

# Lebenslauf

## Dr. Werner Dubitzky

### Inhaltsverzeichnis

---

Inhaltsverzeichnis.....	1
Persönliche Daten.....	2
Karriere- und Profil-Zusammenfassung.....	2
Akademische Ausbildung.....	3
Beruflicher Werdegang.....	3
Motivation, Forschungsinteressen, Wissenschaftliches Leitmotiv.....	4
Motivation.....	4
Forschungsinteressen und Wissenschaftliches Leitmotiv.....	5
Drittmittel für Forschung.....	6
Gutachtertätigkeit.....	6
Wissenschaftskooperationen, Projekt- und Managementenerfahrung.....	7
Editorial Board.....	8
Editor / Herausgeber.....	8
Sammelbände und Enzyklopädien.....	8
Spezialausgaben wissenschaftlicher Zeitschriften.....	9
Organisation von Konferenzen und Workshops.....	10
Betreute Promotionen und Promotionskommission.....	10
Betreute Promotionen.....	10
Mitglied Promotionskommissionen.....	11
Lehrveranstaltungen.....	11
Schriftenverzeichnis.....	11
Begutachtete Veröffentlichungen in Fachzeitschriften.....	11
Begutachtete Veröffentlichungen in Konferenzberichtsbanden.....	14
Begutachtete Veröffentlichungen in Sammelbänden und Nachschlagewerken.....	18
Sonstige wissenschaftliche Veröffentlichungen.....	19
Anhang A: Some Research Interests in Machine Learning and Data Science.....	21
Statistical machine learning.....	21
Statistical learning as experimental science.....	22
Machine learning that matters.....	23
Machine learning for integration/analysis of complex biomedical data sets.....	25
Feature engineering and knowledge integration in machine learning.....	25
Intelligent e-biomedicine.....	25
Other AI / machine learning topics of interest.....	26

## Persönliche Daten

---

Name **Werner Dubitzky**  
Adresse: Genhofen 21, 88167 Stiefenhofen  
E-Mail: [werner@dubitzky.com](mailto:werner@dubitzky.com)  
Handy 0170 2231 654



Researcher ID [orcid.org/0000-0002-5111-8012](https://orcid.org/0000-0002-5111-8012)  
LinkedIn [URL](#)  
ResearchGate [URL](#)  
Google Scholar [URL](#)

## Karriere- und Profil-Zusammenfassung

---

Zur Zeit bin ich wissenschaftlicher Mitarbeiter (Data Analyst und Projektmanager) am [Helmholtz Zentrum München](#) tätig. Die folgende Liste zeigt einen Überblick über meine relevanten Qualifikationen und beruflichen Erfahrungen:

- Promotion in Informatik (künstliche Intelligenz, maschinelles Lernen, Data-Mining) (Jordanstown, GB, 1998) und Dipl.-Ing. (FH) in Elektrotechnik/Nachrichtentechnik (Augsburg, 1991).
- Über 8,5 Millionen Euro akquirierter Drittmittel aus europäischen (EC, ESF, COST Association) und nationalen (GB, Deutschland, Japan) Förderprogrammen.
- Autor von 150 publizierten [wissenschaftlichen Artikeln](#) in internationalen Fachzeitschriften und Konferenzbänden, Herausgeber von neun Sammelbänden, einschließlich der Springer Encyclopedia of Systems Biology, und Herausgeber von zehn Spezialausgaben internationaler Fachzeitschriften.
- 15 Jahre Erfahrung mit den strategischen, finanziellen, organisatorischen und operativen Belangen des Wissenschaftsmanagements:
  - Als Koordinator und Projekt-Manager von großen nationalen und internationalen Forschungsprojekten und Forschungsnetzwerken und Anträgen auf Forschungsförderung;
  - Als formaler Gutachter von großen Forschungsanträgen und Forschungsprojekten europäischer (EC, COST Association), nationaler und internationaler Förderprogramme (in den Bereichen Biomedizin und IKT);
  - Als Leiter der formalen Forschungsgruppe (20 Team-Mitarbeiter) und als Mitglied im Direktorium des Biomedical Science Research Institute (ca. 100 Mitarbeiter und über und ca. 200 Doktoranden) an der University of Ulster (GB);
  - Als Berater von nationalen und internationalen Forschungsförderprogrammen und öffentlichen Einrichtungen;

- Als Koordinator/Manager von Aktivitäten im Bereich Wissenschaftskommunikation und Wissensvermarktung in großen europäischen Verbundprojekten.
- 25 Jahre Erfahrung als Forscher und Akademiker in IKT und Bioinformatik mit Fokus auf Data-Science (Data- und Knowledge-Engineering, maschinelles Lernen, Algorithmen) und computergestützte Modellierung und Simulation von komplexen Systemen (Systembiologie, multiskale Modellierung und Simulation);
- Mitglied im verschiedenen Editorial Boards und Programmkommissionen internationaler Fachzeitschriften und Konferenzen/Workshops.
- Organisation großer internationaler Konferenzen und Workshops und Chair von internationalen Programmkommissionen.
- Entwicklung und Lehre eines MSc-Programms: Computational Systems Biology.
- Erfahrung mit verschiedenen Programmiersprachen, einschließlich C/C++ (8 Jahre), Java (8 Jahre), R (4 Jahre) und Python (2 Monate).
- Erfahrung mit verschiedenen Software-Tools, einschließlich Büroanwendungen, wissenschaftlicher Software und Betriebssystemen (Windows, Linux, MacOS).
- Erfahrung mit der Ausgründung einer IT-Firma im Data-Science-Bereich im Biotechnologiesektor.
- Erfahrung als Berater im Bereich Data-Science (im Biotechnologie-Sektor) und Informations- und IT-Sicherheit (speziell im Kontext der internationalen Normenreihe IOS/IEC 27000).
- Verhandlungssicheres Englisch in Wort und Schrift.

## **Akademische Ausbildung**

---

1998	Promotion in Informatik (Schwerpunkte maschinelles Lernen und künstliche Intelligenz) an der University of Ulster, Jordanstown, GB
1991	Diplom-Ingenieur (FH) in Elektrotechnik / Nachrichtentechnik an der Fachhochschule Regensburg (Vordiplom) und der Fachhochschule Augsburg (Hauptdiplom)

## **Beruflicher Werdegang**

---

Seit 12/2017	Data Analyst / Projektmanager, Helmholtz Zentrum München – Deutsches Forschungszentrum für Gesundheit und Umwelt
09/2017 – 11/2017	Data Scientist bei Blum GmbH, Höchst, Österreich
07/2016 – 09/2016	Senior Consultant für Informations- und IT-Sicherheit, ausecus GmbH, Augsburg
01/2002 – 04/2016	Lehrstuhl/Professor für Bioinformatik an der University of Ulster, Coleraine, GB
03/2001 – 12/2001	IT-Consultant (Datenbanken und Data-Mining) für phase-it Intelligent Solutions AG, Heidelberg
01/2000 – 12/2001	Wissenschaftlicher Angestellter (Data-Mining), German Cancer Research Center, Heidelberg

- 10/1999 – 12/1999 Wissenschaftlicher Angestellter (Data-Mining, maschinelles Lernen, Wissensmanagement), Forschungsinstitut für anwendungsorientierte Wissensverarbeitung, Ulm
- 01/1999 – 09/1999 Lecturer in Informatics, University of Ulster, Jordanstown, GB
- 07/1993 – 12/1998 Research Fellow und Research Associate, University of Ulster, Jordanstown, GB
- 10/1991 – 07/1997 Doktorand, University of Ulster, Jordanstown, GB
- 05/1997 – 09/1993 Software Entwickler für verschiedene Firmen in Deutschland (insgesamt 18 Monate in diesem Zeitraum): Siemens-Nixdorf, Augsburg; Atlas Elektronik, Bremen; CIB Software, München; MEL Mikroelektronik, München
- 07/1976 – 11/1979 Ausbildung zum Energieanlagenelektroniker, KUKA, Augsburg

## **Motivation, Forschungsinteressen, Wissenschaftliches Leitmotiv**

---

### **Motivation**

Ich war schon immer an technisch-wissenschaftlichen Themen interessiert. Dies erklärt, zumindest teilweise, meinen nicht so „linearen“ Werdegang von der Elektrotechnik zur Informatik (Data Science, AI) bis hin zur Systemmedizin. Neben den eigentlichen technisch-wissenschaftlichen Problemstellungen, interessiert mich aber zunehmend die Rolle von Wissenschaft und Technik in der Gesellschaft. Aus meiner Sicht muss die Wissenschaft mehr als bisher Verantwortung übernehmen, um den großen Herausforderungen der Zukunft gerecht zu werden, indem sie

- die negativen sozialen, gesundheitlichen, ökologischen und ökonomischen Folgen viel stärker als bisher als integralen Bestandteil von Forschung und Entwicklung behandelt, und
- einen Großteil ihrer Aktivitäten auf die großen gesellschaftlichen Herausforderungen der Zukunft richtet (Überbevölkerung, Sicherheit, Gesundheit, Ernährung, demografische Veränderungen, Umwelt und Klimawandel, Verteilung von Wohlstand, Digitalisierung und Automatisierung, Massenarbeitslosigkeit, Energie, Finanz- und Wirtschaftskrisen, Massenmigration, etc.).

Damit die Wissenschaft diesen Wandel hin zu mehr gesellschaftlicher Verantwortung vollziehen kann, bedarf es der Beteiligung der Wissenschafts- und Bildungsorganisationen, Forschungsförderern, Politik und Wirtschaft, und Öffentlichkeit. Ich bin überzeugt, dass der Schlüssel zu diesem Wandel in den Computerwissenschaften liegt, denn der Fortschritt im Bereich der Computerwissenschaften wird letztlich den gesamten wissenschaftlichen Prozess verändern zu einem Prozess, den man „intelligent e-science“ nennen könnte.

## **Forschungsinteressen und Wissenschaftliches Leitmotiv**

Seit meinem Diplom in Elektrotechnik/Nachrichtentechnik (1991), ist meine wissenschaftliche Laufbahn geprägt von den Veränderungen in den Naturwissenschaften durch den Einsatz von Informations- und Kommunikationstechnologien (IKT). Seitdem habe ich in folgenden IKT-Bereichen geforscht:

- Objektorientiertes Programmieren und Modellieren von Daten und Informationen
- Künstliche Intelligenz, maschinelles Lernen
- Computational-Statistics, Data-Mining, Big Data
- Computergestütztes Modellieren und Simulation von komplexen Systemen
- E-Science und verteiltes Rechnen
- Bioinformatik, Computational-Biology, Systembiologie

Aus meiner Forschung in diesen Bereichen entstand mein Leitmotiv für Wissenschaft:

### *Intelligent E-Science*

Intelligent E-Science beschreibt den modernen wissenschaftlichen Prozess anhand von fünf sich gegenseitig ergänzenden Paradigmen (bei Experiment und Theorie handelt es um die klassischen Paradigmen):

1. Experiment
2. Theorie
3. Computational-Science (wissenschaftliches Rechnen / E-Science)
4. Datenintensive Wissenschaft (heutzutage: Data-Science oder Big Data)
5. Künstliche Intelligenz

Im Gegensatz zu den bekannten Ansätzen, die oft multi- beziehungsweise interdisziplinär ausgerichtet sind, baut meine Leitidee auf einem *transdisziplinären* Ansatz auf. Im Zentrum dieses Ansatzes befindet sich ein gemeinsames, die Disziplinen übergreifendes konzeptuelles Gerüst (die fünf Paradigmen). Im Kern dieses Gerüsts steht das computergestützte Forschen, welches in dieser Form in der Vergangenheit nicht möglich war. Intelligent E-Science fußt auf dem gezielten Einsatz moderner Methoden und Technologien der Mathematik und Informatik in der Wissenschaft. Die großen Aufgaben der Zukunft erfordern daher eine Art »Lingua franca« (Prinzipien, Konzepte, Methoden, Werkzeuge, Systeme, Infrastrukturen, Organisationsstrukturen), die es den beteiligten Akteuren ermöglicht, komplexe Aufgaben zu lösen, die mit herkömmlichen multi- oder interdisziplinären Ansätzen kaum lösbar sind. Das Fundament eines solchen die Fachbereiche übergreifenden und integrierenden Systems findet sich in den entsprechenden Feldern der Mathematik und IKT (z. B. Datentechnologien, maschinelles Lernen, verteiltes Rechnen), und natürlich in den jeweiligen wissenschaftlichen und technischen Disziplinen. Die Konsequenzen, die sich aus diesem transdisziplinären Ansatz ergeben, sind weitreichend. Sie stellen beispielsweise die über Jahrhunderte gewachsenen Grenzlinien infrage, welche heute noch immer die verschiedenen Fachbereiche in voneinander weitgehend isolierte Silos einordnet. Um die großen Herausforderungen der Gegenwart und Zukunft zu meistern, müssen wir die Mauern dieser Silos nach und nach einreißen und durch einen transdisziplinären wissenschaftlichen Prozess ersetzen.

Anhang A beschreibt einiger meiner Forschungsinteressen im Bereich Data-Science etwas ausführlicher.

## Drittmittel für Forschung

Project Title (Acronym)	Role	Funder	From	To	£	€
Open Multiscale Systems Medicine (OpenMultiMed)	PI	COST	01.05.2016	30.04.2019		560,000
Multiscale Applications on European e-Infrastructures (MAPPER)	PI	EC-FP7	01.01.2010	30.09.2013		216,000
Robustness of Pathways to Invasion and Metastases in Breast Cancer	CI	Sasakawa	01.01.2009	31.07.2009	2,500	
Computational Systems Biology	PI	I.N.I.	13.12.2008	17.12.2008	368	
Biosociation Networks for Creative Information Discovery (BISON)	PI	EC-FP7	01.06.2008	31.05.2011		194,000
A Virtual Environment for Socially Aware knowledge Management (KnowledgeCraft)	PI	I.N.I.	30.01.2008	02.02.2008	630	
Computational Systems Biology	PI	I.N.I.	14.01.2008	28.02.2009	717	
SOLID -- Open, Service-Oriented Infrastructure for Language-Based Information Discovery	PI	I.N.I.	12.12.2007	14.12.2007	166	
Classification of Formalin-Fixed Paraffin-Embedded Tissue Data	PI	Almac	01.02.2007	30.09.2012	22,140	
SystemsBiologyGrid	PI	I.N.I.	27.11.2006	03.12.2006	1,662	
Mining of High-Throughput Data in Functional Genomics	CI	ESF	01.10.2006	30.09.2007	6,716	
Quasi-Opportunistic Supercomputing for Complex Systems in Grid Environments (QosCosGrid)	PI	EC-FP6	01.09.2006	31.05.2009		247,500
Bioinformatics Capability Funding	PI	D.E.L.	01.07.2006	31.07.2007	242,936	
Grid Services Based Environment to Enable Innovative Research (Chemomentum)	PI	EC-FP6	01.07.2006	31.03.2009		215,791
Advanced Methods and Technologies for Bioinformatics	CI	EC-ALFA	02.04.2006	01.12.2006	5,520	
Centre for Metabolomics	CI	D.E.L.	01.10.2005	31.03.2008	3,510,875	
Bioinformatics Capability Funding	PI	D.E.L.	31.08.2005	31.07.2006	234,000	
NUGO -- Impact of Food on Health -- Linking Genomics and Nutrition and Health Research	CI	I.N.I.	11.04.2005	30.04.2011	1,360	
A Regional Network for Post-Genomics and Systems Biology (SB(R)Net)	PI	EPSRC	01.10.2004	07.02.2008	58,182	
Data Mining Tools and Services for Grid Computing Environments (DataMiningGrid)	PI	EC-FP6	01.09.2004	30.09.2006		347,000
Systems Biology and Imaging: An Integrative Approach	CI	MRC/EPSRC	01.09.2004	31.08.2005	47,990	
Bioinformatics Capability Funding	PI	D.E.L.	01.08.2004	31.07.2007	502,000	
Bio 2004	PI	I.N.I.	02.06.2004	06.06.2004	5,995	
NUGO -- Impact of Food on Health -- Linking Genomics and Nutrition and Health Research	CI	EC-FP6	01.01.2004	30.04.2011	15,536	
Evaluation of the Prevalence of Vitamin D Deficiency ... and its Impact on Bone Health	CI	H.E.A.	01.10.2003	30.03.2006	72,816	
Medical Proteomics on the Grid	PI	I.N.I.	09.07.2003	17.17.2003	1,152	
Open Life Science GRID	PI	I.N.I.	20.11.2002	23.11.2002	2,235	
Information Integration of Life Science Data	PI	DWP	08.10.2002	30.10.2005	11,800	
Open Computing GRID for Molecular Science and Engineering (OpenMolGRID)	PI	EC-FP5	01.09.2002	28.02.2005	219,184	
ChemoGRID	PI	I.N.I.	02.02.2002	10.02.2002	280	
Miscellaneous	PI		01.01.2002	31.12.2011	1,916	
ESF COST Action 282: Knowledge and Exploration in Science and Technology (KnowLEST)	PI	ESF	01.12.2001	30.06.2005		240,000
Bioinformatics System for Correlation of Clinical and Molecular-Genetic Data	CI	BMBF		2 years		450,000
					<b>4,968,676</b>	<b>2,470,291</b>

### Key:

EC = European Commission  
 Sasakawa = Great Britain Sasakawa Foundation  
 D.E.L. = Northern Ireland Department of Education and Learning  
 I.N.I. = Invest Northern Ireland  
 EPSRC = Engineering and Physical Sciences Research Council, UK  
 MRC = Medical Research Council, UK  
 ESF = European Science Foundation  
 BMBF = Ministry for Education and Research, Germany  
 COST = COST Association

PI = Principal Investigator  
 CI = Co-Investigator

## Gutachtertätigkeit

Gutachter von Forschungsanträgen und laufender Forschungsprojekte für die Europäische Kommission (FP4, FP5, FP6, FP7, H2020), die European Science Foundation (ESF), COST Association, das Human Science Frontier Program (HSFP), und nationale Förderprogramme aus verschiedenen Ländern (Deutschland, Spanien, Norwegen, Tschechische Republik, Republik Irland, Großbritannien, Zypern, Polen, Niederlande und Kanada).

## **Wissenschaftskooperationen, Projekt- und Managementenerfahrung**

Federführend Rolle bei der Initiation und Umsetzung von nationalen und internationalen Forschungsk Kooperationen und Verbundprojekten:

- Initiator und Action-Chair: COST Action 15120 Open Multiscale Systems Medicine. Wissenschaftsnetzwerk, Partnerorganisationen aus mehr als 20 Ländern.
- Initiator und Mitglied des Management Komitees: Multiscale Applications on European e-Infrastructures. Europäisches Verbundprojekt (FP7, e-Infrastructures), acht Partnerorganisationen aus sechs Ländern.
- Initiator und Lead-Editor-in-Chief der Springer Encyclopedia of Systems Biology. Insgesamt mehrere hundert Forscher weltweit.
- Initiator und Mitglied des Management-Komitees: Bisociation Networks for Creative Information Discovery. Europäisches Verbundprojekt (FP7, FET), sieben Partnerorganisationen aus sechs Ländern.
- Initiator und Koordinator: A Virtual Environment for Socially Aware Knowledge Management (KnowledgeCraft). IP-Antrag unter FP7 mit Partnerorganisationen aus 18 Ländern<sup>1</sup>.
- Initiator und Koordinator (Projektmanager): Quasi-Opportunistic Supercomputing for Complex Systems in Grid Environments. Europäisches Verbundprojekt (FP6), zehn Partnerorganisationen aus neun Ländern (einschließlich Israel und Australien).
- Initiator und Mitglied des Management-Komitees: Grid Services Based Environment to Enable Innovative Research. Europäisches Verbundprojekt (FP6), neun Partnerorganisationen aus sieben Ländern.
- Initiator und Koordinator (Projektmanager): A Regional Network for Post-Genomics and Systems Biology (EPSRC, GB). Wissenschaftsnetzwerk, Partnerorganisationen aus fünf Ländern.
- Initiator und Koordinator (Projektmanager): Data Mining Tools and Services for Grid Computing Environments. Europäisches Verbundprojekt (FP6), fünf Partnerorganisationen aus fünf Ländern (einschließlich Israel).
- Initiator und Mitglied des Management-Komitees: Open Computing GRID for Molecular Science und Engineering. Europäisches Verbundprojekt (FP5), sieben Partnerorganisationen aus fünf Ländern.
- Initiator und Action-Co-Chair: COST Action 282: Knowledge Exploration in Science and Technology. Wissenschaftsnetzwerk, Partnerorganisationen aus mehr als zehn Ländern.
- Als Herausgeber von wissenschaftlichen Sammelbänden, Sonderausgaben von wissenschaftlichen Zeitschriften, Mitglied von Editorial Boards, Organisator von Konferenzen/Workshops, und als Mitglied von Programmkommissionen internationaler Konferenzen/Workshops, habe ich mit Organisationen und Wissenschaftlern weltweit kooperiert. (Details zu diesen Aktivitäten sind weiter unten aufgeführt.)

---

<sup>1</sup> Der Antrag erreichte 13 Punkte und war damit auf Platz 2 nach der Remote-Evaluation. Nach dem Hearing in Luxemburg wurde aber aus Kostengründen auf 12 Punkte herabgestuft.

Neben den oben aufgelisteten Kooperationen und Rollen, war ich in folgenden Bereichen in einer Führungsrolle tätig:

- Leiter der Bioinformatics Research Group (ca. 20 Mitarbeiter), University of Ulster, Coleraine, GB
- Mitglied des Direktorates des Biomedical Sciences Research Institute (ca. 100 Akademiker und 200 Doktoranden), University of Ulster, Coleraine, GB
- Leiter des Data-Mining-Teams am Deutschen Krebsforschungszentrum (DKFZ), Heidelberg
- Leiter des Teams für Data-Mining und maschinelles Lernen am Forschungsinstitut für anwendungsorientierte Wissensverarbeitung (FAW), Ulm

Daneben war ich im Auftrag von Invest NI und dem Department of Employment and Learning (beides Agenturen/Ministerien für Forschung, Entwicklung und Investment) als nordirischer Vertreter auf internationalen Messen (CeBit, Bio) und bei der Europäischen Kommission und der National Science Foundation, USA.

## Editorial Board

---

2016 – 2016	<i>Advances in Bioinformatics</i> , Hindawi Publishing Corporation
2013 – 2016	<i>J. of Complex Systems</i> , Hindawi Publishing Corporation
2011 – present	<i>Network Modeling &amp; Analysis in Health Inf. and Bioinformatics</i> , Springer
2009 – present	<i>Int'l J. of Next-Generation Computing</i> , Perpetual Innovation Media
2008 – 2016	<i>BMC Systems Biology</i> , BioMed Central
2006 – present	<i>Int'l J. of Comp. Intelligence in Bioinf. &amp; Systems Biology</i> , Res. India
2004 – present	<i>OMICS A J. of Integrative Biology</i> , Mary Ann Liebert, Inc.
2003 – present	<i>Briefings in Bioinformatics</i> , Oxford University Press
2001 – present	<i>Online J. of Bioinformatics</i>
2008 – present	<i>The Open Biotechnology J.</i> , Bentham Open
2009 – 2012	<i>J. of Computational Science</i> , Elsevier
2005 – 2010	<i>Int'l J. of Data Mining and Bioinformatics (IJDMB)</i> , InderScience
2005 – 2011	<i>BMC Bioinformatics</i> , BioMed Central

## Editor / Herausgeber

---

### Sammelbände und Enzyklopädien

- Dubitzky W., Wolkenhauer O., Cho K-H., Yokota H. (editors-in-chief) (2013), *Encyclopedia of Systems Biology*, Springer
- Dubitzky W., Kurowski K., Schott B. (editors) (2012), *Large-Scale Computing Techniques for Complex System Simulations*, John Wiley and Sons, Inc.



- Dubitzky W., Southgate J., Fuß H. (editors) (2011), *Understanding the Dynamics of Biological Systems: Lessons Learned from Integrative Systems Biology*, Springer
- Dubitzky W. (editor) (2008), *Data Mining Techniques in Grid Computing Environments*, Wiley-Blackwell, ISBN: 978 0 470 51258 6.
- Dubitzky W., Schuster A., Sloat P., Schroeder M., Romberg M. (editors) (2007), *Proc. of Int'l Workshop Distributed, High-Performance and Grid Computing in Computational Biology (GCCB 2006)*, Eilat, January, 2007, *Lecture Notes in Bioinformatics LNBI 4360*, ISSN 0302-9743, Springer Berlin Heidelberg New York.
- Dubitzky W., Granzow M., Berrar D. (editors) (2007), *Fundamentals of Data Mining in Genomics and Proteomics*, Springer Berlin Heidelberg New York, ISBN: 978-0-387-47508-0.
- Bremer E.G., Hakenberg J., (Sam) Han E-H., Berrar D., Dubitzky W. (editors) (2006), *Proc. of Int'l Workshop on Knowledge Discovery in Life Science Literature (KDLL 2006)*, Singapore, April, 2006, *Lecture Notes in Bioinformatics LNBI 3886*, ISSN 0302-9743, Springer Berlin Heidelberg New York.
- Dubitzky W., Azuaje F. (editors) (2004), *Artificial Intelligence Methods and Tools for Systems Biology*, Kluwer Academic Publishers, Kluwer Academic Publishers, Boston/Dordrecht/London. ISBN 1-4020-2859-8 (hardcover), 1-4020-2959-4 (paperback).
- López J.A., Benfenati E., Dubitzky W. (editors) (2004), *Proc of Int'l Symposium Knowledge Exploration in Life Science Informatics 2004*, Milan, Italy, 25-26 November 2004, *Lecture Notes in Computer Science 3303*, ISBN 3-540-23927-8 Springer Berlin Heidelberg New York.

### **Spezialausgaben wissenschaftlicher Zeitschriften**

- Berrar D., Lopes P., Davis J. & Dubitzky W. (2018). *Machine Learning* (Springer). Special Issue: [Machine Learning for Soccer](#)
- Dubitzky W. (2016). *Briefings in Bioinformatics*, 17(3). Special Issue: Computational Systems Biomedicine
- Dubitzky W. & Wang C. (2014). *Computation*. Special Issue: Multiscale Modeling and Simulation
- Gao J. & Dubitzky W. (2013), *IEEE J. of Biomedical and Health Informatics*
- Dubitzky W. (2010), *Briefings in Bioinformatics*, 10(4). Special Issue: Challenges in Bioinformatics and Computational Biology
- Dubitzky W. & Stankovski V. (2007), *Future Generation Computer Systems*, 23(1).
- Dubitzky W. (2006), *Briefings in Bioinformatics*, 7(4). Special Issue: Understanding the Computational Methodologies of Systems Biology
- Huang C-H., Lanza V., Rajasekaran S. & Dubitzky W. (2005), *J. of Clinical Monitoring and Computing*, 19(4-5)
- Lopez J.A., Azuaje F., Prank K. & Dubitzky W. (2004), *IEEE Transactions on Nano-Bioscience*, 3(3)
- Dubitzky W. (2004), *OMICS: A Journal of Integrative Biology*, 8(2)

- Dubitzky W. & Azuaje F.J. (2003), *Artificial Intelligence Review*, 20(1-2)

## **Organisation von Konferenzen und Workshops**

---

- Program Co-Chair: IEEE Int'l Conference on Bioinformatics & Biomedicine, Belfast, GB, 2-5 November 2014
- Program Co-Chair: Biomedical and Bioinformatics Challenges for Computer Science (BBC 2013) at Int'l Conference on Computational Science 2013 (ICCS 2013), Barcelona, Spanien, 5-7 June 2013
- Program Co-Chair: IEEE Int'l Conference on Bioinformatics & Biomedicine, Philadelphia, USA, 4-7 Oktober 2012
- Program Co-Chair: ESF Int'l Workshop on Mining of High-Throughput Data in Functional Genomics, University of Ulster, Coleraine, GB, May 8-9, 2007
- Program Co-Chair: Int'l Workshop on Distributed, High-Performance and Grid Computing in Computational Biology (GCCB 2006), Eilat, Israel, 2007
- Program Co-Chair: Int'l Workshop on Knowledge Discovery in Life Science Literature (KDLL 2006), Singapur, 2006
- Program Co-Chair: Int'l Program Committee of Int'l Symposium on Knowledge Exploration in Life Science Informatics (KELSI 2004), Mailand, Italien, 2004
- Program Co-Chair: Int'l Conference on Artificial Intelligence 2003 (IC-AI'03), Las Vegas, Nevada, USA, 2003
- European Simulation Multi Conference, Track 8: Environment, Biology, Ecology, Sociology and Medicine, Darmstadt, 2002

Seit 1999 Mitglied von ca. 100 Programmkommissionen internationaler Konferenzen, Workshops und Symposien.

## **Betreute Promotionen und Promotionskommission**

---

### **Betreute Promotionen**

- Multiscale Modeling and Simulation in Systems Biology (PI)
- Computational Biology Study into of Genetic Causes of LQT Syndrome (PI)
- A Data Mining Approach to Protein Unfolding Simulation Data (Adviser)
- Data Exploration and Knowledge Extraction: Their Application to the Study of Endocrine-Disrupting Chemicals (Adviser)
- Analysis and Modelling of Interactions in Kinase/Phosphatase Systems (PI)
- Graph-Based Modeling and Reverse-Engineering of Biochemical Networks (PI)
- Automated Analysis of Biomedical Literature and its Application to Microarray Data Analysis and Interpretation (PI)
- Machine Learning Methods for Analyzing DNA Microarray Data (PI)

- Knowledge Discovery & Automated Decision-Making with Unstructured & Structured Data (PI)

### Mitglied Promotionskommissionen

- Großbritannien (University of Ulster, Coleraine)
- Deutschland (Universität Konstanz)
- Schweiz (Universität Genf)
- Slowenien (Jožef Stefan Institute, Ljubljana)
- Spanien (Universitat Pompeu Fabra, Barcelona)
- Portugal (Universität Coimbra)

### Lehrveranstaltungen

---

- Einführung in System Biologie mit R
- Statistisches und Wissenschaftliches Programmieren in R: Teil I und II
- Modellierung und Simulation von Biologischen Systemen mit R: Teil I und II
- Analyse Biologischer Daten mit R: Teil I und II
- Programmieren mit C/C++

### Schriftenverzeichnis

---

#### Begutachtete Veröffentlichungen in Fachzeitschriften

Berrar, D., Lopes, P., Dubitzky, W. (2018). [Incorporating domain knowledge in machine learning for soccer outcome prediction](#), *Machine Learning*, Springer. doi: 10.1007/s10994-018-5747-8

Dubitzky, W., Berrar, D., Davis, J., Berrar, D. (2018). [The Open International Soccer Database for machine learning](#), *Machine Learning*, Springer. doi: 10.1007/s10994-018-5726-0

Berrar, D., Dubitzky, W. (2018). [Should significance testing be abandoned in machine learning?](#), *Int'l Journal of Data Science and Analytics*, Springer. doi: 10.1007/s41060-018-0148-4

Berrar, D., Lopes, P., Dubitzky, W. (2017). Caveats and pitfalls in crowdsourcing research on soccer referee bias. *Int'l Journal of Data Science and Analytics*, 4(2): 143-151, Springer.

Mizeranschi, A., Swain, M.T., Scona, R., Fazilleau, Q., Bosak, B., Piontek, T., Kopta, P., Thompson, P., Dubitzky, W. (2016). MultiGrain/MAPPER: A distributed multiscale computing approach to modeling and simulating gene regulation networks. *Future Generation Computing Systems*, 63: 1-14, Elsevier.

Mizeranschi, A., Zheng, H., Thompson, P., Dubitzky, W. (2016). A Two-Stage Inference Algorithm for Gene Regulation Network Models. *International Journal of Computational Biology and Drug Design*, 9:1/2, 6-24.

Mizeranschi, A., Zheng, H., Thompson, P., Dubitzky, W. (2015). Evaluating a common semi-mechanistic mathematical model of gene-regulatory networks. *BMC Systems Biology* 2015, 9(Suppl 5):S2.

Borgdorff J., Belgacem M.B., Bona-Casas C., Fazendeiro L., Groen D., Hoenen O., Mizeranschi A., Suter J.L., Coster D., Coveney P.V., Dubitzky W., Hoekstra A.G., Strand P., Chopard B. (2014),

- Performance of distributed multiscale simulations, *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 372(2021).
- Wang C., Beyerlein P., Pospisil H., Krause A., Nugent C. & Dubitzky W. (2012), An efficient method for modeling kinetic behavior of channel proteins in cardiomyocytes. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 9(1) (January 2012), 40-51. doi: 10.1109/TCBB.2011.84.
- Dubitzky W. (2011), Toward a system-centric global knowledge management approach to discovering (organizing and sharing) scientific knowledge from large-scale data, *OMICS: A Journal of Integrative Biology*, 15(4): 243-246.
- Hill T.R., Cotter A.A., Mitchell S., Boreham C.A., Dubitzky W., Murray L., Strain J.J., Flynn A., Robson P.J., Wallace J.M., Kiely M. & Cashman K.D. (2010), Vitamin D status and parathyroid hormone relationship in adolescents and its association with bone health parameters: analysis of the Northern Ireland Young Heart's Project, *Osteoporosis International*, 21(4): 695-700.
- Swain M.T., Mandel, J.J. & Dubitzky, W. (2010), Comparative study of three commonly used continuous deterministic methods for modeling gene regulation networks. *BMC Bioinformatics* 11:459.
- Swain M., Silva C.G., Loureiro-Ferreira N., Ostroptsyky V., Brito J., Riche O., Stahl F., Dubitzky W. & Brito R.M.M. (2010) P-found: Grid-enabling distributed repositories of protein folding and unfolding simulations for data mining. *Future Generation Computer Systems*, 26(3): 424-433
- Stankovski, V., Swain, M., Kravtsov, V., Niessen, T., Wegener, D., Röhm, M., Trnkoczy, J., May M., Franke, J., Schuster, A. & Dubitzky, W. (2008), Digging deep into the data mine with DataMiningGrid, *IEEE Internet Computing*, 12(6), 69-76.
- Cashman K.D., Hill T.R., Cotter A.A., Boreham C.A., Dubitzky W., Murray L., Strain J., Flynn A., Robson P.J., Wallace J.M. & Kiely M. (2008), Low vitamin D status adversely affects bone health parameters in adolescents, *Am J Clin Nutr.*, 87(4), 1039-1044.
- Hill T.R., Cotter A.A., Mitchell S. & Boreham C.A., Dubitzky W., Murray L., Strain J.J., Flynn A., Robson P.J., Wallace J.M., Kiely M. & Cashman K.D. (2008), Vitamin D status and its determinants in adolescents from the Northern Ireland Young Hearts 2000 cohort, *Br J Nutr.*, 99(5), 1061-1067.
- Stankovski V., Swain M., Kravtsov V., Niessen T., Wegener D., Kindermann J. & Dubitzky, W. (2008), Grid-enabling data mining applications with DataMiningGrid: An architectural perspective, *Future Generation Computer Systems*, 24(4), 259–279.
- Fuß H., Dubitzky W., Downes C.S. & Kurth M.J. (2008), Src family kinases and receptors: analysis of three activation mechanisms by dynamic systems modeling, *Biophysical Journal*, 94(6), 1995-2006. Preprint, 30 Nov 2007: doi:10.1529/biophysj.107.115022. Open Access.
- Fuß H., Dubitzky W., Downes C.S. & Kurth M.J. (2007), Deactivation of Src family kinases: Hypothesis testing using a Monte Carlo sensitivity analysis of systems-level properties, *Journal of Computational Biology*, 14(9), 1185–1200. doi: 10.1089/cmb.2007.0095
- Mandel J.J., Fuss H., Palfreyman N.M. & Dubitzky W. (2007), Modeling biochemical transformation processes and information processing with Narrator, *BMC Bioinformatics*, 8:103. doi:10.1186/1471-2105-8-103.
- Mandel J.J., Palfreyman N.M. & Dubitzky W. (2007), Modelling codependence in biological systems, *IET Systems Biology*, 1(1), 18-32. doi:10.1049/iet-syb:20060002
- Gilbert D., Fuß H., Gu X., Orton R., Robinson S., Vyshemirsky V., Kurth M.J., Downes C.S. & Dubitzky W. (2006), Computational methodologies for modelling, analysis and simulation of signalling networks, *Briefings in Bioinformatics*, 7(4), 339-353. doi:10.1093/bib/bbl043.
- Natarajan, J., Berrar, D., Dubitzky, W., Hack, C., Zhang, Y., DeSesa, C., Van Brocklyn, J.R. & Bremer, E.G. (2006), Text mining of full-text journal articles combined with gene expression analysis reveals a relationship between sphingosine-1-phosphate and invasivity of a glioblastoma cell line, *BMC Bioinformatics*, 7:373. doi:10.1186/1471-2105-7-373.
- Fuß, H., Dubitzky, W., Downes, C.S. & Kurth, M.J. (2006), Bistable switching and excitable behaviour in the activation of Src at mitosis, *Bioinformatics*, 22(14), e158-e165. doi:10.1093/bioinformatics/btl201.

- Berrar D., Bradbury I. & Dubitzky W. (2006), Instance-based concept learning from multiclass DNA microarray data, *BMC Bioinformatics*, 7:73. doi:10.1186/1471-2105-7-73.
- Berrar D., Bradbury I. & Dubitzky W. (2006), Avoiding model selection bias in small-sample genomic data sets, *Bioinformatics*, 22(10), 1245-1250. doi:10.1093/bioinformatics/btl066.
- Swain M., Hunniford T., Mandel J., Palfreyman N. & Dubitzky W. (2005), Reverse-engineering gene-regulatory networks using evolutionary algorithms and grid computing, *Journal of Clinical Monitoring and Computing*, 19(4-5), 329-337.
- Berrar D., Stahl F., Goncalves Silva C.S., Rodrigues J.R., Brito R.M.M. & Dubitzky W. (2005), Towards data warehousing and mining of protein unfolding simulation data, *Journal of Clinical Monitoring and Computing*, 19(4-5), 307-317.
- Fuß H., Dubitzky W., Downes S. & Kurth M.J. (2005), Mathematical models of cell cycle regulation, *Briefings in Bioinformatics*, 6(2), 1–15. doi:10.1093/bib/6.2.163.
- Berrar D., Bradbury I., Sturgeon B., Downes C.S. & Dubitzky W. (2005), Survival trees for analyzing clinical outcome in lung adenocarcinomas based on gene expression profiles: Identification of neogenin and diacylglycerol kinase A expression as critical factors, *Journal of Computational Biology*, 12(5), 534–544.
- Natarajan J., Berrar D., Hack C. & Dubitzky W. (2005), Knowledge discovery in biology texts: Applications, evaluation strategies, and perspectives, *Critical Reviews in Biotechnology*, 25(1-2), 31-52.
- Mandel J., Palfreyman N., Lopez J. & Dubitzky W. (2004), Representing bioinformatic causality, *Briefings in Bioinformatics*, 5(3), 270-283. doi:10.1093/bib/5.3.270.
- Dubitzky W., McCourt D., Galushka M., Romberg M. & Schuller B. (2004), Grid-enabled data warehousing for molecular engineering, *Parallel Computing*, 30(9-10), 1019-1035.
- Brito R.M.M., Dubitzky W. & Rodrigues J.R. (2004), Protein folding and unfolding simulations: A new challenge for data mining, in W. Dubitzky (guest editor), *Data mining meets integrative biology – a symbiosis in the making*, *OMICS: A Journal of Integrative Biology*, 8(2), 153-166.
- Berrar D., Dubitzky W., Solinas-Toldo S., Bulashevskaya S., Granzow M., Conrad C., Kalla J., Lichter P. & Eils R. (2001), Design and implementation of a database system for comparative genomic hybridization analysis, *IEEE Engineering in Medicine and Biology*, 20(4), 75-83.
- Azuaje F., Dubitzky W., Black N.D. & Adamson K. (2001), Case retrieval strategies for CBR: A categorized bibliography, *The Knowledge Engineering Review*, 15 (4), 371-379.
- Dubitzky W., Bulashevskaya S., Granzow M., Solinas-Toldo S., Joos S., Lichter P. & Eils R. (2001), A data mining approach to detect and analyse comparative genomic hybridization patterns of cancer cells, *Annales de Génétique*, 44(1), 160.
- Azuaje F., Dubitzky W., Black N.D. & Adamson K. (2000), Discovering relevance knowledge in data: A growing cell structures approach, *IEEE Transactions on Systems, Man and Cybernetics. Part B: Cybernetics*, 30(3), 448-460.
- Azuaje F., Dubitzky W., Black N.D. & Adamson K. (1999), Improving clinical decision support through case-based fusion, *IEEE Transactions on Biomedical Engineering*, 46(10), 1181-1185.
- Dubitzky W., Büchner A.G., Hughes J.G. & Bell D.A. (1999), Towards concept-oriented databases, *Data and Knowledge Engineering*, 30, 23-55.
- Dubitzky W., D. Mariotti, M. Hyland, J. McLaughlin & P. Maguire (1999), Analyzing plasma emission spectra using neural networks, *ISCA Int'l Journal of Computers and Their Applications*, 6, 88-94.
- Patterson D., Anand S.S., Dubitzky W. & Hughes J.G. (1999), Towards automated case knowledge discovery in the M<sup>2</sup> case-based reasoning systems, *Knowledge and Information Systems: An Int'l Journal*, 1, 61-82.
- Azuaje F., Dubitzky W., Lopes P., Black N.D., Adamson K., Wu X. & White J.A. (1999), Predicting coronary disease based on short-term electrocardiogram patterns: A neural network approach, *Artificial Intelligence in Medicine*, 15, 275-297.

Azuaje F., Dubitzky W., Lopes P., Black N.D., Adamson K., Wu X. & White J. (1998), Making clinical data meaningful: Knowledge discovery in coronary disease risk assessment, *The Irish Journal of Medical Science*, 167(4), 274.

Dubitzky W., Schuster A., Hughes J.G. & Bell A.D. (1997), An advanced case-knowledge architecture based on fuzzy objects, *Applied Intelligence: The Int'l Journal of Artificial Intelligence, Neural Networks, and Complex Problem-Solving Technologies*, 7, 187-204.

Dubitzky W., Hughes J.G. & Bell A.D. (1996), A generic, object-oriented case-knowledge representation scheme, and its integration into a wider information management scenario, *Expert Systems: The Int'l Journal of Knowledge Engineering and Neural Networks*, 13(3), 219-233, Blackwell Publishers, UK.

Dubitzky W., Carville F. & Hughes J.G. (1993), Case-level Knowledge modelling in CBR, *Irish Journal of Psychology*, 14(3), 478-479.

## **Begutachtete Veröffentlichungen in Konferenzberichtsbanden**

Mizeranschi, A., Zheng, H., Thompson, P., Dubitzky, W. (2014), A multi-model reverse-engineering algorithm for large gene regulation networks, *The IEEE Int'l Conference on Bioinformatics and Biomedicine 2014 (BIBM 2014)*, pp. 510-514, Belfast, UK, 2-5 November.

Mizeranschi, A., Kennedy, N., Zheng, H., Thompson, P., Dubitzky, W. (2014), Automated reverse-engineering of gene regulatory networks based on semi-mechanistic rate laws, *The 5th International Workshop on Integrative Data Analysis in Systems Biology*, Belfast, UK, 2-5 November 2014.

Mizeranschi A., Kennedy N., Thompson, P., Huiru Zheng, & Dubitzky W. (2014), The influence of network topology on reverse-engineering of gene-regulatory networks, *Procedia Computer Science, International Conference on Computational Science (ICCS 2014)*, 29(2014), 410-421.

Kennedy N., Mizeranschi A., Thompson, P., Zheng H. & Dubitzky W. (2013), Reverse-engineering of gene regulation models from multi-condition experiments, *IEEE Symposium Series on Computational Intelligence 2013 (SSCI 2013)*, Singapore, 16-19 April, pp. 112-119.

Kennedy N., Thompson, P., Zheng H. & Dubitzky W. (2011), Multi-scale modelling of the bile acid and xenobiotic system, in Arabnia H.R. & Quoc-Nam T. (editors) *Proc. of The 2011 Int'l Conference on Bioinformatics and Computational Biology (BIOCOMP'11)*, Volume II, 499-505.

Kurowski, K., Back, W., Dubitzky, W., Gulyás, L., Kampis, G., Mamonski, M., Szemes, G. and Swain, M. (2009), Complex system simulations with QosCosGrid, *Proc. of 9th Int'l Conference on Computational Science, Part I*, G. Allen, J. Nabrzyski, E. Seidel, G.D. Albada, J. Dongarra and P.M. Sloot (eds.). *Lecture Notes in Computer Science*, vol. 5544. Springer-Verlag, Berlin, Heidelberg, 387-396.

Kravtsov, V., Schuster, A., Carmeli, D., Kurowski, K. & Dubitzky, W. (2008), Grid-enabling complex system applications with QosCosGrid: An architectural perspective, in *Proc. of Int'l Conference on Grid Computing and Applications (GCA'08)*, Las-Vegas, USA.

Kravtsov, V., Carmeli, D., Dubitzky, W., Orda, A., Schuster, A., Silberstein, M. & Yoshpa, B. (2008), Quasi-opportunistic supercomputing in grid environments, in *Proc. of Int'l Conference on Algorithms and Architectures*, Cyprus, 233-244.

Kravtsov, V., Swain, M., Dubin, U., Dubitzky, W. & Schuster A. (2008), A fast and efficient algorithm for topology-aware coallocation, in *Proc. of Int'l Conference on Computational Science*, Krakow, Poland, 274-283.

Gulyas L., de Back W., Szemes G., Kurowski K., Dubitzky W. and Kampis G. (2008), Templates for distributed agent-based simulations, *20th European Modeling and Simulation Symposium*, September 2008, Briatico, Italy.

Swain, M., Mandel J.J. and Dubitzky, W. (2008), Comparing grid computing solutions for reverse-engineering gene regulatory networks, in *Proc. of Int'l Conference on Computational Science (ICCS 2008)*, Poland, 106-115.

Kravtsov., V., Swain, M., Schuster, A., Dubitzky, W. and Dubin, U. (2008), Grid computing solutions for distributed repositories of protein folding and unfolding simulations, in *Proc. of Int'l Conference on Computational Science (ICCS 2008)*, Poland, 274-283.

Swain, M., Ostropytskyy, V., Silva, C.G., Stahl, F., Riche, O., Brito, R.M.M. and Dubitzky, W. (2008), Grid computing solutions for distributed repositories of protein folding and unfolding simulations, Proc. of Int'l Conference on Computational Science (ICCS 2008), Krakow, Poland, 70-79.

Silva C.G., Ostropytskyy V., Loureiro-Ferreira N., Berrar D., Swain M., Dubitzky W. & Brito R.M.M. (2006), P-found: The protein folding and unfolding simulation repository, in Proc. of the 2006 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB'06), 101-108.

Berrar D. & Dubitzky W. (2006), Neural plasma, in M. Bramer (editor), Artificial Intelligence in Theory and Practice, Proc. of IFIP 19th World Computer Congress, TC 12: IFIP AI 2006 Stream, Series: IFIP Int'l Federation for Information Processing, Vol. 217, 159-168, XVI, Springer, Boston, ISBN: 0-387-34654-6.

Bi Y. & Dubitzky W. (2006), An evidential approach in ensembles, in H. Haddad (editor), Proc. of the 2006 ACM Symposium on Applied Computing (SAC), , 1-6, Dijon, France, April 23-27, 2006. ACM 2006, ISBN 1-59593-108-2.

Natarajan J., Haines C., Berglund B., DeSesa C., Hack C.J., Dubitzky W. & Bremer E.G. (2006), GetItFull – A tool for downloading and pre-processing full-text journal articles, in E.G., Bremer, J. Hakenberg, E.-H. S., Han, D. Berrar, W. Dubitzky (editors), Knowledge Discovery in Life Science Literature, Proc. of Int'l Workshop, KDLL 2006, Singapore, April 9, 2006, Lecture Notes in Bioinformatics, Vol. 3869, 139-145.

Boreham C.A., Dubitzky W., Robson P.J., Wallace Julie M. & Bi Y. (2005), Three-dimension model for developing a nutrition, lifestyle and health database, Proc. of The 2005 Int'l Conference on Information and Knowledge Engineering, 151-157.

Wang C., Krause A., Nugent C. & Dubitzky W. (2005), Focal activity in simulated LQT2 models at rapid ventricular pacing: Analysis of cardiac electrical activity using grid-based computation, Proc. of 6th Int'l Symposium on Biological and Medical Data Analysis (ISBMDA'05), 305-316, Portugal.

Natarajan J., Mulay N., DeSesa C., Hack C.J., Dubitzky W. & Bremer E.G. (2005), A grid infrastructure for text mining of full text articles and creation of a knowledge base of gene relations, Proc. of 6th Int'l Symposium on Biological and Medical Data Analysis (ISBMDA'05), 101-108, Portugal.

Hui W. & Dubitzky W. (2005), A flexible and robust similarity measure for analogy-based AI methods and analytical tasks, Proc. of the Int'l Joint Conference on Artificial Intelligence (IJCAI'05), 27-30, Edinburgh, Scotland.

Stahl F., Berrar D., Goncalves Silva C.S., Rodrigues J.R., Brito R.M.M. & Dubitzky W. (2005), Grid warehousing of molecular dynamics protein unfolding data, in Proc. of 5th IEEE/ACM Int'l Symposium on Cluster Computing and the Grid, BioGrid 2005, Vol. 1, 496-503.

Swain M., Hunniford T., Mandel J., Palfreyman N. & Dubitzky W. (2005), Modeling gene-regulatory networks using evolutionary algorithms and distributed computing, Proc. of the 5th IEEE/ACM Int'l Symposium on Cluster Computing and the Grid, BioGrid 2005, Vol. 1, 512- 519.

Bremer E.G., Natarajan J., Zhang Y., DeSesa C., Hack C.J. & Dubitzky W. (2004), Text mining of full text articles and creation of a knowledge base for analysis of microarray data, in J.A. López, E. Benfenati & W. Dubitzky (editors), Proc. of Int'l Symposium Knowledge Exploration in Life Science Informatics 2004 (KELSI 2004), 84-95, Milan, Italy, 25-26 November 2004, Lecture Notes in Computer Science 3303, ISBN 3-540-23927-8 Springer Berlin Heidelberg New York, 2004.

Bi Y., Bell D.A., Wang H., Guo G. & Dubitzky W. (2004), Classification decision combination for text categorization: An experimental study, Proc. of Database and Expert Systems Applications, 15th Int'l Conference (DEXA'04), 222-231, Lecture Notes in Computer Science 3180 Springer 2004, ISBN 3-540-22936-1.

Stankovski V., May M., Franke J., Schuster A., McCourt D. & Dubitzky W. (2004), A service-centric perspective for data mining in complex problem solving environments, H.R. Arabnia & J. Ni (editors), Proc. of Int'l Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA'04), Vol. 2, 780-787.

Mazzatorta P., Benfenati E., Schuller B., Romberg M., McCourt D., Dubitzky W., Sild S., Karelson M., Papp A., Bágyi I. & Darvas F. (2004), OpenMolGRID: Molecular science and engineering in a grid

context, H.R. Arabnia & J. Ni (editors), Proc. of Int'l Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA'04), Vol II, 775-779.

McCourt D., Lopez J., Benfenati E., Mazzatorta P., Romberg M., Schuller B. & Dubitzky W. (2003), Towards an intelligent data type for toxicity, Proc. of Int'l Conference on Artificial Intelligence, USA, 328-334.

Ragg T., Granzow M., Menzel W. & Dubitzky W. (2003), Filtering nonlinear intensity dependencies in cDNA-microarray experiments, Proc. of Int'l Conference on Artificial Intelligence, 335-341.

Berrar D., Downes C.S. & Dubitzky W. (2003), A probabilistic neural network for gene selection and classification of high-dimensional microarray data, Proc. of Int'l Conference on Artificial Intelligence, 342-349.

Sturgeon B., McCourt D., Cowper J., Palmer F., McClean S. & Dubitzky W. (2003), Can the grid help to solve the data integration problems in molecular biology?, Proc. of the 3rd IEEE/ACM Int'l Symposium on Cluster Computing and the Grid, BioGrid 2003, 588-593.

Berrar D., Downes C.S. & Dubitzky W. (2003), Multiclass cancer classification using gene expression profiling and probabilistic neural networks, Proc. of the Pacific Symposium on Biocomputing 8, 5-16.

Berrar, D., Granzow M., Dubitzky W., Stilgenbauer S., Wilgenbus K., Döhner H., Lichter P. & Eils R. (2001), New insights in clinical impact of molecular genetic data by knowledge-driven data mining, Proc. 2nd Int'l Conference on Systems Biology, 275-281.

Dubitzky W., Berrar D., Granzow M. & Eils R. (2001), Detecting broad-band and selective correlation patterns among gene expression and drug activity data, Proc. of Critical Assessment of Techniques for Microarray Data Analysis Techniques (CAMDA 2001), 17-22.

Berrar D., Dubitzky W., Granzow M. & Eils R. (2001), Analysis of gene expression and drug activity data by knowledge-based association mining, Proc. of Critical Assessment of Techniques for Microarray Data Analysis Techniques (CAMDA 2001), 23-28.

Dubitzky W., Krebs O. & Eils R. (2001), Minding, OLAPing, and mining biological data: Towards a data warehousing concept in biology, Proc. of Network Tools and Applications in Biology (NETTAB), CORBA and XML: Towards a Bioinformatics Integrated Network Environment, 78-82, Genoa, Italy.

Schuster A., Dubitzky W., Azuaje F., Granzow M., Berrar D. & Eils R. (2000), Tumor identification by gene expression profiles: A comparison of five different clustering methods, Proc. of Critical Assessment of Techniques for Microarray Data Analysis (CAMDA 2000), 34-35.

Dubitzky W., Granzow M., Berrar D., Bulashevskaya S., Conrad C., Gerlich D. & Eils R. (2000), A comparison of symbolic and subsymbolic machine learning approaches to molecular classification of cancer and gene identification, Proc. of Critical Assessment of Techniques for Microarray Data Analysis (CAMDA 2000), 12-13.

Bulashevskaya S., Dubitzky W. & Eils R. (2000), Mining gene expression data using rough set theory, Proc. of Critical Assessment of Techniques for Microarray Data Analysis (CAMDA 2000), 4-5.

Azuaje F., Dubitzky W., Black N.D. & Adamson K. (1999), Automated modelling of case retrieval structures using self-organising neural networks, Proc. of 1st Int'l Information Reuse and Integration Conference, 32-35.

Lester N.G., Wilkie F.G., Dubitzky W. & Bustard D.W. (1999), A knowledge-guided retrieval framework for reusable software artefacts, Proc. of 17th Annual Association of Management / Int'l Association of Management Conf. on Computer Science, 206-211.

Wang H., Dubitzky W., Düntsch I. & Bell A.D. (1999), A lattice machine approach to automated case base design: Marrying lazy and eager learning, Proc. of 16th Int'l Joint Conference on Artificial Intelligence, Sweden, Vol. 1, 254-259.

Wu X.W., Dubitzky W., Azuaje F. & Black N.D. (1999), Discovering relevant knowledge for clustering through incremental growing cell structures, Proc. of 2nd Int'l. Conf. on Information Fusion, 46-52.

Azuaje F., Dubitzky W., Black N.D. & Adamson K. (1999), Discovering and fusing relevant knowledge from databases based on an incremental unsupervised learning approach, Proc. 2nd Int'l. Conf. on Information Fusion, 31-38.



Dubitzky W., Büchner A.G. & Azuaje F. (1999), Viewing knowledge management as a case-based reasoning application, Proc. of AAAI Workshop: Exploring Synergies of Knowledge Management and Case-Based Reasoning, AAAI Press, Technical Report WS-99-10, 23-27.

Dubitzky W., Azuaje F., Lopes P. & Black N.D. (1999), McCullagh P., Song Y., On local and global feature weight discovery for case-based reasoning, Proc. of ISCA 14th Int'l Conference on Computers and their Applications, Mexico, 107-110.

Azuaje F., Dubitzky W., Lopes P., Black N.D., Adamson K., Wu X., White J.A., Discovery of incomplete knowledge in electrocardiographic data, in Proc. of 3rd Int'l Conference of Neural Networks and Expert Systems in Medicine and Healthcare, Italy, E. Ifeachor, A. Sperduti, A. Starita (eds.), World Scientific, Singapore, 286-294, 1998.

Dubitzky W., Mariotti D., Hyland M. & McLaughlin J. (1998), A neural network approach to plasma emission interpretation, Proc. of ISCA 11th Int'l Conference on Computers and their Applications in Industry and Engineering, 47-50.

Dubitzky W., Lopes P., Hughes J.G. & Bell A.D. (1998), Representing incomplete knowledge in case-based reasoning, Proc. ISCA 11th Int'l Conference on Computers and their Applications in Industry and Engineering, USA, 133-136.

Azuaje F., Dubitzky W., Lopes P., Black N.D., Adamson K., Wu X. & White J.A. (1998), Knowledge discovery in electrocardiographic data based on neural clustering algorithms, Proc. of 8th Mediterranean Conference on Medical and Biological Engineering and Computing, (electronic proceedings on CD), Cyprus, ISBN 9963-607-13-6.

Patterson D., Dubitzky W., Anand S.S. & Hughes J.G. (1998), On the automation of case base development from large databases, Proc. of AAAI 1998 Workshop: Case-Based Reasoning Integrations, 126-130, USA.

Azuaje F., Dubitzky W., Wu X., Lopes P., Black N.D., Adamson K. & White J.A. (1997), A neural network approach to coronary heart disease risk assessment based on short-term measurement of RR intervals, Proc. of Computers in Cardiology, Vol. 24, 53-56, Sweden.

Dubitzky W., Schuster A., Bell A.D. & Hughes J.G. (1997), How similar is VERY YOUNG to 43 years of age? On the representation and comparison of polymorphic properties, Proc. of 15th Int'l Joint Conference on Artificial Intelligence, 226-231, Japan.

Schuster A., Dubitzky W., Lopes P., Adamson K., Bell A.D., Hughes J.G. & White J.A. (1997), Aggregating features and matching cases on vague linguistic expressions, Proc. of 15th Int'l Joint Conference on Artificial Intelligence, 252-257, Japan.

Büchner A.G., Dubitzky W., Schuster A., Lopes P., Bell A.D., O'Doneghue P.G., Hughes J.G., Bell A.D., Adamson K, White J.A., Anderson J.M.C.C. & Mulvenna M.D. (1997), Corporate evidential decision making in performance prediction domains, Proc. of 13th Conference on Uncertainty in Artificial Intelligence, 38-45, USA.

Dubitzky W., Hughes J.G. & Bell A.D. (1997), Corporate multi-criteria decision making using evidential reasoning, Proc. of 26th Northeast Decision Sciences Institute Annual Meeting, 135-137, Annapolis, USA.

Schuster A., Dubitzky W., Adamson K., Bell A.D. & Hughes J.G. (1997), Processing similarity between a mix of crisply and fuzzily defined case properties, Proc. of 2nd Int'l ICSC Symposium on Fuzzy Logic and Applications, 247-253, Switzerland.

Dubitzky W., Hughes J.G. & Bell A.D. (1996), A multi-expert scheme for defining reliable case categories, Proc. of Int'l Conference on Knowledge Based Computer Systems, 91-102, India.

Dubitzky W., Hughes J.G. & Bell A.D. (1996), Case memory and the behaviouristic model of concepts, Proc. of Advances in Case-Based Reasoning, 3rd European Workshop, EWCBR-96, 120-134, Switzerland.

Dubitzky W., Lopes P., White J.A., Anderson J.M.C.C., Dempsey G.J., Hughes J.G. & Bell A.D. (1996), A holistic approach to coronary heart disease risk assessment using case-based reasoning, Proc. of 2nd Int'l Conference on Neural Networks and Expert Systems in Medicine and Healthcare, 97-103, UK.

Dubitzky W., Hughes J.G. & Bell A.D. (1996), Multiple opinion case-based reasoning and the theory of evidence, Proc. of 6th Int'l Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, 447-452, Spain.

Dubitzky W., Hughes J.G. & Bell A.D. (1996), Knowledge management via case-knowledge and object-orientation, Proc. of 3rd World Congress on Expert Systems, 1150-1157, Korea.

Dubitzky W., Schuster A., Hughes J.G. & Bell A.D. (1995), Conceptual distance of numerically specified case features, Proc. of 2nd New Zealand Int'l Two-Stream Conference on Artificial Neural Networks and Expert Systems, 210-213, New Zealand.

Dubitzky W., Carville F. & Hughes J.G. (1993), Case-level knowledge modelling in CBR, Proc. of 6th Irish Conference on Artificial Intelligence and Cognitive Science, 217-227, UK.

## **Begutachtete Veröffentlichungen in Sammelbänden und Nachschlagewerken**

Mizeranschi, A., Groen, D., Borgdorff, J., Hoekstra, A.G., Chopard, B., Dubitzky, W. (2016), Chapter 17: Anatomy and physiology of multiscale modeling and simulation in systems medicine, in Ulf Schmits et al. (editors). Systems biology for medicine, Methods Mol Biology 2016; 1386, 375-404.

Berrar D. & Dubitzky W. (2013), Bootstrapping, in W. Dubitzky, O. Wolkenhauer, K.-H. Cho & H. Yokota (editors), Encyclopedia of systems biology, Springer, New York.

Berrar D. & Dubitzky W. (2013), Overfitting, in W. Dubitzky, O. Wolkenhauer, K.-H. Cho & H. Yokota (editors), Encyclopedia of systems biology, Springer, New York.

Berrar D. & Dubitzky W. (2013), Decision trees, in W. Dubitzky, O. Wolkenhauer, K.-H. Cho & H. Yokota (editors), Encyclopedia of systems biology, Springer, New York.

Kennedy N., Thompson P., Schmidt O., Zheng H. & Dubitzky W. (2013), Bile acid and xenobiotic system, in W. Dubitzky, O. Wolkenhauer, K.-H. Cho & H. Yokota (editors), Encyclopedia of systems biology, Springer, New York.

Dubitzky W., Kötter T., Schimdt O. & Berthold M.R. (2012), Towards creative information exploration based on Koestler's concept of bisociation, in M.R. Berthold (editor), Bisociative knowledge discovery, 11-32, Springer, LNCS 7250 2012.

Schmidt O., Kranjc J., Mozetič I., Thompson P. & Dubitzky W. (2012), Bisociative exploration of biological and financial literature with using clustering, in M.R. Berthold (editor), Bisociative knowledge discovery, 438-451, Springer, LNCS 7250 2012.

Swain M., Coti C., Mandel J. & Dubitzky W. (2011), A topology-aware evolutionary algorithm for reverse-engineering gene regulatory networks, in Dubitzky W., Kurowski K., Schott B. (eds.), Large-scale computing techniques for complex system simulations, 141-162, Wiley-VCH.

Sánchez A., Montes J., Dubitzky W., Valdés J.J., Pérez M.S. & de Miguel P. (2008), Data mining meets grid computing: Time to dance?, in W. Dubitzky (editor), Data mining techniques in grid computing environments, Wiley, 1-16.

Romberg M., Benfenati E. & Dubitzky W. (2007), Open computing grid for molecular sciences, in E-G. Talbi & A. Zomaya (editors), Grids for bioinformatics and computational biology, Wiley Book Series on Parallel and Distributed Computing. John Wiley & Sons, New York, 1-22. ISBN: 978-0-471-78409-8.

Berrar D., Sturgeon B., Bradbury I., Downes C.S. & Dubitzky W. (2004), Integration of microarray data for a comparative study of classifiers and identification of marker genes, in J.S. Shoemaker & S.M. Lin (editors), Methods of microarray data analysis IV, 147-162, Springer.

Dubitzky W., Granzow M., Downes C.S. & Berrar D. (2002), Introduction to Microarray Data Analysis, in D. Berrar, W. Dubitzky & M. Granzow (editors), A Practical Approach to Microarray Data Analysis, 1-46, Copyright 2002, Kluwer Academic Publishers, Boston/Dordrecht/London.

Dubitzky W., Granzow M. & Berrar D. (2001), Comparing symbolic and subsymbolic machine learning approaches to classification of cancer and gene identification, in S.M. Lin & K.F. Johnson (editors), *Methods of microarray data analysis: Papers from CAMDA'00*, 151-166, Kluwer Academic Publishers; ISBN: 0792375645.

Dubitzky W., Granzow M. & Berrar D. (2001), Data mining and machine learning methods for microarray analysis, in S.M. Lin & K.F. Johnson (editors), *Methods of microarray data analysis: Papers from CAMDA'00*, 5-22, Kluwer Academic Publishers; ISBN: 0792375645.

Dubitzky W. & Azuaje F. (2000), A genetic algorithm and a growing cell structures approach to learning case retrieval structures, in S.K. Pal, T.S. Dillon & D.S. Yeung (editors), *Soft Computing and Case-Based Reasoning*, 115-146, Springer.

## **Sonstige wissenschaftliche Veröffentlichungen**

Berrar D., Lopes P., Davis J. & Dubitzky W. (2018). [Editorial: Machine Learning for Soccer](#). *Machine Learning* (Springer).

Dubitzky W. (2010), Editorial: Challenges in bioinformatics and computational biology, *Briefings in Bioinformatics*, 343- 344, 10(4).

Dubitzky W., Stankovski V. (2007), Editorial: Data mining in grid and Web services computing environments: Challenges and applications, *Future Generation Computer Systems*, 23(1), 31-33. doi: 10.1016/j.future.2006.05.001.

Dubitzky W. (2006), Editorial: Understanding the computational methodologies of systems biology, *Briefings in Bioinformatics*, 7(4), 315-317. doi:10.1093/bib/bbl044.

Huang C-H., Lanza V., Rajasekaran S. & Dubitzky W. (2005), Editorial: *Journal of Clinical Monitoring and Computing*, 19(4-5), 259-262.

Lopez J.A., Azuaje F., Prank K. & Dubitzky W. (2004), Editorial: Molecular and sub-cellular systems biology, *IEEE Transactions on Nano-Bioscience*, 3(3), 141-143.

Dubitzky W. (2004), Editorial: Data mining meets integrative biology – a symbiosis in the making, *OMICS: A Journal of Integrative Biology*, 8(2), 153-166.

Dubitzky W. & Azuaje F.J. (2003), Editorial – AI in the life sciences, *Artificial Intelligence Review*, 20(1-2), 7-11.

Dubitzky W. (2005), Review of Computational Modeling of Genetic and Biochemical Networks edited by Bower J.M. & Bolouri H. (2001), *BioMedical Engineering OnLine* 2005, 4:56, doi:10.1186/1475-925X-4-56

Dubitzky W. (2003), Review of Advances in the Evolutionary Synthesis of Intelligent Agents edited by Patel M.J., Honavar V. & Balakrishnan K. (2001), *IEEE Transactions on Neural Networks*, 14(4), 970, July 2003.

Berrar D.P., Sturgeon B., Bradbury I., Downes C.S. & Dubitzky W. (2003), Microarray data integration and machine learning techniques for lung cancer survival prediction, *Critical Assessment of Microarray Data Analysis (CAMDA 2003)*, Oral and Poster Presenters Abstracts, 43-54, Durham, North Carolina, USA.

Kochmann T.K., Dubitzky W., Flaig R.M. & Eils R. (2001), Towards a Hermeneutic knowledge management in science, Proc. of 222nd American Chemical Society National Meeting, CINF 36, Chicago, Illinois, USA. (Abstract)

Dubitzky W., Kochmann T.K., Flaig R.M. & Eils R. (2001), Implementation of a global computing scenario in science, Proc. of 222nd American Chemical Society National Meeting, CINF 33, Chicago, Illinois, USA. (Abstract)

Dubitzky W. (1998), Knowledge integration in case-based Reasoning: A concept-centered approach, PhD Thesis, University of Ulster, Northern Ireland, GB.

Montazemi A.R., Bayless S.J., Dubitzky W., Gupta K.M. & Klahr P. (1996), Industrial application of case-based reasoning systems, Proc. of 25th Northeast Decision Sciences Institute Meeting, p587, USA.

Granzow M., Berrar D., Dubitzky W., Schuster A., Azuaje F. & Eils R. (2001), Tumor classification by gene expression profiling: Comparison and validation of five clustering methods, in SIGBIO Newsletter Special Interest Group on Biomedical Computing of the ACM, ACM Press, 21(1), 16-22.

Dubitzky W. (1992), An introduction to case-based reasoning, Expert. Newsletter of the Expert System Centre (University of Ulster, Jordanstown, GB), 5: 5-7, UK.

## Anhang A: Some Research Interests in Machine Learning and Data Science

Below I outline various research areas related to data science/machine learning and biomedicine that I consider important. Essentially, these arise from my prior work which took me from an engineering perspective (computer science, artificial intelligence, machine learning, large-scale computing) to an experimental science perspective (bioinformatics and computational biology), and back (systems view, complexity, multiscale modeling, simulation and data science). These research interests center firmly on the integration and analysis of complex and heterogeneous biomedical data sets using statistical machine learning and modeling and simulation approaches.

In knowledge-rich domains like biomedicine, machine learning could offer a wealth of useful ways to approach problems that otherwise defy solution. Adopting machine learning in this area raises numerous new research challenges. First, partly because of the recent big data developments in the life sciences, we are witnessing a growing demand for novel statistical machine learning solutions. Second, there is a growing need to adopt a systematic experimental-science approach to machine learning research. Third, the bulk of machine learning research focuses on narrowly defined algorithmic and technical advances and data sets of limited scope, but the real-world impact of machine learning research is rarely assessed. Impact assessment is particularly important in biomedical applications. We need more research on machine learning that really matters, i.e. we need to develop and evaluate machine learning solutions in terms of their real-world impact. Fourth, emerging and future progress in biomedicine will increasingly rely on intelligent solutions from artificial intelligence, machine learning and other ICT areas – this is what I call *intelligent e-science*. We need to develop machine learning solutions in the context of the emerging scientific process in biomedical research. Fifth, systems medicine is an emerging interdisciplinary framework that aims to improve our understanding, prevention and treatment of complex diseases by integrating knowledge and data across multiple levels of biomedical organization. Such integration and analysis of complex, multiscale data requires a multiscale approach to systems medicine (multiscale machine learning, data science and modeling and simulation supported by multiscale computing). Sixth, feature engineering and representation learning are highly underestimated research areas in machine learning. Particularly, for scientific machine learning solutions in biomedicine these topics will become increasingly important in the near future.

### Statistical machine learning

Machine learning was originally concerned with the automated acquisition, use and evolution of knowledge for intelligent systems that are able to perform complex tasks such as reasoning, design, diagnosis, problem solving, planning, and language understanding. This type of machine learning typically employs symbolic knowledge structures (e.g. production rules, decision trees and logical formulae) and data sets of small to moderate size. Statistical learning refers to a set of tools for modeling and understanding complex data sets. Statistical learning approaches commonly employ sub-symbolic knowledge structures and techniques (such as lasso and sparse regression, classification and regression trees, support vector machines, artificial neural networks) and are often based on large and very large data sets. Recent developments in the field of big data have led to a fast growth in the field of statistical learning. Nowadays, the distinction between statistical learning and machine learning is blurred and the term statistical machine learning is often used. Statistical machine learning could be viewed as a hybrid discipline borrowing techniques from statistical learning and machine learning. The adoption of statistical machine learning especially in knowledge-rich domains and in science poses many new machine learning research challenges. These challenges include, but are not limited to, high-dimensional data, sparsity, semi-supervised learning, the relation between computation and risk, and structured prediction. Some of these challenges are briefly highlighted below.

Most statistical machine learning theory is based on asymptotic approximations that allow the sample size  $n$  to grow large. When the number of variables  $p$  in the model is large, this theory can be problematic, however. The small- $n$ -large- $p$  problem is known as the curse of dimensionality. Unless certain strong assumptions hold, the sample size  $n$  needs to grow exponentially with  $p$  to achieve good model performance (statistical curse of dimensionality). One important challenge in statistical machine learning is to develop relevant theory and methods when  $p$  grows with  $n$ . Such a theory should yield useful insights for real data sets with moderate sample sizes but large number of variables. Associated with the statistical curse of dimensionality is the computational curse of dimensionality – the computational burden of algorithms typically grows exponentially with  $p$ . We need to develop novel statistical machine learning algorithms and tools that are able to scale gracefully with a growing number of variables.

In a typical statistical machine learning problem, getting raw data is relatively easy, but producing labeled examples is time-consuming and expensive, since the labeling may require expensive experiments, such as clinical trials or human experts. The challenge of semi-supervised learning is to somehow leverage large amounts of unlabeled data in order to improve upon a learning algorithm that uses only labeled data. While some work on this problem exists in the classical machine learning community, little attention has been given to this in the statistical learning field.

Statistical machine learning research is usually aimed at finding algorithms that minimize the expected loss or cost (or error) of the learner. But these algorithms usually ignore the computational costs of machine learning. We need to develop a new statistical machine learning theoretical framework that combines prediction error with computational complexity.

In statistical machine learning, a structured prediction problem can be thought of as a multi-class problem with a large number of class labels, typically exponential in the number of variables. Developing estimators and efficient algorithms for structured prediction problems, the structure of the problem must be taken into account. For example, in part-of-speech tagging in natural-language processing, the type of a word in a sentence depends strongly on the type of previous words in the sentence. More complex forms of structured prediction exist. As more complex problems are being tackled in knowledge-rich domains, more research is needed in the area of structured prediction problems.

## **Statistical learning as experimental science**

Fundamental research in machine learning is inherently empirical, because the performance of machine learning algorithms is determined by how well their underlying assumptions match the structures and conditions in the real world. Hence, no amount of mathematical analysis can determine a priori whether a machine learning model will work well or not (this is known as the *no free lunch theorem* of machine learning). Hence, an experimental-science approach to machine learning is required. A machine learning researcher translates these assumptions into a set of learners (called the hypothesis space) and defines how the performance of a learner is to be evaluated. The researcher then implements these specifications using suitable computer software and hardware, and tests the performance on real-world data sets.

A great advantage of my excursion into life science is that I am very familiar with scientific process in experimental science. Machine learning researchers typically come from computer science, other engineering disciplines or mathematics. Therefore, they are normally not very familiar with the methodologies of experimental science. As a result, a great deal of machine learning research is conducted in a manner that does not fully embrace the principles and methodologies of experimental science. Typically, machine learning studies are designed around the following basic procedure: (1) Implement new algorithm. (2) Compare its performance (typically a kind of accuracy or error measure)

based on benchmark data sets to state-of-the-art algorithms. (3) Publish, if accuracy is higher (in most cases). This design of a machine learning study leaves a lot to be desired. First, it restricts assessment to a small selection of data and methods. Second, it limits assessment to only one criterion (typically a kind of accuracy score possibly associated with a p-value expressing statistical significance). Indeed, one of the main differences between machine learning and statistics is that machine learning places a strong emphasis on accuracy (of model predictions), whereas statistics typically emphasizes interpretability of models. Third, such a study design fails to take into account the idiosyncrasies (context) of many real-world problems. Arguably, the utility of this type of machine learning research methodology is limiting the progress of machine learning research.

Increasingly, because of the strong focus of machine learning on standard data sets (which are essentially viewed as matrices of numbers with their domain context ignored), performance measure typically report a kind of accuracy/error scalar associated with the p-value of a statistical significant test. Given the diversity of potentially interested people (particular researcher, the research community as a whole, end users of applications and, of course, referees for conference and journal papers), it is unreasonable to expect to capture all their concerns in a single scalar measure. With only the outcome of a significance test (usually: reject null hypothesis), we have no idea of the size of the actual difference between our measured values, however small the p-value. We need additional information. We can study the tables of results produced in a paper and make our own estimate of the difference, but this is not a statistically justified procedure. One way this problem could be addressed is by the use of confidence intervals. From these we can judge not statistical significance but also the size of the effect. Each researcher can then decide what effect is sufficiently large to make the approach interesting in the context of their work and problems. Another problem of requiring statistically significant results before a paper is published is that we do not see the whole picture. A survey of the literature would give an impression that there is stronger support for a particular algorithm than there actual is. Another issue is the strong reliance of machine learning research on standard benchmark data sets. The main advantage of commonly used machine learning benchmark data sets is our familiarity with them. However, this is also the major downside, as the very familiarity with these data sets potentially leads to over-fitting. Our knowledge encourages the writing of algorithms that are tuned to them. This is part of a larger concern about how well experimental results will generalize to other yet unseen problems.

Even when an experimental machine learning study has been carried out with great care, the conclusions that can be drawn from it are often very weak. Problematic components of the standard machine learning testing procedures include the performance measures used, the reliance on null hypothesis statistical testing, and the use of benchmark data sets. Hence, future machine learning research should adopt a much wider view in evaluating machine learning solutions. For example, from a *local* perspective, we would like to know how well our testing procedure predicts performance on future applications. From a *global* perspective, we would like to know how well our testing procedure encourages progress in our field. Present machine learning evaluation procedures are not as effective, from the local perspective, as people imagine. But it does not seem to be doing very well from the global perspective. Not only is it ineffective at filtering out dubious theories, the overly strong emphasis on testing discourages a broad, dynamic, and ultimately more fruitful dialogue within the machine learning community.

## **Machine learning that matters**

A large portion of machine learning research is inspired by challenging real-world problems in medicine, education, science, engineering, environment, economy, society, and so on. And yet, we still see a plethora of published machine learning papers that evaluate new solutions on a handful of isolated benchmark data sets. While these data sets may have originated in the real world, the results from the evaluation of the solutions are rarely communicated back to the origin. Quantitative improvements in performance are rarely accompanied by an assessment of whether those gains

matter in the world outside of machine learning research. A large part of current machine learning research is too focused on benchmark data sets and abstract performance metrics, and lacks proper follow-through with the real-world user community.

A growing number of machine learning studies present the results of a new algorithm based on synthetic data and/or standard data sets, such as those available from the UCI machine learning repository. The main advantage of such an approach is better comparison with other algorithms. However, in practice comparisons fail, because there are no standards for reproducibility. Machine learning studies vary considerably in methodology (partitioning of data, performance metrics, parameter settings), implementations (tools, libraries, platforms, programming languages, hardware), and reporting (language, formal specification, tables, charts and plots, intermediate data). Furthermore, interpretation of the result in the context of the domain problem is practically never made. Does a certain increase of classification performance over a small set of state-of-the-art solutions matter in the domain from which the data set comes from? Which classes were predicted well and which not so well, and what does this mean in the context of the domain?

There are also problems with how we measure performance. Commonly used performance metrics include classification accuracy, various forms of errors, F-measure, lift, area under the ROC curve, etc. These metrics are highly abstract in that they do not contain problem-specific details. While this allows for results to be compared across solutions and domains, these metrics tell us nothing about the impact different performance has or may have in the problem domain. For example, a 82% correct classification rate might be sufficient for a bank to decide whether or not to approve a small credit loan application, but to classify a sample as *cancer* or *no cancer*, perhaps a 98% or higher accuracy is required. The assumption of cross-domain comparability is an illusion created by abstract metrics that have the same numeric range, but not the same meaning. Such measures tell us nothing at all useful about generalization of impact across different data sets and across different domains. Beyond abstract measures of performance (and statistical significance and effect size plus confidence interval), we need to measure the true impact a novel machine learning technique has in the real world.

With the strong focus on benchmark data and abstract performance measures, a considerable part of machine learning research in the last 20 years has concentrated on the following simplified study procedure: (1) Set machine learning task. (2) Identify standard data sets. (3) Select or generate features. (4) Choose or develop algorithm. (5) Choose performance metrics. (6) Conduct experiments. (7) Publish results in machine learning journal or conference. Emerging and future machine learning research should be more ambitious and meaningful and design a machine learning study around problems that promise an impact of machine learning in the real world. Future machine learning researchers should adopt a more challenging and potentially more rewarding procedure, which could be stated in simplified form as follows (impact aspects highlighted in italics): (1) Identify a challenging real-world problem for which machine learning could potentially make a crucial contribution. This is difficult and is likely to involve interdisciplinary collaboration. (2) Determine what data needs to be collected. (3) Select and generate relevant features from the data. (4) Choose or develop a relevant learning procedure or algorithm. (5) Choose existing implementation or implement the learning algorithm. (6) Select an evaluation method and procedure. (7) Perform the machine learning experiments. (8) Apply impact measures and interpret results by involving domain experts where necessary. (9) Publicize results both to the machine learning and domain-specific communities. (10) Try to convince users to adopt the new technique and tool. All these steps are a necessary component of any machine learning research program that seeks to have a real impact on the world outside of machine learning.

My experience in evaluating and reviewing EU projects tells me that particularly the ICT community sometimes does not distinguish clearly between innovation and impact. Given the recent (since FP7)



emphasis on impact, this is quite remarkable. To facilitate machine learning research that matters, we need complement traditional performance measures with evaluation measures that enable direct assessment of the impact of novel machine learning solutions. Such methods should measure the money saved, life preserved, disease burden lowered, time conserved, effort reduced, quality of life improved, loss of biodiversity decreased, crimes prevented/solved, environmental pollution reduced, level of poverty reduced, unemployment lowered, effects of catastrophes and crises mitigated, recycling of waste increased, waste of energy and resources reduced, educational learning outcomes improved, quality of goods and services enhanced, economic competitiveness improved, effects of aging populations mitigated, social inequality lowered, and so on. Such impact measures will serve to refocus and restructure machine learning research efforts in terms of the problems we tackle, the data sets we use, the way we design machine learning studies, and the objective functions we define and use. They will also motivate machine learning researchers to report how a novel feature or the performance of a new machine learning solution translates to impact in the originating problem domain.

Clearly, this type of *impact machine learning* research is inherently interdisciplinary (and ultimately transdisciplinary), and will require a change in the way we communicate and publish machine learning research results.

### **Machine learning for integration/analysis of complex biomedical data sets**

Human health and disease are characterized by a complex interplay of multiple factors from the genome to the exposome. For many complex diseases, a sufficiently detailed understanding of the underlying mechanisms has remained elusive, and therefore the development of effective cures continues to be major challenge. As a result, the socioeconomic burden (morbidity, mortality, financial cost) of complex diseases remains high and is likely to grow within Europe's aging population. Systems medicine is an emerging interdisciplinary framework that aims to improve our understanding, prevention and treatment of complex diseases by integrating knowledge and data across multiple levels of biomedical organization. Such integration and analysis of complex, multiscale data requires a multiscale approach to machine learning, data science and modeling and simulation supported by multiscale computing. In 2016 I have led an EU H2020 COST network grant proposal on Open Multiscale Systems Medicine which aims to integration research in this area.

### **Feature engineering and knowledge integration in machine learning**

The selection of relevant features, and the elimination of irrelevant ones, is a central problem in machine learning. Before an induction algorithm can move beyond the training data to make predictions about novel test cases, it must decide which attributes to use for these predictions and which to ignore. Intuitively, one would like the learner to use only those attributes that are *relevant* to the target concept. In other words, the success of machine learning algorithms generally depends on (input) data representation – different data representations can entangle and hide more or less the different explanatory factors of variation behind the data. In knowledge-rich domains such as science, suitable data representations are often created by preprocessing the measured data with specific domain knowledge. This feature engineering approach is one of the most important ways to integrate domain knowledge into the machine learning process. Another approach that aims to automate the construction of relevant features is referred to as feature learning, representation learning or deep learning. I think in particular for complex scientific data sets (e.g. in biomedicine), representation learning research and feature engineering could potentially lead to fundamental breakthroughs.

### **Intelligent e-biomedicine**

Since my degree in electrical/telecommunication engineering, my career has been strongly influenced by my interest in natural science as well as computer science. Reflecting on the increasingly fuzzy boundaries between science and technology, ICT areas such as artificial intelligence and machine learning could be viewed as the key components in the emerging field of enhanced science or e-

*science* (e.g. e-biomedicine). While e-science as an interdisciplinary framework (in its conventional form known as computational science or scientific computing) has been around for many years, there is still a bias towards compute-intensive modeling and simulation approaches (systems dynamics, dynamical and complex systems, control theory, and so on). The field of computational biology is a good example of this type of R&D. Recently, there has been a realization that data-intensive and intelligent technologies (machine learning, AI, computational intelligence, computational creativity, etc.) could usefully complement the standard e-science framework. For instance, the field of bioinformatics and computational biology has a long tradition in incorporating intelligent techniques into the arsenal of tools, albeit usually not in the context of large-scale computing and very large data (this is changing, however). The data-intensive element in e-science is sometimes referred to as the fourth paradigm (complementing the first three science paradigms of experiment, theory and computational science). One could argue that modeling, analysis and prediction based on the methods and tools of intelligent technologies may develop into the fifth paradigm of the evolving scientific process. I refer to this emerging framework or scientific process as: *intelligent e-science*, which could be defined as follows:

1. Experiment (first paradigm)
2. Theory (second paradigm)
3. Computational science (third paradigm)
4. Data-intensive science (fourth paradigm) – nowadays called data science
5. Science enabled by intelligent technologies (fifth paradigm)

The intelligent e-science framework as outlined above should be understood as a transdisciplinary scientific paradigm in which researchers work jointly using a shared conceptual framework and combined disciplinary-specific approaches to address complex R&D problems. Clearly, this paradigm will require considerable changes transcending the mind-set and culture of current science and education environments. The present scientific mind-set and culture is characterized by an interdisciplinary approach (where researchers work jointly but still from a disciplinary-specific basis) which has evolved from multidisciplinary science (researchers working in parallel or sequentially from disciplinary-specific base) in the past century.

It is my view that intelligent e-science – and its domain-specific versions like intelligent e-biomedicine – is likely to be at the heart of a future, highly transdisciplinary science. This type of science is likely to revolve around the processing and analyzing large amounts of data (big data). Research in the field of statistical machine learning will probably play a major role in developing this way of doing science. At present we are designing and constructing the foundation of this kind of future science. I would think that I have the qualification, experience and standing to make a constructive contribution in laying this foundation.

### **Other AI / machine learning topics of interest**

- Deep learning
- Explainable AI and incorporation of domain knowledge in machine learning
  - (a) Incorporate scientific (neurobiological) constraints on information processing and learning in neural network architectures.
  - (b) Equation-constrained modeling
  - (c) Explain the behavior/decisions of AI algorithms.
  - (d) Combine simulation science with machine learning.
- Uncertainty quantification
- Reliable models
- Scalable/distributed learning

- Transfer learning for cross-discipline applications
- Continuous learning
- Multitask learning
- Reinforcement learning
- Generative models
- Ethics in AI
- AI and politics