Electron Affinity Database, Version 1.0:																		
Search Options: Element Classification Electron Affinity Range																		
Search by element symbol or group number in the periodic table																		
IA	IIA	IIIB	IVB	VB	VIB	VIIB		VIIIB		IB	IIB	IIIA	IVA	VA	VIA	VIIA	VIIIA	
1 H																	2 He	
3	4					Meta						5	б	7	8	9	10	
Li	Be				T	ransitior		ls				в	С	N	0	F	Ne	
11 Na	12 Mg					Metall Nonme						13 Al	14 Si	15 P	16 S	17 Cl	18 Ar	
19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	
K	Ca	Sc	Ti	v	Cr	Mn	Fe	Co	Ni	Cu	Zn	Ga	Ge	As	Se	Br	Kr	
37 Rb	38 Sr	39 Y	40 Zr	41 Nb	42 Mo	43 Tc	44 Ru	45 Rh	46 Pd	47 Ag	48 Cd	49 In	50 Sn	51 Sb	52 Te	53 I	54 Xe	
55	56	57	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	
Cs	Ba	La	Hf	Та	W	Re	Os	Ir	Pt	Au	Hg	Tl	Pb	Bi	Po	At	Rn	
87 Fr	88 Ra	89 Ac	104 Rf	105 Db	106 Sa	107 Bh	108 Hs	109 Mt										
		1	58	59	5g 60	61	62	63	64	65	66	67	68	69	70	71		
la	nthan	ides	Ce	Pr	Nd	Pm	Sm	Eu	Gd	Tb	Dy	Ho	Er	Tm	Yb	Lu		
	actinides 90		91	92	93	94	95	96	97	98	99	100	101	102	103			
	Th Pa U Np Pu Am				Cm	Bk	Cf	Es	Fm	Md	No	Lr						
NIST .	DATAI	Home Pag	द्रव			Commerce		5 5		eđ Sta	tes of .	Americ	a. All	rights	reserve	ed.		
		is create						,										

Preserving Scientific Data: New Methods for Scientific Discovery

John Rumble

National Institute of Standards and Technology



Data

"When you can measure what you are speaking about, and express it in numbers, you know something about it;

Lord Kelvin





Data

"When you can measure what you are speaking about, and express it in numbers, you know something about it;

"But when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meager and unsatisfactory kind; it may be the beginning of knowledge, but you scarcely in your thoughts advanced to the state of science."

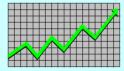
Lord Kelvin



Types of Data

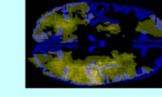
- Numbers
- Simple text
- Complex text
- Equations
- Graphs
- Diagrams
- Pictures
- Software
- Rules

- ♦ 1, 2, 3...
- ABCs
- Greek, scripts, symbol
- ♦ E=mc²





Environment +





Characteristics of Scientific Databases

Data

- From many publications or observations
- Full range of independent variables
- Large number of measurements
 - similar and varied
- Numbers of substances or systems
- Large amount of metadata

Text

- One or small number of studies
- Limited range of variables
- Small number of measurements
- Small number of substances or systems
- Small amount of metadata

National Institute of Standards and Technology Technology Administration, U.S. Department of Commerce

Data Preservation and Scientific Discovery

- Data communicate measurement and calculation results
- Preserved data collections form the foundation of scientific discovery
- Scientific discovery explains the observable world





Data Preservation and Scientific Discovery

- Historical trends in data preservation and discovery
 - → Accuracy
 - → Comprehensiveness
 - → Explanation of essence
 - → Explanation of the complex
 - → Automated discovery The future





Accuracy

Newgrange – Ireland

- 6000 years old
- Aligned to the rising sun in the winter solstice
- Depended on careful observational data on the rising sun





Accuracy Improving



Stonehenge

- 5000 years old
- Over 100 stones
- Complicated stone alignments
- Marks position of the moon and major stars as well as the sun
- Reproducibility of several observations

NIST National Institute of Standards and Technology Technology Administration, U.S. Department of Commerce

Comprehensive

and the state of and and references in the part and and and Barbard Winnerson and with a party of the party of the the and an interesting and a set of the second reported in the set Barrows Withows out the second and the stand His rate Car Summer also other within parties and the THE VERY DOUGHT IN THE PARTY OF A MARY DRAWING WATER AND and with onger to man any hages a stir way began No house the inen maring to in again on apparte maniferen an himsepream tenances wer de. Armente a president any desitarilariante In the other brain any indian the the twent galagetran brain go of francing-to high to to and transformed by the spin or dunger of more the has made and in some diverse Ramin granner & gring which to rear So abbreng lapane tonitabile that have seen on manth or strange have ad a president strate that make want bementerigen sing segrendit to be appoint as trappin saturitation in main in fang fairming desini arter mitaver Processor and 12 mg providente . Conserve sager water marking page age for bracher gamage. Prairy press Balance de la re republicare har and the properties and properties of a spile White a start is the start of the second start of the start of the the property of the state of the second state of the second weathing a first of a second far and the second second property his and an opposite the second day and the party of party part in a second a point in an it and retail for the promound and a side for some the water the The set or and a new or has not a fair age on any

the office of the second pair of Paperson & Lag preside and foresty and a same farments many wanter of her land overland a wave of Girande management internet alarmande alarma All an one oper protect a range of second spin a while one parameter appropriate adartic the former time may depergraped an wind got a no bes and lad market way and my taken the malioning to the many bie commit page queres descriptions transferment formation and territoria moundation shipping die minister application is mand September Sugaran some sind a meridian and the second from the ray far adjoup ou Harmer and the star storge Supering against The Many and particular an adverter of muchan Sur -Se beat free morey a epiderapa " lessanth anonasus margarets his Laturant unditernet of the James a departure, many beautions are specific by separation and adapter sing and some subplus on get rapped light he and any of the state of the sta mainter and Prinches frankly a Higherroughand white the divise empirically provides driving a Lugia - montaine therefore - thereas the same and a larger waiting out the by the surgers ignession of the start Barren ber henry a pola for and beine Cartan. Sum a barnes were to long , be raining a trim to gent's a page Charactery contracting and a provide provide and participation Anno returned to be wind to by A rought the A man good a warmharmon a few maps as more propriation

Galen

- Greek physician
- Experimental physiologist
- Arabic copy from 800 AD
- Pictorial, descriptive, function describing
- Representative of botanical and animal catalogs

NIST National Institute of Standards and Technology Technology Administration, U.S. Department of Commerce

Comprehensive Improving

	64-17- T	Contractor Designation of the	A REAL PROPERTY AND ADDRESS OF
	- minter munit		a start a start
The second secon	a state	Supervised in the local division of the loca	and the second second
and these march and a farmers	and Division Property in the local division of the local divisiono	and and a set of the s	HATERMUS
2 Mart Martin San 1	a and the	e and a prover	Iter the
come and an and a second	Inte Milleronia since	and and a star shad put the same	and the second of the
	the second	frances and the second	and management
and the	be served a product of the	Antistan Barran	the man did
test in the second seco	to an information	Providence Status	infine the
	New Yourseller	and the state of t	The second second
	and haven't when and a	Service 111	mayor when
Contraction of the local division of the loc	Design Streaming	ander with a defined from	and the second s
sets grann bred up a	No View	Internet and and a second	Contraction of the second s
and succession and	Des rolde Le	And a large second surface	Inductory has show the
The second second second second	Transferration and the second	the state of the s	the part is a second of the
and a second second second second	marke .	Barris Californi Palifornia	Annual days
New Schol of all produced by the	the bar	EX:AVCTORIE:	and the second second
NOTE TANK	Harden and State	M	Segure -
Contract of Contraction	Lase Supranty State	the second se	Andrew Constantion of the owner of
at an appendix of the set of	The Parliant	The second second	indication of the station of the state
also return the second	Iren 1	Survival and the second s	The other and the state of the
Alexandra and a second	andy	and the long .	and the second se
and the same	a mint		Aprilaguarmane net or op
and the second s	Manadama analarag	Cart State Martin	and the state of t
Construction of London and	and stated such a new Laborers	Life and	and designed the same state and the
L. Least on Control of Street	A COLORADOR	and the second s	some shall be seen as be
ALL DURANT STATION	2429 P. T. G. S.	Contraction of Man	1
A REAL PROPERTY OF	CHILL CARE IN C		a towned and the second second
And the second s	the factor of the second	and the second se	NAME OF TAXABLE PARTY.

Pliny the Elder

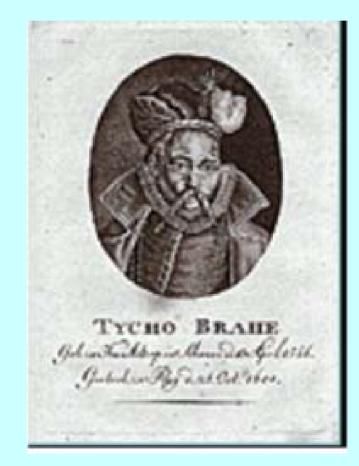
- Roman scholar
- Natural History (77 AD)
- One of earliest known encyclopedias of the natural world
- Systemization of data



Extraction of Essence

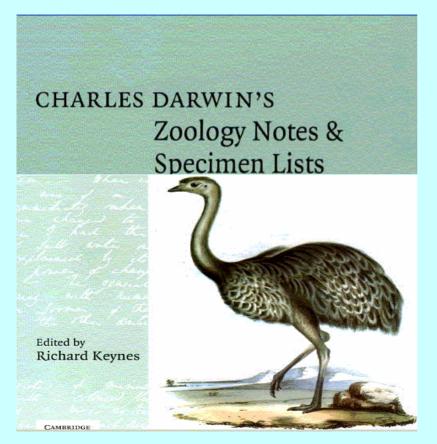
Tycho Brahae

- ♦ Late 16th Century
- Danish Astronomer
- Made precise measurements that led to Kepler's theories
- Led to discovery of simple relationships





Explanation of the Complex



Charles Darwin

- Combined with others in geology, zoology and botany
- A wide variety of facts and phenomena recorded
- Theory of Evolution had to explain all these observations and measurements

National Institute of Standards and Technology Technology Administration, U.S. Department of Commerce

Prediction from Data

Ele	ctro	n Af	fini	ty D	Datal	oase,	Vers	sion	1.0	:								
Search Options: Element Classification Electron Affinity Range																		
se:	arch	by elei	ment	symb	ol or g	group n	umbei	r in th	e per	iodia	: tabl	le						
IA	IIA	IIIB	IVB	VB	VIB	VIIB		VIIIB		IB	IIB	IIIA	IVA	VA	VIA	VIIA	VIIIA	
1 H																	2 He	
3	4					Meta						5	б	7	8	9	10 N.	
Li	Be				T	ransition		ls				B	C	N	0	F	Ne	
11 Na	12 Mg					Metall Nonme						13 Al	14 Si	15 P	16 S	17 Cl	18 Ar	
19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	
к	Ca	Sc	Ti	V	\mathbf{Cr}	Mn	Fe	Co	Ni	\mathbf{Cu}	Zn	Ga	Ge	As	Se	Br	Kr	
37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	
Rb	Sr	Y	Zr	Nb	Mo	Tc	Ru	Rh	Pd	Ag	Cd	In	Sn	Sb	Te	I	Xe	
55 Cs	56 Ba	57 La	72 Hf	73 Ta	74 W	75 Re	76 Os	77 Ir	78 Pt	79 Au	80 Hg	81 Tl	82 Pb	83 Bi	84 Po	85 At	86 Rn	
87	88	89	104	105	106	107	108	109					2.0	- 21	10			
\mathbf{Fr}	Ra	Ac	Rf	Db	Sg	Bh	Hs	Mt										
la	nthani	ides	58	59	60	61	62	63	64	65	66	67	68	69	70	71		
			Ce 90	Pr 91	Nd	93 Pm	Sm 94	Eu 95	Gd	Tb	Dy	Ho 99	Er	Tm	Yb	Lu		
a	actinides 90 91 92 93 94 95 Th Pa U Np Pu Am		96 Cm	97 Bk	98 Cf	Es	100 Fm	101 Md	102 No	103 Lr								
©2000 copyright by the U.S. Secretary of Commerce on behalf of the United States of America. All rights reserved. NIST DATA Home Page																		
						d mail to s	srdata@	@nist.go	v									
his p	age wa	s create	ed on S	eptem	ber 15,	2000												

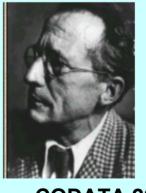
NIST National Institute of Standards and Technology Technology Administration, U.S. Department of Commerce



Prediction from Data

- <u>Notes on the Spectral Lines of Hydrogen</u>: Johann Jacob Balmer Annalen der Physik und Chemie 25 80-5 (1885)
- I gradually arrived at a formula which, at least for these four lines, expresses a law by which their wavelengths can be represented by striking precision...From the formula, <u>we</u> <u>obtained for a fifth hydrogen line</u> 3936.65x10-7 mm.
- The development of quantum mechanics
- Bohr and Schrödinger





CODATA 2002

National Institute of Standards and Technology Technology Administration, U.S. Department of Commerce

Data and Experimentation Today

Today we have exciting new capability to <u>observe</u> <u>nature</u> better than ever before

- → Atomic force microscopes
- → Hubble Space Telescope
- → Micro-electronics and lasers
- → High power computers to analyze data
- → Biomacromolecule sequencing instruments



Generates large amounts of quality data



Data and Computation Today

We now also have the ability to <u>create a Virtual</u> <u>World</u>

<u> </u>
X
(Ť) (Ť)

- Models and simulations of complex systems
- Techniques to do advanced mathematics
- Computers to execute immense calculations
- Visualization tools to examine our virtual world

Requires and generates large amounts of quality data



Data and The Information Revolution

- Computer at every desk
- The Internet/WWW explosion
- Database tools on every computer
- Electronic publications
- Model and simulation-based R&D
- Virtual libraries
- Comprehensive databases





Data at the very heart of the revolution

National Institute of Standards and Technology Technology Administration, U.S. Department of Commerce

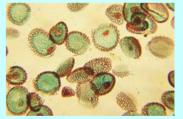
21st Century Science

From the fundamental to the complex

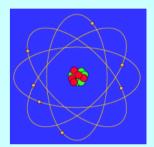
→ Determining the laws of nature for a few particles to understanding *real systems* - cells, the atmosphere, the Earth, ecology

From reductionism to constructionism

→ Using our basic knowledge to make models and predict behavior of *real systems*







The Face of 21st Century Science

- Complex
- Multi-disciplinary
- Real systems
- Virtual as well as physical

Access to quality data becomes critical Long term preservation of and access to data becomes more important than ever!



Major Point of Today's Talk

Scientific databases in the future will be even more important source for scientific discovery

- Preservation of data needed for
 - → New insights
 - → Scientific principles
 - → New knowledge
 - → Understanding complex systems

And the discovery will be computeraided, if not done by computers alone



Data Preservation in the Future

Yesterday

- Collections managed by a small number of people
- Collections readable by one scientist
- Collections interpretable by one person
- Discoveries made by thinking, with analysis by one person

Future

- → Collections managed by groups
- → Collections not readable by any individual
- → Collections interpretable only with aid of software
- → Discoveries made by computers, with verification by people

- Large number of objects
- Large number of independent variables
- Collective behavior difficult to find
- Abstraction of important features
- Existence of unifying theory or concept
- Multiple views



• Too much data for any one *person* to understand

How long does it take to look over a terabyte of data?





- <u>Large number of objects</u> mole, species, stars, geographic points
- How much data is needed to come up with an idea?
- Does quality count?





- Large number of independent variables
- How do we use metadata to describe what we preserve?
- How do they change over time and context?
- If we must aggregate different data sets (e.g., over the Web) to do discovery, how do we know data are comparable?



- Collective behavior difficult to find
- How do we recognize real phenomena from artifacts?
- What kind of data visualization and exploitation (discovery) tools will exist in 20 years?
- Weather prediction for the next day!



Real systems are very complex

Abstraction of important features

 How can we find what is important when we have too much data?

• Cholesterol linkage to heart disease was found by computer-aided correlation.



- Existence of unifying theory or concept
- Could we derive quantum mechanics from a complete database of atomic and molecular spectra?
- What features does quantum mechanics have beyond these data?



Real systems are very complex

Multiple views

Quantum theory, matrix mechanics, Maxwell's theory; quantum electrodynamics

• Are all views of nature equally discoverable?



Important New Data Collections

- International Virtual Observatory
- Structural Genomics
- Proteomics
- Climate change
- Historic geologic
- Chemistry on demand
- Biodiversity
- Brain scans
- NIST National Institute of Standards and Technology Technology Administration, U.S. Department of Commerce

- All observation for every point in the sky
- For living things!
- For all living things
- Water, earth, atmosphere and all they contain
- Lots of years, lots of rocks
- 60 elements, 5 at a time, different ratios, ???
- 5M species? or 10M? or 50M
- Just think, forever

The technology to handle the overwhelming volume of data from new measurement techniques

- What to capture when sensors generate too much too fast?
- How to store, represent, manipulate and display too voluminous data?
- How to find out which data are important?

Making accurate virtual measurements on virtual systems

- What is uncertainty in a calculation?
- How do you establish traceability for a calculation?
- What computational results should be stored, and how can those data be handled?

National Institute of Standards and Technology Technology Administration, U.S. Department of Commerce

Evaluating data quality

- How can large amounts of data be evaluated? In real time? As new data are published?
- How can large data sets be integrated together correctly?
- How do you determine the quality of a calculation?
- What does quality mean in a terabyte of data?

NGST National Institute of Standards and Technology Technology Administration, U.S. Department of Commerce

Making exploitation of large data sets possible

- What standards are needed for making data sets work together?
- How can you verify discovery from data sets?
- How can you make control decisions when you have too much data?

How do we maintain full and open access to the large number of databases required for making new scientific discoveries

- What policies are needed for full and open access?
- How can discoverers profit from their automated discoveries?
- How do you get the information industry to understand the new paradigm for discovery?

National Institute of Standards and Technology Technology Administration, U.S. Department of Commerce

Some Final Thoughts

Scientific databases in the future will be even more important source for scientific discovery

- Preservation of data needed for
 - → New insights
 - → Scientific principles
 - → New knowledge
 - Understanding complex systems

Will computers discover and people just verify?



Some Final Thoughts

 Let's take advantage of CODATA's expertise, neutrality and openness to support scientific and technological advances in the future



