DKG 2015

Institut für Geographie und Regionalforschung
aau.at | geo.aau.at

ALPEN-ADRIA UNIVERSITÄT
KLAGENFURT | WIEN GRAZ

Data, Concepts and Models of a Spatial Data Science by Linking Geography, GI-Science and Data Analytics

Peter Mandl
Institut für Geographie und Regionalforschung,
Universität Klagenfurt

DKG 2015, Berlin, 02.10.2015

RE-FS-14 Theoretische, methodische und disziplinäre Reflexion

Data, Concepts and Models of

- **Geography:** (big) geo-data, concepts, models, location analytics, spatial data mining, visual analytics, …
- GI-Science: semantic web, linked open data, GIS-modeling, …
- Data Analytics: methods from quantitative geography extended by visual, predictive, prescriptive and explorative aspects, business analytics, data science

(Big) Geo-Data today

- Books, **literature** data bases
- Digital **maps** and web maps, media repositories (images, films, videos)
- Volunteered (VGI) or **citizen** data
- **Statistical** data, geodata-repositories
- **Remote sensing** data, (real time) **sensor data**
- **Communication** data and metadata
- **Personalized data** (self tracking data, shopping cards, social networks, …)
- **Geodatainfrastructures**, GeoWeb,
- …

Most having **spatial reference** (locational information, addresses, geotags), **time reference** and **cross-linking**

BIG DATA are today (Definition 2015)

- A valuable and ubiquitous **datapool (structured, unstructured,**
- **Often** useable and **combinable**
- Suitable for **complex** queries and analyses
- Mostly similar to a **population** (not a sample)
- **Blurred**
- Highly **up-to-date**
- Usable for **inductive model building**, for **prediction** and **prescription**
- Primarily useful to answer „**what**“ and less „why“ questions

Modelling categories in (Human-) Geography

- **Description** (visually, statistically, Spatial Data Mining)
- **Classification** (supervised, unsupervised, ...)
- **Location / Allocation** (sites, paths, hinterländer, ...)
- **Assessment** (MADM, MODM, ...)
- **Interaction** (Wilson models)
- **Diffusion** (Markov, differential equations, CA)
- …

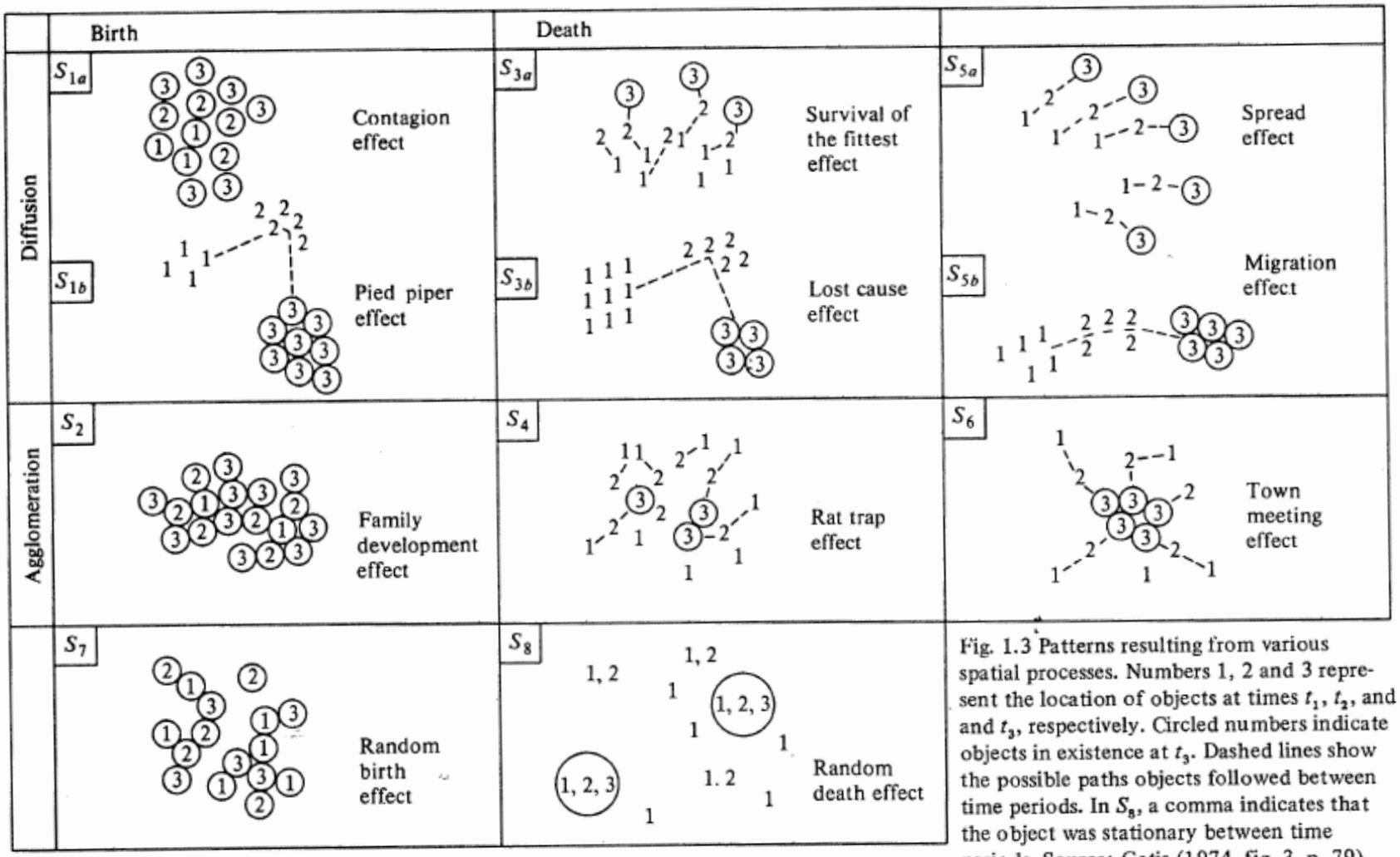
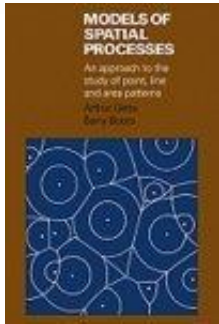


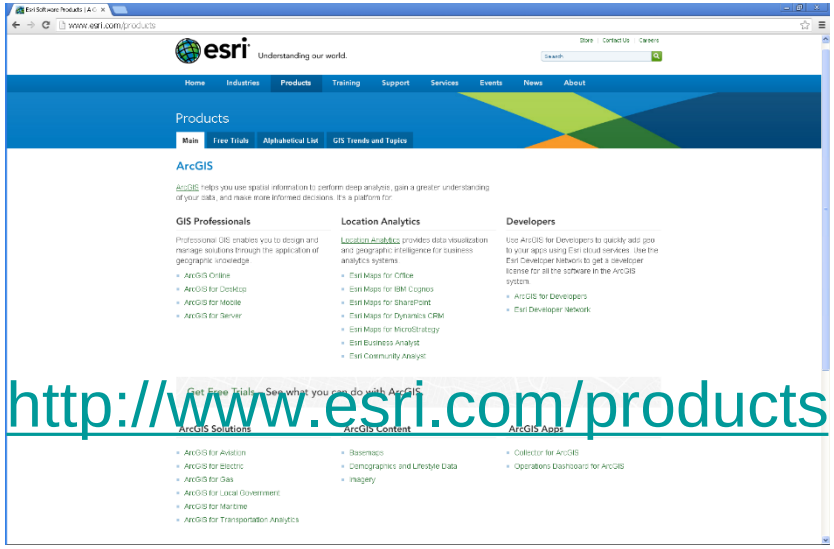
Fig. 1.3 Patterns resulting from various spatial processes. Numbers 1, 2 and 3 represent the location of objects at times t_1, t_2, and and t_3, respectively. Circled numbers indicate objects in existence at t_3. Dashed lines show the possible paths objects followed between time periods. In S_8, a comma indicates that the object was stationary between time periods. Source: Getis (1974, fig. 3, p. 79)

Getis & Boots 1978: Models of Spatial Processes

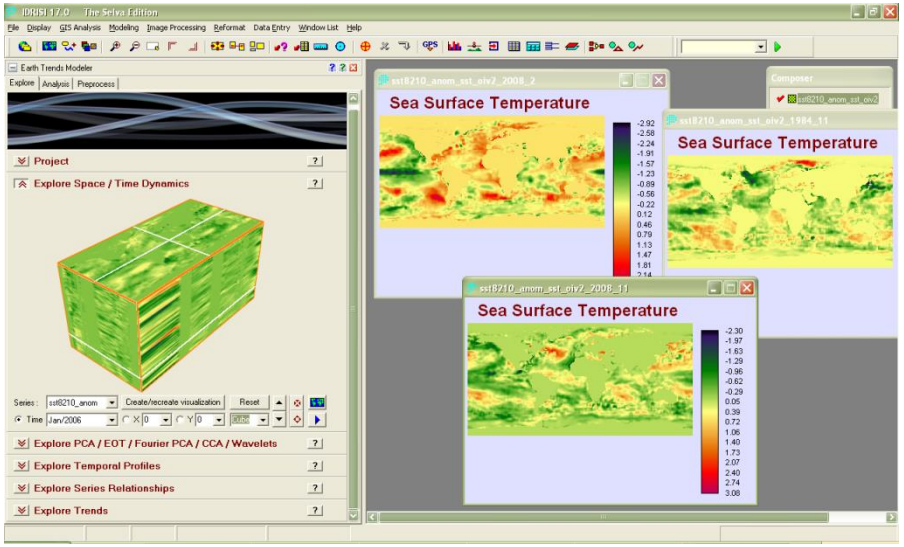


Today: Location Analytics

- Explorative spatial data analysis
- Visual Analytics
- Spatial Data Mining
- Multivariate Quantitative Geography (e.g. Crime Analysis, Spatial Econometrics, Spatial Data Analysis)
- Locational analysis
- Space/Time Analysis (e.g. Earth Trends Modeller)



http://www.esri.com/products

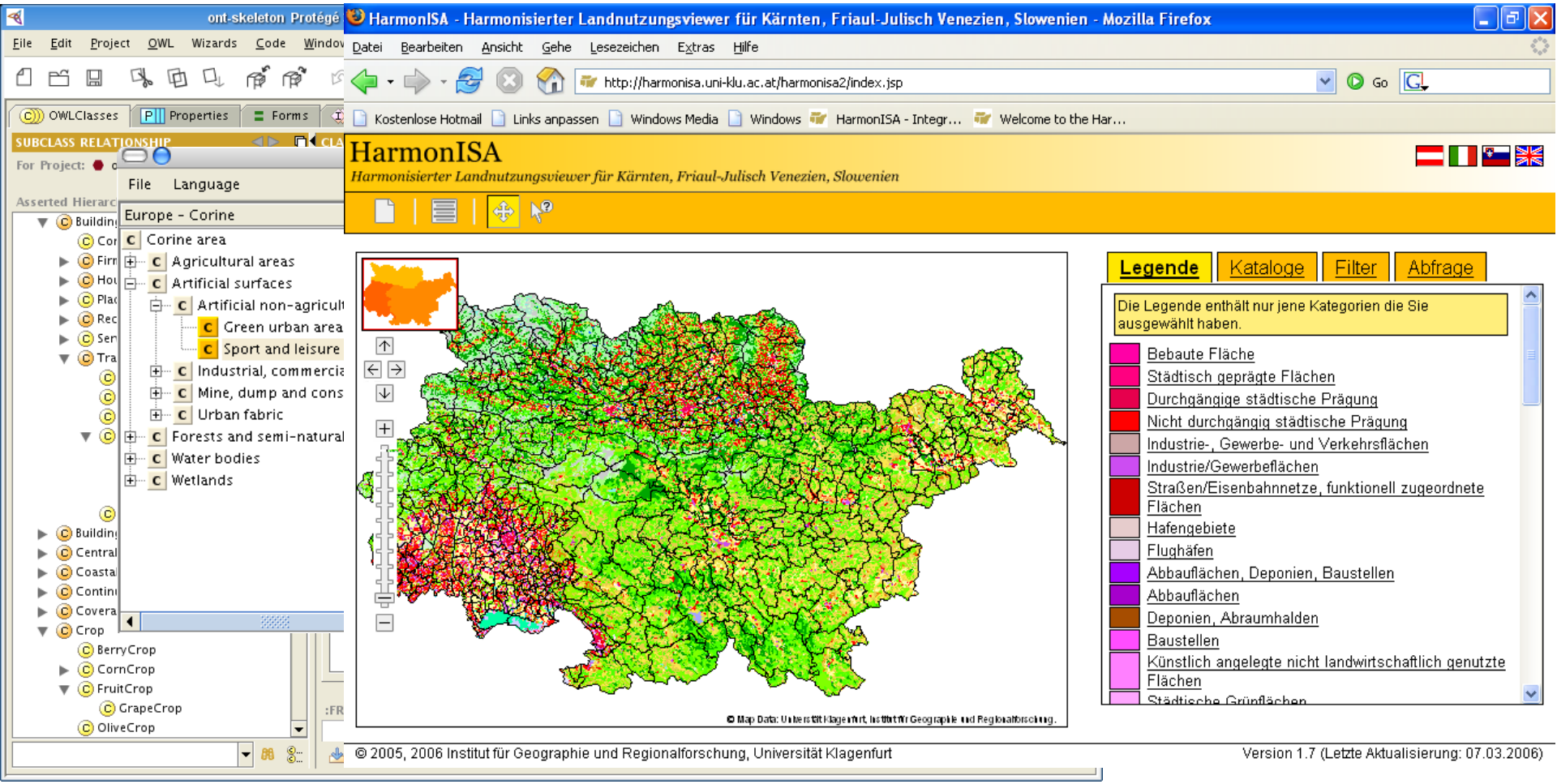


Data, Concepts and Models of

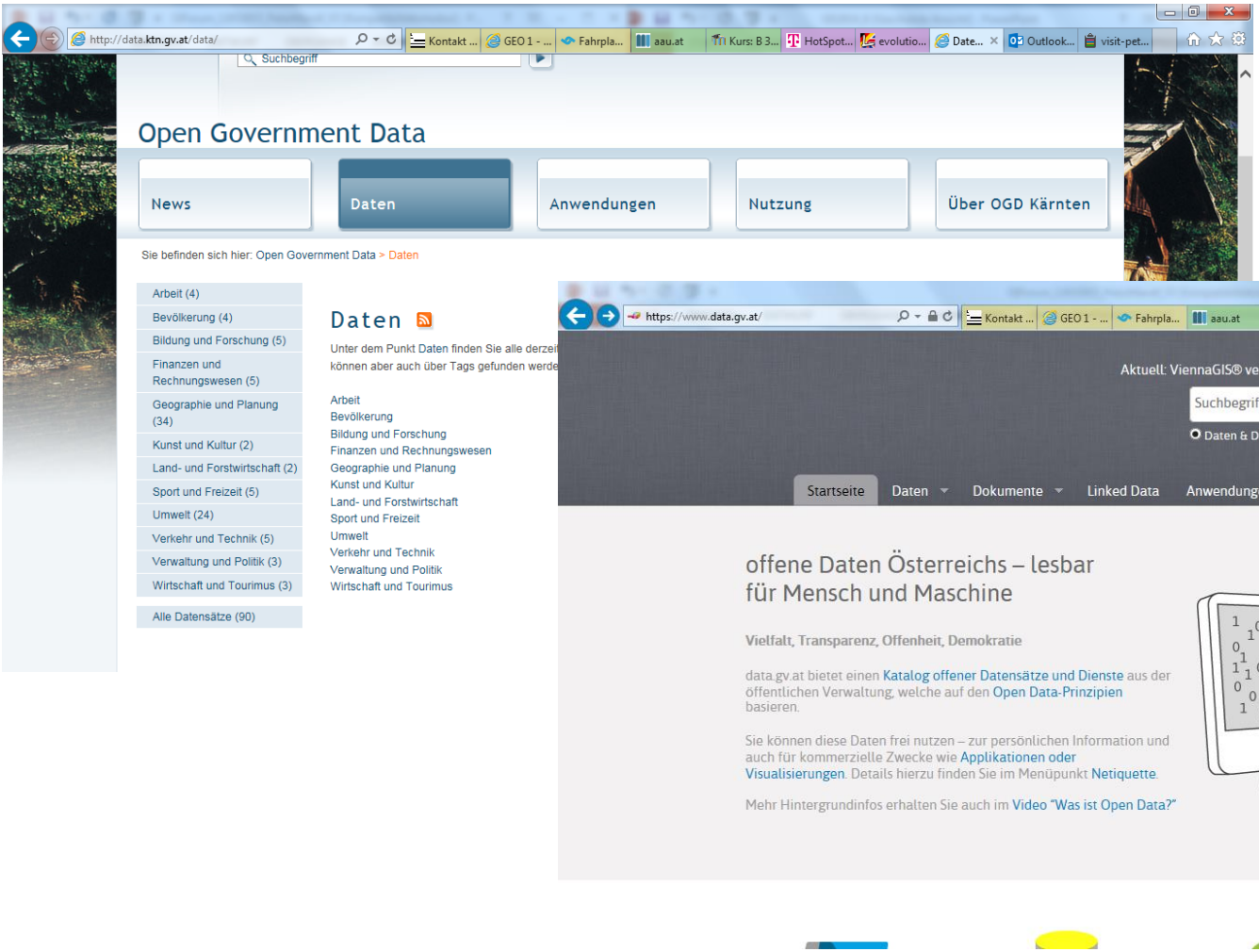
- **Geography:** (big) geo-data, concepts, models, location analytics, spatial data mining, visual analytics, …
- **GI-Science:** semantic web, linked open data, GIS-modeling, …
- Data Analytics: methods from quantitative geography extended by visual, predictive, prescriptive and explorative aspects, business analytics, data science

HarmonISA Project

harmonisa.aau.at

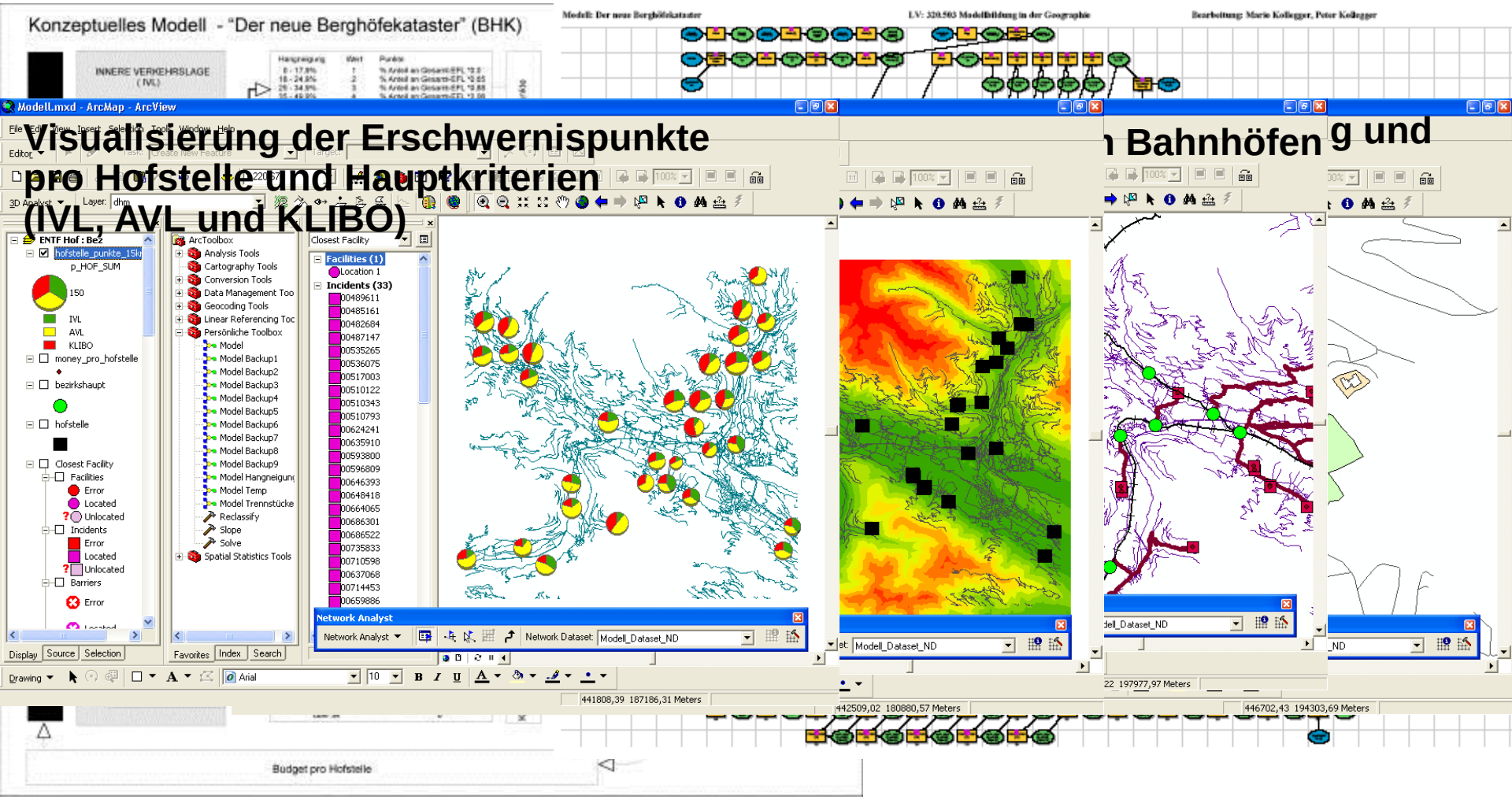


data.ktn.gv.at, data.gv.at



http://www.opengeospatial.org/standards

Model „Der neue Berghöfekataster“
Mario Kollegger, Peter Kollegger



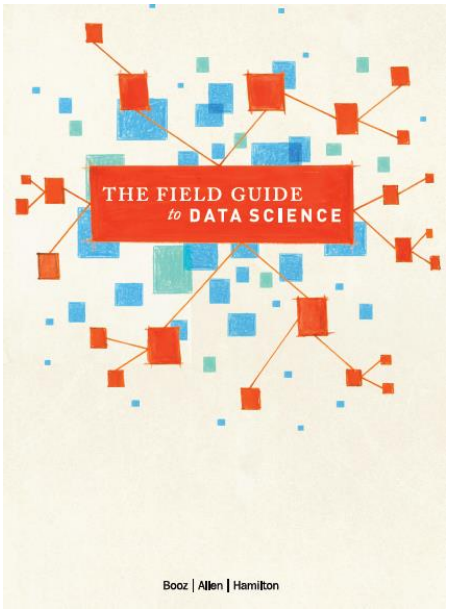
Data, Concepts and Models of

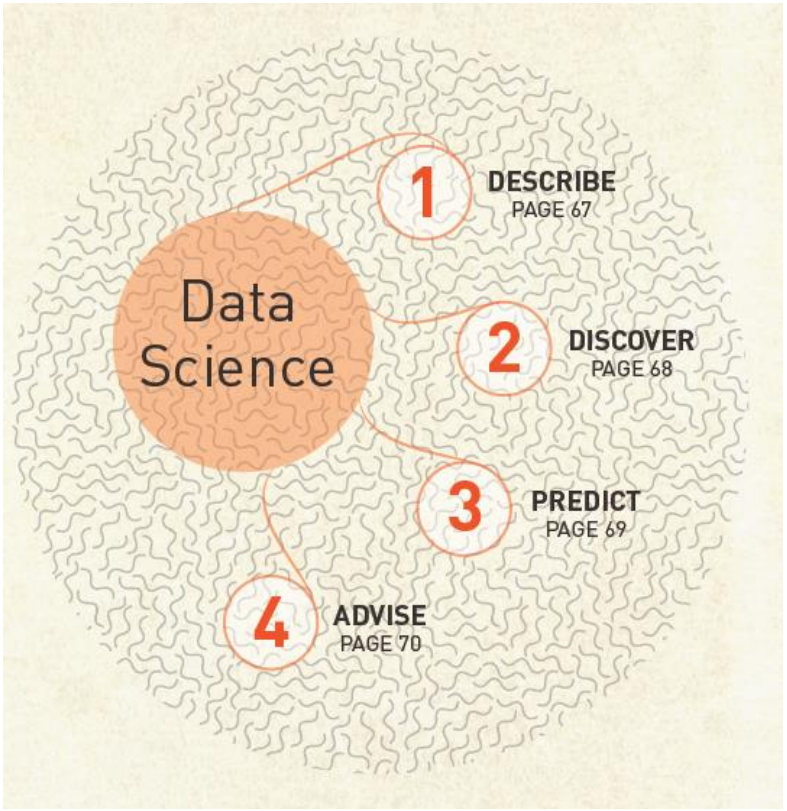
- Geography: (big) geo-data, concepts, models, location analytics, spatial data mining, visual analytics, …
- GI-Science: semantic web, linked open data, GIS-modeling, …
- **Data Analytics:** methods from quantitative geography extended by visual, predictive, prescriptive and explorative aspects, business analytics, **data science**

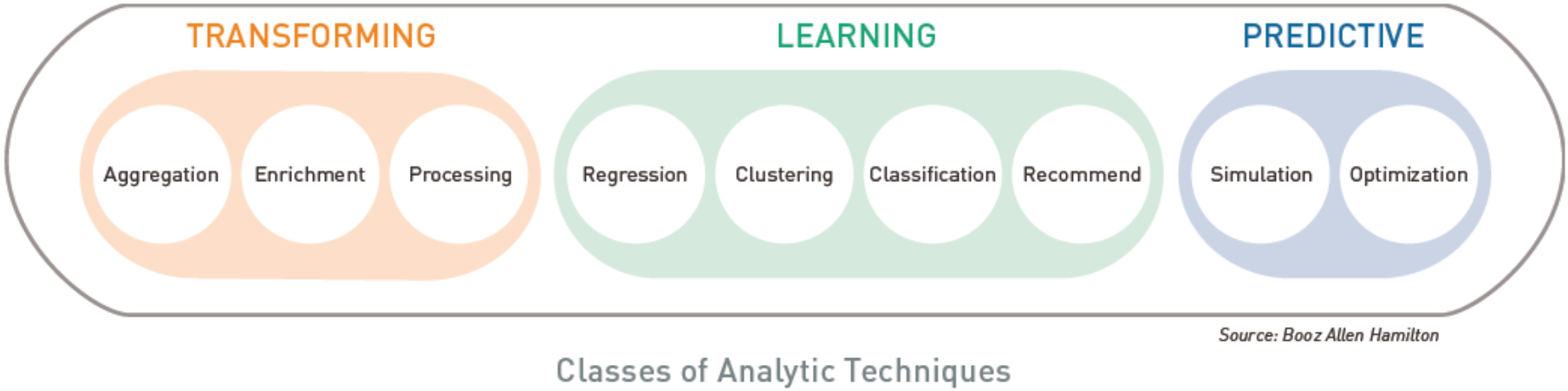
Terms

- Analytics, Forecasting
- Descriptive, predictive, prescriptive analytics and modelling
- Data Warehouse
- Webintelligence, Business Intelligence
- Knowledge discovery
- Data Science

Methods of a „Data Science“

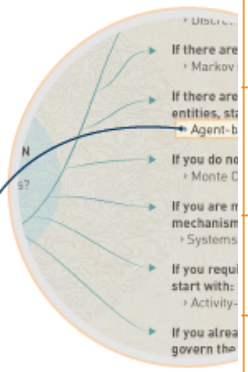






Source: Booz Allen Hamilton

Classes of Analytic Techniques



Algorithms or Method Name	Descri
Agent Based Simulation	Simulates the and interaction autonomous ag
Collaborative Filtering	Also known as 'Recommendat or eliminate ite set by compari of actions agai performed by u similar items b used them or s based on the it
Coordinate Transforma-tion	Provides a diff perspective on
Design of Experiments	Applies control experiments to effects on syst caused by char

Algorithms or Method Name	Description	
Differential Equations	Used to express relationships between functions and their derivatives, for example, change over time.	Differen models themsel tested w system t
Discrete Event Simulation	Simulates a discrete sequence of events where each event occurs at a particular instant in time. The model updates its state only at points in time when events occur.	Discrete analyzin producti determi as differ Optimiz gain effi
Discrete Wavelet Transform	Transforms time series data into frequency domain preserving locality information.	Offers ve localizat transfor and loca
Exponential Smoothing	Used to remove artifacts expected from collection error or outliers.	In compa where p exponen decreasi
Factor Analysis	Describes variability among correlated variables with the goal of lowering the number of unobserved variables, namely, the factors.	If you su influenc try facto
Fast Fourier Transform	Transforms time series from time to frequency domain efficiently. Can also be used for image improvement by spatial transforms.	Filtering more eff noise ca observin
Format Conversion	Creates a standard representation of data regardless of source format. For example, extracting raw UTF-8 encoded text from binary file formats such as Microsoft Word or PDFs.	There ar package can inte notable
Gaussian Filtering	Acts to remove noise or blur data.	Can be u noise fro
Generalized Linear Models	Expands ordinary linear regression to allow for error distribution that is not normal.	Use if th not follo
Genetic Algorithms	Evolves candidate models over generations by evolutionary inspired operators of mutation and crossover of parameters.	Increasi in consi requires Calculat is strong candida
Grid Search	Systematic search across discrete parameter values for parameter exploration problems.	A grid a visualize whether

Algorithms or Method Name	Description
Hidden Markov Models	Models sequential data by determining the discrete latent variables, but the observables may be continuous or discrete.
Hierarchical Clustering	Connectivity based clustering approach that sequentially builds bigger (agglomerative) or smaller (divisive) clusters in the data.
K-means and X-means Clustering	Centroid based clustering algorithms, where with K means the number of clusters is set and X means the number of clusters is unknown.
Linear, Non-linear, and Integer Programming	Set of techniques for minimizing or maximizing a function over a constrained set of input parameters.
Markov Chain Monte Carlo (MCMC)	A method of sampling typically used in Bayesian models to estimate the join distribution of parameters given the data.
Monte Carlo Methods	Set of computational techniques to generate random numbers.
Naïve Bayes	Predicts classes following Bayes Theorem that states the probability of an outcome given a set of features is based on the probability of features given an outcome.
Neural Networks	Learns salient features in data by adjusting weights between nodes through a learning rule.
Outlier Removal	Method for identifying and removing noise or artifacts from data.
Principal Components Analysis	Enables dimensionality reduction by identifying highly correlated dimensions.

Algorithms or Method Name	Description	Tips From the Pros	References and Papers We Love to Read
Regression with Shrinkage (Lasso)	A method of variable selection and prediction combined into a possibly biased linear model.	There are different methods to select the lambda parameter. A typical choice is cross validation with MSE as the metric.	Tibshirani, Robert. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society. Series B (Methodological)* 58.1 (1996): 267-288. Print.
Sensitivity Analysis	Involves testing individual parameters in an analytic or model and observing the magnitude of the effect.	Insensitive model parameters during an optimization are candidates for being set to constants. This reduces the dimensionality of optimization problems and provides an opportunity for speed up.	Saltelli, A., Marco Ratto, Terry Andres, Francesca Campolongo, Jessica Cariboni, Debora Gatelli, Michaela Saisana, and Stefano Tarantola. *Global Sensitivity Analysis: the Primer.* New Jersey: John Wiley & Sons, 2008. Print.
Simulated Annealing	Named after a controlled cooling process in metallurgy, and by analogy using a changing temperature or annealing schedule to vary algorithmic convergence.	The standard annealing function allows for initial wide exploration of the parameter space followed by a narrower search. Depending on the search priority the annealing function can be modified to allow for longer explorative search at a high temperature.	Bertsimas, Dimitris, and John Tsitsiklis. "Simulated Annealing." *Statistical Science*. 8.1 (1993): 10-15. Print.
Stepwise Regression	A method of variable selection and prediction. Akaike's information criterion *AIC* is used as the metric for selection. The resulting predictive model is based upon ordinary least squares, or a general linear model with parameter estimation via maximum likelihood.	Caution must be used when considering Stepwise Regression, as over fitting often occurs. To mitigate over fitting try to limit the number of free variables used.	Hocking, R. R. "The Analysis and Selection of Variables in Linear Regression." *Biometrics*. 32.1 (March 1976): 1-49. Print.
Stochastic Gradient Descent	General-purpose optimization for learning of neural networks, support vector machines, and logistic regression models.	Applied in cases where the objective function is not completely differentiable when using sub-gradients.	Witten, Ian H., Elbe Frank, and Mark A. Hall. *Data Mining: Practical Machine Learning Tools and Techniques*. Massachusetts: Morgan Kaufmann, 2011. Print.
Support Vector Machines	Projection of feature vectors using a kernel function into a space where classes are more separable.	Try multiple kernels and use k-fold cross validation to validate the choice of the best one.	Hsu, Chih-Wei, Chih-Chung Chang, and Chih-Jen Lin. "A Practical Guide to Support Vector Classification." *National Taiwan University*. 2003. Print.
Term Frequency Inverse Document Frequency	A statistic that measures the relative importance of a term from a corpus.	Typically used in text mining. Assuming a corpus of news articles, a term that is very frequent such as "the" will likely appear many times in many documents, having a low value. A term that is infrequent such as a person's last name that appears in a single article will have a higher TD IDF score.	Ingersoll, Grant S., Thomas S. Morton, and Andrew L. Farris. *Taming Text: How to Find, Organize, and Manipulate It*. New Jersey: Manning, 2013. Print.
Topic Modeling (Latent Dirichlet Allocation)	Identifies latent topics in text by examining word co-occurrence.	Employ part-of-speech tagging to eliminate words other than nouns and verbs. Use raw term counts instead of TF/IDF weighted terms.	Blei, David M., Andrew Y. Ng, and Michael I Jordan. "Latent Dirichlet Allocation." *Journal of Machine Learning Research*. 3 (March 2003): 993-1022. Print.
Tree Based Methods	Models structured as graph trees where branches indicate decisions.	Can be used to systematize a process or act as a classifier.	James, G., D. Witten, T. Hastie, and R Tibshirani. *Tree-Based Methods. In An Introduction to Statistical Learning*. New York: Springer, 2013. Print.
Wrapper Methods	Feature set reduction method that utilizes performance of a set of features on a model, as a measure of the feature set's performance. Can help identify combinations of features in models that achieve high performance.	Utilize k-fold cross validation to control over fitting.	John, G. H, R Kohavi, and K. Pfleger. "Irrelevant Features and the Subset Selection Problem." Proceedings of ICML-94, 11th International Conference on Machine Learning. New Brunswick, New Jersey. 1994. 121-129. 59. Conference Presentation.

Compiled by: Booz Allen Hamilton

Cross Industry Standard Process for Data Mining (CRISP-DM)

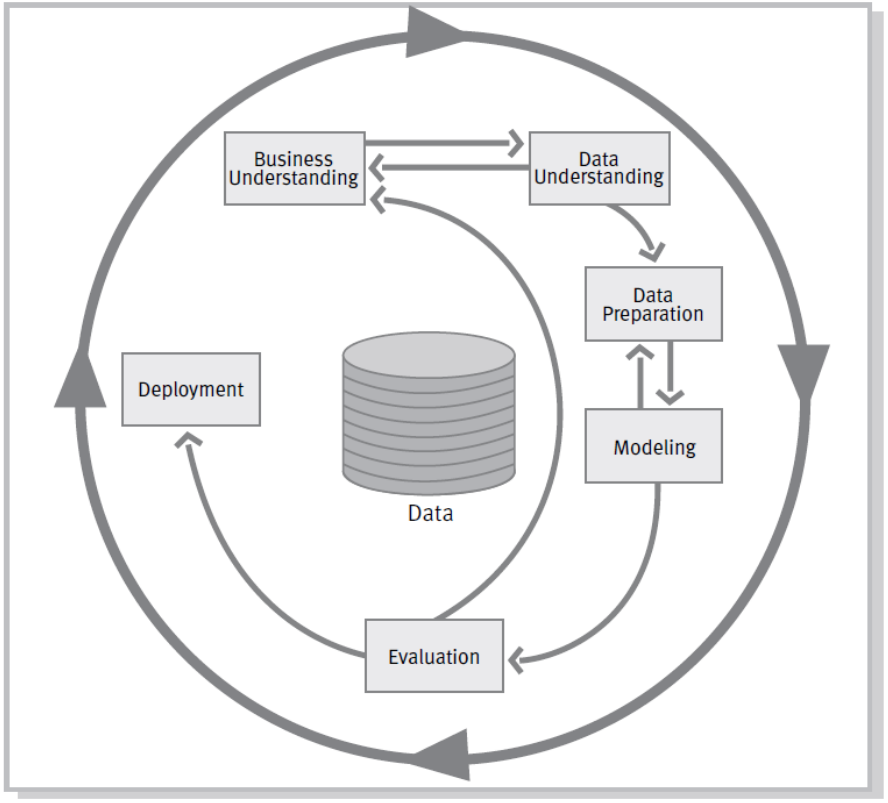


Figure 2: Phases of the CRISP-DM reference model

Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer and Rüdiger Wirth (2000): **CRISP-DM 1.0 Step-by-step data mining guide.** SPSS Inc.

Development of a Spatial Data Science

In a Spatial Data Science

spatio-temporal problems using established and new tools from the mentioned linked research fields are solved

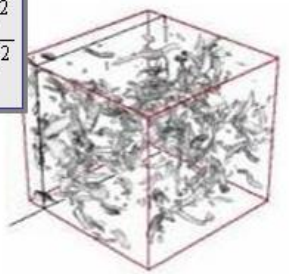
What is to do?

- Use New Science Paradigms
  - 4th scientific paradigm of an e-science (data exploration approach)
  - http://research.microsoft.com/en-us/um/people/gray/talks/NRC-CSTB_eScience.ppt
- Design Suitable Concepts and Frameworks
  - http://www.esri.com/news/releases/12-3qtr/carl-steinitz-explains-geodesign-process-in-new-esri-press-book.html
- Use Analytics Tools and Methods
  - rapidminer.com/
- Integrate Visualization
  - CommonGIS
  - http://www.iais.fraunhofer.de/1871.html
- Apply More Spatio-Temporal Modeling
  - IDRISI or TerrSet
  - www.clarklabs.org/

Use New Science Paradigms

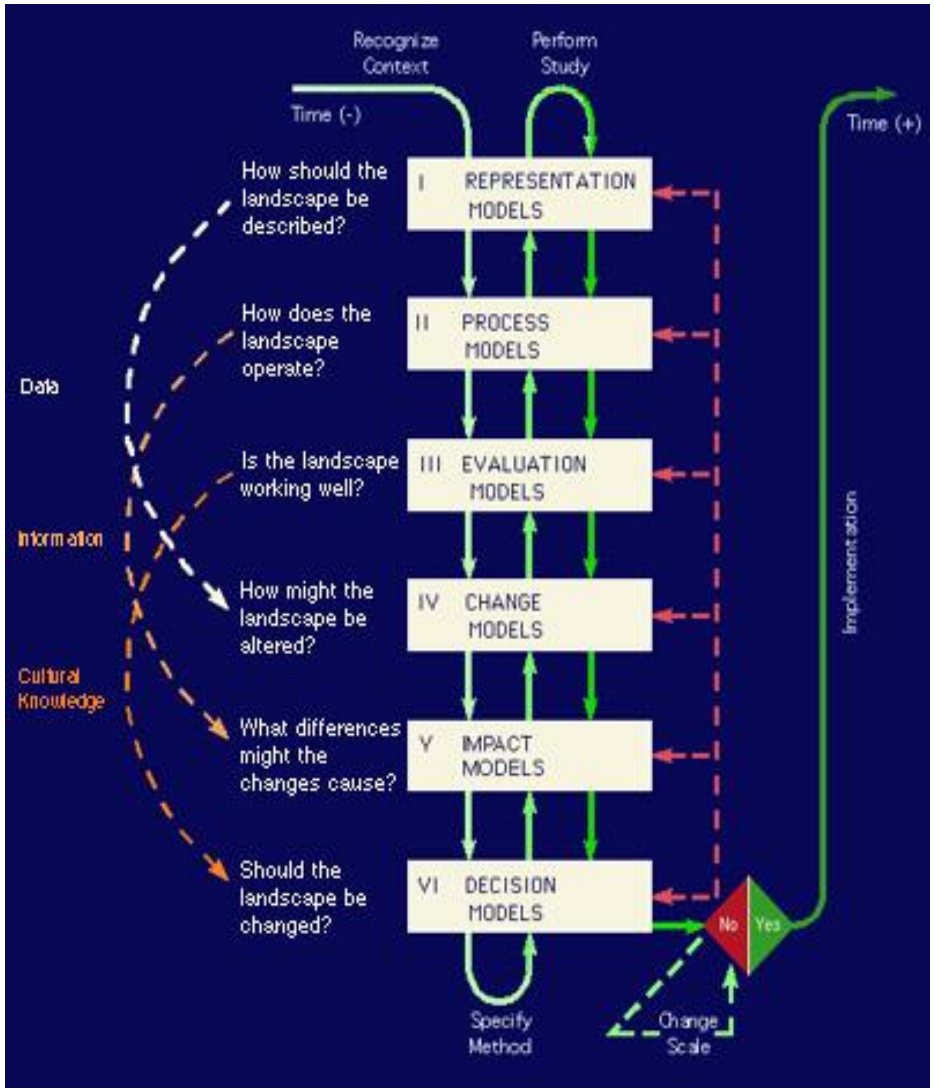
- Thousand years ago:
  science was **empirical**
  describing natural phenomena
- Last few hundred years:
  theoretical branch
  using models, generalizations
- Last few decades:
  a **computational** branch
  simulating complex phenomena
- Today:
  data exploration (eScience)
  unify theory, experiment,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

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



Jim Gray, Alex Szalay, *eScience—A Transformed Scientific Method*, Presentation to the Computer Science and Technology Board of the National Research Council, Mountain View, CA, 11 January 2007; see http://research.microsoft.com/en-us/um/people/gray/talks/NRC-CSTB_eScience.ppt.

Design Suitable Concepts and Frameworks

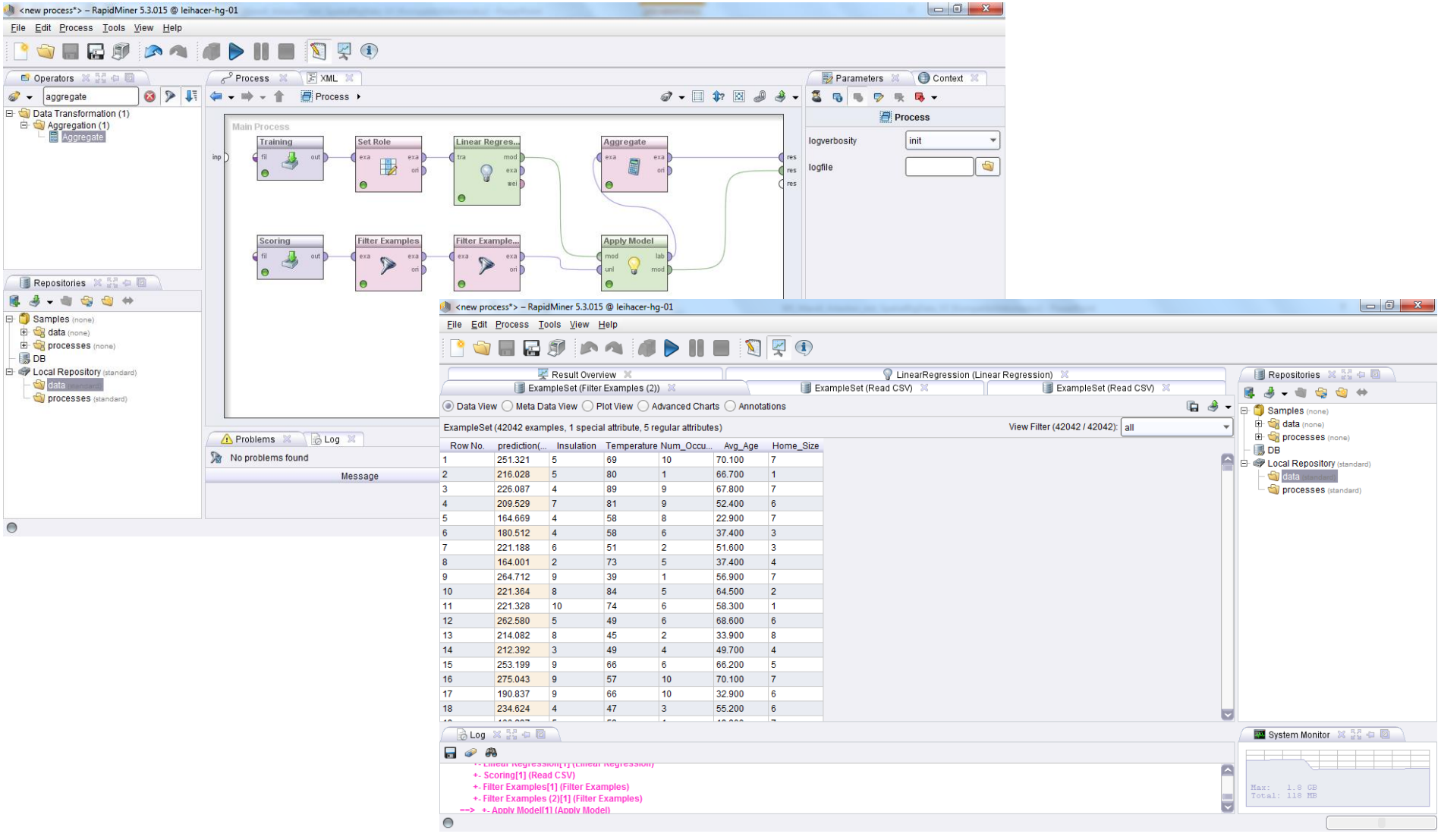


http://www.esri.com/news/arcwatch/0412/a-conversation-with-carl-steinitz.html

http://www.esri.com/news/releases/12-3qtr/carl-steinitz-explains-geodesign-process-in-new-esri-press-book.html

Use Analytics Tools and Methods

rapidminer.com/



Other Geo Tools

„Geo“-Software, which is based on Big Data approaches - like Spatial Data Mining or Geospatial Visual Analytics:

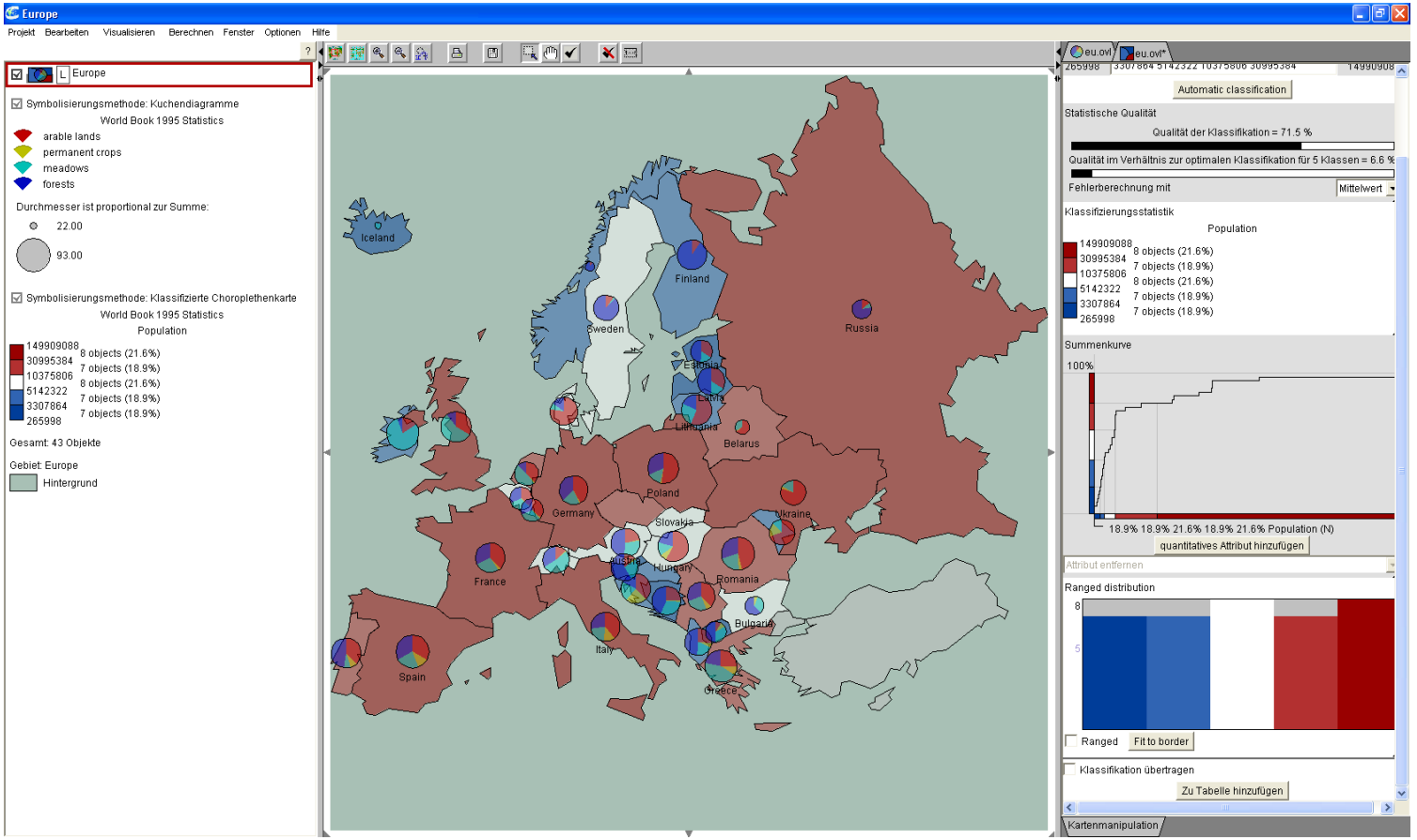
http://www.csiss.org/clearinghouse/select-tools.php

Special example: Common GIS
http://www.iais.fraunhofer.de/1865.html
(Educational Version is free of charge)

Integrate Visualization

CommonGIS

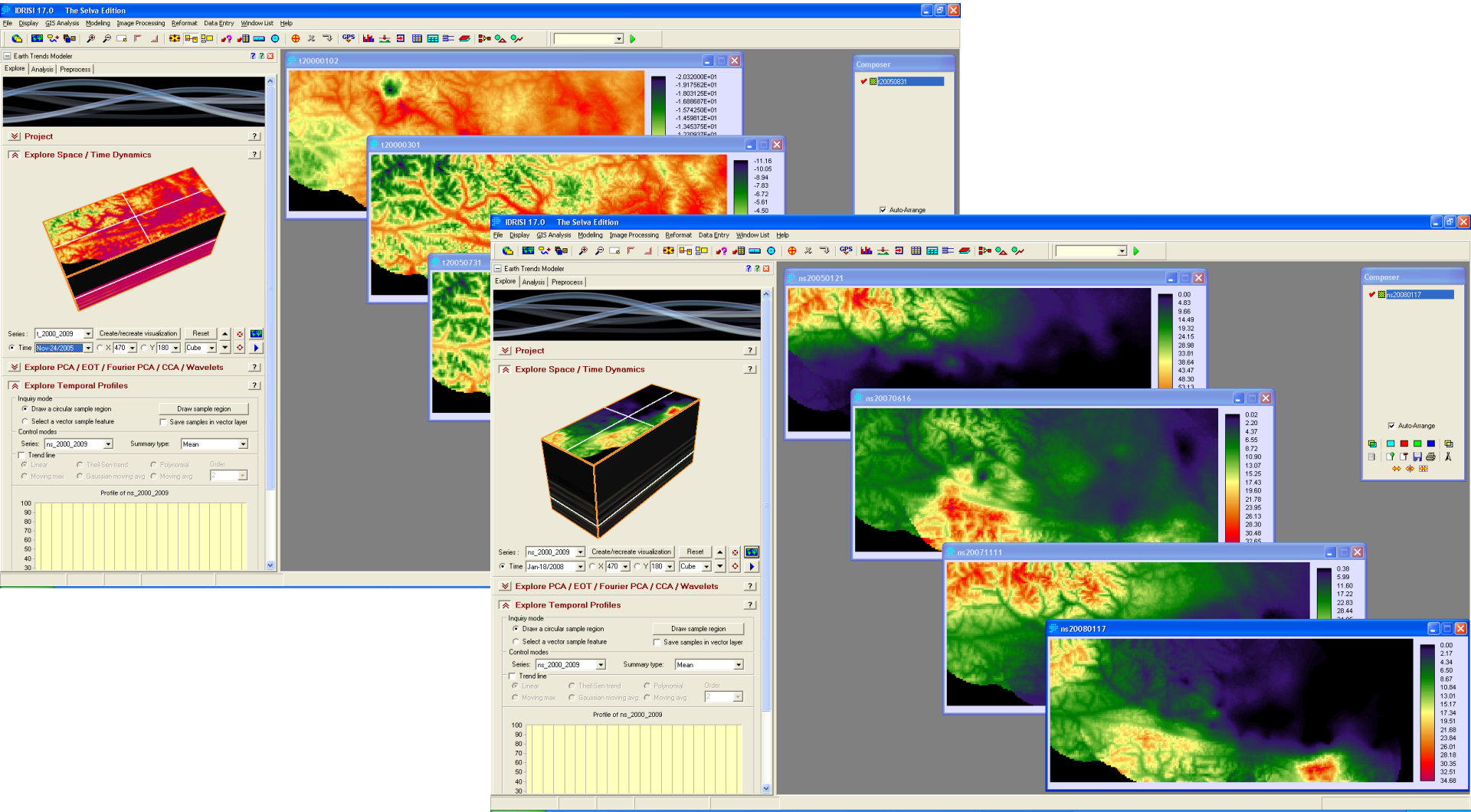
http://www.iais.fraunhofer.de/1871.html

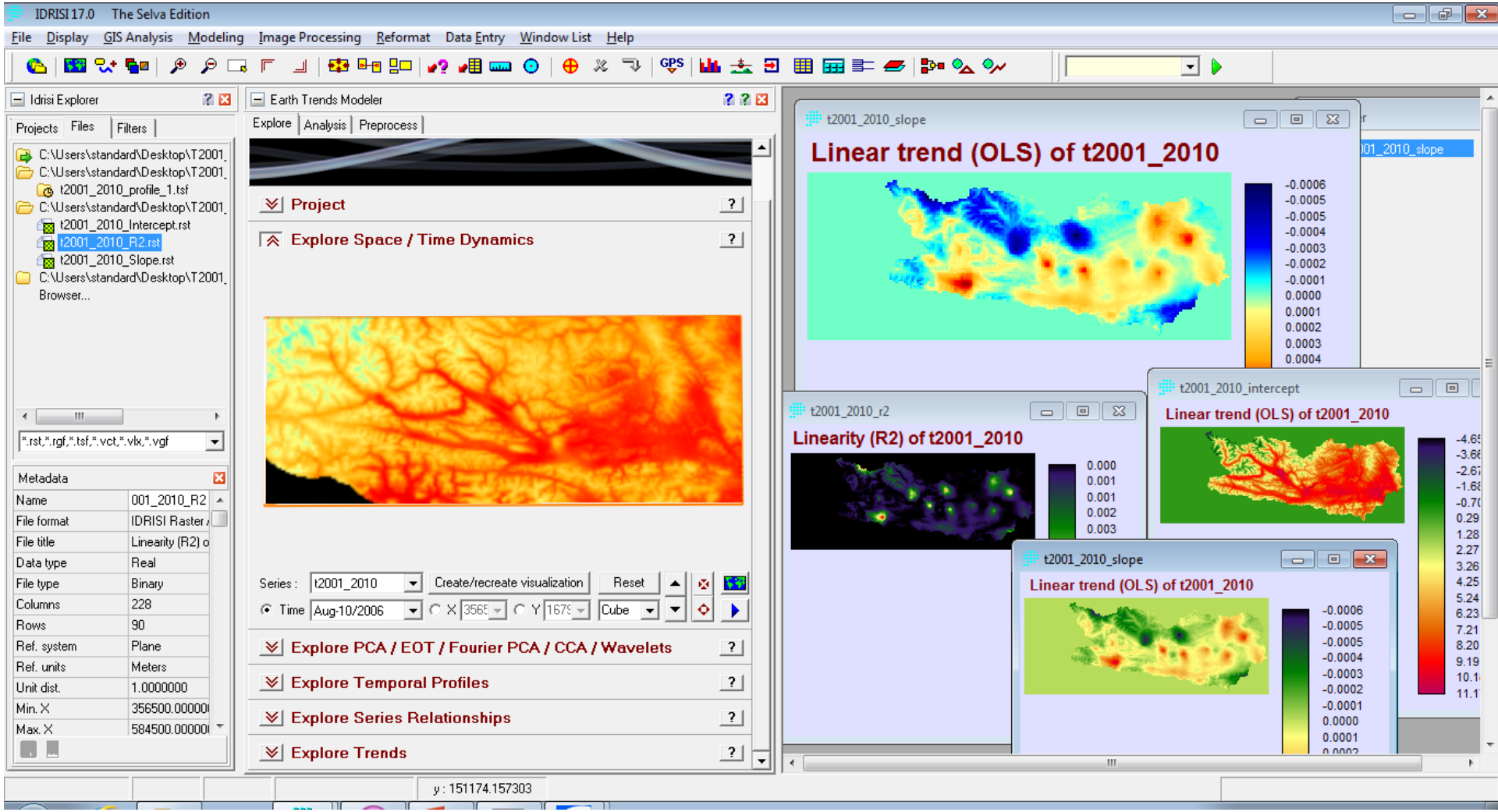


Apply More Spatio-Temporal Modeling

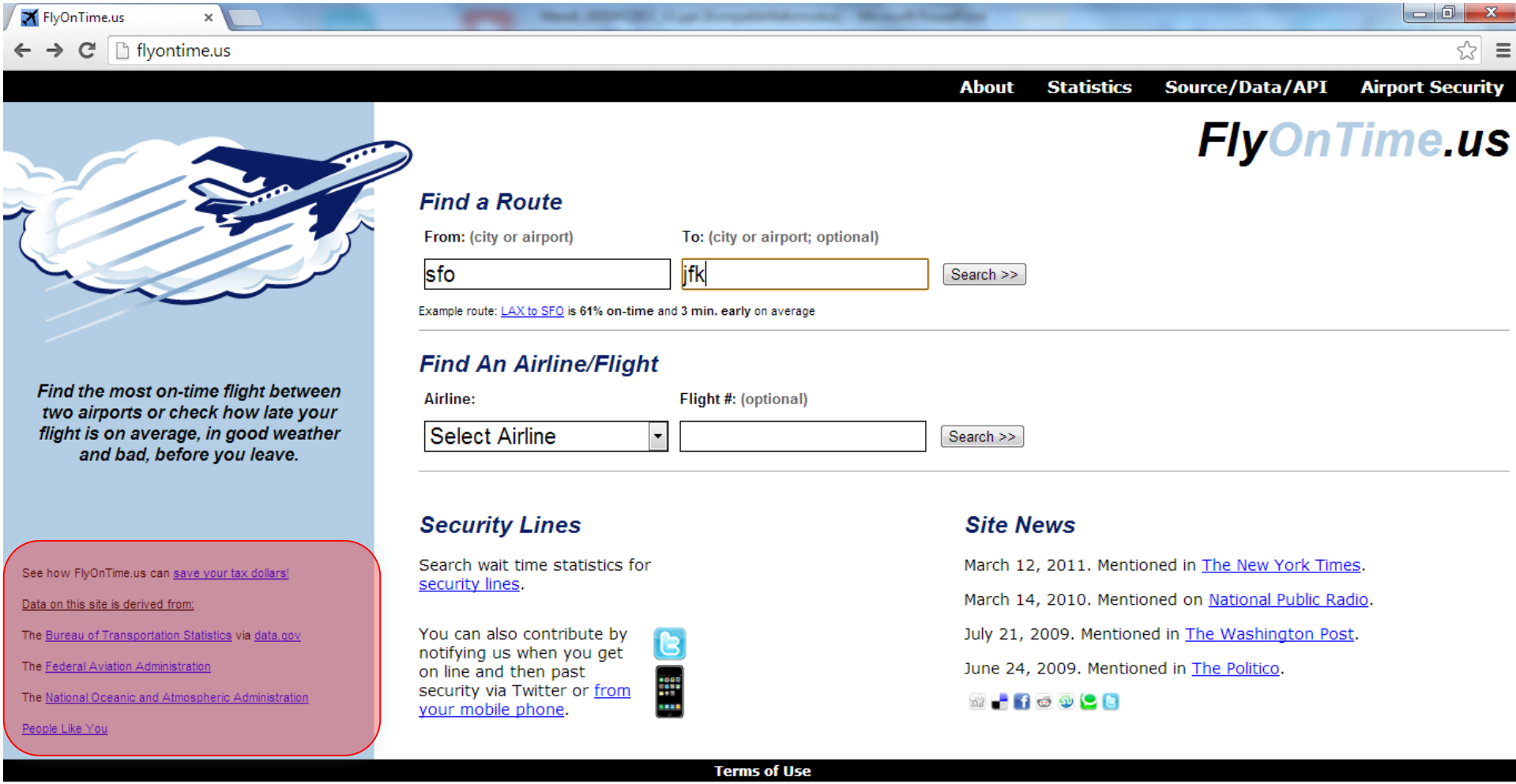
IDRISI or TerrSet

www.clarklabs.org/

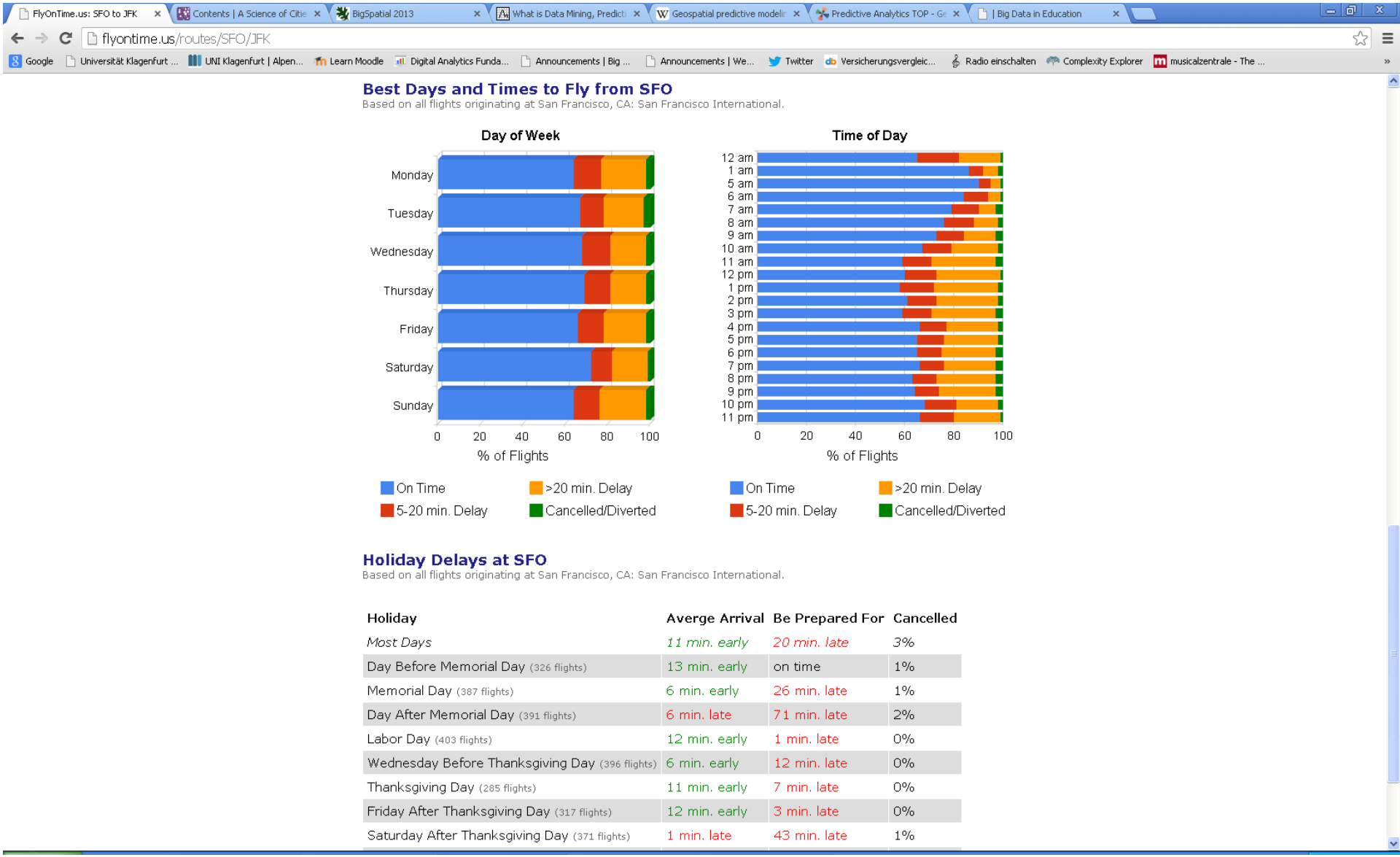



IDRISI 17.0 The Selva Edition
File Display GIS Analysis Modeling Image Processing Reformat Data Entry Window List Help
Idrisi Explorer
Projects Files Filters
C:\Users\standard\Desktop\T2001_
C:\Users\standard\Desktop\T2001_
t2001_2010_profile_1.tsf
C:\Users\standard\Desktop\T2001_
t2001_2010_Intercept.rst
t2001_2010_R2.rst
t2001_2010_Slope.rst
C:\Users\standard\Desktop\T2001_
Browser...
.rst,.rgf,*.tsf,*.vct,*.vlx,*.vgf
Metadata
Name 001_2010_R2
File format IDRISI Raster
File title Linearity (R2) o
Data type Real
File type Binary
Columns 228
Rows 90
Ref. system Plane
Ref. units Meters
Unit dist. 1.0000000
Min. X 356500.000000
Max. X 584500.000000
Earth Trends Modeler
Explore Analysis Preprocess
Project
Explore Space / Time Dynamics
Series : t2001_2010
Create/recreate visualization
Reset
Time Aug-10/2006
Cube
Explore PCA / EOT / Fourier PCA / CCA / Wavelets
Explore Temporal Profiles
Explore Series Relationships
Explore Trends
t2001_2010_slope
Linear trend (OLS) of t2001_2010
-0.0006
-0.0005
-0.0005
-0.0004
-0.0003
-0.0002
-0.0001
0.0000
0.0001
0.0002
0.0003
0.0004
t2001_2010_r2
Linearity (R2) of t2001_2010
0.000
0.001
0.001
0.002
0.003
t2001_2010_intercept
Linear trend (OLS) of t2001_2010
t2001_2010_slope
Linear trend (OLS) of t2001_2010
-0.0006
-0.0005
-0.0005
-0.0004
-0.0003
-0.0002
-0.0001
0.0000
0.0001
y : 151174.157303

FlyOnTime.us



http://flyontime.us/routes/SFO/JFK

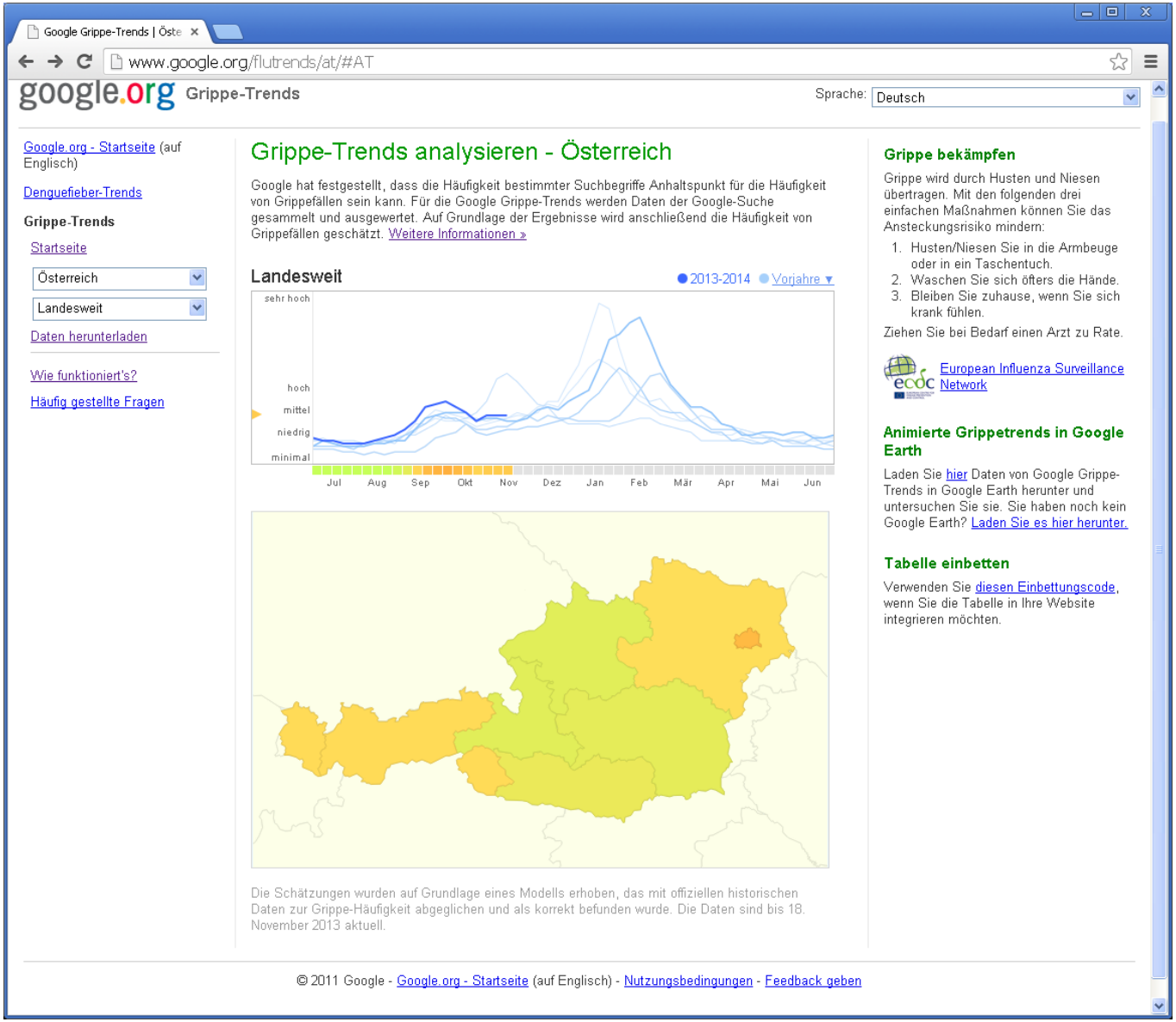


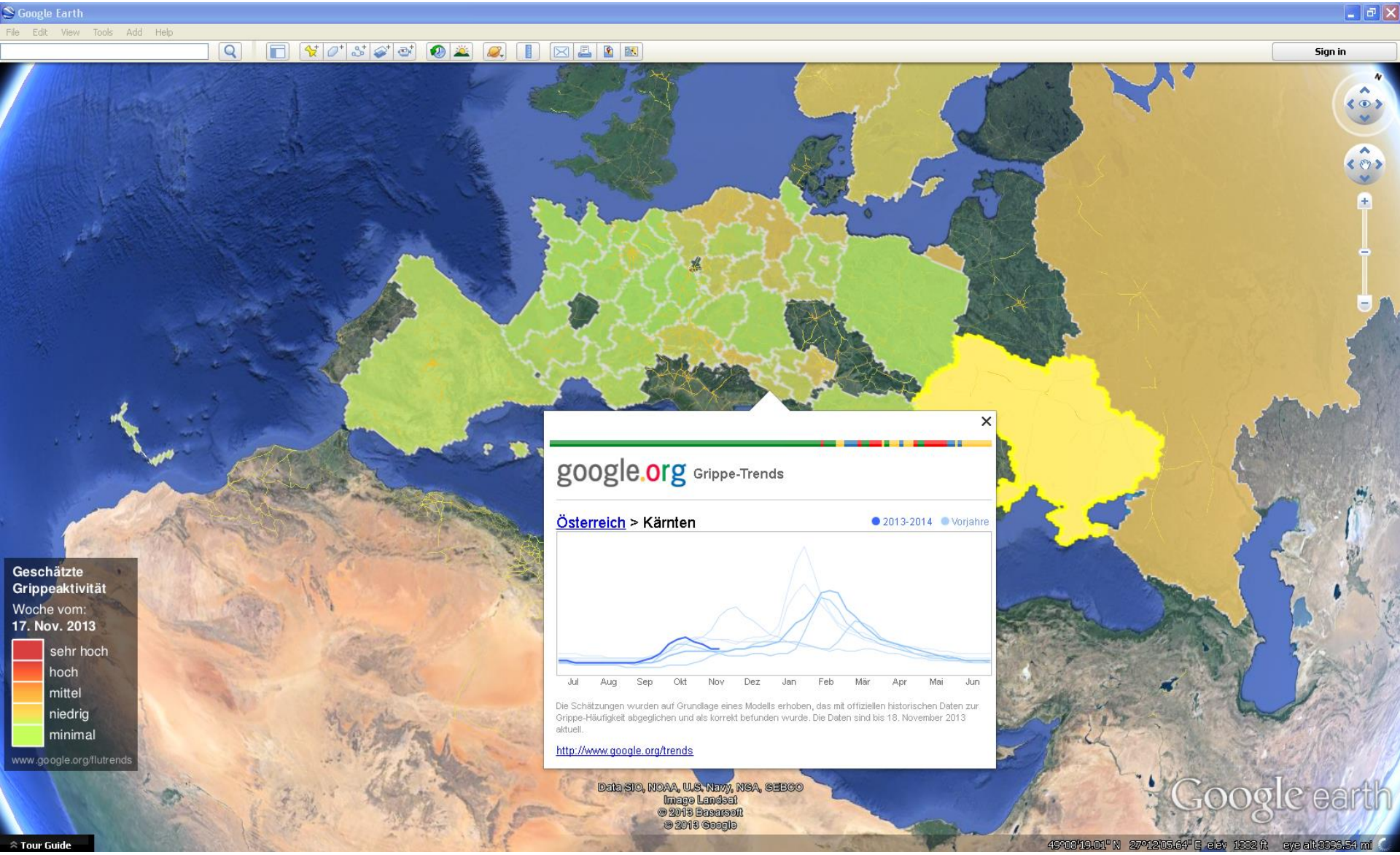
Holiday Delays at SFO

Based on all flights originating at San Francisco, CA: San Francisco International.

Holiday	Averge Arrival	Be Prepared For	Cancelled
Most Days	*11 min. early*	*20 min. late*	*3%*
Day Before Memorial Day (326 flights)	13 min. early	on time	1%
Memorial Day (387 flights)	6 min. early	26 min. late	1%
Day After Memorial Day (391 flights)	6 min. late	71 min. late	2%
Labor Day (403 flights)	12 min. early	1 min. late	0%
Wednesday Before Thanksgiving Day (396 flights)	6 min. early	12 min. late	0%
Thanksgiving Day (285 flights)	11 min. early	7 min. late	0%
Friday After Thanksgiving Day (317 flights)	12 min. early	3 min. late	0%
Saturday After Thanksgiving Day (371 flights)	1 min. late	43 min. late	1%

http://www.google.org/flutrends/at/#AT




Google Earth
File Edit View Tools Add Help
Sign in
Geschätzte Grippeaktivität
Woche vom:
17. Nov. 2013
sehr hoch
hoch
mittel
niedrig
minimal
www.google.org/flutrends
google.org Grippe-Trends
Österreich > Kärnten
2013-2014
Vorjahre
Jul Aug Sep Okt Nov Dez Jan Feb Mär Apr Mai Jun
Die Schätzungen wurden auf Grundlage eines Modells erhoben, das mit offiziellen historischen Daten zur Grippe-Häufigkeit abgeglichen und als korrekt befunden wurde. Die Daten sind bis 18. November 2013 aktuell.
http://www.google.org/trends
Data SIO, NOAA, U.S. Navy, NGA, GEBCO
Image Landsat
© 2013 Basarsoft
© 2013 Google
Google earth
Tour Guide
49°08'19.01" N 27°12'05.64" E elev 1382 ft eye alt 3896.54 mi

What is to do?

- Use New Science Paradigms
  - 4th scientific paradigm of an e-science (data exploration approach)
  - http://research.microsoft.com/en-us/um/people/gray/talks/NRC-CSTB_eScience.ppt
- Design Suitable Concepts and Frameworks
  - http://www.esri.com/news/releases/12-3qtr/carl-steinitz-explains-geodesign-process-in-new-esri-press-book.html
- Use Analytics Tools and Methods
  - rapidminer.com/
- Integrate Visualization
  - CommonGIS
  - http://www.iais.fraunhofer.de/1871.html
- Apply More Spatio-Temporal Modeling
  - IDRISI or TerrSet
  - www.clarklabs.org/

Additional Material

Slides:

- Shashi Shekhar (2014): Spatial Data Science: Challenges & Opportunities http://www.nist.gov/itl/iad/upload/Shashi-Shekhar-14-3-CCC_NIST_WEB.pdf

> Maggi Kelly (2014): Spatial Data Science: the 21st Century Mapping Toolkit

https://www.youtube.com/watch?v=qWQp11CE4S0

To a Data-Driven Geography or Spatial Data Scence?

- In a **Spatial Data Science** we have **to consider**
  - Errors / Calibration / Privacy protection / Prediction without explanation?
- **Granularity** of he data (**resolution**) and **Accuracy**
- **Interoperability** of the programs, data, models and methods is to be established.
- The **Big Data** have to be used **multiple** and have to be **combined** and **linked**.
- Many **Best Practice Examples** have to be done and the **models** have to be **distributed** and **offered** via internet **interoperably**.
- Use a **formal language** or **ontology** to describe the new approach of **Spatial Data Science**.

See also: Harvey J. Miller & Michael F. Goodchild (2014): Data-driven geography. In: GeoJournal
http://link.springer.com/article/10.1007/s10708-014-9602-6

Thank you very much for your attention!

Questions and Suggestions please to:

peter.mandl@aau.at

Informations at:

http://giscience.aau.at/