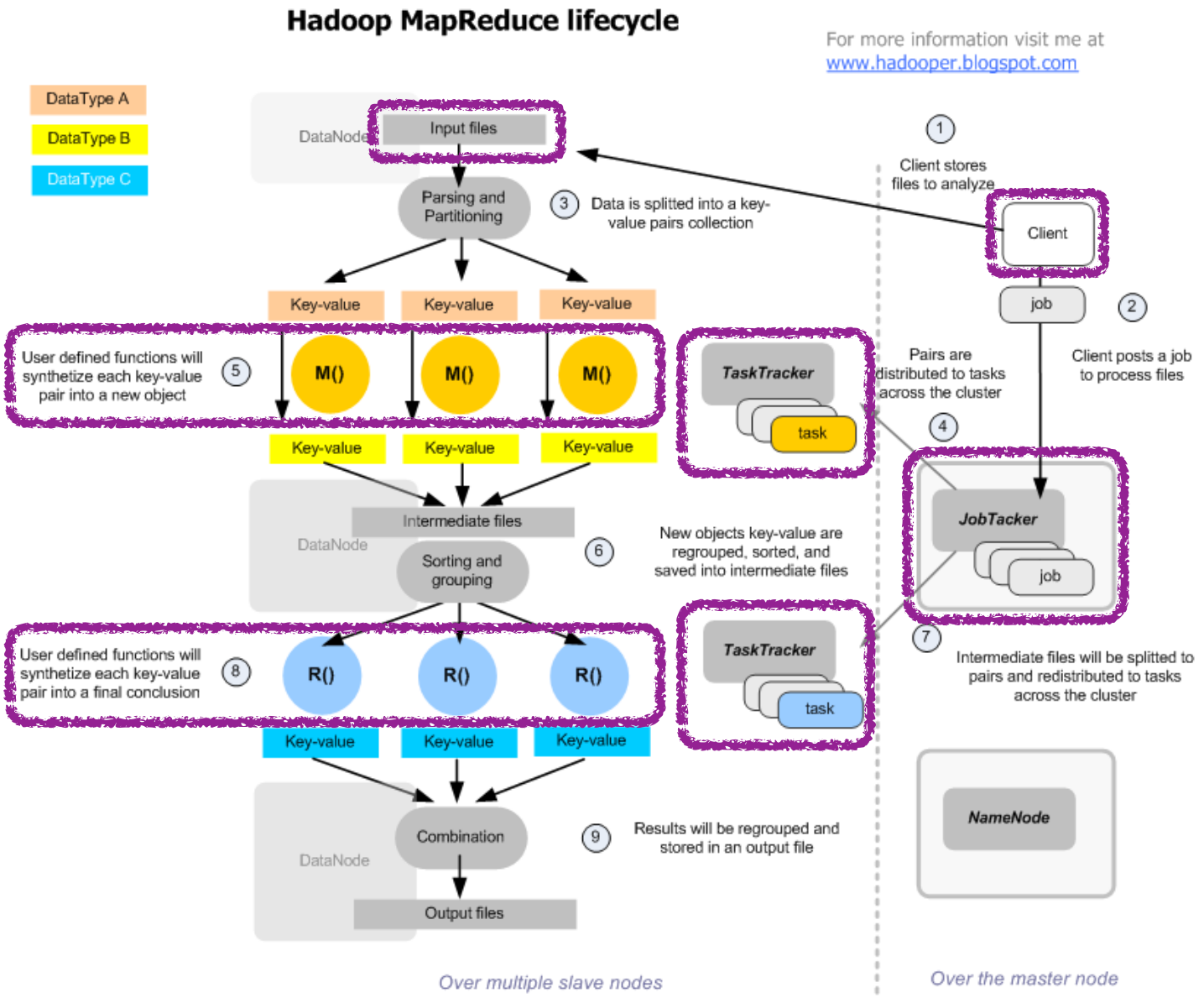
"*Google is living a few years in the future and sending the rest of us messages.*"

-***Doug Cutting***@ O'Reilly Strata Conference in London, 2013

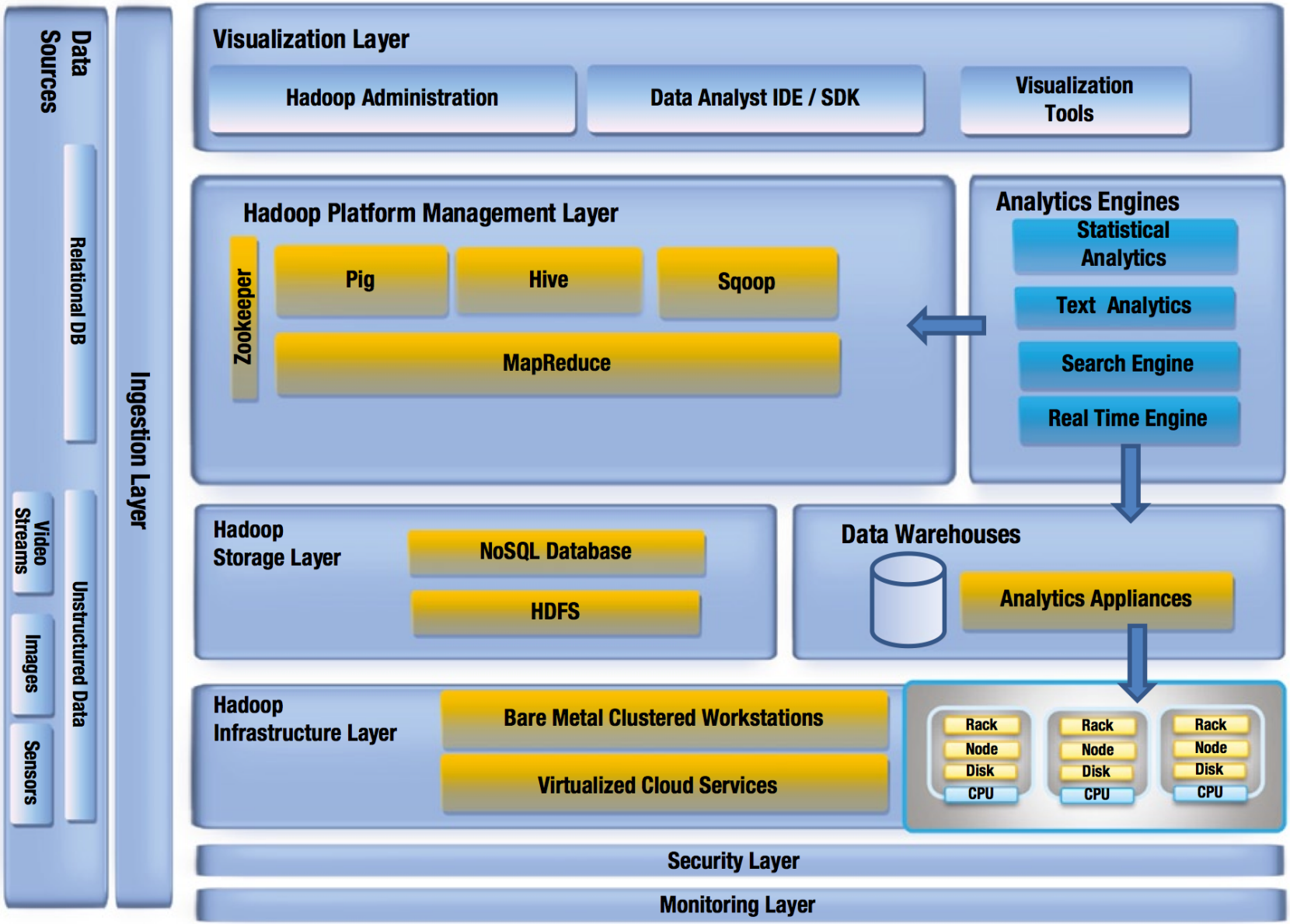
Hadoop Ecosystem



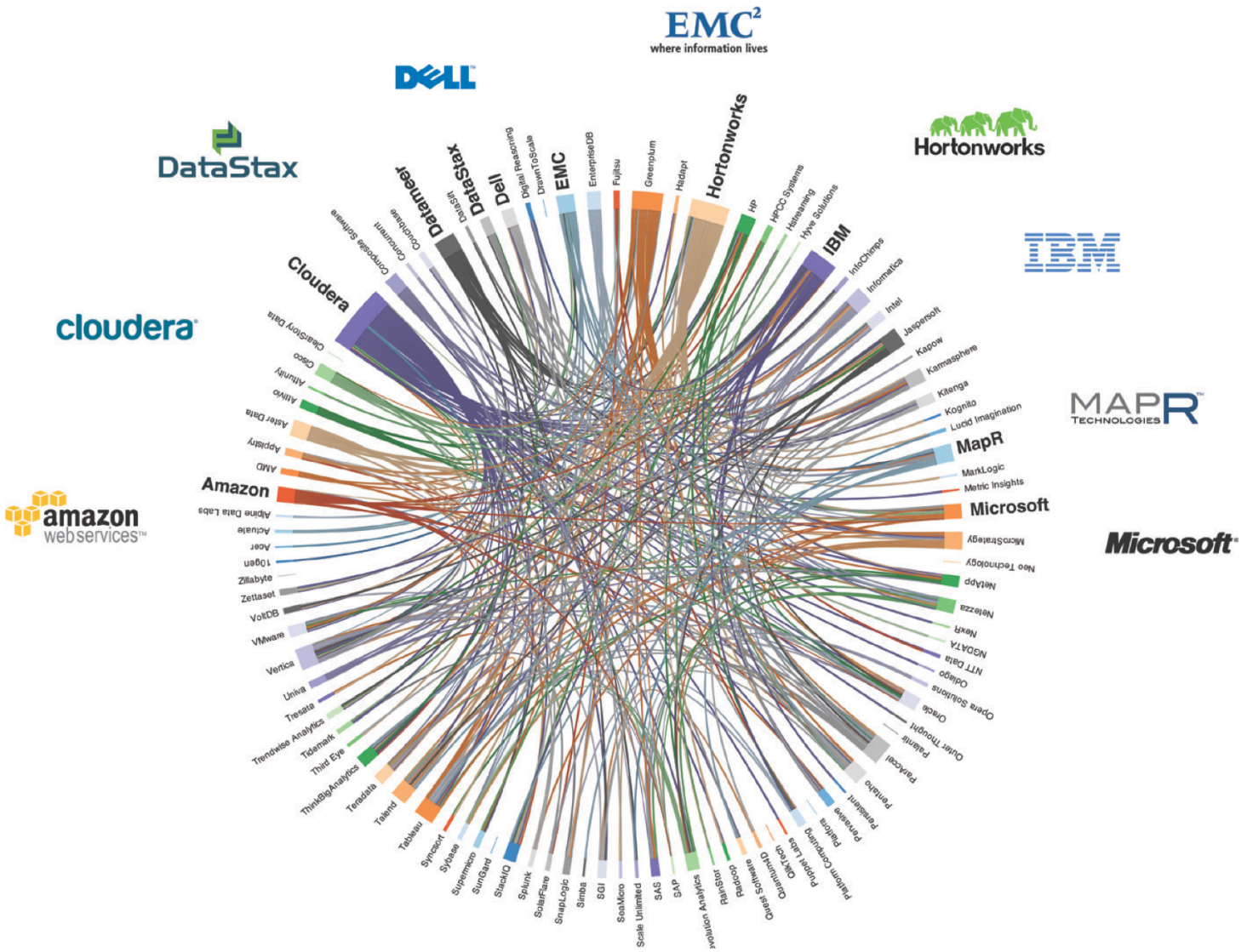
Hadoop MapReduce Lifecycle (MRv1)



The Big Data Architecture



Hadoop Ecosystem Components as Visualized by Datameer



Hadoop ecosystem components as visualized by Datameer.

Summary: Part I

- Big Data
  - IBM's 4Vs : *Volume*, *Velocity*, *Variety*, *Veracity*
  - Big data at rest
  - Big data in motion
- NoSQL
- MapReduce Framework
- Hadoop Ecosystem

- Let's take a break! 15 min


Collaboration is everything!

Brainstorming

- Key Requirements
  - Issue #1: Raw Data Gathering/Construction
    - 날씨, 행사, 매출(지역, 고객층), 카드사, 통신사
  - Issue #2: Business Logic, Decision 기준
    - Prototype 개발, 데이터 구성, 가중치 알고리즘 (sampling, ranking)
    - Action: customized (국내 현실에 맞는 모델이 필요)
    - Measure: 매출 이력 정보 활용
- Business Models
  - Big Data for CRM
    - 프렌차이즈: 고객 성향 분석
  - Twitter API

ETRI Electronics and Telecommunications Research Institute

Part II: Predictive Policing 101

Preventing Crime with Data and Analytics

04 August 2015

Sung-Soo Kim

sungsoo@etri.re.kr

Data Management Research Section

ETRI


THE LION KING
THE LION KING
PEACE
TARZAN
LG
BORN IN THE GLAMOROUS USA
Juicy Couture
DEAL OR NO DEAL
Play only at
TOYS R US
Play for a chance to win
$1,000,000
or a trip to the show or over
100,000 other amazing prizes
Play only at TOYS R US
Check this weekend's Toys"R"Us newspaper insert for a game piece.
Virgin mobile
CHALLENGE:
WATCHING OTHERS PASS YOU
asics
ONE WAY
swatch

References

1. Craig D. Uchida et. al, Data-Driven Crime Prevention - New Tools for Community Involvement and Crime Control, Justice & Security Strategies, Inc., 2013.
2. Walter L. Perry et.al, Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations, Research Report of RAND Safety and Justice Program, 2013.
3. Jennifer Bachner, Predictive Policing; Preventing Crime with Data and Analytics, Improving Performance Series, IBM Center for The Business of Government, 2013.
4. Ryan Gale, An Application of Risk Terrain Modeling to Residential Burglary, TCNJ Journal, 2013.
5. James Byrne, A History of Crime Analysis, Technology and the Criminal Justice Program (44.203 Course Note), UMass Lowell, Fall, 2012.
6. Mohler, G.O., M.B. Short, P.J. Brantingham, F.P. Schoenberg, and G.E. Tita, Self-exciting point process modeling of crime. Journal of the American Statistical Association 106(493):100-108, 2011.
7. Erik Lewis et. al, Self-exciting point process models of civilian deaths in Iraq, Security Journal, 2011.
8. Leslie Kennedy, Joel Caplan and Eric Piza, Risk Clusters, Hotspots, and Spatial Intelligence: Risk Terrain Modeling as an Algorithm for Police Resource Allocation Strategies, Journal of Quantitative Criminology 27, pp. 339–362, 2011.
9. Joel Caplan, Leslie Kennedy, and Joel Miller, Risk Terrain Modeling: Brokering Criminological Theory and GIS Methods for Crime Forecasting, Justice Quarterly 28:2, pp. 360–381, 2011.
10. Graham Farrell and William Sousa, Repeat Victimization and Hot Spots: The Overlap and its Implications for Crime Control and Problem-Oriented Policing, Crime Prevention Studies 12, pp. 221–240, 2001.
11. Mike Egesdal et. al, Statistical and Stochastic Modeling of Gang Rivalries in Los Angeles, SIAM Journal, 2010.
12. AUSTIN C. ALLEMAN, GEOGRAPHIC PROFILING THROUGH SIX-DIMENSIONAL NONPARAMETRIC DENSITY ESTIMATION, Technical Report, DEPARTMENT OF MATHEMATICS AND COMPUTER SCIENCE, SANTA CLARA UNIVERSITY, 2012.
13. P Jeffrey Brantingham, Prey selection among Los Angeles car thieves, Crime Science 2013.
14. George Mohler, MODELING AND ESTIMATION OF MULTI-SOURCE CLUSTERING IN CRIME AND SECURITY DATA, Annals of Applied Statistics, 2013.
15. George Mohler, ESTIMATING THE HISTORICAL AND FUTURE PROBABILITIES OF LARGE TERRORIST EVENTS, Annals of Applied Statistics, 2013.

Outline

- Why is the Predictive Policing Important?
- Background
- Mathematical Frameworks
- Data Used in Predictive Policing
- Predictive Methodologies for Predictive Policing
- Summary

Why is the Predictive Policing Important?

Why is the Predictive Policing Important?

- What is Crime Analysis?
  - Administrative, Tactical and Strategic Problem Solving
  - Forecasting and predictive analysis

August Vollmer
Source: LAPD Web Site

1909

- Principles of problem-oriented policing
- Evidence-based policing
- Real-time crime analysis
- Intelligence-led policing and other proven policing models

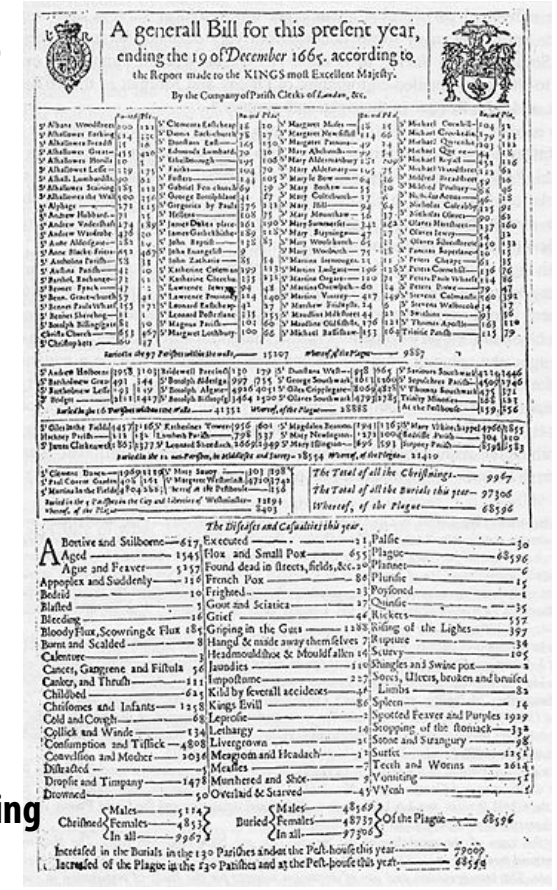
Orlando W. Wilson

1950

History

- *Origin of Crime Analysis*
  - Crime Mapping: 1829, Italian geographer and French Statistician
  - Crime Analysis: 1842, London's Metropolitan Police

- *First Data Miner*
  - John Graunt [Amateur Scientist]
  - Bill of Mortality (1st: 21 December 1592)
    - essentially random set of information
    - the study of the patterns causes and effects of disease
    - *Plague*
      - caused by person-to-person contact
      - tended to increase during the first year of the reign of a new king



A generall Bill for this present year, ending the 19 of December 1665. according to the Report made to the KINGS most Excellent Majesty.

By the Company of Parish Clerks of London, &c.

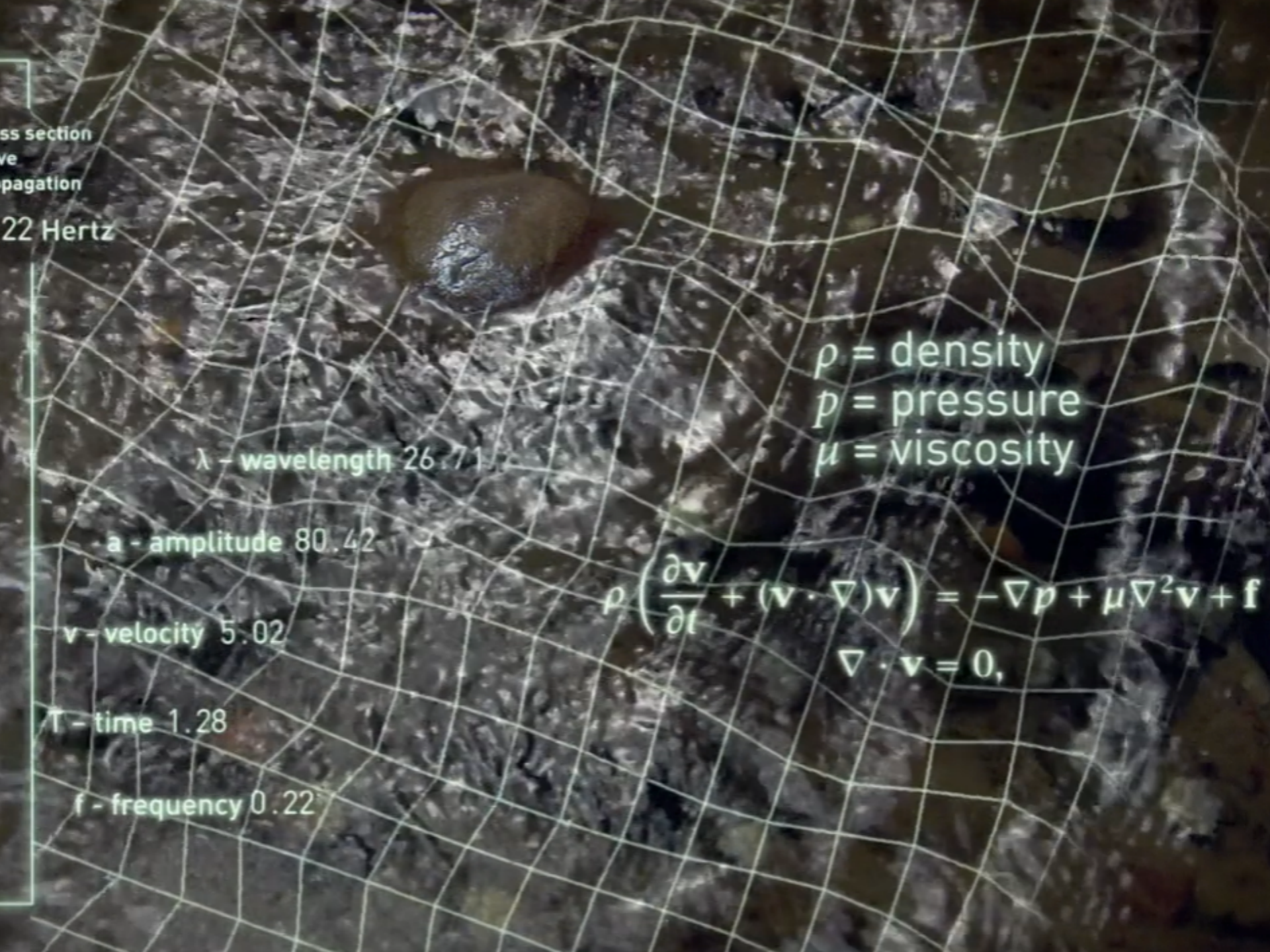
The Diseases and Casualties this year.

Abortive and Stillborne	617	Executed	21	Palsie	50
Aged	1545	Flox and Small Pox	655	Plague	68596
Ague and Feaver	5257	Found dead in streets, fields, &c.	20	Planet	6
Appoplex and Suddenly	116	French Pox	86	Plurisie	15
Bedrid	10	Frighted	13	Poysoned	1
Blasted	5	Gout and Sciatica	17	Quinsie	35
Bleeding	16	Grief	46	Rickets	557
Bloody Flux, Scowring & Flux	185	Griping in the Guts	1288	Riding of the Lights	397
Burnt and Scalded	8	Hangd & made away themselves	7	Rupture	34
Calenture	3	Headmouldshot & Mouldfallen	14	Scurvy	105
Cancer, Gangrene and Fistula	56	Jaundice	110	Shingles and Swine pox	2
Canker, and Thrush	211	Imposthume	227	Sores, Ulcers, broken and bruised	
Childbed	625	Kild by several accidents	54	Limbs	82
Chrisomes and Infants	1258	Kings Evill	86	Spleen	14
Cold and Cough	68	Leprosie	2	Spotted Feaver and Purples	1929
Collick and Winde	134	Lethargy	14	Stopping of the Stomack	332
Consumption and Tisick	4808	Livergrown	31	Stone and Strangury	98
Convulsion and Mother	2036	Megrom and Headach	13	Surfet	1251
Distracted	4	Measles	7	Teeth and Worms	2614
Dropsie and Timpany	1478	Murthered and Shot	9	Vomiting	51
Drowned	50	Overlaid & Starved	45	Wenn	5

Christned	Males 5114	Buried	Males 48363	Of the Plague	68596
	Females 4853		Females 48733		
	In all 9967		In all 97306		

Increased in the Burials in the 130 Parishes and at the Pest-house this year 79009

Increased of the Plague in the 130 Parishes and at the Pest-house this year 68558


ss section
ve
pagation
22 Hertz
λ - wavelength 26.71
a - amplitude 80.42
v - velocity 5.02
T - time 1.28
f - frequency 0.22
ρ = density
p = pressure
μ = viscosity
$\rho\left(\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v}\cdot\nabla)\mathbf{v}\right) = -\nabla p + \mu\nabla^2\mathbf{v} + \mathbf{f}$
$\nabla \cdot \mathbf{v} = 0,$

Clustering Patterns

- Loma Prieta Earthquake (1989)
  - Earthquake analysis
  - Clustering patterns
- MASC Project
  - Clustering patterns are also seen in *crime data*
  - Future events *nearby* in *space* and *time* after shocks of crime
  - Finding and clustering the *patterns* from chaos human behaviours
  - *Self-exciting point* process model *similar* to that used in earthquake analysis
  - Pilot Project : Los Angeles
    - Data: 13 million past crimes
    - Duration: past 80 years

UC MASC PROJECT
MATHEMATICAL AND SIMULATION MODELING OF CRIME

Mathematical and Simulation Modeling of Crime

UC MASC PROJECT
MATHEMATICAL AND SIMULATION MODELING OF CRIME

Jeff Brantingham
UCLA Anthropology

George Mohler
Santa Clara Mathematics

George Tita
UCI Criminology, Law and Society

Mathematical Framework [7]

- **The Hawkes Process Model (1974)**
  - any given event, or collection events can be causally linked to a background Poisson process and foreground self-exciting process.

Self-excitation

the distribution of crimes following an initial event

$$\lambda(t) = \mu + k_0 \sum_{t_k < t} g(t - t_k)$$

background rate of events
stationary Poisson process

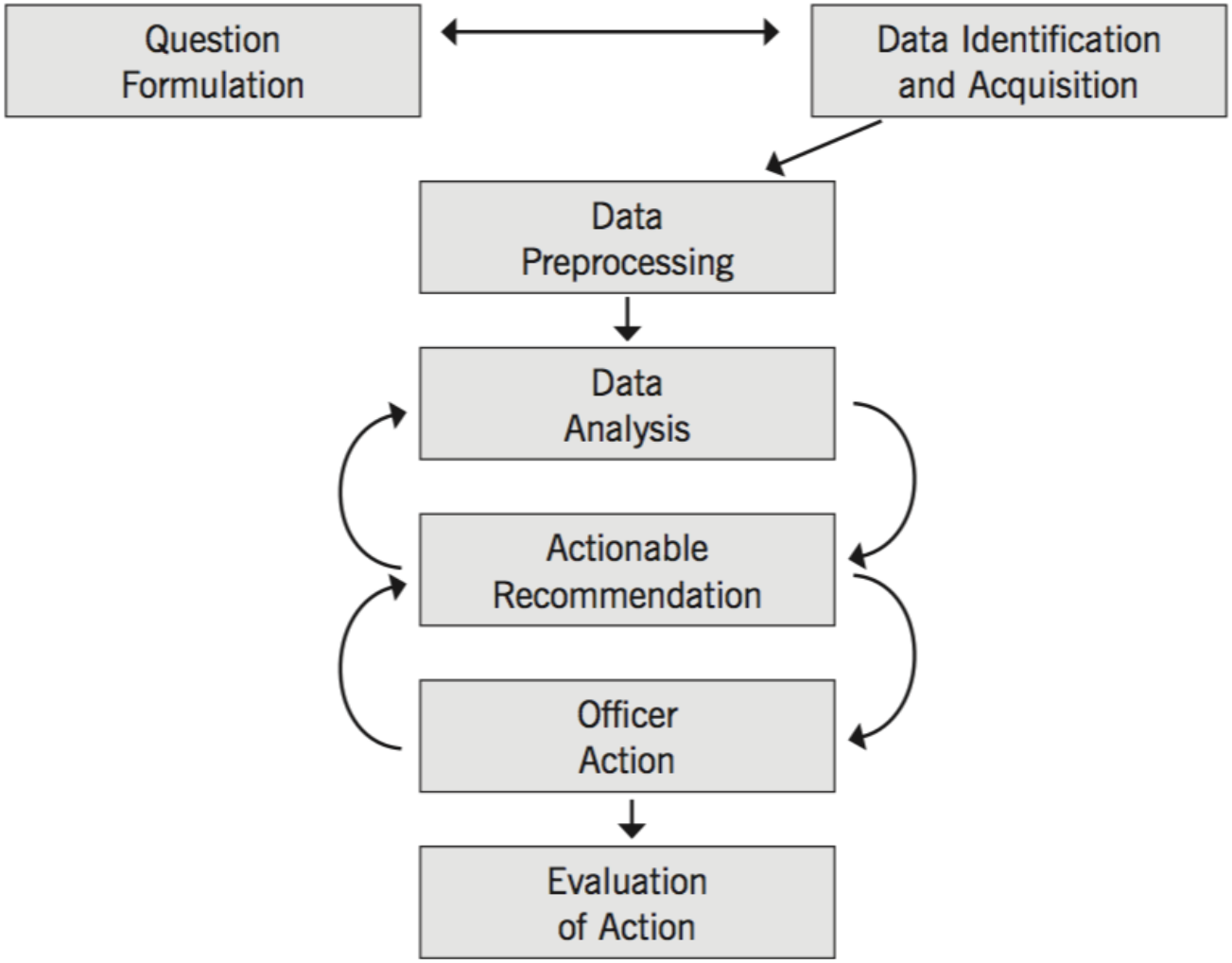
the density of prior events
necessary to trigger excitation

how much excitation is generated
by a collection of prior events

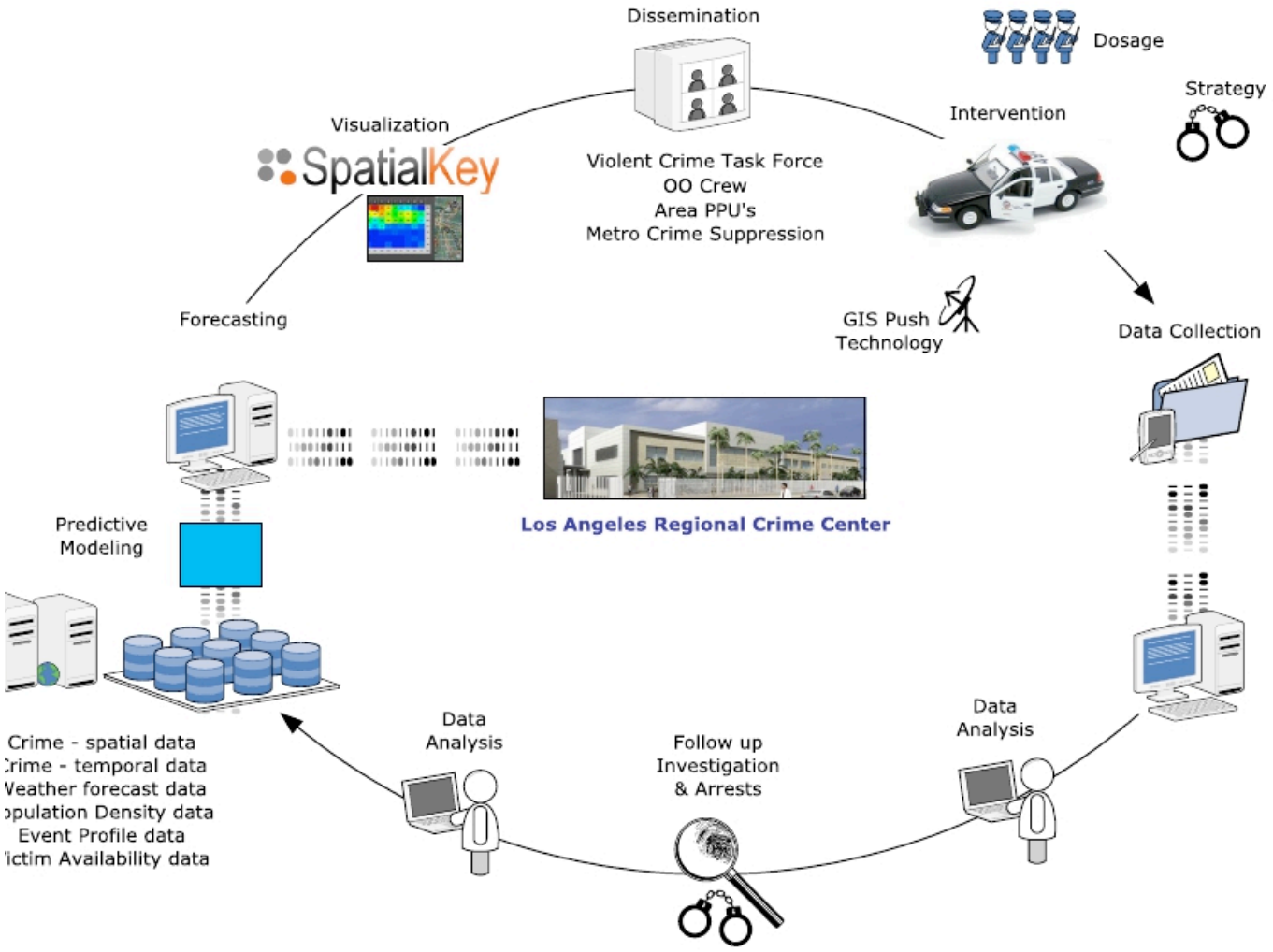
[7] Erik Lewis et. al, Self-exciting point process models of civilian deaths in Iraq, Security Journal, 2011.


BURGLARY

Operational Challenges of Predictive Policing



Case Study: Los Angeles Regional Crime Center



Data Used in Predictive Policing

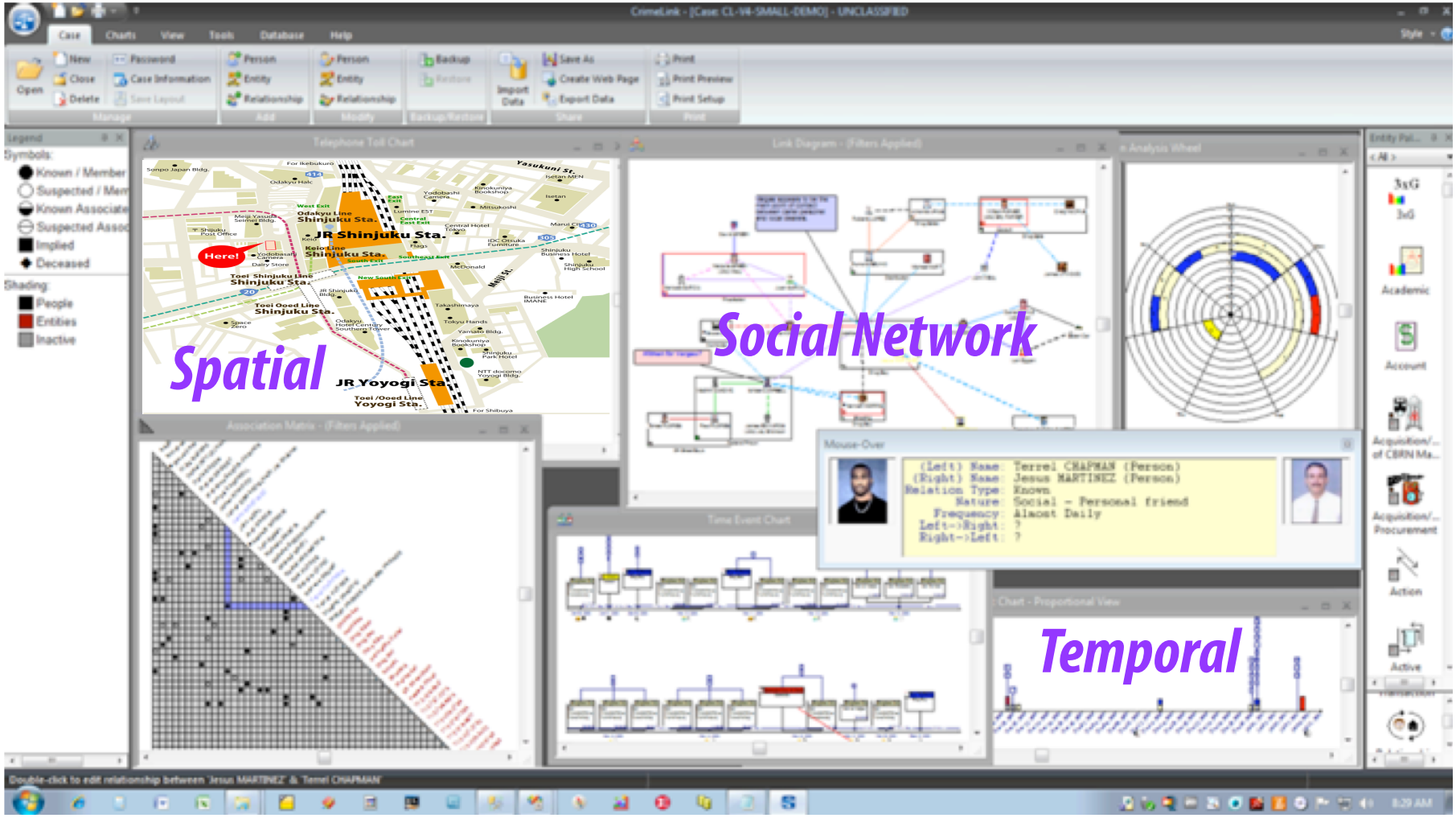
- **Three Categories**
  - **Spatial**
  - **Temporal**
  - **Social Network**

Spatial Variables	Temporal Variables	Social Network Variables
Indicators of Areas with Potential Victims/Targets • Shopping malls • Property values • Hotels • Area demographics • Population density • Residential instability **Indicators of Escape Routes** • Highways • Bridges • Tunnels • Public transportation • Railways • Dense foliage **Indicators of Criminal Residences** • Bars and liquor stores • Adult retail stores • Fast food restaurants • Bus stops • Public health information • Areas with physical decay	• Payday schedules • Time of day • Weekend vs. weekday • Seasonal weather (e.g., hot versus cold weather) • Weather disasters • Moon phases • Traffic patterns • Sporting and entertainment events	• Kinship • Friendship • Affiliation with an organization • Financial transaction • Offender/victim

The Transition from Data to Knowledge in a Police Agency

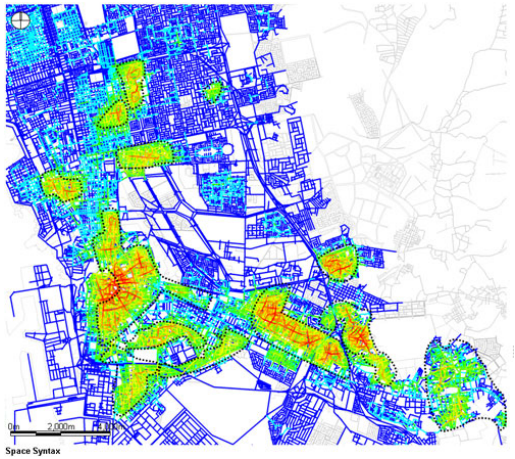
Data		Information		Knowledge		Result
Individual Incident Reports in a records management system	Analysis	Six of the reports are related in a series of robberies	Communication	Robbery series is prime topic of discussion in next detective's meeting	Strategy & Action	Robbery offender is apprehended
Statistics showing number of officers per capita throughout the state		Your police department has 20% fewer officers per capita than average		Chief has this information in mind when preparing his budget proposal		Agency is granted additional officers by city
Crime volume of current year compared to past years; individual records in RMS; jurisdictional information		Auto theft is up 20%, with most of the increase in Police Beat 5 on the midnight shift, probably influenced by new sports arena		Officers internalize this information and consider it when patrolling Beat 5		Auto theft is reduced

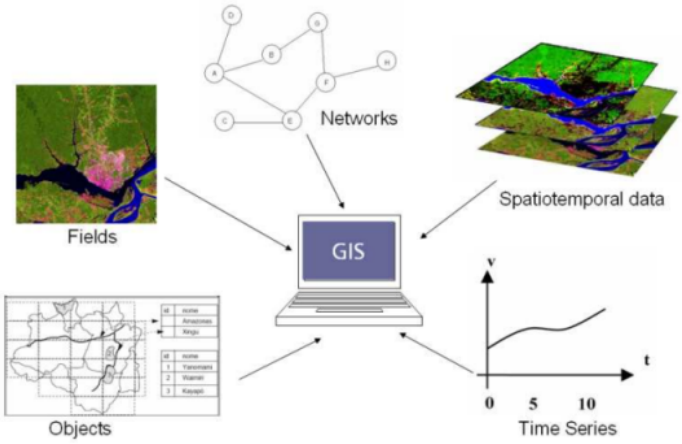
Full Scale Analysis

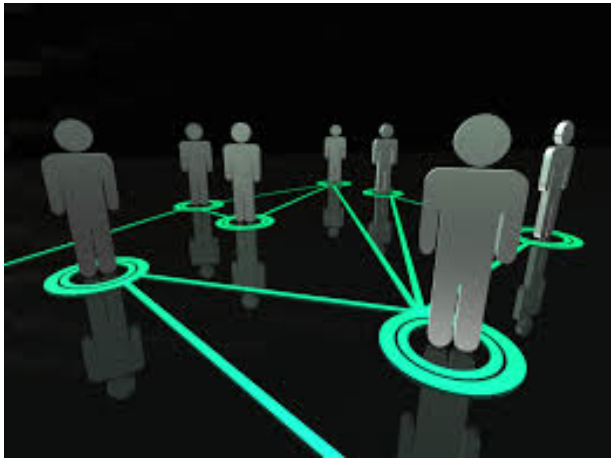


Predictive Methodologies

- Analysis of space
- Analysis of time and space
- Analysis of social networks

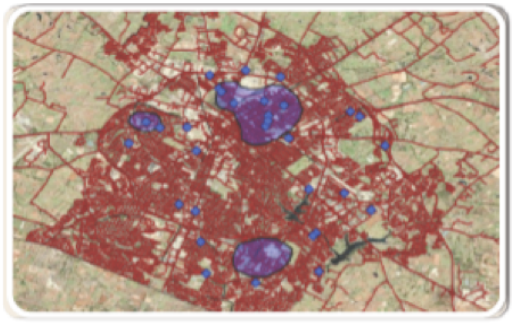




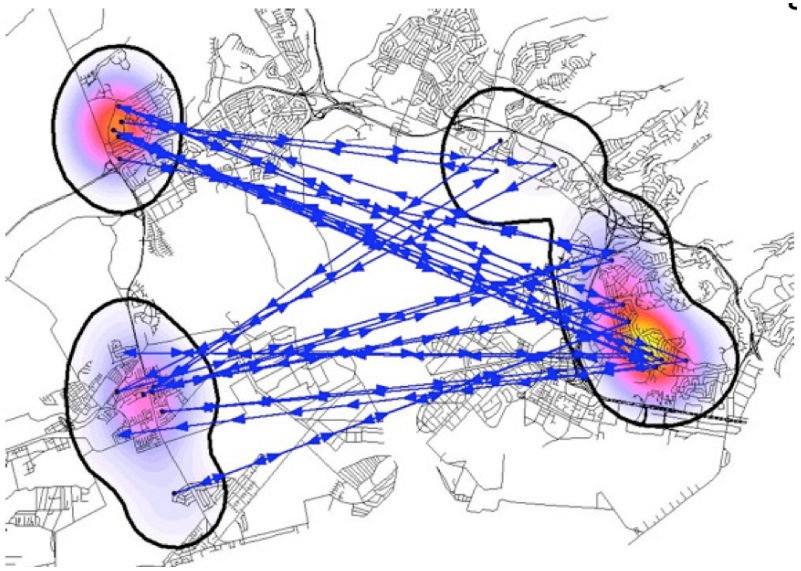


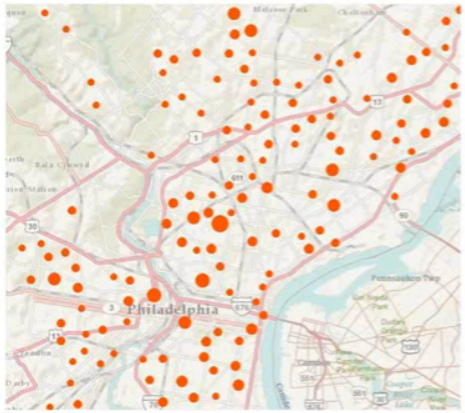
Predictive Methodology One: Analysis of Space

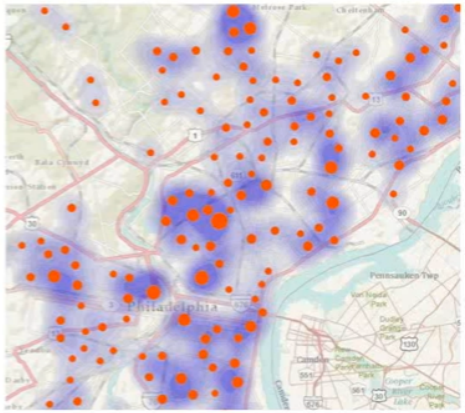
- Point (or offense) locations
- Hierarchical clusters
- Partitioned clusters
- Fuzzy clusters
- Density mapping
- Risk-terrain modeling (RTM) clusters



All crime related gun incidents







Predictive Methodology One: Analysis of Space

- **Point (or offense) locations**
  - ***Theory of repeat victimization, 500x500 feet [PredPol Software]***
- **Hierarchical clusters**
  - **use a nearest-neighbor technique [display the clusters: *ellipses*, *convex hulls*]**
- **Risk-terrain modeling (RTM) clusters**

Gun shootings example

Gun shootings example

Predictive Methodology Two: Analysis of Time and Space

Predictive Methodology Two: Analysis of Time and Space

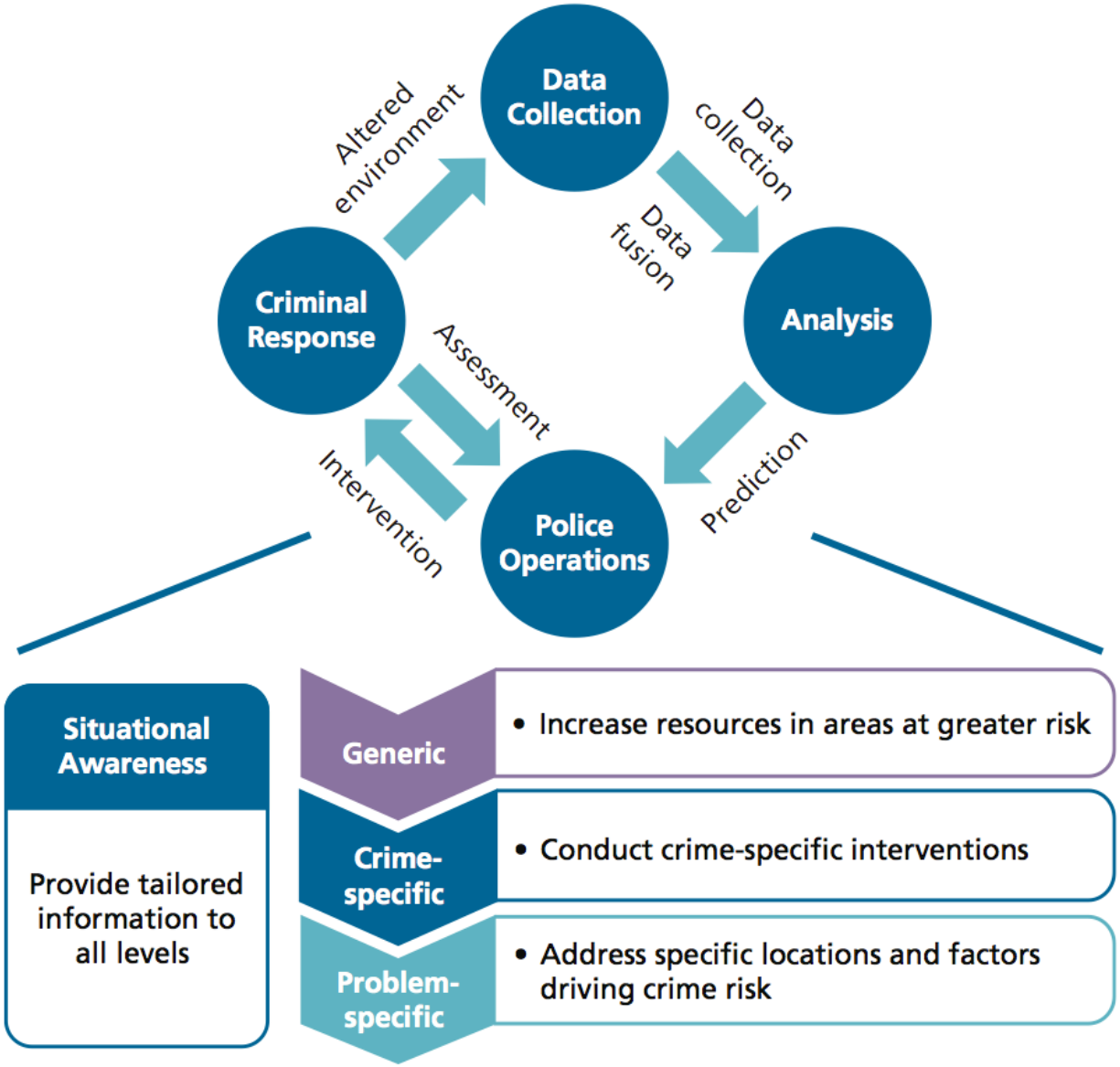
- **CrimStat III : a software program (sociologist + National Institute of Justice)**
  - **spatial-temporal moving average (STMA)**
    - **the *average time and location* for a subset of incidents**
  - **correlated walk analysis (CWA) : temporal and spatial relationships between incidents**
    - **computing the correlation between *intervals* [time, distance, direction between two events]**

Predictive Methodology Three: Analysis of Social Networks

- **Social Network Analysis (SNA) : *cutting edge* of crime analysis**
  - **detect *persons of interest*, as opposed to *locations of interest***
  - **identify individuals that are central to criminal organizations (eg., gangs and drug networks)**

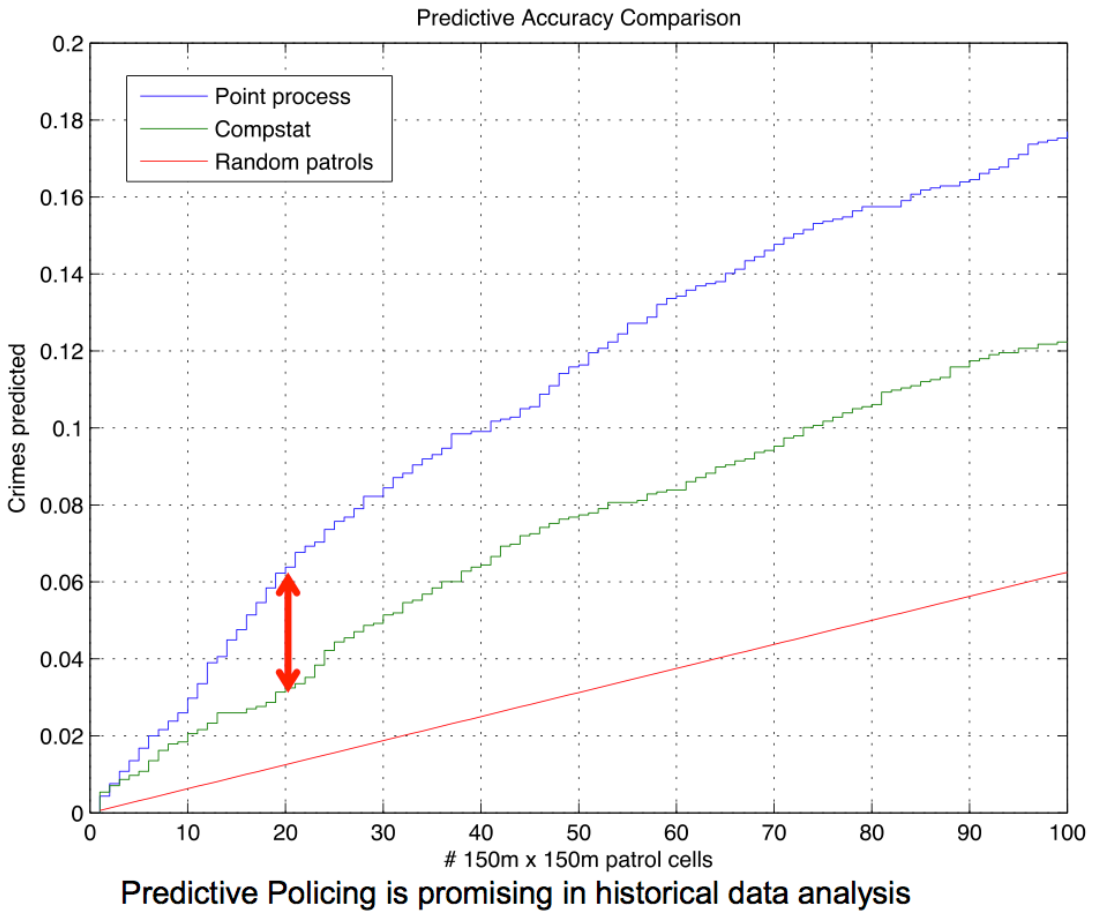
- **Building blocks of a social network: *relationships* between two actors**
- **In crime-fighting applications,**
  - **SNA is frequently used to identify central nodes [high level of connectivity]**
  - **Measures of *centrality***
    - ***degree* : the number of links possessed by a node**
      node's level of connectedness
    - ***closeness* : the total distance from a node to all other nodes in the network**
      ease of obtaining information from the network
    - ***betweenness* : the number of instances a given node appears in the shortest path between other nodes**
      relevance to the passage of information within the network

Review: Prediction-Led Policing Business Process [2]



Places on the Frontier of Predictive Policing in the United States

- *Los Angeles, California*
- Santa Cruz, California
- Baltimore Country, Maryland
- Richmond, Virginia
- Memphis, Tennessee



Predictive Policing is promising in historical data analysis

Summary: Part II

- Predictive Policing *Concept*
- *Mathematical Frameworks* for Predictive Policing
- *Data* Used in Predictive Policing
- *Predictive Methods* for Predictive Policing
- Predictive Policing *Business Process*


Thank you